TÜRKİYE CUMHURİYETİ
KARADENİZ TEKNİK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TIP BİLİŞİMİ ANABİLİM DALI

KARADENİZ TEKNİK ÜNİVERSİTESİ FARABİ HASTANESİ BİYOKİMYA LABORATUVARI TEST SONUÇLARINDAN VERİ MADENCİLİĞİ YOLU İLE ÖRÜNTÜ ÇIKARMA

YASEMİN ZEYNEP ENGİN

YÜKSEK LİSANS TEZİ

Doç. Dr. Kemal TURHAN

TRABZON – 2014

TÜRKİYE CUMHURİYETİ
KARADENİZ TEKNİK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TIP BİLİŞİMİ ANABİLİM DALI

KARADENİZ TEKNİK ÜNİVERSİTESİ FARABİ HASTANESİ BİYOKİMYA LABORATUVARI TEST SONUÇLARINDAN VERİ MADENCİLİĞİ YOLU İLE ÖRÜNTÜ ÇIKARMA

YASEMİN ZEYNEP ENGİN

YÜKSEK LİSANS TEZİ

Doç. Dr. Kemal TURHAN

TRABZON – 2014

ONAY

Bu tez Yüksek Lisans Tezi Standartlarına Uygun Bulunmuştur.
Doç. Dr. Kemal TURHAN
Tıp Bilişimi Anabilim Dalı Başkanı

Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Tıp Bilişimi Anabilim Dalı Yüksek Lisans öğrencisi Yasemin Zeynep Engin'in hazırladığı "KARADENİZ TEKNİK ÜNİVERSİTESİ FARABİ HASTANESİ BİYOKİMYA LABORATUVARI TEST SONUÇLARINDAN VERİ MADENCİLİĞİ YOLU İLE ÖRÜNTÜ ÇIKARMA" başlıklı tez KTÜ Lisansüstü Eğitim - Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca kapsam ve bilimsel kalite yönünden değerlendirilerek Yüksek Lisans Tezi olarak kabul edilmiştir.

Danışman Doç. Dr. Kemal Turhan _____

Yüksek Lisans/Doktora Sınavı Jüri Üyeleri

Doç. Dr. Kemal Turhan _____
Prof. Dr. Asım Örem _____
Prof. Dr. Hamdi Öğüt _____

Tarih: 13/01/2014

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

…….
Prof. Dr. Ahmet KALKAN
Sağlık Bilimleri Enstitüsü Müdürü

BEYAN

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

08.12.2013

Yasemin Zeynep Engin

(imza)

İthaf

Yüksek lisans tezimi, beni her koşulda destekleyen, zorluklar karşısında yılmadan mücadele etmem ve asla doğruluktan ve dürüstlükten ödün vermememem gerektiğini öğreten annem Asuman ENGİN'e ithaf ediyorum.

TEŞEKKÜR

Bu günlere gelmemde kuşkusuz en büyük payın sahibi, maddi ve manevi en büyük destekçilerim olan başta annem ve babam olmak üzere canım aileme, sonsuz sevgi, saygı ve teşekkürlerimi sunarım.

İhtiyaç duyduğum hiçbir zaman desteğini esirgemeyen, yanında çalışmaktan gurur duyduğum tez danışmanım ve saygıdeğer hocam KTÜ Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi Anabilim Dalı Başkanı Doç. Dr. Kemal Turhan' a; birlikte çalışmaktan zevk aldığım çalışma arkadaşlarıma,

Yüksek lisans eğitimim boyunca kıymetli bilgilerini bizlerden esirgemeyen bütün saygıdeğer hocalarıma, özellikle tezim konusunda da yol gösteren ve ufkumu genişleten KTÜ Tıp Fakültesi Biyokimya Anabilim Dalı Öğretim Üyesi Prof. Dr. Asım Örem' e, KTÜ Tıp Fakültesi Biyokimya Anabilim Dalı'ndan Araş.Gör.Dr. Selçuk Yaman' a ve Uz.Dr. Sabiha Kamburoğlu' na sonsuz teşekkürlerimi bir borç bilirim.

Son olarak, hayatıma girdiği günden beri her zaman yanımda olan ve bana birçok güzel duyguyu yaşatan nişanlım Rıdvan Avcı' ya ve tüm dostlarıma sevgi ve teşekkürlerimi sunarım.

Yasemin Zeynep Engin

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

Şekil **Sayfa No**

TABLOLAR DİZİNİ

Tablo **Tablo No**

KISALTMA, SİMGE ve FORMÜLLER DİZİNİ

Kısaltmalar

ARA	Artırılmış Regresyon Ağaçları
DBMS	Database Management Systems
DM	Data Mining
DVM	Destek Vektör Makineleri
GNU	General Public License
Hb	Hemoglobin
ICD	International Classification of Diseases
KDD	Knowledge Discovery in Databases
KTÜ	Karadeniz Teknik Üniversitesi
RBC	Red Blood Count
VBK	Veritabanlarında Bilgi Keşfi
VYS	Veritabanı Yönetim Sistemleri
WBC	White Blood Count
YSA	Yapay Sinir Ağları

ÖZET

Karadeniz Teknik Üniversitesi Farabi Hastanesi Biyokimya Laboratuvarı Test Sonuçlarından Veri Madenciliği Yolu ile Örüntü Çıkarma

Veri madenciliği, tek boyutlu yararlanılan çok miktardaki verinin yararlı bilgi birikimine dönüştürülmesini ve özellikle öngörülemeyen boyutlarıyla kullanılmasını sağlayan popüler bir çalışma alanıdır. Tıp alanında da çok sayıda ve türdeki veri veritabanlarında kayıt altına alınmaktadır. Bu bağlamda, sağlık alanı veri madenciliği çalışmaları için önemli kaynaklardan birisidir. Çalışmada, tıp alanındaki laboratuvar verilerinin veri madenciliği teknikleri ile analiz edilmesi ve örüntüler çıkarılması, elde edilen örüntülerden çıkarımlar yapılması amaçlanmıştır. Çalışma kapsamına Karadeniz Teknik Üniversitesi Tıp Fakültesi Farabi Hastanesi biyokimya laboratuvarında Ocak-Aralık 2010 ve Haziran 2011-Haziran 2012 süreleri arasında tutulan test sonuçları kullanılmıştır. Tez kapsamında, birliktelik kuralı analizi ile test seçiminde canlı arama tekniğinin kullanılması; kanser belirteci testlerin değerlendirilerek destek vektör makineleri yöntemiyle jinekolojik kanserlerin tanı kontrolünün yapılması; lökosit(WBC), eritrosit(RBC) ve hemoglobin(Hb) değerlerinin artırılmış regresyon ağaçları yöntemiyle tahmin edilmesi; Potasyum (K) ve Hemoliz İndeksi (HI) arasındaki ilişkinin kümeleme analizi yöntemi ile irdelenmesi; yaşamsal tahmin yapılması konuları ele alınmıştır. Birliktelik kuralları çıkarılarak sıklıkla birlikte istenen testlerin canlı arama tekniği ile görüntülenmesi ve test seçiminde hekime yol göstermesi; test seçim başarısını artırmaya, hekimin daha hızlı karar verebilmesine ve eksik istemlerin önüne geçilerek maliyetlerin düşürülmesine imkan sunacaktır. Kanser belirteci testlerin değerlendirilerek tanı kontrollerinin yapılması ile kullanıcı hatası sonucu yanlış girildiği bariz olan ICD (International Statistical Classification of Diseases and Related Health Problems) kodlarının tespit edilebilmesi ve yeniden hekim görüşüne sunulması; böylece sisteme girilen tanı kodlarının doğruluğunun ve tutarlılığının artırılması amaçlanmıştır. Elde edilen sonuçlarda 304 hasta içinde yanlış sınıflandırılan iki kayıt gözlenmiş; bunlardan birinin kardiyoloji polikliniğinde kayıtlı ve 30 yaşında olduğu saptanmıştır. Bu ICD kodunun yanlış girildiği düşünülmektedir. WBC, RBC ve Hb parametrelerinin tahmini yapılmış ve WBC testinde gözlenen değerlere yakın sonuçlar elde edilmiştir.

Bu yöntemle başka uygun problem alanlarında da araştırma amaçlı olarak eksik verilerin tamamlanabileceği gösterilmiştir. K ve HI arasında var olduğu bilinen ilişki bu konuda daha önce kullanılmamış bir yöntemle ortaya çıkarılmış ve geçerli değerler içeren iki kümede K ve HI arasında korelasyon değeri 0,91 ve 0,80 hesaplanmıştır. Yöntemin bu gibi problemlerde başarılı olduğu ve başka tıbbi problemlerde kullanılmaya uygun olduğu ortaya konulmuştur. Destek vektör makineleri kullanılarak, laktat ve laktat dehidrogenaz testleri istenen hastalarda yapılan yaşamsal tahmin sınıflandırmasında, %97, %96.6 ve %92.3 oranlarında başarılı sonuçlar elde edilmiştir. Hemogram parametreleri istenen diyabetli hastalarda yapılan yaşamsal tahmin analizinde ise, %96.6 ve %97.5 oranlarında başarılı sınıflandırma yapılmıştır. Bu şekilde risk grubunda olma olasılığı yüksek hastalara dikkat çekilebilir, bu hastaların risk değerlendirmelerinin yapılması önerilebilir. Yapılan analizlerle tıbbi veriden önceden öngörülemeyen ve klasik yöntemlerle elde edilemeyecek sonuçlar çıkarılmıştır. Fakat bu sonuçların anlamlandırılması için tıbbi alan uzmanı desteği kesinlikle gerekmektedir. Bu yüzden tıbbi veri üzerinde yapılacak olan veri madenciliği çalışmaları multidisipliner şekilde yürütülmelidir. Ayrıca kullanılan tıbbi veriler ön işlemeye tabi tutulmuş ve birçok eksik kayıt olduğu görülmüştür. Daha sonra yapılması planlanan çalışmaların başarısı için verilerin daha kaliteli olması gerekmektedir.

Anahtar Sözcükler: Veri madenciliği, biyokimya testleri, birliktelik analizi, destek vektör makineleri, kümeleme analizi, artırılmış regresyon ağaçları, örüntü çıkarma

SUMMARY

Pattern Extraction From Karadeniz Technical University Farabi Hospital Biochemistry Laboratory Medical Tests Using Data Mining Techniques

Data mining is a popular field that enables you to convert the massive amount of data to useful information and allows you to use it with unpredictable dimensions. In the field of medicine, a large number of data in many types are recorded and stored. In this context, health area is one of the sources for the studies of data mining. Aim of the study was analyzing laboratory data of the medical field with data mining techniques and extracting patterns, then making inference from derived patterns. Biochemistry laboratory data of Karadeniz Technical University, Faculty of Medicine, Farabi Hospital collected between dates January-December 2010 and June 2011-June 2012 were included in the study. In the scope of this thesis, live search technique for medical test selection by using association rule mining for more accurate test claims; the control of the diagnosis of gynecologic cancers and detecting ICD codes entered incorrectly; prediction of some complete blood count parameters' values and searching method to complete the missing data in another possible problem area; analysis of a known problem area with previously unused method, analyzing the relation between Potassium (K) and Hemolysis Index (HI) with data mining; prediction of survival and identifying the patients who are in risk group realized. Displaying rules often requested together provides physicians to decide more quickly and lowering costs by avoiding inaccurate claims. Diagnostic checks done by evaluating cancer marker tests aim to detect the incorrect ICD codes as a result of user error obviously, so that to increase the accuracy and consistency of the identification codes entered into the system. Obtained results show misclassified two records observed in 304 patients, one of them was in the cardiology clinic and 30 years old. It is thought that the ICD code was entered incorrectly. Values of WBC, RBC and Hb parameters were predicted and WBC test results were close to the observed values. Missing data was shown to be complete in other suitable research areas with this method. Relationship between K(potassium) and HI(hemolysis index) is known to exist is revealed by a method previously unused in this regard. In two clusters containing valid values of K and HI, correlation value was

calculated 0.91 and 0.80. The success of the method have been proved on such problems and it have been found to be suitable for use in other medical problems. Classification of survival predict in patients which lactate and lactate dehydrogenase tests have requested good results were obtained in the rates of 97%, 96.6% and 92.3%. Survival prediction of hemogram parameters requested of patients with diabetes successful classification was made in rates of 96.6% and 97.5%. In this way patients likely to be at risk can be pointed out, performing risk assessments of these patients can suggested. With the analysis of medical data, unpredictable results can not be obtained by conventional methods was concluded. However, the support of the medical field expert is required for the signification of results. Therefore data mining studies that will be performed on medical data should be conducted in a multidisciplinary way. Furthermore used medical data was preprocessed and many missing records were found. Data needs to have better quality for the success of planned work in future.

Keywords: Data mining, biochemistry, association analysis, support vector machines, cluster analysis, regression trees increased, pattern extraction

3. GİRİŞ ve AMAÇ

Gelişen teknoloji ile günümüzde hemen her alanda bilgisayar sistemleri kullanılmaktadır. Bu sistemlerle beraber, çok sayıda ve türdeki veriler veritabanlarında kayıt altına alınabilir ve saklanabilir olmuşlardır. Toplanan bu veriler gelecekte tahmin ve analizler yapılması için taban oluşturmaktadır. Veri madenciliği, bu veriler analiz edilerek önceden bilinmeyen fakat potansiyel olarak anlamlı bilgiye ulaşılması işidir. Yani veri madenciliği ile veriler içerisindeki örüntüler, bilinmeyen ilişkiler, düzensizlikler ve kurallar tespit edilebilir. Veri madenciliği astronomi, biyoloji, finans, pazarlama, sigortacılık, tıp gibi birçok dalda uygulanmaktadır.

Veri madenciliği büyük veri kümeleri üzerinde gizli kalmış örüntüleri tespit etmeye olanak verir. Bu yönü veri madenciliğini istatistikten ayırır. İstatistikte bilinen faktörler arasındaki ilişkiler incelenir. Veri madenciliğinde ise bilinmeyen faktörler arasındaki tahmin edilemeyecek ilişkiler ortaya çıkarılır. Örneğin, bira alan müşterilerin %30'unun bebek bezi de aldığı ilişkisi önceden tahmin edilebilecek türden bir ilişki değildir. Fakat yapılan veri madenciliği analizleri ile bu tarz ilginç ilişkiler ortaya konulabilmektedirler.

Veri madenciliği uygulanacak olan veriler çok çeşitli türde olabilir; sayılar, metinler, sesler ve hatta görüntüler gibi. Bu verilerin veri madenciliğine uygun hale getirilebilmesi için birtakım ön işlemlerden geçmeleri gerekmektedir. Bu işlemler, veri toplama, gürültülü ve tutarsız verilerin çıkarılması için veri temizleme, analiz ile ilgili verilerin ayrılması için veri seçme, veri madenciliğinde kullanılabilecek hale getirmek için veri dönüştürme şeklinde ilerleyecektir. Ön işleme aşamasından geçen veriler veri madenciliği metotları uygulanarak yorumlanmaya ve yorumlara bağlı olarak bilgiye ulaşılmasına olanak sağlayacaktır.

Tıbbi veriler diğer alanlardaki verilerle karşılaştırıldığında temelde çok fazla farklılık olmasa da, uygulamada ciddi fark teşkil etmektedirler. Verinin biçimi ve boyutunun yanı sıra, hukuki ve ahlaki yönü de önemlidir. Tıbbi veriler, her çeşit görüntüleme cihazı verisi, özel görüşmeler, doktor gözlem ve girişimleri, yardımcı

sağlık personeli gözlem ve girişimleri, laboratuvar verileri gibi çok çeşitli biçimlerde olabilir.

Veri madenciliğinin çeşitli uygulama modelleri vardır. Bu modeller tahmin edici ve tanımlayıcı olmak üzere iki başlık altında tanımlanabilir. Sınıflandırma, regresyon analizi, zaman serileri analizi tahmin edici; kümeleme, sıralı dizi keşfi, birliktelik kuralları çıkarımı ise tanımlayıcı modeller arasındadır. Bu çalışmada birliktelik kuralı analizi, kümeleme analizi, sınıflandırma yöntemlerinden destek vektör makineleri, artırılmış regresyon ağaçları kullanılmıştır.

Bu çalışmada Karadeniz Teknik Üniversitesi (KTÜ) Tıp Fakültesi Farabi Hastanesi biyokimya laboratuvarı verileri kullanılarak, belirlenen problem alanlarında veri madenciliği yöntemleri kullanılarak analizler yapılmıştır. Bu problem alanları;

- eksik veya hatalı biyokimya test istemleri azaltılabilir mi, daha isabetli test istemleri sağlanabilir mi;

- hızlı davranmayı ve karar vermeyi gerektiren hastane şartlarında sağlık uzmanlarınca konulan tanıların tutarlılığı biyokimya test sonuçlarından faydalanılarak kontrol edilebilir mi;

- hızlı çalışma temposunda eksik girilen hasta biyokimya test sonuçları tahmin edilebilir mi;

- istatistiksel yöntemler ile irdelenmeye uygun bilinen tıbbi problem alanlarında veri madenciliği yöntemleri kullanılarak doğru sonuçlara ulaşılabilir mi;

- hastaların çeşitli biyokimya test sonuçları kullanılarak yaşamsal tahmin yapılabilir mi?

şeklindedir. Analizler sonucunda farklı bakış açısı oluşturabilecek bulgulara ulaşılmış, önceden öngörülemeyen sonuçlar elde edilmiştir. Çalışmada iki farklı veri grubu kullanılmış olup; birinci grup hastanın doğum tarihini, cinsiyetini, testin istendiği polikliniği, test kodunu, test sonucunu, testin hangi cihazda işlendiğini, kullanılan materyali, sonuçlarla ilişkilendirilen hastalığın ICD kodunu ve testin istendiği tarih bilgilerini içermektedir ve Ocak-Aralık 2010 tarihleri arasında toplanmıştır. İkinci kısım veri ise Haziran 2011-Haziran 2012 tarihleri arasında hastanenin çeşitli servis ve polikliniklerinde tedavi gören hastalara ait biyokimya test verileridir.

Çalışmanın sonraki bölümünde bilgi ve veri kavramları hakkında genel bilgiler verilmiş, veritabanlarında bilgi keşfi süreci ve sürecin adımları açıklanmıştır. Bilgi keşfi sürecinin en önemli adımı olan ve bazı kaynaklarda veritabanlarında bilgi keşfi süreci anlamında da kullanılan veri madenciliği konusu anlatılmış ve kullanılan yöntemler sunulmuştur. Ayrıca bilgi keşfi süreci kapsamında uygulanan veri önişleme basamaklarından ve veri madenciliğinin ilgilendiği problem alanlarından bahsedilmiştir.

Beşinci bölümde bu çalışmada kullanılan veri madenciliği yöntemleri olan birliktelik kuralı analizi, destek vektör makineleri yöntemi, kümeleme analizi ve artırılmış regresyon ağaçları metotları detaylandırılmıştır. Ayrıca geliştirilen uygulamaların amaç, kapsam, kullanılan algoritma ve programlar gibi detayları verilmiştir. Bu çalışma kapsamındaki beş farklı uygulamanın materyal-metot kısmını da içermektedir.

Altıncı bölümde, yapılan uygulamalar sonucunda elde edilen bulgular açıklanmış ve bulguların değerlendirilmesi yapılmıştır.

Son bölüm olan tartışma ve sonuç kısmında elde edilen sonuçlar irdelenerek katkılar vurgulanmıştır. Ayrıca ileride yapılabilecek benzeri çalışmalar ve uygulama alanları için öneriler tartışılmıştır.

4. GENEL BİLGİLER

4.1. Veri

Veri, bir bilgi sistemine girilen veya diğer cihazlardan alınan, yapısal olmayan, işlenmemiş girdilerin tümüdür. Veri metin, sayı, görüntü gibi çeşitli formatlarda olabilir. Veriler filtreleme, formatlama gibi işlemler uygulanarak yoruma hazır hale getirilebilirler (1).

4.2. Bilgi

Bilgi, belli bir süreçten geçmiş veriler olarak tanımlanabilir. Organizasyonlar için bilgi; müşteriler, ürünler, süreçler, hatalar ve başarılar hakkında sahip olunan enformasyondur (2).

Bilginin elde edilmesinde belirli bir sıra vardır. Sırasıyla imgelerden veriler, verilerden enformasyon, enformasyondan da bilgi elde edilir. Belirli durumlarda yararlı olacak şekilde kullanılan enformasyona bilgi adı verilmektedir (3).

4.3. Veritabanı ve Veritabanı Yönetim Sistemleri

Veritabanı büyük miktardaki veriyi kolayca depolamayı ve düzenlemeyi amaçlayan bir sistemdir (4). Veritabanları verilerin düzenli olarak tutulduğu, güncellenebilir, taşınabilir, erişim imkanı sunan, veriler arasında ilişkilerin tanımlanabildiği kayıt ve dosyalar bütünüdür. Günümüzde veritabanı yazılımları ile büyük miktarlardaki verileri depolamak ve işleyebilmek mümkündür.

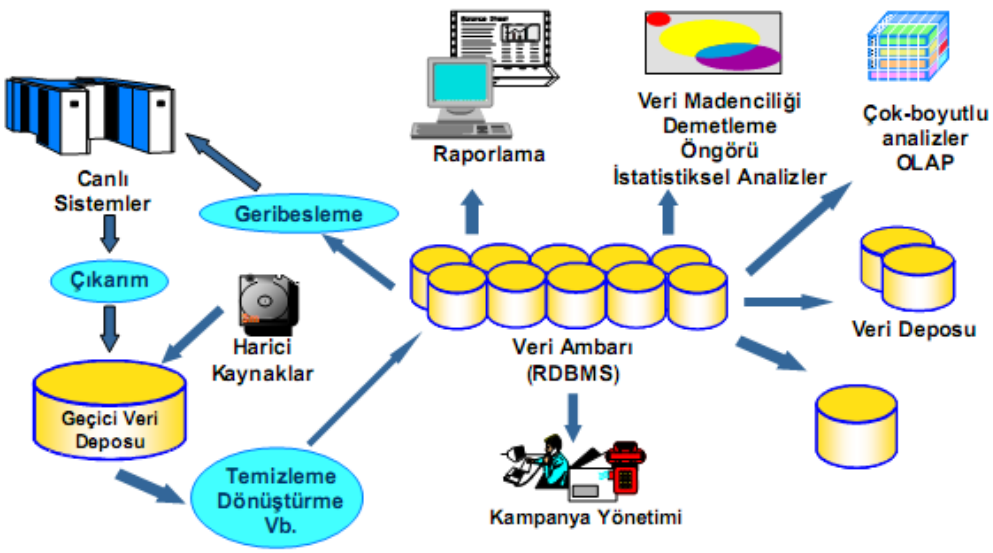
Sistemlerin karmaşıklaşmasıyla gittikçe daha fazla verinin depolanması gereksinimi doğmuştur. Fakat tutulan verilerin boyutları arttıkça el ile kontrol etmek imkansız hale gelmiştir. Bu yüzden veritabanı yönetim sistemleri kullanılmaya başlanmıştır. Bir veritabanını tanımlamak, kullanmak, değiştirmek ve veritabanı sistemleri ile ilgili her türlü işletimsel gereksinimleri karşılamak için kullanılan programlara Veritabanı Yönetim Sistemleri (Database Management Systems - DBMS)

adı verilir (4). Veritabanı Yönetim Sistemi ilişkisel, ağ, hiyerarşik veya nesne yönelimli veritabanı modellerinden birini kullanabilir. Bunlar içinde ilişkisel veri modeli günümüzde en yaygın olarak kullanılan veritabanı modelidir.

İlişkisel veritabanı, her biri özel isimlere sahip tablolardan oluşmaktadır. Burada her tablo bir varlığa veya bir ilişkiye karşılık gelmektedir. Tablonun sütunları nitelikleri; satırları ise bu niteliklerin değerlerini ifade eder. Her bir satır bir "kayıt" olarak da düşünülebilir. Anahtar alan tablonun tanımlayıcısıdır (5).

4.4. Veri Ambarı

Veri ambarı veri kaynaklarında tutulan verilerin, hatalardan arındırılarak üzerinde işlem yapılmaya uygun hale dönüştürülmesini sağlar. Böylece ilişkili veriler sorgulanabilir ve analiz edilebilir hale gelir. Bu işlemlerden başlıcaları veri madenciliği, çok boyutlu analizler, kampanya yönetimi, istatistiksel analizler, sorgulama ve raporlamadır. Başka bir tanımla, farklı kaynaklarda tutulan verilerin ortak bir çatı altında birleştirilerek, verilerin zaman boyutunda birbiri ile konuşmasını sağlayan, tutarlı ve doğru verilerin yer aldığı sistemdir (6).



Şekil 1. Veri Ambarı yapısı (7)

Veri ambarı oluşturma aşamasında verilere temizleme, birleştirme, dönüştürme ve indirgeme gibi bir dizi işlem uygulanır (Şekil 1). Oluşturulduktan sonra veriler periyodik olarak güncelleştirilir.

4.5. Veritabanlarında Bilgi Keşfi Süreci

90'lı yıllardan bu yana bilgisayar bilimlerinde yaşanan gelişmeler, bilişim sistemlerinin hayatın hemen her alanına girmesini sağlamıştır. İşletmeler ve kuruluşlar ihtiyaç duydukları yapıda veritabanları kullanarak çeşitli türlerdeki verileri tutmaktadırlar. Sistemlerle beraber, veritabanlarında tutulan veri türlerinde ve verilerde de büyük artışlar gözlenmiştir. Her an kayıtta olan güvenlik sistemleri, hastanelerde kayıt tutan laboratuvar ve görüntüleme sistemleri, çeşitli işlemlerde kullanılan barkod sistemleri veri tür ve miktarlarının artışına örnek verilebilecek birçok alandan birkaçıdır.

Veritabanları büyük veri kümelerinin kolaylıkla saklanabilmesini sağlar, fakat saklanan verileri analiz etme ve bu verilerden fayda sağlama için başka yollara başvurmak gerekir. Geleneksel sorguların, istatistiksel metotların ve raporlama araçlarının yetersiz kaldığı durumlarda Veritabanlarında Bilgi Keşfi - VBK (Knowledge Discovery in Databases - KDD) olarak adlandırılan süreçlerle bilgi çıkarımı yapılabilmektedir (8). Veritabanlarında bilgi keşfi verilerden doğru, yeni, faydalı ve anlaşılır bilgiler elde edilmesine olanak sağlar.

Keşif, ne olabileceği konusunda önceden belirlenmiş bir fikir ya da hipotez olmadan, veritabanı içerisinde gizli desenleri arama işlemidir. Geniş veri tabanlarında kullanıcının pratik olarak aklına gelmeyecek ve bulmak için gerekli doğru soruları bile düşünemeyeceği birçok gizli desen olabilir. Buradaki asıl amaç, bulunacak desenlerin zenginliği ve bunlardan çıkarılacak bilginin kalitesidir (9).

Veritabanlarında bilgi keşfi, var olan veri içindeki faydalı bilgi ve örüntüleri bulma sürecidir. Süreçte girdi veri, çıktı ise kullanıcı tarafından istenen faydalı bilgilerdir. Süreç sonucunda eldeki veriler belirsiz veya zor anlaşılır olabilir. Sonuçların

yararlılığını ve başarısını arttırmak için süreç akışında konu alanı uzmanları ve teknik uzmanlarla etkileşim içinde olmak gerekebilir.

VBK sürecinin başarı ile tamamlanabilmesi, gerekli olan işlem basamaklarının dikkatle tamamlanmasına bağlıdır. Bu beş basamak aşağıdaki gibidir (8):

- Problemin tanımlanması
- Verilerin Hazırlanması
- Modelin Kurulması ve Değerlendirilmesi
- Modelin Kullanılması
- Modelin İzlenmesi.

Bilgi keşfi sürecinin başarıyla sonuçlanabilmesi, problem alanının iyi tanımlanması, verilerin sağlıklı bir şekilde hazırlanması ve analiz edilmesine bağlıdır.

4.5.1. Problemin Tanımlanması

Bütün uygulama alanlarında olduğu gibi, VBK için de öncelikle problem tanımlanmalıdır. Uygulamanın amacı gereksinimler analiz edilerek açık bir şekilde ifade edilmelidir. Bu aşamaya gerekli özen gösterilmezse istenen sonuçlara ulaşmak mümkün olmayacaktır.

4.5.2. Verilerin Hazırlanması

Veri hazırlamanın amacı veri madenciliği uygulanabilecek formatta veriye ulaşmaktır. Bu aşamanın iyi tasarlanmaması uygulamanın ilerleyen bölümlerinde sık sık bu aşamaya dönmeyi gerektirebilir. Bu aşama bilgi keşfi süreci için ayrılan kaynakların (zaman, enerji,…) büyük bölümünün kullanılmasına neden olur. Çok sayıda uygulamada, veri ön işlemenin bir türünden daha fazlasına ihtiyaç duyulmaktadır. Bu

sebeple veri ön işlemenin türünün belirlenmesi önemli bir iştir. Veri önişleme ile orijinal veriler daha faydalı bir forma dönüştürülür (10). Verilerin hazırlanması toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme alt aşamalarından oluşur (Şekil 2).

4.5.2.1 Veri Toplama

Amaca ulaşmak için gerekli olan veriler birden fazla sayıda kaynak üzerinde olabilir. Bu durumda öncelikle elektronik olmayan ortamdaki veriler elektronik ortama geçirilmeli ve farklı veritabanı veya dosyalardaki veriler birleştirilerek, üzerinde işlem yapılabilecek bir veritabanında toplanmalıdır. Veri madenciliği uygulamaları için verilerin tek bir tabloda toplanmış olması en uygun şekildir.

4.5.2.2. Değer Biçme

Veri toplama aşamasında veriler çeşitli kaynaklardan geldiği için veriler arasında uyumsuzluklar gözlenebilir. Değer biçme aşamasında amaç bu uyumsuzlukların tespiti ve verilerin birbirleriyle ne kadar uyumlu olduklarının değerlendirilmesidir. Uyumsuzluklar; birimler arasındaki farklar, kodlamadan kaynaklanan farklılıklar vb. olabilir.

4.5.2.3. Temizleme ve Birleştirme

Bu aşamada amaç değer biçme aşamasında saptanan uyumsuzlukların giderilmesi ve farklı kaynaklardan gelen verilerin birleştirilmesi ve tek bir veritabanında toplanmasıdır. Veri temizlenirken hatalı girilmiş, tutarsız veya gürültülü veriler silinir, böylece araştırma sonuçlarının olumsuz etkilenmesi en aza indirgenebilir.

Veri temizlemede kullanılabilecek farklı yöntemler vardır. Örneğin eksik verilerin olduğu bir kayıt için boş alanlar birden fazla ise kayıt silinebilir; ya da boş bırakılan alan sayısal bir alan ise diğer kayıtların ortalaması bu alana değer olarak atanabilir. Bazı durumlarda ise gruplama yapmak daha sağlıklı sonuçlar alınmasını sağlayabilir.

Örneğin, gelir durumunu gösteren bir alan için düşük, orta, yüksek gibi sınırlar belirleyerek veriyi sadeleştirmek gibi (11).

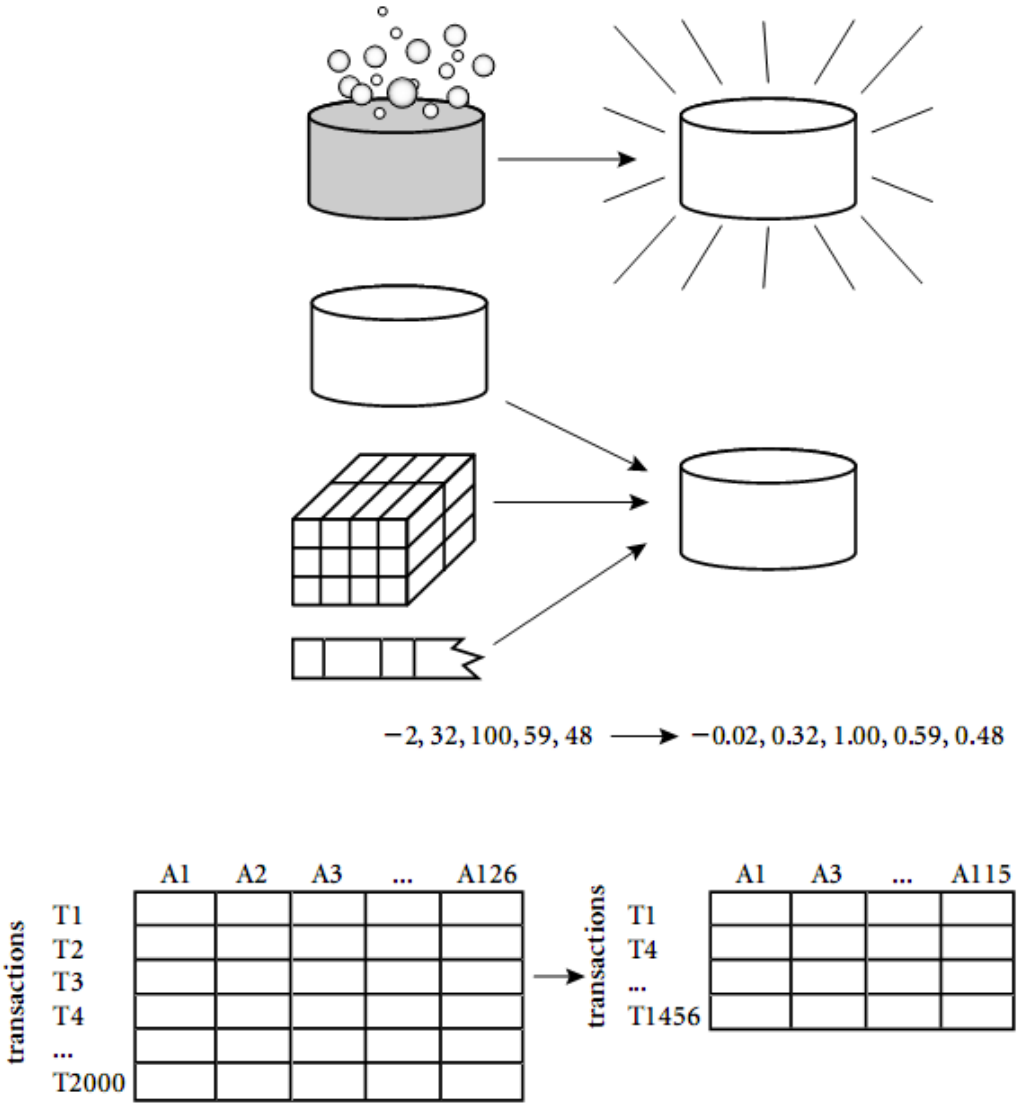
4.5.2.4. Seçme

Bu aşamada geliştirilmesi planlanan veri madenciliği modeline göre veri seçimi yapılır. Örneğin tahmin edici bir model için bu adım, bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır (7).

Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır (7).

4.5.2.5. Dönüştürme

Dönüştürme, veri tiplerinin belirlenen veri madenciliği modelinde kullanılacak algoritmaya uygun biçime dönüştürülmesi işlemidir. Ayrıca kullanılacak verilere ilişkin değişkenlerin uygun formata dönüştürülmesini de kapsar. Örneğin, kredi riskinin tahmini için geliştirilen bir modelde, borç/gelir gibi önceden hesaplanmış bir oran yerine, ayrı ayrı borç ve gelir verilerinin kullanılması tercih edilebilir (12).



Şekil 2. Veri önişleme adımları (5)

4.5.3. Modelin Kurulması ve Değerlendirilmesi

İlk iki aşama başarıyla tamamlandıktan sonra bilgi keşfi sürecinin çekirdeği sayılan model kurma aşamasına geçilebilir. Bu aşama veri madenciliği uygulama adımıdır.

Veri madenciliği (Data Mining-DM); veri ambarlarındaki verileri kullanarak daha önce keşfedilmemiş bilgileri ortaya çıkarma, bunları karar verme ve gerçekleştirme için kullanma sürecidir (9). Başka bir deyişle, veri madenciliği; veritabanı sistemleri, veri depolaması, istatistik, makine öğrenimi gibi alanların kombinasyonundan oluşan multidisipliner bir yaklaşımdır (13). Bilgi keşfi sürecinin çekirdeği sayılan bu adım, son birkaç yıldır bilgi keşfi süreci kavramının yerine kullanılmaktadır; fakat VBK sürecinin aşamalarından yalnızca biridir.

Problemin tanımlanması aşamasındaki analizler ve bu analizler dikkate alınarak belirlenen amaçlar modelin kurulması aşamasında yol gösterici olacaktır. Bu aşamada gerekli olursa verilerin hazırlanması aşamasına tekrar dönülüp, veriler tekniğe daha uygun hale getirilebilir. Veri madenciliği modeli belirlendikten sora modelde çalıştırılacak uygun algoritmanın belirlenmesi gerekmektedir. Her bir veri madenciliği modeli için geliştirilen farklı algoritmalar mevcuttur ve bu algoritmalar denenmeden en uygun olanın seçilmesi zordur. En uygun model ve en uygun algoritma bulunana kadar süreç tekrar eder (14).

Değerlendirme aşamasında modelin etkinliği ölçülür. Elde edilen bilgilerin güvenilir, yeni ve faydalı olup olmadığı incelenir. Bunun için çeşitli yöntemler vardır.

4.5.4. Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak da kullanılabilir.

4.5.5. Modelin İzlenmesi

Zaman içerisinde sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir (8).

4.6. Veri Madenciliği

Büyük miktar ve çeşitlilikte olan ve oldukça hızlı toplanan verilerden çeşitli analizler sonucunda anlamlı bilgiler çıkarılması "veri madenciliği" olarak adlandırılır. Veri madenciliği multidisipliner bir alan olduğu için, net olarak bir tanımı bulunmamakla birlikte, varolan tanımlar da tanımlayıcıların ilgi ve bilgi alanına göre farklılık göstermektedir.

Veri madenciliği; istatistik, veritabanı, örüntü tanıma, makine öğrenmesi alanlarının etkileşimde olduğu yeni bir disiplin ve büyük veritabanlarında önceden tahmin edilemeyen, bilinmeyen ilişkilerin analizidir (15). Başka bir tanıma göre, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır (16).

Diğer bir tanımına göre veri madenciliği, veri sahibi tarafından anlaşılabilecek ve veri sahibine faydalı olacak şekilde, genellikle büyük boyutlu gözlemsel veri setlerinin, var olduğu bilinmeyen ilişkilerin tespit edilmesi ve verinin metinsel bir biçimde özetlenmesi amacıyla analiz edilmesidir. Veri madenciliği işlemi sonucunda bulunan ilişkiler ve elde edilen metinler genellikle model veya örüntü olarak isimlendirilir (17). Çıkarılan örüntüler kullanıldığı alan için önemli, ancak kolay anlaşılmayacak yapıdadır. Örüntüler yapıları gereği içinde ilişkiler, kurallar, değişim düzenleri bulundururlar. Bu örüntüleri veritabanlarında keşfetmek için otomatik veya kimi zaman yarı otomatik yöntemler kullanılır (18).

Veri madenciliği tanımları incelendiğinde, bu tanımlardan ortak olan unsurlardan ilki "çok fazla" miktar ve çeşitlilikte verinin tutulması, ikincisi ise bu verilerden "anlamlı" bilgiler elde edilmesidir (14).

Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır. Veri madenciliği; analistine, iş yapma aşamasında oluşan veriler arasındaki şablonları ve ilişkileri bulması konusunda yardım etmektedir (19).

4.6.1. Veri Madenciliği Kullanım Alanları

Veri madenciliği birçok problem durumuna uygulanabilir farklı yöntemler sunduğundan oldukça geniş bir kullanım alanına sahiptir. Pazarlama, bankacılık, sigortacılık, tıp veri madenciliği uygulamalarının yoğun olarak kullanıldığı alanlardandır. Müşterinin değerlendirilmesi ve satın alma ilişkilerinin analizi, bankalarda kredi taleplerinin değerlendirilmesi ve kredi kartı sahtekarlıklarının belirlenmesi, sigortacılıkta riskli müşteri gruplarının belirlenmesi, tıbbi verinin analiz edilmesi ve karar destek sistemleri geliştirilirken faydalanılan veri madenciliği uygulamaları örnek olarak gösterilebilir.

4.6.2. Veri Madenciliğinde Kullanılan Modeller

Verinin bilgiye dönüştürülmesinde kullanılan modeller tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki başlıkta incelenir (20).

Tahmin edici modellerde, önceden bilinen verilerden yola çıkılarak bir model geliştirilir ve bu model ile yeni durumlar karşısında sonuçlar elde etmek amaçlanır. Örneğin, bir takım şikayetlerle bir polikliniğe başvurarak, hiperlipidemi tanısı konmuş hastaların demografik özellikleri ve plazma lipopreteinleri değerleri gibi klinik bulguları kaydedilir. Burada hastanın özellikleri bağımsız değişken, klinik bulguları ise bağımlı değişken grubundadır. Gelecek dönemlerde hastaneye başvuran kişilerin hastalık risk

analizleri bu veriler kullanılarak hesaplanabilir. Sınıflama ve regresyon, veriyle ilgili, şimdiki veya gelecek değerlerin tahminine yönelik tekniklerdendir.

Tanımlayıcı modellerde ise, verinin genel karakteri keşfedilir, daha sonra keşfedilen örüntülerden faydalanılarak veri hakkında karara varılır. Kümeleme ve birliktelik kuralları tanımlayıcı olarak kullanılan veri madenciliği tekniklerindendir.

Veri madenciliği teknikleri, kullandıkları veri yapılarına ve keşfedebildikleri örüntü biçimlerine göre kategorilere ayrılırlar (5).

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

4.6.2.1. Sınıflandırma ve Regresyon

Sınıflandırma en sık kullanılan veri madenciliği modellerinden birisidir. Sınıflandırma insanın doğasında olan bir eğilimdir; çocuk-genç-yaşlı derken, çalışkan-tembel ayrımı yaparken aslında insanları belli özelliklerine göre ayırarak sınıflandırma yapılmaktadır.

Sınıflamada veri önceden tanımlanmış gruplara ayrıştırılır ve yeni karşılaşılan değişkenin hangi sınıfa atanacağına karar verilir. Sınıflar önceden belirlendiği için denetimli öğrenme tekniklerindendir. Seçilen bir değişken bağımlı değişken, yani tahmin edilen, diğerleri bağımsız değişken, yani tahmin edici, olarak nitelendirilir.

Bağımlı değişken kategorik bir değerse problem sınıflama problemi; sürekli bir değerse regresyon problemi kabul edilir (5). Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen iki veri analiz yöntemidir. Sınıflama ve regresyon modellerinde çok kullanılan tekniklerden bazıları; karar ağaçları, naive-bayes, lojistik regresyon, genetik algoritmalar, k-en yakın komşu, yapay sinir ağlarıdır.

Karar Ağaçları

Karar ağaçları bir nesne ile hedef değeri arasındaki ilişkilerin tahmin edilmesi ve açıklanması konusunda basit ama başarılı bir tekniktir. Karar ağaçları, her bir düğümü bir nitelik değerinin ölçütüne denk gelen, her bir dalı bir ölçütün sonucunu gösteren ve yaprakları da sınıflar veya sınıf dağılımlarını temsil eden akış diyagramı benzeri ağaç yapılardır (5).

Karar ağaçları kolaylıkla sınıflandırma kurallarına dönüştürülebilir. Ayrıca regresyon, kümeleme ve özellik seçimi gibi veri madenciliği işlemlerinde de geniş bir uygulama alanına sahiptir. Karar ağaçlarının kolay okunabilirliği; nominal, nümerik veya metin girdileri kullanabilme esnekliği; hata veya kayıp değer içeren veri kümelerini işlemedeki uyum yeteneği; küçük işlem gücü ile yüksek tahmin performansları gösterebilmesi ve büyük veri kümeleri için kullanışlılığı sayabileceğimiz belli başlı özellikleri arasındadır (21).

Tıp alanında yapılan çalışmalara örnek olarak, bir enfeksiyon hastalığının tedavisinde kullanılan iki ilacın ekonomik yönden değerlendirilmesi için oluşturulan karar ağacı modeli gösterilebilir (22). Bu çalışmaya göre, kullanılan iki ilaçtan ilki, yıllardır bu hastalığın birinci basamak tedavisinde kullanılmaktadır. Ancak bu ilaç bağışıklığı zayıflamış hastalarda yetersiz yanıta neden olmakta ve toksik etkileri bulunmaktadır. Tedavinin ilk aşamada başarısız olduğu durumlarda ise ikinci ilaç, ilk ilaç kadar etkin olup toksisitesi daha düşüktür. Ancak ikinci ilacının fiyatı daha yüksek olduğundan iki ilacın ekonomik açıdan değerlendirilmesi ve tedavide hangi ilacın tercih edileceğine karar verilmesi için bir modele ihtiyaç duyulmaktadır.

Naive-Bayes Yöntemi

Hem tahmin edici, hem de tanımlayıcı bir sınıflama tekniğidir. Her ilişkide koşullu bir olasılık türetmek için bağımlı ve bağımsız değişkenler arasındaki ilişkiyi analiz eder (23).

Geliştirilen bir karar destek sistemi ile hastaların yaş, cinsiyet, kan basıncı ve kan şekeri gibi sağlık profilleri kullanılarak kalp hastalığına yakalanma olasılıkları tahmin edilmiştir (24). Bu karar destek sisteminde kullanılan veri madenciliği tekniği Naive Bayes'tir.

Lojistik Regresyon

Sosyal bilimler alanında yapılan birçok çalışmada, bağımlı olan değişkenin iki mümkün değerden bir tanesine sahip olabileceği varsayılmaktadır. Örneğin; bir hasta akciğer kanseri olabilir veya olmayabilir, tedaviye cevap vermiş olabilir veya olmayabilir. İki mümkün farklı değeri içeren söz konusu veri kümesine iki değerli (binary) denmektedir. İki değerli değişkenler literatürde (0;1) değişkenleri olarak adlandırılmaktadır. Bağımlı değişkenin 0 ve 1 gibi ikili ya da ikiden fazla kategori içeren kesikli değişken olduğu durumlarda, normallik varsayımı bozulmakta olup doğrusal regresyon uygulanamamaktadır.

Çok değişkenli istatistiksel yöntemlerden biri olan lojistik regresyon analizi, sınıflama ve atama işlemlerinde kullanılabilen bir regresyon yöntemi olup normallik, ortak kovaryansa sahip olma gibi çeşitli varsayım bozulmalarının olduğu durumlarda diskriminant ve çapraz tablo analizlerine alternatif olarak uygulanmaktadır.

Lojistik regresyon modelinin tıpta kullanımına örnek olarak, kişinin koroner kalp hastası olma olasılığının değerlendirildiği bir çalışma verilebilir (25). Çalışmaya göre kişinin biyokimya ve hemogram laboratuvar test sonuçlarını analizi edilerek hastalık için tanımlanan referans değerlere göre değişkenlerin önemlilik testleri yapılmış ve önemli bulunan bağımsız değişkenlerin modele dahil edilmesiyle lojistik regresyon

analizi gerçekleştirilmiştir. Elde edilen doğru sınıflama oranları yöntemin umut verici olduğunu göstermektedir.

Genetik Algoritmalar

Evrimsel programlamanın bir parçası olan ve genellikle optimizasyon problemlerinde kullanılan genetik algoritmalar Darvin'in evrim teorisine dayanmaktadır. En iyinin korunumu ve doğal seçilim ilkesinin benzetim yoluyla bilgisayarlara uygulanması ile elde edilen arama yöntemidir. Bugünkü biçimi 1975 yılında J.H.Holland tarafından ortaya konmuştur (26).

Genetik algoritmalarda aday sonuçlar eşit boylu vektörler olarak ifade edilir ve başlangıçta bu vektörlerden bir grup rastlantısal şekilde seçilerek bir popülasyon oluşturulur. Kromozom adı verilen bu vektörler yeni nesiller (jenerasyon) oluşturarak değişikliğe uğrar. Bir kromozomun üzerindeki genler, n boyutlu vektörlerin bir boyutuna denk gelir. Her yeni nesilde kromozomlar amaç fonksiyonuna yerleştirilerek sonuç hesaplanır ve iyiliği ölçülmüş olur. Bir sonraki nesil oluşturulurken tekrarlama, çaprazlama, mutasyon gibi operatörlerden yararlanılarak bazı kromozomlar yeniden üretilir. Bu şekilde en iyi sonuçlar bulunmaya çalışılır (26). Genetik algoritmalar sınıflandırmanın yanı sıra, kümeleme ve birliktelik kurallarında da kullanılabilir (27).

Genetik algoritmaların tıp alanında kullanımına ilişkin birçok örnek bulunmaktadır. Özellikle radyolojik uygulamalarda sık tercih edilen bir tekniktir. Yapılan bir çalışmada abdominal aort anevrizması görülen hastaların radyoloji raporları genetik algoritma ile analiz edilmiştir (28). Ameliyat sonrası hastaların raporları değerlendirilerek ortak desenler tespit edilmiş, istenmeyen durum veya olağandışı özellik gösteren hastaların tespiti bu şekilde kolayca sağlanmıştır.

K-En Yakın Komşu

Bellek tabanlı yöntemler, günümüzde bilgisayar teknolojilerinin gelişimi ile önem kazanmış veri madenciliği teknikleridir. Bu tekniklerin içinde en yaygın olarak bilineni k-en yakın komşu algoritmasıdır.

Veri uzayındaki kayıtların birbirlerine olan uzaklık ölçüleri değerlendirilerek yapılan bir sınıflama tekniğidir. 'k-en yakın komşu' ifadesindeki k, incelenecek en yakın komşuların sayısını verir. Yeni bir kayıt için sınıflama yapılacağı zaman, k tane yakın komşunun davranışı incelenir ve bunların ortalaması nesnenin davranışı kabul edilir (27).

Cleveland Kalp Hastalıkları Veritabanı kullanılarak yapılan bir çalışmada, k-en yakın komşu algoritması ile kalp hastalıklarının teşhisi için bir uygulama geliştirilmiştir. Uygulama sonuçlarına göre %97.4'lere varan doğru sınıflandırma oranları gözlenmiştir (29).

Yapay Sinir Ağları

Yapay sinirin ağları (YSA)'nın kullanımı ilk olarak 1940'larda gündeme gelmiştir; fakat 1980'lere kadar bilgisayarla kullanılmamıştır. Biyolojik sinir ağlarından ayrılması için yapay sinir ağları denmiş ve insan beyni işlevleri göz önüne alınarak modellenmiştir (27). Doğumdan ölüme kadar sürekli olarak öğrenme süreci içerisinde olan insanlar, yaşayıp tecrübe ettikçe sinaptik bağları arasındaki bağlantılar güçlenir, çeşitlenir ve yeni bağlar oluşur. Bu durumun aynısı YSA için de geçerlidir.

YSA, dışarıdan gelen girdilere dinamik olarak yanıt oluşturma yoluyla bilgi işleyen, birbiriyle bağlantılı basit elemanlardan oluşan bilgi işlem sistemidir (30). Bir kuralı veya algoritması olmayan problemlerin çözümünde kullanılır.

YSA herhangi bir girdi vektörünü çıktı vektörüne nasıl dönüştürdüğü hakkında bilgi vermez. Bu özelliğinden dolayı kara kutular olarak düşünülmektedir.

YSA'nın veri madenciliği açısından çeşitli avantaj ve dezavantajları vardır (31). Avantajları, doğru sınıflandırma oranının genelde yüksek olması; kararlı olması yani öğrenme kümesinde hata olması durumunda da çalışması; çıkış vektörünün kesikli, sürekli ya da kesikli veya sürekli değişkenlerden oluşan bir vektör olabilmesi; hem sayısal, hem de kategorik veriler üzerinde işlem yapabilmesidir. Dezavantajları ise, öğrenme süresinin uzun olması ve öğrenilen fonksiyonun anlaşılmasının zor olmasıdır.

YSA öğrenme biçimine göre denetimli (supervised) ve denetimsiz (unsupervised) olmak üzere ikiye ayrılır. Denetimli öğrenmede, ağa giriş-çıkış vektörleri şeklinde detaylı eğitim örnekleri verilmektedir. Denetimsiz öğrenmede ise, ağa giriş bilgileri verilerek, problemin çözümü ağdan beklenmektedir (26).

Akut sağ kasık ağrısı bulunan hastalarda apandisit teşhisinde YSA'nın rolünü değerlendirmek için geliştirilen bir sistemde hasta verileri ağın eğitimi ve test edilmesi için kullanılmıştır. Sonuçlar YSA'nın apandisitin doğru tanısında etkili bir araç olabileceğini ve gereksiz apendektomileri azaltabileceğini göstermiştir (32).

Destek Vektör Makineleri

Destek vektör makineleri (DVM) istatistiksel öğrenme teorisine dayalı parametrik olmayan bir sınıflandırma yöntemidir. DVM ikili sınıflandırmalar için geliştirilmiştir ve bu yöntem ile az sayıda örnekleme verisi ile doğru sınıflandırma sonuçları elde etmek mümkündür. Özellik uzayında sınıflar arasındaki sınırı belirlemek için optimum algoritmanın kullanıldığı bir yöntemdir. Başlangıçta iki sınıflı doğrusal verilerin sınıflandırılması için tasarlanmış olan yöntem daha sonra çok sınıflı ve doğrusal olmayan verilerin sınıflandırılması için geliştirilmiştir. Temel olarak iki sınıfı birbirinden ayırabilen hiper düzlemin belirlenmesi prensibine dayanmaktadır. Bu yöntem kernel fonksiyonları kullanılarak çok boyutlu özellik uzayında çalışma özelliğine sahiptir ve yöntemden elde edilen sonuçlar seçilen kernel ve parametrelerin özelliklerine bağlıdır (26).

DVM'nin tıp alanında kullanımı giderek yaygınlaşmaktadır. Yapılan bir çalışmada tüberküloz tanısı için karar destek sistemi geliştirilmiştir. Tüberküloz tanısı için standart yöntem lekeli balgam sürüntülerinde tüberküloz basilinin görsel olarak tanımlanmasıdır. Görüntü yakalamaya dayalı önerilen teknik görsel nesnenin başarılı sınıflandırılması için kullanılmıştır (33).

Artırılmış Regresyon Ağaçları

Yeni bir veri madenciliği tekniği olan Artırılmış Regresyon Ağaçları (ARA), makine öğrenme tekniklerini ve istatistiği birleştirir. ARA tahmine dayalı veri madenciliği problemlerinde yüksek oranda başarı gösterir. Yöntem, regresyon ağaçları ve artırma (boosting) algoritmalarının güçlü yanlarının birleştirilmesi ile oluşturulmuştur. Regresyon ağaçları karar ağaçları grubundan bir modeldir ve artırma algoritmaları da denetimli öğrenme gerçekleştiren bir makine öğrenmesi algoritmasıdır.

Karar ağaçları sınıflandırma, kümeleme ve tahmin problemlerinde kullanılan bir tahmin aracıdır. Karar ağaçlarının çalışma mantığı 'böl ve yönet' tir; bu nedenle problem çözümü için alt kümelere ayrılmıştır. Karar ağacı bir kök ve hepsi bir soru ile etiketlenmiş olan düğümlerden oluşur. Her düğümden gelen dallar, soruya ilişkin olası cevapları temsil eder. Her yaprak düğümü ise, incelenen problemin çözümü hakkında bir tahmini temsil eder (27). ID3, C4.5 ve CART (sınıflandırma ve regresyon ağaçları) gibi çeşitli yaygın karar ağacı algoritmaları vardır.

Regresyon ağaçları yöntemi bağımlı ve bağımsız değişkenler arasındaki önemli ilişkilerin açıklamasını veren çok değişkenli istatistiksel bir yöntemdir. Regresyon ağaçları genellikle bir veya daha fazla sürekli ve / veya kategorik belirleyici değişkenler ile bir sürekli değişkenin değerini tahmin etmek için kullanılır.

Artırma algoritması, herhangi bir öğrenme algoritmasının performansını artırmak için kullanılabilen genel bir yöntemdir. Yöntem sıralı olarak ağırlıklı bir eğitim kümesini sınıflandırıcı algoritmasına uygulayan ve bu şekilde sıralı olarak ortaya çıkan

sınıflandırıcıların ürettiği sonuçların ağırlıklı çoğunluk oylamasını alan bir sınıflandırma metodolojisidir (34).

Yapılan bir çalışmada, girdi olarak tanı gruplarını ve demografik değişkenleri kullanarak hastaların karışık teşhisleri için sağlık maliyet tahminlerini ayarlayan bir model geliştirilmiştir (35).

4.6.2.2. Kümeleme Analizi

Kümeleme, sınıfların içerisindeki benzerliklerin maksimize edilip, diğer sınıflar ile olan benzerliklerinin ise minimize edilmesi prensibi ile verilerin gruplandırılması olarak tanımlamaktadır. Kümeleme analizine denetimsiz sınıflama da denebilir (8). Başka bir deyişle, kümeleme işlemi heterojen yapıya sahip bir kitleyi daha homojen bir kaç alt gruba ya da kümeye bölme işlemidir. Sınıflamadan farklı olarak, kümeleme analizinde verilerin ayrılacağı sınıflar önceden belli değildir; verinin dağılımına göre oluşturulur. Örneğin, büyük bir mağaza tarafından müşterilerin yaş, gelir, cinsiyet gibi çeşitli özellikleri göz önüne alınarak oluşturulan katalogları müşterilere gönderilecek olsun. Müşteriler kümeleme analizine tabi tutularak benzer özelliktekiler gruplanabilir. Kataloglar bu grupların özelliklerine göre yönlendirilebilir. Tıbbi uygulamada benzer belirtiler gösteren hasta grupları tespit edilebilir.

Literatürde birçok kümeleme algoritması bulunmaktadır. Kullanılacak olan kümeleme algoritmasının seçimi veri tipine ve çalışmanın amacına bağlıdır.

Yapılan bir çalışmada, Kızılderili toplumunda son yıllarda hızla artan AIDS vaka sayısı göz önüne alınarak, farklı HIV risk/koruma gruplarının belirlenmesi için kümeleme analizi yönteminden yararlanılmıştır (36).

4.6.2.3. Birliktelik Analizi

Birliktelik kuralları aynı işlem içindeki nesnelerin birlikte bulunma ya da olayların birlikte gerçekleşme olasılıklarını verir. Yani bir veritabanındaki herhangi bir X, aynı zamanda Y'yi de içeriyorsa, bu bir birlikteliktir (37).

Birliktelik kuralları, market-sepet uygulaması olarak da bilinmektedir. Müşterilerin sepetindeki ürünlerin her biri bir nesne, her bir alışveriş de bir işlemdir. Sepetler analiz edilerek müşterilerin hangi ürünleri birlikte almayı tercih ettikleri keşfedilip, satın alma eğilimleri ortaya konmaktadır. Eğilimler göz önünde bulundurularak raf düzenlemeleri yapılabilir, müşteriler satın almaya özendirilebilirler.

Birliktelik kuralları çıkarılırken kullanıcı tarafından belirlenen eşik değerleri göz önünde bulundurulmalıdır. Dikkate alınması gereken ölçütlerden biri destek değeri; diğeri güven değeridir. Destek ve güven değerlerini sağlayan kurallar değerlendirilmeye alınır. Bir birliktelik kuralı,

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

şeklinde ifade edilir. Bu kuraldan anlaşılması gereken A ürününü satın alanların %60'ı aynı zamanda B ürününü de almıştır. Yapılan tüm satışlar içinde A ve B ürünlerinin birlikte satın alınma oranı da %5'tir. Yani güven değeri, (A ve B'nin bulunduğu satır sayısı) / (A'nın bulunduğu satır sayısı) formülü ile hesaplanır. Destek değeri ise, (A ve B'nin bulunduğu satır sayısı) / (Toplam satır sayısı) formülü ile hesaplanmaktadır (38).

Birliktelik kuralı analizinin tıpta kullanımına ilişkin birçok örnek bulunmaktadır. Biyokimya test sonuçları analiz edilerek hiperlipidemi hastalığının teşhisi için geliştirilen uygulama da bu örneklerden biridir (39).

4.6.2.4. İstisna Analizi

Bir veritabanında verilerin genel davranışlarına uymayan veya veri dağılım modellerinden farklılık gösteren nesneler bulunabilir, bunlara sıra dışı (outlier) denir. Birçok veri madenciliği yöntemi istisnaları gürültü veya aşırı durumlar olarak görür, bu yüzden dikkate almaz. Fakat bazı durumlarda istisna noktalar değerlerine göre çok daha fazla bilgi içerir. Örneğin kredi kartı veya sigorta sahtekarlıklarının tespitinde sık kullanılan bir metottur. Tıp biliminde de yeni bir hastalığın başlangıcını tespit etmede istisna analizi kullanılabilir (40). İstisna analizinde iki yöntem söz konusudur (5).

İstatistik Tabanlı Yöntem

Dağılım analizi ya da standart sapma hesabı gibi istatistiksel yöntemlerle istisna olabilecek noktalar tespit edilir, fakat çok büyük veri yığınlarında yoğun hesaplama gücü gerektirdikleri için performansları sınırlıdır.

Yoğunluk Tabanlı Yöntem

Bu yöntemde her noktanın çevresindeki komşuları ile olan yakınlığı hesaplanır. Yakınlık hesaplamada genelde Öklit uzaklığı kullanılsa da veri türüne göre yakınlık hesaplama yöntemi farklılık gösterebilir. Bu yöntemin temel prensibi 'yeterince komşu olmayan noktaları' tespit etmektir.

4.6.2.5. Evrimsel Analiz

Evrimsel analiz, zamanla davranışları değişen nesnelerin düzenlilik (regularities) ve eğilimlerini (trends) tanımlar ve modeller. Evrimsel analiz ayrımlama, birliktelik ve korelasyon analizi, sınıflama, zamanla ilişkili verilerin kümelenmesi metotlarını içerse de, bu analizin asıl amacı verilerin zamanla ilişkisini analiz etmektir. Bunun için zaman serileri analizi (time-series analysis), ardışık veya periyodik örüntü eşleştirme (sequence or periodicity pattern matching), benzerlik analizi (similarity-based data analysis) gibi yöntemler kullanır (5).

Tıbbi ya da epidemiyolojik kayıtların yokluğunda bulaşıcı hastalıkların yayılım tarihi genellikle patojen genetik sekanslarının evrimsel analizi ile yapılandırılmaktadır. Bir çalışmada 1953'te kayıt altına alınmış Hepatit C virüs ile enfekte olmuş gen sekanslarının evrimsel analizi yapılmıştır (41).

4.6.2.6. Tanımlama ve Ayrımlama

Veriler gösterdikleri ortak özelliklere göre genelleştirilmiş sınıflara ayrılabilir. Bir firma müşteri portföyünü alışveriş ortalaması belirli bir miktardan yüksek olan müşterileri "zengin ", diğerlerini ise "orta halli " ya da "fakir " olarak tanımlayabilir. Bu

tür genellemeler veri kümesini elemanlarının ortak özellikleri ya da veri kümesinin diğer veri kümeleri ile olan farklılıklarını yansıtacak şekilde yapılabilmektedir.

Tanımlama

Bir veri kümesinin elemanlarının genel özelliklerinin özetidir. Tanımlama işleminin çıktıları pasta grafikler, çubuk grafikler, eğriler, çok boyutlu veri küpleri gibi çeşitli şekillerde sunulabilir. Örneğin, büyük bir mağaza yıllık 5000TL ve üzeri harcama yapan müşterilerinin özelliklerini tanımlayıcı bir çalışma yapabilir.

Ayrımlama

Hedef veri kümesi elemanlarının genel özelliklerinin, başka bir ya da birçok veri kümesinin elemanları ile karşılaştırılarak sonuçlar elde edilmesidir.

5. GEREÇ ve YÖNTEM

5.1 Çalışmada Kullanılan Yöntemlerin Detayları

Bu tez çalışması kapsamında gerçekleştirilen uygulamalarda çeşitli veri madenciliği teknikleri kullanılmıştır. Bu teknikler birliktelik kuralı analizi, kümeleme analizi, destek vektör makineleri ve artırılmış regresyon ağaçlarıdır.

5.1.1 Birliktelik Kuralı Analizi

Birliktelik kuralları veriler analiz edilerek aynı işlem içerisinde birlikte görülen nesnelerin çıkarıldığı kurallardır. Geçmiş veriler içerisindeki birliktelik kuralları ortaya konduktan sonra gelecekte olması muhtemel durumlar hakkında da tahmin olanağı sunar. En çok dikkat çeken veri madenciliği tekniklerinden birisidir ve geniş bir kullanım alanına sahiptir. Bu kullanım alanlarının içinde en çok bilineni market sepeti uygulamalarıdır.

Market sepeti analizi ile müşterilerin aldıkları ürünler incelenerek satın alma alışkanlıkları çıkarılır. Böylece market yöneticileri markette müşterilerin alışkanlıkları doğrultusunda ürünlerin raflara dizilimi, indirimli ürünlerin belirlenmesi vb. işlemler yaparak etkili satış stratejileri geliştirebilirler. Mağazadaki tüm ürünlerin kümesi evren olarak düşünülürse, bir işlemde her bir ürünün varlığı ya da yokluğu boolean (ikili) bir değişkenle gösterilir. Her bir alışveriş sepeti de bu değerlerden oluşan bir vektör olarak ifade edilir. Vektörler analiz edilerek hangi ürünlerin sıklıkla birlikte alındığını veya birbiriyle ilişkili olduğunu gösteren satın alma örüntüleri çıkarılabilir (5).

Market sepeti analizi ile birliktelik kuralları çıkarımı ilk olarak Agrawal ve diğerleri tarafından 1993 yılında ele alınmıştır (38). Çalışmada, X ve Y'nin nesne kümesi oluşu,

$X \Rightarrow Y$ (X birliktelik Y)

şeklinde ifade edilmiştir. Bu ifade birliktelik kurallarının matematiksel gösterimidir. Kuralın sol tarafı öncül veya gövde (antecedent) ve sağ tarafı artçıl veya baş (consequent) şeklinde ifade edilebilir. Oluşturulan kuralların gücünü gösteren destek (support) ve güven (confidence) değerleri tanımlanmıştır. Kuralları oluşturabilmek için kullanıcı tarafından belirlenmiş minimum destek (min_destek) ve minimum güven (min_güven) eşik değerlerinden faydalanılır. Market sepet analizinde, nesneler müşteriler tarafından satın alınan ürünlerdir ve bir hareket (transaction) birçok nesneyi içinde bulunduran tek bir satın almadır.

5.1.1.1 Birliktelik Kuralları Temel Kavramlar

Birliktelik Kuralları aynı işlem içinde birlikte gerçekleşen olayları analiz eder. Birliktelik kurallarının matematiksel gösterimi Agrawal, Imielinski ve Swami tarafından 1993 yılında açıklanmıştır (38).

Bu modele göre, $I=\{i_1, i_2, \ldots i_m\}$ nesneler denilen binary (ikili) niteliklerin kümesi olsun. Her i bir nesneyi (ürün) temsil etsin. D veri tabanındaki her bir hareket (transaction) T ile gösterilsin ve $I \supseteq T$ şartını yerine getirecek şekilde nesne küme (itemset) oluştursun. Eğer i_k nesnesi alınmış ise $t_k = 1$, alınmamışsa $t_k = 0$'dır. Yani, $X \subseteq I$ için, X'teki her bir I_k'ya karşılık gelen t_k değeri t_k=1'dir.

$X \Rightarrow I_j$ şeklindeki bir birliktelik kuralından bahsederken; X, I içindeki bazı nesnelerden oluşan bir kümedir, yani I'nın altkümesi ve I_j ise, I içinde bulunan ve X içinde bulunmayan bir elemandır. $X \Rightarrow I_j$ kuralının T işlemler kümesi için geçerli olabilmesi güven faktörünü sağlaması gerekmektedir. Güven değeri (confidence-c) $0 \le c \le 1$ aralığındadır ve T içindeki tüm X'lerin ne kadarının I_j'yi içerdiğinin göstergesidir. Güven seviyesi kuralın gücünün de göstergesidir. $X \Rightarrow I_j \mid c$ şeklinde ifade edilebilir.

Bir kuralın geçerli olabilmesi için önemli olan bir diğer faktör de destek değeridir. Destek değeri de güven gibi sayısal bir değerdir ve T içinde kaç harekette X ve I_j'nin birlikte bulunduğunun ifadesidir.

Daha sade bir ifade ile $A \Rightarrow B$ kuralının geçerliliği için destek ve güven değerleri hesaplanmış olmalıdır. Destek, kuralın öncül ve artçılının birlikte bulunduğu işlemlerin sayısının, veritabanındaki tüm işlemlerin sayısına oranıdır. Güven ise, kuralın öncül tarafında bulunan ürünün/ürünlerin, kuralın artçıl tarafında bulunan ürünü/ürünleri de bulundurma oranıdır. $A \Rightarrow B$ kuralı için güven ve destek değerlerinin matematiksel ifadesi aşağıdaki gibidir (37):

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

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

Destek ve güven değerleri kuralın kullanışlılığını ve kesinliğini ifade eden değerlerdir. Destek ve güven değerleri yüksek olan kurallara güçlü kurallar denir. Kurallar oluşturulurken kullanıcı tarafından minimum destek eşiği (min-destek) ve minimum güven eşiği (min-güven) belirlenir. Üretilen kurallar içinde eşik değerlerinden yüksek destek ve güvene sahip olanlar dikkate alınır.

Örneğin, bir A ürününü alan müşterilerin B ürününü de beraber aldıklarında çıkarılan birliktelik kuralı aşağıdaki gibi gösterilsin:

$A \Rightarrow B$ [destek = %10, güven = %50]

Bu birliktelik kuralı, kullanıcı tarafından belirlenen destek eşiğinin %10'dan, güven eşiğinin de %50'den küçük olması durumunda geçerlidir. Kuraldaki destek ve güven ölçütleri kuralın gücünün göstergesidir. Kural eşitlik (4.1) ve eşitlik (4.2) göz önüne alınarak açıklanırsa, analiz edilen tüm alışverişler içerisinde, %10 oranında A ve B ürünlerinin birlikte satın alındığı ve A ürününü satın alanların %50'sinin aynı alışverişte B ürününü de aldığı söylenebilir. Nesnelerin kümesi nesneküme olarak adlandırılır. Bir nesneküme k tane eleman içeriyorsa, bu kümeye 'k-nesneküme' denir. Eğer bir nesneküme min_destek değerini sağlıyorsa, yaygın nesneküme (frequent itemset) denilir ve L_k şeklinde gösterilir (5).

Birliktelik kuralı analizi iki aşamalı bir süreç olarak nitelendirilebilir (42):

- Destek değeri minimum eşiğin üzerinde kalan tüm nesnekümelerin bulunması: destek değerini sağlayan nesnekümeler yaygın nesnekümelerdir. Bu adımda yaygın nesnekümelerin bulunması hedeflenir. Birliktelik kuralı çıkarım algoritmasının performansı bu adımda belirlenir.

- Yaygın nesnekümelerden güçlü birliktelik kuralları çıkarmak: elde edilen yaygın nesnekümeleri kullanarak min_güven eşiğinin üstünde kalan kuralların çıkarıldığı adımdır.

Birliktelik kurallarında destek ve güven değerlerinin yanısıra, lift (ilgi, kaldırma) değeri de hesaplanır. Lift, basit bir korelasyon ölçüsüdür. Destek ve güven değerlerinin kuralın ilginçliğini ölçmede yetersiz kaldığı durumlarda lift değeri dikkate alınabilir (5). Destek ve güven değeri yüksek çıkmasına rağmen ilginç olmayan kurallar olabilir. Örneğin, cips alan bir müşteri, yüksek ihtimalle kola da alacaktır; fakat bu ilginç bir kural değildir çünkü bilinen bir durumdur (27).

$A \Rightarrow B$ [destek, güven, korelasyon]

Bu birliktelik kuralı destek ve güven faktörlerinin yanı sıra, kuralın korelasyonunun da hesaplandığı durumlarda çıkarılır. Korelasyon ile kuralın içerdiği nesneler arasındaki ilişkiler de ölçülebilir.

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

Lift değerinin 1'den küçük olması, A'nın görülmesinin B'nin görülmesiyle negatif ilişkisinin olduğunu; 1'den büyük olması, A'nın görülmesinin B'nin görülmesi üzerinde pozitif ilişkisinin olduğunu ifade eder -ki bunun anlamı, birinin görülmesi ile diğerinin görülmesi ilişkilidir. Eğer lift değeri 1 çıkmışsa iki taraf birbirinden bağımsızdır.

Kavramların daha anlaşılır olması için bir örnek durum üzerinde açıklayalım. Bir marketten alışveriş yapan müşterin kayıtlarının 10 bin tanesi analiz edilmiş olsun. Bu kayıtların 8 bin tanesinde A ürünü, 6 bin tanesinde de B ürünü alınmış olsun. Ayrıca 4 bin kayıtta da her ikisi birlikte alınmış olsun. Bu analiz için minimum destek değeri %25, güven değeri ise %50 olarak belirlenmiş olsun. Bu şartlarda aşağıdaki birliktelik kuralı elde edilmiş olsun:

$B \Rightarrow A$ [destek = %40, güven = %66, 0.83]

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

Güven ($B \Rightarrow A$) = $\frac{Destek(B \cup A)}{Destek(B)} = \frac{\frac{4000}{10000}}{\frac{6000}{10000}} = 0.66$

Lift ($B \Rightarrow A$) = $\frac{Güven(B \Rightarrow A)}{Destek(A)} = \frac{0.66}{\frac{8000}{10000}}$ = 0.83 olarak bulunur.

Elde edilen birliktelik kuralı min_destek ve min_güven eşiklerinin üzerinde destek ve güven değerine sahip olduğu için güçlü bir kuraldır; fakat A ürününün destek değeri olan %80, kuralın güven değeri %66'dan büyüktür. Yani lift değerinin 1'den küçük çıkmasından da anlaşılacağı gibi, bu kuralda negatif korelasyon vardır (27). Birinin satışı, diğerinin satışını olumsuz etkilemektedir.

Lift değeri hesaplanırken hem öncülün, hem de artçılın destek değerleri göz önünde bulundurulmaktadır. Bu ölçümün olumsuz tarafı simetrik olmasıdır. Yani lift($A \Rightarrow B$) ile lift($B \Rightarrow A$) arasında fark yoktur (27).

Birliktelik kuralı analizinin başarılı ve başarısız olduğu birtakım noktalar vardır (43). Başarılı olduğu noktalar,

- Kolay anlaşılır sonuçlar üretmesi,

- Değişik boyutlardaki veriler üzerinde çalışabilmesi,

- Her ne kadar kayıtların sayısı ve kombinasyon seçimine göre işlem adedi artsa da sepet analizi için her adımda gerekli olan hesaplamalar diğer yöntemlere göre (genetik algoritmalar, yapay sinir ağları vb.) çok daha basittir.

Başarısız olduğu noktalar ise,

- Sorunun boyutu büyüdükçe, gerekli hesaplamaların da üstel olarak büyümesi,

- Kayıtlarda çok az görülen ürünleri göz ardı etmesi,

- Destek ve güven eşikleri bulunan kural sayısına kısıtlama getirmektedir; fakat kullanıcı bu yolla ilgilendiği kuralları kaybedebilir.

5.1.1.2 Birliktelik Kuralı Çıkarımında Kullanılan Algoritmalar

Agrawal ve diğerleri tarafından 1993 yılında ortaya atılan birliktelik analizinin gerçeklenmesi için çeşitli algoritmalar geliştirilmiştir. Bu bölümde bunlardan bazıları işlenecektir.

AIS Algoritması

AIS algoritması Agrawal, Imielinski ve Swami tarafından 1993 yılında veri tabanlarındaki tüm yaygın nesne kümeleri bulmak için geliştirilmiş ve yayınlanmış ilk algoritmadır ve adını geliştiricilerin isimlerinin ilk harflerinden alır (38). Veritabanındaki nesnelerin isimlerinin A'dan Z'ye sıralanması kısıtını taşır.

AIS algoritması veritabanını birçok kez tarar ve her tarama esnasında tüm işlemleri okur. İlk tarama esnasında veritabanındaki nesneleri, teker teker sayarak hangilerinin geniş nesneler (çok rastlananlar, yaygın olanlar) olduğunu belirler.

Bunlardan geniş olanlar aday nesne kümeleri olarak işaretlenir. Bir işlem tarandıktan sonra, bir önceki taramada geniş oldukları belirlenen nesnekümeleriyle, o işlemin nesneleri arasındaki ortak nesne kümeleri belirlenir. Belirlenen bu ortak nesne kümeleri işlemde mevcut olan diğer nesnelerle birleştirilerek yeni aday kümeler oluşturulur. Herhangi bir l nesnekümesi, bir işlemdeki nesnelerle birleşip aday kümelerden birini oluşturabilmesi için, birleşeceği nesnenin hem geniş olması hem de harf sırası açısından nesnekümesi içindeki tüm nesnelerden sonra geliyor olması gerekir (44).

SETM Algoritması

SETM algoritması 1995 yılında Houtsma ve Swami tarafından önerilmiştir. Yaygın kümelerin hesaplanmasında SQL kullanılmasını temel almaktadır (45). Bu algoritmada yaygın nesne kümelerinin her bir üyesi L_k, iki parametreden oluşur; TID (transaction ID) birincil anahtar olmak üzere <TID, nesne kümesi ismi> biçimindedir. Benzer şekilde aday kümelerinin her bir üyesi, C_k da <TID, nesne kümesi ismi> formatında tutulur (45).

AIS algoritması gibi, SETM algoritması da veritabanı üzerinden birçok kez geçerek tarama yapar. İlk taramada, ayrı ayrı her bir nesnenin destek sayısını sayar ve veritabanında hangilerinin büyük veya sık olduğunu bulur. Sonraki taramalarda, bir önceki geçişte işaretlenmiş olan yaygın nesnekümelerini kullanarak aday nesnekümeleri oluşturur. Aday nesnekümelerini oluştururken, aday nesnekümelerinin TID bilgileri de saklanır. Daha sonra aday nesnekümeler isimlerine göre sıralanır ve yaygın olmayanlar silinir. Eğer veritabanı TID numarasına göre sıralanmışsa, bir sonraki tarama esnasında herhangi bir işlemdeki geniş nesnekümeleri L_k 'nın TID numarasına göre sıralanmasıyla elde edilir. Veritabanı bu şekilde yaygın nesnekümelerinin tamamı çıkarılana kadar taranır ve sonlandırılır (44).

Apriori Algoritması

Apriori algoritması, en iyi bilinen ve en çok kullanılan birliktelik analizi algoritmasıdır. Agraval ve Srikant tarafından 1994 yılında geliştirilmiştir (42).

Algoritmanın ismi, bilgileri bir önceki adımdan aldığı için "prior" anlamında Apriori'dir. 'Yaygın nesneküme özelliği' denilen, 'bir yaygın nesnekümenin tüm alt kümeleri de yaygın olmalıdır' ilkesini temel alır (27).

Apriori algoritması, AIS ve SETM algoritmalarından, nesnekümelerin oluşturulması ve sayılacak nesnekümelerin seçilmesi adımlarında farklılaşır. AIS ve SETM algoritmalarında her geçişte, bir önceki geçişteki yaygın nesnekümeler arasındaki ortak nesnekümeleri bulur ve yeni işlemdeki nesneler elde edilir. Ortak nesnekümeler yeni işlemdeki bağımsız nesnelerle genişletilerek aday nesnekümeler oluşturulur. Fakat bağımsız nesneler yaygın olmazlarsa, oluşturulan aday nesnekümeler de yaygın olamazlar. Aprioride ise, bir önceki geçişte oluşturulan yaygın nesnekümeleri birleştirilerek aday nesnekümeleri oluşturulur ve yaygın olmayanlar silinerek yaygın nesnekümeler elde edilir.

Yaygın nesnekümeleri bulan algoritmalar veri üzerinde birçok kez tarama yapar. Aprioride, ilk taramada her bir nesnenin destek sayısı hesaplanır ve bu destek sayıları kullanıcının belirlediği min_destek eşik değeriyle karşılaştırılır. Destek eşiğinin üzerinde kalan nesneler yaygın olarak işaretlenir. Apriori ile k boyutlu bir nesneküme oluşturulurken, 'bir yaygın nesnekümenin tüm alt kümeleri de yaygın olmalıdır' ilkesine dayanılarak (k-1) boyutlu nesnekümeler birleştirilir ve elde edilen alt kümelerden yaygın olmayanlar silinir. Bu şekilde yaygın nesnekümeler elde edilmiş olur.

Apriori bir nesneküme veya işlemdeki nesnelerin alfabetik sırada olduğunu kabul eder. F_k, k boyutlu yaygın nesneküme, C_k da onların adayları olsun. Apriori önce veritabanını 1 boyutlu nesnekümeler için tarar ve belirlenen güven eşik değerini aşanları arar. Daha sonra aşağıdaki 3 adımı tekrar eder (46):

- k boyutlu sık geçen nesnekümelerden, (k+1) boyutlu sık geçen nesnekümelerin adaylarını, C_{k+1}, üret,

- Veritabanını tara ve her bir aday sık geçen nesneküme için destek değerini hesapla,

- Bunlardan min_destek değerini sağlayanları F_{k+1}'e ekle.

Apriori algoritması Şekil 3'te gösterilmiştir. Üçüncü satırdaki apriori-gen fonksiyonu aşağıda verilen iki adımla F_k geniş nesnekümesinden C_{k+1} üretir (46).

Birleşme adımı: k-1 yaygın elemanı olan k boyutlu P_k ve Q_k kümelerini birleştirerek k+1 boyutlu başlangıç aday yaygın nesnekümeler, R_{k+1}, oluşturulur.

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

Burada nesnel < nesne2 < … < $nesne_k$ < $nesne_{k'}$ 'dir.

Budama adımı: k boyutlu R_{k+1} nesnekümelerinin yaygın olup olmadığı kontrol edilir ve bu şartı sağlamayan nesnekümeler çıkarılarak C_{k+1} üretilir. Çünkü C_{k+1}'in k boyutlu nesnekümelerinden yaygın olmayanlar, k+1 elemanlı yaygın olan nesnekümelerin alt kümesi olamazlar.

Algorithm *Apriori*

F_1= (1 elemanlı sık geçen nesnekümeler)
for($k = 1; F_k \neq \phi; k++$) do begin
  C_{k+1} = apriori-gen(F_k); // Yeni adaylar
  for all transactions $t \in$ Database do begin
    C'_t = subset(C_{k+1}, t); // T içindeki adaylar
    for all candidate $c \in C'_t$ do
      $c.count++$;
    end
    $F_{k+1} = \{C \in C_{k+1} \mid c.count \geq$ minimum support $\}$
  end
end
Answer $\cup_k F_k$;

Şekil 3. Apriori algoritması (42)

Beşinci satırdaki altküme fonksiyonu t işlemi içindeki tüm yaygın nesneküme adaylarını bulur. Bundan sonra Apriori veritabanını tarayarak sadece bu şekilde üretilmiş adayların frekanslarını hesaplar (46).

Yaygın nesnekümelerin maksimum boyutu k_{max} ise, Apriori veritabanını en fazla k_{max+1} kere tarar.

Apriori algoritması aday kümelerin sayısını azaltmada iyi bir performans gösterir. Ancak; fazla yaygın nesneküme olması veya çok düşük destek değeri tanımlanması gibi durumlarda, yüksek sayıda aday küme oluşturması ve çok sayıdaki aday nesnekümelerini kontrol etmek için veritabanını tekrar tekrar taraması gibi sorunlarla karşılaşılabilir. Gerçekte, 100 boyutlu bir sık geçen nesnekümenin aday nesnekümeleri 2^{100} adettir.

AprioriTid Algoritması

Birliktelik kuralları hesaplanırken, algoritmalar veritabanı üzerinde tekrar tekrar taramalar yaparlar; fakat bazı aşamalarda tarama yapılmasına gerek yoktur. AprioriTid algoritması bu yaklaşıma uygun olarak, 1994 yılında sunulmuştur (Şekil 4).

1) L_1 = {yaygın 1-nesnekümeler}
2) $\overline{C}_1$ = database $\mathcal{D}$;
3) **for** ($k = 2$; $L_{k-1} \neq \emptyset$; k++) **do begin**
4) C_k = apriori-gen(L_{k-1}); // Yeni adaylar
5) $\overline{C}_k = \emptyset$;
6) **forall** entries $t \in \overline{C}_{k-1}$ **do begin**
7) // Ck'daki aday nesnekümeleri tanımla
// işlemde t.TID ile tanımlanan
$C_t = \{c \in C_k \mid (c - c[k]) \in t.\text{set-of-itemsets} \wedge (c - c[k-1]) \in t.\text{set-of-itemsets}\}$;
8) **forall** candidates $c \in C_t$ **do**
9) c.count++;
10) **if** $(C_t \neq \emptyset)$ **then** $\overline{C}_k$ += $< t.\text{TID}, C_t >$;
11) **end**
12) $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
13) **end**
14) Answer = $\bigcup_k L_k$;

Şekil 4. AprioriTid algoritması (42)

AprioriTid algoritması da aday nesnekümeleri belirlemek için Apriori-gen algoritmasını kullanır. Bu algoritmanın ilginç tarafı, ilk geçişten sonra D veritabanını destek değeri hesaplamak için kullanmamasıdır; onun yerine bu amaçla C_k kümesi kullanılır. C_k'nın her elemanı <TID, $\{X_k\}$> formunda tutulur. Bu açıdan AprioriTid, SETM algoritması ile benzerlik gösterir. TID tanımlayıcı numarası ile bir işlemde yer alan her X_k, potansiyel olarak yaygın k-nesnekümesidir ve k=1 iken C_1, D veritabanına karşılık gelir. k>1 olduğu durumlarda C_k algoritmanın onuncu adımı kullanılarak üretilir. C_k'nın t işlemine karşılık gelen elemanı <t.TID, c> şeklinde gösterilir. Burada $\{ c \in C_k \mid c \}$'dır ve c, t işlemi içerisinde vardır. Eğer bir işlem hiçbir k-nesneküme içermiyorsa, C_k bu işlemden herhangi bir girdi almaz, yani bu işlemin bir TID numarası olmayacaktır. Böylece C_k'daki girdi sayısı, özellikle bu k değerleri için, veritabanındaki işlem sayısından daha küçük olabilir. Bunun dışında yine büyük k değerleri için her girdi kendisine karşılık gelen işlemden daha küçük olabilir. Çünkü o işlemde çok az sayıda aday barınıyor olabilir. Ancak, küçük k değerleri için bunun tersi olacaktır; yani girdiler kendilerine karşılık gelen işlemlerden daha büyük olabilirler (42).

FP-Growth Algoritması

FP-Growth, sık örüntüleri bulmak için kullanılan bir birliktelik algoritmasıdır. Bu algoritmanın önceki çoğu algoritmadan daha etkili bir şekilde çalışarak maliyeti azalttığı görülmüştür. Bunun en büyük nedeni, tüm veritabanını daha küçük ve daha yoğun bir veri yapısı olan sık örüntü ağacı (FP-Tree) içinde tutmasıdır.

Apriori tabanlı algoritmalardan farklı olarak FP-Growth içinde tüm veritabanı sadece iki kez taranır. İlki tüm öğelerin destek değerinin hesaplanması için, ikincisi ise ağaç yapısının oluşturulması içindir. Yeni aday üretimine ve her defasında veritabanının taranmasına gerek kalmadığı için bu algoritma büyük veritabanları için bir kazanımdır. Algoritma esas olarak şu şekilde çalışır: Veritabanı bir kez taranarak her nesnenin destek değeri hesaplanır. Destek değerleri, algoritma içerisinde atanan destek eşik değerine büyük ve eşit olan nesneler büyükten küçüğe sıralanarak bir liste içine konur. Aynı şekilde veritabanında bulunan hareket içerisindeki nesneler de destek değerine göre büyükten küçüğe sıralanır. FP-Tree oluşturmak için öncelikle "root" adında yeni

bir düğüm oluşturulur. Daha sonra her bir hareket (öğeler bahsedilen sırada olmak üzere) ağaç içerisine yerleştirilir. Bu süreç de şu şekilde gerçekleşir: işlem içerisinde yer alan bir öğe eğer ağaçta yoksa o öğe için yeni bir düğüm oluşturulur ve destek değeri 1 yapılır. Destek değerleri de öğelerle beraber tutulur. Eğer o öğe daha önce oluşturulmuşsa sadece o düğümün destek değeri 1 arttırılır. Düğümler arasındaki ilişkiyi tutmak için de bir başlık tablosu (header table) tutulur. Bu tabloda her düğümün başlangıç noktası işaretlenir. Aynı zamanda ağaç içerisindeki aynı düğümler birbirine işaretçilerle bağlanır. Ağaç oluşturulduktan sonra üzerinde FP-Growth algoritması çalıştırılır. Bunun için sıklığı en az olan nesneden başlanır. İçinde o nesnenin geçtiği yollar belirlenir. Her yol için de, o öğenin destek değeri o yolun destek değeri olarak atanır. Bu yollar o nesne için şartlı desen temelini (conditional pattern base) oluşturur. Her bir şartlı desen temelinden şartlı desen ağacı (conditional pattern tree) oluşturulur. Daha sonra bu şartlı desen ağacı üzerinde algoritma rekursif olarak yeniden çalışır. Tablo içindeki her bir öğe için bu süreç tekrarlanır ve böylece sık nesneler kümesi belirlenir (47).

Diğer Algoritmalar

Bu konuda bugüne kadar birçok algoritma geliştirilmiştir. Literatürdeki diğer birliktelik kuralları algoritmaları önceki algoritmalara benzer mantık yürütmektedirler. Aşağıda bu algoritmaların isimleri verilmiştir (48).

- Bu algoritmalardan bir tanesi Apriori ve AprioriTid algoritmalarının bir karışımı olan Apriori-Hybrid algoritmasıdır (42).

- Geniş nesne kümelerini belirlemek için veritabanından alınmış küçük örneklerin çok iyi sonuçlar verebileceği fikrine dayanan OCD (Off-line Candidate Determination-Sıradışı Aday Belirleme) algoritması (49).

- Veritabanını küçük parçalara bölerek, bellekte işgal edilen yeri azaltıp daha hızlı sonuca ulaşma sağlayan bölümleme (partitioning) tekniği (50).

- Kullanıcıya her taramadan sonra oluşan kuralları gösterip, minimum destek ve güven seviyelerini değiştirme olanağı veren CARMA (Continuous Association Rule Mining Algorithm-Sürekli Bağlantı Kuralı Madenciliği) (51).

- Veri paralelliğine dayanan CD (Count Distribution – Sayım Dağılımı) (52).

- PDM (Parallel Data Mining - Paralel Veri Madenciliği) (53).

- DMA (Distributed Mining Algorithm-Dağıtılmış Madencilik Algoritması) (54).

- CCPD (Common Candidate Partitioned Database-Ortak Aday Bölünmüş Veritabanı) (55).

- IDD (Intelligent Data Distribution) (56).

- HPA (Hash-based Parallel Mining of Association Rules-Birliktelik Kurallarının Çırpı Temelli Paralel Madenciliği) (57).

- PAR (Parallel Association Rules-Paralel Birliktelik Kuralları) (58).

5.1.2 Kümeleme Analizi

Kümeleme analizi, sınıflandırma yapılması açısından sınıflandırma ile benzerdir; fakat sınıflandırmanın aksine kümelemede gruplar önceden belli değildir. Kümeleme analizi nesnelerin ortak özellikleri arasındaki mesafenin kümeler içinde minimum, kümeler dışında ise her kümenin diğerleri ile arasındaki mesafenin maksimum olmasını sağlayan bir tekniktir. Başka bir deyişle küme içindeki her bir nesne birbirine çok benzer iken, diğer küme elemanlarına ise benzememektedir. Benzerlikler veya farklılıklar probleme konu olan nesneleri tanımlayan özellikleri aracılığı ile belirlenmektedir. Veritabanlarında toplanan verilerin büyük miktarlarda olması

nedeniyle, kümeleme analizi son zamanlarda veri madenciliği araştırmalarında son derece aktif bir konu haline gelmiştir (5).

Kümeleme analizi, hemen hemen tüm bilim alanlarında yararlanılan bir yöntemdir. Tıp, biyoloji, psikoloji, sosyoloji, arkeoloji gibi belirsizlik koşullarının ve karmaşık oluşumların bulunduğu bilim alanlarında ise daha yoğun olarak yararlanılan bir yöntemdir. Örneğin, tıp alanında; hastalıkların sınıflandırılması, psikiyatride; paranoya, şizofreni gibi semptomların doğru sınıflandırılması (teşhis profilleri oluşturması), laboratuar bulguları ile klinik bulguların oluşturduğu veri matrislerinden hastalık alt gruplamalarının ya da yeni semptomların tanımlanması gibi amaçlarla kümeleme analizinden yararlanılmaktadır (59).

Kümeleme analizi -mesafeye dayalı küme analizine odaklanarak- yaygın olarak uzun yıllar boyunca çalışılmıştır. K-means, k-medoids ve başka birkaç yönteme dayalı kümeleme analizi araçları birçok istatistiksel analiz yazılım paketlerinde yapılandırılmıştır. Makine öğrenmesi alanında, kümeleme analizi bir denetimsiz öğrenme örneğidir (5).

Kümeleme analizi, potansiyel uygulama alanlarının kendine özgü gereksinimleri bulunduğundan zor bir araştırma alanıdır. Bu gereksinimler, büyük veritabanları için ölçeklenebilirlik özelliği, farklı nitelik tipleri ile başa çıkabilme, küresel kümelerin yanı sıra rastgele şekilli kümelerin de keşfi, giriş parametrelerinin belirlenebilmesi için yeterli alan bilgisi gereksinimi, gürültülü veri ile başa çıkabilme, veri boyutunun artması durumuna uyabilme ve giriş verisinin sırasına duyarlı olmama, çok boyutluluk, yorumlanabilirlik ve kullanılabilirliktir (5).

Kümeleme analizi, temel olarak dört değişik amaç için uygulanır. Bu amaçlar aşağıdaki gibi sıralanabilir (60):

- n sayıda birimi, nesneyi, oluşumu, p değişkene göre saptanan özelliklerine göre olabildiğince kendi içinde türdeş ve kendi aralarında farklı alt gruplara ayırmak,

- p sayıda değişkeni n sayıda birimde saptanan değerlere göre ortak özellikleri açıkladığı varsayılan alt kümelere ayırmak ve ortak faktör yapıları ortaya koymak,

- Hem birimleri hem de değişkenleri birlikte ele alarak ortak n birime p değişkene gören ortak özellikli alt kümelere ayırmak,

- Birimleri, p değişkene göre saptanan değerlere göre, izledikleri biyolojik ve tipolojik sınıflamayı ortaya koymak.

Kümeleme analizinin aşamaları aşağıdaki şekilde sıralanabilir (60):

Birimler arasında var olan benzerliğin belirlenebilmesi için kullanılacak ölçülerin ve değişkenlerin belirlenmesi,

- Birimler arasındaki benzerliklerin belirlenmesinden sonra birimlerin kümelenmesi,

- Oluşturulan kümelerin uygun olup olmadığının belirlenmesi,

- Kümelerin uygun olarak elde edildiği varsayımı altında bunun istatistik geçerliliğinin ortaya konması.

Kümeleme dinamik bir veri madenciliği araştırma alanıdır ve birçok kümeleme algoritması geliştirilmiştir. Bunlar bölümlemeli metotlar, hiyerarşik metotlar, yoğunluk tabanlı yöntemler, grid tabanlı yöntemler, modele dayalı yöntemler, çok boyutlu veri için yöntemler ve kısıt tabanlı yöntemler olarak kategorize edilebilir. Bazı algoritmalar birden fazla kategoriye dahil olabilmektedir (5).

5.1.2.1 Kümeleme Analizinde Veri Tiplerine Göre Benzerlik ve Uzaklık Ölçüleri

Başlıca bellek tabanlı kümeleme algoritmaları genellikle data matriksi ve farklılık matriksi olmak üzere iki veri yapısı ile çalışabilir. Data matriksi p tane özelliği bulunan

n tane nesneden oluşurken; farklılık matriksi n objelerinin tüm çiftlerinin kullanabileceği yakınlıklar koleksiyonunu saklar. Farklılık matrisinde d(i,j), i ve j nesnelerinin arasında ölçülen fark ya da farklılık olsun. d(i,j), i ve j nesnelerinin büyük oranda birbirleri ile benzer olmaları durumunda sıfıra yakın negatif olmayan bir sayıdır, farklılık arttıkça daha büyük bir sayı haline gelir (5).

Bir veri setinde yer alan birimlerin kümelenmesi işlemi bu birimlerin birbirleriyle olan benzerlikleri ya da birbirlerine olan uzaklıkları kullanılarak gerçekleştirilmektedir. Değişkenlerin kesikli ya da sürekli olmalarına ya da değişkenlerin nominal, ordinal, aralık ya da oransal ölçekte olmalarına göre hangi uzaklık ölçüsünün ya da hangi benzerlik ölçüsünün kullanılacağına karar verilir. Aşağıdaki alt bölümlerde değişkenlerin türlerine göre kullanılan uzaklık veya benzerlik ölçülerine yer verilecektir.

Aralık ölçekli değişkenler arasındaki karşılaştırma oranıyla değil fark ile yapılır. Oransal ölçekli değişkenlerse; alan, uzunluk gibi aralarında kıyaslama yapılırken bir oran belirtebileceğimiz değişkenlerdir. Aralık ve oransal ölçekteki değişkenleri bulunan birimleri kümelerken, değişik uzaklık ölçüleri kullanılır. Bu ölçülerden bazıları: Öklidyen, Manhattan, Minkowski uzaklık ölçüleridir. Uzaklık formülleri p boyuta sahip 2 birim arasındaki uzaklığa göre verilmiştir.

Öklidyen Uzaklık Ölçüsü

Öklidyen uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık

$$d(i, j) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{ip} - x_{jp}\right)^2} \qquad (5.4)$$

formülüyle hesaplanır. En çok kullanılan ölçü birimidir (5).

Pearson Uzaklık Ölçüsü

Pearson uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık

$$d(i,j)=\sqrt{\left(x_{i1}-x_{j1}\right)^2/S_1^2+\left(x_{i2}-x_{j2}\right)/S_2^2+\cdots+\left(x_{ip}-x_{jp}\right)^2/S_p^2} \quad (5.5)$$

formülü ile hesaplanır. Bu formülde kullanılan S_p, uzaklığın hesaplandığı değişkene ait varyanstır. Bununla birlikte farklı gruplar hakkında önceden bilgi sahibi olunmadığı için, uzaklık hesaplanmasında S değerinin kullanılması doğru olmaz. Bu nedenle Pearson uzaklık ölçüsü yerine genellikle Öklidyen uzaklık ölçüsü tercih edilir. Pearson uzaklık ölçüsüne, "karesel Pearson uzaklık" ya da "standardize öklid uzaklığı" adı da verilir (5).

Manhattan Uzaklık Ölçüsü

Manhattan uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık

$$d(i,j)=\left(\left|x_{i1}-x_{j1}\right|+\left|x_{i2}-x_{j2}\right|+\cdots+\left|x_{ip}-x_{jp}\right|\right) \quad (5.6)$$

formülü ile hesaplanır. Bu ölçü de birimler arasındaki mutlak uzaklık kullanılır. Manhattan uzaklık ölçüsüne, "city block uzaklık ölçüsü" adı da verilir (5).

Minkowski Uzaklık Ölçüsü

Minkowski uzaklık ölçüsü genel bir formüldür. Formülde yer alan m değerinin alacağı farklı değerlere göre yeni formüller türetir. Minkowski uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık aşağıdaki formül ile hesaplanır (5).

$$d(i,j)=\left[\left|x_{i1}-x_{j1}\right|^m+\left|x_{i2}-x_{j2}\right|^m+\cdots+\left|x_{ip}-x_{jp}\right|^m\right]^{1/m} \quad (5.7)$$

Minkowski uzaklık ölçüsündeki m değeri büyük ve küçük farklara verilen ağırlığı değiştirir. m = 1 değerini alırsa, formül Manhattan uzaklık ölçüsünün formülüne; m = 2 değerini alırsa, Öklidyen uzaklık ölçüsü formülüne dönüşür.

Uzaklık Fonksiyonun Özellikleri

1. $d(i, j) \geq 0$; Uzaklık negatif değil
2. $d(i,i) = 0$; Her birim kendisine olan uzaklığı sıfırlar.
3. $d(i, j) = d(j,i)$; Uzaklık fonksiyonu simetriktir.
4. $d(i, j) \leq d(i,h) + d(h, j)$; iki birimin arasındaki uzaklık bu iki birimin üçüncü bir birime olan uzaklıkları toplamından küçük olamaz (üçgen eşitsizliği) (5)

Değişkenlerin Standardizasyonu ve Değişkenlerin Dönüştürülmesi

Veri matrisinde yer alan değişkenlerin ortalamaları ve varyansları birbirilerinden oldukça farklı olduğu durumlarda birimler arası uzaklık hesaplanırken, ortalaması daha büyük ve/veya varyansı daha büyük olan değişkenler, hesaplanılan uzaklık değerine daha büyük etki yapacaktır. Bu durumda kümeler oluşturulurken sistematik bir yanlış yapılmış olacaktır. Ayrıca değişkenlerde yer alan aşırı değerlerde uzaklık değerine etki eden başka bir faktördür. Aşırı değerler kümeleme analizi sonucunda ayrı kümeler olarak karşımıza çıkabilir. Bu gibi durumlar söz konusu olduğunda yapılacak işlem, değişkenlerin dönüştürülmesidir. Verilerin dönüştürülme işlemi, standartizasyon ya da belirli aralıklara indirgeme biçiminde gerçekleşir. Verilerin dönüştürülmesinde kullanılan bazı teknikler; z skorlarına dönüştürme, standart sapma bir olacak şekilde indirgeme, maksimum değer bir olacak biçimde indirgemedir.

Nominal Ölçekli Değişkenler

Nominal ölçekli veriler arasında büyüklük yada küçüklük gibi herhangi bir karşılaştırmanın yapılamayacağı; sadece saç rengi, cinsiyet, medeni durum gibi nitel değişkenlerin şıklarını 0,1,2 vb. kesikli nicel değerlere atanması ile oluşan değişkenlerdir. Nominal ölçekli değişkenleri kümeleme analizinde incelerken, ikili, ikili olmayan nominal ölçekli değişkenler olarak ikiye ayırabiliriz.

İkili (Binary) Değişkenler

Erkek–kadın, evet–hayır gibi salt ikili değer alan nominal ölçekli değişkenlerdir. Bu değişkenlerde birimler arası uzaklık hesaplanırken aralık ve/veya oransal ölçekli verilerin hesaplanmasında kullanılan Pearson, Öklid, Manhattan (City block), Minkowski uzaklık ölçüleri tercih edilmez. Bu tür değişkenlerin kümelenmesinde, kontenjans tablolarının yardımıyla bulunan, benzerlik ölçüleri adı verilen ölçüler kullanılır.

Kontenjans tablosundan elde edilen 1-1, 0-0, 0-1, 1-0 eşleşmelerin frekansları kullanılarak sözü edilen benzerlik ölçüleri hesaplanır (5).

İkili Olmayan Nominal Değişkenler

Şıklarında iki durumdan daha fazla durum bulunduran değişkenlerin uzaklıklarının hesaplanması ise şu formülle sağlanır (5):

$$d(i, j) = \frac{p - m}{p} \tag{5.8}$$

p: toplam değişken sayısı, m: eşleşen değişken sayısı olmak üzere eşleşmeyen değişken sayısı (p-m) toplam değişken sayısına oranlanarak iki birim arasındaki uzaklık bulunur. Bu uzaklık 1' den çıkarılırsa, iki değişken arasındaki benzerlik katsayısı bulunur.

Ordinal Ölçekli Değişkenler

Ordinal değişkenler, aralarında bir sıra bulunan, ama sıralar arasındaki farkın ya da oranın yalnızca hangi birimin ön sırada yer aldığını gösteren kesikli değişkenlerdir. Örneğin, bir kurumdaki kişilerin unvanlarına göre sıralanması ordinal ölçeklidir. Ordinal ölçekli değişkenlere sahip birimlerin kümelenmesi 3 aşamada gerçekleşir (5).

- xf değişkenine ait i. birimin sıralaması M birim içerisinde r_{if} olur.

- Her bir ordinal değişken farklı sayıda durumlara sahip olacağından, her bir değişken aşağıdaki formül ile [0,1] aralığına indirgenir.

$$z_{if} = \frac{r_{if} - 1}{M_f - 1} \tag{5.9}$$

- [0,1] aralığına indirgenen değişkenlere ait birimler arasındaki uzaklıklar oransal ve aralık ölçekte kullanılan uzaklık ölçüleri ile bulunur.

5.1.3 Destek Vektör Makineleri

Destek Vektör Makineleri (DVM) istatistiksel öğrenme teorisine dayalı bir kontrollü sınıflandırma algoritmasıdır (61). Destek vektör ağları, Vapnik öncülüğünde bir dizi çalışma sonucu 1990'lı yıllarda ikili sınıflandırma için geliştirilmiştir. Bu yöntemin amacı, iki sınıfın vektörleri arasındaki en büyük uzaklık ile doğrusal karar fonksiyonu olarak bir optimal hiper düzlem oluşturmaktır (62). DVM'nin sahip olduğu matematiksel algoritmalar başlangıçta iki sınıflı doğrusal verilerin sınıflandırılması problemi için tasarlanmış, daha sonra çok sınıflı ve doğrusal olmayan verilerin sınıflandırılması için genelleştirilmiştir. DVM'nin çalışma prensibi iki sınıfı birbirinden ayırabilen en uygun karar fonksiyonun tahmin edilmesi, başka bir ifadeyle iki sınıfı birbirinden en uygun şekilde ayırabilen hiper-düzlemin tanımlanması esasına dayanmaktadır (63,64).

5.1.3.1 Doğrusal Ayrılabilen Veriler İçin DVM

Destek vektör makineleri ile sınıflandırmada genellikle {-1,+1} şeklinde sınıf etiketleri ile gösterilen iki sınıfa ait örneklerin, eğitim verisi ile elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır. Söz konusu karar fonksiyonu kullanılarak eğitim verisini en uygun şekilde ayırabilecek hiper-düzlem bulunur (61).

İki sınıflı verileri birbirinden ayırabilen birçok hiper-düzlem çizilebilir. Ancak DVM'nin amacı kendisine en yakın noktalar arasındaki uzaklığı maksimuma çıkaran hiper-düzlemi bulabilmektir. Sınırı maksimuma çıkararak en uygun ayrımı yapan hiper-

düzleme optimum hiper-düzlem ve sınır genişliğini sınırlandıran noktalar ise destek vektörleri olarak adlandırılır (61).

Doğrusal olarak ayrılabilen iki sınıflı bir sınıflandırma probleminde DVM'nin eğitimi için k sayıda örnekten oluşan eğitim verisinin $\{x_i,y_i\}$, i = 1,...,k olduğu kabul edilirse, optimum hiperdüzleme ait eşitsizlikler aşağıdaki şekilde olur:

$$w.x_i + b \geq +1 \text{ her } y = +1 \text{ için} \tag{5.10}$$

$$w.x_i + b \leq +1 \text{ her } y = -1 \text{ için} \tag{5.11}$$

Burada $x \in R^N$ olup N-boyutlu bir uzayı, y ∈ {-1, +1} ise sınıf etiketlerini, w ağırlık vektörünü (hiper-düzlemin normali) ve b eğilim değerini göstermektedir (65). Optimum hiper-düzlemin belirlenebilmesi için bu düzleme paralel ve sınırlarını oluşturacak iki hiper-düzlemin belirlenmesi gerekir. Bu hiper-düzlemleri oluşturan noktalar destek vektörleri olarak adlandırılır ve bu düzlemler $w.x_i + b = \pm 1$ şeklinde ifade edilirler. Optimum hiper-düzlemin sınırının maksimuma çıkarılması için $\|w\|$ ifadesinin minimum hale getirilmesi gerekir. Bu durumda en uygun hiperdüzlemin belirlenmesi aşağıdaki sınırlı optimizasyon probleminin çözümünü gerektirir.

$$min\left[\frac{1}{2}\|w\|^2\right] \tag{5.12}$$

Buna bağlı sınırlamalar ise,

$$y_i(w.x_i + b) - 1 \geq 0 \text{ ve } y_i \in \{-1,1\} \tag{5.13}$$

şeklinde ifade edilir (63). Bu optimizasyon problemi Lagrange denklemleri kullanılarak çözülebilir. Bu işlem sonrasında;

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{k} \alpha_i y_i (w.x_i + b) + \sum_{i=1}^{k} \alpha_i \tag{5.14}$$

eşitliği elde edilir. Sonuç olarak, doğrusal olarak ayrılabilen iki sınıflı bir problem için karar fonksiyonu aşağıdaki şekilde yazılabilir (65):

$$f(x) = sign\left(\sum_{i=1}^{k} \lambda_i y_i (x.x_i) + b\right) \tag{5.15}$$

5.1.3.2 Doğrusal Ayrılamayan Veriler İçin DVM

Günlük hayattaki birçok problemde birçok problemde verilerin doğrusal olarak ayrılması mümkün değildir. Bu durumda eğitim verilerinin bir kısmının optimum hiper-düzlemin diğer tarafında kalmasından kaynaklanan problem pozitif bir yapay değişkenin (ξ_i) tanımlanması ile çözülür. Sınırın maksimum hale getirilmesi ve yanlış sınıflandırma hatalarının minimum hale getirilmesi arasındaki denge pozitif değerler alan ve C ile gösterilen bir düzenleme parametresi ($0 < C < \infty$) tanımlanmasıyla kontrol edilebilir (62). Düzenleme parametresi ve yapay değişken kullanılarak doğrusal olarak ayrım yapılamayan veriler için optimizasyon problemi:

$$min\left[\frac{\|w\|^2}{2} + C.\sum_{i=1}^{r} \xi_i\right] \tag{5.16}$$

şeklini alır. Buna bağlı sınırlamalar ise,

$$y_i(w.\varphi(x_i) + b) - 1 \geq 1 - \xi_i \qquad \xi_i \geq 0 \text{ ve } i = \{1, \dots, N\} \tag{5.17}$$

şeklinde ifade edilir. Eşitlik 5.16 ve 5.17'de ifade edilen optimizasyon probleminin çözümü için girdi uzayında doğrusal olarak ayrılamayan veri, özellik uzayı olarak tanımlanan yüksek boyutlu bir uzayda görüntülenir. Böylece verilerin doğrusal olarak ayrımı yapılabilmekte ve sınıflar arasındaki hiper-düzlem belirlenebilmektedir.

DVM ile matematiksel olarak $K(x_i, x_j) = \varphi(x).\varphi(x_j)$ şeklinde ifade edilen bir kernel fonksiyonu yardımıyla doğrusal olmayan dönüşümler yapılabilmekte ve bu şekilde verilerin yüksek boyutta doğrusal olarak ayrımına imkan sağlanmaktadır. Sonuç olarak, kernel fonksiyonu kullanarak doğrusal olarak ayrılamayan iki sınıflı bir problemin çözümü ile ilgili karar kuralı aşağıdaki şekilde yazılabilir (65):

$$f(x) = sign(\sum_i \alpha_i y_i \varphi(x).\varphi(x_i) + b) \quad (5.18)$$

DVM ile gerçekleştirilecek bir sınıflandırma işlemi için kullanılacak kernel fonksiyonu ve bu fonksiyona ait optimum parametrelerin belirlenmesi esastır. Literatürde kernel fonksiyonu olarak en sık kullanılan fonksiyonlar polinom fonksiyonu, radyal tabanlı fonksiyon, doğrusal fonksiyon ve sigmoid fonksiyonudur. Kernel fonksiyonları karşılaştırıldığında polinom ve radyal tabanlı kernellerin daha sade ve anlaşılabilir olduğu ifade edilebilir. Matematiksel olarak basit görünse de, polinomun derecesindeki artış algoritmanın karmaşık bir hal almasına neden olmaktadır. Bu da hem işlem süresini önemli ölçüde artırmakta hem de bir noktadan sonra sınıflandırma doğruluğunu düşürmektedir. Buna karşın radyal tabanlı fonksiyonun kernel boyutu (γ) olarak ifade edilen parametresindeki değişimlerin sınıflandırma performansına etkisinin daha az olduğu görülmüştür (66).

Kernel fonksiyonuna özgü parametrelerin yanında tüm destek vektör makineleri için düzenleme parametresi C'nin kullanıcı tarafından belirlenmesi gerekir. Bu parametre için olması gerekenden çok küçük veya çok büyük değerler seçilmesi durumunda optimum hiper-düzlem doğru belirlenemeyeceğinden sınıflandırma doğruluğunda ciddi düşüş beklenir. Diğer taraftan C = ∞ olması durumunda DVM modeli sadece doğrusal olarak ayrılabilen veri setleri için uygun hale gelir. Buradan da görüleceği üzere parametreler için uygun değerlerin seçimi DVM sınıflandırıcısının performansını direkt olarak etkileyen bir faktör durumundadır. Genellikle deneme ve hata stratejisi kullanılmasına karşın, çapraz doğrulama yaklaşımı başarılı sonuçlara ulaşılmasına olanak sağlamaktadır. Çapraz doğrulama yaklaşımında amaç oluşturulan sınıflandırma modelinin performansının belirlenmesidir. Bu amaçla veri seti iki kısma ayrılır. Birinci kısım sınıflandırmaya esas olan model oluşumunda eğitim verisi olarak kullanılırken ikinci kısım modelin performansının belirlenmesi amacıyla test verisi olarak işleme konur. Eğitim seti ile oluşturulan modelin test veri setine uygulanması sonucunda doğrusınıflandırılan örneklerin sayısı sınıflandırıcının performansını gösterir. Dolayısıyla çapraz geçerlilik yöntemi kullanılarak en iyi sınıflandırma performansının elde edildiği kernel parametrelerinin belirlenmesi ve sınıflandırmaya esas olacak modelin oluşturulmuştur (61).

5.1.4 Artırılmış Regresyon Ağaçları

Artırılmış Regresyon Ağaçları (ARA) birçok modelle uyum sağlayarak tek bir modelin performansını artırmayı ve bunları tahminler için kullanmayı hedefleyen tekniklerden biridir (67). Yeni bir teknik sayılan ARA, makine öğrenmesi ve istatistik alanlarının anlayışlarını ve tekniklerini birleştirir. ARA temelde tahmin performansını iyileştirmek için çok sayıda uyarlanabilir basit ağaç modellerini birleştirmek yerine tek bir en iyi model üreterek geleneksel regresyon yöntemlerinden ayrılır (68,69). ARA iki algoritma kullanır; biri sınıflandırma ve karar ağacı modellerinden regresyon ağaçları, diğeri ise modellerle birleşerek yeni modeller inşa eden artırma (boosting) algoritmasıdır (67).

5.1.4.1 Karar Ağaçları

Ağaç tabanlı modeller, yordayıcılara en homojen cevapları veren bölgeleri belirlemek için birçok kural kullanarak yordayıcı alanları dikdörtgenlere bölerler. Daha sonra, kısıt olarak en olası sınıfa uyan sınıflandırma ağacı ve bu bölgede gözlem için ortalama yanıta uygun regresyon ağaçları ile, hataların normal dağıldığını varsayarak, her bölgeye bir kısıt atanır. Yordayıcılar ve bölünme noktaları tahmin hatalarını minimize edecek şekilde seçilir. Tek bir karar ağacı oluşturulması için büyük bir ağaç oluşturup, daha sonra çapraz doğrulama ile belirlenen zayıf bağlantıları budamak etkili bir yöntemdir (70).

Karar ağaçları veriyi kolay görselleştirebildikleri ve sezgisel yolla sunabildikleri için popülerdir. Aday yordayıcıların hazırlanması kolaydır, çünkü yordayıcı değişkenleri herhangi bir tipte olabilir ve model çıktıları monoton dönüşüm işlemlerinden etkilenmezler ve alakasız yordayıcılar pek seçilmezler. Ağaçlar istisnalara karşı duyarsızdırlar. Bir ağacın hiyerarşik yapısı, bir giriş değişkenine verilen yanıtın ağaçtaki daha yüksek değerli giriş değişkenlerine bağlı olduğu anlamına gelir, böylece yordayıcılar arasındaki etkileşimler otomatik olarak modellenir. Bu özelliklerine karşın, düzgün fonksiyonların modellenmesi güçtür. Ayrıca ağacın yapısı verinin örneklemine bağlıdır ve eğitim verisindeki küçük değişiklikler çok farklı bölünme noktası serileri ile

sonuçlanabilir. Bu faktörler ağaçların dezavantajlarıdır ve tahmin etme gücünü düşürebilirler (70).

5.1.4.2 Artırma Algoritması

Artırma, çok sayıda kaba kural bulup ortalamalarını almanın, doğru tahmin başarısı yüksek tek bir model bulmadan daha kolay olması fikrine dayanan, model başarısını geliştiren bir modeldir (71). Artırma algoritması modeli inşa edip sonuçları birleştiren diğer modellerden farklı olarak; ardışık, ileri dönük etapları olan bir tekniktir. Artırmada, ağaçların mevcut koleksiyonlarına bakarak, kötü modellenmiş gözlemlere vurguyu arttırmak için modeller uygun yöntemler kullanılarak eğitim verilerine yinelemeli şekilde uyarlanırlar. Artırma algoritmaları uyumsuzluğu ölçüp, sonraki yineleme için ayarları değiştirirler. AdaBoost gibi orjinal artırma algoritmaları iki sinifli sınıflandırma problemleri için geliştirilmiştir. Bu algoritmalar kötü modellenmiş olanlara vurgu yaparak, gözlemlerini ağırlıklandırırlar (67).

5.2 Laboratuvar Verilerinde Veri Madenciliği Uygulamaları

Bu bölümde veri madenciliği yöntemlerinin tıbbi verilere uygulanması ile, belirlenen birtakım araştırma alanları irdelenmiş ve çeşitli sonuçlar elde edilmiştir. Problem alanlarının belirlenmesinde ortak payda laboratuar test verileri kullanılarak çözüm arayışına uygun olmalarıdır. Önceki bölümlerde anlatılan teorik bilgiler ışığında hastalardan istenmiş olan test verilerine belirlenen problemin yapısına uygun olarak çeşitli veri madenciliği teknikleri uygulanmıştır. Geliştirilen uygulamalarla elde edilen ilginç sonuçlara dikkat çekmek ve bunların irdelenmesi amaçlanmıştır.

5.2.1. Uygulamaların Amaç ve Kapsamları

Tıp çeşitli türdeki birçok verinin kaydedildiği alanlardan biridir. Tıbbi verilerin artış hızı karşısında, geleneksel veri analiz araçları yetersiz kalmaktadır. Tıbbi veritabanlarında yeni bilgilerin keşfedilmesi, veritabanlarında depolanan verinin etkili bir şekilde kullanılması açısından çok önemlidir. Veri madenciliği teknikleri, özellikle

klasik yöntemlerle elde edilemeyen, önceden kestirilemeyen örüntülerin ve bilgilerin açığa çıkmasında önemlidir.

Bu çalışmada test seçiminde canlı arama (live search) tekniğinin kullanılarak daha isabetli test istemleri yapılması; jinekolojik kanserlerin tanı kontrolü yapılması, bu şekilde yanlış girilen ICD kodlarının tespit edilebilmesi; bazı tam kan testi parametrelerinin değerlerinin tahmin edilmesi ile olası başka problem alanlarında esik verilerin tamamlanabilmesi için yöntem arayışı; Potasyum (K) ve Hemoliz İndeksi (HI) arasındaki ilişkinin veri madenciliği yöntemi ile irdelenmesi, böylece bilinen bir problem alanında daha önce kullanılmamış bir yöntemle analiz yapılması; yaşamsal tahmin yapılması, bu şekilde risk grubuna giren hastaların tespit edilmesi gerçekleştirilmiştir.

Çalışma kapsamına Trabzon ili, Karadeniz Teknik Üniversitesi Tıp Fakültesi Farabi Hastanesi Biyokimya Laboratuvarı'nda 7 Ocak–31 Aralık 2010 ve Haziran 2011-Haziran 2012 tarihleri arasında işlenen testler alınmıştır.

5.2.2. Uygulamalarda Kullanılan Algoritma ve Programlar

Çalışmaya esas oluşturan biyokimya laboratuvarı verilerinin elde edilmesi ve bazı ön işlemlerden geçirilerek analize uygun hale getirilmesi için kullanılan yazılım, Delphi programlama dili kullanılarak hazırlanmıştır.

Uygulamalarda kullanılan veri madenciliği yöntemleri birliktelik kuralları, destek vektör makineleri, artırılmış regresyon ağaçları, kümeleme analizidir. Birliktelik kuralları analizinde en çok bilinen ve kullanılan algoritma olan Apriori algoritması kullanılmıştır. Apriori algoritması sıralı algoritmalar içinde, uygulandığı veri büyüklüğüne göre en iyi performans gösteren algoritmalardan biridir. Analiz için, Yeni Zellanda Waikato Üniversitesi tarafından bir proje olarak başlatılan, bugün birçok kullanıcı tarafından kullanılan WEKA veri madenciliği uygulaması geliştirme programı kullanılmıştır. WEKA java platformu üzerinde geliştirilmiş açık kodlu bir programdır. Program geliştirilmeye devam edilmektedir ve çalışmada programın açıklanan son

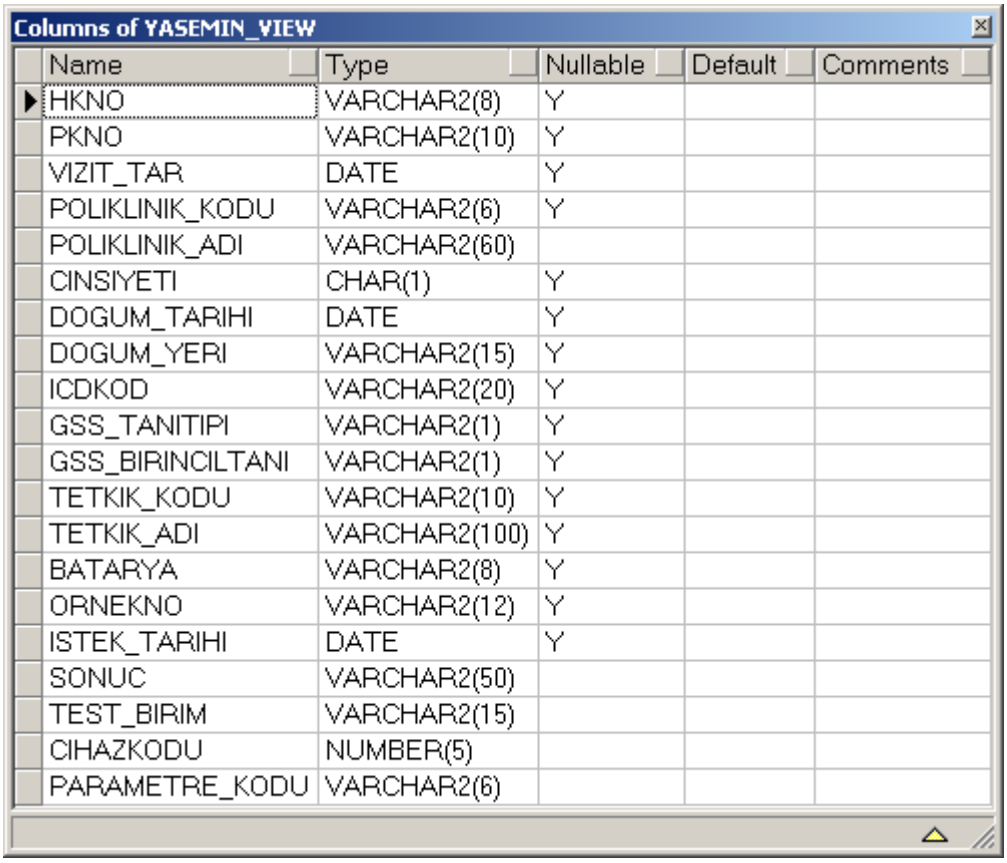
kararlı sürümü olan 3.6.4 kullanılmıştır. WEKA, GNU (General Public License) lisanslı bir yazılımdır. WEKA, Apriori algoritması konusunda iyi bir performans sergileyen, akademik çalışmalarda da kullanılan bir programdır.

5.2.3. Verilerin Hazırlanması

Çalışmada KTÜ Tıp Fakültesi Farabi Hastanesi biyokimya laboratuvarı tarafından sağlanan veriler kullanılmıştır. Bir kısım veriler Farabi Hastanesi Bilgi İşlem Merkezi tarafından yönetilen veritabanından ham haliyle alınmış, birtakım veri işleme aşamalarından geçirilerek veri madenciliği uygulaması için kullanılmaya hazır hale getirilmiştir. Çalışmada kullanılan verilerin bir bölümü ise yine Karadeniz Teknik Üniversitesi Tıp Fakültesi biyokimya laboratuvarı verileri kullanılarak başka bir çalışma (72) için oluşturulmuş veri setidir.

Farabi Hastanesi Bilgi İşlem Merkezi (BİM) tarafından yönetilen hastane veritabanı birçok tablodan oluşan karmaşık bir yapıdadır. Bu çalışma için öncelikle hastane veritabanı tabloları incelenerek, hangi tablolarda bulunan hangi alanların gerekli olacağına karar verilmiş, daha sonra bu alanlar oluşturulan bir ‘view’ ile tek bir tabloda toplanmıştır (Şekil 5). BİM tarafından sağlanan bir kullanıcı adı ve şifre kullanılarak bu view’e erişim sağlanmıştır.

View, veritabanında saklanan bir sorgudur ve sanal tablo oluşturmak için kullanılır. Bir view veri içermez, bu nedenle depolama da gerektirmez. View oluşturmak için kullanılan tablolara ‘temel tablolar’ denir. Kullanıcılar tablolara erişim hakkına sahip olmasalar bile, view’lere erişerek sorgu çalıştırabilirler; böylece veritabanı için daha güvenli bir yapı oluşmuş olur (73). View’ler mantıksal olarak tablolar ile aynı yapıya sahiptir ve aynı şekilde satırlar ve kolonlar içerirler.View’ler üzerinde veriler fiziksel olarak tutulmaz, bunun yerine sadece view’in tanımı tutulur. Bir view her çağrıldığında tanımda yer alan sorgu yeniden çalıştırılır, bu yüzden view içindeki veriler en güncel haldedir. View kullanılarak, uzun ve karmaşık bir sorgu bir ‘select’ işlemi ile yaptırılabilir, birden fazla mantıksal veya fiziksel veritabanından, tek bir veritabanından veri çekiliyor gibi veri çekilebilir.



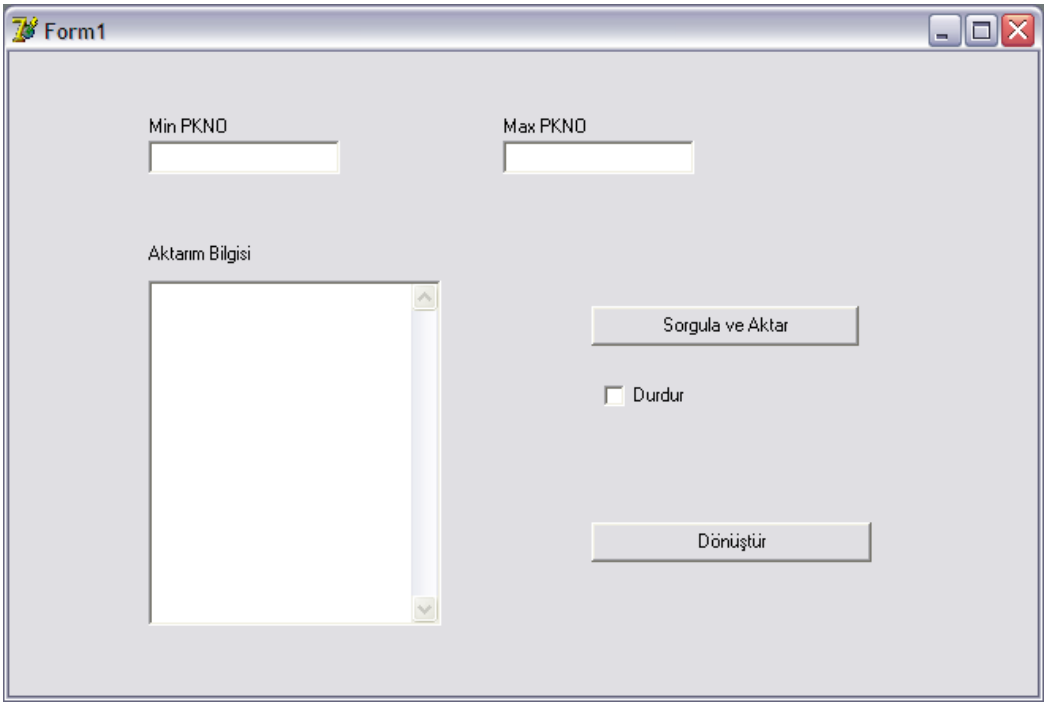
Columns of YASEMIN_VIEW

Name	Type	Nullable	Default	Comments
HKNO	VARCHAR2(8)	Y		
PKNO	VARCHAR2(10)	Y		
VIZIT_TAR	DATE	Y		
POLIKLINIK_KODU	VARCHAR2(6)	Y		
POLIKLINIK_ADI	VARCHAR2(60)			
CINSIYETI	CHAR(1)	Y		
DOGUM_TARIHI	DATE	Y		
DOGUM_YERI	VARCHAR2(15)	Y		
ICDKOD	VARCHAR2(20)	Y		
GSS_TANITIPI	VARCHAR2(1)	Y		
GSS_BIRINCILTANI	VARCHAR2(1)	Y		
TETKIK_KODU	VARCHAR2(10)	Y		
TETKIK_ADI	VARCHAR2(100)	Y		
BATARYA	VARCHAR2(8)	Y		
ORNEKNO	VARCHAR2(12)	Y		
ISTEK_TARIHI	DATE	Y		
SONUC	VARCHAR2(50)			
TEST_BIRIM	VARCHAR2(15)			
CIHAZKODU	NUMBER(5)			
PARAMETRE_KODU	VARCHAR2(6)			

Şekil 5. Oluşturulan ‘view’ yapısı

View yapısı oluşturulduktan sonra, Delphi programlama dili kullanılarak hazırlanan yazılım ile hastane veritabanına erişim ve view üzerinden sorgu yapma imkanı elde edilmiştir (Şekil 6). Programın çalıştığı tarafta Oracle Database 10g R2 kuruludur. Hastane veritabanına erişim, yazılımda kullanılan Oracle Data Access Components (ODAC) bileşeni aracılığı ile sağlanmıştır.

Programda sorgu için daha hızlı sonuç getirdiği tespit edilen PKNO alanı kullanılmıştır. PKNO, polikliniğe başvuran hastaya verilen kodun tutulduğu bir alandır. HKNO ile ifade edilen hastaya özel olan hasta kayıt no alanından farkı, hastanın her bir poliklinik başvurusunda farklı bir PKNO değeri almasıdır. Program çalıştırılıp bağlantı kurulduktan sonra minimum ve maksimum PKNO değerleri belirlenerek sorgu çalıştırılmış, sonuçlar bu çalışma için kurulan Oracle veritabanına aktarılmıştır.



Şekil 6. Veri aktarımı için hazırlanan yazılım arayüzü

Girilen PKNO değerleri ile sorgulanarak aktarılan veriler 7 Ocak–31 Aralık 2010 tarihleri arasında tutulan kayıtları içermektedir. Kayıtların 3553877 tanesi ayaktan hastalara, 4789139 tanesi de yatan hastalara aittir. Verilerin ham haliyle toplam 8343016 tane kayıt aktarılmıştır.

Sorgu sonucunda aktarılan verilerde, hastalardan istenen her bir test, bir satırda gösterilmiştir. Fakat bu yapı veri madenciliği uygulamaları için uygun değildir. Verilerin veri madenciliğine uygun hale getirilmesi için dönüşüm yapılması gerekmektedir. Bu işlem için de, yine aynı yazılım kapsamında geliştirilen dönüşüm fonksiyonu kullanılmıştır. Yapılan dönüşümle, hastaların çalışılan her bir örneğine verilen ORNEKNO alanı temel alınarak iki yeni tablo oluşturulmuştur. Oluşturulan tablolarda, her test yeni bir alan olarak işlenmiştir. Ayrıca hemogram gibi alt parametreleri olan testlerin parametreleri de tabloya alan olarak eklenmiştir. Bu şekilde oluşturulan yeni tablolarda, kişinin yaşı, doğum yeri, doğum tarihi gibi demografik özelliklerinin yanı sıra, tüm test alanları da bulunmaktadır. Kişilerden istenen testler, Tablo1'de 'var–yok' mantığını ifade edecek şekilde '1–0' biçiminde işaretlenip

birliktelik analizine uygun yapı elde edilmiştir. Hastadan istenmiş olan testler ‘1’ ile işaretlenmiştir. Diğer tablo olan Tablo2’de ise testlerin ölçüm sonuçları işlenerek sonuçlara dayalı analizlere uygun hale getirilmiştir (Şekil 7, Şekil 8, Şekil 9).

HKNO	PKNO	VIZIT_TAR	POLIKLINIK_ADI	CINSIYETI	DOGUM_TARIHI	DOGUM_YERI	ICDKOD	TETKIK_KODU	TETKIK_ADI	SONUC	TEST_BIRIM
L0516603	A10167708	27.05.2010	Enfeksiyon Poliklin	E	05.04.1950	RİZE	B18.2	90399044	Trigliserid	144	mg/dL
L1096039	A10167294	26.05.2010	Nöroloji Polikliniği	E	07.06.1953	TRABZON	I63.8	90399044	Trigliserid	132	mg/dL
L1096039	A10167294	26.05.2010	Nöroloji Polikliniği	E	07.06.1953	TRABZON	I10	90399044	Trigliserid	132	mg/dL
L0516635	A10166818	26.05.2010	Hematoloji Poliklini	E	15.05.1960	ŞALPAZARI	D50.8	90126044	Fosfor (P)	4.12	mg/dL
L1096910	A10167085	26.05.2010	Üroloji Polikliniği	E	10.07.1942	RİZE	N42.9	90025044	Alfa- feto prot	2.98	ng/mL
L1086855	A10166372	26.05.2010	Kardiyoloji Poliklinik	K	08.08.1957	ARTVİN	I10	90399044	Trigliserid	116	mg/dL
L0772998	A10166419	26.05.2010	Hematoloji Poliklini	K	01.01.1965	İKİZDERE	D75.2	90126044	Fosfor (P)	3.57	mg/dL
L0803512	A10171344	31.05.2010	Enfeksiyon Poliklin	K	01.01.1986	MAÇKA	B18.1	90025044	Alfa- feto prot	1.67	ng/mL
L1014204	A10170977	31.05.2010	Pediatri Göğüs ve	K	08.12.2003	TRABZON	J30.3	90126044	Fosfor (P)	5.0	mg/dL
L0401693	A10171448	31.05.2010	Endokrinoloji Polikl	K	24.03.1937	GÖRELE	E04.2	90347044	Serbest T3	2.52	pg/mL
L0129015	A10168097	27.05.2010	Endokrinoloji Polikl	K	15.08.1963	ÜZÜMÖZÜ	E05.2	90126044	Fosfor (P)	2.8	mg/dL

Şekil 7. Verilerin aktarımıyla oluşturulan ilk tablo

PKNO	VIZIT_TAR	POLIKLINIK_ADI	CIN	DOGUM_TAF	DOGUM_YE	ICDKOD	ORNEKNO	AC_90126044	AC_90013144	AC_90150044	AC_90018044	AC_90020044	AC_9
Y10020998	05.07.2010	Pediatri (SÜT ÇOCUĞU) Serv	E	01.03.2010	TRABZON	Q24.8	7130981						
Y10020874	04.07.2010	Pediatri (SÜT ÇOCUĞU) Serv	E	11.02.2008	OSMANGAZİ	B00.4	7130982						
Y10020833	04.07.2010	Pediatri (SÜT ÇOCUĞU) Serv	E	15.10.2009	TRABZON	J21.9	7130983						
Y10020644	01.07.2010	Pediatri Enfeksiyon Servisi	E	10.02.2005	TRABZON	W84	7155469	1					
Y10020644	01.07.2010	Pediatri Enfeksiyon Servisi	E	10.02.2005	TRABZON	W84	7155470		1				
Y10020883	05.07.2010	Pediatri Enfeksiyon Servisi	K	04.11.1993	G.HANE	A98.0	7155493	1		1		1	1
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156134						
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156136			1			1
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156140						
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156145			1			1
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156146						

Şekil 8. Verilerin dönüştürülmesinden sonra oluşturulan Tablo1, istenen testler 1 ile işaretlendi

PKNO	VIZIT_TAR	POLIKLINIK_ADI	CINS	DOGUM_TAR	DOGUM_YE	ICDKOD	ORNEKNO	AC_90126044	AC_90013144	AC_90150044	AC_90018044	AC_90020044	AC_S
Y10020998	05.07.2010	Pediatri (SÜT ÇOCUĞU) Serv	E	01.03.2010	TRABZON	Q24.8	7130981						
Y10020874	04.07.2010	Pediatri (SÜT ÇOCUĞU) Serv	E	11.02.2008	OSMANGAZİ	B00.4	7130982						
Y10020833	04.07.2010	Pediatri (SÜT ÇOCUĞU) Serv	E	15.10.2009	TRABZON	J21.9	7130983						
Y10020644	01.07.2010	Pediatri Enfeksiyon Servisi	E	10.02.2005	TRABZON	W84	7155469	5.17					
Y10020644	01.07.2010	Pediatri Enfeksiyon Servisi	E	10.02.2005	TRABZON	W84	7155470		28.1				
Y10020883	05.07.2010	Pediatri Enfeksiyon Servisi	K	04.11.1993	G.HANE	A98.0	7155493			169		13	4.7
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156134	8.57					
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156136			327			2.2
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156140						
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156145			71			2.4
Y10020008	27.06.2010	Anestezi Yoğun Bakım BD.Se	K	12.11.1931	ÇAYELİ	C76.0	7156146						

Şekil 9. Verilerin dönüştürülmesinden sonra oluşturulan Tablo2, testlerin sonuçları ile dolduruldu

Dönüşümden sonra, ORNEKNO alanı taban alınarak elde edilen tablolarda 227432 adet kayıt elde edilmiştir. Bunlardan 140640'ı ayaktan hastalara, 86792'si yatan hastalara ait kayıtlardır.

5.2.4. Birliktelik Kuralları Analizi ile Test Seçiminde Canlı Arama Tekniğinin Kullanılması

Hastalıkların teşhis ve ayrımında beden sıvıları ve salgılarının tetkiki çok yararlıdır. Bu amaçla kan, serum, idrar, dışkı, beyin, omurilik gibi materyalin alınarak laboratuvarda değerlendirilmesi doktorlara önemli ipuçları verebilir. Günümüzde, dünyada hastalık tanılarının % 70' ten fazlası laboratuvar testleriyle konulmaktadır(74). Bu nedenle uygun bir laboratuar testlerinin seçimi hastalıkların doğru tanısı için çok önemlidir.

Test takımları, özel bir hastalık veya organ sistemi için kullanılan testlerden oluşan test gruplarıdır. Kolay seçilebilmesi ve şüphelenilen durumla ilgili geniş çerçevede test içermesi hekim için kullanım rahatlığı sağlar. Test takımları eksiksiz ve isabetli test istemi yapılmasına olanak verir. Böylece özellikle acil durumlarda hekim test istemi yaparken daha başarılı kararlar verebilir; bu da istemlerin tam yapılmasını sağlar ve tekrar test isteminin önüne geçilir. Bu şekilde hem zaman verimli kullanılmış olur, hem de ekstra maliyetlerin önüne geçilmiş olur. Fakat bazen test takımlarının kullanılması, gerekli olmayan bazı testlerin de istenmesine neden olabilmektedir. Bu da hastanın muayene maliyetini arttırmaktadır ve gereksiz istenen bu testler hastaneler için istenmeyen durumdur.

Bu bölümde, test seçiminde canlı arama (live search) tekniği kullanılarak geliştirilmesi önerilen sistemle, hem test takımı kullanır gibi eksiksiz test seçimi yapılmasını sağlamak, hem de test takımlarının getirdiği gereksiz testlerin maliyetini ortadan kaldırmak amaçlanmıştır.

Canlı arama tekniği yeni ve yaygın olarak kullanılan bir tekniktir. Geleneksel arama yöntemlerine göre daha pratiktir. Canlı arama yaparken harfler ya da sözcükler yazıldıkça sonuçlar kısıtlanır. Arama sonuçları, arama sorgusu tamamlanmadan görünmeye başlar. Bu yöntem kullanıcıyı sonuçlara erişmek için 'gönder' düğmesine basmaktan da kurtarır.

Laboratuvar testleri kanıta dayalı tıbbın en önemli araçlarıdır. Testlerin dikkatli ve akılcı seçilmesi hasta bakımında alınan kararların kalitesini arttırır. Bu bölümde geliştirilmesi önerilen sistem için her bir polikliniğin test istemleri göz önünde bulundurulacaktır. Bu şekilde daha isabetli ve özgün test önerileri yapılabilecektir. Önce belirlenen polikliniğin test istemleri birliktelik analizi modeli kullanılarak çıkarılacak, daha sonra elde edilen kurallarla canlı arama yapısı kullanılarak laboratuvar testlerinin seçimi için öneriler sunulacaktır. Testler birlikte istenme yoğunlukları göz önüne alınarak, birlikte en fazla istenen test grubundan, en az istenen test grubuna göre gösterilecek şekilde sıralanacaktır.

5.2.5. Destek Vektör Makineleri ile Jinekolojik Kanserlerin Tanı Kontrolü

Kanser, vücut hücrelerinin kontrolsüz veya anormal bir şekilde büyümesi ve çoğalmasıdır. Tüm dünyada olduğu gibi ülkemizde de kalp-damar hastalıklarından sonra en öldürücü hastalık gurubu kanserdir (75). Kadınlarda erkeklerden farklı olarak gelişen jinekolojik kanserler ovaryan, servikal, endometriyal, vulvar kanserler olup; Uluslararası Kanser Araştırmaları Ajansı'nın 2002 yılı verilerine göre tüm dünyada kadınlardaki yıllık 5.1 milyon yeni kanser vakasının %19'unu oluşturmaktadır. Bu, kanserden ölümlerin 2.9 milyonunu ve 5 yıllık kanser prevalansının da 13 milyonunu oluşturmaktadır (76).

Jinekolojik kanserlerde, özellikle yumurtalıklara ait patolojilerde kan tahlilleri ile elde edilen bazı belirteçler teşhiste ve tedavide yardımcı oldukları gibi, tedavi sonrası hastaların takibinde de rol oynamaktadırlar. Kanserli dokulardan kaynaklanan ve çeşitli vücut sıvılarında artan pek çok biyolojik bileşik tümör belirteci olarak kullanılabilir. Kanser şüphesi ve tanısı konusunda önemli ipucu sağlayan belirteçlerin Total ve Free

PSA, CEA, AFP, CA–125, CA 15–3, CA 19–9 olduğu bilinmektedir (77,78,79). Jinekolojide en sık kullanılanlar CEA, AFP, CA–125, CA 19–9, beta-hCG belirteçleridir (80,81). Bu belirteçler fikir vermeleri açısından önemlidir; ancak kesin tanıya götürmezler. Jinekolojik kanserlerin genellikle 40 yaş üzerinde görüldüğü bilinmektedir (82).

Bu bölümde, Birliktelik Kuralı Madenciliği ve Destek Vektör Makineleri yöntemleri kullanılarak, kanser belirteçlerinin sonuçları analiz edilmiş ve hastaya konulan tanılarla sistem çıktılarının tutarlılıkları incelenmiştir.

5.2.6. Bazı Tam Kan Testi Parametrelerinin Değerlerinin Artırılmış Regresyon Ağacı Yöntemi ile Tahmini

Hemogram testi olarak da bilinen tam kan testi, kandaki hücrelerin otomatikleştirilmiş sayımını içerir. Tam kan testi laboratuvarlarda sıklıkla istenen temel bir tarama testidir. Tam kan testindeki bulgular doktorlara hematolojik ve diğer vücut sistemleri, hastalıkların seyri, tedaviye yanıt ve iyileşme hakkında tanısal bilgi sunar. Tam kan testi kan hücrelerinin sayısı, çeşidi, yüzdesi, konsantrasyonları ve kalitesini belirleyen kan parametreleri olarak anılan bir dizi testten oluşur. Bu testler (83):

- Beyaz Kan Hücresi Sayısı (WBC - Lökositler): Lökositler enfeksiyonlarla mücadele eder ve vücudun bağışıklık hücrelerinin toplamını gösterir. Vücudun sağlık durumunu korumak ve enfeksiyon veya yaralanma gibi nedenlerle mücadele için kullandığı lökositlerin beş farklı türü vardır; bunlar nötrofiller, lenfositler, bazofiller, eozinofiller, monositlerdir.

- Kırmızı Kan Hücresi Sayısı (RBC - eritrosit): Eritrositler 1 mikrolitredeki kırmızı kan hücrelerinin sayısıdır. Artışlar da, azalışlar da anormal koşullara işaret edebilir. Eritrositlerin içinde vücuda oksijen taşıyan hemoglobin proteini vardır.

- Hemoglobin (Hb): Hb kandaki oksijen taşıyan protein miktarını gösterir.

- Hematokrit (Hct): Belli bir tam kan hacmindeki eritrositlerin yüzdesini gösterir.

- Ortalama eritrosit hacmi (MCV): Eritrositlerin ortalama büyüklüklerinin bir ölçümüdür. Eritrositler vitamin B12 eksikliğinin neden olduğu anemideki gibi normalden daha büyükse (makrositik) MCV yükselmiştir. Eritrositler demir eksikliği anemisi ve talasemilerde görüldüğü gibi normalden daha küçükse MCV azalmıştır.

- Ortalama eritrosit hemoglobini (MCH): Bir eritrosit hücresi içinde hemoglobinin ortalama miktarının hesaplanmasıyla belirlenir.

- Ortalama eritrosit hemoglobin konsantrasyonu (MCHC): Bir eritrosit içindeki ortalama hemoglobin konsantrasyonunun hesaplanması yoluyla belirlenir.

- Trombosit (PLT): Trombositler kanın pıhtılaşmasında görev alan özel hücrelerdir.

Bu bölümde WBC, RBC ve Hb parametreleri, bilinen diğer kan parametrelerinin değerleri kullanılarak tahmin edilmeye çalışılmıştır.

5.2.7. Potasyum (K) ve Hemoliz İndeksi (HI) Arasındaki İlişkinin Kümeleme Analizi Yöntemi ile İrdelenmesi

Hemoliz, kırmızı kan hücrelerinin zarlarının parçalanmasıyla, içlerinde bulunan hemoglobin ve diğer iç bileşenlerin hücreleri çevreleyen sıvıya salınmasıdır. Hemoliz iki sebepten oluşabilir:

- In-vivo hemoliz, otoimmün hemolitik anemi veya transfüzyon reaksiyonu gibi patolojik koşullar nedeniyle olabilir (84).

- In-vitro hemoliz, uygunsuz örnek toplama, örnek işleme, ya da santrifüj nedeniyle olabilir (85).

In-vivo hemolizin tüm hemolitik örnekler içinde görülme sıklığı %~3.2 oranındadır (86). Hemoliz klinik laboratuvarlarda en yaygın görülen preanalitik hata kaynağıdır ve reddedilen test örneklerinin %~40-70'e yakınının gerekçesidir (86,87). Hemolitik örnekler, istenen tüm rutin testleri içinde %~3.3 gibi bir paya sahiptir ve yetersiz numune, pıhtılı numune, doğru olmayan numune gibi ret nedenlerine kıyasla 5 kat daha fazla gözlenir (87). Hemoliz, hücre içeriğinde bulunan potasyum (K), demir (Fe), magnezyum (Mg), aspartat amino transferaz (AST), laktat dehidrogenaz (LDH) gibi plazma bileşenlerinde hatalı yükselme veya düşüşlere yol açabilir (84).

Bu çalışmada Haziran 2011-Haziran 2012 tarihleri arasında KTÜ Tıp Fakültesi Farabi Hastanesi yatan hasta verileri kullanılarak, potasyum (K) ve hemoliz indeksi (HI) arasındaki ilişki veri madenciliği tekniklerinden Kümeleme Analizi (Clustering) yöntemi kullanılarak irdelenmiştir.

5.2.8. Yaşamsal Tahmin

5.2.8.1. Laktat ve Laktat Dehidrogenaz ile Mortalite Tahmini

Laktat glukozun oksijensiz solunumla metabolize olması sonucu ortaya çıkan bir metabolittir. Glukozun tamamen enerjiye çevrilememesi de kasa yeteri kadar oksijen aktarılamamasından kaynaklanır. Laktat dehidrogenaz (LDH) ise, vücuttaki doku hasarının nedenini ve yerini tanımlamaya ve hasarın ilerlemesini izlemeye yardımcı olan bir biyobelirteçtir. Günümüzde akut ya da kronik bir doku hasarının varlığının ve şiddetinin göstergesi olarak; bazen de bazı kanserlerin, böbrek hastalığı ve karaciğer hastalığı gibi ilerleyici rahatsızlıkların takibi amacıyla kullanılmaktadır (83).

Laktat ve LDH testlerinin mortalite ile ilişkili olduğu bir çok çalışmada ileri sürülmüştür. Yapılan bir çalışmada ölümcül kanser hastalarında yaşam süresi tahmini için prognostik bir faktör olarak LDH seviyelerini değerlendirilmiş ve elde edilen

sonuçlar, serum LDH düzeyinin ölümcül kanser hastalarında hayatta kalma süresi açısından iyi bir yordayıcı olduğunu göstermiştir (88). Bir başka çalışmada, germ hücreli testis tümörü görülen hastalarda serum LDH-1 testinin ölüm konusunda uluslararası sınıflandırmalarda belirtilen kriterlerden daha iyi bir tahmin edici olduğu ileri sürülmüştür (89). Enfeksiyonlu yoğun bakım servisi hastalarında venöz laktat seviyesinin artan ölüm riski ile ilişkili olduğunun vurgulandığı bir çalışma yapılmıştır (90). Bu konudaki başka bir çalışmada da, kritik hastalarda artan kan laktat seviyesinin artan mortalite ve morbidite ile ilişkili olduğu saptanmıştır (91).

Bu bölümde K.T.Ü. Tıp Fakültesi Farabi Hastanesi'nde 7 Ocak–31 Aralık 2010 tarihleri arasında kayıt edilen yatan hasta verileri kullanılarak, laktat testi değerlerinin DVM ile analiz edilmesiyle yaşamsal tahmin yapılıp yapılamayacağı araştırılmıştır.

5.2.8.2. Diyabet Hastalarında Hemogram Parametreleri ile Mortalite Tahmini

Tam kan testi laboratuvarlarda sıklıkla istenen temel bir tarama testidir. Tam kan testindeki bulgular doktorlara hematolojik ve diğer vücut sistemleri, hastalıkların seyri, tedaviye yanıt ve iyileşme hakkında tanısal bilgi sunar. Tam kan testi kan hücrelerinin sayısı, çeşidi, yüzdesi, konsantrasyonları ve kalitesini belirleyen kan parametreleri olarak anılan bir dizi testten oluşur.

Diabetes mellitus, hastada insülin üretme ve/veya kullanma yetisinin bulunmamasıyla karakterize, kan şekeri (glukoz) düzeylerinde yükselmeye yol açan rahatsızlıklar grubuna verilen isimdir. Kontrolsüz diyabet vücutta kan damarları, sinirler ve diğer organlarda hasara neden olabilmekte, böbrek yetmezliği, görme kaybı, inmeler ve kalp-damar hastalığına yol açabilmektedir (83). Amerikan Diyabet Derneği'nin 2011 yılı verilerine göre, 25.8 milyon çocuk ve yetişkin, yani populasyonun % 8.3'ü diyabet hastasıdır. Aynı derneğin 2007 yılı verilerine göre, diyabet 71382 ölüm vakasının asıl nedeni, 160022 ölüm vakasında da etken faktörlerden biri olarak, toplamda 231404 ölüme neden olmuştur (92).

Yapılan bir çalışmada, yatan hastaların ölüm riski tahmininde hangi tam kan sayımı parametrelerinin etkin olduğu araştırılmış ve çekirdekli kırmızı kan hücreleri ve hastaneye başvuruda mutlak lenfositoz varlığı gibi nedenlerin hastanede 30 gün kalış süresinde ölüm riskini üç kata kadar arttırdığı gözlenmiştir (93). Bir başka çalışmada, genç ve yaşlı diyabet hastalarında hemogram profili incelenmiş, bazı parametrelerde anlamlı farklılıklar olduğu tespit edilmiştir (94).

Bu bölümde K.T.Ü. Tıp Fakültesi Farabi Hastanesi'nde 7 Ocak–31 Aralık 2010 tarihleri arasında kayıt edilmiş ve diyabet teşhisi konulmuş yatan hasta verileri üzerinden, hemogram parametreleri analiz edilerek yaşamsal tahmin yapılıp yapılamayacağı araştırılmıştır.

6. BULGULAR

6.1 Birliktelik Kuralları Analizi ile Test Seçiminde Canlı Arama (Live Search) Tekniğinin Kullanılması Uygulamasına Dair Bulgular

Bu bölümde örnek olarak Endokrinoloji Polikliniği'nde istenen testler incelenmek üzere ele alınmıştır. Endokrinoloji Polikliniği'nde belirlenen süre içinde 7758 hastadan istenen 11483 farklı örnek üzerinde çalışılmıştır.

Birliktelik kuralları uygulanırken destek seviyeleri 0.01–0.1 ve 0.003-0.05 aralıklarında belirlenerek veriden iki kesit şekilde kural çıkarımı yapılmıştır. Güven katsayısı sonuçların kesinlik içermesi açısından 1 olarak tanımlanmıştır. Bu değerlerle oluşturulan ilk 100 kural çıkarılmıştır. Birkaç farklı test istem kombinasyonunda sonuçların nasıl değiştiği aşağıdaki örnekle açıklanmıştır. Kurallar destek değerine göre büyükten küçüğe sıralanacak ve ilk beş kuralı içerecek şekilde verilecektir.

Aşağıdaki örnek FSH(Follicle Stimulating Hormone) testinin istemleri çerçevesinde sunulacaktır. İncelenen kural kesitinde FSH ilk seçildiğinde aşağıdaki kuralların sıralanmıştır.

FSH=1 Prolaktin=1 356 ==> Lüteinleştiren hormon (LH) 356 conf:(1)

Estradiol=1 FSH=1 Prolaktin=1 268 ==> Lüteinleştiren hormon (LH) 268 conf:(1)

FSH=1 Prolaktin=1 Total testesteron=1 231 ==> Lüteinleştiren hormon (LH) 231 conf:(1)

Büyüme hormonu=1 FSH=1 Prolaktin=1 188 ==> Lüteinleştiren hormon (LH) 188 conf:(1)

Estradiol=1 FSH=1 Prolaktin=1 Total testesteron=1 174 ==> Lüteinleştiren hormon (LH) 174 conf:(1)

FSH testi ile Progesteron testinin seçildiği varsayılırsa kurallar aşağıdaki gibi sıralanacaktır:

FSH=1 Progesteron=1 Prolaktin=1 121 ==> Lüteinleştiren hormon (LH) 121 conf:(1)

Estradiol=1 FSH=1 Progesteron=1 Prolaktin=1 111 ==> Lüteinleştiren hormon (LH) 111 conf:(1)

FSH=1 Lüteinleştiren hormon (LH)=1 Progesteron=1 Total testesteron=1 94 ==> Prolaktin 94 conf:(1)

Büyüme hormonu=1 FSH=1 Progesteron=1 Total testesteron=1 61 ==> Lüteinleştiren hormon (LH) 61 conf:(1)

Dehidroepiandrosteron sulfat=1 FSH=1 Progesteron=1 35 ==> Lüteinleştiren hormon (LH) 35 conf:(1)

6.1.1 Değerlendirme

Bu bölümde, doktorlara tıbbi test istemlerini oluştururken kolaylık sağlayacağı düşünülen canlı arama yapısı ortaya konulmuştur. Önce laboratuar test istemleri analiz edilerek birliktelik kuralları çıkarılmış, daha sonra birliktelik kurallarına göre sıralanmış ve seçilen test ile birlikte en çok istenen testleri sıralı bir şekilde listeleyen canlı arama kullanıcı arayüzü önerilmiştir. Bu yapının hekimler için daha hızlı ve eksiksiz test seçimine yardımcı olacağı düşünülmektedir. Özellikle acil durumlarda, en isabetli test isteminin, hızlı ve tam yapılması çok önemli olabilir. Doktorun seçtiği testle ilgili tüm testleri görerek test istemini oluşturması ile unutulan testlerin önüne geçilebilir. Ayrıca bu şekilde tekrar test istemlerinin önlenmesi ile maliyetlere de olumlu katkı sağlanabilir.

6.2 Destek Vektör Makineleri ile Jinekolojik Kanserlerin Tanı Kontrolü Uygulamasına Ait Bulgular

Çalışma kapsamına KTÜ Tıp Fakültesi Farabi Hastanesi veritabanına 7 Ocak–31 Aralık 2010 tarihleri arasında kaydedilen hasta verileri alınmıştır. Kanser belirteci olarak bilinen CEA, AFP, CA-125, CA 15–3, CA 19–9 testlerinden en az birisi istenmiş olan 6628 istem seçilerek üzerinde çalışılmıştır. Bu hastalar yatan ve ayaktan hastalar olarak ayrıştırılmış, ayaktan hasta verileri üzerinde çalışılmaya devam edilmiştir. Ön çalışma olarak ICD10 kodu C51-C58 aralığında olan jinekolojik kanser tanısı konmuş 642 hasta (E) ile geri kalan hastalar (H) ile işaretlenerek gruplandırılmıştır.

DVM doğrusal ya da doğrusal olmayan sınıflandırma problemlerinde oldukça başarılı sonuçlar vermektedir. DVM'nin temel prensibi bir karar düzleminde sınıflar arası en uygun karar sınırlarını belirlemektir. Bu sınırlara göre istenen nesnenin hangi sınıfa ait olduğu belirlenebilmektedir. DVM ile yapılacak olan sınıflamada kanser hastaları için en sık kullanıldığı bilinen CA-125 testi (80, 81) ile birlikte değerlendirilecek diğer test ya da testlerin seçimi için birliktelik kuralları çıkarılmış ve aşağıdaki kural bulunmuştur.

CA-125=1 1401 ==> CA 19- 9=1 1171 conf (0.84), supp (0.23)

Kurala göre, çalışmada CEA, AFP, CA-125, CA 15–3, CA 19–9 testlerinden herhangi birinin istendiği toplam 5105 ayaktan hasta kaydı kullanılmış, bunlar içinde CA-125 testi 1401 kez istenmiş ve bu istemlerin 1171'inde aynı zamanda CA 19–9 testi de istenmiştir (conf = 1171/1401 = 0.84). Ayrıca kural sonuçlarına göre yapılan kümeleme analizi de bu iki testin jinekolojik kanser teşhisi konulan hastalarda önemli olduğunu doğrulamaktadır.

Kümeleme analizi verilerin birimlere veya değişkenlere göre birbirlerine benzerlikleri bakımından ayrık kümelerde toplanmasını sağlayan bir tekniktir (54). Bu çalışmada kümeleme analizi, CA-125 ve CA 19–9 testlerinin birlikte istenmiş olduğu, kanser teşhisi konulmuş hastalardan rastgele oluşturulan 380 kişilik bir veri setine

uygulanmıştır. Analizde kategorik değişkenler olarak poliklinik kodu, cinsiyet ve ICD10 kodu alanları ve sürekli değişkenler olarak CA-125 ve CA 19–9 test değerleri verilerek yapılmıştır. Elde edilen sonuçlar aşağıdaki gibidir (Tablo 1):

Tablo 1. Kanser teşhisi konulmuş 380 hastanın poliklinik kodu, cinsiyet, ICD10 kodu, CA-125 ve CA 19-9 test sonuçları kullanılarak yapılan kümeleme analizi sonuçları.

Servis	**Cinsiyet**	**ICD10**	**CA-125**	**CA 19- 9**	**N**	**%**
Nöroloji	E	D43.0	36.7	77.3	96	25.3
Kadın hast.	K	C56	88.2	27.6	224	58.9
Gast. Ent.	K	C16.9	266.8	1688.3	60	15.8

Tabloda da görüldüğü gibi erkek hastaların yoğunlukta olduğu kümede D43.0 ICD10 kodlu beyin kanseri görülmüştür, bayan hastalarda ise C56 ICD10 kodlu yumurtalık kanseri ve C16.9 kodlu mide kanseri ağırlıklı olarak görülmüştür.

Jinekolojik kanser tanılarının tutarlılığının ölçüldüğü analizde, DVM'ye kesiksiz değişken olarak CA-125 ve CA 19-9 sonuçları ve kategorik değişken olarak poliklinik kodu, cinsiyet, yaş ve ICD10 tanı kodları verilmiştir. DVM sonuçları aşağıda listelenmiştir.

Dataset CANCER_DATA:

Dependent: KADINGK

Independents: CA125_90081044, CA19-9_90083044, POLIKLINIK_KODU, CINSIYET, AGE, ICD10

Sample size = 867 (Train), 304 (Test), 1171 (Overall)

Support Vector machine results:

SVM type: Classification type 1 (capacity=10,000)

Kernel type: Radial Basis Function (gamma=0,167)

Number of support vectors = 320 (0 bounded)

Support vectors per class: 208 (N), 112 (Y)

Class. accuracy (%) = 100,000(Train), 99,342(Test), 99,829(Overall)

Sonuçlara göre toplam 1171 test istemi içerisinde 867'si DVM eğitimi için kullanılmış, 304 istem ise DVM'nin testi için kullanılmıştır. Eğitim başarısı %100, test başarısı %99.3 olduğu görülmektedir. Test setinde yer alan 304 hastanın 166'sı jinekolojik kanser tanısı (H) ve 138'i ise(E) şeklinde iken DVM'nin sınıflamasına göre ise 168 (H), 136 (E) şeklinde oluşmuştur. Buna göre 2 hasta kadın genital bölge kanseri sonucu (E) iken DVM'ye göre (H) olmuştur. Bu iki test istemi incelendiğinde CA-125 ve CA 19-9 test sonuçlarının normal aralıkta olduğu görülmektedir.

6.2.1 Değerlendirme

Bu bölümde "Çıkarılan birliktelik kurallarına göre çalışma kapsamına alınan özelliklerle jinekolojik kanser tanılarının tutarlılığını DVM ile sınamak mümkün müdür?" sorusuna cevap aranmıştır. Elde edilen sonuçlar DVM'nin buna benzer sınıflama ve tahmin için ümit verici bir teknik olduğunu göstermektedir.

Yanlış sınıflandırılan iki test istemi incelendiğinde CA-125 ve CA 19-9 test sonuçlarının normal aralıkta olduğu görülmektedir. Dikkat çeken diğer özellik ise bir hastanın kardiyoloji polikliniğinde kayıtlı ve 30 yaşında olduğudur. Jinekolojik kanserlerin genellikle 40 yaş üzerinde görüldüğü bilinmektedir (82). DVM'nin sınıflandırmayı başarılı yaptığı, yanlış sınıflandırmaların ise araştırılması gerektiği düşünülmektedir. Bu araştırmanın amacına uygun bir durumdur.

Hastanelerde ICD kodları özellikle SGK ödemeleri için zorunlu olduğundan girilmektedir. Bu kodlar hızlı çalışma temposu ve belirsizlikler nedeniyle hatalı

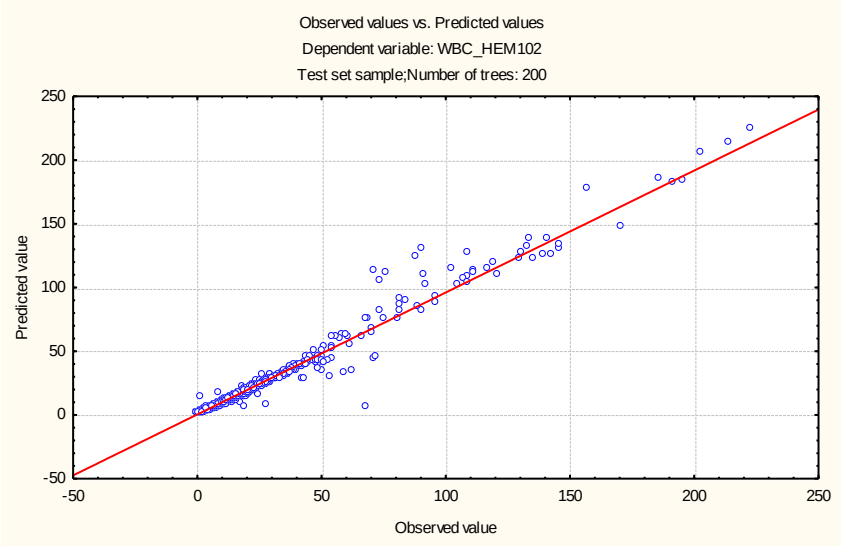
girilebilmektedir. Bu tür çalışmalar girilen tanıların tutarlı ve doğru olması için yararlı sonuçlar verebilir. Gerek ödeyici kurumlar ve gerekse hizmet veren kurumlarda yapılacak veri madenciliği çalışmaları için geniş bir uygulama alanı olduğu gözükmektedir.

6.3 Bazı Tam Kan Testi Parametrelerinin Değerlerinin Artırılmış Regresyon Ağacı Yöntemi ile Tahmini Uygulamasına Ait Bulgular

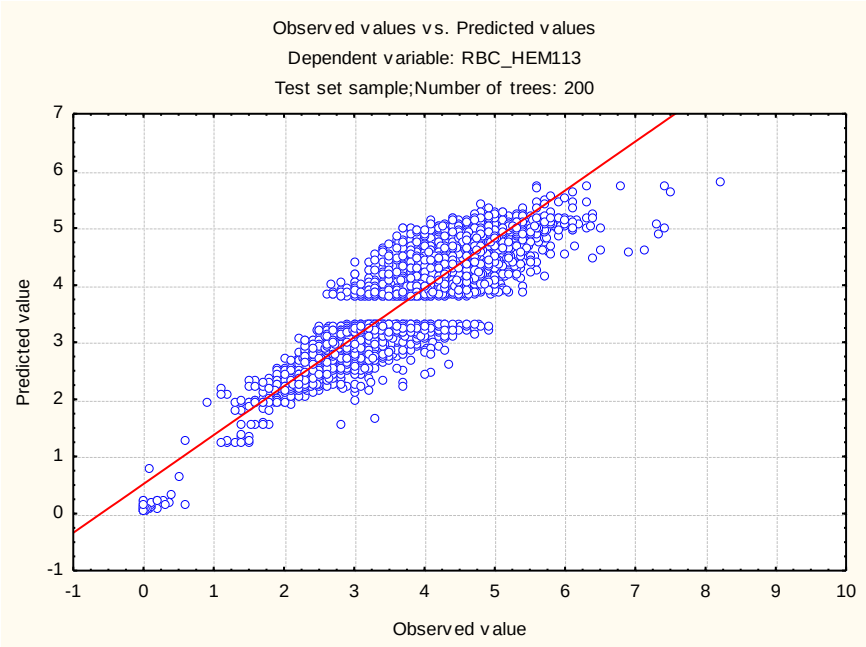
Çalışma kapsamına K.T.Ü. Tıp Fakültesi Farabi Hastanesi'nde 7 Ocak–31 Aralık 2010 tarihleri arasında kaydedilen 86588 hasta verisi alınmış; fakat eksik veriler çıkarıldıktan sonra veri sayısı 46317'e düşmüştür. Bu bölümde WBC, RBC ve Hb parametreleri diğer parametrelerin değerleri kullanılarak tahmin edilmeye çalışılmıştır. İlk olarak, her bir parametre için özellik seçimi ve değişken eleme yapılmış, daha sonra seçilen özellikler kullanılarak Artırılmış Regresyon Ağacı(ARA) yöntemi ile değerler tahmin edilmiştir. ARA yönteminin tahmine dayalı problemlerde başarılı sonuçlar verdiği bilinmektedir.

Özellik seçiminde WBC, RBC ve Hb bağımlı değişkenler olarak seçilmişlerdir. Klinik kodu, ICD 10 kodu, cinsiyet ve ölüm kategorik değişkenler olarak; kalış süresi, yaş ve diğer kan parametreleri sürekli değişkenler olarak verilmiştir. WBC, RBC ve Hb örnekleri en iyi tahmin edici özellikleri bulmak için analiz edilmiştir. WBC için, en iyi 10 tahmin edici, Nötrofil, Monosit, Lenfosit, Bazofil, %Lenfosit, ölüm, % Nötrofil, Eozinofil, PLT, %Monosit olarak bulunmuştur. RBC için en iyi 10 tahmin edici, Hct, Hb, ölüm, PLT, Lenfosit, WBC, Monosit, Nötrofil, kalış süresi, %Monosit olarak bulunmuştur. Ve Hb için, en iyi 10 tahmin edici, Hct, RBC, ölüm, Lenfosit, PLT, WBC, Monosit, Nötrofil, kalış süresi, Bazofil olarak bulunmuştur.

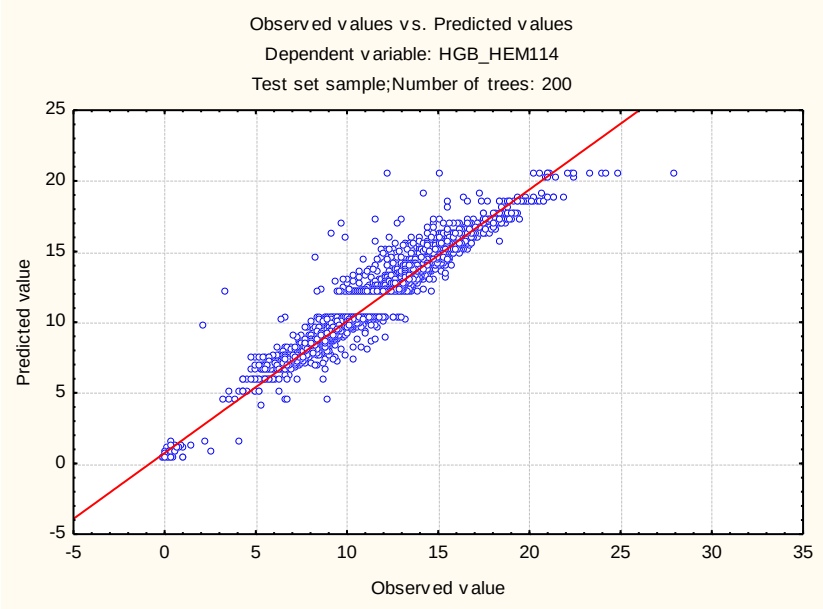
Her bir kan parametresi için sonuçlar ARA yöntemi ile tahmin edilmiştir. Grafiklere bakıldığında, WBC tahmininde en doğru sonuçların bulunduğu görülmektedir (Bkn. Şekil 9, 10, 11).



Şekil 10. WBC için gözlenen ve tahmin edilen değerler



Şekil 11. RBC için gözlenen ve tahmin edilen değerler



Şekil 12. Hb için gözlenen ve tahmin edilen değerler

6.3.1 Değerlendirme

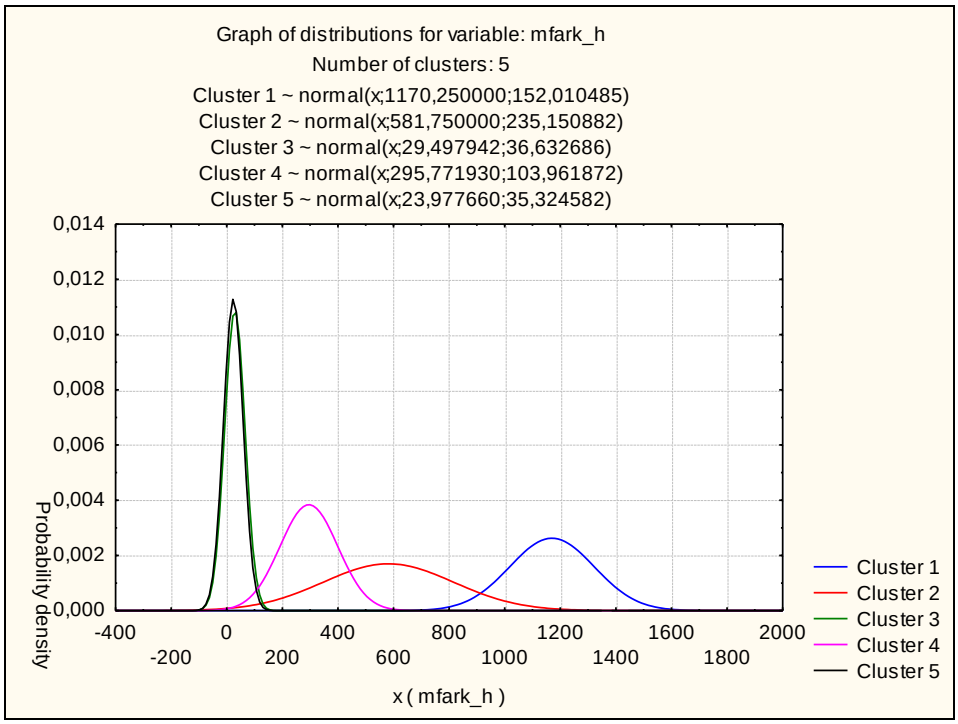
Bu bölümde bir veri madenciliği metodu olan ARA yöntemi tıbbi verilere uygulanarak tahmin gücü ölçülmüştür. Tıbbi veri kaynakları genellikle hızlı çalışma temposu ve belirsizlikler nedeniyle hatalı veya eksik veri içermektedir. Bu tarz çalışmalar geliştirilerek, klinik olarak sağlıklı sonuçlar olmamasına rağmen, veri madenciliği tarzında analizler için verilerdeki eksiklikler giderilebilir. Böylece veri ön işleme basamağında eksik veriler için uygulanan, aynı gruptaki verilerin değerlerinin ortalamasının atanması ya da az sayıda eksik veri içeren kayıtların da tamamen veri setinden çıkarılması gibi çözümlerden daha uygun bir adım sağlanabilir.

6.4 Potasyum (K) ve Hemoliz İndeksi (HI) Arasındaki İlişkinin Kümeleme Analizi Yöntemi ile İrdelenmesi Uygulamasına Ait Bulgular

Haziran 2011-Haziran 2012 tarihleri arasında kan örnekleri tekrarlanmış olan 9777 yatan hastadan elde edilen son iki potasyum (K) ve hemoliz indeksi (HI) ölçümleri çalışmada kullanılmıştır. Oluşturulan veri setinde her bir kayıt satırı, son K ve HI değerleri ve sonuç tarihini, bir önceki K ve HI değerleri ve sonuç tarihini, K ölçüm

çiftinin farkını, HI ölçüm çiftinin farkını içermektedir. Elde edilen veri seti, bir ön elemeye tabi tutulmuş, kayıtlardan ilk ve son K değerleri arasındaki farklar ile ilk ve son HI değerleri arasındaki farkların her ikisinin de sıfırın üzerinde –pozitif- olduğu; ya da sıfırın altında –negatif- olduğu kayıtlar seçilmiş, diğer kayıtlar çalışma kapsamı dışında bırakılmıştır. Bu işlemin amacı, K ve HI değerleri arasındaki farkın ilişkili olduğu, yani aynı yönlü değiştiği verilerin seçilmesidir. Ön eleme adımının ardından kayıt sayısı 4394'e düşmüştür. Kalan kayıtlar alınan kan örneklerinin arasındaki süre farkı gözetilerek tekrar süzülmüş ve kayıt sayısı 1252'ye düşmüştür.

Kümeleme analizi verilerin birimlere veya değişkenlere göre birbirlerine benzerlikleri bakımından ayrık kümelerde toplanmasını sağlayan bir tekniktir (95). Çalışmada kümeleme analizi, sürekli değişken olarak ilk ve son K ve HI farklarının mutlak değerleri tanımlanarak yapılmıştır. Analizde kümeler arası mesafe ölçümü için Euclidean ölçü birimi ve k-means algoritması kullanılmıştır. Kümeleme analizinde sınıf sayısı 5 olarak tanımlanmıştır. Bunun nedeni son HI değerlerine göre yapılan hemoliz sınıflandırmasıdır (HI<=5 mg/dL için 'hemoliz yok', HI<=30 mg/dL için 'belirsiz', HI<=60 mg/dL için 'hafif', HI<=200 mg/dL için 'orta', HI>200 mg/dL için 'ağır') (96, 97). Analizde hata oranı 0,043588 olarak saptanmıştır. Kümeleme sonuçlarına göre gruplardaki HI fark ortalamalarına ait dağılım grafiği aşağıda verilmiştir (Şekil 12).



Şekil 13: Gruplardaki HI fark ortalamalarına ait dağılım grafiği

Kümeler oluşturulduktan sonra her bir küme içerisinde K ve HI fark ortalamaları arasındaki korelasyon ilişkisi sorgulanmıştır. Sonuçlar aşağıdaki gibidir (Tablo 2):

Tablo 2. K ve HI fark değerleri ile yapılan kümeleme analizi sonuçları ve her bir kümede K ve HI fark ortalamaları arasında bulunan korelasyon katsayıları

Kümeler	**K (mmol/L)** (Fark Ortalamaları ,p=0,00)	**HI (mg/dL)** (Fark Ortalamaları,p=0,00)	**n**	**Yüzde (%)**	**r**
1	4.9	1170.2	4	0,3	0,99
2	3.7	581.8	8	0,6	0,80
3	1.1	29.4	243	19,4	0,58
4	1.4	295.8	57	4,5	0,91
5	0.3	23.9	940	75	0,54

Potasyum testi beklenen değerleri 3–6 mmol/L olarak tanımlanmaktadır (98). Kümeleme analizi sonuçlarına bakıldığında, 1 numaralı kümede HI grup ortalaması 1000 (mg/dL)'nin üstünde olduğu için, 3 ve 5 numaralı kümelerde ise 100 (mg/dL)'nin altında kaldığı için dikkate alınmayacaktır.

Sonuçlara bakıldığında kalan iki küme içinde en yüksek korelasyon değeri 0.91 ile 4. kümede hesaplanmıştır. Küme kapsamında verilerin %4.55' lik kısmı olan 57 vaka kaydı vardır. K ve HI fark ortalamalarına bakıldığında, K için 1.445 mmol/L, HI için 295.7 mg/dL olduğu görülmektedir. 2. kümede ise korelasyon değeri 0,80 hesaplanmıştır. Kümede 8 vaka bulunmaktadır.

6.4.1 Değerlendirme

Kümeleme yöntemi HI üzerine yapılan bir çalışmada ilk defa kullanılmıştır ve diğer çalışmaların çoğunluğunda kan örneklerinin çalışmalar için özellikle hemolize edilmesine karşın, bu analizde gerçek hasta verileri kullanılmıştır. K ve HI arasındaki ilişkiyi irdeleyen diğer çalışmalarda (99, 100, 101) genellikle HI'ye göre statik sınıflandırma yöntemleri kullanılmıştır. Kümeleme yöntemi ile veriler doğasına uygun şekilde sınıflandırılmıştır ve klinik bulgularla tutarlı sonuçlar elde edilmiştir. Bu da veri madenciliğinin tıbbi veri analizlerinde klinik çalışmaların ön çalışma aşamasında kullanılmaya uygun olabileceğinin göstergesidir.

6.5 Yaşamsal Tahmin Uygulamasına Ait Bulgular

6.5.1 Laktat ve Laktat Dehidrogenaz ile Mortalite Tahmini Bulguları

Çalışma kapsamına yatan hastalardan toplanmış ve laktat testini içeren 686 adet istem alınmıştır. Bu istemlerin 219 tanesi pediatrik servislerde tedavi gören hastalara, 467'si yetişkin hastalara aittir. Analizler tüm veri seti için ve yetişkin veri seti için ayrı ayrı yapılmıştır.

686 kayıt içeren veri setinde, 331 kayıtta laktat değerleri normal seviyede, 355 kayıtta ise yüksek seviyededir. Yüksek laktat seviyesi gözlenen kayıtlara sahip 89 hastanın durumu ölü olarak kaydedilmiştir. Veri setine birliktelik kuralı analizi uygulanmış, 0,2 destek düzeyi ve 0,50 güven seviyesinde elde edilen ilk iki kural aşağıdaki gibidir:

OLUM == 0 526 ==> LAKTAT == N 356 supp (41.25364), conf (53.9048), corr (65.553)

OLUM == 1 161 ==> LAKTAT == Y 331 supp (12.97376), conf (55.2795), corr (38.55343)

Elde edilen kurallar çok güçlü olmasa da, ölüm ve laktat seviyesi arasında bir ilişkinin varlığından söz edilebilir.

Mortalite tahmini için DVM'ye kesiksiz değişken olarak hastanın yaşı, laktat değeri, LDH değeri, sodyum (Na) değeri; kategorik değişken olarak ise cinsiyet ve ICD kod bilgisi verilmiştir. Burada analize Na testinin eklenmesinin sebebi, laktat testinin istenen testlerle yapılan birliktelik kuralı analizi sonucunda en çok Na ile istenmiş olmasıdır. DVM sonuçları aşağıdaki gibidir:

Dataset laktat yatan:

Dependent: OLUM

Independents: YAS, AC_90225044, AC_90226044, AC_90367044, CINSIYETI...

Sample size = 202 (Train), 67 (Test), 269 (Overall)

Support Vector machine results:

SVM type: Classification type 1 (capacity=10,000)

Kernel type: Radial Basis Function (gamma=0,167)

Number of support vectors = 71 (17 bounded)

Support vectors per class: 30 (0), 41 (1)

Class. accuracy (%) = 96,535(Train), 97,015(Test), 96,654(Overall)

Sonuçlara bakıldığında, toplam 269 istem analizde değerlendirilmiş, 202 istem DVM'nin eğitimi, 67 istem de test için kullanılmıştır. Değerlendirilen istem sayısının 269'a düşmesinin nedeni, kesiksiz değişken olarak tanımlanan üç testin birlikte istendiği test istemlerinin değerlendirilmesidir. Kernel fonksiyonu olarak Radial Basis fonksiyonu seçilmiştir. Test için kullanılan 67 istemden, 55'i hayatta olan, 12 tanesi de ölü hastalara ait istemlerdir. DVM analiz sonucunda hayatta olan 54, ölmüş olan 11 istemi doğru değerlendirmiştir. Yanlış sınıflandırılan iki istem bulunmaktadır. Böyle bir analiz için, % 97 test başarısının iyi bir değer olduğu düşünülmektedir.

Yetişkin veri seti üzerinde de DVM ile yaşamsal tahmin analizi yapılmıştır. Analizde kesiksiz değişken olarak yaş ve LDH değeri; kategorik değişken olarak da poliklinik kodu, ICD kod bilgisi ve laktat düzeyi verilmiştir. Elde edilen sonuçlar aşağıdaki gibidir:

Dataset laktat yatan yetiskin:

Dependent: OLUM

Independents: YAS, AC_90226044, POLIKLINIK_KODU, ICDKOD, LAKTAT

Sample size = 192 (Train), 59 (Test), 251 (Overall)

Support Vector machine results:

SVM type: Classification type 1 (capacity=10,000)

Kernel type: Radial Basis Function (gamma=0,167)

Number of support vectors = 63 (5 bounded)

Support vectors per class: 34 (0), 29 (1)

Class. accuracy (%) = 99,479(Train), 96,610(Test), 98,805(Overall)

Sonuçlara bakıldığında, toplam 251 istem değerlendirilmiş, bunlardan 192 tanesinin DVM'nin eğitimi için, 59'unun da test başarısını ölçmek için kullanıldığı görülmektedir. Hayatta olan 41 hastanın hepsi doğru tahmin edilmiştir. Ölmüş olan 18 hastanın ise, 16 tanesi doğru sınıflandırılmış, iki hasta yanlış sınıflandırılmıştır. Bu grupta test başarısı % 96.6 olarak saptanmıştır.

Yetişkin veri setinin tümünün analiz edilmesi amacı ile, test verisi olarak sadece laktat düzeyinin kullanıldığı bir analiz daha yapılmıştır. Bu analizde kesiksiz değişken olarak hastanın yaşı, kategorik değişken olarak da poliklinik kodu, ICD kodu ve hastanın laktat düzeyi tanımlanmıştır. Elde edilen sonuçlar aşağıdadır:

Dataset laktat yatan yetiskin:

Dependent: OLUM

Independents: YAS, POLIKLINIK_KODU, ICDKOD, LAKTAT

Sample size = 350 (Train), 117 (Test), 467 (Overall)

Support Vector machine results:

SVM type: Classification type 1 (capacity=10,000)

Kernel type: Radial Basis Function (gamma=0,250)

Number of support vectors = 142 (24 bounded)

Support vectors per class: 54 (0), 88 (1)

Class. accuracy (%) = 96,571(Train), 92,308(Test), 95,503(Overall)

Analizde 350'si DVM'nin eğitimi, 117'si test için olmak üzere 467 yetişkin hasta isteminin tümü kullanılmıştır. Test veri setindeki 84 hayatta olan hastadan 80 tanesi başarılı tahmin edilmiş, dört tane hastada sapma görülmüştür. Hayatta olmayan 33

hastadan ise, 28 tanesi doğru tahmin edilmiş, beş tane hasta yanlış sınıflandırılmıştır. Test başarısı % 92.3 kaydedilmiştir.

6.5.2 Diyabet Hastalarında Hemogram Parametreleri ile Mortalite Tahmini Bulguları

Çalışma kapsamına ICD kod bilgisi diabet melitus olarak işlenmiş ve tüm hemogram parametrelerinin eksiksiz istenmiş olduğu 685 test istemi alınmıştır. Bu istemlerden 116 tanesi hastanede kaldığı süre içinde ölmüştür.

Analizde bu tarz sınıflandırma problemlerinde başarılı sonuçlar veren DVM yöntemi kullanılmıştır. Analiz iki kez tekrarlanmış; ilkinde hastaların yaş bilgisi ve istenen hemogram testi değerleri kesiksiz değişken olarak tanımlanmış, ikincisinde ise yaş ve hemogram testi bilgilerine, hastanın hastanede kalış süresi de eklenmiştir.

Mortalite tahmini için DVM'ye kesiksiz değişken olarak hastaların yaşı ve test sonuçlarının verilmesi ile gerçekleştirilen analizde, 520 istem DVM'nin eğitimi, 165 istem ise testi için kullanılmıştır. Analizde kernel fonksiyonu olarak Radial Basis fonksiyonu seçilmiştir. 165 test istemi, 130 sağ ve 35 ex hastanın istemlerini kapsamaktadır. Elde edilen sonuçlarda, 130 sağ hastadan dört tanesi yanlış sınıflandırılmıştır, 35 ex hastanın ise yalnızca bir tanesi yanlış sınıflandırılmıştır.

İkinci analiz, kesiksiz değişkenlere hastanın hastanede kalış süresi de eklenerek yapılmıştır. Aynı şekilde, 520 istem DVM'nin eğitimi, 165 istem ise testi için kullanılmıştır. Farklı olarak, analizde kernel fonksiyonu olarak Polynominal fonksiyon türü seçilmiştir. 165 test istemi, 130 sağ ve 35 ex hastanın istemlerini kapsamaktadır. Elde edilen sonuçlarda, 130 sağ hastadan üç tanesi yanlış sınıflandırılmıştır, 35 ex hastanın ise yalnızca bir tanesi yanlış sınıflandırılmıştır.

6.5.3 Değerlendirme

Bu bölümde farklı iki problem alanı seçilmiştir ve eldeki verilerden yola çıkarak, biyokimyasal test sonuçlarının veri madenciliği yöntemleri ile irdelenmesi sonucu yaşamsal tahmin yapılmıştır.

Laktat ve LDH testlerinin mortalite ile ilişkili oldukları birçok çalışma sonucunda kanıtlanmıştır (91, 89, 90, 88). Yapılan analizler, yetişkin ve pediatrik hasta verileri birlikte kullanılarak ve sadece yetişkin hasta verilerinde kullanılarak tekrarlanmıştır. Tüm kayıtları içeren veri seti üzerinde, laktat ve LDH sonuçları kullanılarak yapılan analizde test başarısı % 97 olarak bulunmuştur. Yetişkin hasta veri seti üzerinde, LDH sonuçları ve laktat düzeyleri kullanılarak yapılan analizde test başarısı % 96.6 olarak bulunmuştur. Son yapılan analizde ise, yetişkin veri kullanılmış; fakat yalnızca laktat düzeyleri değerlendirilerek sınıflandırma yapılmış ve test başarısı % 92.3 kaydedilmiştir. DVM'nin bu tarz sınıflandırma problemlerinde başarılı olduğu düşünülmektedir.

Literatürde diyabet melitus hastalarında hemogram parametreleri ile mortalite riskinin ilişkili olduğuna dair çalışma bulunamamıştır. Yapılan analizde, ICD kodu diyabet melitusu işaret eden ve hemogram testleri eksiksiz istenmiş olan hastaların istemleri toplanarak elde edilen veri seti kullanılmıştır. Hastaların yaş ve hemogram değerleri kullanılarak yapılan analizde eğitim başarısı % 99.8; test başarısı ise % 96.9 olarak hesaplanmıştır. Yapılan ikinci analizde kesiksiz değişkenlere hastanede kalış süresi de eklenmiştir. Elde edilen sonuçlara bakıldığında, eğitim başarısı % 100; test başarısının ise % 97.5 olduğu görülmüştür.

Sağlık hizmeti veren kuruluşlarda bu tarz bulgular ile hastanın risk değerlendirmesi yapılabilir; tanı ve tedaviye bulgular çerçevesinde yön verilebilir. Durumu kötüye giden hastaların tahmini, erken müdahaleler için fırsat sunabilir ve özellikle acil servislerde hastaların mortalite düzeyleri düşürülebilir.

7. TARTIŞMA ve SONUÇ

Veri madenciliği yaklaşımlarıyla tıp dünyasında hastalara ve sağlık çalışanlarına hizmet sunulması günümüzdeki popüler konulardan olmuştur. Bu nedenle tezde bu amaca yönelik olarak, biyokimya laboratuvarı verileri kullanılarak çeşitli analizler yapılmıştır. Analizlerde günümüzde yaygın olarak kullanılan veri madenciliği tekniklerinden, birliktelik analizi, destek vektör makineleri, kümeleme analizi ve artırılmış regresyon ağaçları kullanılmıştır. Bu uygulamalarla tıbbi verinin farklı bakış açıları ve farklı analiz yöntemleri ile irdelenmesi amaçlanmıştır. Yapılan analizlerle;

- Test seçiminde daha önce istenmiş test grupları analiz edilerek ve canlı arama tekniği kullanılarak daha isabetli test istemleri yapılması,
- Jinekolojik kanser vakalarında hastaların biyokimya test sonuçları analiz edilerek kabaca tanı kontrolü yapılması, bu şekilde yanlış girilen ICD kodlarının tespit edilebilmesi,
- Bazı tam kan testi parametrelerinin değerlerinin tahmini yapılması, olası başka problem alanlarında eksik verilerin tamamlanabilmesi için yöntem arayışı,
- Potasyum (K) ve Hemoliz İndeksi (HI) arasındaki ilişkinin veri madenciliği yöntemi ile irdelenmesi; böylece, bilinen bir problem alanında daha önce kullanılmamış bir yöntemle analiz yapılması ve yöntemin doğrulanması,
- Hastaların bazı biyokimya testleri ile yaşam süreleri arasındaki ilişkiler değerlendirilerek yaşamsal tahmin yapılması, bu şekilde risk grubuna giren hastaların tespit edilmesi,

gerçekleştirilmiştir.

Birliktelik analizi ile test seçiminde canlı arama tekniğinin kullanılması ile hekimin test seçerken bulunduğu poliklinik veya serviste daha önce yapılan istemleri görerek daha doğru seçimler yapabilmesine ve hızlı karar verebilmesine katkıda bulunmak amaçlamıştır. Bu yöntemle hekim test takımı kullanır gibi eksiksiz test seçimi yaparken; birlikte seçilme olasılığı düşük olan testler belirlenen destek ve güven eşik değerleri ile süzüleceğinden gereksiz test seçimi de en aza indirgenecektir. Yöntem daha önce aynı tanı konulan hastalara hangi tedavilerin uygulandığının ve ne kadar başarılı

olunduğunun gösterilmesi, bu şekilde daha başarılı sonuç veren tedavinin seçilmesi gibi farklı tıbbi problemlerin çözülmesi için de uyarlanarak kullanılabilir. Analizde en iyi bilinen ve en çok kullanılan birliktelik analizi algoritması olan Apriori algoritması kullanılmıştır. Apriori algoritması aday kümelerin sayısını azaltmada iyi bir performans gösterdiği için seçilmiştir. Yapılan başka bir çalışmada, Santangelo ve ark. yeni klinik test bataryaları oluşturulması için veri madenciliği araçlarından birliktelik analizini kullanmışlardır (102).

DVM ile yapılan tanı kontrolü bölümünde, 1171 kayıttan 304 tanesi sistemin test edilmesi için kullanılmıştır. Test sonucunda jinekolojik kanser tanısı konan 138 hastadan 136 tanesi doğru sınıflandırılmıştır. Yanlış sınıflandırılan iki kayıt incelendiğinde, hastalardan birinin kardiyoloj polikliniğinde kayıtlı ve 30 yaşında olduğu saptanmıştır. Bu sonuç uygulamanın amacına uygundur. Bu tarz kontrollerle hastanelerde yanlış girilen ICD10 kodu sayıları alt limitlere çekilebilir.

WBC, RBC ve Hb parametrelerinin sonuçlarının diğer tam kan testi parametrelerinin değerlerinin değerlendirilmesi ile tahmin edilmesi bölümünde elde edilen sonuçlara bakıldığında, yalnızca WBC parametresinin tahmininde gözlenen ve tahmin edilen değerler birbirlerine çok yakın çıkmıştır. Bu tarz tahminlerin klinik olarak sağlıklı sonuçlar vermediği söylenebilir. Fakat bu uygulama ile amaçlanan örneklemde az sayıda elemanın bulunduğu durumlarda, eksik veri içeren kayıtların silinmesi yerine, eksik alanların tahmin edilerek eleman sayısının daha da azalmasının önüne geçilmesini sağlamaktır.

K ve HI arasındaki ilişkinin, kümeleme analizi yöntemiyle irdelenmesi ile veriler klinik bulgulara yakın şekilde sınıflandırılmıştır. Yapılan bir çalışmada (97), acil servise başvuran hastalarda hemolizin etkisi araştırılmış ve serum örnekleri hemoglobin konsantrasyonuna göre analiz edilmiştir. Oluşturulan kümelerin üç tanesinde HI değerleri belirlenen aralıkların dışında kaldığı için kümeleme sonuçları dikkate alınmamıştır. Kalan iki kümede potasyum ve hemoliz seviyelerinin ilişkili olduğu elde edilen yüksek korelasyon değerleri ile ortaya koyulmuştur. Bu sonuçla yöntemin iyi

sonuç verdiği doğrulanmıştır ve bundan sonra karşılaşılacak uygun problemlerde de verilerin ön sınıflandırmasında kullanılabileceği ispat edilmiştir.

Laktat ve LDH testleri kullanılarak yapılan yaşamsal tahminde; yaş, cinsiyet ve ICD10 kod bilgilerine ek olarak sodyum(Na) sonuçlarının da kullanıldığı analizde 269 kayıttan 67 tanesi sistemin test başırısını ölçmek için değerlendirilmiş ve yöntem %97 oranında doğru sınıflandırmıştır. Bu analizde Na test sonucunun da değişkenlere eklenmesinin sebebi, laktat testinin birlitelik analizi sonucunda en çok Na ile birlikte istenmiş olmasıdır. Fakat sonuçların konu alanı uzmanı ile değerlendirilmesi sonucunda, Na testinin bu kadar yüksek oranda istenmesinin sebebinin veri setini oluşturan yatan hastaların genellikle damar yolu ile besleniyor olmaları ve bu yüzden Na düzeylerinin sık ölçülmesi olduğu anlaşılmıştır. Birliktelik analizi ile elde edilen kurallar yüksek güven ve destek seviyelerine sahip olsalar bile anlamlı olmayabilirler. Sadece yetişkin hastaların verileri kullanılarak yapılan analizde laktat, LDH, yaş, poliklinik kodu ve ICD10 kodu bağımsız değişken olarak tanımlanmıştır. Sonuçlara bakıldığında 251 istemden 59 tanesi test için kullanılmış ve %96.6 oranında doğru sınıflama başarısı elde edilmiştir. Yine yetişkin veri seti ile yapılan ve laktat, yaş, poliklinik kodu ve ICD10 kodu bilgileri kullanılan analizin sonuçlarına bakıldığında, 467 istemden 117 tanesinin test için kullanılmıştır. Sistemin test başarısının %92.3 seviyesine düştüğü gözlenmiştir. Bu konuda daha sonra yapılacak çalışmalarda laktat, LDH ve Na testlerinin ölüm ile ilişkileri irdelenerek modelin kurulmasının daha başarılı sonuçlar elde edilmesini sağlayacağı düşünülmektedir. Yapılan bir çalışmada ölümcül kanser hastalarının LDH seviyeleri değerlendirilmiş ve serum LDH düzeyinin hastaların hayatta kalma süreleri açısından iyi bir yordayıcı olduğu ileri sürülmüştür (88).

Diyabet hastalarında hemogram parametreleri kullanılarak yapılan yaşamsal tahminde, 685 test iseminin 165 tanesi test için kullanılmış ve 130 sağ hastadan dört tanesi, 35 ölü hastadan da bir tanesi yanlış sınıflandırılmıştır. Yapılan ikinci bir analizde hastanın hastanede yatış süresi de eklenerek analiz tekrarlanmış ve 130 sağ hastadan üç tanesi, 35 ölü hastadan bir tanesi yanlış sınıflandırılmıştır. Literatürde diyabet hastalarının hemogram parametreleri ile mortalite riskinin ilişkili olduğuna dair çalışmaya rastlanmamıştır. Fakat genç ve yaşlı diyabet hastalarında hemogram

profillerinin incelendiği bir çalışmada, bazı parametrelerde anlamlı farklılıklar olduğu gösterilmiştir (94). Diyabet hastaları ile yapılacak sonraki çalışmalarda, diyabet süresi gibi parametrelerin de çalışma kapsamına alınması daha sağlıklı sonuçlar elde edilmesine katkı sağlayacaktır.

Yapılan analizlerle tıbbi veriden önceden öngörülemeyen ve klasik yöntemlerle elde edilemeyecek sonuçlar çıkarılmıştır. Fakat bu sonuçların anlamlandırılması ve sağlık çalışanları ve hastalar için fayda sağlaması için tıbbi alan uzmanı desteği kesinlikle gerekmektedir. Bu yüzden tıbbi veri üzerinde yapılacak olan veri madenciliği çalışmaları multidisipliner şekilde yürütülmelidir. Bu çalışmaların planlanması aşamasında da her iki alan uzmanları alan yeterlilikleri kapsamında çalışmanın güvenirliğini sağlamak için birlikte hareket edilmelidir.

Ayrıca kullanılan tıbbi veriler ön işlemeye tabi tutulmuş ve birçok eksik kayıt içerdiği görülmüştür. Hızlı işlem yapılması gerekliliği, hasta yoğunluğu gibi nedenler bu durumu açıklamaktadır. Fakat veri madenciliğinde analizlerin başarısını etkileyen en önemli unsur verinin kalitesidir. Bu yüzden daha sonra yapılması planlanacak çalışmalar için kullanılacak verinin daha kaliteli oluşturulması; eksik-yanlış kayıtlardan ve tekrarlı verilerden arındırılmış olması gerekmektedir. Sağlık çalışanlarının da bu konu hakkında bilgilendirilmesi ve veritabanlarında tutulan büyük miktarlardaki verinin başarılı bir şekilde analiz edilerek fayda sağlanabilmesi için kısıtlı zaman, hasta yoğunluğu gibi olumsuz şartlara rağmen kayıtların itina ile doldurulmasının gerekliliği açıklanmalıdır.

Günümüzde hemen her alanda teknolojinin ve bilişimin kullanılması kaçınılmaz hale gelmiştir. Tıpta da teknoloji büyük miktarlarda bilgiyi depolama, depolanan bilgiye hızlı erişme, alarm ve uyarı sistemlerinin kurulmasına uygun olma, karar destek sistemleri ile kararlara destek olma gibi insan yeteneklerini aşan çözümler sunarak hasta güvenliğini ve bakım kalitesini arttırmaktadır. Bilginin kullanımı, hasta güvenliği ve bakım kalitesinin geliştirilmesinde en iyi teknolojinin seçilmesinde, uygulama ve planlama stratejilerinin geliştirilmesinde ve anlaşılmasında sağlık çalışanlarına yardımcı olacaktır. Bakım kalitesini arttırmak ve hasta güvenliğini geliştirmek için klinik bilişim

sistemlerinin kullanımı artık bir gereklilik olarak karşımıza çıkmaktadır (103). Nitekim Amerika Birleşik Devletleri'nde 2012 güz döneminden itibaren Klinik Bilişim tıbbi yandal alanı için kurul sınavları yapılmaya başlanmıştır. Ülkemizde de bireysel ve toplum sağlığı sonuçlarını geliştirmek, hasta bakımını geliştirmek ve klinisyen-hasta ilişkisini güçlendirmek için klinik bilişim alanındaki çalışmaların özendirilmesi ve desteklenmesi gerekmektedir.

Bu tez çalışması kapsamında yapılan uygulamalar esnek bir yapıya sahip olduğundan, farklı veri girdileri ile farklı durumlar için de uyarlanabilme özelliği taşımaktadır.

8. KAYNAKÇA

1. Özdemir A (2004). Veritabanlarında Bilgi Keşfi ve Veri Madenciliği-Sağlık Sektöründe Bir Uygulama. Doktora Tezi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü, Erzurum.

2. Bollenger AS, Smith RD (2001). Managing Organizational Knowledge as a Strategic Asset. Journal of Knowledge Management 5(1):9.

3. Bailey C, Martin C (2000). How Do Managers Use Knowledge About Knowledge Management?. Journal of Knowledge Management 4(3):237.

4. Wikipedia. (2011) [online]. http://en.wikipedia.org

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

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

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

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

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

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

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

12. Oğuzlar A (2003). Veri Ön İşleme. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 21:67–76.

13. Dolgun MÖ (2006). Büyük Alışveriş Merkezleri İçin Veri Madenciliği Uygulamaları. Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimler Enstitüsü, Ankara.

14. Özmen Ş (2008). İş Hayatı Veri Madenciliği İle İstatistik Uygulamalarını Yeniden Keşfediyor. Marmara Üniversitesi İ.İ.B.F. http://idari.cu.edu.tr/sempozyum/bil38.htm

15. Hand DJ (1998). Data mining: statistics and more?. The American Statistician 52(2):112–118.

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

17. Kumar AVS, Wahidabanu RSD (2009). Data Mining Association Rules for Making Knowledgeable Decisions. H. Rahman (Ed.), Data Mining Applications for Empowering Knowledge Societies. New York, USA: IGI Global.

18. Kavurkacı Ş, Aydın ZG, Şamlı R (2011). Büyük Ölçekli Veritabanlarında Bilgi Keşfi. Akademik Bilişim'11 Konferansı, Malatya, Türkiye.

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

20. Zhong N, Zhou L (1999). Methodologies for Knowledge Discovery and Data Mining. Third Pacific-Asia Conference, Springer Verlag, China, pp:13-15.

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

22. Wenzel R, Del Favero A, Kibbler C, Rogers T, Rotstein C, Mauskopf J, Morris S, Schlamm H, Troke P, Marciniak A (2005). Economic evaluation of voriconazole compared with conventional amphotericin B for the primary treatment of aspergillosis in immunocompromised patients. J Antimicrob Chemother 55(3):352-61.

23. Aydoğan KE (2008). Veri Madenciliğinde Sınıflandırma Problemleri için Evrimsel Algoritma Tabanlı Yeni Bir Yaklaşım: ROUGH-MEP Algoritması. Doktora Tezi, Gazi Üniversitesi Fen Bilimler Enstitüsü, Ankara.

24. Subbalakshmi G, Ramesh K, Chinna Rao M (2011). Decision support in heart disease prediction system using Naive Bayes. Indian Journal of Computer Science and Engineering (IJCSE) 2(2):170–176.

25. Akyol K, Şen B, Çalık E (2012). Biyokimya ve Hemogram Laboratuvar Test Sonuçlarının Lojistik Regresyon Yöntemiyle Analizi. Akademik Bilişim Konferansları-AB2012, Uşak, Türkiye.

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

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

28. Patton RM, Beckerman BG, Potok TE, Treadwell JN (2010). Genetic Algorithm for Analysis of Abdominal Aortic Aneurysms in Radiology Reports. Genetic and Evolutionary Computation Conference-GECCO2010, 1:1931-1936, New York, ABD.

29. Shouman M, Turner T, Stocker R (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. International Conference on Knowledge Discovery-ICKD 2012, IPSCIT Singapore: IACSIT Press.

30. Caudill M (1989). Neural Network Primer: Part I. Al Expert 2(12):47.

31. Elmas Ç (2003). Yapay Sinir Ağları (Kuram, Mimari, Eğitim, Uygulama), Seçkin Yayınları: Ankara.

32. Prabhudesai SG, Gould S, Rekhraj S, Tekkis PP, Glazer G, Ziprin P (2008). Artificial neural networks: Useful aid in diagnosing acute appendicitis. World J Surg 32:305-9.

33. Veropoulos K, Cristianini N, Campbell C (1999). The application of support vector machines to medical decision support: a case study. ACAI 99, Workshop on Support Vector Machines Theory and Applications, Crete, Greece.

34. Demirkır C, Kargın S, Sankur B. Look-Up Table Tabanlı Boosting Algoritması ile Karmaşık Sahnelerde Yüz Sezimi. IEEE 13. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1567621&userType=inst

35. Robinson JW (2008). Regression Tree Boosting to Adjust Health Care Cost Predictions for Diagnostic Mix. Health Serv Res 43(2):755–772.

36. Mitchell CH, Kaufman CE, Beals J, Pathways of Choice and Healthy Ways Project Team (2004). Identifying Diverse HIV Groups Among American Indian Young Adults: The Utility of Cluster Analysis. AIDS and Behavior 8(3):263-275.

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

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

39. Doğan Ş, Türkoğlu İ (2008). Diagnosing Hyperlipidemia Using Association Rules. Mathematical and Computational Applications 13(3):193-202.

40. Anbarasi MS, Ghaayathri S, Kamaleswari R, Abirami I (2011). International Journal of Computer Science and Information Technologies 2(1):512-516.

41. Gray RR, Tanaka Y, Takebe Y, Magiorkinis G, Buskell Z, Seeff S, Alter HJ, Pybus OG (2013). Evolutionary analysis of hepatitis C virus gene sequences from 1953. Philos Trans R Soc Lond B Biol Sci. 368(1626):20130168.

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

43. Tantuğ AC (2002). Veri Madenciliği ve Demetleme. (Yayınlanmamış) Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

44. Silahtaroğlu G (2008). Kavram ve Algoritmalarıyla Temel Veri Madenciliği. Papatya Yayıncılık, İstanbul.

45. Houtsma M, Swami A (1995). Set-Oriented Mining for Association Rules in Relational Databases. Eleventh International Conference on Data Engineering (11. ICDE 95), Taipei.

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

47. Özdoğan GÖ, Abul O, Yazıcı A (2009). Paralel Veri Madenciliği Algoritmaları. BAŞARIM'09, 1. Ulusal Yüksek Başarım ve Grid Konferansı, ODTÜ-KKM, Ankara.

48. Silahtaroğlu G (2008). Kavram ve Algoritmalarıyla Temel Veri Madenciliği. Papatya Yayıncılık, İstanbul.

49. Mannila H, Toivonen H, Verkamo I (1994). Efficient Algorithms for Discovering Association Rules. AAAI'94 Workshop Knowldege Discovery in Databases, pp:181–192.

50. Toivonen H (1996). Sampling Large Databases for Association Rules. 22nd International Conference on Very Large Data Bases (VLDB), pp:134–145.

51. Hidber C (1999). Online Association Rule Mining. ACM SIGMOD International Conference on Management of Data, pp:145–156.

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

53. Park JS, Chen MS, Yu PS (1995). An effective hash based algorithm for mining association rules. ACM-SIGMOD International Conference Management of Data, San Jose, California.

54. Cheung D, Han J, Ng VT, Fu AW, Fu Y, Yongjian AW (1996). A Fast Distributed Algorithm for Mining Association Rules. The 4th International Conference on Parallel and Distributed Information Systems.

55. Parthasarathy S, Zaki MJ, Ogihara M, Li W (2001). Parallel Data Mining for Association Rules on Shared-Memory Systems. Knowledge and Information Systems 3:1–29.

56. Han E, Karypis G, Kumar V (1997). Scalable Parallel Data Mining for Association Rules. 1997 ACM-SIGMOD International Conference On Management Of Data.

57. Masaru TS, Kitsuregawa M (1996). Hash Based Parallel Algorithms for Mining Association Rules. 4th Int. Conf. on Parallel and Distributed Information Systems, pp:19–30.

58. Zaki MJ, Parthasarathy S, Ogihara M, Li W (1997). Parallel Algorithms for Discovery of Association Rules. Data Mining and Knowledge Discovery, 1(4):343–373.

59. Özdamar K (1999). Paket Programlarla İstatistiksel Veri Analizi 2, Kaan Kitabevi, Eskişehir.

60. Bülbül S, Güler MF, Kandemir AŞ (2009). Propensity Skor Uygulamalarında Kümeleme Analizinin Test Amaçlı Kullanımı. 10.Ekonometri ve İstatistik Sempozyumu, Erzurum, Türkiye.

61. Kavzoğlu T, Çölkesen İ (2010). Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. Harita Dergisi 144:73-82.

62. Cortes C, Vapnik VN (1995). Support-vector networks. Machine Learning 20(3):273-297.

63. Vapnik VN (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York.

64. Vapnik VN (2000). The Nature of Statistical Learning Theory. 2. Baskı, Springer-Verlag, New York.

65. Osuna EE, Freund R, Girosi F (1997). Support Vector Machines: Training and Applications. A.I. Memo No. 1602, C.B.C.L. Paper No. 144, Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts.

66. Hsu CW, Chang CC, Lin CJ (2010). A Practical Guide to Support Vector Classification. Son erişim: 20.06.2013

http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

67. Elith J, Leathwick JR, Hastie T (2008). A working guide to boosted regression trees. Journal of Animal Ecology 77:802–813.

68. Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberon J, Williams S, Wisz MS, Zimmermann NE (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129–151.

69. Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecology Progress Series 321:267–281.

70. Hastie T, Tibshirani R, Friedman JH (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

71. Schapire R (2003). The boosting approach to machine learning – an overview. MSRI Workshop on Nonlinear Estimation and Classification, 2002 (eds Denison DD, Hansen MH, Holmes C, Mallick B, Yu B). Springer, New York.

72. Kamburoğlu S (2013). Klinik biyokimya otoanalizörlerinde hemoliz indeks kullanımı. Tıpta Uzmanlık Tezi, Karadeniz Teknik Üniversitesi Tıp Fakültesi, Trabzon.

73. Powell G, McCullough-Dieter C (2005). Oracle SQL Jumpstart with Examples. Elsevier Digital Press, USA, ISBN: 1-55558-323-7.

74. Sağlık Aktüel (2010) [online]. http://www.saglikaktuel.com/ , Göktaş P.: Laboratuvar testleri fazla mı isteniyor?. Son erişim: 27.01.2012.

75. İstanbul Jinekolojik Onkoloji Grubu. http://jinekolojionkoloji.com/ Son erişim: 20.06.2013

76. International Agency for Research on Cancer (IARC). http://www.iarc.fr/ Son erişim: 20.06.2013

77. Chun FK-H, Hutterer GC, Perrotte P, Gallina A, Valiquette L, Benard F, McCormack M, Briganti A, Ionescu C, Jeldres C, Guay JP, Saad F, Karakiewicz PI (2007). Distribution of prostate specific antigen (PSA) and percentage free PSA in a contemporary screening cohort with no evidence of prostate cancer, BJU International 100(1):37-41.

78. Atalay C (2008). Are CA 15-3, CEA and CA 125 predictors for lymphatic and vascular spread in breast cancer?, Erciyes Tıp Dergisi 30(4):218-224.

79. Bekci TT, Senol T, Maden E (2009). The Efficacy of Serum Carcinoembryonic Antigen (CEA), Cancer Antigen 125 (CA125), Carbohydrate Antigen 19-9 (CA19-9), Carbohydrate Antigen 15-3 (CA15-3), alpha-Fetoprotein (AFP) and Human Chorionic Gonadotropin (hCG) Levels in Determining the Malignancy of Solitary Pulmonary Nodules, Journal Of International Medical Research 37(2):438-445.

80. Alanbay İ, Çoksüer H, Ercan CM (2011). Jinekolojik Onkolojide Tümör Belirteçleri: Literatür Derleme, Kocatepe Tıp Dergisi 12: 157-163.

81. Tümör Belirteçleri, Tümör Markırları. http://www.doktornevra.com/jinekoloji/yumurtalik_kistleri/tumor_belirtecleri.asp Son erişim: 20.06.2013

82. Kadın Hastalıkları ve Doğum - Jinekolojik Kanserler. http://www.istanbultipmerkezi.com.tr/content.asp?PID=3#kadin-hastaliklari-ve-dogum Son erişim: 20.06.2013

83. Lab Tests Online. http://labtestsonline.org/ Son erişim: 20.06.2013

84. Lemery L (1998). Oh, No! It's Hemolyzed! What, Why, Who, How? Advance for Medical Laboratory Professionals 15: 24-25.

85. Koseoglu M, Hur A, Atay A, Çuhadar S (2011). Effects of hemolysis interferences on routine biochemistry parameters. Biochemia Medica 21(1):79–85.

86. Carraro P, Servidio P, Plebani M (2000). Haemolyzed specimens: a reason for rejection or clinical challenge? Clin Chem 46: 306–7.

87. Lippi G, Blanckaert N, Bonini P, Green S, Kitchen S, Palicka V, et al (2008). Haemolysis: an overview of the leading cause of unsuitable specimens in clinical laboratories. Clin Chem Lab Med 46: 764-72.

88. Suh SY, Ahn HY (2007). Lactate dehydrogenase as a prognostic factor for survival time of terminally ill cancer patients: A preliminary study. Eur J Cancer 43(6):1051-59.

89. Von Eyben FE, Blaabjerg O, Hyltoft-Petersen P, Madsen EL, Amato R, Liu F, Fritsche H (2001). Serum lactate dehydrogenase isoenzyme 1 and prediction of death in patients with metastatic testicular germ cell tumors. Clin Chem Lab Med. 39(1):38-44.

90. Shapiro NI, Howell MD, Talmor D, Nathanson LA, Lisbon A, Wolfe RE, Weiss JW (2005). Serum lactate as a predictor of mortality in emergency department patients with infection. Ann Emerg Med. 45(5):524-8.

91. Bakker J, Gris P, Coffernils M, Kahn RJ, Vincent JL (1996). Serial blood lactate levels can predict the development of multiple organ failure following septic shock. Am J Surg 171:221–226.

92. American Diabetes Association. http://www.diabetes.org/ Son erişim: 20.06.2013

93. Kho AN, Hui S, Kesterson JG, McDonald CJ (2007). Which observations from the complete blood cell count predict mortality for hospitalized patients? J Hosp Med. 2(1):5-12.

94. Chowta MN, Chowta NK, Adhikari P, Shenoy AK (2013). Analysis of hemogram profile of elderly diabetics in a tertiary care hospital. Int J Nutr Pharmacol Neurol Dis 3:126-30.

95. Çakmak Z. Kümeleme Analizinde Geçerlilik Problemi ve Kümeleme Sonuçlarının Değerlendirilmesi, Dumlupınar Üniversitesi Sosyal Bilimler Dergisi 3:187-205.

96. Yiğitbaşı T, Şentürk BA, Baskın Y, Öney M, Üstüner F (2010). Hemolizin Rutin Acil Biyokimya Testlerine Etkisi. Türk Klinik Biyokimya Dergisi 8(3): 105-110.

97. Turhan K, Engin YZ, Kamburoğlu S, Örem A, Kurt B (2012). Potasyum (K) ve Hemoliz İndeksi (HI) Arasındaki İlişkinin Kümeleme Analizi Yöntemi ile İrdelenmesi. IX. Ulusal Tıp Bilişimi Kongresi, Antalya Türkiye.

98. T.C. Sağlık Bakanlığı Kayseri Eğitim ve Araştırma Hastanesi. Laboratuvar Test Rehberi 2011. http://www.kdh.gov.tr/pdf/Lab_Test_Rehberi.pdf Yayın Tarihi: Eylül 2011, Rev Tarihi: Mayıs 2012. Son erişim: 20.06.2013

99. Shepherd J, Warner MH, Poon P, Kilpatrick ES (2006). Use of Haemolysis Index to Estimate Potassium Concentration in In-Vitro Haemolysed Serum Samples. Clin Chem Lab Med 44(7):877–879.

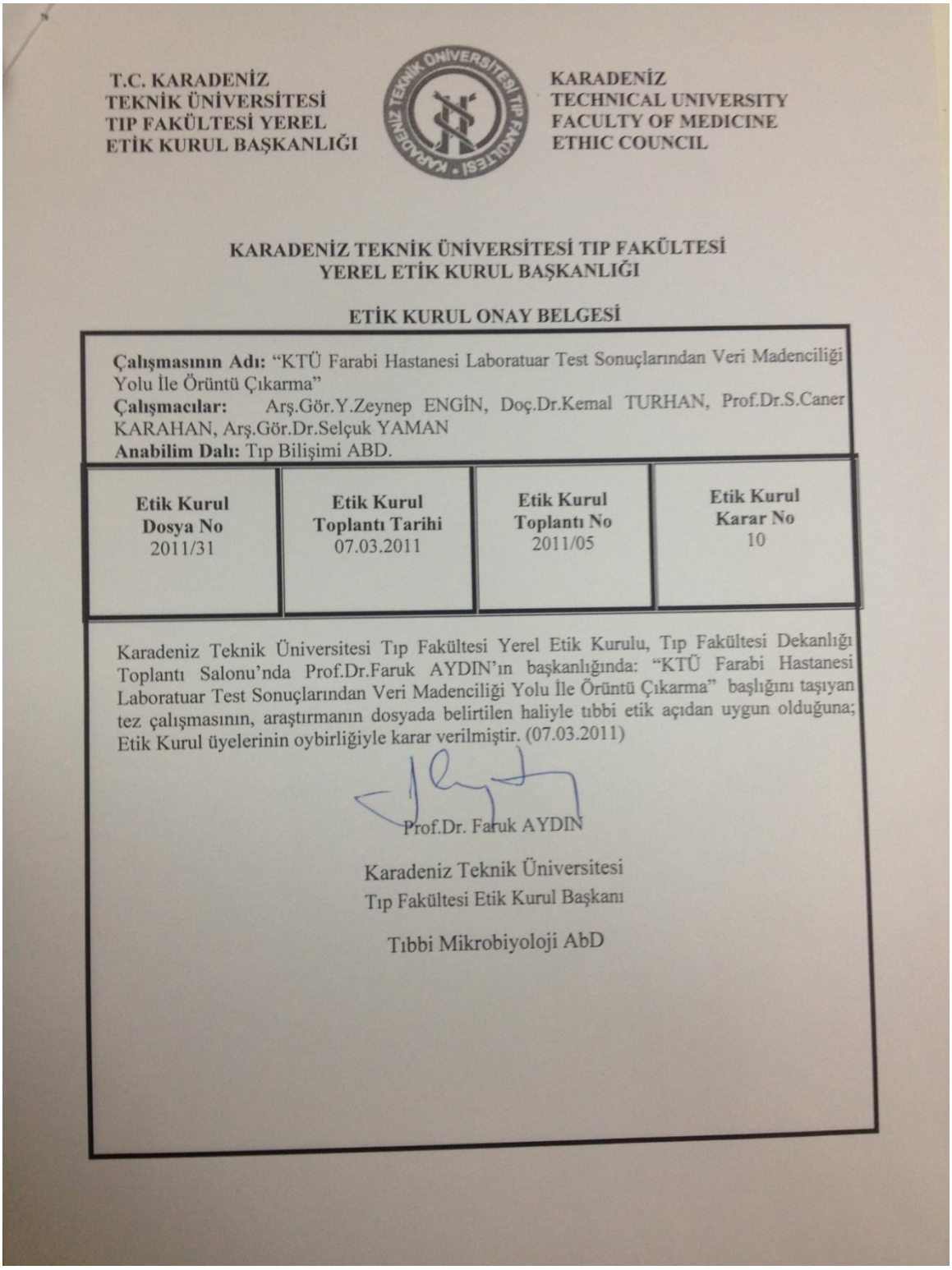
100. Jeffery J, Sharma A, Ayling RM (2009). Detection of Haemolysis and Reporting of Potassium Results in Samples from Neonates. Ann Clin Biochem 46: 222–225.

101. Mansour MM, Azzazy HM, Kazmierczak SC (2009). Correction Factors for Estimating Potassium Concentrations in Samples With In Vitro Hemolysis. Arch Pathol Lab Med 133(6):960-6.

102. Santangelo J, Rogers P, Buskirk J, Mekhjian HS, Liu J, Kamal J (2007). Using Data Mining Tools to Discover Novel Clinical Laboratory Test Batteries. AMIA 2007 Symposium Proceedings:1101.

103. Sucu G, Dicle A, Saka O (2010). Bakım Kalitesi ve Hasta Güvenliğini Geliştirmede Klinik Bilişim Sistemlerinin Kullanımına Yönelik Örnekler. VII. Ulusal Tıp Bilişimi Kongresi Bildirileri, Mağusa, KKTC.

ETİK KURUL ONAYI



T.C. KARADENİZ
TEKNİK ÜNİVERSİTESİ
TIP FAKÜLTESİ YEREL
ETİK KURUL BAŞKANLIĞI

KARADENİZ
TECHNICAL UNIVERSITY
FACULTY OF MEDICINE
ETHIC COUNCIL

KARADENİZ TEKNİK ÜNİVERSİTESİ TIP FAKÜLTESİ
YEREL ETİK KURUL BAŞKANLIĞI

ETİK KURUL ONAY BELGESİ

Çalışmasının Adı: "KTÜ Farabi Hastanesi Laboratuar Test Sonuçlarından Veri Madenciliği Yolu İle Örüntü Çıkarma"
Çalışmacılar: Arş.Gör.Y.Zeynep ENGİN, Doç.Dr.Kemal TURHAN, Prof.Dr.S.Caner KARAHAN, Arş.Gör.Dr.Selçuk YAMAN
Anabilim Dalı: Tıp Bilişimi ABD.

Etik Kurul Dosya No	Etik Kurul Toplantı Tarihi	Etik Kurul Toplantı No	Etik Kurul Karar No
2011/31	07.03.2011	2011/05	10

Karadeniz Teknik Üniversitesi Tıp Fakültesi Yerel Etik Kurulu, Tıp Fakültesi Dekanlığı Toplantı Salonu'nda Prof.Dr.Faruk AYDIN'ın başkanlığında: "KTÜ Farabi Hastanesi Laboratuar Test Sonuçlarından Veri Madenciliği Yolu İle Örüntü Çıkarma" başlığını taşıyan tez çalışmasının, araştırmanın dosyada belirtilen haliyle tıbbi etik açıdan uygun olduğuna; Etik Kurul üyelerinin oybirliğiyle karar verilmiştir. (07.03.2011)

Prof.Dr. Faruk AYDIN

Karadeniz Teknik Üniversitesi
Tıp Fakültesi Etik Kurul Başkanı

Tıbbi Mikrobiyoloji AbD

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Soyadı, Adı : Engin, Yasemin Zeynep
Uyruğu : T.C.
Doğum tarihi ve yeri : 1985 - Erzurum
Medeni hali : Bekâr
E-Posta : ysmnzynp.engin@gmail.com
Yazışma adresi : Karadeniz Teknik Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

EĞİTİM BİLGİLERİ

- Devam: Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Tıp Bilişimi AbD. Tezli Yüksek Lisans
  3,64 / 4
- 2008 : Atatürk Üniversitesi, K.K.E.F Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü
  69,9 / 100
- 2003 : Erzurum Anadolu Lisesi

AKADEMİK/MESLEKİ DENEYİMİ

- Kasım 2009 – Kasım 2010: Akgün Yazılım, programcı
- Kasım 2010 - Aralık 2013: Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Tıp Bilişimi Anabilim Dalı, araştırma görevlisi

YABANCI DİL

- KPDS İngilizce (2012 İlkbahar Dönemi): 67.5 Puan
- ÜDS İngilizce (2011 İlkbahar Dönemi): 65 Puan