İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİNİ KULLANARAK ANEMİ SINIFLANDIRILMASINA YÖNELİK BİR UYGULAMA

YÜKSEK LİSANS TEZİ

Betül Merve FAKI

Endüstri Mühendisliği Anabilim Dalı

Mühendislik Yönetimi Programı

OCAK 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİNİ KULLANARAK ANEMİ SINIFLANDIRILMASINA YÖNELİK BİR UYGULAMA

YÜKSEK LİSANS TEZİ

Betül Merve FAKI
(507111204)

Endüstri Mühendisliği Anabilim Dalı

Mühendislik Yönetimi Programı

Tez Danışmanı: Yrd. Doç. Dr. Başar ÖZTAYŞİ

OCAK 2015

İTÜ, Fen Bilimleri Enstitüsü'nün **507111204** numaralı Yüksek Lisans Öğrencisi **Betül Merve FAKI,** ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"VERİ MADENCİLİĞİ YÖNTEMLERİNİ KULLANARAK ANEMİ SINIFLANDIRILMASINA YÖNELİK BİR UYGULAMA"** başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmustur.

Tez Danışmanı :	**Yrd. Doç. Dr. Başar ÖZTAYŞİ** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Prof. Dr. Selim ZAİM** İstanbul Teknik Üniversitesi	.....
	Doç. Dr. Selçuk ÇEBİ Karadeniz Teknik Üniversitesi	.....

Teslim Tarihi : **15 Aralık 2014**
Savunma Tarihi : **21 Ocak 2015**

Babama,

ÖNSÖZ

Hemen her disiplin tarafından farklı amaçlarla kullanılan veri madenciliği, her geçen gün gelişen bilişim teknolojisi sayesinde daha da önem kazanmaktadır. Bu çalışma, veri madenciliği yöntemleri uygulanarak anemi hastalıklarının sınıflandırılması amacı ile gerçekleştirilmiştir.

Öncelikle çalışmam süresince benden yardımlarını, desteğini, sabrını ve bilgisini esirgemeyen tez danışmanım Yrd. Doç. Dr. Başar ÖZTAYŞİ'ye teşekkürü borç bilirim. Verilerin sağlanmasında, tıbbi bilgilendirme ve destek konusunda her zaman yardımcı olan Prof. Dr. Rıdvan ALİ'ye ve veri toplanması süresince kısıtlı olan vaktini bana ayıran Dr. Aybüke MUTİ'ye teşekkür ediyorum.

Tüm zor zamanlarımda ve her türlü kararımda yanımda olan, sevgi ve desteğini hiçbir zaman eksik etmeyen anneme, kardeşime ve nişanlıma tez çalışmalarım sırasında da sıkıntılarımı paylaşıp, anlayış göstererek bana yardımcı oldukları için minnetarım.

Son olarak, bugünlere gelmemdeki en büyük destekcim, varlığını her an yanımda hissettiğim babama çok teşekkür ediyorum.

Aralık 2014

Betül Merve FAKI
(İşletme Mühendisi)

İÇİNDEKİLER

Sayfa

ÖNSÖZ ..... **vii**
İÇİNDEKİLER ..... **ix**
KISALTMALAR ..... **xi**
ÇİZELGE LİSTESİ ..... **xiii**
ŞEKİL LİSTESİ ..... **xv**
ÖZET ..... **xvii**
SUMMARY ..... **xix**
1. GİRİŞ **1**
2. VERİ MADENCİLİĞİ **3**
2.1 Veri Madenciliği Tanımı 3
2.2 Veri Madenciliğinin Gelişimi 4
2.3 Veri Madenciliğinin Önemi 5
2.4 Veri Madenciliğinin Uygulama Alanları 6
2.5 Veri Madenciliği Süreci 8
2.5.1 İş sorusunu anlama aşaması (Business Understanding) 9
2.5.2 Verileri anlama aşaması (Data Understanding) 10
2.5.3 Verileri hazırlama aşaması (Data Preperation) 10
2.5.4 Modelleme aşaması (Modelling) 12
2.5.5 Değerlendirme aşaması (Evaluation) 15
2.5.6 Uygulama aşaması (Deployment) 16
2.6 Veri Madenciliği Yöntemleri 16
2.6.1 Sınıflandırma modelleri 17
2.6.2 Regresyon 18
2.6.3 Kümeleme 19
2.6.4 İlişkilendirme Kuralları 20
2.6.5 Sıralı Diziler 21
3. TIBBİ KARAR DESTEK SİSTEMLERİ VE TIPTA VERİ MADENCİLİĞİ UYGULAMALARI **23**
3.1 Tıbbı Karar Destek Sistemleri 23
3.2 Tıp Alanında Veri Madenciliği Uygulamaları 26
4. METODOLOJİ **37**
4.1 K-En Yakın Komşu Algoritması 37
4.2 Bayesyen Sınıflandırma 38
4.3 Karar Ağaçları 39
4.3.1 ID3 Algoritması 40
4.3.2 C4.5 Algoritması 41
4.3.3 CART Algoritmaları 42
4.3.4 SLIQ 42
4.3.5 SPRINT 43
4.4 Yapay Sinir Ağları 44
5. UYGULAMA **47**

5.1 Çalışmada Kullanılan Yazılım Programı 47
5.2 Anemi 48
5.2.1 Demir eksikliği anemisi 51
5.2.2 Aplastik anemi 51
5.2.3 Talasemi 52
5.2.4 Romatolojik hastalıklara bağlı anemi 52
5.2.5 Böbrek hastalıklarına bağlı anemi 53
5.3 Veri Seti 53
5.3.1 Veri seti ön işleme prosedürü 56
5.4 Uygulama 57
6. SONUÇ ve ÖNERİLER 67
KAYNAKLAR 69
EKLER 79
ÖZGEÇMİŞ 83

KISALTMALAR

VM : Veri Madenciliği
CRISP-DM : Cross Industry Standard Process for Data Mining
ROC : Receiver Operating Curve
TKDS : Tıbbi Karar Destek Sistemleri
ID3 : Induction of Decision Trees
CART : Classification and Regression Trees
SLIQ : Supervised Learning in Quest
SPRINT : Scalable Parallelizable Induction of Decision Trees
THD : Türk Hematoloji Derneği
MCHC : Ortalama Alyuvar Hemoglobin Yoğunluğu
MCV : Ortalama Eritrosit Hacmi

ÇİZELGE LİSTESİ

Sayfa

Çizelge 2. 1: Karışıklık matrisi. 13
Çizelge 3. 1: Literatür araştırması. 28
Çizelge 4. 1: Bazı ateşleme fonksiyonları. 45
Çizelge 5. 1: Veri seti. 54
Çizelge 5. 2: Kan parametreleri ve açıklamaları. 55
Çizelge 5. 3: Test grupları. 57
Çizelge 5. 4: Test sonuçları. 59
Çizelge 5. 5: Karışıklık matrisi. 62

ŞEKİL LİSTESİ

Sayfa

Şekil 2. 1: Veri madenciliği ile diğer disiplinler arası ilişki (Turban ve diğ, 2011). ... 5
Şekil 2. 2: Veri madenciliği metodolojilerinin kullanımları (Turban ve diğ, 2011). ... 8
Şekil 2. 3: CRISP-DM süreci (Larose, 2005). ... 9
Şekil 2. 4: Veri hazırlama aşaması (Turban ve diğ, 2011). ... 11
Şekil 2. 5: Örnek ROC eğrisi (Han ve Kamber, 2006). ... 15
Şekil 3. 1: Tıbbi karar destek sistemi bileşenleri (Mendonça, 2004). ... 24
Şekil 4. 1: Örnek karar ağacı (Silahtaroğlu, 2013) ... 39
Şekil 4. 2: Yapay sinir ağı örneği. ... 45
Şekil 5. 1: RapidMiner ekran görüntüsü. ... 48
Şekil 5. 2: Hemoglobin yapısı (Url-2) ... 48
Şekil 5. 3: Kırmızı kan hücresi sayısı (Url-3). ... 49
Şekil 5. 4: AVICENNA ekran görüntüsü. ... 54
Şekil 5. 5: Tam kan sayımı testi örneği. ... 55
Şekil 5. 6: AVICENNA Hastane Bilgi Yönetim Sistemi. ... 56
Şekil 5. 7: Yapay sinir ağları modeli. ... 61
Şekil 5. 8: Karar ağacı yapısı. ... 63

VERİ MADENCİLİĞİ YÖNTEMLERİNİ KULLANARAK ANEMİ SINIFLANDIRILMASINA YÖNELİK BİR UYGULAMA

ÖZET

Günümüzde gelişen bilgi teknolojileri sayesinde toplanan, saklanan veri miktarı her geçen gün artmaktadır. Veri ve elde edilen bilgi gittikçe önem kazansa da artan veri hacmi insan algısı ile verileri analiz etmeyi imkansız kılmaktadır. Büyük veri tabanlarından istenilen bilgilere klasik sorgulama yöntemleri ile kolayca ulaşılabilse de verilerin barındırdığı gizli bilgilerin öneminin anlaşılması veri madenciliği uygulamalarını gerekli hale getirmektedir.

Veri madenciliği, büyük miktarlardaki verinin içinden anlamlı ve yararlı, ilişki ve kuralların bilgisayar programları aracılığıyla analiz edilmesidir. Böylece eldeki verilere uygulanan belirli yöntemlerle mevcut veya geleceğe yönelik anlamlı bilgiler elde edilebilmektedir. Artan veri hacmine rağmen; verilerin toplanması ve saklanmasındaki kolaylık, veri tabanı yönetim sistemlerindeki teknolojik gelişmeler, büyük miktarda ve komplike veri içerisinden gizli kalmış yararlı bilgilerin elde edilebilmesi veri madenciliği uygulamalarına olan ilgiyi arttırmaktadır.

Veri madenciliği uygulamaları, verinin bilgiye dönüşümünün öneminin anlaşılması ile telekomünikasyon, bankacılık, sigorta, tıp gibi çeşitli alanlarda yaygın olarak kullanılmaktadır. Tıp alanındaki verinin büyüklüğü ve hayati önem taşıması, veri madenciliğinin bu alanda da uygulanmasını gerekli kılmaktadır. Veri madenciliği yöntemleri tıp alanında kronik hastalıklarda erken uyarı sinyallerinin geliştirilmesi, tıbbi tedavi süreçlerinin optimizasyonu, hasta akış planlarının yapılması gibi birçok farklı amaçla kullanılmaktadır. Bu doğrultuda veri madenciliği yöntemleri ile önceden bilinmeyen, ilk bakışta fark edilemeyen, gizli kalmış anlamlı ve değerli bilgilerin elde edilmesi etkili tedaviyi destekleyici bir unsur oluşturmaktadır.

Çalışmamızın amacı veri madenciliği yöntemleri ile anemi hastalıklarının sınıflandırılması üzerine analizler yaparak hastaların anemi tipinin belirlenmesine yardımcı olmaktır. Çalışmamızda Uludağ Üniversitesi Tıp Fakültesi Hematoloji Anabilim Dalı bünyesinde farklı tip anemi hastalarına ve hematolojik açıdan sağlıklı bireylere ait tam kan sayımı testleri incelenerek bir veri seti oluşturulması hedeflenmektedir. Bu veri seti temel alınarak farklı test grupları oluşturulması planlanmaktadır. Test gruplarına veri madenciliği yöntemlerinden Karar Ağaçları, Yapay Sinir Ağları, Bayesyen Sınıflandırma ve K-En Yakın Komşu yöntemleri uygulanarak anemi hastalıklarının sınıflandırılması planlanmaktadır. Bu kapsamda uygulanan dört farklı yöntemde parametre değişikliklerine gidilerek hem yöntem içinde parametreler arası, hem de testlerde yöntemler arası karşılaştırmalar yapılacaktır. Böylece tam kan sayımı testi sonuçlarına veri madenciliği yöntemleri uygulanması ile anemi hastalıklarının sınıflandırılmasında doktorlara yardımcı olacak bir sistem geliştirilmesi planlanmaktadır.

CLASSIFICATION OF ANEMIA USING DATA MINING METHODS: AN APPLICATION

SUMMARY

Nowadays, as a result of evolving information technology, amount of collected and stored data is increasing day by day. Thanks to information technology, companies have huge data stores obtained from different sources. The important point is to convert data into knowledge, otherwise they have no meaning, just data pollution. Even if data and information gain importance, increasing volumes of data makes impossible to analyze the data with human perception. Although accessing the desired data in huge databases is easy by classical polling methods, data mining applications are necessary to identify the meaning of the importance hidden information in datasets.

There are various definitions of data mining in literature. Data mining, however, generally aims to gain useful knowledge and hidden patterns from large amount of data using computer programs. Data mining converts data into knowledge and useful information. Thus, it is possible to acquire present or prudential meaningful knowledge by means of data mining applications. This knowledge gained from data has an important role in decision making process. "Unknown" knowledge, which refers to hardness of estimation, can be obtained by data mining. Data mining applications, help to discover valuable knowledge even if they hidden in datasets. As a result of these applications, databases of companies work as a valuable asset. The demand for data mining applications is increasing due to easiness of collecting and storing data, technological development in database management system, and gaining useful information from huge and complex data.

As a result of understanding the importance of the conversion data into knowledge, data mining applications used in various areas such as marketing, telecommunication, banking, insurance, and etc. In these areas, data mining applications can be used for several purposes such as specifing customer profile, estimating sales, offering personalized services, preventing machine errors, determining criminal profiles. Additionally, medical data can be easily obtained and stored with the help of electronical medical instruments and hospital information management systems. Usage of data mining applications in medical field is crucial due to the huge amount and vital importance of data. Data mining applications are attained for various purposes in medical field such as recognizing early symptoms of chronic diseases, optimizing process of medical treatment, and planning patient flow. In this respect, obtaining valuable hidden information with data mining applications encourages effective treatment.

Medical decision support systems (MDSS) are computer based systems that help physicians to diagnose and decide process by analzing medical informations and

patients' history. MDSS are used to improve health and healthcare systems which offer wisely filtered medical informations or patient specialized datas to physicians on time. Different data mining applications are used to improve and develop MDSS. MDSS have lots of advantages such as prodiving accurate diagnosis, timely screening of diseases, preventing erroneous drup interactions, imporving effectiveness of the decision-making process and minimizing possible incorrect interventions to the patient. One of the other significant advantages of MDSS is ensuring early diagnosis which helps to reduced time on treatment, preventing usage of unnecessary tests and drugs, and decreased healthcare costs. Additionally, MDSS helps to reduce the erroneous diagnosis ratio based on false self-expression of patient or inadequate knowledge of physicians.

CRISP-DM (Cross Industry Standard Process for Data Mining) which breaks data mining process into six major phases, is a commonly used approach in data mining applications. The process begins by identifing the problem and determining aims and goals of study. It, then, continues with collecting, understanding and exploring data. Generally, this step necessitates the support of experts. Further, data proccessing step includes the operations applied from raw data to final data, which is really labor and time intensive. Following the preparation of data, modeling step is applied which is usually a comparision between different models. If final data does not fit with selected model, be turned to previous step and data is rearranged according to the model. In evaluation process, model is evaluated in terms of productivity and quality and, in turn results are compared with the evaluation criteria defined at the start of the project. On the other hand, in this step it is evaluated that the model meets the goals of the study or not. Finally, in deployment process, succesful model is choosed and implemented according to the strategy concluded.

Data mining methods can be seperated into two groups: predictive and descriptive models. Predictive modeling is a technique used to predict future behavior and anticipate the consequences of change. Classification and regression are predictive models. In classification models, target groups are known at the begining, however analyzing, and placing data into appropriate group are the critical steps of this model. In regression models, the main objective is to choose the rigth model that can correlate inputs and outputs accurately, and predict possible results.

On the other hand, in descriptive models, results are obtained for taking an action by defined pattern from the avaliable data. Clustering, association rules and sequence patterns are descriptive models. In clustering models, data is analyzed and converted to inherent, homogenous clusters. In association rules, the common grounds between different products and behaviours are identified. In sequence patterns, trends or relationships that emerging in time are defined.

In literature review, various articles and studies are examined. Result of this research, some articles related to data mining applications in the medical field are selected, and will be presented in a summary table based on applications and methods used. In addition to the summary table, a few studies have specifically focused on anemia will be discussed in more detailed.

The main purpose of our study is to classify anemia with various data mining applications and compare them to generate optimal classification method. Hemoglobin is a main part of red blood cells , binds to oxygen. When blood lacks

enough red blood cells or hemoglobin, the condition called anemia occurs. Anemia is the most common disease of blood disorders and is a global health problem affecting the public health in developed or developing countries. In our study, we will classify the five different types of anemia: iron-deficiency anemia, aplastic anemia, thalasemia, renal anemia and rheumatoid anemia. In scope of our study, not only the these types of anemic pations's complete blood count test results but also hematologically healty indivudials results are examined.

Mentioned complete blood count test results belonging to different types of anemia patients and hematological healthy individuals are collected within the scope of Uludağ University Internal Medicine Division of Hematology Department. Uludağ University Patient Management Information System AVICENNA is used for collecting data. This system is based on manuel logic and, in turn, collecting data process is very difficult. In collecting data process, physicians gave support due to requirement of expert knowledge in medical data. Also, our study is presented to Uludağ University Ethical Committee for approval. Approximately, 25 hematologic inputs in patients' blood count test are eliminated to 8, because of their strong relation with anemia. However, patients' age and gender input variables are also used. Therefore, in our study 10 different inputs are used for classification.

Different test groups are planned to be created by changing inputs of generated dataset. There will be no changes in the blood inputs, but the only differences will be made in gender and age inputs. Classification of anemia diseases will be obtained by applying various data mining applications such as Decision Tree, Neural Networks, Bayesian classification and K-nearest Neighbor by using RapidMiner software program. In this context, comparisons between methods in test groups as well as parameters in methods will be attained by changing parameters in four different methods applied in our research. In this way, developing a system which assisting the physicians about classifying anemia disease is planned by applying data mining methods to complete blood test results.

1. GİRİŞ

Günümüzde teknolojinin ve bilgi sistemlerinin gelişimi ile pazarlama, bankacılık, telekomünikasyon ve birçok alanda olduğu gibi tıp ve sağlık alanında da birçok veri sayısal ortamda saklanabilir ve kolayca erişilebilir hale gelmiştir. Gelişen bilgisayar teknolojisi ve veri tabanları, saklanan ve işlenebilen veri miktarını artırmış fakat böylece kontrol de zorlaşmıştır. Veri miktarı arttıkça bu verilerin insanlarca anlaşılması ve yararlı bilginin elde edilmesi herhangi bir araç kullanılmadığı takdirde imkansız hale gelmiştir. Bu duruma çözüm olarak "Veri Madenciliği (Data Mining)" kavramı geliştirilmiştir.

Veri madenciliği, verileri farklı birçok açıdan analiz etme ve o verilerden anlamlı bilgiler elde edilmesi işlemi olarak tanımlanabilir. Veri madenciliği sayesinde birçok verinin içerisinde anlamlı ve yeni ilişkiler keşfedilmesi ile değerli bilgiye ulaşılabilmektedir. Bilgisayar ortamında toplanan ve saklanan tıbbi bilgiyi analiz etmek klasik sorgulama yöntemleriyle veya basit gözlemlerle mümkün olmadığı için anlamlı ve yararlı örüntüleri tespit edebilmek için geliştirilmiş olan veri madenciliği teknikleri gerekli bir araç haline gelmiştir.

Bu çalışmada veri madenciliğinin tıp alanında kullanımı incelenmiş ve bu alanda bir uygulama yapılması amaçlanmıştır. Çalışma kapsamında farklı tip anemi hastalarına ve hematolojik açıdan sağlıklı bireylere ait veriler, veri madenciliği sınıflandırma yöntemlerinden olan Karar Ağaçları, Yapay Sinir Ağları, Bayesyen Sınıflandırma ve K-En Yakın Komşu yöntemlerinin kullanılması ile sınıflandırılmaya çalışılacak ve doğruluk oranları kıyaslanacaktır.

Tez çalışmasının amacı, veri madenciliği ile anemi hastalıklarının sınıflandırılması üzerine bir analiz yapmak ve hastaların anemi tipinin belirlenmesine yardımcı olmaktır. Bu sayede, tıp alanında hizmet veren tüm uzmanlara yardımcı olacak ve kolaylık sağlayacak bir çalışma yapılması planlanmaktadır.

Veri madenciliği bölümünde veri madenciliğinin tanımı ve ilgili kavramlar hakkında genel bilgiler verilmiş, uygulama alanları, süreç ve yöntemlerine değinilmiştir. Tezin üçüncü bölümünde kısaca tıbbi karar destek sistemlerinden bahsedilmiş, tıp alanındaki veri madenciliği uygulamalarını içeren literatür taramasına yer verilmiştir. Tezin dördüncü bölümünde çalışma içerisinde kullanılan veri madenciliği methodları hakkında detaylı bilgi verilmiştir. Çalışmanın beşinci bölümü olan uygulama bölümünde çalışmada kullanılan yazılım programına ve veri seti elde edilmesi sürecine değinilerek anemi hastalığı ve çeşitleri hakkında genel bilgi verilmiştir. Son olarak yapılan çalışmada elde edilen bulgular sunularak doğruluğu gösterilmiş, gelecek çalışmalar için önerilerde bulunulmuştur.

2. VERİ MADENCİLİĞİ

2.1 Veri Madenciliği Tanımı

Veri madenciliğinin literatürde birçok farklı tanımı bulunmaktadır. Genel anlamda veri madenciliği eldeki verilerin analiz edilmesi ile verilerin anlamlı bilgilere dönüştürülmesi sürecidir.

Özkan (2008) veri madenciliğini "değeri olan" bilgiyi birçok veri arasından elde etme işi olarak tanımlamaktadır. Böylece eldeki verilere uygulanan belirli yöntemlerle mevcut veya geleceğe yönelik anlamlı bilgiler elde edilebilir. Elde edilen bu bilgiler karar verme aşamasında önemli rol oynamaktadırlar.

Veri madenciliğinin literatürde bulunan birkaç tanımı şöyledir:

"Veri madenciliği, depolanmış yüksek miktardaki veriden istatistiksel ve matematiksel teknikler gibi desen tanımlayıcı teknolojiler kullanarak anlamlı ve yeni ilişkiler, desenler ve trendler keşfetme prosesidir" (Akpınar, 2000).

Veri madenciliği, çok miktarda veri içerisinden anlamlı yollar, desenler ve kurallar keşfedilmesi işidir (Berry ve Linoff, 2011).

"Veriler içinde desen aramak insan hayatının başlamasından beri süre gelen bir süreçtir. Bilim insanlarının işi ise veriler içinde fiziksel dünyanın nasıl çalıştığını anlatan modeller aramak ve bulacağı modeller yardımıyla ortaya çıkacak yeni durumlarda neler olacağını tahmin edecek teoriler geliştirmektir" (Chakrabarti ve diğ, 2008).

Dönmez (2008)'e göre "Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönden farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır."

Gerekli verilere, gerektiği zamanda hızla erişilebilmesi gerekmektedir. Geleneksel dosyalama sisteminin yetersiz kalması ile veriyi saklama ve erişim konusunda kolaylık sağlayan veri tabanları oluşturulmuştur. Bu sayede veriler belli bir düzen

çerçevesinde toplanmış, düzenlenmiş, saklanmış ve işlenmesi sağlanmıştır. Günümüzde yaygın olarak kullanılan veri ambarları, günlük kullanılan verilerin işlemeye uygun hale getirilmiş özetini veri tabanlarında saklamayı amaçlamaktadır (Baykal, 2006).

2.2 Veri Madenciliğinin Gelişimi

Veri madenciliği, kavramsal olarak 1960'lı yılların sonunda, bilgisayarların veri analiz problemlerini çözmek için kullanılmasıyla ortaya çıkmış, bu yönde basit öğrenmeli bilgisayarlar geliştirilmiştir. Bu işlemler için veri taraması (*data dredging*) ve veri yakalaması (*data fishing*) gibi isimler kullanılmıştır.

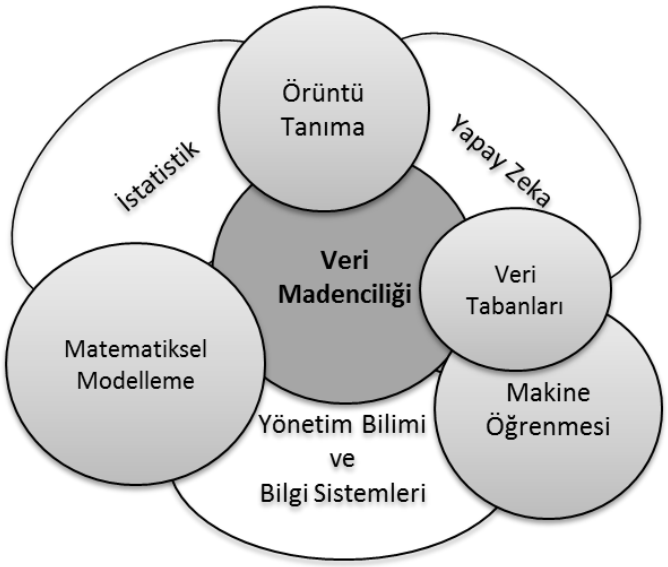
1970'lerde İlişkisel Veri Tabanı Yönetim Sistemleri uygulamaları kullanılmaya başlanmıştır. Bilgisayar uzmanları bununla beraber basit kurallara dayanan uzman sistemler geliştirmişler ve basit anlamda makine öğrenimini sağlamışlardır.

1980'lere gelindiğinde veri tabanı yönetim sistemleri yaygınlaşmış ve bilimsel alanlarda, mühendisliklerde vb. alanlarda uygulanmaya başlanmıştır. Şirketler, müşterileri, rakipleri ve ürünleri ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Bu yıllarda geleneksel algoritmalara dayalı istatistiksel araçlar kullanılmış ve mükemmel sonuçlar alınmıştır, fakat bilgi gereksiniminden dolayı sadece istatistik uzmanları tarafından kullanılabilmiştir. Aynı zamanda veri boyutu arttıkça güvenilirliğin azaldığı gözlemlenmiştir.

1990'larda veri miktarı katlanarak artan veri tabanlarından, faydalı bilgilerin nasıl çıkarılabileceği düşünülmeye başlanmıştır. Veri madenciliği ismi bilgisayar mühendisleri tarafından kullanılmaya başlanmıştır.

2000'li yıllarda veri madenciliği sürekli gelişmiş ve hemen hemen tüm alanlara uygulanmaya başlanmıştır. Alınan sonuçların faydaları görüldükçe, bu alana ilgi artmıştır.

Veri madenciliğinin gelişmesinde ve ilerlemesinde birçok disiplinin katkısı olmuştur. Veri madenciliği ile ilişkili disiplinler Şekil 2.1'de gösterilmiştir:



Şekil 2. 1: Veri madenciliği ile diğer disiplinler arası ilişki (Turban ve diğ, 2011).

2.3 Veri Madenciliğinin Önemi

Teknolojinin gelişimi ve buna bağlı olarak maliyetlerin azalması sonucunda önceden çok daha özet halinde bilgi depolayabilen veri tabanlarının yerini günümüzde daha detaylı veri depolayabilen böylece kapsamlı veri elde etmemize olanak sağlayan veri tabanları almıştır. Geçmişte veriler özet olduğundan aralarındaki bağlantıların basit yöntemlerle tahmin edilebilmesi mümkündü. Günümüzde ise büyük miktarda ve detaylı olarak elde edilen verilerin içerisindeki bağlantı ve yararlı bilgileri edebilmek ancak bilgisayar algoritmaları kullanılarak mümkün olmaktadır. Diğer bir sebep ise verilerin daha detaylı tutulması ile değişken sayısının artması, böylece sorgulamanın daha komplike hale gelmesidir. Veri madenciliği, çeşitli algoritmalar kullanarak büyük miktarda, komplike veri içerisinden yararlı bilgilerin elde edilmesi sürecidir.

Verilerin toplanması ve saklanmasındaki kolaylık, veri tabanı yönetim sistemlerindeki teknolojik gelişmeler, bilgi işlem maliyetlerindeki meydana gelen düşüşler ile birlikte kullanılabilecek analitik araçların artması da veri madenciliği uygulamalarına olan ilgiyi arttırmaktadır (Park ve diğ, 2001).

Geçmişte kararların reaktif (bir olay ya da durum sonucuna bağlı) olarak verilmesinde bir sakınca görülmezken, günümüzde bir olay gerçekleşmeden onu fark etmenin ve koruyucu önlemler almanın önemi bilinmektedir. Rekabetin de etkisi ile müşterilerin ihtiyaçlarını fark edebilmek ve bunlar doğrultusunda ürün veya hizmet

sunabilmek organizasyonlara esneklik kazandırmakta ve büyük avantaj sağlamaktadır. İhtiyaç ve beklentilerin doğru ve güvenilir tahmin edilmesi ancak doğru bilgi bağlantılarının elde edilmesi ile sağlanabilir. Veri madenciliğinin eldeki veriler ile gerçekleştirilen analizler sayesinde gelecekteki eğilim ve davranışları öngörmesi, proaktif (önceden görerek, sonuç yaşanmadan) yaklaşım için gerekli bilginin elde edilmesine imkan sağlamaktadır.

Kullanılan veri tabanlarının sayısındaki hızlı artış, her geçen gün teknolojik olarak daha da geliştirilen güçlü ve ekonomik veri tabanı sistemlerinin varlığı ile açıklanabilmektedir. Veri ve veri tabanı sistemlerindeki bu hızlı artış, işlenmiş veriyi kullanışlı bilgiye dönüştürebilecek yeni teknolojilere olan ihtiyacı arttırmıştır. Bu nedenle veri madenciliği her geçen gün önemi artan bir araştırma alanı haline gelmiştir (Sund, 2002).

2.4 Veri Madenciliğinin Uygulama Alanları

Veri madenciliği uygulamaları, verinin bilgiye dönüşümünün öneminin anlaşılması ile telekomünikasyon, bankacılık, sigorta, tıp gibi çeşitli alanlarda yaygın olarak kullanılmaktadır.

Veri madenciliği analizi sonucunda elde edilen bilgi; pazar analizi, sahtekârlık tespiti, müşteri elde tutma, üretim kontrol ve bilimsel keşiflere kadar pek çok alanda çeşitli uygulamalar için kullanılabilir (Han ve Kamber, 2006).

Veri madenciliği kullanılan bazı uygulamalar (Turban ve diğ, 2011):

- **Pazarlama:** Günümüzde rekabetin gittikçe yoğunlaştığı pazarlama sektöründe müşterinin elde tutulması ve yeni müşterilerin kazanılmasında, müşterilerin satın alma alışkanlıkların belirlenmesinde, pazar payını ve kar getirisini artırmak amacı ile kampanyaların belirlenmesinde kullanılmaktadır. Özellikle en çok kazanç sağlayan müşterilerin belirlenmesi ve ilişkileri güçlendirmek adına bu müşterilerin öncelikli ihtiyaçlarının belirlenmesi gibi konularda müşteri ilişkileri yönetiminde birçok veri madenciliği uygulaması kullanılmaktadır.

- **Perakende ve Lojistik:** Doğru envanter miktarını belirlemek adına spesifik noktaların satışlarının öngörülmesi, farklı ürünler arasındaki bağlantıların belirlenerek mağazada ürünlerin yerleşimi veya çeşitli promosyonların geliştirilmesi,

taşıma ve satışı optimize edebilmek adına farklı ürün tiplerinin tüketim seviyelerinin belirlenmesi, tedarik zincirindeki farklı örüntülerin tanımlanması gibi farklı konularda veri madenciliği tekniklerinden faydalanılmaktadır.

- **Banka ve Sigortacılık:** Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi, sahtekârlık ve dolandırıcılık tespiti, kredi taleplerinin ve geri ödeme durumlarının belirlenmesinde kullanılmaktadır. Aynı zamanda hangi müşterilerin hangi poliçeleri satın alabileceğinin tahmini, yanlış hasar ödemelerinin veya sahtekarlıkların belirlenerek önlenmesi gibi sigortacılık temelli problemlerin çözümünde de veri madenciliği yöntemlerinden yararlanılmaktadır.

- **Telekomünikasyon:** Telekomünikasyon sektörü gibi olgun bir pazarda müşteriyi elde tutmanın yeni müşteri kazanımından çok daha az maliyetli olduğunun kabullenilmesi ile müşteri profillerinin belirlenmesi, uygun kampanyalar ve teklifler sunulması, kaybedilen müşterinin geri kazanılması gibi birçok alanda kullanılmaktadır.

- **Üretim:** Veri madenciliği duyusal verilerin kullanımı ile makine hatalarının önceden tahmini, üretim kapasitesini optimize etmek adına üretim sistemindeki anormallik ve genellemelerin belirlenmesi, ürün kalitesini artırmak adına farklı örüntülerin keşfedilmesi gibi konularda kullanılmaktadır.

- **Tıp:** Hastalıkların tahmin edilerek önlemlerin alınması, hastalığın erken teşhisinin sağlanmasında ve pek çok alanda kullanılmaktadır. Çalışmanın üçüncü bölümünde veri madenciliğinin tıp alanında kullanımına ayrıntılı bir şekilde değinilmiştir.

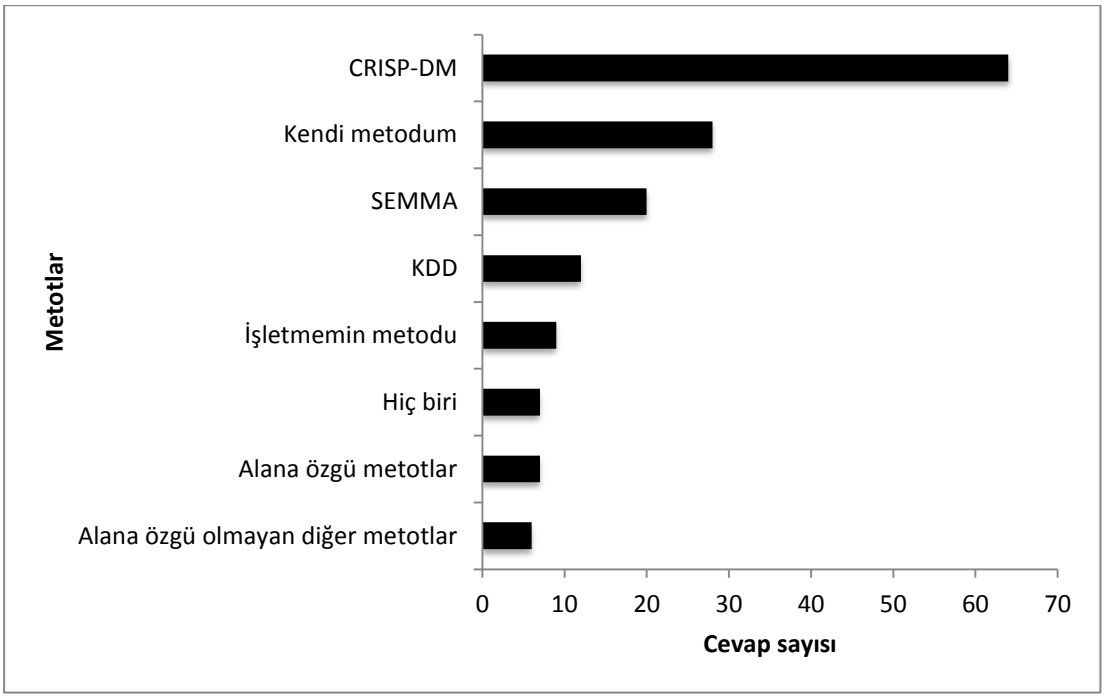
- **Devlet ve Savunma:** Ülkemizde uygulanmaya başlanan e-devlet uygulaması ile veri madenciliği yöntemleri daha kolay uygulanabilmektedir. Bu sayede emniyet güçleri için şüpheli durumların tespiti, vergi ya da vergi ile ilgili yolsuzlukların belirlenmesi, suç oranları ve suçlu profillerinin çıkarılması gibi birçok konuda veri madenciliğinden yararlanılabilmektedir.

- **Turizm:** Turizm sektöründeki birçok şirketin başlıca konularından olan farklı servislerin satışlarının tahmin edilerek en çok kazanç getirecek şekilde optimal fiyatların belirlenmesi, farklı lokasyonlardaki talebin tahmin edilmesi ile kısıtlı kaynakların doğru yerlere tahsis edilmesinin sağlanması, yüksek kazanç sağlayan müşterilerin belirlenmesi ve onları uzun süre elde tutabilmek için kişiselleştirilmiş hizmetler sunulması gibi konularda veri madenciliğinden yararlanılmaktadır.

- **Eğlence:** İzleyici verilerinin analiz edilmesi ile televizyonun en çok izlediği saatteki programlara karar verilerek reklamların en çok getiri getirecek şekilde yerleştirilmesinde, film ve dizilerin başarılarının öngörülerek yatırım kararları alınmasında, etkinliklerin planlanmasında ve talep tahmini yapılarak kaynak planlanmasında veri madenciliği yöntemleri kullanılmaktadır.

2.5 Veri Madenciliği Süreci

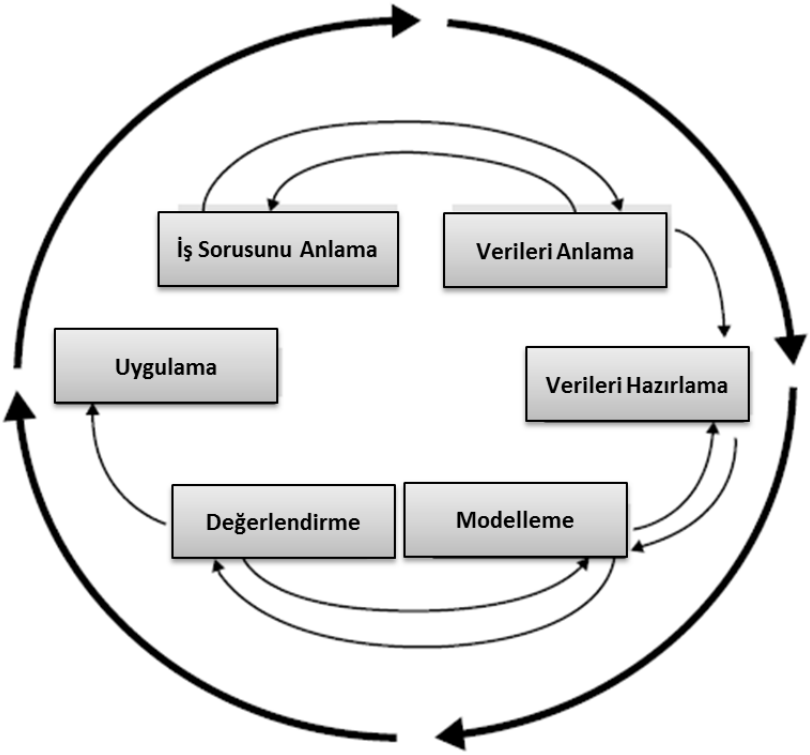
Veri madenciliği süreciyle ilgili yaygın kullanılan modelllerden biri Çapraz-Endüstri Standardı Süreci (CRISP-DM: Cross Industry Standard Process for Data Mining) modelidir. CRISP-DM, 1996 yılının sonlarına doğru henüz olgunlaşmamış veri madenciliği pazarının üç deneyimli analisti (DaimlerChrysler, SPSS ve NCR markalarının temsilcileri) tarafından tasarlanmıştır (CRISP-DM 1.0, 2010). Şekil 2.2'de veri madenciliğinde temel olarak hangi metodolojiyi kullanıyorsunuz sorusuna verilen cevapların oranından da en yaygın olarak CRISP-DM metodolojisinin kullanıldığı anlaşılabilmektedir.



Şekil 2. 2: Veri madenciliği metodolojilerinin kullanımları (Turban ve diğ, 2011).

CRISP-DM'e göre, bir veri madenciliği projesinde altı aşamadan oluşan bir yaşam döngüsü vardır. Fazlar sıra ile ilerlemektedir yani sonraki faz önceki fazın çıktılarına

bağlıdır. Fazlar arasındaki en önemli bağımlılıklar oklar ile belirtilmiştir. Şekil 2.3'te CRISP-DM süreci gösterilmektedir (Larose, 2005).



Şekil 2. 3: CRISP-DM süreci (Larose, 2005).

CRISP-DM'in tekrarlayıcı doğası dış daire ile sembolize edilmektedir. Genellikle belirli bir iş veya araştırma sorununun çözümü, ileride genel olarak aynı işleyiş kullanılarak tekrar çözüm elde edilebilmesine katkı sağlar.

Söz konusu süreç aşağıda belirtilen adımları içermektedir (Larose, 2005):

2.5.1 İş sorusunu anlama aşaması (Business Understanding)

Araştırma problemini anlama evresi olarak da adlandırılabilen bu evre veri madenciliği sürecinin ilk adımını oluşturmaktadır. Projenin hedef ve ihtiyaçlarının bir bütün olarak kesin şekilde ifade edilmesi, veri madenciliği formülasyonu açısından amaç ve kısıtlamaların belirlenmesi ve bu doğrultuda bir ön strateji oluşturulması süreçlerini kapsamaktadır. Amaç, işletme problemi üzerine odaklanmış ve açık dille ifade edilmiş olmalıdır. Çalışma sonunda elde edilecek sonuçların başarı oranlarının ne şekilde ölçüleceği bu adımda tanımlanmalıdır. Bu safhada sonuçların nasıl değerlendirilerek kullanılacaklarını bilmek önemlidir. Yapılan çalışma sonunda

doğru cevaplanmış birçok yanlış soru elde edilmek istenmiyorsa, çalışmanın cevap aranılan soru ile uyumlu olduğu kesinleştirilmelidir (Argüden ve Erşahin, 2008).

2.5.2 Verileri anlama aşaması (Data Understanding)

Veri anlama aşaması veri toplama adımı ile başlamaktadır. Veriler iç ve dış kaynaklardan elde edilebilir. İç kaynaklar müşteri kayıtları, satın almalar gibi işletmenin veri tabanlarıdır. Dış kaynaklar ise nüfus sayımı, araştırma şirketleri, veri tabanları gibi işletme dışından elde edilen verilerdir. Daha sonra ön bilgiler elde etmek adına verilerin keşfedilmeye başlanması, verilerin niteliklerinin tanımlaması, verilerin kalitesinin değerlendirilmesi ile süreç devam etmektedir. Projenin başarısı için analist veriler hakkında yeterli bilgiye sahip değilse bilgi sahibi olan bir kişiden yardım almalıdır.

Amaç, çalışmada kullanılacak olan verilere aşinalık kazanmaktır. Verileri tanıma aşaması ile araştırma probleminin tanımlanması aşaması iç içe geçmiş alt süreçlerdir. Problemi anladıkça farklı verilere bakmak, verilere baktıkça problem ile ilgili farklı bakış açıları kazanmak mümkün olmaktadır.

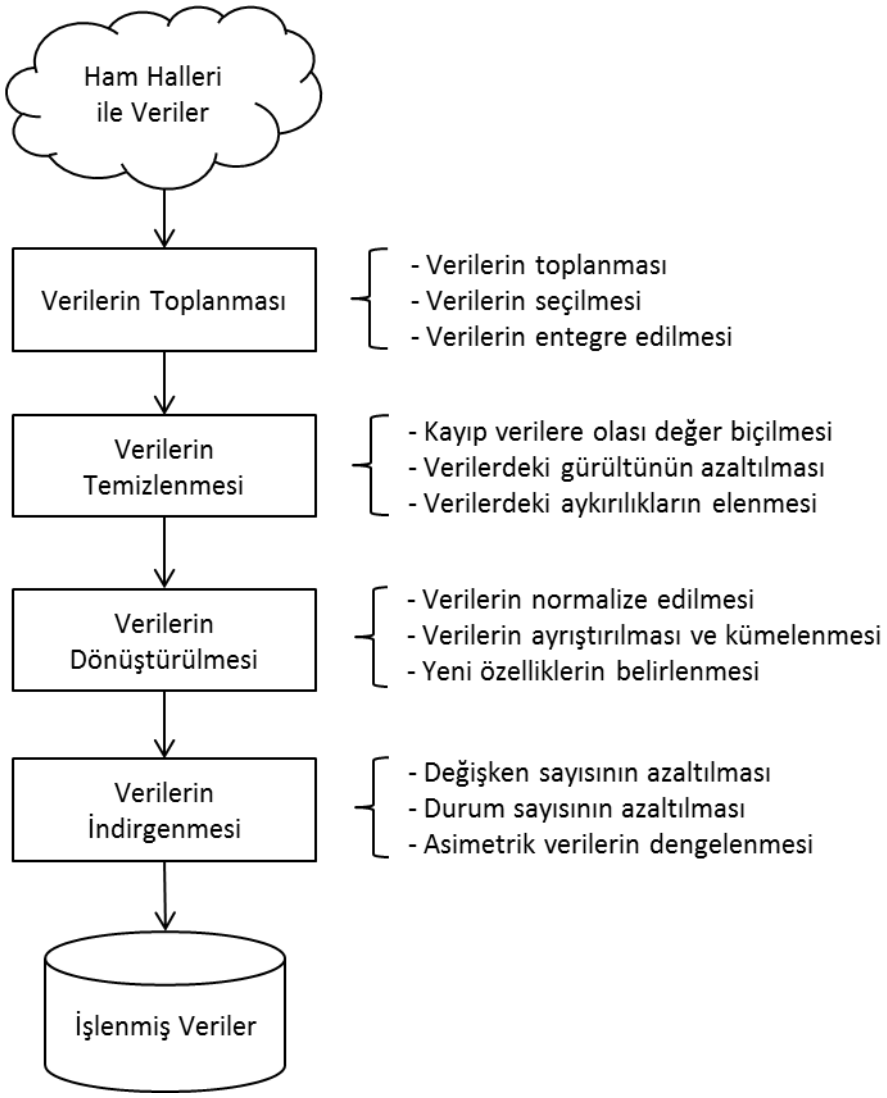
2.5.3 Verileri hazırlama aşaması (Data Preperation)

Bu aşama ham veriden başlanarak işlenecek son veriye kadar yapılması gerek tüm düzenlemeleri içeren emek olarak oldukça yoğun bir aşamadır. Analiz için uygun olan durum ve değişkenlerin belirlenmesi, verilerin dönüşümü ve temizlenmesi gibi süreçleri içermektedir.

Verilerin ön işlemden geçirilmesi aşaması veri madenciliği sürecinde, genellikle veri kümelerinin de büyüklüğü kaynaklı, en çok zaman alan aşamadır. CRISP-DM'in diğer aşamaları ile kıyaslandığında en çok efor ve zamanın (%80'inin) verilerin hazırlanması aşamasında harcandığı belirtilmektedir. (Turban ve diğ, 2011).

Şekil 2.4'te verilerin hazırlanması aşaması aşağıdaki gibi özetlenmiştir:

- *Verilerin Toplanması*: Verileri anlama aşamasında toplanmaya başlanan veriler, kayıtlar ve değişkenler seçilir. Verinin kalitesine ve hedefler ile ilişkisine, aynı zamanda var olan teknik limitlere dikkat etmek gerekir. Kullanılan kayıt sayısı ise bir diğer önemli noktadır. Çalışmanın eksik kalmaması veya tam tersi süreci uzatacak veri kirliliği olmaması adına kayıt sayısı değerlendirilmelidir. Farklı veri kaynaklarından gelen kayıtlar bu aşamada entegre edilmektedir.



Şekil 2. 4: Veri hazırlama aşaması (Turban ve diğ, 2011).

Verilerin farklı kaynaklardan elde edilmesi veri formatlarında farklılıkların olmasına neden olmaktadır. İleriki aşamalarda durumu çarpıtacak sonuçlar çıkmaması adına titiz davranılarak verilerin uyumlarının değerlendirilmesi ve mevcut olan uyumsuzlukların giderilmesi gerekmektedir.

- *Verilerin Temizlenmesi*: Hatalı ve tutarsız verilerin çıkarılarak verinin kalitesini artırma aşamasıdır. Yanlış girişlerden veya istisnai durumlardan kaynaklanan verilerin, sonucu çarpıtmasını önlemek adına çıkarılması tercih edilir. Kimi durumlarda veri kirliliğinin önlenmesi adına tüm popülasyonu temsil edebilecek bir örneklem ile ilerlenmesi tercih edilebilir. Kimi durumlarda ise eksiklik bulunan veriyi değerlendirmeden direkt çıkarmak yerine, eksik olan verilerin tamamlanması adına farklı modellemeler uygulanabilir.

- *Verilerin Dönüştürülmesi*: Eldeki verilerin belirli aralıklara indirilmesi işlemidir. Bir değişkenin çok yüksek değere sahip olmasından dolayı düşük değerli değişkenleri domine etmesi ile oluşabilecek potansiyel sapmaları hafifletmek için kullanılmaktadır.

- *Verilerin İndirgenmesi*: Veri madenciliği uygularken genelde büyük veri setleri tercih edilir fakat bu durum aynı zamanda problemleri de beraberinde getirmektedir. Özellikle bazı uygulamalarda değişken ve durum sayıları oldukça fazla olmaktadır, uygulayıcının durum ve değişken sayılarını yönetilebilir rakamlara indirgemesi gerekmektedir. Çözümlemelerin sadece belli boyutlar için yapılmasıyla, belli boyutların veri setinden çıkarılmasıyla veya büyük veri toplulukları yerine onu temsil eden daha küçük veri kümelerinin oluşturulması ile veri daraltılmasına gidilebilmektedir.

2.5.4 Modelleme aşaması (Modelling)

Modelleme aşaması problem için istenirse birkaç farklı yöntem uygulanması adına uygun modelleme tekniklerinin seçilip uygulandığı aşamadır. Bazı yöntemler veri veya problem tipi için uygun olmayabilir, bu yüzden gerekli olursa bir önceki aşama olan veri hazırlama aşamasına geri dönülebilir. Bu yüzden modelleme aşaması ile veri hazırlama aşamaları en uygun model elde edilene kadar yinelenen süreçlerdir.

Model kuruluş çalışmalarına başlamadan önce en uygun tekniğe karar verilmesi zordur. Farklı modeller kurulup doğruluk dereceleri kıyaslanarak en uygun modeli bulmak için çalışmalar yapılmasında yarar bulunmaktadır. Fakat kurulan modelin doğrululuk derecesi oldukça yüksek olsa bile, gerçek dünyayı tam anlamıyla modellediğini garanti edebilmenin mümkün olmadığı bilinmelidir.

Modelin doğruluğunun test edilmesinde kullanılan en kolay yöntem basit geçerlilik (*Simple Validation*) testidir. Verilerin %5 ile %33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde model öğrenimi gerçekleştirilir. Kurulan model test verileri ile test edilerek doğruluk oranı hesaplanır. Bir sınıflama modelinde doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile modelin doğruluk derecesi (*Accuracy*) hesaplanmaktadır (Akpınar, 2000).

Sınıflandırma problemlerinde modelin doğruluk derecelerinin değerlendirilmesinde karışıklık matrisi (*Confusion Matrix*) temel olarak kullanılmaktadır. Karışıklık matrisi modelin tahmin yeteneğini ölçen bir matristir. Çizelge 2.1'de ikili

sınıflandırma problemini gösteren bir karışıklık matrisi bulunmaktadır. Sol üst ve sağ altta bulunan değerler doğru kararları, diğerleri ise hataları göstermektedir. Faydalı bir araç olarak yaygın kullanılmasında tablonun yorumlanmasının kolay olması da etkili olmaktadır.

Çizelge 2. 1: Karışıklık matrisi.

	Tahmin Edilen Sınıf		
Gerçek Sınıf		C_1 (+)	C_2 (-)
	C_1 (+)	Doğru pozitif **TP**	Yanlış negatif **FN**
	C_2 (-)	Yanlış positif **FP**	Doğru negatif **TN**

Karışıklık matrisinde doğruluk, kesinlik, duyarlılık ve F-ölçümü oranları aşağıdaki şekilde hesaplanmaktadır:

$$Doğruluk\ Oranı\ (Accuracy) = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{(2.1)}$$

$$Kesinlik\ (Precision) = \frac{TP}{TP+FP} \quad \textbf{(2.2)}$$

$$Duyarlılık\ (Recall) = \frac{TP}{TP+FN} \quad \textbf{(2.3)}$$

$$F-ölçümü\ (F-measure) = \frac{2TP}{2TP+FN+FP} \quad \textbf{(2.4)}$$

Veri sayısının sınırlı olması durumunda modelin doğruluğunun test edilebilmesi için kullanılan bir diğer yöntem çapraz geçerlilik (*Cross Validation*) testidir. Bu yöntemde veri kümesi rastgele iki eşit parçaya ayrılır. İlk veri kümesi ile model

kurulur, ikinci küme ile test gerçekleştirilir ve doğruluk oranı hesaplanır. Daha sonra ikinci küme ile model kurulur, ilk veri kümesi ile test gerçekleştirilerek doğruluk oranı hesaplanır. Hesaplanan doğruluk oranlarının ortalaması kullanılır (Akpınar, 2000).

Diğer bir doğrulama tekniği K-kere çapraz doğrulama (*K-Fold Cross Validation*) yöntemidir. Bu yöntemde veri kümesi her biri yaklaşık olarak eşit büyüklükte *K* adet gruba ayrılmaktadır. Her zaman bir grup test için ayrılırken geriye kalan gruplarla model kurulur ve test için ayrılan verilerle doğruluk oranı hesaplanır. Bu durum yinelemeli olarak *K* defa tekrarlanır. Böylece her bir grup eğitim ve test için aynı sayıda kullanılmış olur. Modelin doğruluk oranı, *K* adet doğrululuk oranının ortalamasıdır (Han ve Kamber, 2006).

Leave-One-Out yöntemi K-kere çapraz doğrulama tekniğinin özel bir halidir. Bu teknikte *K* sayısı veri kümesinin örnek sayısı olan *N* sayısına eşittir. Model *N-1* adet örnek üzerinde öğrenilir ve dışarıda bırakılmış olan 1 örnek üzerinde test edilir. Böylece her veri bir kere test olarak ayrılarak veri sayısı kadar model kurulmuş olur. Zaman açısından verimli bir yöntem olmasa da az verinin olduğu durumlarda bir seçenek olarak yer almaktadır (Turban ve diğ, 2011).

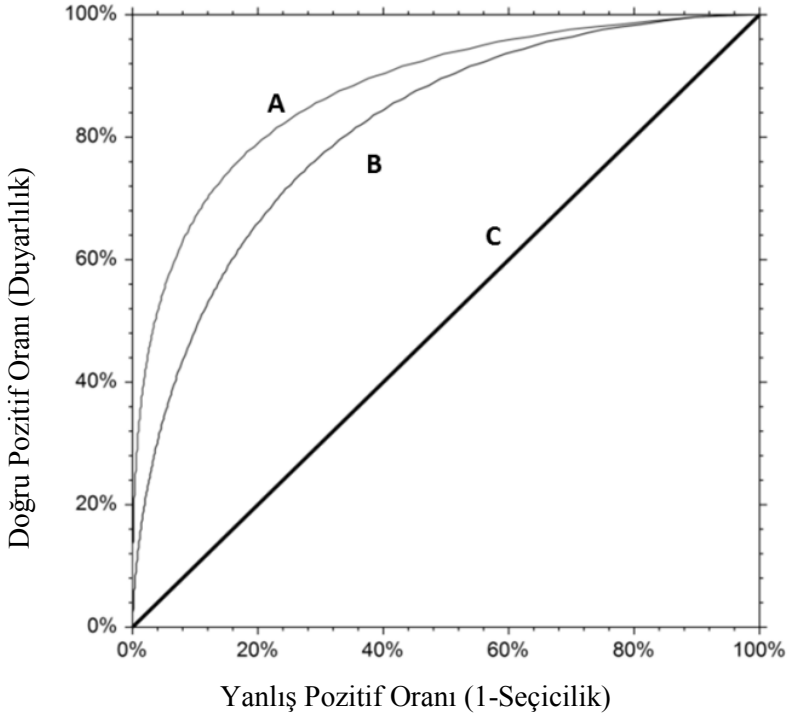
Özellikle küçük veri kümelerinin doğruluk tahmininde etkin çalışan bir başka teknik *bootstrapping* yöntemidir. Çapraz geçerlilik testinde olduğu gibi model tüm veri kümesi üzerine kurulmuştur. *N* adet örnekten oluşan veri kümesinden yerine koyarak *N* örnek seçilir ve bu küme öğrenim kümesi olarak ayrılır. Öğrenme kümesinde yer almayan örnekler sınama kümesi olarak kullanılır. Bu durum istenildiği kadar tekrarlanarak devam eder (Turban ve diğ, 2011).

Jackknifing yöntemi de *leave-one-out* yöntemi ile benzerlik göstermektedir. Bu yöntemde de her süreçte bir değer ayrı tutularak tahmin yapılır ve sonrasında o veri bilinen veriler arasına konularak tahmin işlemi bir sonraki ayrı tutulan veri üzerine kaydırılır, böylece model doğruluğu yinelemeli olarak hesaplanmaktadır.

ROC eğrisi (*Receiver Operating Curve*), modellerin doğruluk değerlendirilmesinde yaygın olarak kullanılan yöntemlerden biridir. Farklı sınıflandırıcıları karşılaştırmak için kullanılmaktadır. ROC eğrisi üzerindeki her bir nokta bir sınıflandırıcının oluşturduğu bir modele karşılık gelmektedir. ROC eğrisi, hassaslık ve özgüllük oranlarını kullanarak birimleri sınıflarına ayıran en uygun kesim noktasını belirler.

Yapılan sınıflamanın doğruluğu, ROC eğrisi altında kalan alanın büyüklüğüne bağlıdır.

ROC eğrisinin koordinat düzleminde Y ekseninde doğru pozitif (hassaslık), X ekseninde ise yanlış pozitif (1- özgüllük) olarak çizilmektedir. Şekil 2.5'te örnek olarak gösterilen ROC eğrisini incelendiğinde A'nın sınıflandırma performansı B'den daha iyidir, C ise 0.5'lik rastgele bir şansla aynı performanstadır.



Şekil 2. 5: Örnek ROC eğrisi (Han ve Kamber, 2006).

Karar verilen modelin doğrudan bir uygulama olarak kullanılabileceği gibi, yapılan başka bir uygulamanın alt parçası olarak kullanılması da mümkündür. Modelleme aşaması sonunda başarı kriterleri, daha önceki tecrübeler ve test sonuçlarına göre model teknik olarak değerlendirilir.

2.5.5 Değerlendirme aşaması (Evaluation)

Modelleme aşamasında uygulanan bir veya birden fazla modelin kalite ve verimliliklerinin ve modelin ilk fazda belirlenen amaçlara uygun olup olmadığının değerlendirilmesi sürecidir. Değerlendirme aşaması uygulanan veri madenciliği süreçlerinin tekrar gözden geçirilmesi ve sonuçlarının kullanımı için karar verilmesi süreçlerini içermektedir. Modelin belirlenen hedefler doğrultusunda olması

önemlidir, kapsanmamış bir konu olup olmadığı değerlendirmeye alınmalıdır. Aynı zamanda bu aşamada gelecekte hangi farklı verilerin kullanılabileceği ve çalışmanın geldiği yerin yeterli olup olmadığı da değerlendirilmektedir.

2.5.6 Uygulama aşaması (Deployment)

Son aşama olan uygulama aşamasında süreç sonucunda üretilen bilgiler belirlenen amaç doğrultusunda problem çözümü için kullanılmaktadır. Sonuçlar değerlendirilir, yorumlanır ve elde edilen sonuçlar neticesinde belirlenen stratejiler, kararlar gerçek hayatta uygulanır. Karar verilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak da kullanılabilir. Modelin amacı eldeki veriler hakkında bilinenleri artırmak olsa da, elde edilen sonuç verileri kullanılacak biçimde organize edilmeli ve sunulmalıdır. Aynı zamanda bu aşama, araştırma raporunun yazılması ve projenin tekrar gözden geçirilmesi işlemlerini kapsamaktadır. Raporlar, projelerin sonuçlarının aktarılması ve gerekirse başkaları tarafında da tekrarlanabilirliğinin sağlanması açısından önemlidir.

Zaman içerisinde sistemlerde ortaya çıkan değişiklikler, kurulan modellerin izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Modelin yanlış veriler kullanarak çalışmasının önüne geçmek için izleme önemlidir.

2.6 Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri kestirime dayalı (*predictive*) ve tanımlayıcı (*descriptive*) modeller olmak üzere ikiye ayrılmaktadır (Tsiptsis ve Chorianopoulos, 2010).
Kestirime dayalı modellerde amaç var olan veriler içindeki desenleri ortaya çıkarma veya var olan veriler ışığında gelecekteki bir durumu öngörmektir. Sınıflandırma modellerinde hedef gruplar en baştan bilinmektedir. Eldeki veriler analiz edilerek bilinen sınıflara ayrılmakta ve gelecekteki bir durum öngörülmektedir. Tahmin modellerinde ise mevcutta var olan veriler üzerine çalışılmakta ve gözlemlenen bu veriler içerisinden bir sonuç tahmini yapılmaktadır. Sınıflandırma (*Classification*) ve Regresyon (*Regression*) kestirime dayalı modellerdendir. Tanımlayıcı modellerde ise eldeki verilerden örüntüler tanımlanarak, aksiyon almak adına ortaya sonuçlar çıkarılmaktadır. Kümeleme modellerinde eldeki veriler analiz edilerek doğal bir kümeleme yapılmaktadır. İlişkilendirme kurallarında farklı durumlar, ürünler veya

davranışlar arasındaki ortak paydalar, ilişkiler belirlenmekte; sıralı dizilerde ise zamanla ortaya çıkan ilişkiler ortaya çıkarılmaktadır. Kümeleme (*Clustering*), İlişkilendirme Kuralları (*Association Rules*) ve Sıralı Diziler (*Sequence Patterns*) tanımlayıcı modeller arasında yer almaktadır.

2.6.1 Sınıflandırma modelleri

Mevcut verilerin içerdikleri ortak özelliklere göre gruplara ayrıştırılmasına sınıflandırma adı verilmektedir. Sınıflandırma birçok farklı alanda kullanılabilir. Kredi başvurularının değerlendirilmesi, kullanıcı davranışlarının belirlenmesi veya hastalık teşhisi gibi birçok farklı konuda sınıflandırma örneklerine rastlamak mümkündür. Sınıflandırma bir öğrenme algoritmasına bağlı olarak uygulanır. Bir diğer deyişle sınıflandırma hangi sınıfa ait olduğu bilinmeyen mevcut verilerin, sınıflarının belirlenmesi işlemidir (Özkan, 2008).

Sınıflandırma için kullanılan başarı değerlendirme kriterleri, doğru sınıflandırma, anlaşılabilirlik, ölçeklenebilirlik, kararlılık ve kuralların yapısıdır (Han ve Kamber, 2006). Doğru sınıflandırma başarısı, oluşturulan modelin daha önce görülmemiş örnekleri sınıflandırmada ne kadar etkin olduğunun bir ölçüsüdür. Anlaşılabilirlik, oluşturulan modelin kullanıcılar tarafından ne kadar yorumlanabilir olduğunun göstergesidir. Ölçeklenebilirlik, modelin büyük miktarda veri kümeleri çalışabilmesidir. Kararlılık veri kümesindeki eksik veya gürültülü verilerin varlığında da modelin doğru sonuç vermesidir. Kuralların yapısı ise birbirleri ile örtüşmeyen kuralların model içerisinde bulunma durumunu göstermektedir.

Verilerin sınıflandırılması işlemi iki adımdan oluşmaktadır (Han ve Kamber, 2006):

İlk adım veri tabanında bulunan verilerin nitelikleri incelenerek veri kümelerine uygun bir modelin ortaya konulmasıdır. Veri tabanında bulunan verilerin bir kısmı rassal olarak seçilerek eğitim verileri olarak kullanılır. Geri kalan veriler de test verileri olarak kullanılır. Belirlenen eğitim verileri üzerinden bir algoritma uygulanarak sınıflama modeli elde edilir.

İkinci adım söz konusu kuralların test verilerine uygulanarak sınanmasıdır. Test sonucunda elde edilen modelin doğruluğu kabul edilecek olursa, model diğer veriler üzerinde uygulanır.

2.6.2 Regresyon

Regresyon analizinde gözlenen olay değerlendirilirken hangi olaylardan etkilenildiğini belirlemek esastır. Herhangi bir değişkenin, bir veya daha fazla farklı değişkenler arasındaki ilişkinin matematiksel olarak ifade edilmesidir (Silahtaroğlu, 2013). Regresyonda amaç girdiler ile çıktılar arasındaki ilişkiyi kurmada en doğru tahmini yapabilecek modeli oluşturmaktır.

Regresyonu sınıflandırmadan ayıran özellik tahmin edilen bağımlı değişkenin süreklilik gösteren sayısal değişken olmasıdır.

Regresyon analizinde sonuç "bağımlı değişken", girdiler ise "bağımsız değişken" olarak adlandırılmaktadır. Regresyon analizi yapılırken modelde yer alan değişkenler bir bağımlı değişken ve bir veya birden çok bağımsız değişkenden oluşmaktadır. Değişkenler sayılabilir veya ölçülebilir nitelikte olmaktadır (Argüden ve Erşahin, 2008). Tek değişkenli modeller basit doğrusal regresyon, birden fazla bağımsız değişkenli modeller çoklu regresyon modeli olarak ifade edilmektedir. Aynı zamanda oluşturulan denklemin türüne göre de regresyon analizi doğrusal regresyon ve doğrusal olmayan regresyon olmak üzere ikiye ayrılmaktadır (Silahtaroğlu, 2013).

En küçük kareler yöntemi ile elde edilen doğrusal bir regresyon denklemi aşağıdaki şekilde ifade edilmektedir:

$$y = a + bx + e \tag{2.5}$$

Bu denklemi veri madenciliği açısından incelersek bağımsız değişken olan x nitelikleri, y ise sınıfları temsil etmektedir. Böylece x nitelik değeri yerine konularak y değeri hesaplanmakta, yani x'in hangi sınıfa ait olacağı tahmin edilmektedir. Eğer problemde birden fazla nitelik yani x değeri mevcut ise çoklu regresyon olarak adlandırılan denklem aşağıdaki gibi gösterilmektedir:

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_ix_i + e \tag{2.6}$$

Gerçek hayattaki problemlerin neredeyse hepsinde doğru tahmine ulaşabilmek için birden fazla girdiden faydalanmak gerekmektedir. Önemli olan konu girdilerin tahminin doğruluğuna yaptıkları katkıdır. Kimi zaman sonuca katkısı kısıtlı olan girdileri modelden çıkarmak, modelin etkinliğini artırmak adına gerekli hale gelmektedir (Argüden ve Erşahin, 2008).

Regresyon analizi birçok alanda kullanılmaktadır. Finansal tahminler, müşterinin yarattığı değer, fiyat değerlendirmeleri, ilaç reaksiyonları gibi birçok farklı problemin çözümünde regresyon analizinden yararlanılabilmektedir (Argüden ve Erşahin, 2008).

2.6.3 Kümeleme

Kümeleme (*clustering*) veri tabanında bulunan heterojen verilerden kendi içlerinde homojen veriler içeren gruplar oluşturma tekniğidir. Benzer özellikler taşıyan verileri ortak bir küme altında birleştirmektedir. Kümeleme modellerinde amaç veri tabanındaki kayıtların birbirinden çok farklı özelliklere sahip kümelere, onlarla benzer özelliklere sahip verilerle ortak kümede yer alacak şekilde bölünmeleridir (Akpınar, 2000).

Kümeleme modellerinin temel amaçları arasında; geniş veri yığınları için tanımlayıcı veriler belirleyerek işlenecek veri hacmini daraltmak, veri yığınlarındaki doğal kümeleri ortaya çıkarak sahip oldukları ortak özelliklerden dolayı aynı kümede olması gereken verileri belirlemek ve belirlenmiş kümelerin dışında kalan istisna durumları tanımlamak sayılabilir (Argüden ve Erşahin, 2008).

Sınıflama modelleri ile arasındaki en büyük fark kümeleme modellerinde sınıfların önceden tanımlanmamış olmasıdır. Kümeleme modelinde önceden belirlenmiş bir sınıf yoktur. Veri tabanında bulunan veriler, modeli kuran uzman tarafından belirlenmiş belli özelliklere göre kendi içlerinde gruplanırlar. Böylece kendi içlerinde homojen olan fakat birbirlerine göre farklı özellikler taşıyan heterojen kümeler oluşur.

Kümeleme kimi zaman başka veri madenciliği yöntemlerinin uygulanması öncesinde veri tabanındaki kayıtları daraltmak veya özelleştirmek adına kullanılabilir. Planlanan veri madenciliği yöntemi kümelerden yalnızca birine uygulanabilir veya oluşan farklı kümelere farklı yöntemler uygulanması gerekebilir.

Kümeleme modelleri özellikle müşteri özellikleri belirlenmesinde uygulanmaktadır, bu yüzden perakende sektöründe yaygın olarak kullanılmaktadır. Kümeleme modeli sayesinde müşteriler karlılıklarına, kullanım oranlarında veya pazar potansiyellerine göre gruplanabilirler. Müşterilerin bu şekilde kümelenmesi benzer özellikler barındıran dolayısıyla ortak tepkiler vermesi beklenen müşterilerle doğru iletişim yapılabilmesini mümkün kılmaktadır.

2.6.4 İlişkilendirme Kuralları

Birliktelik kuralları da denilen ilişkilendirme kuralları (*association rules*) veri tabanlarında bulunan veriler arasındaki bağlantıyı bulmak için kullanılmaktadır. İlişkilendirme kuralları sayesinde veriler arasındaki bağlantılar belirlenmekte ve olayların birlikte, eş zamanlı olma olasılıkları hesaplanmaktadır.

İlişkilendirme kuralı, veri tabanı içerisinde bulunan belirli nesnelerin eş zamanlı olarak, birlikte görülme olasılığının ifadesidir (Hand ve diğ, 2001).

İlişkilendirme kuralları veriler arasında fark edilmesi kolay olmayan bağlantıları kurmaktadır. Bu bağlantılar firmalar için birçok fayda sağlamakta, müşteri memnuniyetinden stratejik karar almaya kadar farklı aşamalarda kullanılmaktadır.

Çocuk bezi alan birinin bebek maması da alacağı kolay tahmin edilebilir fakat ilişkilendirme kuralları bebek maması ve yumuşatıcı arasındaki bağlantıyı kurmayı sağlamaktadır (Argüden ve Erşahin, 2008).

Yöntem veriler arasındaki ilişkileri ortaya çıkarmayı amaçladığından özellikle perakende sektöründe pazar sepeti analizi olarak sıklıkla kullanılmaktadır. Bu sayede perakende sektöründe ürünlerin raflara yerleşiminden, avantajlı ürün paketleri oluşturulmasına, katalogların hazırlanmasından, kullanılacak alanın belirlenmesine kadar birçok konuda ilişkilendirme kuralları kullanılmaktadır.

Sadece perakende sektöründe değil, otomotivde ürün özelliklerinin belirlenmesinde, depolamalarda yakın olması gereken ürünlerin belirlenmesinde veya tıp alanında da birçok uygulaması bulunmaktadır.

Hu (2010), "İlişkilendirme Kurallarına Dayalı Tıbbi Veri Madenciliği" başlıklı araştırmasında ilişkilendirme kuralları uygulanan veri madenciliği sonucunda büyük ve kötü huylu tümör bulunan hastalarda nüks ihtimalinin yüksek, daha küçük tümör bulunan ve radyoterapi gören hastalarda ise nüks ihtimalinin daha düşük olduğu gözlenmiştir.

Çok sayıda veri içerisinden kolaylıkla fark edilemeyen anlamlı ilişkiler oluşturulmasının basit bir süreç olmaması ve kimi zaman bazı ilişkilerin şans eseri ortaya çıkması ilişkilendirme kurallarında dikkat edilmesi ve değerlendirilmesi gereken noktalardır (Tan ve diğ, 2006).

2.6.5 Sıralı Diziler

Ardışık zaman örüntüleri olarak da adlandırılabilen sıralı diziler (*sequence patterns*) zaman içerisinde gerçekleşen olayları incelemektedir. Eldeki kayıtlar incelenerek trend ve döngüler belirlenir. Belli zaman aralıkları içerisinde belli frekanslarda olan olaylar incelenerek ilişkiler kurulur. Sıralı dizilerin zaman içerisinde gerçekleşen olaylara dayanması bu algoritmayı ilişkilendirme kurallarından ayırmaktadır.

Wojciechowski (2003) tanımına göre " Ardışık örüntüler, meydana gelen bir dizi olay içinde en sıklıkla rastlanan alt dizilerdir.".

Örneğin telekomünikasyon sektöründe eldeki kayıtlar incelendiğinde akıllı telefon alan kişilerin mevcutta yoksa internet paketi aldıkları veya mevcut olan paketlerinin kapasitelerini yükselttikleri gözlemlenmiştir. Kurulan bu ilişki ile akıllı telefon aldığı tespit edilen kişilere daha uygun fiyattan fakat taahhüt süresi ile bağlı tutmak amaçlı teklifler sunulması mümkün olabilir.

Perakende gibi rekabet yüksek bir sektörde de müşteri davranışlarını tahmin edebilme büyük bir avantaj sağlamaktadır. Barkod teknolojisinin gelişmesi ile firmaların elinde oldukça büyük veriler toplanmaktadır. Firmalar sadakat kartları gibi ürünler kullanarak kişi özelinde elde edebildikleri bu veriler üzerinden sıralı dizileri kullanarak müşterinin zaman içerisindeki hareketlerini inceleyebilir, ilişkilendirebilir ve kişiye özel kampanyalar sunarak kendileri için bir avantaja dönüştürebilirler. Yine aynı şekilde bu kayıtlar sayesinde kaybedilen müşterinin davranışları belirlenerek, mevcut müşterileri elde tutmak için farklı yollara başvurmak adına sıralı diziler kullanılabilmektedir.

3. TIBBİ KARAR DESTEK SİSTEMLERİ VE TIPTA VERİ MADENCİLİĞİ UYGULAMALARI

3.1 Tıbbı Karar Destek Sistemleri

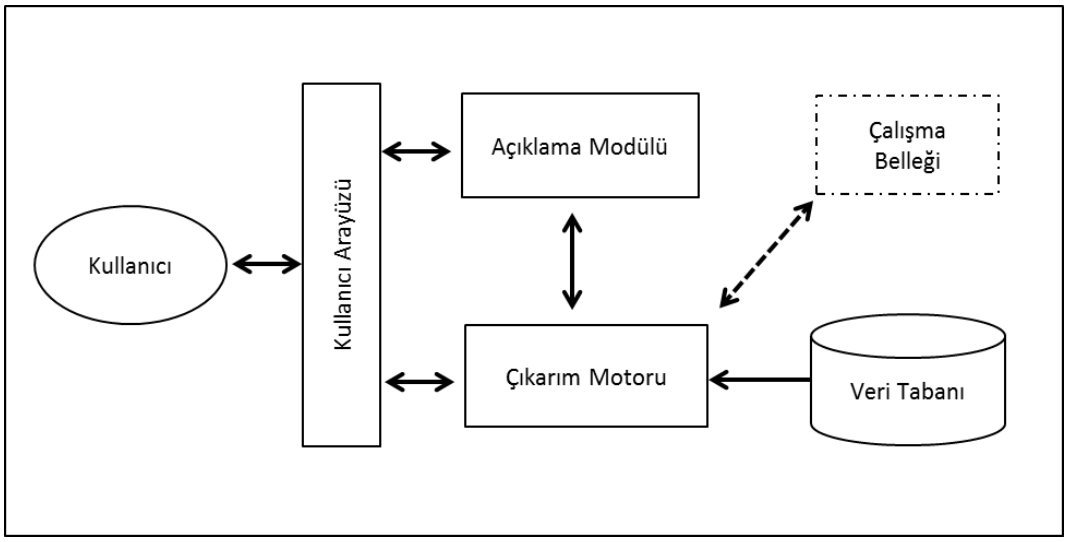
Tıbbi tanı koyma; hastaya ait gerçekçi veriler, doğru tıp literatürü bilgisi ve klinik tecrübeler gerektiren önemli ve karmaşık bir süreçtir. Tıbbi teşhis süreci, içinde birçok beklenmeyen durum bulundurduğundan dolayı başka sektörlerdeki tanımlama süreçlerine göre çok daha karmaşık bir yapıdadır. Klinik kararlar çoğu zaman doktorların algı ve deneyim temeline dayanılarak alınır (Parthiban ve Subramanian, 2008). Ancak her doktor farklı uzmanlık dallarının detaylarında yeterince detaylı bilgiye sahip olmayabilir veya hastalar her zaman şikayetlerini ve hissettiklerini doğru olarak ifade edemeyebilirler. Aynı zamanda günümüzde bilginin çok daha hızlı eskimesi ve veri miktarının gitgide artması hekimlere karar verme aşamasında güçlük oluşturmaktadır.

Tıbbı karar destek sistemleri (TKDS), tıbbi bilgileri ve hastaya ait verileri analiz ederek hekimlere teşhis ve karar verme aşamasında yardımcı olan bilgisayar tabanlı sistemlerdir. Sağlık ve sağlık hizmetlerini geliştirmek için kullanılan tıbbi karar destek sistemleri hekimlere; tıbbi ya da hasta özelindeki bilgileri, istenilen zamanda akıllıca filtrelenmiş olarak sunmaktadır (Anooj, 2012). Tıbbi karar destek sistemleri idari işlem ve sevklerde, klinik karmaşıklık ve ayrıntıların yönetilmesinde, maliyet kontrolünde, klinik tanıyı desteklemekte, iyi uygulamaların teşvik edilmesinde ve tıbbi uygulamaların verimliliğinin artırılmasında kullanılmaktadır (Uzoka, 2011). Bu sistemler tıp dünyasında diş hekimliği, eczacılık gibi farklı disiplinlerde de kullanılmaktadır (Mendonça, 2004).

Bu tür sistemlerin oluşturulmasına 1970'li yıllarda başlanmış olup, 1980'lerde yapay zeka yöntemlerinin geliştirilmesi ile bu sistemler yaygınlaşmaya başlamıştır. TKDS'ler geliştirilirken yapay sinir ağları, Bayesian yaklaşımı, bulanık mantık ve kural tabanlı yaklaşımlar tercih edilmiştir (Koç ve diğ., 2012).

Doğru teşhis sağlanması, önlenebilir hastalıklar için zamanında tarama yapılması veya hatalı ilaç etkileşimlerinin önlenmesi gibi klinik gereksinimlere hitap etmesi tıbbi karar destek sistemlerinin en önemli faydalarındandır (Garg ve diğ, 2005). TKDS'ler sayesinde sağlık bakım hizmetleri kalitesini ve karar verme sürecinin etkinliğini arttırmak ve hastaya uygulanabilecek olası yanlış müdahaleleri minimize etmek mümkündür. Böylece erken teşhisin sağlanması ile hekimlerin tedaviye harcadıkları süre azaltılmakta, gereksiz tahlillerin ve ilaç kullanımlarının önüne geçilmektedir. Bunların sonucunda düşük maliyetlerle sağlık bakım giderlerinin azaltılması sağlanmaktadır. Ayrıca TKDS'ler ile hastanın kendini yanlış ifade etmesine veya hekimin bilgisinin yetersiz olmasına bağlı yaşanan hatalı teşhis oranlarının düşmesi mümkün olmaktadır.

Genellikle tıbbi karar destek sistemleri 4 temel içerikten oluşmaktadır: çıkarım motoru, veri tabanı, açıklama modülü ve çalışma belleği (Şekil 3.1).



Şekil 3. 1: Tıbbi karar destek sistemi bileşenleri (Mendonça, 2004).

Çıkarım motoru sistemin ana parçasıdır. Çıkarım motoru sistemdeki tıbbı bilgileri ve hasta hakkındaki verileri kullanarak bir sonuca varır. Çıkarım motoru ne gibi aksiyonların alınması gerektiğini kontrol etmektedir. Teşhis sistemlerinde sonuca varmak veya uyarı sistemlerinde uyarı ve hatırlatmaların rotasına karar vermek gibi aksiyonlar çıkarım motorları sayesinde kontrol edilebilir.

Çıkarım motorunda kullanılan bilgiler veri tabanından elde edilmektedir. Veri tabanları uzmanlar yardımı ile veya otomatik bir süreç ile oluşturulabilir. Uzman

yardımı ile ilerlenen süreçte, uzmanın veri tabanındaki bilgiyi oluşturması, düzenlemesi ve sürekli takip etmesi gerekmektedir. Otomatik süreçte ise bilgiler kitap ve makale gibi dış kaynaklardan bilgisayar uygulamaları ile elde edilmektedir. Oldukça karmaşık bir işlem olan veri tabanı oluşturma evresi otomatik sistemler sayesinde biraz daha kolaylaşmıştır.

Hastalar ile ilgili veriler çalışma belleğinde depolanmaktadır. Bu bilgiler cinsiyet, yaş gibi demografik veya hastanın bulunan alerjileri, geçirmiş olduğu hastalıklar ve ameliyatlar gibi tıbbı bilgiler olabilir.

Açıklama modülü tüm karar destek sistemlerinde bulunmamaktadır. Bu modül veri tabanındaki bilgiyi, çalışma belleğindeki hasta bilgisi ile birlikte kullanarak sonuca varmak için nedenler oluşturmakla sorumludur.

Karar destek sistemleri hekim reçete yazarken ilaç etkileşimlerine veya hastanın mevcut alerjilerine karşı hekimi uyaran bir sistem olarak senkronize veya düzenli kontrolleri hatırlatmak adına senkronize olmayan rutin bir şekilde çalışabilir (Mendonça, 2004).

Tıbbi karar destek sistemlerin bir diğer önemli çeşidi de durum görüntüleme, konsültasyon sistemleri ve tıbbi yönetmeliklerdir (Mendonça, 2004).

Durum görüntüleme bir kurumda elektronik formatta mevcut olan tüm verileri kopyalayıp, uygun gördüğü durumlarda hekimler için uyarılar ve hatırlatmalar göndermek üzere bu veri tabanını kullanan bir yazılım uygulamasıdır.

Konsültasyon sisteminde hekim hastanın tıbbi geçmişi, demografik bilgileri ve laboratuvar sonuçları gibi vaka ayrıntılarını sisteme girer ve sistem vakadaki problemlerin açıklamalarını yaparak alınabilecek aksiyonları önerir.

Klinik uzman grupları tarafından geliştirilip hükümet veya meslek kuruluşları tarafından yayınlanan ve en iyi uygulama örneklerini gösteren tıbbi yönetmelikler de tıbbi karar destek sistemlerinin içine katılabilir.

Bu sistemler hasta özelinde bilgi sağlamanın yanı sıra kalite güvencesi ve tıp eğitimi alanlarında da kullanılabilir (Mendonça, 2004).

TKDS'lerin karar verme, hastalık yönetimi, teşhis ve tedavi üzerine etkileriyle ilgili birçok olumlu etkisi tespit edilmiştir (Raymond ve Dold, 2002). Hem hekimlere kolaylık sunulmakta hem de hasta memnuniyetsizliklerinin önüne geçilmektedir.

Aynı şekilde TKDS'lerin, gerekli olan tıbbi bilginin kolayca elde edilmesinde etkili olduğu birkaç sağlık kurumu ve uygulama alanlarında yapılan çalışmalarda gözlemlenmiştir (Merijohn ve diğ, 2008).

TKDS alanında günümüze kadar pek çok çalışma yapılmış olup, çalışmaların başarı yüzdesi gittikçe artmakta böylece yapılan çalışmalar yükselen bir ivme kazanmaktadır (Yurtay ve diğ, 2013). Buna rağmen TKDS'lerin yeterince yaygınlaşamamasının altında yatan sebep hekimlerin bu sistemleri kullanmakta göstermiş oldukları dirençtir. Oysaki kabul edilmelidir ki TKDS'ler yaptıkları analizler ile sadece mevcut verilerden anlamlı bilgiler elde ederek potansiyel önerileri sağlarlar, bu önerileri kendi bilgi süzgecinden geçirip nasıl bir aksiyon alacaklarına karar verecek olan hekimlerdir.

3.2 Tıp Alanında Veri Madenciliği Uygulamaları

Veri madenciliği, birçok farklı analiz aracı yardımıyla veri içerisindeki örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir (Two Crows Coorporation, 2005).

Önceden kullanılan kağıt üzerinde veri toplanan klasik hastane bilgi sistemlerinin yerini elektronik ortamda düzenlenen veri sistemleri almıştır. Özellikle günümüzde neredeyse tüm tıbbi cihazların dijital hale gelmesi bu durumu kolaylaştırmıştır. Biyoloji ve tıp teknolojilerindeki muazzam gelişmeler sayesinde büyük miktarlarda biyomedikal veriye ulaşılabilmektedir (Giannapoulou, 2008). Tıpta veri biriktirilmesine neden olan uygulamalara görüntüleme, teşhis koyma, tedavi, prognoz (hastalığın sonucunun tahmini), biyomedikal ve biyolojik analizler, epidemiyolojik çalışmalar örneklerini verebiliriz (Abdel-Aal, 2005). Klasik bilgi sisteminde biriken verilerin bireysel olarak incelenip, çalışılması artan veri sayıları ile neredeyse imkansız hale gelmişken, elektronik ortamdaki veriler bu durumu kısmen kolaylaştırmaktadır.

Veri madenciliği yöntemleri klasik yöntemlerle bulunması çok zor belki de imkansız olan anlamlı ilişkilerin ve bilgilerin eldesini mümkün kılmaktadır. Böylece veri madenciliği yöntemleri elde edilen çok sayıdaki tıbbi verileri anlamlandırmak, verinin doğası gereği kesin ve doğrusal olmayan verileri düzenlemek işlenir hale getirmek, teşhis sırasında karşılaşılan çok sayıda hastalık belirtisi ile hasta ve yakınlarındaki farkındalığa cevap vermek gibi ihtiyaçlara cevap vermiştir. Aynı

zamanda farkına varılması daha güç detayların tespit edilip hastalıkların teşhisinin yapılması ve erken müdahaleler veya daha uygun olduğu görülen tedavi değişiklikleri veri madenciliği yaklaşımı ile daha olası hale gelmiştir. Bu sayede veri madenciliği yöntemleri tıbbi araştırmaların vazgeçilmez araçlarından biri haline gelmiştir (Kaya ve diğ, 2003).

Basit veri tabanları hasta bilgileri (ad, soyad, adres, kan grubu vb.) gibi temel verileri saklarken, günümüzde hastaların hasta yerleşiminden finansal verilerine, ilaçlarından ziyaret planına kadar pek çok veri de bu sistemlerce kayıt altına alınmaktadır (Cios ve Moore, 2002). Bu kapsamda tıp alanında da oldukça fazla veri bulunmaktadır. Veri madenciliği teknikleri ile hayati önem taşıyan bu verilerden yararlı bilgiler edinmek mümkündür.

Tıp alanında bulunan hayati öneme sahip veriler hastane bilgi sistemleri sayesinde düzenli olarak tutulmaktadır. Bu hayati öneme sahip verilerin işlenmesi ve aynı derece önemli bilgiler elde edilmesi veri madenciliği çalışmaları ile mümkün olabilmektedir. Bu durum hem uzmanlar hem de hastalar açısından fayda sağlamaktadır.

Tıbbi veriler üzerinde çalışma yapmak ciddi bir emek ve uzmanlık gerektirmektedir. Verilerin yorumlanması önemli ve uzmanlık gerektiren bir iştir. Terimler karışık ve uzmanlık sahibi olmayan biri tarafında kolayca yanlış anlaşılmaya açıktır. Diğer bir yandan tıp alanında bulunan terimlerde standardın olmaması veya farklılık göstermesi veri düzenleme işini daha da zorlaştırmaktadır. Bu nedenle işlemler yapılırken uzman görüşleri dikkate alınmalıdır.

Ülkemizde Sağlık Bakanlığı, hayati öneme sahip verilerin toplanmasında, saklanmasında ve analiz edilmesinde ulusal veya uluslararası standartların olmadığı, sağlık alanında veri toplama konusunda ciddi bir karmaşanın mevcut olduğu tespitinde bulunmuş ve bu konuda çalışmalar başlatmıştır (Koyuncugil ve Özgülbaş, 2009).

Literatür taramasında incelenen tıp alanında uygulanan veri madenciliği çalışmalarından bir kısmı Çizelge 3.1'de özetlenmiştir:

Çizelge 3. 1: Literatür araştırması.

No	Referans	Uygulama Alanı	Karar Ağaçları	Yapay Sinir Ağları	Destek Vektör Makineleri	K-En Yakın Komşu Alg.	Naive-Bayes	Diğer
1	Saichanma ve diğ. (2014)	RBC morfolojisi incelenmesi	√					
2	Niemann ve diğ. (2014)	Hepatik steatoz sınıflandırması	√					√
3	Khan, Westin ve Dougherty (2014)	Parkinson hastalığında konuşma anlaşılırlığı sınıflandırması			√			
4	Zhou ve diğ. (2014)	Karaciğer hastalıklarının erken teşhisi			√			√
5	Yurtay ve diğ. (2013)	Anemisi tanısı						√
6	Omiotek, Burda ve Wojcik (2013)	Haşimato hastalığı tanısı	√	√				
7	Shukla ve Barot (2013)	Meme kanseri tanısı						√
8	Parshutin ve Kirshners (2013)	Atrofik gastrit görüntülemesi	√				√	
9	Uçar, Karahoca A. ve Karahoca D. (2013)	Tüberküloz hastalığının teşhisi		√				√
10	Köktürk (2012)	Sınıflandırma başarı karşılaştırması	√	√			√	
11	Yılmaz ve Bozkurt (2012)	Demir eksikliğine bağlı anemi tanısı		√				
12	Elshami ve Alhalees (2012)	Talasemi teşhisi	√	√			√	
13	Azarkhish ve diğ. (2012)	Demir eksikliğine bağlı anemi teşhisi		√				√
14	İşler ve Narin (2012)	Konjektif kalp yetmezliği teşhisi						√
15	Ahmed ve Hannan (2012)	Kalp hastalığı riski tahmini	√	√	√			√
16	Abdullah ve Rajalaxmi (2012)	Koroner kalp hastalıklarının tahmini	√					√
17	Soni ve diğ. (2011)	Kalp hastalıkları tahmini	√	√		√	√	
18	Yeh ve diğ. (2011)	Beyin ve damar hastalıkları tahmini	√	√			√	

Çizelge 3. 1 (devamı) : Literatür araştırması.

No	Referans	Uygulama Alanı	Karar Ağaçları	Yapay Sinir Ağları	Destek Vektör Makineleri	K-En Yakın Komşu Alg.	Naive-Bayes	Diğer
19	Kumar ve diğ. (2011)	Tıbbi tanı	√					
20	Patil ve diğ. (2011)	Yanık hastalıklarının hayatta kalma durumu	√	√	√		√	
21	Sanap ve diğ. (2011)	Anemi sınıflandırması	√		√			
22	Coşkun ve Baykal (2011)	Göğüs kanseri hastalarının hayatta kalma tahmini	√				√	√
23	Chen ve diğ. (2011)	Erken doğum risk faktörleri tahmini	√	√				
24	Oztekin, Kong ve Delen (2011)	Karaciğer transplantasyonu analizi	√	√			√	
25	Jovic ve Bogunovic (2011)	Elektrokardiyogram analizi	√		√		√	√
26	Kiranyaz ve diğ. (2011)	Elektrokardiyogram sınıflandırması						√
27	Uçar ve Karahoca (2011)	Mikobakterium tüberküloz tespiti		√				√
28	Lee ve diğ. (2011)	Hastaların düşmelerindeki kritik faktörlerin belirlenmesi		√				
29	Danacı ve diğ. (2010)	Meme kanseri tahmin ve teşhisi	√					
30	Karaolis ve diğ. (2010)	Koroner kalp hastalıkları risk faktörleri	√					
31	Srinivas ve diğ. (2010)	Koroner kalp hastalıkları analizi	√	√			√	
32	Das (2010)	Parkinson hastalığı tanısı	√	√				√
33	Patil, Joshi ve Toshniwal (2010)	Diyabet hastalığı tahmini	√					√
34	Sarvestani ve diğ. (2010)	Meme kanseri hastalarının yaşam süresi tahmini		√				√
35	Srinivas, Rani ve Govrdhan (2010)	Kalp hastalığı ile ilgili faktörlerin belirlenmesi	√	√			√	
36	Chang ve Chen (2009)	Dermotolojik tanı	√	√				

Çizelge 3. 1 (devamı) : Literatür araştırması.

No	Referans	Uygulama Alanı	Karar Ağaçları	Yapay Sinir Ağları	Destek Vektör Makineleri	K-En Yakın Komşu Alg.	Naive - Bayes	Diğer
37	Taşdelen ve diğerleri (2009)	Yaşlılarda baş ağrısı tahmini		√				
38	Karabatak ve İnce (2009)	Meme kanseri teşhisi		√				√
39	Gregori ve diğ (2009)	Diabet süreci izlenmesi		√				
40	Oztekin, Delen ve Kong (2009)	Kalp ve karaciğer transplantasyonu sonrası yaşam süresi tahmini	√	√				√
41	Doğan ve Türkoğlu (2008)	Anemi teşhisi	√					
42	Jonsdottir ve diğ. (2008)	Meme kanseri tahmin ve teşhisi	√				√	
43	Kahramanli ve Allahverdi (2008)	Diyabet ve kalp hastalıklarının sınıflandırılması		√				
44	Bellazzi ve Zupan (2008)	Tıbbi veri madenciliği modeli	√	√	√	√	√	√
45	İşler ve diğ. (2008)	Konjektif kalp yetmezliği teşhisi için değişken analizi						√
46	Gören ve diğ. (2008)	Siklosporin A kan düzeyinin tahmini		√				√
47	İşler ve Kuntalp (2007)	Konjektif kalp yetmezliği teşhisi				√		√
48	Wongseree ve diğ. (2007)	Talasemi sınıflandırması		√				√
49	Karahoca, Kucur ve Aydin (2007)	Emboli tespiti		√			√	√
50	Delen,Sharda ve Bessonov (2006)	Trafik kazalarındaki yaralanma ciddiyeti tahmini		√				
51	Bellaachia ve Guven (2006)	Meme kanseri hastalarının yaşam süresi tahmini	√	√			√	
52	Abdel-Aal (2005)	Tıbbi veri sınıflandırması		√				
53	Delen, Walker, Kadam (2005)	Meme kanseri hastalarının yaşam süresi tahmini	√	√				√
54	Li ve diğ. (2004)	Kanser teşhisi			√			√
55	Bircan (2004)	Doğum ağırlığını etkileyen risk faktörlerinin belirlenmesi						√

Tabloda özet olarak gösterilen çalışmaların bir kısmı aşağıda detaylı olarak açıklanmaktadır:

"Karar Ağacı Yöntemini Kullanarak Biyokimya Verileri ile Anemi Teşhisi" (Doğan ve Türkoğlu, 2008) isimli çalışmada, kan biyokimya parametreleri ile demir eksikliği anemisi teşhisinde, hekime yardımcı olacak ve kolaylık sağlayabilecek bir karar destek sistemi oluşturulmuştur. Sistemin işleyişi veri madenciliği tekniklerinden olan karar ağaçları yapısı ile sağlanmaktadır. Sisteme giriş olarak, biyokimya parametrelerinden demir eksikliği anemisi hastalığı için temel belirleyiciler olan Serum demiri, Serum demir bağlama kapasitesi (SDBK) ve Ferritin enzimleri kullanılarak, çıkış olarak da Anemi(+) ve Anemi(-) değerlendirmelerinde bulunulmuştur. 96 hasta verisi değerlendirildiği karar destek sisteminin sonuçları, doktorun verdiği kararlarla tamamen örtüşmüştür.

"Dermatolojik Tanı Kalitesini Artırmak için Karar Ağacı ve Sinir Ağları Yöntemlerinin Uygulanması" (Chang ve Chen, 2009) başlıklı makalede, başlıca altı cilt hastalığına odaklanan beş deney yapılmıştır. Çalışmada dermatolojide en iyi tahmin modelini oluşturmak için sinir ağları sınıflandırma yöntemleri ile birleştirerek karar ağacı metodu uygulanmıştır. Sonuçlar en yüksek doğruluk oranının %92.62 ile sinir ağları metodu ile tahminlemede bulunduğunu göstermiştir. Karar ağacı modeline duyarlılık analizi uygulandığı zaman ise tam tersi olarak %80.33 ile en düşük doğruluk oranı elde edilmiştir. Bu sonuçlar, veri madenciliği sayesinde kullanılan sınıflandırma teknolojisi ile hizmet ve sağlık kalitesini artırmak için tanı kadar önemli ve yararlı referanslar sağlanabileceğini doktorlara göstermiştir.

"Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi" (Danacı ve diğ, 2010) isimli çalışmada Irvine California Üniversitesi'nin havuzundan alınan 569 hastaya ait meme altı doku örnekleri tanı ve teşhis amacı ile C4.5 karar ağacı algoritması kullanılarak hastalık tahmini yapılmıştır. Veri madenciliği yöntemlerinden olan C4.5 karar ağacının hastalık tanı ve teşhisinde %97,4359 yüksek doğruluk oranı sonucunu vermesi önemli ve günümüzde oldukça sık rastlanan meme kanseri hastalığının erken tanısında önemli rol oynamıştır.

"Koroner Kalp Hastalıkları Risk Faktörleri'nin Veri Madenciliği Yöntemlerinden Karar Ağaçları ile Değerlendirmesi" (Karaolis ve diğ, 2010) adlı makalede koroner kalp hastalıklarının öneminden bahsedilmiş, teşhis ve tedavisinde ilerlenmiş olsa da

daha fazla araştırma gerektirdiği vurgulanmıştır. Bu kapsamda koroner kalp hastalıklarının önlenmesi için risk faktörlerinin belirlenmesini sağlayacak bir veri madenciliği modeli geliştirilmesi hedeflenmiştir. 3 farklı vaka ve bu vakalar öncesi ve sonrasında gerçekleşen risk faktörleri belirlenmiştir. Bahsedilen üç vaka, belirlenen beş farklı kriter ile C4.5 karar ağacı algoritması kullanılarak veri madenciliği analizi gerçekleştirilmiştir. Sınıflandırma kurallarından miyokardiyal enfarktüs için, yaş, sigara, hipertansiyon; perkütan koroner girişim için aile öyküsü, hipertansiyon ve diyabet tarihin geçmişi; koroner arter baypas greft cerrahisi için yaş, hipertansiyon öyküsü, sigara en önemli risk faktörleri olarak analiz edilmiştir. Doğru sınıflandırma oranları ise sırasıyla 66%, 75%, ve 75% olarak gerçekleşmiştir.

"Kömür Madeni Bölgelerinde Veri Madenciliği Teknikleri Kullanılarak Koroner Kalp Hastalığı Analizi ve Kalp Krizi Tahmini" (Srinivas ve diğ, 2010) adlı makalede kalp hastalıklarının modern toplumlardaki en önemli ölüm sebeplerinden biri oluşundan ve teşhisinin doğru yapılması gereken karmaşık bir yapıda olduğundan bahsedilmiştir. Çalışmada Hindistan'da bir bölge olan Andhra Pradesh'te Singareni isimli kömür madenciliği bölgesinde bildirilen kardiyovasküler hastalık oranının, diğer risk faktörleri de göz önüne alındıktan sonra diğer bölgelerle kıyaslanması sonucu gerçekten yüksek olup olmadığını incelemişlerdir. Kardiyovasküler hastalıklara işaret eden bağımlı değişkenler olarak göğüs ağrısı, inme ve kalp krizini belirlemişler ve hastalığın varlığının tahmini ile ilgili ise 15 farklı özellik belirtmişlerdir. Düzenli kriterler dışında daha genel kriterler olan Beden Kütle İndeksi, hekim görüşü, yaş, etnik köken, eğitim seviyesi, gelir düzeyi ve buna benzer diğer kriterler de tahmin yürütmede kullanılmıştır. İsabetli tahmin yürütme için veri madenciliği tekniklerinden karar ağaçları, Naive-Bayes ve yapay sinir ağları algoritmaları kullanılmış ve en yüksek hastalık tahmin doğruluk oranı %89 ile karar ağaçları yönteminden elde edilmiştir.

"Tıbbi Tanı için Öngörücü Veri Madenciliği ile Kalp Hastalıkları Tahmini'ne Bakış" (Soni ve diğ, 2011) başlıklı araştırmada günümüzde tıbbi araştırmalarda kullanılmakta olan veri madenciliği tekniklerini ile veri tabanlarından bilgi keşfine yardımcı olacak, özellikle kalp hastalığı tahmini üzerine bir anket sağlanması çalışılmıştır. Birçok deney uygulanarak, kimi veri setleri ve çıktılar üzerinde veri madenciliği tekniklerinin performansları kıyaslanmıştır. Bu sonuçlar göstermiştir ki bazen Bayesyen sınıflandırmaları karar ağacı tekniği ile aynı doğrulukları gösterse de

k-en yakın komşu, yapay sinir ağları ve sınıflandırma bazlı kümeleme yöntemleri gibi yöntemler iyi performans gösterememiştir. İkinci olarak ise kalp hastalıkları tahmininde optimal özellikleri göz önünde bulundurmak ve ham veri büyüklüğünü azaltmak adına genetik algoritmaların uygulanmasının Bayesyen sınıflandırma ve karar ağacı metotlarının performansı artırdığı gözlemlenmiştir.

"Veri Madenciliği Kullanılarak Beyin ve Damar Hastalıkları için Bir Tahmin Modeli" (Yeh ve diğ, 2011) adlı makalede, beyin ve damar hastalıklarının önlenmesi ve tedavi edilmesine yönelik uygulanan bir programdan alınan 493 veri örneğine karar ağacı, Bayesyen sınıflandırma ve geri beslemeli sinir ağları gibi 3 farklı metot uygulanmıştır. Uygulanan sınıflandırma metotlarının verimlilikleri doğruluk ve duyarlılık açısından karşılaştırılmış, karar ağacı metodu uygulanarak oluşturulan model beyin ve damar hastalıklarının tahmin edilmesine yönelik en uygun model olarak seçilmiştir. Bu modelde, duyarlılık ve doğruluk sırasıyla % 99.48 ve% 99.59 elde edilmiştir ve beyin ve damar hastalıkları tahmininde etkisi olan sekiz önemli faktör ve 16 tanı sınıflama kuralı çıkarılmıştır. Beş deneyimli beyin ve damar hastalıkları doktoru bu kuralları değerlendirmiş ve modelin yararlı olduğunu doğrulamıştır.

"Veri Madenciliği Kullanılarak Tıbbi Tanı İçin Karar Destek Sistemi" (Kumar ve diğ, 2011) isimli makalede sağlık alanında çok fazla veriye sahip olunduğundan fakat verilerde gizlenen ilişkilerin ve desenlerin keşfedilmesinin gerekliliğinden bahsedilmiştir. Araştırma diyabet, hepatit ve kalp hastalıklarından toplanan veriler ile doktorlara yardımcı olmak amaçlı akıllı bir tıbbi karar destek sistemi oluşturmaya odaklanmıştır. Makalede, C4.5, CART, ID3 karar ağaçları algoritmaları kullanılmış ve etkinlikleri karşılaştırılmıştır. Çalışmanın sonucunda algoritmaların başarı sırası CART, C4.5 ve ID3 şeklinde olmuştur.

"Yanık Hastalarının Hayatta Kalma Durumu Tahmininde Veri Madenciliğinin Rolüne Yönelik Yeni Bir Yaklaşım" (Patil ve diğ, 2011) adlı makale çalışmasında sağlık alanında birçok farklı veri madenciliği algoritmasının performansları farklı ölçütler alınarak karşılaştırılmıştır. 2002-2006 yıllarında arasında yanık şikayeti olan 180 hastanın verileri Hindistan'ın en büyük kırsal hastanelerinin birinden alınmıştır. Alınan veriler hastalarının yaşlarını, cinsiyetlerini ve vücudun 8 farklı bölgesinden alınan yanık yüzdesi oranlarını kapsamaktadır. Çalışmada Naive Bayes, karar ağaçları, destek vektör makineleri ve geri beslemeli sinir ağları sınıflandırma

teknikleri kullanmıştır. Uygulanan sınıflandırma tekniklerinin performans değerlendirmesi yapıldığında, Naive Bayes %97,78 doğruluk yüzdesi ile en iyi tahminleyici, karar ağaçları ve destek vektör makineleri %96,12 ile ikinci en iyi tahminleyici olmuş, geri beslemeli sinir ağları ise %95 doğruluk oranı göstermiştir.

"Veri Madenciliği Teknikleri Kullanarak Anemi Sınıflandırması Yapılması" (Sanap ve diğ, 2011) adlı çalışmada 191 sağlıklı, 323 anemi teşhisi konulmuş toplam 514 durum incelenerek karar ağaçları ve destek vektör makineleri uygulanmış, iki yöntem arasında karşılaştırma yapılmıştır. Tam kan sayımında bulunan 22 parametreden önemli olan 10 parametre belirlenerek 7 farklı anemi çeşidi için sınıflandırma yapılmıştır. Yapılan araştırmada karar ağaçları yöntemi %92,42 doğruluk oranı ile en başarılı yöntem olarak belirlenmiştir.

"K-En Yakın Komşuluk, Yapay Sinir Ağları ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması" (Köktürk, 2012) adlı doktora tez çalışmasında, veri madenciliğinin tıp alanındaki öneminden bahsedilmiştir ve çeşitli yöntemlerin sınıflandırma başarıları karşılaştırılmıştır. Bu doğrultuda Bülent Ecevit Üniversitesi Uygulama ve Araştırma Hastanesi Kadın Hastalıkları ve Doğum Polikliniği'ne başvuran erken ve zamanında doğum yapan gebelerden elde edilen veri setine k-en yakın komşu, yapay sinir ağları ve karar ağaçları yöntemleri uygulanmıştır. En yüksek doğruluk oranı % 90.8 ile yapay sinir ağı tekniğinden elde edilmiştir, bunu % 82.5 ile karar ağacı yöntemi, % 78.3 ile de k-en yakın komşu analizi takip etmiştir.

"Kadınlarda Demir Eksikliğine Bağlı Kansızlık Tanısına İlişkin Bir Veri Madenciliği Çalışması" (Yılmaz ve Bozkurt, 2012) adlı makalede aneminin en yaygın çeşidi olan demir eksikliği anemisinin, kadın hastalar tanısına ilişkin farklı yapay sinir ağları yöntemleri geliştirilmiştir. Zonguldak Devlet Hastanesi'nden alınan 2600 kadın hastaya ait kan tahlili sonuçları bilgisayara yüklenmiş ve bunların 2000'i öğrenme, 600'ü ise test için kullanılmıştır. Kan tahlili verilerinden 6 parametre belirlenmiş ve 6 farklı yapay sinir ağları çeşidi düzenlenen bu verilere uygulanmıştır. Çalışma sonucunda test için kullanılan 600 verinin 478 adeti sağlıklı, 122 adeti ise anemi olarak belirlenmiştir. %97,6 duyarlılık ve %99,16 doğruluk oranı elde edilmesi ise tıbbı verilerin sınıflandırmasında yapay sinir ağları yöntemlerinin başarılı olduğu gösterilmiştir.

"Veri Madenciliği Sınıflandırma Yöntemlerine Dayalı Talasemi Teşhisi" (Elshami ve Alhalees, 2012) adlı çalışmada 46920 örnek veri değerlendirilerek, tam kan sayımı sonuçlarından 11 farklı parametre belirlenmesi ile talasemi çeşitleri üzerinde sınıflandırma methotları uygulanmıştır. Veri madenciliği yöntemlerinden karar ağaçları, yapay sinir ağları ve Bayesyen sınıflandırma yöntemleri kullanılmış ve elde edilen sonuçların doğruluk oranları kıyaslanmıştır. Yapay sinir ağları yöntemi en başarılı yöntem olarak kabul edilmiştir.

Azarkhish ve arkadaşları tarafından yapılan bir çalışmada Tahran'da bir hastaneden alınan 203 hasta verisinden 149'u öğrenme, 54'ü ise test için kullanılmıştır. Çalışmada 4 laboratuvar verisine dayandırılarak demir eksikliği anemisini ve serum demiri seviyesini tahmin edebilmek için yapay sinir ağları ve bulanık mantığa dayanan bir yöntem geliştirilmiştir. Çalışma sonunda elde edilen sonuçlara göre yapay sinir ağları yönteminin, demir eksikliği anemisi tanısı koyulmasında ve serum demiri seviyesi tahmininde, bulanık mantığa dayalı yöntemden ve lojistik regresyon yönteminden daha üstün olduğu belirtilmiştir (Azarkhish ve diğ, 2012).

"Kansızlık Tanısına İlişkin Bir Veri Madenciliği Uygulaması" (Yurtay ve diğ,2013) adlı çalışmada 9312 kansızlık tanısı olmayan, 1457 ise kansızlık tanısına sahip 10769 durum verisi üzerinde veri madenciliği uygulanmıştır. Veriler içerisinden demir eksikliği anemisi tanısına yönelik olarak laboratuvar kan tahlili sonuçları seçilerek, 6 parametre kullanılmıştır. Verilere kümeleme algoritmaları uygulanmış ve sonuçlar hemogram testlerinden elde edilen sonuçlarla karşılaştırılarak, uzman doktor yardımı ile yorumlanmıştır.

Saichanma ve arkadaşları tarafından gerçekleştirilen bir çalışmada otomatik kan sayacından toplanan 13 hematolojik parametre içeren bir veri seti kullanılarak 1362 öğrenciden alınan periferik kan yaymasındaki anormallikleri tahmin edebilmek için karar ağaçlarına dayanan J48 algoritması uygulanmıştır. Çalışma sonucunda J48 algoritması ile elde edilen karar ağacının, 4 hematolojik parametre kullanarak alyuvar morfolojisinin tahmin edilmesinde pratik bir kılavuz olarak kullanılabileceği ortaya çıkarılmıştır (Saichanma ve diğ, 2014).

4. METODOLOJİ

Sınıflandırma; örüntü tanıma, pazarlama, kalite kontrol çalışmaları, hastalık tanıları ve dolandırıcılık tespiti gibi bir çok farklı alanda kullanılan, en bilinen veri madenciliği tekniklerinden biridir (Silahtaroğlu, 2013). Veri madenciliğinde sınıflandırma işlemi, elde edilen verilerin belirlenen özelliklere göre sınıflandırılması ve yeni verinin sınıfının tahmin edilmesidir. Sınıflandırma yöntemleri verilerin içinde gizli kalan kuralları ortaya çıkararak, veriyi en doğru sınıfa yerleştirmektedir.

Bu bölümde, çalışmada kullanılmış olan K-En Yakın Komşu algoritması, Bayesyen Sınıflandırma, Karar Ağaçları ve Yapay Sinir Ağları yöntemleri detaylı şekilde anlatılmaktadır.

4.1 K-En Yakın Komşu Algoritması

K-en yakın komşu (*K-Nearest Neighbor,k-NN*) algoritması sınıfları belirli olan mevcut verileri inceleyerek gelecek yeni kaydın hangi sınıfa ait olacağını tahmin etmek için kullanılmaktadır (Giannopoulou, 2008). *k-NN* yönteminde sınıfı belirlenmek istenen verinin, veri tabanındaki her bir değere olan uzaklıkları hesaplanmaktadır. Belirlenen *k* adet en yakın noktanın belirlenmesi ile yeni gelen verinin sınıfı şekillenmektedir. Algoritmada *k* değeri hesaplama öncesinde belirlenir, değerin uygun olmasına dikkat edilmelidir. *K* değerinin çok yüksek olması aslında birbirine benzemeyen noktaların bir arada toplanmasına sebep olabileceği gibi, değerin çok düşük olması da birbirine benzeyen noktaların ayrı sınıflarda olmasına neden olmaktadır (Han ve Kamber, 2006). *k-NN* algoritması özellikle coğrafi bilgi sistemlerinde en yakın şehir veya istasyon gibi birimlerin belirlenmesinde kullanılmaktadır (Beyer ve diğ, 1999). Ayrıca *k-NN* sınıflandırma işlemlerinin dışında kümeleme işlemlerinde de kullanılmaktadır (Silahtaroğlu, 2013).

k-NN'de en yaygın olarak kullanılan Öklid uzaklık ölçüsüdür. Öklid uzaklık ölçüsüne ek olarak kullanılan bir diğer uzaklık ölçüsü de Manhattan uzaklık ölçüsüdür.

Öklid uzaklık ölçüsü aşağıdaki formül ile hesaplanmaktadır:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \quad \textbf{(4.1)}$$

Manhattan uzaklık ölçüsü aşağıdaki formül ile hesaplanmaktadır:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}| \quad \textbf{(4.2)}$$

4.2 Bayesyen Sınıflandırma

Bayesyen sınıflandırma tekniği elde var olan sınıflanmış verileri kullanarak yeni bir verinin herhangi bir sınıfa ait olma olasılığını tahmin eden istatistiksel bir yöntemdir.

Bayes sınıflandırıcılar belirli bir değişkenler grubunun belirli bir sınıfa ait üyelik olasılıklarını tahmin eden istatistiksel sınıflandırıcılardır ve bu sınıflandırma Thomas Bayes'in teoremine dayanmaktadır (Han ve Kamber, 2006). Bu teorem; belirsizlik içeren bir durumun modelinin oluşturularak, bu durumla ilgili gerçekçi gözlemler ve evrensel doğrular ışığında belli sonuçlar elde edilmesine olanak sağlamaktadır (Argüden ve Erşahin, 2008). Algoritma önce kendisine verilen öğrenme kümesindeki değerlerin ve sınıfların bulunma sıklığını hesaplayarak yeni gelecek olan verinin belli bir sınıfta olma olasılığını tahmin etmektdir.

Bayesyen sınıflandırma belirsiz durumlarda tahmin ve sınıflandırma yapmak için kullanılmaktadır. Belirsizlik içeren durumlarda karar verme konusunda oldukça kullanışlı bir tekniktir. Bayesyen sınıflandırmanın büyük veri tabanlarına uygulandığında kısa zamanda yüksek doğruluk oranı elde ettiği gözlemlenmiştir (Han ve Kamber, 2006). En büyük dezavantajı değişkenlerin birbirinden tamamen bağımsız olduğunu varsayması ve değişken arası ilişkinin modellenmiyor olmasıdır (Argüden ve Erşahin, 2008).

Koşullu olasılık durumları ile ilgili olan Bayesyen sınıflandırmada *m* adet sınıf olduğu düşünülürse x_i değerinin C_1 sınıfında olma olasılığı aşağıdaki formül ile hesaplanır:

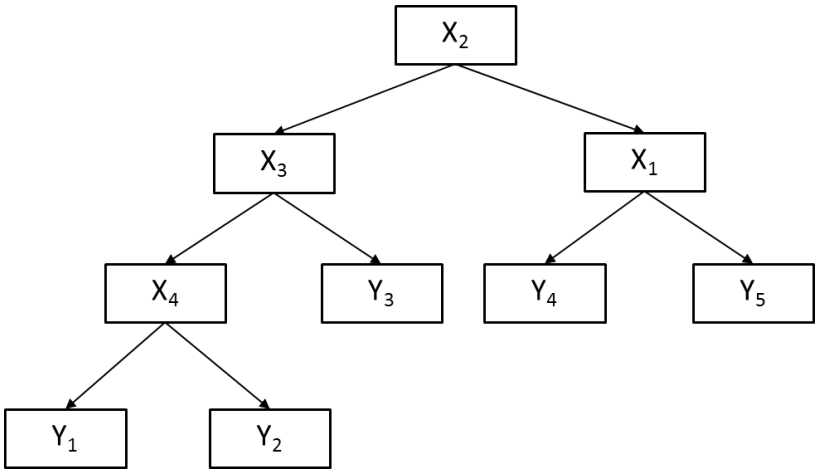
$$P(C_1|x_i) = \frac{P(x_i|C_1)P(C_1)}{P(x_i)} \quad \textbf{(4.3)}$$

4.3 Karar Ağaçları

Karar ağaçları, sınıflandırma problemlerinde sıkça kullanılan yöntemlerden biridir. Karar ağaçları karar vericinin en doğru karara ulaşması için algoritmalar uygulayabilmek adına uygun bir altyapı sağlamaktadır. Yorumlanabilmelerinin kolay olması ve veri tabanlarına kolayca entegre edilebilmeleri sebebiyle karar ağaçlarının kullanımı oldukça yaygındır.

Karar ağaçları akış şemalarına benzeyen yapılar olarak tanımlamaktadır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı yaprak, en üst yapı kök ve bunların arasında kalan yapılar ise dal olarak adlandırılmaktadır (Quinlan, 1993).

Karar ağacı işlemi kök düğümden başlar ve yukarıdan aşağıya doğru olarak ardışık düğümleri takip ederek devam eder (Han ve Kamber, 2006). Eğer dalın ucunda sınıflama işlemi gerçekleşemiyorsa, o dalın sonucunda bir karar düğümü oluşur. Fakat eğer sınıflama işlemi gerçekleşebiliyorsa dalın ucunda belirlenmek istenen sınıfı ifade eden yaprak oluşur. Şekil 4.1'de örnek olarak gösterilen karar ağacında X_i'lerin her biri bir düğümü oluşturmaktadır. Y_i'lerin her biri birer yapraktır ve aynı zamanda sınıfı temsil etmektedirler.



Şekil 4. 1: Örnek karar ağacı (Silahtaroğlu, 2013).

Karar ağacı ve karar kuralları oluşturulduktan sonra sınıfı bilinmeyen bir veriyi sınıflamak için verinin nitelik değerleri, karar ağacı üzerinde test edilir. Karar ağacı üzerinde kökten yaprağa doğru gidilerek verinin hangi sınıfta olduğu belirlenir.

Karar ağaçlarının; belirli bir sınıfın potansiyel üyelerinin belirlenmesi (s*egmentation*), olayların risk gruplarına göre kategorize edilmesi (s*tratification*), gelecekteki olayları tahmin edebilmek için belli kurallar oluşturulması, parametrik model kurulması için çeşitli veriler arasından faydalı olacakların seçilmesi, kategorilerin birleştirilerek sürekli değişkenlerin kesikli hale dönüştürülmesi gibi alanlarda kullanımı oldukça yaygındır (Akpınar, 2000).

Karar ağaçları oluşturulurken en önemli noktalardan biri de dallanmanın hangi ölçütlere göre yapılacağının belirlenmesidir. Kullanılan algoritmaya göre ağacın şekli değişebilir, farklı ağaçlar farklı sınıflandırma sonuçları vereceklerdir. Bu yüzden algoritma tespiti önemlidir.

Karar ağaçlarında uygulanabilen birçok farklı algoritma vardır. Bu algoritmalar birbirlerinden kök, düğüm ve dallanma kriterleri seçimlerinde izledikleri yöntemler açısından farklılık göstermektedirler (Silahtaroğlu, 2013).

4.3.1 ID3 Algoritması

ID3 (Induction of Decision Trees) algoritması ilk olarak J. Ross tarafından Sydney Üniversitesi'nde geliştirilmiştir (Quinlan, 1986). ID3, verilen örnekler içinde en ayırt edici olan değişkeni makine öğrenmesi ve bilişim teknolojisine dayanarak bulan ve bu işlem sırasında da entropiden yararlanan bir algoritmadır.

Entropi, verilerin birbirlerinden farklılıklarıdır. Entropi bir veri kümesi içerisinde belirsizlik ve rastgeleliği ölçmek için kullanılmaktadır (Quinlan, 1986). Entropi, sistemdeki belirsizliğin ölçüsüdür ve bir alanın entropi ölçüsü ne kadar fazla ise, o alan kullanılarak oluşturulan sonuçlar da bir o kadar belirsiz ve kararsızdır (Özkan, 2008). Bu yüzden doğru sonuca varmak adına karar ağacının kökünde entropi ölçüsü en az belirsizlik ve kararsızlık içeren değer kullanılır.

Entropi hesaplaması yapmak için, S'nin bir kaynak olduğunu ve bu kaynağın m_1,m_2…m_n olmak üzere *n* adet mesaj ürettiğini varsayalım. Birbirinden bağımsız olarak üretilen bu m_i mesajlarının gerçekleşme olasılıkları p_i'dir. P= {p_1,p_2….p_n } olasılık dağılımına sahip mesajları üreten *S* kaynağının entropisi H(S) aşağıdaki şekilde hesaplanır (Shannon, 1948):

$$\mathrm{H(S)} = \sum_{t=1}^{n} p_i * \log_2(p_i) \qquad \textbf{(4.4)}$$

H: Entropi

S: Kaynak

p: Olasılık

ID3 algoritması öncelikli düğüm ve dallanmalara karar verme aşamasında veri tabanı bölünmeden önce doğru sınıflandırma yapmak için gelen bilgi ile veri tabanı bölündükten sonra doğru sınıflandırma için gelen bilgi arasındaki farkı kullanır. Bu fark kazanım olarak adlandırılmaktadır.

Kazanım *(K)*, başlangıçtaki entropi ile her bir alt bölümün entropilerinin ağırlıklı toplamları arasındaki fark alınarak Formül 4.5 ile hesaplanmaktadır. Fark hangi alt bölüm için daha büyük ise dallanma o bölüme doğru yapılmaktadır (Silahtaroğlu, 2013).

$$Kazanım(D) = \mathrm{H(D)} - \sum_{i=1}^{n} \mathrm{P}(D_i)H(D_i) \quad \textbf{(4.5)}$$

D: Veri seti

H: Entropi

P: Olasılık (Ağırlık)

4.3.2 C4.5 Algoritması

Kategorik nitelikler yerine sayısal niteliklerin sınıflandırılmasının söz konusu olmasıyla 1993 yılında Quinlan tarafından ortaya atılan C4.5, algoritması birkaç farklı konuda ID3 algoritmasına üstünlük sağlamaktadır. C4.5 algoritmasında karar ağacı oluştururken kayıp veriler hesaba katılmaz, kazanım oranı hesaplanırken yalnızca verileri eksik olmayan kayıtlar kullanılmaktadır. Algoritma kayıp verileri diğer veri ve değişkenler yardımıyla öngörerek kazanım oranının hesaplanmasında kullanılmakta, böylece daha duyarlı ve anlamlı kurallar çıkarabilen bir ağaç üretebilmektedir (Silahtaroğlu, 2013). ID3 algoritmasının değişkenleri birçok alt bölüme ayırmasının aşırı öğrenmeye sebebiyet vermesini önlemek adına Quinlan tarafından kullanılan kazanım oranı aşağıdaki formül ile hesaplanmaktadır:

$$Kazanım\ Oranı\ (D;S) = \frac{Kazanım(D,S)}{Ayırma\ Bilgisi(D,S)} \quad \textbf{(4.6)}$$

$$Ayırma\ Bilgisi(D;S) = H\left(\frac{|D_1|}{|D|}, \ldots . \frac{|D_S|}{|D|}\right) \quad \textbf{(4.7)}$$

D: Veri seti

S: Kaynak

H: Entropi

4.3.3 CART Algoritmaları

Sınıflandırma ve regresyon ağaçları (*CART - Classification and Regression Trees*), 1984'te Breiman tarafından ortaya atılmıştır. CART algoritması da ID3 algoritmasında olduğu gibi en iyi dallara ayırma kriterini seçmek için entropiden faydalanır (Dunham, 2003). Fakat ayırma kriterini belirlemek için C4.5 ve ID3 algoritmalarından farklı bir formül kullanır ve yine ID3 algoritmasından farklı olarak dallara ayırma kriterini hesaplarken kaybolan verileri önemsemez.

CART algoritmasında her bir karar düğümünün iki dala ayrılması söz konusudur. Aşağıdaki şekilde hesaplanan Ψ değerleri içinden en büyük değere sahip nokta düğüm olarak seçilir ve aynı işlemler tüm yapraklara ulaşıncaya kadar devam ettirilir (Silahtaroğlu, 2013).

Herhangi bir *t* düğümündeki *s* dallara ayrılma kriteri Ψ *(s/t)* olarak gösterilirse:

$$\Psi(\mathrm{s/t}) = 2P_L P_R \sum_{J=1}^{M} \left|P\left(C_J|t_L\right) - P\left(C_J|t_R\right)\right| \quad \textbf{(4.8)}$$

t: Dallanmanın yapılacağı düğüm

c: Kriter

L: Ağacın sol tarafı

R: Ağacın sağ tarafı

P_L, P_R: Öğrenim kümesindeki bir kaydın sağda veya solda olma olasılığı

$P\left(C_J|t_L\right)$ ve $\mathrm{P}\left(C_J|t_R\right)$: C_jsınıfındaki bir kaydın sağda veya solda olma olasılığı

4.3.4 SLIQ

Araştırmada Denetimli Öğrenme (*SLIQ - Supervised Learning in Quest*) algoritması 1996 yılında Mehta M., Agrawal R. ve Rissanen J. tarafından ortaya konulmuştur (Mehta ve diğ, 1996).

Hem sayısal hem de kategorik verilerin sınıflandırılmasında kullanılabilen SLIQ algoritması, sayısal verilerin değerlendirilmesi aşamasındaki maliyeti azaltmak adına önceden sıralanma tekniğini kullanmaktadır.

SLIQ algoritmasında verileri sıralama işlemi her düğümde uygulanmaz, sadece öğrenme verilerine ve ağacın büyüme aşamasının başlangıcında sadece bir kere uygulanır. "Önce derinlik" ilkesiyle çalışan ID3 ve C4.5 gibi algoritmalar en iyi dallara ayırma kriterine ulaşabilmek için her düğümde verileri sürekli olan sıraya dizerlerken, bunlardan farklı olarak SLIQ algoritması "önce genişlik" ilkesi ile çalışarak aynı anda birçok yaprak oluşturur (Silahtaroğlu, 2013).

SLIQ algoritmasında düğümlerden alt dallara ayırma kriteri için Formül 4.9'da gösterilen *Gini* indeksini kullanılır, burada p_j, K kümesi içinde *j* sınıfının sıklığıdır (Shafer ve diğ, 1996):

$$gini(\mathrm{K}) = 1 - \sum p_j^2 \qquad \textbf{(4.9)}$$

SLIQ algoritmasının diğer bir avantajı ise bellekte tutulması güç olan büyük verilere, veriler belleğe alınmadan doğrudan kolayca uygulanabilmesidir.

4.3.5 SPRINT

SPRINT (*Scalable Parallelizable Induction of Decision Trees*) algoritması da SLIQ algoritması gibi "önce genişlik" ilkesiyle çalışmaktadır. SPRINT algoritması verileri sıralama işlemini bir kez yapması ile SLIQ algoritmasıyla benzerlik gösterse de farklı veri yapılarını kullanmaktadır (Shafer ve diğ, 1996).

SPRINT algoritması her bir değişken için ayrı liste hazırlamaktadır. Değişken sayısı kadar olan her tabloda, kullanılacak olan değişken için sınıf ve sıra numaraları yer almaktadır. Sürekli değer taşıyan tablolar sürekli değer değişkenine göre sıraya dizilmekte, kategorik veriler taşıyan tablolar ise sıra numaralarına göre sıralı olarak kalmaktadır. Eğitim kümelerinden elde edilen ilk listeler ağacın kökleriyle ilişkilendirilmektedir. Ağaçlar büyüyüp düğümler yeni dallara bölündükçe her düğüme ait değişken listeleri de bölünür ve yeni dallarla ilişkilendirilir. Bölünme aşamasına gelmiş düğümler için düğümlerdeki sınıf dağılımlarını elde etmek amaçlı kullanılan $C_{üst}$ ve C_{alt} adı verilen histogramlar belirlenmektedir (Silahtaroğlu, 2013). SLIQ algoritmasında olduğu gibi SPRINT algoritmasında da düğümlerden alt dallara ayırma kriteri için *Gini* indeksi kullanılmaktadır.

4.4 Yapay Sinir Ağları

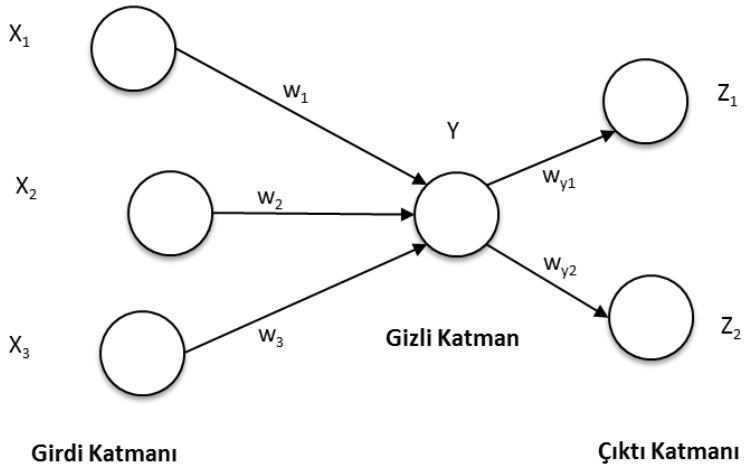
Yapay Sinir Ağları (*Artificial Neural Networks*) biyolojik sinir ağlarından esinlenerek ve insan beyninin işleme mantığını temel alarak geliştirilmiş bir bilgi işleme sistemidir (Larose, 2005). Tecrübelerden yararlanarak belirsiz ve karmaşık verilerden bilgi edinebilirler. Katmanlar şeklinde bulunan Yapay Sinir Ağları (YSA) birbirine bağlı nöronlardan oluşmaktadır. YSA bu nöronların matematiksel olarak modellenmesidir. Böylece süreç davranışını anlamak için kurulan model öğrenme faaliyetini temel alarak geliştirilmektedir. Yani bu yöntemde parametrik bir model öngörülmemektedir, her uygulamadan sonra bir ders çıkarılmakta süreç o şekilde ilerlemektedir (Argüden ve Erşahin, 2008).

YSA sadece sınıflandırma işlemlerinde değil kümeleme ve örüntü tanımlama işlemlerinde de kullanılmaktadır (Silahtaroğlu, 2013). Aynı zamanda yapay sinir ağlarının endüstriyel uygulamalar, finans uygulamaları, askeri ve savunma uygulamaları, tıp ve sağlık uygulamaları, mühendislik uygulamaları, robotbilim, görüntü işleme, örüntü tanıma dışında iletişim sanayi, eğlence amaçlı tahmin gibi birçok alanda uygulamaları bulunmaktadır (Uğur ve Kınacı, 2008). Hava tahminleri, müşteri profillerinin belirlenmesi, ses ve şekil tanıma, kredi değerlendirmeleri, biyomedikal modellemeler, satış tahminleri, tıbbi teşhisler yapay sinir ağlarının kullanıldığı örnek uygulama alanlarından bazılarıdır.

YSA'nın veri madenciliğinde kullanımı birçok avantaj sağlamaktadır. Davranış öğrenme özelliği sayesinde komplike veriye sahip olan birçok alanda problem çözümü yapabilmektedir. Öğrenme süresinin uzun olması bir dezavantaj gibi görünse de öğrenme sağlandıktan sonra ortaya çıkan sonuçlarla çok duyarlı bir sınıflandırma yapmak mümkündür. Diğer bir dezavantajı ise kullanımının bir uzmanlık gerektirmesi ve çıkan sonucun ifade edilmesinin güç olmasıdır.

YSA'nın en küçük birimi nöronlardır. Nöronlar bir araya gelerek ağları oluşturur. YSA giriş, çıkış ve gizli katmanlardan oluşan katmanlı bir yapıdadır. Ağdaki sinirler katmanlar boyunca düzenlenmişlerdir. Tüm sinir ağlarında giriş ve çıkış katmanları bulunmak zorundadır, gizli katman ise zorunlu değildir. Gizli katmanın aktif hale gelebilmesi için fonksiyon değerinin belirli bir eşik değerinin üzerinde olması gerekmektedir (Silahtaroğlu, 2013). Daha fazla örüntü tanımlanmasını mümkün kıldığından gizli katman ağı daha güçlü kılmaktadır.

Her nöronun bir iç hali yani aktivasyon seviyesi bulunmaktadır, bu seviye gelen girdileri tanımlayan bir fonksiyondur. Ağ içerisinde bulunan nöronlar diğer nöronlara işaret gönderirler, bu işaretler gönderilen nöronların giriş fonksiyonunu ifade etmektedir. YSA'da bilgi, nöronlar arasındaki bağlantı ağırlıkları üzerinden dağıtılır. Şekil 4.2'de bulunan yapay sinir ağı örneğinde X_1, X_2 ve X_3 nöronları ve onları Y nöronuna bağlayan ağırlıklar w_1,w_2 ve w_3 olarak gösterilmektedir.



Şekil 4. 2: Yapay sinir ağı örneği (Silahtaroğlu, 2013).

Y nöronunun girdisi aşağıdaki şekilde ifade edilmektedir:

$$y\text{-}girdi = w_1x_1 + w_2x_2 + w_3x_3 \qquad \textbf{(4.10)}$$

Y nöronunun etkin hale gelebilmesi için uygulanan bir ateşleme fonksiyonu ile belirli eşik değerini aşması gerekmektedir. Uygulanabilecek örnek fonksiyonlar aşağıdaki Çizelge 4.1'de gösterilmiştir:

Çizelge 4. 1: Bazı ateşleme fonksiyonları.

Fonksiyon	**Formül**
Lojistik Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
Hiperbolik Tanjant	$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$

Model kurulduktan sonra eğitim verileri sürekli olarak modele girilerek sonuçlar gerçek sonuçlar ile karşılaştırılır. YSA'da kullanılan *w* ağırlıkları her seferinde Δw

kadar düzeltilerek modelde iyileştirmeler yapılır (Silahtaroğlu, 2013). Böylece her defasında ortaya çıkan hatayı en aza indirecek bir yapı kullanılmaktadır. Algoritma hatanın minimum olduğu noktada tamamlanmış olacaktır.

5. UYGULAMA

Bu çalışma kapsamında Uludağ Üniversitesi Tıp Fakültesi Hematoloji Anabilim Dalı bünyesinde anemi teşhisi alan hastaların ve hematolojik açıdan sağlıklı bireylerin tam kan sayımı testleri incelenerek bir veri seti oluşturulmuştur. Bu veri setine veri madenciliği yöntemlerinden Karar Ağaçları, Yapay Sinir Ağları, Bayesyen Sınıflandırma ve *K*-en Yakın Komşu yöntemleri uygulanarak sınıflandırma yapılmıştır. Elde edilen doğruluk (*accuracy*) oranları karşılaştırılmıştır. Sınıflandırma modellerinin uygulanmasında RapidMiner 6.0.008 yazılım programından yararlanılmıştır.

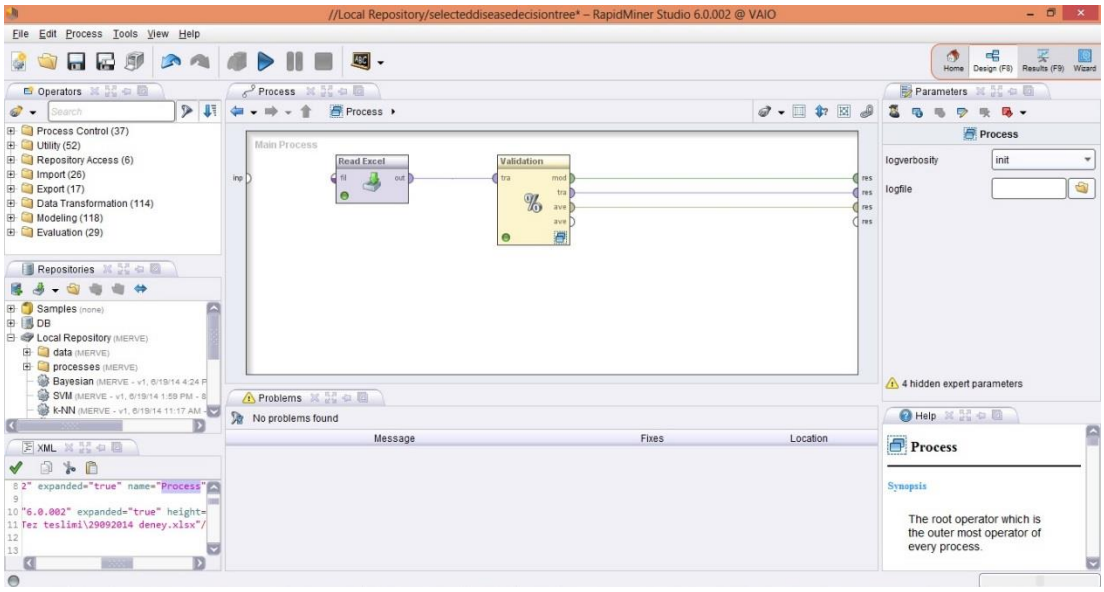
Bu bölümde, veri setinin eldesi ve işlenmesi aşaması ayrıntılı olarak açıklanmıştır. Kullanılan yazılım programı ve anemi hastalığı çeşitleri hakkında genel bir bilgi verilmiştir. Çalışmanın uygulama detaylarına bu bölüm içerisinde yer verilmektedir.

5.1 Çalışmada Kullanılan Yazılım Programı

Veri madenciliği yöntemlerini uygulamak için birçok farklı yazılım bulunmaktadır. IBM SPSS Clementime, SAS Enterprise Miner, STATISTICA Data Miner ve RapidMiner bunlardan bazılarıdır. Bu çalışma kapsamında RapidMiner yazılımı kullanılmıştır.

RapidMiner veri madenciliği için kullanılan açık kodlu ücretsiz bir yazılım programıdır. İlk olarak 2001 yılında Dortmund Teknik Üniversitesi Yapay Zeka bölümündeki Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer tarafından YALE (Yet Another Learning Environment) adıyla geliştirilmeye başlanmış, 2007 yılında Rapidminer ismini almıştır (Url-1). RapidMiner karmaşık veri setlerinden bir örüntü elde etmeye yardımcı olan veri madenciliği yöntemlerini içerir ve kullanıcı için kolay bir arayüze sahiptir.

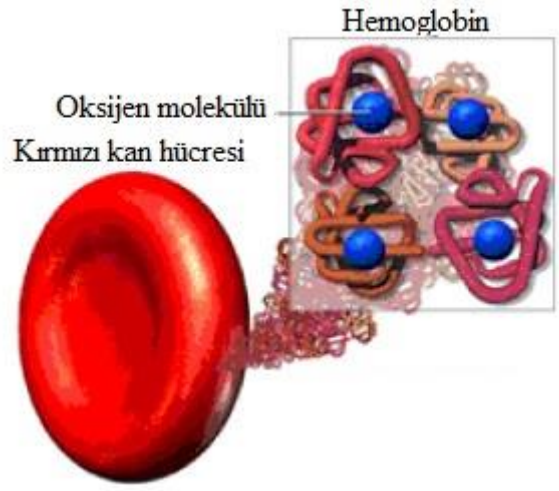
RapidMiner sınıflandırma ve değerlendirme için kolay bir arayüz sağladığı gibi veri madenciliği uygulamaları kullanımını geliştirmeye de yardımcı olmaktadır (Shukla ve Barot, 2013). RapidMiner'a ait bir ekran görüntüsü Şekil 5.1'de gösterilmektedir.



Şekil 5. 1: RapidMiner ekran görüntüsü.

5.2 Anemi

Kanda bulunan kırmızı kan hücrelerinin (alyuvarlar) yapısında oksijenin taşınmasını ve bu hücrelerin kırmızı olmasını sağlayan hemoglobin bulunmaktadır (Şekil 5.2). Nefes alırken akciğerdeki oksijen, bu hemoglobinin yapısına bağlanarak taşınır. Bu hemoglobinin kanda bulunması gereken miktarın altında olması sonucu kansızlık (anemi) diye tanımlanan kan hastalığı ortaya çıkmaktadır.



Şekil 5. 2: Hemoglobin yapısı (Url-2).

Anemi, hem gelişmiş hem de gelişmekte olan ülkelerde halk sağlığına etki eden küresel bir halk sağlığı sorunudur. Anemi, kan hastalıkları içinde en sık rastlanılan rahatsızlıktır (Sanap ve diğ,2011). Anemi, toplam eritrosit sayısının azalması veya eritrositlerin içindeki hemoglobin miktarının azalması veya her ikisinin birlikte olması sonucu oluşan bir hastalıktır. Kısaca anemi kan hücrelerinin sayısının normalden düşük olması durumudur (Şekil 5.3).

Şekil 5. 3: Kırmızı kan hücresi sayısı (Url-3).

Anemi, kan testi ile tespit edilebilir ve tanı için hekimin kan testi sonuçlarını görmesi gerekmektedir. Dünya Sağlık Örgütü'ne göre anemi kan sayımı testlerinden sağlanan hemoglobin ve serum hematokrite (kırmızı kan hücrelerinin oluşturduğu hacmin, toplam kan hacmine oranı) bağlı olarak saptanabilmektedir. Ancak, tüm bu işlemler özellikle gelişmemiş ülkelerde pahalı ve zor olabilmektedir.

Göreceli olmakla birlikte, genelde her ülke için belirli bir ortalama değer vardır ve bunun altındaki hemoglobin düzeyi anemi kabul edilmektedir. Hematokrit değeri erkeklerde %39'dan, kadınlarda ise %36'dan düşük olduğunda hemoglobinin erkeklerde 13'ten, kadınlarda ise 12'den az olması aneminin açık bir göstergesidir (Yılmaz ve Bozkurt, 2011). Dünyada kadınlarda görülme sıklığı %30-40, erkeklerde ise yaklaşık %20′dir (THD Ulusal Tedavi Klavuzu, 2011).

Anemiler farklı şekillerde sınıflandırılabilmektedirler. Bunlar aşağıdaki şekilde sıralanabilir (THD Ulusal Tedavi Klavuzu, 2011):

- *Sitometrik düzen ile sınıflandırma:* Bu sınıflandırma hücre boyutlarına ve hemoglobin içerme parametrelerine göre gerçekleştirilmektedir.

- *Eritrokinetik düzen ile sınıflandırma:* Bu sınıflandırma kırmızı kan hücrelerinin üretim ve yıkım oranları hesap edilerek gerçekleştirilmektedir.
- *Biyokimyasal /Moleküler düzen ile sınıflandırma:* Bu sınıflandırma düzeninde moleküler düzeyde anemi etiyolojisi dikkate alınarak ilerlenmektedir.

Sitometrik parametreler diğerlerine göre çok daha ucuz ve kolay uygulanabilir olduklarından anemi sınıflandırmasında genellikle sitometrik düzen uygulanmaktadır. Bu düzene göre anemiler 3 temel sitometrik kategoriye ayrılmaktadır:

- **Normokromik, normositik anemi**
- Kandaki ortalama alyuvar hemoglobin yoğunluğu (MCHC) ve ortalama eritrosit hacmi (MCV) değerlerinin normal aralıkta olduğu kategoridir.
- Kronik hastalık anemileri, hemolitik anemiler ve aplastik anemiler bu kategoride bulunmaktadır.
- **Hipokromik, mikrositik anemi**
- Kandaki ortalama alyuvar hemoglobin yoğunluğu (MCHC) ve ortalama eritrosit hacmi (MCV) değerlerinin normalin altında olduğu kategoridir.
- Demir eksikliği anemisi ve talasemi bu kategoride yer almaktadır.
- **Normokromik, makrositik anemi**
- Kandaki ortalama alyuvar hemoglobin yoğunluğu (MCHC) normal değerlerde olsa da, ortalama eritrosit hacmi (MCV) değerinin normalin üstünde olduğu kategoridir.
- Megaloblastik anemiler bu kategoride yer almaktadır.

Sınıflandırmada da görüldüğü gibi aneminin birçok farklı çeşidi bulunmaktadır. Çalışmamız kapsamında incelenen anemi çeşitleri aşağıdaki şekilde sıralanabilir:

- Demir eksikliği anemisi
- Aplastik anemi
- Talasemi
- Romatolojik hastalıklara bağlı anemi
- Böbrek hastalıklarına bağlı anemi

5.2.1 Demir eksikliği anemisi

Demir eksikliği anemisi vücutta yeterli demirin bulunmamasından kaynaklanan, sık görülen ve kolay tedavi edilebilen bir anemi türüdür (Yılmaz ve Bozkurt, 2011). Demir eksikliği anemisi, demirin yiyeceklerle alımının ya da bağırsaklardan emiliminin az olması sonucu, hemoglobin yapımının azalmasına sebep olmasıyla tanımlanmaktadır (Doğan ve Türkoğlu, 2008). Bu durumda eritrositler normalden daha küçük olup, görevlerini tam ve başarıyla yerine getirememektedirler.

Eksiklik, fizyolojik olarak artan ihtiyacın karşılanamaması ya da kaybın artması ile ortaya çıkabilmektedir. Demir eksikliği anemisinin en sık nedenleri aşağıdaki şekilde sıralanabilir (THD Ulusal Tedavi Klavuzu, 2011):

- Fizyolojik kan kayıpları
- Patolojik kan kayıpları
- Gıda demirinin yetersiz emilimi

Demir eksikliği anemisi halsizlik, nefes darlığı, göğüs ağrısı vb. belirtilere neden olabilmektedir. Ağır demir eksikliği anemisi kalp sorunlarına, enfeksiyonlara, çocuklarda büyüme ve gelişme ile ilgili problemlere ve diğer komplikasyonlara yol açabilmektedir (THD Ulusal Tedavi Klavuzu, 2011).

5.2.2 Aplastik anemi

Aplastik anemi, kemik iliğinin yeteri kadar veya hiç yeni hücre üretememesi durumudur.

Aplastik terimi yeni dokular meydana getirebilme yeteneğinden yoksun olan, fonksiyonunu tam icra edemeyen manasına gelir ve burada kemik iliğinin fonksiyonunu tam olarak icra edememesini işaret eder (THD Ulusal Tedavi Kılavuzu, 2011). Tipik olarak anemi düşük eritrosit sayımını tanımlar ama aplastik anemi hastalarında üç kan hücresi tipinin (trombosit, eritrosit, lökosit) de sayımları genel olarak düşüktür.

Genel olarak sadece gençlerde veya 20'li yaşlarda değil, yaşlılarda da oldukça sık görülmektedir. Aplastik anemiler doğuştan olabildikleri gibi sonradan da oluşabilirler.

Kimyasallara, ilaçlara, radyasyona maruz kalmak veya enfeksiyonlar, kalıtımsal durumlar aplastik anemiye sebebiyet verebilmektedir. Aplastik aneminin bilinen sebeplerinden biri de bir tür otoimmün bozukluktur. Bu bozuklukta lökositler (akyuvarlar) kemik iliğine saldırmakta ve ona zarar vermektedir (THD Ulusal Tedavi Klavuzu, 2011).

Aplastik anemi tedavisi bağışıklık sisteminin günlük ilaç alımıyla bastırılması ile gerçekleştirilmekte, ciddi vakalarda ise kemik iliği naklini içermektedir.

5.2.3 Talasemi

Talasemi, anne ve babadan çocuklara kalıtsal olarak geçen, önlenebilir bir kan hastalığıdır. Anemi oluşmasına neden olan etmen, kanda alyuvarların yapısında yer alan hemoglobin molekülünün yapısındaki bozukluktur (Elshami ve Alhalees, 2012). Akdeniz anemisi olarak da bilinen talasemi, Türkiye'nin de içinde olduğu Akdeniz ülkelerinde önemli bir halk sağlığı sorunudur. Taşıyıcıların saptanması, genetik danışma ve doğum öncesi tanı konabilmesiyle engellenebilir bir hastalık olmasına rağmen, dünyada her yıl en az 365.000 talasemi hastası doğmakta ve tedavi görmektedir. Türkiye'de yaklaşık 1.300.000 talasemi taşıyıcısı ve 4.500 kadar talasemi hastası vardır (THD Ulusal Tedavi Klavuzu, 2011). Talasemi hastalığı, tedavisi düzgün sürdürülmezse yaşam süresini belirgin kısaltan ve yaşam kalitesini son derece olumsuz etkileyen ağır bir hastalıktır (THD Ulusal Tedavi Kılavuzu, 2011).

5.2.4 Romatolojik hastalıklara bağlı anemi

Bazı kronik hastalıkların anemiye sebep oldukları bilinmektedir. Romatoid artrit bu hastalıklardan biridir. Romatoid artrit, en sık rastlanan sistemik otoimmün, eklemlerde artrite yol açan simetrik olarak kronik inflamatuar bir hastalıktır (Kınıklı, 2011). Romatoid artritli hastalarda sistemik hastalığın başlamasından 1-2 ay sonrasında kronik hastalık anemisi gelişir ve anemi, hastalığın seyri ile paralellik gösterir (Turgay, 2011).

Romatoid artritte anemi farklı bir çok neden ile gelişebilir. Bunlardan başlıcaları, asıl hastalığın etkisine bağlı olarak gelişen kronik hastalık anemisi, demir eksikliği anemisi ve ilaç yan etkisi nedeni ile gelişen anemilerdir (THD Ulusal Tedavi Klavuzu, 2011).

5.2.5 Böbrek hastalıklarına bağlı anemi

Anemi böbrek hastalıklarında sıklıkla görülmektedir. Böbrek yetmezliğinde anemi görülmesinin birçok nedeni vardır; ancak en sık nedeni eritropoietin yetersizliğine bağlı azalmış eritropoezdir (Knott, 2014).

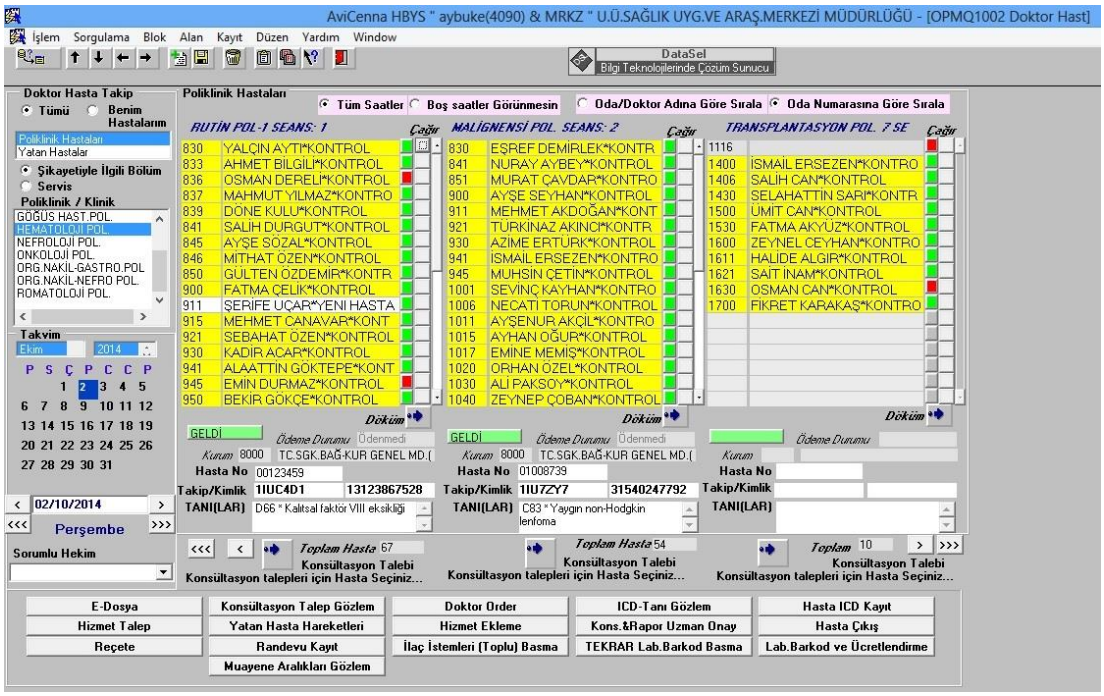
Kronik böbrek hastalığı olan bireylerde, hastalığın erken evrelerinden itibaren anemi ortaya çıkabilmekte, hastalıklık derinleştikçe anemi sıklığı da artmaktadır. Evre 3 kronik böbrek hastaların %15, Evre 5 prediyalitik hastaların %50, diyaliz hastalarının ise % 80'den fazlasının anemik olduğu saptanmıştır (Şengül ve Ertürk, 2008).

Diyaliz Hekimleri Derneği'ne göre kronik böbrek hastalarında anemi nedenleri eritropoietin yetersizliği, demir eksikliği, yetersiz diyaliz, azalmış kemik iliği fonksiyonu, kan kaybı, beslenme bozuklukları olarak sıralanabilir (Url-4).

5.3 Veri Seti

Çalışma kapsamında kullanılan veriler Uludağ Üniversite Tıp Fakültesi Hematoloji Anabilim Dalı'na gelen hastaların tam kan sayımı testlerinin incelenmesi ile elde edilmiştir. Uludağ Üniversitesi Hasta Yönetim Bilgi Sistemi AVICENNA programından faydalanarak 2008-2014 yılları arasında hematoloji birimine gelerek anemi teşhisi konulan hastalar incelenmiştir (Şekil 5.4). Hamileler, çocuklar ve kanser hastaları gibi kan değerlerini etkileyebilecek hastalıklar çalışma kapsamının dışında tutulmuştur. Tüm veriler ve sistemde yazan teşhisler Uludağ Üniversitesi Tıp Fakültesi Dahiliye alanında ihtisasını yapan Dr. Aybüke Muti tarafından gözden geçirilmiş ve onay alınmıştır. Çalışmanın yapılabilmesi için 15 kişiden oluşan bir hekim heyetine öneri formu sunulmuş ve çalışmasının yapılması için onay alınmıştır. Uludağ Üniversitesi Tıp Fakültesi'nden alınan etik kurul kararı EK-A'da verilmiştir.

Bu kapsamda incelenen yüzlerce veri arasından %75'i bayan, %25'i erkek toplam 323 kişinin tam kan sayımı testleri çalışma için uygun bulunmuştur. Seçilen 323 kişinin 221'i farklı anemi çeşitlerinden teşhis almış, 102'si ise hematolojik açıdan sağlıklı kişilerdir (Çizelge 5.1).



Şekil 5. 4: AVICENNA ekran görüntüsü.

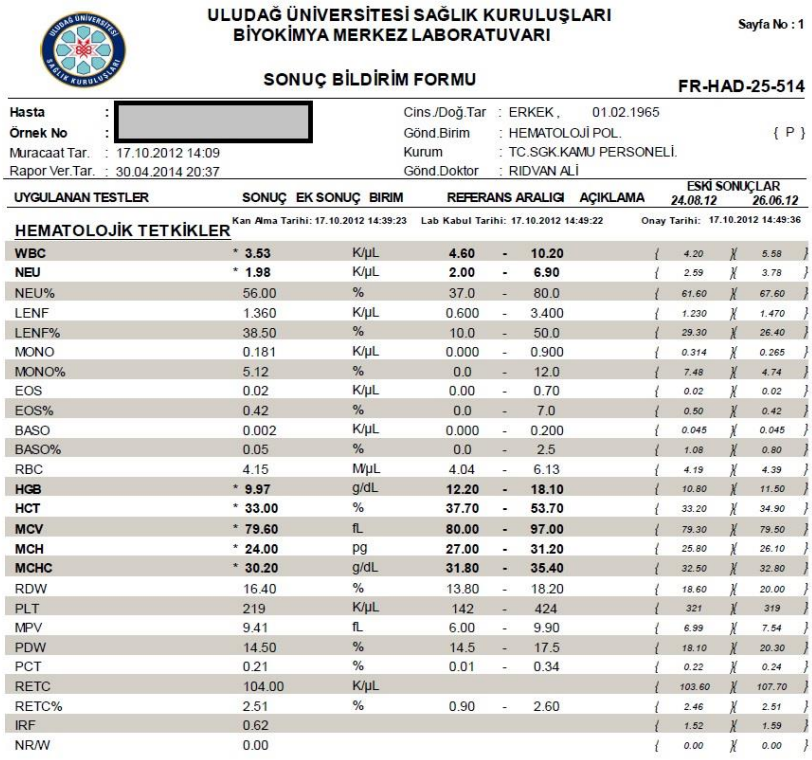
Çizelge 5. 1: Veri seti.

Veri Seti		**Sayı**
Anemi	Böbrek hast. bağlı anemi	44
	Demir eksikliği anemisi	52
	Talasemi	44
	Romatolojik hast. bağlı anemi	45
	Aplastik anemi	36
Sağlıklı		102
TOPLAM		**323**

Çalışmada incelen anemi çeşitleri aşağıda sıralanmıştır:

- Aplastik anemi
- Demir eksiliği anemisi
- Talasemi
- Böbrek hastalıklarına bağlı anemi
- Romatizmal hastalıklara bağlı anemi

Tam kan sayımı testlerinde yaklaşık 25 farklı kan parametresi bulunabilmektedir (Şekil 5.5).



ULUDAĞ ÜNİVERSİTESİ SAĞLIK KURULUŞLARI
BİYOKİMYA MERKEZ LABORATUVARI

Sayfa No : 1

SONUÇ BİLDİRİM FORMU

FR-HAD-25-514

Hasta : [redacted]
Örnek No :
Muracaat Tar. : 17.10.2012 14:09
Rapor Ver.Tar. : 30.04.2014 20:37

Cins./Doğ.Tar : ERKEK , 01.02.1965
Gönd.Birim : HEMATOLOJİ POL.
Kurum : TC.SGK.KAMU PERSONELİ.
Gönd.Doktor : RIDVAN ALİ

{ P }

UYGULANAN TESTLER	SONUÇ	EK SONUÇ	BIRIM	REFERANS ARALIGI			AÇIKLAMA	ESKİ SONUÇLAR 24.08.12		26.06.12
HEMATOLOJİK TETKİKLER	Kan Alma Tarihi: 17.10.2012 14:39:23			Lab Kabul Tarihi: 17.10.2012 14:49:22			Onay Tarihi: 17.10.2012 14:49:36			
WBC	* **3.53**		K/µL	**4.60**	-	**10.20**		4.20		5.58
NEU	* **1.98**		K/µL	**2.00**	-	**6.90**		2.59		3.78
NEU%	56.00		%	37.0	-	80.0		61.60		67.60
LENF	1.360		K/µL	0.600	-	3.400		1.230		1.470
LENF%	38.50		%	10.0	-	50.0		29.30		26.40
MONO	0.181		K/µL	0.000	-	0.900		0.314		0.265
MONO%	5.12		%	0.0	-	12.0		7.48		4.74
EOS	0.02		K/µL	0.00	-	0.70		0.02		0.02
EOS%	0.42		%	0.0	-	7.0		0.50		0.42
BASO	0.002		K/µL	0.000	-	0.200		0.045		0.045
BASO%	0.05		%	0.0	-	2.5		1.08		0.80
RBC	4.15		M/µL	4.04	-	6.13		4.19		4.39
HGB	* **9.97**		g/dL	**12.20**	-	**18.10**		10.80		11.50
HCT	* **33.00**		%	**37.70**	-	**53.70**		33.20		34.90
MCV	* **79.60**		fL	**80.00**	-	**97.00**		79.30		79.50
MCH	* **24.00**		pg	**27.00**	-	**31.20**		25.80		26.10
MCHC	* **30.20**		g/dL	**31.80**	-	**35.40**		32.50		32.80
RDW	16.40		%	13.80	-	18.20		18.60		20.00
PLT	219		K/µL	142	-	424		321		319
MPV	9.41		fL	6.00	-	9.90		6.99		7.54
PDW	14.50		%	14.5	-	17.5		18.10		20.30
PCT	0.21		%	0.01	-	0.34		0.22		0.24
RETC	104.00		K/µL					103.60		107.70
RETC%	2.51		%	0.90	-	2.60		2.46		2.51
IRF	0.62							1.52		1.59
NR/W	0.00							0.00		0.00

Şekil 5. 5: Tam kan sayımı testi örneği.

Uzman görüşleri alınarak anemi için önemi olan parametreler sekize indirilmiş, yaş ve cinsiyet kriterleri de eklendiğinde veri setinde 10 farklı parametre incelenmiştir. İncelenen kan parametreleri aşağıda sıralanmıştır (Çizelge 5.2):

Çizelge 5. 2: Kan parametreleri ve açıklamaları.

Parametre	**Açıklaması**
RBC	Kırmızı kan hücreleri
HGB	Hemoglobin
HCT	Hematokrit
MCV	Ortalama eritrosit hacmi
MCH	Ortalama hemoglobin hacmi
MCHC	Ortalama hemoglobin hacmi yoğunluğu
RDW	Kırmızı kan hücreleri dağılım genişliği
PLT	Platelet

5.3.1 Veri seti ön işleme prosedürü

Verilere uygulanan ön işlemler işlenecek veriyi oldukça etkilemektedir. Çalışmada elde edilen veriye uygulanan ön işleme aşamalarından aşağıda bahsedilmektedir.

- **Verilerin toplanması**

2008-2014 yılları arasında Uludağ Üniversitesi Tıp Fakültesi Hematoloji kliniğine gelen ve çeşitli anemi teşhisleri alan hastaların kayıtları AVICENNA hasta takip programı kullanılarak incelenmiştir (Şekil 5.6). İnceleme bayan ve erkek yetişkin hastaları kapsamaktadır. Hamileler, 18 yaşından küçükler ve kanser hastaları sonuçları saptırmamak adına çalışma kapsamı dışında tutulmuştur.

Şekil 5. 6: AVICENNA Hastane Bilgi Yönetim Sistemi.

- **Verilerin temizlenmesi**

Verilerin toplanması U.Ü Tıp Fakültesi Dahiliye alanında ihtisas yapan Dr. Aybüke Muti gözetiminde gerçekleştirilmiştir. Farklı hastalıklara sahip olan ve farklı tedaviler alan kişilerin verileri, kan değerlerini çarptırmaması adına veri seti dışına çıkarılmıştır. Datadaki gürültüler temizlenmiş ve anemi hastalığı teşhisinde önemsiz olduğu uzmanlar yardımı ile belirlenen parametreler veri setinden çıkarılmıştır. Herhangi bir parametresinde eksiklik bulunan birkaç veri için aynı sınıftaki diğer verilerin ilgili parametrelerinin ortalamaları alınarak eksik veriler tamamlanmıştır.

- **Verilerin dönüştürülmesi**

Kan parametreleri belirli aralıklar içerisinde değiştiğinden aralıkları daraltmak adına bir işlem uygulanmamıştır. Dönüştürme işlemi hastane kayıtlarında doğum tarihi olarak geçen verilerin, yaş olarak düzenlenerek veri setine alınması şeklinde uygulanmıştır. Buna ek olarak, yapay sinir ağları yöntemi uygulanmadan önce veri setinde nominal olan cinsiyet parametresi nümerik olarak düzenlenmiştir.

5.4 Uygulama

Çalışmada oldukça yaygın olan anemi hastalığının sınıflandırılması amacı ile toplanıp, gerekli ön işlemlerden geçirilen verilere veri madenciliği methodları uygulanmıştır. Farklı girdilerden oluşan dört farklı test grubu oluşturulmuş, algoritmalar çalıştırılırken test yöntemi olarak "10-kere çapraz doğrulama" metodu kullanılmıştır. Oluşturulan test gruplarına uygulanan yöntemlerin kriter parametrelerinde değişiklikler yapılarak yöntemler kendi içinde karşılaştırıldığı gibi, aynı testlere farklı yöntemler uygulanarak yöntemler arası doğruluk (*accuracy*) oranları karşılaştırmaları gerçekleştirilmiştir.

Hazırlanan dört farklı test grubu hakkındaki bilgiler aşağıdaki tabloda açıklanmaktadır (Çizelge 5.3):

Çizelge 5. 3: Test grupları.

	Test0	**Test1**	**Test2**	**Test3**
Durum	Böbrek hast.bağlı anemi	Böbrek hast.bağlı anemi	Böbrek hast.bağlı anemi	Böbrek hast.bağlı anemi
	Romatolojik hast.bağlı anemi	Romatolojik hast.bağlı anemi	Romatolojik hast.bağlı anemi	Romatolojik hast.bağlı anemi
	Talasemi	Talasemi	Talasemi	Talasemi
	Demir eksikliği anemisi	Demir eksikliği anemisi	Demir eksikliği anemisi	Demir eksikliği anemisi
	Aplastik anemi	Aplastik anemi	Aplastik anemi	Aplastik anemi
	Sağlıklı	Sağlıklı	Sağlıklı	Sağlıklı
Girdiler	Cinsiyet	Cinsiyet	Yaş	RBC
	Yaş	RBC	RBC	HGB
	RBC	HGB	HGB	HCT
	HGB	HCT	HCT	MCV
	HCT	MCV	MCV	MCH
	MCV	MCH	MCH	MCHC
	MCH	MCHC	MCHC	RDW
	MCHC	RDW	RDW	PLT
	RDW	PLT	PLT	
	PLT			

Belirlenen bu dört test grubuna, RapidMiner yazılım programından yararlanılarak veri madenciliği yöntemleri uygulanmıştır. Uygulanan sınıflandırma yöntemlerinin

parametreleri değiştirilerek hem en yüksek doğruluk oranı veren parametreler hem de testlerde en yüksek doğruluk oranına sahip methodlar belirlenmiştir.

Test gruplarına uygulanan methodlar ve parametre değerleri aşağıdaki tabloda özetlenmektedir:

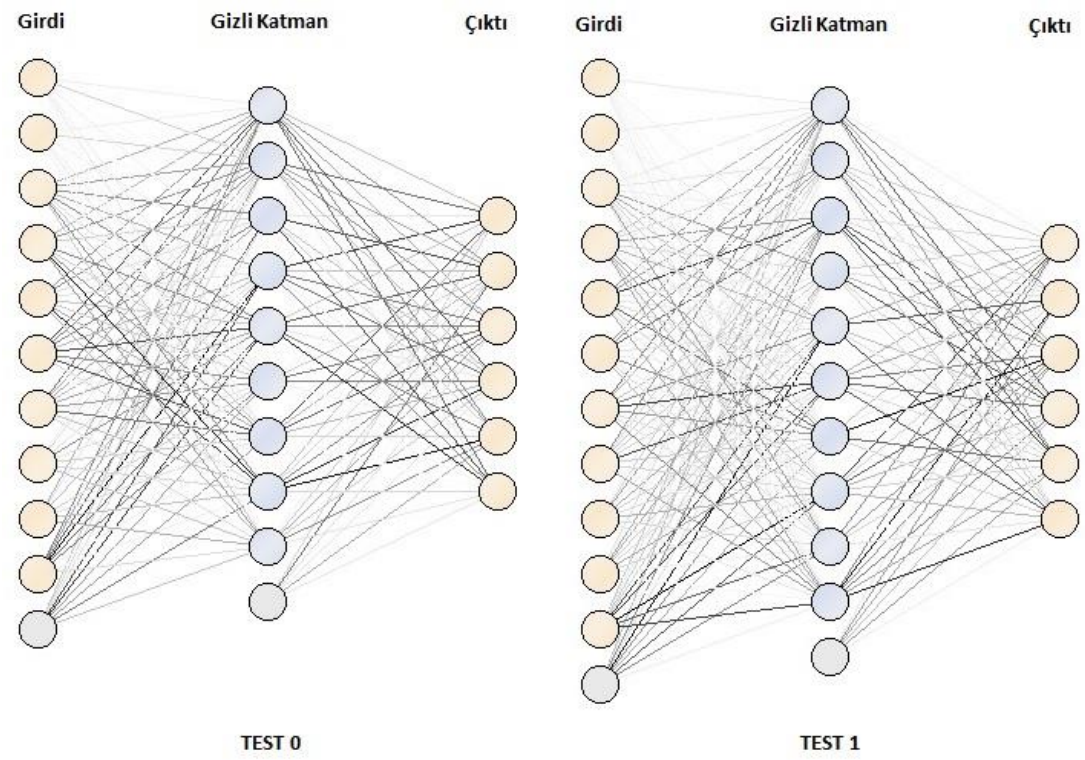
Çizelge 5. 4: Test sonuçları.

			TEST0	**TEST1**	**TEST2**	**TEST3**
Veri Madenciliği Yöntemleri	**Parametre**	**Değer**	***Doğruluk(%)***	***Doğruluk(%)***	***Doğruluk(%)***	***Doğruluk(%)***
Yapay Sinir Ağları	Öğrenme katsayısı Momentum katsayısı Gizli katman	0,1 0,1 2	71,45	69,3	69,28	69,23
	Öğrenme katsayısı Momentum katsayısı Gizli katman	0,1 0,1 1	75,53	76,48	**74,64**	**75,51**
	Öğrenme katsayısı Momentum katsayısı Gizli katman	0,1 0,2 1	76,16	76,48	74,02	**75,51**
	Öğrenme katsayısı Momentum katsayısı Gizli katman	0,1 0,3 1	**76,77**	76,16	74,01	75,21
	Öğrenme katsayısı Momentum katsayısı Gizli katman	0,1 0,4 1	75,84	**76,77**	72,48	74,6
	Öğrenme katsayısı Momentum katsayısı Gizli katman	0,2 0,4 1	73,69	**76,77**	72,17	72,76
Bayesyen Sınıflandırma			71,79	71,8	72,42	71,81

Çizelge 5. 4 (devamı) : Test sonuçları.

			TEST0	**TEST1**	**TEST2**	**TEST3**
Veri Madenciliği Yöntemleri	**Parametre**	**Değer**	***Doğruluk(%)***	***Doğruluk(%)***	***Doğruluk(%)***	***Doğruluk(%)***
Karar Ağaçları	Kriter Min. bölünme Max. derinlik	Kazanım oranı 4 20	60,38	60,07	59,74	59,42
	Kriter Min. bölünme Max. derinlik	Gini indeksi 20 20	**73,67**	72,75	**73,34**	**73,35**
	Kriter Min. bölünme Max. derinlik	Gini indeksi 40 10	73,34	**73,34**	73,03	73,03
	Kriter Min. bölünme Max. derinlik	Gini indeksi 60 60	72,14	72,14	72,14	72,14
k-NN	Manhattan Uzalık Ölçümü	k=1	61,28	64,38	62,2	63,77
		k=3	63,76	**66,88**	63,76	**66,59**
		k=5	62,51	64,68	62,5	64,99
		k=7	**65,32**	66,22	64,7	66,22
		k=9	65,3	66,23	**65,62**	66,24
	Öklid Uzaklık Ölçümü	k=1	56,9	63,13	57,22	**63,45**
		k=3	56,89	61,89	56,89	61,59
		k=5	57,27	**63,47**	57,27	63,16
		k=7	**59,11**	63,14	**58,8**	63,14
		k=9	57,86	61,87	57,55	61,87

Sonuçlar göz önünde bulundurulduğunda yapay sinir ağları yöntemi en başarılı yöntem olarak değerlendirilmektedir. En yüksek doğruluk yüzdesini tüm testlerde yapay sinir ağları yöntemi vermiştir. Bu yöntem ile elde edilen en yüksek doğruluk oranları Test 0 ve Test 1 için %76,77 iken, Test 2 ve Test 3 için sırasıyla %74,64 ve %75,51'dir. Test 0 ve Test 1'de %76,77 ile aynı ve en yüksek doğruluk oranı elde edilmiş olsa da iki testteki öğrenme ve momentum katsayı değerleri ile gizli katmandaki düğüm (*node*) sayıları farklılık göstermektedir. Test 0 ve Test 1'e uygulanan YSA modelinin görüntüsü Şekil 5.7'de gösterilmektedir. Çizelge 5.4'te ayrıntıların görüldüğü gibi yapay sinir ağları yönteminde gizli katman sayısı arttığında düşük doğruluk oranı elde edilmiştir.



Şekil 5. 7: Yapay sinir ağları modeli.

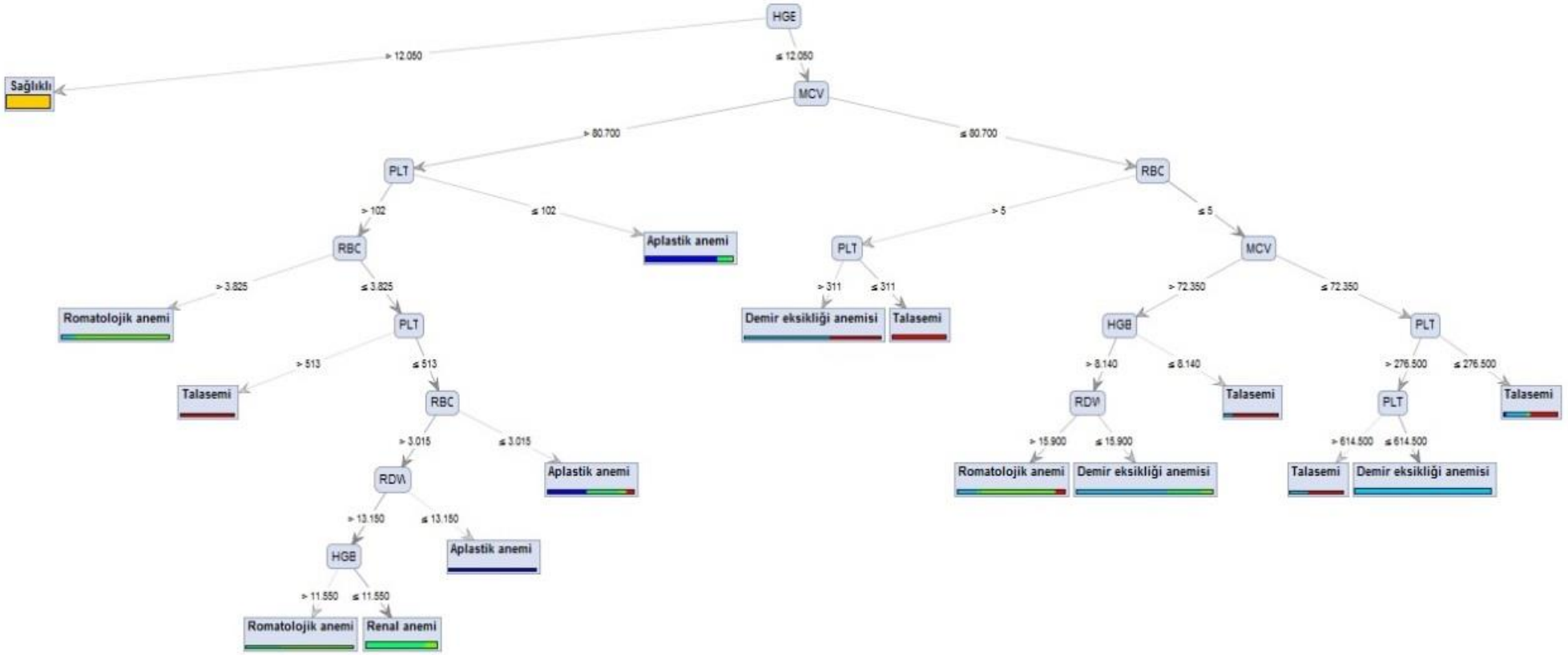
Test 0'a uygulanan yapay sinir ağları yönteminin sonuçlarını gösteren karışıklık matrisi Çizelge 5.5'te verilmiştir. Bu matrise göre sağlıklı bireylerden sonra en yüksek duyarlılık (*recall*) oranı talasemi sınıflandırmasında, en yüksek kesinlik (*precision*) oranı ise aplastik anemi sınıflandırmasında elde edilmiştir. F-ölçümü (*F-measure*) hesaplandığında en yüksek oran %76 ile aplastik anemi ve talasemi sınıflandırılmasında elde edilmiştir.

Çizelge 5. 5: Karışıklık matrisi.

Doğruluk oranı (accuracy): 76.77% +/- 5.98% (mikro: 76.78%)

Gerçek sınıf / *Tahmin edilen sınıf*	Aplastik anemi	Demir eksikliği anemisi	Renal anemi	Romatolojik anemi	Sağlıklı	Talasemi	**Kesinlik (precision)**
Aplastik anemi	25	0	4	1	0	0	83.33%
Demir eksikliği anemisi	1	32	0	9	0	8	64.00%
Renal anemi	9	2	30	8	0	2	58.82%
Romatolojik anemi	1	11	6	25	0	0	58.14%
Sağlıklı	0	0	2	0	102	0	98.08%
Talasemi	0	7	2	2	0	34	75.56%
Duyarlılık (recall)	69.44%	61.54%	68.18%	55.56%	100.00%	77.27%	

Karar ağacı yöntemi uygulanması ile en yüksek doğruluk oranı %73,67 olarak Test 0'da elde edilmiştir. Bu doğruluk oranın elde edildiği karar ağacı yapısı Şekil 5.8'de görülmektedir. Bu yöntem ile Test 1 ve Test 2'de %73,34, Test 3'te ise %73,35 doğruluk oranları elde edilmiştir. Çizelge 5.4'te görüldüğü gibi aynı parametre değerleri uygulandığında Test 1, Test 2 ve Test 3'ün doğruluk oranlarının birbirine bu kadar yakın olması ile bu yöntemde yaş ve cinsiyet parametrelerinin sınıflandırmada farklılık oluşturmadığı gözlemlenmiştir.



Şekil 5. 8: Karar ağacı yapısı.

Yapısı Şekil 5.8'de yapısı görülen karar ağacı sonucunda 4 kural elde edilmiştir. Bu kuralların adımları aşağıda gösterilmektedir:

Kural 1

HGB ≤ 12.050

| MCV > 80.700

| | PLT > 102

| | | RBC ≤ 3.825

| | | | PLT > 513: Talasemi {Aplastik anemi=0, Demir eksikliği anemisi=0, Renal anemi=0, Romatolojik anemi=0, Sağlıklı=0, Talasemi=7}

Kural 2

HGB ≤ 12.050

| MCV > 80.700

| | PLT > 102

| | | RBC ≤ 3.825

| | | | PLT ≤ 513

| | | | | RBC > 3.015

| | | | | | RDW ≤ 13.150: Aplastik anemi {Aplastik anemi=2, Demir eksikliği anemisi=0, Renal anemi=0, Romatolojik anemi=0, Sağlıklı=0, Talasemi=0}

Kural 3

HGB ≤ 12.050

| MCV ≤ 80.700

| | RBC ≤ 5

| | | MCV ≤ 72.350

| | | | PLT > 276.500

| | | | | PLT ≤ 614.500: Demir eksikliği anemisi {Aplastik anemi=0, Demir eksikliği anemisi=25, Renal anemi=0, Romatolojik anemi=0, Sağlıklı=0, Talasemi=0}

Kural 4

HGB ≤ 12.050

| MCV ≤ 80.700

| | RBC > 5

| | | PLT ≤ 311: Talasemi {Aplastik anemi=0, Demir eksikliği anemisi=0, Renal anemi=0, Romatolojik anemi=0, Sağlıklı=0, Talasemi=15}

Bayesyen sınıflandırma algoritması uygulanarak yapılan sınıflandırmada dört test için sırasıyla %71,79, %71,8, %72,42 ve %71,8 doğruluk oranları elde edilmiştir.

K-en yakın komşu yöntemi ile yapılan sınıflandırmada kısmen düşük doğruluk oranları elde edilmiş olup, dört test için elde edilen değerler sırası ile %65,32, %66,88, %65,62 ve %66,59'dur.

6. SONUÇ ve ÖNERİLER

Günümüzde artan veri hacmi ile verilerin toplanması ve işlenmesi güçleşmiştir. Veri madenciliği yüksek miktardaki verilerin işlenmesine ve anlamlı örüntüler elde edilmesine olanak sağlamaktadır. Bir çok farklı sektörde kullanılan veri madenciliği sağlık sektöründe de görüntüleme, teşhis koyma, tedavi gibi amaçlarla kullanılmaktadır.

Veri madenciliği tekniklerinden biri olan sınıflandırma tekniği, verinin önceden belirlenen çıktılara uygun olarak ayrıştırılmasını sağlayan bir tekniktir. Çıktılar, önceden bilindiği için sınıflama, veri kümesini denetimli olarak öğrenir. Çalışmamızda da veri setine farklı sınıflandırma yöntemleri uygulanarak, elde edilen doğruluk oranları karşılaştırılmıştır.

Çalışmamızda anemi hastalıklarının sınıflandırılması amacı ile Uludağ Üniversitesi Tıp Fakültesi Hematoloji Anabilim Dalı bünyesinde bir çok hasta verisi incelenerek, 323 hastanın kan parametre değerleri, cinsiyet ve yaş bilgileri derlenmiştir. Beş farklı çeşit anemiye sahip hastaların ve hematolojik açıdan sağlıklı bireylerin tam kan sayımı testi değerlerinin bulunduğu veriler, ön işleme aşamasından geçirilmiş ve girdilerinde farklılıklara gidilerek dört ayrı veri seti oluşturulmuştur. Bu veri setlerine veri madenciliği yazılımlarından olan RapidMiner programı kullanılarak Yapay Sinir Ağları, Karar Ağaçları, K-En Yakın Komşu ve Bayesyen Sınıflandırma yöntemleri uygulanmıştır.

Yapılan sınıflandırma sonucu tüm testlerde en yüksek doğruluk oranı yapay sinir ağları sınıflandırma yöntemi ile elde edilmiştir. Böylece yapay sinir ağları yöntemi ile; demir eksikliği anemisi, talasemi, aplastik anemi, böbrek hastalıklarına bağlı anemi ve romatizmal hastalıklara bağlı anemiye sahip hastalara ve hematolojik açıdan sağlıklı kişilere ait kan parametrelerinin bulunduğu veri setlerinde farklı öğrenme ve momentum katsayıları ile Test 0 ve Test 1'de en yüksek doğruluk oranı olan %76,77 elde edilmiştir.

Geliştirme çalışmaları devam etse de ülkemizde henüz genel bir hasta veri tabanı oluşturulmamıştır. Bu durum veri toplama sürecini zorlaştırmakta ve araştırmacıları ciddi bir karmaşanın içerisine sürüklemektedir. Tüm hastalara ait verilerin ulusal veya uluslararası standartlar gözetilerek tek bir havuzda toplanması, saklanması ve araştırmacılara erişiminin sağlanması ile sağlık sektöründe veri madenciliği uygulamalarının ülkemizdeki sayısı da artacaktır.

Diğer bir açıdan, tıbbi veriler üzerinde çalışma yapılırken bilgi gereksinimi ve uzman desteği ihtiyacı hat safhadadır. Bu durum disiplinler arası çalışma ile olumlu yönde değiştirilebildiği gibi çok daha verimli sonuçların elde edilmesine olanak sağlayabilmektedir. Ülkemizde sağlık sektöründe uygulanacak olan veri madenciliği çalışmalarında disiplinler arası çalışma gruplarının teşvik edilmesinin bu uzman desteği ihtiyacını karşılayacağı düşünülmektedir.

Gelecekte yapılacak daha uzun süreli ve kapsamlı çalışmalar ile daha yüksek doğruluk oranına sahip sonuçlar elde edilebilir. Verilerin standart bir düzende, tek bir havuzda olması ile erişiminin kolayca sağlanması böylece kısa sürede çok daha fazla sayıda verinin analiz edilmesi bu durumun iyileştirilmesine yardımcı olacaktır. Bu bağlamda, ileride yapılacak daha kapsamlı çalışmalar ile kanser gibi komplike hastalıkların erken teşhisine veya tedavi yöntemi karar aşamasına yardımcı olacak karar destek sistemleri geliştirilebilir.

KAYNAKLAR

Abdel-Aal, R. E. (2005). GMDH-based Feature Ranking and Selection for Improved Classification of Medical Data, *Journal of Biomedical Informatics*, Vol. 38, No. 6, s. 456-468.

Abdullah, A. S., Rajalaxmi, R. R. (2012). A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier, IJCA Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls (ICON3C 2012) ICON3C(3):22-25, April 2012.

Ahmed, A., Hannan, S. A. (2012). International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-1, Issue-4, 18-23.

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, *İÜ. İşletme Fakültesi Dergisi,* İstanbul.

Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, *Journal of King Saud University – Computer and Information Sciences*, Vol. 24, s. 27–40.

Argüden, Y. ve Erşahin, B. (2008). "Veri Madenciliği: Veriden Bilgiye, Masraftan Değere", *ARGe Danışmanlık Yayınları*, No:10, İstanbul.

Azarkish, I., Raoufy, M. R., Gharibzadeh, S. (2012). Artificial Intelligence Models for Predicting Iron Deficiency Anemia and Iron Serum Level Based on Accessible Laboratory Data, Journal of Medical Systems, 36:2057–2061.

Baykal, A. (2006). Veri madenciliği uygulama alanları, *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 7, 96.

Belazzi, R., Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines, International Journal of Medical Informatics, 77, 81-97.

Bellaachia, A., Guven, E. (2006). Predicting Breast Cancer Survivability Using Data Mining Techniques, Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining, April 22, Bethesda, USA.

Berry, M. J. A. ve Linoff, G. S. (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management Second Edition, *Wiley Publishing*, Indianapolis.

Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. (1999). "When Is 'Nearest Neighbor' Meaningful?", 7. International Database Theory Conference (ICDT'99), pp. 217- 235, Israel.

Bircan, H. (2004). Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama, Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2 : 185-208.

Chakrabarti, S. ve diğ. (2008). Data Mining Know It All, *Morgan Kaufmann Publishers*, Massachusetts.

Chang, C-L.,Chen, C-H. (2009). "Applying decision tree and neural network to increase quality of dermatologic diagnosis", *Expert Systems with Applications*, 36, 4035–4041.

Chen, H-Y., Chuang, C-H., Yang, Y-J., Wu, T. P. (2011). Exploring the risk factors of preterm birth using data mining, Expert Systems with Applications, 38: 5384–5387.

Cios, K.J. ve Moore, G.W. (2002). "Uniqueness of Medical Data Mining", *Artificial Intelligence in Medicine Journal*, 26(1-2):1-24.

Coşkun, C., Baykal, A. (2011). Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması, XIII. Akademik Bilişim Konferansı Bildirileri, Malatya.

CRISP-DM 1.0. (2010). Step-by-step data mining guide, Registered by SPSS Inc.

Danacı, M., Çelik, M. ve Akkaya, A.E. (2010). "Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi", *Akıllı Sistemlerde Yenilikler ve Uygulama Sempozyumu*, 21-24 Haz. 2010, Kayseri, 9-12.

Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease, Expert Systems with Applications, 37:1568–1572.

Delen, D., Sharda, R., Bessonov, M. (2006) Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks, Accident Analysis and Prevention, 38, 434-444.

Delen, D., Walker, G., Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods, Artificial Intelligence in Medicine, 34, 113-127.

Doğan, Ş. ve Türkoğlu, İ. (2008). "Iron-Deficiency Anemia Detection From Hematology Parameters By Using Decision Trees", *International Journal of Science & Technology*, Cilt 3, No. 1, 85-92.

Dönmez, Z. S. (2008). Bayi Performans Değerlendirmesinde Bir Veri Madenciliği Uygulaması, *Yüksek Lisans Tezi,* İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Dunham, M. H. (2003). Data Mining Introductory and Advanced Topics, *Pearson Education Inc.*, New Jersey.

Elshami, E.H. ve Alhalees, A.M. (2012). "Automated Diagnosis of Thalassemia Based on Data Mining Classifiers", *In Proceeding of The International Conference on Informatics & Applications (ICIA2012),* University Sultan Zainal Abidin, Malesia, June 3-5.

Garg, A.X. ve diğ. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes, *Journal of the American Medical Association*, PubMed 293 (10), 1223–1238.

Giannopoulou, E. G. (2008). Data Mining in Medical and Biological Research, *In-Teh Press*, Croatia.

Gören, S., Karahoca, A., Onat, F. Y., Gören, M. Z. (2008). Prediction of cyclosporine A blood levels: an application of the adaptive-network-based fuzzy inference system (ANFIS) in assisting drug therapy, Eur J Clin Pharmacol, 64:807–814.

Gregori, D., Petrinco, M., Bo, S., Rosato, R., Pagano, E., Berchialla, P., Merletti, F. (2011). Using Data Mining Techniques in Monitoring Diabetes Care. The Simpler the Better?, Journal of Medical System, 35:277–281.

Han, J. ve Kamber, M. (2006). Data Mining Concepts and Techniques 2nd Ed., *Morgan Kaufmann Publishers*, Massachusetts.

Hand, D., Mannila, H. ve Smyth, P. (2001). Principles of Data Mining [Elektronik Sürüm], *The Mit Press*, England.

Hu, R. (2010). Medical Data Mining Based on Association Rules, *Computer and Information Science*, Vol. 3, No. 4, s. 104-108.

İşler, Y., Avcu, N., Kocaoğlu, A., Kuntalp, M. (2008). Investigating the Effects of the Methods Used in Frequency Domain Measures in Heart Rate Variability Analysis, IEEE 16th Signal Processing and Communications Applications Conference, 20-22 April, Aydin, Turkey.

İşler, Y., Kuntalp, M. (2007). Combining classical HRV indices with wavelet entropy measures improves to performance in diagnosing congestive heart failure, Computers in Biology and Medicine, 37, 1502-1510.

İşler, Y., Narin, A. (2012). WEKA Yazılımında k-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi, SDU Teknik Bilimler Dergisi, Cilt:2, Sayı:4, 21-29.

Jonsdottir, T., Hvannberg, E. T., Sigurdsson, H., Sigurdsson, S. (2008). The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining, Expert Systems with Applications, 34, 108-118.

Jovic, A., Bogunovic, N. (2011). Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features, Artificial Intelligence in Medicine, 51, 175-186.

Kahramanli, H., Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases, Expert Systems with Applications, 35, 82-789.

Karabatak, M., Ince, M.C. (2009). An expert system for detection of breast cancer based on association rules and neural network, Expert Systems with Applications, 36: 3465–3469.

Karahoca, A., Kucur, T., Aydin, N. (2007). Data Mining Usage in Emboli Detection, 2007 ECSIS Symposium on Bio-inspired, Learning, and Intelligent Systems for Security, August 9-10, 2007 Edinburgh, UK.

Karaolis M. A., Moutiris J. A., Hadjipanayi D. ve Pattichis C. S. (2010). Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees, *IEEE Transactions On Information Technology In Biomedicine*, Vol. 14, No. 3.

Kaya, E., Bulun, M. ve Arslan, A. (2003). Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları, *Akademik Bilişim Konferansları*, Adana, 3-5 Şubat 2003.

Khan, T., Westin, J., Dougherty, M. (2014). Classification of speech intelligibility in Parkinson's disease, Biocybernetics and Biomedicalengineering, 34, 35-45.

Kınıklı, G. (2011). Romatoid Artrit. Ed: Düzgün N. Sf. 124-140. Ankara.

Kiranyaz, S., Ince, T., Pulkkinen, J., Gabbouj, M. (2011). Personalized long-term ECG classification: A systematic approach, Expert Systems with Applications, 38, 3220-3226.

Knott, L. (2014). Anaemia in Chronic Kidney Disease, Egton Medical Information Systems (EMIS) Group, 4 pages.

Koç, E., Atılgan Şengül, Y., Özkaya Uyar, A., Gökçe, B. (2012). Klinik Karar Destek Sistemleri Kullanımına Yönelik Bir Araştırma: Acıbadem Hastanesi Örneği, IX. Ulusal Tıp Bilişimi Kongresi, 15-17 Kasım 2012, Antalya.

Koyuncugil, A.S., Özgülbaş, N. (2009). Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları, Bilişim Teknolojileri Dergisi, Cilt: 2, Sayı: 2.

Köktürk, F. (2012). "K-En Yakın Komşuluk, Yapay Sinir Ağları Ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması", *Doktora Tezi*, Zonguldak Karaelmas Üniversitesi Fen Bilimleri Enstitüsü, Zonguldak.

Kumar, D.S., Sathyadevi, G. ve Sivanesh, S. (2011). Decision Support System for Medical Diagnosis Using Data Mining, *International Journal of Computer Science Issues*.

Larose, D.T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining, *Wiley Publishing*, USA.

Lee, T-T., Liu, C-Y., Kuo, Y-H., Mills, M. E., Fong, J-G., Hung, C. (2011). Application of data mining to the identification of critical factors in patient falls using a web-based reporting system, International Journal of Medical Informatics, 80, 141-150.

Li, L., Tang, H., Wu, Z.,Gong, J., Gruidl, M., Zou, J., Tockman, M., Clark, R. A. (2004). Data mining techniques for cancer detection using serum proteomic profiling, Artificial Intelligence in Medicine, 32:71-83.

Mehta, M., Agrawal, R., ve Rissanen, J. (1996). SLIQ: A Fast Scalable Classifier for Data Mining, 5. International Extending Database Technology Conference, Avignon, France.

Mendonça, E.A. (2004). Clinical Decision Support Systems: Perspectives in Dentistry, *Journal of Dental Education*, 589-597.

Merijohn, G. K., Bader, J. D., Frantsve-Hawley, J. ve Aravamudhan, K. (2008). "Clinical decision support chairside tools for evidence-based dental practice", *The Journal of Evidence-Based Dental Practice*, 8 (3), 119–132.

Niemann, U., Völzke, H., Kühn, J-P., Spiliopoulou, M. (2014). Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis, Expert Systems with Applications, 41, 5405-5415.

Omiotek, Z., Burda, A., Wójcik, W. (2013). The use of decision tree induction and artificial neural networks for automatic diagnosis of Hashimoto's disease, Expert Systems with Applications, 40:6684–6689.

Oztekin, A., Delen, D., Kong, Z. J. (2009). Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology, International Journal of Medical Informatics, 78, e84-e96.

Oztekin, A., Kong, Z.J., Delen, D. (2011). Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations, Decision Support Systems, 51, 155-166.

Özkan, Y. (2008). Veri Madenciliği Yöntemleri, *Papatya Yayıncılık*, İstanbul.

Park, S. C., Selwyn, P. ve Shaw, M. J. (2001). Dynamic Rule Refinement in Knowledge-Based Data Mining Systems, *Decision Support Systems*, 31:205-222.

Parshutin, S., Kirshners, A. (2013). Research on clinical decision support systems development for atrophic gastritis screening, Expert Systems with Applications, 40:6041–6046.

Parthiban, L. ve Subramanian, R. (2008). "Intelligent heart disease prediction system using CANFIS and genetic algorithm", *International Journal of Biological, Biomedical and Medical Sciences*, 3, 3.

Patil, B.M., Joshi, R.C., Toshniwal, D. (2010). Hybrid prediction model for Type-2 diabetic patients, Expert Systems with Applications, 37, 8102-8108.

Patil, M.B., Joshi, R.C., Toshniwal, D. ve Biradar, S. (2011). "A New Approach: Role of Data Mining in Prediction of Survival of Burn Patients", *Journal of Medical Systems*, Springer.

Quinlan, J.R. (1986). Induction of Decision Trees, *Journal of Machine Learning*, Vol. 1, pp.81-106.

Quinlan, J.R. (1993). C4.5: Programs For Machine Learning, *Morgan Kaufmann Publishers*, Massachusetts.

Raymond, B. ve Dold, C.(2002). Clinical Information Systems: Achieving the Vision, (Report) Kaiser Permanente Institute for Health Policy, Oakland.

Saichanma, S., Chulsomlee, S., Thangrua, N., Pongsuchart, P., Sanmun, D. (2014). The Observation Report of Red Blood Cell Morphology in Thailand Teenager by Using Data Mining Technique, Hindawi Publishing Corporation Advances in Hematology, Volume 2014, Article ID 493706, 5 pages.

Sanap, S.A., Nagori, M. ve Kshirsagar, V. (2011). "Classification of Anemia Using Data Mining Techniques", In proceeding of: Swarm, Evolutionary, and Memetic Computing - Second International Conference, SEMCCO 2011, India, December 19-21**.**

Sarvestani, A. S., Safavi, A. A., Parandeh, N. M., Salehi, M. (2010). Predicting Breast Cancer Survivability Using Data Mining Techniques, 2010 2nd International Conference on Software Technology and Engineering(ICSTE), 03 Oct - 05 Oct 2010, San Juan, PR, USA.

Shafer, J., Agrawal, R. ve Mehta, M. (1996). SPRINT: A Scalable Parallel Classifier for Data Mining, 22. International Conference on Very Large Databases, Mumbai (Bombay), India.

Shannon, C.E. (1948). 1A Mathematical Theory of Communication", *The Bell System Technical Journal*, Vol. 27, pp. 379-423,623-656.

Shukla, V., Barot, M. (2013). Optimized classification method for prognosis of breast cancer using RapidMiner, International Journal of Artificial and Knowledge Discovery, 3:14-17.

Silahtaroğlu, G. (2013). Veri Madenciliği Kavram ve Algoritmaları 2. Basım, *Papatya Yayıncılık*, İstanbul.

Soni, J., Ansari, U., Sharma, D. ve Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, *International Journal of Computer Applications*, Volume 17– No.8. (0975 – 8887).

Srinivas K., Rao G.R. ve Govardhan A. (2010). Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques, The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 1344.

Srinivas, K., Rani, B. K., Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.

Sund, R. (2002). "Utilization of Administrative Registers using Statistical Knowledge Discovery", International Workshop on "Mining Official Data", National Research and Development Centre for Welfare and Health, Helsinki- Finland.

Şengül, Ş., Ertürk, Ş. (2008). Renal Anemia, Turkiye Klinikleri J Nephrol-Special Topics, 1(2):18-23.

Tan, P.N., Steinbach, M. ve Kumar, V. (2006). Introduction to Data Mining, *Pearson Education Inc.*, Boston.

Taşdelen, B., Helvacı, S., Kaleağası, H., Özge, A. (2009). Artificial Neural Network Analysis for Prediction of Headache Prognosis in Elderly Patients, Journal of Medical System, 39 (1): 5-12.

THD Ulusal Tedavi Kılavuzu. (2011). Türk Hematoloji Derneği.

Tsiptsis, K. ve Chorianopoulos, A. (2010). Data Mining Techniques in CRM: Inside Customer Segmentation, *Wiley Publishing*, United Kingdom.

Turban, E., Sharda, R. ve Delen, D. (2011). Decision Support and Business Intelligence Systems, *Pearson Education Inc.*, New Jersey.

Turgay, M. (2011). Romatolojik Hastalıkların Tanısında Kullanılan Laboratuvar Testleri. Ed: Düzgün N. Sf. 28-37. Ankara.

Two Crows Coorporation. (2005). Introduction to Data Mining and Knowledge Discovery, Third Edition, Two Crows Corporation, Maryland.

Uçar, T., Karahoca, A. (2011). Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches, Procedia Computer Science, 3, 1404-1411.

Uçar, T., Karahoca, A., Karahoca, D. (2013). Tuberculosis disease diagnosis by using adaptive neuro fuzzy inference system and rough sets, Neural Computing & Application, 23:471–483.

Uğur, A. ve Kınacı, A.C. (2008). Yapay Zeka Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması, Inet-tr 2006, XI. Türkiye'de İnternet Konferansı, TOBB Ekonomi ve Teknoloji Üniversitesi, Ankara, 21-23 Aralık 2006.

Uzoka, F-M. E., Osuji, J. ve Obot, O. (2011). Clinical decision support system (DSS) in the diagnosis of malaria: A case comparison of two soft computing methodologies, *Expert Systems with Applications*, 38, 1537-1553.

Wojciechowski, M. (2003). Supporting Interactive Sequential Pattern Discovery in Databases, Emerging Database Research in East Europe, VLDB 2003 Workshop, Berlin, Germany.

Wongseree, W., Chaiyaratana, N., Vichittumaros, K., Winighagoon, P., Fucharoen, S. (2007). Thalassaemia classification by neural networks and genetic programming, Information Sciences, 177, 771-786.

Yeh D-Y., Cheng C-H. ve Chen, Y-W. (2011). "A predictive model for cerebrovascular disease using data mining", *Expert Systems with Applications*, 38, 8970–8977.

Yılmaz, Z. Ve Bozkurt, M.R. (2012). "Determination of Women Iron Deficiency Anemia Using Neural Networks", *Journal of Medical Systems*, Vol. 36, Issue 5, pp. 2941.

Yurtay, Y., Ak, G. ve Bacınoğlu, N.Z. (2013). "Tıbbi Karar Destek Sistemlerinin Yöntemsel Olarak Değerlendirilmesi Üzerine Bir Çalışma", ISITES2013 International Symposium On Innovate Technologies In Engineering Science, Sakarya Üniversitesi, 7-9 Haziran 2013.

Zhou, X., Zhang, Y., Shi, M., Shi, H., Zheng, Z. (2014). Early detection of liver disease using data visualisation and classification method, Biomedical Signal Processing and Control, 11, 27-35.

Url-1 <*http://rapidminer.com/about-us*>, Erişim Tarihi: 15.05.2014.

Url-2 <*http://www.masimo.fr/hemoglobin/anemia.htm*>, Erişim Tarihi: 26.03.2014.

Url-3 <*http://www.medindia.net/patients/patientinfo/iron-deficiency-anemia.htm*>, Erişim Tarihi: 26.03.2014.

Url-4 <*http://www.dihed.org.tr/index.php/hemodiyaliz/17-8-anemi*>, Erişim Tarihi: 26.03.2014.

EKLER

EK A: ULUDAĞ ÜNİVERSİTESİ TIP FAKÜLTESİ ETİK KURUL ONAY YAZISI

T.C.
ULUDAĞ ÜNİVERSİTESİ
Tıp Fakültesi Klinik Araştırmalar Etik Kurulu

Sayı : B.30.2.ULU.0.20.70.02-050.99/148
Konu : Etik Kurul kararı

28 Nisan/2014

Sayın Yrd.Doç.Dr.Başar ÖZTAYŞİ
İstanbul Teknik Üniversitesi İşletme Fakültesi
Endüstri Mühendisliği Bölümü Öğretim Üyesi

Kurulumuza başvurusunu yaptığınız ve sorumlu araştırmacısı olduğunuz "Veri madenciliği yöntemleri kullanarak anemi sınıflandırmasına yönelik bir uygulama" konulu araştırmanıza ilişkin Kurulumuzun 15 Nisan 2014 tarih ve 2014-8/3 nolu kararı ekte gönderilmektedir.

Gereği için bilgilerinize sunulur.

Prof.Dr.Mine Sibel GÜRÜN
Kurul Başkan

EK:
- Karar (1 adet)

Uludağ Üniversitesi Tıp Fakültesi Dekanlığı Rektörlük Binası, Görükle Kampusü 16059 Nilüfer/BURSA
Tel: 0-224-2950020 Fax: 0-224-2950029
e-posta: uukaek@uludag.edu.tr Elektronik Ağ: www.tip.uludag.edu.tr

ULUDAĞ ÜNİVERSİTESİ TIP FAKÜLTESİ İLAÇ DIŞI KLİNİK ARAŞTIRMALARI ETİK KURULU KARAR FORMU

BAŞVURU BİLGİLERİ	ARAŞTIRMANIN ADI	Veri madenciliği yöntemleri kullanarak anemi sınıflandırmasına yönelik bir uygulama
	SORUMLU ARAŞTIRMACI UNVANI/ ADISOYADI	Yrd.Doç.Dr.Başar Öztayşi
	SORUMLU ARAŞTIRMACININ BULUNDUĞU MERKEZ	İTÜ İşletme Fakültesi Endüstri Mühendisliği Bölümü
	YARDIMCI ARAŞTIRMACI	Arş.Gör.Betül Merve Fakı Bursa Teknik Üniv.Doğa Bilimleri, Mimarlık ve Müh.Fak. Prof.Dr.Rıdvan Ali UÜ.Tıp Fakültesi İç Hast.AD Hematoloji BD
	DESTEKLEYİCİ	-
	ARAŞTIRMANIN NİTELİĞİ	Dosya, görüntü kayıtları ve/veya insanlardan elde edilen arşiv materyalleri kullanılarak yapılan retrospektif araştırma
	ARAŞTIRMANIN YAPILIŞ AMACI	Yüksek lisans tez çalışması
	ARAŞTIRMANIN TAHMİNİ SÜRESİ İNCELENECEK DOSYA SAYISI	3 ay 500

	Belge Adı	Tarihi	Dili
DEĞERLENDİRİLEN İLGİLİ BELGELER	Araştırma ilk başvuru ön yazısı	25.03.2014	Türkçe
	GİRİŞİMSEL OYMAYAN ARAŞTIRMALAR İÇİN BAŞVURU FORMU	03.04.2014	Türkçe
	BİLGİLENDİRİLMİŞ GÖNÜLLÜ OLUR FORMU	20.03.2014	Türkçe
	ARAŞTIRICILAR İÇİN TAAHHÜTNAME FORMU	20.03.2014	Türkçe

KARAR BİLGİLERİ

Karar No : 2014-8/3 **Tarih : 15 Nisan 2014**

İstanbul Teknik Üniversitesi İşletme Fakültesi Endüstri Mühendisliği Bölümü Öğretim Üyesi Yrd.Doç.Dr.Başar Öztayşi'nin sorumluluğunda yürütülmesi planlanan, yukarıda başvuru bilgileri verilen araştırma başvurusu dosyası ve ilgili belgeler, araştırmanın gerekçe, amaç, yaklaşım ve yöntemleri dikkate alınarak incelenmesi sonucunda;

1- Araştırmanın yapılmasının uygun olduğuna,

2-Araştırmanın başlama tarihinin bildirilmesi ve araştırma tamamlandığında özet bir sonuç raporunun hazırlanarak kurulumuza iletilmesine,

3- Araştırma protokolünde ve başvuru formunda yapılacak tüm değişiklikler için Etik Kuruldan izin alınması gerektiğinin sorumlu araştırıcılara iletilmesine oybirliği ile karar verildi.

ULUDAĞ ÜNİVERSİTESİ TIP FAKÜLTESİ KLİNİK ARAŞTIRMALAR ETİK KURULU

ÇALIŞMA ESASI	Klinik Araştırmalar Hakkında Yönetmelik, İyi Klinik Uygulamalar Kılavuzu
BAŞKANIN UNVANI/ADI SOYADI	Prof.Dr.Mine Sibel GÜRÜN

ÜYELER

Unvanı / Adı / Soyadı EK Üyeliği	Uzmanlık Alanı	Kurumu	Cinsiyeti	Araştırma ile İlişki	Katılım (*)	İmza
Prof. Dr. Mine Sibel GÜRÜN Başkan	Farmakoloji	U.Ü.T.F. Farmakoloji ve Klinik Farmakoloji AD.	K	☐E ■H	■E ☐H	
Prof.Dr.Mustafa HACIMUSTAFAOĞLU BaşkanYardımcısı	Çocuk Sağlığı ve Hastalıkları	U.Ü.T.F. Çocuk Sağlığı ve Hastalıkları AD.	E	☐E ■H	■E ☐H	
Prof.Dr.Necdet KARLI Üye	Nöroloji	U.Ü.T.F. Nöroloji AD.	E	☐E ■H	■E ☐H	
Prof.Dr.Elif BAŞAĞAN MOĞOL Üye	Anesteziyoloji	U.Ü.T.F. Anesteziyoloji ve Reanimasyon AD.	K	☐E ■H	■E ☐H	
Doç.Dr.Emel İRGİL Üye	Halk Sağlığı	U.Ü.T.F. Halk Sağlığı AD.	K	☐E ■H	■E ☐H	

Sayfa 1

ULUDAĞ ÜNİVERSİTESİ TIP FAKÜLTESİ İLAÇ DIŞI KLİNİK ARAŞTIRMALARI ETİK KURULU KARAR FORMU

Doç.Dr.Mehmet CANSEV Üye	Farmakoloji	U.Ü.T.F. Tıbbi Farmakoloji AD	E	☐ E ■ H	■ E ☐ H	
Yrd.Doç.Dr.Tuna GÜLTEN Üye	Tıbbi Genetik	U.Ü.T.F. Tıbbi Genetik AD.	K	☐ E ■ H	■ E ☐ H	
Yrd.Doç.Dr.Pınar VURAL Üye	Psikiyatri	U.Ü.T.F. Çocuk ve Ergen Ruh Sağlığı ve Hastalıkları AD.	K	☐ E ■ H	■ E ☐ H	
Yrd.Doç.Dr.Çiğdem Mine YILMAZ Üye	Hukuk	U.Ü.Hukuk Fakültesi	K	☐ E ■ H	☐ E ■ H	Görevli
Yrd.Doç.Dr.Engin SAĞDİLEK Raportör	Biyofizik	U.Ü.T.F. Biyofizik AD.	E	☐ E ■ H	■ E ☐ H	
Yrd.Doç.Dr.Sezer ERER Üye	Tıp Tarihi ve Etik	UÜ.T.F. Tıp Tarihi ve Etik AD.	K	☐ E ■ H	■ E ☐ H	
Uz.Dr.Serhat YALÇINKAYA Üye	Göğüs Cerrahisi	Bursa Yüksek İhtisas EAH Göğüs Cerrahisi Kliniği	E	☐ E ■ H	☐ E ■ H	Görevli
Uz.Dr.Kağan HUYSAL Üye	Biyokimya	Bursa Yüksek İhtisas EAH Biyokimya	E	☐ E ■ H	☐ E ■ H	Görevli
Ecz.Zeynep Gözde SÖZER Üye	Eczacı	UÜ.SUAM	K	☐ E ■ H	■ E ☐ H	
Ahmet GÖREN Üye	Sağlık mesleği mensubu olmayan üye	Serbet Meslek	E	☐ E ■ H	■ E ☐ H	

: Toplantıda Bulunma

ÖZGEÇMİŞ

Ad Soyad: Betül Merve FAKI

Doğum Yeri ve Tarihi: Bursa / 03.01.1989

Adres: Beşevler Mah. Yıldırım Cad. No:309/A D:7 Nilüfer/BURSA

E-Posta: merve.faki@gmail.com

Lisans: İstanbul Teknik Üniversitesi – İşletme Mühendisliği (2011)

Yüksek Lisans: İstanbul Teknik Üniversitesi – Mühendislik Yönetimi (devam)

Mesleki Deneyim:

Pfizer İlaçları – Marketing Part-time
(12.2011 – 03.2013)
Avea - Uzman
(03.2013 – 08.2013)
Bursa Teknik Üniversitesi – Araştırma Görevlisi
(02.2014 – devam)