



**VERİ MADENCİLİĞİNDE BİRLİKTELİK
KURALI VE HEPATİT HASTALIĞI ÜZERİNE
BİR UYGULAMA
Yüksek Lisans Tezi**

Birsen RENCÜZOĞULLARI

Eskişehir, 2019

**VERİ MADENCİLİĞİNDE BİRLİKTELİK KURALI VE HEPATİT
HASTALIĞI ÜZERİNE BİR UYGULAMA**

Birsen RENCÜZOĞULLARI

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Levent TERLEMEZ

Eskişehir

Eskişehir Teknik Üniversitesi

Lisansüstü Eğitim Enstitüsü

Ekim 2019

JÜRİ VE ENSTİTÜ ONAYI

Birsen RENCÜZOĞULLARI'nın “**VERİ MADENCİLİĞİNDE BİRLİKTELİK KURALI VE HEPATİT HASTALIĞI ÜZERİNE BİR UYGULAMA**” başlıklı tezi 11/09/2019 tarihinde aşağıdaki jüri tarafından değerlendirilerek “Eskişehir Teknik Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği”nin ilgili maddeleri uyarınca, İstatistik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

<u>Jüri Üyeleri</u>	<u>Unvanı Adı Soyadı</u>	<u>İmza</u>
Üye (Tez Danışmanı)	Dr. Öğr. Üyesi Levent TERLEMEZ	
Üye	Doç. Dr. Cengiz BAL	
Üye	Dr. Öğr. Üyesi Alper BEKKİ	

Prof. Dr. Murat TANIŞLI
Lisansüstü Eğitim Enstitüsü Müdürü

ÖZET

VERİ MADENCİLİĞİNDE BİRLİKTELİK KURALI VE HEPATİT HASTALIĞI ÜZERİNE BİR UYGULAMA

Birsen RENCÜZOĞULLARI

İstatistik Anabilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Ekim 2019

Danışman: Dr. Öğr. Üyesi Levent TERLEMEZ

Bu çalışmanın amacı Apriori algoritması yöntemi ile hastalık ve belirtilerinin birliktelik ilişkilerini test ederek hepatit rahatsızlığının tanısının konulmasında hekimler için karar destek sistemi geliştirmektir. Hepatit hastası 155 kişiye ait toplam 17 semptom analiz edilmiştir.

Veri madeninden elde edilen hepatit veri setine Apriori algoritması uygulanması sonucunda hepatit hastalarının gösterdiği bazı belirtiler arasında birliktelik ilişkileri kurulmuştur. Algoritma sonucu elde edilen birliktelik kurallarının güven düzeylerinin %95 ile %100 aralığında oldukları gözlemlenmiştir. Elde edilen verilerin analiz edilmesinden sonra belirtilerin birlikte görülme olasılıkları hesaplanmıştır. Bu kurallarla kişiden kişiye hastalık seyri ve belirtileri farklı olabilen hepatit hastalığının teşhisinde istatistiksel bakımdan doktorlara karar verme desteği sunması hedeflenmektedir.

Anahtar sözcükler: Veri madenciliği, Apriori algoritması, Birliktelik kuralları, Hepatit

ABSTRACT

ASSOCIATION RULE IN DATA MINING AND AN APPLICATION ON HEPATITIS

Birsen RENCÜZOĞULLARI

Department of Statistics

Eskişehir Technical University, Intitute of Graduate Programs, October 2019

Supervisor: Ass. Prof. Dr. Levent TERLEMEZ

This study aimed to develop a decision support system for physicians in diagnosing hepatitis by testing the association relationships of the disease and symptoms with Apriori algorithm. Totally 17 symptoms of 155 hepatitis patients were analyzed.

Application of Apriori algorithm to hepatitis data set, association relations have been established among some of the symptoms of hepatitis patients. The confidence levels of the association rules was observed between 95% and 100%. After analyzing, the probability of coexisting symptoms were calculated. These rules may provide statistically decision-making support to physicians in the diagnosis of hepatitis, which may have different prognosis and symptoms.

Key words: Data mining, Apriori algorithm, Association rules, Hepatitis

11/09/2019

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Eskişehir Teknik Üniversitesi tarafından kullanılan “bilimsel intihal tespit programı” ile tarandığını ve hiçbir şekilde “intihal içermediğini” beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Birsen RENCÜZOĞULLARI

TEŞEKKÜR

Tez konumun belirlenmesi, yürütülmesi ve bütün çalışmalarım boyunca bana rehber olan, ilgi ve yardımlarını esirgemeyen, öğrencisi olmaktan onur ve mutluluk duyduğum Sayın Hocam **Dr. Öğr. Üyesi Levent TERLEMEZ**'e sonsuz teşekkürlerimi sunarım.

Klinisyen olarak katkılarından dolayı Gastroenteroloji Uzmanı Dr. Öğr. Üyesi Sezgin BARUTÇU'ya teşekkür ederim. Apriori algoritmasının dosya dönüştürme işlemlerinde teknik destekleri ve yardımları için Ceyhun Enki Aksan'a teşekkür ederim.

Birsen RENCÜZOĞULLARI



İÇİNDEKİLER

	<u>Sayfa</u>
BAŞLIK SAYFASI.....	i
JÜRİ VE ENSTİTÜ ONAY SAYFASI.....	ii
ÖZET.....	iii
ABSTRACT.....	iv
ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER.....	vii
ÇİZELGELER DİZİNİ.....	ix
ŞEKİLLER DİZİNİ.....	x
SİMGELER VE KISALTMALAR DİZİNİ.....	xi
1. GİRİŞ.....	1
2. VERİ MADENCİLİĞİ.....	4
2.1. Veri madenciliğinin Temel Aşamaları.....	7
2.1.1 Amacın tanımlanması.....	7
2.1.2. Verilerin derlenmesi ve ön işlemlerin yapılması.....	8
2.1.3. Modelin kurulması.....	13
2.1.4. Modelin kullanılması ve yorumlanması.....	15
2.1.5. Modelin izlenmesi.....	16
2.2. Veri Madenciliğinde Kullanılan Teoriler.....	16
2.2.1. Bulanık mantık teorisi (Fuzzy theory).....	16
2.2.2. Kaba küme teorisi (Rough set theory).....	16
2.2.3. Bulut teorisi (Cloud theory).....	17
2.2.4. Dempster–Shafer teorisi (Dempster–Shafer theory).....	17
2.2.5. Yapay sinir ağları (Artificial neural network).....	17
2.2.6. Genetik algoritma (Genetic algorithm).....	18
2.2.7. İndükleyici öğrenme teorisi (Inductive learning theory).....	18
2.3. Veri Madenciliği Alanında En Çok Kullanılan Yöntemler.....	18
2.3.1. Bayesian ağı (Bayesian network).....	18
2.3.2. Karar ağacı (Decision tree).....	19
2.3.3. Örüntü tanıma (Pattern recognition).....	19

2.3.4. Yüksek performanslı hesaplama (High performance computing).....	20
2.3.5. İstatistiksel analiz (Statistical analysis).....	20
2.4. Veri Madenciliğinin Uygulama Alanları.....	21
2.4.1. Sağlık ve genetik alanında veri madenciliği.....	21
2.4.2. Gıda sektöründe veri madenciliği.....	22
2.4.3. Kimya endüstrisinde veri madenciliği.....	22
2.4.4. İletişim endüstrisinde veri madenciliği.....	23
2.4.5. Pazarlama, bankacılık ve finansal alanlarda veri madenciliği.....	23
2.4.6. Astroloji biliminde veri madenciliği.....	23
2.4.7. Sosyal bilimlerde veri madenciliği.....	23
3. BİRLİKTELİK ANALİZİ VE APRİORİ ALGORİTMASI	24
3.1. Birliktelik Kuralı.....	24
3.2. Apriori Algoritması.....	27
3.3. Birliktelik Kuralı ve Apriori Algoritması İle İlgili Yapılan Çalışmalar..	30
3.3.1. Sağlık dışı alanlarda birliktelik kuralı ve Apriori algoritması ile ilgili yapılan çalışmalar.....	30
3.3.2. Sağlık alanında birliktelik kuralı ve Apriori algoritması ile ilgili yapılan çalışmalar.....	32
3.3.3. Hepatit ve hepatit hastalıklarında veri madenciliği çalışmaları...	34
3.4. Weka Explorer Paket Programı.....	37
3.5. Apriori Algoritmasının Hepatit Verilerine Uygulanması.....	38
4. BULGULAR VE YORUM.....	46
5. SONUÇLAR VE TARTIŞMA.....	55
KAYNAKÇA.....	59
EKLER	
ÖZGEÇMİŞ	

ÇİZELGELER DİZİNİ

	<u>Sayfa</u>
Çizelge 2.1. R-mice programının değerlendirme teknikleri.....	12
Çizelge 2.2. Denetimli ve denetimsiz öğrenmenin karşılaştırılması.....	13
Çizelge 3.1. Sık görülen hepatit tipleri.....	35
Çizelge 3.2. Verilerin rassal dağılımının saptanması amacıyla uygulanan Little's MCAR test sonucu.....	40
Çizelge 3.3. Kan parametrelerinin normal değer aralıkları.....	41
Çizelge 3.4. Eksik verilerin özelliklere göre frekans değerleri.....	42
Çizelge 4.1. Ortalama değerler ve standart hata.....	47
Çizelge 4.2. Nitel değişkenlerin karşılaştırmalı frekans değerleri.....	49
Çizelge 4.3. Apriori algoritması sonucu elde edilen kurallar.....	52

ŞEKİLLER DİZİNİ

	<u>Sayfa</u>
Şekil 2.1. Veri madenciliğinde veri ambarı kullanımı.....	4
Şekil 2.2. Veri madenciliğinin birlikte çalıştığı disiplinler.....	5
Şekil 2.3. Veri tabanından veri madenciliğine veri analiz işlemleri.....	7
Şekil 2.4. Veri ön işleme aşamaları.....	8
Şekil 2.5. Regresyon, sınıflandırma ve kümeleme.....	14
Şekil 2.6. Denetimli ve denetimsiz öğrenme arasındaki fark.....	15
Şekil 3.1. Dört nesne içeren bir kümenin farklı alt nesne setlerinin oluşumu.....	25
Şekil 3.2. Apriori algoritması kodu.....	28
Şekil 3.3. Apriori gen fonksiyonu.....	29
Şekil 3.4. Apriori algoritması uygulama işlemleri.....	43
Şekil 4.1. Ortalama değere sahip verilerin frekanslarının karşılaştırılması.....	48
Şekil 4.2. Nitel değere sahip verilerin frekanslarının karşılaştırılması.....	50
Şekil 4.3. Apriori algoritması, minMetric ve numRules ayarları.....	51

SİMGELER VE KISALTMALAR DİZİNİ

ALP	: Alkalen fosfataz (Safra kanallarıyla ilişkili enzim)
ALT	: Alanin aminotransferaz (Karaciğer enzimi, hepatiti saptamaya yarayan test)
ANN	: Artificial neural network
ARFF	: Attribute relationship file format
AST	: Aspartat aminotransferaz (Karaciğer ve kalp kaslarında bulunan enzim)
EEG	: Elektroensefalograf
EKG	: Elektrokardiyograf
ETS	: Eritroblast transformasyon sistemi
FDA	: Food and Drug Administration
Hbg	: Hemoglobin
HPC	: Yüksek performanslı bilgi işlem
KMP	: Kapsamlı Metabolik Panel (Hepatit saptamaya yarayan rutin test)
KOAH	: Kronik obstrüktif (tıkayıcı) akciğer hastalığı
MCHC	: Ortalama korpüsküler hemoglobin konsantrasyonu
OLAP	: On line analytical processing, çevrimiçi analitik işleme
OSCC	: Oral skuamöz hücreli karsinoma
PT	: Protrombin zamanı
PubMed	: Biyomedikal veri tabanı, https://www.ncbi.nlm.nih.gov/pubmed/
Rbc	: Red blood cells
SGOT	: Serum glutamik oxaloasetik transaminaz (hasarlı karaciğer hücrelerinden salınan enzim)
TROK	: Tamamıyla rassal olarak kayıp veri (Missing Completely At Random, MCAR)
YSA	: Yapay sinir ağları

1. GİRİŞ

Verilerin toplanması, düzenlenmesi, özetlenmesi, sunulması, analiz edilmesi, sonuçların yorumlanması ve elde edilen sonuçlara göre karar verilmesi işlemleriyle ilgilenen İstatistik Bilimi küreselleşen dünyada ve artan veri yığınlarında gittikçe daha fazla önem kazanmaya devam etmektedir.

Dünyada ilk defa 17.yüzyılda istatistik çalışmalar yapıldığı fakat ilk istatistiki çalışmaların dar amaçlı ve sadece bilgi toplamaya yönelik yapılmış olduğu bildirilmiştir (Akdeniz ve Yıldırım, 2011). Halbuki, Akdeniz (2015) yayınladığı bir çalışmada 50 yüzyıl önce eski Mısır'da yaşamış bir kralın mezarındaki şekillerin incelenmesinde savaş esirlerinin ve ele geçirilen hayvanların sayılarının ilk istatistik veriler olarak düşünülebileceğini bildirmiştir. 20. Yüzyıldan sonra ise olasılık teorisi ve bu alanda gerçekleştirilen çalışmalar sonucu kaydedilen gelişmeler ile birlikte istatistik bilimi geliştirilmiştir (Akdeniz ve Yıldırım, 2011). Ülkemizde Cumhuriyet ile beraber istatistikçiler etkin olmaya başlamış ve özellikle son 30 yıla damgasını vuran çalışmalar istatistiği evrensel bir konuşma dili konumuna getirmiştir (Akdeniz, 2015).

Son yıllarda veri kümelerindeki artış inanılmaz boyutlara ulaşmıştır (Coşkun ve Baykal, 2011; Zhang vd., 2016). Bilgi teknolojilerinin gelişmesi de istatistiksel analizlerin daha kolay uygulanabilmesinin önünü açmıştır. Büyük veri hareketi son birkaç yılda ivme kazandığından, sağlık hizmeti ile üretilen büyük veriyi analiz etmek için veri madenciliği teknikleri ve yöntemlerinin kullanımıyla ilgili özel bir ilginin ortaya çıkmış olduğu görülmektedir (Househ ve Aldosari, 2017).

Veri madenciliği amaca göre farklı tanımlansa da genel anlamda büyük veri yığınları içerisinde amacımızla ilgili tahmin yapmayı sağlayacak bağlantı ve kuralların bilgisayar programları kullanarak aranması olarak tanımlanır (Bellinger vd., 2017; Zhang vd., 2016; Farboudi, 2009). Daha geniş bir tanımlama Bellinger vd. (2017) tarafından yapılmıştır. Bellinger vd. (2017) yayınladıkları raporlarında veri madenciliğini, genellikle büyük veri kümelerini analiz etmek, kalıpları keşfetmek, işlem yapılabilir bilgiyi çıkarmak ve gelecekteki veya bilinmeyen olayların sonuçlarını tahmin etmek için uygulanan hesaplama işlemi olarak tanımlamışlardır. Aynı araştırmacılar bu süreçte kullanılan yöntemlerin bir kombinasyon halinde olduğunu ve yapay zeka, istatistik, matematik, makine öğrenimi ve veri tabanı sistemlerinden oluştuğunu belirtmişlerdir.

Veri madenciliği uygulamaları Fen, Sosyal ve Sağlık Bilimleri de dâhil hayatın her aşamasına uygulanabilmektedir. Köroğlu (2016) veri madenciliğinin tıp ve sağlık

alanında önemli katkılar sunarak birçok çözüm yolu gösterebileceğini bildirmiştir. Househ ve Aldosari (2017), 1990'ların ortasından itibaren, sağlık verilerindeki kalıpları ve ilişkileri araştırmak ve bulmak için veri madenciliği yöntemlerinin başarıyla kullanıldığını saptamışlardır.

Klinik karar vermede veri madenciliğinin amacı, klinik ortamdaki özellik ve kalıp arasındaki ilişkiyi tanımak ve sonucu tahmin etmek, hekimleri karar verirken istatistiksel bakımdan desteklemek olarak açıklanmıştır (Chen ve Fawcett, 2016). Birçok hastalık oluşum sırasında birbirine benzerlik gösterdiğinden hastalık teşhisini, en önemli ve en kritik aşamalardan birincisi içine koymaktadır. Çünkü yapılacak yanlış hastalık teşhisi, yanlış tedaviler sonucunda, hastayı daha zor bir duruma düşürebilmektedir. Ayrıca hastalıkların seyri teşhisi de etkileyerek tanı konulması sürecini de etkilemektedir. Bu nedenle, veri madenciliği karar vermedeki hataları en aza indirmenin bir yolu olarak geliştirilmiştir ve dolayısıyla kullanılan modeller klinik karar verme stratejilerine benzeyecek şekilde tasarlanmıştır.

Sağlık ve tıp alanındaki veri madenciliği uygulamalarında başta kanser olmak üzere birçok hastalık ile ilgili verilere ulaşmak mümkündür (http-1). Bu veri tabanında tüm sağlık alanı ile ilgili verilerle birlikte hepatit enfeksiyonlarına yakalanan insanlara ait veriler de mevcuttur.

Hepatit yaş ve cinsiyet ayırt etmeksizin tüm insanları etkileyebilen, kansere ve ölüme kadar götürebilen bir karaciğer hastalığıdır. Uzun yıllar semptom vermeden insanlarda varlığını sürdürürken bazı insanlarda farklı tip ve şiddette semptomlar vererek hayatı zorlaştırabilmektedir (McMahon vd., 2005; Aydın vd., 2016). En fazla virütik hepatitlere rastlansa da (Miguet vd., 1990; Briese vd., 2009; Gallegos-Orozco ve Rakela-Brödner, 2010; Weiss ve Leibowitz, 2011; Okano ve Gross, 2012), bakteriyel hepatitlere (Selmi ve Gershwin, 2004; Heyman vd., 2010) ve hatta parazitik veya toksinler (alkol, ilaçlar, diğer endüstriyel kimyasallar) gibi enfeksiyöz olmayan hepatitlere de rastlanmaktadır (Seitz, 1995; Romero ve Razonable, 2011).

Veri madenciliğinde birliktelik analizlerinin yapılıp birliktelik kurallarının ortaya çıkartılması konusunda en çok bilinen ve kullanılan algoritmalarından birisi Apriori algoritmasıdır (Duru, 2005; Köroğlu, 2016). Bu algoritma geniş nesne kümelerinin ortaya çıkartılması işlemleri için güvenle kullanılmaktadır (Maria Luna vd., 2018; Somek ve Hercigonja-Szekeres, 2017). Chen vd. (2011) yaptıkları çalışmalarda hepatit hastalığının teşhisi için bir sınıflandırma algoritması olan “LFDA-SVM” adlı yeni bir karma yöntemi

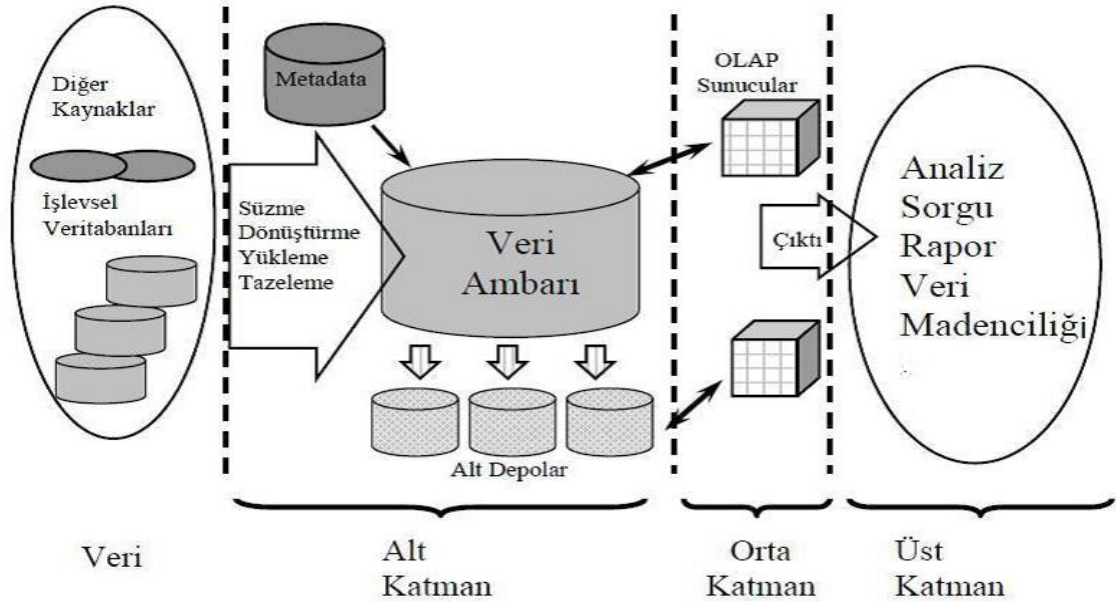
denedikleri ve “LFDA-SVM” yönteminin iyi bir sınıflandırma yöntemi olduğunu bildirmişlerdir. Araştırmacılar ayrıca bu yöntemin hepatit teşhisinde iyi bir performans gösterebileceğini önermişlerdir. Mahmoodi vd. (2016) mide kanserine sahip kişilerde veri madenciliği algoritması yöntemlerini kullanarak kuralların daha sezgisel, çekici ve anlaşılır hale getirilebildiği, israf ve işe yaramaz kuralların ortadan kaldırılabildiğini belirtmişler ve Apriori algoritmasının bu alanda önemli avantajlar sağladığını saptamışlardır. Hatta Parraga-Alava vd. (2018) biyolojik verilerin birliktelik kurallarına göre ilişkilendirilmesinde kullanılan farklı teknikleri karşılaştırmak için yaptıkları çalışmalarda Apriori algoritmasının birlikte ifade, biyolojik tutarlılık, kompaktlık ve ayrıştırma açısından mevcut olan diğer mikrodizi kümeleme tekniklerinden daha yüksek kalitede çözümler bulmasını sağladığını saptamışlardır.

Sağlık bilimlerinde teknolojinin gelişmesi birçok hastalığın belirtilerinin detaylı bir şekilde saptanmasını mümkün kılmıştır. Bu durum birçok hastalığın belirtilerinin benzer olmayabildiğini ortaya çıkarmıştır. Örneğin kanser olgularında karın boşluğunda asitlenme, kan düzeyinin düşmesi, halsizlik gibi belirtiler ayrıca hepatit hastalarında da ortaya çıkmaktadır (Shao vd., 2017). Hepatit hastalarında görülen bilirubin yüksekliği ayrıca yeni doğan sarılığı olarak da bilinir (Yu vd., 2019). Bir çeşit kalıtsal kan hastalığı olan orak hücre ve akdeniz hastalığında da kırmızı kan hücrelerinin yıkımı sonucu bilirubin düzeyi yükselmektedir (Yahouédéhou vd., 2019). Bütün bu hastalarda halsizlik de görülmektedir.

PubMed veri tabanında “data mining and hepatit” ve “apriori and hepatit” anahtar kelimeleri ile yapılan aramada kompleks hastalıklardan bir tanesi olan ve farklı fiziksel belirtiler veren hepatit enfeksiyonu ile ilgili Apriori algoritması uygulamasına dair herhangi bir çalışmaya rastlanmamıştır. Dolayısıyla bu çalışmanın amacı Apriori algoritması yöntemi ile hastalık ve belirtilerinin birliktelik ilişkilerini test ederek hepatit rahatsızlığının tanısının konulmasında hekimler için karar destek sistemi geliştirmektir.

2. VERİ MADENCİLİĞİ

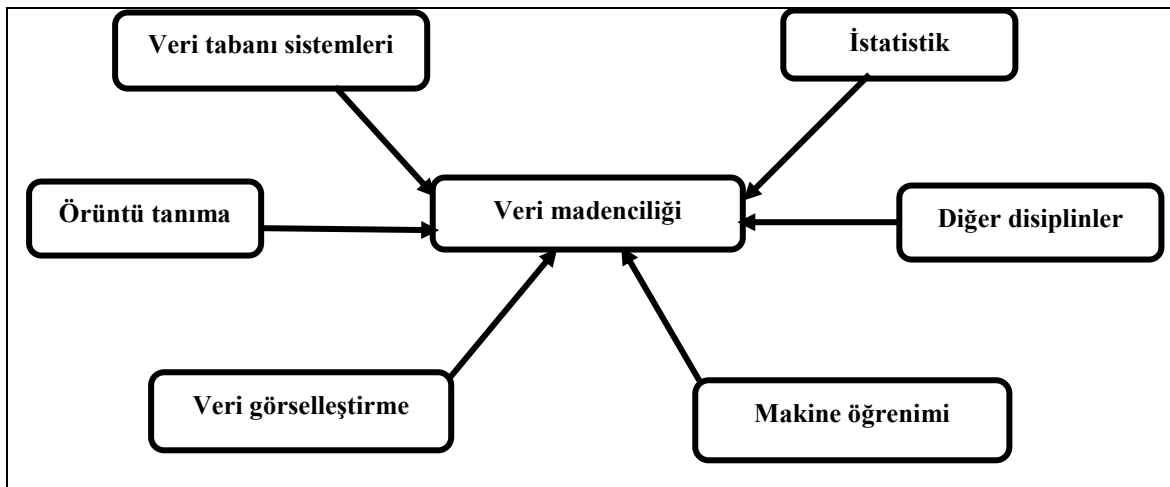
Geçmiş faaliyetlerin verilerini ele alarak tahmine yönelik karar verme amacıyla uygulanan veri madenciliğine yönelik ilk çalışmalar 1970’li yıllarda böbrek hastalığını teşhis etmek amacıyla olasılık ölçütleriyle birlikte karar teorisi kullanılması ile başlamıştır (Koyuncugil ve Özgülbaş, 2009; Poyraz, 2012; Finlay, 2014; Chen ve Fawcett, 2016). Finlay (2014) veri madenciliğini “bireysel tercihler hakkında fikir veren veriler arasındaki ilişkileri tanımlamak” şeklinde izah etmiştir. Veri madenciliği veri ambarlarındaki veri kümelerini kullanır (Şekil 2.1). Veri ambarları, var olan tüm verilerin analiz için özel olarak modellenen ve tarih akışı içinde biriktirilen veri depolama sistematiğidir. Bu tanımlama ile birlikte daha geniş anlamı ile veri madenciliği, veri ambarlarındaki çeşitli verilerin amaca uygun olarak ortaya çıkarılması, kullanılması ve analiz edilmesi süreci olarak tanımlanabilir. Yani, aslında veri madenciliği bir istatistiksel süreçtir. İlerleyen yıllarda bilgi teknolojisindeki ilerlemelerle beraber veri madenciliği çalışmaları her alana uygulanmaya başlanmıştır. Buradaki amaç, verilerin analizinde yeni yaklaşımlar getirecek ve dolayısıyla istatistiksel yöntemlerin yetersizliklerini giderecek bilgisayar algoritmalarını kullanarak istenilen analizleri hızlı, güvenli ve sağlıklı bir şekilde gerçekleştirmektir (Lezki, 2016).



Şekil 2.1. Veri madenciliğinde veri ambarı kullanımı (Farboudi, 2009).

Farklı kaynaklardan sağlanan veriler çeşitli işlemlerden geçirilerek sisteme yüklenirler ve dolayısıyla veri ambarları oluşturulur (Şekil 2.1). Veri ambarları 3 katmandan oluşurlar: Alt katman, orta katman ve üst katman. Alt katman, veri tabanının sunucusu konumundadır. Veri tabanı günlük bilgilerin depolandığı ve işlevsel olan veri havuzudur. Bundan dolayı güncel bilgileri içerir. Veri ambarı ise işlevsel veri tabanındaki bilgilerin güncelliklerini yitirmiş veri havuzu olarak tanımlanır. Yani aslında veri tabanı günlük işlemler için kullanılırken veri ambarı ise uzun süreli veri analizi ve geleceğe yönelik tahminler elde etmek için kullanılmaktadır. Bilgi işlem sistemleri aracılığıyla alt katmandan sağlanan bilgiler orta katmana aktarılarak çok boyutlu analiz yöntemleri uygulanır ve sonuçlar üst katmana iletilir. Orta katmandaki OLAP (On line analytical processing, Çevrimiçi analitik işleme) veriyi çok boyutlu bir şekilde analiz eder ve yöneticilere stratejik karar vermelerinde destek olacak bilgiler sunar (Lezki, 2016). Üst katmanda veriler veri madenciliği araçları da dâhil analiz ve sorgulama araçlarını da içerir.

Veri madenciliği aslında bilgisayar sistemlerinin verilerin analizine imkân verecek ölçülerde gelişmesi ve dolayısıyla problemlerin çözümü amacıyla kullanılmaya başlamasıyla birlikte ortaya çıkmıştır. Burada amaçlanan, aslında geleneksel istatistik yöntemler yerine bunların eksikliklerini giderecek yeni yaklaşımları bilgisayar algoritmaları yardımıyla uygulamak ve istenilen analizleri hızlı ve güvenli bir şekilde yapmaktır (Lezki, 2016). Dolayısıyla veri madenciliği birçok disiplin ile çalışmaktadır (Şekil 2.2).



Şekil 2.2. Veri madenciliğinin birlikte çalıştığı disiplinler

İstatistik, verilerin analizi ve değerlendirilmesi ile uğraşan bir disiplindir. Bilgisayar sistemlerinin gelişmesi istatistik alanını da etkilemektedir.

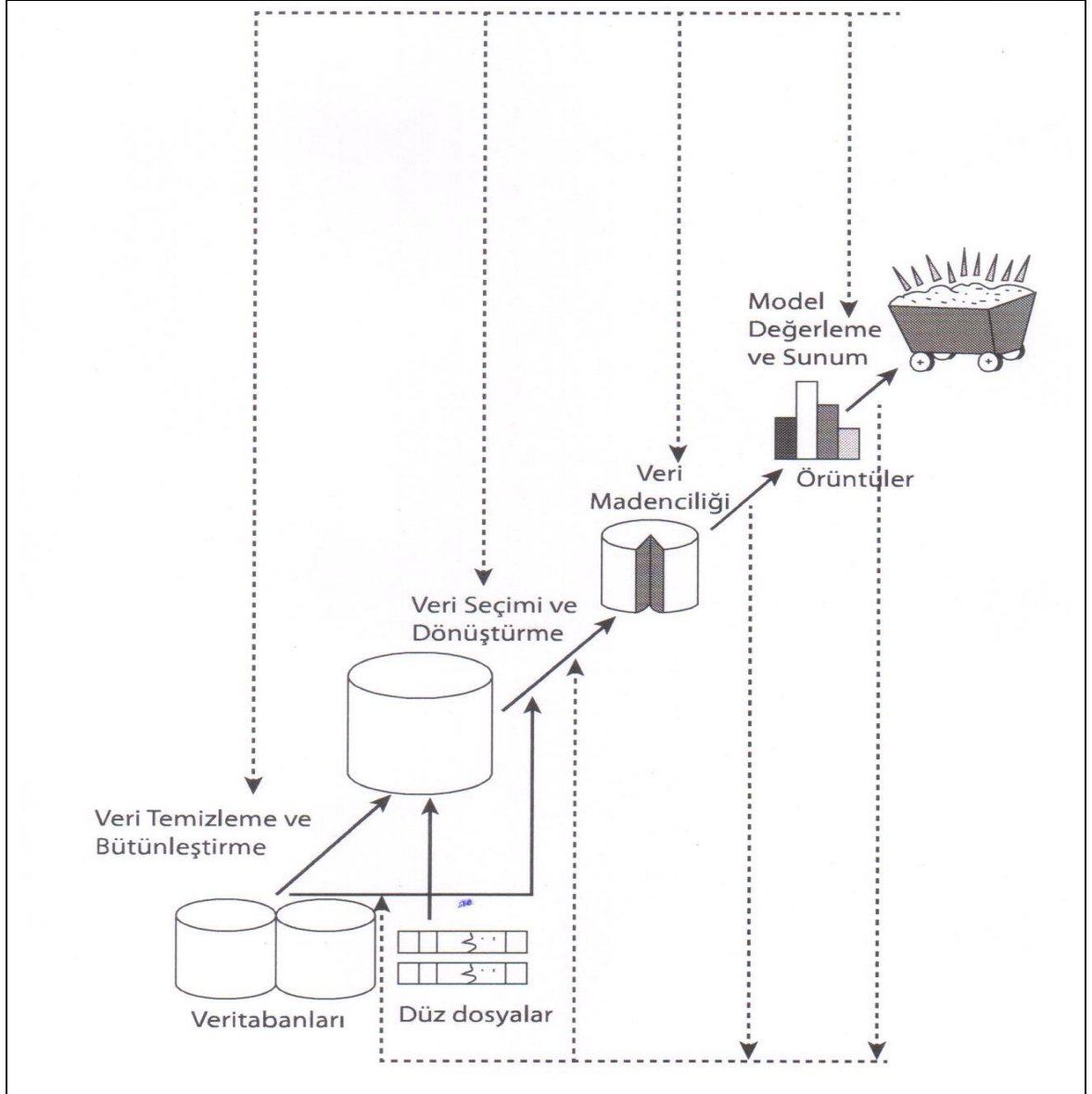
Makine öğrenimi bilgisayarların bazı işlemlerden yola çıkarak yeni işlemler üretmesi şeklinde tanımlanabilir. İnsanların öğrenebildiklerini bilgisayarlara da öğretme ve dolayısıyla bilgisayarların da insanlar gibi öğrenebilme düşüncesini temel alan bir disiplindir.

Görselleştirme ise verilerin görsel olarak sunulması temeline dayanır. Sonuçlar ya grafik ya da tablolar şeklinde sunulurken verilerin daha kolay anlaşılmasını amaçlayan bir disiplindir.

Veri madenciliğinin faydalandığı diğer bir disiplin ise veri tabanlarıdır. Veri tabanı günlük bilgilerin depolandığı ve işlevsel olan güncel bir veri havuzudur. Burada toplanan veriler, gelişigüzel veriler olmayıp birbirleriyle ilişkili ve amaca uygun düzenlenmiş veri yığıntılarından oluşmaktadır.

Olaylar ve nesneler arasında düzenli olarak tekrarlanan ilişki modellerini tanımlayan örüntü kavramı, örüntü tanıma teknolojisi ile daha önceden tanımlanmış bir örüntünün veri tabanındaki benzerlerini arama ve bulma yazılımlarını ifade eder.

Veri madenciliğinde veri tabanının büyüklüğü çok fazla olduğundan, başarılı bir analiz için veri madenciliği sürecinin düzgün işletilmesi önemlidir. Veri madenciliği sürecinde ilk aşama sorunun tanımlanması ve buna uygun olarak verinin hazırlanması aşamalarıdır (Şekil 2.3). Veriler hazırlanırken yapılan işlemler sırasıyla, verilerin toplanması ve uygun hale getirilmesi, birleştirilmesi, temizlenmesi, verilerin seçilmesi ve dolayısıyla en son analiz edilerek sonuçların yorumlanması aşamalarını kapsamaktadır (Ngai et al., 2009). Veritabanında depolanmış verinin sorgulama ve analiz için uygun hale getirilmesi işlemleri yürütülür. Lezki, (2016) raporunda veri madenciliği çalışmalarında bilgi keşfi sürecinde izlenmesi gereken temel aşamaları 5 ana başlık altında tanımlanmıştır.



Şekil 2.3. Veri tabanından veri madenciliğine veri analiz işlemleri (Lezki, 2016)

2.1. Veri madenciliğinin Temel Aşamaları

2.1.1. Amacın tanımlanması

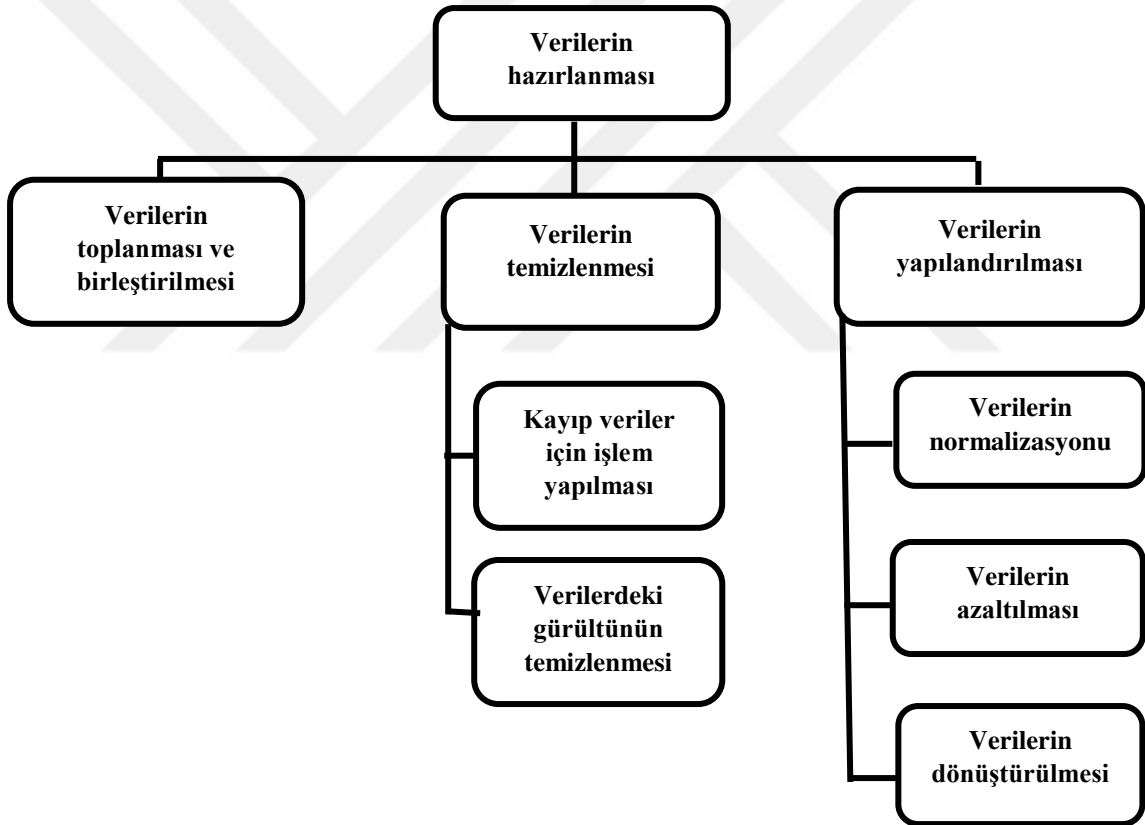
Veri madenciliğinde ilk adım olarak kabul edilen sınıflandırma işlemlerinde uygulama amacının belirlenmesi ve sonuçların başarı düzeylerinin saptanması işlemleri gerçekleştirilmelidir. Sorun ile tam örtüşmeyen veri kümelerini çalışmak sorunu çözemediği gibi farklı handikaplara da sebep olabilmektedir.

Amacın iyi tanımlanması çalışmanın başarısının anahtarını oluşturur. Amaç, var olan bir problemi çözmek odaklı olmalıdır. Bunun için öncelikle problemin tam, açık ve net biçimde tarif edilmesi gerekir. Elde edilecek verilerin ölçüm şekilleri belirlenmeli ve

olası sonuçların başarı düzeylerinin yanlış veya doğru değerlendirilmeleri durumunda maliyetlerin veya kazanımların ne olacağının öngörülmesi gerekir.

2.1.2. Verilerin derlenmesi ve ön işlemlerin yapılması

Bu aşamada verilerin seçimi, işlenmesi, birleştirilmesi, temizlenmesi ve dolayısıyla verilerin analiz için hazır hale getirilmesi işlemleri gerçekleştirilmektedir (Sönmez, 2016; Lezki, 2016). Veri madenciliğinin en önemli araçlarını veri tabanları oluşturur. Veri tabanlarında yer alan verilerin kalitesi hedeflenen sonuçların başarısında önemli bir yer tutmaktadır. Dolayısıyla verilerin öncelikle analize hazır hale getirilmesi gerekmektedir (Şekil 2.4).



Şekil 2.4. Veri ön işleme aşamaları (Lezki, 2016)

Bu aşamada verilerin toplanması, birleştirilmesi, temizlenmesi, amaca ve analiz metoduna uygun hale getirilmesi için yeniden yapılandırılması işlemleri gerçekleştirilmektedir. Veriler toplandıktan sonra şayet veriler farklı kaynaklardan toplanmışsa öncelikli olarak verilerin birleştirilmesi gerekir. Birleştirilen veriler

temizlenme aşamasına alınırlar. Temizlenme işlemlerinde kayıp veriler ve gürültülü veriler temizlenir. Son işlem olarak da verilerin yapılandırılması işlemi yapılmalıdır.

Verilerin hazırlanma aşamasında sorunu çözmeye odaklı olarak verilerin toplanma aşaması özenle gerçekleştirilmelidir. Bu aşama verilerin toplanacağı kaynağın belirlenmesi ile başlamaktadır. Verilerin az veya fazla olması doğru karar vermeyi veya karmaşıklığa sebep olabileceği unutulmamalıdır. Şayet veriler birden fazla kaynaktan toplanacaksa bu sefer veri uyumsuzluğu sorunu oluşabilecektir. Bu uyumsuzluklar arasında verilerin toplanma zamanları, kaydedilme formatları, kodlama değerlerinin farklılığı örneğin cinsiyet için “E veya K” ya da “1 veya 2” değerlerinin kullanılması gibi uyumsuzlukların hassasiyetle düzeltilmesi gerekmektedir.

Veriler toplanıp uyumsuzluklar giderildikten sonra birleştirme aşamasında verilerin tümü bir bütün haline getirilmelidir. Bir sonraki seçim aşamasında kullanılacak modele göre bağımlı veya bağımsız değişkenlerin sadeleştirilmesi veya temizlenmesi aşamaları yapılmalıdır. Örneğin sıra numarası veya kimlik numarası gibi bazı değişkenlerin dâhil edildiği bir analizde esas değişkenlerin modeldeki ağırlıklarını azalmasına neden olabilmektedir. Bu yüzden analiz aşamasına geçilmeden verilerin temizlenmesi amaca hizmet edebilmektedir.

Veri tabanında bazı veriler eksik olabileceği gibi (kayıp veri), bazı veriler de hatalı da olabilir (gürültülü veri). Veriler içinde tamamıyla yanlış veya hatalı veri olasılığını da göz ardı etmemek gerekir. Kayıp, gürültülü veya uyumsuz verilerin temizlenmesi ve uyumsuzlukların giderilmesi farklı şekillerde yapılabilmektedir. Örneğin eksik veri varsa ve bunlar sonuç üzerinde önemsiz derecede etkili ise tamamen silinebileceği gibi, eksiklikler tek tek de düzeltilebilir ya da ortalama değerler verilerek ortalama üzerinde etkisiz bir şekilde düzeltilebilirler. Hatta kayıp veriler diğer değişkenler yardımıyla tahmin edilebilir. Bunun için regresyon analizi, karar ağaçları ve diğer istatistiksel analiz metotlarından faydalanmak mümkündür (Chavan ve Verma, 2013; Lezki, 2016). Kayıp verilerin neden olacağı olumsuzlukları ortadan kaldırmak için kullanılan metotlar aşağıda detaylı olarak açıklanmıştır (Lezki, 2016).

Kayıp verileri silmek kullanılan yaklaşımlardan en kolay olan yaklaşımdır. Şayet toplam veri içinde kayıp veriler az oran tutuyorsa ve kayıp verilerin sonuç üzerine etkisi önemsenmeyecek derecede az ise kayıp veriler silinebilirler. Ama kayıp veriler çok olursa bunların silinmesi sonucu değiştirebileceği için hatalı sonuçların elde edilmesine sebep olabilecektir.

Kayıp veriler tek tek yazılabilirler. Bu yaklaşım kayıp veri sayısının az olduğu ve kayıp verilere ulaşmanın mümkün olduğu durumlarda uygulanabilir. Bu yaklaşım kayıp verilere mutlaka ihtiyaç duyulması durumunda uygulanabilir fakat gereksiz zaman kaybına neden olacaktır.

Kayıp verilerin hepsi için aynı veri girilerek kayıp verilerin neden olduğu olumsuzluklar giderilebilir. Veri sağlayan kişiler gelir veya yaş gibi bazı kişisel bilgilerini vermeyebilirler. Bunun yerine bilgi yok anlamında “Y” ya da bilgi eksik anlamında “E” girilerek analiz yapılabilir. Bu durum bazen algoritmalarda hatalara sebep verebileceği için eksik veri yerine standart bir şekilde örneğin “9” rakamı girilerek kayıp verilerden kaynaklı olumsuzluklar giderilebilir. Bazen bu uygulama analizlerde beklenmeyen sonuçlara da neden olabilecektir. Örneğin yaşını belirtmeyen kişilerin alışveriş sepetlerinde yaşlanma karşıtı ürünlerin var olması durumunda yaş-ürün tercihi ilişkisinin kurulmasını sağlayabilir. Kayıp veri yerine verilen harf veya rakam kodunun veriler içinde anlamlı olabilecek başka veriler ile çakışmamasına özen gösterilmelidir.

Kayıp veri yerine tüm verilerin ortalama değerlerinin girilmesi de kayıp verilerin neden olduğu olumsuzlukları gidermenin başka bir yöntemi olarak kullanılabilir. Bu yöntemde kayıp veri dışında bulunan tüm verilerin ortalaması alınır ve kayıp verilerin yerine bu değer girilir. Fakat bu yöntem verilerin rassal dağılım göstermesi durumunda kullanılabilir. Şayet eksik veriler tamamıyla rassal olarak kayıp ise (TROC, Missing Completely At Random, MCAR) eksik verilerin tamamlanmasında ortalamanın kullanılabileceği bildirilmektedir (Köse ve Öztumur, 2014). Değişkenler tamamen bağımsız olduklarında yani değişkenlerin gözlenme olasılıklarının birbirlerinden etkilenmemesi durumunda veriler tamamen rassal olarak kayıp durumu oluşturdukları kabul edilir.

Veri kümesindeki diğer veriler yardımıyla kayıp verilerin tahmin edilmesi yaklaşımının da kullanılması mümkündür. Bunun için regresyon analizi, karar ağaçları gibi çeşitli analiz yöntemlerinin kullanılması gerekmektedir.

Gürültülü veriler de farklı yaklaşımlar ile düzeltilebilir. Bu yaklaşımlar arasında bölümleme yöntemi, sınır değerlerin kullanılması, kümeleme ve regresyon yöntemleri sayılabilir (Lezki, 2016).

Bölümleme yönteminde veriler büyükten küçüğe sıralanır ve eşit biçimde bölümlere ayrılır. Her bölümün ortalaması alınarak her veri yerine ortalamalar yazılabilir. Bu durum aşağıda bir örnekle anlatılmıştır:

Veriler: 3, 7, 11, 12, 23, 25, 35, 36, 43 şeklinde olsun. Bölümleme;

Bölüm 1: 3, 7, 11

Bölüm 2: 12, 23, 25

Bölüm 3: 35, 36, 43 olacaktır. Ortalama bölüm 1 için 7, bölüm 2 için 20 ve bölüm 3 için ise 38 olacaktır. Buna göre bölüm verileri;

Bölüm 1: 7, 7, 7

Bölüm 2: 20, 20, 20

Bölüm 3: 38, 38, 38 olarak yazılabilir. Dolayısıyla her bir grubun değeri ortalama değer ile değiştirilerek düzeltilebilir.

Gürültülü veriler sınır değerleri kullanılarak da düzeltilebilir. Bu yöntemde de veriler sıralanır ve eşit olarak bölümlere ayrılır. Her bölümün alt ve üst verileri sınır değerleri olarak kabul edilerek bölüm içindeki veriler hangi sınır değerine daha yakınsa o sınır değeri ile değiştirilebilir. Sınır değerleri yaklaşımı ile veri düzeltme aşağıdaki gibi bir örnekle açıklanabilir:

Veriler: 3, 7, 11, 12, 23, 25, 35, 36, 43 şeklinde olsun. Bölümleme;

Bölüm 1: 3, 7, 11 $\Rightarrow$ alt sınır 3, üst sınır 11,

Bölüm 2: 12, 23, 25 $\Rightarrow$ alt sınır 12, üst sınır 25

Bölüm 3: 35, 36, 43 $\Rightarrow$ alt sınır 35, üst sınır 43

Sınır değerlerine göre düzeltme yapıldığında veriler aşağıdaki gibi olacaktır. Bölüm 1'de ortadaki değer her iki sınır değerinden eşit derecede farklıdır. Düzeltme ya alt sınır ya da üst sınır değerine göre yapılabilir.

Bölüm 1: 3, 3, 11

Bölüm 2: 12, 25, 25

Bölüm 3: 35, 35, 43

Sınır değer yönteminde başka bir yaklaşım da sınır değer farkının ortalamasının alınması ve tüm bölüm değerlerinin elde edilecek değer ile değiştirilmesi şeklinde de uygulanabilir.

Bölüm 1: 3, 7, 11 $\Rightarrow (11-3)/3=2,66$

Bölüm 2: 12, 23, 25 $\Rightarrow (25-12)/3=4,33$

Bölüm 3: 35, 36, 43 $\Rightarrow (43-35)/3=2,66$

Sonuçlar;

Bölüm 1: 2,66 2,66 2,66

Bölüm 2: 4,33 4,33 4,33

Bölüm 3: 2,66 2,66 2,66 şeklinde düzeltilebilir.

Kümelenme yöntemiyle gürültünün temizlenmesi yaklaşımı aykırı değerlerin bulunması ve düzeltilmesinde kullanılır. Veriler önce benzerliklerine göre kümelere ayrılırlar. Kümelemede uç değer olarak belirlenen veriler hiçbir kümede yer almayacak şekilde düzeltme yapılır. Bu şekilde belirlenen aykırı değerler en yakın olduğu kümenin ortalama değeri ile değiştirilir veya en küçük ya da en büyük değer atanarak aykırı değerler temizlenmiş olur.

Regresyon yöntemine göre gürültünün temizlenmesi yaklaşımında değişken değerler bir fonksiyon yardımıyla ilişkilendirilmektedir. Burada regresyon iki değişken arasında en uygun doğruyu bulur. Dolayısıyla bir değer kullanılarak diğerinin tahmin edilmesi sağlanabilir. Çoklu regresyonda ise birden fazla değişken söz konusu ise kullanılan doğrusal regresyonun genişletilmiş şeklidir.

Yapılandırma aşaması ise eldeki verilerin kullanılacak analiz yöntemlerine uygun hale getirilmesi işlemidir. Bazı algoritmalar sadece numerik değerler üzerinden çalışırken bazıları “0” ve “1” değerleri veya kategorik değerler üzerinden çalışabilmektedir. Yapılandırma aşaması verilerin azaltılması, normalleştirilmesi ve dönüştürülmesi işlemlerini kapsamaktadır (Imberman. ve Domanski, 2001; Lezki, 2016; Sönmez, 2016).

Çizelge 2.1. *R Core Team paketi değerlendirme teknikleri (R Core Team, 2019)*

Metod	Açıklama	Ölçek tipi	Varsayılan
pmm	Tahmini ortalama eşleştirme	sayısal	Evet
norm	Bayesian doğrusal regresyon	sayısal	
norm.nob	Doğrusal regresyon, non-Bayesian	sayısal	
mean	Koşulsuz ortalama doldurma	sayısal	
2L.norm	İki-seviyeli doğrusal model	sayısal	
logreg	Loistik regresyon	faktör, 2 seviye	Evet
polyreg	Çok terimli logit model	faktör, >2 seviye	Evet
polr	Düzenli logit model	düzenli, >2 seviye	Evet
lda	Doğrusal diskriminant analiz	faktör	
sample	Gözlenen verilerden rastgele örnek	hiç	

R Core Team (2019) geliştirdikleri ve R adını verdikleri bir istatistik programına eklenen mice 2.9 paketinde, eksik verilerin doldurulma tekniklerini açıklamışlardır (Çizelge 2.1). Tek değişkenli değerlendirme teknikleri olarak tanımlanan bu teknikler

amaca göre kullanılabilir. Nitel değişkenlerde logreg (mice.impute.logreg) doldurma tekniğinin iyi sonuç verdiği belirtilirken, nicel değişkenlerde pmm, norm, norm.nob gibi tekniklerin kullanılabileceğini ve bu teknikler içinde özellikle pmm'nin (mice.impute.pmm) iyi sonuç verdiği saptanmıştır (R Core Team, 2019).

2.1.3. Modelin kurulması

Modelin kurulması veri madenciliğinde önemli adımlardan birisidir. Bunun için en uygun model bulununcaya kadar birçok model incelenmeli ve denenmelidir. Supervised (denetimli) ve unsupervised (denetimsiz) öğrenme modellerine göre model kurulumu farklılık gösterebilir. Bu farklılık, denetimli öğrenmenin girdiden temel çıktıya kadar eşlemeyi içermesinden kaynaklı farklılık olarak kendini gösterebilir (Çizelge 2.2) (http-2).

Çizelge 2.2. Denetimli ve denetimsiz öğrenmenin karşılaştırılması (http-2)

Karşılaştırma için temel	Denetimli Öğrenme	Denetimsiz Öğrenme
Temel	Etiketli veri ile ilgilenir.	Etiketlenmemiş verileri yönetir.
Hesaplamalı karmaşıklık	Yüksek	Düşük
Analiz yapma	Çevrimdışı	Gerçek zaman
Doğruluk	Doğru sonuçlar üretir	Orta derecede sonuç üretir
Alt etki	Sınıflandırma ve regresyon	Kümeleme ve Dernek kural madenciliği

Denetimli ve denetimsiz öğrenmeği aşağıdaki şekilde bir örnek ile açıklamak mümkündür (http-3):

X bir veri girişi olsun $(X_1, X_2, \dots, X_n)$. Y de verilerin hedefi olsun $(Y_1, Y_2, \dots, Y_n)$. Eğitim hedefi olan N katmanlı veri kümesi de $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)\}$ şeklinde olacaktır.

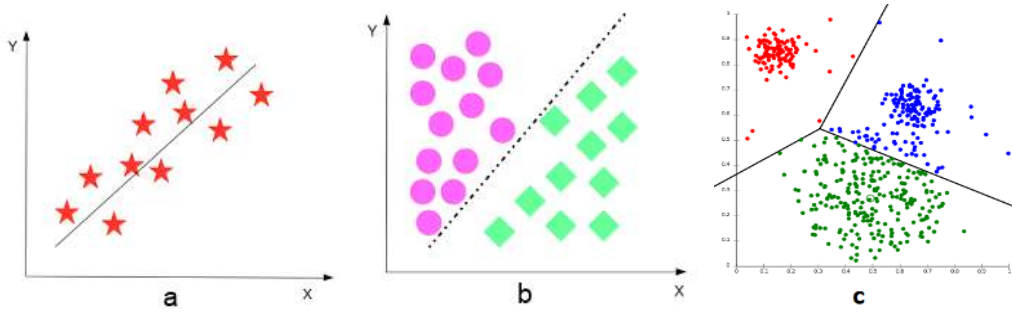
Denetimli öğrenmede belirli bir X için Y olasılığı öğrenilmeye çalışılır. Bu olasılık $P(Y|X)$ şeklinde gösterilir. Yani aslında burada bir deneyden veya anketten elde edilen kanıtlara göre rastgele değişken olarak bilinmeyen bir değişkenin olasılık dağılımı hesaplanmaktadır. Dolayısıyla giriş olarak bir (X) ve çıkış olarak bir (Y) değişkeni varsa ve girişten çıkışa kadar öğrenmek amacıyla eşleştirme işlevini öğrenmek için analiz yapılıyorsa bu bir denetimli öğrenme işlemidir. Fonksiyon olarak basitçe " $Y = f(X)$ " olarak gösterilir.

Denetimsiz öğrenmede ise her veriye karşılık gelen hedeflerin " $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)\}$ " biçiminde yeni bir veri seti oluşturmak üzere

birleştirilmiş olduğu gözetimsiz bir işleme dönüştürülmesi analizleri yapılmaktadır. Bu olasılık da " $P(X,Y)$ " şeklinde gösterilir. Denetimsiz öğrenmede sadece giriş olarak (X) verilerini içerirken çıkış olarak hiçbir veriye karşılık verilmemektedir.

Lezki (2016) hazırladığı raporda veri madenciliği modellerini tahmin edici (predictive) ve tanımlayıcı (descriptive) olarak ikiye ayırmıştır. Tahmin edici alt başlığında ise regresyon ve sınıflandırma modelleri yer almaktadır.

Regresyon modeli bağımlı ve bağımsız değişkenler arasındaki ilişkiyi net bir biçimde açıklayan fonksiyonel bir istatistik analiz yöntemidir. Regresyon probleminde sonuçlar sürekli bir çıktı içinde tahmin edilir. Örneğin evlerin büyüklüklerine göre fiyatlarının tahmin edilmesi bir regresyon problemidir. Çünkü fiyat evin büyüklüğünün bir fonksiyonudur (Şekil 2.5a). Aynı şekilde firmanın reklam harcamalarının satışları nasıl etkilediği de bir regresyon problemidir. Sınıflandırma modeli ise veri kümesini tanımlama sürecidir ve ayırmayı sağlayan en iyi modeli bulma aşamasıdır. Sınıflandırma probleminde sonuçlar ayrı ayrı çıktılarda tahmin edilir (Şekil 2.5b). Örneğin tümörlü bir hastada tümörün iyi veya kötü huylu olarak ayrılması bir sınıflandırma problemidir. Terlemez (2016) raporunda sınıflandırma işlemini verileri önceden tanımlanmış birçok sınıftan birisine atayan bir model olarak tanımlamıştır.



Şekil 2.5. Regresyon (a), sınıflandırma (b) ve kümeleme (c)

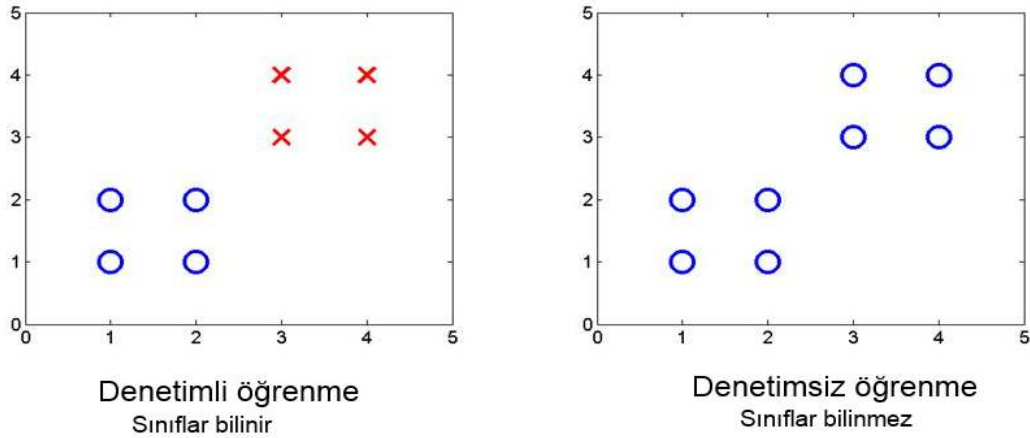
Lezki (2016) tanımlayıcı modeller arasında kümeleme, birliktelik kuralları, sıra örüntü analizi, özetleme tanımlayıcı istatistik ve istisna analizi modellerini sıralamıştır. Bu modellerden kümeleme analizi verileri birbirlerine olan benzerliklerine göre gruplara ayıran bir modeldir (Şekil 2.5c). Tahmin edici modellerden olan sınıflandırma analizinde veriler bağımlı bir değişkene göre gruplandırılırken tanımlayıcı model olan kümeleme analizinde veriler öznitelik değerlerine göre gruplandırılır. Kümeleme analizi denetimsiz

öğrenme başlığı altında yer alır çünkü bu analiz yönteminde önceden belirlenmiş sınıflar bulunmamaktadır.

Denetimli yapılan sınıflandırmada, sınıfların sayısı ve hangi nesnenin hangi sınıfta olduğu bilinmektedir. Yani ilgili sınıflar daha önceden belirlenen ölçütlere göre ayrılarak her sınıf ilgili alana dağıtılır. Bu sınıflandırma, örneklerden hareket ederek her bir sınıfa ait özelliklerin belirlenmesine ve sürecin sonunda oluşturulan kurallar yardımıyla yeni örneklerin hangi sınıfa dâhil edilmesi gerektiğini belirlemektedir

Denetimsiz yapılan sınıflandırmada ise nesnelerin özellikleri verilirken nesnelerin isimleri verilmez (Lezki, 2016). Dolayısıyla sınıf sayısı bilinmediği gibi nesnelerin hangi sınıfta oldukları da bilinmemektedir. Buradaki amaç örneklerin özellikleri arasındaki benzerliklerden yola çıkarak ait oldukları sınıfların tanımlanmasıdır.

Özetlemek gerekirse denetimli öğrenmede sınıfın özelliklerinden yola çıkarak örneklerin sınıfları belirlenirken, denetimsiz öğrenmede örneklerin özelliklerinden yola çıkarak sınıfların belirlenmesi aşamalarını kapsamaktadır (Şekil 2.6).



Şekil 2.6. Denetimli ve denetimsiz öğrenme arasındaki fark

2.1.4. Modelin kullanılması ve yorumlanması

Bu aşama kurulan ve geçerliliği denen modelin kullanıldığı aşamadır. Sonuçlar değerlendirilerek çalışmanın amacına uygun olup olmadığı incelenmeli ve problemin çözümüne katkı sağlayıp sağlamadığı belirlenmelidir. Amaçta belirlenen sorunlara çözüm getirmiyorsa farklı modelin kurulması amacıyla süreç yenilenmelidir (Lezki, 2016).

2.1.5. Modelin izlenmesi

Model ne olursa olsun iyi çalışan bir model süreç gelişiminden etkilenebilir. Dolayısıyla modelin çalışmasının sürekli izlenmesi ve ölçülmesi zaman içinde problemlerin çözümünden uzaklaşılmasını önleyecektir.

2.2. Veri Madenciliğinde Kullanılan Teoriler

Zhang vd. (2016) veri madenciliği temel teorileri 7 başlık altında derlemiştir. Chen ve Fawcett, (2016) ise yayınladıkları bir derleme çalışmasında yine veri madenciliği modellerini sınıflandırmışlardır. Bu bölümde veri madenciliğinde kullanılan modeller kısaca incelenecektir.

2.2.1. Bulanık mantık teorisi (Fuzzy theory)

Bulanık mantık teorisi bir denetimli öğrenme modelidir (Devillez, 2004). Klinik tıpta yaygın olarak kullanılan ve belirli kurallara sahip verilere uygulanabilen bir teoridir. Bulanık mantık teorisinin matematik temeli belirsiz kümeleri ve belirsiz mantıkları içermektedir. Bu matematiksel temel sayesinde veri fonksiyonları saptanarak ve belirli kurallara dayalı muhakeme yapılarak ilgili grupların ilişkisi belirlenebilir (Zhang vd., 2016). Bu teoriye örnek olarak Kuo ve arkadaşlarının (Kuo vd., 2015) prostat kanserinin seyrini (prognoz) tahmin etmek için yaptıkları çalışma örnek verilebilir. Araştırmacılar belirsiz kümeler sinir ağları yöntemi ile klinik verilerden prostat kanserinin seyrini tahmin edilebileceğini bildirmişlerdir.

Bulanık mantık, karmaşık ve zor birçok problemin çözümüne yaklaşım şekli ve getirdiği çözüme yönelik uygulamaları sayesinde başta sağlık olmak üzere fen, mühendislik, matematik ve sosyal bilimler alanlarında kullanım alanı bulmaktadır (Keskenler ve Keskenler, 2017).

2.2.2. Kaba küme teorisi (Rough set theory)

Bir denetimli öğrenme modeli olarak bu teori yanlış problemleri bulmak ve çözmek için kullanılan matematiksel bir araçtır (Qian vd., 2018). Veri madenciliği alanında kullanılan önemli metotlardan birisidir. Verilerin tanımlanmasında, azaltılmasında, verilerin değerlendirilmesinde ve analizlerinde geniş bir uygulama alanı bulmaktadır (Zhang vd., 2016). Kaba küme teorisinin avantajı veri analizinde herhangi bir ön bilgi gerektirmemesidir.

2.2.3. Bulut teorisi (Cloud theory)

Bulut teorisi bir denetimsiz öğrenme modelidir (Usama vd., 2017). , Bulanık mantık teorisinin uzantısı şeklindedir. Belirsiz dönüşüm modeli olarak kabul edilmektedir (Zhang vd., 2016). İlk defa Liu and Xio (2014) bulut teorisine dayalı yeni bir model tasarlamışlardır. Araştırmacılar sağlık yönetim sisteminde ankete tanımlı çalışmalar yapmışlardır. Sistemin hastalığı engelleme, geciktirme ve hatta tersine çevirme potansiyeline sahip olabileceğini bildirmişlerdir.

2.2.4. Dempster–Shafer teorisi (Dempster–Shafer theory)

Siemiatkowska ve harasymowicz-Boggio (2017) kesin olmamakla birlikte bu teoriyi denetimli öğrenme modeli olarak tanımlamışlardır. Fakat Ranoeliarivao vd. (2014) Dempster–Shafer teorisini denetimsiz öğrenme modeli olarak tanımlamışlardır. Klasik olasılık teorisinin uzantısı olarak kabul edilen Dempster-Shafer teorisi veri işlemlerini iyileştirmek amacıyla rough, cloud ve artificial neural network (yapay sinir ağları) teorileriyle birleştirilmiştir. Bu teori klinik tıpta geniş olarak kullanılmaktadır (Zhang vd., 2016). Bu teori PET analiz sonuçları analiz edilerek kanser sonuçlarının değerlendirilmesi ve hem akciğer kanseri hem de özofagus (yemek borusu) kanseri gruplarında yüksek tahmin doğruluğu sağlanması amacıyla kullanılmıştır (Lian ve Denoeux, 2015).

2.2.5. Yapay sinir ağları (Artificial neural network)

Bir denetimli öğrenme modeli olan bu algoritma insan beyin simülasyonlarından geliştirilen bir algoritmadır (Lezki, 2016). Beynin hafızasına benzer bir şekilde istatistiksel veri yasalarını öğrenir ve örneğin özelliklerini tanımlayabilen bir veri modeli yaratır (Zhang vd., 2016). Öğrenilen veri modeli daha sonra, yeni verilerin sınıflandırılmasında kullanılmaktadır. Bulanık mantık log sistemi ile kıyaslandığında bu sistemin daha akıllıca olduğu görülmektedir. İnan (2015) yapay sinir ağları (YSA) metodunu basit bir sinir sistemine benzeterek açıklamaya çalışmıştır. Sinir sistemi nöronlar içerir ve bunlar birbirlerine bağlanarak farklı ağ sistemlerini oluştururlar. Bu ağlar öğrenme, hafızaya alma ve veriler arasındaki ilişkiyi ortaya çıkarma kapasitesine sahiptirler.

Bu sistem son yıllarda tıp alanında bir atılım gerçekleştirmiştir. Son yıllarda Sunkaria ve arkadaşları (Sunkaria vd., 2014) hastaların kalp sağlığı seyrini değerlendirmek için ANN tabanlı bir kalp sınıflandırma algoritması geliştirmişlerdir.

Araştırmacılar, toplam 46 hastanın elektrokardiyografi sonuçlarını algoritmaya kaydetmişler ve elde edilen sonuçları kalp sorunu olan hastaların verileri ile karşılaştırmışlardır. Hem algoritmadan elde edilen verilerin hem de deneysel verilerin yüksek doğrulukla uyumlu olduğu gözlenmiştir.

2.2.6. Genetik algoritma (Genetic algorithm)

Lezki (2016) tarafından bir denetimli öğrenme modeli olarak sınıflandırılan genetik algoritma sınıflandırma ve optimizasyon problemlerinde geniş bir şekilde kullanılan kolay bir algoritmadır. Bu algoritma yönteminde biyolojik evölasyon simülasyonu yoluyla elde edilen veriler hesaplanarak kodlanır ve seleksiyon, krossingover, mutasyon gibi verilerin filtrelenmesine olanak verir. Sonunda doğal seçim kuralında gereksiz bilgileri ayıklayarak diğer bilgileri kaydeder (Zhang vd., 2016). Johnson ve arkadaşları (Johnson vd., 2014) bu yöntemi alzheimer hastalarına uygulayarak hastalık ilerlemesi gibi değişkenlerin kombine etkisinin tek bir değişkenden daha önemli olduğunu belirlemişlerdir.

2.2.7. İndükleyici öğrenme teorisi (Inductive learning theory)

Birçok veri parçasını işleyerek genel kuralları ve kalıpları ayıklayan makine öğrenme kategorisine ait bir teoridir (Zhang vd., 2016). Bir denetimli öğrenme modelidir (Hovsepien vd., 2011). Bu yöntem 2013 yılında hızlandırılmış (inductive) ve zamana yayılmış (interleaving) öğrenme arasındaki ilişkileri saptamak amacıyla uygulanmıştır (Birnbaum vd., 2013). Interleaving öğrenmenin indüktif öğrenmeyi artırdığı bildirilmiştir.

2.3. Veri Madenciliği Alanında En Çok Kullanılan Yöntemler

Zhang vd. (2016) veri madenciliği alanında en çok kullanılan yöntemleri 5 başlık altında sınıflamışlardır. Bu bölümde veri madenciliği alanında en çok kullanılan ve 2016 yılında yayınlanan derleme çalışmasında saptanan yöntemler özetlenerek anlatılacaktır. Bu yöntemler 5 ana başlık altında sınıflandırılmıştır (Zhang vd., 2016).

2.3.1. Bayes ağı (Bayesian network)

Bir denetimli öğrenme modeli olarak tanımlanan Bayes ağı (BN) (Lezki 2016), belirsiz bir bilgi ifade modelidir ve olasılık analizine dayalı muhakeme teorisidir. Nesnelerin grafik şeklindeki görüntüsünü verir (Zhang vd., 2016). Bu ağ sayesinde hastalık ve bu hastalığın belirtileri (semptomları) arasında olasılık ilişkileri kurulabilir ve dolayısıyla semptomların görülmesi durumunda bireylerin hastalığa sahip olma

olasılıkları hesaplanabilir. Akciğer kanseri ve radyoterapi gören hastalarda radyasyonun zatürre riskini belirlemek (Lee vd., 2015) ve tip 1 diyabetli hastalarda düşük glisemik indeksi ve hipoglisemi risklerini incelemek amacıyla (Dain vd., 2015) BN analizi yöntemi kullanılmıştır.

2.3.2. Karar ağacı (Decision tree)

Lezki (2016) tarafından bir denetimli öğrenme modeli olarak tanımlanan karar ağaçları değişken şekilleri yönünden regresyon modellerine benzer algoritmalar. Bu modelde tek bağımlı değişken ve çok bağımsız değişkenler bulunmaktadır. Çelik (2009) yaptığı tez çalışmasında karar ağaçlarının birçok avantaja sahip olduğunu ve bu avantajları (1) kolayca kural cümleciklerine çevrilebilir olmaları, (2) sürekli ya da kesikli veriler ile çalışabilmeleri ve (3) eksik veya hatalı veriler ile tahmin yürütebilme özelliklerine sahip olmaları şeklinde tanımlamıştır.

Parametrik olmayan yöntemler arasındadır. Bu ifade, bu karar ağaçlarının uzay dağılımı veya sınıflayıcı yapısı ile ilgili varsayımlara uymak zorunda olmadığı anlamına gelir. Bununla birlikte, eksik veya hatalı verilere duyarsız olması ve yaprak düğümlerde mükerrerlik içermesi de dezavantajı olabilmektedir.

Karar ağacı algoritması verileri bir dizi kurala göre sınıflandırır ve veri teorisindeki bilgiler ışığında veri tabanındaki en büyük bilgi alanını bulur. Daha sonra karar ağacının bir düğümü kurulur ve ağacın dalı alanın farklı değerlerine göre ayarlanır. Sonuçta karar ağacının alt düğümleri ve dalları oluşturulur (Terlemmez, 2016). Bir karar ağacı, kolayca anlaşılabilen kurallar oluşturan açık ve belirgin bir özellik gösterir (Zhang vd., 2016). Yang vd. (2015) halk sağlığında risk analizlerini saptamak amacıyla karar ağacı yöntemini kullanmıştır. Aynı şekilde Yaël vd. (2015) ortalama korpüsküler hemoglobin (MCHC) konsantrasyonunun doğruluğunu artırmak amacıyla kan örneklerini analiz etmek için karar ağacı metodunu kullanmışlardır. Sonuçta kanda bulunan bazı parametrelerin (rbc=red blood cells ve hbg=hemoglobin) tamamen anormal MCHC seviyelerini belirlediği saptanmıştır. Dolayısıyla karar ağacı metodu ile hızlı takip sağlanabileceği gibi biyolojik yorumlama kurallarının da oluşumuna katkı sağlayabilir.

2.3.3. Örüntü tanıma (Pattern recognition)

Örüntü tanıma, öğelerin tanımlanması, sınıflandırılması ve yorumlanması için farklı öğelerin özelliklerinin (sayısal, gerçek ve mantıksal ilişkiler) çeşitli bilgi işlem biçimlerini ve analizlerini ifade eder. Kalıplar, zaman ve alan dağılımına sahip verilerdir

ve farklı veriler, farklı kalıplara sahiptir. Örüntü tanıma yöntemi teorik karar yöntemi, sözdizimi yöntemi ve ANN yöntemi gibi alt bölümlere ayrılabilir (Zhang vd., 2016). Örüntü tanıma metodu insan yüzünün tanınması, ses ve parmak izinin tanınması, görüntü sinyali analizi, ürün hatasının tanınması ve diğer tanıma ve takip alanlarında yaygın olarak kullanılmaktadır.

Tıp ve sağlık alanlarında, örüntü tanıma yöntemi özellikle klinik otomatik test ve analizlerinde, EKG ve elektroensefalografın (EEG) uygulanması ve analizlerinde, tıbbi görüntü işleme ve analizlerinde, otomatik tedavi planlaması ve yardımcı tanı konularında büyük başarı elde etmiştir.

2.3.4. Yüksek performanslı hesaplama (High performance computing)

Yüksek performanslı bilgi işlem (HPC) genellikle hesaplama işlemlerini gerçekleştirmek için birkaç işlemcinin veya birkaç bilgisayar sisteminin kümesinin kullanılmasını ifade eder (Zhang vd., 2016). Dolayısıyla HPC birden fazla sunucunun gücünü yüksek performanslı bilgisayarlar sistemiyle birleştirerek işlemlerin daha hızlı tamamlanmasını sağlayan teknolojik bir uygulamadır. Bu işlemler için sadece donanım ve cihaz gelişimi yeterli değildir. Bilgisayar işlemcileri, sunucular ve diğer ağlar ile birlikte alt yapı, sistem yapısı, benzer algoritmalar ve belki de özel yazılımların birlikte geliştirilmesine bağlı bir sistemler bütünüdür. Bir milyon insanın yaklaşık 360 binden fazla gen çeşitlerinin analizinde ve protein-protein ilişkileri gibi biyolojik konformasyonların devasa analizlerinde güvenle kullanılmıştır (De Coninck vd., 2014; Fleck vd., 2015).

2.3.5. İstatistiksel analiz (Statistical analysis)

En eski veri madenciliği tekniği olan istatistiksel analiz olasılık teorisini ve istatistik prensiplerini kullanarak verileri analiz eder ve olaylar arasındaki korelasyonu belirler. Lineer ve non-lineer analiz yöntemi, sürekli ve lojistik regresyon analiz yöntemi, çok değişkenli analiz yöntemi ve zaman seri analiz yöntemi en yaygın kullanılan istatistiksel analiz yöntemleri arasında sayılabilir (Zhang vd., 2016). Bununla birlikte kümeleme analizi, diskriminant (discriminant) analizi ve tanımlayıcı istatistik gibi yöntemlerde yaygın olarak kullanılan yöntemler arasında sayılmaktadır.

Kümeleme analizi, karmaşık ve büyük verilerin yer aldığı veri setinde verileri gruplandırmak ve oluşan grupları karşılaştırmak amacıyla kullanılan çok değişkenli istatistik yöntemlerinden birisidir (Bal, 2016). Buradaki amaç ait olduğu küme belli

olmayan verilerin sınıflandırılarak anlamlı hale getirilmesidir. Bu yöntemde veriler önce uzaklık ve benzerliklerine göre kümelenirler. Kümeleme sonunda birbirine çok benzeyen veriler aynı kümede toplanmaktadır. Birbirine çok benzeyen verileri içeren kümeler homojen küme diğer farklı verileri içeren kümelere de heterojen kümeler denir (Bal, 2016). Başta sağlık olmak üzere ziraat bilimlerde, biyolojik, psikoloji ve hatta sosyolojik bilimlerde elde edilen verilerin sınıflandırılmasında sıklıkla kullanılan yöntemdir. Kendi arasında aşamalı ve birleştirici kümeleme yöntemleri olarak ikiye ayrılır.

Diskriminant analizi, nesne veya gözlemlerin ayrı kümelere ayrıştırılması esasına dayanan çok değişkenli istatistiksel tekniklerden birisidir (Oğuzlar, 2006). Bu yöntem iki ya da daha fazla grubun çoklu değişkenlere bağlı olarak karşılaştırılmasını sağlar. Burada amaç, grupların birbirinden farklılaşmasına neden olan değişkenleri saptamaktır. Yani diskriminant analiz yönteminde grupların ayırıcı özellikleri saptanmaktadır (Atakan ve Karabulut, 2003; Oğuzlar, 2006).

Tanımlayıcı istatistiksel yöntem ise, bir veya birden fazla dağılımı karşılaştırmak için kullanılan veya verilerden yola çıkarak frekans dağılımlarını sayısal değerler olarak tanımlayan bir yöntemdir. Dolayısıyla veri tiplerine göre hesaplamalarda farklı formüller kullanılmaktadır. Frekans dağılımları ortalama veya yüzde cinsinden standart sapma gibi değerler belirlenerek tablolar veya grafikler şeklinde sunulmaktadır (http-4). Tanımlayıcı istatistik analiz yöntemleri yer ölçüleri, değişkenlik ölçüleri, çarpıklık ölçüleri ve basıklık ölçüleri alt başlıklarına ayrılır.

2.4. Veri Madenciliğinin Uygulama Alanları

Günümüzde veri madenciliği gıda sektöründe, sağlık alanında, genetik çalışmalarda, kimya biliminde, iletişim endüstrisinde, pazarlama, bankacılık ve finansal alanlarda ve hatta astroloji biliminde dahi kullanılabilmektedir (Poyraz, 2012). Sosyal bilimlerde de veri madenciliği yöntemlerin yoğun kullanıldığı bilinmektedir. Bu alanlar aşağıdaki gibi kategorize edilmiş ve kısaca tanımlanmıştır:

2.4.1. Sağlık ve genetik alanında veri madenciliği

Sağlık ve genetik alanında veri madenciliği uygulamaları hem teşhis ve tedavi hizmetlerinin etkinliğinin artırılması hem de kurumlarda daha etkin bir yöntem sağlamak amacıyla kararları destekleyen bir sistem olarak değerlendirilebilir (Koyuncugil ve Özgülbaş, 2009; Poyraz, 2012). Bu alanda yapılacak veri analizler doktorların teşhiste karar vermelerine yardımcı olacağı gibi tedavi sürelerinin kısaltılmasına ve sağlık

hizmetlerinin etkinliğinin artırılmasına da katkı sağlayacaktır (Kocamaz, 2007; Koyuncugil ve Özgülbaş, 2009). Bununla birlikte hastalıkların seyri ve ilerleme yöntemlerinin anlaşılmasında da veri madenciliği analizlerinin etkin rol oynadığı bildirilmiştir (Kuo vd., 2015; Alonso vd., 2018). Zhang ve arkadaşları 2016 yılında yayınladıkları makalelerinde veri madenciliğinin sağlık alanında kararın desteklenmesinde, seyrinin tahmin edilmesinde, pratik ilaç kullanım rehberinin oluşturulmasında, tıbbi yöntemlerin geliştirilmesinde, kanıta dayalı ilaç kullanımının özendirilmesinde ve sağlık alanında gelecekteki öngörülerini görme alanlarında önemle kullanılabildiğini bildirmişlerdir (Zhang vd., 2016).

Genetik alanında ise veri madenciliği yöntemleri ıslah çalışmalarında yoğun olarak kullanılırken, kriminolojide ise suç işleme yatkınlığı ile genlerin ilişkilendirilmesi alanında da kullanılmaktadır. Vishweswaraiah vd. (2018) yaptıkları çalışmada insan-fare veri madenciliği yaklaşımı yoluyla mevcut veri setlerini kullanarak yeni KOAH aday genlerinin tanımlanabildiğini göstermişlerdir.

2.4.2. Gıda sektöründe veri madenciliği

Veri madenciliği gıda sektöründe karar verici kurumların güvenli rapor hazırlama ve karar verme aşamalarında yardımcı olmaktadır. Özellikle gıda kaynaklı olumlu veya olumsuz tüm özelliklerin (gıda kaynaklı alerjik reaksiyonlar, yaralanmalar ölümcül olaylar, kullanım hataları veya tehlikeli ürünler) analiz edilerek Gıda ve İlaç İdaresi (Food and Drug Administration-FDA) gibi kurumların güvenlik raporlarının hazırlanmasında veri madenciliği uygulamaları etkin bir şekilde kullanılmaktadır (Duggirala vd., 2016; Min vd., 2018). Fardet vd. (2018) raporlarında besleyici olmayan besin özellikleri (doku, su aktivitesi, glisemik ve doyma potansiyelleri ve raf ömrü gibi) ile işleme derecesi arasında bağlantı kurmak, beslenme özellikleri ile bu özellikler arasındaki ilişkileri araştırmak, bütünsel ve kantitatif teknolojik endeksleri bulmak ve buna bağlı olarak kantitatif gıda kurallarını saptamak amacıyla da veri madenciliği yöntemlerin kullanılabileceğini bildirmişlerdir.

2.4.3. Kimya endüstrisinde veri madenciliği

Veri madenciliği yöntemleri kimya ve proses endüstrisinde de kullanılabilmektedir. Özellikle yeni kimyasal ürünlerin bulunması, yeni ilaç türlerinin keşfi ve bunların sınıflandırılmasında kullanım alanı bulduğu gibi maddelerin korozyon etkilerinin tahminlerinde de kullanılabilmektedir (Doğan, 2007; Poyraz, 2012).

2.4.4. İletişim endüstrisinde veri madenciliği

Telekomünikasyon çağında yaşanan bu günlerde firma tercihleri, hat yoğunlukları, kullanım süreleri ve nitelikleri hesaplanarak firmaların altyapı yatırımları için bilgi amaçlı olarak veri madenciliği yöntemleri kullanılabilmektedir. Firmalar bu vesileyle müşteri bazlı kampanyalar da düzenleyebilmektedirler (Poyraz, 2012). Çelik (2009) yayınladığı çalışmada telekomünikasyon şirketlerinin veri madenciliği yoluyla alandaki iş akışını çözebileceklerini, iletişim ağlarını tanımlayarak sahtekârlıkları önleyebileceklerini ve hatta bu yolla veri kaynaklarını daha iyi kullanabileceklerini, dolayısıyla hizmet kalitesini arttırabileceklerini belirtmiştir.

2.4.5. Pazarlama, bankacılık ve finansal alanlarda veri madenciliği

Bankacılık ve pazarlama alanlarında da veri madenciliği yoğun olarak kullanılabilmektedir (Çelik, 2009; Poyraz, 2012). Özellikle kredi risk analizi ve kredi sahtekarlıklarının saptanmasında yoğun olmak üzere kar-zarar tespitlerinde ve dolayısıyla müşteri odaklı pazarlama tekniklerinin geliştirilmesinde veri madenciliği yöntemleri kullanılmaktadır. Kredi pazarlamalarında kredi için uygunluğun saptanmasında ve geri ödemelerde olası sorunların yaşanmasında ve dolayısıyla krediye uygun müşterilerin saptanmasında ön plana çıkan veri madenciliği yöntemleri müşteriye farklı seçeneklerin sunulması aşamalarının saptanmasında önem arz etmektedir. Poyraz (2012) sigortacılık alanında özellikle dolandırıcılık işlemlerinde ve riskli müşteri analizlerinde, bireylerin kültürel ve bölgesel potansiyeli dikkate alınarak analizlerin yapılabileceğini ve dolayısıyla risk olasılıklarının hesaplanabileceğini belirtmiştir.

2.4.6. Astroloji biliminde veri madenciliği

Veri madenciliği yöntemleri astroloji bilimi alanına giren tüm olasılık hesaplamalarında kullanılabilmektedir.

2.4.7. Sosyal bilimlerde veri madenciliği

Sosyal bilimler alanında da etkin kullanım alanına sahip veri madenciliği yöntemleri, çoğunlukla kamuoyu yoklamaları şeklinde uygulanmaktadır. Elde edilen olasılıklara göre strateji belirlenebilmektedir.

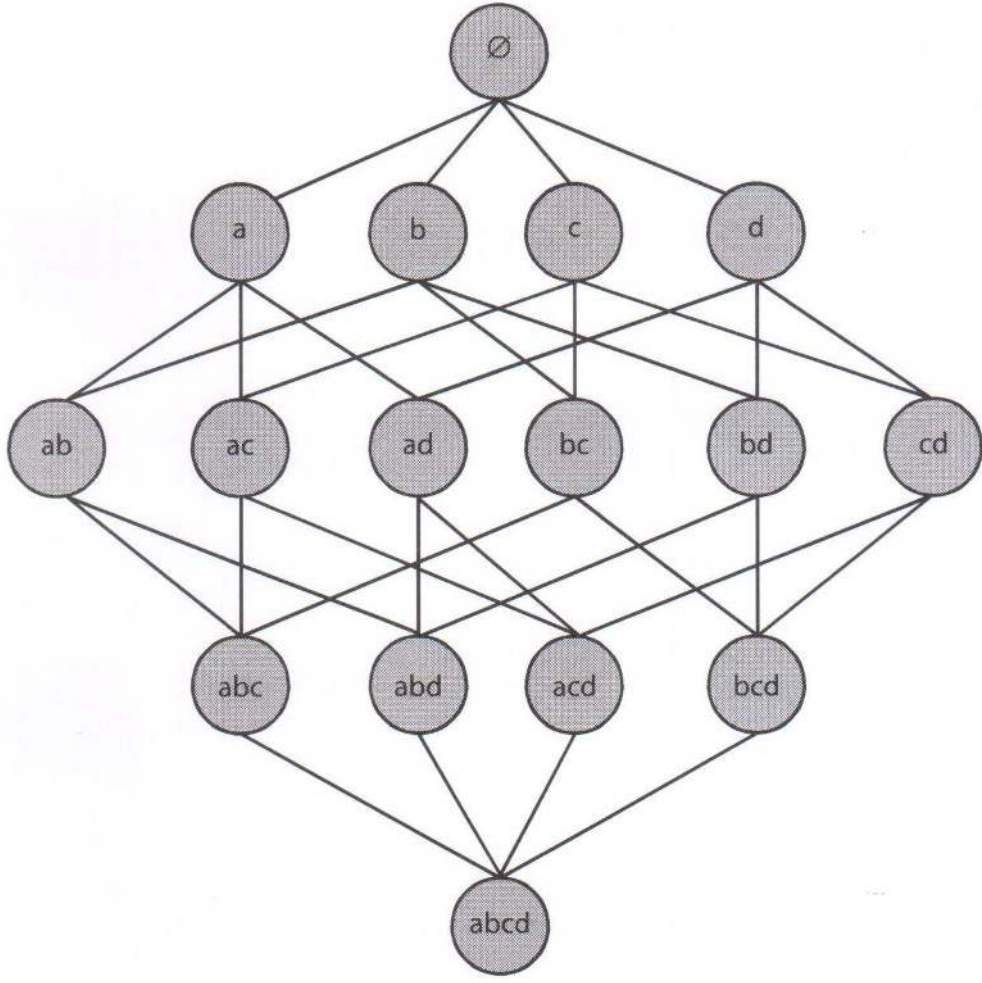
3. BİRLİKTELİK ANALİZİ VE APRİORİ ALGORİTMASI

3.1. Birliktelik Kuralı

Birliktelik kuralı verilerin analiz edilerek birlikte hareket eden verilerin davranışlarını ortaya çıkarmak ve bu birliktelik kurallarından yola çıkarak gelecekte bu aynı işlemler için çıkarım yapmak olarak tanımlanır. Yani kısaca birliktelik kuralı nesnelerin veya niteliklerin bir arada olma durumlarını belirlemede kullanılmaktadır. Birliktelik kuralı, mobilya sektöründen, inşaat sektörüne kadar her alanda, market zincirlerinde ve hatta sağlık alanında dahi kullanım yeterliliğine sahiptir (Kumdereli, 2012; İnan, 2015; Bekki, 2016; Sözen vd., 2017).

İlk defa 1993 yılında Agrawal ve arkadaşları tarafından kullanılmıştır (Agrawal vd., 1993; Özçakır ve Çamurcu, 2007). Araştırmacılar çalışmalarında Pazar sepetini örnek vererek birliktelik kuralını açıklamaya çalışmışlardır. Pazar sepetinde amaç markete giden bireylerin satın aldıkları ürünler arasında birliktelik ilişkisini kurmak ve dolayısıyla tüketicinin satın alma alışkanlıklarını belirlemektir. Örneğin A ürününü alan müşterilerin çoğunlukla B ve C ürünlerini de tercih ettikleri saptanması durumunda tüketiciye bu konuda karar verme konusunda yönlendirme yapılabilir. Özellikle bebek bekleyen tüketicilerin organik gıda ve sağlıklı beslenme reyonlarına yöneldikleri, bunların çocuk malzemeleri ile daha çok ilgilendikleri görülmüştür.

İlişki kuralı elde etmek için ilk olarak nesne küme sayısının bilinmesi gerekir (Bekki, 2016). Nesne sayısı “S” olan bir nesneler kümesi “K” ise bu nesnelerden oluşturulacak tüm nesnelerin sayısı $2^S - 1$ kadar olacaktır. Bu nesneler kümesinden bir tanesi boş küme olduğundan toplam nesneler kümesine dahil edilmemektedir. Örneğin $K = \{a, b, c, d\}$ sahipse ise $S=4$ demektir. Bu durumda nesne saetlerinin sayısı $2^4 - 1 = 15$ olacaktır (Şekil 3.1). Şekil incelendiğinde ilk nesne kümesinin boş olduğu dolayısıyla toplamda 15 kümenin oluştuğu görülecektir.



Şekil 3.1. Dört nesne içeren bir kümenin farklı alt nesne setlerinin oluşumu (Bekki, 2016)

İlişki kuralında ikinci adımda kural belirlemede kullanılacak nesne setlerinin içerisinde kaç tanesinin nesne içeren nesne seti olduğunun bilinmesidir (Bekki, 2016). Örneğin S adet nesne içeren bir K kümesinin ilişki kuralını oluşturmada kullanılacak nesne kümelerinin sayısı da aşağıdaki formül yoluyla hesaplanabilir (3.1) (Bekki, 2016):

$$C_K^S = \frac{4!}{2!(4-2)!} \quad (3.1)$$

İlişki kuralında son adım ise kural sayısının bilinmesi gerekir. Örneğin “S” adet nesne içeren bir “K” nesneler kümesinde toplam aşağıdaki (3.2) formüle göre hesaplanabilir (Bekki, 2016). Bu sonuç nesne kümesindeki nesne sayısının artmasıyla ilişki kuralı sayısının da katlanarak artacağı anlamını içermektedir. Dolayısıyla oluşturulabilecek tüm kurallar içinden işe yarayacak bilgiyi açıkça ortaya koyan kuralları

belirlemek amacıyla kural sayısında sadeleştirme yoluna gidilebilir. Bu sadeleştirme aşamalarında en çok destek ve güven eşik değerleridir.

$$3^S - 2^{S+1} + 1 \quad (3.2)$$

Birliktelik kuralında ürünler (öğeler) küme şeklinde sunulmaktadır. Örneğin X kümesinde n sayıda ürün olsun: A ve B de yapılan işi temsil etsin. Birliktelik kuralı ise aşağıdaki şekilde özetlenebilir (Özçakır ve Çamurcu, 2007; http-5).

$$X = \{X_1, X_2, X_3, \dots, X_n\}$$

$$A_1, A_2, A_3, \dots, A_n \Rightarrow B_1, B_2, B_3, \dots, B_n$$

Bu denklem $A_1, A_2, A_3, \dots, A_n$ işi yapıldığında aynı zamanda $B_1, B_2, B_3, \dots, B_n$ işinin de yapılmış olduğunu ifade etmektedir. Dolayısıyla A ve B kümelerinin birliktelik kuralı " $A \Rightarrow B$ " şeklinde ifade edilebilir. Bu önermeden yola çıkarak A ve B kuralı için destek (D) tanımlamaları;

$$D(A \Rightarrow B) = \frac{\text{A ve B'nin bulunduğu satır sayısı}}{\text{Toplam satır sayısı}} \quad (3.3)$$

şeklinde formüle edilebilir (3.3) (Bekki, 2016).

A ürününü içeren hareketlerin B ürününü de içermesi güven değerini vermektedir. Örneğin bir kural %95 güven değerine sahipse A ürününü alan bir kişinin veya A ürününde bir hareketin B ürününü de kapsaması ihtimali %95 olacak demektir. $A \Rightarrow B$ için güven (G);

$$G(A \Rightarrow B) = \frac{\text{A ve B'nin bulunduğu satır sayısı}}{\text{A'nın bulunduğu satır sayısı}} \quad (3.4)$$

şeklinde hesaplanabilir (3.4). Güven değerinin artması A ve B ürünlerinin beraber hareket etme olasılıklarını da artırır. Güven değeri %100 olduğunda işlemler kesin bir şekilde açıklanabilir.

Kural belirleme ölçütleri olarak destek ve güven değerlerine ilaveten lift (kaldıraç), leverage, conviction (inanç), Brute-Force gibi ölçütler de bulunmaktadır. Lift kuralı öncül (A) ve sonuç (B) nesneleri arasındaki ilişkiyi belirleme temeline dayanmaktadır. Bu durum güçsüz kurallar elendiğinde geri kalan ve belki de çok fazla olan güçlü kurallar içinde problemin amacına en uygun kuralın seçilmesini veya kuralların güçlüden itibaren sıralanmasını sağlamaktadır (Bekki, 2016).

Lift kuralı güven ve destek değerlerini birlikte kullanır:

$$Lift(A \Rightarrow B) = \frac{\text{Güven değeri } (A \Rightarrow B)}{\text{Destek değeri } (B)} = \frac{\text{Destek } (A \cup B)}{\text{Destek } (A) \times \text{Destek } (B)} \quad (3.5)$$

şeklinde hesaplanır (3.5). Sonuç olarak Lift=1 ise A ve B nesneleri arasında ilişkinin olmadığını, 1’den küçük olması durumunda $Lift(A \Rightarrow B) < 1$ pozitif bir ilişki olduğunu ve 1’den büyük olması durumunda ise $Lift(A \Rightarrow B) > 1$ negatif yönde ilişki olduğunu gösterir (Bekki, 2016).

Leverage ise, maddelerin istatistiksel olarak bağımsız olması durumunda, kuralın olasılığı ile beklenen olasılık arasındaki farkı belirlemektedir (The BigML Team, 2019). Denklem 3.6’da açıklandığı gibi formüle edilebilir.

$$Leverage(A \Rightarrow B) = \text{Destek } (A \Rightarrow B) - (\text{Destek } (A) \times \text{Destek } (B)) \quad (3.6)$$

Birliktelik kuralına ilişkin olarak geliştirilen bazı algoritmalar arasında AIS (Agrawal vd., 1993), Apriori (Agrawal ve Srikant, 1994), Partition (Savasere vd., 1995), CHARM, Rapid Association Rule Mining ve SETAM metotları sayılabilir (Özçakır ve Çamurcu, 2007).

Birliktelik kuralı veri madenciliği teknikleri arasında sınıflandırılmaktadır. Bu çalışmada birliktelik kuralı yönteminden yararlanarak tıp alanında özellikle hepatit hastalığına yakalanan bireylerin semptomlarının analiz ederek doktorların teşhis ve tedavisine yönelik kurallar geliştirilebilmesi amaçlanmıştır. Birliktelik kurallarında en fazla kullanılan algoritmalarından bir tanesi de Apriori algoritmasıdır (Özçakır ve Çamurcu, 2007; Maria Luna vd., 2018; Somek ve Hercigonja-Szekeres, 2017).

3.2. Apriori Algoritması

Apriori felsefe dilinde “a priori” kelimesinden türetilmiştir. A priori, bir bilginin ispata gerek duymadan doğruluğunun kabul edilmesi anlamında kullanılmaktadır. Örneğin “bir bütünün parçalarının bütünden küçük olması” kesin bir ifadedir (http-5). Bu algoritma çok büyük veri madenciliği çalışmalarında etkin olarak kullanılmaktadır. Amaç veri satırları (attributes) arasındaki ilişkiyi saptamak olarak açıklanabilir. Birliktelik kuralı oluşturabilmek için en sık kullanılan ve en çok bilinen algoritmadır (Bekki, 2016).

Apriori algoritmasının temelinde tüm verileri analiz etmek için geniş veri kümelerini sürekli taraması özelliği bulunmaktadır (Köroğlu, 2016). Bu analizde önce veri kümesinin destek seviyesi hesaplanır ve minimum destek seviyesi ile karşılaştırılır. Veri kümesinin her bir nesnesinin geniş olup olmadığı taranır. Bir nesne kümesinin geniş olarak adlandırılabilmesi için o nesne kümesinin kullanıcı tarafından verilen minimum

destek seviyesinin üzerinde bir destek seviyesine sahip olması gerekir. Bu işlem veri kümesinin en başından en sonuna kadar tekrarlanır ve tüm geniş veri kümeleri belirlenene kadar sürer (Agrawal ve Srikant, 1994; Koroğlu, 2016). Klasik bir Apriori algoritması kodu Şekil 3.2’de sunulmuştur.

```

 $L_1 = \{\text{sık geçen 1-öge kümesi}\};$ 
for ( $k=2$ ;  $L_{k-1} \neq \emptyset$ ;  $k++$ ) do begin
     $C_k = \text{apriori-gen}(L_{k-1});$  // Yeni adaylar
    forall transactions-hareketler  $t \in D$  do begin
         $C_t = \text{subset}(C_k, t);$  // Adaylar t içindedir
        forall candidates – adaylar  $c \in C_t$  do
             $c.\text{count}++;$ 
    end
     $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$ 
end
Answer =  $\bigcup_k L_k;$ 

```

Şekil 3.2. Apriori algoritması kodu (Özçakır ve Çamurcu, 2007).

Yukarıda özetlenen Apriori algoritması kodunda L , veri tabanı veya veri kümemizde bulunan öğeleri temsil etmektedir. Üretilen her bir C_k kümesi için, t bir işlemi olmak üzere, üretilen bu C_k kümesinin t işlemini tutan sayacı, her durumda bir arttırılarak algoritma çalıştırılmaktadır (http-5). Yukarıdaki algorithmada belirsiz olan apriori-gen fonksiyonu ise Şekil 3.3’te verilmiştir: Dolayısıyla bahsedilen bu algoritma, bütün verilerin üretilmesini sağlamak için tüm elemanların veri kümesine konulmasını öngörmektedir.

Apriori gen fonksiyonunda (Şekil 3.3) $(k-1)$ adet öğeye sahip L_{k-1} sık geçen öge kümesini kullanarak k adet öğeye sahip aday kümeleri oluşturulur. Bu fonksiyon ile ilk önce, L_{k-1} sık geçen öge kümesine kendisi ile birleştirme işlemi uygulanır. Birleştirme işleminde L_{k-1} sık geçen öge kümesinin her satırında yer alan son öge haricinde diğer öğelerin çapraz olarak benzerliği aranır ve son öge haricinde diğer öğelerle yakalanan benzerliklerden yeni aday öge kümeleri oluşturulur. Oluşan kümeler budama adımı ile budanarak fonksiyondan dönülür (Özçakır ve Çamurcu, 2007). Burada L_{k-1} öge kümesinde yer almayan tüm işlemler silinir. Aslında L_{k-1} öge kümesinde yer almayan öğelerin tüm alt öğelerinin C_k aday kümesinden silinmesi işlemi uygulanmaktadır.

Gölce (2010) ve İnan (2015) hazırladıkları Lisansüstü tezlerinde Apriori algoritmasını daha anlaşılır bir ifade ve geniş anlamı ile tanımlamıştır. Apriori algoritmasının temel yaklaşımı destek sağlayan alt kümelerin de destek sağlaması esasına dayanır. Yani bir öge kümesi en düşük düzeyde destek ölçütünü sağlarsa bu kümenin alt kümeleri de en düşük düzeyde destek ölçütünü sağlamalıdır mantığına dayanır (Gölce, 2010; İnan, 2015).

```

insert into  $C_k$ 
select  $p.items_1, p.items_2, \dots, p.items_{k-1}, q.item_k$ 
from  $L_{k-1} p, L_{k-1} q$ 
where  $p.item_1 = q.item_1, \dots, p.item_{k-2} = q.item_{k-2},$ 
       $p.item_{k-1} < q.item_k$ ;
for all itemsets  $c \in C_k$  do
  for all  $(k-1)$ -subsets  $s$  of  $c$  do
    if  $(s \notin L_{k-1})$  then
      delete  $c$  from  $C_k$ 

```

Şekil 3.3. Apriori gen fonksiyonu (Özçakır ve Çamurcu, 2007).

Çok kullanılan bir nesne küme oluşturuyorsa bu kümenin alt kümeleri de aynı şekilde çok kullanılan nesneler kümesi olmalıdır. Yani bir ürün ögesi minimum destek ölçütünü sağlıyorsa bu kümeye ait tüm alt kümeler de aynı ölçütü sağlıyor demektir. Örneğin X bahsedilen minimum destek kümesi ise Y de bunun alt kümesini oluşturuyorsa Y kümesinin eleman sayısı X kümesinin eleman sayısına ya eşit ya da onlardan az olmalıdır. $(X \supset Y) \Rightarrow s(X) \geq s(Y)$ şekline formüle edilir.

Bunun tam tersi olarak çok kullanılmayan nesneleri içeren kümelere farklı ürünler ilave edilmesi durumunda küme çok kullanılan nesneler kümesi olamayacağı bildirilmiştir (Gölce, 2010; İnan, 2015). Örneğin X kümesine yeni bir ürün eklendiğinde (örneğin A ürünü) oluşan yeni küme $(X \cup A)$ çok kullanılan ürünleri temsil edemez demektir.

Apriori algoritmasında kümelerin destek değerleri tarama sırasında hesaplanmaktadır. Destek değeri minimum ölçütü değerlerine sahip kümeler en çok kullanılan (sık kullanılan) öge kümeleri olarak kaydedilmektedir (Gölce, 2010). Bu taramalar en son sık kullanılan ögeler kümesi bulununcaya kadar devam etmektedir.

3.3. Birliktelik Kuralı ve Apriori Algoritması İle İlgili Yapılan Çalışmalar

Çok farklı sektörde kullanılan veri madenciliği çalışmalarında birliktelik kuralı ile ilgi önceki çalışmalar, sağlık dışı alanlarda ve sağlık alanında yapılan çalışmalar şeklinde ikiye ayrılarak sunulmuştur. PubMed bilimsel veri tabanında “apriori algorithm” anahtar kelimesi ile yapılan taramalarda 2.600’den daha fazla bilimsel çalışmaya rastlanmıştır. Dolayısıyla bu kısımda hem dünyada hem de ülkemizde Apriori algoritması çalışmalarından örnekler verilmiştir.

3.3.1. Sağlık dışı alanlarda birliktelik kuralı ve Apriori algoritması ile ilgili yapılan çalışmalar

Sağlık dışı alanlarda veri madenciliğinin en çok kullanıldığı alanlardan biri sigortacılık ve pazarlama alanlarıdır. Bu çalışmalarda müşterilerin poliçelerde isteklerini ve beraber satın aldıkları poliçeler saptanmaya çalışılmıştır. Doğan vd. (2014) bu alanda yayınladıkları raporlarında bu hususların saptandığını ve sigorta pazarlamalarında çalışanların müşterilere yaklaşımında yardımcı kuralların oluşturulabildiğini belirtmişlerdir. Aytaç ve Bilge (2013) Apriori algoritmasını pazarlama verilerine uygulamışlar ve pazarlama stratejilerinde önemli bilgileri saptayabildiklerini rapor etmişlerdir. Pastanede pazarlama verileri ile (Özçakır ve Çamurcu, 2007) ve marketlerde raf dizilimi ile ilgili (Timor ve Şimşek, 2008) çalışmalarla pazarlama alanında müşteri odaklı analizler yapılmıştır. Padilla ve García (2018) marketlerde çok satan ürünlerin göz önüne alınması ile ilişkili mekânsal değişikliklerin gelecekte hatayı en aza indirmek ile ilişkili çalışmalarında Apriori algoritması analiz sonucunda mekânsal tahminin doğru yapılması ile satışta iyileşmeyi ilişkilendirdiklerini bildirmişlerdir.

Özçalıcı (2017) ikinci el araba piyasasında veri madenciliği yöntemlerinden birliktelik kuralını analiz ederek ikinci el araç piyasasının işleyiş mekanizmalarına ilişkin faydalı bilgiler elde etmeyi hedeflediği bir çalışma gerçekleştirmiştir. Apriori algoritması sonuçlarına göre “dizel araçlar az yakar, fazla hız yapmazlar, torkları yüksektir, yeni ve pahalı araçların vergisi yüksektir” kurallarının ön plana çıktığını saptamıştır. Bu çalışma müşteri tercihinin neler olabileceğinin tahmin etmemize yardımcı olmaktadır. Otomobil satış ve tercih verilerinde hem Apriori hem de başka analiz yöntemlerinin kullanılması sonucunda müşterilerin otomobil tercihlerinin ve otomobil alırken beraber aldıkları ürünlerin saptanabileceği bildirilmiştir (Erpolat, 2012). Bu çalışmaların dışında ikinci el ve yeni araçların özellikleri, fiyatları ve satın alma tercihleri ile ilgili birçok çalışma mevcuttur (Asilkan ve Irmak, 2009; Daştan, 2016; Akçi, 2016).

Söylemez vd. (2016) trafik kazalarında birliktelik kuralını analiz ederek Ankara ilinde bir çalışma gerçekleştirmişlerdir. Araştırmacılar çalışmalarının amaçlarını elde edecekleri kurallar sayesinde farklı hava koşullarında, kazanın gerçekleştiği yollar, yolun bölünmüş veya bölünmemiş olma durumlarında kaza olasılıklarının ve yaralanma veya ölüm riskinin belirlenmesi olarak açıklamışlardır. Çalışma sonucunda bazı koşullarda alınacak tedbirlerin hem maddi hasar oluşturan hem de ölümcül kazaların azaltılması konusunda önemli tahminlerin yapılmasına imkan verdiğini bildirmişlerdir. Antalya ilinde de ve Türkiye genelinde de benzer çalışmalar yapılmış ve kazaların demografik özellikleri incelenmiştir (Eken vd., 2005; Temel ve Özcebe, 2006; Sungur vd., 2014). Dünyanın bir çok ülkesinde de benzer çalışmaların gerçekleştirildiği, trafik kazalarının demografik verilerinin incelendiği, başta Fuzzy teorisi ve diğer birliktelik kuralı analiz yöntemlerinin kullanılarak trafik kazalarının oluşma olasılıklarının hesaplandığı çalışmalar mevcuttur (Vlahogianni ve Golias, 2012; De Oña vd., 2013; Tavassoli Hojati vd., 2013; Mirzaei vd., 2014; Dimitriou ve Vlahogianni, 2015; Hong vd., 2015; Kar vd., 2015).

Apriori algoritmasının yukarıda bahsedilen alanlar dışında sosyal medya kullanımından, öğrencilerin başarılarının sebeplerini analizlerine kadar, üniversiteyi tercih sebeplerinden telekomünikasyon sektöründe müşteri analizine kadar çok farklı alanlara uygulanabileceğini gösteren çalışmalar bulunmaktadır (Lin, 2007; Karabatak ve İnce, 2004; Naboulsi, 2015; Kurt Pehlivanoğlu ve Duru, 2015). Hatta Demirel 2010 yılında hazırlamış olduğu yüksek lisans tezinde SGK veri tabanından elde ettiği verileri kullanarak veri madenciliği yöntemlerini uygulamıştır (Demirel, 2010). Sözen vd. (2017) Türkiye’de mobilya sektöründe Apriori algoritması analizi sonucunda birliktelik kurallarının çıkarımını sağlamışlar ve sonuçta farklı tüketicilerin mobilya seçiminde tercihlerini ve hangi ölçütü öncelikli olarak istediklerini belirleyebilmişlerdir.

Robertson vd. (2019) sporda antrenman ortamlarının iyileştirilmesi amacıyla görsel öğrenmenin önemini ölçmek için Apriori algoritması analiz yöntemlerini uygulamışlardır. 2015 yılında Avustralya Futbol Ligi sezonundan 9005 tekme örneğini basınç tipi, vuruş mesafesi, pozisyon kaynağı, pozisyon süresi, hız ve vuruş hedefi şeklinde sınıflandırmışlardır. Araştırma sonucunda bu çalışmanın sporda centilmenliğin derecesini belirlemek için bir ölçü yaklaşımı sağladığını ve diğer tüm spor dallarına uygulanabileceğini saptamışlardır.

3.3.2. Sağlık alanında birliktelik kuralı ve Apriori algoritması ile ilgili yapılan çalışmalar

Wei vd. (2019) Çin’de koroner kalp hastalıklarında kullanılan ilaçların özelliklerini ve reçetelerin düzenliliklerini incelemek için Apriori algoritması yöntemlerini kullanmışlardır. Bunun için Çin’de koroner kalp hastalığı ve anjina pektorisin (koroner arter hastalığı) tedavisi için kullanılan tüm ilaçların bilgileri toplanmış, bunların arasına Çin’de kullanılan tıbbi bitkiler de dâhil edilmiş ve analize tabi tutulmuştur. Çalışmada koroner kalp hatalıklarının tedavisinde yeni ilaçların geliştirilmesi için fikir veren ilaç özellikleri ve reçeteleme kuralları belirlenmiştir.

You vd. (2019) böbrek hastalıklarında kemik iliği baskılamasının (BMS) tedavisi için, bitki ve otların klinik uygulamasına dair kanıt ve referans sağlamak amacıyla yapılan ve Apriori algoritması analiz yöntemlerinin kullanıldığı çalışmalarında veri tabanından elde ettikleri toplam 239 formül ve 202 bitki ile analiz çalışmaları yapmışlardır. Bitkiler kan aktive edici, idrar söktürücü, sindirimi kolaylaştırıcı gibi kategorileri ayrılmış ve sonuçta Apriori algoritması ile 26 bitkinin birlikte kullanım kuralları belirlenmiştir.

Xia vd. (2019) tıbbi kaliteyi arttırmak, hastanede kalış süresini kısaltmak ve gereksiz tedaviyi azaltmak için birliktelik analizinin yapılabileceğini bildirmişlerdir. Bu amaçla tanı ve tedavi verileri ile yaptıkları çalışmada saptanan kuralların uygulanması ile hastalıkların teşhisi ve tedavisinde herhangi bir maliyet artışı olmadan ciddi anlamda tıbbi kalitede belirgin bir iyileşme sağlandığını göstermişlerdir.

Chen vd. (2018) akciğer kanserli hastalarda beklenmeyen, ters (advers) olaylar ve farmakoterapi arasındaki ilişkileri araştırmak için Apriori algoritması analiz yöntemleri kullanmışlardır. Çalışmanın sonucunda ilaçlar ve yan etkiler arasındaki anlamlı ilişki kurallarına rastlandığını ve kanser tedavisi sırasında yan etkilerin risk faktörlerini ortaya çıkarmak için umut verici bir yol gösterdiğini saptamışlardır.

Nassar ve Richter (2018) diyabetik olmayan gastroparezi (mide tembelliği) ile ilişkilendirilen hastalık durumlarını analiz etmek amacıyla mide tembelliği ile eşlik eden hastalıkların verilerini toplamışlar ve Apriori algoritması yöntemiyle analiz etmişlerdir. Çalışma sonucunda mide tembelliği ile en yakından ilişkili olduğu tespit edilen belirgin koşulların sırasıyla kronik pankreatit, son dönem böbrek hastalığı, irritabl bağırsak sendromu, sistemik lupus eritematozus (otoimmün bağ doku hastalığı), fibromiyalji (yumuşak doku romatizması) ve venöz tromboemboli (pıhtılaşma) hastalıklarının olduğunu göstermişlerdir.

Yi vd. (2018) demans (unutkanlık) hastalarının tedavisinde antik tıp kayıtlarına dayanarak Çin tıbbi reçetelerin ilaç özelliklerini ve düzenliliğini analiz etmek için eski tıbbi kayıtları ele aldıkları çalışmalarında en sık ve düzenli yazılan reçeteler arasında kurulan ilişki neticesinde en sık kullanılan bitkisel ilaçların listesini belirleyebilmişlerdir. Araştırmacılar çalışma sonucunda Apriori algoritmasının unutkanlık tedavisinde Çin tıbbının klinik uygulamasına rehberlik edebilecek derecede yararlı olduğunu bildirmişlerdir.

Guan vd. (2018) diyet yapan insanların yiyecek tercihlerini belirlemek için bir çalışma tasarlamışlardır. 12 ay boyunca diyet uygulayanların veri geçmişleri toplanmıştır. Karmaşık hiyerarşik yiyecek gruplarını sınıflandırmak için Apriori algoritması kullanılmıştır. Toplamda 432 kahvaltı, 428 öğle yemeği, 432 akşam yemeği ve 433 kişi ilave öğün alım verilerinden toplam 433 katılımcı tespit edilmiştir. Üç gıda grubu düzeyinde toplam birbiriyle yakından ilişkili gıda kümeleri olmak üzere toplam 142 madde belirlenmiştir. Örneğin birinci alt gıda grubu düzeyinde, ekmek öğle yemeğinde besin kombinasyonlarının merkezi olarak ortaya çıkmıştır, ancak bunun için işlenmemiş et akşam yemeğinde ortaya çıkmıştır. Akşam yemeği genel olarak daha fazla sebze ve yiyecek çeşidi ile karakterize edilmiştir. Sonuçta yiyecek gruplarının tanımları, ana öğünlerde yiyecek seçim kalıplarının belirlenmesinde çok önemli bir rol oynadıkları görülmüştür. Dolayısıyla yeme alışkanlıklarının anlaşılması, bireysel düzeyde kilo kaybı için yeni diyet stratejilerinin belirlenmesine ve geliştirilmesine yardımcı olabileceği saptanmıştır.

Prostat kanseri hastalarında patolojik faktörler ile eritroblast (olgunlaşmamış alyuvar hücresi) transformasyon sistemi (ETS) ilgili gen (ERG) ekspresyonu arasındaki ilişkileri değerlendirmek amacıyla Apriori algoritmasının kullanıldığı bir çalışma gerçekleştirilmiştir (Lee vd., 2018). 60 radikal prostatektomide (prostat ameliyatında) ERG ekspresyonu Apriori algoritması yöntemiyle patolojik bulgularla karşılaştırılmıştır. Tüm DNA ve RNA dizilimi, üç ERG pozitif ve negatif prostat kanseri numunesinde gerçekleştirilmiştir. DNA analizleri ve RNA dizilimine göre ilişkili genlerde toplam 97 mutasyon saptanmıştır. Bu genler NOTCH1, MEF2C, STAT3, LCK, CACNA2D3, PCSK7, MEF2A, PDZD2, TAB1 ve ASGR1 şeklinde tanımlanmıştır. NOTCH1 geni ana şalter geni olduğundan ERG ile NOTCH1 pozitifliği arasında anlamlı bir ilişki bulunmuştur. Sonuçta ERG ekspresyonu ile patolojik faktörler arasındaki ilişkinin

değerlendirilmesi, düzenleyici ağın ve önemli merkezin belirlenmesi, ERG pozitif prostat kanserinin klinik öneminin anlaşılmasına yardımcı olabileceği sonucuna varılmıştır.

Tang vd. (2013) tedavi edilen oral yassı hücreli kanser (OSCC) hastalarında 5 yıllık sağ kalım ile ilişkili klinik-patolojik faktörlerin katkısını ortaya çıkarmak ve sağ kalım süresini uzatmak için Apriori algoritmasını kullanmışlardır. Çalışma sonucunda tek değişkenli demografik analizler, klinik tümör boyutunun, patoloji tümör boyutunun ve tedavi gruplamasının 36 aydan uzun süren sağ kalım ile ilişkili olduğu gösterilmiştir. Sonuç olarak, erken tedavi OSCC hastalarının sağ kalımını arttırmanın anahtarı olduğu bildirilmiştir.

Erdem ve Özdağoglu (2008) Ege Bölgesinde yer alan bir hastanenin acil servisine başvuran hastaların verisini alarak Apriori algoritmasını analiz yöntemlerini uygulamışlardır. Hastalara ilişkin anlamlı kurallar ortaya çıkmıştır. Araştırmacılar çalışmadan elde edilen sonuçların, acil servislere gelen hastaların başvuru nedenleri ve hasta profilleri hakkında bazı ipuçları verdiğini, ayrıca acil servis bölümlerinin yeniden yapılanma çalışmalarına da farklı bir açıdan yol göstererek katkıda bulunduğunu saptamışlardır.

3.3.3. Hepatit ve hepatit hastalıklarında veri madenciliği çalışmaları

Hepatit çoğu kez virüslerle oluşan, ancak kimyasal maddeler, ilaçlar, alkol, kalıtsal hastalıklar veya otoimmün hastalıkların da neden olabildiği bir karaciğer hastalığıdır (http-6; Miguet vd., 1990; Seitz, 1995; Selmi ve Gershwin, 2004; Briese vd., 2009; Heyman vd., 2010; Gallegos-Orozco ve Rakela-Brödner, 2010; Weiss ve Leibowitz, 2011; Romero ve Razonable, 2011; Okano ve Gross, 2012). Hastalık bazen hafif seyredebilmekte ve belirti vermeden çok uzun süre devam edebilmektedir. Bazen kronikleşerek başta siroz ve karaciğer kanseri olmak üzere çeşitli semptomlara neden olmaktadır ve hatta ölüme de sebebiyet verebilmektedir (Çizelge 3.1).

Belirtiler (Semptomlar)

Hepatit belirtileri genel anlamda tüm bireylerde aynı olmakla birlikte nadir durumlarda kişiden kişiye değişebilmektedir. Çoğu durumlarda hafif seyreden hepatit belirtileri soğuk algınlığı ile karıştırılmaktadır. Yorgunluk, halsizlik, bulantı, karın ağrısı, eklem ağrıları, kaşıntı ve sarılık en sık görülen semptomlardır. Bazı kişilerde iştah kaybı, koyu renkli idrar ve açık renkli dışkı gibi ilave belirtiler göstermekle birlikte daha ciddi durumlarda karında sıvı birikimi (asit) ve zihin bulanıklığı da görülmektedir (http-6).

Çizelge 3.1 incelendiğinde, başta virüsler hepatit hastalığına neden olabileceği gibi kullanılan ilaçlar veya maruz kalınan kimyasal maddeler de hepate neden olabilmektedir. Hatta bazı bireylerde hepatit nedenleri kalıtsal da olabilmektedir. Bazı olgularda karaciğer yağlanması veya bağışıklık sisteminde görülen bazı hasarlar sonucunda da karaciğer yetmezliği yani hepatit hastalığı oluşabilmektedir. Fiziki muayenede karaciğer büyümesi önemli bulgular arasında sayılır. Bazı durumlarda sadece halsizlik gibi kendini gösteren kronik hepatit karaciğeri yavaş yavaş tahrip edebilmekte ve karaciğer yetmezliğine neden olabilmektedir.

Çizelge 3.1. Sık görülen hepatit tipleri (http-6)

Hepatit tipi	Tanımı	Nedenleri
Viral	Hepatit virüslerinin biriyle oluşan enfeksiyonlar	Hepatit A, B, C virüsleri
Kimyasal maddeler veya ilaçlar	Karaciğer detoksifikasyon (arınma) organı olduğu için kimyasal maddeler toksik etki gösterebilirler	Aşırı tüketilen tüm kimyasal maddeler
Kalıtsal	Gen mutasyonları karaciğeri hasara uğratarak hepatite neden olan olabilirler	Wilson hastalığı, , hemokromatoz, alfa-1 antitripsin
Karaciğer yağlanması	Alkolik ve non alkolik olabilir. Karaciğerde aşırı yağ depolanması sağlıklı karaciğer dokusunun azalmasına yol açabilirler	Lizozomal depo hastalıkları
Otoimmün	Vücudun bağışıklık sistemi karaciğer dokusuna uygunsuz biçimde yönlendirilmiş antikorlar üretir.	Tip I diyabet, Sjögren sendromu

Testler

Kan analizi ile yapılan birçok test hepatit teşhisinde yardımcı olabilmektedir (http-6). Bu testler amaca göre 4 farklı başlık altında sınıflandırılabilir:

A-Biyokimyasal testler: Karaciğer iltihaplanması veya hasarını saptamak için kullanılan testler.

B-Elisa testleri: Viral hepatiti saptamak için kullanılan testler.

C-Arka plan testleri: Altta yatan nedene tanı koymaya yardımcı olmak için kullanılan testler.

D-İzlem testleri: Hepatitin seyrini izlemek veya tedaviyi yönlendirmeye yardımcı olmak için kullanılan izlem testleri.

Kapsamlı Metabolik Panel (KMP) olarak tanımlanan testlerde hiçbir belirtisi olmayan veya belli belirsiz hastalık belirtileri olan hastada hepatiti saptamaya yarayan rutin tarama testidir.

Bütün bu testlerin dışında karaciğer hasarının ciddiyetini saptamak için ilave testler uygulanmaktadır:

A-Alanin aminotransferaz (ALT): Başlıca karaciğerde bulunan bir enzim olup hepatiti en iyi saptayan testtir.

B-Aspartat aminotransferaz (AST): Karaciğerde ve özellikle kalp ve başka kaslarda bulunan bir enzimdir.

C-Alkalen fosfataz (ALP): Safra kanallarıyla ilişkili, kanallar tıkalı olduğu zaman yükselen, kemik hastalıklarında da artan bir enzimdir.

D-Bilirubin: Sarılığa neden olan ve kan hücrelerinin parçalanması sonucu oluşan üründür.

E-Albümin: Karaciğer tarafından üretilen başlıca proteindir. Karaciğerin bu proteini ne kadar iyi yaptığının ölçüsüdür.

F-Total Protein: Kanda bulunan tüm proteinlerdir.

G-Protrombin zamanı (PT): Hepatit hastası olan veya hepatiti olduğundan kuşku edilen kişiler için bu test istenebilir. Kan pıhtısı oluşumu için kullanılan proteinler (pıhtılaşma faktörleri) çoğunlukla karaciğer tarafından üretilmekte olup uzamış protrombin zamanı karaciğer hasarının şiddet derecesini gösterebilir.

H-Karaciğer biyopsisi: Biyopsi hastalığa tanı koymanın en kesin yöntemidir. Bu bir girişimsel işlem olduğundan en son istenen testlerden birisidir.

I-USG ve görüntüleme teknikleri: Hepatiti belirlemek, tanı koymaya ve karaciğer hasarının nedenini belirlemeye yardımcı olmak için ultrasonografi ve özel görüntüleme teknikleri kullanılabilir.

Bulaşma ve korunma

Virütik hepatitlerde kan yolu en önemli bulaşma şeklidir. Bunun dışında cinsel ilişki, vücut sıvıları ve hatta virüs bulaşmış sulardan da bulaşma riski mevcuttur. Virütik hepatit bulaşmış gebe annelerin de çocuklarına bulaştırma riskleri göz ardı edilmemelidir. Gebe annelerin hepatit enfeksiyonu yönünden taranmaları ve yeni doğanların aşılanmaları korunma programında önemli yer tutmaktadır.

Tedavi

Hepatit A hastaları herhangi bir komplikasyon ile karşılaşmadan iyileşebilmektedir. Fakat belirtilerin görüldüğü hastalarda çoğunlukla bol istirahat ve sıvı alımı ile birlikte iyi beslenme en önemli tedavi edici önerilerdir. Hepatit E, B veya daha riskli türü olan

hepatit C türleri de söz konusudur. Hepatit E, hepatit A gibi bağışıklık sistemi ile yok edilerek hastalar iyileşebilir. Hepatit B hastalarının yaklaşık %90'ı tamamen iyileşebilir fakat geri kalan kısmının siroza yakalanma riskleri söz konusudur. Hepatit C ise siroz ve karaciğer kanserinin ana nedenlerinden birisidir (Rosen, 2011). Siroz veya karaciğer kanseri gibi olgularda ise hastalar sürekli kontrol altında tutulmaktadırlar. Bazı durumlarda antiviral ilaçlar ve bağışıklık sistemini güçlendiren interferonlar gibi biyoteknolojik ilaçlar tedavi amaçlı kullanılabilir.

Hepatit veri madenciliğinden elde edilen veriler kullanılarak hepatit teşhisinin daha iyi yapılabilmesini sağlayan tekniklerin saptanması ile ilgili az sayıda çalışma bulunmaktadır. Bu çalışmalarda farklı test teknikleri kullanılmış veya entegre test sistemleri ile çalışmalar yapılmıştır (Chen vd., 2011; Başçıl ve Temurtaş, 2011; İnan, 2015). Dolayısıyla yapılan bu çalışma veri madenciliği yoluyla elde edilen hepatit verileri ile Apriori algoritması analiz yöntemlerinin kullanıldığı ilk çalışma olma özelliğine sahiptir.

Veri madenciliği sağlık alanında kararın desteklenmesinde, prognoz tahmin edilmesinde, pratik ilaç kullanım rehberinin oluşturulmasında, tıbbi yöntemlerin geliştirilmesinde, kanıta dayalı ilaç kullanımının özendirilmesinde ve sağlık alanında gelecekteki öngörülerini görme alanlarında önemle kullanılmaktadır (Zhang vd., 2016). Veri madenciliği çalışmalarında biyolojik verilerin birliktelik kurallarına göre ilişkilendirilmesinde kullanılan farklı teknikler içinde Apriori algoritmasının daha yüksek kalitede çözümler bulabildiği de saptanmıştır (Parraga-Alava vd., 2018). Bundan dolayı bu çalışmada hepatit hastalarına ait ve veri madenciliği yoluyla elde edilen verilerin birliktelik kurallarına göre analiz edilerek hepatit hastalığının teşhisine yönelik kurallar oluşturulmuştur. Apriori uygulaması için Weka Explorer firmasının geliştirdiği paket program kullanılmıştır.

Dolayısıyla uygulama aşamalarına geçmeden önce Weka Explorer Apriori algoritması programının tanıtımı yapılmıştır.

3.4. Weka Explorer Paket Programı

Machine Learning Software in Java tarafından geliştirilen Weka Explorer ücretsiz olarak bilgisayara kurulmuştur (<http://7>). Masaüstü modülü ile çalıştırılan programın kendi veri tabanında birçok dosyanın eğitim amaçlı yüklenmiş olduğu görülmektedir. Sadece “.arff” uzantılı dosyaları çalıştırabildiğinden öncelikle tüm veri dosyalarının dönüştürülmesi gerekmektedir.

Weka üzerinden istatistik analizlerin nasıl yapıldığı Weka eğitim serisinde detaylı bir şekilde anlatmıştır (http-8). Program çalışmaya başladığında Explorer modülü yardımıyla uygulama çalıştırılır. Explorer açıldığında “open file” modülü ile veri tabanında bulunan herhangi bir dosya açılabilir veya kendi dosyamız arff olarak sisteme yüklenebilir. Veri dosyası yüklendikten sonra program dosya üzerinde değişiklik yapılmasına izin verebilmektedir. Edit modülünden eksik veya hatalı veriler düzeltilebilmektedir. Hatta istenmesi durumunda aşırı uç veriler dahi düzeltilebilir. Bu programda ayrıca istenilen veri bilgileri de silinebilmektedir.

Associate paneli Weka Explorer paket programında tanımlı algoritmaların kullanılmasına imkân verir. Veri tabanında “associations” modülü altında yer alan algoritmalarından biri de Apriori algoritmasıdır. Apriori algoritması bu çalışmada kullanıldığından özellikleri ve kullanım yöntemleri bulgular kısmında anlatılmıştır.

“Select attributes” modülü veri kümesinde seçme ve işleme özelliklerini ayarlamaya yarayan bir paneldir. Gerekli ayarlar yapıldıktan sonra veriler dönüştürülebilir ve bu veriler görselleştirme ekranında görülebilir. Veriler ayrıca grafik şeklinde de görüntülenebilir.

3.5. Apriori Algoritmasının Hepatit Verilerine Uygulanması

Uygulamada kullanılan hepatit ile ilgili veriler Dua ve Graff (2019) tarafından yayınlanmıştır. Veriler birliktelik kuralına göre Apriori Algoritması yöntemi kullanılarak analiz edilmiştir. Dua ve Graff (2019) tarafından Carnegie-Mellon Üniversitesi, Jozef Stefan Institute’den elde edilen ve hepatit hastalarına ait veriler aşağıdaki özellikleri içermektedir:

Attribute informations:

1. CLASS (SINIF): DIE, LIVE
2. AGE (YAŞ): 10, 20, 30, 40, 50, 60, 70, 80
3. SEX (CİNSİYET): male, female
4. STEROID (HORMON): no, yes
5. ANTIVIRALS (ANTİVİRAL İLAÇLAR): no, yes
6. FATIGUE (YORGUNLUK): no, yes
7. MALAISE (KIRGINLIK): no, yes
8. ANOREXIA (PSİKOLOJİK SORUN): no, yes
9. LIVER BIG (KARACİĞER BÜYÜMESİ): no, yes

10. LIVER FIRM (KARACİĞERDE KİTLE): no, yes
11. SPLEEN PALPABLE (DALAK BÜYÜMESİ): no, yes
12. SPIDERS (SİROZ): no, yes
13. ASCITES (ASİTLENME): no, yes
14. VARICES (VARİSLENME): no, yes
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
17. SGOT: 13, 100, 200, 300, 400, 500,
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. HISTOLOGY (BİYOPSİ): no, yes

Veri madenciliği yoluyla elde edilen verilerde toplam 155 kişinin özelliklerinin tamamının numerik değerler şeklinde olduğu görülmüştür. Dolayısıyla yukarıda görüldüğü gibi nominal değer olan verilerin tamamı numerik değer şeklinde sunulmuştur (Ek-1). Nominal değerlerde “NO, 1” ve “YES, 2” olarak belirtilmiştir. İlk sütunda görülen Class değeri ise “DIE, 1” ve “LIVE, 2” şeklindedir. “AGE” satırı hariç, “CLAS” satırından itibaren “VARICES” satırına kadar olan ve en son “HISTORY” satırı olan 1-14 ve 20 numaralı veri satırları nitel değişkenler oldukları için bilgi amaçlı olarak bu verilerin frekans tabloları hazırlanmış ve çizelge şeklinde sunulmuştur (bakınız BULGULAR VE YORUM, Çizelge 4.1 ve 4.2, Şekil 4.1 ve 4.2). Daha önce açıklandığı gibi eksik veriler birçok yöntem kullanılarak giderilebilmektedir.

Kayıp verilerin rassal olup olmadığını test etmek için Little (1988) tarafından geliştirilen MCAR testi kullanılmıştır (Little’s MCAR test). Test, kayıp veriye sahip alt grupların aritmetik ortalamaları ile genel ortalama arasındaki farkın standartlaştırılması ve ağırlıklı ortalamaların toplanması esasına dayanır. Şayet alt grupların aritmetik ortalamaları genel ortalamanın örnekleme hatasına ilişkin aralıkta yer alırsa kayıp veri dağılımı rastgele (rassal) olarak kabul edilmektedir (Akbaş ve Tavşancıl, 2015).

Bu çalışmada tüm verilerin Little's MCAR test yöntemi kullanılarak analiz edilmesi sonucunda tamamen rassal dağılım göstermedikleri ve dolayısıyla değişkenlerin birbirlerini etkiledikleri saptanmıştır ($P=0,021$) (Çizelge 3.2). Testin güven değeri %95 olarak kabul edilmiştir. P değeri %5’ten küçük olduğu için ($p < 0,05$) eksik verilerin dağılımının tamamen rastlantısal olmadıkları kabul edilmiştir (3.7). Bu nedenle bu çalışmada eksik veriler R yazılımında, mice 2.9 paketi kullanılarak doldurulmuştur. Nitel

değerler logreg (mice.impute.logreg) tekniğine göre doldurulurken nicel değerler pmm (mice.impute.pmm) tekniğine göre tamamlanmıştır.

$$\chi^2 = 374,851, p = 0,021 > \alpha = 0,05 \quad (3.7)$$

Çizelge 3.2. Verilerin rassal dağılımının saptanması amacıyla uygulanan Little's MCAR test sonucu

EM Means^a			
YAŞ	BİLİRUBİN	ALK FOSFAT	SGOT
41,200	1,430	105,195	86,301
a. Little's MCAR test: Chi-Square = 374,851, DF = 321, Sig. = ,021			
EM Means devam			
ALBUMİN	PROTROMBİN TESTİ	SINIF	CİNSİYET
3,819	63,024	1,794	1,103
EM Means devam			
STEROİD	ANTİVİRAL İLAÇLAR	YORGUNLUK	KIRGINLIK
1,507	1,845	1,351	1,604
EM Means devam			
ANOREKSİA	KARACİĞER BÜYÜMESİ	KARACİĞERDE KİTLE	VARİSLER
1,792	1,829	1,590	1,883
EM Means devam			
SİRÖZ	ASİTLENME	DALAK BÜYÜMESİ	BİYOPSİ
1,659	1,868	1,803	1,452

Apriori algoritması anılan özellikleri birlikte gösteren bireyleri dikkate alarak analiz yapıldığından veriler tamamlandıktan sonra analiz değerleri şeklinde verilen age, bilirubin, alk phosphate, SGOT, albümine ve protime (pıhtılaşma zamanı) verilerinin değerleri ortalamanın altında (1) ve ortalamaya eşit veya ortalamanın üstünde (2) şeklinde değiştirilmiştir (Köroğlu, 2016). Bu uygulama “edit” modülü sayesinde verileri değiştirme imkânı sunmaktadır fakat tüm verilerin tek tek düzeltilmesi zorunluluğu bulunmaktadır. Bundan dolayı yukarıda bahsedilen değişiklikler Excel tablosunda yapıldıktan sonra verilerin “.arff” şekline dönüştürülmesi yapılmıştır.

Kan analizi sonucu elde edilen verilerin normal sınır değerleri bilgi amaçlı olarak Çizelge 3.3’te verilmiştir. Fakat referans değerlerinin aslında analizin yapıldığı laboratuvara ve laboratuvarında kullanılan cihazlara bağlı olarak değişebileceğini unutmamak gerekir. Alkalen fosfat ve SGOT gibi bazı kan değerleri gelişim yaşına ve cinsiyete bağlı olarak da değişebilmektedir (http-9).

Çizelge 3.3. Kan parametrelerinin normal değer aralıkları

Kan parametreleri	Yaş	Normal değer aralığı
BiLiRUBİN	5 gün ve üstü yaşlar	0,3 – 1,2 mg/dL
ALKALEN FOSFAT	7-9 yaş	Erkek 86 – 315 U/L Kadın 69 – 325 U/L
	10-12 yaş	Erkek 42 – 362 U/L Kadın 51 – 332 U/L
	13-15 yaş	Erkek 74 – 390 U/L Kadın 50 – 162 U/L
	16-18 yaş	Erkek 52 – 171 U/L Kadın 47 – 119 U/L
	19 yaş ve üstü	30 – 120 U/L
SGOT	0-1 yaş	25 – 75 U/L
	2-19 yaş	15 – 60 U/L
	20 yaş ve üstü	Erkek 0 – 35 U/L Kadın 0 – 31 U/L
ALBUMİN	0-14 yaş	3,8 – 5,4 g/dL
	20-59	3,5 – 5,2 g/dL
	60 yaş ve üstü	3,2 – 4,6 g/dL
PROTROMBİN TESTİ*	-	-

*Protrombin testi kan pıhtılaşma süresini belirler. Parmak lanset ile delindiğinde en fazla 13 saniyede kanamanın durması gerekir. Bazı testlerde kan pıhtılaşmasına sebep olan ve kan serumunda bulunan proteinlerin miktarı ölçülerek pıhtılaşma zamanı hesaplanır. Dolayısıyla bu parametre için laboratuvar koşullarına göre farklı hesaplamalar yapılabilmektedir.

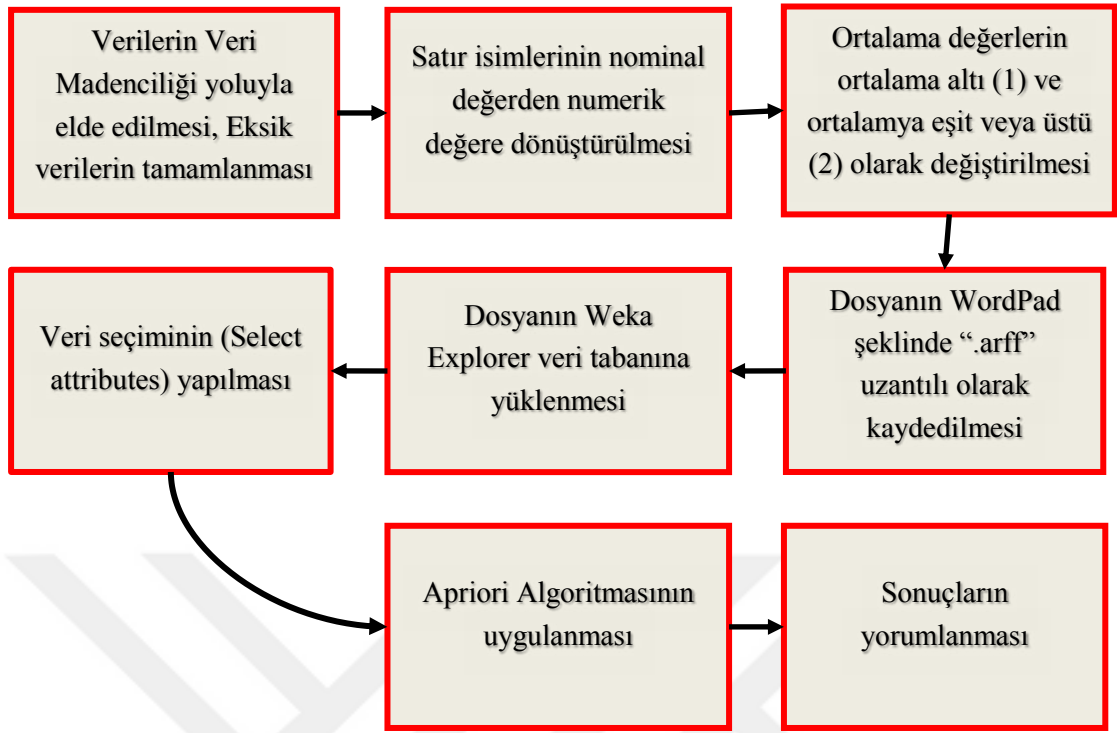
Eksik verilerin özelliklere göre frekansları incelendiğinde en fazla eksik verinin %43,22 oranında kan pıhtılaşma zamanının belirlenmesi özelliğinde (protime) olduğu görülmektedir (Çizelge 3.4). %18,70 oranında eksik veri içeren alkalen fosfat özelliğini %10,32 oranında eksik veri içeren albümin özeliği takip etmiştir. Hasta ifadesine bağlı olan özelliklerde eksik veri ya yoktur veya %0,64 gibi bir orana sahiptir. Fakat, özellikle hekimin fiziki muayenesi, kan tetkikleri ve görüntüleme tetkikleri sonucu ortaya konulan özelliklerde eksik veri frekansları dikkat çekmektedir. Bu durum hekimin tetkik istemi sırasında hepatit hastalığı olasılığından şüphelenmediğini ve dolayısıyla hepatit hastalığını tanımlayıcı tetkikleri isyemediğini göstermektedir.

Çizelge 3.4. Eksik verilerin özelliklere göre frekans değerleri

ÖZELLİKLER	Eksik veri frekansı (f)*	%
CLAS (Sınıf)	0	0
AGE (Yaş)	0	0
SEX (Cinsiyet)	0	0
STEROID (Hormon)	1	0,64
ANTIVIRALS (Antiviral ilaçlar)	0	0
FATIGUE (Yorgunluk)	1	0,64
MALAISE (Kırgınlık)	1	0,64
ANOREXIA (Psikolojik sorun)	1	0,64
LIVER BIG (Karaciğer büyümesi)	10	6,45
LIVER FIRM (Karaciğerde kitle)	11	7,09
SPLEEN PALPABLE (Dalak büyümesi)	5	3,22
SPIDERS (Siroz)	5	3,22
ASCITES (Asitlenme)	5	3,22
VARICES (Varisler)	5	3,22
BILIRUBIN	6	3,87
ALKALEN PHOSPHATE	29	18,70
SGOT	4	2,58
ALBUMIN	16	10,32
PROTIME(Protrombin testi)	67	43,22
HISTOLOGY (Doku bilimi, Biyopsi)	0	0

*Toplam 155 veri

Apriori algoritması uygulanmadan önce hepatit veri setinde aşağıdaki işlemler uygulanmış ve dosya Weka analizi için uygun hale getirilmiştir (Şekil 3.4). Her ne kadar satır özellikleri “die, live, male, female, no, yes” şeklinde verilmiş olsa da Ek-1’de görüldüğü gibi değerlerin tamamı numerik veriler olarak sunulmuştur. Bundan dolayı tabloda ilgili sütundaki verilerin nominal değerlere dönüştürülmesi daha zahmetli ve zaman alıcı işlem olduğundan satır özelliklerinin “1, 2” gibi numerik değerlere dönüştürülmesi tercih edilmiştir. Yukarıda belirtildiği gibi nominal değerlerde “NO, DIE, MALE” değerleri “1” ve “YES, LIVE, FEMALE” değerleri de “2” olarak belirtilmiştir.



Şekil 3.4. Apriori algoritması uygulama işlemleri

Bundan sonraki işlem ise dosyanın Weka veri tabanına yüklenmesi işlemidir. Weka sadece “.arff” uzantılı dosyaları kabul etmektedir. Dolayısıyla verilerin öncelikle Weka veri tabanının kabul edeceği ARFF (Attribute Relationship File Format) uzantılı dosya şekline dönüştürülmesi gerekir.

Kısaca verilere Apriori algoritması uygulanmadan önce aşağıdaki işlemler sırası ile yapılmıştır:

- 1-Nominal değerler numerik değerlere kodlanmıştır
 - 2-Eksik veriler tamamlanmıştır
 - 3-Ortalama değer şeklinde verilen verilerin değerleri ortalama altı ve ortalamaya eşit veya üstü olarak değiştirilmiştir
 - 4- Dosya “.arff” şekline dönüştürülmüştür
 - 5- Bazı özellikler silinmiştir
 - 6-Dosya Weka Explorer veritabanına yüklenerek Apriori algoritması uygulanmıştır
- Hepatit veri seti aşağıdaki gibi WordPad şeklinde kaydedilmiş ve işlemler tarif edildiği gibi yapılmıştır:

```

@relation data2
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-
last-weka.filters.unsupervised.attribute.StringToNominal-Rlast

@attribute AGE {1,2}
@attribute STEROID {1,2}
@attribute ANTIVIRALS {1,2}
@attribute FATIGUE {1,2}
@attribute MALAISE {1,2}
@attribute ANOREXIA {1,2}
@attribute LIVER_BIG {1,2}
@attribute LIVER_FIRM {1,2}
@attribute SPLEEN_PALPABLE {1,2}
@attribute SPIDERS {1,2}
@attribute ASCITES {1,2}
@attribute VARICES {1,2}
@attribute BILIRUBIN {1,2}
@attribute ALK_PHOSPHATE {1,2}
@attribute SGOT {1,2}
@attribute ALBUMIN {1,2}
@attribute PROTIME {1,2}

@data
1,1,2,2,2,2,1,2,2,2,2,2,1,1,1,2,2
2,1,2,1,2,2,1,2,2,2,2,2,1,2,1,1,2
2,2,2,1,2,2,2,2,2,2,2,2,1,1,1,2,1
1,2,1,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
2,1,2,1,2,1,2,2,1,1,2,2,2,2,1,1,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,1,2,2,2,1,2,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,2,2,2,1

```

“@relation” bölümünde dosya adı (*data2*) yazılmıştır.

“@attribute” sonrasında ise özellikler ve veri tipleri yazılmıştır. Özellikler “AGE” ile başlamış ve “PROTIME” ile sonlanmış, toplam 17 özellik olarak yazılmıştır. Özellik adlarından sonra parantez içinde veri tipleri numerik değer olarak kaydedilmiştir.

“@data” sonrasında ise tüm veriler aralarında virgöl olacak şekilde ve hiç aralık bırakmadan her satır için aşağıya doğru yazılmıştır. Yukarıda görülen “.arff” uzantılı dosya örneğinde ilk 10 kişinin verileri örnek olarak gösterilmiştir. Data2.arff olarak kaydedilen dosya Weka veri tabanına yüklenerek istatistiki analizler yapılmış ve sonuçlar kaydedilmiştir.



4. BULGULAR VE YORUM

İstatistiksel analizlere başlamadan önce veri seçimi (select attributes) yapılmıştır. CLASS, SEX ve HISTOLOGY veri değerleri silinmiştir. Çünkü hepatit cinsiyete bağlı olmadan tüm canlıları etkileyebilen viral bir hastalıktır. Ayrıca canlının hastalık belirtilerini gösterebilmesi için hayatta kalması gerekir. Bununla birlikte HISTOLOGY sütununda bireylerin biyopsi yapıp yapmadığı sorulmuştur. Biyopsi hastalığın son aşamasında yapılan ve hasar durumunu belirlemeye yarayan girişimsel bir işlemdir. Hastaların yarısından fazlasının (%54,83) biyopsi yapmadıkları görülmektedir. Bundan dolayı anılan 3 veri analiz tablosundan çıkarılmışlardır. Geriye kalan 17 özellik Türkçeleştirilerek çalışmaya dâhil edilmiş ve birliktelik kuralları çalışılmıştır (Ek-2). Bireylere ait frekans tabloları ile eksik veriler tamamlanmadan önce ve tamamlandıktan sonra ortalama değerlerin sonuçları karşılaştırmalı olarak Çizelge 4.1 ve Çizelge 4.2’de görülmektedir.

Çalışmaya dâhil edilen veri setlerinin frekans grafikleri Şekil 4.1 ve Şekil 4.2’de sunulmuştur. Grafikler ortalama değerin altında ve ortalama değere eşit veya üstünde olan verileri ve bu verilerde kaç kişinin bulunduğunu göstermektedir. Grafik incelendiğinde hastalığın her yaş grubunu etkileyebildiği görülmektedir. Ortalama yaş $41,20 \pm 1,00$ olarak saptanmıştır. 86 kişinin yaşı ortalamanın altında (%55,48) iken 69 kişinin yaşı da ortalamaya eşit veya ortalamanın üstünde olarak belirlenmiştir. Her yaş grubundan insanların hepatite yakalanma riski eşittir. Fakat hepatit hastalığının bulgularının ortaya çıkması ve hastalık etkilerinin görülmesi zaman içinde ortaya çıkmaktadır. Fakat yaş ile birlikte viral enfeksiyonlara maruz kalma veya hepatit hastalığına neden olan diğer ajanlar ile karşılaşma olasılığı arttığından ileri yaşlarda hepatit olma olasılığı artmaktadır. Bu çalışmada 7 yaşında bir kişinin hepatite yakalandığı görülmektedir. Dolayısıyla hepatitin en çok hangi yaş grubunu etkileyebileceği konusunda yorum yapılması mümkün değildir.

Kan analizi sonucu bilirubin değerinin ortalama $1,42 \pm 0,09$ olduğu ve 110 kişinin bilirubin değerinin ortalama altında olduğu görülmüştür. Toplam 109 kişinin alkalen fosfataz değerleri ortalamanın ($105,32 \pm 3,72$) altında olduğu, SGOT değerleri de ortalamanın ($85,89 \pm 7,10$) altında olan 98 kişi saptanmıştır. Albumin ve protime (pıhtılaşma zamanı) ortalama değerleri sırasıyla $3,81 \pm 0,04$ ve $61,85 \pm 1,38$ olarak hesaplanırken 100 kişinin albümin değerlerinin ve 71 kişinin de pıhtılaşma zamanı

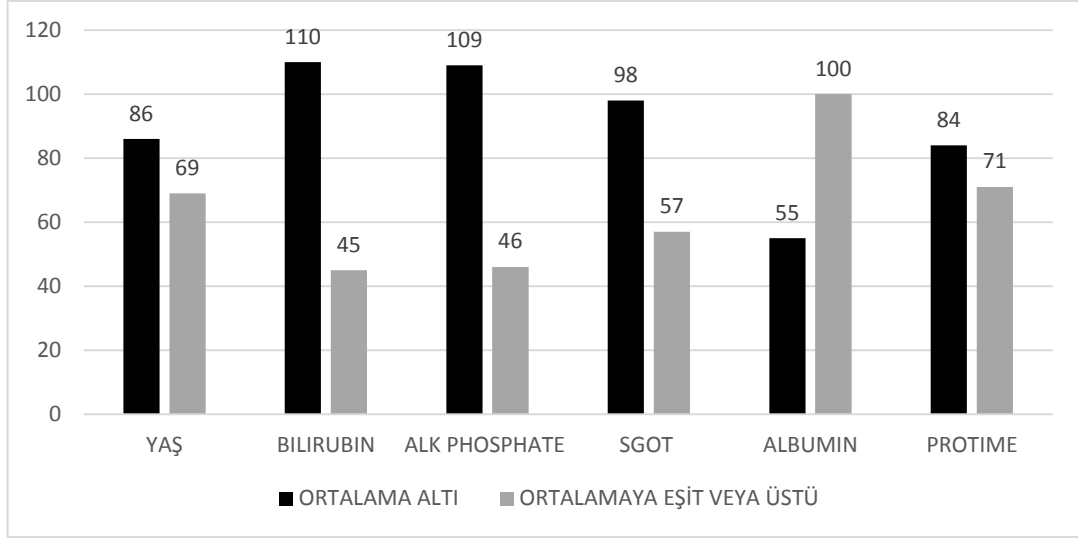
değerlerinin ortalamaya eşit veya ortalamanın üstünde olduğu saptanmıştır (Çizelge 4.1 ve Şekil 4.1).

Eksik verilerin silinmesi veya nitel verilerin mice paketinde yer alan mice.impute.logreg ve nicel verilerin mice.impute.pmm fonksiyonları yardımı ile tamamlanması işlemlerinin ortalamayı etkileyip etkilemediği incelenmiştir (Çizelge 4.1). Eksik veriler tamamlanmadan önce ve tamamlandıktan sonra ortalamaları alınmış ve t-testi ile birbirleriyle karşılaştırılmıştır. Eksik veriler silindikten sonra veya tamamlandıktan sonra ortalamanın istatistiksel anlamda önemli derecede değiştiği görülmüştür. Testin güven değeri 0,95 ve hata payı ise 0,05 olarak kabul edilmiştir. P değeri 0,05'ten büyük olduğu için ($p>0,05$) eksik veriler tamamlanmadan önce ve eksik veriler tamamlandıktan sonra ortalamaları arasında farkın önemli olmadığı ve dolayısıyla eksik veriler tamalandıktan sonra ortalamanın anlamlı bir şekilde değişmediği sonucuna varılmıştır.

Çizelge 4.1. Ortalama değerler ve standart hata

Değişkenler	İşlem yapılan	Veri sayısı*	%	Ortalama± SE*	Veri sayısı**	Ortalama± SE**	P değeri
Yaş	Cevap veren	155	100,0	41,20±1,00	155	41,200±1,00	1,00
	Eksik veri	-	-	-	-	-	-
Bilirubin	Analizi yapılan	149	96,12	1,42±0,09	155	1,49±0,10	0,51
	Eksik veri	6	3,87	-	-	-	-
Alkalen fosfat	Analizi yapılan	126	81,29	105,32±3,72	155	106,20±4,00	0,81
	Eksik veri	29	18,70	-	-	-	-
SGOT	Analizi yapılan	151	97,41	85,89±7,10	155	85,00±7,12	0,90
	Eksik veri	4	2,58	-	-	-	-
Albumin	Analizi yapılan	139	89,67	3,81±0,04	155	3,78±0,05	0,45
	Eksik veri	16	10,32	-	-	-	-
Protrombin testi	Analizi yapılan	88	56,77	61,85±1,38	155	59,15±1,79	0,99
	Eksik veri	67	43,22	-	-	-	-

*Eksik veri silindikten sonra ortalama değerler; **Eksik veriler R-mice pmm programına göre düzeltildikten sonra ortalama değer



Şekil 4.1. Ortalama değere sahip verilerin frekanslarının karşılaştırılması

Çizelge 4.2, nitel verilerde eksik veriler tamalanmadan önce ve tamamlandıktan sonra frekans değerlerini göstermektedir. Araştırmaya dâhil edilen 76 birey steroid kullanmadığını, 79 birey ise steroid kullandığını beyan etmiştir. Frekans değeri sırasıyla %49,03 ve %50,96 olarak hesaplanmıştır. Bazı steroid tipi hormonlar kas yapımı amacıyla fazla miktarlarda kullanıldığından karaciğerde tahribata neden olarak hepatit oluşturabilmektedir. Antiviral ilaçları kullananların 131 kişi olduğu, yorgunluk hissedenlerin sayısının 54 olduğu saptanmıştır. Katılımcılardan 93 bireyin kırgınlık hissettiği, anoreksia tip psikolojik rahatsızlık gösterenlerin 123 kişi oldukları, karaciğer büyümesi ve karaciğerde kitle görülenlerin sayısının da sırasıyla 129 ve 92 kişi oldukları saptanmıştır. Aynı şekilde toplam 102 kişide siroz, 133 kişide asitlenme ve 137 kişide de varis oluşumu gözlenmiştir (Çizelge 4.2 ve Şekil 4.2).

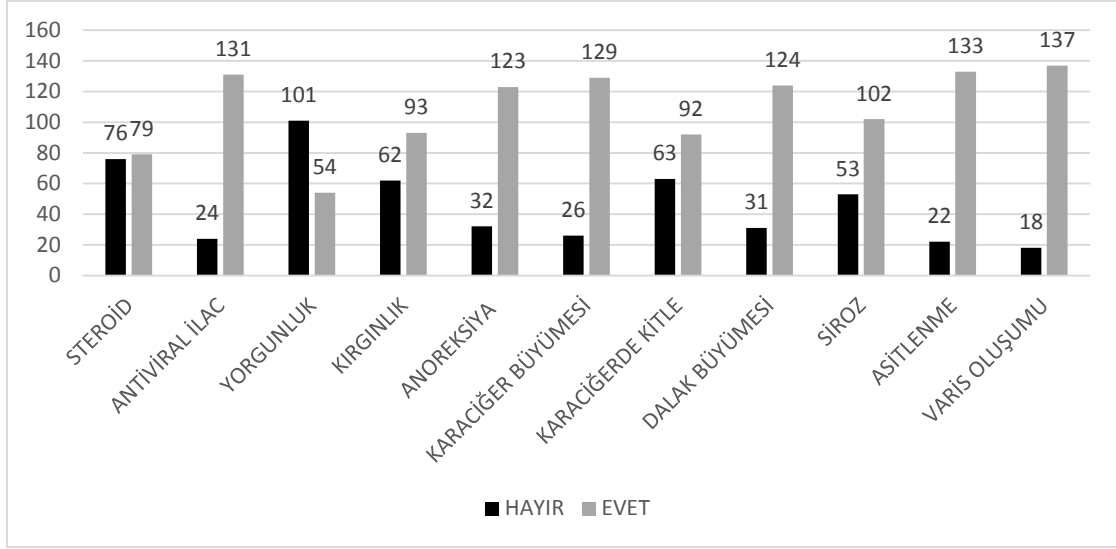
Karaciğer büyümesi, kitle oluşumu ve buna bağlı olarak karaciğer yetmezliğine sebep olan siroz gelişimi, dalak büyümesi ve karaciğer yetmezliğinin ileri aşamalarındaki belirtiler olan varislenme ve karın boşluğunda asitlenme bulgularını gösteren bireylerin frekansı oldukça yüksek olarak belirlenmiştir. Dolayısıyla antiviral ilaç kullananların frekansı da yüksek olarak saptanmıştır. Bu hastalığın başlangıç aşamasında görülen halsizlik gibi belirtileri gösterenlerin frekansının az olduğu fakat kırgınlık belirtisi olanların frekansının da nispeten yüksek olduğu görülmektedir. Dolayısıyla bu hastalığın belirtilerinin başında kırgınlık hissedildiği, ilerleyen aşamalarda diğer belirtilerin ortaya çıktığı söylenebilir. Anorexia denilen bir çeşit psikolojik sorunlar her ne kadar hepatit ile direkt ilişkili olmasa da hepatit hastalığının hastalar üzerinde oluşturduğu olumsuz yaşam

standardı ve ölüm korkusu sonucu gelişen depresif bir rahatsızlıktır. Bu çalışmada hastaların büyük bir kısmının (%79,35) Anoreksiya sıkıntısı yaşadıkları görülmektedir.

Çizelge 4.2. Nitel değişkenlerin karşılaştırmalı frekans değerleri

ÖZELLİK	HASTANIN CEVABI	Eksik veri tamamlanmadan önce		Eksik veri tamamlandıktan sonra	
		f*	%	f	%
Hormon (Steroid)	Hayır (1)	76	49,03	76	49,03
	Evet (2)	78	50,32	79	50,96
	Eksik	1	0,64	-	-
Antiviral ilaçlar (Antivirals)	Hayır (1)	24	15,48	24	15,48
	Evet (2)	131	84,51	131	84,51
	Eksik	-	-	-	-
Yorgunluk (Fatigue)	Hayır (1)	100	64,51	101	65,16
	Evet (2)	54	34,83	54	34,83
	Eksik	1	0,64	-	-
Kırgınlık (Malaise)	Hayır (1)	61	39,35	62	40,00
	Evet (2)	93	60,00	93	60,00
	Eksik	1	0,64	-	-
Psikolojik sorun (Anorexia)	Hayır (1)	32	20,64	32	20,64
	Evet (2)	122	78,70	123	79,35
	Eksik	1	0,64	-	-
Karaciğer büyümesi (Liver big)	Hayır (1)	25	16,12	26	16,77
	Evet (2)	120	77,41	129	83,22
	Eksik	10	6,45	-	-
Karaciğerde kitle (Liver firm)	Hayır (1)	60	38,71	63	40,64
	Evet (2)	84	54,19	92	59,35
	Eksik	11	7,09	-	-
Dalak büyümesi (Splee palpable)	Hayır (1)	30	19,35	31	20,00
	Evet (2)	120	77,41	124	80,00
	Eksik	5	3,22	-	-
Siroz (Spiders)	Hayır (1)	51	32,90	53	34,19
	Evet (2)	99	63,87	102	65,80
	Eksik	5	3,22	-	-
Asitlenme (Ascites)	Hayır (1)	20	12,90	22	14,19
	Evet (2)	130	83,87	133	85,80
	Eksik	5	3,22	-	-
Varisler (Varices)	Hayır (1)	18	11,61	18	11,61
	Evet (2)	132	85,16	137	88,38
	Eksik	5	3,22	-	-

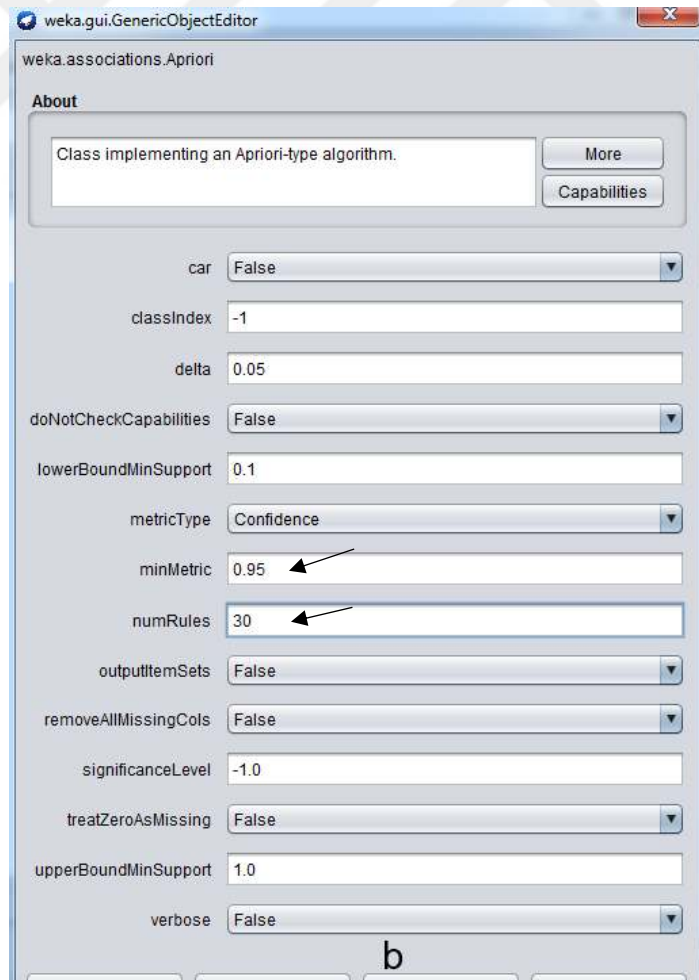
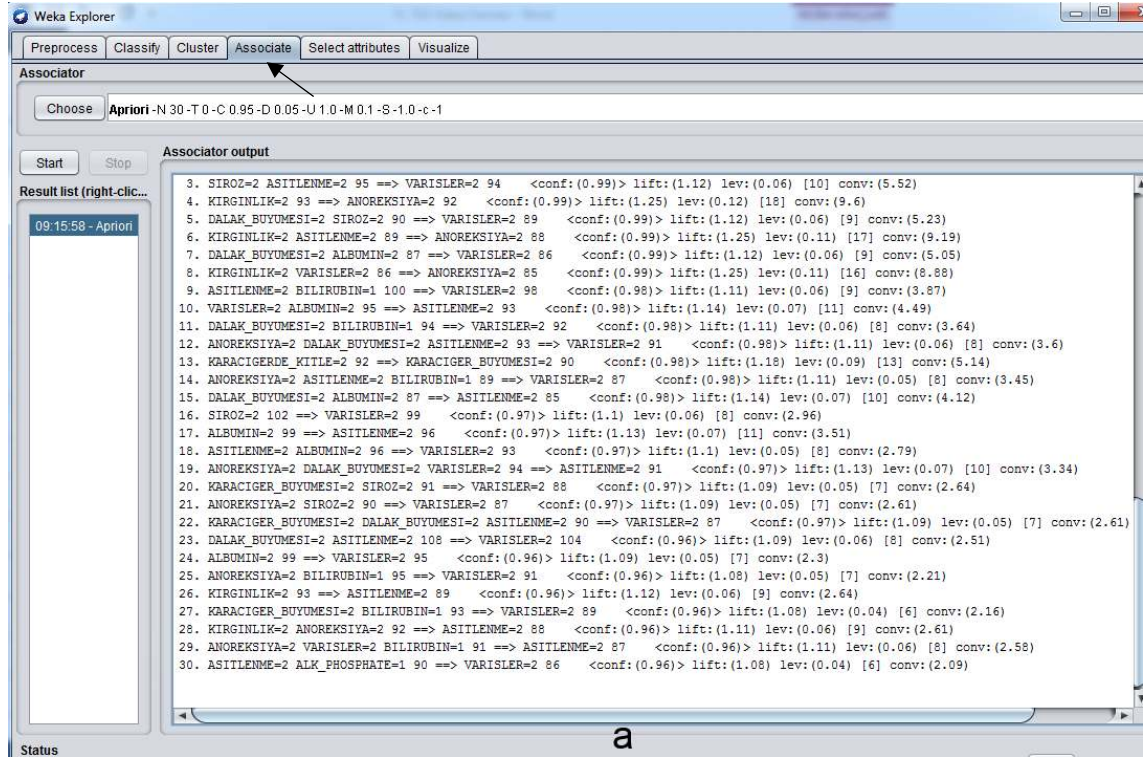
*Toplam 155 hasta



Şekil 4.2. Nitel değere sahip verilerin frekanslarının karşılaştırılması

Weka paket programında dosya yüklendikten sonra “associate” modülünden “Apriori” modülü açılmıştır. Analize başlamadan önce “minMetric” ve “numRules” ayarları yapılmıştır (Şekil 4.3a ve b).

Bekki (2016) nesne kümesindeki nesne sayısının artmasıyla ilişki kuralı sayısının da katlanarak artacağını, dolayısıyla oluşturulabilecek tüm kurallar içinden işe yarayacak bilgiyi açıkça ortaya koyan kuralları belirlemek amacıyla kural sayısında sadeleştirme yoluna gidilebileceğini belirtmiştir. Ayrıca bu sadeleştirme aşamalarında da en çok destek ve güven eşik değerlerinin ön planda olacağını da belirtmiştir. Bu çalışmada da en iyi kuralları elde etmek amacıyla önem derecesi alt sınırı (minumum metrik puan, minMetric) 0,95 olarak ayarlanmıştır. Önem derecesi alt sınırı düşürüldüğünde kural sayısı çok fazla olmakta ve dolayısıyla uygun verilerin seçilmesinde karışıklık oluşmaktadır. Bu önem derecesinde bulunacak kural sayısı da (numRules) 30 olarak belirlenmiştir. Elde edilen en iyi 30 kurala ait sonuçlar Çizelge 4.3’te verilmiştir. Kural sayısı numRules ayarlarından istenildiği kadar çoğaltılabilir veya azaltılabilir.



Şekil 4.3. a) Apriori algoritması, b) minMetric ve numRules ayarları

Çizelge 4.3. Apriori algoritması sonucu elde edilen kurallar

Minimum support: 0.55 (85 instances)		
Minimum metric <confidence>: 0.95		
Number of cycles performed: 9		
Generated sets of large itemsets:		
Size of set of large itemsets L(1): 15		
Size of set of large itemsets L(2): 47		
Size of set of large itemsets L(3): 38		
Size of set of large itemsets L(4): 7		
Bulunan en iyi kurallar:		
1.	DALAK BÜYÜMESİ=2 ASİTLENME=2 BİLİRUBİN=1 87 ==> VARİSLER=2 87	<conf:(1)> lift:(1.13) lev:(0.07) [10] conv:(10.1)
2.	DALAK BÜYÜMESİ=2 SİROZ=2 ASİTLENME=2 85 ==> VARİSLER=2 85	<conf:(1)> lift:(1.13) lev:(0.06) [9] conv:(9.87)
3.	SİROZ=2 ASİTLENME=2 95 ==> VARİSLER=2 94	<conf:(0.99)> lift:(1.12) lev:(0.06) [10] conv:(5.52)
4.	KIRGINLIK=2 93 ==> ANOREKSİYA=2 92	<conf:(0.99)> lift:(1.25) lev:(0.12) [18] conv:(9.6)
5.	DALAK BÜYÜMESİ=2 SİROZ=2 90 ==> VARİSLER=2 89	<conf:(0.99)> lift:(1.12) lev:(0.06) [9] conv:(5.23)
6.	KIRGINLIK=2 ASİTLENME=2 89 ==> ANOREKSİYA=2 88	<conf:(0.99)> lift:(1.25) lev:(0.11) [17] conv:(9.19)
7.	DALAK BÜYÜMESİ=2 ALBUMİN=2 87 ==> VARİSLER=2 86	<conf:(0.99)> lift:(1.12) lev:(0.06) [9] conv:(5.05)
8.	KIRGINLIK=2 VARİSLER=2 86 ==> ANOREKSİYA=2 85	<conf:(0.99)> lift:(1.25) lev:(0.11) [16] conv:(8.88)
9.	ASİTLENME=2 BİLİRUBİN=1 100 ==> VARİSLER=2 98	<conf:(0.98)> lift:(1.11) lev:(0.06) [9] conv:(3.87)
10.	VARİSLER=2 ALBUMİN=2 95 ==> ASİTLENME=2 93	<conf:(0.98)> lift:(1.14) lev:(0.07) [11] conv:(4.49)
11.	DALAK BÜYÜMESİ=2 BİLİRUBİN=1 94 ==> VARİSLER=2 92	<conf:(0.98)> lift:(1.11) lev:(0.06) [8] conv:(3.64)
12.	ANOREKSİYA=2 DALAK BÜYÜMESİ=2 ASİTLENME=2 93 ==> VARİSLER=2 91	<conf:(0.98)> lift:(1.11) lev:(0.06) [8] conv:(3.6)
13.	KARACİĞERDE KİTLE=2 92 ==> KARACİĞER BÜYÜMESİ=2 90	<conf:(0.98)> lift:(1.18) lev:(0.09) [13] conv:(5.14)
14.	ANOREKSİYA=2 ASİTLENME=2 BİLİRUBİN=1 89 ==> VARİSLER=2 87	<conf:(0.98)> lift:(1.11) lev:(0.05) [8] conv:(3.45)
15.	DALAK BÜYÜMESİ=2 ALBUMİN=2 87 ==> ASİTLENME=2 85	<conf:(0.98)> lift:(1.14) lev:(0.07) [10] conv:(4.12)
16.	SİROZ=2 102 ==> VARİSLER=2 99	<conf:(0.97)> lift:(1.1) lev:(0.06) [8] conv:(2.96)
17.	ALBUMİN=2 99 ==> ASİTLENME=2 96	<conf:(0.97)> lift:(1.13) lev:(0.07) [11] conv:(3.51)
18.	ASİTLENME=2 ALBUMİN=2 96 ==> VARİSLER=2 93	<conf:(0.97)> lift:(1.1) lev:(0.05) [8] conv:(2.79)
19.	ANOREKSİYA=2 DALAK BÜYÜMESİ=2 VARİSLER=2 94 ==> ASİTLENME=2 91	<conf:(0.97)> lift:(1.13) lev:(0.07) [10] conv:(3.34)
20.	KARACİĞER BÜYÜMESİ=2 SİROZ=2 91 ==> VARİSLER=2 88	<conf:(0.97)> lift:(1.09) lev:(0.05) [7] conv:(2.64)
21.	ANOREKSİYA=2 SİROZ=2 90 ==> VARİSLER=2 87	<conf:(0.97)> lift:(1.09) lev:(0.05) [7] conv:(2.61)
22.	KARACİĞER BÜYÜMESİ =2 DALAK BÜYÜMESİ=2 ASİTLENME=2 90 ==> VARİSLER=2 87	<conf:(0.97)> lift:(1.09) lev:(0.05) [7] conv:(2.61)

23.	DALAK BÜYÜMESİ=2 ASİTLENME=2 108 ==>	<conf:(0.96)> lift:(1.09) lev:(0.06) [8] conv:(2.51)
24.	ALBUMİN=2 99 ==> VARİSLER=2 95	<conf:(0.96)> lift:(1.09) lev:(0.05) [7] conv:(2.3)
25.	ANOREKSİYA=2 BİLİRUBİN=1 95 ==> VARİSLER=2 91	<conf:(0.96)> lift:(1.08) lev:(0.05) [7] conv:(2.21)
26.	KIRGINLIK=2 93 ==> ASİTLENME=2 89	<conf:(0.96)> lift:(1.12) lev:(0.06) [9] conv:(2.64)
27.	KARACİĞER BÜYÜMESİ =2 BİLİRUBİN=1 93 ==> VARİSLER=2 89	<conf:(0.96)> lift:(1.08) lev:(0.04) [6] conv:(2.16)
28.	KIRGINLIK=2 ANOREKSİYA=2 92 ==> ASİTLENME=2 88	<conf:(0.96)> lift:(1.11) lev:(0.06) [9] conv:(2.61)
29.	ANOREKSİYA=2 VARİSLER=2 BİLİRUBİN=1 91 ==> ASİTLENME=2 87	<conf:(0.96)> lift:(1.11) lev:(0.06) [8] conv:(2.58)
30.	ASİTLENME=2 ALK FOSFAT=1 90 ==> VARİSLER=2 86	<conf:(0.96)> lift:(1.08) lev:(0.04) [6] conv:(2.09)

Apriori algoritması uygulamasından sonra işlem gören verilerde destek değerinin 0,55 ve güven sınırının 0,95 olduğu görülmüştür. İlk iki kuralın güven değerinin 1 olduğu (conf:1), diğer kuralların ise 0,99 ile 0,96 arasında bir güven değerine sahip oldukları saptanmıştır. Çizelge 4.3'te saptanan kurallar şu şekilde açıklanabilir:

Bir numaralı kuralda, toplam 87 kişide hem dalak büyümesi, hem asitlenme hem de düşük bilirubin seviyesi görülmüştür. Bunların da ayrıca varis olma olasılığı da %100 oranındadır.

İki numaralı kuralda, dalak büyümesi görülen, siroz olan ve asitlenme gösteren kişilerin ayrıca varislere sahip olma olasılığı %100 demektir. Veri tablosunda bu bulguların tümünü gösterenlerin sayısı 85 kişidir.

Bu iki kural dışında yer alan 3-8 numaralı kuralların güven değerleri 0,99 olarak belirlenmiştir. Örneğin 3 numaralı kuralda yer alan verilere göre toplam 95 kişi siroz ve asitlenme özelliğine sahipken aynı zamanda bunlarda varis oluşumu da gözlenme ihtimali söz konusudur. Birliktelik kuralına göre bu bireylerde varis oluşumu olasılıkları da %99 olarak görülmektedir. Bu varis oluşumu bulgusu gösterenlerin sayısı ise 94 olarak sayılmıştır. Güven değeri 0,99 olan gruba bir başka örnek 7 numaralı kural verilebilir. Yedi numaralı kuralda 87 kişi karaciğer büyümesi ve yüksek albümin değerlerine sahiptir. Bunların ayrıca varis oluşumu bulgusuna sahip olma olasılıkları da %99 olarak hesaplanmıştır. Çalışmada verilen örnekte bu kuralda 86 kişide varis oluşumu gözlenmiştir.

Kural sayısı 9-15 arasında olanların ise güven değerleri 0.98 olarak hesaplanmıştır. Bu grupta yer alan kişilerin birliktelik kuralına göre bulguları gösterme olasılıkları %98'dir. Örneğin 12 numaralı kuralda anoreksia ve dalak büyümesi bulgularına sahip,

ayrıca asitlenme özelliği gösteren toplam 93 kişinin varislere de sahip olma olasılığı %98 olarak ön görülmüştür. Bu kuralda 91 kişi varislenme göstermektedir. 13 numaralı kuralda karaciğerinde kitle görülen 92 kişinin karaciğerlerinin de büyümüş olma olasılığı %98 olarak tahmin edilmiştir. Bulgularda karaciğerinde kitle görülen toplam 92 kişinin 90'ında karaciğer büyümesi gözlenmiştir.

Çizelge 4.3'te görülen 16-22 numaralı kuralların güven değeri %97'dir. Örneğin, 19 numaralı kuralda anoreksiye, dalak büyümesi ve varislenme bulgularının gösterenlerin aynı zamanda asitlenme özelliğini gösterme olasılığı %97 olarak ön görülmektedir. Anoreksiye, dalak büyümesi ve varislenme bulgularının hepsini gösteren 94 kişinin 91 kişisinde ayrıca asitlenme görülmüştür. Aynı şekilde 22 numaralı kuralda 90 kişi karaciğer büyümesi, dalak büyümesi ve asitlenme bulgularını gösterirken 87 kişi de varislenme bulgularını göstermektedir. Bunların birliktelik ihtimali %97 olarak hesaplanmıştır.

23-30 nolu kuralların güven değeri %96 olarak hesaplanmıştır. 23 nolu kuralda dalak büyümesi ve asitlenme bulgularını taşıyan 108 kişinin 104'ünde ayrıca varislenme şikayetleri de ortaya çıkmıştır. Birliktelik kuralına göre bu 3 bulgunun bir kişide beraber bulunma olasılığı %96 olarak hesaplanmıştır. 30 nolu kuralda ise asitlenme ve düşük alkali fosfat değerlerine sahip olanların ayrıca varislenme ihtimalleri hesaplanmıştır. Bu üç belirtinin beraber ortaya çıkma olasılığı %96 olarak ön görülmektedir.

Bu çalışmada kural sayısı 30 olarak ayarlanmıştır. Daha fazla kural görülmek istenmesi durumunda Şekil 4.3'de görüldüğü gibi kural sayısı (numRules) arttırılabilir.

5. SONUÇLAR VE TARTIŞMA

Veri madeninden elde edilen hepatit veri setine Apriori algoritması uygulanması sonucunda hepatit hastalarının gösterdiği bazı bulgular arasında ilişkilerin kurulabildiği görülmektedir. Algoritma veri tabanında güven sınırının 0,95 olarak verilmesi durumunda %95 ile %100 güven aralıklarında birliktelik kurallarının hesaplandığı görülmektedir. Elde edilen verilerin analiz edilmesinden sonra bulguların birlikte görülme olasılıkları hesaplanmıştır. Bu kurallar ile kişiden kişiye ilerleyişi ve bulguları farklı olabilen hepatit hastalığının teşhisinde doktorlara karar verme desteği sunabilmesi hedeflenmektedir.

Birliktelik kuralları algoritmalarından birisi olan Apriori algoritmasının günümüzde kullanılan algoritmalar içinde en kullanışlı algoritma olduğu, elde edilen sonucun tüm algoritmalar içinde en tutarlı sonuçların bu algoritma tarafından sunulduğu bilinmektedir (Duru, 2005; Köroğlu, 2016; Parraga-Alava vd., 2018). Dolayısıyla bu algoritmanın sağlık alanlarında uygulanması durumunda sonuçların daha sezgisel ve anlaşılabilir olarak sunulduğu ve karmaşıklığa sebep olan birçok kuralın sadeleştirildiği bildirilmiştir (Mahmoodi vd., 2016). Bu sebepten dolayı birçok araştırmada Apriori algoritmasının güvenle kullanılabileceği ve geniş veri kümelerinin analizinde iyi sonuçlar veren bir algoritma olduğu tavsiye edilmektedir (Maria Luna vd., 2018; Somek ve Hercigonja-Szekeres, 2017).

Bu çalışmada bazı semptomların birlikte ortaya çıkma olasılıklarının incelenmesiyle bireyin hepatit olup olmadığı şeklinde teşhisin ortaya konulması kapsamında hekimlere karar vermede yardımcı olabileceği sonucuna ulaşılmıştır. Hekimler hastanın şikâyetlerini aldıklarında ve fiziki muayenelerde hepatit belirtilerinin birini veya birkaçını saptadıklarında algoritmanın sonuçlarına göre hepatit olma olasılığını tahmin edebilecek ve daha ileri araştırmada hangi tetkikleri özellikle istemesi yönünde karar vermesine yardımcı olacaktır. Bu durum hekimlere hastalıkla ilgili nereye bakmaları gerektiğinin işaretlerini verecektir. Bu çalışmada hastanın ifadesine bağlı olan özelliklerde eksik verinin olmaması veya çok düşük frekanslarda olması, tam aksine özellikle hekimin fiziki muayenesi, kan tetkikleri ve görüntüleme tetkikleri sonucu ortaya konulan özelliklerde eksik veri frekanslarının fazla olması bu uygulamanın hekimlere hepatit hastalığının olma olasılığını hatırlatması ve ona göre tetkik istemini yönlendirmesi açısından oldukça faydalı olacaktır.

Apriori algoritmasının uygulandığı birçok çalışmada sağlık alanında kararın desteklenmesinde, prognozun tahmin edilmesinde, pratik ilaç kullanım rehberinin oluşturulmasında, tıbbi yöntemlerin geliştirilmesinde, kanıta dayalı ilaç kullanımının özendirilmesinde ve sağlık alanında gelecekteki öngörülerini görme alanlarında önemli bilgiler sunduğu ve karar vericilerin karar vermelerine yol gösterici olduğu belirlenmiştir (Yi vd., 2018; Wei vd., 2019; Xia vd., 2019). Başka araştırmacılar elde edilen bu sonuçlar bizim çalışmamızdan elde edilen sonuçlarımızı destekler niteliktedir. Bu uygulamanın bir başka önemi de prognozu benzer hastalıkları birbirlerinden ayırmada ve doğru teşhisinde aradaki çok küçük ayrıntıların fark edilmesini sağlamaktır. Zira Yıldırım vd. (2014) yayınladıkları makalelerinde birçok hastalığın onlarca tanısının aynı veya benzer olduğunu bildirmişlerdir. Bu husus birliktelik kurallarının karar vermedeki önemini göstermektedir. Çünkü hekimler hastaya, semptomlarının tıbbi öyküye, fiziki muayeneye ve bazı testlere dayanan geçici hipotezler oluşturma yoluyla tanı koymaktadırlar. Bununla birlikte, özellikle çok fazla bilgi içeren karmaşık bir hastalık durumunda klinik karar verme hataya sebep olabilir. Bu nedenle, veri madenciliği karar vermedeki hataları en aza indirmenin bir yolu olarak geliştirilmiştir ve verilerden değerli bilgilerin elde edilmesinde etkili bir yöntemdir (Chen ve Fawcett, 2016).

Bu çalışma karaciğer hastalarının semptomları dikkate alınarak gerçekleştirilmiştir. Semptomlardan biri de varislenmedir. Varis toplardamarların deri altında mavi renkte, genişlemiş ve kıvrımlar yapmış olarak görülmesidir. Birçok durumda varis ile karaciğer hastalığı arasında ilişki kurulmasında sorunlar yaşanabilmektedir. Özellikle ilerlemiş kronik karaciğer hastalıklarında varislenme önemlidir ve hastaların yaklaşık %14'ünde kanamaya sebep olarak ciddi sonuçlar doğurabilmektedir (Çelik vd., 2017).

Apriori algoritması birlikte hareket eden öğelerin birliktelik olasılıklarını hesaplayabilmektedir. Ortalama değer şeklinde sunulan verilerde neredeyse hiçbir veri bir diğeriyle aynı değeri taşımamaktadır. Yani destek değeri hesaplanırken toplam satır içinde beraber olan öğelerin sayısı önemlidir $D(A \Rightarrow B) = (A \text{ ve } B'nin \text{ bulunduğu satır sayısı}) / (\text{Toplam satır sayısı})$. Bu durumda birliktelik kuralları işlevsiz kalmaktadır. Bizim çalışmamızda age, bilirubin, alkalophosphate, SGOT, albümin ve protime gibi değerler ortalama değerler olduğundan analiz öncesi dönüştürülmüşlerdir. Değerler nominal değerler olarak “ortalama altı” veya “ortalama üstü” şeklinde yazılabileceği gibi (Köroğlu, 2016), bizim çalışmamızda yapıldığı gibi değerler numerik olarak “1 (ortalama altı)” veya “2 (ortalama üstü)”

şeklinde de yazılabilmektedirler. Bu çalışmada ortalama değerler numerik değer olarak girilmiş ve Apriori algoritmasında analiz işlemleri yapılmıştır.

Veri madenciliği hemen hemen her konuda kullanılmakla birlikte bazı zorluklarının ve sınırlamalarının olduğunu söylemek mümkündür. Househ ve Aldosari (2017) bu konuda yaptıkları derleme çalışmalarında sağlık hizmetlerinde tıbbi verilerin güvenilirliğini, sağlık kuruluşları arasında veri paylaşımını ve yanlış tahminlere yol açan uygunsuz modellemeyi içermeye olasılığına bağlı olarak veri madenciliğinin dikkatli analiz edilmesi gerektiğinin altını çizmişlerdir. Çalışmalarının devamında araştırmacılar, sağlık hizmetlerinde veri madenciliğinin kullanımında pek çok sorunların olduğunu ve sağlık hizmeti sağlayıcılarının, yöneticilerin ve politika yapıcıların sağlık hizmeti ile ilgili karar vermesini kolaylaştırmak amacıyla daha fazla çalışmanın gerekli olduğunu ve veri madenciliğinin geneline ilişkin daha fazla kanıtın bulunmasının zorunlu olduğu sonucuna varmışlardır.

Veri madenciliğinde hepatit veri setlerinin Apriori algoritması analiz yöntemlerinin kullanılması ile analizi sonucunda istatistiksel olarak hekimlere karar destek yardımına yönelik birçok kararın oluşturulduğu görülmüştür. Fakat veri madenciliğine birliktelik kurallarının kullanılmasında sayılan birçok faydası ile birlikte yukarıda bahsedilen kısıtlamaların ve olumsuzlukların göz ardı edilmemesi gereklidir. Bununla birlikte hekimlere tanı koyma, tetkik isteme, reçetelendirme aşamasında ve sonrasındaki karar verme aşamalarında yardım edebilecek bazı öneriler aşağıda sunulmuştur. Bu öneriler birliktelik kuralları analizinde ortaya çıkan sonuçlar doğrultusunda öne çıkmaktadır. Çünkü hastalıkla ilgili verilerin fazla olması en iyi sonucun alınmasını sağlayabilecektir. Özellikle birçok hastalığın belirtilerinin benzer olması aradaki çok küçük bir farkın bile göz ardı edilmemesi gerektiğinden algoritmanın önemini bir kez daha ortaya çıkarmaktadır. Karar verme algoritmalarının tüm ülkede ve mümkünse her poliklinikte hayata geçirilmesi durumunda hastanın kliniğe başvurduğu andan itibaren tedavisinin tamamlanması aşamasına kadar geçen süreçte tüm hekimler arasında karar birliği oluşturabileceği için üzerinde önemle durulması gerektiği düşünülmektedir.

Bu durum hastalıklarla ilgili hem geniş bir veri tabanının oluşmasını sağlayacak hem de tüm polikliniklerde karar birliği sağlayacaktır. Sonuçta hekimlerin olası hatalarının önüne geçilmiş olacaktır. Ayrıca maliyet etkinliği (cost effectivity) oluşturarak gereksiz tetkik istemini ve dolayısıyla devlet harcamalarını azaltabilecek, sadece gerekli tetkiklerin istenmesini mümkün kılacaktır.

Sapiens kitabının yazarı Prof. Yuval Noah Harari 2018 yılında “Davos Annual Meeting” toplantısında “Questioning Our Human Future” adlı seminerinde veriyi kontrol edenlerin sadece insanlığı değil, yaşamın kendisinin de geleceğini kontrol edebileceğini belirtmiştir. Harari seminerinde veriler sayesinde artık banka hesaplarının değil insanın kendisinin zarar gördüğünü belirtmiş ve bunun organizmaların her birisinin aslında birer biyokimyasal algoritma olmalarından kaynaklandığını söylemiştir (<http-10>).

Buradan çıkarım yapıldığında, istatistik bilimi ve dolayısıyla veri madenciliğinin geleceğin en önemli çalışma konularından birisi olacağını söylemek mümkündür.



KAYNAKÇA

- Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)* Santiago, Chile, pp 487-499. B.J. Bocca, M. Jarke, and C. Zaniolo (Eds).
- Agrawal, R., Imieliński, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD'93. pp. 207.* CiteSeerX 10.1.1.40.6984. doi:10.1145/170035.170072. ISBN 978-0897915922.
- Akbaş, U. ve Tavşancıl, E. (2015). Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde Ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 6(1), 38-57.
- Akçi, Y. (2016). İkinci el otomobil: Tüketici bakışıyla. *Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 8 (22), 329-362.
- Akdeniz F. (2015). New trends and developments. *Statistics Social Sciences Research Journal*, 4 (4), 1-11.
- Akdeniz F. ve Yıldırım F. (2011). Türkiyede Osmanlıdan günümüze istatistik tarihine bakış. *TÜİK, İstatistik Araştırma Dergisi*, 8 (3), 1-11.
- Alonso, S.G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D.C., Nozaleda, L.M., Franco, M. (2018). Data mining algorithms and techniques in mental health: A systematic review. *J Med Syst.*, 42 (9),161-176.
- Asilkan, Ö. ve Irmak, S. (2009). İkinci el otomobillerin gelecekteki fiyatlarının Yapay Sinir Ağları ile tahmin edilmesi. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 14 (2), 375-391.
- Atakan, C. ve Karabulut, İ. (2003). Derinliğe dayalı diskriminasyon. *S Ü Fen Ed. Fak. Fen Dergisi*, 22, 53- 63.
- Aydın, N., Kırdar, S., Uzun, N., Eyigör, M., Sayan, M. (2016). Atypical serological profiles in hepatitis B infections: investigation of S gene mutations in cases with

- concurrently positive for HBsAg and anti-HBs. *Mikrobiyol Bul.*, 50 (4), 535-543.
- Aytaç, M.B. ve Bilge, H.Ş. (2013). Tele pazarlama verilerinin birliktelik kurallarıyla ve Crisp-Dm yöntemiyle analiz edilmesi. *Aksaray Üniversitesi İİBF Dergisi*, 5 (2), 25-40.
- Bal, C. (2016). Kümeleme analizi. F. Er (Ed.), *Veri madenciliği* (s. 137-173). Eskişehir: Anadolu Üniversitesi Açıköğretim Yayınları.
- Başçıl, M.S. ve Temurtaş, F. (2011). A study on hepatitis disease diagnosis using Multilayer Neural Network with Levenberg Marquardt Training algorithm. *J Med Syst.*, 35(3), 433-436.
- Bekki, A. (2016). İlişki kuralları. F. Er (Editör), *Veri madenciliği* (s. 103-125). Eskişehir: Anadolu Üniversitesi Açıköğretim Yayınları.
- Bellinger, C., Mohamed Jabbar, M.S., Zaiane, O., Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, 17(1), 907-926.
- Birnbaum, M.S., Kornell, N., Bjork, E.L., Bjork, R.A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Mem Cognit.*, 41 392-402.
- Briese, T., Paweska, J.T., McMullan, L.K., Hutchison, S.K., Street, C., Palacios, G., Khristova, M.L., Weyer, J., Swanepoel, R., Egholm, M., Nichol, S.T., Lipkin, W.I. (2009). Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog.*, 5(5), e1000455.
- Chavan, A. and Verma, V.K. (2013). Treatment of missing values for association rules: A recent survey. *International Journal of Computer Applications*, 70 (26), 1-4.
- Chen, H.L., Liu, D.Y., Yang, B., Liu, J., Wang, G. (2011). A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis. *Expert Syst Appl.*, 38(9), 11796-11803.
- Chen, L.A. and Fawcett, T.N. (2016). Using Data Mining strategies in clinical decision making: A literature review. *Comput Inform Nurs.*, 34(10), 448-454.

- Chen, W., Yang, J., Wang, H.L., Shi, Y..F, Tang, H., Li, G.H. (2018). Discovering associations of adverse events with pharmacotherapy in patients with non-small cell lung cancer using modified Apriori Algorithm. *Biomed Res Int.*, 2018, 1245616
- Coşkun, C. ve Baykal, A. (2011). Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması. *Akademik Bilişim'11-XIII. Akademik Bilişim Konferansı* bildirileri. Malatya: İnönü Üniversitesi.
- Çelik, F., Tekin, F., Ünal, N.G., Özütemiz, Ö. (2017). Karaciğer sirozlu 225 hastanın retrospektif irdelenmesi: Tek merkez deneyimi. *Akademik Gastroenteroloji Dergisi*, 16 (2), 47-53.
- Çelik, M. (2009). *Veri madenciliğinde kullanılan sınıflandırma yöntemleri ve bir uygulama*. Yayınlanmamış Yüksek Lisans Tezi, İstanbul: İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü.
- Dain, A., Ruiz, M., Rista, L., Flores, A., Muratore, C., Ladino, C. (2015). Basal and bolus rates in insulin accu-check combo pump type 1 diabetes patients and its correlation with age, frequency of self monitoring and metabolic control. *Diabetes Technol Ther.*, 17, A56-A56.
- Daştan, H. (2016). Türkiye’de ikinci el otomobil fiyatlarını etkileyen faktörlerin hedonik fiyat modeli ile belirlenmesi. *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 18 (1), 33-327.
- De Coninck, A., Fostier, J., Maenhout, S., De Baets, B. (2014). DAIRRY-BLUP: A high-performance computing approach to genomic prediction. *Genetics*, 197, 813-822.
- De Oña, J., López, G., Mujalli, R., Calvo, F.J. (2013) Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis and Prevention*, 51, 1-10.
- Demirel, B. (2010). *Veri madenciliğinde Chaid algoritmasının Sosyal Güvenlik Kurumu veri tabanına uygulanması*. Yayınlanmamış Yüksek Lisans Tezi, Ankara: Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

- Devillez, A. (2004). Four fuzzy supervised classification methods for discriminating classes of non-convex shape. *Fuzzy Sets and Systems*, 141 (2), 219-240.
- Dimitriou, L. and Vlahogianni, E.I. (2015) Fuzzy Modeling of freeway accident duration with rainfall and traffic flow interactions. *Analytic Methods in Accident Research*, 5, 59-71.
- Doğan, B., Erol, B. ve Buldu, A. (2014). Sigortacılık sektöründe müşteri ilişkileri yönetimi için birliktelik kurallarının kullanılması. *Marmara Fen Bilimleri Dergisi*, 3, 105-114.
- Doğan, Ş. (2007). *Veri madenciliği kullanarak biyokimya verilerinden hastalık teşhisi*. Yayınlanmamış Yüksek Lisans Tezi, Elazığ: Fırat Üniversitesi, Fen Bilimleri Enstitüsü.
- Dua, D. and Graff, C. (2019). UCI machine learning repository, Irvine, CA: University of California, *School of Information and Computer Science* sunulan bildiri. . <https://archive.ics.uci.edu/ml/datasets/hepatitis> (Erişim tarihi: 14.01.2019)
- Duggirala, H.J., Tønning, J.M., Smith, E., Bright, R.A., Baker, J.D., Ball, R., Bell, C., Bright-Ponte, S.J., Botsis, T., Bouri, K., Boyer, M., Burkhart, K., Condrey, G.S., Chen, J.J., Chirtel, S., Filice, R.W., Francis, H., Jiang, H., Levine, J., Martin, D., Oladipo, T., O'Neill, R., Palmer, L.A., Paredes, A., Rochester, G., Sholtes, D., Szarfman, A., Wong, H.L., Xu, Z., Kass-Hout, T. (2016). Use of data mining at the Food and Drug Administration. *J Am Med Inform Assoc.*, 23 (2), 428-434.
- Duru, N. (2005). An application of Apriori Algorithm on a diabetic database. *Springer-Verlag Berlin Heidelberg LNAI*, 3681, 398.404.
- Eken, C., Görmez, H., Başhekim, M., Akyürek, S., Kartal, M. (2005) Motorlu taşıt kazaları ve risk faktörleri: Antalya ilinde meydana gelen 12185 trafik kazasının analizi. *Türkiye Acil Tıp Dergisi*, 5 (4), 175-180.
- Erdem, S. ve Özdağoğlu, G. (2008). Ege bölgesindeki bir araştırma ve uygulama hastanesinin acil hasta verilerinin Veri Madenciliği ile analiz edilmesi, *Anadolu Üniversitesi Bilim Ve Teknoloji Dergisi*, 9 (2) , 261-270.

- Erpolat, S. (2012). Otomobil yetkili servislerinde Birliktelik Kurallarının belirlenmesinde Apriori ve Fp-Growth Algoritmalarının karşılaştırılması. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, 12 (1), 37-146.
- Farboudi, S. (2009). *Tıp Bilişiminde istatistiksel veri madenciliği*. Yayınlanmamış Yüksek Lisans Tezi, Ankara: Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.
- Fardet, A., Lakhssassi, S. and Briffaz, A. (2018). Beyond nutrient-based food indices: a data mining approach to search for a quantitative holistic index reflecting the degree of food processing and including physicochemical properties. *Food Funct.*, 9 (1), 561-572.
- Finlay, S. (2014). Predictive analytics, Data Mining and big data myths. *Misconceptions and Methods*. Basingstoke, Hampshire, UK: Palgrave Macmillan. <http://www.palgraveconnect.com.ezproxy> (Erişim tarihi 23.03.2019)
- Fleck, M., Polyansky, A.A. and Zagrovic, B. (2015). Parent: A parallel software for the calculation of conformational entropy in biomolecular systems. *Eur Biophys J.*, 44, 243-248.
- Gallegos-Orozco, J.F. and Rakela-Brödner, J. (2010). Hepatitis viruses: not always what it seems to be. *Rev Med Chil.*, 138 (10), 1302-1311.
- Guan, V.X., Probst, Y.C., Neale, E.P., Batterham, M.J., Tapsell, L.C. (2018). Identifying usual food choices at meals in overweight and obese study volunteers: implications for dietary advice. *Br J Nutr.*, 120 (4), 472-480.
- Gölce, A.C. (2010). *Veri Madenciliğinde Apriori Algoritması ve Apriori Algoritmasının farklı veri kümelerinde uygulanması*. Yayınlanmamış Yüksek Lisans Tezi, Edirne: Trakya Üniversitesi, Fen Bilimleri Enstitüsü.
- Heyman, P., Cochez, C., Hofhuis, A., van der Giessen, J. (2010). A clear and present danger: tick-borne diseases in Europe. *Expert Rev Anti Infect Ther.*, 8 (1), 33-50.
- Hong, K., Lee, K.M. And Jang, S.N. (2015). Incidence and related factors of traffic accidents among the older population in a rapidly aging society. *Archives of Gerontology and Geriatrics*, 60 (3), 471-477.

- Househ, M. and Aldosari, B. (2017). The hazards of Data Mining in healthcare. *Stud Health Technol Inform.*, 238, 80-83.
- Hovsepian, K., Anselmo, P. ve Mazumdar, S. (2011). Supervised inductive learning with Lotka–Volterra derived models. *Knowl Inf Syst*, 26, 195-223.
- Imberman, S. and Domanski, B. (2001). Finding association rules from quantitative data using data booleanization. *Americas Conference on Information Systems*'de sunulan bildiri. New York, USA.
- İnan, O. (2015). *Veri madenciliği uygulamaları için veri indirgeme algoritmalarının geliştirilmesi ve resim madenciliğine uygulanması*. Yayınlanmamış Doktora Tezi, Konya: Selçuk Üniversitesi, Fen Bilimleri Enstitüsü.
- Johnson P., Vandewater, L., Wilson, W., Maruff, P., Savage, G., Graham, P, Macaulay, L.S., Ellis, K.A., Szoeki, C., Martins, R.N., Rowe, C.C., Masters, C.L., Ames, D. and Zhang, P. (2014). Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. *BMC Bioinformatics*, 15(16), 11-25.
- Kar, S., Das, S.C., Tiwari, A., Pharveen, I. (2015). Pattern of road traffic accidents in bhubaneswar, Odisha. *Clinical Epidemiology and Global Health*, 4 (3), 115-119.
- Karabatak, M. ve İnce, M.C. (2004). Apriori Algoritması ile öğrenci başarısı analizi. *Eleco Elektrik-Elektronik ve Bilgisayar Mühendisleri Sempozyumunda* sunulan poster. Bursa: Bursa Uludağ Üniversitesi.
- Keskenler, M.F. ve Keskenler, E.F. (2017). Bulanık Mantığın tarihi gelişimi. *Takvim-i Vekayi*, 5 (1), 1-10.
- Kocamaz, K. (2007). *Hastane yönetim bilgi sistemlerinde veri madenciliği*. Yayınlanmamış Yüksek Lisans Tezi, Konya: Selçuk Üniversitesi, Sosyal Bilimler Enstitüsü.
- Koyuncuoglu, A.S. ve Özgülbaş, N. (2009). Veri Madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. *Bilişim Teknolojileri Dergisi*, 2 (2), 21-32.
- Köroğlu, F. (2016). *Tıp sektöründe apriori algoritması ile bir veri madenciliği uygulaması*. Yayınlanmamış Yüksek Lisans Raporu, Sakarya: Sakarya Üniversitesi, Fen Bilimleri Enstitüsü.

- Köse, İ.A. ve Öztemur, B. (2014). Kayıp veri ele alma yöntemlerinin t-testi ve ANOVA parametreleri üzerine etkisinin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 14 (1), 400-412.
- Kumdereli, Ü.C. (2012). *Tıp bilişimi ve veri madenciliği uygulamaları: EEG sinyallerindeki epileptiform aktiviteye veri madenciliği yöntemlerinin uygulanması*. Yayınlanmamış Yüksek Lisans tezi, Edirne: Trakya Üniversitesi, Fen Bilimleri Enstitüsü.
- Kuo, R.J., Huang, M.H., Cheng, W.C., Lin, C.C., Wu, Y.H. (2015). Application of a two-stage fuzzy neural network to a prostate cancer prognosis system. *Artif Intell Med.*, 63, 119-33.
- Kurt Pehlivanoğlu, M. ve Duru, N. (2015). Veri madenciliği teknikleri kullanılarak ortaokul öğrencilerinin sosyal ağ kullanım analizi: Kocaeli ili örneği. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 3, 508-517.
- Lee, S., Ybarra, N., Jeyaseelan, K., Faria, S., Kopek, N., Brisebois, P. (2015). Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. *Med Phys.*, 42, 2421-2430.
- Lee, S.R., Choi, Y.D. and Cho, N.H. (2018). Association between pathologic factors and ERG expression in prostate cancer: finding pivotal networking. *J Cancer Res Clin Oncol.*, 144 (9), 1665-1683.
- Lezki, Ş., (2016). Temel Kavramlar. F. Er (Editör), *Veri madenciliği* (s. 2-31). Eskişehir: Anadolu Üniversitesi Açıköğretim Yayınları.
- Lian, C. and Denoeux, T. (2015). Cancer therapy outcome prediction based on Dempster-Shafer Theory and PET imaging. *Med Phys.*, 42, 3549.
- Lin, Q. (2007) Mobile customer clustering analysis based on call detail records. *Communications of the IIMA*, 7 (4), 95-100.
- Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198-1202.

- Liu, K.H. and Xiao, L.F. (2014). New health management model based on cloud platform and big data. *J Public Health Prev Med.*, 25, 89-91.
- Mahmoodi, S.A., Mirzaie, K. and Mahmoudi, S.M. (2016). A new algorithm to extract hidden rules of gastric cancer data based on ontology. *Springerplus*, 5, 312.
- Maria Luna, J., Padillo, F., Pechenizkiy, M., Ventura, S. (2018). Apriori versions based on MapReduce for mining frequent patterns on big data. *IEEE Trans Cybern.*, 48 (10), 2851-2865.
- McMahon, B.J., Bruden, D.L., Petersen, K.M., Bulkow, L.R., Parkinson, A.J., Nainan, O., Khristova, M., Zanis, C., Peters, H., Margolis, H.S. (2005). Antibody levels and protection after hepatitis B vaccination: results of a 15-year follow-up. *Ann Intern Med.*, 142 (5), 333-341.
- Miguet, J.P., Coaquette, A., Bresson-Hadni, S., Lab, M. (1990). [The other types of viral hepatitis. *Rev Prat.*, 40 (18), 1656–1659.
- Min, J., Osborne, V., Kowalski, A., Prosperi, M. (2018). Reported adverse events with painkillers: Data mining of the US Food and Drug Administration Adverse Events Reporting System. *Drug Saf.*, 41 (3), 313-320.
- Mirzaei, R., Hafezi-Nejad, N., Sabagh, M.S., Moghaddam, A.A., Eslami, V., Rakhshani, F., Rahimi-Movaghar, V. (2014). Dominant role of drivers, attitude in prevention of road traffic crashes: A study on knowledge, attitude, and practice of drivers in Iran. *Accident Analysis and Prevention*, 66, 36-42.
- Naboulsi, D. (2015). *Analysis and exploitation of mobile traffic datasets. Networking and Internet Architecture*. Yayınlanmamış Doktora tezi, Fransa: INSA de Lyon, Institut National des Sciences Appliquées de Lyon.
- Nassar, Y. and Richter, S. (2018). Gastroparesis in non-diabetics: Associated conditions and possible risk factors. *Gastroenterology Res.*, 11 (5), 340-345.
- Ngai, E.W.T., Xiu, L. and Chau, D.C.K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36, 2592–2602.
- Oğuzlar, A. (20016). Hanehalkı tipi ve kır kent ayırımının diskriminant analizi ile incelenmesi. *Akdeniz İ.İ.B.F. Dergisi*, 11, 70-84.

- Okano, M. and Gross, T.G. (2012). Acute or chronic life-threatening diseases associated with Epstein-Barr virus infection. *Am. J. Med. Sci.*, 343 (6), 483–489.
- Özçakır, F.C. ve Çamurcu, A.Y. (2007). Birliktelik kuralları yönetimi için bir veri madenciliği yazılımı tasarımı ve uygulaması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 6 (12), 21-37.
- Özçalıcı, M. (2017). Veri madenciliğinde birliktelik kuralları ve ikinci el otomobil piyasası üzerine bir uygulama. *Ordu Üniversitesi Sosyal Bilimler Araştırmaları Dergisi*, 7 (1), 45-58.
- Padilla, W.R. and García, J. (2018). Data association methodology to improve spatial predictions in alternative marketing circuits in ecuador. *Comput Intell Neurosci.*, 2018, 6587049.
- Parraga-Alava, J., Dorn, M. and Inostroza-Ponta, M. (2018). A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies. *BioData Min.*, 11, 16-38.
- Poyraz, O. (2012). *Tıp'da veri madenciliği uygulamaları, Meme kanseri veri seti*. Yayınlanmamış Yüksek Lisans Tezi, Edirne: Trakya Üniversitesi, Fen Bilimleri Enstitüsü.
- Qian, Y., Liang, X., Wang, Q., Liang, J., Liu, B., Skowron, A., Yao, Y., Ma, J., Dang, C. (2018). Local rough set: a solution to rough data analysis in big data. *International Journal of Approximate Reasoning*, DOI: 10.1016/j.ijar.2018.01.008.
- R Core Team (2019). R.A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Ranoelirivao, S., de Morsier, F., Tuiaz, D., Rakotoniaina, S., Borgeaudy, M., Ph. Thiran, J., Rakotondraompiana, S. (2014). Multisource clustering of remote sensing images with entropy-based Dempster-Shafer fusion. <https://www.researchgate.net/publication/50417825> (Erişim tarihi 01.07.2019).

- Robertson, S., Spencer, B., Back, N., Farrow, D. (2019). A rule induction framework for the determination of representative learning design in skilled performance. *J Sports Sci.*, 37 (11), 1280-1285.
- Romero, F.A. and Razonable, R.R. (2011). Infections in liver transplant recipients. *World J Hepatol.*, 3 (4), 83–92.
- Rosen, H.R. (2011). Clinical practice. Chronic hepatitis C infection. *The New England Journal of Medicine*, 364 (25), 2429–2438.
- Savasere, A., Omiecinski, E. and Navathe, S., (1995). An efficient algorithm for mining association rules in large databases. *20th International Conference on VLDB*'de sunulan poster. Swizerland: Zurich.
- Seitz, H.M. (1995). Parasitic diseases of the liver. *Verh Dtsch Ges Pathol.*, 79, 241–248.
- Selmi, C. and Gershwin, M.E. (2004). Bacteria and human autoimmunity: the case of primary biliary cirrhosis. *Curr Opin Rheumatol.*, 16 (4), 406-410.
- Shao, Y.H., Kao, C.C., Tang, S.H., Cha, T.L., Tsao, C.W., Meng, E., Yu, D.S., Sun, G.H., Wu, S.T. (2017). Unusual presentation of direct intraperitoneal metastases complicated with massive ascites from plasmacytoid variant of bladder cancer and adenocarcinoma of colon: A case report and literature review. *Medicine (Baltimore)*, 96 (7), e5816. doi: 10.1097/MD.00000000000005816.
- Siemiatkowska, B. and Harasymowicz-Boggio. B. (2017). Place classification using Dempster-Shafer theory foundations of computing and decision sciences. 42 (3), 257-273.
- Somek, M. and Hercigonja-Szekeres, M. (2017). Decision support systems in health care-Velocity of Apriori Algorithm. *Stud Health Technol Inform.*, 244, 53-57.
- Sönmez, H. (2016). Verinin hazırlanması. F. Er (Editör), *Veri madenciliği* (s. 56-83). Eskişehir: Anadolu Üniversitesi Açıköğretim Yayınları.
- Söylemez, İ., Doğan, A. ve Özcan, U. (2016). Association rules on traffic accident: Case of Ankara. *Ege Academic Review*, 16 (5), 11-20.
- Sözen, E., Bardak, T., Peker, H., Bardak, S. (2017). Apriori Algoritması kullanılarak mobilya seçiminde etkili olan faktörlerin analizi. *4th International Furniture and Decoration Congress*'de sunulan bildiri. Turkey: Düzce University.

- Sungur, İ., Akdur, R. and Piyal, B. (2014). Türkiye'deki trafik kazalarının analizi. *Ankara Medical Journal*, 14 (3), 114-124.
- Sunkaria, R.K., Kumar, V., Saxena, S.C., Singhal, A.M. (2014). An ANN-based HRV classifier for cardiac health prognosis. *Electron Healthc.*, 7, 315-330.
- Tang, J.Y., Chuang, L.Y., Hsi, E., Lin, Y.D., Yang, C.H., Chang, H.W. (2013). Identifying the association rules between clinicopathologic factors and higher survival performance in operation-centric oral cancer patients using the Apriori algorithm. *Biomed Res Int.*, 2013, 359634.
- Tavassoli Hojati, A., Ferreira, L., Washington, S., Charles, P. (2013). Hazard based models for freeway traffic incident duration. *Accident Analysis and Prevention*, 52, 171-181.
- Temel, F. ve Özcebe, H. (2006). Türkiye'de karayollarında trafik kazaları. *Sted*, 15(11), 192-200.
- Terlemez, L. (2016). Karar Ağaçları. F. Er (Editör), *Veri madenciliği* (s. 137-173). Eskişehir: Anadolu Üniversitesi Açıköğretim Yayınları.
- The BigML Team (2019). Association discovery with the big ML dashboard. <http://bigml.com>. (Erişim tarihi 17.07.2019).
- Timor, M. ve Şimşek, U.T. (2008). Veri madenciliğinde sepet analizi ile tüketici davranışı modellenmesi. *Yönetim*, 19, 3-10.
- Usama, M., Qadir, J., Raza, A., Arif, H., Lim, K., Yau, A., Elkhatib, Y., Hussain, A., Al-Fuqaha, A. (2017). Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. <https://arxiv.org/pdf/1709.06599> (Erişim tarihi 01.07.2019).
- Vishweswaraiah, S., George, L., Purushothaman, N., Ganguly, K. (2018). A candidate gene identification strategy utilizing mouse to human big-data mining: "3R-tenet" in COPD genetic research. *Respir Res.*, 19 (1), 92-103.
- Vlahogianni, E.I. and Golias, J.C. (2012). Bayesian modeling of the microscopic traffic characteristics of overtaking in two-lane highways Transportation Research Part F: *Traffic Psychology and Behaviour*, 15 (3), 348-357.

- Wei, P.L., Gu, H., Guan, S., Su, Y.Y., Liu, Q., Liu, J., Wang, Z. (2019). Core prescription excavation of launched Chinese patent medicines for treating different syndromes of angina pectoris. *Zhongguo Zhong Yao Za Zhi.*, 44 (5), 1041-1047.
- Weiss, S.R. and Leibowitz, J.L. (2011). Coronavirus pathogenesis. *Adv. Virus Res.*, 81, 85-164.
- Xia, K., Zhong, X., Zhang, L., Wang, J. (2019). Optimization of diagnosis and treatment of chronic diseases based on association analysis under the background of regional integration. *J Med Syst.*, 43 (3), 46-54.
- Yaël, B., Pérol, J. and Dignat-George, F. (2015). Managing samples with implausibly high mean cell hemoglobin concentration (MCHC) values by using red blood cell (RBC) and hemoglobin (HGB) values obtained by fluorescence flow cytometry. *Clin Chem Lab Med.*, 53 (1), S156.
- Yahouédéhou, S.C.M.A., da Guarda, C.C., Figueiredo, C.V.B., Santiago, R.P., Carvalho, S.P., Fiuza, L.M., Ndidi, U.S., Oliveira, R.M., Carvalho, M.O.S., Nascimento, V.M.L., Rocha, L.C., Lyra, I.M., Adorno, E.V., Goncalves, M.S. (2019). Hydroxyurea alters hematological, biochemical and inflammatory biomarkers in Brazilian children with SCA: Investigating associations with β S haplotype and α -thalassemia. *PLoS One*, 14 (7), e0218040. doi: 10.1371/journal.pone.0218040. eCollection 2019.
- Yang, Y., Sun, H., Kang, Z., Wu, Q.H. (2015). Application of decision tree model ID3 algorithm in risk assessment of public health emergencies. *Chin Prev Med.*, 16, 60-64.
- Yi, Y.Q., Fang, R., Ge, J.W., Cheng, S.W., Wang, G.Z., Liu, L. (2018). Analysis on medication rules for treatment of dementia by ancient physicians based on data mining methods. *Zhongguo Zhong Yao Za Zhi.*, 43 (16), 3376-3381.
- Yildirim, P., Majnarić, L., Ekmekci, O., Holzinger, A. (2014). Knowledge discovery of drug data on the example of adverse reaction prediction. *BMC Bioinformatics*, 15(6), S7.
- You, X., Xu, Y., Huang, J., Zhi, Y., Wu, F., Ding, F., Liang, J., Cui, Z., Xu, J., Zhang, J. (2019). A data mining-based analysis of medication rules in treating bone

marrow suppression by Kidney-Tonifying Method. *Evid Based Complement Alternat Med.*, 2019, 1907848

Yu, H., Zou, L., He, Y., Luo, L., Dong, W., Zhang, Y., Lei, X. (2019). Associations between neonatal serum bilirubin and childhood hypertension. *PLoS One*, 14 (7), e0219942. doi: 10.1371/journal.pone.0219942. eCollection 2019.

Zhang, Y., Guo, S.L., Han, L.N., Li, T.L. (2016). Application and exploration of big data mining in clinical medicine. *Chin Med J.*, 129 (6), 731-738.

http-1: <https://archive.ics.uci.edu/ml/machine-learning-databases/> (Eriřim tarihi 04. 02. 2019).

http-2: <http://bilgisayarkavramlari.sadievrenseker.com/2011/08/13/apriori-malum-aposteriori-afortiori/> (Eriřim tarihi 20.04.2019).

http-3: <https://deepinsightr.com/2018/02/05/supervised-unsupervised-machine-learning/> (Eriřim tarihi 01.07.2019).

http-4: [kisi.deu.edu.tr/userweb/s.ucdogruk/ ist1Tanımlayıcı%20İstatistikler\(Bolum2\).pdf](http://kisi.deu.edu.tr/userweb/s.ucdogruk/ist1Tanımlayıcı%20İstatistikler(Bolum2).pdf) (Eriřim tarihi 16.07.2019).

http-5: <https://tr.gadget-info.com/difference-between-supervised> (Eriřim tarihi 01.07.2019)

http-6: <https://www.labtestsonline.org.tr/conditions/hepatit> (Eriřim tarihi 23.04.2019).

http-7: <https://weka.informer.com/3.9/> (Eriřim tarihi 28.01.2019).

http-8: <http://bilgisayarkavramlari.sadievrenseker.com/2009/06/01/weka/> (Eriřim tarihi 28.01.2019).

http-9: <http://merkezlab.erciyes.edu.tr/biyokimya/referans.htm> (Eriřim tarihi 31.07.2019)

http-10: <https://www.youtube.com/watch?v=bg27nJb7Rck> (Eriřim tarihi 26.04.2019)

EKLER

EK-1. Veritabanından elde edilen verilerin Excel tablosunda gösterilmesi

CLASS	AGE	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER BIG	LIVER FIRM	SPLEEN PALPABLE	SPIDERS	ASCITES	VARICES	BILIRUBIN	ALK PHOSPHATE	SGOT	ALBUMIN	PROTIME	HISTOLOGY
2	30	2	1	2	2	2	2	1	2	2	2	2	2	1	85	18	4	?	1
2	50	1	1	2	1	2	2	1	2	2	2	2	2	0,9	135	42	3,5	?	1
2	78	1	2	2	1	2	2	2	2	2	2	2	2	0,7	96	32	4	?	1
2	31	1	?	1	2	2	2	2	2	2	2	2	2	0,7	46	52	4	80	1
2	34	1	2	2	2	2	2	2	2	2	2	2	2	1	?	200	4	?	1
2	34	1	2	2	2	2	2	2	2	2	2	2	2	0,9	95	28	4	75	1
1	51	1	1	2	1	2	1	2	2	1	1	2	2	?	?	?	?	?	1
2	23	1	2	2	2	2	2	2	2	2	2	2	2	1	?	?	?	?	1
2	39	1	2	2	1	2	2	2	1	2	2	2	2	0,7	?	48	4,4	?	1
2	30	1	2	2	2	2	2	2	2	2	2	2	2	1	?	120	3,9	?	1
2	39	1	1	1	2	2	2	1	1	2	2	2	2	1,3	78	30	4,4	85	1
2	32	1	2	1	1	2	2	2	1	2	1	2	2	1	59	249	3,7	54	1
2	41	1	2	1	1	2	2	2	1	2	2	2	2	0,9	81	60	3,9	52	1
2	30	1	2	2	1	2	2	2	1	2	2	2	2	2,2	57	144	4,9	78	1
2	47	1	1	1	2	2	2	2	2	2	2	2	2	?	?	60	?	?	1
2	38	1	1	2	1	1	1	2	2	2	2	1	2	2	72	89	2,9	46	1
2	66	1	2	2	1	2	2	2	2	2	2	2	2	1,2	102	53	4,3	?	1
2	40	1	1	2	1	2	2	2	1	2	2	2	2	0,6	62	166	4	63	1
2	38	1	2	2	2	2	2	2	2	2	2	2	2	0,7	53	42	4,1	85	2
2	38	1	1	1	2	2	2	1	1	2	2	2	2	0,7	70	28	4,2	62	1
2	22	2	2	1	1	2	2	2	2	2	2	2	2	0,9	48	20	4,2	64	1
2	27	1	2	2	1	1	1	1	1	1	1	2	2	1,2	133	98	4,1	39	1
2	31	1	2	2	2	2	2	2	2	2	2	2	2	1	85	20	4	100	1
2	42	1	2	2	2	2	2	2	2	2	2	2	2	0,9	60	63	4,7	47	1
2	25	2	1	1	2	2	2	2	2	2	2	2	2	0,4	45	18	4,3	70	1
2	27	1	1	2	1	1	2	2	2	2	2	2	2	0,8	95	46	3,8	100	1
2	49	1	1	1	1	1	1	2	1	2	1	2	2	0,6	85	48	3,7	?	1
2	58	2	2	2	1	2	2	2	1	2	1	2	2	1,4	175	55	2,7	36	1
2	61	1	1	2	1	2	2	1	1	2	2	2	2	1,3	78	25	3,8	100	1
2	51	1	1	1	1	1	2	2	2	2	2	2	2	1	78	58	4,6	52	1
1	39	1	1	1	1	1	2	2	1	2	2	2	2	2,3	280	98	3,8	40	1
1	62	1	1	2	1	1	2	?	?	2	2	2	2	1	?	60	?	?	1

EK-1. Devam

2	41	2	2	1	1	1	1	2	2	2	2	2	2	0,7	81	53	5	74	1
2	26	2	1	2	2	2	2	2	1	2	2	2	2	0,5	135	29	3,8	60	1
2	35	1	2	2	1	2	2	2	2	2	2	2	2	0,9	58	92	4,3	73	1
1	37	1	2	2	1	2	2	2	2	2	1	2	2	0,6	67	28	4,2	?	1
2	23	1	2	2	1	1	1	2	2	1	2	2	2	1,3	194	150	4,1	90	1
2	20	2	1	2	1	1	1	1	1	1	1	2	2	2,3	150	68	3,9	?	1
2	42	1	1	2	2	2	2	2	2	2	2	2	2	1	85	14	4	100	1
2	65	1	2	2	1	1	2	2	1	1	1	1	2	0,3	180	53	2,9	74	2
2	52	1	1	1	2	2	2	2	2	2	2	2	2	0,7	75	55	4	21	1
2	23	1	2	2	2	2	2	?	?	?	?	?	?	4,6	56	16	4,6	?	1
2	33	1	2	2	2	2	2	2	2	2	2	2	2	1	46	90	4,4	60	1
2	56	1	1	2	1	2	2	2	2	2	2	2	2	0,7	71	18	4,4	100	1
2	34	1	2	2	2	2	2	2	2	2	2	2	2	?	?	86	?	?	1
2	28	1	2	2	1	1	2	2	2	2	2	2	2	0,7	74	110	4,4	?	1
2	37	1	1	2	2	2	2	2	1	2	1	2	2	0,6	80	80	3,8	?	1
2	28	2	2	2	1	1	2	2	1	2	2	2	2	1,8	191	420	3,3	46	1
2	36	1	1	2	2	2	2	2	2	1	2	2	2	0,8	85	44	4,2	85	1
2	38	1	2	1	1	1	1	2	2	2	1	2	2	0,7	125	65	4,2	77	1
2	39	1	1	2	2	2	2	2	2	2	2	2	2	0,9	85	60	4	?	1
2	39	1	2	2	2	2	2	2	2	2	2	2	2	1	85	20	4	?	1
2	44	1	2	2	2	2	2	2	2	2	2	2	2	0,6	110	145	4,4	70	1
2	40	1	2	1	1	2	2	2	1	1	2	2	2	1,2	85	31	4	100	1
2	30	1	2	2	1	2	2	2	2	2	2	2	2	0,7	50	78	4,2	74	1
2	37	1	1	2	1	1	1	2	2	2	2	2	2	0,8	92	59	?	?	1
2	34	1	1	2	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1
2	30	1	2	1	2	2	2	2	2	2	2	2	2	0,7	52	38	3,9	52	1
2	64	1	2	1	1	1	2	1	1	2	2	2	2	1	80	38	4,3	74	1
2	45	2	1	2	1	1	2	2	2	1	2	2	2	1	85	75	?	?	1
2	37	1	2	2	2	2	2	2	2	2	2	2	2	0,7	26	58	4,5	100	1
2	32	1	2	2	2	2	2	2	2	2	2	2	2	0,7	102	64	4	90	1
2	32	1	2	2	1	1	1	2	2	2	1	2	1	3,5	215	54	3,4	29	1
2	36	1	1	2	2	2	2	1	1	1	2	2	2	0,7	164	44	3,1	41	1
2	49	1	2	2	1	1	2	2	2	2	2	2	2	0,8	103	43	3,5	66	1
2	27	1	2	2	2	2	2	2	2	2	2	2	2	0,8	?	38	4,2	?	1
2	56	1	1	2	2	2	2	2	2	2	2	2	2	0,7	62	33	3	?	1
1	57	1	2	2	1	1	1	2	2	2	1	1	2	4,1	?	48	2,6	73	1
2	39	1	2	2	1	2	2	2	2	2	2	2	2	1	34	15	4	54	1
2	44	1	1	2	1	1	2	2	2	2	2	2	2	1,6	68	68	3,7	?	1
2	24	1	2	2	2	2	2	2	2	2	2	2	2	0,8	82	39	4,3	?	1

EK-1. Devam

1	34	1	1	2	1	1	2	1	1	2	1	2	2	2,8	127	182	?	?	1
2	51	1	2	2	1	1	1	?	?	?	?	?	?	0,9	76	271	4,4	?	1
2	36	1	1	2	1	1	1	2	1	2	2	2	2	1	?	45	4	57	1
2	50	1	2	2	2	2	2	2	2	2	2	2	2	1,5	100	100	5,3	?	1
2	32	1	1	1	1	1	2	2	2	2	2	2	2	1	55	45	4,1	56	1
1	58	1	2	2	1	2	2	1	1	1	1	2	2	2	167	242	3,3	?	1
2	34	2	1	1	2	2	2	2	1	2	2	2	2	0,6	30	24	4	76	1
2	34	1	1	2	1	2	2	1	1	2	1	2	2	1	72	46	4,4	57	1
2	28	1	2	2	2	2	2	2	2	2	2	2	2	0,7	85	31	4,9	?	1
2	23	1	2	2	1	1	1	2	2	2	2	2	2	0,8	?	14	4,8	?	1
2	36	1	2	2	2	2	2	2	2	2	2	2	2	0,7	62	224	4,2	100	1
2	30	1	1	2	2	2	2	2	2	2	2	2	2	0,7	100	31	4	100	1
2	67	2	1	2	1	1	2	2	2	?	?	?	?	1,5	179	69	2,9	?	1
2	62	2	2	2	1	1	2	2	1	2	1	2	2	1,3	141	156	3,9	58	1
2	28	1	1	2	1	1	1	2	1	2	2	2	2	1,6	44	123	4	46	1
1	44	1	1	2	1	1	2	2	2	1	2	2	1	0,9	135	55	?	41	2
1	30	1	2	2	1	1	1	2	1	2	1	1	1	2,5	165	64	2,8	?	2
1	38	1	1	2	1	1	1	2	1	2	1	1	1	1,2	118	16	2,8	?	2
2	38	1	1	2	1	1	1	1	1	2	2	2	2	0,6	76	18	4,4	84	2
2	50	2	1	2	1	2	2	1	1	1	1	2	2	0,9	230	117	3,4	41	2
1	42	1	1	2	1	1	1	2	2	1	1	2	1	4,6	?	55	3,3	?	2
2	33	1	2	2	2	2	2	?	?	2	2	2	2	1	?	60	4	?	2
2	52	1	1	2	2	2	2	2	2	2	2	2	2	1,5	?	69	2,9	?	2
1	59	1	1	2	1	1	2	2	1	1	1	2	2	1,5	107	157	3,6	38	2
2	40	1	1	1	1	1	1	1	1	2	2	2	2	0,6	40	69	4,2	67	2
2	30	1	1	2	1	1	2	2	1	2	1	2	2	0,8	147	128	3,9	100	2
2	44	1	1	2	1	1	2	1	1	2	1	2	2	3	114	65	3,5	?	2
1	47	1	2	2	2	2	2	2	2	2	1	2	1	2	84	23	4,2	66	2
2	60	1	1	2	1	2	2	1	1	1	1	2	2	?	?	40	?	?	2
1	48	1	1	2	1	1	2	2	1	2	1	1	1	4,8	123	157	2,7	31	2
2	22	1	2	2	2	2	2	2	2	2	2	2	2	0,7	?	24	?	?	2
2	27	1	1	2	1	2	2	2	1	2	2	2	2	2,4	168	227	3	66	2
2	51	1	1	2	1	1	1	2	1	1	1	2	1	4,6	215	269	3,9	51	2
1	47	1	2	2	1	1	2	2	1	2	2	1	1	1,7	86	20	2,1	46	2
2	25	1	2	2	2	2	2	2	2	2	2	2	2	0,6	?	34	6,4	?	2
1	35	1	1	2	1	2	2	?	?	1	1	1	2	1,5	138	58	2,6	?	2
2	45	1	1	2	1	1	1	2	2	2	2	2	2	2,3	?	648	?	?	2
2	54	1	1	1	2	2	2	1	1	2	2	2	2	1	155	225	3,6	67	2
1	33	1	1	2	1	1	2	2	2	2	2	1	2	0,7	63	80	3	31	2

EK-1. Devam

2	7	1	2	2	2	2	2	2	1	1	2	2	2	0,7	256	25	4,2	?	2
1	42	1	1	1	1	1	2	2	2	2	1	2	2	0,5	62	68	3,8	29	2
2	52	1	1	2	1	2	2	2	2	2	2	2	2	1	85	30	4	?	2
2	45	1	1	2	1	2	2	2	1	1	2	2	2	1,2	81	65	3	?	1
2	36	1	1	2	2	2	2	2	2	2	2	2	2	1,1	141	75	3,3	?	2
2	69	2	2	2	1	2	2	2	2	2	2	2	2	3,2	119	136	?	?	2
2	24	1	1	2	1	2	2	2	2	2	2	2	2	1	?	34	4,1	?	2
2	50	1	2	2	2	2	2	2	2	2	2	2	2	1	139	81	3,9	62	2
1	61	1	1	2	1	1	2	?	?	2	1	2	2	?	?	?	?	?	2
2	54	1	2	2	1	2	2	1	1	2	2	2	2	3,2	85	28	3,8	?	2
1	56	1	1	2	1	1	1	1	1	2	1	2	2	2,9	90	153	4	?	2
2	20	1	1	2	1	1	1	2	2	2	1	1	2	1	160	118	2,9	23	2
2	42	1	2	2	2	2	2	2	2	1	2	2	2	1,5	85	40	?	?	2
2	37	1	1	2	1	2	2	2	2	2	1	2	2	0,9	?	231	4,3	?	2
2	50	1	2	2	2	2	2	2	1	1	1	2	2	1	85	75	4	72	2
2	34	2	2	2	1	1	1	1	1	2	1	2	2	0,7	70	24	4,1	100	2
2	28	1	2	2	1	1	1	?	?	2	1	1	2	1	?	20	4	?	2
1	50	1	2	2	1	2	2	2	1	1	2	1	1	2,8	155	75	2,4	32	2
2	54	1	1	2	1	1	2	2	2	2	2	1	2	1,2	85	92	3,1	66	2
1	57	1	1	2	1	1	2	2	2	2	1	1	2	4,6	82	55	3,3	30	2
2	54	1	2	2	2	2	2	2	2	2	2	2	2	1	85	30	4,5	0	2
1	31	1	1	2	1	1	1	2	2	1	2	2	2	8	?	101	2,2	?	2
2	48	1	2	2	1	1	1	2	1	2	1	2	2	2	158	278	3,8	?	2
2	72	1	2	1	1	2	2	2	1	2	2	2	2	1	115	52	3,4	50	2
1	38	1	1	2	2	2	2	2	1	2	2	2	2	0,4	243	49	3,8	90	2
2	25	1	2	2	1	2	2	1	1	1	1	1	1	1,3	181	181	4,5	57	2
2	51	1	2	2	2	2	2	1	1	2	1	2	2	0,8	?	33	4,5	?	2
2	38	1	2	2	2	2	2	2	1	2	1	2	1	1,6	130	140	3,5	56	2
1	47	1	2	2	1	1	2	2	1	2	1	1	1	1	166	30	2,6	31	2
2	45	1	2	1	2	2	2	2	2	2	2	2	2	1,3	85	44	4,2	85	2
2	36	1	1	2	1	1	1	1	1	2	1	2	1	1,7	295	60	2,7	?	2
1	54	1	1	2	1	1	2	?	?	1	2	1	2	3,9	120	28	3,5	43	2
2	51	1	2	2	1	2	2	2	1	1	1	2	1	1	?	20	3	63	2
1	49	1	1	2	1	1	2	2	2	1	1	2	2	1,4	85	70	3,5	35	2
1	45	1	2	2	1	1	1	2	2	2	1	1	2	1,9	?	114	2,4	?	2
2	31	1	1	2	1	2	2	2	2	2	2	2	2	1,2	75	173	4,2	54	2
1	41	1	2	2	1	2	2	2	1	1	1	2	1	4,2	65	120	3,4	?	2
1	70	1	1	2	1	1	1	?	?	?	?	?	?	1,7	109	528	2,8	35	2
2	20	1	1	2	2	2	2	2	?	2	2	2	2	0,9	89	152	4	?	2

EK-1. Devam

2	36	1	2	2	2	2	2	2	2	2	2	2	2	0,6	120	30	4	?	2
1	46	1	2	2	1	1	1	2	2	2	1	1	1	7,6	?	242	3,3	50	2
2	44	1	2	2	1	2	2	2	1	2	2	2	2	0,9	126	142	4,3	?	2
2	61	1	1	2	1	1	2	1	1	2	1	2	2	0,8	75	20	4,1	?	2
2	53	2	1	2	1	2	2	2	2	1	1	2	1	1,5	81	19	4,1	48	2
1	43	1	2	2	1	2	2	2	2	1	1	1	2	1,2	100	19	3,1	42	2



EK-2. Çalışmaya dahil edilen ve eksik verileri tamamlanmış olan 17 özellik

```
@relation data2-
weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-
last-weka.filters.unsupervised.attribute.StringToNominal-Rlast

@attribute YAS {1,2}
@attribute STEROID {1,2}
@attribute ANTIVIRAL_KULLANIMI {1,2}
@attribute YORGUNLUK {1,2}
@attribute KIRGINLIK {1,2}
@attribute ANOREKSIYA {1,2}
@attribute KARACIGER_BUYUMESI {1,2}
@attribute KARACIGERDE_KITLLE {1,2}
@attribute DALAK_BUYUMESI {1,2}
@attribute SIROZ {1,2}
@attribute ASITLENME {1,2}
@attribute VARISLER {1,2}
@attribute BILIRUBIN {1,2}
@attribute ALK_FOSFAT {1,2}
@attribute SGOT {1,2}
@attribute ALBUMIN {1,2}
@attribute PROTROMBIN_TESTI {1,2}

@data
1,1,2,2,2,2,1,2,2,2,2,2,1,1,1,2,2
2,1,2,1,2,2,1,2,2,2,2,2,1,2,1,1,2
2,2,2,1,2,2,2,2,2,2,2,2,1,1,1,2,1
1,2,1,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
2,1,2,1,2,1,2,2,1,1,2,2,2,2,1,1,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,1,2,2,2,1,2,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,2,2,2,1
1,1,1,2,2,2,1,1,2,2,2,2,1,1,1,2,2
1,2,1,1,2,2,2,1,2,1,2,2,1,1,2,1,1
1,2,1,1,2,2,2,1,2,2,2,2,1,1,1,2,1
1,2,2,1,2,2,2,1,2,2,2,2,2,1,2,2,2
2,1,1,2,2,2,2,2,2,2,2,2,1,2,1,1,1
1,1,2,1,1,1,2,2,2,2,1,2,2,1,2,1,1
2,2,2,1,2,2,2,2,2,2,2,2,1,1,1,2,1
1,1,2,1,2,2,2,1,2,2,2,2,1,1,2,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,1,1,2,2,2,2,1,1,2,2,2,2,1,1,1,2,2
1,2,1,1,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,1,1,1,1,1,1,1,2,2,1,2,2,2,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,1
1,1,1,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,1,2,1,1,2,2,2,2,2,2,2,1,1,1,1,2
2,1,1,1,1,1,2,1,2,1,2,2,1,1,1,1,2
2,2,2,1,2,2,2,1,2,1,2,2,1,2,1,1,1
2,1,2,1,2,2,1,1,2,2,2,2,1,1,1,2,2
```

EK-2. Devam

2,1,1,1,1,2,2,2,2,2,2,2,1,1,1,2,1
1,1,1,1,1,2,2,1,2,2,2,2,2,2,2,1
2,1,2,1,1,2,2,2,2,2,2,2,1,1,1,2,1
1,2,1,1,1,1,2,2,2,2,2,2,1,1,1,2,2
1,1,2,2,2,2,2,1,2,2,2,2,1,2,1,2,2
1,2,2,1,2,2,2,2,2,2,2,2,1,1,2,2,2
1,2,2,1,2,2,2,2,2,1,2,2,1,1,1,2,1
1,2,2,1,1,1,2,2,1,2,2,2,1,2,2,2,2
1,1,2,1,1,1,1,1,1,1,2,2,2,2,1,2,1
2,1,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
2,2,2,1,1,2,2,1,1,1,1,2,1,2,1,1,2
2,1,1,2,2,2,2,2,2,2,2,2,1,1,1,2,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2,2
2,1,2,1,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2,1
1,2,2,1,1,2,2,2,2,2,2,2,1,1,2,2,2
1,1,2,2,2,2,2,1,2,1,2,2,1,1,1,2,2
1,2,2,1,1,2,2,1,2,2,2,2,2,2,1,1
1,1,2,2,2,2,2,2,1,2,2,2,1,1,1,2,2
1,2,1,1,1,1,2,2,2,1,2,2,1,2,1,2,2
1,1,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,1
2,2,2,2,2,2,2,2,2,2,2,2,1,2,2,2,2
1,2,1,1,2,2,2,1,1,2,2,2,1,1,1,2,2
1,2,2,1,2,2,2,2,2,2,2,2,1,1,1,2,2
1,1,2,1,1,1,2,2,2,2,2,2,1,1,1,2,1
1,1,2,1,1,2,2,2,2,2,2,2,2,1,1,1,1
1,2,1,2,2,2,2,2,2,2,2,2,1,1,1,1,1
2,2,1,1,1,2,1,1,2,2,2,2,1,1,1,2,2
2,1,2,1,1,2,2,2,1,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,2,2,1,1,1,2,2,2,1,2,1,2,2,1,1,1
1,1,2,2,2,2,1,1,1,2,2,2,1,2,1,1,1
2,2,2,1,1,2,2,2,2,2,2,2,1,1,1,1,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,1
2,1,2,2,2,2,2,2,2,2,2,2,1,1,1,1,2
2,2,2,1,1,1,2,2,2,1,1,2,2,2,1,1,2
1,2,2,1,2,2,2,2,2,2,2,2,1,1,1,2,1
2,1,2,1,1,2,2,2,2,2,2,2,2,1,1,1,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2,1
1,1,2,1,1,2,1,1,2,1,2,2,2,2,1,1,1
2,2,2,1,1,1,2,2,2,2,1,2,1,1,2,2,1
1,1,2,1,1,1,2,1,2,2,2,2,1,1,1,2,1
2,2,2,2,2,2,2,2,2,2,2,2,2,1,2,2,2
1,1,1,1,1,2,2,2,2,2,2,2,2,1,1,1,2,1
2,2,2,1,2,2,1,1,1,1,2,2,2,2,2,1,2
1,1,1,2,2,2,2,1,2,2,2,2,1,1,1,2,2
1,1,2,1,2,2,1,1,2,1,2,2,1,1,1,2,1
1,2,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2
1,2,2,1,1,1,2,2,2,2,2,2,2,1,1,2,2
1,2,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2,2

EK-2. Devam

1,1,2,2,2,2,2,2,2,2,2,2,1,1,1,2,2
2,1,2,1,1,2,2,2,1,1,2,2,2,2,1,1,1
2,2,2,1,1,2,2,1,2,1,2,2,1,2,2,2,1
1,1,2,1,1,1,2,1,2,2,2,2,2,1,2,2,1
2,1,2,1,1,2,2,2,1,2,2,1,1,2,1,1,1
1,2,2,1,1,1,2,1,2,1,1,1,2,2,1,1,2
1,1,2,1,1,1,2,1,2,1,1,1,1,2,1,1,1
1,1,2,1,1,1,1,1,2,2,2,2,1,1,2,2
2,1,2,1,2,2,1,1,1,1,2,2,1,2,2,1,1
2,1,2,1,1,1,2,2,1,1,2,1,2,2,1,1,2
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2
2,1,2,2,2,2,2,2,2,2,2,2,2,1,1,1
2,1,2,1,1,2,2,1,1,1,2,2,2,2,2,1,1
1,1,1,1,1,1,1,1,2,2,2,2,1,1,2,2
1,1,2,1,1,2,2,1,2,1,2,2,1,2,2,2,2
2,1,2,1,1,2,1,1,2,1,2,2,2,2,1,1,1
2,2,2,2,2,2,2,2,2,1,2,1,2,1,1,2,2
2,1,2,1,2,2,1,1,1,1,2,2,2,1,1,2,2
2,1,2,1,1,2,2,1,2,1,1,1,2,2,2,1,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2
1,1,2,1,2,2,2,1,2,2,2,2,2,2,1,2
2,1,2,1,1,1,2,1,1,1,2,1,2,2,2,2,1
2,2,2,1,1,2,2,1,2,2,1,1,2,1,1,1,1
1,2,2,2,2,2,2,2,2,2,2,2,1,1,2,2
1,1,2,1,2,2,1,1,1,1,2,2,2,2,1,1,2
2,1,2,1,1,1,1,1,1,2,1,2,2,2,1,2,2,1
1,1,2,1,1,2,2,2,2,2,1,2,1,1,1,1,1
1,2,2,2,2,2,2,2,1,1,2,2,2,1,2,2,2
2,1,1,1,1,2,2,2,2,1,2,2,1,1,1,2,1
2,1,2,1,2,2,2,2,2,2,2,2,1,1,1,2,2
2,1,2,1,2,2,2,1,1,2,2,2,1,1,1,1,1
1,1,2,2,2,2,2,2,2,2,2,2,1,2,1,1,1
2,2,2,1,2,2,2,2,2,2,2,2,2,2,2,1
1,1,2,1,2,2,2,2,2,2,2,2,1,1,2,1
2,2,2,2,2,2,2,2,2,2,2,2,1,2,1,2,2
2,1,2,1,1,2,2,2,2,1,2,2,1,1,1,1,1
2,2,2,1,2,2,1,1,2,2,2,2,1,1,2,1
2,1,2,1,1,1,1,1,1,2,1,2,2,2,1,2,2,1
1,1,2,1,1,1,2,2,2,1,1,2,1,2,2,1,1
2,2,2,2,2,2,2,2,2,1,2,2,2,2,1,1,1
1,1,2,1,2,2,2,2,2,1,2,2,1,2,2,2,1
2,2,2,2,2,2,2,1,1,1,2,2,1,1,1,2,2
1,2,2,1,1,1,2,1,2,1,1,2,1,1,1,2,2
2,2,2,1,2,2,2,1,1,2,1,1,2,2,1,1,1
2,1,2,1,1,2,2,2,2,2,1,2,1,1,2,1,2
2,1,2,1,1,2,2,2,2,1,1,2,2,1,1,1,1
2,2,2,2,2,2,2,2,2,2,2,2,1,1,2,1
1,1,2,1,1,1,2,2,1,2,2,2,2,2,1,1
2,2,2,1,1,1,2,1,2,1,2,2,2,2,2,1
2,2,1,1,2,2,2,1,2,2,2,2,1,2,1,1,1
1,1,2,2,2,2,2,2,1,2,2,2,2,1,2,2,2

EK-2. Devam

1,2,2,1,2,2,1,1,1,1,1,1,1,2,2,2,1
2,2,2,2,2,2,1,1,2,1,2,2,1,1,1,2,1
1,2,2,2,2,2,2,1,2,1,2,1,2,2,2,1,1
2,2,2,1,1,2,2,1,2,1,1,1,1,2,1,1,1
2,2,1,2,2,2,2,2,2,2,2,2,1,1,1,2,2
1,1,2,1,1,1,1,1,2,1,2,1,2,2,1,1,1
2,1,2,1,1,2,2,2,1,2,1,2,2,2,1,1,1
2,2,2,1,2,2,2,1,1,1,2,1,1,1,1,1,2
2,1,2,1,1,2,2,2,1,1,2,2,1,1,1,1,1
2,2,2,1,1,1,2,2,2,1,1,2,2,2,2,1,1
1,1,2,1,2,2,2,2,2,2,2,2,1,1,2,2,1
1,2,2,1,2,2,2,1,1,1,2,1,2,1,2,1,1
2,1,2,1,1,1,2,1,2,1,1,2,2,2,2,1,1
1,1,2,2,2,2,2,2,2,2,2,2,1,1,2,2,2
1,2,2,2,2,2,2,2,2,2,2,2,1,2,1,2,2
2,2,2,1,1,1,2,2,2,1,1,1,2,2,2,1,1
2,2,2,1,2,2,2,1,2,2,2,2,1,2,2,2,1
2,1,2,1,1,2,1,1,2,1,2,2,1,1,1,2,1
2,1,2,1,2,2,2,2,1,1,2,1,2,1,1,2,1
2,2,2,1,2,2,2,2,1,1,1,2,1,1,1,1,1

ÖZGEÇMİŞ

Adı Soyadı : Birsen RENCÜZOĞULLARI

Yabancı dil : İngilizce, YDS 2014: 82,5

e-posta : birsenrencuz@gmail.com

Doğum Yeri ve Yılı : Adana, 1992

Eğitim Bilgisi:

2010- Seyhan Danişment Gazi Anadolu Lisesi

2015- Orta Doğu Teknik Üniversitesi Fen Edebiyat Fakültesi, Matematik Bölümü

2019- Eskişehir Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü İstatistik Anabilim Dalı,
Uygulamalı İstatistik Tezli Yüksek Lisans

Meslek Bilgisi:

2017- Özel Çağdaş Koleji, Matematik Öğretmeni