
GAZİ ÜNİVERSİTESİ
Gazi
1926

AKUT PANKREATİT HASTALARININ MORTALİTE RİSKLERİNİN KARAR AĞACI YÖNTEMİ İLE BELİRLENMESİ

Zeynep Hilal GÖKÇEN ALIÇ

YÜKSEK LİSANS TEZİ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

HAZİRAN 2014

Zeynep Hilal ALIÇ tarafından hazırlanan "AKUT PANKREATİT HASTALARININ MORTALİTE RİSKLERİNİN KARAR AĞACI YÖNTEMİ İLE BELİRLENMESİ" adlı tez çalışması aşağıdaki juri atrafından OY BİRLİĞİ / OY ÇOKLUĞU ile Gazi Üniversitesi Yönetim Bilişim Sistemleri Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman : Prof. Dr. İnan GÜLER

Elektronik-Bilgisayar Eğitimi, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum

Başkan : Doç.Dr. Diyar AKAY

Endüstri Mühendisliği, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum

Üye : Doç.Dr. Hasan Şakir BİLGE

Bilgisayar Mühendisliği, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum

Tez Savunma Tarihi: 27/06/2014

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

Doç.Dr. Nurettin TOPALOĞLU

Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Zeynep Hilal GÖKÇEN ALIÇ
27/06/2014

AKUT PANKREATİT HASTALARININ MORTALİTE RİSKLERİNİN KARAR AĞACI YÖNTEMİ İLE BELİRLENMESİ

(Yüksek Lisans Tezi)

Zeynep Hilal GÖKÇEN ALIÇ

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

Haziran 2014

ÖZET

Akut pankreatit; pankreasın ani şekilde ortaya çıkan iltihabıdır, pankreas hem iç salgı bezi hem de dış salgı bezi olarak görev yapar ve pankreasın salgıladığı enzimler pankreasta iken inaktif haldedir. Yağların, proteinlerin, karbonhidratların sindirimini sağlayan bu enzimler, akut pankreatit hastalarında daha pankreastayken aktif hale geçer ve dokuların parçalanmasına yol açar. Bu tezde, Ankara'da bir kamu hastanesindeki 206 akut pankreatit hastasına ait veriler, IBM PASW (Predictive Analytics Software) Modeler 14.0'da analiz edilerek, hastaların mortalite (ölüm) riskleri tespit edilmeye çalışılmıştır. Pek çok sınıflandırıcı karar ağacı yöntemi denenmiş, en iyi performans C5.0 karar ağacı yöntemiyle sağlanmıştır. Elde edilen karar ağacı kuralları, hastalara uygulanacak tedavi yöntemlerinin belirlenmesi ve doğru tedavinin hızlı bir şekilde öngörülmesi bakımından hekimlere önemli karar desteği sağlayabilecektir.

Bilim Kodu : 902.1.071

Anahtar Kelime : Akut Pankreatit, veri madenciliği, karar ağacı

Sayfa Adedi : 69

Danışman : Prof. Dr. İnan GÜLER

DETERMINATION OF MORTALITE RISKS OF ACUTE PANCREATITIS PATIENTS BY USING DECISION TREE METHOD

(M. Sc. Thesis)

Zeynep Hilal GÖKÇEN ALIÇ

GAZİ UNIVERSITY

INFORMATICS INSTITUE

June 2014

ABSTRACT

Acute pancreatitis is an inflammation of the pancreas that arises suddenly. Pancreas serves as both endocrine and exocrine glands and while in the pancreas that secrete pancreatic enzymes are inactivated form. These enzymes enables the digestion of fats, proteins and carbohydrates. In patients with acute pancreatitis these enzymes are activated while in the pancreas and lead to breakdown of tissues. In this thesis, 206 data which are belong to patients with acute pancreatitis obtained from a public hospital in Ankara, are analysed by using IBM PASW (Predictive Analytics Software) Modeler 14.0 and mortality risks of the patients are tried to be predicted. The best performance among the decision tree methods which are used is obtained with the C5.0 method. The resulting decision tree rules will be able to provide an important decision support to physicians in terms of determining the treatment methods to be applied to patient and predicting the proper treatments quickly.

Science Code : 902.1.071

Key Words : Acute pancreatitis, data mining, decision trees

Page Number : 69

Supervisor : Prof. Dr. İnan GÜLER

TEŞEKKÜR

Çalışmalarım süresince yardım ve katkılarıyla beni yönlendiren değerli Hocam Prof. Dr. İnan GÜLER'e, yine benden desteğini esirgemeyen babam Prof. Dr. Hadi GÖKÇEN'e ve Doç.Dr. Diyar AKAY hocama, son olarak bu süreçte beni yalnız bırakmayan sevgili eşime sonsuz şükranlarımı sunarım.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklama**
S	Hedef sınıf
A	Nitelik
DP	Doğru pozitif örnek sayısı
DN	Doğru negatif örnek sayısı
YP	Yanlış pozitif örnek sayısı
YN	Yanlış negatif örnek sayısı
P	Pozitif toplam örnek sayısı
N	N negatif toplam örnek sayısı
M	Model

Kısaltmalar	**Açıklama**
SQL	Structured Query Language-Yapısal Sorgulama Dili
CART	Classification And Regression Trees
PAM	Partitioning Around Medoids
YSA	Yapay Sinir Ağları
GA	Genetik Algoritma
SVM	Supported Vector Machines-Destek vektör makinesi
AP	Akut Pankreatit
PASW	Predictive Analytics Software
WBC	White Blood Cell-Beyaz kan hücrelerinin(lökosit)-sayısı

Kısaltmalar	Açıklama
HCT	Hematokrit
NE	Nötrofil
PLT	Platelet-Trombosit sayısı
RDW	Redcell Distribution Width
MPV	Mean Platelet Volüme/ortalama trombosit hacmi
BIL	Total Bilirubin
LDH	Laktatdehidrogenaz
AST	Aspartataminotransferaz
ALT	Alaninaminotransferaz
GLU	Glukoz/kan şekeri
BUN	Blood Urea Nitrogen/kan üre azotu
ALB	Albümin
CA	Kalsiyum
AİK	Alıcı İşletim Karakteristiği

1. GİRİŞ

Büyük miktarlarda tutulan verilerden, aslında var olan ancak farkedilemeyen gizli desenlerin/anlamların çıkarılması son zamanlarda odaklanılan konuların başlarında gelmektedir. Veri madenciliği adı verilen bu alan, büyük miktarda veri içinden gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılması ve bu bilgiler üzerinden gelecekle ilgili tahmin yapılmasını sağlayacak bağıntı ve kuralların aranması süreci olarak adlandırılabilir.

Günümüzde en fazla bilgi birikiminin yaşandığı alanlardan biri olan tıp alanındaki bilgilerden faydalanılarak önemli bilgiler elde etmek mümkündür. Veri madenciliği, sağlık ve tıp alanındaki büyük veri tabanlarından değerli bilgileri ortaya çıkartarak, hem tıp hem de hizmet kalitesinin artırılması açısından büyük katkılar sağlamaktadır. Söz konusu insan sağlığı olduğu için bu alandaki veri madenciliği çalışmaları önemli bir uygulama alanı bulacaktır. Bu konudaki önemi son yıllarda giderek artan çalışmalar ile ortaya koyulmaya başlamıştır [1].

Bu tezde uygulama olarak ele alınan tıbbi alan Akut pankreatit hastalığıdır. Akut pankreatit pankreas hastalıklarının geniş bir bölümünü oluşturmaktadır. Hastalık klinik olarak hafif abdominal ağrıdan, sıvı sekestrasyonu, hipotansiyon, metabolik bozukluklar sepsis, çoklu organ yetmezliği ve ölümle sonuçlanan ağır formlara kadar değişen bir yelpaze içermektedir. Klinik olarak hafif-geçici tip ve hızlı seyreden-fatal tip olmak üzere iki şekilde karşımıza çıkabilmektedir. Hastaların %90'ı hafif ya da orta şiddette hastalıkla karşı karşıya kalırlarken, %10 hasta, şiddetli forma tutulmaktadır [2].

Çalışmada, Akut pankreatit hastalarının mortalite oranlarının belirlenmesine yönelik olarak karar ağacı yöntemi kullanılarak bir tıbbi veri madenciliği çalışması gerçekleştirilmiş ve elde edilen karar kuralları tartışılmıştır. Elde edilen karar kuralları, hastalara uygulanacak tedavi yöntemlerinin belirlenmesi ve doğru tedavinin hızlı bir şekilde öngörülmesi bakımından hekimlere önemli karar desteği sağlayabilecektir.

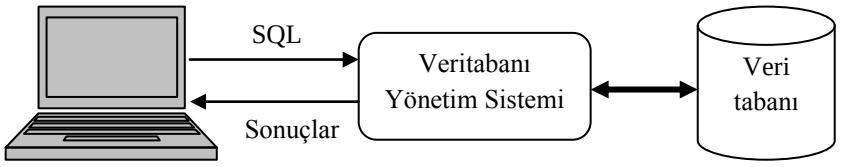
Tezin ikinci bölümünde, veri madenciliği ve veri madenciliğinde kullanılan popüler yöntemler detaylı olarak ele alınmış, örneklerle açıklanmıştır. Üçüncü bölümünde tıpta veri madenciliği konusunun önemi tartışılmış ve tıbbi veri madenciliğine dair detaylı bir literatür çalışması sunulmuştur. Dördüncü bölüm, çalışmanın uygulama bölümüdür. Ankara'da bir kamu hastanesindeki 206 adet Akut Pankreatit (AP) hastasına ait veriler düzenlenmiş ve IBM PASW (Predictive Analytics Software) Modeler 14.0 kullanılarak hastalığın mortalite riskinin belirlenmesi hedeflenmiştir. Bu kapsamda, hastalık risk sınıflandırma probleminin çözümünde en iyi performansı veren sınıflandırıcı algoritması C5.0 olarak belirlenmiş ve karar ağacı oluşturularak karar ağacından elde edilen karar kuralları tartışılmıştır. Tezin beşinci bölümünde ise sonuç ve değerlendirmeler yer almaktadır.

2. VERİ MADENCİLİĞİ

Bilgi teknolojilerindeki gelişmeler, bu çerçevede kuruluşların bilgi sistemlerini kurma gereklilikleri, çok fazla verinin elde edilmesine ve depolanmasına neden olmuştur. Söz konusu verilerdeki gizli örüntülerin (pattern) keşfedilmesi, onlardan anlamlı ve faydalı bilgilerin elde edilerek karar mekanizmalarına sunulması gerekliliği veri madenciliğinin önemli bir araştırma alanı olmasına neden olmuştur.

Veri madenciliği son yıllarda, toplumda ve bilgi endüstrisinde büyük bir ilgi odağı haline gelmiştir. Büyük miktarlardaki verinin mevcudiyeti ve bunların faydalı bilgi (information) ve üstbilgiye (knowledge) dönüştürülmesi gereklilik arz etmiştir. Veri madenciliği bilgi teknolojisi evriminin doğal bir sonucu olarak görülebilir.

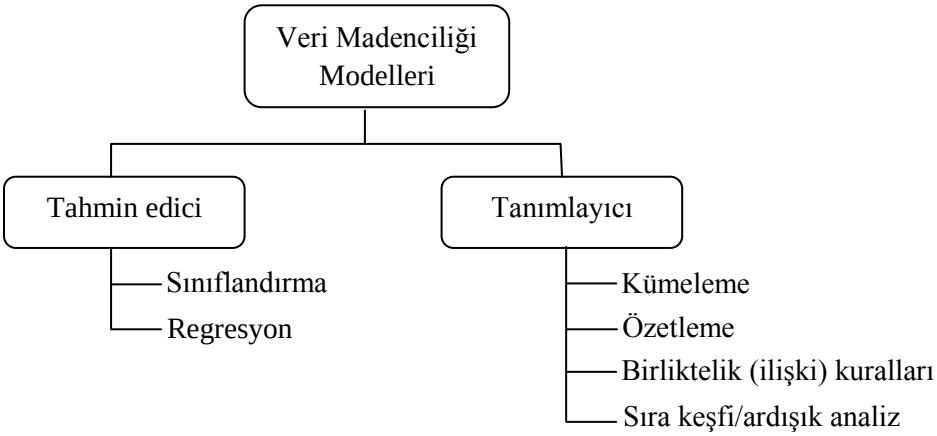
Geleneksel veritabanı sorgularında, örneğin SQL gibi bir sorgulama diliyle iyi-tanımlı (well-defined) bir sorgu kullanılarak veritabanına erişilebilir. Sorgunun çıktısı, veritabanından elde edilen ve sorguyu sağlayan verilerden oluşacaktır. Yani çıktı, genellikle veritabanının bir alt kümesidir, çıkarılmış bir görüntü ya da birleştirilmiş durumları içerebilmektedir. Şekil 2.1' de, veritabanı erişimi verilmektedir.



Şekil 2.1. Veritabanı erişimi [3]

Bir veritabanına veri madenciliği erişimi ise, bu geleneksel erişimden farklılıklar arz etmektedir; (a) Veri madenciliği sorgusu iyi biçimlendirilmemiş ya da tam ifade edilmemiş olabilir. Veri madencisi, ne görmek istediğinden tam olarak emin olmayabilir. (b) Erişilen veri, genellikle orijinal operasyonel veritabanının farklı bir versiyonundadır. Madencilik sürecini daha iyi desteklemesi bakımından veri temizlenmiş ve düzenlenmiştir. (c) Veri madenciliği sorgusunun çıktısı, muhtemelen veritabanının bir alt kümesi değildir, veritabanı içeriğinin bazı analizlerinin çıktısıdır [3].

Veri madenciliği modelleri, tahmin edici (predictive) ya da tanımlayıcı (descriptive) olabilirler. Şekil 2.2'de, model tipleri ve altında model tipinin kullandığı veri madenciliği görevleri verilmektedir.



Şekil 2.2. Veri madenciliği modelleri ve görevleri [3]

Bir tahmin edici model, farklı verilerden elde edilen bilinen sonuçları kullanarak, veri değerleri hakkında tahmin yapar. Tahmin edici modelleme, diğer tarihsel verinin kullanımına göre de yapılabilir. Bir tanımlayıcı model ise, verilerdeki örüntüleri ya da ilişkileri tanımlar. Tahmin edici modellerin aksine yeni özellikler (properties) tahmin etmek yerine, sınanan verinin özelliklerini keşfetmeye çalışır.

2.1. Temel Veri Madenciliği Görevleri

Şekil 2.2'de, bu görevler listelenmiştir. Bu görevlerin, gelişmiş veri madenciliği uygulamaları için birleştirilmeleri de mümkündür.

2.1.1. Sınıflandırma

Verinin, önceden tanımlanmış kategorik sınıf etiketleri içerisinde sınıflandırılması işlemidir. "Sınıf", bir veri kümesinde kullanıcıların çoğunlukla ilgilendiği nitelik ya da özelliktir [4] Sınıflar, veri sınanmadan önce belirlendiğinden, sınıflandırma, genellikle denetimli öğrenme (supervised learning) olarak da adlandırılmaktadır. Örneğin, banka kredisi verilip verilmeme kararı, kredi risklerinin tanımlanması vb. Sınıflandırma algoritmaları, veri özellik değerlerine göre tanımlanan sınıflar gerektirmektedir. Sınıflar, o sınıfa ait olduğu bilinen verinin karakteristiğine bakılarak tanımlanmaktadır. Örneğin; bir

havayolu güvenlik tarama istasyonunda yolcuların potansiyel terörist ya da suçlu olup olmadığının belirlenmesinde kullanıldığını düşünelim. Bunun için her bir yolcunun yüzü taranır ve temel örüntü (gözler arasındaki mesafe, ağız şekli ve büyüklüğü, kafa şekli vb.) tanımlanır. Bu örüntü, bilinen suçlarla ilgili örüntülerle eşleşip eşlenmediğinin tespiti için veritabanındaki kayıtlarla karşılaştırılır [4]. Tıp alanında veri madenciliğinin bu tipi, teşhis ve tedavi amaçlı karar vermede oldukça önemlidir. Bir tıp alanındaki denetimli veri kümesi, hastanın semptomlarına göre geçmiş hastalık bilgilerini içermektedir. Öğrenilen bilgi, gelecek hastaların (hastalık semptomlarına sahip) riskinin değerlendirilmesinde yardımcı olan bir tıbbi uzman sistem gibi kullanılabilir [5].

2.1.2. Regresyon

Üzerinde durulan değişkenlerden birisinin bağımlı (y), diğerinin de bağımsız (x) olması durumunda, y'nin, x'in bir fonksiyonu olarak ifade edildiği ilişkidir. Regresyon analizi, bağımlı değişkenle, bir ya da birden fazla bağımsız değişken arasındaki ilişkinin modellenmesi için kullanılan analiz yöntemidir. Bu analizle, değişkenler arasındaki neden-sonuç ilişkileri bulunabilir ve farklı değerler için tahmin de yapılabilir.

Gelecek değerler, zaman serileri analizi ya da regresyon teknikleriyle tahmin edilebilmesine rağmen, başka yaklaşımlar da kullanılabilmektedir. Örneğin; su baskını (sel) tahmini oldukça zor bir problem olarak ifade edilmektedir. Buna dair bir yaklaşım, nehirdeki çeşitli noktalara, sel tahmini için uygun verileri toplayacak (su seviyesi, yağmur miktarı, zaman, nem, vb.) monitörlerin yerleştirilmesidir. Nehirdeki potansiyel bir sel noktasındaki su seviyesi, nehir boyunca bulunan sensörlerden toplanan verilerden tahmin edilebilir. Tahmin, verilerin toplandığı zamana göre yapılmak durumundadır [3].

2.1.3. Kümeleme

Kümeleme, sınıflandırmayla benzerdir. Grupların önceden belirlenememesi, kümelemenin farkı olarak ifade edilmektedir. Kümeleme, denetimsiz öğrenme (unsupervised learning) ya da segmentasyon olarak da adlandırılmaktadır, verinin, ayrık ya da ayrık olmayan gruplara bölümlenmesi olarak düşünülebilir. En fazla benzerlik gösteren veriler, kümelerde gruplanırlar. Kümeler önceden tanımlı değildir, genellikle bir alan uzmanının, oluşturulan kümeleri yorumlaması gerekmektedir. Örneğin; bir mağazanın kataloglarının gönderilmesi

için, müşterilerin gelir durumu, yeri vb. niteliklerine ve potansiyel müşteri fiziksel niteliklerine (yaş, boy, kilo vb) göre farklı demografik grupların belirlenmesi durumu [3]. Kümeleme normalde, veriler hakkında çok az bilgi olması ya da olmaması durumunda uygulanır, kümeleme algoritmaları nümerik/kategorik verilerde kullanılabilir [4].

2.1.4. Özetleme

Veritabanından temsili/karakteristik bilginin çıkarılması (veri bölümlerinin getirilmesi) ya da türetilmesidir (bazı nümerik niteliklerin ortalaması gibi özet tip bilgi, veriden türetilebilir). Özetleme, veritabanının içindekileri rafine bir şekilde karakterize etmektedir. Örneğin; üniversiteleri karşılaştırmak için kullanılan pek çok kriterden birisi, ortalama SAT (Scholastic Assesment Test) ya da ACT (American College Testing) skorlarıdır. Bu, öğrencinin tipi ve entellektüel seviyesini tahmin etmede kullanılan bir özetlemedir [3].

2.1.5. Birliktelik (ilişki) kuralları

Burada amaç, beraber giden, birlikte hareket eden şeyleri (markette birlikte satılan ürünler gibi) belirlemektir. Örneğin, "Diş fırçası alan müşterilerin %40'ı makarna da satın alır" örüntüsü gibi. Buradaki amaç, mallar arasındaki pozitif veya negatif korelasyonları bulmaktır. Diş fırçası alan müşterilerin diş macunu da satın alacağını tahmin edebiliriz ancak otomatik bir analiz bütün olasılıkları göz önüne alır ve kolay düşünülemeyecek, örneğin diş fırçası ve makarna arasındaki bağıntıları da bulur. Perakendeciler bu bilgileri raf düzenlemeleri, katalog tasarlaması ve çapraz-satış fırsatları yaratmak için kullanmaktadır [6].

2.1.6. Sıra keşfi

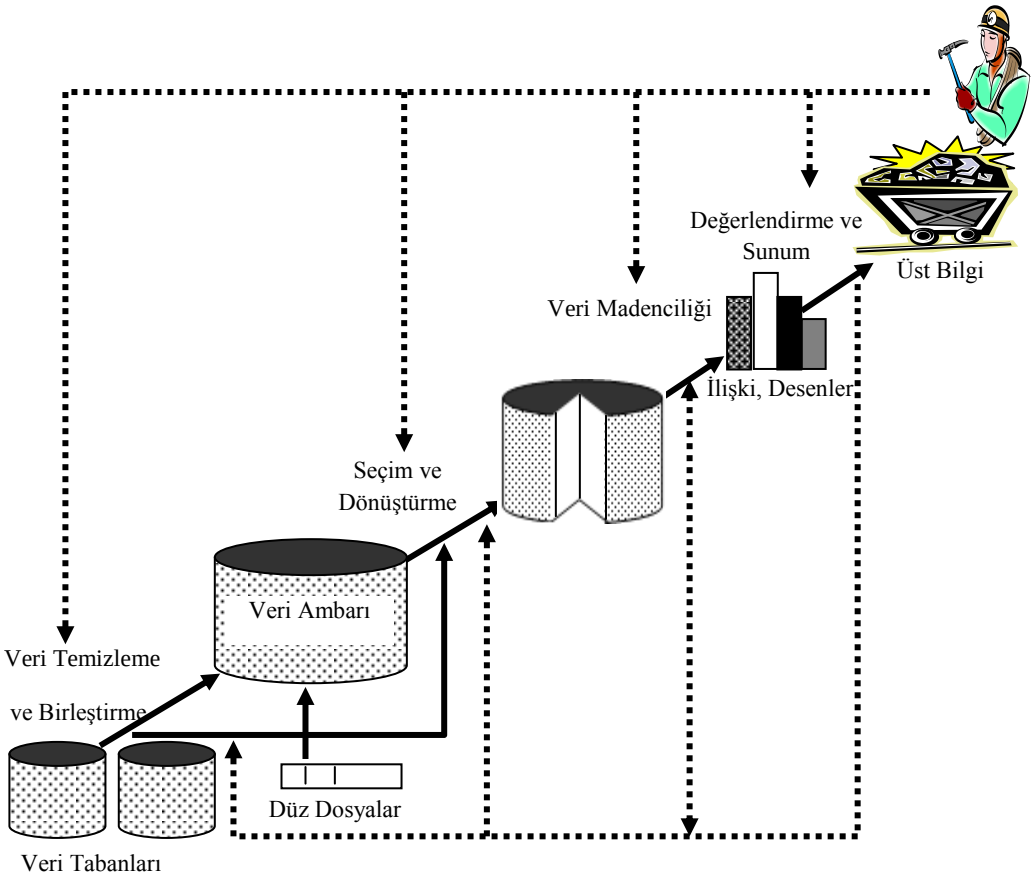
Ardışık analiz olarak da ifade edilen sıra keşfi, verideki ardışık örüntülerin belirlenmesinde kullanılmaktadır. Bu örüntüler, eylemlerin bir zaman sırasına dayanmaktadır. "CD çalar satın alan çoğu müşterinin, bir hafta içerisinde CD satın alacağı" durumu, ardışık örüntüye örnek olarak verilebilir [3].

2.2. Veritabanlarında Bilgi Keşfi ve Veri Madenciliği

Veritabanı, sistematik erişim imkânı sunan, birleştirilmiş ve koordine edilmiş dosyalar kümesi olarak tanımlanabilir. Veri ambarları ise veri madenciliği (iş zekâsı) uygulamaları için alt yapı oluştururlar ve klasik veri depolama yöntemleri ile toplanan verilerin uzun süreli saklandıkları, ilişkili verilerin sorgulanabildiği ve analizlerinin yapılabildiği veri depolarıdır. Başka bir ifadeyle veri ambarı, bir işletmenin ya da kuruluşun değişik birimleri tarafından canlı sistemler aracılığı ile toplanan bilgilerin, gelecekte kullanılabilecek ya da değerlendirilebilecek olanlarının arka planda üst üste yığılarak birleştirilmesinden oluşan büyük çaplı bir veri deposudur [6]. Veri ambarı kavramı, karar vermede kullanılabilecek yapısal kaliteli bilgiye kolay erişimi sağlama ihtiyacından ortaya çıkmıştır. Kurumların büyük miktarda verileri olmasına rağmen, ne yazık ki bu verilere erişmek ve kullanmak, veri miktarı arttıkça daha da zorlaşmaktadır. Bunun sebebi, değişik zamanlarda ve değişik kimseler tarafından geliştirilmiş veri tabanı sistemlerinin ve kütük yapılarının veriyi tutmak için kullanılması, bunun sonucu olarak da çok miktardaki veriye farklı düzlem ve farklı biçimlerden erişme gereksiniminin ortaya çıkmasıdır. Veri ambarındaki veri, daha sonra sorgulama, raporlama ve veri çözümlemede kullanılır [7].

Bir veri ambarı, genellikle boyutların ve bu boyutlarla ilgili özelliklerin belirlediği değerleri barındıran hücrelere sahip birçok boyutlu veritabanı yapısıyla modellenir ve belirli bir amaç için organize edilir. Veri ambarları, değişik kaynaklarda bulunan verilerle ilgili veri temizleme, aktarım, birleştirme, yükleme ve periyodik olarak güncelleme süreçleri ile inşa edilirler. Bu işlem ETL olarak bilinmektedir (Extract Transform Load). Veri madenciliği (VM) genel bir kabulle insan merkezli bir süreçtir. VM, veritabanı sahibine anlaşılır ve faydalı sonuçlar üretmek amacıyla büyük miktardaki verilerin, daha önceden bilinmeyen ilişki ve kuralların keşfedilebilmesi için modelleme, çıkarım ve seçim sürecidir [6].

Veritabanında bilgi keşfi (VBK), verilerdeki faydalı bilgi ve örüntüleri bulma sürecidir. Veri madenciliği ise, VBK sürecinden türetilen bilgi ve örüntülerin çıkarılması için algoritmaların kullanılmasını ifade etmektedir. Şekil 2.3'de, bir bilgi keşfi süreci verilmektedir.



Şekil 2.3. Bilgi keşfi süreci [6-7]

VBK süreci, aşağıdaki beş adımdan oluşmaktadır [3]:

a) *Seçim (selection):* Veri madenciliği süreci için gerekli olan veri, pek çok farklı ve heterojen veri kaynaklarından elde edilebilir. Bu adım veriyi, değişik veritabanlarından, dosyalardan ve elektronik olmayan kaynaklardan elde etmektedir.
b) *Ön işleme (preprocessing):* Sürecin kullandığı veriler yanlış ya da eksik olabilir. Farklı veri tipi ve metrikleri içeren, birden fazla kaynaktan elde edilen anormal veriler de olabilir. Bu adımda, hatalı veriler düzeltilirler ya da kaldırılırlar. Eksik veriler, temin edilmek ya da veri madenciliği araçları kullanılarak tahmin edilmek durumundadır.
c) *Dönüşüm (transformation):* Farklı kaynaklardan elde edilen veriler, işlenmek üzere ortak/uygun bir formata dönüştürülmek zorundadır.
d) *Veri madenciliği:* Gerçekleştirilen veri madenciliği görevine bağlı olarak, istenilen sonuçların üretilebilmesi için algoritmaların dönüştürülmüş veriye uygulanmasıdır.

e) *Değerlendirme ve sunum:* Veri madenciliği sonuçlarının görsel ve bilgi sunum teknikleri kullanılarak kullanıcılara aktarılmasını ifade etmektedir.

2.3. Veri Madenciliği Teknikleri

Veri Madenciliği; büyük miktarda veri içinden gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılması ve bu bilgiler üzerinden gelecekle ilgili tahmin yapılmasını sağlayacak bağıntı ve kuralların aranması süreci olarak adlandırılabilir. Bir başka deyişle; veri ambarlarında tutulan çok çeşitli ve çok miktarda veriye dayanarak daha önce keşfedilmemiş anlamlı kuralları ortaya çıkarmak, bunları karar verme ve eylem planını gerçekleştirmek için kullanma sürecidir. Veri madenciliğinde icra edilen değişik görevlerle ilgili kullanılan tekniklerden bazıları aşağıda verilmiştir [6-7]:

- *Sınıflandırma Teknikleri*
    - Regresyon
    - Bayes sınıflandırma
    - K en yakın komşular
    - ID3-Karar ağacı algoritması
    - C4.5 ve C5.0- Karar ağacı algoritması
    - CART
    - Yapay sinir ağları
    - Genetik algoritma
- *Kümeleme Teknikleri*
    - K-ortalamalar kümeleme
    - En yakın komşu algoritması
    - PAM algoritması
    - Yapay sinir ağları
    - Genetik algoritma
- *Birliktelik Kural Teknikleri*
    - Apriori algoritması
    - Örnekleme algoritması
    - Bölümleme
    - Genetik algoritma

2.3.1. Sınıflandırma teknikleri

Sınıflandırma, veri sınıflarını ya da kavramlarını tanımlayan ve farklılaştıran bir model ya da bir fonksiyonun keşfedilmesi süreci olarak tanımlanabilir. Model keşfi, bir denetimli veri kümesinin analiz edilmesiyle gerçekleştirilebilir [8]. Sınıflandırma, iyi bilinen ve popüler bir veri madenciliği tekniğidir. Sınıflandırma uygulama örnekleri; görüntü ve örüntü tanıma, tıbbi tanı, kredi onay, sanayi uygulamalarında hata tespiti, finansal pazar eğilimlerinin sınıflandırılması vb. olarak ifade edilebilir. Tahmin, sınıflandırmanın bir tipi olarak görülebilir [3]. Regresyon, Bayes sınıflandırma, K en yakın komşular, ID3, C4.5 ve C5.0, CART, Yapay sinir ağları ve Genetik algoritma, sınıflandırmada kullanılan bazı teknikler olarak ifade edilebilir. Tez kapsamında sınıflandırma tekniklerinden karar ağaç/karar kurallarının oluşturulması algoritmaları(ID3, C4.5 ve C5.0, CART) kullanılacağı için, bu algoritmalar diğerlerine göre daha detaylı anlatılmaya çalışılmıştır.

Regresyon

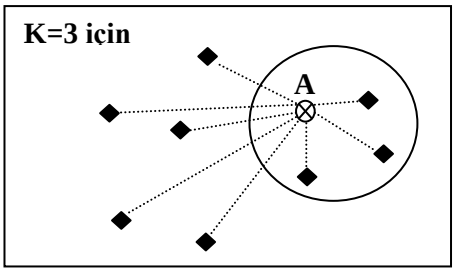
Regresyon problemleri, girdi değerlerine bağlı olarak bir çıktı değerinin tahminiyle ilgilenmektedir. Regresyon, sınıflandırma için kullanıldığında, girdi değerleri veritabanından alınan değerlerdir. Çıktı değerleri ise sınıfları göstermektedir. Regresyon, sınıflandırma problemlerinin çözümünde kullanılabileceği gibi, tahmin gibi diğer uygulamalar için de kullanılabilir [3].

Saf Bayes sınıflandırma

Bayes teoremini kullanan istatistiksel bir sınıflandırma tekniğidir ve koşullu olasılık durumları ile ilgilidir. Saf Bayes algoritmasının uygulanmasında bir takım kabuller yapılır. Bunlardan en önemlisi niteliklerin birbirinden bağımsız olduğudur. Eğer nitelikler birbirini etkiliyorsa burada olasılık hesaplamak zordur. Niteliklerin hepsinin aynı derecede önemli olduğu kabul edilir. Bayes sınıflandırma, istatistiksel verileri olasılık değerlerine göre sınıflara atama ve ayırma işlemidir. Eğer x veriler, s de sınıf etiketiyse, Bayes sınıflandırıcısı p(s/x) sonsal olasılığını en yüksek kılan sınıf etiketini seçmektedir. Bayes kuralı şu şekilde tanımlanır: $p(X|Y) = \frac{p(Y|X).p(X)}{P(Y)}$ bu ifade, Y'nin gerçekleşmesi durumunda, X'in gerçekleşme olasılığının ne olduğunu ifade etmektedir.

K en yakın komşular

Mesafe ölçümünün kullanılmasına dayalı yaygın bir sınıflandırma yöntemidir. Bu yöntem, sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden faydalanarak, örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunun belirlenmesi amacıyla kullanılmaktadır. Yöntem, örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip "K" sayıda gözlemin seçilmesi esasına dayanmaktadır [9]. Şekil 2.4'de, K=3 için A noktasına en yakın komşuların belirlenmesi gösterilmektedir.



Şekil 2.4. k=3 için A noktasına en yakın komşular

ID3- Karar agacı algoritması

Karar ağacı yaklaşımı, sınıflandırma problemlerinde oldukça faydalıdır. Karar ağaçlarının oluşturulması sırasında dallanmaya/bölümlenmeye hangi nitelikten başlanacağı oldukça önemlidir. Zira, sınırlı sayıda kayıttan oluşan bir eğitim kümesinden faydalanarak mümkün tüm ağaç yapılarını ortaya çıkarmak ve içlerinden en uygun olanını seçerek ondan başlamak çok fazla alternatiften dolayı kolay bir iş değildir. O nedenle karar ağacı algoritmalarının çoğu, daha başlangıçta bir takım değerleri hesaplayarak ona göre ağaç oluşturma yoluna gitmektedirler. Bu amaçla entropi kavramı kullanılabilir ve ağacın dallanması entropinin alacağı değere göre gerçekleştirilebilir [9].

ID3 (Iterative Dichotomiser 3), karar ağacı tabanlı bir algoritmadır ve Ross Quinlan tarafından 1979 yılında geliştirilmiştir [10]. Algoritma, bir veri setinden bir karar ağacı oluşturulmasıyla ilgilidir. Algoritmada, hedef sınır değerlerini içeren nitelik belirlenir (bu niteliğin değerleri S kümesidir) ve bu niteliğin kümesi için "Entropi" [11] hesaplanır (Eş. 2.1).

$$Entropi\ (S) = -\sum_{x \in X} p(x) log_2 p(x) \qquad (2.1)$$

Burada, X, S'deki sınıfların kümesini, p(x), x sınıfındaki eleman sayısının S'deki elemanların sayısına oranıdır. Entropi (S)=0 ise, S kümesi tam olarak sınıflandırılmış demektir.

Karar ağacının hangi nitelikten dallanacağının belirlenmesi için, hedef için kullanılan entropi değeri kullanılarak her bir diğer niteliğin "Bilgi kazancı ya da Kazanç" hesaplanması (Eş. 2.2) gereklidir [12].

$$Kazanç\ (S,A) = Entropi(S) - \sum_{v \in A} \frac{|S_v|}{|S|}.Entropi(S_v) \quad (2.2)$$

Bu kazançlardan en yükseğine sahip olan nitelik, dallanacak nitelik olarak belirlenir. Dallanılan niteliğin her bir sınıfı için dallanma seçenekleri belirlenmeye çalışılır ve karar ağacı oluşturulur.

Söz konusu algoritma, C4.5 ve C5.0'a da temel oluşturduğundan burada bir örnekle açıklanmasının faydalı olacağı düşünülmüştür. Örnek; Havanın beyzbol oynamak için elverişli olup olmadığına karar vermek isteniyor. Toplanan iki haftalık veriler Çizelge 2.1'de verilmektedir. ID3 algoritmasıyla karar ağacı oluşturulması istenmektedir. Söz konusu örneğin çözümü [9,13-14] den düzenlenerek alınmıştır.

Çizelge 2.1. İki haftalık hava verileri

Gün	Dışarı	Sıcaklık	Nem	Rüzgar	Oyun
D1	Güneşli	Sıcak	Yüksek	Hafif	Hayır
D2	Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
D3	Kapalı	Sıcak	Yüksek	Hafif	Evet
D4	Yağmurlu	Ilık	Yüksek	Hafif	Evet
D5	Yağmurlu	Soğuk	Normal	Hafif	Evet
D6	Yağmurlu	Soğuk	Normal	Kuvvetli	Hayır
D7	Kapalı	Soğuk	Normal	Kuvvetli	Evet
D8	Güneşli	Ilık	Yüksek	Hafif	Hayır
D9	Güneşli	Soğuk	Normal	Hafif	Evet
D10	Yağmurlu	Ilık	Normal	Hafif	Evet
D11	Güneşli	Ilık	Normal	Kuvvetli	Evet
D12	Kapalı	Ilık	Yüksek	Kuvvetli	Evet
D13	Kapalı	Sıcak	Normal	Hafif	Evet
D14	Yağmurlu	Ilık	Yüksek	Kuvvetli	Hayır

Hedef sınıf değerlerini içeren "Oyun" dur. Oyun nitelik değerlerinden oluşan küme S kümesidir. Kümede "9 Evet" ve "5 Hayır" bulunmaktadır.

$$Entropi\ (OYUN) = -\left[\frac{9}{14} log_2 \frac{9}{14} + \frac{5}{14} log_2 \frac{5}{14}\right] = 0{,}940$$

Herbir nitelik için kazanç değerleri hesaplanır;

$$\left|DIŞARI_{Güneşli}\right| = 5\ kayıt\ var\ \longrightarrow\quad Karşılık\ gelen\ 2\ Evet, 3\ Hayır$$

$$\left|DIŞARI_{Yağmurlu}\right| = 5\ kayıt\ var \longrightarrow\quad Karşılık\ gelen\ 3\ Evet, 2\ Hayır$$

$$\left|DIŞARI_{Kapalı}\right| = 4\ kayıt\ var\ \longrightarrow\quad Karşılık\ gelen\ 4\ Evet, 0\ Hayır$$

$$Entropi\ \left|DIŞARI_{Güneşli}\right| = -\left[\frac{2}{5} log_2 \frac{2}{5} + \frac{3}{5} log_2 \frac{3}{5}\right] = 0{,}971$$

$$Entropi\ \left|DIŞARI_{Yağmurlu}\right| = -\left[\frac{3}{5} log_2 \frac{3}{5} + \frac{2}{5} log_2 \frac{2}{5}\right] = 0{,}971$$

$Entropi\ \left|DIŞARI_{Kapalı}\right| = -\left[\frac{4}{4} log_2 \frac{4}{4}\right] = 0$ olarak bulunur.

$$Kazanç\ (DIŞARI, OYUN) = 0{,}940 - (\frac{5}{14}(0{,}971) + \frac{5}{14}(0{,}971) + \frac{4}{14}(0)) = 0{,}246$$

Benzer şekilde diğer nitelikler için de entropi/kazanç hesaplamaları yapıldığında aşağıdaki tablo değerleri (Çizelge 2.2) elde edilecektir.

Çizelge 2.2. Kazanç değerleri

Nitelik	Kazanç
Dışarı	0,246
Sıcaklık	0,029
Nem	0,151
Rüzgar	0,048

Tablo 2'den görüleceği üzere, en büyük kazançlı nitelik "Dışarı" dır. Dolayısıyla dallanma da bu nitelikten olacaktır. Kök düğüm "Dışarı" dır ve 3 dala sahiptir (Güneşli, Kapalı ve Yağmurlu). Şimdiki soru, "Güneşli" dal düğümünde hangi niteliğin test edileceğidir. $S_{Güneşli}$={D1, D2, D8, D9, D11}. Yeni tablo (Çizelge 2.3), sadece 5 satıra karşılık gelen değerlerden oluşacaktır.

Çizelge 2.3. Dışarı=Güneşli için veriler

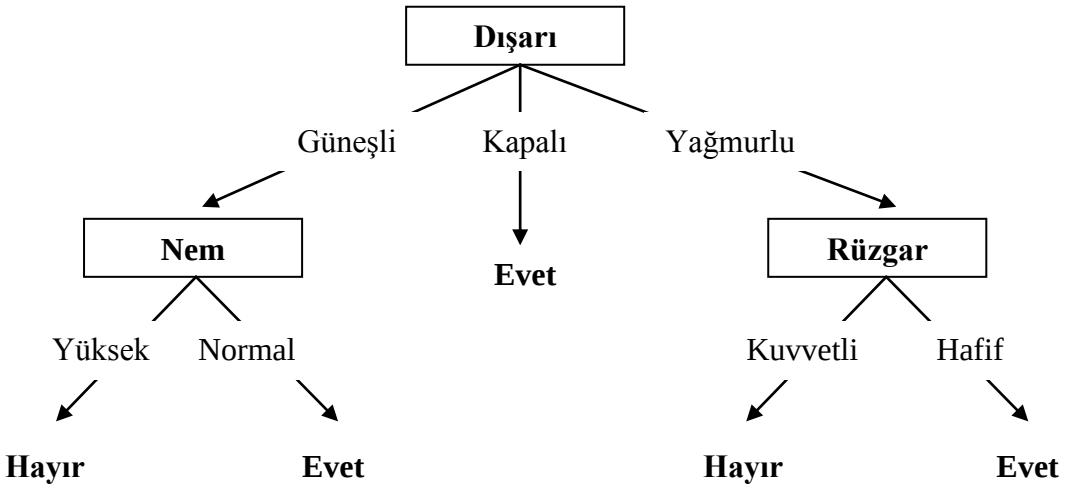
Gün	Dışarı	Sıcaklık	Nem	Rüzgar	Oyun
D1	Güneşli	Sıcak	Yüksek	Hafif	Hayır
D2	Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
D8	Güneşli	Ilık	Yüksek	Hafif	Hayır
D9	Güneşli	Soğuk	Normal	Hafif	Evet
D11	Güneşli	Ilık	Normal	Kuvvetli	Evet

Burada önce "Oyun" için yine entropi hesaplanır ve her bir nitelik için önceden olduğu gibi kazanç değerleri hesaplanır. Benzer hesaplamalar tekrarlandığında Çizelge 2.4'deki değerlere ulaşılabilir:

Çizelge 2.4. Kazanç değerleri

Nitelik	Kazanç
Sıcaklık	0,570
Nem	0,970
Rüzgar	0,019

Çizelge 2.4'den görüleceği üzere, en büyük kazançlı nitelik "Nem" dir. Dolayısıyla dallanma da bu nitelikten olacaktır. Süreç, tüm veriler sınıflandırılana kadar devam edecektir. Algoritma sonucu elde edilen karar ağacı Şekil 2.5'de verilmektedir.



Şekil 2.5. ID3 Nihai karar ağacı

Karar ağacı kural formatında da ifade edilebilir (IF Dışarı=Güneşli AND Nem=Yüksek THEN Oyun=Hayır vb.)

C4.5 ve C5.0 Karar ağacı algoritması

C4.5 karar ağacı algoritması [15], ID3'ün geliştirilmiş hali, üst versiyonudur. Algoritma, en iyi bilinen ve en yaygın kullanılan öğrenme algoritmalarından birisidir. Karar ağacı oluşturulmasında ID3 algoritması, "Bilgi kazancı" ölçüsünü kullanırken, C4.5 ise, "Kazanç Oranı" ölçüsünü kullanmaktadır. C4.5 algoritması, ID3'e ilaveten (ID3 kısıtlamalarını kaldırmak için) bazı ilave durumlar içermektedir [16]:

a) Sayısal (kesikli ve sürekli) değerli nitelikleri de izin verir.
b) Bilinmeyen (eksik) değerleri kümelerin olmasına izin verir.
c) Gereksiz alt ağaçları budayarak ağacın yapısını basitleştirir.

Bilgi bölümleme aslında, A niteliğinin değerlerine göre S nin entropisidir. S kümesi de A niteliğinin belirlenmesi için gereken bilgi miktarını vermektedir ve "Eş. 2.3", "Eş. 2.4" ve " Eş. 2.5" deki gibi hesaplanmaktadır:

$$Bilgi\ Bölümleme\ (S,A) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \quad (2.3)$$

$$Kazanç\ Oranı\ (S,A) = \frac{Kazanç\ (S,A)}{Bilgi\ Bölümleme\ (S,A)} \quad (2.4)$$

$$Kazanç\ Oranı\ (S,A) = \frac{Kazanç\ (S,A)}{-\sum_{i=1}^{v} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}} \quad (2.5)$$

Kazanç oranı, bilgi kazancını normalize eder. Kazanç oranı, dal sayısının kısıtlanmasını sağlamakta, bir nitelikten dallanan dal sayısını belli bir dengede tutarak kazancı optimize etmeye çalışmaktadır.

Örnek: ID3 karar ağacı algoritmasıyla ilgili kullanılan örneği dikkate alalım ve "Dışarı" niteliği ile ilgili olarak kazanç oranını hesaplayalım. {S_1, S_2, ...S_v} değerleri, "Dışarı" niteliğinin her bir değerine karşılık gelen hedef değerleridir. Yani;

$S_{güneşli}$=S_1={hayır, hayır, hayır, evet, evet}= 5 adet (14 kayıttan)
$S_{bulutlu}$=S_2={ evet, evet evet, evet}= 4 adet (14 kayıttan)
$S_{yagmurlu}$=S_3={ evet, evet, hayır evet, hayır}= 5 adet (14 kayıttan)
"Dışarı" niteliğine dair A kümesi={güneşli, bulutlu, yağmurlu} olarak yazılabilir.

$$Bilgi\ Bölümleme\ (S,A) = -\left[\frac{5}{14}log_2\left(\frac{5}{14}\right) + \frac{4}{14}log_2\left(\frac{4}{14}\right) + \frac{5}{14}log_2\left(\frac{5}{14}\right)\right] = 1{,}577$$

Kazanç değeri, önceki hesaplamalarda $Kazanç\ (DIŞARI, OYUN) = 0{,}246$ bulunmuştu. Bu durumda kazanç oranı;

$Kazanç\ Oranı\ (DIŞARI, OYUN) = \frac{0{,}246}{1{,}577} = 0{,}155$ olarak bulunacaktır.

Niteliğin kesikli ya da sürekli sayısal değerlere sahip olması durumunda, niteliğin bölünme noktalarının belirlenmesi (eşik değer için en iyi değerin nasıl seçileceği) önemlidir. Örneğin, A niteliği {20, 78, 90, 150, 160, 170, 180, 200, 250} değerlerine sahipse, 160 değeri, bölümleme için eşik değeri olarak alınabilir. Bu durumda $A_{\leq 160}$, $A_{>160}$ dallanmaları denenebilir. Başka bir durum, B niteliği değerlerinden tekrarlı olanlar kaldırıldığında {60, 65, 70, 73, 75, 80, 85, 90} elde edilmişse, (73+75)/2=74 bölümleme eşik değeri olarak belirlenebilir. Bu durumda, $B_{\leq 74}$, $B_{>74}$ için kazanç değerleri hesaplanır ve en iyi kazanç değerine sahip bölünme noktası belirlenerek uygun dallanmalar gerçekleştirilir. Nominal niteliklerin aksine, her bir nitelik birden fazla bölünme noktasına sahip olabilir.

C4.5 algoritmasının WEKA gurubu tarafından Javada kodlanan versiyonu J4.8 olarak ifade edilmiştir.

C5.0 (windows üzerinde See5 olarak da adlandırılmaktadır), C4.5 in özel lisanslı ticari bir sürümüdür ve Clementine ve RuleQuest gibi pek çok veri madenciliği paket yazılımlarında kullanılmaktadır. C4.5'e göre daha hızlı olup daha az hafıza kullanmaktadır

CART

CART (Classification And Regression Trees- Sınıflandırma ve Regresyon Ağaçları), 1984 yılında Bremian ve Arkadaşları [17] tarafından geliştirilen, nümerik ya da kategorik nitelikler için sınıflandırma ya da regresyon ağaçları oluşturan ve ağacın ikili (binary) ve tekrarlı dallanmasına/bölünmesine izin veren bir metodolojidir. Dallanmalar hep ikili olarak gerçekleştirilir. Metodolojide, bir düğümde belli bir kriter uygulanarak bölünme gerçekleştirilir. *Gini* algoritması bunlardan birisidir. Her bir nitelik için hesaplanan *gini* değerleri arasından en düşük değerli nitelik, dallanacak nitelik olarak belirlenir. CART,

C4.5 le oldukça benzerdir (C4.5 de ikili dallanma sınırı yok), C4.5, bilgi tabanlı kriterleri kullanırken, CART gini bölünme indeksini kullanmaktadır. En iyi bölünme şartlarının belirlenmesindeki hesaplamaların karmaşıklığı CART'ın dezavanatajı olarak bildirilmiştir.

gini indeksi hesaplamaları "Eş. 2.6" ve "Eş. 2.7" de verilmektedir:

$$gini\ (S1) = 1 - \sum_{i=1}^{k} \left(\frac{S1_i}{|S1|}\right)^2; \qquad gini\ (S2) = 1 - \sum_{i=1}^{k} \left(\frac{S2_i}{|S2|}\right)^2 \tag{2.6}$$

$$gini_{bölünme}(S) = \frac{|S1|}{n}\ gini\ (S1) + \frac{|S2|}{n}\ gini\ (S2) \tag{2.7}$$

Burada k, sınıfların sayısını, $|S1|$, 1. gruptaki örneklerin sayısını, $|S2|$, 2. gruptaki örneklerin sayısını, $S1_i$, 1. gruptaki i kategorisindeki örneklerin sayısını, $S2_i$, 2. gruptaki i kategorisindeki örneklerin sayısını, n de, eğitim setindeki (çizelgedeki) satırların sayısını ifade etmektedir.

Örnek; Çizelge 2.1'deki verileri kullanarak gini hesaplamalarının nasıl yapıldığını gösterelim:

DIŞARI={güneşli, kapalı, yağmurlu}, SICAKLIK={sıcak, ılık, soğuk}, NEM={normal, yüksek} ve RÜZGAR ={hafif, kuvvetli} ve hedef nitelik OYUN={evet,hayır} şeklindeydi. Gini hesaplamasında 2'li dallanma grupları oluşturulur. DIŞARI niteliği için S1={güneşli}, S2={kapalı,yağmurlu}, SICAKLIK niteliği için S1={sıcak}, S2={ılık, soğuk}, NEM niteliği için S1={normal}, S2={yüksek}, RÜZGAR niteliği için S1={hafif}, S2={kuvvetli} gibi. Çizelge 2.5'de, nitelik değerlerinin ikili gruplandırılması verilmektedir.

Çizelge 2.5. Nitelik değerlerinin ikili gruplanması

Dışarı		Sıcaklık		Nem		Rüzgar		Oyun
S1	S2	S1	S2	S1	S2	S1	S2	
2	7	2	7	6	3	6	3	Evet
3	2	2	3	1	4	2	3	Hayır
5	9	4	10	7	7	8	6	

Örneğin, 1. ve 2. sütun sırasıyla, (S1)=Dışarı={güneşli} olan 2 kaydın, (S2)=Dışarı={kapalı,yağmurlu} olan 7 kaydın hedef değerinin "evet" olduğunu göstermektedir. Son satır toplam satırıdır.

Her bir nitelik için gini indeks hesaplamaları aşağıda verilmektedir:

$$gini\left(Dışarı_{güneşli}\right) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right] = 0{,}480\ ;$$

$$gini\left(Dışarı_{kapalı,yagmurlu}\right) = 1 - \left[\left(\frac{7}{9}\right)^2 + \left(\frac{2}{9}\right)^2\right] = 0{,}346$$

$$gini\left(Sıcaklık_{sıcak}\right) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0{,}500$$

$$gini\left(Sıcaklık_{ılık,soguk}\right) = 1 - \left[\left(\frac{7}{10}\right)^2 + \left(\frac{3}{10}\right)^2\right] = 0{,}420$$

Her bir niteliğe ait gini bölünme hesaplamaları da aşağıda verilmektedir:

$$gini_{bölünme}(Dışarı) = \frac{5}{14}\ (0{,}480) + \frac{9}{14}\ (0{,}346) = 0{,}394$$

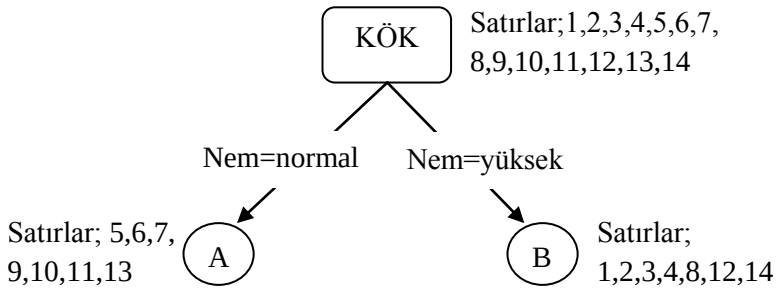
$$gini_{bölünme}(Sıcaklık) = \frac{4}{14}\ (0{,}500) + \frac{10}{14}\ (0{,}420) = 0{,}443$$

Tüm hesaplamaların özeti Çizelge 2.6'da verilmektedir.

Çizelge 2.6 . Gini hesaplama özeti

	Dışarı		Sıcaklık		Nem		Rüzgar	
	S1	S2	S1	S2	S1	S2	S1	S2
gini	0,480	0,346	0,500	0,420	0,245	0,490	0,375	0,500
gini, bölünme	0,394		0,443		0,367		0,429	

Çizelge 2.6'dan görüleceği üzere en küçük gini bölünme değeri 0,367 ile "Nem" niteliğine aittir. Bu nitelik kök düğümden dallanacak olan nitelik olarak seçilir. Nem niteliğinden dallanma, S1={normal}, S2={yüksek} olarak gerçekleşecektir. S1={normal} e karşılık gelen satırlar {5,6,7,9,10,11,13}, S2={yüksek} e karşılık gelen satırlar ise {1,2,3,4,8,12,14} dür (Şekil 2.6). Yeni tabloda ilgili satırlardan oluşacaktır ve A ve B den dallanmalara benzer şekilde devam edilecektir.



Şekil 2.6. İlk dallanma

Yapay sinir ağları

Yapay sinir ağları (YSA), biyolojik sinir ağlarından esinlenerek geliştirilmiş bir bilgi işleme sistemidir. YSA, yapay sinir hücrelerinin birbirleriyle çeşitli şekillerde bağlanmasından oluşur ve genel olarak katmanlar şeklinde düzenlenir. En belirgin özellikleri, birbirlerine bağlı nöronlar, bağlantılar arasındaki ağırlıkların belirlenmesi ve aktivasyon fonksiyonudur [18]. YSA, insan beynindeki sinir hücrelerinin öğrenmesini model almaktadır. Sistem verilerinden olayla ilgili öğrenmeleri gerçekleştirmekte, daha sonra hiç görmediği durum karşısında öğrendiği bilgileri kullanabilmekte ve o durumla ilgili karar verebilmektedir. YSA, sınıflandırma için önemli bir araç olarak ortaya çıkmıştır, kümeleme ve birliktelik kurallarında da kullanılabilmektedir.

Genetik algoritma

Genetik algoritmalar (GA), evrimsel hesaplama yöntemlerinin örnekleridir ve optimizasyon tipi algoritmalardır. Potansiyel problem çözümlerinin popülasyonu verildiğinde, evrimsel hesaplama bu popülasyonu yeni ve potansiyel olarak daha iyi çözümlerle genişletir. Evrimsel hesaplama algoritmalarının esası, zaman boyunca en iyi ya da en güçlü bireylerin üretildiği biyolojik evrimdir. DNA dizisindeki kromozomlar, canlı bir organizmanın özet modelini oluşturmaktadır. Koromozomların alt bölümleri olan

genler ise, bireylerin farklı özelliklerini tanımlamada kullanılırlar. Üreme esnasında ebeveynden elde edilen genler, çocuk genlerin üretimi için birleştirilirler [3]. Genetik algoritmaların çoğu, seçim (selection), çaprazlama (crossover) ve mutasyon (mutation) gibi üç ana operatörden oluşmaktadır. Algoritma, veri madenciliği uygulamalarında çözüm uzayının hepsi yerine belirli bir kısmını incelediği için diğer yöntemlere göre daha hızlı çalışmaktadırlar [19]. Veri madenciliğinde GA, sınıflandırma, kümeleme, tahmin ve birliktelik kurallarında kullanılabilmektedir [3].

2.3.2. Kümeleme teknikleri

Kümeleme, veriyi, birbirine benzeyen elemanlardan oluşan sınıflara (kümelere) ayırarak, heterojen bir veri grubundan homojen alt veri grupları elde edilmesidir [20]. Kümeleme, verilerin gruplanması bakımından sınıflandırmayla benzerdir. Ancak sınıflandırmanın aksine, gruplar önceden tanımlı değildir (kayıtların hangi kümelere ayrılacağı ya da kümelemenin neye göre yapılacağı belirsizdir). Bunun yerine gerçek verilerde bulunan özelliklere göre, veriler arasındaki benzerlikler bulunarak gruplama gerçekleştirilir (bu gruplar kümelerdir), yani, en fazla benzerlik gösteren veriler, kümelerde gruplanırlar. Bazı araştırmacılar kümelemeyi, sınıflandırmanın özel bir tipi olarak görmüşlerdir [3]. Kümelemede kullanılan tekniklerden bazıları; K-ortalamalar kümeleme, En yakın komşu algoritması, PAM (Partitioning Around Medoids) algoritması, Yapay sinir ağları, Genetik algoritma vb. olarak ifade edilebilir.

Kümeleme algoritmaları, küme içinde benzerliğin en büyüklenmesi (küme içi uzaklıkların en azlanması), kümeler arası benzerliğin en azlanması (kümeler arası uzaklıkların en büyüklenmesi) kavramına dayanmaktadır [20].

Kümeleme'de temel olmaları bakımından K-ortalamalar kümeleme ve En yakın komşu algoritması kısaca açıklanacaktır.

K-ortalamalar kümeleme

Arzu edilen sete erişilene kadar, birimlerin kümeler arasında hareket ettiği iteratif bir kümeleme algoritmasıdır [3]. K-ortalamalar yöntemi, sadece kümenin ortalamasının tanımlanabildiği durumlarda kullanılabilir, kullanıcıların K değerini, yani oluşacak küme sayısını belirtme gerekliliği bir dezavantaj olarak görülebilir. Esas önemli olan dezavantaj

ise, dışarıda kalanlar (outliers) olarak adlandırılan birimlere/nesnelere karşı olan duyarlılıktır. Değeri çok büyük olan bir nesne, dahil olacağı kümenin ortalamasını ve merkez noktasını büyük bir derecede değiştirebilir. Bu değişiklik kümenin hassasiyetini bozabilir. Bu sorunu gidermek için kümedeki nesnelerin ortalamasını almak yerine, küme içerisindeki elemanların benzemezliği en az olan nesne anlamındaki "medoid" kullanılabilir [21]. Algoritma, hem sayısal hem de kategorik verilerde kullanılabilir. Başlangıçta kaç küme/bölümleme oluşturulacağı (örn: k=2) ve başlangıç kümeler (örn: K1,K2) belirlendiğinde, K1 ve K2 kümelerinin ortalamalarına yakınlıklarına göre küme elemanları bu kümelere dahil edilmektedir. Süreç, kümelerde değişiklik olmayana kadar devam eder. Örneğin; A Kümesi={1,4,10,14,18,6,2} olsun, k=2 ve K1={1}, K2={4} olsun, öklit mesafe kullanıldığında K1={1,2}, K2={4,6,10,14,18}, ortalama hesaplandığında (1.5) ve (10.4) bulunur, yeni kümeler K1={1,2,4}, K2={6,10,14,18}dir. Ortalama hesaplandığında (2.3) ve (12.0) bulunur, yeni kümeler K1={1,2,4,6}, K2={10,14,18} dir.

En yakın komşu algoritması

Birimler, en yakın mevcut kümeler içerisine iteratif olarak eklenirler. Bu algoritmada, birimlerin mevcut kümeye eklenip eklenmeyeceği ya da yeni bir kümede tanımlanıp tanımlanmayacağı kararı eşik değer (t) ye göre verilmektedir [3]. Algoritmada, bir başlangıç kümesiyle (K1={A} olsun) başlanır. B biriminin K1 kümesine eklenip eklenmeme durumuna bakılır, B biriminin A ya olan mesafesi eşik değerden küçükse, K1={A,B} elde edilir, değilse K2={B} şeklinde ifade edilir. Benzer şekilde diğer birimlerin küme/lere eklenip eklenmeme durumları eşik değerle belirlenir.

2.3.3. Birliktelik (ilişki) kural teknikleri

Birliktelik kurallarının analizi süreci market sepeti analizi olarak da adlandırılır. Market sepeti analizinde müşteri ile ilgili veri hareketlerinden gelecekte müşterinin nasıl bir tercih yapacağına dair sonuçlar tahmin edilmektedir. Çok sayıda verinin depolandığı bir veri tabanı içinde çeşitli nitelikler arasında hemen fark edilmeyen birtakım ilişkilerin ortaya çıkartılması stratejik kararların alınmasına yardımcı olabilir [22]. Bir ürün satın alındığında başka bir ürünün de satın alınması bir birliktelik (ilişki) kuralını ifade etmektedir. İlişki kuralları perakende satış mağazalarında, pazarlama, reklam, ürün yerleştirme, stok kontrol gibi amaçlara sıklıkla hizmet etmektedirler. Bu kurallar, veri birimleri arasındaki ilişkinin

gösterilmesinde kullanılmaktadır [3]. Örneğin, market sistemlerinde, hangi zaman dilimlerinde hangi ürünlerin, hangi ürünlerle birlikte satıldıkları bilgisi (örneğin, fıstık ezmesinin satıldığı zaman dilimlerinin %33 ünde jölenin de satılması gibi), önemli yönetimsel kararların verilebilmesini tetikleyebilir. Normalde görülemeyen ancak kayıtların taranması sonucunda ortaya çıkan örüntüler, akla gelmeyen birimler arasında var olan ilişkileri, bize söyleyebilmektedir. Bu tür Birliktelik ilişkilerinde kullanılan tekniklerden bazıları; Apriori algoritması, Örnekleme algoritması, Bölümleme, Genetik algoritma vb. olarak ifade edilebilir. Birliktelik ilişkilerinde çok temel ve popüler olmasından dolayı sadece Apriori algoritması verilecektir.

Apriori algoritması

Apriori algoritması, Agrawal ve arkadaşları [23] tarafından geliştirilen, çok iyi bilinen ve pek çok ticari üründe kullanılan önemli bir birliktelik kuralı algoritmasıdır [3]. Geniş nesne kümelerinin (itemset) ortaya çıkarılmasını sağlar. Algoritmada, "kural destek" ve "kural güven" denilen iki ölçütten faydalanılır. $Destek(A\ birlikte\ B) = \frac{Sayı\ (A,B)}{N}$ ve $Güven(A\ birlikte\ B) = \frac{Sayı\ (A,B)}{Sayı(A)}$ şeklindedir. Bununla birlikte, hesaplanan bu değerleri karşılaştırmak üzere bir de "eşik değer" gereksinimi vardır. Elde edilen hesaplama sonuçlarının bu eşik değerlere eşit ya da büyük olması istenir. Veritabanı taranarak, analize dahil edilecek her bir ürün için tekrar sayıları (destek sayıları) hesaplanır. Bu değer eşik destek sayısı ile karşılaştırılır. Eşik destek sayısından küçük değerlere sahip satırlar analizden çıkarılır ve koşula uygun kayıtlar dikkate alınır. Bu sefer önceki adımda seçilen birimler ikişerli gruplandırılarak tekrar sayıları (destek sayıları) hesaplanır. Bu değerler, eşik destek sayısı ile karşılaştırılır. Eşik destek sayısından küçük değerlere sahip satırlar analizden çıkarılır. Daha sonra, üçerli dörderli gruplamalar yapılarak bu grupların destek sayıları hesaplanır ve eşik değerle karşılaştırılır. Eşik değerlere uygun olduğu sürece de işlemlere devam edilir [9].

2.4. Veri Madenciliği Uygulama Alanları

Veri madenciliği, iş etkinliğinin geliştirilebilmesinde gerekli olan yeni ve saklanmış bilginin bulunması için çok büyük miktarlardaki verileri analiz eden bir süreçtir. Pek çok endüstri, işlerini geliştirmek ve rekabet avantajı kazanmak için veri madenciliğini misyonlarına ve iş süreçlerine dahil etmektedirler. Günümüzde veri madenciligi birçok

alanda kullanılmaktadır. Aşağıda satış/pazarlama, bankacılık/finans, sağlık ve sigorta, taşıma ve tıp alanındaki bazı veri madenciliği uygulamaları verilmektedir [24].

- *Satış/Pazarlama*
    - Piyasa sepet analiziyle, hangi ürünlerin birlikte satıldığı, ne zaman ve hangi sırada satın alındığına dair bilgilerin sağlanması.
    - Perakende firmalarında müşterilerin satın alma davranışlarının belirlenmesi.
    - Müşteri özellikleri/nitelikleri arasındaki ilişkilerin belirlenmesi.
- *Bankacılık/Finans*
    - Kredi kartı dolandırıcılığının tespiti.
    - Müşteri satın alma faaliyetlerinin analiziyle müşteri sadakatinin belirlenmesi.
    - Kredi kartı müşterilerinin kaybedilmemesinin (muhafazasının) sağlanması.
    - Müşteri grupları tarafından yapılan kredi kartı harcamalarının belirlenmesi.
    - Farklı finansal göstergeler arasındaki gizli ilişkinin keşfedilmesi.
    - Geçmiş piyasa verilerinden faydalanılarak, hisse senedi alım satım kurallarının belirlenmesi, hisse senedi fiyat tahmini.
- *Sağlık ve Sigorta*
    - Yeni poliçeleri satın alacak potansiyel müşterilerin tahmin edilmesi.
    - Sigorta şirketleri için riskli müşterilerin davranış örüntülerinin belirlenmesi.
    - Hileli davranışların belirlenmesi.
- *Taşımacılık*
    - Depolar ve mağazalar arasındaki dağıtım programlarının belirlenmesi ve yükleme örüntülerinin analiz edilmesi.
- *Tıp-İlaç*
    - Farklı hastalıklar için başarılı tıbbi teşhis ve tedavi örüntülerinin belirlenmesi, risk değerlendirmesi.
    - İlaç dozajına hastanın tepkisinin tahmini, ilaç dozaj ve yan etkileri arasındaki ilişkinin belirlenmesi.
- *Diğer uygulamalar*
    - Uydu görüntüleriyle, yıldız gibi gök cisimlerinin sınıflandırılması.
    - Vergi kaçakçılığının tespiti.
    - Kara para aklama takibi.
    - Bilgisayarların, yazılımların ve Web portalların değerlendirilmesi.

- Telekomünikasyonda hile tespiti, hatların yoğunluk tahminleri.
- Amprik verilere göre modeller kurularak bilimsel problemlerin çözülmesi.
- Eğitim planlama, personel değerlendirmesi.
- Daha pek çok uygulama alanı

Veri madenciliğinin yukarıda listelenen pek çok kullanım alanının olması, konunun popüler olmasına ve bu konuda pek çok bilimsel çalışmaların yapılmasını da beraberinde getirmiştir. Veri madenciliği alanında üretilen yüzlerce teorik/uygulama çalışmalarına rastlanmaktadır. Bu konuda iki araştırma çalışması incelenebilir [25,26].

3. TIP'TA VERİ MADENCİLİĞİ

Tıbbi teşhislerin subjektif olduğu, sadece mevcut verilere değil, aynı zamanda doktorun tecrübesine ve hatta psikolojik şartlarına da bağlı olduğu bilinmektedir. Yapılan araştırmalar göstermektedir ki, bir hastaya konulacak tanı, farklı doktorlarda ve farklı zamanlarda farklılık gösterebilmektedir [27].

3.1. Amaç ve Gereklilik

Tıbbi veri madenciliğinin amacı, veri madenciliği tekniklerini kullanarak tıbbi alandaki gizli bilgilerin çıkarılması, önemli örüntülerin keşfedilmesidir. Bu örüntülerin arkasındaki nedensel mekanizmalar tam olarak anlaşılamamış olsa dahi, örüntülerin tanımlanması mümkündür.

Büyük miktarlardaki biyolojik, klinik ve yönetimsel verileri içeren klinik veri depoları, hastalığın ilerlemesi ve yönetimi çalışmalarında faydalı olacak ilişkilerin ve örüntülerin keşfedileceği veritabanlarıdır. Bu keşif, veri madenciliği tekniklerince gerçekleştirilmektedir [5].

Başarılı bir tıbbi veri madenciliği uygulaması, tanı süreci, tedavi seçeneklerinin seçimi, hastalık sonucunun tahmin edilmesi vb. gibi klinik karar verme faaliyetleri ve personel tahmini, sigorta, demografik eğilimler, kalite güvence ve süreç etkinliği gibi sağlık hizmetlerinin sunumuna dair yönetimsel karar verme faaliyetlerinin desteklenmesinde etkili bir şekilde kullanılabilen biyomedikal ve sağlık bakım (healt care) bilgilerini sağlayabilmektedir. Veri madenciliğinden elde edilen bilginin kullanımı, hastalara daha iyi hizmet verilebilmesi bakımından bu hizmeti sunanlara yardımcı olacaktır [4].

Tıbbi verilerin başlıca kaynağı hastalardır. Her bir hasta verisi benzersiz olarak değerlendirilir. Biyomedikal ve sağlık alanlarındaki verilerin kalitesi, diğer alanlardaki veri kalitesinden genellikle daha düşük olmaktadır [4]. Söz konusu kalitesizliğin nedenlerinden birisi, tıbbi verilerin pek çok eksik veri içermesidir. Farklı yaş, semptom, aile geçmişi ve/veya komplikasyon riski gibi nedenlerden dolayı, hastaların, aynı hastalıkta dahi aynı tetkiklere ve laboratuvar testlerine tabi tutulamaması, veri üretiminde farklılıkların olmasına neden olmaktadır. Ayrıca, tıbbi veriler çoğunlukla zaman-serisi özelliğine sahip olmalıdır (yani, muayene/tetkik, laboratuvar test tarihleri, klinik bakımdan oldukça

önemlidir). Bu nedenle veri kümelerinin, zaman ögesini içerecek şekilde toplanması zorunluluktur [28].

Hastane bilgi sistemleri ya da hastane veritabanları, tıbbi/klinik amaçlı değil, öncelikli olarak finansal/faturalama amaçlı tasarlanmıştır. Bu nedenle klinik veri madenciliği için yüksek kalitede verilerin sağlanması oldukça zordur. Amerika'da, örneğin pek çok hastanede klinik verilerin üretilmesi/kullanılması bakımından komple elektronik kayıt sistemi kullanılmamaktadır. Dolayısıyla çoğu tıbbi veri (özellikle laboratuvar test sonuçları) kağıt esaslıdır, yani elektronik kayıtta bu bilgiler eksik olarak görülecektir. Tarihsel hasta verilerinin çoğu kağıt esaslıdır ya da taranmış dijital formattadır. Bu nedenle bu verilerle ilgili esaslı bir veri hazırlama/ön işleme yapılmadan, veri madenciliği için kullanılması mümkün değildir [4].

Tıbbi kaynakların son derece kısıtlı olması, var olan kaynakların da etkin kullanılamaması sonucu dünyada her yıl yüz binlerce kişi hayatını kaybetmektedir. Tıpta ve sağlık sistemlerinde sayısal (kantitatif) tekniklerin kullanılması ile hasta kayıpları azaltılabilmektedir. Kanser, DNA'nın hasarı ile hücrelerin programdan çıkması sonucu hücrelerin kontrolsüz bir şekilde veya anormal bir şekilde büyümesi ve çoğalması sonucu oluşan genetik bir hastalıktır. Kanser ne kadar erken teşhis edilirse, tedavisi de o düzeyde başarılı olacaktır. Tıp, istatistik ve veri madenciliği gibi teknikleri kendi alanlarında kullanabilirse gelecekte kanser gibi birçok hastalık erken teşhis sayesinde ilaçla tedavi edilebilecektir. Böylece pahalı ameliyatlara gerek kalmayabilecektir. Günümüzde kansere yakalanan kişilerin çoğu hastalığın ilerlemiş safhalarında hastanelere başvurmakta ve bu sebeple geç teşhis edilmektedir. Bunun sonucunda tedaviler çoğu zaman işe yaramamakta ve hasta kısa zamanda ölmektedir. Sağlam kişilerde ileriye yönelik kanser hastalığının teşhisi, üzerinde durulması gereken en mühim konulardan birisidir. Ülkemizde kanser hasta kayıtlarının düzenli bir ortamda tutulmadığı açıktır. Halbuki tutulacak kayıtlar sayesinde ileriye yönelik daha hızlı karar verme teknikleri oluşturulabilir [29].

Tıp alanında bulunan mevcut veri oldukça fazla ve hayati öneme sahiptir. Sağlık alanında yapılan birçok veri madenciliği araştırmasında, hastaların elektronik tıbbi kayıtları ve idari işleri belgeleyen verilerden yararlanılarak farklı tahminler yapılabilir. Bunlardan bazıları şunlardır:

- Belirli bir hastalığa sahip kişilerin ortak özelliklerinin tahmin edilmesi
- Tıbbi tedaviden sonra hastaların durumlarının tahmin edilmesi
- Hastane maliyetlerinin tahmin edilmesi
- Ölüm oranları ve salgın hastalıkların tahmin edilmesidir.

Günümüzde en fazla bilgi birikiminin yaşandığı alanlardan biri olan tıp alanındaki bilgilerden faydalanılarak önemli bilgiler elde etmek mümkündür. Veri madenciliği, sağlık ve tıp alanındaki büyük veri tabanlarından değerli bilgileri ortaya çıkartarak, hem tıp hem de hizmet kalitesinin artırılması açısından büyük katkılar sağlamaktadır. Söz konusu insan sağlığı olduğu için bu alandaki veri madenciliği çalışmaları önemli bir uygulama alanı bulacaktır. Bu konudaki önem, son yıllarda giderek artan çalışmalar ile ortaya koyulmaya başlamıştır [1].

3.2. Uygulama Örneği

Bazı hastalıkların %100 kesin teşhisi mümkün olmamaktadır. Örneğin, gebelik esnasında çocukta oluşabilecek herhangi bir "down sendromu" riskinin kesin tanısı dış bulgularla sağlanamamaktadır. Buradaki dış bulgulardan kasıt, anneden alınacak kan örneği, ultrason ile bebeğin görüntülenmesi, anne adayının yaşı, hamilelik ayı, aldığı kilo vb. bulgulardır. Ancak bu bulguların hemen hiçbiri hekime %100 tanı koyma olanağı vermez; %100 veya %100'e çok yakın bir tanı için anne karnından alınacak sıvının da incelenmesi gerekmektedir. Oysa bu işlemde de 1/300 oranında bir düşük riski vardır. Dolayısıyla bu işleme girmeden önce hekimin anne karnındaki bebekte "down sendromu" olduğundan kuşkulanması gerekmektedir. Bu aşamada yukarıda söz edilen dış bulgular ve veri madenciliği teknikleri devreye girmektedir.

Daha önce bu işlem uygulanmış, dış bulguları ve operasyon sonucu kaydedilmiş hasta adaylarına ait veritabanı, veri madenciliği algoritmaları tarafından incelenerek, bir makine öğrenmesi, sınıflandırma, karar ağacı vb. gerçekleştirilir. Daha sonra gerçekleştirilen bu sisteme -örneğin karar ağacı- mevcut anne adayının bilgileri girilerek bebekteki risk oranı belirlenir. Bu oranın büyüklüğüne bağlı olarak hekimin bir fayda risk analizi yapıp operasyona karar vermesi kolaylaşır.

Tıp alanında bunun gibi ameliyat riski taşıyan ancak, ameliyat öncesinde gerçekten ameliyat olması gerektiği tam olarak anlaşılamayan hasta ve hastalıklar için de veri madenciliği yöntemi kullanılır.

Ayrıca parmak izi tespiti, yüz şeklinden kimlik tespiti, insan sesinin bilgisayar ve diğer elektronik aygıtlarda komut olarak kullanılması konularında da yapay zeka teknikleri veri madenciliği için de geçerlidir [18].

3.3. Tıbbi Verilerin Yapısı

Veri madenciliğinin tıp alanında uygulanması, tıbbi verilerin yapısından kaynaklanan çeşitli zorluklar nedeniyle diğer alanlardan daha sonra olmuştur. Tıbbi verilerin kendine has yapısından kaynaklanan zorluklar maddeler halinde aşağıda incelenmiştir [20,30]:

i) Tıp alanındaki veriler genellikle farklı kaynaklardan toplanmaktadır. Örneğin hastanın laboratuvar ile ilgili verileri ile hastanın teşhis bilgileri farklı kaynaklarda ve farklı şekillerde tutulmaktadır. Bu nedenle verilerin anlamlı bir bütün haline getirilmeleri için uzman desteğine ihtiyaç vardır

ii) Tıbbi veriler, metinden tıbbi görüntüye (röntgen vb), EEG (Elektroensefalografi), EKG (Elektrokardiyogram) gibi sinyallerden diğer görüntüleme yöntemlerine kadar pek çok farklı formatta tutulmaktadır. Farklı formattaki tıbbi verilerin bir arada kullanılması gerektiği durumlarda anlamlı bir model oluşturmak zorlaşmaktadır.

iii) Tıbbi veriler homojen değildir, çok fazla nümerik ya da kategorik değerler alabilir. Bu nedenle veri kümesi oluşturulurken her bir verinin değerine ayrıca dikkat edilmesi gerekmektedir. Tıbbi veriler kişiye özel olduğundan veri değişim aralığının çok dışında veri değerlerine de rastlanabilmektedir.

iv) Tıbbi veriler, özellikle zaman bazlı tutulan veriler ve görüntü verileri çok büyük boyutlara ulaşabilmektedir. Bu durum, bu verilerin depolanması için çok büyük boyutlu veri depoları ve verileri işlemek için büyük bir işlemci gücü ve yeni teknikler gerektirmektedir.

v) Tıbbi veriler tek başlarına anlam taşımazlar, mutlaka bir uzman tarafından yorumlanması gereklidir. Bu nedenle tıp alanında yapılacak veri analizi çalışmalarında insan faktörü büyük öneme sahiptir.

vi) Tıbbi veriler kesin değildir, her zaman içinde hata ihtimali barındırır. Hata ölçümden kaynaklanabileceği gibi verinin kişiye özel olmasından da kaynaklanabilir. Bu nedenle tıbbi verilerin veri madenciliği çalışmasında kullanılmadan önce, yeterli güvenilirlikte kesinlik taşıması sağlanmalıdır.

vii) Tıbbi veriler, fiziksel verilere göre daha fazla metinsel karakter taşır. Tıbbi verilerin, metin yapısından nümerik yapıya dönüştürülmesi, veri madenciliği uygulama sürecini zorlaştırmaktadır.

viii) Tıbbi veriler insanlardan elde edildiği için yasal kurallara ve gizlilik kurallarına uygun olarak sağlanmalı ve kullanılmalıdır.

ix) Tıp alanında belli bir standardın olmaması ve var olan standartlar arasında tam bir uyumun olmaması, bu alanda bir veri ambarının oluşturulmasını zorlaştırmaktadır.

3.4. Tıpta Veri Madenciliği Literatürü

Özellikle son yıllarda, teknolojideki gelişmeler, bilişim sistemlerinin faydalarının farkına varılması ve buna paralel olarak kurumların bilişim sistemlerini kurmaları ve dolayısıyla verilerin kayıt saklama sistemlerinde saklanabilme durumu, insan odaklı olmanın gereği insan sağlığına yönelik çalışmalar ve veri madenciliğinin bu yöndeki başarısı, bu konuda çok fazla sayıda çalışmanın yapılmasına neden olmuştur. Tıp alanında yapılan çalışmalar o kadar fazladır ki, bu tez kapsamında bu çalışmaların tamamının sınıflandırılması mümkün olmamıştır. Bunun yerine tıp alanında yapılan bir kaç çalışma sunulmuş, özellikle literatür taraması şeklinde olan çalışmalar ön plana çıkarılarak verilmeye çalışılmıştır. Söz konusu literatür tarama çalışmaları, bu konuda çalışacak araştırmacılara ciddi anlamda yardımcı olacağı düşünülmüştür.

Khan ve Arkadaşları (2004), tıbbi görüntü analizi için karar ağacı veri madenciliği algoritmasına odaklanmışlardır. Röntgen görüntülerinin sınıflandırılmasıyla akciğer kanseri teşhisi üzerine çalışmışlardır [31].

Wren ve Garner (2005), Tip II Diyabet hastalığını veri madenciliği yardımıyla incelemiş, vücut içerisindeki epigenetik (genetik olmayan ırsi) değişimlerin Tip II diyabetin ortaya çıkmasında etkisinin olduğunu belirlemişlerdir [32].

Wang ve Arkadaşları (2005), tıbbi görüntülere bulanık küme analizini uygulamışlardır. Mamografiyi normal ve anormal olarak gruplayan bir karar ağacı algoritması kullanmışlardır [33].

Cheng ve Arkadaşları (2006), kardiyovasküler hastalıkları teşhis etmek için bir sınıflandırma algoritması kullanmışlardır. Sınıflandırmanın etkinliği için, otomatik özellik seçme ve uzman görüşü olmak üzere iki özellik çıkarma tekniği üzerine odaklanmışlardır [34].

Bethel ve Arkadaşları (2006), meme kanseri hastalarının geçmiş bilgilerinden elde edilen kriterlere dayaklı olarak bir ilişki kural öğrenicisi geliştirmişler, bu öğreniciyi, CTAES isimli bir uzman sistem içerisinde kullanmışlardır [35].

Xue ve Arkadaşları (2006), koroner kalp rahatsızlığı olarak bilinen bir hastalığın teşhisinde Bayes ağı algoritmasını önermiş ve kullanmışlardır [36].

Aftarczuk (2007), tez çalışmasında tıbbi karar destek sistemlerinde uygulanan veri madenciliği tekniklerinden üç tanesinin (Saf Bayes, C4.5 ve çok boyutlu algılama) beş farklı tıbbi veri kümesi (Kalp hastalığı, hepatit, gögüs kanseri, dermatoloji ve diyabet) üzerindeki performanslarını karşılaştırmıştır [37].

Xing ve Arkadaşları (2007), koroner kalp hastası hastalarının daha uzun yaşama olasılığının tahmini için veri madenciliği tekniklerini kullanmışlardır. Bu amaçla destek vektör makinesi (SVM), yapay sinir ağları ve karar ağaçları (C4.5 ya da ID3, CART ve C5.0) gibi üç tahmin modelini birleştirmişlerdir [38].

Floyd (2007), pankreas kanserinin prognozu için veri madenciliği teknikleri üzerine bir yüksek lisans tez çalışması gerçekleştirmiştir. Tezde, pankreas kanseri olan hastaların beklenen hayatta kalma süresinin araştırılması için veri madenciliği tekniklerinin (yapay sinir ağları, bayes ağları ve SVM) kullanımına odaklanmıştır [39].

Potter (2007), sınıflandırma algoritmasının meme kanseri prognozu ve teşhis veri seti üzerinde en iyi sınıflandırma tutarlılığını sağlayan bir sınıflandırma algoritmasının olup olmadığının tespiti için araştırma yapmışlardır. 56 sınıflandırma algoritmasının kullanıldığı

araştırmada, her veri seti için tek bir en iyi sınıflandırma algoritmasının olmadığını tespit etmişlerdir [40].

Bach ve Cosic (2008), sağlık yönetiminde veri madenciliğinin uygulanabilirliğini amaçlamışlardır. İlk önce konuyla ilgili veri madenciliği uygulamaları üzerine yayımlanan 221 makale tespit etmiş (2007 de sorgulanmıştır) ve kategorik olarak listelemişlerdir. Çalışmada daha sonra doğum kontrol metotlarının seçimine yönelik olarak karar ağacı metodunu başarılı bir şekilde uygulamışlardır [41].

Wu ve Arkadaşları (2008), IEEE 2006 Uluslararası veri madenciliği konferansında belirlenmiş olan en yaygın 10 veri madenciliği algoritmasını (C4.5, K-ortalamalar, SVM, Apriori, EM, PageRank, Adaboost, K en yakın komşu, Bayes ve CART) ele almış, algoritmaların etkisini tartışmış ve algoritmalarla ilgili mevcut ve gelecek araştırmalara dair bilgiler sunmuştur [42].

Lavindrasana ve Arkadaşları (2009), tıbbi uygulamalarda çalışmaların, veri madenciliği amacına, veri tipine, veri madenciliği fonksiyonuna (sınıflandırma, kümeleme, vb) ve kullanılan veri madenciliği algoritmalarına göre sınıflandırıldığı oldukça kapsamlı bir literatür araştırması sunmuşlardır [43].

Tu ve Arkadaşları (2009), C4.5 karar ağacı ve Bayes algoritmalarıyla, kalp rahatsızlıklarının teşhisine imkan veren zeki bir tıbbi karar destek sistemi önermişlerdir [44].

Sapna ve Tamilarasi (2010), diabetik durumun tahmin doğruluğunun test edilmesinde bulanık sinir ağları algoritmalarını kullanmışlardır [45].

Morra ve Arkadaşları (2010), otomatik hippocampal segmentasyonuyla Alzheimer hastalığının tespiti çalışmalarını yapmıştır [46].

Lopes ve Arkadaşları (2011), prostat kanser MRI görüntülerini kullanarak prostat kanser teşhisine yönelik çalışmalar yapmışlar, bu amaçla Adaboost meta modelini kullanmışlardır [47].

Soni ve Arkadaşları (2011), kalp hastalığı tahminine yönelik çalışmışlardır. Analizde üç farklı veri madenciliği algoritması (Bayes sınıflandırma, Karar ağacı ve K en yakın komşu) önermişlerdir. Aynı veri kümesi üzerinde tekniklerin performansı karşılaştırılmış, karar ağacı algoritmasının diğerlerine göre daha iyi sonuç verdiğini göstermişlerdir [48].

Deri kanserinin tahmini için geliştirilen sınıflama, kümeleme ve ilişki (birliktelik) metotlarının kullanımına dair pek çok çalışma bulunmaktadır. Bu çalışmaların detayları için Barati ve arkadaşlarının 2011 yılı tarama makalesine bakılabilir [5].

Yoo ve arkadaşları (2012) , biyomedikal ve sağlık alanlarında, sınıflandırma, kümeleme ve bağıntı tanımlamada kullanılan çeşitli veri madenciliği algoritmalarının, sağlık alanında kullanımı ile ilgili geniş bir literatür taraması yapmışlardır. Çalışmada, sınıflandırma algoritmaları, kullanılan yöntemlere göre de tasnif edilmiştir [4].

Padhy ve Arkadaşları (2012), çok sayıdaki veri madenciliği uygulama alanını açıklamış ve ayrıca gelecek araştırmalarda faydalı olabilecek veri madenciliği çalışma alanlarını bildirmişlerdir [49].

Kolçe ve Frasheri (2012), çeşitli hastalıkların teşhisi ve prognozu için kullanılan veri madenciliği tekniklerinin uygulandığı mevcut araştırmaları değerlendirmiş, tıbbi veri tabanları üzerinde en iyi performans gösteren veri madenciliği algoritmalarının belirlenmesini amaçlamışlardır [50].

Khaleel ve Arkadaşları (2013), özellikle de kalp rahatsızlıkları, akciğer kanseri, meme kanseri vb. lokal olarak sıklıkla görülen hastalıkların keşfedilmesinde kullanılan tıbbi veri madenciliği tekniklerinin analizini gerçekleştirmiş, konu ile ilgili geniş bir literatür taraması yapmışlardır. Çalışmada, hastalık bazında kullanılan veri madenciliği teknikleri de listelenmiştir [51].

Vijiyarani ve Sudha (2013), veri madenciliği tekniklerinin, farklı tip hastalıkların tahmininde nasıl kullanıldığını analiz etmişlerdir. Makale, özellikle kalp hastalıkları, diyabet ve meme kanserinin tahmin edilmesine yönelik geniş bir literatür çalışmasıdır [52].

Taşbaş, Çalık ve Dicle (2012), kronik pankreatit araştırmasında kullanılmak üzere bir bilgi sistemi oluşturmuş ve sisteme girilen hasta bilgileri kullanılarak, hekime kararında destek olacak bir yapı geliştirmeyi hedeflemişlerdir [53].

4. UYGULAMA: AKUT PANKREATİT HASTALARININ MORTALİTE RİSKLERİNİN BELİRLENMESİ

Ankarada bir kamu hastanesindeki 206 adet Akut Pankreatit (AP) hastasına ait verilerin, hastalığın riskinin (mortalite/hayatta kalma) ortaya konulması probleminin çözümünde IBM PASW (Predictive Analytics Software) Modeler 14.0 kullanılmıştır. Hastalıkla ilgili genel tanımlamalar verildikten sonra, problemin çözümüne dair adım adım analiz işlemleri, karar kurallarının oluşturulması ve değerlendirilmesi bu bölüm kapsamında gerçekleştirilmektedir.

4.1. Akut Pankreatit

Pankreas, karın boşluğunda, omurganın bel bölümü önünde yer alan bir salgı bezidir. Ortalama 15-25 cm uzunluğunda ve kadınlarda 55 gr, erkeklerde ise 70 gr ağırlığındadır. Önden arkaya doğru yassılaşan pankreasın düzensiz olan biçimi çengele benzetilebilir. Şişkin olan sağ ucuna baş, daha dar olan orta bölümüne gövde, gövde ile başın birleştiği ince bölüme boyun, ince uzun olan son ucuna da kuyruk denir. Kuyruk bölümü dalağa dek uzar. Pankreas, dalak, karaciğer ve üst mezanter atardamarlarıyla beslenir. Pankreasın boşaltıcı kanalları, Wirsung kanalı ve Santorini kanalıdır [54].

Akut pankreatit (AP), 1992 yılında Atlanta'daki uluslararası konsensus konferansında; pankreasın çeşitli derecelerde etkilendiği lokal doku ve organ sistemlerinin iştirak edebildiği inflamatuvar bir proses olarak tanımlanmıştır. Klinik olarak akut pankreatit, karın ağrısına, serumdaki normal düzeyin 3 katı seviyesinde amilazeminin (veya lipazemi) iştirak etmesi olarak tanımlanır [55].

Bu hastalık şiddetli kanamaya yol açabilen oldukça ciddi bir hastalıktır. Pankreasın ani şekilde ortaya çıkan iltihabıdır. Pankreas hem iç salgı bezi hem de dış salgı bezi olarak görev yapar. Salgıladıkları enzimler pankreasta iken inaktif haldedir. Bu inaktif enzimler sindirim sisteminin bazı bölümlerinde aktifleşerek yağların, proteinlerin, karbonhidratların sindirimini yani parçalanmasını sağlarlar. Akut pankreatitte ise bu enzimler daha pankreastayken aktif haldedir ve dokuların parçalanmasına yol açarlar [56].

AP fizyopatolojisi net anlaşılamamış olup kompleks bir hastalıktır. Hastalığın prognozunu hastaneye başvuru anında belirlemek zor olup özgül bir tedavisi yoktur. Hastalık

konservatif tedavi ile büyük bir oranda düzelmektedir. Şiddetli AP formunda erken yoğun bakım takibi, enteral besleme, ERCP, sfinkterotomi, geniş spektrumlu antibiyoterapi gibi yöntemler kullanılmaktadır. AP potansiyel olarak fatal bir hastalık olup, mortalitesi % 2.1 ile % 7.8 arasında bildirilmektedir. Multiorgan yetmezlik ve pankreatik nekroz prognozu belirleyen en önemli faktörlerdir. Mortalite, olguların yarısında ilk iki haftalık süreçte gözlenir. Nekrotizan pankreatit tüm olguların % 10-20'sinde görülmekte olup mortalite oranı % 14-25' dir. İyileşen olguların 1/3-5'inde diyabetus mellitus gibi fonksiyonel hastalıklar gelişebilmektedir [55].

Amilaz, hastalığın tanısıyla ilgili en sık kullanılan biyolojik parametre olarak bilinmektedir. Amilaz değerinin 1000 IU seviyesinin üzerine çıkmasının akut pankreatit için tanısal olduğu kabul edilmektedir. Kan lipaz düzeyindeki ileri derecede artış da pankreatit için patognomonik kabul edilir. Lipaz/amilaz oran ise alkolik pankreatitin ayırıcı tanısında işe yarar. Diğer laboatuvar tetkiki olarak, AP'nin gerek etiyolojisinin araştırılmasında, gerekse de seyrinin takibinde kullanılan bazı biyokimyasal parametreler vardır. Bunlar, tam kan sayımı, ALT ve AST, alkalen fosfataz, bilirubinler, kan şekeri, kan üre ve kreatinini, serum elektrolitleri ve tanısal periton lavajı sıvısının incelenmesidir. Bu tetkikler doğrudan hastalıkla ilgili olmasalar dahi hekime oldukça değerli bilgiler verirler [57].

4.2. Uygulamada Kullanılacak Hastalık Verileri

Uygulama kapsamında kullanılacak olan verilerin ait olduğu nitelikler (atributes), başka bir ifadeyle satırları ve sütunları olan bir tablonun sütunlarını oluşturan nitelikler aşağıda listelenmektedir. Bu nitelikler, hastalık konusunda uzman hekimlerin, hastalık kapsamında değerlendirmek üzere hastalardan temin ettikleri ve hastalıkta belirleyici olduğunu düşündükleri nitelikleri içermektedir. Oluşturulacak karar kuralları, bu nitelikler arasındaki anlamlı ve ilginç ilişkilerin oluşturulmasına yönelik olacaktır.

Cinsiyet: Hastalara ait cinsiyet kodları erkek hastalar için 1, bayan hastalar için 2 olacak şekilde kodlanmştır.

Yaş: 17-89 yaş arasında dağılım gösteren 206 adet hastaya ait yaş verileridir.

WBC (White Blood Cell/ Beyaz kan hücrelerinin(lökosit)-sayısı): Vücudun savunmasında ve bağışıklığında görevlidir. WBC kemik iliğinde üretilir. Bunlar renksizdir ve şekilleri asimetriktir. WBC değişik sağlık durumlarının teşhisinde önemli bir kriter olarak görülmektedir. Normal bir insanda yaşla değişmekle birlikte, normal akyuvar sayısı 4.4-11.3 (10x3/mikrolitre birimi, ilgili hastane hemogram sonuçlarındaki referans değere göre)

HCT (Hematokrit): Kandaki hemoglobin ve eritrosit miktarını gösterir. Bir başka ifadeyle kanın şekilli elemanlarının tüm kana oranıdır. Anemi ve kan kaybı gibi durumlarda miktarı azalır. Buna karşılık vücut su kaybederse (kusma v.b.) ya da yüksek rakımda hematokrit miktarı artar (Referans değer: %42-50).

NE (Nötrofil): Nötrofil kısa adıyla *Neu* olarakta bilinmektedir. Bir akyuvar (WBC) hücresi olan nötrofil, nötrofilgranülosit olarak da adlandırılmaktadır. Akyuvar hücreleri arasında en sık bulunanıdır. Nötrofillerin sahip olduğu granüller, boyalara özel bir afinite (bağlanma eğilimi) göstermediği için "*nötrofil*" olarak adlandırılmıştır. Nötrofillerin dışındaki granülositler, Eozinofil ve Bazofillerdir (Referans değer: %45.5-73.1).

PLT (Platelet/ Trombosit sayısı): Pıhtılaşmayı sağlayan hücrelerdir. Koagülasyon sistemi ve hemostaz bozukluklarının değerlendirilmesinde kullanılır. Demir eksikliği anemisi ve akut enfeksiyonlarında trombosit sayısı artarken lösemiler, bazı enfeksiyonlar ve kemik iliğinin baskılanması ile trombosit sayısı düşer (Referans değer: 10-450 10x3/mikrolitre).

RDW (Redcell Distribution Width): Alyuvar (eritrosit) dağılım genişliğini gösterir, tam olarak alyuvarların büyüklüklerinin dağılmını gösteren bir parametredir (Referans değer: %11.5-14.5).

MPV (Mean platelet volüme/ortalama trombosit hacmi): Trombositlerin ortalama büyüklüklerini ölçen makineyle hesaplanan bir sayıdır. Yeni trombositler daha büyüktür. Yeni trombositler üretildikçe MPV artar. MPV doktora kemik iliğindeki trombosit üretimi hakkında bilgi verir (Referans değer: 7-12 femtolitre).

BIL (Total Bilirubin): Kanda dolaşan eritrositlerin ortalama ömrü 120 gün kadardır. Bu süre sonunda parçalanan eritrositlerin içindeki hemoglobin bir dizi kimyasal reaksiyonla bilirubine dönüştürülür. Hemoglobin kırmızı renkli bir bileşik olduğu halde, bilirubin sarı

renklidir. Hemoglobinin yıkılmasıyla oluşan bilirubin serumda albumine bağlanarak karaciğere taşınır. Bu tip bilirubine indirekt bilirubin denilmektedir. Karaciğere gelen bu bilirubin glukuronik asitle birleşerek suda erime özelliği kazanır ve safra yoluyla bağırsağa atılır; bu tip bilirubine de direkt bilirubin denilmektedir. İkisi birlikte toplam bilirubini oluştururlar. Serumdaki normal değerleri: direkt bilirubin için % 0.1-0.4 mg, indirekt bilirubin için % 0.1-0.6 mg ve toplam bilirubin için %0.2-l mg arasındadır (Referans değer: <1.4 mg/dl).

AMİLAZ: Amilaz, nişastanın sindiriminde rol oynayan protein yapılı bir enzimdir. Bu gibi gıdaların sindirimi ağız içinde amilaz sayesinde başlar ve ince bağırsakta devam edip orada sonlanır. Amilazın en önemli oluşum yeri pankreastır. Bununla birlikte karaciğer, tükürük bezleri, yumurtalık ve böbreklerde az miktarda da olsa amilaz üretilmektedir. Amilaz bedenin doğası gereği idrar ve kanda az miktarda bile olsa bulunur. Amilazın hafif yüksek olması her zaman önemli bir hastalığa işaret etmez. Kandaki amilazın genellikle üçte biri pankreas, üçte ikisi ise tükürük bezleri kaynaklıdır. Amilazın pankreastan geçiş yolunun herhangi bir sebeple tıkanması hallerinde, kan serumu ve idrardaki amilaz enzimi düzeyi yükselir. Bu duruma amilaz yüksekliği denir (Referans değer: 28-100 U/L).

LDH (Laktatdehidrogenaz): Laktik ve pirüvik asidin birbirlerine dönüşümünü iki yönlü olarak kataliz eden hücre içerisine yerleşmiş bir enzimdir. Hücre hasarının olduğu tüm durumlarda düzeyi artar (Referans değer: 240-480 U/L).

AST-ALT (Aspartataminotransferaz-Alaninaminotransferaz): Karaciğerde oluşan hasarın ilk belirleyicisi karaciğer hücreleri tarafından kana salınan enzimlerdir. Normal koşullarda bu enzimler karaciğer hücreleri tarafından depo edilmektedirler. Ancak karaciğer hücrelerinde meydana gelen hasar sonucu bu enzimler kana karışır ve kan testleri ile tespit edilebilirler. Karaciğere özgü olan ve karaciğer hasarını belirlemek için sıklıkla kullanılan enzimler aminotranferazlardır. Bunlar Aspartataminotransferaz (AST - SGOT) ve alaninaminotransferaz (ALT - SGPT) dir. Bu enzimler normalde karaciğer hücreleri olan hepatositlerde bulunurlar. Karaciğerde bir hasar meydana geldiğinde kana karışırlar ve kandaki seviyeleri yükselir (Referans değer: <50 U/L).

GLU (Kan şekeri): Glukoz kan şekeri düzeyini gösteren tahlildir. Glikoz vücutta enerji ihtiyacını karşılayacak başlıca yakıt olarak kullanılan 6 karbonlu moleküldür. Glukoz

tahlili (kan şeker düzeyi) esas olarak diyabet yani şeker hastalığı taraması, teşhisi ve takibinde kullanılır. Bunun dışında bazı metabolik ve hormonal hastalıklara bağlı glikoz düzeyinde değişiklikler görülebilir (Referans değer: 70-109 mg/dl).

ÜRE: Protein metabolizmasının bir ürünüdür ve böbrekler yoluyla idrarla atılır. Böbrek fonksiyonlarını değerlendirmede önemli bir ölçüttür. Ancak böbrek fonksiyonları dışında vücuttaki azot yükü, günlük sıvı alımı ve idrar akım hızından da etkilendiği için tek başına karar verdirici değildir (Referans değer: 10-50 mg/dl).

BUN (Blood urea nitrogen/kan üre azotu): BUN karaciğerde üretilen ve kanda belli düzeylerde bulunan bir atık üründür. Böbrek fonksiyonlarını değerlendiren başlıca tahlillerden birisidir. Akut ve kronik böbrek yetmezlikleri ve benzeri hastalıkların, diyaliz tedavilerinin takibinde sıklıkla kullanılır.

ALB (Albümin): Albümin plazmada en çok bulunan proteindir. Albuminin en önemli görevi kan damarlarından dışarı sıvı sızmasını önlemesidir. Albümin karaciğerde üretilmektedir. Bu nedenle karaciğer hasarından etkilenmektedir. Karaciğer fonksiyon bozukluğu, nefrotik sendroma neden olan böbrek hastalığı ve beslenme bozukluklarında albümin düzeyi düşer (Referans değer: 35-52 g/l).

Total CA (Kalsiyum): Vücutta en çok bulunan mineraldir.%99 u kemik dokusuna bağlı olarak bulunur. Geri kalan kalsiyum tüm dokulara ve sıvılara dağılmış olarak bulunur. Vücudun hayati fonksiyonlarını devam ettirmede temel rol alır. Kan kalsiyum düzeyi çeşitli metabolik ve endokrin fonksiyonları değerlendirmede kullanılır. Akut pankreatit hastalarında kan kalsiyum düzeyi düşük çıkar (Ref değer: 8.5-10.5 mg/dl).

4.3. Kullanılan Veri Madenciliği Aracı

Akut Pankreatit (AP) riskinin sınıflandırılması probleminin çözümü için IBM PASW (Predictive Analytics Software) Modeler 14.0 kullanılmıştır.

Daha önce SPSS Clementine olarak bilinen PASW Modeler (SPSS tarafından 2009 yılında bu ismi almıştır), IBM tarafından geliştirilen ve verilerdeki desenlerin keşfedilmesini kolaylaştıran SPSS 'in kurumsal veri madenciliği yazılımıdır. Kullanıcılara, programlama bilgisi gerektirmeden istatistiksel ve veri madenciliği algoritmalarını kullanabilmelerine

izin veren görsel bir arayüze sahip olan yazılım, windows işletim sisteminde çalıştırılabilmektedir. Yazılım ticari lisansa sahiptir ve en popüler veri madenciliği paketlerinden birisi olarak bilinmektedir [58].

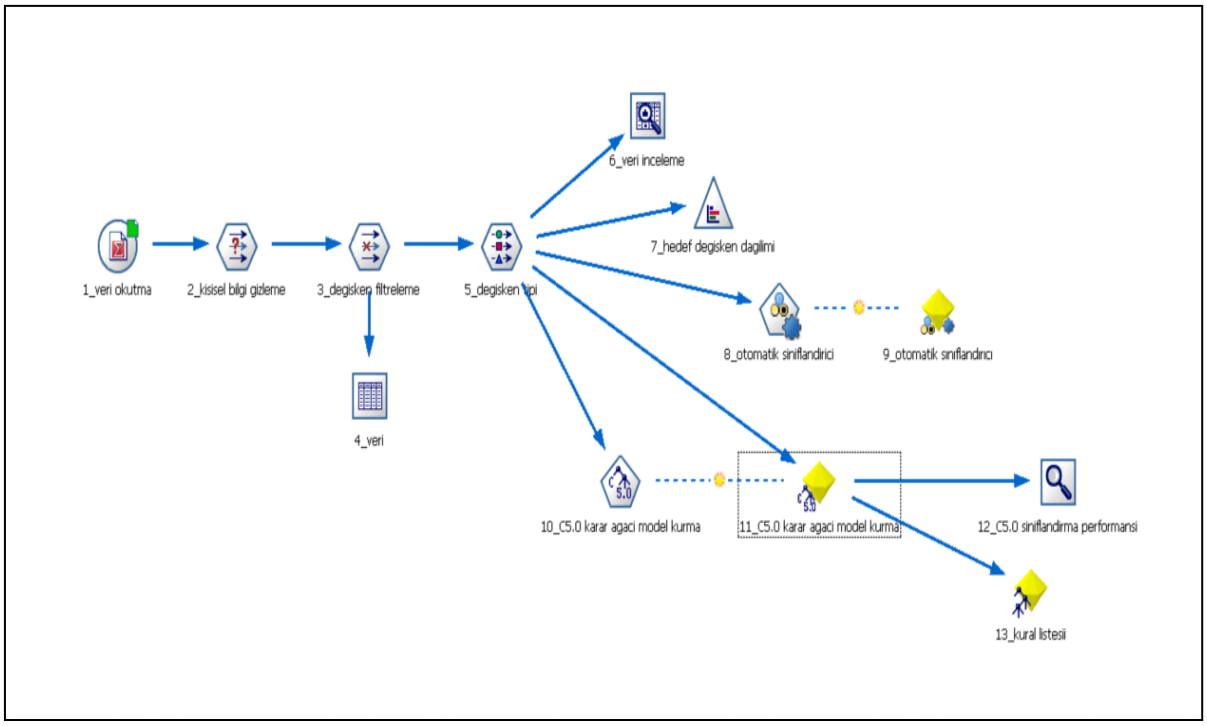
Karar verme süreçlerinin etkinliği artırmak için tasarlanmış tahmine dayalı bir analitik platformu olan PASW'ın önemli üstünlükleri; veri ambarları, veritabanları, düz dosyalar gibi verinin depolandığı yerden bağımsız olarak karmaşık ve zor analiz işlemlerini (tahmin, sınıflandırma, kümeleme, istatistiksel analizler, metin madenciliği, karar yönetimi) gerçekleştirerek gizli kalmış öngörüleri ve modelleri yüksek hız ve verimlilikte ortaya çıkarabilmesi ve diğer IBM yazılımları ile (SPSS Statistics, Cognos Business Intelligence, Infosphere vd.) bütünleşik olarak çalışabilmesidir.

4.4. Modelin Oluşturulması ve Sonuç Üretimi

Modelin oluşturulmasından kuralların üretilmesine kadar olan iş akışı bu bölümde verilmektedir.

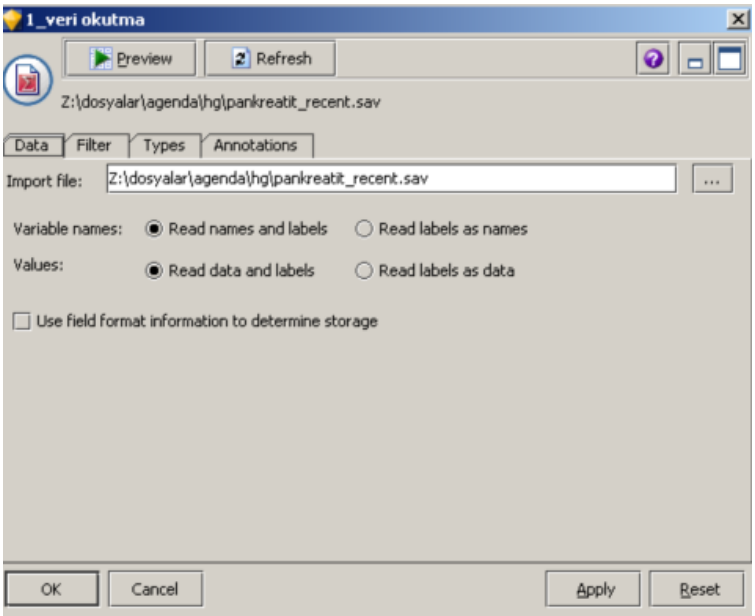
4.4.1. Verilerin girilmesi

AP risk sınıflandırma probleminin çözümüne ilişkin PASW akışı (stream) Şekil 4.1'de verilmiştir. Her bir düğüm (node) işlevi sırasıyla aşağıda açıklanmıştır.



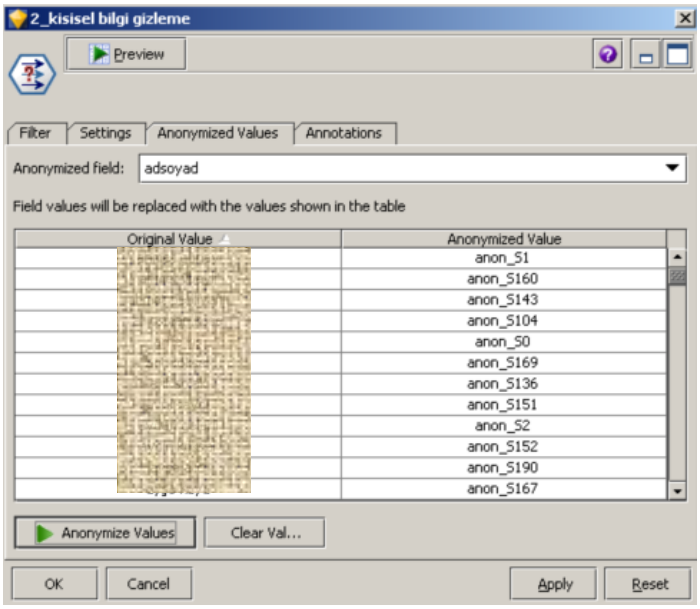
Şekil 4.1. AP risk sınıflandırma probleminin çözümüne ilişkin PASW akışı

Verilerin Okunması: PASW, SQL ve SAS gibi farklı veri tabanları, Excel, XML, SPSS istatistik dosyası, sabit ve değişken dosya gibi farklı kaynaklardan verileri alma kapasitesine sahiptir. AP hastalarına ait kayıtlar SPSS istatistik dosyasından (*.sav) alınmıştır (Şekil 4.2).



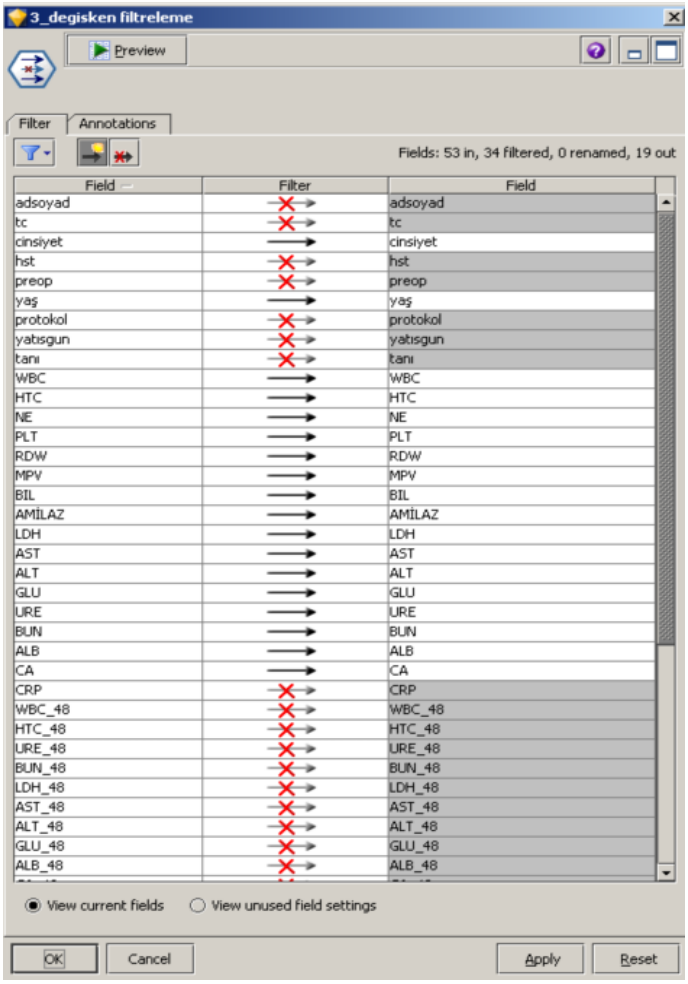
Şekil 4.2. Verilerin alınması

Kişisel Bilgilerin Gizlenmesi: Analiz aşamasına geçmeden önce anonymize düğümü kullanılarak kişisel bilgiler (ad-soyad ve tc kimlik numarası) şifrelenmiştir (Şekil 4.3).



Şekil 4.3. Kişisel bilgilerin şifrelenmesi

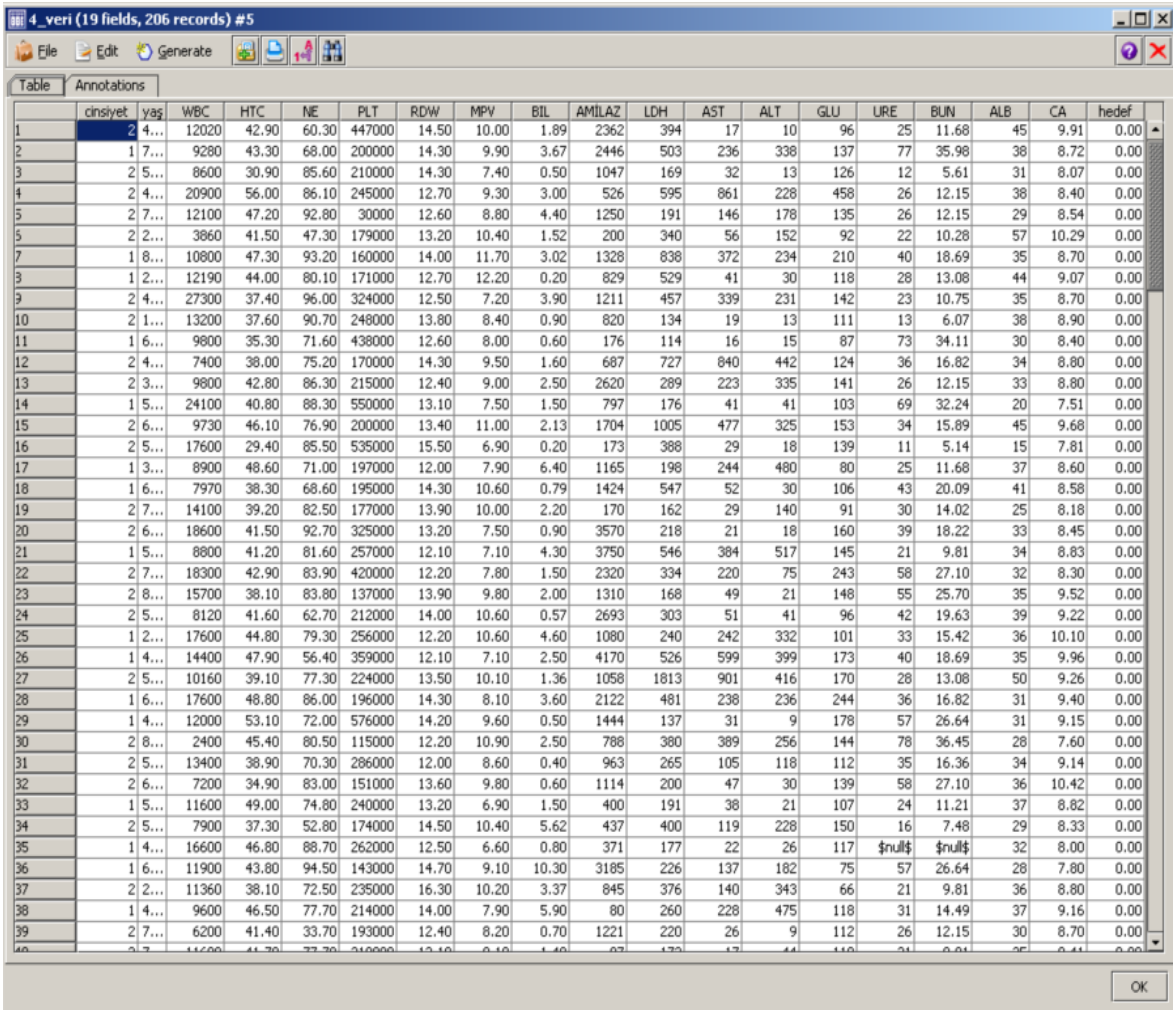
Kullanılmayacak Değişkenlerin Filtrelenmesi: AP hastalarına ilişkin kayıtlarda yer alan fakat AP riskinin sınıflandırmasında dikkate alınmayacak değişkenler filtrelenmiştir (Şekil 4.4).



Şekil 4.4. Kullanılmayacak değişkenlerin filtrelenmesi

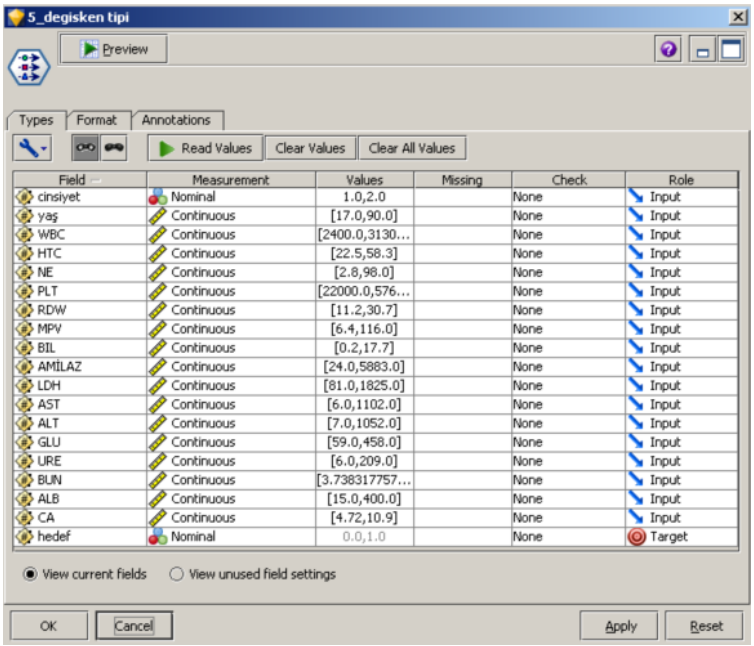
Veri Tablosu: Tablo (table) düğümü ile mevcut veri görüntülenebilir bir yapıda sunulmuştur. Şekil 4.5'den görüleceği üzere veritabanında 19 değişken (field) ve 206 kayıt (record) bulunmaktadır.

Değişken Tiplerinin Belirlenmesi: Sınıflandırma probleminin çözümü için öncelikle hangi değişkenlerin girdi (input), hangi değişkenin ise çıktı (target) olduğunun belirlenmesi gereklidir. Ayrıca değişkenlerin ölçüm düzeyleri de (measurement levels - nominal, ordinal, continuous vb.) tanımlanmış olmalıdır. Tip (type) düğümü kullanılarak bu işlemler yapılmıştır (Şekil 4.6). Hedef değişkeni target, diğer değişkenler ise input olarak atanmıştır. Cinsiyet ve hedef nominal düzeyde, diğer değişkenler ise sürekli düzeyde ölçülmüştür.



	cinsiyet	yaş	WBC	HTC	NE	PLT	RDW	MPV	BIL	AMİLAZ	LDH	AST	ALT	GLU	URE	BUN	ALB	CA	hedef
1	2	4...	12020	42.90	60.30	447000	14.50	10.00	1.89	2362	394	17	10	96	25	11.68	45	9.91	0.00
2	1	7...	9280	43.30	68.00	200000	14.30	9.90	3.67	2446	503	236	338	137	77	35.98	38	8.72	0.00
3	2	5...	8600	30.90	85.60	210000	14.30	7.40	0.50	1047	169	32	13	126	12	5.61	31	8.07	0.00
4	2	4...	20900	56.00	86.10	245000	12.70	9.30	3.00	526	595	861	228	458	26	12.15	38	8.40	0.00
5	2	7...	12100	47.20	92.80	30000	12.60	8.80	4.40	1250	191	146	178	135	26	12.15	29	8.54	0.00
6	2	2...	3860	41.50	47.30	179000	13.20	10.40	1.52	200	340	56	152	92	22	10.28	57	10.29	0.00
7	1	8...	10800	47.30	93.20	160000	14.00	11.70	3.02	1328	838	372	234	210	40	18.69	35	8.70	0.00
8	1	2...	12190	44.00	80.10	171000	12.70	12.20	0.20	829	529	41	30	118	28	13.08	44	9.07	0.00
9	2	4...	27300	37.40	96.00	324000	12.50	7.20	3.90	1211	457	339	231	142	23	10.75	35	8.70	0.00
10	2	1...	13200	37.60	90.70	248000	13.80	8.40	0.90	820	134	19	13	111	13	6.07	38	8.90	0.00
11	1	6...	9800	35.30	71.60	438000	12.60	8.00	0.60	176	114	16	15	87	73	34.11	30	8.40	0.00
12	2	4...	7400	38.00	75.20	170000	14.30	9.50	1.60	687	727	840	442	124	36	16.82	34	8.80	0.00
13	2	3...	9800	42.80	86.30	215000	12.40	9.00	2.50	2620	289	223	335	141	26	12.15	33	8.80	0.00
14	1	5...	24100	40.80	88.30	550000	13.10	7.50	1.50	797	176	41	41	103	69	32.24	20	7.51	0.00
15	2	6...	9730	46.10	76.90	200000	13.40	11.00	2.13	1704	1005	477	325	153	34	15.89	45	9.68	0.00
16	2	5...	17600	29.40	85.50	535000	15.50	6.90	0.20	173	388	29	18	139	11	5.14	15	7.81	0.00
17	1	3...	8900	48.60	71.00	197000	12.00	7.90	6.40	1165	198	244	480	80	25	11.68	37	8.60	0.00
18	1	6...	7970	38.30	68.60	195000	14.30	10.60	0.79	1424	547	52	30	106	43	20.09	41	8.58	0.00
19	2	7...	14100	39.20	82.50	177000	13.90	10.00	2.20	170	162	29	140	91	30	14.02	25	8.18	0.00
20	2	6...	18600	41.50	92.70	325000	13.20	7.50	0.90	3570	218	21	18	160	39	18.22	33	8.45	0.00
21	1	5...	8800	41.20	81.60	257000	12.10	7.10	4.30	3750	546	384	517	145	21	9.81	34	8.83	0.00
22	2	7...	18300	42.90	83.90	420000	12.20	7.80	1.50	2320	334	220	75	243	58	27.10	32	8.30	0.00
23	2	8...	15700	38.10	83.80	137000	13.90	9.80	2.00	1310	168	49	21	148	55	25.70	35	9.52	0.00
24	2	5...	8120	41.60	62.70	212000	14.00	10.60	0.57	2693	303	51	41	96	42	19.63	39	9.22	0.00
25	1	2...	17600	44.80	79.30	256000	12.20	10.60	4.60	1080	240	242	332	101	33	15.42	36	10.10	0.00
26	1	4...	14400	47.90	56.40	359000	12.10	7.10	2.50	4170	526	599	399	173	40	18.69	35	9.96	0.00
27	2	5...	10160	39.10	77.30	224000	13.50	10.10	1.36	1058	1813	901	416	170	28	13.08	50	9.26	0.00
28	1	6...	17600	48.80	86.00	196000	14.30	8.10	3.60	2122	481	238	236	244	36	16.82	31	9.40	0.00
29	1	4...	12000	53.10	72.00	576000	14.20	9.60	0.50	1444	137	31	9	178	57	26.64	31	9.15	0.00
30	2	8...	2400	45.40	80.50	115000	12.20	10.90	2.50	788	380	389	256	144	78	36.45	28	7.60	0.00
31	2	5...	13400	38.90	70.30	286000	12.00	8.60	0.40	963	265	105	118	112	35	16.36	34	9.14	0.00
32	2	6...	7200	34.90	83.00	151000	13.60	9.80	0.60	1114	200	47	30	139	58	27.10	36	10.42	0.00
33	1	5...	11600	49.00	74.80	240000	13.20	6.90	1.50	400	191	38	21	107	24	11.21	37	8.82	0.00
34	2	5...	7900	37.30	52.80	174000	14.50	10.40	5.62	437	400	119	228	150	16	7.48	29	8.33	0.00
35	1	4...	16600	46.80	88.70	262000	12.50	6.60	0.80	371	177	22	26	117	$null$	$null$	32	8.00	0.00
36	1	6...	11900	43.80	94.50	143000	14.70	9.10	10.30	3185	226	137	182	75	57	26.64	28	7.80	0.00
37	2	2...	11360	38.10	72.50	235000	16.30	10.20	3.37	845	376	140	343	66	21	9.81	36	8.80	0.00
38	1	4...	9600	46.50	77.70	214000	14.00	7.90	5.90	80	260	228	475	118	31	14.49	37	9.16	0.00
39	2	7...	6200	41.40	33.70	193000	12.40	8.20	0.70	1221	220	26	9	112	26	12.15	30	8.70	0.00

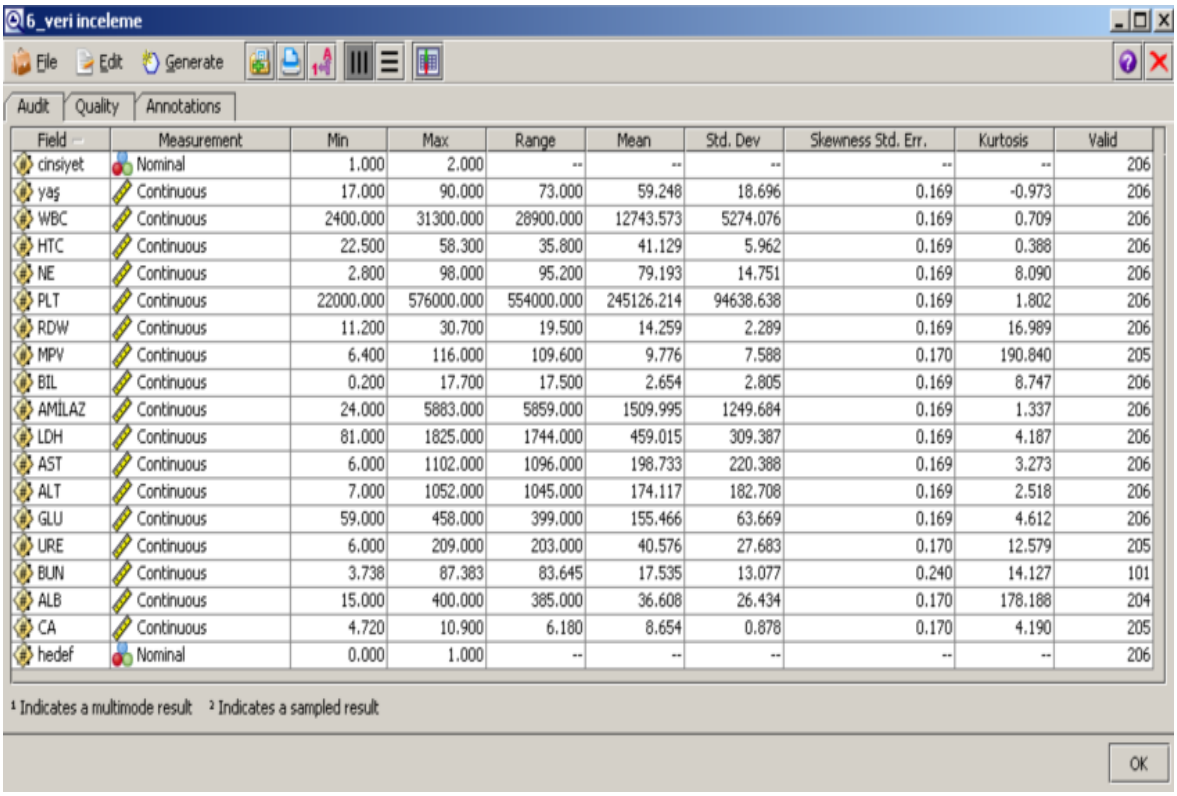
Şekil 4.5. Veri tablosu



Field	Measurement	Values	Missing	Check	Role
cinsiyet	Nominal	1.0,2.0		None	Input
yaş	Continuous	[17.0,90.0]		None	Input
WBC	Continuous	[2400.0,3130...		None	Input
HTC	Continuous	[22.5,58.3]		None	Input
NE	Continuous	[2.8,98.0]		None	Input
PLT	Continuous	[22000.0,576...		None	Input
RDW	Continuous	[11.2,30.7]		None	Input
MPV	Continuous	[6.4,116.0]		None	Input
BIL	Continuous	[0.2,17.7]		None	Input
AMİLAZ	Continuous	[24.0,5883.0]		None	Input
LDH	Continuous	[81.0,1825.0]		None	Input
AST	Continuous	[6.0,1102.0]		None	Input
ALT	Continuous	[7.0,1052.0]		None	Input
GLU	Continuous	[59.0,458.0]		None	Input
URE	Continuous	[6.0,209.0]		None	Input
BUN	Continuous	[3.738317757...		None	Input
ALB	Continuous	[15.0,400.0]		None	Input
CA	Continuous	[4.72,10.9]		None	Input
hedef	Nominal	0.0,1.0		None	Target

Şekil 4.6. Değişken tiplerin belirlenmesi

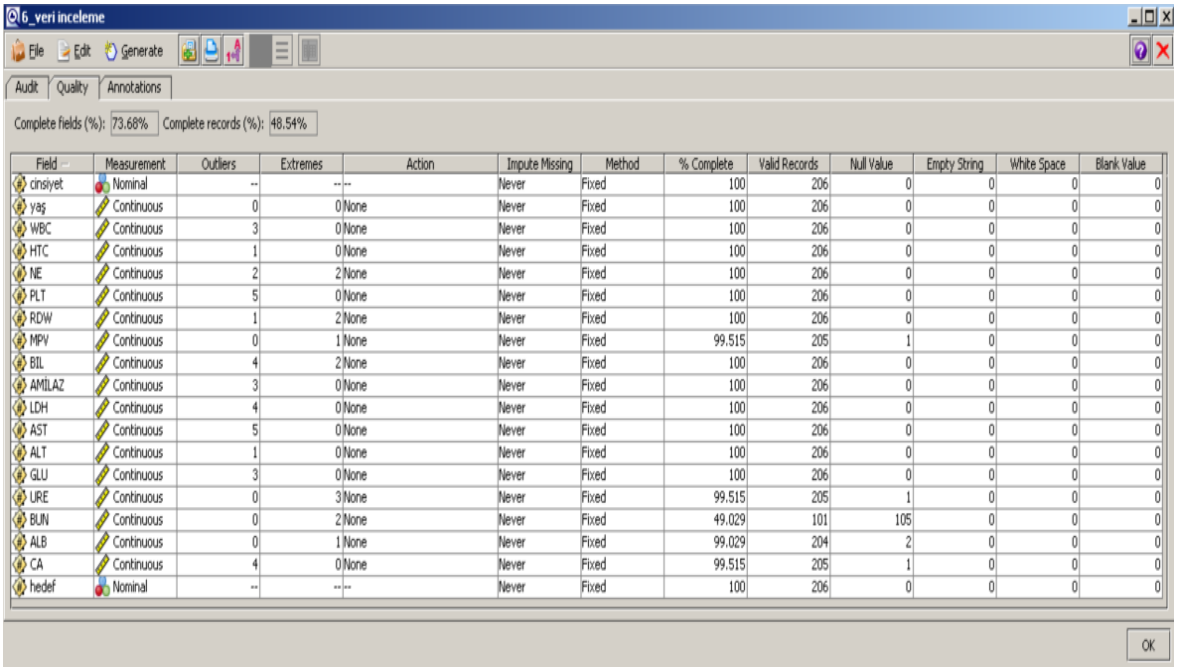
Veri İnceleme/Önişleme: Veri denetim (data audit) düğümü ile veri inceleme/ön işleme işlemleri yapılmıştır (Şekil 4.7, Şekil 4.8). BUN değişkeni dışında kayıp değer (missing values) oranları çok düşüktür. BUN değişkenindeki kayıp değerler herhangi bir yöntem (rassal doldurma, fonksiyon tanımlama, kayıp değerler için tahmin algoritması çalıştırma) kullanılarak doldurulmamıştır. Karar ağacında dallanma kriteri olarak seçilen değişkene ilişkin değerin kayıp olması durumunda, bu değişken yerine kullanılabilecek alternatif/vekil/yedek değişkenlerden (surrogate field) dallanılması, kayıp değerlerin doldurulması zorunluluğunu ortadan kaldırmaktadır.



Field	Measurement	Min	Max	Range	Mean	Std. Dev	Skewness Std. Err.	Kurtosis	Valid
cinsiyet	Nominal	1.000	2.000	--	--	--	--	--	206
yaş	Continuous	17.000	90.000	73.000	59.248	18.696	0.169	-0.973	206
WBC	Continuous	2400.000	31300.000	28900.000	12743.573	5274.076	0.169	0.709	206
HTC	Continuous	22.500	58.300	35.800	41.129	5.962	0.169	0.388	206
NE	Continuous	2.800	98.000	95.200	79.193	14.751	0.169	8.090	206
PLT	Continuous	22000.000	576000.000	554000.000	245126.214	94638.638	0.169	1.802	206
RDW	Continuous	11.200	30.700	19.500	14.259	2.289	0.169	16.989	206
MPV	Continuous	6.400	116.000	109.600	9.776	7.588	0.170	190.840	205
BIL	Continuous	0.200	17.700	17.500	2.654	2.805	0.169	8.747	206
AMİLAZ	Continuous	24.000	5883.000	5859.000	1509.995	1249.684	0.169	1.337	206
LDH	Continuous	81.000	1825.000	1744.000	459.015	309.387	0.169	4.187	206
AST	Continuous	6.000	1102.000	1096.000	198.733	220.388	0.169	3.273	206
ALT	Continuous	7.000	1052.000	1045.000	174.117	182.708	0.169	2.518	206
GLU	Continuous	59.000	458.000	399.000	155.466	63.669	0.169	4.612	206
URE	Continuous	6.000	209.000	203.000	40.576	27.683	0.170	12.579	205
BUN	Continuous	3.738	87.383	83.645	17.535	13.077	0.240	14.127	101
ALB	Continuous	15.000	400.000	385.000	36.608	26.434	0.170	178.188	204
CA	Continuous	4.720	10.900	6.180	8.654	0.878	0.170	4.190	205
hedef	Nominal	0.000	1.000	--	--	--	--	--	206

[1] Indicates a multimode result [2] Indicates a sampled result

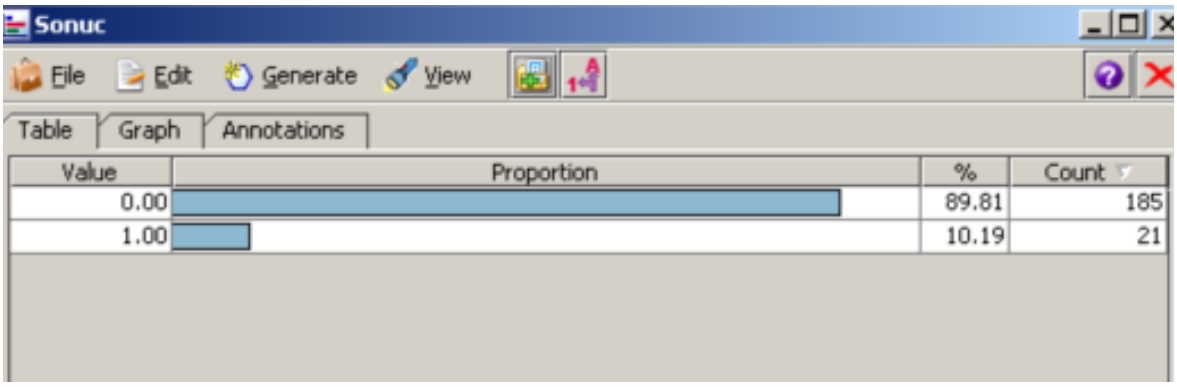
Şekil 4.7. Veri inceleme



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
cinsiyet	Nominal	--	--	--	Never	Fixed	100	206	0	0	0	0
yaş	Continuous	0	0	None	Never	Fixed	100	206	0	0	0	0
WBC	Continuous	3	0	None	Never	Fixed	100	206	0	0	0	0
HTC	Continuous	1	0	None	Never	Fixed	100	206	0	0	0	0
NE	Continuous	2	2	None	Never	Fixed	100	206	0	0	0	0
PLT	Continuous	5	0	None	Never	Fixed	100	206	0	0	0	0
RDW	Continuous	1	2	None	Never	Fixed	100	206	0	0	0	0
MPV	Continuous	0	1	None	Never	Fixed	99.515	205	1	0	0	0
BIL	Continuous	4	2	None	Never	Fixed	100	206	0	0	0	0
AMİLAZ	Continuous	3	0	None	Never	Fixed	100	206	0	0	0	0
LDH	Continuous	4	0	None	Never	Fixed	100	206	0	0	0	0
AST	Continuous	5	0	None	Never	Fixed	100	206	0	0	0	0
ALT	Continuous	1	0	None	Never	Fixed	100	206	0	0	0	0
GLU	Continuous	3	0	None	Never	Fixed	100	206	0	0	0	0
URE	Continuous	0	3	None	Never	Fixed	99.515	205	1	0	0	0
BUN	Continuous	0	2	None	Never	Fixed	49.029	101	105	0	0	0
ALB	Continuous	0	1	None	Never	Fixed	99.029	204	2	0	0	0
CA	Continuous	4	0	None	Never	Fixed	99.515	205	1	0	0	0
hedef	Nominal	--	--	--	Never	Fixed	100	206	0	0	0	0

Şekil 4.8. Veri kalitesi

Hedef Değişken Dağılımı: Hedef değişken dağılımı Şekil 4.9'da görülmektedir. 0 tbc durumunu, 1 ex durumunu ifade etmektedir.

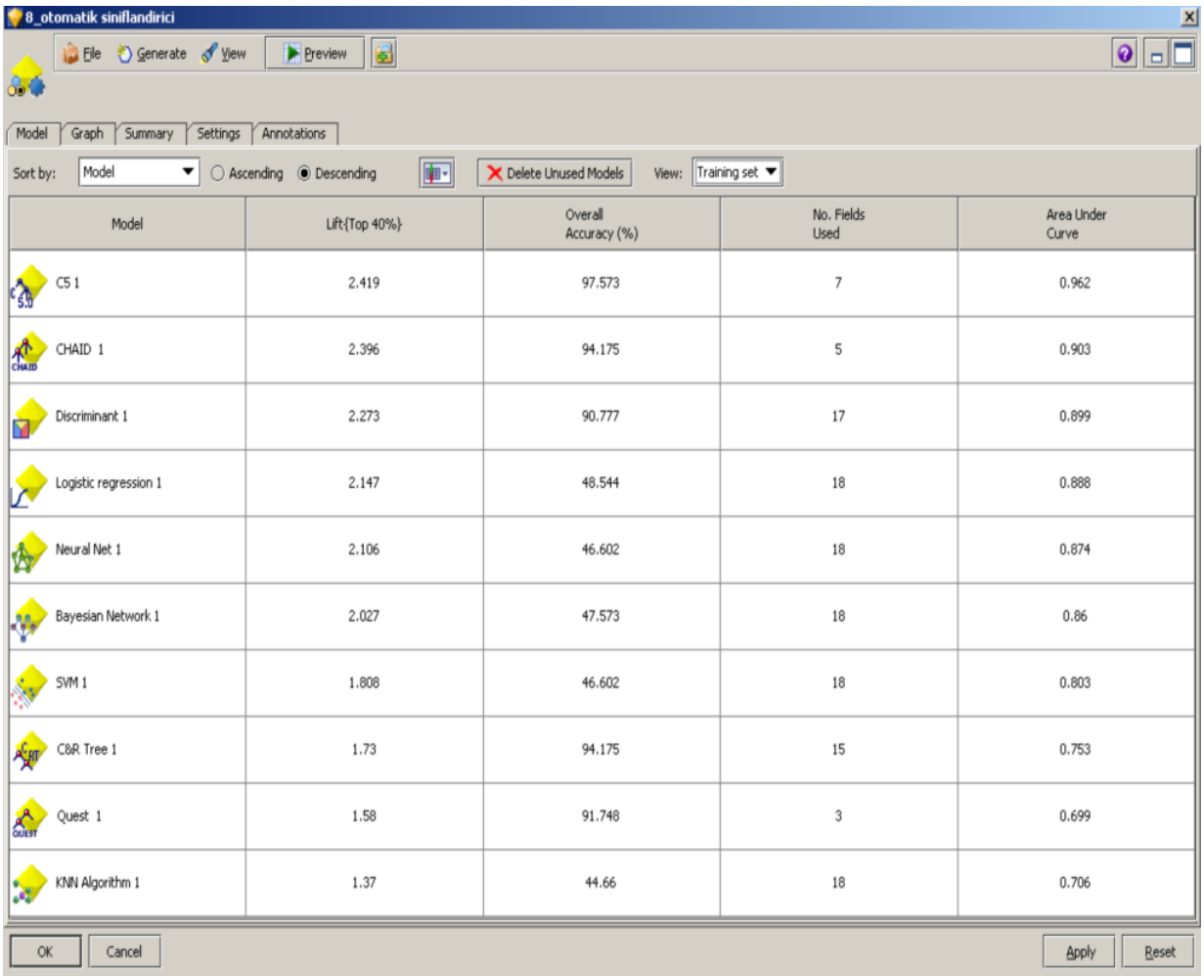


Şekil 4.9. Hedef değişken dağılımı

4.4.2. Uygun yöntemin belirlenmesi

AP risk sınıflandırma probleminin çözümünde en iyi performansı veren sınıflandırıcı algoritmasını tespit etmek için Otomatik Sınıflandırıcı (auto classifier) düğümünden yararlanılmıştır. C5.0, C&R, CHAID, QUEST, Diskriminant Analizi, Lojistik Regresyon,

Bayes Ağı, Destek Vektör Makinası, Yapay Sinir Ağı ve En Yakın Komşu sınıflandırıcıları ile elde edilen sonuçlar Şekil 4.10'da görülmektedir.



Model	Lift(Top 40%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
C5 1	2.419	97.573	7	0.962
CHAID 1	2.396	94.175	5	0.903
Discriminant 1	2.273	90.777	17	0.899
Logistic regression 1	2.147	48.544	18	0.888
Neural Net 1	2.106	46.602	18	0.874
Bayesian Network 1	2.027	47.573	18	0.86
SVM 1	1.808	46.602	18	0.803
C&R Tree 1	1.73	94.175	15	0.753
Quest 1	1.58	91.748	3	0.699
KNN Algorithm 1	1.37	44.66	18	0.706

Şekil 4.10. Sınıflandırıcı performansları

Bu sınıflandırıcıların performansı 3 kriter dikkate alınarak karşılaştırılmıştır:

- Kaldıraç Değeri (Lift ratio)
- Doğru Sınıflandırma Yüzdesi (Overall Accuracy)
- Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan (Area under Curve)

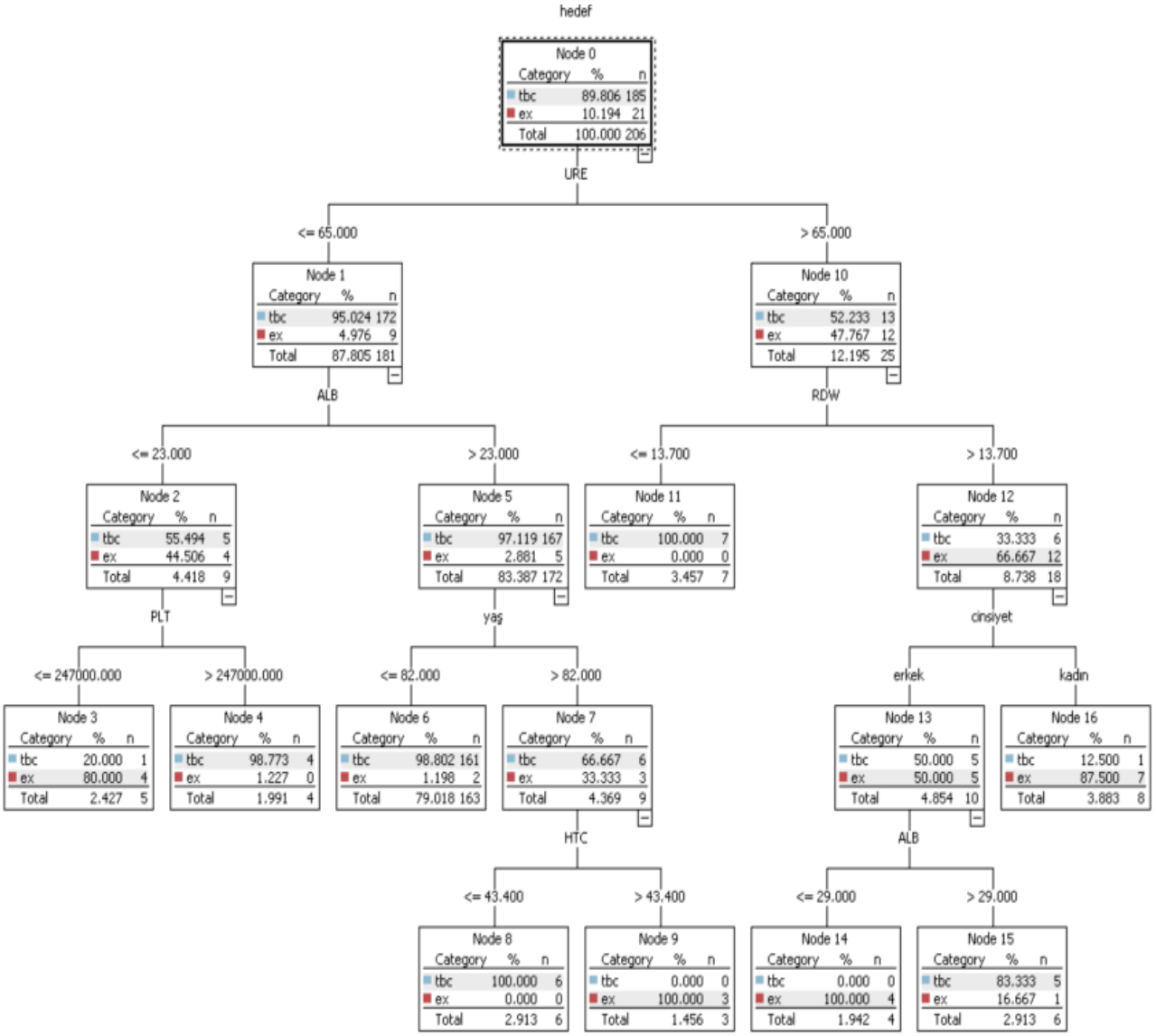
Söz konusu kriterlerle ilgili tanım bilgileri EK-1'de verilmektedir.

Kaldıraç değeri grafiğinde karar %40'lık değere karşılık gelen kaldıraç oranı ile alınmıştır. Şekil 4.10 incelendiğinde her üç kriter içinde en iyi performans C5.0 karar ağacı ile elde

edilmiştir. C5.0'ın bir diğer üstünlüğü de daha yüksek performansı daha basit bir karar ağacı (toplam 18 girdi değişkeninin 7'si ağaçta yer almıştır) ile elde etmiş olmasıdır.

4.4.3. Karar ağacının ve kuralların üretilmesi

C5.0 karar ağacı Şekil 4.11'de verilmektedir.



Şekil 4.11. C5.0 Karar Ağacı

Analize toplam 18 girdi değişkeniyle/nitelikle başlanmıştır. URE, ALB, RDW, PLT, YAŞ, CINSIYET ve HTC olmak üzere toplam 7 değişkenin karar ağacında yer aldığı, diğer 11

değişkenin yer almadığı, ağaçta yer alan bu değişkenlerle üretilen karar kurallarının yeterli tutarlılıkta olduğu görülmektedir. Başlangıç dallanma ÜRE niteliğinden olmuştur.

Sınıflandırıcının tutarlılığını gösterir C5.0 Karşılaştırma matrisi çizelge 4.1' de verilmektedir:

Çizelge 4.1. Karşılaştırma matrisi

		Tahmini Sınıf		**Toplam Örnek**
		0	1	
Gerçek Sınıf	0	183	2	185
	1	3	18	21

Çizelge 4.7 incelendiğinde, gerçekte "0" sınıfında olup, model tarafından da "0" sınıfına atanan veri sayısı 183 olarak belirlenmiştir. Yani 185 örnekten 183 tutarlı tahmin yapılmıştır. Yine gerçekte "1" sınıfında olup, model tarafından da "1" sınıfına atanan veri sayısı 18 olarak belirlenmiştir. Yani 21 örnekten 18 tutarlı tahmin yapılmıştır. Dolayısıyla sınıflandırıcının (C5.0 karar ağacı) tüm verinin %95,57 ((183+18)/206)sini doğru olarak sınıflandırdığı görülebilir.

C5.0 karar ağacı sınıflandırıcısından elde edilen kurallar listesi aşağıda listelenmektedir. Taburcu durumu için 5 adet kural, Ölüm/EX durumu için de 4 kural olmak üzere toplam 9 adet kural elde edilmiştir.

Taburcu durumu için C5.0 kuralları:

Kural 1:

ifbasURE<= 65
andbasALB<= 23
andbasPLT> 247,000
then 0.000

Eğer başlangıç üre değeri 65'ten küçük veya 65'e eşitse ve başlangıç albümin değeri 23'ten küçük veya 23'e eşitse aynı zamanda platelet değeri de 247000'den büyükse bu durumda kural hastanın taburcu olacağını göstermektedir.

Kural 2:

ifbasURE<= 65
andbasALB> 23
and yas <= 82
then 0.000

Eğer başlangıç üre değeri 65'ten küçük veya 65'e eşitse ve başlangıç albümin değeri 23'ten büyükse ve aynı zamanda hastanın yaşı 82 ise veya 82'den küçükse bu durumda kural hastanın taburcu olacağını göstermektedir.

Kural 3:

ifbasURE<= 65
andbasALB> 23
and yas > 82
andbasHTC<= 43.40
then 0.000

Eğer başlangıç üre değeri 65'ten küçük veya 65'e eşitse ve başlangıç albümin değeri 23'ten büyükse ve hastanın yaşı 82'den büyükse ve aynı zamanda başlangıç hematokrit değeri 43.40'dan küçük veya 43.40'a eşitse bu durumda kural hastanın taburcu olacağını göstermektedir.

Kural 4:

ifbasURE> 65
andbasRDW<= 13.70
then 0.000

Eğer başlangıç üre değeri 65'ten büyükse ve başlangıç alyuvar dağılım genişliği değeri 13.70'den küçük veya 13.70'e eşitse bu durumda kural hastanın taburcu olacağını göstermektedir.

Kural 5:

ifbasURE> 65
andbasRDW> 13.70
and cins = 1
andbasALB> 29
then 0.000

Eğer başlangıç üre değeri 65'ten büyükse ve başlangıç alyuvar dağılım genişliği değeri 13.70'den büyükse ve hasta erkekse (1) ve aynı zamanda başlangıç albümin değeri 29'dan büyükse bu durumda kural hastanın taburcu olacağını göstermektedir.

Ölüm/Ex durumu için C5.0 kuralları:

Kural 1:

ifbasURE<= 65
andbasALB<= 23
andbasPLT<= 247,000
then 1.000

Eğer başlangıç üre değeri 65'ten küçükse veya 65'e eşitse ve başlangıç albümin değeri 23'e eşit veya 23'den küçükse ve aynı zamanda başlangıç platelet değeri 247000'den küçükse veya 247000'e eşitse bu durumda kural hastanın Ex olacağını göstermektedir.

Kural 2:

ifbasURE<= 65
andbasALB> 23
and yas > 82
andbasHTC> 43.40
then 1.000

Eğer başlangıç üre değeri 65'ten küçükse veya 65'e eşitse ve başlangıç albümin değeri 23'den büyükse ve hastanın yaşı 82'den büyükse ve aynı zamanda başlangıç hematokrit değeri 43.40'dan büyükse bu durumda kural hastanın Ex olacağını göstermektedir.

Kural 3:

ifbasURE> 65
andbasRDW> 13.70
and cins = 1
andbasALB<= 29

then 1.000

Eğer başlangıç üre değeri 65'ten büyükse ve başlangıç alyuvar dağılım genişliği değeri 13.70'den büyükse ve hasta erkekse (1) ve aynı zamanda başlangıç albümin değeri 29'a eşit veya 29'dan küçükse, bu durumda kural hastanın Ex olacağını göstermektedir.

Kural 4:

ifbasURE> 65
andbasRDW> 13.70
and cins = 2
then 1.000

Eğer başlangıç üre değeri 65'ten büyükse ve başlangıç alyuvar dağılım genişliği değeri 13.70'den büyükse ve aynı zamanda hasta bayansa(2), bu durumda kural hastanın Ex olacağını göstermektedir.

4.5. Analiz Sonuçlarının Değerlendirilmesi

Analize toplam 18 girdi değişkeniyle/nitelikle başlanmıştır. Bu girdi değişkenleriyle sınıflandırıcı yöntemlerin performansları denenmiştir. Sınıflandırıcılar, 3 kriter dikkate alınarak karşılaştırılmıştır: a) Kaldıraç Değeri (Lift ratio), b) Doğru Sınıflandırma Yüzdesi (Overall Accuracy) ve c) Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan (Area under Curve). Kaldıraç değeri grafiğinde karar, %40'lık değere karşılık gelen kaldıraç oranı ile alınmıştır. Her üç kriter için de C5.0 karar ağacı yöntemi en iyi performans sergileyen yöntem olarak belirlenmiştir. Performans kriterlerinden kaldıraç değeri 2,419, Doğru sınıflandırma yüzdesi %97,573 ve AİK eğrisi altında kalan alan da 0,962 olarak elde edilmiştir. Kaldıraç değerinin 1'den büyük olması, ilgili sınıflandırıcı performansının hiç bir sınıflandırıcı modeli kullanılmaması durumunda elde edilecek sonuca göre ne kadar üstün olduğunu göstermektedir. AİK eğrisi altında kalan alanın bire yakın olması, o modelin performansının çok iyi olduğunu göstermektedir, zira bu alan en fazla 1 değerini almaktadır ve bu da mükemmel sınıflandırıcı anlamına gelmektedir. Aynı zamanda sınıflandırıcının, tüm verinin %95,57'sini doğru olarak sınıflandırdığı da görülmüştür. Toplam 18 girdi değişkenini 7'ye indirerek bu sonuçları elde etmesi, yani basit bir karar ağacına indirgemesi, C5.0 karar ağacı yönteminin performans üstünlüğünü göstermektedir.

C5.0 Karar ağacı yöntemiyle elde edilen kurallar, temin edilen veriler doğrultusunda teorik olarak doğru ve tutarlı sonuçlar/kurallar ürettiği ifade edilebilmekte olup, bu kurallar, hasta verilerinin temin edildiği uzman hekimlerle paylaşılmış ve elde edilen kuralların uygunluğu/tutarlılığıyla ilgili yorumlarına başvurulmuştur. Uzman görüşleri, taburcu durumu için elde edilen kurallardan kural 2'nin ve Ex durumu için elde edilen kurallardan da kural 3'ün önemli olduğu yönündedir ve iyi kurallar üretildiği değerlendirilmiştir. Elde edilen kuralların, hastalık üzerinde çalışan hekimlere önemli farkındalıklar ve açılımlar sağlayacağı düşünülmektedir.

5. SONUÇ VE ÖNERİLER

Bilgi teknolojilerindeki gelişmeler, bu çerçevede kuruluşların bilgi sistemlerini kurma gereklilikleri, çok fazla verinin elde edilmesine ve depolanmasına neden olmuştur. Söz konusu verilerdeki gizli örüntülerin (pattern) keşfedilmesi, onlardan anlamlı ve faydalı bilgilerin elde edilerek karar mekanizmalarına sunulması gerekliliği veri madenciliğinin önemli bir araştırma alanı olmasına neden olmuştur.

Tıbbi veri madenciliğinin amacı, veri madenciliği tekniklerini kullanarak tıbbi alandaki gizli bilgilerin çıkarılması, önemli örüntülerin keşfedilmesidir. Bu örüntülerin arkasındaki nedensel mekanizmalar tam olarak anlaşılamamış olsa dahi, örüntülerin tanımlanması mümkündür [5].

Başarılı bir tıbbi veri madenciliği uygulaması, tanı süreci, tedavi seçeneklerinin seçimi, hastalık sonucunun tahmin edilmesi vb. gibi klinik karar verme faaliyetleri ve personel tahmini, sigorta, demografik eğilimler, kalite güvence ve süreç etkinliği gibi sağlık hizmetlerinin sunumuna dair yönetimsel karar verme faaliyetlerinin desteklenmesinde etkili bir şekilde kullanılabilen biyomedikal ve sağlık bakım (healtcare) bilgilerini sağlayabilmektedir. Veri madenciliğinden elde edilen bilginin kullanımı, hastalara daha iyi hizmet verilebilmesi bakımından bu hizmeti sunanlara yardımcı olacaktır [4].

Çalışmada, Ankarada bir kamu hastanesindeki 206 adet Akut Pankreatit (AP) hastalarına ait verilerin kullanılarak, hastaların riskinin (mortalite/hayatta kalma) ortaya konulmasına yönelik olarak karar ağacı yöntemi tabanlı bir tıbbi veri madenciliği çalışması gerçekleştirilmiş ve elde edilen karar kuralları tartışılmıştır. Probleminin çözümünde IBM PASW (PredictiveAnalytics Software) Modeler 14.0 kullanılmıştır. Analize toplam 18 girdi değişkeniyle/nitelikle başlanmıştır. Uygulama kapsamında kullanılacak olan verilerin ait olduğu nitelikler (attributes), hastalık konusunda uzman hekimlerin, hastalık kapsamında değerlendirmek üzere hastalardan temin ettikleri ve hastalıkta belirleyici olduğunu düşündükleri nitelikleri içermektedir. AP risk sınıflandırma probleminin çözümünde, C5.0, C&R, CHAID, QUEST, Diskriminant Analizi, Lojistik Regresyon, Bayes Ağı, Destek Vektör Makinası, Yapay Sinir Ağı ve En Yakın Komşu sınıflandırıcıları arasından en iyi performansı veren sınıflandırıcının C5.0 olduğu görülmüştür. C5.0 karar ağacı yönteminin, sınıflandırıcıların değerlendirildiği üç kriterin tamamında da iyi performans sergilediği görülmüştür. Yöntem, tüm verinin %95,57'sini doğru olarak

sınıflandırmaktadır ve 18 niteliği, 7 niteliğe (URE, ALB, RDW, PLT, YAŞ, CINSIYET ve HTC) indirgeyerek karar ağacı oluşturmuştur. Başlangıç dallanma ÜRE niteliğindendir.

C5.0 karar ağacı sınıflandırıcısından, taburcu durumu için 5 adet kural, Ölüm/Ex durumu için de 4 kural olmak üzere toplam 9 adet kural elde edilmiştir. C5.0 Karar ağacı yöntemiyle elde edilen kurallar, temin edilen veriler doğrultusunda teorik olarak doğru ve tutarlı sonuçlar/kurallar ürettiği ifade edilebilmekte olup, bu kurallar, hasta verilerinin temin edildiği uzman hekimlerle paylaşılmış ve elde edilen kuralların uygunluğu/tutarlılığıyla ilgili yorumlarına başvurulmuştur. Uzman görüşleri, taburcu durumu için elde edilen kurallardan kural 2 nin ve Ex durumu için elde edilen kurallardan da kural 3'ün önemli olduğu yönündedir ve iyi kurallar üretildiği değerlendirilmiştir. Elde edilen kuralların, hastalık üzerinde çalışan hekimlere önemli farkındalıklar ve açılımlar sağlayacağı düşünülmektedir. Elde edilen karar kuralları, hastalara uygulanacak tedavi yöntemlerinin belirlenmesi ve doğru tedavinin hızlı bir şekilde öngörülmesi bakımından hekimlere önemli karar desteği sağlayabilecektir.

Söz konusu hastalıkla ilgili olarak ülkenin sağlık kuruluşlarının sahip olduğu verilerin tamamının merkezi bir sistemle toplanması ve bu önemli hastalığa dair geniş kapsamlı tıbbi veri madenciliği çalışmalarının gerçekleştirilmesi önerilmektedir.

KAYNAKLAR

1. Köktürk, F., Ankaralı, H., Sümbüloğlu, V. (2008). Veri madenciliği yöntemlerine genel bakış. *Türkiye Klinikleri Bioistatistik Dergisi*, 1(1), 20-25
2. Koyuncu, A., Gökgöz, Ş. (2001). Akut Pankreatitte Görülen Sistemik Komplikasyonlar. *C. Ü. Tıp Fakültesi Dergisi*, 23 (1), 65 - 72
3. Dunham, M.H. (2003). *Data mining: introductory and advanced topics*, (First edition). Prentice Hall, Pearson Education Inc.
4. Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., and Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. *Journal of Medical Systems*, 36, 2431-2448
5. Barati, E., Saraee,A., Mohammadi, A., Adibi, N., Ahamadzadeh,R. (2011). A survey on utilization of data mining approaches for dermotological (skin) diseases prediction. *Cyber Journals: Multidisiplinary Journals in Science and Techology*, Journal of Selected Areas in Healt Informatics (JSHI), March Edition. 1-11
6. Gökçen, H. (2011). *Yönetim bilgi/bilişim sistemleri: analiz ve tasarım* (Birinci Baskı). Türkiye: AFŞAR matbaacılık.
7. Türkiye Bilişim Derneği (TBD). (2010) Kamuda karar destek sistemlerinin kullanımı ve bir model önerisi. *Kamu Bilgi işlem Merkezleri Yöneticileri Birligi, Kamu Bilişim Platformu XII, Çalışma Grubu 2 Raporu* , Ankara
8. Han, J., Kanber (2001). *Data Mining: Concepts and Techniques*, (Second Edition). Morgan Kaufmann.
9. Özkan, Y. (2008). *Veri madenciliği yöntemleri* (Birinci Basım). Türkiye: PAPATYA Yayıncılık
10. Quinlan, J. R. (1979). *Discovering rules by induction from large collections of examples.* Michie, D. (Ed), Expert systems in the micro electronics age. Edinburgh University Press
11. Ian H. Witten, Frank Eibe, Mark A. Hall. (2011). *Data mining: practical machine learning tools and techniques* (Thirth Edition). Morgan Kaufmann Publishers, Elsevier Morgan Kaufmann Publishers is an imprint of Elsevier
12. Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 81-106

13. İnternet: *The ID3 Algorithm*, URL:http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm. http://www.webcitation.org/query?url=http%3A%2F%2Fwww.cise.ufl.edu%2F%7Eddd%2Fcap6635%2FFall-97%2FShort-papers%2F2.htm&date=2014-05-23, Son Erişim Tarihi: 23.05.2014
14. İnternet: *Building Classification Models: ID3 and C4.5*, URL:http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html. http://www.webcitation.org/query?url=http%3A%2F%2Fwww.cis.temple.edu%2F%7Eingargio%2Fcis587%2Freadings%2Fid3-c45.html&date=2014-05-23, Son Erişim Tarihi: 23.05.2014
15. Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo
16. Adhatrao, K., Gaykar, A., Dhavan,A., Jha,R., Honrao,V. (2013). Prediction students' performance using ID3 and C4.5 classification algorithms. *International Journal of DataMining and Knowledge Management Process (IJDKP)*, 3(5), 39-52
17. Bremian, L., Freidman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks / Cole Advanced Books and Software.
18. Silahtaroğlu, G. (2008). *Kavram ve algoritmalarıyla temel veri madenciliği* (Birinci Basım). Türkiye: PAPATYA Yayıncılık
19. Aytaç, M.B. (2013). *Doğrudan pazarlama için veri madenciliği çözümleri: Banka müşterileri üzerine bir uygulama*, Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Yönetim Bilişim Sistemleri ABD, Ankara.
20. Altıntaş, Y.Y. (2010). *Veri madenciliğinin tıpta kullanımı ve bir uygulama: hemodiyaliz hastaları için risk seviyelerine göre risk faktörlerinin etkileşimlerinin incelenmesi*, Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.
21. Özekes, S. (2003). Veri madenciliği modelleri ve uygulama alanları. *İstanbul Ticaret Üniversitesi dergisi*, 3, 65-82.
22. Albayrak, A.S., Yılmaz, Ş.K. (2009). Veri madenciliği: karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama. *Süleyman Demirel Üniversitesi İİBF Dergisi*, 14(1), 31-52.
23. Agrawal, R.,Imielinski, T., Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOID, International Conference on the Management of Data*. ACM Washington DC, 207-216

24. Sumathi, S., Sivananda, S.N. (2006). *Introduction to data mining and its application*, Springer, 178,629
25. Hilage, T.A., Kulkarni,R.V. (2012). Review of literature on data mining, *International* Journal *of Research and Reviews in Applied Sciences* 10(1), 107-114
26. Galvao, N.D., Marin, H.F. (2009). Data mining: a literature review. Acta *Paulista de Enfermagem*, 22(5), 686-690
27. Zhang, R., Katta, Y. (2002). Medical data mining. *Data Mining and Knowledge Discovery*, 305-308
28. Ichise, R., Nuamo Learning, M. (2001). First-order rules to handle medical data. *NII Journal*, 2, 9-14
29. Güllüoğlu, S.S. (2011). Tıp ve sağlık hizmetlerinde veri madenciliği çalışmaları: kanser teşhisine yönelik bir ön çalışma. *Online Academic Journal of Information Technology (AJIT)*, 2(5), 1-7
30. Kaya, E., Bulun, M., Arslan, A. (2003). Tıpta veri ambarları oluşturma ve veri madenciliği uygulamaları, *Akademik bilişim 2003*, Çukurova Üniversitesi, Adana.
31. Khan, S. M. , Islam, M.R., Chowdhury, M.U. (2004). Medical Image Classification Using an Efficient Data Mining Technique. *in Proceedings of International Conference on Machine Learning and Applications* (ICMLA'04), Louisville, KY, USA,
32. Wren, J., Garner, H. (2005). Data mining analysis suggests an epigenetic pathogenesis for Type II diabets. *Journal of Biomedicine and Biotechnology*, 2, 104-112.
33. Wang, S., Zhou, M., Geng, G. (2005). Application of fuzzy cluster analysis for medical image data mining. *Proceedings of the IEEE International Conference on Mechatronics and Automation*, Niagara Falls, Canada
34. Cheng, T.H., Wei, C.P., Tseng, V.S. (2006). Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 165-170
35. Bethel, C.L., Hall, L.O., Goldgof, D.B. (2006). Mining for Implications in Medical Data. *18th International Conference on Pattern Recognition (ICPR'06)* (1),1212-1215

36. Xue, W., Sun, Y., Lu, Y. (2006). Research and Application of Data Mining in Traditional Chinese Medical Clinic Diagnosis. *Signal Processing*, 8th International Conference on (4), Beijing.
37. Aftarczuk, K. (2007). *Evaluation of data mining algorithms implementing in medical decision support system*. MSc Thesis, School of Engineering at Blekinge Institute of Technology, Ronneby, Sweden.
38. Xing, Y., Wang, J., Zhao, Z., Gao, Y. (2007). Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. *Convergence Information Technology*, International Conference,
39. Floyd, S. (2007). *Data mining techniques for prognosis in pancreatic cancer*. MSc. Master Thesis, Worcester Polytechnic Institute, Computer Science
40. Potter, R. (2007). Comparison of classification algorithms applied to breast cancer diagnosis and prognosis, advances in data mining. *7th Industrial Conference, ICDM 2007*, Leipzig, Germany, 40-49.
41. Bach, M.P., Cosic, D. (2008). Data mining usage in healt care management: literature survey and decision tree application. *Medicinski Glasnik*, 5(1), 57-64
42. Wu, X., Kumar,V., Quinlan,J.R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. *Knwledge Information Systems*, 14, 1-37
43. Lavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R. (2009). Clinical data mining: a review. *IMIA Yearbook of medical Informatics*, 1-3
44. Tu, M.C., Shin, D., Shin, D. (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, 183-187
45. Sapna, S., Tamilarasi, A. (2010). Data mining fuzzy neuro in predicting diabetes. *Int. J. Computational Intelligence Research*, 6(3), 431-442
46. Morra, J.H., Tu, Z.,Apostolova, L.G., Green, A.E., Toga, A.W., Thompson, P.M. (2010). Comparison of Adaboost and suppot vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. *IEEE Transactions* on *Medical Imaging*, 29(1), 30-43
47. Lopes, R., Ayache, A., Makni, N., Puech, P., Villers, A., Mordon, S., et al. (2011). Prostate cancer characterization on MR images using fractal features. *Medical Physics*, 38, 83-95

48. Soni, J., Ansari, U., Sharma, D., Soni, S. (2011). Predictive data mining for medical diagnosis: an overview of heart disease. *International Journal of Computer Applications*, 17(8), 43-48
49. Padhy, N., Mishra, P., Panigrahi, R. (2012). The survey of data mining applications and feature scope. *Int. J. Comp. Sci. Engin. and Inf. Techn. (IJCSEIT),* 2(3), 43-58
50. Kolçe, E., Frasheri, N. (2012). A literature review of data mining techniques used in healtcare databases. *ICT innovations 2012 Web Proceedings*, Poster sessions, 577-582.
51. Khaleel, M.A., Pradham, S.K., Dash, G.N. (2013). A survey of data mining techniques on medical data for finding locally frequent diseases. *International Journal* of *Advanced Research* in *Computer Science* and *Software Engineering (IJARCSSE)*, 3(8), 149-153
52. Vijiyarani, S., Sudha, S. (2013). Disease prediction in data mining technique-A survey. *International Journal* of *Computer Applications and Information Technology (IJCAIT)*, 2(1), 17-21
53. Taşbaş, E., Çalık, E., Dicle, O. (2012). Kronik pankreatit evrelemesinde karar destek sistemi uygulaması. *IX. Ulusal Tıp Bilişimi Kongresi*, 15-17 kasım 2012, Antalya
54. İnternet:*Pankreas*,URL:http://tr.wikipedia.org/wiki/Pankreas#B.C3.B6l.C3.BCmleri.http://www.webcitation.org/query?url=http%3A%2F%2Ftr.wikipedia.org%2Fwiki%2FPankreas%23B.C3.B6l.C3.BCmleri&date=2014-05-23, Son Erişim Tarihi: 23.05.2014
55. Demiral,G., ve ark. (2011). Akut pankreatitli hastalarımızın retrospektif olarak değerlendirilmesi. *Göztepe Tıp Dergisi*, 26(1), 4-9
56. İnternet: Engin, B. *Akut Pankreatit: Nedenleri, Belirtileri, Tanısı ve Tedavisi* URL:http://www.xn--salk-1wa3i.net/akutpankreatit.html. http://www.webcitation.org/query?url=http%3A%2F%2Fwww.xn--salk-1wa3i.net%2Fakutpankreatit.html&date=2014-05-23, Son Erişim Tarihi: 23.05.2014.
57. Pekmezci, S. (2002). Akut pankreatitte yaklaşım ve tedavi. *Hepato-Bilier Sistem ve Pankreas Hastalıklar Sempozyum Dizisi No: 28,* 239-262
58. İnternet: *SPSS Modeler*, URL:http://en.wikipedia.org/wiki/SPSS_Modeler. http://www.webcitation.org/query?url=http%3A%2F%2Fen.wikipedia.org%2Fwiki%2FSPSS_Modeler&date=2014-05-23, Son Erişim Tarihi: 23.05.2014.

59. Tan P.N. Steinbach M, Kumar V. (2006). *Introduction to Data Mining*. *Pearson*.

60. Bramer, M. (2007). *Principles of Data Mining*. *Springer*, 173-185

EKLER

EK-1. Sınıflandırıcıların performans kriterleri

- Kaldıraç Değeri (Lift ratio)
- Doğru Sınıflandırma Yüzdesi (Overall Accuracy)
- Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan (Area under Curve)

Sınıflandırıcı performansı, model tarafından tahmin edilen doğru ve yanlış sayısı hesaplanarak bulunur. Bu hesaplar aşağıda karşılaştırma matrisi olarak adlandırılan Çizelge E.1'de gösterilir.

Çizelge E.1. Karşılaştırma matrisi formatı

		Tahmini Sınıf		**Toplam Örnek**
		+	-	
Gerçek Sınıf	+	DP	YN	P= DP+YN
	-	YP	DN	N= YP+DN

Karşılaştırma matrisi 4 hücreden oluşmaktadır ve hücrelerin anlamları aşağıdaki gibidir. DP, pozitif sınıf olarak sınıflandırılan pozitif örnek sayısı; YP, pozitif sınıf olarak sınıflandırılan negatif örnek sayısı; YN, negatif sınıf olarak sınıflandırılan pozitif örnek sayısı; DN, negatif sınıf olarak sınıflandırılan negatif örnek sayısı; P pozitif toplam ve N negatif toplam örnek sayısıdır.

Karşılaştırma matrisi model sonuçlarını özetlese de matrisin sonuçlarını tek bir sonuca dönüştürerek performans karşılaştırması yapmak daha uygun olacaktır. Model performanslarının karşılaştırılmasında şu ölçütler kullanılmıştır: DSY, Hata oranı, Doğru pozitif oranı, Kesinlik değeri, F1 skoru, Yakınlık katsayısı ve AİK eğrisi altında kalan alan.

Doğru sınıflandırma yüzdesi (DSY)

DSY hesaplaması aşağıdaki verilmiştir: [59,60]:

$$\text{DSY}=\frac{\text{Toplam doğru tahmin sayısı}}{\text{Toplam tahmin sayısı}}=\frac{\text{DP+DN}}{\text{P+N}}$$

EK-1. (devam) Sınıflandırıcıların performans kriterleri

Doğru pozitif oranı

Doğru pozitif (DP) oranı, gerçek sınıfı DP olarak sınıflandırılan örneklerin, gerçek tüm pozitif (P) sınıfların toplamına oranıdır. Yüksek DP oranı iyidir; çünkü gerçekte pozitif sınıfta yer alması gereken örneklerin yanlış sınıflandırılması önlenecektir. DP oranı aşağıdaki gibi hesaplanır [59,60].

$$\text{DP oranı} = \frac{\text{DP}}{\text{DP+YN}} = \frac{\text{DP}}{\text{P}}$$

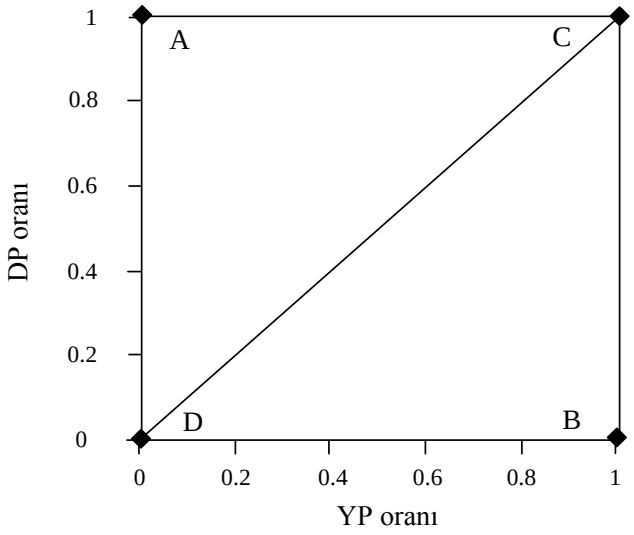
Alıcı işletim karakteristiği (AİK) grafiği

Alıcı İşletim Karakteristiği (AİK) grafiği farklı sınıflandırıcı performanslarını karşılaştırmak için geliştirilen yöntemlerdendir. AİK grafiğinde, DP oranı dikey eksende ve yanlış pozitif (YP) oranı yatay eksende gösterilir.

$$\text{YP oranı} = \frac{\text{YP}}{\text{DN+YP}}$$

AİK grafiğinde gösterilen her bir (x, y) noktası sırasıyla YP oranını ve DP oranını göstermektedir. AİK grafiği üzerindeki noktalar (0, 1), (1, 0), (1, 1) ve (0, 0) Şekil E.1'de gösterilmiştir. Bu noktalar dört özel duruma (A, B, C ve D) karşılık gelmektedir. A, mükemmel sınıflandırıcı olarak tanımlanmaktadır ve grafik üzerinde, sol üst köşe, en iyi noktaya karşılık gelmektedir. B, en kötü olası sınıflandırıcı olarak tanımlanmaktadır ve grafik üzerinde sağ alt köşede yer almaktadır. C, ultra-liberal sınıflandırıcı olarak tanımlanmaktadır, anlamı sınıflandırıcının nesneleri her zaman pozitif sınıfa ayırması durumudur; D, ultra-muhafazakar sınıflandırıcı olarak tanımlanmaktadır, anlamı ise sınıflandırıcının nesneleri her zaman negatif sınıfa ayırması durumudur. Sonuçta, sol üst köşeye yakın olan grafik daha iyi sınıflandırıcı performansına sahiptir [60].

EK-1. (devam) Sınıflandırıcıların performans kriterleri

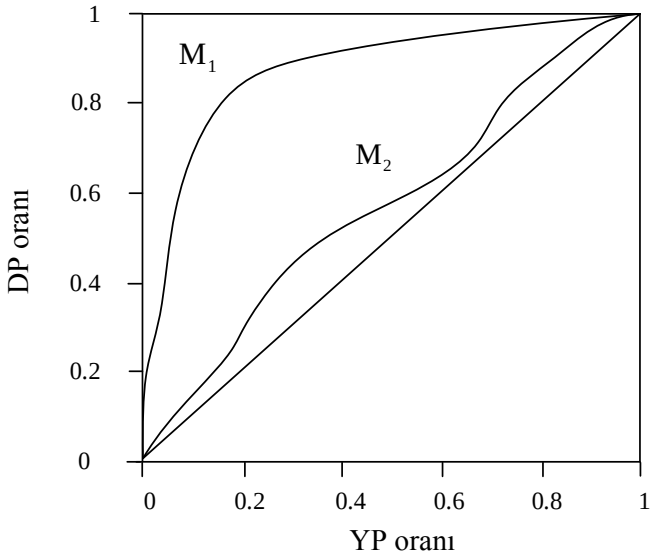


Şekil E.1. AİK grafiğinde önemli noktalar [60]

AİK eğrisi altında kalan alan

AİK grafiği sınıflandırıcı performansını görsel olarak yansıtmada iyidir; fakat sayısal olarak modellerin karşılaştırılması da gerekmektedir. AİK eğrisi altında kalan alanın hesaplanması ile sınıflandırıcı modellerin performansının karşılaştırılması daha uygun olmaktadır. Eğri altında kalan alan bire yakın ise o modelin performansı daha iyidir. AİK eğrisi altında kalan alan en fazla 1 değerini almaktadır, bu mükemmel sınıflandırıcı anlamına gelmektedir. Şekil E.2'de iki farklı modele ait AİK eğrileri bulunmaktadır. Şekilde görüldüğü üzere M_1 modeli daha iyi performansa sahiptir. M_1 modelinin AİK eğrisi altında kalan alan, diğer modelin (M_2) AİK eğrisi altında kalan alandan büyüktür [59].

EK-1. (devam) Sınıflandırıcıların performans kriterleri

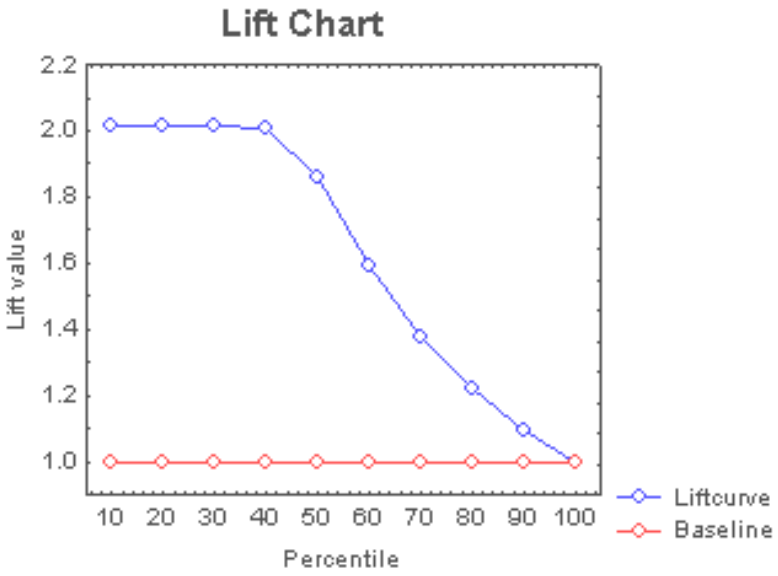


Şekil E.1. Farklı modeller için AİK eğrileri [59]

Kaldıraç (Lift) Grafiği

Kaldıraç oranı, sınıflandırıcı ile tahmin edilen hedef değerin ilgilenen değerinin (hit) ele alınan yüzdelik değer (quantile) içerisindeki oranının hedef değerin ilgilenilen değerinin tüm veri içerisindeki oranına bölümünü ifade eder. Bölme işlemi sonucu elde edilen değerin birden büyük olması ilgili sınıflandırıcı performansının rassal sınıflandırıcı modele (hiç bir sınıflandırıcı modeli kullanılmaması durumunda elde edilecek sonuç) göre ne kadar üstün olduğunu gösterir. Bu nedenle, kümülatif olarak çizilen bu grafikte (x ekseni yüzde, y ekseni kaldıraç oranı), üstte seyreden sınıflandırıcıya ait eğri, o sınıflandırıcının performansının yüksek olduğunu belirtir (Şekil E.3).

EK-1. (devam) Sınıflandırıcıların performans kriterleri



Şekil E.2. Kaldıraç grafiği

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	:	GÖKÇEN ALIÇ, Zeynep Hilal
Uyruğu	:	T.C.
Doğum tarihi ve yeri	:	23/06/1986 Ankara
Medeni hali	:	Evli
Telefon	:	05070443708
Faks	:	
e-posta	:	zeynephilalgokcen@gmail.com

Eğitim Derecesi	Okul/Program	Mezuniyet yılı
Yüksek lisans	Gazi Üniversitesi/Yönetim Bilişim Sist.	Devam ediyor
Lisans	Başkent Üniversitesi/Biyomedikal Müh.	2009
Lise	Yıldırım Beyazıt Anadolu Lisesi	2004

İş Deneyimi, Yıl	Çalıştığı Yer	Görev
2013-Devam ediyor	Sağlık Bakanlığı	Ürün Denetmen Yard.

Yabancı Dili

İngilizce

Hobiler

Kitap okuma, spor

GAZİ GELECEKTİR...