KIRIKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİNİN NEFROLOJİ ALANINA UYGULANMASI

Yunus KÖKVER

HAZİRAN 2012

Bilgisayar Mühendisliği Anabilim Dalında Yunus KÖKVER tarafından hazırlanan VERİ MADENCİLİĞİNİN NEFROLOJİ ALANINA UYGULANMASI adlı Yüksek Lisans Tezinin Anabilim Dalı standartlarına uygun olduğunu onaylarım.

Prof. Dr. Hasan ERBAY
Anabilim Dalı Başkanı

Bu tezi okuduğumu ve tezin **Yüksek Lisans Tezi** olarak bütün gereklilikleri yerine getirdiğini onaylarım.

Doç. Dr. Necaattin BARIŞÇI
Danışman

Jüri Üyeleri

Başkan	: Prof. Dr. Hasan ERBAY	_____
Üye (Danışman)	: Doç. Dr. Necaattin BARIŞÇI	_____
Üye	: Yrd. Doç. Dr. Taner TOPAL	_____

27/06/2012

Bu tez ile Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onaylamıştır.

Doç. Dr. Erdem Kamil YILDIRIM
Fen Bilimleri Enstitüsü Müdürü

ÖZET

VERİ MADENCİLİĞİNİN NEFROLOJİ ALANINA UYGULANMASI

KÖKVER, Yunus
Kırıkkale Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi
Danışman: Doç. Dr. Necaattin BARIŞÇI
Haziran 2012, 82 sayfa

Günümüzde bir hastalığa yapılacak doğru ve hızlı tespit büyük önem taşımaktadır. Hekimlere etkin bir şekilde teşhis koyabilmeleri hususunda yardımcı olabilmek için, veri madenciliği son zamanların en gözde yöntemlerinden birisidir. Bu tez çalışmasında retrospektif yöntemle 150 hastadan alınan veriler, veri madenciliği sınıflandırma algoritmalarıyla incelenmiştir. Normal veya Hasta olacak şekilde iki farklı sınıf vardır. Böylelikle hipertansiyon hasta adaylarının hipertansiyon olup olmadığını tahmin edecek bir teşhis sistemi geliştirilmiştir. Ayrıca elde edilen sonuçlara göre bir karar ağacı oluşturularak, hipertansiyona doğrudan ve dolaylı olarak etki eden faktörler belirlenmiştir.

Bu çalışma veri madenciliğinin hipertansiyon alanında da faydalı bir araç olabileceğini ortaya koymuştur. Böylece veri madenciliği, tedavi karar aşamasında doktorun kısa sürede objektif kararlar almasına yardımcı olabilecektir.

Anahtar Kelimeler: Veri Madenciliği, Nefroloji, Hipertansiyon, Karar Tablosu, Saf Bayes, IB1, J48, Çok Katmanlı Algılayıcı

ABSTRACT

APPLICATION OF DATA MINING TO THE FIELD OF NEPHROLOGY

KÖKVER, Yunus
Kırıkkale University
Graduate School of Natural and Applied Sciences
Depertment of Computer Engineering, M.Sc. Thesis
Supervisor: Assoc. Prof. Dr. Necaattin BARIŞÇI
June 2012, 82 pages

Nowadays, accurate and instant detection of an illness is very significant. Data mining is a popular technique that enables doctors to diagnose illnesses efficiently. In this study, data gathered from 150 patients are analyzed with data mining classification algorithms. There are two different classes which are normal and ill. Thus, a diagnostic system is developed which predicts whether a candidate patient has hypertension or not. Besides, a decision tree is created and factors affecting hypertension directly and indirectly are determined.

This study shows that data mining is a useful tool in the field of hypertension. Thus, data mining can help doctors for making objective decisions in treatment.

Keywords: Data Mining, Nephrology, Hypertension, Decision Table, Naive Bayes, IB1, J48, Multilayer Perceptron

TEŞEKKÜR

Tezimin hazırlanması esnasında yardımlarını esirgemeyen tez danışmanım Sayın Doç. Dr. Necaattin BARIŞÇI'ya, Veri Madenciliği konusunda engin bilgilerinden yararlanma fırsatı bulduğum hocam Sayın Doç. Dr. Suat ÖZDEMİR'e, kullandığım verilerin elde edilmesindeki yardımlarından ve fedakârlıklarından dolayı hakkını ödeyemeyeceğim Uzman Doktor Aydın ÇİFTÇİ'ye, bu çalışmanın ana taslağını oluşturmamızda yardımlarıyla bize ışık tutan Doç. Dr. Yakup EKMEKÇİ'ye, yardımlarını hiçbir zaman esirgemeyen hocam Sayın Prof. Dr. Hasan ERBAY'a, tezin yazımı sırasında bana yardımcı olan ve benimle beraber uykusuz kalan arkadaşım Volkan ATEŞ'e teşekkürlerimi sunarım.

Varlıkları ve destekleriyle her zaman yanımda olan aileme ayrıca teşekkürü bir borç bilirim. İyi ki varsınız.

İÇİNDEKİLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

ŞEKİL Sayfa

ÇİZELGELER DİZİNİ

ÇİZELGE Sayfa

1. GİRİŞ

Hipertansiyon hastalığı, günümüzde önemli bir sağlık sorunu olarak karşımıza çıkmaktadır. Hipertansiyonu olan bir hasta yıllarca bu hastalığı fark etmeyebilir. Hipertansiyon yavaş ve sinsice yıllarca herhangi bir belirti vermeden hastanın böbrek, kalp ve diğer damarlarına zarar verebilir. Toplumlarda hipertansiyona maruz kalma sıklığı yaşla beraber artar. İlerleyen yıllarda çoğu kimsede spot olarak yapılan ölçümlerde rastlanabilir. Bu bağlamda hipertansiyonun erken teşhis ve tedavisi büyük önem arz etmektedir. Bu önemin sebebi hipertansiyonun organlarda yaptığı hasar ve bunun sonucunda ortaya çıkan komplikasyonların tedavi maliyetleri ile iş gücü kaybının yüksek olmasıdır.

Hipertansiyonu olan hastaların yaklaşık %95'inde herhangi bir neden bulunmaz, ancak risk faktörlerinden söz edilebilir. Bu gruba bilim dilinde esansiyel (primer), günlük kullanım dilinde sinirsel (asabi) hipertansiyon denilir. %5'lik grupta kalan hastalarda hipertansiyonu başlatan bir neden söz konusudur. İkincil (sekonder) hipertansiyon denilen bu grubun büyük kısmını böbrek ve böbrek damar hastalıkları oluşturmaktadır. Bu hasta grubu ile Nefroloji bilim dalı ilgilenmektedir [1].

Ülkemizde hipertansiyon konusunda henüz yeterince farkındalık oluşmamıştır. Bunun sonucu olarak hipertansiyon tedavi ve komplikasyonları önlemede kritik önem arz eden erken teşhis istenilen düzeyde değildir. Bu açıdan bakılacak olursa erken teşhis çok önemlidir. Hipertansiyonun erken teşhisinde kamuoyu oluşturma yanında yeni metotlara da ihtiyaç vardır. Bu çalışmada, hipertansiyonun erken teşhisinde yeni bir yaklaşım olarak düşünülebilecek Veri Madenciliği yöntemleri incelenmiştir.

Özellikle tıp alanındaki verinin büyüklüğü ve hayati önem taşıması bu alandaki uygulamaları daha da önemli kılmaktadır. Tıp alanında veri madenciliği, tıbbi verilerin heterojen yapıda olması, özel etik ve hukuki kurallar gerektirmesi ve hasta sırlarını temel alan kurallar içermesi, istatistik metotların bu heterojenite ve sosyal

konuları adres etmek zorunda olması ve tıbbın insan hayatında özel bir yerinin olması gibi nedenlerle diğer alanlardan farklılık gösterir.

Veri madenciliğinde birinci ve en basit analitik adım, tanımlayıcı istatistikleri kullanarak, grafik ve şekillerle görsel inceleme yaparak ve değişkenler arası potansiyel anlamlı bağlantılara bakarak veriyi tanımlamaktır. Veri madenciliği ile istatistik disiplini, sınıflama ve yapı tanımlama problemleri üzerine çalışmaktadır. Bilgisayarların gücündeki artış ve fiyatlarının düşmesi, mümkün olan çözümlerin incelenmesi yoluyla ortaya çıkan yeni tekniklerin gelişmesine imkân sağlamıştır. Yeni teknikler içinde yapay sinir ağlarının algoritmalarına benzer yeni algoritmalar, karar ağaçları ve diskriminant analizi gibi eski algoritmalara yeni yaklaşımların getirildiği teknikler yer alır. Bu yöntemlerin çoğunluğu tıpta tanı koyma ve sınıflama amaçlarıyla kullanılmaktadır [2].

Sağlık ve tedavi verilerinin elektronik ortamlarda saklanması sonucunda, tıp alanında cevapları Veri Madenciliği teknikleri ile bulunabilecek sorular ve uygulama alanları şunlardır:

- Hastalıkları etkileyen faktörlerin ortaya çıkartılması.
- Hastalıklara erken teşhis koyularak sağlığın korunması ve doğru tedavi yöntemlerinin seçilmesi.
- Sağlık hizmetlerinin kalitesinin artırılması ve geleceğe dönük doğru sağlık politikalarının oluşturulması.
- Koruyucu hekimliğin yaygınlaştırılması ve sağlık harcamalarının düşürülmesi.
- Salgın hastalıkların tespit edilmesi gerekli önlemlerin alınması.
- Sağlık harcamalarındaki hileli işlemlerin ortaya çıkartılması, maliyetlerin düşürülmesi.
- İlaç geliştirici firmaların, sağlık veri tabanlarından yararlanarak doğru ilaçları geliştirmesi.
- Sağlık hizmetlerinde kalitenin artırılması [3].

Günümüzde tıp literatürü üzerinde yapılan veri madenciliği çalışmaları aşağıda kısaca özetlenmektedir.

A. Kusiak ve arkadaşları tarafından akciğer deki tümörün iyi huylu olup olmadığına dair, karar destek amaçlı bir çalışma yapılmıştır. İstatistiklere göre Amerika da 160.000 den fazla akciğer kanseri vakasının olduğu ve bunların %90'ının öldüğü belirlenmiştir. Bu bağlamda bu tümörün erken ve doğru olarak teşhisi önem kazanmaktadır. Beden dışı (non-invasive) testler ile elde edilen bilgi sayesinde %40-60 oranında doğru teşhis konabilmektedir. İnsanlar kanser olup olmadıklarından emin olmak için biyopsi yaptırmayı tercih etmektedirler. Biyopsi gibi beden içi (invasive) testler hem maliyeti yüksek hem çeşitli riskler taşımaktadır. Farklı yerlerde ve farklı zamanlarda kliniklerde toplanan beden içi test verileri arasında yapılan veri madenciliği çalışmaları teşhiste %100 oranında doğruluk sağlamıştır [4].

Yine A. Kusiak ve arkadaşları tarafından hemodiyalize bağlı böbrek hastalarının sonuçlarını düzeltmek ve bakımının maliyetini düşürmek için çalışmalar yapılmıştır. Bu araştırmada ölçülen bu birçok parametrenin hastanın hayatta kalması ile etkileşimi; veri önişleme, veri dönüşümü ve veri madenciliği yaklaşımı ile ortaya çıkarılmaktadır [5].

Kore Tıbbi Sigorta Kurumu (The Korea Medical Insurance Corporation) tarafından hazırlanan bir veri tabanı üzerinde yüksek tansiyon ile ilgili bir çalışma yapılmıştır. Bu çalışma 1998 yılına ait 127886 kayıt üzerinde yapılmıştır. İlk aşamada yüksek tansiyona sahip 9103 kayıt üzerinde, daha sonra aynı sayıda yüksek tansiyonu olmayan kayıtlar üzerinde çalışılmıştır. Bu örnek 13689 kayıttan oluşan öğrenme ve 4588 kayıttan oluşan test setine bölünerek modelin eğitimi yapılmıştır. Öğrenim algoritmasında karar ağaçları algoritmalarından CHIAD, C4.5, C5.0 kullanılmıştır. Bu çalışmalar sonucunda yüksek tansiyon tahmininde etkili değerler BMI, idrar proteini (urinary protein), kan glikozu, kolesterol değerleridir. Yaşam koşullarının (diyet, alınan tuz miktarı, alkol, tütün gibi) hiçbirinin tahminde etkili olmadığı ayrıca grafiksel değerlerde de yalnızca yaşın etkili olduğu saptanmıştır [6].

Razali ve Ali [7], Malezya'da farklı sağlık merkezlerinden toplanan hasta verileri üzerinde veri madenciliğinin C5.0 karar ağacı algoritmasını uygulayarak hastalarda

en sık rastlanan akut üst solunum yolları enfeksiyonu için bir tedavi planı geliştirme modeli oluşturmuştur.

Persson ve Lavesson [8], hasta verileri ile cerrahi operasyon verilerinden hareketle hastanenin cerrahi ihtiyaçlarını (cerrahi müdahale olasılığı, ameliyat süresi, iyileşme süresi vb.) belirlemekte kullanılacak bir otomatik tahmin modelinin veri madenciliği teknikleri yardımıyla geliştirilme olanaklarını incelemiştir.

Tsumoto ve Hirano [9], çalışmasında tıbbi müdahale hatalarının nedenlerine ilişkin analiz yapmak amacıyla tıbbi hata verilerinden oluşan bir veri kümesi üzerine karar ağacı algoritması kullanmıştır. Sonuçta hemşirelerin çalışma saatlerinin tıbbi müdahale hatalarının ortaya çıkışında etkisi olduğu, ayrıca hemşirelere yardımcı olmayan hastaların hemşirelerin iş konsantrasyonunu azaltmaları ile tıbbi müdahale hatalarına yol açtığı tespit edilmiştir.

Yeh ve Wu [10], veri madenciliği teknikleri (karar ağacı ve birliktelik kuralları analizi) ile zaman soyutlama işlemini birleştirerek hastaneye yatma oranının tahmini için hemodiyaliz hastalarının biyokimyasal verilerini analiz eden bir karar destek sistemi geliştirmiştir.

Huang ve arkadaşları [11], kronik hastalıkların tahmin ve teşhisinde yardımcı olacak veri madenciliği teknikleri ile durum tabanlı muhakemeyi bir araya getiren bir karar destek sistemi önermiştir. Sistemin amaçları arasında sağlık tarama verilerinin veri madenciliği teknikleri (sınıflandırma ve diğer) ile analizi sonucunda elde edilecek kuralların kronik hastalık tahmininde kullanımı, yorumlanan verilerden kronik hastalıkların teşhisinde yararlanılması, durum tabanlı muhakeme ile kronik hastalıkların teşhis ve tedavisine yardımcı olunması yer almaktadır.

Wren ve Garner [12], Tip II Diyabet hastalığını veri madenciliği yardımıyla incelemiş ve vücut içinde epigenetik (genetik olmayan ırsi) değişimlerin Tip II diyabetin ortaya çıkmasında etkisi olduğunu belirtmişlerdir.

Le Duff ve arkadaşları [13] çalışmasında kalp durması geçiren hastaların hastaneden taburcu olduktan sonra kurtulma oranını etkileyen en önemli faktörün tespiti ve hasta profillerinden kurtulma olasılığının çıkarımını sağlayan bir model geliştirmek için Bayes ağları tekniğini kullanmıştır.
Palaniappan ve Awang [14], çeşitli veri madenciliği teknikleri (Karar ağaçları, Saf Bayes ve Sinir ağları) kullanarak geleneksel karar destek sistemlerinin cevaplayamadığı "what if" (ya olursa) sorgularına cevap verebilen Akıllı Kalp Hastalığı Tahmin Sistemi (Intelligent Heart Disease Prediction System- IHDPS) adlı prototip bir sistem geliştirmiştir.

Antoine ve arkadaşları [15] çalışmasında dijital meme filmlerinden tümör tespit edilmesini sağlamak amacıyla sinirsel ağlar ve birliktelik kuralları algoritmaları gibi çeşitli veri madenciliği tekniklerinin kullanım imkanlarını incelemiştir.

Sleeman ve arkadaşları [16], hemodiyaliz seanslarında hemodiyaliz cihazları tarafından toplanan ve hastanın süreçteki kan basıncı, kan değerleri ve diğer fizyolojik durumunu belirten veriler üzerinde hiyerarşik kümeleme ve Bayes Ağları gibi makine öğrenimli teknikler kullanarak analiz etmiş ve sonuçların tekil hastaların klinik durumları ile ilgisini araştırmıştır.

Almazyad ve arkadaşları [17], hipertansiyonla ilgili 15-64 yaş arası erkeklerde yaptıkları çalışmada, belirledikleri 5 tedavi türünün, hangi yaş grubunda daha etkili olduğu sonuçlarının analizinde veri madenciliği algoritmalarından yararlanmışlardır.

Cheng-Ding Chang ve arkadaşları[18],ortak risk faktörlerine göre Hipertansiyon ve Hiperlipidemi çoklu hastalık tahmini modellemesi için Veri Madenciliği teknikleri kullanmışlardır. Bu çalışma aynı anda Hipertansiyon ve Hiperlipidemi hastalık tahmini için iki aşamalı bir analiz yöntemi önermektedir. Öncelikle, bu iki hastalığın bireysel risk faktörlerini belirlemek için altı veri madenciliği yaklaşımı kullanılmıştır ve sonra oy ilkesi ile ortak risk faktörleri belirlenmiştir. Önerilen analiz yöntemi, bu iki hastalığın ortak risk faktörlerinin Sistolik Kan Basıncı, Trigliserid, Ürik Asit, Glutamat Piruvat Transaminaz(GPT) ve cinsiyet olduğunu göstermektedir.

Mevlut Türe ve arkadaşları [19], yaptıkları çalışmada, hipertansiyon hastalığı riskini tahmin etmek amacıyla, sınıflandırma teknikleri performansı karşılaştırılmıştır. 694 veri üzerinde retrospektif bir analiz yapılmıştır. 3 Karar Ağacı, 4 İstatistiksel Algoritma ve 2 Yapay Sinir Ağı performansları karşılaştırılmıştır. Belirleyici değişkenler; yaş, cinsiyet, hipertansiyon aile öyküsü, sigara alışkanlığı, lipoprotein(a), trigliserid, ürik asit, total kolesterol ve vücut-kütle indeksidir. Yapay Sinir Ağı algoritması olan Çok Katmanlı Algılayıcı (ÇKA), hipertansiyon tahmininde diğer yöntemlere göre daha iyi bir performans sergilemiştir.

Tezin ikinci bölümünde, veri madenciliği hakkında genel bilgiler verilmiştir. Veri madenciliğiyle ilgili bilim dallarından bahsedilip, veri madenciliği tekniklerine değinilmiştir. Nefroloji, böbreğin yapısı, böbreğin işlevleri, böbrek hastalıkları, hipertansiyon ve hipertansiyonun böbrek ile ilişkisinden bahsedilmiştir. Bu çalışmada kullanılan veri madenciliği ara yüzü, veri madenciliği algoritmaları ve hastalardan alınan değerler tanıtılmıştır.

Tezin üçüncü bölümünde tezle ilgili genel sonuçlar, veri madenciliği süreçlerine uygun olarak incelenmiştir.

Tezin son bölümünde ise tezle ilgili elde edilen sonuçlara ve değerlendirmelere değinilmiştir.

2. MATERYAL VE YÖNTEM

2.1. Veri Madenciliği Tanımı

Verilerin dijital ortamda saklanmaya başlanması ile birlikte, yeryüzündeki bilgi miktarının her geçen gün katlanarak arttığı günümüzde, veri tabanlarının sayısı da benzer, hatta daha yüksek bir oranda artmaktadır. Akıllı veri işleme metodu olan veri madenciliği, dünya üzerinde artan veri miktarının etkili bir biçimde kullanılmasının neredeyse tek çözümü olarak görülmektedir.

Veri madenciliği, veri ambarlarındaki tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş verileri ortaya çıkarmak, bunları, karar vermek ve gerçekleştirmek için kullanma sürecidir [20].

Shalvi ve DeClaris'e göre "Veri Madenciliği, belirli bir alanda ve belirli bir amaç için toplanan veriler arasındaki gizli kalmış ilişkilerin (desenlerin, modellerin vb.) ortaya konulmasıdır" [21].

Veri Madenciliğinin daha detaylı bir tanımı şöyledir [22] :
"Veri madenciliği, depolanmış yüksek miktardaki veriden istatistiksel ve matematiksel teknikler gibi desen tanımlayıcı teknolojiler kullanarak anlamlı ve yeni ilişkiler, desenler ve trendler keşfetme sürecidir."

Veri madenciliği, büyük miktarda veri içinden gelecekle ilgili tahmin yapmayı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır. Başka deyişle veri madenciliği, organizasyonların karar aşamaları için yeni bilgiler üreten ya da gelecekle ilgili tahminler ve planlar yapılmasını sağlayan bir dizi teknikler ve anlayışlar bütünüdür [23].

Yakın geleceğin, geçmişten çok fazla farklı olmayacağını varsayarsak geçmiş veriden çıkarılmış olan kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahminler yapmayı sağlayacaktır.

Günümüzde finansal uygulamalar, şirket kaynak yönetimi (ERP), müşteri ilişkileri yönetimi (CRM) gibi ticari yazılımlardan ve web logları gibi kaynaklardan giderek daha fazla veri toplanmakta ve veri tabanlarında saklanmaktadır. Bu verileri işleyemeyen şirketler veri yönünden zengin fakat bilgi bakımından fakir duruma düşmektedir. Bu veri tabanlarından faydalı bilgileri bulmak günümüzde birçok şirketin ve uygulamanın odaklandığı ana noktalardan biri olmuştur. Bu tür bilgi keşifleri için veri madenciliği süreci bir anahtardır. Veri madenciliğinin ana amacı elde bulunan veriden gizli kalmış örüntüleri çıkarmak, bu verinin değerini arttırmak ve veriyi bilgiye dönüştürmektir.

1960'larda veri toplama sistemleri popülerdi, 1970'lerde ise ilişkisel veri tabanları kullanılmaya başlandı. 1980'lerde ilişkisel veri tabanı modellerinin, 1990'larda ve 2000'li yıllarda ise dünya gündeminde veri madenciliğinin, veri ambarlarının, multimedya ve web veri tabanlarının oldukça popüler olduğu görülmüştür.

Karar aşamalarında çok kritik bazı bilgiler vardır ki, sonuçların etkileri bu bilgilerin doğruluğuyla orantılıdır. Birçok durumda cevabı tam olarak verilemeyen sorular doğrultusunda karar verilebilir. Müşterilerin ilgi alanları, firmaya karşı olan bakış açıları, rakip firmalara olan ilgileri, markalara olan bağlılıkları, gelir düzeyleri gibi bilgiler onlara sağlanan mal ve hizmetlerin kalitesi üzerinde çok net etkiler yapacaktır. Bu tür bilgiler teorik olarak her ne kadar şirketlerin sistemlerinde kayıt altında olsa da, kullanılabilir bir şekilde açık ve net cevaplara ulaşabilme, mevcut kullanımdaki sistemlerle çok zordur [23].

Çok büyük veri yığınları altında saklı olan bu bilgilere ulaşmak için uzun yıllar boyu yapılagelen çalışmaların neticesinde bir dizi yöntem geliştirilmiştir. Veri madenciliği uzun yıllardır özellikle Batı ülkelerinde üzerinde çalışılan bir konudur.

Veri madenciliği, veri tabanları, istatistik ve yapay öğrenme gibi bilgi yönetimi alt bileşenlerinin kavramlarına dayalı teknikler kullanır ve bu bir bilgi yönetim işinin diğer modülleri ile uyumlu bir fikirsel yapıdadır.

Veri madenciliği bir sorgu işleme mekanizması, yapay öğrenme sistemi değildir. Zaman zaman veri madenciliği, istatistiki programlar ya da müşteri takip programları gibi yanlış tanımlanmaktadır [23].

Madencilik terimi, topraktan küçük ve değerli maden parçalarının çıkarılması ile yüksek miktarda veri içinden değerli bilginin çıkarılması arasındaki benzerlik nedeniyle kullanılmıştır [24].

Veriler içinde desen aramak insan hayatının başlamasından beri devam eden eski bir süreçtir. Avcılar hayvanların göç etme davranışlarında, çiftçiler ekin gelişiminde, politikacılar seçmen görüşlerinde tahmin edici modeller aramışlardır. Bilim insanlarının işi ise veriler içinde fiziksel dünyanın nasıl çalıştığını anlatan modeller aramak ve bulacağı modeller yardımıyla ortaya çıkacak yeni durumlarda neler olacağını tahmin edecek teoriler geliştirmektir.

Veri tabanı kavramı ve veri tabanı sistemi teknolojisi geliştirilme sürecinde verilerin belli bir amaca yönelik düzenli olarak toplanmasını, saklanmasını ve işlenmesini sağlayarak anlamlı bilgi elde edilmesi için gerekli altyapıyı sağlamıştır.

1960'lı yıllarda verilerin ilkel dosyalara işlenmesi ile başlayan verileri toplama girişimleri sonraki 20 yıllık süreçte önce verilerin sınıflara ayrılması ardından bu sınıflar arasında ilişkisel bağlantılar kurulması ile veri tabanı kavramının ortaya çıkmasını sağlamıştır. Veri tabanı kavramı ardından kullanıcı ile veri tabanı arasındaki etkileşimi sağlayan arayüzler geliştirilmiş, 1980'lerin ortasından itibaren çok farklı yapı ve boyutlarda verileri işleyebilecek gelişmiş veri tabanı sistemleri tasarımı çalışmaları günümüzde de devam etmektedir.

Geliştirilen veri tabanı sistemlerinde çok büyük miktar ve boyutlarda veriler toplanması, bu verilerden etkin ve yeni bilgiler elde edilmesinde zorluklara yol açmıştır. 1990'lardan itibaren gelişen internet ve bilgisayar teknolojisi devasa boyutlara ulaşan verilerin analizinde işlemci gücü yardımıyla yeni matematiksel tekniklerin kullanılabilmesini dolayısıyla veri analizinin kolaylaşmasını sağlamıştır. Veri analizi ile çok miktarda bilgi elde edilmesi bu bilgilerin ne kadar anlamlı olduğu

sorusuna önem kazandırmıştır. Veri Madenciliği kavramı bu ortamda ortaya atılmış ve bu başlık altında anlamlı bilgi elde edilmesi için büyük boyutlu ve çok farklı yapıdaki veri kümelerinde desen araştırma teknikleri geliştirilmiştir.

Özetle veri madenciliği kavramının gelişimi ve bu konudaki uygulamalarının yaygınlaşması bilgi teknolojisinin gelişiminin bir sonucudur.

İnsanoğlu geçmişten bugüne her zaman verileri yorumlayıp bilgi edinmeye çalışmıştır ve bunun için çeşitli donanımlar oluşturmuştur. Bu donanımlar bilginin taşınmasını sağlamıştır. Zamanla her alanda bilgi toplanmaya başlanmış ve kronolojik olarak gelişimi Çizelge 2.1.'de özetlenmiştir [25].

Çizelge 2.1. Veri madenciliği tarihçesi

Gelişim Adımları	**Cevaplanan Karar Problemi**	**Kullanılabilen Teknolojiler**	**Ürün Sağlayıcıları**	**Karakteristikler**
Veri Toplama (1960'lar)	"Toplam kârım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM, CDC	Geriye Dönük, Statik Veri Dağıtımı
Veri Erişimi (1980'ler)	"Türkiye'de mart ayında birim satışları ne kadar?"	İlişkisel Veritabanları, SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Geriye Dönük, Dinamik Veri Dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"Türkiye'de geçen mart ayında birim satışları ne kadardı?"	Çok Boyutlu Veritabanı Sistemleri, Veri Ambarları	Pilot, Comshare, Arbot, Cognos, Microstrategy	Çoklu Düzeylerde, Geriye Dönük Dinamik Veri Dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Kırıkkale'deki birim satışları muhtemelen ne olabilir? Niçin?"	İleri Düzeyde Algoritmalar, Çok İşlemcili Bilgisayarlar, Büyük Veri tabanları	Pilot, Lockheed, IBM, SGI, SPSS, SAS, Microsoft vs.	Geleceğe Dönük, Proaktif Enformasyon Dağıtımı

2.2. Veri Madenciliğinin Kullanım Alanları

Veri madenciliğinin asıl amacı, çeşitli kaynaklar kullanılarak elde edilen birçok veriyi anlamlı bilgiler elde etmek ve bunu eyleme dönüştürecek kararlar için kullanmaktır.

Veri madenciliğinin uygulama alanları konu başlıkları itibariyle aşağıdaki gibi sınıflandırılabilir.

Veri madenciliğinin pazarlama alanındaki kullanım amaçları;

- Müşteri bölünmesinde,
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında,
- Çeşitli pazarlama kampanyalarında,
- Mevcut müşterileri elde tutmada,
- Pazar sepeti analizinde,
- Çapraz satış analizleri,
- Müşteri değerleme,
- Müşteri ilişkileri yönetiminde,
- Çeşitli müşteri analizlerinde,
- Satış tahminlerinde.

Veri madenciliğinin bankacılık alanındaki kullanım amaçları;

- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- Kredi kartı dolandırıcılıklarının tespitinde,
- Müşteri bölünmesinde,
- Kredi taleplerinin değerlendirilmesinde,
- Usulsüzlük tespiti,
- Risk analizleri,
- Risk yönetimi,
- Sigortacılık,
- Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Riskli müşteri tipinin belirlenmesinde.

Veri madenciliğinin perakendecilik alanındaki kullanım amaçları;

- Satış noktası veri analizleri,
- Alış-veriş sepeti analizleri,
- Tedarik ve mağaza yerleşim optimizasyonu.

Veri madenciliğinin borsa alanındaki kullanım amaçları;

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,
- Alım-satım stratejilerinin optimizasyonu.

Veri madenciliğinin telekomünikasyon alanındaki kullanım amaçları;

- Kalite ve iyileştirme analizlerinde,
- Hisse tespitlerinde,
- Hatların yoğunluk tahminlerinde,
- İletişim desenlerinin belirlenmesi,
- Kaynakların daha iyi kullanılması,
- Servis kalitesinin arttırılması.

Veri madenciliğinin sağlık ve ilaç alanındaki kullanım amaçları;

- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tıbbi teşhis,
- Tedavi sürecinin belirlenmesinde,
- Semptomlara göre hastalık tespiti,
- Magnetik rezonans verileri ile sinir sistemi bölge ilişkilerinin belirlenmesi.

Veri madenciliğinin endüstri alanındaki kullanım amaçları;

- Kalite kontrol analizlerinde,
- Lojistik,
- Üretim süreçlerinin optimizasyonunda.

Çizelge 2.2.'de 2003 yılında veri madenciliğinin sektörler bazında kullanımına ilişkin bir araştırmanın sonuçları yer almaktadır. Toplamda 421 şirketin katıldığı araştırmada veri madenciliği kullanım yüzdeleri görülmektedir [26].

Çizelge 2.2. Veri madenciliğinin uygulandığı alanlar

Sektör	**Veri Madenciliği Kullanımı**
Bankacılık	12%
Bioteknik / Genetik	3%
Kredi Skorlama	8%
CRM	12%
Doğrudan Pazarlama	8%
E-Ticaret	3%
Eğlence / Müzik	1%
Sahtekarlık Tespiti	7%
Şans Oyunları	0,3%
Kamu Uygulamaları	3%
Sigortacılık	6%
Yatırım / Hisse Senedi	1%
Junk E-Mail / Anti Spam	1%
Sağlık / İK	4%
İmalat	5%
Tıp / Farmakoloji	3%
Perakende	6%
Bilim	4%
Güvenlik / Anti Terörizm	1%
Telekomünikasyon	5%
Seyahat	2%
Web	2%
Diğer	3%

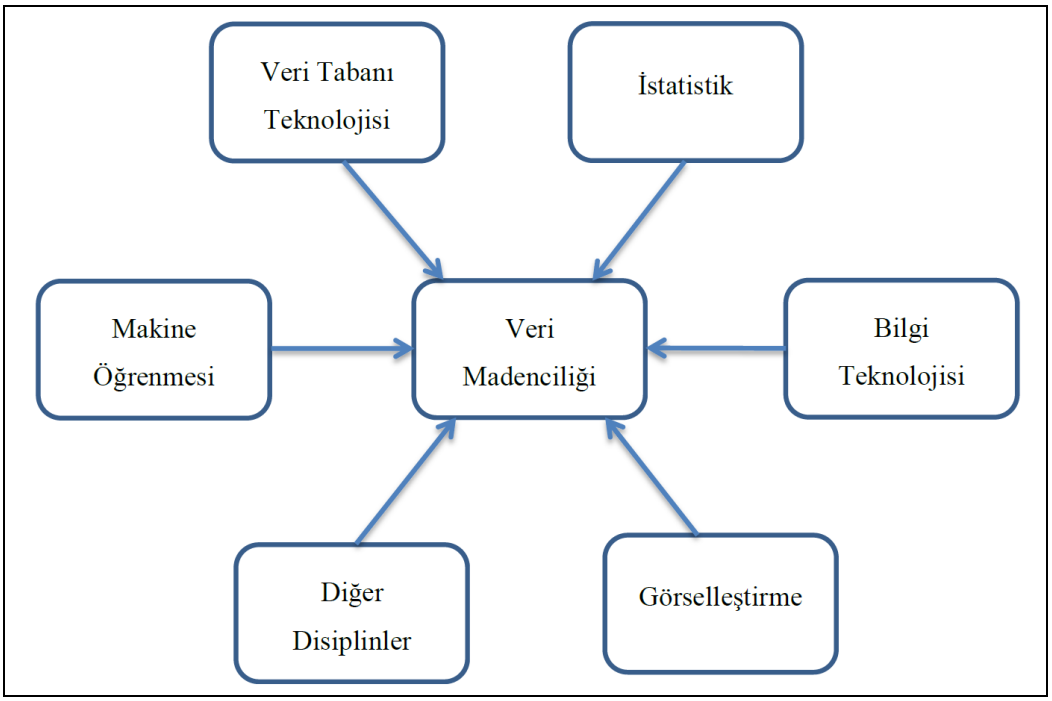
2.3. Veri Madenciliği İle İlişkili Bilim Dalları

Veri Madenciliği, Veritabanı Sistemleri, İstatistik, Makine Öğrenmesi, İnsan-Makine Etkileşimi, Veri Görselleştirme gibi farklı disiplinlerden faydalanan disiplinler arası bir alandır [27].

Kullanılan Veri Madenciliği yaklaşımına bağlı olarak, Veri Madenciliğinde farklı disiplinlerde kullanılan tekniklerden de yararlanılmaktadır. Bu tekniklere örnek olarak Sinirsel Ağlar, Bulanık Küme Teorisi, Tümevarımsal Mantık Programlama, yüksek performanslı hesaplamalar verilebilir.

İşlenen verinin yapısına bağlı olarak Veri Madenciliği, Uzaysal Veri Analizi, Resim Analizi, Sinyal İşleme, Bilgisayar Grafikleri, Web teknolojisi vb. tekniklerden de faydalanmaktadır.

Farklı disiplinlerden elde edilen veriler üzerinde yapılan Veri madenciliği uygulamalarında farklı modeller kurulması gerekliliğini ortaya çıkarmıştır. Veri madenciliği için farklı modelleme gerektiren disiplinlere Ekonomi ve Biyoinformatik gibi disiplinler örnek verilebilir [27]. Şekil 2.1.'de Veri Madenciliğinin diğer disiplinlerle ilişkisi verilmiştir.



Şekil 2.1. Veri madenciliği ile diğer disiplinler arası ilişki

Makine öğrenmesi, istatistik ve veri madenciliği birbirleriyle yakından ilişkilidir [28]. Bu üç disiplin veri içindeki bağıntıları ve örüntüleri bulmayı amaçlar. Makine öğrenmesi yöntemleri, veri madenciliği algoritmalarında kullanılan yöntemlerin temelini oluşturur. Makine öğrenmesi ve yapay zekâ uygulamalarında kullanılan karar ağaçları, kural çıkartımı, sınıflama ve kümeleme gibi pek çok veri madenciliği algoritmasında da kullanılmaktadır. Ancak makine öğrenmesinde ve istatistiksel yöntemlerde kullanılan örnekleme genişliği, veri madenciliği algoritmalarında kullanılan örneklem boyutuna göre çok daha küçüktür. Veri madenciliği, makine öğrenmesi yöntemlerine göre gürültülü, eksik ve boş değerleri işlemede daha başarılıdır. Veri madenciliği, istatistiksel yöntemleri veri setindeki değişkenler arasındaki bağımlılığın derecesini ölçmek, veriyi tanımlamak, verinin özetini çıkarmak ve veri setindeki eksik değerlerin tahminlerini yapmak gibi konularda kullanmaktadır.

Veri madenciliği ve veri tabanı teknolojisi arasında da önemli bir ilişki vardır. Veri madenciliği yöntemlerinin uygulanacağı veriler genellikle büyük boyutlu veri

tabanlarında tutulmaktadır. Veri tabanları sorgu dili SQL (Structured Query Language) ise veri tabanlarındaki var olan ve bilinen ilişkileri ortaya koymak için kullanılmaktadır. Veri madenciliği ise, veri tabanlarında bulunan veriler arasındaki bilinmeyen ilişkileri ortaya çıkarmaktadır.

2.4. Veri Madenciliği Süreci

Veri madenciliği kısaca gizli bilgilerin keşfi sürecidir. Veri madenciliğinin bir süreç olarak tanımlanabilmesi için, sürecin her bir aşamasının dikkatle izlenmesi gerekmektedir. Bir aşamanın sonucu, diğer bir aşamanın girdisidir. Bu sebeple her aşama bir önceki aşamanın sonuçlarına bağımlıdır.

Veri madenciliği için belirlenen standart bir süreç söz konusudur. Bu standart süreç bir konsorsiyum tarafından belirlenmiştir. The Cross- Industry Standard Process for Data Mining (CRISP-DM) konsorsiyumu, 1996 yılının sonlarına doğru genç ve olgunlaşmamış veri madenciliği pazarında üç firma tarafından kurulmuştur.

Bu üç firmanın ilki olan Daimler Chrysler birçok endüstriyel ve ticari organizasyona, veri madenciliği tekniklerini uygulama konusunda öncü olmuştur. SPSS (Statistical Package for the Social Sciences) firması 1990 yılından beri veri madenciliği üzerine çeşitli hizmetler sağlamış ve ilk ticari veri madenciliği çalışma platformu olan Clementine'i 1994 yılında harekete geçirmiştir. NCR (National Cash Register), müşterilerine değer katma işini sağlayabilmek ve alıcılarının ihtiyaçlarına hizmet edebilmek için birçok veri madenciliği danışmanlığı ve teknoloji uzmanlığı takımları kurmuştur [29].

Bu gelişmelerden bir yıl sonra, sözcüklerin baş harfleri "Cross- Industry Standard Process for Data Mining" açılımında olan CRISP-DM konsorsiyumu oluşturulmuş, Avrupa Komisyonundan fon elde edilmiş ve başlangıç fikirleri oluşturulmaya başlanmıştır [29].

2.4.1. Problemin Tanımlanması

Veri madenciliği sürecinin en önemli aşamalarından biridir. Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı, işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçülebileceği tanımlanmalıdır.

Problemin tanımlanması aşamasında, veri madenciliği uygulayan kişi öncelikle işletmenin hangi amacı gerçekleştirmek istediğini dikkate almalıdır. Analizi yapan kişinin hedefi, veri madenciliği uygulamasının sonuçlarını etkileyebilecek önemli faktörleri ortaya çıkarmaktır. Veri madenciliği projesinin başarılı olması, projenin dikkatle planlanması ve spesifik, gerçekleştirilebilir, ölçülebilir bir hedefin olmasına bağlıdır.

Müşteri değerini ve bağlılığını artırmak isteyen bir perakendecinin; çapraz satış tekniklerini kullanmayı hedeflemek, pazarlama kampanyasına cevap verme olasılığı yüksek olan müşterileri belirlemek, karı arttırmak, maliyetleri düşürmek, yeni ürün stratejileri geliştirmek veya pazar payını arttırmayı istemek gibi hedefleri bulunabilir.

2.4.2. Verinin Hazırlanması

Verinin hazırlanması veri madenciliğinin en önemli aşamasıdır, çünkü modelin kurulması sırasında ortaya çıkacak sorunlar, sonradan bu aşamaya geri dönülmesine sebep olacaktır. Bu aşamada daha önceden tespit edilen probleme uygun olarak veri tabanından veya veri ambarından problemle ilgili değişkenler seçilir. Bu değişkenlere uygun olarak veriler temizlenir. Veri ambarında veya veri tabanında toplanan veri hatalar, aşırı değerler içerebilir. Veriler içerisinde uyumsuzluk söz konusu olabilir. Hatta veriler tamamlanmamış olabilir ve eksik verilerin toparlanması gerekebilir. Tüm bu işlemlerin gerçekleştirilmesi verinin hazırlanması aşamasını oluşturmaktadır. Kısaca, verinin hazırlanması; verinin toplanması, verinin

birleştirilmesi ve temizlenmesi ile verinin seçilmesi aşamalarından oluşmaktadır [29].

2.4.2.1. Verinin Toplanması

Veri madenciliği modeli oluşturma sürecinde, verinin hazırlanması sürecindeki ilk adım verinin toplanmasıdır. Bu aşamada, elde var olan ve toplanması gereken verilerin belirlenmesi gereklidir. Kullanılacak veri belirlenirken veri analizi yapacak kişi, veri madenciliği hedeflerini ve işletme amaçlarını da dikkate almalıdır.

Tanımlanan sorun için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesinde hangi veri kaynaklarından yararlanılacağı önemli bir karardır. Çünkü gereğinden az veri kaynağı veri madenciliği çalışmasını eksik bırakacağı gibi, gereğinden fazla veri kaynağı sürecin uzamasına neden olabilecek ve veri kirliliğine yol açabilecektir. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında çeşitli veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir.

Verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır. Güvenilir olmayan veri kaynaklarının kullanımı tüm veri madenciliği sürecinin güvenilirliğini etkileyecektir. Bu nedenlerle, iyi sonuç alınacak veri madenciliği çalışmaları ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

2.4.2.2. Verinin Birleştirilmesi ve Temizlenmesi

Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorun ve uyumsuzluklar mümkün olduğu ölçüde giderilerek, veriler tek bir veri tabanında toplanır. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır.

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına sebep olacaktır. Bu uyumsuzlukların belli başlıları verilerin farklı zamanlara ait olmaları, kodlama farklılıkları veya farklı ölçü birimlerine sahip olmalarıdır. Farklı kaynaklardan toplanan verinin birleştirilmesi ve temizlenmesi gerekmektedir. Eğer bir veri tabanında kayıtlı kişilerin hem yaşları, hem de doğum tarihleri bulunuyorsa, bu değişkenlerden bir tanesi fazladır. Bu durumda mevcut değişkenlerin birleştirilmesi ve tek bir değişken gibi işleme girmesi mümkün olabilmektedir.

Ayrıca veride bulunan eksik veya kayıp bilgiler gözden geçirilmelidir. Örneğin bir veri tabanında yer alan kişilerin medeni hali belirliyken, bazı kayıtlarda bu bilgi eksik olabilir veya bu kayıt hiç girilmemiş olabilir. Bu durumdaki eksiklik kayıp veriler olarak tanımlanabilir. Bunun haricinde, bazı kayıtlarda aşırı uç değerler (outliers) veya yanlış girilmiş değerler olabilir. Bu tür bilgilere de gürültü (noise) adı verilmektedir. Veri setinde fazla değişken bulunabilir. Bir veri setinde bulunan doğum tarihi ve yaş değişkenleri de fazla veriye örnek verilebilir.

İdeal olan eksik verilerin zaman içinde tamamlanması yoluna gitmektir. Diğer bir alternatif, eksik bilgilerin tahmin yöntemiyle tamamlanmasıdır. Veri girme aşamasında yanlış girilen değerler olabileceğinden, bu değerlerin göz ardı edilmesi analiz sonuçlarını fazlasıyla değiştirebilir. Bu nedenle aykırı değerlerin gözden geçirilip atılmasıyla beraber, analizde yapabileceği değişiklikler hesaplandıktan sonra, uygun görüldüğü taktirde veri tabanından silinmesi gerekir. Bütün bu düzenlemeler yapıldıktan sonra, tüm veri tek bir veri tabanında, düzenli bir şekilde tutulmalıdır [29].

2.4.2.3. Verinin Seçilmesi

Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için bu adım, bağımlı ve bağımsız değişkenlerin ve modelde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır.

Sıra numarası, kimlik numarası gibi anlamlı olmayan değişkenlerin modele girmemesi gerekmektedir. Çünkü bu tip değişkenler, diğer değişkenlerin modeldeki ağırlığının azalmasına ve veriye ulaşma zamanlarının uzamasına neden olabilmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır.

Verilerin görselleştirilmesine olanak sağlayan grafik araçlar ve bunların sunduğu ilişkiler, bağımsız değişkenlerin seçilmesinde önemli yararlar sağlayabilir. Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin, veri kümesinden atılması tercih edilir. Veri seçimi, konuyla ilgisi olmayan gereksiz verilerin çıkartılması işlemini içermektedir [29].

2.4.3. Modelin Kurulması

Veri madenciliğinde bilgi kaynaklarından en fazla verimin alınabilmesi için, modelin kurulması aşaması çok önemlidir. İyi kurulmuş bir model, analiz sonucunda elde edilecek sonuçların kalitesini de etkileyecektir. İyi bir veri madenciliği uygulayıcısı, analiz sonucunda hangi örüntülerin bulunabileceğini tahmin edebilmelidir. Model kurma süreci, analiz için hangi verilerin elde hazır bulunduğunu kullanıcıya sunar. Eğer model doğru kurulmazsa, veri seti içerisinde bulunabilecek kritik ilişkiler doğru bir şekilde sunulamaz ve önemli örüntüler tespit edilemez. Dolayısıyla modelden başarılı sonuç elde etme olasılığı da azalır.

Veri madenciliği çalışmasında geliştirilen modelde kullanılan veri tabanının çok büyük olması durumunda, rastgeleliği bozmayacak şekilde örnekleme yapılması uygun olabilir. Ayrıca burada seçilen örnek kütlenin ana kütleyi temsil edip etmediği de kontrol edilmelidir. Halen kullanılan işletim sistemleri ve paket programlar ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeniyle mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak birkaç model denemek yerine, rasgele örneklenmiş bir veri tabanı parçası üzerinde birçok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü

modelin seçilmesi daha uygun olacaktır. Diğer bir deyişle modellerin performansları uygun bir karar yöntemi ile sınanmalıdır [29].

2.4.4. Modelin Değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varıncaya kadar yinelenen bir süreçtir.

Model kuruluş süreci, denetimli ve denetimsiz öğrenmenin kullanıldığı modellere göre farklılık göstermektedir.

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenmede, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı, verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir.

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu, kurulan model tarafından belirlenir.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

Denetimli öğrenmede seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenilmesi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenir.

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamıyla modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir [29].

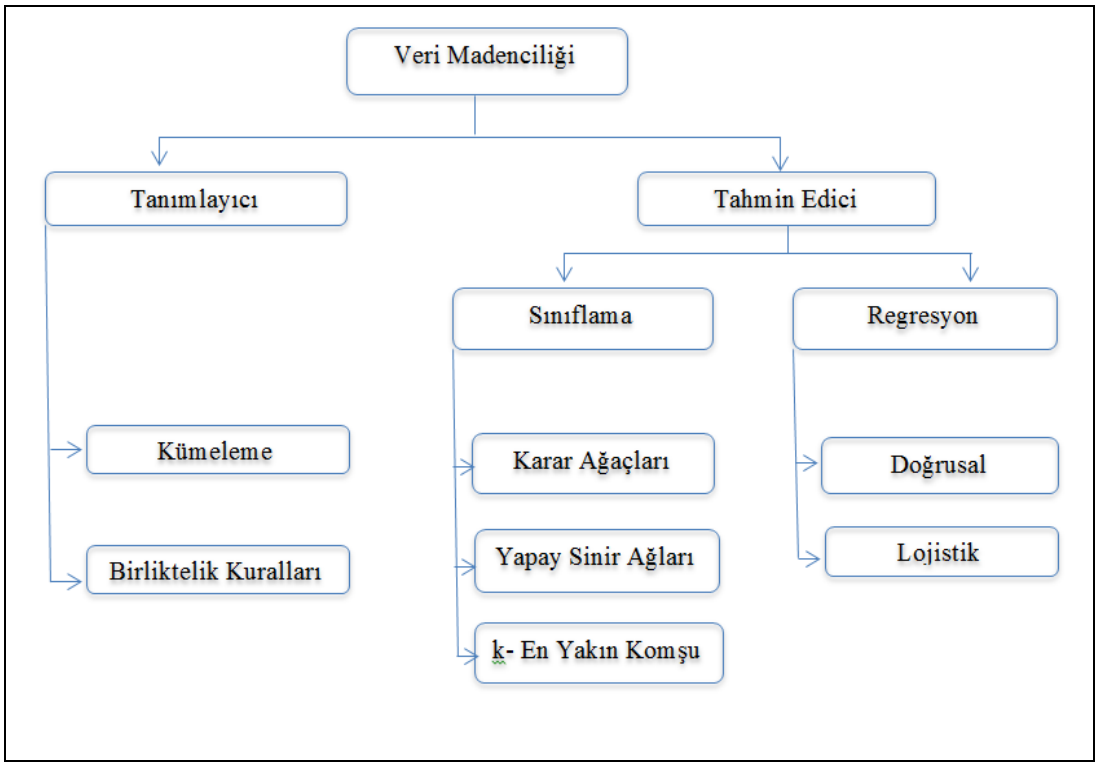
2.4.5. Modelin Kullanılması

Veri madenciliği sürecinin son aşaması, kurulan ve geçerliliği kabul edilen modelin kullanılmasıdır. Bu doğrudan bir uygulama olabileceği gibi, bir başka modelin alt parçası olarak da kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir. Kullanılan modelin zaman içerisinde izlenip ortaya çıkan değişikliklerin modele yansıması, yaşayan bir süreç olması açısından vazgeçilmez bir koşuldur [29].

2.5. Veri Madenciliği Teknikleri

Veri madenciliğinde kullanılan modeller, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir.

Tahmin edici modeller veri tabanındaki bazı değişkenleri veya alanları kullanarak, ilgilenilen diğer değişkenlerin bilinmeyen veya gelecekteki değerlerini öngörmeyi hedeflemektedir. Tanımlayıcı modeller ise, verinin tanımlanmasında, insanlar tarafından yorumlanabilen örüntüler bulmaya odaklanmıştır. Modeller arasındaki sınırlar çok kesin olmamasına rağmen, ayrım bütün keşif hedefinin anlaşılması için yararlıdır [30]. Veri madenciliği modellerine ait sınıflandırma Şekil 2.2.'de sunulmuştur.



Şekil 2.2. Veri madenciliği modelleri

Şekil 2.2.'de sunulan sınıflandırmanın çizgileri kesin sınırlar değildir. Örneğin bazen yapay sinir ağları regresyon sınıfının altında incelenebilir.

2.5.1. Tahmin Edici Modeller

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Örneğin bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır.

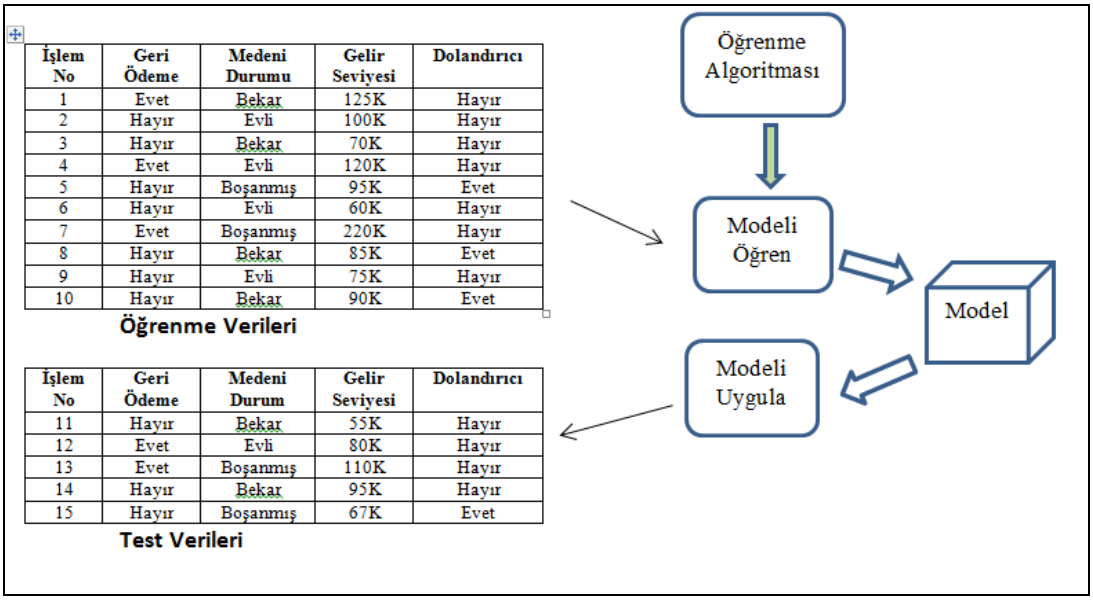
Bazı test sonuçlarına dayanarak hastaya tanı konması, bilinen bir grup ürünle müşterinin A ürününü de alma olasılığının tahmin edilmesi, geçmiş altı aylık indekse bakılarak Dow Jones indeksinin tahmin edilmesi tahmin edici modellere örnek olarak verilebilir.

2.5.1.1. Sınıflama

Sınıflama bir veri maddesini önceden belirlenmiş sınıflardan birine eşleyen, öğrenen bir fonksiyondur. Sınıflama bölümsel değerlerin tahmin edilmesinde kullanılmaktadır. Tahmin edilmesinde öğrenme verisine dayanmaktadır. Sınıflamaya ait model Şekil 2.3.'te sunulmuştur.

Sınıflamanın kullanılma amaçları arasında

- Tümör hücrelerinin iyi veya kötü huylu olmasına göre bölümlenmesinde,
- Kredi kartı işlemlerinin yasal veya sahte olarak ayrılmasında,
- Haber içeriklerinin finans, hava, magazin ve spor olarak sınıflandırılmasında kullanılır.



Şekil 2.3. Sınıflamanın model gösterimi

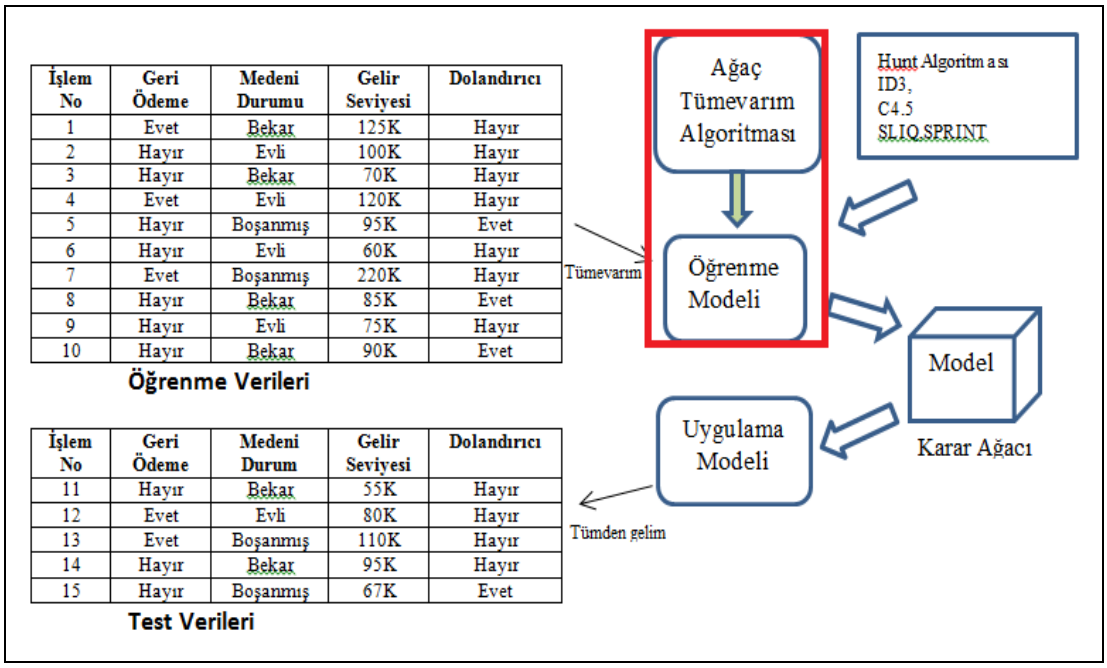
Bir kaydın önceden belirlenmiş bir gruba girebilmesi için sınıflama algoritması ile öğrenme verileri kullanılarak hangi sınıfların var olduğu ve bu sınıflara girmek için bir kaydın hangi özelliklere sahip olması gerektiği otomatik olarak keşfedilir. Test verileriyle de bu öğrenmenin testi yapılarak ortaya çıkan kurallar optimum sayısına getirilir. Sınıflama algoritması, denetimli öğrenme kategorisine giren bir öğrenme biçimidir. Denetimli öğrenme, öğrenme ve test verilerinin hem girdi hem de çıktıyı içerecek şekilde olan verileri kullanmasıdır [31].

2.5.1.2. Karar Ağaçları

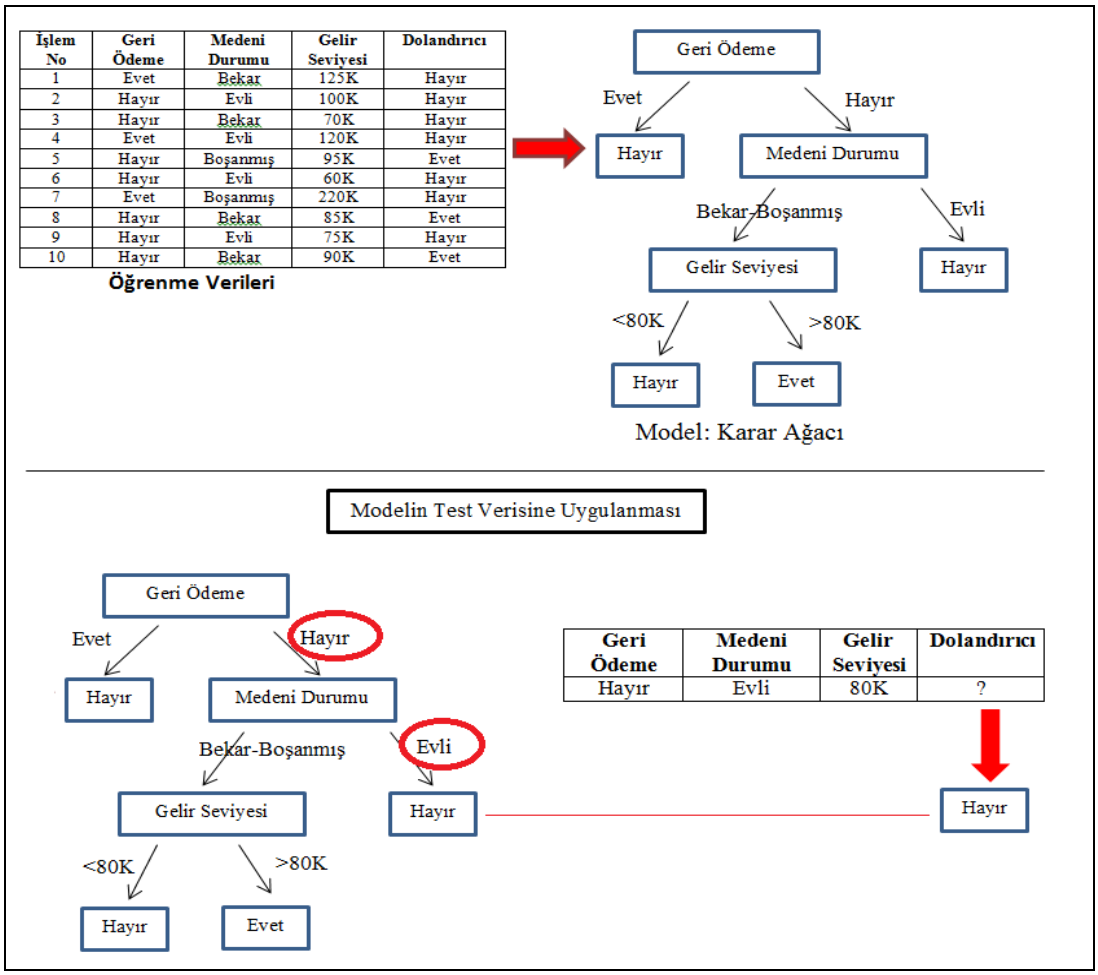
Karar ağacı, görünümünün ağaç şeklinde olmasından dolayı bu şekilde isimlendirilmiş, tahmin edici bir tekniktir. Yapısından dolayı programlamaya uygulanması kolaydır. Görsel açıdan da uygulama sonrası problemin çözümüne ait kurallar kullanıcılara rahatlıkla sunulabilir. Uygulanması fazla maliyet gerektirmez, güvenirlikleri iyi seviyededir. Bu yüzden veri madenciliğinde geniş kullanım alanı bulmaktadır.

Karar ağacı kök, karar düğümleri, dallar ve yapraklardan oluşmaktadır [32]. Kök, veri alanlarının içinde en önemli olandır. Karar düğümü, yapılacak testi temsil eder. Bu testin sonucu, veri, ağacın dalları arasında ayrılır. Ayrılma işlemleri bütün düğümlerde ardışık olarak gerçekleşir. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, orada bir karar düğümü oluşur. Ancak gruba ayrılma sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir [33].
Karar ağacı tekniği sınıflamanın genel özelliklerini taşır. İki aşamalı bir tekniktir. Birinci basamakta öğrenme gerçekleştirilir. Öğrenme işlemi, sınıflama algoritması tarafından öğrenme verisi üzerinde gerçekleştirilir. Bu aşamadan sonra uygulanacak model oluşturulur. Öğrenilen model karar ağacını oluşturur. Model oluşturulurken çıkan anlamsız dallar incelenerek modelden çıkarılır. Bu işleme budama adı verilir. Modelin oluşturulmasından sonra sıra test edilecek verinin sınıflanmasına gelir. Test

verisi sınıflama kurallarının doğruluğunu test etmek amacıyla kullanılır. Eğer elde edilen sonuç belirlenen sınırlar içinde ise model daha büyük çapta yeni verilerin sınıflandırılması amacıyla kullanılabilir. Şekil 2.4. ve Şekil 2.5.'de sınıflamanın model gösterimi ve Karar Ağacının Oluşturulması ve Modelin Test Verisine Uygulanması sırasıyla gösterilmiştir.



Şekil 2.4. Sınıflamanın model gösterimi (Ağaç Tümevarım Algoritması)



Şekil 2.5. Karar ağacının oluşturulması ve modelin test verisine uygulanması

Karar ağaçlarının bakımı ve anlaşılması verinin karmaşıklığının artmasıyla zorlaşır. Eksik verilerin olması durumunda bölünme, değişkenlerinin birisinin değeri bilinmiyorsa karara varılması mümkün değildir. Karar ağacı algoritması elimizdeki veriyi bölümlere ayırırken dikkat edilecek nokta, bağımlı değişkenin değerini en çok belirleyen bağımsız değişkenleri ayırmaktır. Algoritmaya ait adımlar:

- Veri içinden ilgilendiğimiz bağımlı ve bağımsız değişkenlerin belirlenmesi,
- Hedef bağımlı değişkeni en çok etkileyen bağımsız değişkenin bulunması, bu maksatla her değişkenin hedefi ne kadar etkilediğinin bulunması ve en çok etkileyen değişkenin seçilmesi (Amaç bölünmeden sonra kalan parçaların bölünme öncesine oranla daha sade olmasını sağlamaktır.),

- Bölünme sonrasında kalan verilere aynı bölünme testlerinin yapılması ve daha sade gruplara ulaşılıncaya kadar bu işleme devam edilmesidir [34].

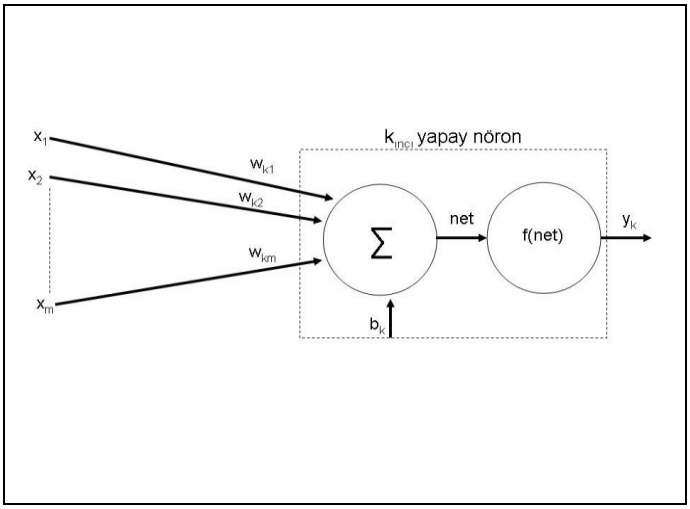
2.5.1.3. Yapay Sinir Ağları

Yapay Sinir Ağları (YSA)'nda amaç fonksiyon birbirine bağlı basit işlemci birimlerden oluşan bir ağ üzerine dağıtılmıştır. YSA sinir ağlarında kullanılan öğrenme algoritmaları, veriden birimler arasındaki bağlantıya ait ağırlık değerlerini hesaplar, uygulama alanı geniştir ve bellek tabanlı yöntemler kadar yüksek miktarda işlem ve bellek gerektirmeyen bir modeldir [35].

YSA üzerindeki çalışma, insan beyninin bilinen dijital bilgisayardan tamamen farklı bir şekilde çalıştığının farkına varılması ile başlamıştır. Farklı alanlardan birçok bilim adamı insan beyninin işlemsel sürecini modelleme konusunda araştırmalar yapmıştır. Beyin oldukça karmaşık, doğrusal olmayan, bilgiyi paralel işleyebilen bir sistemdir. İçeriğinde bulunan bileşenleri düzenleyerek bazı alanlarda bugün bilinen en hızlı bilgisayardan birkaç kat hızlı ve yüksek kalitede işlem yapma kapasitesine sahiptir. Bu alanlar arasında örüntü tanıma, algılama ve motor kontrol işlevleri sayılabilir. YSA insan beyninin soyut bir hesaplama modelidir. İnsan beyni yaklaşık 10^{11} adet, nöron adı verilen küçük birime sahiptir. Bu birimler arasında yaklaşık olarak 10^{15} adet bağlantı bulunmaktadır. Gerçek yapıya benzer olarak YSA'da yapay nöronlardan (işlem birimleri) ve bunlar arasındaki bağlantılardan oluşmaktadır. Bu ağa grafiksel olarak yaklaşırsak, nöronların düğüm noktaları ve bağlantıların da kenarlar olarak görüntülendiğini söyleyebiliriz. YSA adından da anlaşılacağı gibi birçok düğüm noktasının birbirine yönsel bağlantılarla bağlandığı bir ağ yapısıdır. Bu yapıda her düğüm noktası bir işlemci birimi, düğüm noktaları arasındaki bağlantılar da nedensel ilişkileri temsil ederler. Her düğüm noktası uyarlanabilir özelliktedir. Uyarlanabilir olması, bu düğüm noktalarının çıktılarının, düğüm noktaları ile ilgili değiştirilebilir parametrelere dayalı olduğunu göstermektedir. YSA küçük işlemcilerden oluşmuş büyük dağıtık paralel bir işlemcidir. YSA birimleri arası bağlantı güçleri ile ifade edilen deneysel bilgiyi öğrenme ve bu bilgiyi kullanıma sunma kabiliyetine sahiptir [36].

Yapay bir nöron YSA'nın temelini oluşturan bilgi işleme ünitesidir. Şekil 2.6.'da gösterildiği gibi yapay nöron üç ana birimden meydana gelmektedir.

- Değişik girdilerden gelen bir grup bağlantı $\mathbf{x_i}$ (veya sinapsler), bağlantılar bir ağırlık veya güç değeri ile karakterize edilmiştir. Bu değer w_{ki} ile gösterilmektedir. Birinci indeks ilgili nörona, ikinci indeks ise ilgili bağlantının ağırlık değerine karşılık gelmektedir. Genelde yapay nöronun ağırlık değeri eksi veya artı değerler arasında olabilir.
- Girdi sinyallerin ağırlık değerlerine göre toplanmasını sağlayan bir toplama fonksiyonu bulunmaktadır. Bu işlem doğrusal bir birleştiricidir.
- Nöronun çıktısının büyüklüğünün belirli limitler içinde olmasını sağlayan *f* fonksiyonu vardır.
- Ayrıca b_k olarak adlandırılan ve önyargıyı temsil eden bir etki değeri bulunmaktadır. Artı veya eksi değerde olmasına göre f fonksiyonuna girdi değerine arttırıcı veya azaltıcı etki yapar [36].



Şekil 2.6. Yapay Nöron Modeli

2.5.1.4. k-En Yakın Komşu

En yakın komşu sınıflandırıcıları benzerlik yöntemi ile öğrenmeyi esas alır. Eğitim örnekleri n-boyutlu sayısal nitelik ile tanımlanırlar. Her bir örnek n-boyutlu uzayda bir noktayı temsil eder. Bu şekilde tüm eğitim örnekleri n-boyutlu uzayda depolanır. Bilinmeyen bir örnek geldiğinde, bir k-en yakın komşu sınıflandırıcısı bilinmeyen örneğe en yakın k eğitim örneğini bulmak için örüntü uzayını tarar. K eğitim örnekleri bilinmeyen örneğin k-en yakın komşularıdır. Yakınlık Öklit mesafesi kullanılarak ölçülür. Öklit mesafesi X = (x_1;x_2; ;x_n) ve Y= (y_1;y_2; ;y_n) olarak adlandırılan iki nokta arasında;

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2.1)$$

formülü ile bulunur.

Bilinmeyen örnek, örüntü uzayında kendisine en yakın eğitim örnekleri kümesine atanır. En yakın komşu sınıflandırıcıları tüm eğitim örneklerini depoladıkları için örnek tabanlıdır. Sınıflandırılmamış bir örnek karşılaştırılmak istendiğinde eğer olası komşularının sayısı fazlaysa hesaplama zamanı oldukça yüksektir. Bu durumda indeksleme tekniklerinin kullanılması gerekebilir. Karar ağacındaki tümevarım ve tümdengelimin sürecinde uygulananın aksine, en yakın komşu sınıflandırıcıları her bir niteliğe eşit ağırlık verirler. Bu durum, veride çok fazla ilgisiz nitelik bulunduğunda karışıklığa sebep olabilir [32].

2.5.1.5. Regresyon Analizi

Regresyon Analizi süreklilik gösteren değerlerin tahmin edilmesinde kullanılır [37]. Regresyon sınıflamaya benzemektedir. Regresyon Analizi'nin sınıflamadan en önemli farkı tahmin edilebilir değişkenin sürekli bir sayı olmasıdır. Regresyon Analizi teknikleri yüzyıllardan beri istatistiğin geniş çapta çalışmalar yapılan bir alanıdır. Doğrusal regresyon, lojistik regresyon en çok bilinen regresyon çeşitleridir. Dağıtım metotları, hava ısısına bağlı rüzgâr hızının ve nem oranının tahmininde Regresyon Analizi kullanılmaktadır [38].

Regresyon Analizi var olan veriye formüllerin uygulanması ile tahminler yapmada kullanılmaktadır. Doğrusal veya lojistik regresyon tekniklerini kullanarak, var olan veriden fonksiyon elde edilmektedir. Yeni veri var olan fonksiyona uygulanarak tahmin yapmada kullanılmaktadır. Değerleri bilinen değişkenlerin kullanılarak diğer değişkenlerin tahmininde kullanılır. Regresyon terminolojisinde, tahmin edilecek olan değişken "bağımlı değişken", bağımlı değişkeni tahmin etmek için kullanılan değişken ya da değişkenler de "bağımsız değişken" olarak adlandırılır [39].

a) Doğrusal Regresyon

Bir bağımlı bir bağımsız değişkenden oluşan en basit regresyon analizi "basit doğrusal regresyon", iki ya da daha fazla bağımsız değişken içeren regresyon analizi de "çoklu regresyon analizi" olarak adlandırılır.

b) Lojistik Regresyon

Doğrusal regresyon sürekli değer fonksiyonlarını modellemede kullanılır. Genelleştirilmiş regresyon modelleri ise kategoriksel cevap değişkenlerinin modellemesinde doğrusal regresyon yaklaşımının kullanılmasının kuramsal altyapısını temsil eder. Genelleştirilmiş doğrusal regresyon modelinin en yaygın tipi lojistik regresyon modelidir. Lojistik regresyon bir grup tahmin edici değişkenin doğrusal fonksiyonu olarak gerçekleşmiş bir olayın olasılığını modeller. Bağımlı değişkenin değerini hesaplamak yerine, bağımlı değişkenin verilen bir değeri alma olasılığını hesaplar. Örneğin bir müşterinin kredibilitesinin iyi veya kötü olduğunu tahmin etmek için, lojistik yöntem iyi kredibilite olasılığını tahmin etmeye çalışır. Bağımlı değişkenin güncel durumu tahmin edilen olasılığa bakarak tespit edilir. Eğer tahmin edilen olasılık değeri 0.50 değerinden büyükse tahmin EVET (iyi kredibilite), diğer durumda ise HAYIR (kötü kredibilite) değerine yakındır. Bu yüzden lojistik regresyonda kredibilite "p" başarı olasılığı olarak adlandırılır. Diğer yönden, girdilerin bazılarının sayısal olması veya olmaması lojistik regresyon modeli için önemli değildir. Bu yüzden lojistik regresyon daha genel veri çeşitlerinin kullanımını destekler.

2.5.2. Tanımlayıcı Modeller

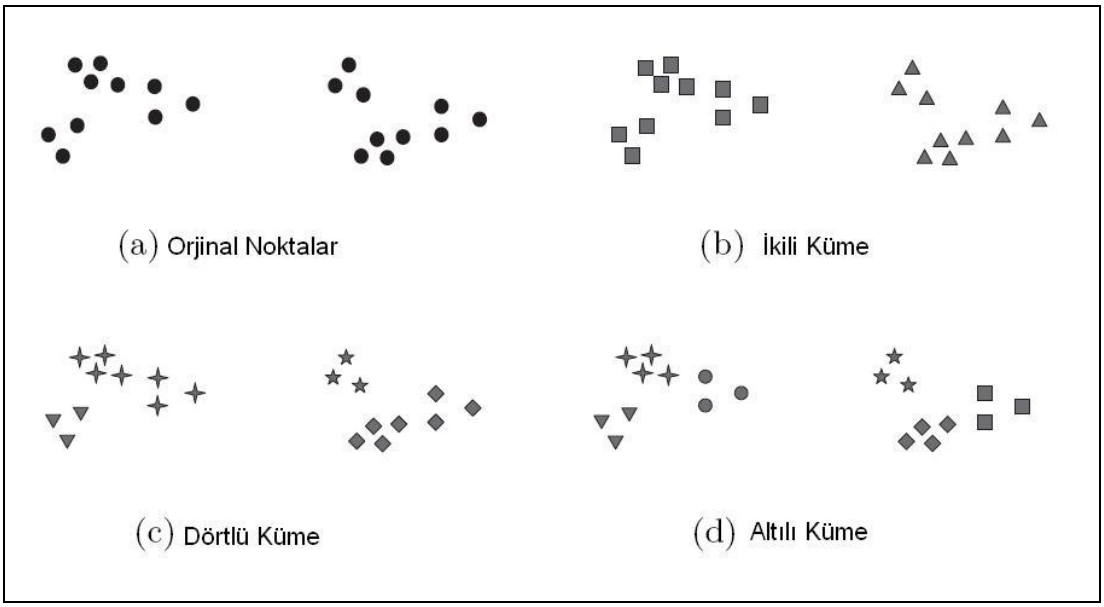
Tanımlayıcı modeller ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasında kullanılmaktadır. Tanımlayıcı modeller arasında Kümeleme (Clustering), Birliktelik Kuralları (Association Rules), Ardışık Zamanlı Örüntüler sayılabilir.

2.5.2.1. Kümeleme Analizi

Fiziksel veya soyut nesnelerden oluşan bir grubun, birbirine benzer nesnelerin aynı sınıflarda kalacak şekilde daha küçük gruplara ayrılması işlemine kümeleme adı verilir. Küme içinde bulunduğu gruptaki nesnelere benzeyen fakat diğer gruptaki nesnelere benzemeyen veri nesnelerinin bir araya toplanmış halidir. Aynı kümede bulunan veri nesneleri üzerinde birçok uygulama daha kolay ve hızlı olarak yapılabilir.

Kümeleme günlük hayatımızda sıklıkla başvurduğumuz bir yöntemdir. Bitki hayvan sınıflandırılmasının yapılması, hayvanların etçil, otçul olarak sınıflandırılması vb. şekilde yapılan sınıflandırmalar sürekli, gelişen ve daha ayrıntılı olacak şekilde yapılmaya devam eder. Kümeleme yöntemi örüntü tanıma, veri analizi, görüntü işleme, pazar analizi gibi sayısal uygulamalarda yaygın olarak kullanılmıştır. Kümeleme yapılarak kalabalık ve seyrek bölgeler ortaya çıkarılabilir. Böylece veri alanları arasındaki ilginç korelasyon ve örüntü dağılımları ortaya çıkarılabilir. World Wide Web (WWW)'de bulunan belgelerin bilgiye daha kolay ulaşım için gruplanması, benzer işlevlere sahip genlerin sınıflandırılması, bitki ve hayvan cinslerinin kökünün araştırılması, bir yerleşim birimindeki benzer özellikteki evlerin gruplanması gibi birçok alanda kullanılabilir. Sınıflamanın aksine, kümeleme sınıfların önceden belirlenmesine veya öğrenme verilerine ihtiyaç duymaz [32].

Şekil 2.7.'de aynı verinin farklı şekilde kümelenme şekilleri görülmektedir.



Şekil 2.7. Aynı grup verinin farklı şekillerde gruplanması

2.5.2.2. Birliktelik Analizi (Association Rules)

Birliktelik kuralları, büyük veri kümelerinde, veriler arasındaki ilginç yerel örüntüleri, bağlantıları ve kuralları bulmaya yönelik, denetimsiz (unsupervised) veri madenciliği şeklidir [32]. Örüntü verinin belli bir yönüyle ilgili bilgiler veren bir yerel kavram, model ise verinin tüm betimlemesini yapan genel bir kavramdır.

Birliktelik analizi insanların veri madenciliğini anlamak için kafalarında tasarladıkları veri madenciliğine en yakın olanıdır. Birliktelik analizi büyük bir veri tabanında altın aramaktır. Veri tabanındaki altın, bizim veri tabanı ile ilgili daha önce bilmediğimiz ve muhtemelen açıkça ifade edemediğimiz bir kuraldır. Bu yöntembilim veri tabanındaki tüm ilginç örüntüleri bulur. Her taşın altına bakılır. Bu aynı zamanda modelin zayıflığıdır. Kullanıcı ortaya çıkan bilginin büyüklüğü karşısında bunalabilir ve bu miktardaki bilginin kullanılabilirliğinin analizi zor ve zaman alıcıdır [36].

Veri madenciliği sonucu ortaya çıkarılan ilişkiler birliktelik kuralları olarak gösterilebilir. Birliktelik analizi, satın alma eğilimlerinin tanımlanması, müşterinin

hangi mal veya hizmetleri almaya eğilimli olduğunun saptanması ve bu yolla müşteriye daha fazla mal satılmasını sağlamak için sıklıkla başvurulan bir modeldir. Çizelge 2.3.'de birliktelik kurallarının en tipik örneklerinden biri olan pazar sepeti uygulamasına ait veriler bulunmaktadır.

Çizelge 2.3. Alışveriş Verileri

İşlem No	Satın Alınan Ürün Alt Kümeleri
1	Ekmek, Süt
2	Ekmek, Çocuk Bezi, Meyve Suyu, Yumurta
3	Süt, Çocuk Bezi, Meyve Suyu, Kola
4	Ekmek, Süt, Çocuk Bezi, Meyve Suyu
5	Ekmek, Süt, Çocuk Bezi, Kola

Açığa çıkarılan örüntüler birliktelik kuralları şeklinde gösterilebilirler. Aşağıdaki kural Çizelge 2.3.'den elde edilmiştir.

- {Süt, Çocuk Bezi}→{Meyve Suyu}

Bu kural çocuk bezi satışı ile meyve suyu satışı arasında güçlü bir ilişki olduğunu göstermektedir. Aşağıdaki örnekler birliktelik kurallarına örnek olarak gösterilebilir.

$$Yaş\ (X, 30\text{-}34) \wedge Gelir\ (X, 42\text{K} - 48\text{K}) \Rightarrow satın\ alınan\ (X, \text{Plazma TV})$$

$$Satın\ alınan\ (X; "Bilgisayar") \Rightarrow$$
$$satın\ alınan\ (X; \text{"finansal\_yönetim\_yazılımı"})$$

Birliktelik analizi Pazar sepeti analizinin yanı sıra aşağıda sunulan alanlarda da kullanılmaktadır:

- Biyoinformatik, tıbbi teşhis,
- Web madenciliği,
- Bilimsel veri analizi (Dünya'ya ait okyanus, kara ve atmosferik verilerin bilimsel analizinde),
- Terör olaylarının tespit edilmesi.

2.5.2.3. Birliktelik Analizi ile İlgili Tanımlar

$I = \{i_1, i_2, i_3, ..., i_d\}$ kümesinin pazar sepeti analizinde kullanılan tüm öğelerin kümesi ve $T = \{t_1, t_2, t_3, ..., t_N\}$ bu analizdeki tüm işlemlerin kümesi olsun. Her bir t_i işlemi I öğe kümesinin alt kümelerini barındırmaktadır. Birliktelik analizinde öğe kümesi bir veya daha fazla öğe barındıran kümeyi temsil etmektedir.

- **Birliktelik Kuralı:** $X \rightarrow Y \; iken \; X \subset I, Y \subset I ve \; X \cap Y = \emptyset$
- **Öğe Kümesi** bir veya daha fazla öğeden meydana gelen kümedir.
  {Süt, Çocuk Bezi, Meyve Suyu, Kola}
- **k-öğe kümesi** içinde k adet öğe bulunduran kümedir.
- **Destek sayısı (σ)** öğe kümesinin görülme sıklığıdır.
  Örnek: σ ({Süt, Ekmek, Çocuk Bezi})=2
- **Destek (Support)** X → Y kuralındaki X ve Y öğelerinin/öğe setlerinin her ikisini de kapsayan işlemlerin toplam işlemlere oranıdır.

  Örnek: {Süt, Çocuk Bezi}→{ Meyve Suyu } s = ({Süt, Çocuk Bezi, Meyve Suyu})/ |T| = 2/5=0.4

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \qquad (2.2)$$

- **Güven (Confidence)** X → Y kuralında Y öğe kümesindeki elemanların X öğe kümesi elemanlarının bulunduğu işlemlerde hangi sıklıkta bulunduğunu göstermektedir.

  {Süt, Çocuk Bezi}→{ Meyve Suyu }

$$C = \frac{\sigma(süt, Çocuk\ Bezi, Meyve\ Suyu)}{\sigma(süt, Çocuk\ Bezi)} = \frac{2}{3} \quad (2.3)$$

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma X} \quad (2.4)$$

Destek önemli bir ölçü birimidir. Eğer bir kural düşük desteğe sahipse bu durumda kural şans eseri ortaya çıkmış olabilir. Destek kuralın kullanılabilirliğini, güven ise doğruluğunu gösterir.

Birliktelik kurallarını bulmak için sık tekrarlanan öğelerin bulunması, bu öğelerin önceden belirlenen minimum destek sayısı kadar tekrarlanması gerekir. Daha sonra tekrarlanan öğelerden güçlü birliktelik kuralları oluşturulur. Bu kurallar minimum destek ve minimum güven değerlerini karşılamalıdır. Sık tekrarlanan öğeleri bulmak için kullanılan en temel yöntem Apriori Algoritmasıdır [33].

2.5.2.4. Appriori Algoritması

Apriori Algoritması 1994 yılında bulunmuştur [40]. Appriori Algoritması birkaç döngü ile veri tabanında sık bulunan öğe kümelerini hesaplar. Döngü *i,* i eleman sayısına sahip kümeleri hesaplar. Her bir döngü iki aşamadan oluşur: birinci aşamada aday kümeler üretilir, ikinci aşamada ise aday kümeler sayılır ve belirlenen değerlere göre seçim yapılır.

İlk döngünün ilk aşamasında, yaratılan aday alt kümelerin hepsi veri tabanındaki tüm öğeleri kapsamaktadır. Sayma aşamasında, algoritma bu adayların destek sayılarını belirlemek için tüm veri tabanını tarar. Sonunda sadece belirtilen destek değerinden

yukarıda destek değerine sahip olanlar sık tekrarlananlar grubuna seçilirler. Böylece ilk döngü sonunda tüm sık kullanılan öğe setleri saptanmış olur.

İkinci basamakta, tüm ikili alt kümeler seçilmek için adaydır. Fakat önceki döngülerden elde edilen bilgilerin ışığında Apriori algoritması destek sayısı az olanları budar. Budama eğer bir ürün kümesi sık tekrarlanıyorsa, bu kümenin tüm alt kümeleri sık tekrarlanıyor prensibi doğrultusunda yapılır. Aynı şekilde eğer bir ürün kümesi sık tekrarlanmıyorsa bu ürünün alt kümeleri de sık tekrarlanmıyor anlamına gelir ve bu alt kümeler atılır [36].

Apriori algoritması Çizelge 2.4.'te görülen pazar sepeti işlemleri üzerinde uygulamalı olarak gösterilmiştir.

Çizelge 2.4. Basit İşlem Veri Tabanı Modeli

İşlem No	**Satın Alınan Ürün Alt Kümeleri**
001	A, C, D
002	B, C, E
003	A, B, C, E
004	B, E

Veri tabanı birliktelik kuralları iki aşamada incelenebilir. Bu aşamalar, önceden belirlenmiş belli eşik değeri üzerinde olan kümelerin bulunması ve bulunan bu kümelerden önceden belirlenmiş belli bir güven değeri üzerindeki kuralların üretilmesidir.

Apriori algoritmasının uygulanmasına ait aşamalar Çizelge 2.5.'te sunulmuştur.

Çizelge 2.5. Apriori Algoritmasının Uygulanması

Apriori Algoritması Birinci Tekrar

1 elemanlı alt küme C_1	1 elemanlı alt küme	Adet	s (%)	Büyük 1 elemanlı alt küme L_1	Adet	s (%)
{A}	{A}	2	50		2	50
{C}	{C}	3	75		3	75
{D}	{D}	1	25			
{B}	{B}	3	75		3	75
{E}	{E}	3	75		3	75
a. Oluşturma Aşaması	b.1 Sayma Aşaması			b.2 Seçme Aşaması		

Apriori Algoritması İkinci Tekrar

2 elemanlı alt küme C_2	2 elemanlı alt küme	Adet	s (%)	Büyük 2 elemanlı alt küme	Adet	s (%)
{A, B}	{A, B}	1	25			
{A, C}	{A, C}	2	50		2	50
{A, E}	{A, E}	1	25			
{B, C}	{B, C}	2	50		2	50
{B, E}	{B, E}	3	75		3	75
{C, E}	{C, E}	2	50		2	50
a. Oluşturma Aşaması	b.1 Sayma Aşaması			b.2 Seçme Aşaması		

Apriori Algoritması Üçüncü Tekrar

3 elemanlı alt küme C_3	3 elemanlı alt küme	Adet	s (%)	Büyük 3 elemanlı alt küme	Adet	s (%)
{B, C, E}	{B, C, E}	2	50	{B, C, E}	2	50
a. Oluşturma Aşaması	b.1 Sayma Aşaması			b.2 Seçme Aşaması		

Varsayalım ki, istenen minimum destek değeri s = % 50, minimum güvenilirlik değeri c = % 50 olsun. İlk aşamaya ait değerler Çizelge 2.5.'teki appriori algoritması birinci tekrar bölümünde görülmektedir. İlk aşamada 1 elemanlı kümeler yaratılmaktadır. İkinci aşamanın birinci bölümünde bunların kaç adet olduğu ve minimum destek oranları hesap edilmekte, ikinci bölümde ise % 50 destek değerinin altındaki kümeler atılmaktadır. Bu hesaplamalar algoritmanın gerektirdiği döngü sayısı kadar yapılır.

Bu aşamadan sonra yapılması gereken güvenilirlik ölçütlerine göre kuralların bulunmasıdır. Üçüncü tekrar sonucu elde edilen 3-elemanlı alt kümeden ve ikinci tekrar sonucu elde edilen iki elemanlı alt kümeden faydalanılarak güvenilirlik katsayısı hesaplanır.

$$c\ (\{B,C\} \rightarrow) = \frac{s(B,C,E)}{s(B,C)} = \frac{2}{2} = 1\ (\%100) \qquad (2.5)$$

Güvenilirlik arzu edilen %50'lik orandan daha yüksek olduğu için kural kabul edilir.

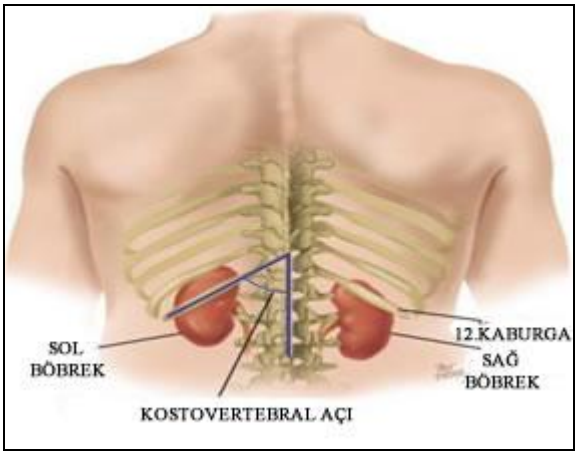
2.6. Nefroloji Bilimi

Böbreklerin sağlığı ve hastalıklarıyla ilgilenen iç hastalıklarına bağlı bir bilim dalıdır, dahili tıp bilimlerinden birinin uygulama alanıdır.

2.6.1. Böbrek Anatomisi

Böbrekler, omurgalılarda bulunan fasulye biçiminde boşaltım organlarıdır. 10 cm boyuna kadar olabilen böbrekler, boşaltım sisteminin bir bölümünü oluştururlar. Bu organlar, başta üre olmak üzere atıkları kandan süzer ve onları su ile birlikte idrar olarak boşaltırlar. Böbrekleri ve böbreklere etki eden hastalıkları inceleyen tıbbi dal nefrolojidir. Nefroloji, adını Yunanca "böbrek" anlamına gelen *nephros* sözcüğünden alır. Böbrek(ler) ile ilgili anlamında kullanılan *renal* sözcüğü ise Latince *renalis* sözcüğünden gelir. Böbreklerin içindeki süzme birimlerine nefron denir. Her

böbrekte yaklaşık 1 milyon nefron bulunur. Böbreklerin vücuttaki konumu Şekil 2.8.'de gösterilmiştir.

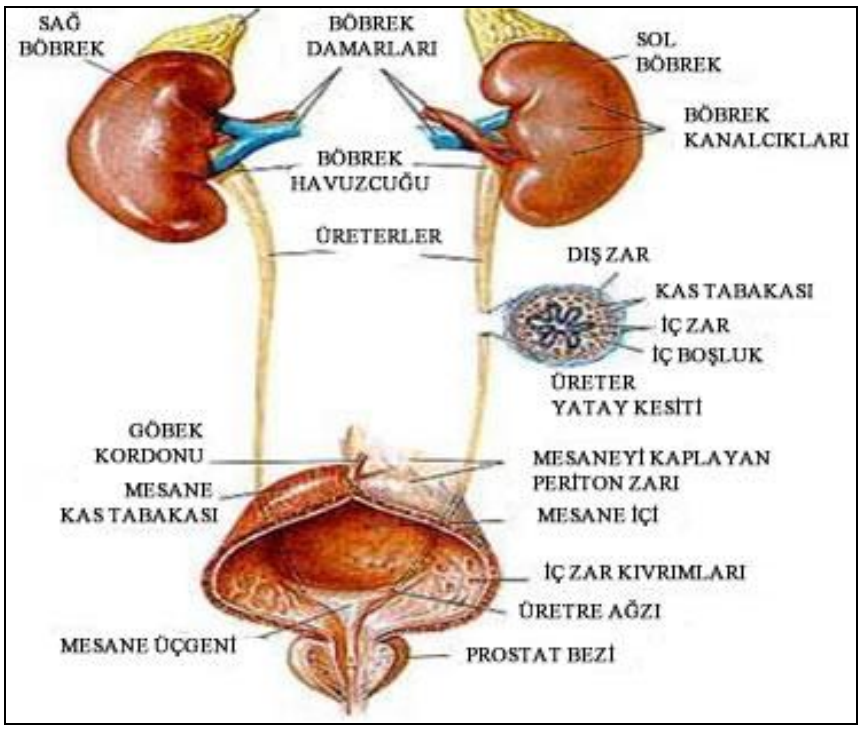


Şekil 2.8. Böbreklerin vücuttaki konumu

İnsanlarda, böbrekler karın bölgesinin arka bölümünde, bir başka deyişle karınzarı arkası (retroperitonal) bölgesinde yer alırlar. İki tane bulunan böbreklerden sağda olanı diyaframın hemen altında ve karaciğerin arkasında (posterior), solda olanı ise diyaframın altında ve dalağın arkasında yer almaktadır. Böbreklerin ikisinin de üstünde böbreküstü bezleri yer almaktadır. Böbreklerin konumları bakımından bakışımsız olmalarının nedeni karın boşluğunda büyük bir yer kaplayan karaciğerin, sağda bulunan böbreğin soldakine göre 1-2 santimetre daha aşağı bir konumda (inferior) bulunmasına neden olmasıdır.

Karınzarı arkasında bulunan böbreklerin boyutları 9 ila 13 cm arasında değişmekte ve sol böbrek sağdakinden az da olsa biraz daha büyüktür. Yaklaşık 12. göğüs omuru ile 3. bel omurlarının (T12-L3) düzeyleri arasında yer almaktadırlar. Böbreklerin üst bölgeleri 11. ve 12. kaburgalarca korunmaktadır. Böbreküstü bezleriyle birlikte böbrekler, yağ dokuyla çevrelenip (buna *pararenal yağ* denilmektedir), bu yapı da böbrek zarı (*renal fasiya* olarak da bilinir) ile bütünüyle sarılmış durumdadır.

Böbrekler, süzülmemiş kanı karın bölgesi aorttan ayrılan sol ve sağ böbrek atardamarları yoluyla almaktadırlar. Böbrekten dönen süzülmüş kan ise sağ ve sol böbrek toplardamarları yoluyla alt ana toplardamara döner. Böbreğe giden kan, kalbin pompaladığı toplam kanın (kardiyak debi) üçte birine ulaşabilir [41]. Böbreklerin bağlantıları Şekil 2.9.' da gösterilmiştir.



Şekil 2.9. Böbreklerin bağlantıları

2.6.2. Böbreklerin İşlevleri

Böbreklerin işlevleri beş çatı altında toplanabilir:

- Metobolizma atık ürünleri olan üre, kreatinin, ürik asit, ilaç ve toksinlerin vücuttan atılmasını sağlamak
- Vücut sıvı elektrolit dengesini düzenlemek
- Vücudun asit baz dengesini düzenlemek
- Kan basıncını ayarlamak
- Alyuvar yapımını uyarmak.

2.6.3. Böbrek Hastalıkları

1. Glomeruler hastalıklar (Akut ve kronik böbrek iltihabı olarak bilinen hastalıklar)
2. Damar hastalıkları
3. Doğuştan gelen hastalıklar
4. Kalıtımsal hastalıklar
5. Damarsal böbrek hastalıkları
6. Gebelik zehirlenmesi
7. Taş hastalıkları
8. Hipertansiyon ve diyabet (şeker) hastalığı gibi sistemik hastalıklara bağlı böbrek hastalıkları.
9. İlaçlara bağlı gelişen böbrek hastalıkları vs. şeklinde sıralanabilir.

Böbrekler, bu fonksiyonları bozulunca görevlerini yerine getiremeyecek; kanda atılması gereken maddeler atılamayacak, kanda birikerek semptomlar verecektir, idrar miktarında azalma olacaktır. Böbrek hastalıklarının çoğu sinsi ve ağrısız seyreder. Fonksiyonu bozulan böbrek idrarla atılması gereken zehirli maddeleri süzemeyerek kanda çeşitli semptomlar vereceklerdir. Böbreklerin fonksiyonlarının azalması ve kaybolması anı ise (akut) geriye dönüşebilen; yıllar içinde sessizce devam ederek gelişiyorsa (Kronik) geri dönüşü olmayan bir tablo ortaya çizer. Bunların oluşturacağı klinik belirti ve bulgular her hastalığa göre farklılık gösterir. İdrar yollarında akımın engellenmesi iltihaplanması dışında böbrek hastalıklarında ağrı sık görülmez. İdrar yapmada bozukluklar; sık idrara çıkma, gece idrara çıkma, ağrılı idrar uykuda idrarı tutamama gibi bozukluklar görülür. İdrar miktarında azalma veya çok idrara çıkma veya hiç çıkmama gibi belirtiler olabilir. Ayrıca böbrek hastalığının genel bulguları vücutta sıvının birikmesi (ödem) gözlerde, ayaklarda vs görülebilir. Kan basıncı yükselmesi ve ishal gibi sindirim sistemi belirtileri bunun yanında burun kanaması ağızdan kan gelmesi dışkı ile kan gelmesi gibi kanama belirtileri görülür [42].

2.6.4. Hipertansiyonun Tanımı

Hipertansiyon iki şekilde tanımlanmaktadır:

- Sistemik arteriyel kan basıncının normal sınırların üstünde olması (sistolik kan basıncı (SKB)≥140 mmHg, diyastolik kan basıncı (DKB) ≥ 90 mmHg) ya da kişinin antihipertansif ilaç kullanıyor olması,
- Bir sağlık profesyoneli tarafından en az iki kere yüksek kan basıncı olduğunun söylenmesi olarak tanımlanmaktadır [43].

Başka bir tanıma göre hipertansiyon "insan sağlığını, yaşam kalitesini ve yaşam süresini kötü yönde etkileyebilecek kadar yüksek olan kan basıncı değerleri" olarak ifade edilmektedir. Hipertansiyon tanısı almış hastalarda SKB, DKB veya her ikisi birden yüksek olabilir; bu değerlerdeki heyecan, korku veya egzersiz gibi durumlar nedeniyle oluşabilecek geçici yükselmeler hipertansiyon olarak kabul edilmez [44].

2.6.5. Hipertansiyon ve Böbrek Arasındaki İlişki

Yüksek tansiyonun nedenlerinin en başında böbrek hastalıkları gelir. Bu hastalıklar, ya böbreği ilgilendiren nefrit, kist, tümör, taş vb. olabildiği gibi, damarlardaki bir daralma veya böbrek üstü bezinin hastalıkları ile ilgili olabilir. Her yüksek tansiyonlu hastada yapılabilecek bir idrar tahlili, üre ve kreatinin tayini veya böbrek ultrasonografisi ile bu hastalıkların önemli bir kısmına teşhis konulabilir.

Hipertansiyonun en önemli hedef organlarından birisi böbreklerdir. Esansiyel olarak adlandırdığımız nedeni belli olmayan yüksek tansiyonlu hastaların, eğer tedavi edilmezlerse, %15'i böbrek yetmezliğinden vefat eder. Ayrıca henüz diyaliz uygulanmayan kronik böbrek hastalarının tansiyonu kontrol altına alınmazsa; hastalıkları daha hızlı ilerler.

Bilindiği gibi, böbrek hastalarında koroner kalp hastalığı ihtimali normale göre yüksektir. Kontrolsüz hipertansiyon bu ihtimali daha da arttırır. Yapılan çalışmalar,

yüksek kan basıncının kontrolü ile böbrek hastalarında kalp komplikasyonlarının azaldığını göstermiştir [45].

2.6.6. Hipertansiyonun Sınıflandırılması

Hipertansiyon sınıflamasının amacı her hastanın durumuna uygun bir profil elde etmede güvenilir ve kolay bir yöntem sunmaktır. Sınıflama ile hastalığın ciddiyeti hakkında değerlendirme yapılabilir ve risk tanımlanarak tedavi sağlanabilir [46].

Erişkinlerde kan basıncı derecesinin sınıflandırılması niteldir. Pratikte tedaviye yaklaşım kolaylığı sağlamak için kan basıncı değerleri dikkate alınmaktadır. Amerika Birleşik Devletleri Ulusal Komitesi(JNC)-7 raporunda 18 yaş ve üstündeki erişkinlerin kan basınçları optimal, normal, yüksek- normal ve hipertansiyon olarak dört dereceye ayrılmıştır. Çizelge 2.6.'da hipertansiyonun sınıflandırılması gösterilmiştir.

Çizelge 2.6. Hipertansiyonun sınıflandırılması

Kan Basıncı Derecesi	**Sistolik**	**Diastolik**
Optimal	**<120**	**<80**
Normal	**<130**	**<85**
Yüksek normal	**130-139**	**85-89**
Evre1	**140-159**	**90-99**
Evre2	**>160**	**>100**

Hipertansiyon tanısı için iki veya daha fazla muayene sırasında, en az iki ölçümün ortalamasına dayandırılmalıdır. Sınıflandırmada sistolik ve diyastolik kan basınçları

farklı sınıflara düşerse, kişinin kan basıncı değerlendirilirken daha yüksek olan kan basıncı derecesi dikkate alınmaktadır [47].

2.6.7. Hipertansiyonun Epidemiyolojik Özellikleri

Hipertansiyon, dünyada önlenebilir ölüm nedenleri içerisinde bir numaralı risk faktörüdür. 2000 yılı itibariyle dünyada erişkin nüfusun % 26,4' ünün hipertansiyonu olduğu ve bu oranın 2025 yılında %29,2'ye çıkacağı öngörülmüştür. Bir diğer deyişle, halen 972 milyon insanın hipertansiyonu vardır ve 12 yıl sonra bu rakam 1,5 milyarı aşacaktır [48]. Hipertansiyonu olan bireylerin çoğu, ekonomik olarak gelişmekte olan ülkelerde yaşamaktadır. Bu ülkelerde hipertansiyonun bu denli sık olması ve giderek artması, "epidemiyolojik geçiş" sürecine bağlanmaktadır [48].

2.6.8. Hipertansiyonun Belirtileri

Hastaların büyük çoğunluğunda ciddi bir şikâyet yoktur. Sabah erken saatlerde başın arkasında hissedilen değişen şiddetlerdeki ağrı en sık şikâyettir. Aşırı yorgunluk, sersemlik, bulanık görme diğer şikâyetler olabilir. Komplikasyon gelişmiş hipertansiflerde, zarar gören organa göre, göğüs ağrısı, nefes darlığı, görme sorunları vb. şikâyetler oluşabilir [49].

2.7. Weka

Yeni Zelanda Waikato Üniversitesi'nde bir proje olarak başlayan Weka (Waikato Environment for Knowledge Analysis), bugün dünya üzerinde birçok araştırmacı tarafından kullanılmaya başlanan bir Veri Madenciliği programıdır [50].

Alanında uzman biri tarafından, elle analiz edilemeyecek kadar büyük veri tabanlarından faydalı bilgi türetmek için kullanılmaktadır. Ayrıca eğitim amaçlı olarak da yaygın olarak kullanılmaktadır. Şekil 2.10.'da ara yüzü görünen Weka'nın,

dünya çapındaki veri madenciliği problemlerini çözmek için geliştirilen tüm algoritmaların bir koleksiyonunu sunması ve kolay ulaşılabilirliği önemli avantajlarıdır [51].

Weka'daki birincil öğrenme metotları, sınıflandırıcılardır ve bu sınıflandırıcılar veriyi modelleyen kural kümeleri veya karar ağaçları üretirler. Weka aynı zamanda, birliktelik kuralları öğrenme ve veriyi kümeleme algoritmalarını da içerir. Tüm uygulamalar tek tip bir komut satır ara yüzüne sahiptir. Veriyi ön işlemek için araçlar veya filtreler, diğer önemli kaynaklardır. Öğrenme planları gibi, filtreler ortak bir komut satır seçimleri kümesiyle, standart ortak bir komut satır ara yüzüne sahiptir. Weka yazılımı, bilgisayar platformu ne olursa olsun veri madenciliği araçlarının kullanılabilirliğini kolaylaştırmak için tamamen Java'da yazılmıştır. Sistem, özetle Java paketlerinin bir takımıdır [51].

Weka ara yüz görünümü Şekil 2.10.'da gösterilmiştir.

Şekil 2.10. Weka ara yüz görünümü

2.8. Kullanılan Veri Madenciliği Sınıflandırma Algoritmaları

2.8.1. Saf Bayes (Naive Bayes) Algoritması

Eldeki verilerin belirlenmiş olan sınıflara ait olma olasılıklarını öngören bir algoritmadır. Temeli, istatistikteki Bayes teoremine dayanır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Belirsizlik taşıyan durumlarda karar verme konusunda oldukça başarılıdır.

Genellikle belirsizlik durumlarında sınıflandırma ve tahmin yapmak için kullanılır. En önemli dezavantajı değişkenler arası ilişkinin modellenmemesi ve değişkenlerin birbirinden tamamen bağımsız olduğu varsayımıdır [52].

Bayes Kuralı;
A ve B rastgele sayılar olsun;

$$P(A \mid B) = P(B \mid A)P(A) / P(B) \quad (2.6)$$

P(A) : A olayının bağımsız olasılığı prior (öncül) olasılık
P(B) : B olayının bağımsız olasılığı
P(B| A) : A olayının olduğu bilindiğinde B olayının olasılığı likelihood
(şartlı olasılık)
P(A | B) : B olayının olduğu bilindiğinde A olayının olasılığı posterior
(artçıl) olasılık
Bayes kuralına dayanarak P(A | B)' yi maksimum yapan durumlar hesaplanabilir.

"E" A olayının bütün durumlarının kümesi;

$$A_{MAX} = \text{argmax}\, A \in E\, P(A|B)$$

$$= \text{argmax}\, A \in E \frac{P(B|A)P(A)}{P(B)}$$

$$= \text{argmax}\, A \in E\, P(B|A)P(A) \tag{2.7}$$

2.8.2. Karar Tablosu (Decision Table) Algoritması

Decision Table, karar tablosu çoğunluk sınıflandırıcısını oluşturur. Karar Tablosu, gelecekteki altkümeleri ilk en iyi arama yöntemini kullanarak değerlendirir ve değerlendirme için çapraz yetki (cross-validation) de kullanabilir. Opsiyonel olarak, her örnek için sınıf belirlemede en yakın komşuluk metodunu kullanabilir [53].

Karar Tablosu; şartlar setini ve mantık tablo formatında kolayca ifade edilebilir olduğunda onların iş sonuçlarını göstermek için kullanılır. Karar Tablosu aynı zamanda ve farklı şartlar altında farklı işler içeren işlemlerin tutarlılığını ve bütünlüğünü doğrulamak için de kullanılır. Karar Tablosu, çok kompleks ve geniş şart setleri için Karar Ağaçları'ndan daha iyi çalışır [54].

2.8.3. IB1 Algoritması

IB1 algoritması basit bir örnek tabanlı öğrenme algoritmasıdır. Test örneklerinden kalan eğitim örneklerini bulmak amacıyla basit bir uzaklık ölçüsü kullanır ve eğitilen örnekle aynı sınıfı tahmin eder. Eğer birden çok örnek test örneğine aynı uzaklıkta ise, ilk bulunan örnek kullanılır [55].

IB1, özniteliklerin dağılımını normalize etmesinin dışında tipik bir en yakın komşuluk (nearest neighbor) algoritması olarak ifade edilebilir. IB1, örnekleri giderek artan şekilde işler ve boş değerleri toleranse eden bir yol izler [56].

Söz konusu bu 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. Uzaklıkların hesaplanmasında, i ve j noktaları için Öklit Uzaklık Formülü kullanılır.

$$d(i,j) = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2} \tag{2.8}$$

olarak ifade edilir.

IB1 algoritması, gözlem değerlerinden oluşan bir küme için aşağıdaki adımların uygulanmasından ibarettir.

1. K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır.
2. Bu algoritma verilen bir noktaya en yakın komşuları belirleyeceği için, söz konusu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanır.
3. Hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir.
4. Seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir.
5. Seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir.

2.8.4. Çok Katmanlı Algılayıcı (Multilayer Perceptron) Algoritması

Bu sınıflandırıcı, örnekleri sınıflandırmak için arka planda yayılma yöntemini kullanır. Bu ağ elle oluşturulabileceği gibi bir algoritmayla ya da her ikisiyle de oluşturulabilir. Eğitim zamanı sürecinde bu ağ görüntülenebilir ve değiştirilebilir. Düğümlerin eşik olmayan lineer birimler haline geldiği ve sınıfın nümerik (sayısal) olduğu durumlar dışında, ağdaki bu düğümler sigmaması (kıvrımlı) biçimdedir [57]. Çok Katmanlı Algılayıcı (ÇKA), yapay sinir ağları yapısının bir sınıfıdır. ÇKA, paralel bağlı ağ modeli kurarak insan beyninin öğrenme mekanizmasını taklit eden

akıllı kodlar oluşturmak üzerine yoğunlaşır. Bir ÇKA modelinde önce sistem eğitilir ve ağ, en son güncellenen ağ parametreleri kullanılarak fonksiyonel haritalayıcı olarak çıkışları hesaplayabilir [58].

Çok Katmanlı Algılayıcı, standart geri yayılım algoritması ile sürekli geri beslenerek eğitilir. ÇKA denetimli(supervised) ağlardır ki bu yüzden ÇKA istenilen cevabın eğitilmesi ihtiyacı duyar. ÇKA giriş verisinin istenilen cevap verisine nasıl dönüştürüleceğini öğrenir ve bu yüzden ÇKA örüntü sınıflamasında yaygınca kullanılır. Bir veya iki gizli katmanla ÇKA giriş ve çıkış haritasını sanal olarak birbirine yaklaştırır. Birçok yapay sinir ağları uygulaması ÇKA içerir [59].

ÇKA, algılayıcılar olarak adlandırılan basit sinir ağıdır. Algılayıcı, giriş ağırlıklarına göre lineer kombinasyon formuna dönüştürülerek çoklu gerçek değerli girişlerden ve sonra da bazı lineer olmayan aktivasyon fonksiyonları içinde yer almış mümkün çıkışlardan tek bir çıkış hesaplar [60].

Genellikle dış dünya bilgileri analog bilgi olduğundan bu bilgilerin, sayısal hale dönüştürülmesi ve bu sayısal bilginin 0–1 arası skalaya indirgenmesi gerekmektedir. Bu durum eşitlik 2.9' da gösterilmiştir.

$$F_k = G_k \tag{2.9}$$

Burada;

F_k : Giriş katmanındaki k. nöronun çıkışını ifade eder.

G_k : Giriş katmanına dış dünyadan gelen bilgiyi ifade eder.

Ara katmanda ve çıkış katmanındaki nöronların çıkısının hesaplanabilmesi için ilgili nörona gelen net girdinin hesaplanması gerekir. Bunun için gelen bilgi ve ağırlık çarpımı kullanılır. Bu durum eşitlik 2.10' da gösterilmiştir.

$$Net_j = \sum_{i=1}^{n} W_{iJ} F_i \tag{2.10}$$

Burada;

Net_j : *j.* prosesin net girdisini ifade eder.

F_i : *j.* nörona bilgi gönderen nöronların çıkış bilgisidir.

W_{ij} : *j.* nörona bilgi gönderen *i.* nöron *j.* nöron arası ağırlığı ifade eder.

Ara katman ve çıkış katmanı nöronları için net bilginin sigmoid fonksiyonu kullanılarak nöron çıkışına dönüştürülmesi işlemi yapılır. Bu işlem eşitlik 2.11' de gösterilmiştir.

$$F_j = \frac{1}{1+e^{(-Net_j+\beta)}} \tag{2.11}$$

Çıkış katmanında elde edilen çıkış bilgisi ile olması gereken çıkış bilgisi arası fark hatayı oluşturur. Hata kavramı eşitlik 2.12' de gösterilmiştir.

$$E_m = B_m - C_m \tag{2.12}$$

Ağın test başarısı aşağıdaki formülle hesaplanır. Ağın test başarısı eşitlik 2.13' te görüldüğü üzere hesaplanmaktadır.

$$P = \frac{D}{T} x100 \tag{2.13}$$

2.8.5. J48 Algoritması

Weka'nın sınıflandırma algoritmalarından birisi olan J48 algoritması, temel olarak C4.5 Karar Ağacı algoritmasını kullanır. Yani C4.5 algoritmasının Weka sınıflandırma aracındaki gerçeklenmesi J48 olarak bilinir. C4.5 karar ağacının oluşturulması aşağıdaki gibi özetlenebilir.

1. Çıkış değerlerini en fazla farklılaştıran öznitelik seçilir.

2. Seçilen özniteliğin her değeri için farklı bir dal yaratılır.
3. Seçilen düğümdeki öznitelik değerlerini yansıtacak şekilde örnekler alt gruplara ayrılır.
4. Her alt grup için öznitelik seçimi durdurulur; Eğer
   a. Alt gruptaki tüm üyeler aynı çıkış değerini üretiyorsa, ağacın ilerlemesi durdurulur ve çıkış değeri olarak son belirlenen değeri atanır.
   b. Alt grupta tek düğüm kaldıysa veya ayırt edici öznitelikler belirlenemiyorsa ağacın ilerlemesi durdurulur.
5. 3.aşamada belirlenen her alt grup için yukarıdaki işlem tekrarlanır.

C4.5 algoritması ile sayısal değerler içeren veri tabanları üzerinde karar ağaçlarının oluşturulma olanağı sağlamıştır.

Sayısal nitelikleri belirli aralıklara bölme konusunda bazı zorluklar görülebilir. Ancak en uygun t eşik değerini hesaplamak için çeşitli yöntemler bulunmaktadır. Nitelik değerleri sıralanır ve $\{V_1, V_2, \dots, V_n\}$ şeklini alır. Nitelik değerler kümesi iki parçaya ayrılır ve Eşik değeri olarak $[V_i, V_{i+1}]$ aralığının orta noktası alınabilir:

$$t_i = (V_i + V_{i+1}) / 2 \qquad (2.14)$$

Örneğin bir eğitim kümesinde yer alan:
(70, 90, 85, 95, 70, 90, 78, 65, 75, 81, 70, 81, 70, 96)

Değerleri nitelik kümesi {65, 70, 75, 81, 85, 90, 95, 96} değerlerine sahiptir. Kümenin orta noktaları olan (81, 85) aralığının orta noktası olan:
$t_i = (V_i + V_{i+1}) / 2 = (81 + 85) / 2 = 83$ noktası eşik değer olarak alınabilir.

Bu durumda bu değerleri içeren değişken için (nitelik $\leq$ 83) veya (nitelik > 83) testi uygulanarak hesaplamalara devam edilir ve entropi, kazanım değerleri bulunur [61].

2.9. Sınıflandırma Algoritmalarının Başarısını Test Etme Yöntemi

Bütün veri madenciliği modellerinin performansını hesaplamak için standart bir ölçütün kullanılması önemlidir. Veri madenciliğinde sınıflama modellerinin karşılaştırılması için en sık kullanılan yöntem hata oranını hesaplamaktır. İki sınıflı model için sınıflama matrisi Çizelge 2.7.'de verilmiştir.

Çizelge 2.7. Sınıflama Matrisi

	Gerçek Sınıf		
Modelin Sınıf Tahmini		**Pozitif**	**Negatif**
	Pozitif	Doğru Pozitif Sayısı (DP)	Yanlış Pozitif Sayısı (YP)
	Negatif	Yanlış Negatif Sayısı (YN)	Doğru Negatif Sayısı (DN)

Modeli oluşturan toplam örnek sayısı = N = DP + YP + YN + DN (2.14)

Modelin doğru sınıflama oranı (Doğruluk) = (DP + DN) / N (2.15)

Doğru Pozitif Oranı(Duyarlılık) = DP / (DP + YN) (2.16)

Doğru Negatif Oranı (Belirlilik) = DN / (DN + YP) (2.17)

Hassaslık (Precision) = DP / (DP + DN) (2.18)

2.10. Kullanılan Etkin Değişkenlerin Tanımlanması

2.10.1. Yaş ve Cinsiyet:

Hipertansiyon, yaşla artan toplumsal bir sağlık sorunudur. Erken tanı ile kontrol altına alınabilen, geç kalındığında ise kalp hastalıkları, serebrovasküler hastalıklar gibi ölümcül seyreden komplikasyonlara yol açan ciddi bir hastalıktır.

Hipertansiyonun menopoz yaşına kadar kadınlarda daha az, erkeklerde daha fazla görüldüğü belirlenmiştir [62]. Fakat menopoz yaşından sonra kadınlarda ve erkeklerde benzer sıklıkta görülmektedir. Yaş ilerledikçe, damar sertliğinin artmasına bağlı olarak hipertansiyon daha sık görülüyor olup, %20 civarında olan oran ileri yaşlarda %50'lere ulaşmaktadır. Birçok sanayileşmiş ülkede kan basıncı yaşla birlikte artmaktadır [63-64].

2.10.2. Vücut Kütle İndeksi:

Obezite hipertansiyon gelişiminde etkili bir risk faktörüdür. Hipertansiyonla beraber obezitenin koroner kalp hastalıklarının oluşumunu hızlandırdığı birçok araştırmada gösterilmiştir. Özellikle vücut üst ve orta bölge tipi obezitelerde bu durum daha belirgindir [65].

Artan bel kalça oranının (BKO) kan basıncı üzerine önemli derecede etkisi vardır. Birçok çalışmada vücut kütle indeksi (VKİ) artışıyla hipertansiyon yaygınlığının anlamlı ölçüde yükseldiği gösterilmiştir [66]. Otuz iki ülkedeki 52 merkezde gerçekleştirilen geniş çaplı ve çok uluslu yeni bir çalışma (Intersolt Study); obezite, alkol ve mineral alımının, yaşla birlikte artan kan basıncı üzerindeki rolünü ortaya koymuştur [67]. Obezite ve hipertansiyon, özellikle gelişmiş ülkelerde sık rastlanılan sağlık sorunlarıdır. Obez kişilerde hipertansiyonun yaygınlığı % 25-50 arasındadır ve genel popülasyona göre daha sıktır. Hipertansif bireylerde ise % 15-20 arasında obezite görülmektedir. Bu oran, hipertansiyon olmayan (normotansif) bireylerde saptanan % 5'lik obezite oranından çok daha yüksektir [68].

2.10.3. Lipid Profili:

Hiperlipidemi, plazma lipoproteinlerinin artması sonucu olmaktadır (kolesterol, trigliserid). Bir veya daha fazla lipoprotein tipinin dolaşımda artmış akışı (sekresyonu) ve üretimi sonucu meydana gelebildiği gibi, bazı vakalarda dolaşımdan atılmasında ya da temizlenmesinde azalma sonucunda da olabilir. Bazen her iki süreç birlikte de görülebilir [69,70]. Kolesterol, hücre membranlarında bulunan bir lipiddir ve safra asitleri ve steroid hormonlarının öncülüdür. Kolesterol, lipid ve proteini içeren (lipoprotein) farklı partiküller şeklinde kanda dolaşmaktadır. Kolesterol için, <200 mg/dL, istenen kolesterol, 200-239 mg/dl, sınırda kolesterol ve 240 mg/dL ve üzeri yüksek kolesterol olarak sınıflandırılır [70]. Lipoproteinlerin 3 ana sınıfı bulunmaktadır: Düşük dansiteli lipoprotein (LDL), yüksek dansiteli lipoprotein (HDL) ve çok düşük dansiteli lipoprotein (VLDL). VLDL ve LDL arasında olan diğer bir lipoprotein orta dansiteli lipoprotein (IDL), klinik pratikte LDL ölçümü içinde değerlendirilir [71]. LDL kolesterol total serum kolesterolün %60-70'ini oluşturur. LDL, en önemli düzensiz gelişen lipoprotein olduğu için, kolesterol düşürücü tedavide primer hedef olarak belirlenmektedir. LDL reseptörler yoluyla kolesterolü karaciğerden başka dokulara taşır. HDL'ler, karaciğerde ve ince bağırsak duvarında sentezlenir ve total serum kolesterolünün %20-30'unun oluşturan bir lipoproteindir. HDL, kolesterolü dokulardan karaciğere taşır [71].

Çizelge 2.8.' de erişkin açlık kanı lipidleri için tavsiye edilen sınırlar verilmiştir.

Çizelge 2.8. Erişkin Açlık Kanı Lipidleri İçin Tavsiye Edilen Sınırlar

Plazma/Serum Lipidleri	Optimal, mg/dL	Sınırda-yüksek, mg/dL	Yüksek risk, mg/dL
Total kolesterol	< 200	200-239	>240
Trigliserid	< 150	150-199	> 200
LDL	< 100	130- 159	>160
HDL	>60		< 40

Aşırı kilo ve obezite yüksek kan basıncı ile ilişkilidir (aşırı kilolu bireylerin yaklaşık %23'ünde, obezlerin yaklaşık %35'inde hipertansiyon mevcuttur) ve şeker hastalığına da neden olabilir (diyabetiklerin %67'sinde VKİ >27 kg/m2, %46'sında VKİ >30 kg/m2'dir). VKİ değerleri normalden obeze doğru yükseldikçe total kolesterol, LDL ve trigliserid değerleri belirgin olarak yükselmekte, HDL değerleriyse düşmektedir [71].

2.10.4. Ürik Asit:

Ürik asit proteinlerden oluşur. Özellikle kırmızı et yiyenlerde ürik asit daha fazla artar. Alkol ve bira ürik asit seviyesini artırır. Ürik asit pürin metabolizmasının son ürünüdür. Hiperürisemi ile hipertansiyon ve kardiyovasküler hastalık gelişim riskinde artış olduğu bilinmektedir. Ayrıca diyabetik hastalarda yüksek ürik asit düzeyleri, ilerleyici böbrek hastalığı gelişimi için de bir risk faktörüdür.

Ürik asit yüksekliği, hipertansiyona ve böbrekle bağlantılı damarlarda renal hasara neden olur.

Serum ürik asit (SÜA) seviyeleri hipertansiyonda artar. Tedavi edilmeyen hastalarda %25, diüretik alanlarda %50, kötü huylu (maling) hipertansiyon olanların %75'inde fazlasında SÜA seviyeleri artmıştır. [72,73].

2.10.5. Sigara Kullanımı:

Sigara içiminin kan basıncında yaklaşık 15-30 dakika süreyle ve tekrarlandığı takdirde geçici olarak 5-10 mmHg kadar akut bir yükselmeye yol açtığı, sigara alışkanlığı olan hipertansiyon olmayan kişilerde bu etkinin günün ilk sigarasında daha belirgin olduğu ve günün ilk sigarasından sonra sistolik kan basıncında 20 mmHg'ye kadar yükselme olabildiği bildirilmiştir [74].

3. ARAŞTIRMA BULGULARI

Bir kişiye hipertansiyon hastası diyebilmek için, farklı zamanlarda ölçülen tansiyonunun en az iki defa normal sınırların üzerinde çıkması gerekir. Tansiyon ölçümü yapmadan bir kişiye hipertansiyon teşhisi konulması oldukça zordur. Ancak insanların demografik bilgileri, kan ve idrar değerleri hipertansiyonla yakından ilgilidir. Hastanın HDL, LDL, Trigliserid, Ürik Asit değerlerinin normal değerlerinden farklı olması durumunun hipertansiyonu tetiklediği bilinmektedir. Ayrıca yaş, cinsiyet, kilo ve sigara kullanımı da hipertansiyon ile doğru orantılı olarak değişim göstermektedir.

Bu bölümde, yapılan uygulama çalışması ve bulgularına yer verilmiştir. Uygulamada veri madenciliği süreci adımları sırasıyla tek tek ele alınmış ve elde edilen bulgular belirtilmiştir. Tez uygulama çalışmasının temel amacı olan hipertansiyona etki eden faktörler, Weka veri madenciliği aracı kullanılarak incelenmiştir.

3.1. Verinin Tanımlanması ve Hazırlanması

Yukarıda tanımlanmış bilgiler ışığında, hipertansiyona etki etmesi muhtemel faktörler, uzman hekimlerle yapılan ortak bir çalışmayla belirlenmiş ve bu faktörlerin; Yaş, Cinsiyet, Boy, Kilo, Lipid Profili(HDL ve LDL), Trigliserid, Ürik Asit ve Sigara Kullanımı(günde içtiği paket sayısı ve kaç yıldır içtiği bilgisi) olması gerektiği vurgulanmıştır. Hastalardan bu değerler alınırken uyulması gereken kriterler ise, hastanın 30 yaş ve üzeri olması, hamile olmaması ve herhangi bir ilaç tedavisine başlamamış olmasıdır.

Bu kriterlere uygun olarak, Kırıkkale Yüksek İhtisas Hastanesi Dahiliye Bölümü'nde 184 kişiden değerler alınmış, sonuçta bu kişiler hipertansiyon hastası ve sağlıklı olarak 2 sınıfa ayrılmıştır.

Veriler üzerinde yapılan incelemeler sonucunda;

- Veri tabanında sadece gerekli alanlar bırakılmıştır(Örneğin hastanın ismi gibi bilgiler sınıflandırmada bir önem teşkil etmeyeceği için silinmiştir).
- Veri madenciliğinde verilerin rahat modellenebilmesi için bazı alanların yapısı değiştirilmiştir(Örneğin Sigara Kullanımı, günde içtiği paket sayısı ile kaç yıldır içtiği bilgisi çarpılarak tek bir alana indirilmiştir. Aynı zamanda boy ve kilo tanımları birleştirilerek Vücut Kütle İndeksi kullanılmıştır).
- Madencilikte verilerin rahat modellenebilmesi için bazı alanlarda tutulan değerlerin yapısı değiştirilmiştir(Örneğin Cinsiyet verisinde Erkekler için 1 değeri, Bayanlar için ise 0 değeri kullanılmıştır).
- Alan bilgilerinin çoğunluğu boş olan 34 kaydın modellemeye doğru etkisi olmayacağı için bu kayıtlar silinmiştir. Son haliyle veri setinde geriye 150 kayıt kalmıştır.

Veri temizleme işleminden sonra veri madenciliği çalışması için kullanılacak veri seti üzerindeki verilerin dağılımı Çizelge 3.1.'de gösterilmiştir.

Çizelge 3.1. Veri madenciliği çalışması için kullanılacak verilerin dağılımı

Sınıf	**Sayı**
Normal	65
Hasta	85
Toplam	150

Kalan 150 adet kayıttan 85 tanesi hasta olarak belirlenmiş, kalan 65 kayıt ise normal olarak sınıflandırmaya katılmıştır.

Weka veri madenciliği aracını kullanmak için veriler kullanılmadan önce Attribute-Relation File Format (ARFF) formatına dönüştürülmesi gerekir. Arff dosyaları, değişken tanımlamasına izin veren ASCII metin dosyasıdır. Arff dosyasının başlık

kısmında, değişkenler(veri tabanındaki her bir kolonun ismi), bunlar arasındaki ilişkiler ve her bir değişkenin türü ve alacağı değer vs. bulunur.

Bu çalışmada kullanılan verileri Arff formatına dönüştürmek için Microsoft Excel programından faydalanılmıştır. Şekil 3.1.'de, oluşturulan Arff dosyasının başlık kısmı verilmiştir.

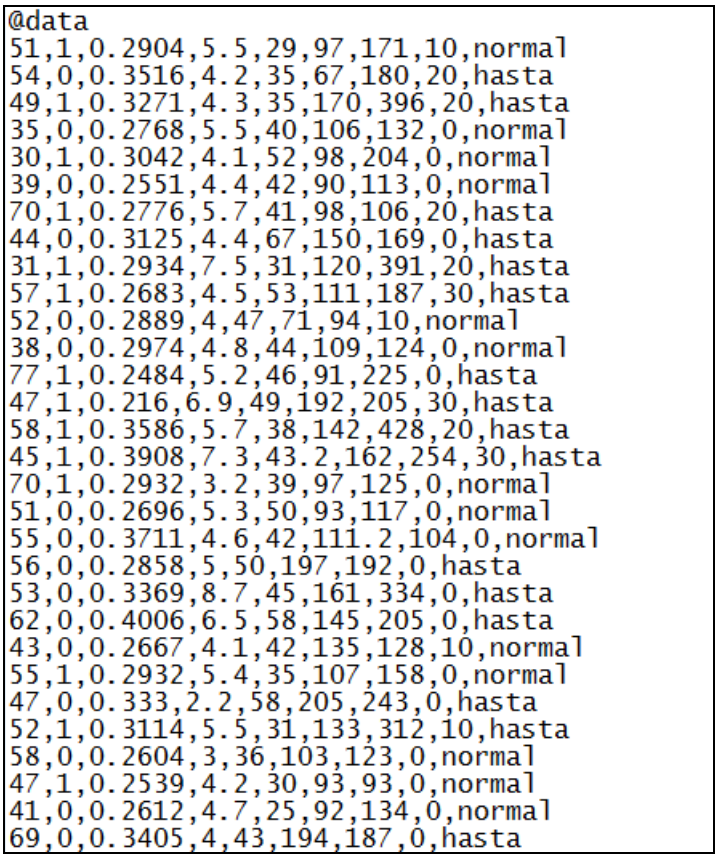
```
@relation hipertansiyon

@attribute yas numeric
@attribute cinsiyet numeric
@attribute bmi numeric
@attribute urikAsit numeric
@attribute ldl numeric
@attribute hdl numeric
@attribute trigliserid numeric
@attribute sigara numeric


@attribute Sonuc {hasta,normal}
```

Şekil 3.1. Çalışmada oluşturulan Arff dosyasının başlık kısmı

Verilerin bulunduğu kısım ise @DATA satırından sonra gelir ve Şekil 3.2.'deki gibi gösterilir.



Şekil 3.2. Çalışmada oluşturulan Arff dosyasında verilerin bulunduğu kısım

3.2. Modelin Kurulması

Arff formatına dönüştürülmüş veri dosyası, Weka programı üzerinden açılmıştır ve önce önişleme (Preprocess) işleminden geçilmiştir. Bu işlem adımında Weka'nın "Explorer" ara yüzü kullanılarak arff dosyası haline dönüştürülen .arff dosyası açılmıştır. Model kurulurken verilerin belli bir kısmı eğitilmiş(training), eğitilen verilerin oluşturduğu örüntüler kullanılarak geri kalan veriler test edilmiştir. Örüntü bulma ve test veri sınıflarının tahmini işlemlerini Weka sınıflandırma algoritmaları yapmıştır. Verileri işlemede tahmin edilecek alan için doğruluk performansı yüksek olan algoritma ele alınmıştır.

Veriler işlenmeden önce uzman görüşü de alınarak verilerde tahmin edilecek çıkış alanları belirlenmiştir. Buna göre alanların kullanım türleri Çizelge 3.2'de verilmiştir.

Çizelge 3.2. Veri işlemede alanların kullanım türü

Veri İşlemede Kullanım Türü	**Alan Adı**	**Veri İşlemede Kullanım Türü**	**Alan Adı**
GİRİŞ	yas cinsiyet bmi uricAsit Ldl Hdl trigliserid sigara	ÇIKIŞ	Hasta Normal

Sınıflandırma algoritmalarının performanslarını test etmek için Weka'daki sınıflandırma algoritmalarından Naive Bayes Algoritması, Çok Katmanlı Algılayıcı (Multilayer Perceptron), IB1 Algoritması, Karar Tablosu (Decision Table) Algoritması ve J48 Algoritması kullanılmıştır.

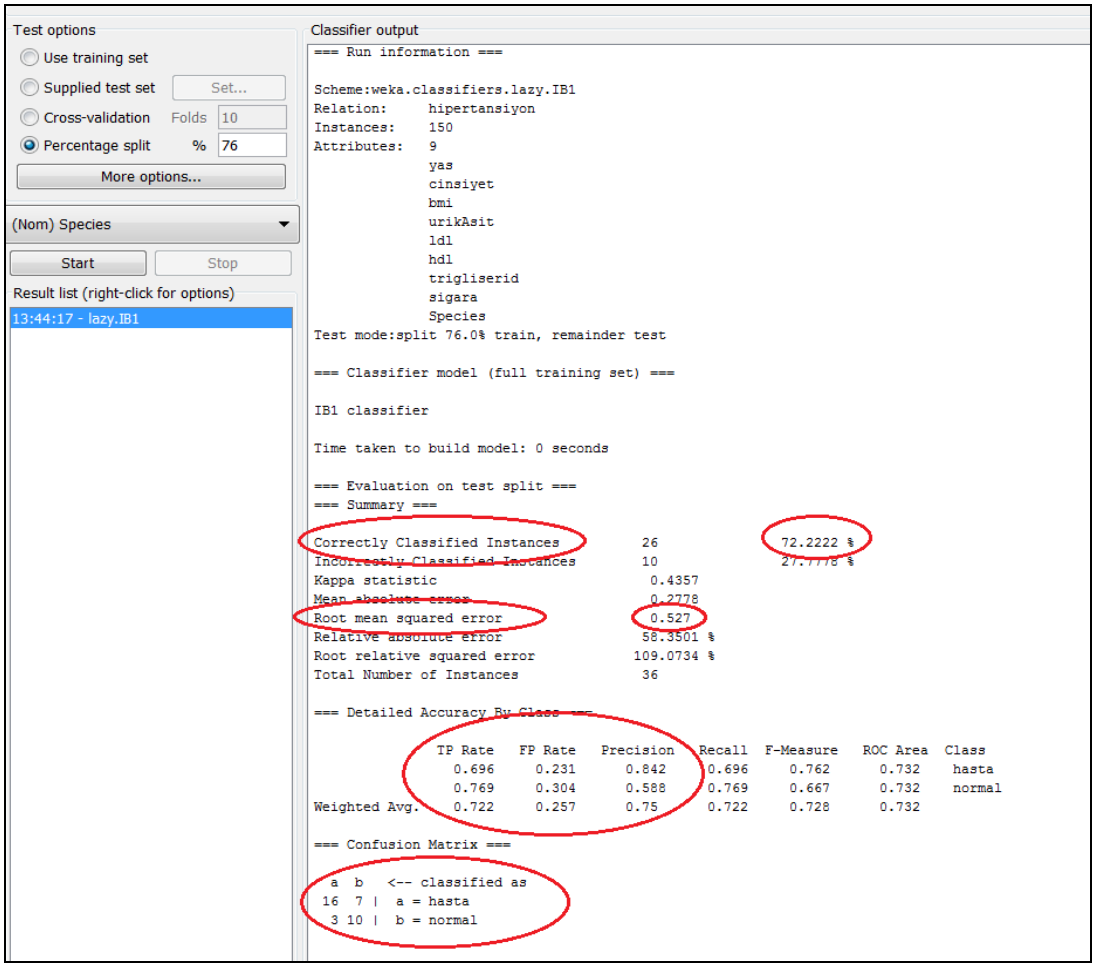
Çalışmada kullanılan Weka 3.6.6 sürümünde yer alan sınıflandırıcılar sırayla seçilmiştir. Test seçeneği olarak tüm sınıflandırıcılar için yüzde ayırma (percentage split) yöntemi kullanılmıştır.

Yüzde ayırma yönteminde verilerin %76'sı eğitim (training) için, geri kalan %24'ü de test verisi olarak kullanılmıştır. Yani bir nevi seçilen algoritma, verilerin %76'sı arasında ilişki ve kuralları belirleyerek çeşitli örüntüler oluşturmuştur. Oluşan örüntü desenlerine göre de algoritmanın doğruluğu, verilerin kalan %24'ü üzerinde test edilmiştir.

3.3. Modelin Değerlendirilmesi

Şekil 3.3.'te IB1 algoritmasının hipertansiyon tahmini için veri modellemesi sonuç ekranı görülmektedir.

Doğru model seçimi için uygulama çalışmasında doğruluk yüzdesi ve düzensizlik matrisi kriterleri değerlendirilmiştir.



Şekil 3.3. Verinin IB1 algoritmasıyla modellenmesi

Şekil 3.3.'teki çıkış ekranı örnek alınarak sınıflandırma algoritmasının doğruluk yüzdesi şöyle hesaplanmıştır:

150 kaydın %76'sı modelin eğitimi için kullanılmıştır:
150 * (76 / 100) = 114
150 – 114 = 36

36 adet veri test amaçlı kullanılmıştır. Şekil 3.3.'de sınıflandırıcı çıkış ekranında görülen Düzensizlik Matrisine göre (Confusion Matrix) 36 verinin düzensizlik matrisi Çizelge 3.3.'de görüldüğü gibidir.

Çizelge 3.3. IB1 Algoritması için düzensizlik matrisi

a	**b**	
16	7	a=hasta
3	10	b=normal

Çizelge 3.3.'e göre gerçekte 23 "hasta" değerli test verisinin 16 tanesi hasta, 7 tanesi "normal" ve gerçekte 13 "normal" değerli test verisinin 10 tanesi "normal", 3 tanesi "hasta" olarak tahmin edilmiştir.

Buna göre toplam 7 + 3 = 10 veri yanlış sınıflandırılmış olup, diğer 16 + 10 = 26 veri doğru sınıflandırılmıştır. Buna göre bu modelin doğru sınıflandırma yüzdesi şöyle hesaplanmıştır:

Doğruluk Yüzdesi = (26 / 36) * 100 = %72,2222'dir.

Şekil 3.3.'te sınıflandırıcı çıkış ekranında görülen sınıflara göre 36 verinin detaylandırılmış doğruluk tablosu Çizelge 3.4.'te verilmiştir.

Çizelge 3.4. Detaylandırılmış doğruluk tablosu

Duyarlılık (DP Oranı)	**Yanlış Hassaslık Oranı (YP Oranı)**	**Hassaslık (Precision)**	**Sınıf (Class)**
0,696	0,231	0,842	a = hasta
0,769	0,304	0,588	b = normal

Çizelge 3.3.'e göre Çizelge 3.4.'teki değerler aşağıdaki şekilde hesaplanmıştır:

Duyarlılık (Doğru Pozitif Oranı) a sınıfı için 16 / 23 = 0,696
b sınıfı için 10 / 13 = 0,769
Yanlış Hassaslık Oranı (Yanlış Pozitif Oranı) a sınıfı için 3 / 13 = 0,231
b sınıfı için 7 / 23 = 0,304
Hassaslık (Precision) a sınıfı için 16 / 19 = 0,842
b sınıfı için 10 / 17 = 0,588

Çizelge 3.5.'te Çok Katmanlı Algılayıcı (ÇKA) algoritmasının hipertansiyon tahmini için Düzensizlik Matrisi verilmiştir.

Çizelge 3.5. ÇKA Algoritması için düzensizlik matrisi

a	**b**	
19	4	a=hasta
1	12	b=normal

Çizelge 3.5.'te verilen Düzensizlik Matrisine göre gerçekte 23 "hasta" değerli test verisinin 19 tanesi hasta, 4 tanesi "normal" ve gerçekte 13 "normal" değerli test verisinin 12 tanesi "normal", 1 tanesi "hasta" olarak tahmin edilmiştir.

Çizelge 3.5.'te verilen Düzensizlik Matrisine göre 36 verinin detaylandırılmış doğruluk tablosu Çizelge 3.6.'da verilmiştir.

Çizelge 3.6. ÇKA için detaylandırılmış doğruluk tablosu

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
%86,11	0,826	0,077	0,95	a = hasta
	0,923	0,174	0,75	b = normal

Çizelge 3.6.'ya göre Çok Katmanlı Algılayıcı için doğruluk yüzdesi, %86,11 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, hasta ve normal sınıfları için ayrı ayrı hesaplanmıştır.

Çizelge 3.7.'de Karar Tablosu (Decision Table) algoritmasının hipertansiyon tahmini için Düzensizlik Matrisi verilmiştir.

Çizelge 3.7. Karar Tablosu Algoritması için düzensizlik matrisi

a	b	
22	1	a=hasta
2	11	b=normal

Çizelge 3.7.'de verilen Düzensizlik Matrisine göre gerçekte 23 "hasta" değerli test verisinin 22 tanesi hasta, 1 tanesi "normal" ve gerçekte 13 "normal" değerli test verisinin 11 tanesi "normal", 2 tanesi "hasta" olarak tahmin edilmiştir.

Çizelge 3.7.'de verilen Düzensizlik Matrisine göre 36 verinin detaylandırılmış doğruluk tablosu Çizelge 3.8.'de verilmiştir.

Çizelge 3.8. Karar Tablosu Algoritması için detaylandırılmış doğruluk tablosu

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
%91,67	0,957	0,154	0,917	a = hasta
	0,846	0,043	0,917	b = normal

Çizelge 3.8.'e göre Karar Tablosu Algoritması için doğruluk yüzdesi, %91,67 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, hasta ve normal sınıfları için ayrı ayrı hesaplanmıştır.

Çizelge 3.9.'da Saf Bayes (Naive Bayes) algoritmasının hipertansiyon tahmini için Düzensizlik Matrisi verilmiştir.

Çizelge 3.9. Saf Bayes Algoritması için düzensizlik matrisi

a	b	
20	3	a=hasta
0	13	b=normal

Çizelge 3.9.'da verilen Düzensizlik Matrisine göre gerçekte 23 "hasta" değerli test verisinin 20 tanesi hasta, 3 tanesi "normal" ve gerçekte 13 "normal" değerli test verisinin 13 tanesi de "normal" olarak tahmin edilmiştir.

Çizelge 3.9.'da verilen Düzensizlik Matrisine göre 36 verinin detaylandırılmış doğruluk tablosu Çizelge 3.10.'da verilmiştir.

Çizelge 3.10. Saf Bayes Algoritması için detaylandırılmış doğruluk tablosu

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
%91,67	0,87	0	1	a = hasta
	1	0,13	0,813	b = normal

Çizelge 3.10.'a göre Saf Bayes Algoritması için doğruluk yüzdesi, %91,67 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, hasta ve normal sınıfları için ayrı ayrı hesaplanmıştır.

Çizelge 3.11.'de J48 algoritmasının hipertansiyon tahmini için Düzensizlik Matrisi verilmiştir.

Çizelge 3.11. J48 Algoritması için düzensizlik matrisi

a	b	
22	1	a=hasta
2	11	b=normal

Çizelge 3.11.'de verilen Düzensizlik Matrisine göre gerçekte 23 "hasta" değerli test verisinin 22 tanesi "hasta", 1 tanesi "normal" ve gerçekte 13 "normal" değerli test verisinin 11 tanesi "normal", 2 tanesi ise "hasta" olarak tahmin edilmiştir.

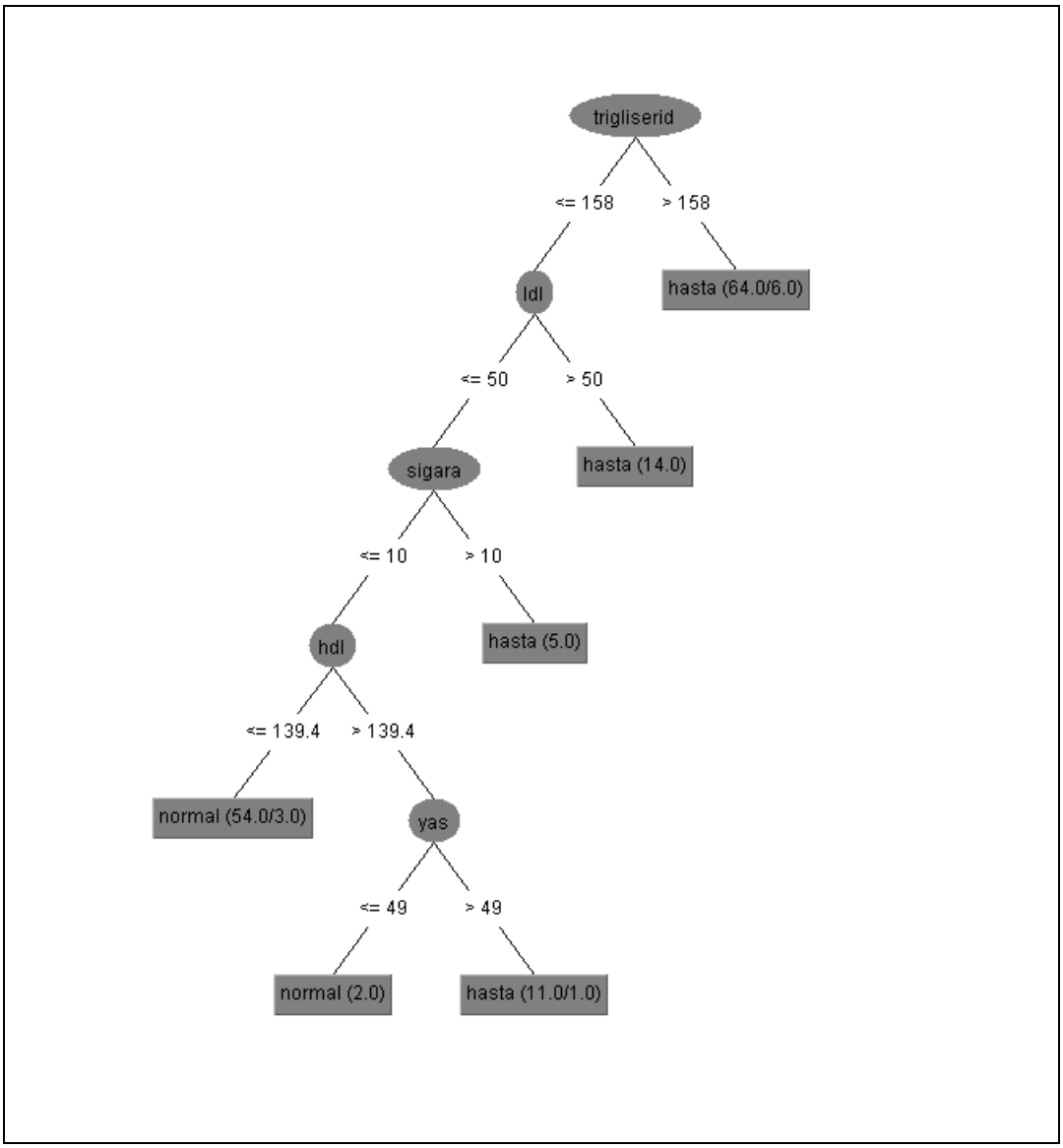
Çizelge 3.11.'de verilen Düzensizlik Matrisine göre 36 verinin detaylandırılmış doğruluk tablosu Çizelge 3.12.'de verilmiştir.

Çizelge 3.12. J48 Algoritması için detaylandırılmış doğruluk tablosu

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
%91,67	0,957	0,154	0,917	a = hasta
	0,846	0,043	0,917	b = normal

Çizelge 3.12.'ye göre J48 Algoritması için doğruluk yüzdesi, %91,67 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, hasta ve normal sınıfları için ayrı ayrı hesaplanmıştır.

Şekil 3.4.'te J48 algoritmasının hipertansiyon tahmini için karar ağacı sonuç ekranı görülmektedir.



Şekil 3.4. Verinin J48 algoritması için karar ağacı ekranı

Şekil 3.4.'e göre trigliserid düzeyinin 158'den büyük olması durumu, doğrudan hipertansiyonun varlığına işaret etmektedir. Trigliserid düzeyine bağımlı olarak LDL değerinin 50'den büyük olması durumu, yine hipertansiyonun varlığına işaret etmektedir. Sigara kullanımının 10'dan büyük olması da hipertansiyona işaret eden farklı bir durum olarak karşımıza çıkmaktadır. Hipertansiyona etki eden bir diğer etken ise yaş değişkenidir. Şekil 3.4.'e göre yaş değişkeni, doğrudan değil, dolaylı olarak hipertansiyona etki etmektedir.

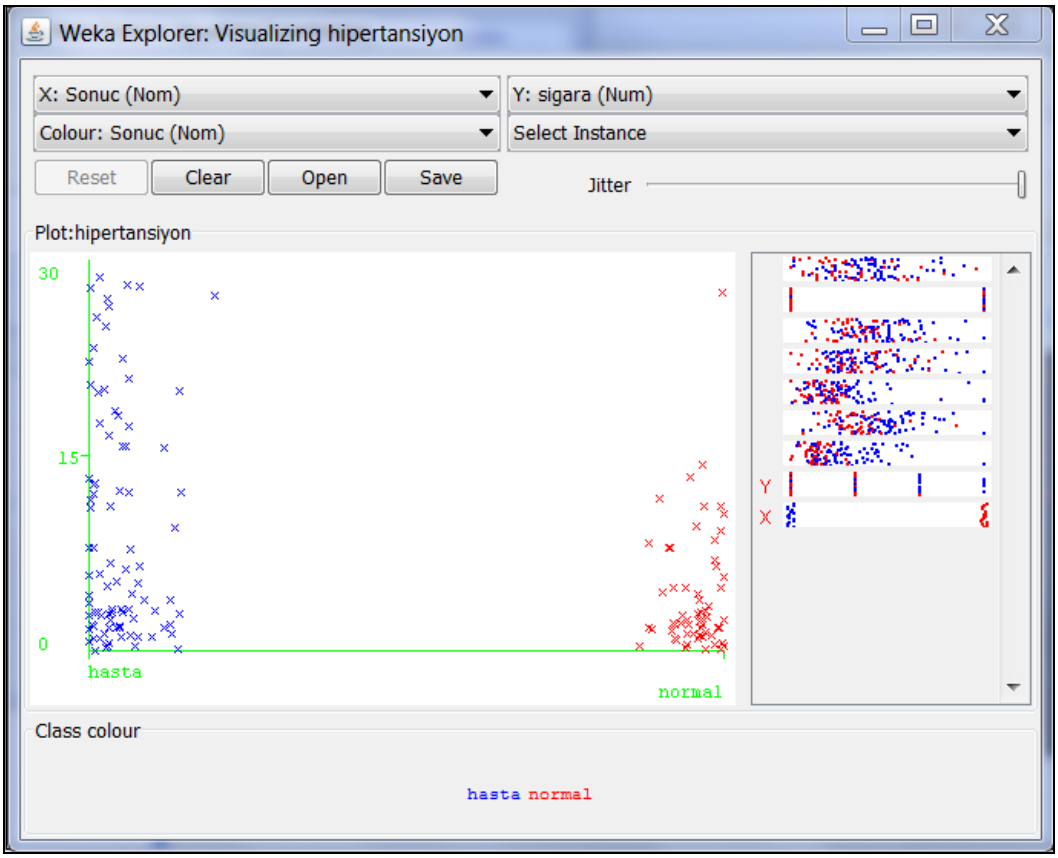
Weka'daki verilere uygun algoritmalar uygulanarak doğru sınıflandırma yüzdeleri ayrı ayrı bulunmuş ve bu sonuçlar Çizelge 3.13.'te karşılaştırılmıştır.

Çizelge 3.13. Seçilen sınıflandırma algoritmaları ve doğruluk yüzdeleri

Algoritma Adı	**Doğruluk Yüzdesi**
IB1 Algoritması	72,222
Çok Katmanlı Algılayıcı (Multilayer Perceptron)	86,111
Karar Tablosu (Decision Table)	91,667
Saf Bayes (Naive Bayes)	91,667
J48 Algoritması	91,667

Tahmin sonuçları Weka'da Şekil 3.5.'de grafiksel olarak raporlanmıştır.

Şekil 3.5.'de ilgili açılır listeden seçenekler değiştirilerek farklı veri alanı özelliklerine göre gruplandırılmış sonuçları izlemek mümkündür.



Şekil 3.5. Grafiksel tahmin aracı

Şekil 3.5.'te sigara kullanımı ile hipertansiyon arasındaki ilişki grafiksel olarak raporlanmıştır. Bu şekle göre sigara kullanımı arttıkça hipertansiyon hastalığı görülme sıklığının da aynı oranda arttığı görülmektedir.

4. TARTIŞMA VE SONUÇ

Günümüzde birçok sebepten dolayı insanların kan basıncında bazı düzensizlikler gözlemlenmektedir. Bu düzensizlikler tansiyon yüksekliği ya da tansiyon düşüklüğü olarak kendini göstermektedir. Hipertansiyon adını verdiğimiz tansiyon yüksekliği, birçok hastalığın da oluşmasına zemin hazırlayarak kişilerin yaşamını olumsuz bir şekilde etkilemektedir.

Birçok insan tansiyon hastası olduğunun farkına bile varmadan günlük yaşamına devam etmekte, olası hipertansiyon belirtilerini ise önemsememektedir. Oysaki çok basit bir baş dönmesi ya da baş ağrısı ile başlayan bu belirtiler; kalp krizleri, beyin kanamaları, böbrek yetmezliği gibi birçok ağır vakaya sebep olmaktadır.

Bu çalışmada kullanılan yöntem, hastaların demografik değerlerini ve kan değerlerini inceleyerek, hastaya cerrahi bir tetkik yapmadan hipertansiyon teşhisi yapmayı sağlamaktadır. Çalışma sayesinde hastaneye başka bir şikâyetinden dolayı giden bir kişi bile hipertansiyon hastası olduğunu öğrenebilir. Böylece birçok ciddi hastalıkların önlenmesi ve ömür boyu ilaç kullanılmasına gerek kalmadan sadece yaşam şeklini değiştirerek hipertansiyon ile başa çıkılabilir.

Çeşitli insan grupları üzerinde yapılan bu çalışmada yaş, cinsiyet, VKİ, Ürik Asit, HDL, LDL, Trigliserid düzeylerinin ve sigara kullanımının kan basıncını nasıl etkilediği incelenmiştir. Kişilerden elde edilen verilere, veri madenciliği sınıflandırma algoritmaları uygulanarak başarılı sonuçlar elde edilmiştir.

Veri madenciliği sınıflandırma algoritmalarından Karar Tablosu (Decision Table), Saf Bayes (Naive Bayes) ve J48, %91,667 ile en yüksek başarı oranını veren algoritmalar olmuştur. Bu üç algoritmanın başarılarını kıyaslamak için ise Düzensizlik Matrisi'ne bakılması gerekir. Düzensizlik Matrisine göre Karar Tablosu ve J48 algoritması, 1 kişiyi gerçekte "hasta" iken "normal" olarak sınıflandırmış, Saf Bayes (Naive Bayes) algoritması ise 3 kişiyi gerçekte "hasta" iken "normal" olarak sınıflandırmıştır. Gerçekte hasta olan bir kişiye sağlıklı teşhisi koymak, tıbben daha

sakıncalı olduğu için, Karar Tablosu ve J48 algoritmaları, Saf Bayes algoritmasına göre daha başarılıdır.

J48 algoritmasının karar ağacı ekranına göre hipertansiyona etki eden faktörler incelendiğinde, Trigliserid düzeyinin 158'den büyük olması durumu, doğrudan hipertansiyon riskini ortaya koymaktadır. LDL'nin 50'den büyük olması durumu da hipertansiyon riskini ortaya koymaktadır. Hipertansiyona etki eden diğer bir faktör ise sigara kullanımıdır. Trigliserid ve LDL'ye bağlı olarak sigara kullanımının 10'dan büyük olması durumunda hipertansiyondan bahsedilebilir. HDL seviyesinin 139,4'ten büyük olması durumunda ise bakılması gereken faktör yaştır. Böyle bir durumda yaşın 49'dan büyük olması durumunda hipertansiyondan bahsedilebilir.

Bu durumda hipertansiyona en çok etki eden faktörlerin Trigliserid düzeyi, LDL değeri ve sigara kullanımı olduğu görülmüş, HDL değeri ve yaşın ise dolaylı olarak hipertansiyona etki ettiği saptanmıştır. Çalışmada kullanılan Vücut Kütle İndeksi (BMI) ve Ürik Asit değerlerinin ise herhangi bir etkisi gözlemlenememiştir. Bu durum ise, çalışmada kullanılan veri sayısının ve çeşitliliğinin azlığına yorulmuştur.

İleriki dönemlerde, bu çalışmada kullanılan veri miktarı artırılarak algoritmaların daha kapsamlı örüntüler oluşturup veri madenciliğinden daha iyi sonuçlar alınması sağlanabilir. Gelecekte bu uygulama, daha geniş bir veri seti kullanılarak, daha yüksek bir doğruluk oranıyla gerçekleştirilebilir.

KAYNAKLAR

[1] Anonim, http://www.ozalpdh.gov.tr/haber_detay.asp?haberID=18 (Erişim Tarihi:24.11.2011)

[2] Köktürk,F., Ankaralı,H., Sümbüloğlu,V., "Veri Madenciliği Yöntemlerine Genel Bakış", Turkiye Klinikleri J Biostat, 1(1):20-5 (2009)

[3] Anonim, http://www.scribd.com/doc/47275291/Tıpta-Veri-Madenciliği-Murat-Azimli (Erişim Tarihi: 12.12.2011)

[4] Shah,S., Kusiak,A., "Cancer gene search with data-mining and genetic algorithms", Computers in Biology and Medicine, 37 : 251 – 261 (2007)

[5] Kusiak, A., Dixon, B., Shah, S., 2005. Predicting Survival Time for Kidney Dialysis Patients: A Data Mining Approach. Computers in Biology and Medicine, 35, 311–327.

[6] Chae,Y.M., Ho,S.E.,Cho,K.W.,Lee,D.H.,Ji,S.E., **"**Data mining approach to policy analysis in a health insurance domain**" ,** International Journal of Medical Informatics, Volume 62, Issues 2-3, 103-111 (2001)

[7] Razali, A.M., Ali, S., " Generating Treatment Plan in Medicine: A Data Mining Approach", Am J of Applied Sciences, 6 (2):345-351 (2009)

[8] Persson, M., Lavesson, N., "Identification of Surgery Indicators by Mining Hospital Data: A Preliminary Study", 20. DEXA Workshops 2009: Linz, Austria, First International Workshop on Database Technology for Data Management in Life Sciences and Medicine DBLM 2009, 323-327 (2009)

[9] Tsumoto, S., Hirano, S., "Data Mining as Complex Medical Engineering", Journal of Japanese Society for Artificial Intelligence, 22 (2):201-207 (2007)

[10] Yeh, J. , Wu, T. , " A Decision Support System for PredictingHospitalization of Hemodialysis Patients ", World Academy of Science,Engineering and Technology, 53:1128-1138 (2009)

[11] Huang, M. J., Chen, M.Y., Lee, S.C., "Integrating data mining with casebased reasoning for chronic diseases prognosis and diagnosis", Expert Systems with Applications, 32:856-867 (2007)

[12] Wren, J., Garner, H., " Data mining analysis suggests an epigenetic pathogenesis for Type 2 Diabetes ", Journal of Biomedicine and Biotechnology, 2:104–112 (2005)

[13] Le Duff, F., Muntean, C., Cuggia, M., Mabo, P., " Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", MEDINFO, 1256-1259, (2004)

[14] Palaniappan, S., Awang, R., " Intelligent Heart Disease Prediction System Using Data Mining Techniques", Int. J. of Computer Science and Network Security, 8 (8):108 - 115 (2008)

[15] Antonie, M., Zaiane, O., Coman, A., " Application of Data Mining Techniques for Medical Image Classification", 2nd Int. Workshop on Multimedia Data Mining, San Fransisco, USA, 94-101 (2001)

[16] Sleeman, D., Fluck, N., Gyftodimos, E., Moss, L., Christie, G., "An Intelligent Aide for Interpreting a Patient's Dialysis Data Set", Blood Pressure Prediction & Clustering, 89:169-179 (2007)

[17] Almazyad,A.S., Ahamad,M.G., Siddiqui,M.K., "Effective Hypertensive Treatment Using Data Mining In Saudi Arabia", Journal of Clinical Monitoring and Computing, 24:391-401 (2010)

[18] Chang,C.D.,Wang,C.C.,Jiang,B.C., “Using Data Mining Techniques For Multi-Diseases Prediction Modeling of Hypertension and Hyperlipidemia by Common Risk Factors”, Expert Systems with Applications 38:5507-5513 (2011)

[19] Ture,M.,Kurt,I.,Kurum,A.T.,Ozdamar,K., “Comparing Classification Techniques for Predicting Essential Hypertension”, Expert Systems with Applications 29:583-588 (2005)

[20] Anonim, http:// www.bayar.edu.tr/bilisim/dokuman/inceoglu.doc (Erişim Tarihi: 02.02.2012)

[21] Kaya, E., Bulun, M., Arslan, A., " Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları", AKADEMİK BİLİŞİM 2003, Çukurova Üniversitesi, Adana, (2003)

[22] Prof.Dr. Akpınar, H., " Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi, 29 (1):1-22 (2000)

[23] Karakas, M., 2002. Veri Madenciliği Üzerine. http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=132. Erisim Tarihi:20.02.20012.

[24] Han, J., Kamber, M., " Data Mining Concepts and Techniques 2nd Ed.", Editor : Jim Grey, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann 2,8,12,14,15,29,30,398-403, (2006)

[25] Anonim, http://yunus.hacettepe.edu.tr/hcingi/ist376a/6Bolum.doc (Erişim Tarihi:15.02.2012)

[26] Anonim, Current data mining applications / percentage in different industries, http://www.kdnuggets.com/polls/2003/data_mining_applications_industries.html (Erişim Tarihi:20.05.2012).

[27] Han, J., Kamber, M., " Data Mining Concepts and Techniques 2nd Ed.", Editor : Jim Grey, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann 2,8,12,14,15,29,30,398-403, (2006)

[28] ZHOU,Z. (2003). Three Perspectives Of Data Mining, Elsevier Science Publishers Ltd. Essex, UK, s.:139-146.

[29] Gürsoy, T.Ş., "Veri Madenciliği ve Bilgi Keşfi", Pegem Akademi Yayınları, Ankara, 2009

[30] FAYYAD, Usama, Gregory Piatetsky-Shapiro, ve Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, c.17,S.3(1996). s.44

[31] AYDOĞAN, Fatih, E-Ticarette Veri Madenciliği Yaklaşımlarıyla Müşteriye Hizmet Sunan Akıllı Modüllerin Tasarımı ve Gerçekleştirimi, Hacettepe Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Ankara, 2003, s.17

[32] HAN, Jiawei ve Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000 S.221

[33] ÖZEKEŞ, Serhat, "Veri Madenciliği Modelleri ve Uygulama Alanları", İstanbul Ticaret Üniversitesi Dergisi, S.3 (Haziran 2003), s.65-82

[34] YILMAZ, Emrah, Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği ile Tüketici Profilinin Çıkarılmasına Yönelik Bir Uygulama, Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Yüksek Lisans Tezi, Kütahya, 2006 s.104

[35] UĞUZ, H. vd., "Apriori Algoritması Kullanılarak Web Kullanım Madenciliği Yönteminin Web Log Kayıtlarına Uygulanması", IJCI Proceeding of International Conference on Signal Processing, C.1, S.2 (2000), s: 499-501

[36] KANTARDZIC, Mehmed, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, Danvers, 2003

[37] ALTINTOP, Ümmühan, Internet Tabanlı Öğretimde Veri Madenciliği Tekniklerinin Uygulanması, Kocaeli Üniversitesi Fen bilimleri Enstitüsü Yüksek Lisans Tezi, Kocaeli, 2006

[38] TANG, ZhaoHui ve Jamie MacLennan, Data Mining with SQL Server 2005, Wiley Publishing, Indianapolis, 2005

[39] HAIR, Joseph F. vd., Multivariate Data Analysis, Prentice Hall, New Jersey, 1998, s.680-695

[40] AGRAWAL, Rakesh ve Ramakrishnan Srikant, "Fast Algorithms For Mining Association Rules", Proceedings of the 20th VLDB Conference, Santiago, Şili, 1994, s.487-499

[41] Anonim, Türk Nefroloji Derneği, http://www.tsn.org.tr (Erişim Tarihi:12.4.2012)

[42 Anonim, http://www.beh.gov.tr/index.php?option=com_content&id=100&Itemid=103. (Erişim tarihi: 17.04.2012)

[43] Heart Disease and Stroke Statistics-2007 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation. 2007;115: 69-171.

[44] Summary of 1993 World Health Organisation-International Society of Hypertension Guidelines for The Management of Mild Hypertension Subcommittee of WHO/ISH Mild Hypertension Liaison Committee. BMJ. 1993;307: 1541-1546.

[45] Anonim, http://www.steteskop.net/hipertansiyon-ve-bobrek-hastaligi-Makale-No-594.html (Erişim Tarihi: 30.03.2012)

[46] ŞARLI, Ş., Hipertansiyon Hastalığı Olanlarda Tedaviye Uyum, Etkileyen Faktörler ve Yaşam Kalitesinin Değerlendirilmesi, Erciyes Üniversitesi Tıp Fakültesi Halk Sağlığı Anabilim Dalı, Tıpta Uzmanlık Tezi, Kayseri, 2011

[47] Joint National Committee on Detection, Evaluation and Treatment of High Blood Pressure (JNC VI).Arch Inter Med ;157:2413-46,1997.

[48] Türk Hipertansiyon Prevalans Çalışması. Türk Hipertansiyon ve Böbrek Hastalıkları Derneği 2003

[49] Kaplan N., Measurement of blood pressure. Clinical Hypertension. Williams and Wilkins, USA 7th ed. 19-39, 181-248,1998.

[50] Dener, M., vd., 2009. Açık kaynak kodlu veri madenciliği programları: WEKA'da örnek uygulama, 787-796. Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, Şubat 11-13, 2009, Harran Üniversitesi, Şanlıurfa.

[51] Alfred, R.,2005. Knowledge Discovery: Enhancing Data Mining and Decision Support Integration. The University of York, United Kingdom

[52] Wang, J. (Editor), "Encyclopedia of Data Warehousing and Mining", Information Science Reference, 49, 140 (2006)

[53] Witten, I.H. and Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

[54] Borysowich, C., 2007. Building Decision Tables, http://it.toolbox.com/blogs/enterprise-solutions/building-decision-tables-15903 (Erişim Tarihi: 15.3.2012)

[55] Anonim, http://kent.dl.sourceforge.net/sourceforge/weka/weka-3-4-5jre.exe. (Erişim Tarihi: 21.12.2011)

[56] D.W., Kibler, D., Albert, M.K., 1991. Instance-based learning algorithms. Machine Learning, 6, 37-66.

[57] Weka, 2005. Class MultilayerPerceptron. http://weka.sourceforge.net/doc/ weka/classifiers/functions-classifiers/MultilayerPerceptron.html. (Erişim Tarihi:05.03.2012)

[58] Göktepe, A.B., Agar, E., Lav, A.H., 2004. Comparison of Multilayer Perceptron and Adaptive Neuro-Fuzzy System on Backcalculating the Mechanical Properties of Flexible Pavements. ARI The Bulletin of the Istanbul Technical University, 54(3), 65-77.

[59] NeuroDimension, 2008. Multilayer Perceptron. http://www.nd.com/ definitions/mlp.htm (Erisim Tarihi: 13.04.2012)

[60] Haykin, S., "Neural Networks and Learning Machines", PHI Learning Private Limited, 2010.

[61] Özkan, Y., "Veri Madenciliği Yöntemleri", Papatya Yayıncılık, 2008.

[62] Önal E, Tümerdem Y. , Yaslılıkta Hipertansiyon. İstanbul Tıp Fakültesi Halk Sağlığı ABD. Türk Geriatri Dergisi. İstanbul. 4**,** 4,2001.

[63] Özcan N, Tüzün A, Baykal Y. , Yaslılık ve Hipertansiyon. Türkiye Klinikleri Tıp Bilimleri. 15, 207-209.,1995.

[64] Hipertansiyonda Bireyselleştirilmiş Tedavi. Hekimler Yayın Birliği. (2.Baskı). 9-25.1997.

[65] Erdine S., Türkiye Hipertansiyon Haritası. Hipertansiyon ve Ateroskleroz Derneği. Pfizer. İstanbul.,1993.

[66] Pehlivan E, Karaoglu L, Günes G, Genç M, Kurçer MA.,Malatya ili Güzelyurt Kasabası 30 Yas ve üzerindeki Kisilerde Hipertansiyon Prevalansı ve Hipertansiyonu Etkileyen Faktörler. Sağlık ve Toplum. 2, 46-49.2002.

[67] Atasoy Hİ, Tataroglu C, Tutucu KN, Yeniçerioglu Y., Yurt E., Ergazi köyü 40 ve üzeri popülasyonda Hipertansiyon Prevelansına İliskin Tarama Çalışması. Hacettepe Üniversitesi Tıp Fak. Halk Sağlığı Bölümü İntern Araştırması. Ankara,1992.

[68] Hsueh WA , Buchman TA. Obesity and Hypertension. Endocrin Met Clin NorthAm. 23, 405-27.

[69] Mahley RW , Weisgraber KH, Bersot TP. Disorders of Lipid Metabolism.Williams Endocrinology 11th ed. 3693-3836.

[70] Doğan A.S.. , hipertansif kişilerin ailesinde genetik polimorfizm ve lipid profili, Cumhuriyet üniversitesi,Sivas,2008.

[71] NCEP Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults: Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA , 285:2486– 2497,2001.

[72] Johnson,RS , Kang DH, a Feig D, Kivlifhn J, Watanabc S, Tutle KR: Is there a pathogenetic role for uric acid in hypertension and cardiovasculer and renal disease?Hypertansion , 41(6):1183-1190,2003.

[73] İnci H., Tip 2 Diyabet Hastalarında Kan Ürik Asit Düzeyi İle İdrar Albumin Atılım Düzeyi Arasındaki İlişki, 41(6):1183-1190, Cumhuriyet Üniversitesi, Sivas, 2007.

[74] Arıcı M., Altun B, Erdem Y ve Ark. , Afyonkarahisar, Türk Hipertansiyon Prevalans Çalısması, Türk Hipertansiyon ve Böbrek Hastalıkları Dernegi. www.turkhipertansiyon.org/pdf/Turk_Hipertansiyon_Prevalans_Calisması ozeti-1.pdf (Erişim Tarihi:29.05.2012)