T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ İLE KANSER TANISI

Zehra KARAPINAR ŞENTÜRK

ELEKTRİK EĞİTİMİ ANABİLİM DALI

ARALIK 2011
DÜZCE

T.C.
DÜZCE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ İLE KANSER TANISI

Zehra KARAPINAR ŞENTÜRK

ELEKTRİK EĞİTİMİ ANABİLİM DALI

ARALIK 2011
DÜZCE

Zehra KARAPINAR ŞENTÜRK tarafından hazırlanan Veri Madenciliği ile Kanser Tanısı adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Yrd. Doç. Dr. Resul KARA ……
Tez Danışmanı, Elektrik Eğitimi Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile Elektrik Eğitimi Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Yrd. Doç. Dr. Resul KARA ……
Elektrik Eğitimi Anabilim Dalı, Düzce Üniversitesi

Yrd. Doç. Dr. Pakize ERDOĞMUŞ ……..……
Elektrik Eğitimi Anabilim Dalı, Düzce Üniversitesi

Yrd. Doç. Dr. Atilla BÜYÜKGÜÇLÜ …….
Düzce Ü. Düzce Meslek Yüksekokulu Teknik Programlar Bölümü

Tarih: …/…/2011

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

Doç. Dr. Haldun MÜDERRİSOĞLU ……
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Zehra KARAPINAR ŞENTÜRK

ÖNSÖZ

Tez çalışmamın her aşamasında büyük emekleri bulunan değerli hocam Sayın Yrd. Doç. Dr. Resul KARA'ya, sevgili eşime ve manevi desteklerinden ötürü her zaman güç aldığım sevgili aileme sonsuz teşekkürler…

Aralık 2011 **Zehra KARAPINAR ŞENTÜRK**

İÇİNDEKİLER Sayfa

ŞEKİL LİSTESİ

Savfa

SEMBOL LİSTESİ

VTYS : Veri Tabanı Yönetim Sistemi

DBMS : DataBase Management Systems

VM : Veri Madenciliği

OLAP : Online Analytical Processing

OLTP : Online Transactional Processing

PHP : Hypertext PreProcessor

BMI : Body Mass Index

KDD : Knowledge Discovery from Databases

CART : Classification And Regression Tree

ACE : Alternating Condition Expectation

MARS : Multivariate Adaptive Regression Spline

CLARANS : Clustering Large Applications based upon RANdomized Search

PAM : Partitioning Around Medoid

CLARA : Clustering Large Applications

MST : Minimum Spanning Tree

MAFIA : Merging of Adaptive Finite Intervals

DBSCAN : Density-Based Spatial Clustering of Applications with Noise

ROCK : Robust Clustering using linKs

YSA : Yapay Sinir Ağları

SQL : Structured Query Language

ID3 : Iterative Dichotomiser 3

CHAID : Chi Square Automatic Interaction Detection

VERİ MADENCİLİĞİ İLE KANSER TANISI

(Yüksek Lisans Tezi)

Zehra KARAPINAR ŞENTÜRK

DÜZCE ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Aralık 2011

ÖZ

Veri kavramı tüm dünyada her zaman oldukça önemli bir mesele olmuştur. Bilimsel ya da günlük yaşamın her diliminde, sık sık, faydalı bilgi elde etmek üzere veriler kullanılmıştır. Önceleri eldeki veriyi gözeterek bir analiz yapmak oldukça kolay, bazı giriş verileri mevcutken bir sonucu (çıktıyı) tahmin etmek daha basitti. Fakat şimdilerde, elde edilen veriye bakarak bir sonuca varmak zorlaşmıştır. Bu da demek oluyor ki, son zamanlarda, veriyi bilgiye çevirmek biraz daha karmaşık bir hal almıştır. Sürekli büyüyen dünya veri yığınlarını da büyütmüş ve onu kullanmayı, depolamayı ve yönetmeyi de aynı oranda zorlaştırmıştır. Günümüzde, ucuzlayan teknoloji sayesinde veriler kolaylıkla depolanabilmekte ve pek çok veritabanı yönetim sistemi ile de mümkün olduğunca yönetilebilmektedir. Bu gelişmelerden sonraki problem ise bunca verinin nasıl analiz edileceği ve böylesine kütlelerdeki veri yığınları arasından nasıl bir sonuca varılacağı konusudur. Veri madenciliği kavramının doğuşunu işte bu problem tetiklemiştir.

Bu çalışmada, sağlık sektörüne veri madenciliği ile katkıda bulunmak amaçlanmıştır. Çalışmada hastalara kanser tanısı koymaya yönelik bir analiz yapılmıştır. Bu amaçla, öncelikle, daha önceden kanser teşhisi konmuş hastaların verileri toplanmış ve düzenlenmiş, daha sonra da bu verilerden yararlanılarak

başka hastaların hangi kansere yakalandıkları tahmin edilmeye çalışılmıştır. Verilerin düzenlenmesi ve madenciliğe hazırlanması aşamasında MySQL VTYS ve PHP programlama dili kullanılmıştır. Veriler hazırlandıktan sonra istenen algoritmaya göre veri madenciliğini gerçekleştirmek üzere RapidMiner 5.0 aracı kullanılmıştır.

Bilim Kodu :
Anahtar Kelimeler : Veri, veri madenciliği, kanser tanısı
Sayfa Adeti : 58
Tez Yöneticisi :Yrd. Doç. Dr. Resul KARA

CANCER DIAGNOSES VIA DATA MINING

(M.Sc. Thesis)

Zehra KARAPINAR ŞENTÜRK

DUZCE UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

December 2011

ABSTRACT

The concept "data" was always an important issue around the world. In every slice of scientific or daily life, data is frequently used to obtain useful information. Before, it was quite easy to make an analysis considering the data on hand, i.e. it was simpler to predict a result when some input data is known, but nowadays, it is not so much easy to come to a point after looking the data obtained. This means nowadays the conversion from data to information is some more complicated. Growing world makes the stack of data greater, and makes it much more difficult to store, manage, and use. Thanks to the cheapening technology, data is now stored easily and with several DBMSs data is managed as much as possible. The two disadvantages of the growing world and growing data stack is now eliminated via those improvements. The problem after those improvements is the topic how to analyze that much of data or how to conclude among a huge amount of data. Here is the concept that triggers the rise of data mining.

In this study, it is intended to contribute to the health sector with data mining. There is an analysis on cancer diagnoses for the patients in the study. For this

purpose, first of all, data about the patients whose cancers' have been diagnosed before and they are arranged, and then which cancer the other patients suffer from is tried to be predicted under cover of those data. MySQL Database Management System and PHP web programming language are used in the arrangement and preparation of the data on hand to data mining. After the data is prepared, RapidMiner 5.0 tool is used to apply data mining with the desired algorithm.

Science Code :
Key Words : Data, data mining, cancer diagnoses
Page Number: 58
Adviser : Yrd. Doç. Dr. Resul KARA

1. GİRİŞ

Geçmişten günümüze doğru gelindikçe, sahip olunan verilerde oldukça hızlı bir artış gözlenir. Hızla artan verileri kontrol edebilmek ve yönetebilmek ise aynı oranda zorlaşmaktadır. Önceleri kâğıt kalemle tutulan hesaplar verileri depolamaya yetmeyince ve de aranan veriyi bulmak giderek zorlaşınca, daha büyük boyutlarda ve kolay yönetilebilir sistemlere ihtiyaç duyuldu. Bilgisayar kullanımının yaygınlaşmasıyla, hızla artan hacimdeki veriler bilgisayar belleklerinde tutulmaya başlandı. İlk bakışta salt bilgisayar hafızalarının kullanımı çözümmüş gibi görünse de belleklerde yüksek boyutlarda yer kaplayan verilere ulaşma ve belli verilerde değişiklik yapma gibi işlemlerin hayli zorlaşması veritabanı yönetim sistemleri fikrini doğurmuştur. Aynı zamanda, depolanan veriler üzerinde kolayca işlemler yapabilme imkânı bu sistemler sayesinde sağlanmıştır. Veritabanı yönetim sistemleri ile normalde yapılması çok fazla zaman alabilecek olan işlemler kısa sürede ve hata oranı da minimize edilerek yapılmaya başlanmıştır. Ancak, ihtiyaçlar, verilerden bilgi elde etmeyi gerektirince, mevcut veritabanı yönetim sistemleri yetmemeye başladı. Verilerden, daha çok faydalı bilgi edinme ihtiyacı hemen her alanda hissedildi ve bu ihtiyaca yönelik yöntemler geliştirildi. Geliştirilen yöntemlerde eldeki veriler üzerinde analiz yapılarak geleceğe dair gerçekçi tahminlerde bulunuldu. Verilerden bilgi elde etme sürecine de "veri madenciliği" denildi.

Veri madenciliği, veriyi farklı açılardan analiz etme ve onu kullanışlı bilgi elde etmek üzere özetleme süreci olarak tanımlanabilir. Buradaki bilgi, geliri artırma ya da harcamaları azaltma gibi çeşitli amaçlarla kullanılabilir. Teknik açıdan bakıldığında veri madenciliği, çok büyük ilişkisel veritabanlarında, düzinelerce alan arasından belli bir model ya da bağlantılar bulma sürecidir.

Bu tez çalışmasının amacı, veri madenciliği ile kanser hastalığının tanısında kullanılmak üzere bir analiz yapmak ve hazırlanan ara yüzlerle, hastaların sahip oldukları olası kanser çeşitlerini kolayca tahmin etmektir. Bu sayede, kanser hastalıkları için hayati

önem taşıyan zaman kaybı problemi ortadan kalkacak ve kazanılan zaman tedavi için kullanılabilecektir.

Tezin birinci bölümünde veri madenciliği ile ilgili kavramlar hakkında genel bilgiler verilmiş, amaç ve kapsamına değinilmiştir. İlerleyen bölümlerde veri madenciliği ile kanser tanısı hakkında geniş bilgiler verilerek literatürde yer alan örnek çalışmalara yer verilmiştir. Veri madenciliği konusu detaylı olarak incelenip, veri madenciliğine neden ihtiyaç duyulduğu, hangi problemlerle karşılaşıldığı ve veri madenciliğinde hangi yöntemlerin kullanıldığı açıklanmıştır. Son kısımda ise yapılan çalışmanın bulguları sunulup doğruluğu gösterilmiştir.

2. GENEL KISIMLAR

2.1. VERİ TABANI YÖNETİM SİSTEMLERİ

Veritabanı, düzenli şekilde bilgisayar ortamında tutulan veriler topluluğudur. Günümüzde teknolojinin gelişip ucuzlaması sebebiyle, her çeşit verinin kaydedilmesi veri yığınları oluşturmuştur. Böylelikle, aynen üç boyutlu resimlere ilk bakışta hiç bir şey anlaşılmadığı gibi veri yığınları da karmaşık bilgi çöplüklerine dönüşmüştür. Böyle yüksek boyutlardaki veriler arasından istenen veriye manüel olarak ulaşmak kimi durumlarda imkânsız olabilmektedir. Hatta günümüzde, hemen hemen her yerde bu olanaksızdır. Bu problemi çözmek için ilk olarak 1980'li yıllarda ortaya atılan "Veri Tabanı Yönetim Sistemi" kavramı, veritabanı sistemlerini tanımlamak, oluşturmak, kullanmak, değiştirmek ve veri tabanı sistemleriyle ilgili her türlü işletimsel gereksinimi karşılamak için kullanılan geniş kapsamlı yazılım sistemidir (Mesut, 2011).

Veritabanı yönetim sistemlerinin tercih edilmesinin bir sebebi, veri girişi ve depolanmasının veriye erişen uygulama programlarından bağımsız olmasıdır. Klasik dosyalamada ise, örneğin, dosya yapılarında ortaya çıkabilecek en ufak bir değişiklik bile uygulama programlarının değiştirilmesine sebep olmaktadır. VTYS'nin diğer bir üstünlüğü, gereksiz veri tekrarını önlemesidir. Bir başka üstünlüğü ise, veri doğruluğu ve tutarlılığını sağlamaktır. Örneğin, sipariş bilgileri girilirken, teslim tarihinin siparişin verildiği günden önce olması durumunda hatalı veri girişi sebebi ile talebin yerine getirilmemesi istenebilir.

Ayrıca, VTYS, her kullanıcıya çeşitli yetkiler vererek verilerin güvenliğini sağlar. Böylelikle istenmeyen kullanıcılar istenmeyen verilere erişemezler.

Bu faydaları ve daha başka üstünlükleri sebebi ile VTYS kavramı hemen her yerde kullanılmaya başlanmış ve kısa geçmişine rağmen işletim sistemlerinden sonra en popüler ve en çok gelir getiren yazılımlar VTYS'ler olmuşlardır.

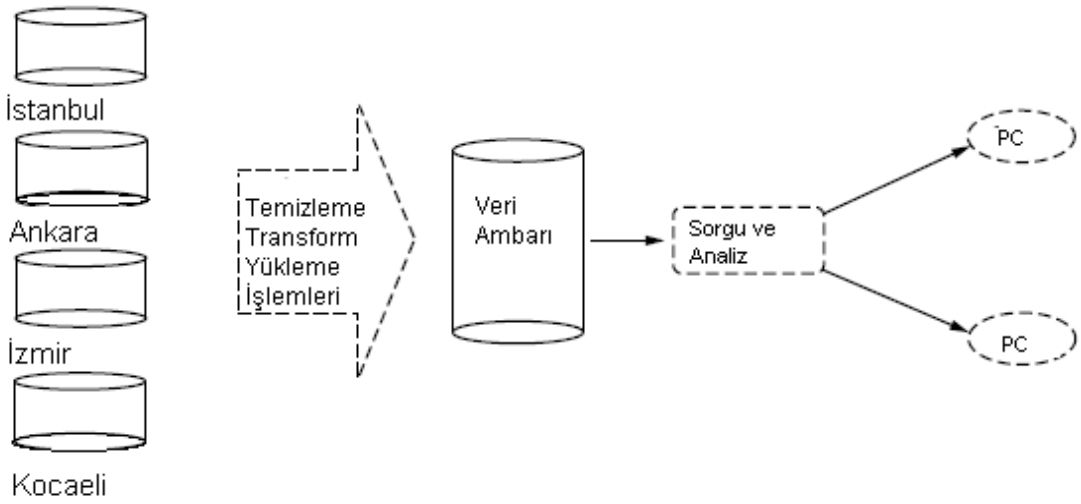
VTYS'ler veri madenciliği sürecine, ilgili verilerin seçilmesi, arındırılması, vs. noktasında katkıda bulunurlar. Yani, veriyi veri madenciliğine hazırlayarak, VM sürecini başlatırlar.

2. 2. VERİ AMBARLARI

Veri ambarları, tüm operasyonel işlemlerin, en alt düzeyindeki verilerine kadar inebilen etkili analiz yapılabilmesi için özel olarak modellenen ve tarihsel derinliği olan veri depolama sistematiğidir (Yaralıoğlu, 2011). Burada operasyonel işlemlerden kasıt, operasyonel veriler üzerinden yapılan işlemlerdir. Operasyonel veri, uygulamaya yönelik, düzensiz, kısa zamanda oluşup tekrarlayabilen veri demektir.

Veri ambarı, bir ürün değil ortamdır. Aynı zamanda, çoklu veri tabanı veya diğer bilgi sistemlerinden ilgili veriyi elde etmek için gerekli olan algoritmaları, araçları içeren mimari topluluğudur. İlgili veriler, veri sorgulama ve raporlama amaçlı elde edilir.

Bir veri ambarı ilgili veriyi kolay, hızlı ve doğru biçimde analiz etmek için gerekli işlemleri yerine getirir, işletimsel sistemlerdeki veriyi kopyalayıp, karar verme işlemleri için uygun formda saklar.



Şekil 2.1: Farklı şehirlerde veritabanı olan bir şirket için örnek veri ambarı gösterimi (Gürünlü, 2009).

Veri ambarlarının dezavantajlarından bazıları şunlardır;

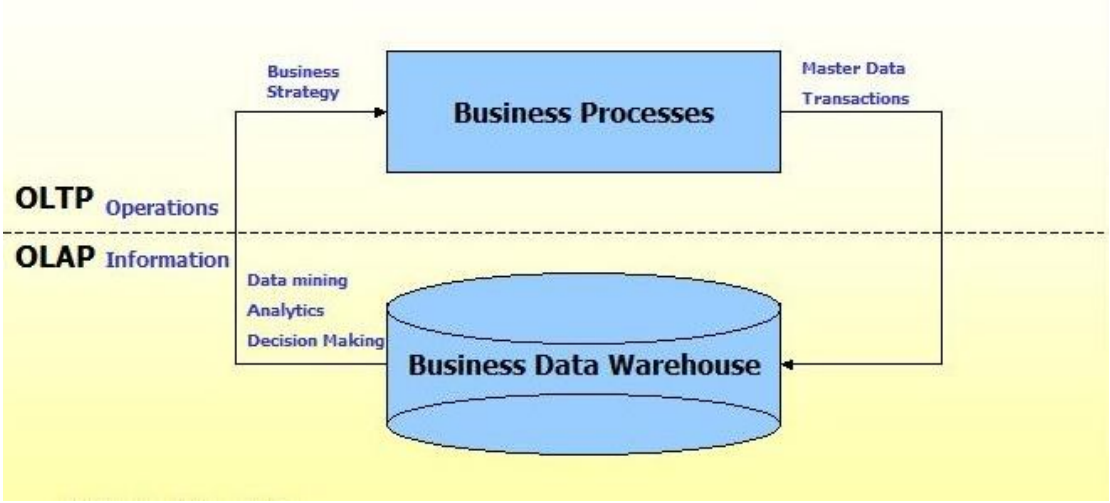
- Veri ambarları, güncelleme yapılma oranına bağlı olarak gereksiz bilgi içerebilir.
- Eksik veri içerebilir ve tüm sorulara cevap veremeyebilir.

Veri ambarlarının avantajları;

- Sorgular doğrudan veri ambarları üzerinden cevaplandığından, yüksek sorgulama performansına sahiptirler.
- Konular iyi ayrılmıştır.
- Tarihsel veriler madenlenebilir.
- Tarihsel veriler sorgulanabilir / analiz edilebilir (OLAP).
- Yerel süreçlerin bir aksama / duraklama zamanı olduğu varsayılır ve veri ambarı güncellemesi bu muhtemel duraksama zamanı boyunca yapılır. Karmaşık sorgular veri ambarlarında, OLTP sorguları kaynak sistemlerde çalışır. Böylelikle, veri ambarlarındaki işlemler, kaynaktaki yerel işlemlerle karışmaz.
- OLTP, veri ambarlarından bağımsız olarak bilgi kaynaklarından yapılır.
- Veri, veri ambarlarına kopyalanır. Böylece, değiştirilebilir, notlar alınabilir, özetlenebilir, silinebilir, vs. ve orijinal veri güvendedir.

2.3. OLAP VE OLTP

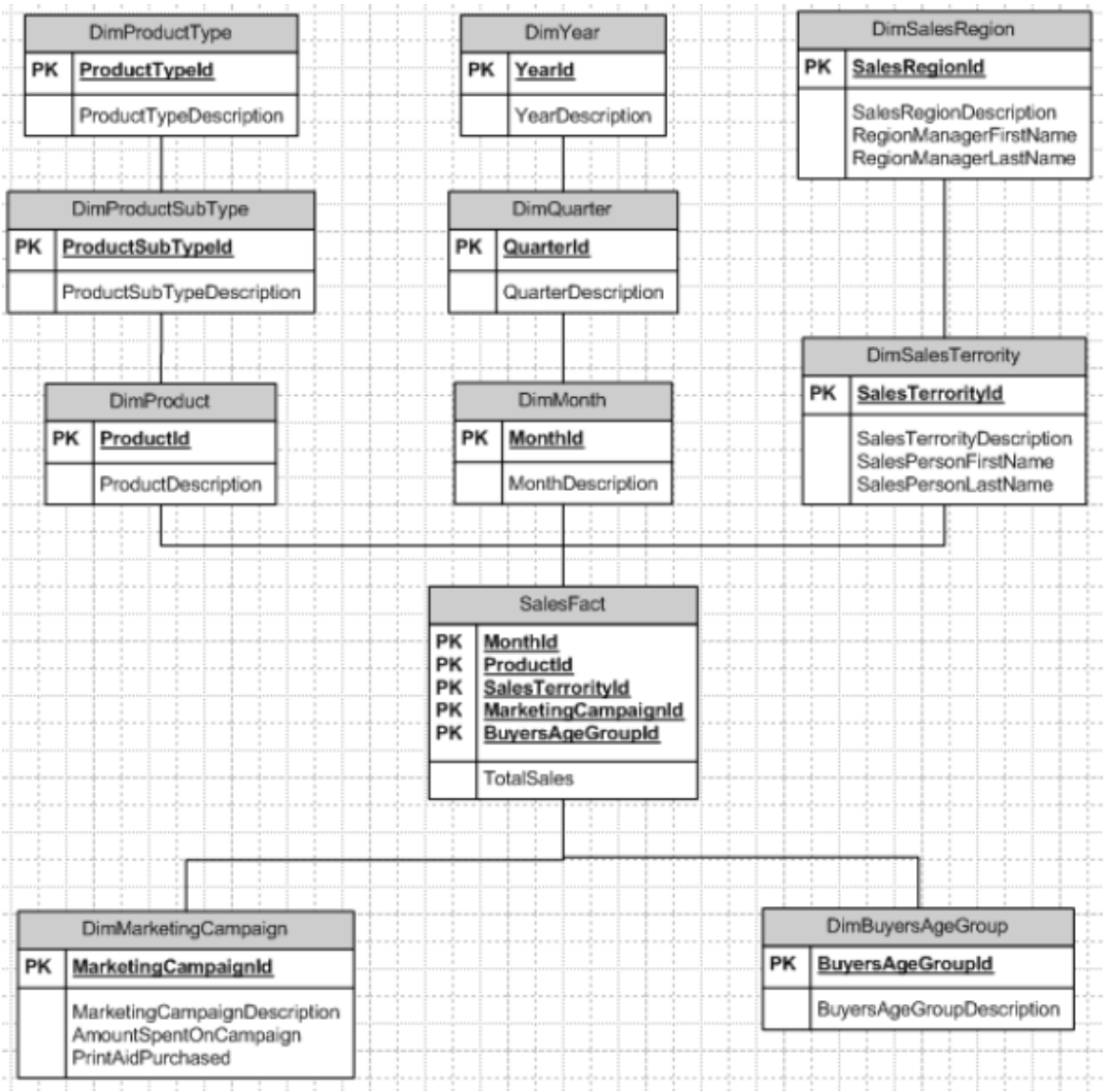
Bilgi teknolojilerini, OLTP (işlemsel) ve OLAP (çözümsel) olarak ikiye ayırabiliriz. Genellikle, OLTP sistemlerinin veri ambarlarına kaynak sağlarken, OLAP sistemlerininse bunu analiz etmeye yardımcı olduğu varsayılır.



Şekil 2.2: OLAP veOLTP (Datawarehouse4u.info, 2011).

OLTP, çok sayıda çevrim-içi işlemden (transaction) (INSERT, UPDATE, DELETE, v.s.) oluşur. OLTP sistemlerindeki temel vurgulanması gereken noktalar, sorgu sürecinin çok hızlı olması, çoklu erişim ortamındaki veri bütünlüğünün sürdürülmesi ve ölçülen, saniyedeki işlem sayısındaki verimliliktir. OLTP veritabanında detaylı ve güncel veri vardır.

OLAP ise, kısmen daha az hacimdeki işlemlerden oluşur. Sorgular çoğu zaman oldukça karmaşıktır ve içinde toplamalar vardır. OLAP sistemleri için verimlilik ölçüsü tepki süresidir (response time). OLAP uygulamaları veri madenciliği tekniklerinde yaygın olarak kullanılır. OLAP veri tabanında, çok boyutlu şemalarda (genellikle yıldız şeması) tutulan, toplanmış tarihsel veriler vardır.



Şekil 2.3: Örnek Yıldız Şema (Sümerkent, 2010).

3. MATERYAL VE YÖNTEM

Bu tez çalışmasında, veri madenciliği yöntemi sağlık sektörüne uygulanmıştır. Hastane veritabanlarında tutulan teşhisi konmuş kanserli hasta bilgileri kullanılarak, henüz teşhisi konmamış fakat diğer verileri (laboratuar sonuçları) mevcut olan yeni hastalar için muhtemel kanser tanıları tahmin edilmiştir. Tahmin etme işleminde çeşitli algoritmalar kullanılmış ve güvenilirliği en yüksek olanı tercih edilmiştir.

Verilerin veri madenciliğine hazırlanması sürecinde PHP programlama dili ile MySQL veri tabanı yönetim sisteminden yararlanılmıştır. Düzce Üniversitesi Eğitim ve Araştırma Hastanesi'nden kanser hastaları ile ilgili gerçek veriler alınmıştır. Ancak, hastaların ad, soyad, cinsiyet, yaş, iletişim bilgileri, vs. kişisel bilgileri alınmamış ve dosya numaraları da değiştirilerek kullanılmıştır. Excel dosyası olarak alınan kanserli hastalara ait tüm verilerdeki hatalar düzeltilmiş ve dosyalar veritabanında ayrı ayrı oluşturulan ilgili tablolara aktarılmıştır. Daha sonra çeşitli veritabanı operasyonları ile istenen tablolar birleştirilmiş ve ilgili noktalar görüntülenmiştir.

Temizlenen veriler daha sonra RapidMiner veri madenciliği aracına aktarılıp teşhisi konmayan hastaların da kanser türleri tahmin edilmiştir. Aracın doğru çalıştığından emin olmak için teşhisi konmuş hastaların verileri kullanılmış ve aynı sonuçların araç tarafından da alındığı görülmüştür.

3.1. VERİ MADENCİLİĞİ İLE KANSER TANISI

Veriler dijital ortamda saklanmaya başladığından beri, yeryüzündeki bilgi miktarının her yirmi ayda bir, iki katına çıktığı günümüzde, veritabanının sayısı da benzer hatta daha yüksek oranda artmaktadır (Vahaplar ve İnceoğlu, 2001). Giderek artan veri yığınları çözümlemeyi gerektirmektedir. Bunun çözümü ise son zamanların en gözde veri analiz teknolojisi olan veri madenciliğidir. Bilgisayar teknolojisindeki gelişmeler veri

madenciliği yöntemleri ve programları, büyük miktardaki verileri etkin ve verimli bir hale getirmektedir. Bilgi ve tecrübeyi geliştirmek için veri madenciliği konusunda geliştirilmiş yazılımların kullanılması gerekmektedir.

Veri madenciliği, geleceğe yönelik tahminlerde kullanabilecek saklı bilgilerin, bir çok veriyi barındıran veritabanlarından, değişik tekniklerle elde edilmesidir. Veri madenciliği teknikleriyle gelecekte oluşabilecek davranış ve olaylar konusunda tahmin yürütülüp, beklenen süreçler ortaya çıkmadan kararlar alınarak süreçler yönlendirilebilmektir (Işık, 2008).

Bilgi keşfine giden yol olarak ifade edilen veri madenciliğinde, değişik kaynaklardan veri toplanır, ön işlemeye tabi tutularak hatalı veriler ayıklanır, eksik yada kayıp veriler tamamlanır, veriler ortak bir formata dönüştürülür. Gerekli sonuçların elde edilmesi için uygun algoritmalar uygulanır ve sonuçta anlaşılır bir şekilde (grafik, tablo, vs.) sunulur. Yapılan işlemlerin ortak hedefi, eldeki verilerin incelenerek gerçeğe en yakın modele oturtulmasını sağlamaktır. Bu modeller, tahmin edici veya tanımlayıcı olabilir.

Veri madenciliği, çok değişik alanlara uygulanmış ve halen uygulanmaktadır. Eğitim–Kültür projeleri, ticari uygulamalar, kartlı alışverişler, hava ve afet tahmini gibi alanlarda veri madenciliği yöntemleri kullanılmıştır. Ayrıca, özellikle son zamanlarda veri madenciliği tıp alanında da sıklıkla kullanılan yöntemlerden biri haline gelmiştir. Bunun sebebi, tıbbi verilerin sürekli artıyor olmasının yanında, bu verilerin çözümlenmesinin hayati önem teşkil etmesidir.

Hayati önem taşıyan bu çözümleme işleminde kullanılacak verilere, uygulamada hastane bilgi sistemleri aracılığı ile ulaşılmaktadır. Hastane bilgi sistemlerinde, hasta demografik bilgilerine, hastalık ve tedavi durumlarına, yapılan tetkiklere, faturalama ve idari işlere ait verilere ulaşmak mümkündür.

Sağlık alanında veri madenciliği, belli hastalıklara sahip hastaların ortak özelliklerinin tahmin edilmesi, tıbbi tedavilerin sonuçlarının tahmin edilmesi, hastane maliyetinin tahmin edilmesi, ölüm oranları ve salgın hastalıkların tahmin edilmesi gibi alanlarda kullanılabilir.

Tıp alanında veri madenciliğine örnek diğer uygulamaları, antipsikotik ilaçların kalp kası hastalıklarının üzerine etkisi (Coulter ve diğ, 2001), solunum fonksiyon testlerinin analizi (Ganzert ve diğ, 2002), genetik bozuklukların tespiti (Ponomarenko ve diğ, 2001), ilaç yan etkilerinin tanınması (Honigman ve diğ, 2001) şeklinde sayabiliriz.

Bir başka örnek, akciğerdeki tümörün iyi huylu olup olmadığına dair karar destek amaçlı bir çalışmadır. İstatistiklere göre, Amerika'da yılda 160.000'den fazla akciğer kanseri vakasının olduğu ve bunların % 90'ının öldüğü belirlenmiştir. Bu noktada, bu tümörün erken ve doğru teşhisi önem kazanmaktadır. Testlerde elde edilen veriler sayesinde % 40-60 oranında doğru teşhis konulabilmektedir (Işık, 2008). İnsanlar kanser olup olmadıklarını öğrenebilmek için biyopsi yaptırmayı tercih etmektedirler. Bu gibi testlerin ise hem maliyeti yüksektir, hem de uygulanması bazı riskleri beraberinde getirir. Farklı yerlerde ve farklı zamanlarda kliniklerde toplanan bu türden test verileri arasında yapılan veri madenciliği çalışmaları teşhiste % 100 oranında doğruluk sağlamıştır (Işık, 2008).

Diğer bir çalışma, Kore Tıbbi Sigorta Kurumu tarafından hazırlanan bir veri tabanı üzerinde yapılan, yüksek tansiyon ile ilgili bir çalışmadır. Bu çalışma 1998 yılına ait 127.886 kayıt üzerinde yapılmıştır. İlk aşamada yüksek tansiyona sahip 9.103 kayıt, daha sonra aynı sayıda yüksek tansiyon olmayan kayıtlar üzerinde çalışılmıştır. Bu model 13.689 kayıttan oluşan öğrenme ve 4.588 kayıttan oluşan test setine bölünerek modelin eğitimi yapılmıştır. Bu çalışmalar sonucunda yüksek tansiyon tahmininde dikkate alınan değerler, vucüt kütle indeksi (BMI), idrar proteini, kan glukozu ve kolesterol değerleridir. Hayat şartlarının (diyet, tuz kullanımı, sigara vb.) hiç birinin bu tahminde rolü olmadığı, ancak yaşın önem arzettiği gösterilmiştir (Işık, 2008).

Bu tez çalışmasında veri madenciliği teknolojisi kanser hastalığının tanısında kullanılmıştır. Bunun için öncelikle hastane bilgi sisteminde daha önceden kayıtlı olan hastaların bilgileri excel tablolar halinde alınmıştır. Alınan veriler daha sonra veri madenciliğine hazırlanmak üzere ön işlemlerden geçirilmiştir. Excel tabloları düzeltildikten sonra veri tabanında oluşturulan ilgili tablolara aktarılıp, çeşitli SQL sorgularıyla istenen tablolarda değişiklik ya da birleştirmeler yapılıp elde edilen son (düzeltilmiş) veriler RapidMiner'a aktarılmıştır. RapidMiner'da teşhisi konmuş ve

teşhisi henüz konmamış fakat laboratuvar sonuçları mevcut olan hastaların verileri çekilip, çeşitli algorimalarla modellemesi yapılarak çalıştırılmış ve doğrulukları test edilmiştir.

Daha önceden teşhisi konmuş hastaların verileri, üzerinde çalışalacak veri seti (training data) olarak ve henüz tanısı konmamış hastaların verileri ise tahmin veri seti (prediction data) olarak belirlenerek modellenmiştir.

3.2. VERİ MADENCİLİĞİ

Veri madenciliği, veriyi, faydalı bilgi elde etmek üzere masaya yatırmaktır. Veri madenciliği çok büyük miktarlardaki veriler arasındaki ilişkileri gün ışığına çıkarır. Veri miktarı arttıkça onu bilgiye dönüştürmek zorlaşır. Hatta bazen terabaytlarca veri arasında bir ilişki kurmak imkânsız olabilir. Ancak, veri madenciliği bu problemi çözmeyi büyük ölçüde başarır.

Veri madenciliği pek çok bilim insanı tarafından tanımlanmıştır. Bunlardan bazıları şunlardır:

Fayyad (1996)'a göre veri madenciliği, veritabanlarından bilgi keşfi süresince (KDD), girdilerin ağırlıklı olarak temizlendiği, verilere dönüştürüldüğü, algoritmalar kullanılarak veriler üzerinde aramalar yapıldığı ve çıktı model ve ilişkilerinin değerlendirildiği bir basamaktır.

David Hand ve diğ. (2001) veri madenciliğini, büyük veri kümeleri ya da veritabanlarından kullanışlı bilgilerin çıkarıldığı bilim olarak tanımlamışlardır.

Berry ve Linoff(1997)'a göre veri madenciliği, büyük miktarlardaki verilerin, içindeki anlamlı model ve kuralların çıkarılması için, otomatik ve yarı otomatik araçlarla analiz ve keşfidir.

Swift (2001)'e göre veri madenciliği, veri ambarlarında tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş verileri ortaya çıkarmak, bunları karar vermek ve gerçekleştirmek için kullanma sürecidir.

Belirtildiği üzere veri madenciliğinden veriyi analiz etmek ve veri kümesi içindeki gizli modelleri keşfetmek için yararlanılır. Daha sonraki adımda, keşfedilen model, veriyi daha detaylı yorumlamak ve ileriye dönük tahminler yapmak için kullanılır. Yani asıl amaç, veriyi bilgiye dönüştürmektir.

Veri madenciliği, veri kümelerine karar ağaçları (decision trees), gruplama (clustering), ilişkilendirme (association), zaman serileri (time series) gibi algoritmalar uygular ve içeriklerini analiz eder. Bu analizler, değerli bilginin keşfi için modeller üretir. Kullanılan algoritmaya bağlı olarak üretilen model, ağaçlar, kurallar, gruplar ya da basit bir matematik formülü olabilir. Model içerisinde bulunan veri, satış stratejisi oluşturmaya rehberlik etmesi ve en önemlisi tahmin için raporlamada kullanılabilir.

Görüldüğü gibi yapılan tanımlamalar hemen hemen eş anlamlı ve veri madenciliğinin istatistiksel bir süreç olduğuna işaret etmektedir.

3.3. VERİ MADENCİLİĞİ İHTİYACI

Teknolojinin günümüzde giderek ucuzlamasıyla birlikle, özellikle büyük şirketler, disk boyutlarını da büyüttüler. Bu sayede kayıt altındaki veri miktarı da büyüdü. Şirketler, eldeki bu kadar veriyi, iş stratejilerine ışık tutması açısından kullanmak isterler.

İnternetin, hayatın hemen her alanında kullanılır duruma gelmesi yeni bir satış yöntemi olan internet üzerinden satışı da beraberinde getirdi. Diğer yöntemlerden daha kolay olan bu metot, hem müşteri portföyünü genişletti, hem de ulusal ve uluslar arası rekabeti arttırdı. Şirketlerin bu rekabetle başa çıkmaları, var olan müşterilerini elde tutmaları ve onlara yenilerini eklemelerine bağlıdır. Burada devreye veri madenciliği girer ve eldeki verileri analiz ederek şirketlere, hedeflerine ilerlemeleri hususunda ışık tutar.

Bunlar gibi ihtiyaçlar, günümüzde veri madenciliği uygulamalarını, akademik dokümanlardan çıkarıp endüstride de kullanılır hale getirmeyi zorunlu kıldı (Uslu,2008).

3.4. VERİ MADENCİLİĞİNDE KARŞILAŞILAN PROBLEMLER

Veri madenciliğinin besin kaynağı olan OLTP sistemler başta olmak üzere birçok veritabanında bulunan sıkıntılardan ötürü kimi durumda, veri madenciliğinde problemlerle karşılaşılmaktadır (Kocabaş,2010). Bir başka sorun da eldeki verinin konu ile uyumsuzluğundan doğabilir.

Veri madenciliğindeki problemlerden genel hatları ile aşağıda bahsedilecektir.

3.4.1. Veritabanı Boyutu

Geliştirilen pek çok makine öğrenimi algoritması birkaç yüz kayıtlık örneklem üzerine çalışabilecek şekilde tasarlanmıştır. Günümüzdeki veritabanlarını düşündüğümüzde bu miktarın eldekinden çok çok küçük olduğu görülür. O kadar bir veri yığınından nispeten çok küçük olan bir örnek uzay oluşturmak ise iyi bir istatistiksel araştırma için şarttır.

3.4.2. Gürültülü Veri

Veri madenciliğindeki bir başka problem de gürültülü verilerdir. Veritabanındaki pek çok niteliğin değeri yanlış olabilir. Bu problem, kullanıcıların yanlış veri girmesi ya da girilen değerin yanlış ölçülmesinden kaynaklanıyor olabilir. Her ne kadar günümüzdeki gelişmiş VTYS'ler böyle yanlışlıkları en aza indirme yeteneğine sahip olsalar da, bu tür problemler hala yaşanmaktadır.

3.4.3. Null Veri

Bir veritabanındaki boş değerlerdir. Boş veri, hiçbir değeri olmayan veri demektir. Bir niteliğin değerinin boş olması ile boşluk olması aynı şey değildir. Bir hücre hiç bir değer içermiyorsa onun değeri NULL' dır.

3.5. VERİ MADENCİLİĞİ YÖNTEMLERİ

Çalışmanın bu bölümünde, günümüzde diğer algoritmalara oranla çok daha yaygın olarak kullanılan veri madenciliği algoritmaları tanımlanmıştır.

Günümüzde hem yazılım seçenekleri hem de teorik bağlamda kullanılmak üzere çok çeşitli veri madenciliği yöntemleri mevcuttur. Bu yöntemler kullanıcılara bireyler ve firmalar tarafından toplanan verilerden, sağlanan değişik araçlarla, kullanışlı bilgiyi

çıkarma şansını verirler. Büyük miktarlardaki veri, bir ya da birden fazla konudaki çeşitli faktörlerin kararında kullanılabilir. Bu veri madenciliği yöntemlerinden çoğunlukla dolandırıcılıktan koruma, pazarlama ve denetim/gözetim gibi alanlarda yararlanılır.

Konulardan belli bilgileri çekip çıkarmada veri madenciliği yüzyıllardır kullanılmaktadır. Buna rağmen, günümüzün modern teknikleri var olan verileri, bilgisayar kaynakları (otomasyon) aracılığı ile otomatik bir kapsamda kullanmaktadır. 20. yüzyılda bilgisayar bilimlerinin gün ışığına çıkmasıyla beraber, toplanan çok büyük miktardaki veriler içindeki gizli modelleri keşfetmek gayesiyle veri madenciliği yöntemleri fikri geliştirildi. Bir çevrim-içi müşterinin alışveriş modelinin çıkarılarak daha sonraki satışlarda kişinin alması kuvvetle muhtemel olan ürünlerin firma tarafından öncelikli olarak pazarlanması bu gelişmeye iyi bir örnektir.

Veri madenciliği yöntemleri temelde dört görevi yerine getirir. Bunlar, sınıflandırma, kümeleme, regresyon ve birlikteliktir (kural). Sınıflandırma, var olan bilgiyi alır ve onu daha önceden tanımlanmış gruplara katar. Kümeleme, daha önce tanımlanmış gruplaşmayı kaldırır ve verinin kendisini benzerlik özelliklerine göre sınıflandırmasını sağlar. Regresyon, bilginin fonksiyonuna ve verinin modellenmesine odaklanır. Son olarak birliktelik ise değişik veri öğeleri arasında ilişkiler bulmaya dayanır.

3.5.1. İstatistiksel Yöntemler

Veri Madenciliği'nin yoğun olarak kullanılmasının temel sebebi, çok fazla miktarda verinin istenildiği gibi kullanılabilmesidir. Veri madenciliğinde kullanılan yöntemler esasında uzun yıllardan beridir istatistikte kullanılmaktadır. Veri madenciliğinin getirdiği avantaj bilişim teknolojisini kullanarak yapılan analizlerin çok kısa sürede ve daha az maliyetle yapılabilmesidir (Özmen, 2001).

İstatistik de veri madenciliği gibi verinin yapısını keşfetmeyi amaçladığından veri madenciliği, istatistiğin alt dalıymış gibi düşünülmektedir. Ancak, elbette veri madenciliği ve istatistiği ayıran özellikler mevcuttur.

Öncelikle, CART, sinir ağları ve en yakın komşu gibi klasik veri madenciliği teknikleri, hem karmaşık gerçek dünya verileri hem de az uzman kişilerin kullanabilmeleri için

güçlendirilmişlerdir. İstatistik, verilerin toplanması ve tanımlanması ile ilgili bir matematik dalıdır. Genelde fizik ve kimya da korkulan konulardan biridir; fakat aslında her gün kullanmamız sebebiyle matematiğin daha arkadaşçıl konularından biridir. İstatistiğin gerçek dünya problemlerine girişi kumardan, biyolojiye ve iş dünyasına kadar geniş bir yelpazede olmuştur.

Peki, istatistik "tahmin"de nasıl kullanılıyor? Tahmin terimi başka yerlerde daha çok regresyon olarak anılan değişik türlerdeki analizler için kullanılır. Ayrıca, regresyon istatistikte yaygın olarak kullanılan güçlü bir araçtır.

İstatistiksel yöntemler, verinin bir örnek kümesine bir kestirici oturtmayı amaçlamıştır. İstatistik literatüründe son elli yılda bu amaç için değişik yöntemler önerilmiştir. Bu teknikler genellikle çok boyutlu analiz başlığı altında toplanır ve genelde verinin parametrik bir modelden geldiğini varsayar. Bu varsayım altında sınıflandırma (classification, discriminant analysis), regresyon, öbekleme (clustering), boyut azaltma (dimensionality reduction), hipotez testi, varyans analizi, bağıntı (association, dependency) kurma için teknikler istatistikte uzun yıllardır kullanılmaktadır (Rencher, 1995). Bu teknikler içerisinden veri madenciliğinde yaygın olarak kullanılan bazı yöntemler ilerleyen bölümlerde anlatılmıştır.

3.5.1.1. Sınıflandırma

Veri madenciliğinde kullanılan sınıflandırma teknikleri ağırlıklı olarak makine öğrenimi disiplinine ait tekniklerdir. Özellikle veri madenciliğinin bir uygulama dalı olan sepet analizinde sınıflandırma kurallarının farklı bir türü olan birliktelik kuralı kullanılmaktadır(Özdamar, 2002). Kullanılan teknikler ise hız, anlaşılabilirlik ve öğrenme zamanı gibi kriterlere göre değerlendirilmektedir.

İstatistikteki sınıflandırma teknikleri modern ve klasik teknikler olarak ikiye ayrılabilir. Klasik teknikler, lineer diskriminant kuralına dayanan ve 1936'da Fisher tarafından ortaya atılan tekniklerdir. Fisher'in bu kuralı, sınıfları, en küçük kareler veya maksimum olabilirliği kullanarak, aralarındaki mesafe maksimum olarak ayrılır. Daha sonra değişik diskriminant fonksiyonları da geliştirilmiştir. Bir diğer klasik sınıflandırma tekniği de Bayes Ağları'dır.

İstatistikte sınıflandırmada kullanılan modern tekniklerin en eskisi Fix ve Hodges tarafından geliştirilen, parametrik olmayan bir teknik olan yoğunluk tahminidir. Diğer modern teknikler ise k-en yakın komşu, Projection Sınıflandırma, Naive Bayes ve rastlantısal ağlardır. Ayrıca, ACE (The Alternating Condition Expectation) ve MARS (Multivariate Adaptive Regression Spline) algoritmaları, istatistikçiler tarafından geliştirilmiş sınıflandırma algoritmalarıdır(Özdamar, 2002). Bu yöntemlerden, veri madenciliğinde en yaygın olarak kullanılanlar ilerleyen bölümlerde verilmiştir.

Bayes Ağları

Bir Bayes Ağı, bir grup rastsal değişkeni ve onların koşullu bağlılıklarını yönlü dönüşsüz grafiklerle ifade eden, olasılığa dayalı bir grafiksel modeldir. Örneğin, bir Bayes ağı, hastalıklar ve belirtileri arasında olası ilişkileri gösterebilir. Verilen belirtiler ile ağ, değişik hastalıkların varlık olasılıklarını hesaplayabilir.

Daha resmi bir tanımla, Bayes ağları, düğümleri (nodes) rastsal değişkenleri Bayes bağlamında gösteren yönlü dönüşsüz grafiklerdir. Bayes bağlamından kasıt, gözlemlenen miktarlar, gizli değişkenler, bilinmeyen parametreler veya hipotezler olabilir. Kenarlar koşullu bağlılıkları, bağlantılı olmayan düğümler ise koşullu olarak birbirinden bağımsız olan değişkenleri gösterir. Her bir düğüm, düğümün alt/üst değişkenleri için bir grup değeri girdi olarak alan olasılık fonksiyonları ile ilişkilendirirler ve düğüm tarafından gösterilen değişkenin olasılığını verirler(Wikipedia, 2011).

Bayes ağlarının daha geniş hali belirsiz karar ağaçlarıdır (uncertain decision trees). Bayes ağlarını daha iyi anlayabilmek için Bayes teorilerini hatırlatmakta yarar vardır:

$$P(Ai\backslash B) = \frac{P(B\backslash A_i).P(A_i)}{P(B\backslash A_2)+\cdots+P(B\backslash A_n)} = \frac{P(BA_i)}{P(B)} \qquad [1]$$

[1] eşitliği Bayes teorimidir. $P(A_i)$ terimine A_i için önsel olasılık veya marjinal olasılık denir. Önseldir, çünkü B olayı hakkında önceden herhangi bir bilgi içermemektedir. $P(A_i\backslash B)$ terimi, verilmiş bir B olayı için A_i'nin koşullu olasılığıdır. $P(B\backslash A_i)$, verilmiş bir A_i için B'nin koşullu olasılığıdır. Son olarak P(B) terimi ise B olayı için önsel olasılıktır.

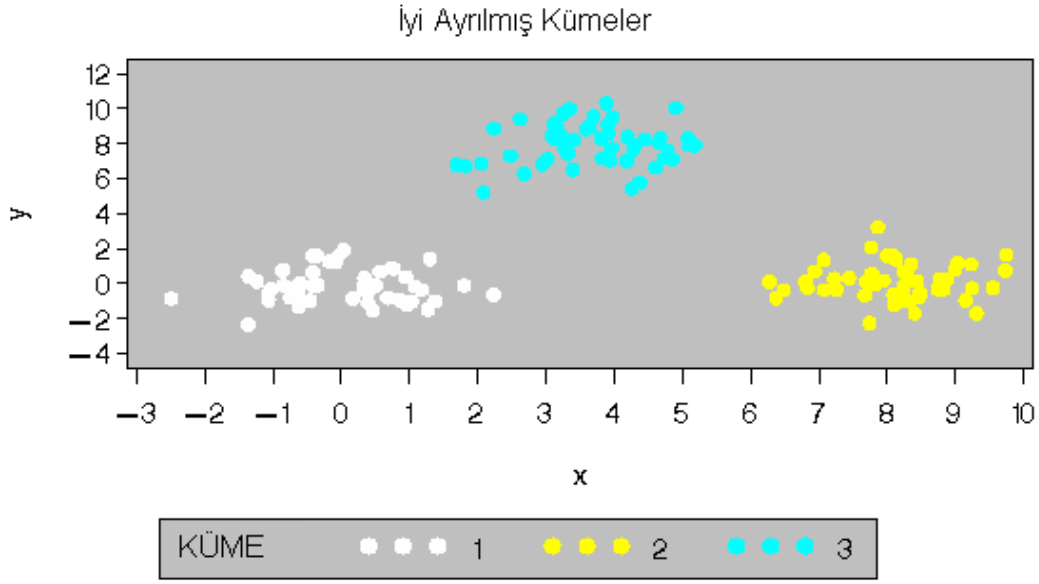
3.5.1.2. Kümeleme

Kümeleme (clustering), verilen bir veri içinde göreceli olarak alt gruplar oluşturmayı amaçlar. Kanser tanısı koyma bağlamında ise, bireylerin kümelenmesi, potansiyel olarak benzer bulgulara sahip bireylerin oluşturduğu alt grupların belirlenmesi şeklinde bir yaklaşımdır. Böyle bir kümeleme, bireylerin laboratuar sonuçları, şikâyetleri vs. gibi özelliklerine dayanır. Başarılı bir kümelemeden sonra, tanısı konulmayan hastaların, ilgili özellikleri gözetilerek, belli bir kümeye dâhil edilmeleri çok daha kolay olacaktır.

Nesnelerin kendilerini veya diğer nesnelerle olan ilişkilerini tarif eden bilgileri kullanarak nesneleri gruplara ayırma işlemleri "kümeleme" olarak ifade edilir. Amaç, grup içinde nesneleri, diğer gruptaki nesnelerden olabildiğince ayrı (bağımsız) ve kendi aralarında da birbirine benzer (bağımlı) olacak şekilde oluşturmaktır (Zaiane ve diğ., 2002).

Kümeleme için farklı algoritmalar kullanılabilir. Ancak, temelde beş küme tanımı vardır:

1.İyi Ayrılmış Küme: Bir küme içinde herhangi bir nokta, o küme içindeki diğer tüm noktalara, küme dışındaki tüm noktalardan daha yakındır (benzerdir). Kimi zaman küme içindeki nesnelerin birbirine yeterince benzer olduklarını belirlemek için bir eşik değer kullanılır. Kümenin bu ideal tanımı sadece verinin doğal sınıfları (birbirinden yeterince uzak olan sınıfları) içermesi durumunda gerçekleşir. Farklı gruplardaki herhangi iki nokta arasındaki uzaklık aynı grup içinde herhangi iki nokta arasındaki uzaklıktan daha fazladır.



Şekil 3.1: İyi Ayrılmış Kümeler

2.Merkeze Dayalı Küme: Bir küme içindeki her nokta, o kümenin merkezine, diğer kümenin merkezlerine olduklarından daha yakındır (benzerlik). Bir kümenin merkezi, centroid veya medoid gibi o kümeyi temsil eden bir noktadır. Böyle bir küme nesneler setidir. Öyle ki, küme içindeki her bir nesne kümeyi tanımlayan prototipe benzer ya da yakın iken diğer küme prototiplerinden farklı / uzaktır. Sürekli özelliğe sahip veriler için prototip bir ağırlık merkezidir (noktaların ortalaması). Ağırlık merkezinin anlamlı olmadığı durumlarda, örnek olarak veri kategorik özelliklere sahipse, bu durumda prototip bir medoid'dir. Yani, kümeyi en iyi temsil edecek noktadır. Kümeler doğal olarak küresel şekle sahip olma eğilimindedirler(Özdamar, 2002).

3.En Yakın Komşu Küme: Bir küme içindeki herhangi bir nokta, küme içindeki diğer bir noktaya ya da noktalara başka bir kümedeki noktalardan daha yakındır (benzerlik). Kümenin bu tanımı, kümeler düzensiz veya birbirine geçmiş durumda iken yararlıdır; fakat bununla beraber gürültünün olması durumunda sorunlarla karşılaşılabilir (Özdamar, 2002).

4.Yoğunluğa Dayalı Küme: Bir küme, noktaların yoğunluklarına göre diğerlerinden ayrılır. Yüksek yoğunluklu olanlar, düşük yoğunluklu olanlar tarafından ayrılır. Veri

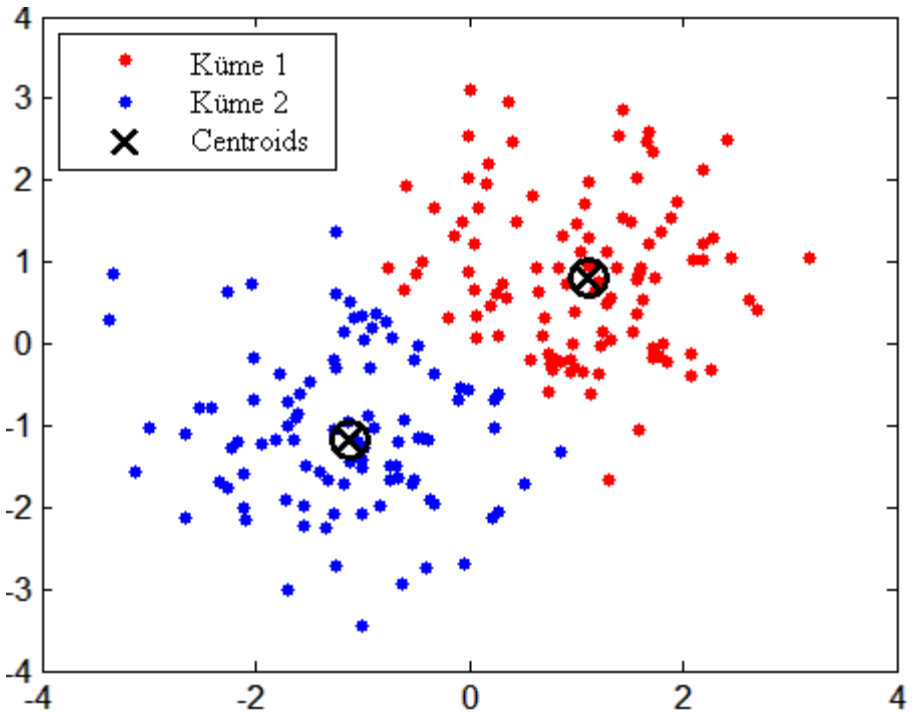
setinde gürültü ve aykırı gözlem olması durumunda bu tanım önem kazanır (Özdamar, 2002).

5.Benzerliğe Dayalı Küme: Bir küme, birbirine benzer nesnelerden oluşur. Kendisi dışındaki kümelerin nesneleri, kendi nesnelerine benzemez. Bu tanım özellikle noktaların oluşturdukları yoğunluk ya da şekle göre, kümeleri farklı olarak tanımlar (Özdamar, 2002).

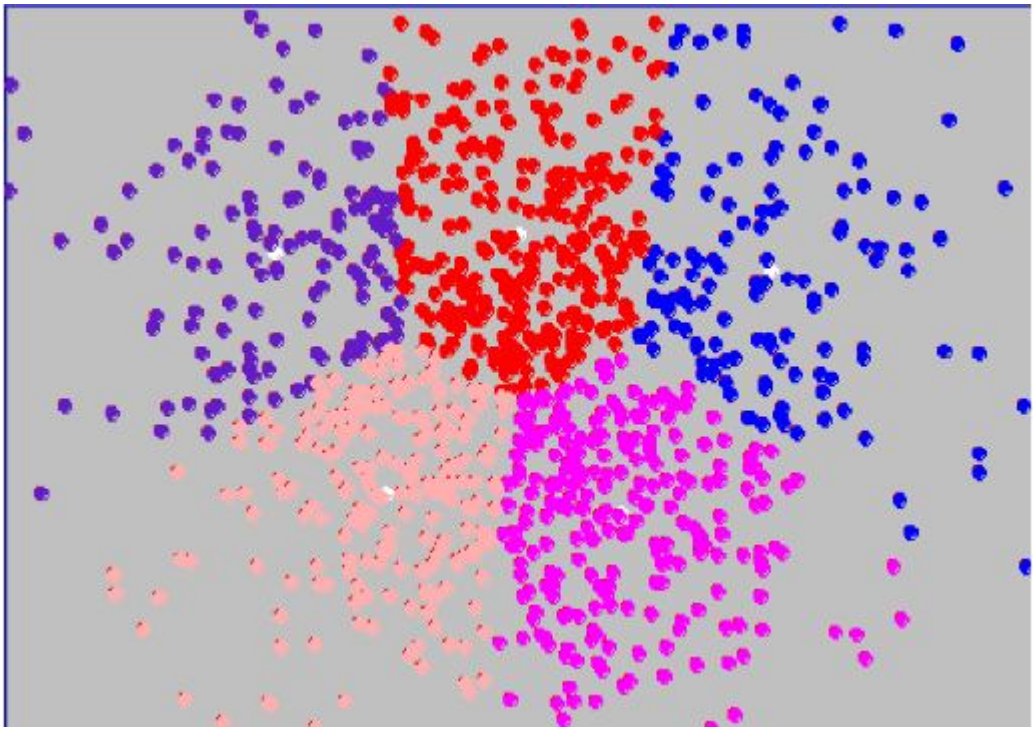
Veri madenciliğinde kullanılan kümeleme teknikleri şunlardır;

- Merkeze dayalı ayırıcı kümeleme teknikleri
- Hiyerarşik kümeleme teknikleri
- Yoğunluğa dayalı kümeleme teknikleri
- Grafik esaslı kümeleme teknikleri

Birçok ayrıcı kümeleme tekniği vardır ve bunlar hemen hemen aynı sonucu verirler. En sık kullanılan merkeze dayalı, ayrıcı kümeleme teknikleri k-means ve k-medoid yaklaşımlarıdır. Her iki yaklaşım da bir kümenin ağırlık merkezinin o kümeyi temsil ettiğini varsayar. K-means ağırlık merkezi olarak kümenin ortalamasını veya medyanını alır ve merkez nokta (centroid) kabul eder. K-medoid ise o kümeyi en iyi şekilde temsil eden nokta olan medoid kavramına dayanır. Medoid, kümenin orta noktasıdır ve gerçek bir noktadır (Özdamar, 2002).



Şekil3.2: K-means Kümeleme



Şekil 3.3: K-medoids Kümeleme(Velmurugan ve Santhanam, 2011)

K-means kümeleme algoritması oldukça basit bir algoritmadır. Algoritmanın uygulama adımları:

1. K tane başlangıç merkezi noktası (centroid) seçilmesi,
2. Tüm noktaların en yakınındaki merkezi noktaya atanması,
3. Her küme için merkezi noktaların yeniden hesaplanması,
4. 2. ve 3. adımların merkezi noktalar değişmeyinceye kadar tekrarlanması şeklindedir (Özdamar, 2002).

K-medoid algoritmasının çalışma adımları ise;

1. K tane başlangıç noktası seçilir. Kendi kümelerinin en orta noktaları oldukları varsayılan bu noktalar aday medoidlerdir.
2. Seçilen bu noktaların ayrı ayrı seçilmiş bir noktayla yer değiştirmesi durumu incelenir. Yani, seçilen her noktanın en yakın aday medoide uzaklığı hesaplanır. Toplam uzaklık maliyet olarak adlandırılır. Daha sonra seçilmiş olan nokta başka bir seçilmemiş noktayla değiştirilir ve maliyet hesaplanır. Tüm olası değişimler için maliyetlerin hesaplanmasından sonra 3.adıma geçilir.
3. En düşük maliyetli durum seçilir.
4. En düşük maliyetli durum, seçilmemiş bir noktayla sağlanıyorsa, nokta, en yakın seçilen nokta olan medoide atanır (Özdamar, 2002).

Veri madenciliği için geliştirilmiş bir başka algoritma CLARANS'tır. İstatistikte kullanılan PAM (Partitioning Around Medoid) ve CLARA (Clustering LARge Applications)'ya dayanır. PAM, 4 aşamalı k-medoid algoritmasını uygular. CLARA ise PAM'ın büyük veri setlerine uyarlanmış halidir. CLARANS algoritmasının çalışma adımları:

1. Rastsal olarak k tane aday medoid seçilir.
2. Seçilen noktalardan bir tanesinin seçilmemiş olanlardan bir tanesi ile değiştirilme durumu irdelenir.

3. Yeni durum daha iyi ise 2. adım yeni durumdan başlanarak tekrarlanır.

4. Yeni durum daha kötü ise 2. adım mevcut durumla tekrarlanır. Tekrar sayısı maksimum (250, K(m-k)) olacak şekilde bir parametredir.

5. Ulaşılan son durum daha önceki durumla karşılaştırılır.

6. 2. adımda optimal durum kararı alınmamışsa 1. adıma geri dönülür.

Hiyerarşik kümeleme teknikleri ise ikiye ayrılır;

- Ayırıcı kümeleme teknikleri
- Toplayıcı kümeleme teknikleri

Ayırıcı kümeleme teknikleri toplayıcı kümelemenin tersine çalışır. Bu tekniklerde her adımda hangi kümenin ayrılacağına karar verilir. Çok yaygın olarak kullanılır.

Ayırıcı kümeleme tekniklerinden Minimum Yayılma Ağaç (MST: Minimum Spanning Tree) Algoritması aşağıdaki gibi uygulanır.

1. Yakınlık grafiği için minimum yayılmalı ağaç hesaplanır.

2. En büyük uzaklığa göre kümeler arası var olan bağıntılar ayrılarak yeni kümeler oluşturulur.

3. Tüm kümeler tek noktadan oluşacak şekilde 2.adım tekrarlanır.

Toplayıcı kümelemede her bir veri en altta birer küme olarak kabul edilir. Yani, en altta veri sayısı kadar küme vardır. Yukarı doğru çıkıldıkça küçük kümeler birleşerek büyük kümeler oluştururlar. En üstte tüm verileri / kümeleri kapsayan tek bir küme vardır.

Yoğunluğa bağlı kümeleme teknikleri, yoğunluğa bağlı kümeleme tanımından yola çıkarak geliştirilmiştir. Bazı algoritmalar, CLIQUE, MAFIA (Merping of Adaptive Finite Intervals) ve DBSCAN'dir. Özellikle CLIQUE ve MAFIA çok boyutlu veri setleri üzerine rahat kümeleme yapabilirler. DBSCAN kümeleme algoritması, veri setindeki her noktayı farklı uzaklık ölçülerine göre ya kümelere atar ya da bir gürültülü nokta olarak kabul eder.

Grafik esaslı kümeleme teknikleri, oluşturdukları kümeleri basit grafikler veya hipergrafiklerle gösterirler. Kullanılan grafik türlerine göre farklı algoritmalar vardır:

- Grafiğe dayalı kümeleme tekniği
- Hipergrafiğe dayalı kümeleme tekniği
- ROCK (Robust Clustering Using linKs)

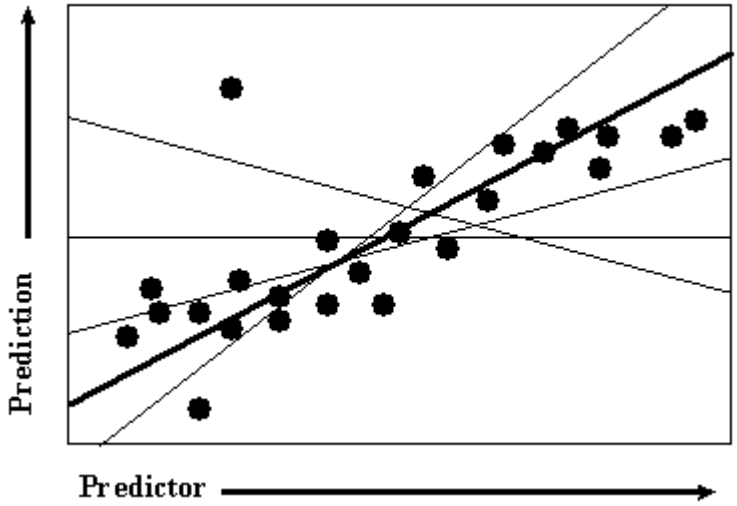
Grafiğe dayalı kümeleme teknikleri, oluşturdukları kümeleri dendogramlarla gösteren hiyerarşik kümeleme teknikleridir. En yakın komşu kavramından yola çıkılarak kümeleme yapılır. Kümelemenin nasıl yapıldığı ilerleyen bölümlerde açıklanmıştır.

3.5.1.3 Regresyon

Regresyon, sınıflandırmanın aksine, kategorik değil, süreklilik gösteren verilerin tahmininde kullanılan, önemli veri sınıflarını belirten ve gelecekteki veri sonuçlarını tahmin eden modeller kuran veri analiz yöntemidir.

Tahmin ederken kullanılan modeller birer madencilik fonksiyonudur ve nesnelerin muhtemel değerlerini tahmin eder.

İstatistikte, regresyon kelimesi çoğu zaman tahmin kelimesi yerine telaffuz edilir. İstatistikte değişik regresyon tipleri vardır, fakat temel düşünce, tahmin yapılırken en az hata payına sahip olacak şekilde, tahmincilerce alınan değerleri haritalayan bir model oluşturmaktır. Regresyonun en basit hali, lineer (doğrusal) regresyondur ve sadece bir tahmin edici ile bir tahminden oluşur. Tahmin ve tahmin edici (prediction – predictor) arasındaki ilişkiyi gösteren grafik iki boyutlu uzayda X ekseninde tahmin edici, Y ekseninde tahmin olmak üzere gösterilebilir. Öyle ki, bu doğru, gerçek tahmin değeri ile doğru üzerindeki nokta (modelin tahmin değeri) arasındaki hata oranını minimuma indirir. Bu durum grafiksel olarak Şekil 3.4'deki gibidir.



Şekil 3.4: Regresyon

Tahmin yöntemi olarak doğrusal regresyon dışında, başka yöntemler de kullanılabilir. Bunlar, doğrusal olmayan regresyon, genelleştirilmiş doğrusal regresyon ve lojistik regresyon olabilir (Yaralıoğlu, 2007).

Regresyon ile veri analizi kredi skor uygulamalarında sıklıkla kullanılmaktadır. Kredi, bir finansal kurumu tarafından müşteriye ödünç olarak verilen ve faiz eklendikten sonra genelde düzenli aralıklı taksitler halinde geri ödenmesi gereken paradır. Bir kredi başvurusunda müşterinin krediyi geri ödememesi olasılığını hesaplamaya kredi skorlama denir. Skorlama yaparak yüksek riskli müşterilere kredi vermeyi reddetmek finansal kurumun olası zararını azaltacaktır (Küçüksille, 2009). Regresyon analizi bu gibi durumlarda kullanılan oldukça basit bir istatistiksel tahmin yöntemidir.

3.5.1.4. Birliktelik Kuralı

Birliktelik kuralının amacı, büyük veri kümeleri arasında birliktelik ilişkileri bulmaktır. Gün geçtikçe veri miktarının hızla artmasıyla birlikte şirketler veritabanlarındaki veriler arasındaki birliktelik kurallarını ortaya çıkarmak isterler. Bunu istemedeki amaç, karar alma sürecinde olumlu etkileri olacak olan, depolanan verilerdeki farklı birliktelik ilişkilerini bulmaktır.

Birliktelik kuralları yaygın olarak market sepeti analizinde kullanılır. Sepet analizindeki birliktelik kuralı, müşterilerin birlikte aldıkları ürünler arasındaki ilişkileri bulur ve müşterilerin satın alma alışkanlıklarını analiz eder. Satın alma alışkanlıkları öğrenildikten sonra şirket yöneticileri daha etkili satış stratejileri geliştirilebilir. Örneğin, bir müşterinin kuruyemiş ile kola alma olasılığı bilinirse, raflarda yapılacak düzenlemelerle satış oranları artırılabilir. Örneğin, eğer olasılık yüksekse kuruyemiş ve içecek reyonlarının yan yana olarak ayarlanacak olması içecek satışlarını arttırabilir.

Sepet analizi, promosyon çalışmalarında, ürün kataloglarının hazırlanmasında ya da mağazaların düzenlenmesinde önemli rol oynar. Sepet analizi çok kullanışlı olmasına rağmen yorumlanmasında bazı problemler olabilir. Örneğin, bir mağazada ekmek alanların %30'unun süt aldığının bilinmesi önemli bir bilgidir. Eğer bu mağazadan yapılan alışverişlerin %50'sini süt oluşturuyorsa, bu iki ürünün birlikte alınmadıkları sonucuna bile varılabilir. Günümüzde yapılan sepet analizi araştırmalarının çoğunda bu rastlantısal ilişkiler dikkate alınmadan sonuçlar çıkartılmaktadır(Yılmaz, 2006).

3.5.2. Bellek Tabanlı Yöntemler

Bellek tabanlı veya örnek tabanlı yöntemler, istatistikte 1950'li yıllarda önerilmiş olmasına rağmen o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılmamış ama günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yönteme en iyi örnek k-en yakın komşu algoritmasıdır (Mitchell, 1997).

3.5.2.1. K-En Yakın Komşu (K-NN)

En yakın komşu tahmin yöntemi, veri madenciliğinde kullanılan en eski yöntemlerden biridir. En yakın komşu yöntemi kümelemeye oldukça benzer ve bir noktanın en yakınındaki noktayla beraber aynı kümeye dâhil olacağı prensibine dayanır.

Ortak en yakın komşu kümesinin iki farklı tanımı vardır ve bu algoritmalar farklı algoritmaların yaratılmasına neden olmuştur. Ortak en yakın komşu kümelemesinin ilk tanımına göre oluşturulan algoritmanın adımları,

1. Her nokta için en yakın komşusu bulunur.
2. Her noktanın kendisi ve komşusundan oluşan çift karşılaştırılır. Karşılaştırmada şu iki nokta dikkate alınır ve üçüncü adıma geçilir.
    a. Eğer iki nokta belli bir komşu sayısından fazla ise,
    b. İki nokta da birbirinin k-en yakın komşularına dâhilse,
3. Bu iki nokta ve bulundukları kümeler birleştirilir.

Farklı yoğunluktaki kümeleri ayırt edebilen bu kümeleme tekniğinin geçişli bir yapısı vardır. Eğer p noktasından q noktasına ortak komşu varsa ve aynı zamanda q noktasının r noktasına ortak komşusu varsa, p, q ve r noktaları aynı kümeye atanırlar. Böylece, algoritma farklı boyutlarda ve şekillerdeki kümeleri de fark edebilir.

Ortak en yakın komşu kümesinin ikinci tanımı, noktaların en yakın komşularının sıralanarak sıra sayılarının toplanmasıyla elde edilen ortak komşuluk değerine dayanır.

Ortak komşuluk değerinin birbirine yakın olduğu noktalar ortak en yakın komşu sayılırlar. Bu tanımdan yola çıkarak oluşturulan algoritmanın adımları;

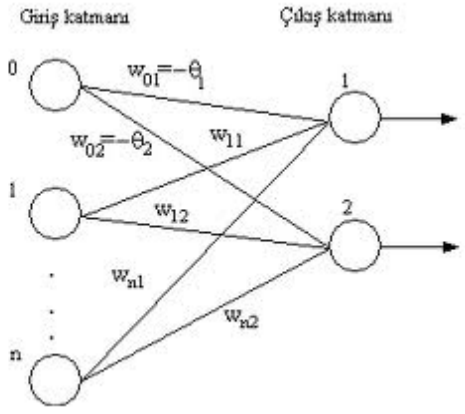
1. Her nokta için k-en yakın komşular bulunur.
2. Her noktanın bütün k-en yakın komşuları için nokta ve komşusu arasındaki ortak komşuluk değeri hesaplanır.
3. En düşük ortak komşuluk değerine sahip olan kümeler birleştirilir.
4. Üçüncü adım istenen küme sayısı elde edilinceye ya da kümeler artık birleştirilinceye kadar devam edilir.

Bu tanıma göre yapılan kümeleme, farklı yoğunluk, büyüklük ve şekillerdeki kümeleri fark edebilir.

3.5.3. Yapay Sinir Ağları

Yapay sinir ağları, beyin hücreleri olan nöronların çalışma prensibini modelleyen, öğrenebilen algoritmalardır. İleriye yönelik ya da geri besleme alabilen yenilemeli ağlar olmak üzere iki tür yapısı vardır. Yapay sinir ağları (YSA), ağırlıklı bağlantılar denilen tek yönlü iletişim kanalları sayesinde birbiriyle haberleşen, her biri kendi hafızasına sahip birçok nörondan oluşan paralel ve dağıtık bilgi işleme yapılarıdır. YSA'lar gerçek dünyaya ait ilişkileri tanıyabilir, sınıflandırma, kestirim ve işlev uydurma gibi görevleri yerine getirebilirler. Desen tanıma tekniğinin gerekliliği, gerçek dünya ile bilgisayar ilişkisinin başlaması ile ortaya çıkmıştır. Bu durumda YSA'nın çok güçlü örnek tanıma tekniği olarak ortaya çıkmasına ve gelişmesine neden olmuştur. (Erdem ve diğ., 2005)

Yapı kurulduktan sonra sinir ağacı eğitilir. Giriş verilerine karşılık çıkış verileri alınır. Bu değer gerçek değerlerle karşılaştırılır ve ağın içerisindeki nöron fonksiyonlarının bu sonuçtaki hata miktarına göre ayarlanması sağlanır. Bu şekilde birçok değer ağa verilir ve ağın eldeki verinin yapısının öğrenilmesi sağlanır. Öğrenme işlemi tamamlandıktan sonra sinir ağı kullanıma hazır hale gelir (Yılmaz, 2006). Bir sinir ağı yapısı aşağıdaki gibidir:

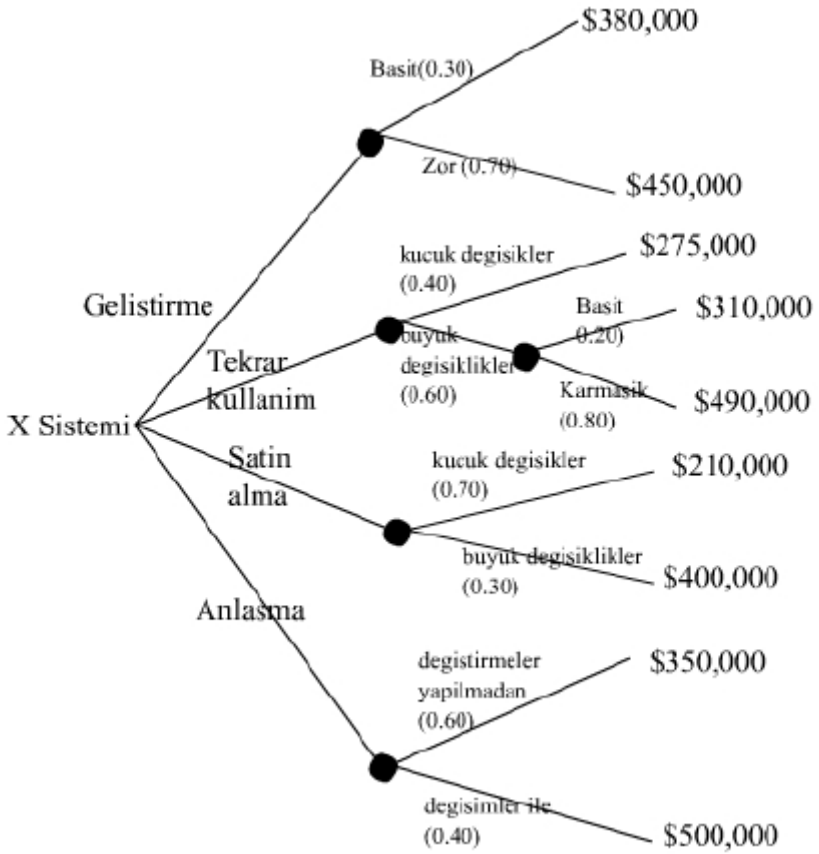


Şekil 3.5: Örnek Bir Yapay Sinir Ağı Yapısı

3.5.4. Karar Ağaçları

Karar ağaçları, kurulmalarının kolay olması, kolay yorumlanabilmesi, veritabanı sistemleriyle kolayca entegre edilebilmesi nedeniyle veri madenciliğinde oldukça yaygın olarak kullanılmaktadır. Bir tahmin algoritması olmalarının yanında, karar ağaçları, bir yandan da kural çıkarma algoritmalarıdır. Kullanıcıların çok rahat anlayabileceği "IF – THEN" türündeki kuralları, bu algoritmalar ağaç yapısında üretebilirler(Uyan ve Çay, 2008).

Bir karar ağacında örnekler, sonlu sayıda sınıfa ayrılırlar. Bir ağaçta, düğümler nitelik isimleri ile uçlar ise bu niteliklerin muhtemel değerleri ile ve her bir yaprak farklı sınıflarda etiketlenir. Bu nesne uçlarındaki niteliklerin değerleri ağacın en alttaki sınıflarıdır.



Şekil 3.6: Örnek Karar Ağacı (Yegül, 2011)

Ali Serhan Koyuncugil (Koyuncugil, 2006), karar ağacının temel özelliklerini

- Verinin herhangi bir kayba yol açmadan her bir dala ayrılabilmesi,
- Modelin nasıl yapılandırıldığının diğer modellerin aksine çok kolay anlaşılması,
- Oluşturulan modelin kolayca kullanılması, ayrıca bazı sezginlerin de modelde yapılandırılmasının mümkün olması,
- Karar ağaçlarının iç içe geçmiş if/then kurallarının dizisi olması, görsel olması nedeniyle oldukça kolay anlaşılması ve aynı zamanda kolaylıkla SQL sorgusuna dönüştürülebilir olması,
- Değişken tiplerine göre, farklı yöntemler kullanabilmesi şeklinde sıralanmaktadır.

Karar ağacı teknolojisi, veri kümesi ve iş problemlerinin keşfinde kullanılabilir. Kredi başvuruları, risk değerlendirilmesi gibi uygulamalarda sıklıkla kullanılmaktadırlar. En sık kullanılan karar ağacı modelleri ID3 ve daha gelişmiş modeli C4.5, Sınıflandırma ve Regresyon Ağacı (CART) ve Otomatik Ki-kare Etkileşim Keşfedicisi (CHAID)'dir.

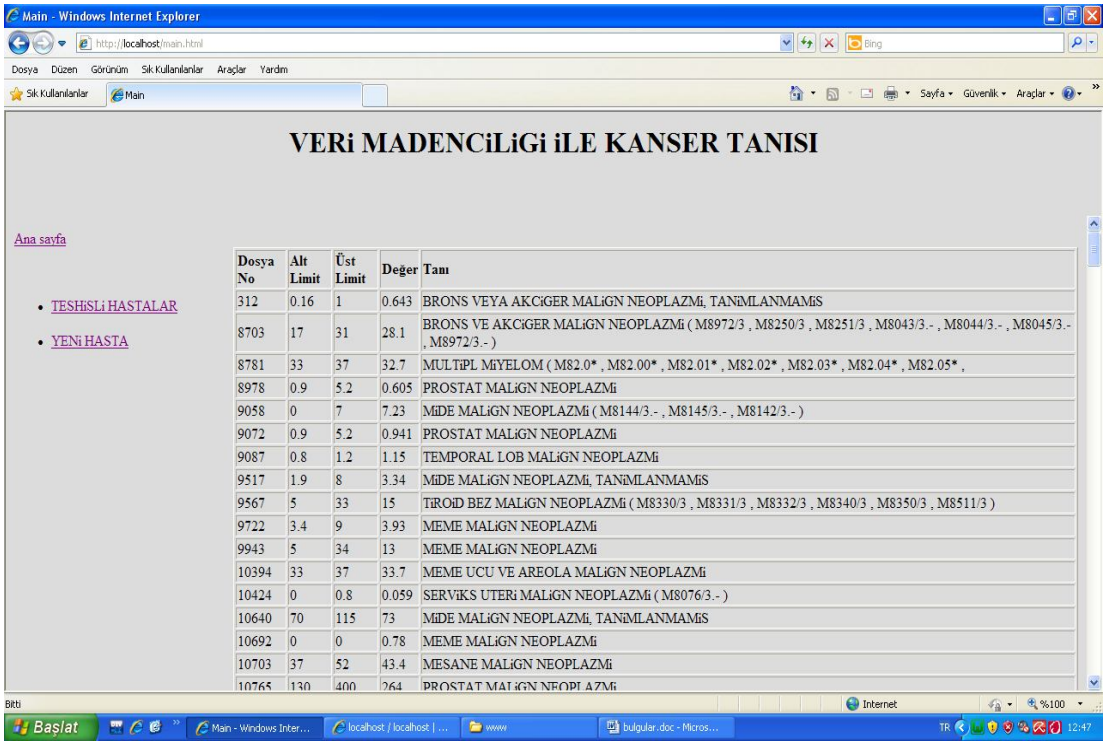
CHAID ile diğer yöntemler arasındaki en önemli fark, ağaç türetim biçimidir. ID3, C4.5 ve CART ikili ağaçlar türetirken, CHAID ikili olmayan, çoklu ağaçlar türetir. CHAID sayısal ve sözel tüm veri türleriyle çalışabilir. CHAID, Ki-Kare yöntemi vasıtasıyla ilgi düzeyine göre farklılık içeren grupları ayrı ayrı sınıflamaktadır. Dolayısıyla ağacın yaprakları ikili değil, verideki farklı yapı sayısı kadar dallanmaktadır.

4. BULGULAR

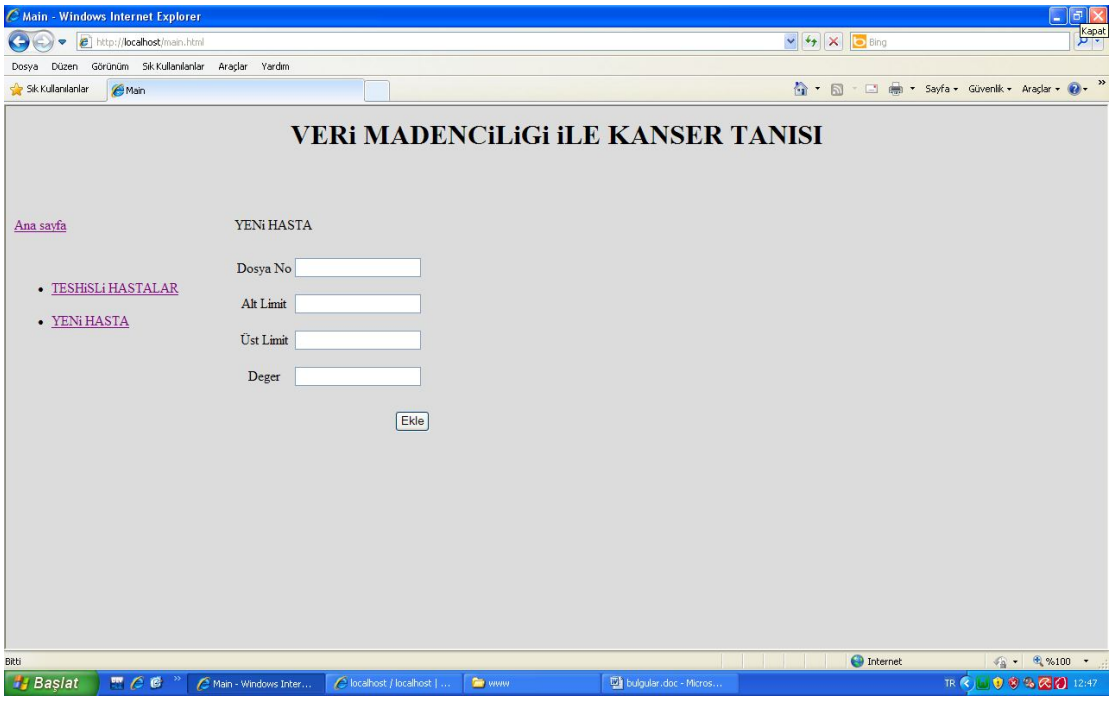
Bu tez çalışmasında önerilen kanser tanı sisteminin veritabanı boyutundaki işlemler için MySQL VTYS ve Wamp Server kullanılmıştır. Ayrıca, veritabanı işlemlerini yapabilmeyi kolaylaştıracak, görsel ara yüzler de hazırlanmış ve bunun için de PHP web programlama dili kullanılmıştır. Aşağıdaki şekilde sistem kullanıcısını karşılayan ara yüz görülmektedir.

Şekil 4.1: Karşılama Ekranı

Şekil 4.1'deki ara yüzden teşhisli hastalar bağlantısı tıklandığında, veritabanında daha önceden kayıtlı olan ve teşhisi konmuş olan hastaların bilgilerine (laboratuar sonuçları ve tanılarına) erişilmektedir. Şekil 4.2, bu ara yüzü göstermektedir.



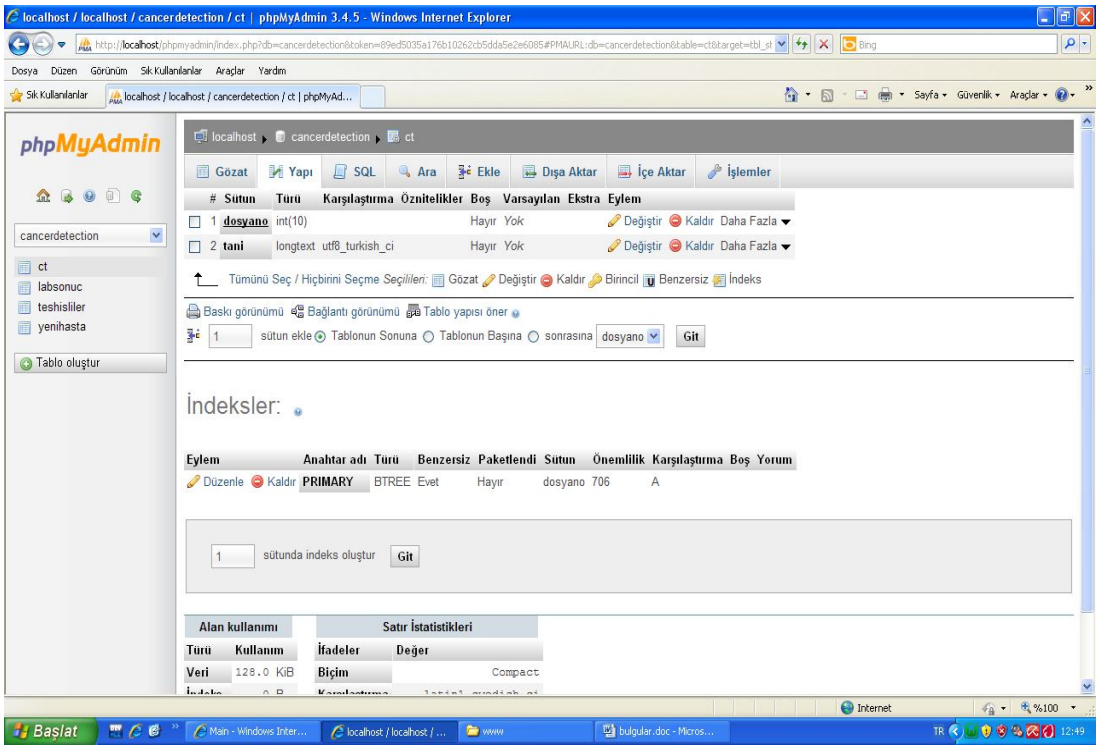
Şekil 4.2: Teşhisli Hastalar Ara Yüzü



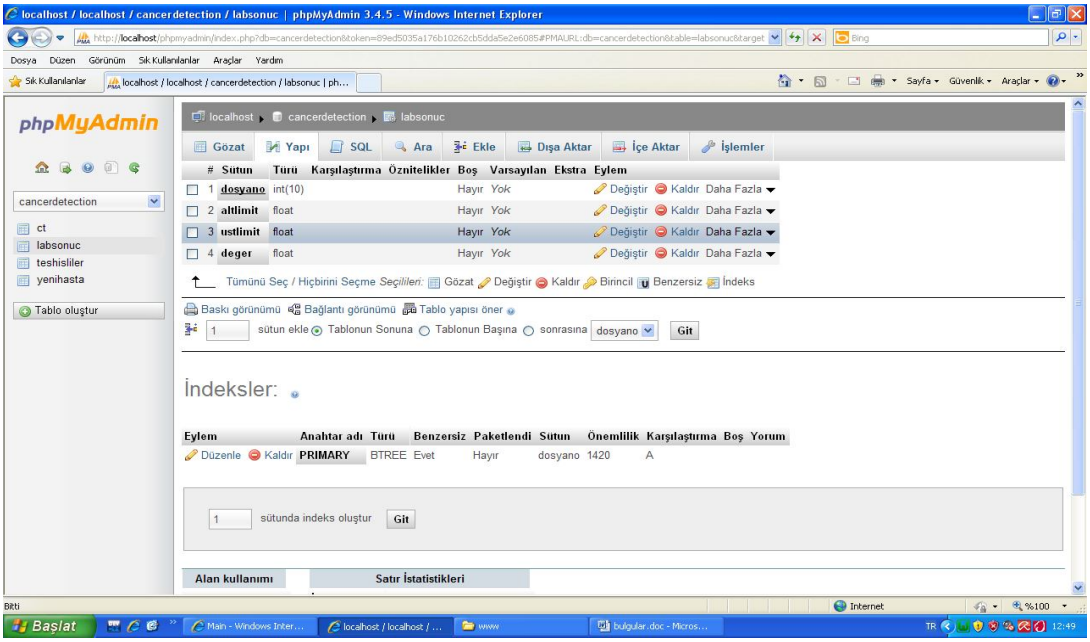
Şekil 4.3: Tanısı Tahmin Edilecek Hastanın Verilerinin Alındığı Ara Yüz

Karşılama ekranından “Yeni Hasta” bağlantısı tıklandığında ise henüz teşhisi konmamış ve tanısının tahmin edilmesi beklenen hastanın verilerinin girilmesini sağlayan bir form ekranı görülmektedir. Şekil 4.3’de veri girişi bekleyen form ekranı görülmektedir.

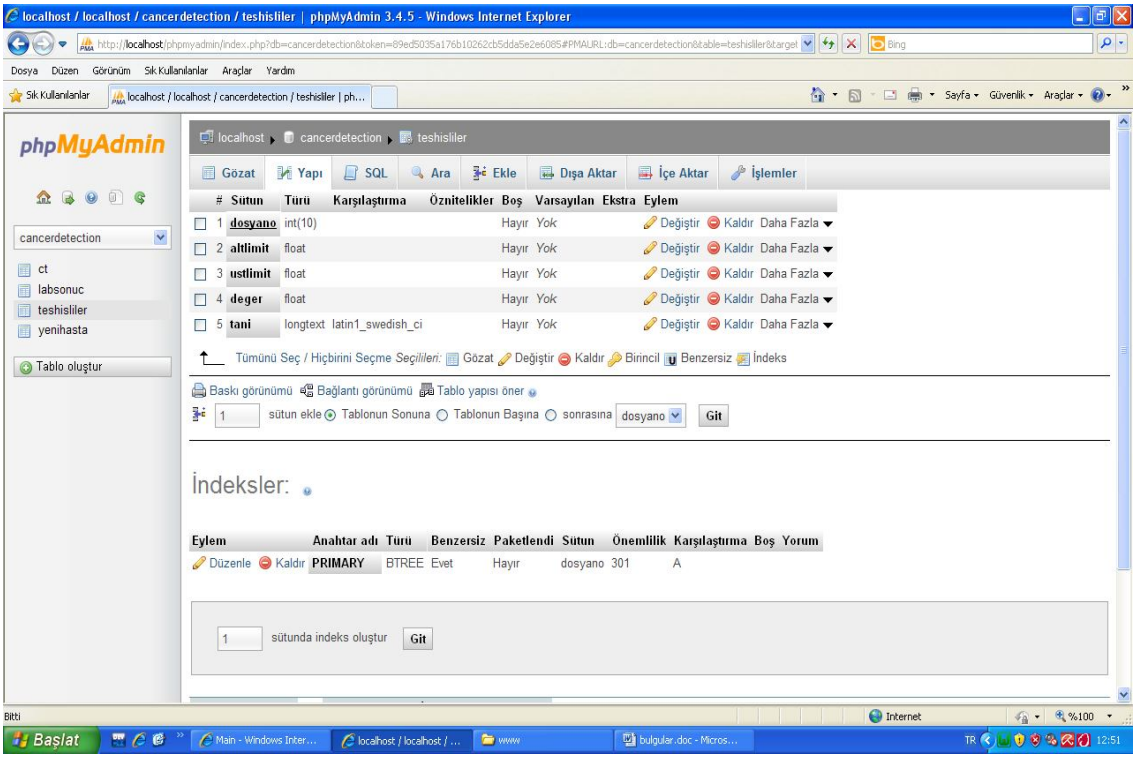
Yukarıda bahsedilen ara yüzlerin arka planındaki veritabanı ise takip eden şekillerde gösterilmeye çalışılmıştır. Veritabanında “ct”, “labsonuc”,”teshisliler” ve “yenihasta” tabloları oluşturulmuştur. Tabloların ortak birincil anahtarları, hastalara ait ve tek olan dosya numarasıdır. Tabloların yapıları Şekil 4.4, Şekil 4.5, Şekil 4.6 ve Şekil 4.7’de gösterilmiştir.



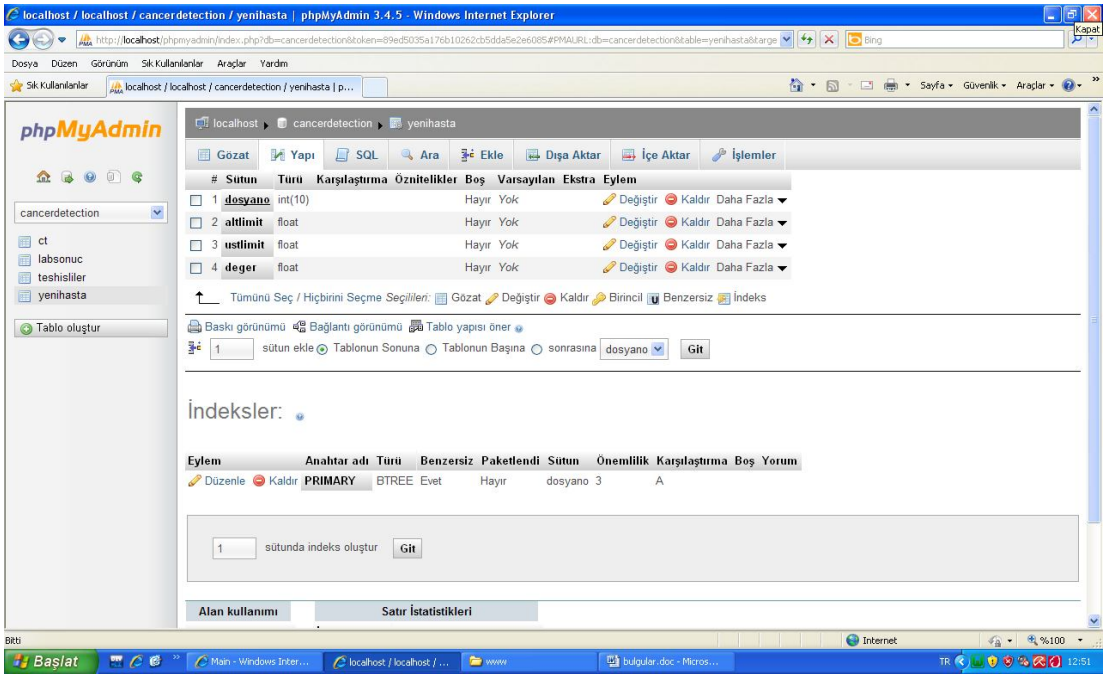
Şekil 4.4: “ct” tablosu



Şekil 4.5: "labsonuc" tablosu

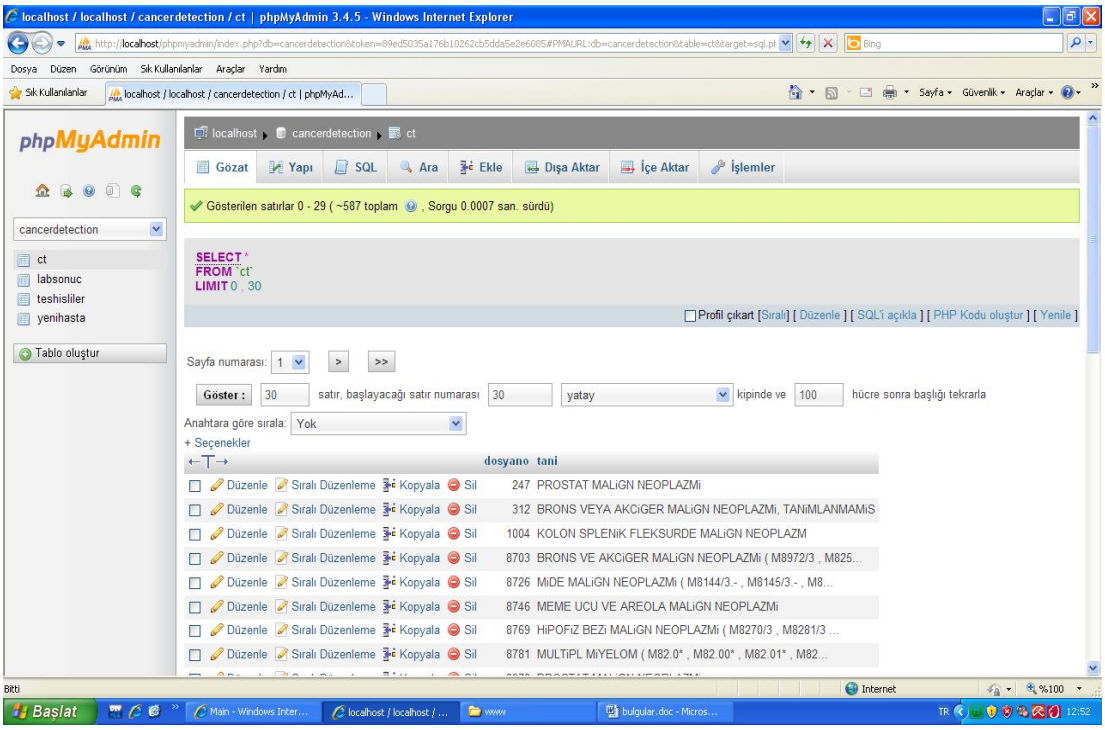


Şekil 4.6: "teshisliler" tablosu



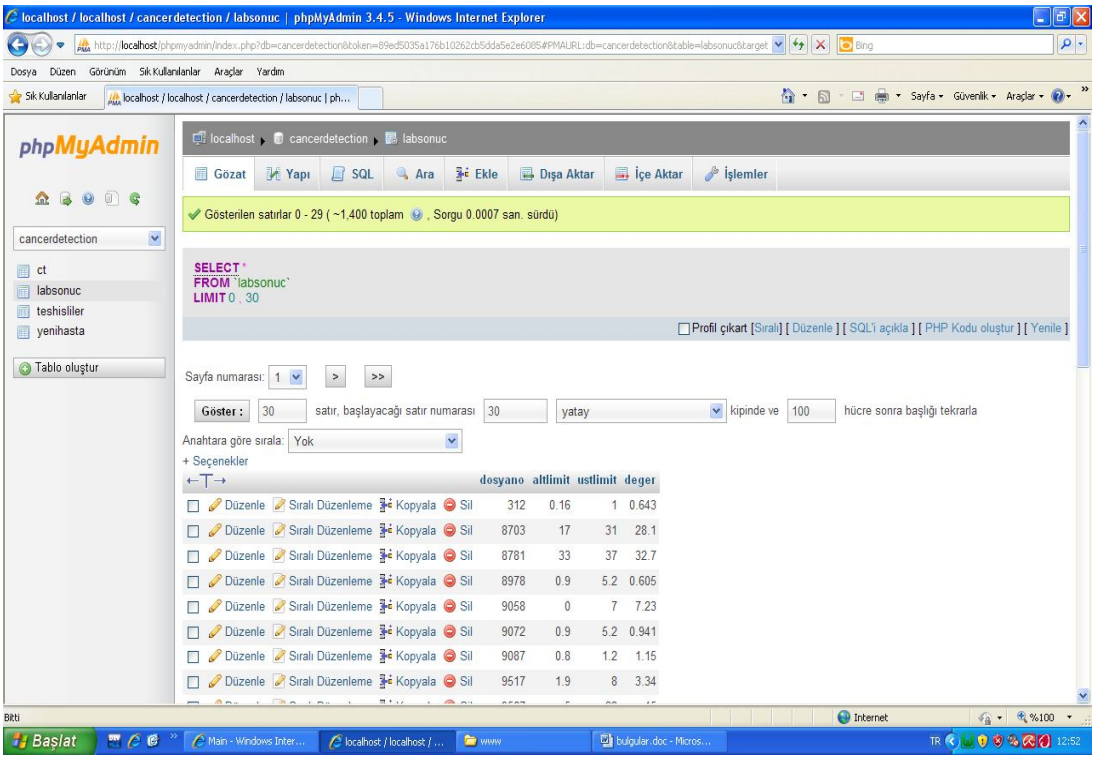
Şekil 4.7: "yeni hasta" tablosu

"ct" tablosu daha önceden kanser tanısı konmuş olan hastaların dosya numaraları ve tanı bilgilerinden oluşmaktadır (Şekil 4.8).



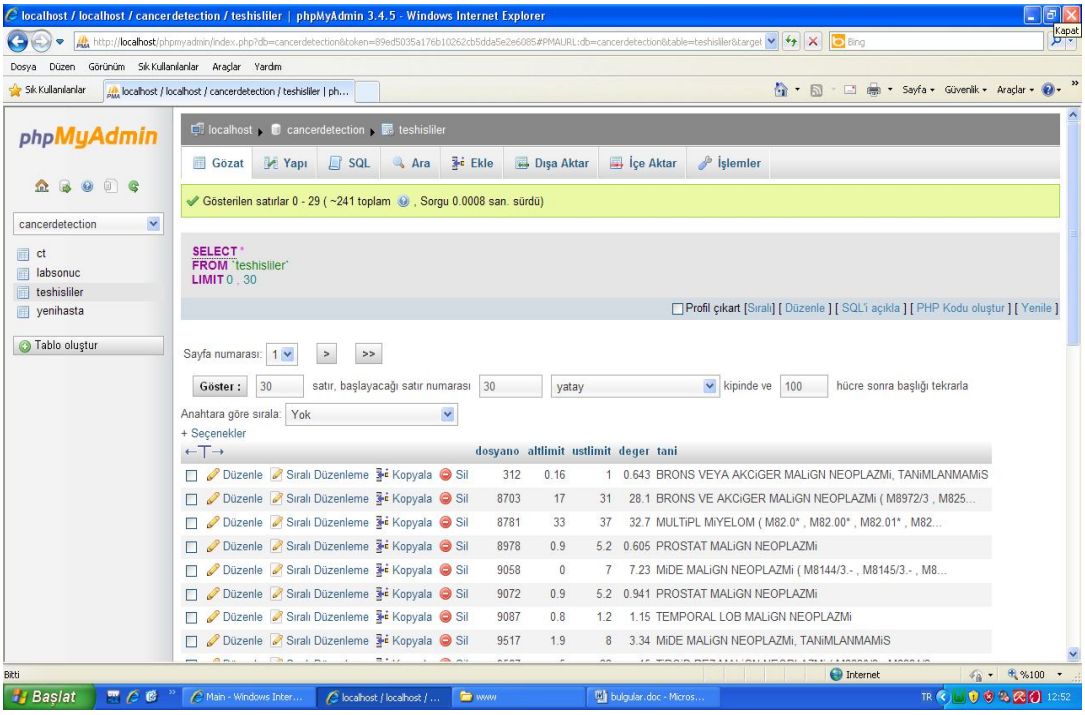
Şekil 4.8: "ct" tablosunun görünümü

"labsonuc" tablosu hastaların dosya numaraları ile birlikte laboratuardan aldıkları sonuçların tutulduğu tablodur. Laboratuar sonuçları alt limit, üst limit ve değer olmak üzere üç sayısal ifadeden oluşur(Şekil 4.9).



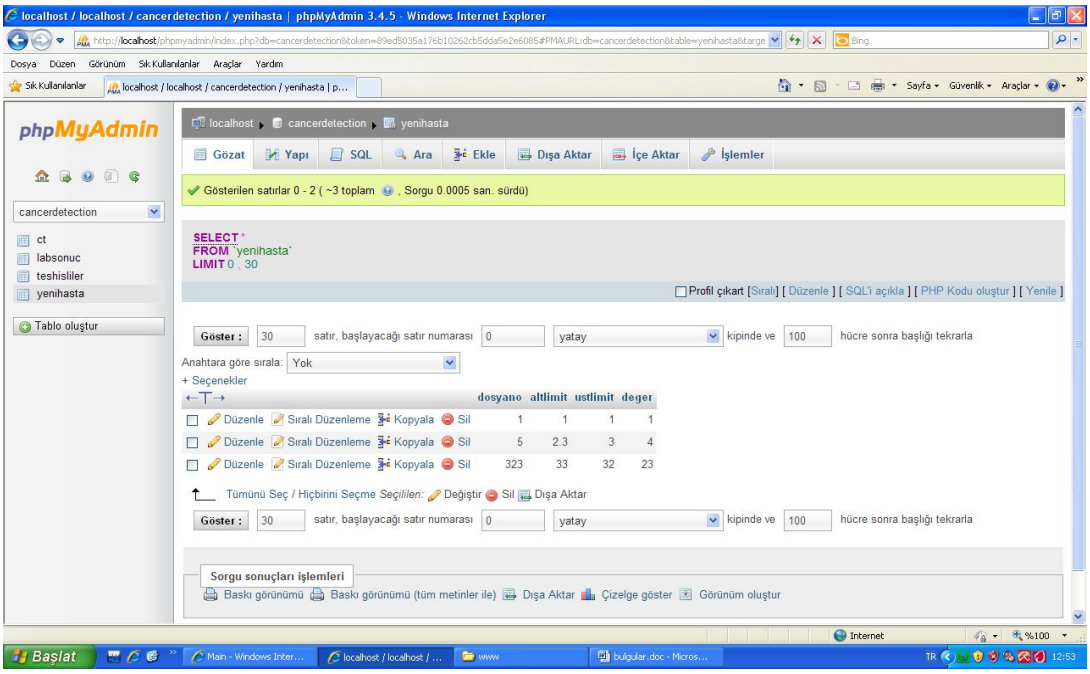
Şekil 4.9: "labsonuc" tablosunun görünümü

"teshisliler" tablosunda kanser türleri kesinleşmiş olan hastaların laboratuar sonuçları ve tanıları bulunmaktadır. Bu tablo "ct" tablosu ve "labsonuc" tablosunun uygun şekilde birleştirilmesiyle oluşturulmuştur (Şekil 4.10).



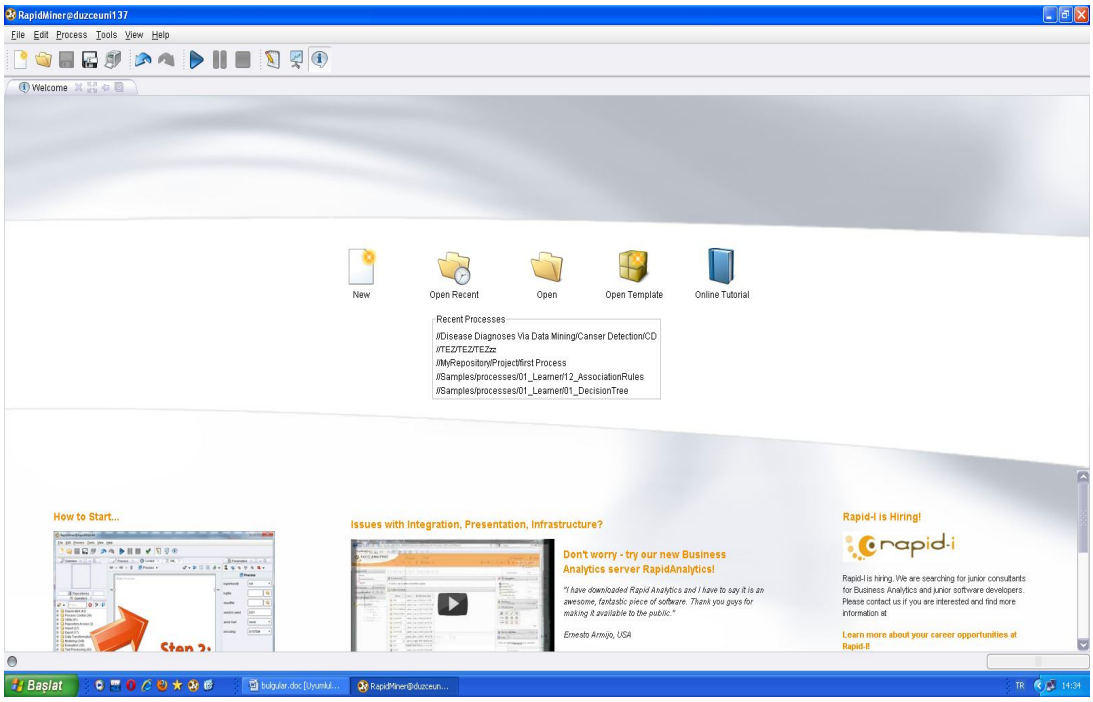
Şekil 4.10: "teshisliler" tablosunun görünümü

Son olarak, yeni hasta tablosu, kanser türü tahmin edilecek hastanın dosya numarası ve laboratuar sonuçlarını içeren bir tablodur ve Şekil 4.11'de gösterilmiştir.



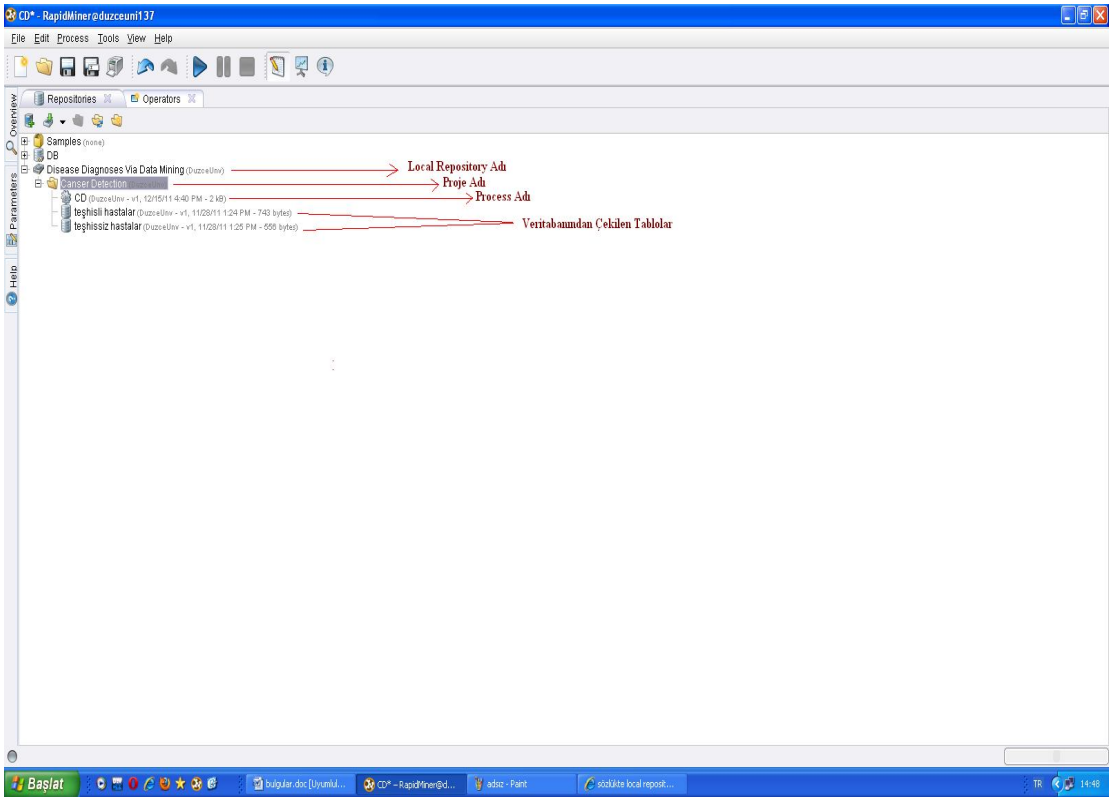
Şekil 4.11: "yenihasta" tablosu

Veritabanı işlemleri başarıldıktan sonra sıra eldeki hastanın tanısını tahmin etmeye gelmektedir. Bunun için RapidMiner veri madenciliği aracı kullanılmıştır. RapidMiner'ın karşılama ekranı Şekil 4.12'de gösterilmiştir.



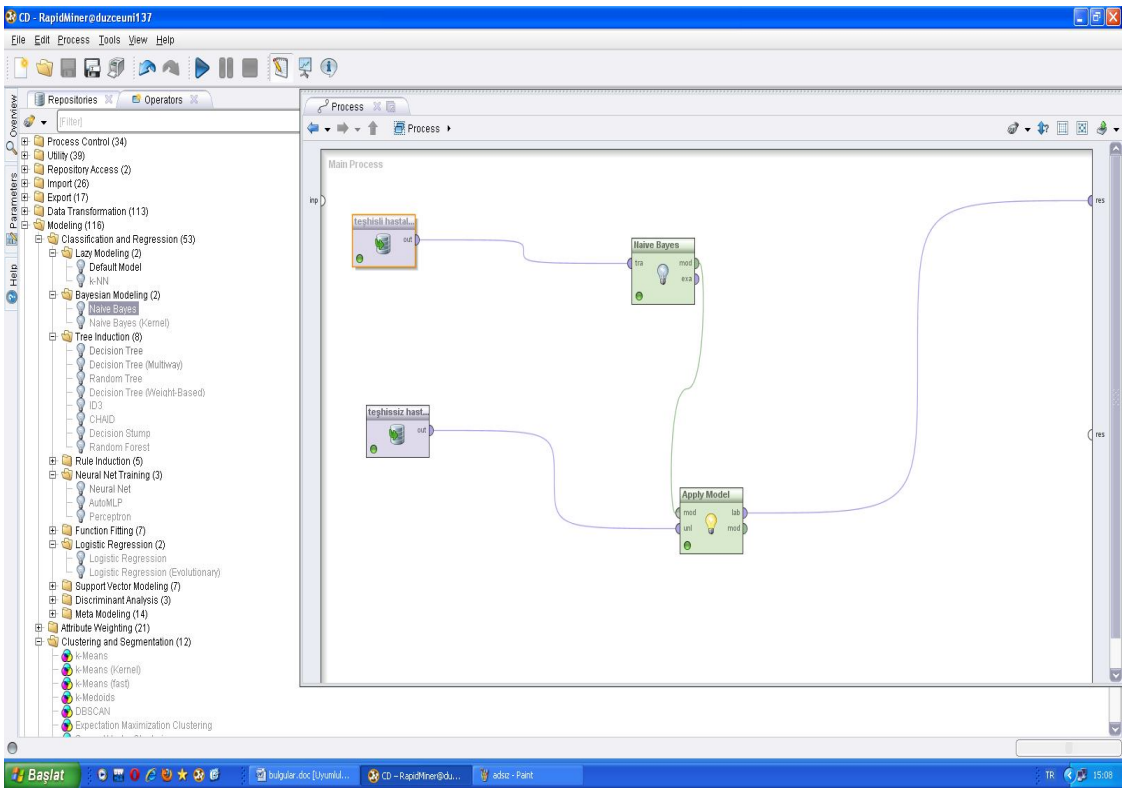
Şekil 4.12: RapidMiner'ın Karşılama Ekranı

RapidMiner'da öncelikle bir local repository oluşturulup, daha sonra onun altında yeni bir proje açılır. Sonra, projenin altında process oluşturulup, ilgili tablolar veritabanından çekilir. Şekil 4.13'de gösterildiği gibi local repository'nin adı "Disase Diagnoses via Data Mining", projenin adı "Cancer Detection", process'in adı "CD" ve veritabanından çekilen tablolar da "teşhisli hastalar" (teshilsiler tablosu) ve "teşhissiz hastalar" (yenihasta tablosu) olarak isimlendirilmiştir.

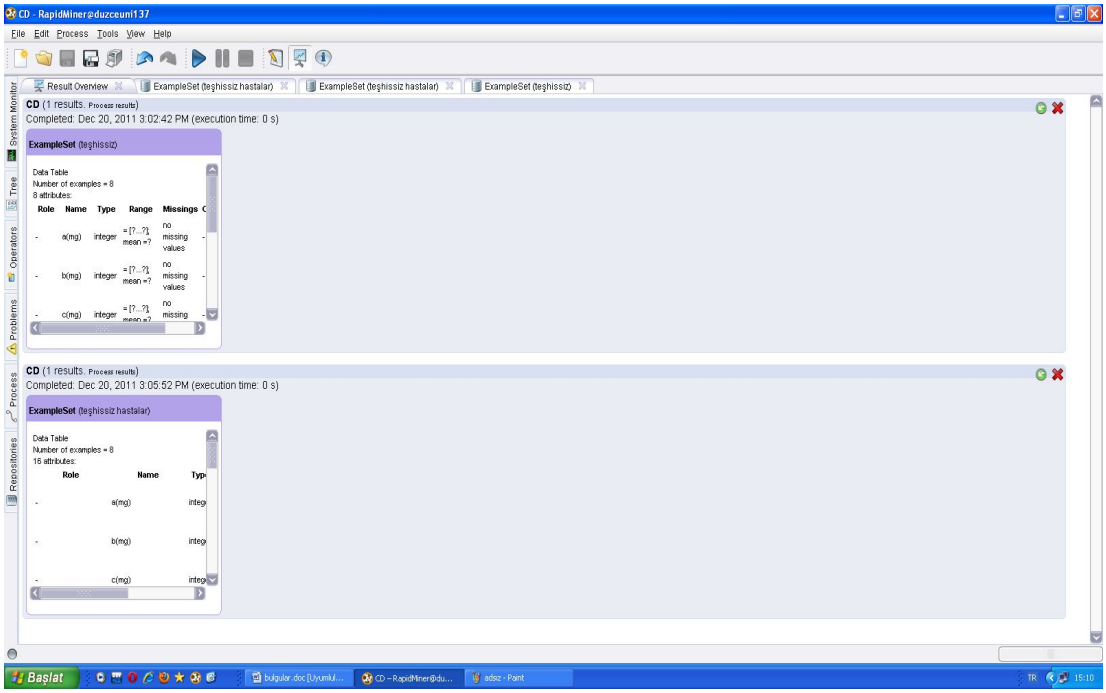


Şekil 4.13: RapidMiner'ın Repository Ekranı

Veritabanından istenen tablolar çekildikten sonra süreç (process) başlar. RapidMiner'ın zenginleştirilmiş görsel ara yüzleri yardımıyla, düşünülen proje "Process" ekranından, sürükle-bırak yöntemiyle tasarlanır. İstenen algoritma, "Operators" sekmesi altındaki "Modelling" sekmesinin sunduğu seçeneklerle seçilir. İstenen algoritma ve veritabanı tabloları sürüklenerek "process" ekranına getirilir ve birbirine bağlanır(Şekil 4.14). Daha sonra tasarlanan process çalıştırılır ve Şekil 4.15'deki görüntü alınır.

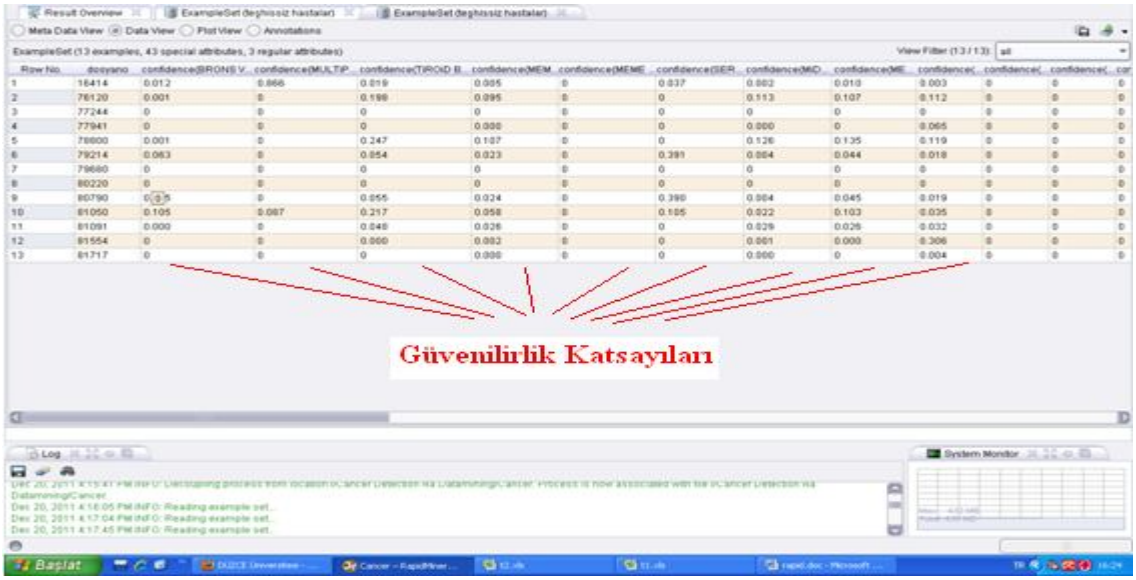


Şekil 4.14: RapidMiner'da modelleme

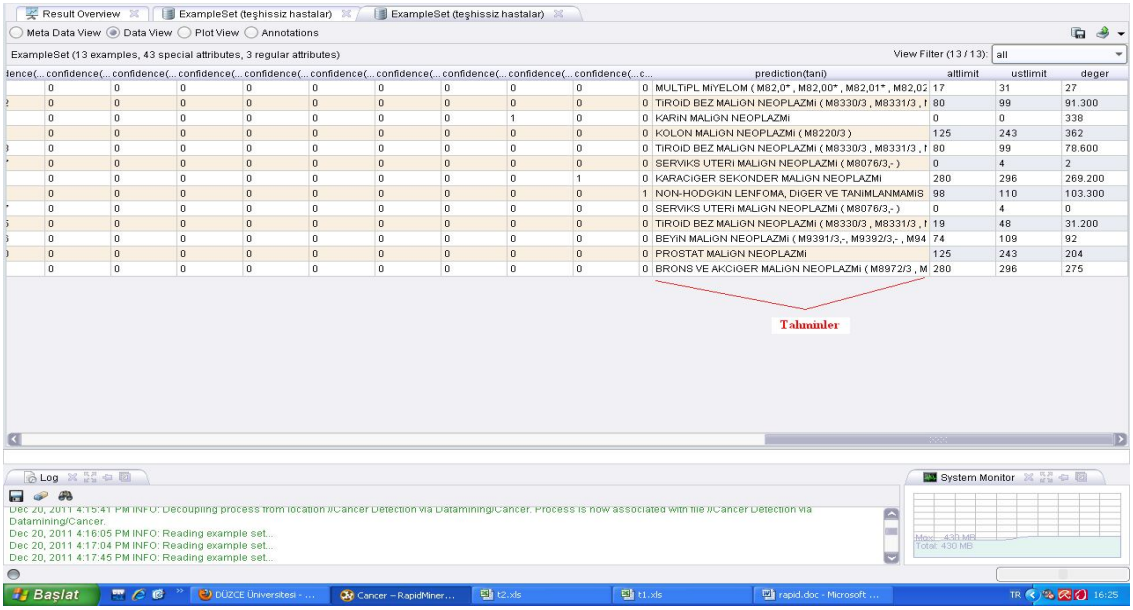


Şekil 4.15: Sonuç Gösterim Ekranı

Sonuç gösterim ekranından ilgili process'in sonucu daha da açılır ve her bir tahmin değeri (RapidMiner'a "label" olarak girilen değerler) için güvenilirlik katsayılarını içeren bir ekran alınır. Sonuçta her bir satır için bu güvenilirlik katsayılarına bakarak, en yüksek katsayıya sahip etiket hastanın kanser tahmini olarak verilir(Şekil 4.16-4.17).



Şekil 4.16: Her Tahmin İçin Güvenilirlik Değerleri



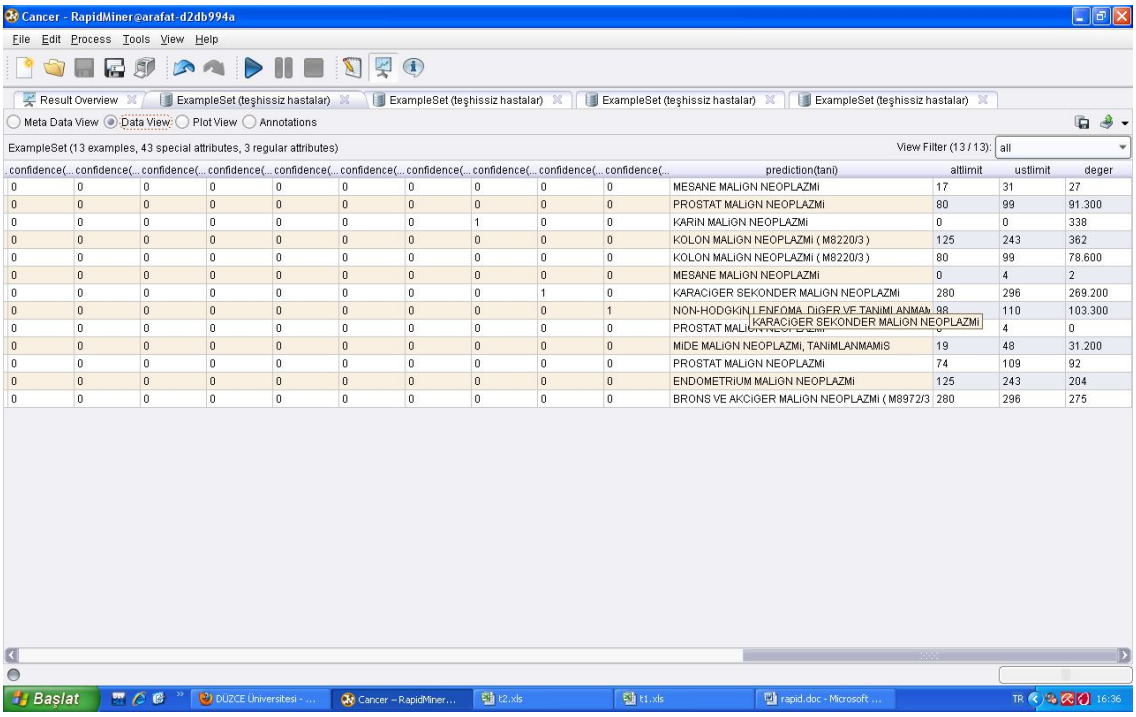
Şekil 4.17: Her Hasta İçin Tahmin Değerleri

5. TARTIŞMA VE SONUÇ

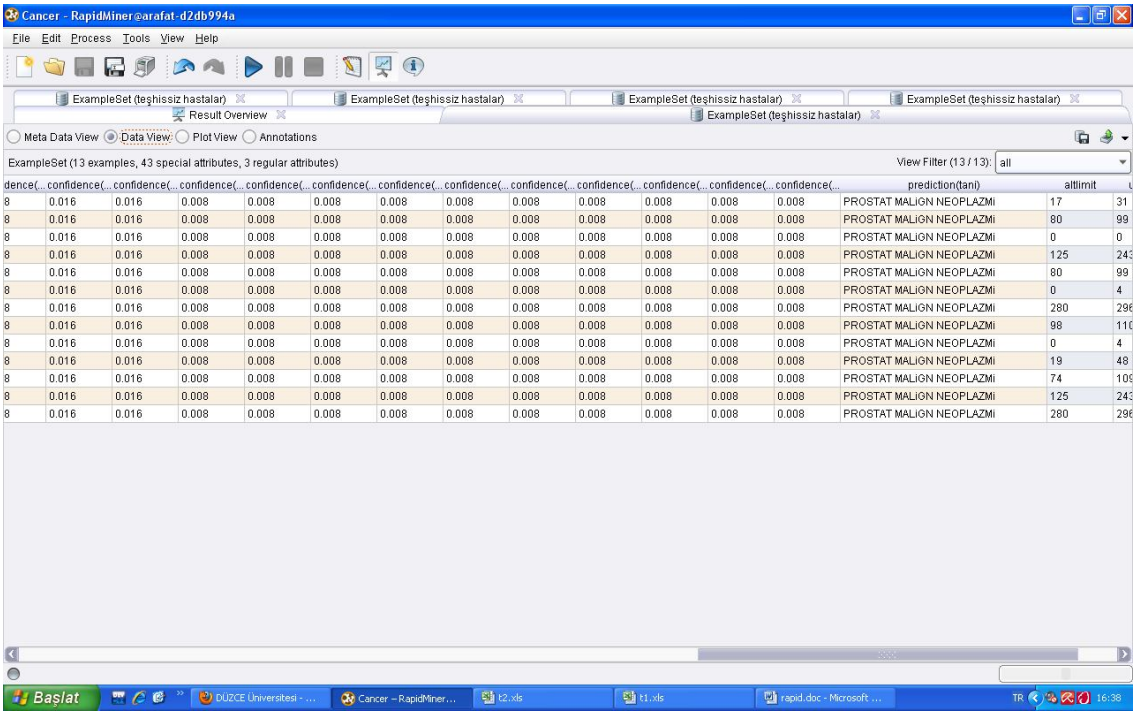
Günümüzde hızla artan veri yığınları arasında ve rekabetin, zaman kavramının giderek önem kazandığı bir çağda, hayli önem arz eden, ham veriden altın veriye ulaşma işi, son zamanlarda veri madenciliği teknikleriyle oldukça kolay ve güvenilir olarak yapılabilmektedir. Özellikle de tıp alanında, verilerin hem diğer alanlara göre daha fazla birikmesi hem de verilerin çözümlenmesinin yaşamsal derecede önemli olması sebebiyle veri madenciliği teknikleri oldukça yaygın olarak kullanılmaya başlamıştır.

Bu tez çalışmasında, günümüzde giderek yaygınlaşmakta olup, çoğunlukla ölümlerle sonuçlanan kanser hastalığının erken tanısına yönelik bir çalışma sunulmuştur. Amacın gerçekleşmesi için hastane bilgi sisteminden alınan hasta bilgileri kullanılmıştır. Daha önceden teşhisi konmuş olan kanser vakaları kullanılarak, seçilen tahmin modeline göre henüz tanısı koyulmamış hastalara, belli güvenilirlik oranlarına göre tanı koyulmuştur.

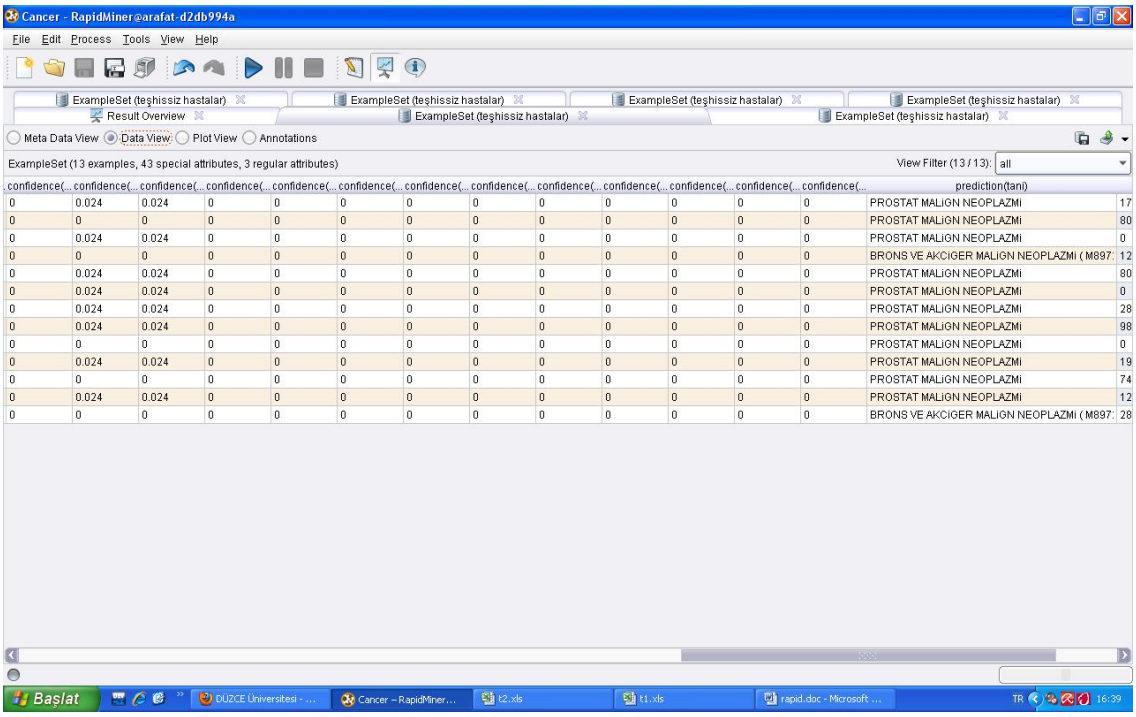
Seçilen modelin güvenilirliğini test etmek amaçlı olarak da kanser türü belli olan hastalar, tanıları bilinmiyormuş gibi modele dâhil edilmiş ve sistemin gerçeğe ne derece uyumlu olduğu gözlemlenmiştir. Bu gözlem sırasında farklı algoritmalar da kullanılarak hangi algoritmanın bu konuda daha etkin olduğu gösterilmiştir.



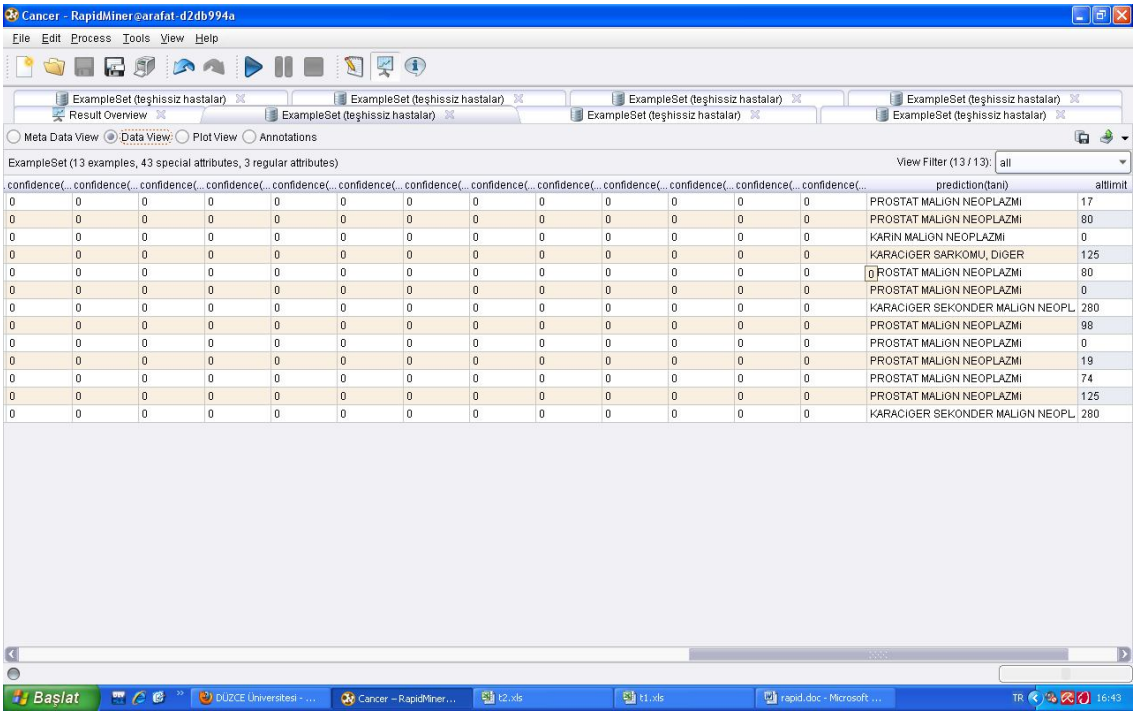
Şekil 5.1: K-En Yakın Komşu Algoritması İle Tahmin



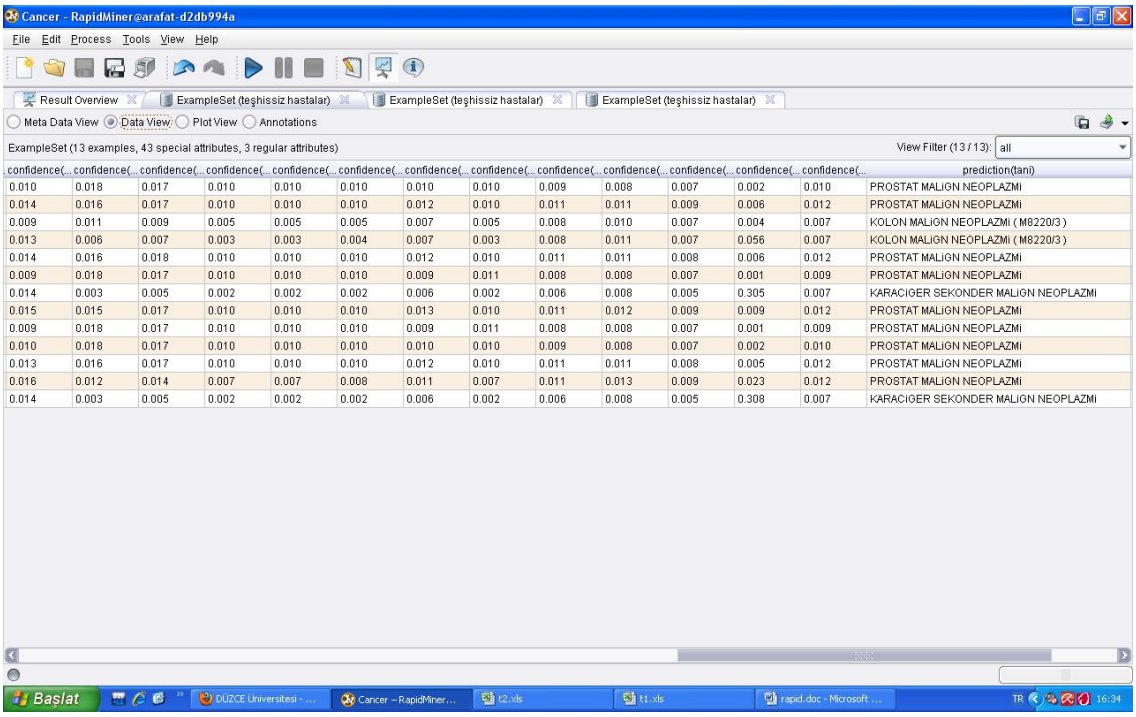
Şekil 5.2: Karar Ağacı İle Tahmin



Şekil 5.3: Birliktelik Kuralına Göre Tahmin



Şekil 5.4: Diskriminant Analizine Göre Tahmin



Şekil 5.5: Sinir Ağları Yaklaşımı ile Tahmin

Şekillerden de görüleceği üzere, tez çalışmasında güdülen amaç doğrultusunda en doğru sonuçlar k-en yakın komşu algoritması kullanılarak yapılan modellemeden elde edilmiş ve en yanlış sonuçlar ise karar ağaçları yöntemi ile alınmıştır.

KAYNAKLAR

BERRY, J.A., LINOFF, G., 1997, *Data Mining Techniques For Marketting, Sales and Customer Support*, John Willey & Sons, Inc., New York, 5.

COULTER, M.D., BATE, A., MEYBOOM R. H. B., LINDQUIST, M., EDWARDS, I. R., 2001, Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study, *BMJ*, 322, 1207-1209.

ERDEM, O. A. ve UZUN, E., 2005, Yapay Sinir Ağları ile Türkçe Times New Roman, Arial ve Elyazısı Karakterleri Tanıma, *Gazi Üniv. Müh. Mim. Fak. Dergisi*, 20, 1.

FAYYAD, U.M., PİATETSKY-SHAPIRO, G., SMYTH, P., 1996, From Data Mining To Knowledge Discovery: An Overview, *Advances in Knowledge Discovery and Data Mining*, 1-30

GANZERT, S.,GUTTMANN J., KERSTING, K., KUHLEN, R., PUTENSEN, C., SYDOW, M., KRAMER, S., 2002, Analysis of Respiratory Pressure-Volume Curves in Intensive Care Medicine Using Inductive Machine Learning, *Artificial Intelligence in Medicine*, 26, 69-86.

GÜRÜNLÜ, B. , 2009, *Veri Madenciliği Projelerinin Yaşam Döngüsü-1* [online], http://www.yazgelistir.com/Makaleler/1000002145.ygpx/print[Ziyaret Tarihi:2 Ağustos 2011].

HAND, D., 1998, Data Mining: Statistics and More?, *The American Statistician*, 52, 112-118.

HAND, D., MANILLA, H., SMYTH, P., 2001, *Principles of Data Mining*, MIT Press ,Cambridge, MA, 0-262-08290-X.

HONIGMAN, B., PARTICE, L., PULLING, R. M., BATES, D. W., 2001, A computerized method for identifying incidents associated with adverse drug events in outpatients, *International Journal of Medical Informatics*, 61, 21-32.

IŞIK, A., 2008, Veri Madenciliği, *Sızıntı*, 352 (8), 28-31.

İnternet kaynaklarından derleme: *"OLAP vs. OLTP"* , http://datawarehouse4u.info/OLTP-vs-OLAP.html

KOCABAŞ, K. , 2010, *Veri Madenciliği Sık Karşılaşılan Problemler* [online], http://www.misjournal.com/?p=2579 [Ziyaret Tarihi: 24 Ağustos 2011].

KOYUNCUGİL, A.S., 2006, *Bulanık Veri Madenciliği ve Sermaye Piyasalarına Uygulanması*, Doktora, Ankara Üniversitesi Fen Bilimleri Enstitüsü.

KÜÇÜKSİLLE, E., 2009, *Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve İMKB Hisse Senetleri Piyasasında Bir Uygulama*, Doktora, Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü.

MESUT, A., 2011,*Veri Tabanı Yönetimi* [online], Edirne, http://altanmesut.trakya.edu.tr/bmg/Ders10.ppt [Ziyaret Tarihi: 10 Temmuz 2011].

MITCHELL, T., 1997, *Machine Learning*, McGraw-Hill, USA, 0-07-042807-7.

ÖZDAMAR, E.Ö., 2002, *Veri Madenciliğinde Kullanılan Teknikler ve Bir Uygulama*, Yüksek Lisans, Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü.

ÖZMEN, S., 2001, İş Hayatı Veri Madenciliği İle İstatistik Uygulamalarını Yeniden Keşfediyor, *V.Ulusal Ekonometri ve İstatistik Sempozyumu,* 19-22 Eylül 2001, Çukurova Üniversitesi, Adana.

PONOMARENKO, J., MERKULOVA, T., ORLOVA, G., FOKIN, O., GORSHKOV, E., PONOMARENKO, M., 2002, Mining DNA sequences to predict sites which mutations cause genetic diseases, *Knowledge-Based Systems*, 15, 225-233.

"RapidMiner Video Tutorial", Rapid-I, http://rapid-i.com
RENCHER, A.C., 1995, *Methods of Multivariate Analysis*, John Wiley & Sons, Inc., USA, 0-471-41889-7.

SÜMERKENT, K. ,2010, *OLAP Sisteminin Özellikleri* [online], http://www.iszekam.net/author/Kadir%20S%C3%BCmerkent.aspx?page=2 [Ziyaret Tarihi: 25 Temmuz 2011].

SWIFT, R.S., 2001, *Accelerating Customer Relationships*, Prentice Hall, London,

THEARLING, K., *An Introduction to Data Mining* [online], http://www.thearling.com/dmintro/dmintro_frame.htm [Ziyaret Tarihi: 13Temmuz2011].

USLU, E. , 2008, *Veri Madenciliği Nedir ve Neden Veri Madenciliği* [online], http://www.yazgelistir.com/Makaleler/1000001858.ygpx [Ziyaret Tarihi: 2 Ağustos 2011].

UYAN, M., ÇAY, T., 2008, *Mekansal Uygulamalar için Veri Madenciliği Yaklaşımı* [online], http://www.uzalcbs2008.org/pdf/77.pdf [Ziyaret Tarihi: 10 Eylül 2011].

VELMURUGAN, T., SANTHANAM, T., 2011, A Comparative Analysis Between K-Medoids And Fuzzy C-Means Clustering Algorithms For Statistically Distributed Data Points, *Journal of Theoretical and Applied Information Technology*, 27(1), 19-30.
VAHAPLAR, A., İNCEOĞLU, M.M., 2001, Veri Madenciliği ve Elektronik Ticaret, *VII. Türkiye'de İnternet Konferansı*, 1-3 Kasım 2001.

Wikipedia, (2011). *Classification.* Available: http://en.wikipedia.org/wiki/Classification [Ziyaret Tarihi: 14 Eylül 2011].

YARALIOĞLU, K., 2011, *Veri Madenciliği* [online], İzmir, http://www.deu.edu.tr/userweb/k.yaralioglu/dosyalar/ver_mad.doc [Ziyaret Tarihi: 25 Temmuz 2011].

YILMAZ, E., 2006, *Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği İle Tüketici Profilinin Çıkarılmasına Yönelik Bir Uygulama,* Yüksek Lisans, Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü.

ZAIANE, O.R., FOSS, A., LEE, C.H., WANG, W., 2002, On Data Clustering Analysis: Scalability, Constraints and Validation, *Proc. of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02)*, May 2002 Taipei, Taiwan, 28-39.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Soyadı, Adı	: KARAPINAR ŞENTÜRK, Zehra
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 24.01.1987, Ankara
Medeni hali	: Evli
Telefon	: 0380 542 10 36 / 4661
Faks	: 0380 542 10 36
E-mail	: zehrakarapinar@duzce.edu.tr

EĞİTİM

Derece	Eğitim Birimi	Mezuniyet Tar.
Yüksek Lisans	Düzce Üniversitesi/Elektrik Eğitimi Bölümü	2011
Yan Dal	Atılım Üniversitesi/ Matematik Bölümü	2009
Lisans	Atılım Üniversitesi/Bilgisayar Mühendisliği Bölümü	2009
Lise	Mimar Sinan Lisesi	2004

İŞ DENEYİMİ

Yıl	Yer	Görev
2010-	Düzce Üniversitesi	Araştırma Görevlisi

YABANCI DİL

İngilizce (KPDS-Fen Bilimleri: 88)