T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

VERİ MADENCİLİĞİ YÖNTEMLERİ İLE KARDİYOVASKÜLER HASTALIK TAHMİNİ YAPILMASI

Yüksek Lisans Tezi

SERAP ERKUŞ

İSTANBUL, 2015

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

VERİ MADENCİLİĞİ YÖNTEMLERİ İLE KARDİYOVASKÜLER HASTALIK TAHMİNİ YAPILMASI

Yüksek Lisans Tezi

SERAP ERKUŞ

Tez Danışmanı: DOÇ. DR. M. ALPER TUNGA

İSTANBUL, 2015

T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİ TEKNOLOJİLERİ

Tezin Adı: Veri Madenciliği Yöntemleri ile Kardiyovasküler Hastalık Tahmini Yapılması
Öğrencinin Adı Soyadı: Serap ERKUŞ
Tez Savunma Tarihi: 07.01.2015

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğu Fen Bilimleri Enstitüsü tarafından onaylanmıştır.

Doç Dr. Nafiz ARICA
Enstitü Müdürü

Bu tezin Yüksek Lisans tezi olarak gerekli şartları yerine getirmiş olduğunu onaylarım.

Doç. Dr. M. Alper TUNGA
Program Koordinatörü

Bu Tez tarafımızca okunmuş, nitelik ve içerik açısından bir Yüksek Lisans tezi olarak yeterli görülmüş ve kabul edilmiştir.

Jüri Üyeleri	İmzalar
Tez Danışmanı Doç. Dr. M. Alper TUNGA	---
Üye Yrd. Doç. Dr. Batu SALMAN	---
Üye Yrd. Doç. Dr. Tevfik AYTEKİN	---

TEŞEKKÜR

Tüm tıbbi bilgi ve veri desteği için Uzm. Dr. M. Ertuğrul MERCAN'a, fikir aşamasından bitimine kadar bu tez çalışmasına desteğini esirgemeyen danışman hocam Doç. Dr. M. Alper TUNGA'ya, çalışma süresince ihmal etmek zorunda kaldığım ailem ve arkadaşlarıma ve attığım her adımda yanımda olan, tüm zorluklara karşı dimdik ayakta durabilmem için elinden geleni yapan sevgili eşime gösterdikleri sabır ve anlayış için teşekkürlerimi sunarım.

İstanbul, 2015 Serap ERKUŞ

ÖZET

VERİ MADENCİLİĞİ YÖNTEMLERİ İLE KARDİYOVASKÜLER HASTALIK TAHMİNİ YAPILMASI

Serap ERKUŞ

Bilgi Teknolojileri

Tez Danışmanı: Doç. Dr. M. Alper TUNGA

Ocak 2015, 62 sayfa

Bu çalışmada biyomedikal veriler incelenerek dünyanın bir numaralı ölüm sebebi olan kalp ve damar hastalıklarının erken teşhisine katkıda bulunabilecek başarılı bir model oluşturmak hedeflenmiştir.

Çalışmada kullanılan veri kümesi 604 kayıt içermektedir. Üç farklı yöntem kullanılarak referans değer aralıklarına göre dönüştürülen bu veri ile üç veri kümesi elde edilmiştir. Oluşan bu üç veri kümesi üzerinde nitelik seçim işlemleri ile belirlenen parametrelere, on sınıflandırma yöntemi uygulanmıştır. Veri kümeleri ve kullanılan algoritmaların başarı durumları incelenmiş ve bu incelemeyi desteklemek amacıyla diğer bazı performans ölçme metrikleri de kullanılarak en başarılı veri kümesi ve algoritma belirlenmiştir. Bu çalışma, oluşan modeli kullanarak hasta laboratuvar sonuçlarından otomatik olarak tanı üreten bir program yazılması ile geliştirilebilir.

Anahtar Kelimeler: Veri Madenciliği, Sınıflandırma, HNB Algoritması, Kardiyovasküler Hastalık Tahmini, Weka

ABSTRACT

CARDIOVASCULAR DISEASE PREDICTION USING DATA MINING TECHNIQUES

Serap ERKUŞ

Information Technologies

Thesis Supervisor: Doç. Dr. M. Alper TUNGA

January 2015, 62 page

In this study the main purpose is to build a successful model using a biomedical data set that will have a contribution to the diagnosis of cardiovascular disease which is the most common cause of death in the world.

The data set used in this study contains 604 records. This data set was transformed according to the reference ranges of parameters using three different methods. Three new data sets were obtained after the transformation process. Future selection methods were applied on each data set to get correct parameter group for modelling. Ten different classification techniques were applied to these data sets to build a model. The most successful data set and algortithm were detected by examining and comparing accuracy of the models. Additional performance evaluation metrics were also used to support the mentioned comparisons. This study can be improved by implementing an application for automatically diagnosis of cardiovascular disease using the model.

Keywords: Data Mining, Classification, HNB Algorithm, Cardiovascular Disease Diagnosis, Weka

İÇİNDEKİLER

TABLOLAR

ŞEKİLLER

KISALTMALAR

BKİ : Beden kitle indeksi

BMI : Body mass index (Beden kitle indeksi)

FN : Yanlış negatif

FP : Yanlış pozitif

HT : Hiper Tansiyon

KVH : Kardiyovasküler hastalık

TF : Doğru negatif

UA : Ürik Asit

TP : Doğru pozitif

VTBK : Veri tabanından bilgi keşfi

1. GİRİŞ

Bilgisayarların yaygınlaşması ile ortaya çıkan veri kavramı, gitgide günlük yaşantımızın içerisinde farkında olmadan kullandığımız bir araç haline gelmiştir. Veri başlangıçta yalnızca bilgisayar ortamında oluşturulup depolanmakta iken, günümüzde teknolojinin gelişmesi, internetin ve mobil cihazların hayatımızın her noktasına girmesi ile daha fazla alanda depolanabilen veriler üretilmeye başlanmıştır. Veri üretiminin artması veri depolama için alternatif ortamların aranmasına ve veri depolama yöntemlerinin gelişmesine sebep olmuştur.

Geçmişten bu güne hızla büyüyen veri, tek başına yalnızca bir anlam ifade ederken başka veriler ile birleştiğinde daha anlamlı hale gelir ve bilgiye dönüşür. Verinin her an her yerde üretilmesi, ayrıştırılması zor veri öbeklerinin oluşması problemini beraberinde getirir. Verinin bilgiye dönüşümü olmaksızın işleyen bir sistem, yalnızca veri kirliliği üretmekte ve veri sahibine hiçbir katkı sağlamamaktadır. Verilerin operasyonların işleyişine yönelik olarak depolanması, karar almaya yardımcı bilgiler ve analiz ihtiyacını karşılamak amacıyla veri ambarı yapısını ortaya çıkarmıştır.

Elde bulunan verilerden anlamlı bilgiler ortaya çıkarabilmek, geçmiş tecrübeleri analiz edip, gelecekte katma değer elde edebilmek mümkündür. İşte bu noktada veri madenciliği farklı analiz yöntemleri ile gerçekte var olan fakat ilk bakışta görülemeyen bilgilerin gün ışığına çıkarılmasına olanak sağlamaktadır.

Veri Madenciliği, veri tabanından bilgi keşfi (VTBK) sürecinde, veri örüntülerinin ortaya çıkarılması için akıllı yöntemlerin uygulanması aşamasıdır (Camacho ve Borges 2005). Bu keşif süreci, bilgi teknolojilerinin doğal evriminin bir sonucu olarak görülmektedir (Han ve Kamber 2006). Günümüzde birçok özel ve kamu kuruluşu veri madenciliğini süreçlerinin bir parçası haline getirmektedir.

Veri madenciliğinin kullanım alanları incelendiğinde, birbirinden farklı birçok sektörde karar alma sürecinin kaynağında yer aldığı görülmektedir. Telekomünikasyon,

pazarlama ve reklam, bankacılık, sigortacılık, perakendecilik gibi alanlarda çeşitli amaçlarla kullanılan veri madenciliği, tıp alanında da kullanılmaya başlanmıştır.

Günümüzde neredeyse tüm tıbbi cihazlar sayısal veriler üretmekte, bu nedenle tıp alanında var olan analiz ihtiyacına çözüm olarak veri madenciliği büyük umut vaat etmektedir. Birikmiş, anlamsız gibi görünen tahlil, görüntüleme, ilaç kullanımı vb. verilerin otomatik olarak işlenip yorumlanması, tıbbi araştırmalarda klasik yöntemler ile bulunması imkânsız olan sonuçlara ulaşmayı mümkün hale getirmiştir.

Hastalıkların erken teşhisi, hastalığa sebep olan etmenlerin tespiti, ilaç yan etkileri, tedavi yönteminin belirlenmesi gibi çalışmalar veri madenciliğinin tıpta kullanılabileceği alanlara örnek olarak verilebilir (Karahoca ve Tunga 2012, Koyuncugil ve Özgülbaş 2009, Frawley ve diğ. 1992).

Dünya Sağlık Örgütü (WHO) verilerine göre kalp ve damar hastalığı ölüm sebeplerinin en başında gelmektedir. 2012 yılı içerisinde 17,5 milyon kişi kalp ve damar hastalığı sebebiyle hayatını kaybetmiştir. Bu da 2012 yılı boyunca yaşanan ölümlerin yüzde otuzuna denk gelmektedir (WHO 2012). Kalp ve damar hastalıklarının bu denli yüksek oranda ölüm sebebi olması bu alanda yapılması gereken araştırmalar olduğunu açıkça ortaya sermektedir.

Öte yandan hastalar üzerinde yapılan tam kan sayımı gibi laboratuvar sonuçları birçok hastalık için klasik teşhis yöntemi olarak kullanılmaktadır. Bu verilerin elektronik ortamda saklanmaya başlanması ile veri madenciliği yöntemlerini uygulamak mümkün olmuştur. Kalp ve damar tıkanıklığı tespiti için günümüzde tanı yöntemi olarak anjiyografi kullanılmaktadır. Anjiyografi işleminin pahalı ve birçok yan etkisinin olması, araştırmacıları kardiyovasküler hastalık (KVH) tanısı için veri madenciliği kullanma konusunda motive etmiştir (Alizadehsania ve diğ. 2013).

Bu çalışmada, biyomedikal veriler incelenerek dünyanın bir numaralı ölüm sebebi olan kalp ve damar hastalıklarının erken teşhisine katkıda bulunabilecek başarılı bir model oluşturmak hedeflenmiştir.

Çalışma içerisinde, KVH tanısı konulmuş hastalara ait bazı laboratuvar sonuçları ile henüz KVH tanısı konulmamış hastalara ait bazı laboratuvar sonuçları birleştirilmiş ve bir veri tabanı oluşturulmuştur. Verilerin kalitesi incelenmiş ve gerekli veri temizliği yapılmıştır. En yüksek doğruluk oranına ulaşabilmek için, parametre sayısal değerlerinin anlamsal ifadelere dönüştürülmesi sırasında üç farklı gruplama mantığı esas alınmıştır. Buna göre her bir parametrenin bu üç farklı grup için referans aralıklarına göre grup ifadeleri oluşturulmuştur. Model oluşturulabilecek en doğru parametre grubunun belirlenebilmesi için nitelik seçme yöntemleri kullanılmıştır. Bu parametre grubuna birçok veri madenciliği yöntemi uygulanarak sonuçlar elde edilmiş, içlerinden en yüksek doğruluk oranına sahip model sonuçları Kalp ve damar hastalıkları uzmanı Dr. M. Ertuğrul MERCAN ile birlikte incelenmiş ve sonuçları tartışılmıştır. Çalışmada doktor görüşüne başvurduğumuz her konuda görüşü alınan kişi Uzm. Dr. Ertuğrul MERCAN'dır.

2. LİTERATÜR TARAMASI

2.1 VERİ MADENCİLİĞİ

Günlük yaşantımıza dair ürettiğimiz birçok verinin hayatımızı kolaylaştırmak için kullanıldığı bir teknoloji döneminde yaşamaktayız. Ürettiğimiz bu dağınık veri kümeleri içerisinde gizlenmiş birçok bilgi keşfedilmeyi beklemektedir. Müşteri memnuniyetini arttıracak bir çözümün ya da bir hastalığın sebebinin eldeki veri dağının içerisinde gizli olduğunu bilmek kurumları veri madenciliğine yönlendirmektedir.

Veri madenciliği ile büyük veri tabanı içinde yapılan araştırmalar sonucu elde edilen gizli örüntüler, ilgileşim ve bağımlılık ilişkileri geleneksel bilgi toplama yöntemleri ile göz ardı edilebilir. Örneğin; rapor oluşturma, grafik üretme, kullanıcı sorgulama, karar destek sistemleri vb. Veri madenciliği işlemi sırasında kullanıcılar tarafından sorulan soruların otomatik olarak cevaplanmasına yardımcı olabilecek çok çeşitli algoritma yelpazesine sahip bir araç kullanılır. Bu araç içerisindeki her bir model sezgisel, açıklanması kolay, anlaşılır ve kullanımı kolaydır (Gargano ve Raggad 1999).

2.1.1 Veri Madenciliği Tanımı

Bugün veri madenciliği olarak bildiğimiz çalışma alanının bir disiplin haline gelmesi 1990'lı yılların başlarına dayanmaktadır (Hong 1997).

Veri madenciliğinin büyük veri ile başa çıkabilen yüksek performanslı algoritmaları, paralel hesaplama yetenekleri, diskteki yerleşik verilere etkin erişimi gibi kullanışlı özellikleri duyuruldukça, kullanımı yaygınlaşmaya başlamıştır (Hong 1997).

Bugüne dek veri madenciliğinin pek çok tanımı yapılmıştır. Bunlardan bazıları şöyledir:

Veri madenciliği;

a. Anahtar değişkenlerin binlerce potansiyel değişkenden izole edilmesini sağlama yeteneğidir (Kitler ve Wang 1999).

b. İşlenmemiş veriden tek başına elde edilemeyen bilginin, veri analiz süreci ile ortaya çıkarılmasıdır (Jacobs 1999).

c. Bir bilgisayar programı yardımı ile büyük veri yığınları içerisinde bulunan gelecekle ilgili tahminleme yapabilmemize olanak sağlayan ilişkilerin ve bilginin aranması işidir (Doğan ve Türkoğlu 2008).

d. İstatistik, veri tabanı teknolojisi, örüntü tanıma, makine öğrenme gibi alanları temel alan, veri sahiplerine çıkar ve değer elde edebilmeleri için büyük veri içerisinden önceden tahmin edilemeyecek ilişkilerin tespit edilmesini sağlayan analiz işlemidir (Hand 1998).

Gorunescu (2011) veri madenciliği tanımına dair bu eşdeğer yaklaşımları aşağıdaki maddelerle ifade etmiştir:

Veri madenciliği;

a. Büyük veri tabanları içerisindeki örüntülerin istatistik, yapay zekâ ve örüntü tanıma için kullanılan hesaplama teknikleri ile otomatik olarak aranmasıdır.

b. Eldeki veriden, var olduğu bilinmeyen, gizli ve yararlı olma ihtimali olan bilginin çıkarılmasıdır.

c. Büyük veri kümeleri ya da veri tabanlarından işe yarar bilgi çıkarma bilimidir.

d. Büyük ölçekli verinin anlamlı örüntüler / ilişkiler keşfetmek için incelenmesi ve inceleme işleminin otomatikleştirilerek analiz edilmesidir.

e. Veri içerisindeki gizli örüntü ve ilişkilerin tanımlanması ve bilginin otomatik olarak keşfedilmesi sürecidir.

f. Veri ile fırsat yaratmaktır.

Veri madenciliği, “samanlıkta iğne aramak” deyiminde geçen arama işleminin hızlandırılması ve otomatikleştirilmesi için bir metal detektörü kullanılmasına benzetilebilir (Gorunescu, 2011).

Veri madenciliğinin amacı, bazı bilgi alanları içerisindeki çoğunlukla denetlenemeyen büyük miktardaki veriyi mantıklı hale getirmektir. Bu tanımda bahsedilen “mantıklı hale getirmek” ifadesi, veriye kullanıcı deneyimlerine dayanan farklı anlamlar katılabilmeyi anlatmaktadır. Anlamlı hale gelmiş olması ve veri sahiplerinin keşfedilen bilgiyi avantaja çevirebilmesi için elde edilen yeni bilginin anlaşılır, geçerli, yeni ve faydalı olması gibi bazı temel nitelikleri taşıması gerekmektedir (Cios ve diğ. 2007).

2.1.2 Bilgi Keşfi Sürecinde Veri Madenciliği

Veri tabanından bilgi keşfi (VTBK) olarak da bilinen bilgi keşfi süreci bazı veri alanlarında yeni bir bilginin aranması işlemleri olarak tanımlanmaktadır (Cios ve diğ. 2007).

Verinin saklanış ve erişim biçimi, yüksek ölçekli veri üzerinde bir analiz yapabilmek için algoritmaların verimli ve ölçülebilir bir şekilde nasıl kullanılacağı, sonuçların nasıl yorumlanacağı ve görselleştirileceği, insan ve makine etkileşiminin nasıl modelleneceği ve destekleneceği VTBK sürecinin birer parçasıdır ve bu bağlamda VTBK bilginin elde edilme sürecinin tamamı ile ilgilidir (Cios ve diğ. 2007).

Verinin hazırlanması, temizlenmesi, elde edilmiş diğer bilgilerle birleştirilmesi, en uygun veri madenciliği yönteminin uygulanması ve veri madenciliği sonuçlarının doğru yorumlanması, VTBK sonucunda faydalı bilgi elde edilmesi için izlenmesi gereken en önemli adımlardır (Fayyad ve Stolorz 1997).

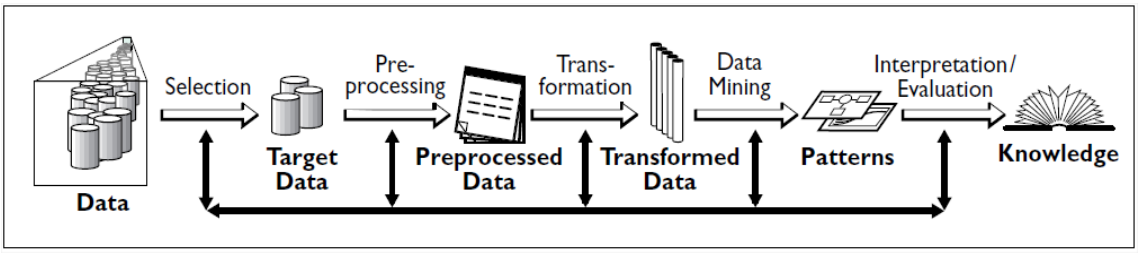
Veriden anlamlı bir bilgi elde etmeye çalışmadan önce bilgi keşfi süreci genel yaklaşımının anlaşılması gerekir. Başarılı bir veri madenciliği çalışması için yalnızca veri analizinde kullanılan algoritmaları bilmek yeterli değildir (Cios ve diğ. 2007).

VTBK sürecini Fayyad ve diğ. (1996a) aşağıdaki gibi aşamalandırmışlardır;

a. Bilgi keşfi süreci uygulanacak verinin anlaşılması
b. Veri kümesinden odaklanılacak kısmın seçilmesi
c. Gürültü temizleme, eksik verilerin tamamlanması ya da süreç dışı bırakılması gibi önişleme ve veri temizliği işlemleri
d. Hedefe ulaşmak için kullanılacak parametrelerin ya da verinin azaltılması, düzenlenmesi
e. Model oluşturmak için kullanılacak olan veri madenciliği yönteminin belirlenmesi
f. Uygulanacak veri madenciliği algoritmasının seçilmesi
g. Veri madenciliğinin uygulanması
h. Model sonucu elde edilen bilginin yorumlanması
i. Keşfedilen bilginin kullanılması (Fayyad ve diğ. 1996a)

VTBK süreci Fayyad ve diğ. tarafından Şekil 2.1'de görüldüğü gibi ifade edilmiştir;

Şekil 2.1: Veri tabanından bilgi keşfi süreci



Kaynak: Fayyad U., Piatetsky-Shapiro G. ve Smyth P., 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of The Acm. 39 (11), ss.27-34.

Veri madenciliği aşaması süreç içerisinde çok önemli bir yere sahiptir. Veri madenciliği için bir program kullanılır ve program içerisinde birçok veri madenciliği algoritması barındırır. İhtiyaca uygun algoritma seçimi ile veriden anlamlı bilgi elde etme işlemi bu aşamada gerçekleştirilir (Pal ve Jain 2005).

2.1.3 Veri Madenciliği Aşamaları

Veri madenciliği, VTBK içerisinde bir aşama olmasına rağmen veri madenciliği olmadan bilgi keşfi yapılamamaktadır. Bu sebeple VTBK süreci veri madenciliğini ifade etmektedir. Bahsedilen VTBK süreçlerinin tamamı veri madenciliği süreçleri olarak da düşünülebilmektedir. Günümüzde birçok kaynakta VTBK süreci, veri madenciliği süreçleri olarak incelenmektedir (Gorunescu 2011, Olson ve Delen 2008).

VTBK sürecinde detaylı olarak gruplandırdığımız aşamalar sürecin doğru ve aynı kalitede uygulanabilmesi için bazı farklı standartlar oluşturularak incelenmiştir. Bu standartlardan en yaygın olarak SEMMA (Sample, Explore, Modify, Model, Assess) ve CRISP-DM (CRoss-Industry Standard Process for Data Mining) yöntem bilimleri (methodology) kullanılmaktadır.

2.1.3.1 SEMMA

SEMMA standardında, örnekleme (Sample), araştırma (Explore), değiştirme (Modify), modelleme (Model) ve değerlendirme (Assess) aşamaları yer almaktadır. Her bir aşamayı ifade eden kelimelerin baş harfleri birleştirilerek isimlendirilmiştir. Bu süreç veri madenciliği projesinin basit ve anlaşılır, aynı zamanda geliştirme ve yönetime uygun olmasını hedeflemektedir. Süreci oluşturan beş aşama aşağıda incelenmiştir;

Örnekleme (Sample), veri madenciliğinin temelini oluşturan verinin örneklenmesi, SEMMA sürecinin ilk aşamasıdır. Bu aşamada önemli bilgileri içerecek kadar büyük, hızlı işlem yapabilecek kadar küçük olmalıdır (Santos ve Azevedo 2005).

Araştırma (Explore), yeni fikir ve bakış açıları kazandırmak amacıyla veri kümesi içerisinde beklenmeyen eğilimler ve aykırılıkların (anomaly) arandığı aşamadır (Santos ve Azevedo 2005).

Değiştirme (Modify), modelleme aşamasına hazırlık amacıyla verinin seçilmesi, oluşturulması ve dönüştürülmesi işlemlerinin yapıldığı aşamadır (Santos ve Azevedo 2005).

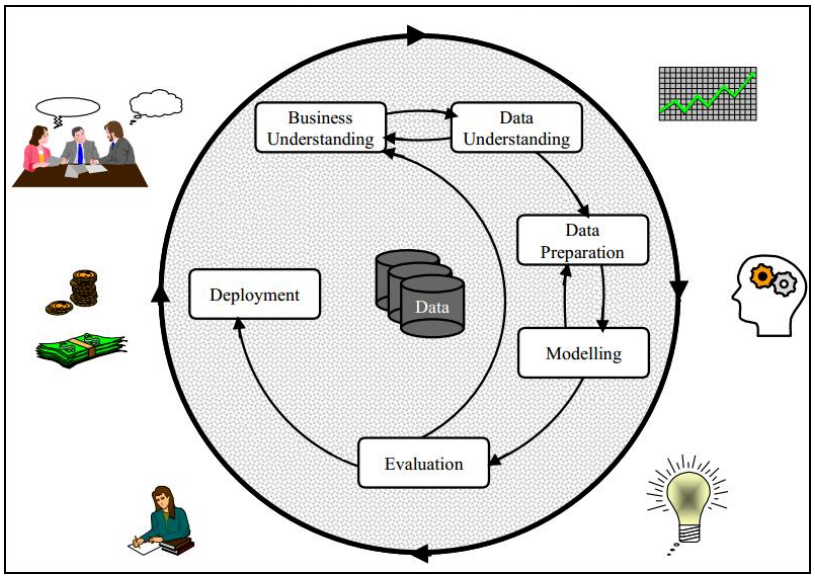
Modelleme (Model), veri madenciliği algoritmalarının veriye uygulanabilmesini sağlayan yazılımlar yardımıyla, hazırlanan veri kümesi ile güvenilir bir model oluşturulur (Santos ve Azevedo 2005).

Değerlendirme (Assess), oluşan modelin yararlılığı ve güvenilirliği değerlendirilir ve performansı gözden geçirilir (Santos ve Azevedo 2005).

2.1.3.2 CRISP-DM

1997'de SPSS, NCR, Daimler-Chrysler ve OHRA firmaları birleşerek bir konsorsiyum oluşturmuşlar ve 1999 yılında sanayiler arası standart veri madenciliği süreci (CRoss-Industry Standard Process for Data Mining) adı ile veri madenciliği projeleri için standart bir veri madenciliği yöntem bilimi (methodology) ve süreç modelini kamuoyuna sunmuşlardır (Anonymous 1999). Şekil 2.2'de CRISP-DM süreci gösterilmiştir.

Şekil 2.2: CRISP-DM süreci



Kaynak: Chapman P., 1999,The CRISP-DM User Guide, http://lyle.smu.edu/~mhd/8331f03/crisp.pdf. [erişim tarihi 28 Ekim 2014]

Chapman ve diğ. (2000) tarafından belirlenen CRISP_DM süreci aşamaları aşağıda incelenmiştir.

İşin anlaşılması (Business Understanding), proje hedeflerinin ve ihtiyaçlarının iş birimi bakış açısıyla anlaşılmaya çalışılması ve veri madenciliği için sonuç aranacak bir problem şekline dönüştürülmesi ve hedefe ulaşmak amacıyla planlanması aşamasıdır.

Verinin anlaşılması (Data understanding), veri toplamaya ve veriyi anlamaya yönelik işlemlerin yapıldığı aşamadır. Bu aşamada veri kalitesi incelenir, veri ile ilgili ilk görüşler elde edilir, gizli bilgileri keşfedebilmek için ilk hipotezler oluşturulur ve ilginç alt kümeler saptanmaya çalışılır.

Verinin hazırlanması (Data preparation), ham verinin veri madenciliği yöntemleri uygulanabilecek bir veri kümesine dönüştürülmesi işlemlerini kapsar. Bu aşamada tablo, kayıt ve nitelik seçimleri, veri temizliği, dönüştürme, birleştirme ve indirgeme işlemleri yapılarak veri modelleme için hazırlanır. (Larose 2005)

a. Verinin seçilmesi, modellemede kullanılacak tablo, kayıt ya da niteliklere karar verilir (CRISP 2011).

b. Veri temizliği, modelin başarısına negatif yönde etki edebileceğini düşündüğümüz verilerin belirlendiği ve bu negatif etkiyi pozitife çevirebilmek için yapılabilecek işlemlerin uygulandığı aşamadır.

Veri kümesi içerisindeki eksik veya uygun olmayan veriler gürültü (noise) olarak isimlendirilir. Veri kümesinin gürültüden arındırılabilmesi için gürültü olarak kabul edilen verileri içeren kayıtları silinebilir, gürültü yerine sabit bir değer yazılabilir, niteliğin ortalaması olacak şekilde yeni bir değer verilebilir ya da uygun tahmin yöntemleri kullanılarak olabilecek değer belirlenip gürültü yok edilebilir (Özkan 2013).

c. Veri inşası, eldeki veriler kullanılarak yeni niteliklerin oluşturulması işidir. Örneğin; en ve boy niteliklerinden alan isimli yeni bir niteliğin oluşturulması (CRISP 2011).

d. Veri birleştirme, farklı kaynaklardan elde edilen verilerin bir araya getirilmesi ya da aynı nesne ile ilgili farklı bilgileri içeren tabloların birleştirilmesi gibi işlemlerin yapılmasını ifade eder (Chapman ve diğ. 2000).

e. Verinin biçimlendirilmesi, veri içerisindeki niteliklerin analiz edilebilecek formata göre düzenlenmesi işlemidir (CRISP 2011).

Modelleme (Modeling), işin anlaşılması aşamasında belirlenen hedefler ışığında veri madenciliği tekniklerinin uygulandığı ve modelin oluşturulduğu aşamadır. Bu aşama kendi içinde modelleme tekniği seçimi, test verisi hazırlanması, modelin oluşturulması ve modelin değerlendirilmesi olarak dört kısımdan oluşur (Chapman ve diğ. 2000).

Bu aşamada modelleme teknikleri içerisinden kullanıldıkları amaçlara göre en uygun olan teknik seçilir. Kullanım amaçlarına göre algoritmalar tanımlayıcı ve tahminleyici modeller olmak üzere ikiye ayrılır.

a. Tanımlayıcı modeller (Descriptive models), bir veri kümesi içerisindeki her bir nesnenin birbiri ile olan ilişkilerinden, benzerliklerinden bağıntılar çıkarmamızı sağlayan modellerdir. Bir öğrenme süreci içermez.

b. Tahminleyici modeller (Predictive models), elde edilmiş olan veriden, tecrübelerden ya da bir başka deyişle geçmişten, gelecek ile ilgili bir sonucun tahmin edilmesi için kullanılan modellerdir. Bir öğrenme süreci ile başlar ve öğrenme işlemi sırasında kazanılan yetenek, hedef veri kümesi için uygulanarak tahminleme gerçekleştirilir.

Değerlendirme (Evaluation), modelin işin anlaşılması aşamasında belirlenen amaçlar açısından iş sahipleri ile birlikte değerlendirilmesi işlemi bu aşamada yapılır. Süreç gözden geçirilir ve bir sonraki aşamada veri madenciliği süreçlerinden hangisinin uygulanacağına karar verilir (CRISP 2011, Chapman ve diğ. 2000).

Sonuçlandırma (Deployment), modelin uygulamaya geçmesi aşamasıdır. Model kurum içerisindeki problemi gidermek için ilgili yere konumlandırılır ve proje hayata geçirilmiş olur. Bakım ve gözlemleme işlemleri için planlama yapılır ve proje sonuçları raporlanır (CRISP 2011).

2.1.4 Veri Madenciliği Fonksiyonları

Veri madenciliği fonksiyonlarını denetimli / eğiticili (supervised) ve denetimsiz / eğiticisiz (unsupervised) olmak üzere iki kısımda inceleyebiliriz.

2.1.4.1 Denetimli / eğiticili (supervised) veri madenciliği

Denetimli / eğiticili veri madenciliğinde bir öğrenme süreci vardır. Bu öğrenme süreci parametrelerin önceden bilinen ilişkilerine ya da çalışmanın hedefine bağlı olarak hazırlanır. Genellikle tahminleyici modeller ile sonuçlanır (Oracle 2014).

Denetimli model bir eğitim süreci içerir. Bu eğitim sürecinde model, hedef değere ulaşılabilmek için eldeki verilerin niteliklerini inceleyerek mantığı öğrenir ve bu mantığı tahmin yapmak için kullanır. Modelin başarısı tahminin kalitesi ile ölçülür (Argüden ve Erşahin 2008).

2.1.4.2 Denetimsiz / eğiticisiz (unsupervised) veri madenciliği

Denetimsiz / eğiticisiz veri madenciliğinde isminde de yer aldığı gibi bir öğrenme süreci yoktur. Model oluşurken algoritmayı yönlendirebilecek önceden bilinen sonuçlar yoktur. Genelde tanımlayıcı modellerle sonuçlansa da tahminleyici de olabilirler (Oracle 2014).

Bu veri madenciliği fonksiyonlarında başlangıçta tanımlanmış sınıflar olmadığı için, verilerin davranışları incelenerek birbirine yakın olanlar gruplanır ve sınıflar model çıktısında elde edilir.

2.1.5 Veri Madenciliğinde Kullanılan Yöntemler

Veri madenciliğinde kullanılan yöntemler sınıflandırma, kümeleme, birliktelik kuralları olmak üzere üç ana kategoriye ayrılmaktadır.

2.1.5.1 Sınıflandırma (Classification)

Sınıflandırma algoritmaları denetimli / eğiticili (supervised) algoritmalardır ve sonuç olarak bir tahmin yapılır. Sınıflandırma işlemi öğrenme ve sınıflandırma olmak üzere iki aşamadan oluşur.

a. Öğrenme: Verinin tamamına ait örnekler bulunacak şekilde basitleştirilmiş bir kısmi veri kümesi oluşturulur (eğitim kümesi – training set) ve bu veri sınıflandırma algoritması ile analiz edilir.

Algoritma uygulanan öğrenme verisinde bulunan her kayıt sınıf bilgisi ile birlikte birçok başka nitelikten oluşur. Algoritma sınıf bilgisi ve diğer nitelikler arasındaki ilişkileri saptar ve tüm veri üzerinde uygulamak üzere bir model oluşturur.

b. Sınıflandırma: Birinci adımda oluşturulan model sınıflandırma için kullanılır. Modelin veri kümesine uygulanması sonucu elde edilen sınıf bilgileri veri kümesinde bulunan sınıf bilgileri ile karşılaştırılarak tahminleme doğruluğu hesaplanır. Eğer doğruluk oranı kabul edilebilir oranda ise model gelecekteki verilerde sınıf bilgisini elde etmek için kullanılabilir (Han ve Kamber 2006).

Kullanılan sınıflandırma algoritması tarafından oluşturulan modelin başarısını ölçmek için doğruluk (accuracy), hata oranı (error rate), hassaslık (sensitivity) ve özel etken oranı (specificity) kullanılır.

Doğruluk; doğru sınıflandırılmış kayıt sayısının toplam kayıt sayısına bölünmesi ile elde edilir.

Hata oranı; 1-doğruluk ile hesaplanır.

Hassaslık ve özel etken oranları Tablo 2.1'de görülen karışıklık matrisi üzerinden bulunur.

Tablo 2.1: Karışıklık matrisi (Confusion matrix)

		Öngörülen sınıf		
		C_1 (Pozitif)	C_2 (Negatif)	
Gerçek Sınıf	C_1 (Pozitif)	Doğru Pozitif TP	Yanlış Negatif FN	Pozitif P
	C_2 (Negatif)	Yanlış Pozitif FP	Doğru Negatif TN	Negatif N

Kaynak: Cios K. J., Pedrycz W., Swiniarski R. W. & Kurgan L. A., 2007. A Knowledge discovery approach, New York:Springer Science+Business Media

C_1 (Pozitif): Pozitif sınıf etiketi

C_2 (Negatif): Negatif sınıf etiketi

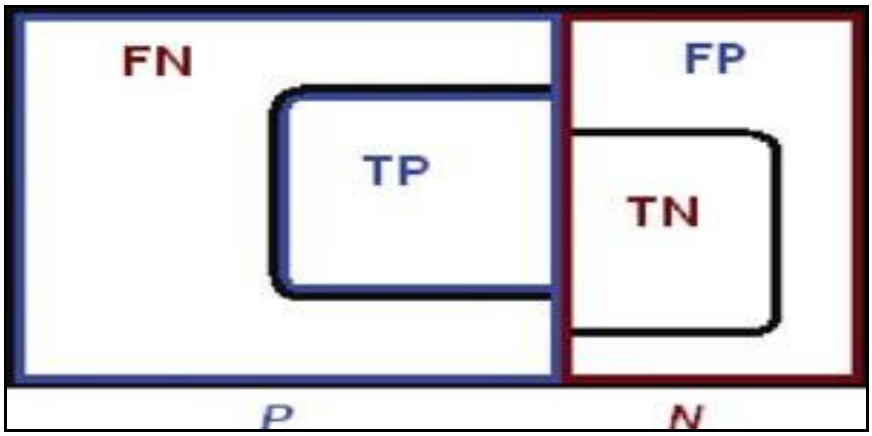
Doğru Pozitif (True Positive TP) sınıflandırıcı tarafından doğru etiketlenmiş olan kayıt pozitif etiket sayısını ifade ederken, Doğru Negatif (True Negative TN) ise sınıflandırıcı tarafından doğru etiketleme yapılmış negatif etiket adedini gösterir. Aynı şekilde Yanlış Pozitif (False Positive FP), yanlış etiketlenen pozitif etiket sayısını, Yanlış Negatif (False Negative FN) ise yanlış etiketlenen negatif etiket adedini ifade eder (Han ve Kamber 2006).

Hassaslık (sensitivity) = TP / (TP + FN)

Özel etken oranı (specificity) = TN / (FP + TN)

Cios ve diğ. (2007) tüm veri üzerinde yapılmış etiketlemeleri Şekil 2.3'de gösterildiği gibi anlaşılır bir biçimde ifade etmişlerdir.

Şekil 2.3: Sınıflandırma sonucu etiketleme durumu gösterimi



Kaynak: Cios K. J., Pedrycz W., Swiniarski R. W. & Kurgan L. A., 2007. A Knowledge discovery approach, New York:Springer Science+Business Media

Sınıflandırma yöntemlerini kullanılan algoritmalara göre beşe ayırabiliriz (Albayrak 2012);

a. Karar ağaçları (Decision Trees): ID3, C4.5, CART karar ağacı algoritmalarına örnek olarak verilebilir (Han ve Kamber 2006).

b. Bellek tabanlı yöntemler (Instance based): k-en yakın komşu algoritmaları örneklerdendir (Kollios 2005).

c. Bayes sınıflandırıcı (Bayes Classifier): Naive Bayes en temel algoritmasıdır (Albayrak 2012).

d. Yapay sinir ağları (Artificial Neural Networks)

e. Genetik algoritmalar (Genetic Algorithms)

2.1.5.2 Kümeleme (Clustering)

Gruplanmamış veriler içerisinden birbirine yakın ilişkide olanların gruplanması işlemidir. Bu işlem herhangi bir öğrenme olmaksızın tamamen benzerlik kullanılarak yapılmaktadır. Veri kümesi içerisindeki küme adedi esnektir.

Sınıflandırma, veri üzerinde sınırlı bir dizi kategori ya da sınıf tanımlamayı hedefleyen benzerlik tanımlayıcı bir araçtır (Jain ve Dubes 1988).

Sınıflandırma, heterojen bir veri grubundaki birbirine benzeyen elemanların kümelere ayrılarak, homojen alt veri grupları oluşturulması işlemidir (Argüden ve Erşahin 2008).

Kümeleme sürecinin başarılı olarak kabul edilebilmesi için sınıf içi benzerlik ve sınıflar arası farklılaşma en üst seviyede olmalıdır. Bu noktada şu sonuca varılmaktadır (Gorunescu 2011);

a. Aynı küme içerisindeki nesneler birbirine daha çok benzerdir.
b. Farklı kümeler içerisindeki nesneler birbirine daha az benzerdir.

Kümeleme algoritmalarında farklılıkları ve benzerlikleri saptayabilmek için genellikle mesafe ölçme teknikleri kullanılır. Kümeleme algoritmaları aşağıdaki gibi beş farklı grupta incelenebilir (Han ve Kamber 2006);

a. Bölümlerine ayırma yöntemleri (Partitioning): k-means ve k-medoids algoritmaları örnek verilebilir (Han ve Kamber 2006).
b. Hiyerarşik yöntemler (Hierarchical) : DIANA, AGNES algoritmaları örneklerdendir (Albayrak 2012).
c. Yoğunluk tabanlı yöntemler (Density-based): DBSCAN, OPTICS, DENCLUE, CLIQUE algoritmaları örnek olarak verilebilir (Kollios 2005).
d. Izgara tabanlı yöntemler (Grid-based): STING, WaveCluster ve CLIQUE örnek algoritmalardır. CLIQUE algoritması hem yoğunluk hem ızgara tabanlı olarak kabul edilmektedir (Kollios 2005).
e. Model tabanlı yöntemler (Model-based): COBWEB, SOM, EM algoritmaları model tabanlı algoritmalardır (Han ve Kamber 2006, Kollios 2005).

2.1.5.3 Birliktelik kuralları (Association rules)

Büyük veri kümeleri içerisindeki ilginç bağımlılık ya da ilişkilerin keşfedildiği, denetimsiz / eğiticisiz (unsupervised) veri madenciliği yöntemleridir (Cios ve diğ. 2007).

Bir veri kümesi içerisinden, destek (support) ve güven düzeyi (confidence) önceden tanımlanmış eşik değerlere göre geçerli sayılabilecek tüm kuralların keşfedildiği veri madenciliği yöntemidir (Gorunescu 2011).

Birliktelik kuralları, geçmiş verilerin analiz edilerek, eş zamanlı gerçekleşmiş nitelikler arasındaki ilişkilerin ortaya çıkarılması işidir (Çıngı 2008).

Birliktelik kuralları büyük veri yığınları içerisinden çıkardığı kurallar ile çapraz pazarlama ve satış, katalog tasarımı gibi birçok önemli noktada iş kararı alınmasına yardımcı olur. Birliktelik kurallarını anlayabilmek için en iyi örnek olarak, pazar sepet analizi verilebilir. Örneğin; bir marketten süt satın alan müşteriler aynı alışverişlerinde ekmek de satın alıyorlar mı? Bu sorunun cevabına göre süt ve ekmek reyonları yan yana yerleştirilebilir (Yin ve diğ. 2011).

Birliktelik kuralları tek boyutlu ve çok boyutlu olarak ikiye ayrılabilir. Örneğin; tek boyutlu birliktelik kuralında iki ürünün birlikte satın alınması durumu incelenebilirken, çok boyutluda bir ürünün hava durumu, yer ve gün gibi birden fazla niteliğe göre satın alınma durumu incelenir. Birliktelik kuralı çalışmalarında Apriori ve FP-Growth algoritmaları kullanılabilir.

2.1.6 Veri Madenciliği Kullanım Alanları

Günlük yaşantımızın her aşamasının, veri üretir hale gelmesiyle, bu verilerin kurumlar ve insanlar yararına kullanılacak şekilde konumlandırılmasına neden olmuştur. Veri madenciliği işte bu sebeple birçok sektörde kullanılmaya başlanmıştır. Özellikle pazarlama alanında tüm sektörlere yönelik çalışmalarda kullanılmaktadır.

Finans ve sigorta sektörü bilgiye dayalı yönetime yüksek oranda ihtiyaç duyan sektörlerdir. Müşteri kazanımı ve mevcut müşterinin elde tutulması, müşteri memnuniyeti sağlanması, çapraz pazarlama ve satış, alternatif satış kanallarının oluşturulması, maliyetlerin azaltılması, risk değerlendirme, kayıp ve kaçakların engellenmesi gibi konular için veri madenciliği kullanımı oldukça uygundur.

Haberleşme sektörü için müşteri kaybı büyük bir sorun teşkil etmektedir. Müşteri kaybını engellemek için kişisel paketler/ürünler çıkarılması, rakiplerin kampanyalarının takip edilmesi ile stratejiler geliştirilmesi gibi yeni müşteri kazanımı ve müşteriyi elde tutma için de veri madenciliği büyük avantajlar sağlamaktadır.

Sağlık sektörü veri madenciliği kullanımı açısından belki de en önemli sektördür. Doğru ve zamanında karar alabilmenin hasta sağlığı üzerindeki etkileri tartışılmaz derecede önem taşımaktadır. Klinik araştırmalar ve hastanelerde oluşan hasta verileri, uygulanan tedaviler ve tedavi sürecindeki veriler hastalıkların risk faktörlerinin belirlenmesi, başarılı tedavi sonuçları için etmenlerin tespit edilmesi, yeni tedavi yöntemlerinin geliştirilmesi ya da tedavi yöntemlerinde yaşanan sorunların giderilmesi, ilaç etkileşimleri ve yan etkileri gibi hayati önem taşıyan konularda veri madenciliği uygulanabilir. Bunların yanında veri madenciliği hastanelerdeki servislerin başarısı, kaynak kullanımı, hastalara ve hastalıklara eğilim tahminleri gibi konularda da hastane yönetimine yardımcı olacak şekilde konumlandırılabilir.

Kamu kurumlarında da ticari kuruluşlar gibi vatandaşlara özel hizmet sunabilmek amacıyla veri madenciliği uygulamaları kullanılabilir. Kaynakların doğru kullanımının sağlanması ve planlanması, vergi ile ilgili güveni kötüye kullanımların tespit edilebilmesi, yolsuzlukların belirlenebilmesi gibi konular için veri madenciliği kullanılabilir. Emniyet birimlerinde suç eğiliminin tespit edilmesi ve suç engelleme politikalarının oluşturulması gibi alanlarda veri madenciliği uygulanabilir (Işıklı 2009).

Bunların dışında eğitim, bilimsel çalışmalar, taşımacılık, e-ticaret, sosyal medya alanlarında veri madenciliği uygulamaları yapılması mümkündür.

2.2 TIP ALANINDA VERİ MADENCİLİĞİ

2.2.1 Tıp Alanında Veri Madenciliği Uygulamaları

Sağlık kuruluşlarında oluşan verilerin, elektronik ortamda tutulmaya başlanması ve veriye erişimin kolaylaşması ile bu verilerin birçok sağlık probleminin bilinmeyenleri için cevap arama çalışmalarında kullanılması gündeme gelmiştir. Hastalıkların sebepleri, risk faktörleri, tedavi yöntemleri, kullanılan ilaçlar ve etkileri, hasta laboratuvar ve demografik verileri bir araya geldiğinde, içlerinden faydalı bilgi elde etmek zorlaşmakta ve bu büyük veri içerisinde bir çok değerli bilgi kaybolmaktadır. Biriken bu büyük veriden hayati önem taşıyan bilgiler ancak veri madenciliği yöntemlerinin yetenekleri kullanılarak elde edilebilir.

Tıp alanındaki ilk veri madenciliği uygulaması 1854 yılında John Snow tarafından kağıt kalem ile yapılmıştır. Londra'da başlayan kolera salgınında, bir harita üzerinde ölenlerin konumlarını işaretleyerek ilk kümeleme çalışmasını yapmış, bu sayede ölümlerin belli bölgelerde toplandığını fark etmiş ve su pompalarından kaynaklanan salgının giderilmesini sağlamıştır (Jacquez ve diğ. 1996).

Delen ve diğ. (2005), göğüs kanseri hastalarının verilerine veri madenciliği teknikleri uygulayarak, hastaların ölüm ve hayatta kalma ihtimalini tespit eden bir model geliştirmişlerdir.

Özekes (2006), doktora tezi çalışmasında ileri örüntü tanıma ve görüntü işleme yöntemlerini kullanarak mamografi görüntülerindeki kitlelerin ve akciğer BT görüntülerindeki nodüllerin tespit edilmesine yardımcı olan bilgisayar destekli tespit yazılımları ve teknikler araştırılmıştır.

Demirel (2008) yüksek lisans tezinde, meme kanseri hastalarının verisi üzerinde veri madenciliği teknikleri uygulayarak bir model elde etmiş ve onkoloğa meme kanseri hastalarına uygulanması gereken tedavi yöntemleri konusunda yardımcı olacak bir yazılım geliştirmiştir.

Patil ve diğ. (2011) tarafından yazılmış olan makalede 180 yanık hastasının 2002 – 2006 yılları arasındaki verileri üzerinde sınıflandırma teknikleri uygulanarak, hastaların hayatta kalma durumuna ilişkin yüksek oranlarda tahminleme yapabilen model oluşturulmuştur.

Doğan (2007) yüksek lisans tezi çalışmasında, biyokimya verilerinden seçilen parametreleri temel alarak, kalp krizi (miyokard enfarktüsü), hiperlipidemi, demir eksikliği anemisi ve hipertiroidi-hipotiroidi hastalıklarının teşhisinde kullanılacak bir karar destek sistemi tasarlanmasını sağlamıştır.

Tahminciler (2014) yüksek lisans tezinde, bir internet sitesinde yapılan yorumları çeşitli araçlar ile okuyarak bir veri tabanı oluşturmuş, kural tabanlı metin ayrıştırma, eşleştirme ve çeşitli veri madenciliği teknikleri uygulayarak Erythromycin ilacı yan etkileri üzerine bir araştırma yapmıştır.

Kumar ve diğ. (2011) yaptıkları araştırmada toplanan verilerden desen (pattern) çıkararak diyabet, hepatit ve kalp hastalıklarında hekime yardımcı olacak akıllı tıbbi karar destek sistemleri geliştirilmiş ve bazı algoritmaların etkinlikleri karşılaştırılmıştır.

Sağlık hizmeti veren kurumların yönetimsel faaliyetleri açısından veri madenciliği yöntemleri uygulanarak çeşitli kazanımlar sağlanmıştır. İngiltere'de St. George Hastanesi'nde yapılan bir veri madenciliği çalışması ile yoğun bakım ünitesinden çıktıktan sonra yaşamını yitiren hastaların, yüzde otuz dokuzunun 48 saat daha yoğun bakımda tutularak ölüm riskinin ortadan kaldırılabileceği tespit edildi (SPSS 2014).

San Francisco Kalp Enstitüsü'nde mali yapının güçlenmesi, hasta bakım kalitesinin arttırılması, hastaların hastanede kalış sürelerinin kısaltılması, çalışan performanslarının arttırılması amacıyla veri madenciliği çalışmaları başlatılmış ve sonucunda oluşan modeller amaçların gerçekleştirilmesini sağlamıştır (SPSS 2014).

Boehringer Ingelheim İtalya dünyadaki önemli ilaç şirketlerinden biridir. Eczanelere yönelik farklılaştırılmış satış politikaları üretmek ve en iyi müşterilerini belirleyebilmek

için bir veri madenciliği çalışması başlatmıştır. Oluşan sınıflandırma modeli sayesinde karlı müşterilerin izlenmesi, en karlı eczaneler hedef listesinin oluşturulması, müşteri ilişkilerinin doğru yönetilmesi ve pazarlama faaliyetlerinin etkinliğinin değerlendirilmesi gibi kazanımlar elde edilmiştir (SPSS 2014).

2.2.2 KVH İle İlgili Veri Madenciliği Çalışmaları

2.2.2.1 Çalışma 1

Chaurasia ve Pal (2013) "Early Prediction of Heart Diseases Using Data Mining Techniques" isimli makalede 303 satır içeren veri üzerinde CART(Classification and Regression Tree), ID3 (Iterative Dichotomized 3) ve DT (Decision Table) algoritmalarını uygulamışlardır.

Veri içerisindeki yaş, cinsiyet, göğüs ağrısı tipi, dinlenme kan basıncı, serum kolesterol, açlık kan şekeri, elektrokardiyografi sonucu, maksimum kalp atışı, egzersiz eğim tipi parametreleri kullanılmıştır. Sonuçta elde edilen üç model ile ilgili bilgiler Tablo 2.2 üzerinde görülmektedir.

Tablo 2.2: Çalışma 1 model sonuçları

	Sınıflandırıcı		
	CART	ID3	DT
TP (Doğru Pozitif)	0.99	0.85	0.98
FP (Yanlış Pozitif)	0.98	0.82	0.98
TN (Doğru Negatif)	0.02	0.18	0.02
FN (Yanlış Negatif)	0.01	0.15	0.02
Doğruluk oranı (accuracy)	**83.49%**	**72.93%**	**82.50%**

Kaynak: Chaurasia V., Pal S., 2013. Early Prediction of Heart Diseases Using Data Mining Techniques. Caribbean Journal of Science and Technology. 1. ss.208-217

Tablo 2.2 incelendiğinde, çalışmada yalnızca tree sınıflandırma yöntemlerinin kullanıldığı görülmektedir. CART algoritması, ID3 ve DT algoritmalarına göre daha

başarılıdır. Aynı zamanda ID3 algoritması kullanılan diğer yöntemlerin yüzde onun altında bir başarı elde edebilmiştir.

2.2.2.2 Çalışma 2

Mishra ve diğ. (2013) tarafından "Novel Approach to Predict CARDIOVASCULAR DISEASE using Incremental SVM" adlı bir çalışma yapılarak, sınıflandırma yöntemi ile KVH tahmini yapılmak istenmiştir.

Bu çalışmada UCI web sitesinden içerisinde 14 parametre (yaş, cinsiyet, göğüs ağrısı tipi, dinlenme kan şekeri, serum kolesterol, açlık kan şekeri, elektrokardiyografi sonucu, maksimum kalp atışı, egzersiz eğim tipi, floroskopi ile renklendirilen damar sayısı, sınıf) ve 303 kayıt içeren bir veri kümesi indirilmiştir.

Veri kümesi üzerinde SVM(Support Vektor Machine) ve ISVM(Incremental Support Vektor Machine) sınıflandırma algoritmaları uygulanmıştır. Çalışma sonucu elde edilen modeller ile ilgili sonuçlar Tablo 2.3'de verilmiştir.

Tablo 2.3: Çalışma 2 model sonuçları

	Hassaslık (Sensitivity)	Özel etken (Specivicity)	**Doğruluk (Accuracy)**
SVM	90.01%	77.22%	**84.12%**
ISVM	92.21%	78.92%	**89.39%**

Kaynak: Mishra B.K., Lakkadwala P. ve Shrivastava N.K., 2013. Novel Approach to Predict Cardiovascular Disease Using Incremental SVM, 2013 International Conference on Communication Systems and Network Technologies, ss.55-59.

Bu çalışmada fonksiyon sınıflandırma yöntemleri kullanılmıştır. ISVM yönteminin SVM ile karşılaştırıldığında daha başarılı bir model oluşturduğu ortaya koyulmuştur.

2.2.2.3 Çalışma 3

Amini ve diğ. (2013) Esfahan Al‑Zahra and Mashhad Ghaem Hastanesi'nde 2010 – 2011 arasında 50 risk faktörüne ait veriler toplanarak elde edilmiştir. 807 adet veri elde edilmiş ve bu veriye K‑en yakın komşu (K‑nearest neighbor) ve C4.5 karar ağacı (C4.5 decision tree) algoritmaları uygulanmıştır. Çalışma sonucunda C4.5 karar ağacı algoritması ile elde edilen model yüzde 95.42 doğruluk oranı (accuracy) verirken, k-en yakın komşu algoritması ile yüzde 94.18 doğruluk oranı (accuracy) elde edilmiştir.

2.2.2.4 Çalışma 4

Yeh ve diğ. (2011) tarafından kalp ve damar hastalıkları teşhisi konulmuş 493 kayıt içeren bir veri kümesine veri madenciliği sınıflandırma algoritmaları uygulanmıştır. Kalp ve damar hastalığı çeşidi tahmin edebilen bir model elde etmek amacıyla başlatılan bu çalışmada, veriler aşağıda belirtilen şekilde üç farklı veri kümesi olarak bölünmüştür.

Veri kümesi 1: hastalık teşhis verileri (sınıf – hastalık çeşidi) + fiziksel muayene verileri + kan testi sayısal verileri

Veri kümesi 2: hastalık teşhis verileri (sınıf – hastalık çeşidi) + fiziksel muayene verileri

Veri kümesi 3: hastalık teşhis verileri (sınıf – hastalık çeşidi) + kan testi sayısal verileri

Çalışma sonucu elde edilen modellere ilişkin bilgiler Tablo 2.4'de görülebilir.

Tablo 2.4: Çalışma 4 model sonuçları

Algoritma	Veri kümesi	Hassaslık (Sensitivity)	Doğruluk oranı (Accuracy)
Decision tree	1	95.29%	98.01%
Bayesian classifier	1	87.10%	91.30%

BPNN	1	94.82%	97.87%
Decision tree	2	94.68%	98.01%
Bayesian classifier	2	86.30%	91.36%
BPNN	2	93.80%	98.05%
Decision tree	3	62.81%	66.93%
Bayesian classifier	3	66.60%	71.83%
BPNN	3	64.21%	69.32%

Kaynak: Yeh D. Y., Cheng C. H., Chen Y. W., 2011. A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications. 38 (7). ss.8970–8977.

Fiziksel muayene sonuçları ile oluşturulan model yüzde doksan sekiz başarılı olurken aynı veri kümesine kan testi verileri eklendiğinde başarı oranı değişmemektedir.

Model oluşturulurken yalnızca kan testi verisi kullanıldığında ise kalp ve damar hastalığı çeşidini tahmin etme başarısı yüzde yetmiş iki civarında kalmıştır.

Bu da çalışmada kullanılan veri kümesindeki kan testi verilerinin model başarısına etki etmediğini açıkça ifade etmektedir.

2.2.2.5 Diğer çalışmalar

Shouman ve diğ. (2012) tarafından yapılan çalışmada farklı veri kümeleri üzerinde daha önce yapılmış olan veri madenciliği çalışmalarına yer verilmiştir.

Sonuçları belirtilen çalışmalarda farklı veri kümeleri kullanıldığından doğruluk oranlarını birbiri ile kıyaslamak doğru olmayacaktır.

Şekil 2.4'de çalışmalarında belirtilmiş olan doğruluk oranlarına ait bilgiler görülebilir.

Şekil 2.4: Farklı veri kümelerine uygulanmış veri madenciliği çalışmaları

Author	Year	Technique	Accuracy
Yan, et al.	2003	Multilayer Perceptron	63.6%
Andreeva, P.	2006	Naïve Bayes	78.563 %
		Decision Tree	75.738 %
		Neural network	82.773 %
		Kernel density	84.444 %
Palaniappan, et al.	2007	Naïve Bayes	95%
		Decision Trees	94.93%
		Neural Network	93.54%
De Beule, et al.	2007	Artificial neural network	82%
Tantimongcolwata, et al.	2008	Direct kernel self-organizing map	80.4%
		Multilayer Perceptron	74.5%
Hara, et al.	2008	Automatically Defined Groups	67.8%
		Immune Multi-agent Neural Network	82.3%
Sitar-Taut, et al.	2009	Naïve Bayes	62.03%
		Decision Trees	60.40%
Rajkumar, et al.	2010	Naive Bayes	52.33%
		KNN	45.67%
		Decision list	52%
Srinivas, et al.	2010	Naïve Bayes	84.14%
		One Dependency Augmented Naïve Bayes classifier	80.46%
Kangwanariyakul, et al.	2010	Back-propagation neural network	78.43%
		Bayesian neural network	78.43%
		probabilistic neural network	70.59%
		linear support vector machine	74.51%
		polynomial support vector machine	70.59%
		radial basis function kernel support vector machine	60.78%
Anbarasi, et al.	2010	Genetic with Decision tree	99.2%
		Genetic with Naïve Bayes	96.5%
		Genetic with Classification via clustering	88.3%

Kaynak: Shouman M., Turner T., Stocker R., 2012. Using data mining techniques in heart disease diagnosis and treatment. Japan-Egypt Conference on Electronics, Communications and Computers ss.189-193

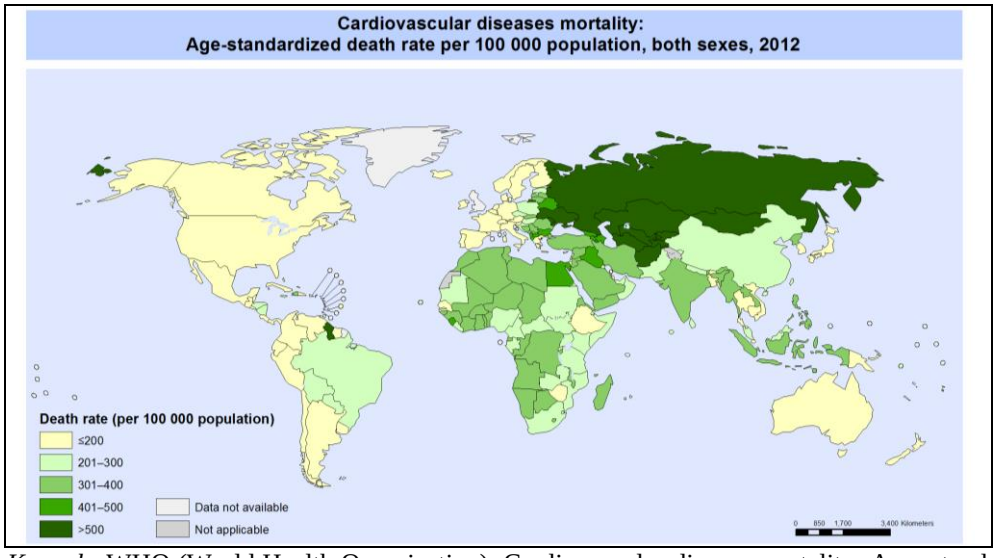
3. KARDİYOVASKÜLER HASTALIKLAR

3.1 KARDİYOVASKÜLER HASTALIK TANIMI VE ÖNEMİ

Kardiyovasküler hastalık (Kalp ve Damar Hastalığı; KVH), kalp (kardiyo) ve / veya vücutta bulunan kan damarları (vasküler) sistemini olumsuz etkileyebilen hastalıkları kapsayan genel bir terimdir (Lab Test Online 2011).

Dünyada ölüm sebeplerinin en başında yüzde otuz oranla KVH gelmektedir. 2012 yılı içerisinde 17.5 milyon kişi kalp ve damar hastalığı sebebiyle hayatını kaybetmiştir. Bunlardan 7.4 milyonu iskemik kalp hastalığı sebebiyle ölürken, 6.7 milyon kişi de felç sebebiyle ölmüştür. Şekil 3.1'te 2012 yılı dünya üzerindeki KVH sebebiyle ölüm dağılımı görülebilir (WHO 2014a).

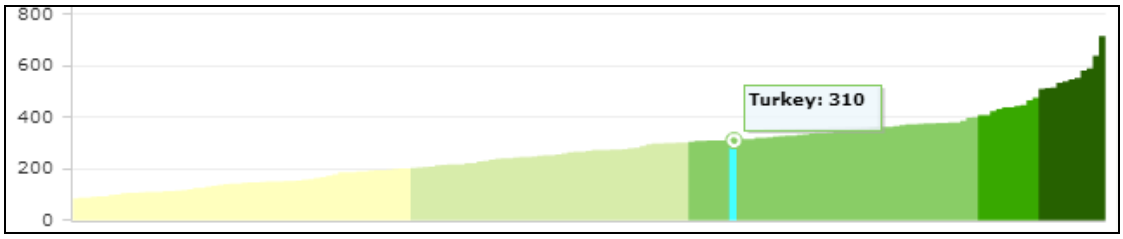
Şekil 3.1: 2012 yılı KVH sebebiyle ölümlerin, dünya üzerindeki dağılımı



Kaynak: WHO (World Health Organization), Cardiovascular disease mortality: Age- standardized death rate per 100.000 population both sexes 2012 , 2014. http://gamapserver.who.int/mapLibrary/Files/Maps/Global_NCD_mortality_CVD_2012.png. [erişim tarihi 31.10.2014]

Türkiye de KVH sebebiyle ölümlerin yüksek olduğu ülkelerden biridir. 2012 yılında Türkiye'deki KVH ölümü 31,000,000 ile diğer ülkeler arasında Şekil 3.2' de görülebileceği gibi yer almıştır (WHO 2014b).

Şekil 3.2: 2012 Türkiye KVH ölüm oranının diğer ülkeler arasındaki yeri



Kaynak: WHO (World Health Organization), Cardiovascular disease mortality, 2014. http://gamapserver.who.int/gho/interactive_charts/ncd/mortality/cvd/atlas.html. [erişim tarihi 31.10.2014]

Türkiye İstatistik Kurumu verilerine göre 2013 yılında, ölümlerin yüzde 39,8'inin sebebi KVH'tır. Şekil 3.3'da görüldüğü gibi kadınlarda bu oran 44,6'ya çıkmaktadır (TÜİK 2014).

Şekil 3.3: 2013 yılı ölümlerin cinsiyete göre dağılımı (%)

	Toplam	Erkek	Kadın
Toplam	**100,0**	**100,0**	**100,0**
Dolaşım sistemi hastalıkları	39,8	35,8	44,6
İyi huylu ve kötü huylu tümörler (malign ve benign neoplazmlar)	21,3	25,3	16,5
Solunum sistemi hastalıkları	9,8	10,7	8,8
Endokrin (iç salgı bezi), beslenme ve metabolizmayla ilgili hastalıklar	5,6	4,3	7,2
Dışsal yaralanma nedenleri ve zehirlenmeler	5,5	7,3	3,3
Sinir sistemi ve duyu organları hastalıkları	4,1	3,4	4,9
Diğer (enfeksiyon ve parazit hastalıkları, mental ve davranışsal bozukluklar, kas-iskelet sistemi ve bağ dokusunun hastalıkları vb.)	13,9	13,2	14,8

Not: Tablodaki rakamlar, yuvarlamadan dolayı toplamı vermeyebilir.

Kaynak: TUİK Türkiye İstatistik Kurumu. (2014). Ölüm nedeni istatistikleri 2013. http://www.tuik.gov.tr/PreHaberBultenleri.do?id=16162. [erişim tarihi 31.10.2014]

KVH nedeniyle gerçekleşen ölümlerin yüzde 38,8'ini iskemik kalp hastalığı oluştururken, yüzde 25,2'sini serebrovasküler hastalıklardan kaynaklanmıştır, Şekil 3.4'de diğer detayları görülebilir.

Şekil 3.4: 2013 yılı KVH sebebiyle ölümlerin cinsiyete göre dağılımı (%)

	Toplam	Erkek	Kadın
Dolaşım sistemi hastalıkları	**39,8**	**35,8**	**44,6**
Alt grupları	**100,0**	**100,0**	**100,0**
İskemik kalp hastalığı	38,8	45,8	32,0
Serebro-vasküler hastalık	25,2	22,9	27,5
Diğer kalp hastalığı	17,7	16,0	19,3
Hipertansif hastalıklar	12,8	10,0	15,6
Diğer	5,4	5,2	5,5

Not: Tablodaki rakamlar, yuvarlamadan dolayı toplamı vermeyebilir.

Kaynak: TUİK Türkiye İstatistik Kurumu. (2014). Ölüm nedeni istatistikleri 2013. http://www.tuik.gov.tr/PreHaberBultenleri.do?id=16162. [erişim tarihi 31.10.2014]

3.2 KARDİYOVASKÜLER HASTALIK ÇEŞİTLERİ

Kardiyovasküler hastalıklar kalp ve damar hastalıklarından oluşan bir grup hastalıktır. Bu hastalıklar WHO (2013) tarafından aşağıdaki gibi açıklanmıştır;

Koroner arter hastalığı (coronary artery disease); kalp kasını besleyen kan damarları hastalığıdır ve bu hastalığın sonucunda koroner kalp hastalığı (kalp krizi - coronary heart disease) ortaya çıkar.

Serebrovasküler hastalık (cerebrovascular disease); beyni besleyen kan damarları hastalığıdır, felç olarak bilinir.

Hipertansiyon (Hypertension); kan basıncının referans değerlere göre yüksek olmasıdır (high blood pressure).

Periferik arter hastalığı (peripheral arterial disease); kolları ve bacakları besleyen kan damarları hastalığıdır.

Romatizmal kalp hastalığı (rheumatic heart disease); kalp kası ve kalp kapakçıklarının streptokok bakterinin sebep olduğu romatizmal ateş sebebiyle zarar görmesidir.

Konjenital kalp hastalığı (congenital heart disease); doğuştan kalp yapısında bozukluk olmasıdır.

Derin ven trombozu (DVT) ve akciğer emboli (deep vein thrombosis and pulmonary embolism); DVT vücudumuzdaki derin bir vende (toplardamar) kan pıhtısı oluşmasıdır. Genelde alt bacak ya da uylukta ortaya çıkar. Bacak damarlarındaki kan pıhtıları yerinden çıkıp akciğerlere ya da kalbe gitmesi ile oluşur. (Hematoloji Uzmanlık Derneği 2014, WHO 2013)

Kalp krizi ve felç (heart attack and stroke), kan pıhtısı sebebiyle kalp ya da beyne kan akışının engellenmesi ile oluşan, akut (hızlı başlayan ve / veya kısa süreli) hastalıklardır. Bunun en yaygın nedeni kalp ve beyni besleyen kan damarlarının iç duvarlarında yağ birikintilerinin oluşmasıdır. Felç ise beyinde bulunan bir damarda kanama ya da bir pıhtı olması sebebiyle oluşabilir.

3.3 ÖNEMLİ KVH RİSK FAKTÖRLERİ

KVH ölümleri incelendiğinde, daha önce teşhis konulmamış ya da bir belirti (symptom) olmayan hastaların ölümlerin yarısından fazlasını oluşturduğu görülmektedir. Bu nedenle KVH risk faktörlerinin belirlenmesi büyük önem kazanmaktadır (Ni ve diğ. 2009). KVH risk faktörlerinin azaltılması hem KVH hem de bu sebeple ölümlerin azalmasını sağlamaktadır (Rothwell ve diğ. 2004).

3.3.1 Değiştirilemeyen Risk Faktörleri

a. Yaş: Yaşın ilerlemesiyle KVH artış göstermektedir. Erkeklerde 45 yaş ve üzeri, kadınlarda ise 55 yaş ve üzeri riskli olarak kabul edilmektedir (Thom ve diğ. 2006)

b. Cinsiyet: Genç erkeklerde KVH ve bu sebeple ölüm oranı kadınlardan 4-5 kat fazladır. 65 yaş altında erkeklerin felç olma riski kadınlara göre 2 kat daha fazladır (Rosengren ve diğ. 2009).

c. Aile öyküsü: Birinci derece akrabalarda kadınlarda 65 erkeklerde 55 yaş öncesi KVH tanısı konulmuş olması, gelecekteki KVH için öngörücü olarak kabul edilmektedir. Fakat günümüzde KVH yatkınlığı için klinik pratikte kullanımı kabul edilmiş bir tarama testi yoktur (Gülel 2012).

3.3.2 Değiştirilebilen risk faktörleri

a. Sigara: KVH için ana risk faktörlerinden biri sigara kullanımıdır ve KVH saptanmış kişilerin sigara içmeye devam etmesi ölüm oranlarını arttırmaktadır.(Abacı 2011)

b. Hipertansiyon (yüksek tansiyon): Dünya genelinde oldukça yaygın bir sağlık problemidir. Kan basıncının 140/90 mmHg eşit veya yüksek olması hipertansiyon göstergesidir. Hipertansiyonun neden olduğu hastalıklar içerisinden en sık rastlananlar ise KVH ve felçtir (Ercan Toptaner 2013). Hipertansiyon tedavisi ile kardiyovasküler risk azalmaktadır (Turkiyedoktorlari.com 2014).

c. Diyabet (şeker hastalığı): KVH başlıca risk faktörlerinden biri de diyabettir. Diyabet hastalıklarında en önemli ölüm sebebi KVH'tır (Ercan Toptaner 2013).

d. Hiperlipidemi (kan yağlarının yüksekliği) : Kan yağlarından en az birinin artmasıdır. Trigliserid ve kolesterol olmak üzere iki çeşit kan yağı vardır (Saglikpark.com 2008). Yapılan tüm araştırmalarda hiperlipidemi ile KVH arasında bir ilişki çıkması sonucu önemli ve düzeltilebilir KVH faktörlerinden biri olarak kabul edilen hiperlipidemi diyet, egzersiz ve ilaç tedavisi ile kontrol altına alınabilir (Turkiyedoktorlari.com 2014).

e. Obezite: Yağ kitlesinin yağsız vücut kitlesine göre yüksek olmasına obezite denir. Beden Kitle İndeksi (BKİ – Body Mass Index - BMI) obezite tespiti için kullanılan en

bilinen yöntemdir. BKİ değerinin 30 üzerinde olması KVH riskini artırır. BKİ hesaplaması aşağıdaki gibi yapılmaktadır (Ercan Toptaner 2013).

BKİ = Kilo (kg) / Boy² (m)

Kalp ve damar hastalarında yukarıda belirtilen başlıca risk faktörlerinden birkaçının aynı anda bulunması dikkat çekmektedir. Bahsedilen risk faktörlerinden üç ya da daha fazlasının bir hastada şans eseri bir arada bulunma ihtimali, üç ya da daha fazla risk faktörüne sahip KVH tanısı konmuş hastalar ile karşılaştırıldığında, ihtimale göre dört kat daha fazla gerçekleştiği görülmektedir (Güleç 2009)

3.4 KVH RİSK HESAPLAMA YÖNTEMLERİ

KVH riskinin tahmin edilebilmesi, risk faktörlerini ortadan kaldırabilmek, önlem alabilmek ve tedaviyi sağlayabilmek için çok önemlidir. Tahmin için ise KVH tanısı konmuş hastalardaki risk faktörlerinin bir arada bulunduğu gerçeği yol göstermektedir. Günümüzde KVH belirlemek için bazı risk hesaplama yöntemleri geliştirilmiştir. Bu yöntemler aşağıda belirtilmiştir (Kültürsay 2011).

Framingham risk hesaplama sistemi: Amerika'nın Massachusetts eyaletine bağlı olan Framingham kasabasında gerçekleştirilen bir izlem çalışmasına dayanmaktadır. Kasabada yaşayan 5209 erişkin ile 1948 yılında başlanmıştır, şu anda üçüncü kuşak izlenmektedir. Amerikan Kalp Birliği (AHA) bu veriler üzerinden bir risk değerlendirme sistemi geliştirmiştir. Geliştirilen sistemde Tablo 3.1'de belirtilen risk faktörleri ile 10 yıl içindeki KVH riski hesaplanır.

Tablo 3.1: Framingham risk hesaplama sistemindeki KVH risk faktörleri

Risk faktörleri
Yaş (Kadın >55, erkek >45)
Sigara kullanımı
Hipertansiyon

Ailede erken koroner kalp hastalığı (Birinci derece kadın akrabalar için <65, erkek akrabalar için <55)
HDL kolesterol düşüklüğü (Kadın için <50 mg/dl, erkek için <40 mg/dl)

Kaynak: Güleç S., 2009. Kalp damar hastalıklarında global risk ve hedefler, Türk Kardiyol Dern Arş - Arch Turk Soc Cardiol. 37(2). ss.1-10.

SCORE çalışması: Framingham çalışmasının bölgesel bir uygulama olması sebebiyle Avrupa Kardiyoloji Derneği 205178 katılımcıdan elde ettikleri verileri kullanarak SCORE çalışmasını yapmıştır. Bu çalışmada kardiyovasküler hastalıkların on yıllık gelişim riski hesaplanır. Tablo 3.2'de bulunan risk faktörleri ile yapılan risk hesaplamasında düşük, orta, yüksek ve çok yüksek olmak üzere dört risk sonucundan birine ulaşılır.

Tablo 3.2 SCORE risk faktörleri

Risk faktörleri
Yaş
Cinsiyet
Toplam kolesterol
HDL kolesterol
Büyük tansiyon
Sigara kullanımı

Kaynak: Yavuz R., Yavuz D., Tontuş H. Ö., 2013. Artan mortalite ve morbidite nedeni olarak kardiyovasküler risk faktörlerine sistematik yaklaşım. Journal of Experimental and Clinical Medicine Deneysel ve Klinik Tıp Dergisi. 30(1). ss.47-53

Bu yöntemler dışında PROCAM, Reynolds Risk Score, QRISK, WHI/ISH ve çeşitli ulusal risk hesaplama yöntemleri de bulunmaktadır (Kültürsay 2011).

TEKHARF çalışması: 1990 yılında Türk halkının sağlık niteliği hakkında bilgi elde edebilmek amacıyla Türk Erişkinlerinde Kalp Hastalığı ve Risk Faktörleri (TEKHARF) çalışması başlatıldı. Çalışmada, Türkiye'nin yedi coğrafi bölgesinden ve 59 farklı yerleşim biriminden 20 yaş üzeri 3687 kişi düzenli olarak tarandı. Sonuçta birçok

istatistiki bilgi ile Türk toplumunun metabolik sendroma çok yatkın olduğu ortaya konulmuştur (Tavsanoglu 2014, Yavuz ve diğ. 2013).

3.5 KVH TANI YÖNTEMLERİ

Kalp ve damar hastalıkları açısından risk taşıyan hastalara KVH tanısı konulabilmesi için aşağıdaki yöntemler uygulanır (Florence.com.tr 2014):

a. Elektrokardiyografi (EKG)
b. Ekokardiyografi (EKO)
c. Egzersiz stres testleri – Treadmill
d. Myokard perfüzyon sintigrafisi
e. Positron yayınlayıcı tomografi (Positron Emission Tomography-PET)
f. Koroner anjiyo CT (Çok Kesitli BT Anjiyografi) (Multislice Kardiyak BT)
g. Kardiyak MR

4. VERİ VE YÖNTEM

4.1 VERİ KÜMESİ TANIMI

Bu çalışmada analiz edilen veri kümesi, Uzm. Dr. M. Ertuğrul Mercan tarafından KVH ihtimali ile hastaneye başvuran hastaların biyokimya sonuçları toplanarak oluşturulmuştur.

Veri KVH tanısı konulan hastalar ve KVH tanısı konulmayan hastalar olmak üzere iki farklı MS Excel dosyasında bulunmaktadır. KVH tanısı konulmuş 297 hasta verisi içeren dosya bundan sonra HASTA adı ile, KVH tanısı konulmamış 307 hasta verisi içeren dosya ise HASTA_DEGIL ismi ile bahsedilecektir.

4.2 VERİNİN HAZIRLANMASI

HASTA ve HASTA_DEGIL dosyaları farklı dönemlerde toplanmış farklı parametrelere ait verilerden oluşmaktadır. Bu nedenle, öncelikle, ortak parametrelerin belirlenmesi ve her iki dosyanın bir bileşiminin hazırlanması gerekmektedir. İki dosya içerisindeki parametreler incelenerek her iki dosyada da bulunan 30 parametre tespit edilmiştir.

4.2.1 Veri Kümesi İçeriği

Tablo 4.1'de iki dosya içerisinde ortak olan alanların listesi ve kısa tanımları bulunmaktadır.

Tablo 4.1: Veri dosyalarında bulunan ortak alanlar

#	HASTA_DEGIL dosyası alan adı	HASTA dosyası alan adı	Çalışmada verilecek alan adı	Açıklama
1	YAŞ	YAŞ	YAS	Yaş Bilgisi
2	CİNSİYET	CİNSİYET	CINSIYET	Cinsiyet

3	BMI	BMI	BMI	Beden Kitle İndeksi
4	AİLE ÖY.	AİLE Ö.	AILE_OY KUSU	Ailesinde kalp hastalığı olup olmadığı bilgisi
5	HT	HT	HT	Hiper Tansiyon hastalığı durumu
6	HL	HL	HL	Yüksek Kolesterol durumu
7	DM	DM	DM	Diyabet
8	SİGARA	SİGARA	SIGARA	Sigara Kullanımı
9	ASA	ASA	ASA	Asprin Kullanımı
10	RAAS	RAAS İNH	RAAS	RAAS Inhibitörü kullanımı
11	BB	BB	BB	Beta-Adrenerjik Reseptör Kullanımı
12	KKB	KKB	KKB	Kalsiyum Kanal Blokeri kullanım durumu
13	STATİN	STATİN	STATIN	Statin Kullanımı
14	METFOR.	METFOR MİN	METFOR MIN	Metformin Kullanımı
15	SULFON.	SULFANÜ RE	SULFONI LURE	Sülfonilüre kullanımı
16	İNSÜLİN	İNSÜLİN	INSULIN	Insülin Kullanımı
17	DİÜRETİK	DİÜRETİK	DIURETIK	Diüretik Kullanımı
18	TK	T KOLES.	TK	Total kolesterol Seviyesi
19	LDL	LDL-K	LDL	Kötü Huylu Kolesterol Seviyesi
20	HDL	HDL-K	HDL	İyi Huylu Kolesterol Seviyesi
21	TRİG.	TRİG.	TRIGLISE RID	Trigliserid Seviyesi
22	BUN	BUN	BUN	Kan Üre Nitrojen Miktarı(Üre Azotu)
23	KREATİ.	KREATİNİ N	KREATINI N	Kreatinin Seviyesi
24	Ü.A.	ÜA	UA	Ürik Asit Seviyesi
25	WBC	WBC	WBC	Beyaz Küre Sayısı
26	HGB	HGB	HGB	Hemoglobin
27	PLT	PLT	PLT	Trombosit Sayısı
28	PDW	PDW	PDW	Trombosit dağılım aralığı
29	RDW	RDW	RDW	Kırmızı küre dağılım genişliği
30	HbA1C	HbA1C	HBA1C	3 Aylık Kan Şekeri Ortalaması

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

4.2.2 Veri Tabanı Oluşturulması

Verinin bir araya getirilmesi, temizlenmesi, dönüştürülmesi ve dışa aktarılabilmesi işlemleri için MS SQL Server üzerinde TEZ adı ile bir veri tabanı oluşturulmuştur.

HASTA dosyasında bulunan veriler "Import Data" yöntemi ile TEZ veri tabanı içerisinde HASTA adlı bir tablo oluşturularak aktarılmıştır.

HASTA_DEGIL dosyasında bulunan veriler "Import Data" yöntemi ile TEZ veri tabanı içerisinde HASTA_DEGIL adlı bir tablo oluşturularak aktarılmıştır.

4.2.3 Parametrelerin Değerlendirilmesi ve Veri Temizliği

4.2.3.1 Kapsam dışı bırakılacak parametrelerin belirlenmesi

ASA, RAAS, BB, KKB, STATIN, METFORMIN, SULFONILURE, INSULIN ve DIURETIK alanları ilaç kullanımı durumunu ifade etmektedir. Bu çalışmadaki amaç, bir ilaç kullanımına bağlı olmaksızın, herhangi bir tanı konulmamış kişilerde dahi KVH tespiti yapabilecek bir model oluşturulabilmesidir. Oluşan modelin uygulanabilirliğini düşüreceği için veri kümesi içerisindeki ilaç kullanımına ilişkin bilgilerin çıkarılmasına karar verilmiştir.

4.2.3.2 NULL kayıtların temizlenmesi / kapsam dışı bırakılması

AILE_OYKUSU, HT, HL, DM, SIGARA alanları içerisinde "Evet" ifadesi için 1, "Hayır" ifadesi için 0 verisi bulunmaktadır. Bu alanlar içerisinde bulunan NULL kayıtların "Hayır" bilgisini ifade ettiği veri sahibi uzman doktor tarafından bildirilmiştir. Bu alanlar içerisindeki NULL kayıtlar "Hayır" ifadesi için 0 olacak şekilde düzeltilmiştir.

TK, LDL, HDL, TRIGLISERID, BUN, KREATININ, UA, WBC, HGB, PLT, PDW, RDW alanlarındaki NULL kayıt sayıları 3 ile 7 arasında değişmekte ve aynı kayıtlara

denk geldiği görülmektedir. NULL kayıtların az sayıda olması sebebiyle her bir alan için ortalama değerler hesaplanarak NULL değerler yerine kullanılmıştır.

HBA1C alanı HASTA tablosunda 232 adet NULL kayıt içermektedir. Toplamda 297 kayıt içerisinden 232 kaydı NULL olan bir alanın modele bir katkıda bulunması söz konusu olmayacağından bu alan kapsam dışı bırakılmıştır.

Bu işlemler sonucunda oluşan HASTA ve HASTA_DEGIL tablo yapıları Tablo 4.2'de belirtilmiştir.

Tablo 4.2: HASTA ve HASTA_DEGIL tablo yapıları

#	Alan Adı	Veri Tipi	Değer Aralıkları
1	AILE_OYKUSU	float	0, 1
2	BMI	float	20 – 46.6
3	BUN	float	7.1 – 61.3
4	CINSIYET	varchar(255)	E, K
5	DM	float	0, 1
6	HDL	float	17 - 3029
7	HGB	float	1.6 - 132
8	HL	float	0, 1
9	HT	float	0, 1
10	KREATININ	float	0.5 - 102
11	LDL	float	9 - 261
12	PDW	float	8 - 23
13	PLT	float	21.9 - 815
14	RDW	float	1 - 20
15	SIGARA	float	0, 1
16	TK	float	50 - 419
17	TRIGLISERID	float	10 - 1627
18	UA	float	0.5 – 10.2
19	WBC	float	0.2 – 17.8
20	YAS	float	30 - 89

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

4.2.4 Referans Değerler

Parametrelere ait veriler incelendiğinde tamamının sayısal değerler içerdiği görülmektedir. Sayısal değerlerin ifade ettiği bilgiyi elde etmek amacıyla her bir parametrenin referans değer aralıkları uzman görüşü alınarak oluşturulmuş ve bu değerler kullanılarak dönüştürme işlemi yapılmıştır.

Referans değerlere göre yorumlama işlemi sırasında 3 farklı yorum tekniği kullanılmıştır. Kullanılan her bir yorum tekniği farklı bir "Grup" olarak ifade edilmiş ve her bir grup için farklı referans aralıkları belirlenmiştir.

Bu farklı gruplara ait referans aralıklarını belirtmek ve verinin dönüştürülmesi işlemini gerçekleştirebilmek için REFERANS_DEGERLER adlı bir tablo oluşturulmuştur. REFERANS_DEGERLER tablo yapısı Tablo 4.3'te belirtilmiştir.

Tablo 4.3: REFERANS_DEGERLER tablo yapısı

#	Alan Adı	Alan Açıklaması	Veri Tipi
1	PARAMETRE	Parametre Adı	nvarchar(50)
2	ALT_SINIR	Referans Aralığı Alt Sınır Değeri	numeric(10,2)
3	UST_SINIR	Referans Aralığı Üst Sınır Değeri	numeric(10,2)
4	GRUP_NO	Dönüştürme işlemi Grup Numarası	tinyint
5	GRUP	Dönüştürülecek Değer	nvarchar(50)
6	CINSIYET	Referans Aralığın Kabul Edildiği Cinsiyet	nvarchar(2)

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

4.2.4.1 Referans değerler hazırlanırken kullanılan yöntemler

Grup 1 (Detaylı Gruplama): Bu grup referans değerleri hazırlanırken, normal aralık değerlerinin altında kalan verilerin DÜŞÜK, aralık içerisinde olan verilerin NORMAL ve aralık üzerinde olan verilerin ise YÜKSEK olarak gruplanması hedeflenmiştir.

Grup 2 (Doktor Görüşü ile Gruplama): Bazı parametreler için, değerin referans aralığına göre düşük ya da yüksek olması, ek bir anlam yaratmayabilir. Bu sebeple doktor görüşü ile bazı parametrelerin düşük ya da yüksek değerleri NORMAL ya da NORMAL DEGIL olarak kabul edilecek şekilde Grup 2 oluşturulmuştur.

Grup 3 (Genel Gruplama): Bu grupta ise, tüm parametreler için belirlenen referans aralığı içerisinde olan değerler NORMAL ve bu aralık dışında kalan tüm değerler NORMAL DEĞİL olacak şekilde yeni bir grup hazırlanmıştır.

Her bir parametre için yukarıda tanımlanmış olan farklı gruplarda referans aralıkları REFERANS_DEGERLER tablosuna yazılmıştır.

4.2.4.2 Parametreler ve referans değer aralıkları

Yaş: Yaş KVH risk faktörleri içerisinde değiştirilemeyen bir risk faktörü olarak yer almaktadır. Tablo 4.4'de yaş parametresi için çalışmada kullanılmış olan referans değer aralıkları yer almaktadır.

Tablo 4.4: Yaş parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	ORTA	30.00	49.99
1	YASLI_ADAYI	50.00	54.99
1	ERKEN_YASLI	55.00	64.99
1	YASLI	65.00	85.00
1	ILERI_YASLI	85.01	110.00
2	<40	0.00	39.99
2	40-50	40.00	49.99

2	50-60	50.00	59.99
2	60-70	60.00	69.99
2	>=70	70.00	10000000.00
3	OLGUN	30.00	64.99
3	YASLI	65.00	110.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Aile Öyküsü: Bu parametre yakın aile üyelerinde KVH hastalığı görülüp görülmediğini ifade etmektedir. Değiştirilemeyen KVH risk faktörleri içerisinde yer almaktadır. Herhangi bir yönteme göre ifadesi değişemeyeceği için üç grupta da 1 değeri AO_VAR (Aile öyküsünde KVH var), 0 değeri ise AO_YOK (Aile öyküsünde KVH yok) olarak değerlendirilmiştir.

BMI (BKİ): Beden kitle indeksi değiştirilebilen KVH risk faktörlerinden biridir, kadın ve erkek için referans değer aralıkları değişmemektedir. Tablo 4.5'de gruplara göre belirlenmiş referans değer aralıkları görülebilir.

Tablo 4.5: BMI parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	ZAYIF	0.00	18.49
1	BMI_NORMAL	18.50	24.99
1	OBEZ1	25.00	29.99
1	OBEZ2	30.00	34.99
1	OBEZ3	35.00	39.99
1	OBEZ4	40.00	100.00
2	BMI_NRM	0.00	26.99
2	FAZLA_KILO	27.00	29.99
2	OBEZ	30.00	34.99
2	MORBID_OBEZ	35.00	10000000.00
3	BMI_ZAYIF	0.00	18.49
3	BMI_NORMAL	18.50	24.99
3	BMI_OBEZ	25.00	100.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

BUN: Kandaki üre azotu miktarını gösteren bu parametrede referans değer aralıkları kadın ve erkek için değişmemektedir. Tablo 4.6'da belirlenmiş olan referans değer aralıkları belirtilmiştir.

Tablo 4.6: BUN parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	BUN_DUSUK	0.00	5.99
1	BUN_NORMAL	6.00	20.00
1	BUN_YUKSEK	20.01	1000000.00
2	BUN_DUSUK	0.00	5.99
2	BUN_NRM	6.00	20.00
2	BUN_YUKSEK	20.01	1000000.00
3	BUN_NORM_DEG	0.00	5.99
3	BUN_NORMAL	6.00	20.00
3	BUN_NORM_DEG	20.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

DM: Diyabet parametresi diyabet hastalığı olup olmadığını 1 ve 0 değerleriyle ifade etmektedir. KVH değiştirilebilen risk faktörlerinden biri olarak kabul edilmektedir. Bu parametre de gruplara göre değişiklik göstermediğinden 1 değeri DM_VAR (diyabet hastalığı var), 0 değeri DM_YOK (diyabet hastalığı yok) olarak nitelendirilir.

HDL: Framingham ve Score risk hesaplama yöntemlerinde kullanılan iyi huylu kolesterol seviyesi için belirlenmiş olan referans değer aralıkları kadın ve erkek için değişmemekte ve Tablo 4.7'de gösterilmektedir.

Tablo 4.7: HDL parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	HDL_DUSUK	0.00	39.99
1	HDL_NORMAL	40.00	60.00
1	HDL_YUKSEK	60.01	1000000.00
2	HDL_ND	0.00	39.99
2	HDL_NRM	40.00	1000000.00
3	HDL_NORM_DEG	0.00	39.99
3	HDL_NORMAL	40.00	60.00
3	HDL_NORM_DEG	60.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

HGB: Hemoglobin için belirlenen referans değer aralıkları kadın ve erkeklerde değişmektedir. Kullanılan referans değer aralıkları Tablo 4.8'de yer almaktadır.

Tablo 4.8: HGB parametresi referans değer aralıkları

GRUP_NO	GRUP	Erkek		Kadın	
		ALT_SINIR	UST_SINIR	ALT_SINIR	UST_SINIR
1	HGB_DUSUK	0.00	13.49	0.00	11.99
1	HGB_NORMAL	13.50	17.50	12.00	16.00
1	HGB_YUKSEK	17.51	1000000.00	16.01	1000000.00
2	HGB_ND	0.00	13.49	0.00	11.99
2	HGB_NRM	13.50	17.50	12.00	16.00
2	HGB_ND	17.51	1000000.00	16.01	1000000.00
3	HGB_NORM_DEG	0.00	13.49	0.00	11.99
3	HGB_NORMAL	13.50	17.50	12.00	16.00
3	HGB_NORM_DEG	17.51	1000000.00	16.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

HL: Bu parametre, hastada yüksek kolesterol hastalığı olup olmadığını 1 ve 0 değerleri ile ifade etmektedir. Gruplara göre değişemeyeceğinden her üç grup için de 1 değeri için grup ifadesi HL_VAR (yüksek kolesterol var), 0 değeri için HL_YOK (yüksek kolesterol yok) olarak belirlenmiştir.

HT: Hastada hipertansiyon (yüksek tansiyon) hastalığı bulunup bulunmadığı bu parametre ile belirtilmiştir. 1 ve 0 değerleri ile ifade edilir. Her bir grup için değişmeyeceğinden HT_VAR (yüksek tansiyon var) ve HT_YOK (yüksek tansiyon yok) olarak düzenlenmiştir.

Kreatinin: Tablo 4.9'da belirtilen referans değer aralıkları kadın ve erkek için değişir.

Tablo 4.9: Kreatinin parametresi referans değer aralıkları

GRUP_NO	GRUP	Erkek		Kadın	
		ALT_SINIR	UST_SINIR	ALT_SINIR	UST_SINIR
1	KRTN_DUSUK	0.00	0.74	0.00	0.64
1	KRTN_NORMAL	0.75	1.20	0.65	1.00
1	KRTN_YUKSEK	1.21	1000000.00	1.01	1000000.00
2	KRTN_NRM	0.00	1.20	0.00	1.00
2	KRTN_ND	1.21	1000000.00	1.01	1000000.00
3	KRTN_NORM_DEG	0.00	0.74	0.00	0.64

3	KRTN_NORMAL	0.75	1.20	0.65	1.00
3	KRTN_NORM_DEG	1.21	1000000.00	1.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

LDL: Kötü huylu kolesterol için kullanılan referans değer aralıkları kadın ve erkek için değişmemekte ve değerler Tablo 4.10'da gösterilmektedir.

Tablo 4.10: LDL parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	LDL_NORMAL	0.00	129.99
1	LDL_YUKSEK	130.00	1000000.00
2	LDL_NRM	0.00	129.99
2	LDL_ND	130.00	1000000.00
3	LDL_NORMAL	0.00	129.99
3	LDL_NORM_DEG	130.00	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

PDW: Trombosit dağılım aralığını ifade eden PDW için belirlenmiş olan referans değer aralıkları kadın ve erkek için değişmektedir. İlgili referans değer aralıkları Tablo 4.11'de görülebilir.

Tablo 4.11: PDW parametresi referans değer aralıkları

GRUP_NO	GRUP	Erkek		Kadın	
		ALT_SINIR	UST_SINIR	ALT_SINIR	UST_SINIR
1	PDW_DUSUK	0.00	9.29	0.00	9.79
1	PDW_NORMAL	9.30	14.30	9.80	16.00
1	PDW_YUKSEK	14.39	1000000.00	16.01	1000000.00
2	PDW_NRM	0.00	14.29	0.00	16.00
2	PDW_ND	14.30	1000000.00	16.01	1000000.00
3	PDW_NORM_DEG	0.00	9.29	0.00	9.79
3	PDW_NORMAL	9.30	14.30	9.80	16.00
3	PDW_NORM_DEG	14.39	1000000.00	16.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

PLT: Trombosit sayısını ifade eder PLT parametresi için belirlenen referans değer aralıkları kadın ve erkek için değişmemekte ve Tablo 4.12'de gösterilmektedir.

Tablo 4.12: PLT parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	PLT_DUSUK	0.00	149.99
1	PLT_NORMAL	150.00	399.00
1	PLT_YUKSEK	399.01	1000000.00
2	PLT_DUSUK	0.00	149.99
2	PLT_NRM	150.00	399.00
2	PLT_YUKSEK	399.01	1000000.00
3	PLT_NORM_DEG	0.00	149.99
3	PLT_NORMAL	150.00	399.00
3	PLT_NORM_DEG	399.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

RDW: Kırmızı küre dağılım değişikliğini içeren bu parametre için referans değer aralıkları kadın ve erkek için değişmemektedir. Tablo 4.13'de ilgili referans değer aralıkları belirtilmiştir.

Tablo 4.13: RDW parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	RDW_DUSUK	0.00	11.49
1	RDW_NORMAL	11.50	14.50
1	RDW_YUKSEK	14.51	1000000.00
2	RDW_NRM	0.00	14.50
2	RDW_ND	14.51	1000000.00
3	RDW_NORM_DEG	0.00	11.49
3	RDW_NORMAL	11.50	14.50
3	RDW_NORM_DEG	14.51	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Sigara: Hastanın sigara içme durumu 1 ve 0 ile ifade edilmektedir. Bu parametre için grup ifadeleri değişmeyeceğinden tüm gruplarda 1 değeri ICIYOR, 0 değeri ICMIYOR olarak düzenlenmiştir.

TK: Total kolesterol parametresi için referans değer aralıkları kadın ve erkek için değişmektedir. Referans değer aralıkları Tablo 4.14'de belirtilmiştir.

Tablo 4.14: TK parametresi referans değer aralıkları

GRUP_NO	GRUP	Erkek		Kadın	
		ALT_SINIR	UST_SINIR	ALT_SINIR	UST_SINIR
1	TK_DUSUK	0.00	81.99	0.00	91.99
1	TK_NORMAL	82.00	200.00	92.00	200.00
1	TK_YUKSEK	200.01	1000000.00	200.01	1000000.00
2	TK_NRM	0.00	200.00	0.00	200.00
2	TK_ND	200.01	1000000.00	200.01	1000000.00
3	TK_NORM_DEG	0.00	81.99	0.00	91.99
3	TK_NORMAL	82.00	200.00	92.00	200.00
3	TK_NORM_DEG	200.01	1000000.00	200.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Trigliserid: Referans değer aralıkları Trigliserid için kadın ve erkekte değişmemektedir. Tablo 4.15'de referans değer aralıkları gösterilmektedir.

Tablo 4.15: RDW parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	TRG_DUSUK	0.00	29.99
1	TRG_NORMAL	30.00	149.00
1	TRG_YUKSEK	149.01	1000000.00
2	TRG_NRM	0.00	149.00
2	TRG_ND	149.01	1000000.00
3	TRG_NORM_DEG	0.00	29.99
3	TRG_NORMAL	30.00	149.00
3	TRG_NORM_DEG	149.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

UA: Ürik asit değerini gösteren bu parametre için referans değer aralıkları kadın ve erkek için değişmektedir. Tablo 4.16'da referans değer aralıkları gösterilmektedir.

Tablo 4.16: UA parametresi referans değer aralıkları

GRUP_NO	GRUP	Erkek		Kadın	
		ALT_SINIR	UST_SINIR	ALT_SINIR	UST_SINIR
1	UA_DUSUK	0.00	3.49	0.00	2.59
1	UA_NORMAL	3.50	7.70	2.60	6.80

1	UA_YUKSEK	7.71	1000000.00	6.81	1000000.00
2	UA_NRM	0.00	6.99	0.00	6.99
2	UA_ND	7.00	10000000.00	7.00	10000000.00
3	UA_NORM_DEG	0.00	3.49	0.00	2.59
3	UA_NORMAL	3.50	7.70	2.60	6.80
3	UA_NORM_DEG	7.71	1000000.00	6.81	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

WBC: Kandaki beyaz küre sayısını ifade eden bu parametre için belirlenen referans değer aralıkları kadın ve erkek için değişmemektedir. Referans değer aralıkları Tablo 4.17'de belirtilmiştir.

Tablo 4.17: WBC parametresi referans değer aralıkları

GRUP_NO	GRUP	ALT_SINIR	UST_SINIR
1	WBC_DUSUK	0.00	3.99
1	WBC_NORMAL	4.00	10.00
1	WBC_YUKSEK	10.01	1000000.00
2	WBC_NRM	0.00	10.00
2	WBC_ND	10.01	1000000.00
3	WBC_NORM_DEG	0.00	3.99
3	WBC_NORMAL	4.00	10.00
3	WBC_NORM_DEG	10.01	1000000.00

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

4.2.5 Parametrelerin Dönüştürülmesi

Veri kümesi içerisindeki verilerin, oluşturulan referans değer aralıklarına göre dönüştürülmesi işlemi için DATA_OLUSTUR adlı prosedür oluşturulmuştur (Bkz Ek 1: DATA_OLUSTUR Prosedürü). Prosedür çalıştırıldığında, HASTA ve HASTA_DEGIL tablolarında bulunan veriler birleştirilir. Her bir grup için dönüştürme işlemi yapılır ve DATA adlı yeni bir tablo oluşturularak içine yazılır. Burada dönüştürme işleminin kayıt bazında yapılabilmesi için bir fonksiyon geliştirilmiştir (Bkz Ek 2: GRUP_GETIR Fonksiyonu). Bu işlemler sonucunda DATA adlı tablo içerisinde 3 farklı grup için dönüştürülmüş veri bulunur.

4.2.6 Verinin Weka Programı İçin Hazırlanması

Bu çalışmada model oluşturulması için Weka veri madenciliği programı kullanılacaktır (Hall ve diğ. 2009). Weka programında analiz yapabilmek için, verinin belirlenmiş bir formatta oluşturulması gerekmektedir. Weka arff uzantılı bir dosya formatı ile çalışır. Bir arff dosyası iki kısımdan oluşur;

a. Tanım: Bu kısımda veri kümesi içerisinde bulunan niteliklerin ve nitelik içerisinde yer alan değerlerin tanımlanması gerekmektedir. Örnek; Grup 2 için YAS niteliği aşağıdaki gibi bir tanımlanır.

@attribute YAS {<40, >=70, 40-50, 50-60, 60-70}

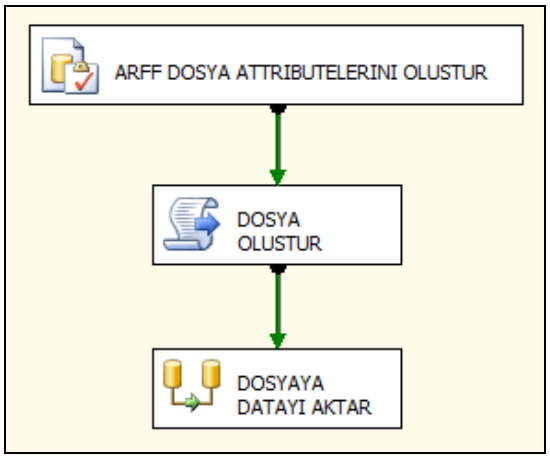
b. Data: Tanım kısmından sonra @data ifadesi ile veri kısmı başlamış olur. İncelenecek veri, niteliklerin tanım kısmında belirtilmiş olan sıralanışına göre virgül (,) ile ayrılarak yazılır. Örnek; bir hasta verisinin veri kısmında ifade edilişi aşağıda gösterilmiştir.

<40,E,AO_VAR,ICIYOR,BMI_NRM,DM_YOK,HGB_NRM,HDL_ND,HL_YOK,HT_YOK,PDW_ND,PLT_DUSUK,RDW_NRM,WBC_ND,TRG_NRM,KRTN_NRM,LDL_NRM,BUN_NRM,TK_NRM,UA_NRM,HASTA

Arff dosyası oluşturmak amacı ile Microsoft SQL Server Integration Services (SSIS) kullanılmıştır. REFERANS_DEGERLER tablosundaki her bir GRUP için bir arff dosyası oluşturulması ve analizlerin bu üç farklı veri kümesi için yapılması gerekmektedir. Arff dosyalarının oluşturulabilmesi için 3 adet SSIS paketi oluşturulmuştur. Bu kısımda grup 2 için oluşturulmuş olan paket üzerinden anlatım yapılmaktadır.

ARFF_OLUSTUR_GRUP2 SSIS paketi hazırlanışı: SSIS paketi bir arff dosyasında bulunması gereken kısımları oluşturacak şekilde tasarlanmıştır. Oluşan SSIS paketi görünümü Şekil 4.1'de belirtilmiştir.

Şekil 4.1: SSIS paket görünümü



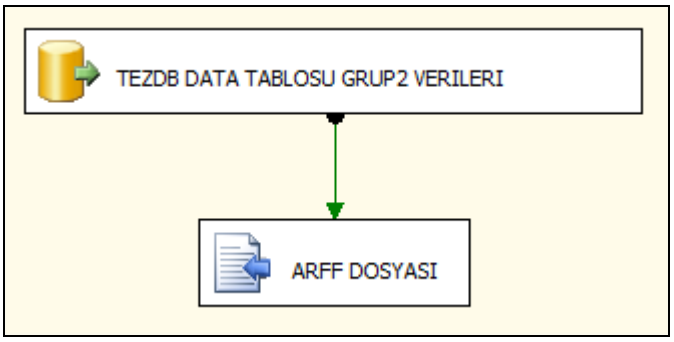
Kaynak: Bu şekil Serap ERKUŞ tarafından hazırlanmıştır.

a. Tanım kısmının hazırlanması; tanım kısmında bulunması gereken bilgileri elimizdeki veriyi kullanarak oluşturabilmek için TEZ veri tabanı içerisinde bir prosedür oluşturulmuştur (Bkz Ek 3: DOSYA_ICERIGI_OLUSTUR Prosedürü). "ARFF DOSYA ATTRIBUTELERINI OLUSTUR" Execute SQL Task'ı içerisinde, bu prosedür grup parametresi gönderilerek çalıştırılmaktadır. Prosedür verilen grup için belirtilmesi gereken nitelikleri hazırlar ve geriye döndürür. Task geriye dönen değeri DOSYA_ICERIGI adı ile oluşturulmuş olan değişkene yazar.

b. Arff dosyasının oluşturulması ve tanım kısmının eklenmesi; bir sonraki aşama için Şekil 4.1'de görülen DOSYA_OLUSTUR komut dizisi elemanı hazırlanmıştır (Bkz Ek 4: DOSYA_OLUSTUR Komut Dizisi). Bu komut dizisi içerisinde arff dosyası fiziksel olarak oluşturulur ve içerisine bir önceki aşamada DOSYA_ICERIGI değişkenine yüklenmiş olan bilgiler yazılır.

c. Verinin arff dosyasına yazılması; Şekil 4.1'de 3. aşamada görülen DOSYAYA DATAYI AKTAR veri aktarım bileşeni içeriği Şekil 4.2'de belirtilmiştir. İlk kısım olan kaynaktan verinin alınması işlemi sırasında "OLEDB Source" bileşeni kullanılmıştır. Bu kaynak bileşeni TEZ veri tabanı DATA tablosundan Grup_No bilgisi 2 olan verileri alacak şekilde düzenlenmiştir. Bu bileşen ile alınmış olan bilgiler, ARFF DOSYASI adlı hedef bileşeni ile bir önceki aşamada oluşturulmuş olan arff dosyasına yazılır.

Şekil 4.2: DOSYAYA DATAYI AKTAR veri aktarım bileşeni



Kaynak: Bu şekil Serap ERKUŞ tarafından hazırlanmıştır.

Paket yukarıda bahsedilen üç kısımdan oluşur. Bu paket çalıştırıldığında, bileşenler sırası ile çalışarak, verilen grup bilgisine ait veriyi gerekli formatlama işlemleri sonrası arff dosyasını oluşturup, içerisine yazar.

4.3 VERİ KÜMESİ SEÇİMİ

Bölüm 4.2.4.1'de veri kümesinin gruplama yöntemlerinden bahsedilmişti. Bu yöntemler ile elde edilen üç farklı veri kümesinden başarılı model oluşturabilmek için doğru algoritmanın, doğru parametreler ile, doğru veri kümesinde uygulanması gerekmektedir. Bu noktada doktor görüşü ile oluşturulmuş olan grup 2 veri kümesi parametre seçimi sırasında kullanılmak üzere seçilmiştir. Bir doktor gözü ile parametrelerin referans değerlere göre nitelendirilmesi sonucu oluşturulan, grup 2 veri kümesi üzerinde başarılı olan parametreler ile seçilen algoritma diğer veri kümelerine de uygulanacaktır.

4.4 NİTELİK SEÇİMİ

Weka programı, nitelik seçimi için kullanılan yöntemleri bir araya getirmiştir. Nitelik seçim kısmında yapılan çalışmalarda, grup 2 veri kümesi üzerinde birçok yöntemin yirmi parametre içerisinden yaklaşık olarak aynı parametreleri seçtiği gözlemlenmiştir. Nitelik seçimi yöntemleri içerisinden InfoGainAttributeEval yöntemi esas alınarak parametre seçim işlemi yapılmıştır. InfoGainAttributeEval yöntemi, bir niteliğin bilgi kazancının ölçülmesi ile sınıflandırmaya katkısını değerlendirir (Weka 2014).

Tablo 4.18'de InfoGainAttributeEval algoritması ile grup 2 için elde edilen sonuçlar verilmiştir.

Tablo 4.18: InfoGainAttributeEval yöntemi ile nitelik seçim işlemi sonuçları

#	nitelik (attribute)	derece (rank)	#	nitelik (attribute)	derece (rank)
1	CINSIYET	0.133	11	HT	0.014
2	BMI	0.111	12	PLT	0.013
3	SIGARA	0.059	13	UA	0.006
4	TK	0.053	14	DM	0.004
5	HDL	0.04	15	PDW	0.002
6	HGB	0.04	16	TRIGLISERID	0.001
7	LDL	0.03	17	RDW	0.001
8	YAS	0.018	18	HL	0.001
9	BUN	0.018	19	AILE_OYKUSU	0
10	WBC	0.016	20	KREATININ	0

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Bu yöntem ile elde ettiğimiz sonuçların en iyi modele ulaşabilmek için ihtiyacımız olan parametre grubu olduğuna emin olabilmek ve en doğru parametre grubuna sahip olabilmek için beş farklı nitelik seçim algoritması için nitelik seçim işlemi yeniden yapılmıştır. Elde edilen sonuçlarda her bir parametrenin konumuna göre Tablo 4.19'da görülebilecek bir liste oluşturulmuştur.

Tablo 4.19: Parametrelerin farklı nitelik seçim yöntemlerine göre sıralaması

#	nitelik (attribute)	Cfs Subset Eval	Classifier Subset Eval	Filtered Subset Eval	Gain Ratio Attribute Eval	ReliefF Attribute Eval	Average
1	CINSIYET	1	1	1	1	1	1.0
2	BMI	2	3	2	2	2	2.2
3	HGB	5	6	5	5	6	5.4
4	TK	3	11	3	4	7	5.6
5	HDL	6	5	6	6	5	5.6
6	SIGARA	4	20	4	3	4	7.0
7	YAS	10	4	10	13	3	8.0
8	LDL	11	2	11	8	9	8.2

9	BUN	7	10	7	9	15	9.6
10	WBC	8	14	8	7	14	10.2
11	PLT	9	12	9	10	20	12.0
12	PDW	17	7	17	16	13	14.0
13	DM	14	18	14	14	11	14.2
14	UA	13	17	13	12	16	14.2
15	AILE_OYKUSU	16	13	16	19	8	14.4
16	HT	12	19	12	11	19	14.6
17	TRIGLISERID	18	8	18	17	12	14.6
18	HL	19	9	19	18	10	15.0
19	RDW	15	16	15	15	17	15.6
20	KREATININ	20	15	20	20	18	18.6

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Tablo 4.19'da verilen liste ile Tablo 4.18'de bulunan InfoGainAttributeEval yöntemi sonuçları karşılaştırıldığında, yirmi parametre içerisinden trigliserid, RDW, HL, aile öyküsü ve kreatinin parametreleri her iki tabloda da ortak bir şekilde düşük oranlarda kalmaktadır, bu nedenle ilk etapta bu parametreler çalışma dışında bırakılmıştır. HT parametresi ortalamaya göre düşük oran sebebiyle çalışmadan çıkarılmış olan aile_oykusu parametresinin de altında kaldığından bu parametre de hedef parametre grubu dışında tutulmuştur.

Elde edilen nitelikler hakkında uzman görüşü alınmıştır. Veri kümesi içerisinde kolesterol ile ilgili bilgi veren 4 nitelik bulunmaktadır. HL (yüksek kolesterol) parametresi önceki paragrafta belirtildiği gibi çalışmadan çıkarılmıştır. Geriye kalan kolesterol ile ilgili parametrelerden HDL (iyi huylu kolesterol) ve LDL (kötü huylu kolesterol) niteliklerinin yeterli bilgiyi vereceği ve TK (total kolesterol) niteliğinin de çalışmadan çıkarılmasına doktor görüşü ile karar verilmiştir. Bu bilgiler ışığında nitelik seçim işlemi tamamlanmıştır. Elde edilen son nitelik listesi Tablo 4.20'de belirtilmiştir.

Tablo 4.20: Modelde kullanılacak nitelikler

#	**nitelik (attribute)**	#	**nitelik (attribute)**
1	CINSIYET	8	BUN
2	BMI	9	WBC
3	SIGARA	10	PLT
4	HDL	11	UA

5	HGB	12	DM
6	LDL	13	PDW
7	YAS		

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

4.5 KULLANILAN VERİ MADENCİLİĞİ YÖNTEMLERİ

Çalışmada biyokimya veri kümesi üzerinden KVH tahminlemesi yapılması hedeflenmiştir. Oluşturulan veri kümelerinde veriler gerçekte belirlenmiş olan sınıfları ile birlikte yer almaktadır. Dolayısıyla yapılmak istenen, eldeki sınıflanmış veriden bir öğrenme süreci sonrası tahmini en yüksek doğruluk oranıyla veren bir model üretmektir. Bu modeli oluşturabilmek için sınıflandırıcı veri madenciliği yöntemleri kullanılmıştır.

Sınıflandırıcı yöntemler, sekiz farklı kategoriye ayrılır. Bu kategorilerden MI ve MISC sınıflandırıcı kategorilerindeki algoritmalar ve aşağıda açıklanan algoritmalar dışında kalan yöntemler de denenmiş fakat diğer yöntemlere göre başarısız kaldıkları için çalışmaya dahil edilmemişlerdir.

Bayes sınıflandırıcılar içerisinden AODE (Averaged One-Dependence Estimators), WAODE (Weightily Averaged One-Dependence Estimators) ve HNB (Hidden Naive Bayes) algoritmaları çalışmada kullanılmıştır. Bayes algoritmalarının temelinde Naive Bayes vardır. AODE yöntemi, Naive Bayes'e göre daha zayıf bağımsızlık varsayımları olan ve Naive Bayes'e alternatif olarak küçük alanların ortalamaları ile sonuca yüksek başarı ile ulaşan bir yöntemdir (Webb ve diğ. 2005). AODE yönteminde tüm bağımlılık sınıflandırıcılarına aynı ağırlık verilirken, WAODE algoritmasında her bir sınıflandırıcı farklı ağırlıklandırılır (Zhang ve diğ. 2012). HNB yöntemi temelde bir Naive Bayes yöntemi olsa da, geliştirme sırasında karmaşık olan öğrenme sürecini en uygun hale getirmek hedeflenmiştir (Zhang ve diğ. 2005).

META kategorisinde bulunan algoritmalar öğrenme işlemi için en uygun parametre grubunu bulur ve sınıflandırmayı buna göre yapar (Neethu 2012). Bu kategoriden Dagging ve Classification via Regression yöntemleri kullanılmıştır.

Fonksiyon sınıflandırıcı yöntemler (Functions Classifiers) sinir ağları ve regresyon yöntemlerinden oluşur. Çalışmada bu kategoriden RBF (Radial Basis Function) Network ve Simple Logistic algoritmaları kullanılmıştır. RBF sınıflandırıcısı doğrusal olmayan fonksiyonları kolaylıkla modelleyebilir (Panda ve Patra 2008).

Ağaç sınıflandırıcılar (Tree classifiers) en popüler sınıflandırıcı yöntemlerindendir. Sonucunda üretilen model ağaç yapısına benzer bir akış diyagramı olarak sunulur. Bu nedenle karar ağaçları (Decision Trees) olarak da bilinirler (Garg ve Khurana 2014). Çalışmada karar ağaçları içerisinden LMT (Logistic Model Trees) algoritması kullanılmıştır. Bu yöntemde, dallara ayırma işlemi sırasında lojistik regresyon (logistic regression) fonksiyonu kullanılır.

Tembel sınıflandırıcılar (Lazy classifiers) basit ve etkilidir, tüm öğrenme verisini saklar ve yeni örnekler içeren bir veri kümesinden model oluşturamaz (Garg ve Khurana 2014). Çalışmada bu yöntemlerden LBR (Lazy Bayesian Rules) algoritması kullanılmıştır.

Kurallar sınıflandırıcılar (Rules classifiers) kategorisindeki yöntemlerde ise doğru tahminleme için birliktelik kuralları (Association Rules) kullanılır (Garg ve Khurana 2014). Çalışmada bu kategoriden Decision Table algoritması kullanılmıştır.

5. BULGULAR

Bu çalışmada, bir biyokimya veri kümesi, farklı bakış açıları ile üç farklı veri kümesine dönüştürülmüştür. Veri kümesi içerisindeki niteliklerin oluşturulacak olan modele etkisi incelenerek, en doğru parametre grubu elde edilmeye çalışılmıştır. Bu parametre grubu dışındaki nitelikler veri kümelerinden çıkarılmıştır. Belirlenmiş olan veri madenciliği yöntemleri, her bir veri kümesine uygulanmıştır. Bu işlemler sırasında elde edilen bulgular bu bölümde anlatılmaktadır.

Tablo 4.19 incelendiğinde CINSIYET parametresinin tüm nitelik seçim işlemlerinde ilk sırada yer aldığı görülmektedir. Aynı tabloda BKI (BKİ – Beden Kitle İndeksi) yüzde seksen oranında ikinci sırada yer aldığı görülebilir.

Her bir veri kümesi için bir önceki bölümde anlatılan veri madenciliği yöntemleri Tablo 4.20'de yer alan parametrelere uygulanmıştır. Tablo 5.1'de Grup 1 (detaylı gruplama) veri kümesi için uygulanan yöntemlerin ürettiği sonuçlara ait karışıklık matrisi bulunmaktadır.

Tablo 5.1: Grup 1 veri kümesi modelleme yöntemlerine göre karışıklık matrisi

#	Classifier		TP	FP	FN	TN
1	bayes	HNB	76	23	26	86
2	trees	LMT	74	25	24	88
3	functions	Simple Logistic	74	25	24	88
4	meta	Classification Via Regression	74	25	24	88
5	lazy	LBR	76	23	28	84
6	meta	Dagging	79	20	32	80
7	bayes	AODE	76	23	30	82
8	bayes	WAODE	75	24	29	83
9	functions	RBFNetwork	69	30	30	82
10	rules	Decision Table	76	23	37	75

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Grup 1 veri kümesinde ilk dört yöntemin de 211 adet test verisinde 162 adet doğru tahminleme yaparak yaklaşık yüzde yetmiş yedi başarılı olduğu görülmektedir. Grup 1 için uygulanan modelleme sonuçları Tablo 5.2'de verilmiştir. Sıralama önce en yüksek başarı oranı (accuracy), sonra en düşük RMSE değerlerine göre yapılmıştır.

Tablo 5.2: Grup 1 veri kümesi modelleme sonuçları

#	Classifier	Accuracy	RMSE	Sensitivity	Specificity	Precision
1	HNB	76.78	0.3937	0.77	0.77	0.77
2	LMT	76.78	0.4017	0.77	0.77	0.77
3	Simple Logistic	76.78	0.4017	0.77	0.77	0.77
4	Classification Via Regression	76.78	0.4133	0.77	0.77	0.77
5	LBR	75.83	0.3937	0.76	0.76	0.76
6	Dagging	75.36	0.4231	0.75	0.76	0.77
7	AODE	74.88	0.3934	0.75	0.75	0.75
8	WAODE	74.88	0.3965	0.75	0.75	0.75
9	RBFNetwork	71.56	0.4203	0.72	0.71	0.72
10	Decision Table	71.56	0.4247	0.72	0.72	0.72

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Aynı oranda doğruluk elde edilebilmesine rağmen içlerinden RMSE (Root Mean Squared Error) değeri en düşük olan HNB algoritması olmuştur.

Grup 2 (doktor görüşü ile gruplama) veri kümesinde yine Tablo 4.20'de belirtilmiş olan parametreler için Bölüm 4.5'de belirtilen yöntemler uygulanmıştır. Tablo 5.3'de Grup 2 veri kümesine uygulanan yöntemlerin ürettiği sonuçlara ait karışıklık matrisi verilmiştir.

Tablo 5.3: Grup 2 veri kümesi modelleme yöntemlerine göre karışıklık matrisi

#	Classifier		TP	FP	FN	TN
1	bayes	HNB	88	15	17	91
2	bayes	WAODE	84	19	24	84
3	bayes	AODE	84	19	24	84
4	meta	Dagging	81	22	21	87
5	functions	RBFNetwork	79	24	20	88
6	trees	LMT	79	24	20	88
7	functions	Simple Logistic	79	24	20	88

8	meta	Classification Via Regression	80	23	23	85
9	lazy	LBR	80	23	26	82
10	rules	Decision Table	79	24	27	81

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

Karışıklık matrisi incelendiğinde 179 doğru tahminleme ile HNB yöntemi yüzde seksen beşe yakın bir başarı ile ilk sırada yer almaktadır. Tablo 5.4'de grup 2 veri kümesi modelleme sonuçları gösterilmektedir.

Tablo 5.4: Grup 2 veri kümesi modelleme sonuçları

#	Classifier	Accuracy	RMSE	Sensitivity	Specificity	Precision
1	HNB	84.83	0.3758	0.85	0.85	0.85
2	WAODE	79.62	0.3764	0.80	0.80	0.80
3	AODE	79.62	0.3825	0.80	0.80	0.80
4	Dagging	79.62	0.385	0.80	0.80	0.80
5	RBF Network	79.15	0.3877	0.79	0.79	0.79
6	LMT	79.15	0.4019	0.79	0.79	0.79
7	Simple Logistic	79.15	0.4019	0.79	0.79	0.79
8	Classification Via Regression	78.20	0.4002	0.78	0.78	0.78
9	LBR	76.78	0.3996	0.77	0.77	0.77
10	Decision Table	75.83	0.4141	0.76	0.76	0.76

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

HNB algoritması doğruluk oranına bakıldığında en yakın yönteme göre yüzde beş daha iyi sonuç vermiştir. Aynı zamanda RMSE (Root Mean Squared Error) en düşük olan yöntem yine HNB'dir.

Genel gruplama sonucu elde edilen Grup 3 veri kümesi için Tablo 4.20'de belirtilmiş olan parametreler için Bölüm 4.5'de belirtilen yöntemlere göre elde edilen karışıklık matrisi Tablo 5.5'de gösterilmektedir.

Tablo 5.5: Grup 3 veri kümesi modelleme yöntemlerine göre karışıklık matrisi

#	Classifier		TP	FP	FN	TN
1	bayes	HNB	78	25	21	87
2	bayes	AODE	79	24	22	86

3	lazy	LBR	78	25	23	85
4	bayes	WAODE	79	24	25	83
5	meta	Classification Via Regression	74	29	21	87
6	meta	Dagging	81	22	28	80
7	trees	LMT	72	31	20	88
8	functions	Simple Logistic	72	31	20	88
9	functions	RBFNetwork	77	26	26	82
10	rules	Decision Table	64	39	23	85

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

HNB ve AODE algoritmaları 165 doğru tahminleme ve yüzde yetmiş sekiz başarı ile ilk sıralarda bulunmaktadırlar. Tablo 5.6'da Grup 3 modelleme sonuçlarına yer verilmiştir.

Tablo 5.6: Grup 3 veri kümesi modelleme sonuçları

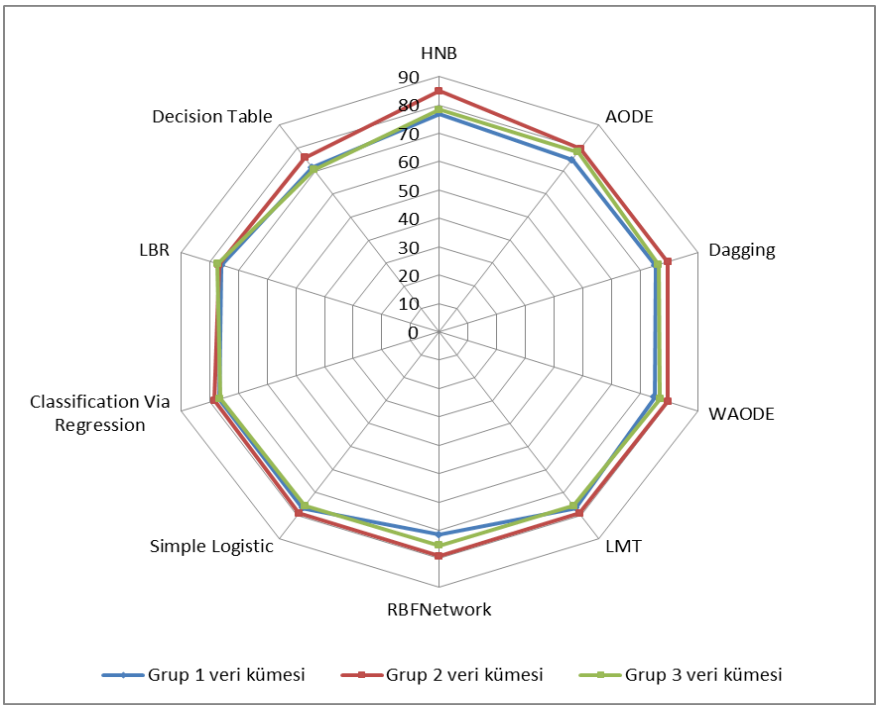
#	Classifier	Accuracy	RMSE	Sensitivity	Specificity	Precision
1	HNB	78.20	0.3981	0.78	0.78	0.78
2	AODE	78.20	0.4003	0.78	0.78	0.78
3	LBR	77.25	0.4077	0.77	0.77	0.77
4	WAODE	76.78	0.3976	0.77	0.77	0.77
5	Classification Via Regression	76.30	0.4124	0.76	0.76	0.76
6	Dagging	76.30	0.4136	0.76	0.76	0.76
7	LMT	75.83	0.4088	0.76	0.76	0.76
8	Simple Logistic	75.83	0.4088	0.76	0.76	0.76
9	RBFNetwork	75.36	0.4141	0.75	0.75	0.75
10	Decision Table	70.62	0.437	0.71	0.70	0.71

Kaynak: Bu tablo Serap ERKUŞ tarafından hazırlanmıştır.

HNB ve AODE algoritmaları doğruluk bakımından aynı oranda başarılı olsalar da RMSE (Root Mean Squared Error) en düşük olan HNB diğer iki veri kümesinde olduğu gibi en başarılı yöntem olmuştur.

Tüm veri kümelerine uygulanan yöntemlerin doğruluk oranlarını ele alınarak hazırlanmış olan Şekil 5.1'de veri kümelerinin başarısı karşılaştırılmaktadır.

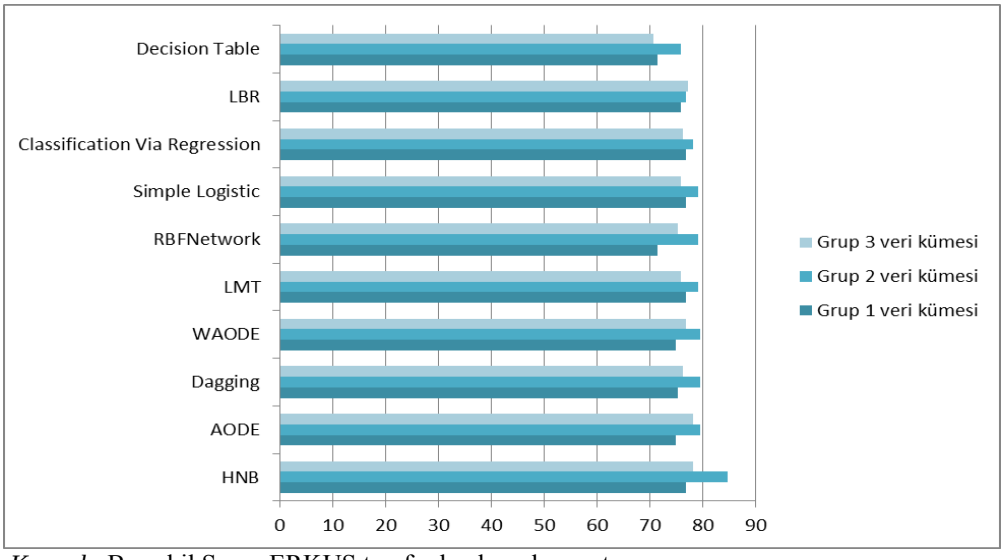
Şekil 5.1: Veri kümelerinin başarı durumu



Kaynak: Bu şekil Serap ERKUŞ tarafından hazırlanmıştır.

Her bir yöntemin veri kümeleri üzerindeki başarı durumu Şekil 5.2'de görselleştirilmiştir.

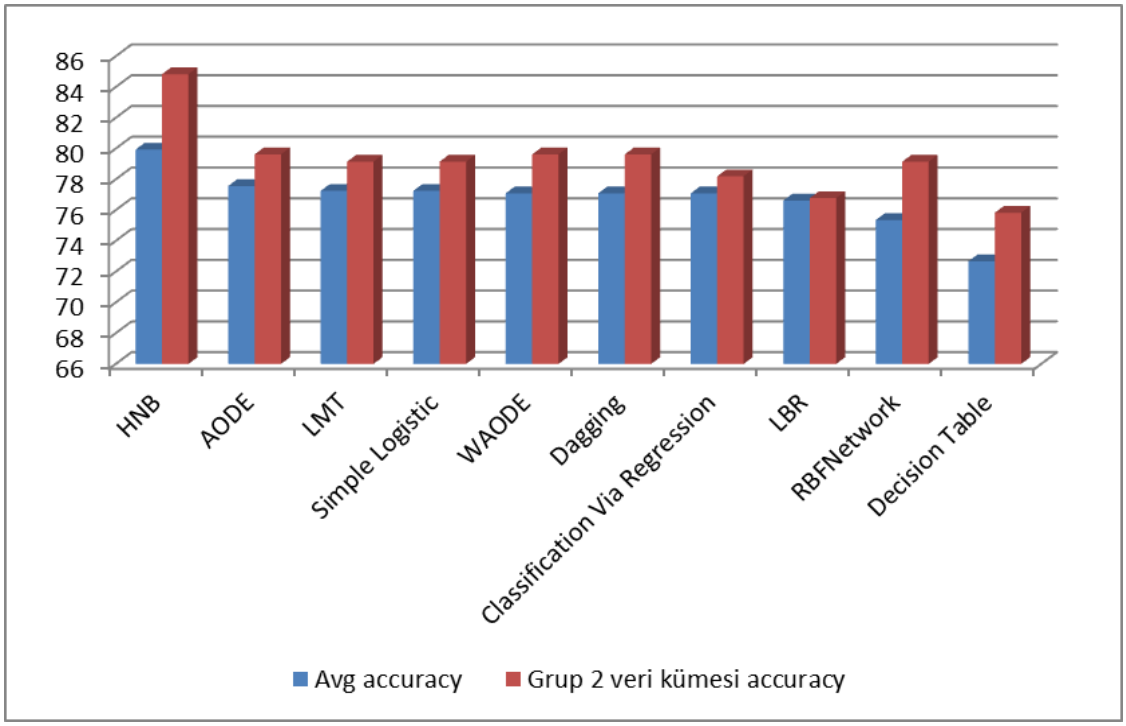
Şekil 5.2: Veri kümeleri üzerinde uygulanan yöntemlerin başarı durumu



Kaynak: Bu şekil Serap ERKUŞ tarafından hazırlanmıştır.

Her bir algoritmanın üç veri kümesi üzerindeki başarısının ortalaması bulunmuş ve Grup 2 veri kümesinin bu ortalamaya göre doğruluk durumu Şekil 5.3'de gösterilmiştir.

Şekil 5.3: Ortalama başarı ile Grup 2 veri kümesi başarı karşılaştırması



Kaynak: Bu şekil Serap ERKUŞ tarafından hazırlanmıştır.

6. TARTIŞMA ve SONUÇ

Bu çalışmada, kalp ve damar hastalıkları teşhisi sırasında anjiyografi gibi zorlu yöntemlere başvurmadan tanı hakkında ağırlıklı bir fikir oluşmasını sağlayabilecek bir model oluşturmak hedeflenmiştir.

Çalışmada KVH teşhisi konmuş ve KVH bulunmayan toplam 604 adet hastaya ait biyokimya sonuçları veri olarak kullanılmıştır. Her iki veri içerisinde otuz ortak parametre bulunmaktadır. Bu parametreler incelenerek çalışmaya katkı sağlayabilecek yirmi parametre kapsam içerisine alınmıştır. Her bir parametre için referans değer aralıkları belirlenmiştir. Veri içerisinde bulunan sayısal ifadeler referans değerlerine göre üç faklı gruplama yöntemi belirlenerek dönüştürülmüş ve sonuç olarak üç farklı veri kümesi elde edilmiştir.

Model oluşturma sırasında kullanılacak parametrelerin belirlenmesi aşamasında InfoGainAttributeEval yöntemi seçilmiş olsa da beş farklı yöntem (ClassifierSubsetEval, CfsSubsetEval, FilteredSubsetEval, GainRatioAttributeEval, ReliefFAttributeEval) daha kullanılarak parametreler nitelik seçim işlemine tabi tutulmuş ve çıkan sonuçlardaki parametre sıralamalarının ortalamaları da göz önünde bulundurularak InfoGainAttributeEval yöntemi sonuçları ile karşılaştırılarak en uygun parametre grubu belirlenmiştir.

Cinsiyet parametresi hangi nitelik seçim yöntemi kullanılırsa kullanılsın, ilk sırada yer almıştır. Bu da KVH tanısı konulabilmesi için gereken bilgilerin cinsiyete göre kesinlikle değişiklik gösterdiğini ifade etmektedir.

Beden kitle indeksi ise nitelik seçim işlemlerinde yüzde seksen oranında ikinci sırada yer almıştır. Bu da çağımızın kontrol edilmesi en zor problemlerinden biri olan obezitenin KVH üzerindeki etkisini açıkça göstermektedir.

Elde edilmiş olan parametre grubu başlıca KVH risk faktörleri ile karşılaştırıldığında yaş, cinsiyet, sigara kullanımı, diyabet, obezite ve hiperlipidemi parametrelerinin bu çalışmada KVH hastalığı belirleyici etkileri olduğu sonucu ortaya çıkmaktadır. Fakat hiper tansiyon ve aile öyküsü parametreleri risk faktörlerinde yer almasına rağmen, çalışılan veri kümelerinde, oluşturulan modelin başarısını düşürdüğü saptanmıştır.

Parametrelerin belirlenmesi sonrasında veri kümelerine, her bir sınıflandırma kategorisinden (bayes, lazy, meta, trees, functions, rules) en az bir yöntem olacak şekilde on farklı algoritma (HNB, AODE, LBR, WAODE, Classification Via Regression, Dagging, LMT, Simple Logistic, RBFNetwork, Decision Table) uygulanmıştır. Uygulanan algoritmaların doğruluk oranları incelendiğinde tüm veri kümeleri üzerinde **HNB** algoritmasının en başarılı algoritma olduğu görülmektedir. HNB algoritması ile elde edilen yaklaşık yüzde seksen beş (84,8) doğruluk oranının uzman doktor görüşü doğrultusunda tanı koymak için oldukça yeterli olduğu sonucuna ulaşılmıştır.

Veri kümeleri modelleme sonuçları incelendiğinde doktor görüşü ile referans değer aralıkları değerlendirilerek gruplanmış olan **Grup 2 Veri Kümesi** en yüksek başarı oranını sağlayan veri kümesi olmuştur. Yalnızca HNB algoritması ile değil neredeyse tüm algoritmalarda en başarılı veri kümesi olduğu gözlemlenmiştir. Bu da veriyi tanımanın ve nasıl işlenmesi gerektiğini bilmenin veri madenciliği açısından önemini ortaya koymuştur.

Çalışma sonucunda cinsiyet, yaş, beden kitle indeksi, sigara kullanımı, diyabet, HDL, LDL, HGB, BUN, WBC, PLT, Ürik Asit ve PDW parametreleri HNB algoritması ile çalışıldığında KVH tanısı koymaya yetecek doğrulukta bir model oluşturulabileceği ortaya konmuştur.

Veri kümesi içerisinde genç yaşa ait (kırk yaş altı) yalnızca 26 kayıt bulunmaktadır. Bu da çalışmanın genç yaşta KVH riski taşıyan kişiler için yönlendirici olabilirliğini tartışılır yapmaktadır. Aile öyküsü parametresinin bu veri grubu için başarıyı düşürme

sebebi, kişilerin yaşı sebebiyle aile öyküsünden çok hastalığa dair diğer risklerin ortaya çıkması ile modele etkisinin azalması olarak düşünülmüştür.

Gelecekte, hasta biyokimya sonuçları elde edilir edilmez, modelin otomatik olarak uygulanması ile tanı koyma sırasında doktora yardımcı olacak bir yazılım yapılması bu çalışmanın devamı olarak başlatılabilir. KVH hastalığında teşhis sırasında kullanılan zorlu yöntemler olmadan bu gibi bir tanı yöntemi sayesinde, hem hastaya kolay tanı ile daha hızlı tedavi uygulanmaya başlanacak, hem de uygulanan teşhis yöntemleri maliyetleri düşürülerek, hasta, devlet ve sigorta kurumları tarafından teşhis için ödenen giderlerde düşüş yaratılacaktır.

Çalışma, bu alanda yapılabilecek çalışmalara ışık tutması açısından doktor, hasta ve giderlerin düşürülmesi ile birçok farklı kuruma ve hastaya fayda sağlayacak bir potansiyele sahiptir.

KAYNAKÇA

Kitaplar

Argüden, Y. & Erşahin, B., 2008. *Veri madenciliği veriden bilgiye, masraftan değere*. Arge Danışmanlık.

Cios, K. J., Pedrycz, W., Swiniarski, R. W. & Kurgan, L. A., 2007. *A Knowledge discovery approach*. New York: Springer Science+Business Media.

Gorunescu, F., 2011. *Data mining: concepts, models and techniques*. Berlin: Springer Berlin Heidelberg.

Han, J. & Kamber, M., 2006. *Data mining concepts and techniques second edition*.2. San Francisco: Morgan Kaufmann Publishers.

Jain, A. K. & Dubes, R. C., 1988. *Algorithms for clustering data*. Englewood Cliffs, N.J.: Prentice-Hall.

Larose, D. T., 2005. *Discovering knowledge in data: an ıntroduction to data mining*. John Wiley & Sons, Inc.

Olson, D. L. & Delen, D., 2008. *Advanced data mining techniques*. Berlin:Springer.

Özkan, Y., 2013. *Veri madenciliği yöntemleri*, 2.Baskı. İstanbul:Papatya.

Rosengren, A., Perk, J. & Dallongeville, J., 2009. Prevention of cardiovascular disease. *ESC textbook of cardiovascular medicine*. New York: Oxford University Press, ss.403-435.

Santos, M. F. & Azevedo, C., 2005. *Data mining – descoberta de conhecimento em bases de dados*. FCA Publisher.

Yin, Y., Kaku, I., Tang, J., Zhu, J.M. *Data mining concepts, methods and applications in management and engineering design*. London:Springer

Süreli Yayınlar

Abacı, A., 2011. Kardiyovasküler risk faktörlerinin ülkemizdeki durumu. *Türk Kardiyol Dern Arş - Arch Turk Soc Cardiol*. **39**(4). ss.1-5

Agrawal, R. & Shafer J. C., 1996. Parallel Mining of Association Rules. *IEEE Transactions on Knowledge and Data Engineering.* **8** (6), ss.962-969

Alizadehsania, R., Habibia, J., Hosseinia, M. J., Mashayekhia, H., Boghratia, R. , Ghandehariouna, Bahadorianb, A. B. & Alizadeh, Sani Z., 2013. A data mining approach for diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine*. **111**(1). ss.52-61.

Amini, L., Azarpazhouh, R., Farzadfar, M. T., Mousavi, S. A., Jazaieri, F., Khorvash, F., Norouzi, R. & Toghianfar, N., 2013. Prediction and control of stroke by data mining.*International Journal of Preventive Medicine (IJPM)*. **4** (2). ss.245-249.

Chaurasia, V. & Pal, S., 2013. Early prediction of heart diseases using data mining techniques. *Caribbean Journal of Science and Technology*. **1**. ss.208-217

Delen, D., Walker & G., Kadam, A., 2005. Predicting breast cancer survivability:a comparison of three data mining methods. *Artificial Intelligence in Medicine*.**34** (2), ss.113-127

Dogan, S. & Turkoglu, I., 2008. Extraction association rules from the biochemistry parameters for diagnosing hyperthyroidi, *IEEE*. ss.1.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. The KDD process for extracting useful knowledge from volumes of data.*Communications of The Acm*. **39** (11), ss.27-34.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth P., 1996. From data mining to knowledge discovery indatabases. *AI Magazine*. **17** (3), ss.44-45

Fayyad, U. & Stolorz, P., 1997. Data mining and KDD: promise and challenges. *Future Generation Computer Systems*. **13** (2-3), ss.99-115.

Frawley, W. J., Piatetsky-Shapiro, G. & Matheus, C. J., 1992. Knowledge discovery in databases:an overview. *AI Magazine*. **13** (3), s.67

Gargano, M. L., & Raggad, B. G., 1999. Data mining - a powerful information creating tool. *OCLC Systems and Services*. **15**(2), 81-90.

Güleç, S., 2009. Kalp damar hastalıklarında global risk ve hedefler, *Türk Kardiyol Dern Arş - Arch Turk Soc Cardiol*. **37**(2), ss.1-10.

Gülel, O., 2012. Kardiyovasküler risk faktörleri. *Deneysel ve Klinik Tıp Dergisi - Journal of Experimental and Clinical Medicine*. **29**(3), ss.107-116

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H., 2009. The WEKA data mining software: an update. *SIGKDD Explorations*. **11** (1), ss.10–18.

Hong, S. J., 1997. Data mining. *Future Generation Computer Systems*. **13**,(2-3), ss.95-97

Jacobs, P., 1999. Data mining: what general managers need to know. *Harvard Management Update*, **4**(10), ss.8.

Jacquez, G.M., Grimson, R. & Waller, L.A.,1996. The analysis of disease clusters, part ıı: ıntroduction to 198 techniques. *Infect Control Hosp Epid*,**17**, ss.385-97

Karahoca, A. & Tunga, M. A., 2012. Dosage planning for type 2 diabetes mellitus patients using indexing HDMR. *Expert Systems with Applications*. **39**(8), ss.7207-7215.

Kitler, R. & Wang, W., 1999. The emerging role of data mining. *Solid State Technology*, **42**(11), ss.45.

Koyuncugil, A. S. & Özgülbaş, N., 2009. Veri madenciliğinin tıp ve sağlık alanında kullanımı. *Bilişim Teknolojileri Dergisi*. **2 (**2).

Kumar, D. S., Sathyadevi, G. & Sivanesh, S., 2011. Decision support system for medical diagnosis using data mining .*International Journal of Computer Science Issues.* **8** (3), ss.147-153.

Kültürsay, H., 2011. Kardiyovasküler hastalık riski hesaplama yöntemleri. *Türk Kardiyol Dern Arş - Arch Turk Soc Cardiol*. **39** (4), ss.6-13.

Neethu, B., 2012. Classification of intrusion detection dataset using machine learning approaches. *International Journal of Electronics and Computer Science Engineering.* **1**, ss.1044-1051.

Ni, H., Coady, S., Rosamond, W., Folsom, A.R., Chambless, L., Russell, S.D. & Sorlie, P.D., 2009. Trends from 1987 to 2004 in sudden death due to coronary heart disease. *The Atherosclerosis Risk In Communities (A.R.I.C.) Study. Am. Heart J.* **157**, ss.46-52.

Panda, M. & Patra, M.R., 2008. A Comparative study of data mining algorithms for network intrusion detection. *IEEE*. ss.504-507.

Patil, B.M., Joshi, R.C., Toshniwal, D. & Biradar, S., 2011. A new approach: role of data mining in prediction of survival of burn patients. *Journal of Medical Systems*. **35** (6), ss. 1531-1542.

Rothwell, P.M., Coull, A.J., Giles, M.F., Howard, S.C., Silver, L.E., Bull, L.M., Gutnikov, S.A., Edwards, P., Mant, D., Sackley, C.M., Farmer, A., Sandercock, P.A., Dennis, M.S., Warlow, C.P., Bamford, J.M. & Anslow, P., 2004. Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004. *Oxford Vascular Study*. *Lancet*. **363**(9425), ss.1925-1933.

Shouman, M., Turner, T. & Stocker, R., 2012. Using data mining techniques in heart disease diagnosis and treatment. *Japan-Egypt Conference on Electronics, Communications and Computers.* ss.189-193

Sundell, J., 2005. Obesity and diabetes as risk factors for coronary artery disease:from the epidemioligical aspect to the initial vascular mechanisms. *Diabetes, Obesity and Metabolism*. **7**, ss.9-20.

Thom, T., Haase, N., Rosamond, W., Howard, V.J., Rumsfeld, J., Manolio, T., Zheng, Z.J., Flegal, K., O'Donnell, C., Kittner, S., Lloyd-Jones, D., Goff, D.C.Jr., Hong, Y., Adams, R., Friday, G., Furie, K., Gorelick, P., Kissela, B., Marler, J., Meigs, J., Roger, V., Sidney, S., Sorlie, P., Steinberger, J., Wasserthiel-Smoller, S., Wilson, M. & Wolf, P., 2006. American heart association statistics committee and stroke statistics subcommittee. *Heart Disease and Stroke Statistics-2006 update*: *A report from the American heart association statistics committee and stroke statistics subcommittee. Circulation*. **113**(14), ss.85-151.

Webb, G., Boughton, J. & Wang, Z., 2005. Not so naive bayes: aggregating one-dependence estimators. *Machine Learning*. **58**(1), ss.5-24.

Yavuz, R., Yavuz, D. & Tontuş, H. Ö., 2013. Artan mortalite ve morbidite nedeni olarak kardiyovasküler risk faktörlerine sistematik yaklaşım. *Journal of Experimental and Clinical Medicine Deneysel ve Klinik Tıp Dergisi*. **30**(1), ss.47-53.

Yeh, D. Y., Cheng, C. H. & Chen, Y. W., 2011. A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*. **38** (7), ss.8970–8977.

Zhang, H., Jiang, L., Wang, D. & Cai, Z., 2012. Weighted average of one-dependence estimators. *Journal of Experimental & Theoretical Artificial Intelligence*. **24** (2), ss.219-230.

Diğer Yayınlar

Acar Şaylan, Ç. (2013). Böbrek nakli geçirmiş hastalarda akıllı yöntem tabanlı yeni öznitelik seçme algoritması geliştirilmesi. *Yüksek Lisans Tezi*. İstanbul: Kadir Has Üniversitesi Fen Bilimleri Enstitüsü.

Albayrak, S., 2012, Veri madenciliği sınıflama ve kümeleme yöntemleri, https://www.ce.yildiz.edu.tr/personal/songul/file/332/Veri+Madencili%C4%9Fi-S%C4%B1n%C4%B1flamaKumeleme.ppt. [erişim tarihi 28 Ekim 2014]

Anonymous 1999, CRISP-DM methodology brings data mining to the masses, Intelligent Systems Report, [Online ProQuest Central], **16** (8), ss.10. http://search.proquest.com/docview/219796754?accountid=15407. [erişim tarihi 28 Ekim 2014]

Camacho, R. & Borges, J. L., 2005, Extracção de conhecimento (LEIC, MEI e PRODEI) introdução à disciplina de data mining. Porto. http://paginas.fe.up.pt/~ec/files_0506/slides/01_IntroDM.pdf. [erişim tarihi 15 Ekim 2014], ss.15.

Chapman, P., 1999,The CRISP-DM user guide. http://lyle.smu.edu/~mhd/8331f03/crisp.pdf. [erişim tarihi 28 Ekim 2014]

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R., 2000. CRISP-DM 1.0 Step-by-step data mining guide. http://www.statoo.com/CRISP-DM.pdf. [erişim tarihi 28 Ekim 2014]

CRISP, 2011, http://blog.visilabs.com/crisp/. [erişim tarihi 28 Ekim 2014]

Çıngı, H., 2008. Veri madenciliğine giriş. http://yunus.hacettepe.edu.tr/~hcingi/ist376a/6Bolum.doc. [erişim tarihi 29 Ekim 2014]

Demirel, B., (2008). Meme kanseri tedavi yöntemlerinin veri madenciliği ile belirlenmesi.*Yüksek Lisans Tezi*. Isparta: Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü.

Doğan, Ş., (2007). Veri madenciliği kullanarak biyokimya verilerinden hastalık teşhisi.*Yüksek Lisans Tezi*. Elazığ: Fırat Üniversitesi Fen Bilimleri Enstitüsü.

Ercan Toptaner, N., (2013). Menopoz dönemindeki kadınlara uygulanan koroner kalp hastalıklarından korunmaya yönelik programın kalp sağlığı üzerine etkileri. *Doktora Tezi*. İstanbul: Marmara Üniversitesi Sağlık Bilimleri Enstitüsü.

Florence.com.tr, 2014. Kardiyovasküler Hastalıklarda Non İnvaziv Tanı Yöntemleri. http://www.florence.com.tr/non-invaziv-tani-yontemleri.html. [erişim tarihi 02 Kasım 2014]

Hall, M., Weka attribute selection class InfoGainAttributeEval, http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html. [erişim tarihi 09 Kasım 2014]

Hematoloji Uzmanlık Derneği, Derin Ven Trombozu, 2014, http://kanhastaliklari.net/icerik.php?id=180&alt_id=153. [erişim tarihi 31.10.2014]

Işıklı, B, 2009. https://burakisikli.wordpress.com/tag/veri-madenciligi/. [erişim tarihi 29 Ekim 2014]

Kollios, G., 2005, Cluster Analysis, http://www.cs.bu.edu/fac/gkollios/ada05/LectNotes/lect28-05.pdf .[erişim tarihi 28 Ekim 2014]

Lab Tests Online, 2011, http://labtestsonline.org.tr/understanding/conditions/cvd/. [erişim tarihi 30.10.2014]

Mishra, B.K., Lakkadwala, P. & Shrivastava, N.K., 2013. Novel Approach to Predict Cardiovascular Disease Using Incremental SVM, *2013 International Conference on Communication Systems and Network Technologies*, ss.55-59.

Oracle, 2014, Oracle Data Mining Concepts, OracleDatabase Online Documentation Library, http://docs.oracle.com/database/121/DMCON/intro_basics.htm#DMCON620 .[erişim tarihi 28 Ekim 2014]

Özekes, S., (2006). Tıbbi görüntülemede bilgisayar destekli tespit. *DoktoraTezi*. İstanbul: Marmara Üniversitesi Fen Bilimleri Enstitüsü.

Saglikpark.com, 2008. http://www.saglikpark.com/yazdir/kan_kolesterol_duzeyi_neden_yukselir_.htm. [erişim tarihi 01 Kasım 2014]

SPSS, http://sbu.saglik.gov.tr/Ekutuphane/kitaplar/biyoistatistik%20(6).pdf. [erişim tarihi 30 Ekim 2014]

Tahminciler, E. (2014). Erythromcin ilacının yan etkilerinin araştırılması üzerine veri madenciliği çalışması.*Yüksek Lisans Tezi*. İstanbul: Okan Üniversitesi Fen Bilimleri Enstitüsü.

Tavsanoglu, L., 2014. Prof. Dr. Altan Onat Türk insanının sağlık davranışlarını inceledi. *Cumhuriyet Gazetesi*, [online gazete] 5 Ocak 2014, http://www.cumhuriyet.com.tr/koseyazisi/25863/Prof._Dr._Altan_Onat_Turk_insaninin_saglik_davranislarini_inceledi_.html

TUİK Türkiye İstatistik Kurumu, Ölüm nedeni istatistikleri 2013, 2014. http://www.tuik.gov.tr/PreHaberBultenleri.do?id=16162. [erişim tarihi 31.10.2014]

Turkiyedoktorlari.com. Kardiyovasküler Risk Faktörleri. http://www.turkiyedoktorlari.com/hastalik-rehberi/branslar/kalp-damar/kalp-/383-kardiyovaskueler-risk-faktoerleri-.html. [erişim tarihi 01 Kasım 2014]

Türk Kalp ve Damar Cerrahisi Derneği, http://www.tkdcd.org/public/uploads/files/pdf/saglikli_yasam/koroner_arter_hastaliklari.pdf.

WHO (World Health Organization), The Top 10 Causes of Death, 2012, http://www.who.int/mediacentre/factsheets/fs310/en [erişim tarihi 15 Ekim 2014], s.3.

WHO (World Health Organization), Cardiovascular diseases (CVDs), 2013, http://www.who.int/mediacentre/factsheets/fs317/en/ [erişim tarihi 30 Ekim 2014]

WHO (World Health Organization), Cardiovascular disease mortality: Age-standardized death rate per 100.000 population both sexes 2012 , 2014, http://gamapserver.who.int/mapLibrary/Files/Maps/Global_NCD_mortality_CVD_2012.png. [erişim tarihi 31.10.2014]

WHO (World Health Organization), Cardiovascular disease mortality, 2014. http://gamapserver.who.int/gho/interactive_charts/ncd/mortality/cvd/atlas.html. [erişim tarihi 31.10.2014]

Zhang, H., Jiang, L., & Su, J. 2005. Hidden Naive Bayes. *In Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA. Cambridge. MA. London. AAAI Press. MIT Press. **20**(2) ss.919

EKLER

EK 1: DATA_OLUSTUR PROSEDÜRÜ

```
CREATE PROCEDURE [dbo].[DATA_OLUSTUR]
AS
BEGIN

IF OBJECT_ID (N'DATA', N'U') IS NOT NULL
DROP TABLE DATA

IF OBJECT_ID (N'tempdb..##DATA', N'U') IS NOT NULL
DROP TABLE ##DATA
SELECT IDENTITY(INT,1,1) AS ID ,
          A.*
INTO ##DATA
FROM(
        SELECT AD_SOYAD, YAS,
            UPPER(CINSIYET) AS CINS,
            ISNULL(AILE_OYKUSU,0) AS AO,
            ISNULL(SIGARA,0) AS SGR,
            BMI,
            ISNULL(DM,0) AS DM,
            HGB,
            HDL,
            ISNULL(HL,0) AS HL,
            ISNULL(HT,0) AS HT,
            PDW,
            PLT,
            RDW,
            WBC,
            TRIGLISERID ,
            KREATININ,
            LDL,
            BUN,
            TK,
            UA_1 AS UA,
            1 AS HASTA
        FROM HASTA
        UNION
        SELECT AD_SOYAD, YAS,
            UPPER(CINSIYET) AS CINS,
            ISNULL(AILE_OYKUSU,0) AS AO,
            ISNULL(SIGARA,0) AS SGR,
            BMI,
            ISNULL(DM,0) AS DM,
```

```
        HGB,
        HDL,
        ISNULL(HL,0) AS HL,
        ISNULL(HT,0) AS HT,
        PDW,
        PLT,
        RDW,
        WBC,
        TRIGLISERID,
        KREATININ,
        LDL,
        BUN,
        TK,
        UA,
        0
    FROM HASTA_DEGIL
    ) AS A

    IF OBJECT_ID (N'tempdb..##GRUPLAR', N'U') IS NOT NULL
    DROP TABLE ##GRUPLAR

    CREATE TABLE ##GRUPLAR
    (GRUP_NO TINYINT)

    INSERT INTO ##GRUPLAR
    SELECT DISTINCT GRUP_NO FROM REFERANS_DEGERLER

    UPDATE ##DATA
    SET TK = (SELECT AVG(TK) FROM ##DATA WHERE TK IS NOT NULL)
    WHERE TK IS NULL

    UPDATE ##DATA
    SET LDL = (SELECT AVG(LDL) FROM ##DATA WHERE LDL IS NOT
NULL)
    WHERE LDL IS NULL

    UPDATE ##DATA
    SET HDL = (SELECT AVG(HDL) FROM ##DATA WHERE HDL IS NOT
NULL)
    WHERE HDL IS NULL

    UPDATE ##DATA
    SET TRIGLISERID = (SELECT AVG(TRIGLISERID) FROM ##DATA
WHERE TRIGLISERID IS NOT NULL)
    WHERE TRIGLISERID IS NULL

    UPDATE ##DATA
```

```
	SET BUN = (SELECT AVG(BUN) FROM ##DATA WHERE BUN IS NOT
NULL)
	WHERE BUN IS NULL

	UPDATE ##DATA
	SET KREATININ = (SELECT AVG(KREATININ) FROM ##DATA WHERE
KREATININ IS NOT NULL)
	WHERE KREATININ IS NULL

	UPDATE ##DATA
	SET WBC = (SELECT AVG(WBC) FROM ##DATA WHERE WBC IS NOT
NULL)
	WHERE WBC IS NULL

	UPDATE ##DATA
	SET HGB = (SELECT AVG(HGB) FROM ##DATA WHERE HGB IS NOT
NULL)
	WHERE HGB IS NULL

	UPDATE ##DATA
	SET PLT = (SELECT AVG(PLT) FROM ##DATA WHERE PLT IS NOT
NULL)
	WHERE PLT IS NULL

	UPDATE ##DATA
	SET PDW = (SELECT AVG(PDW) FROM ##DATA WHERE PDW IS NOT
NULL)
	WHERE PDW IS NULL

	UPDATE ##DATA
	SET RDW = (SELECT AVG(RDW) FROM ##DATA WHERE RDW IS NOT
NULL)
	WHERE RDW IS NULL

	SELECT ID,
		AD_SOYAD,
		GRUP_NO,
		dbo.GRUP_GETIR('YAS',YAS,CINS,GRUP_NO) AS YAS,
		CINS AS CINSIYET,
		dbo.GRUP_GETIR('AILE_OYKUSU',AO,CINS,GRUP_NO) AS
AILE_OYKUSU,
		dbo.GRUP_GETIR('SIGARA',SGR,CINS,GRUP_NO) AS SIGARA,
		dbo.GRUP_GETIR('BMI',BMI,CINS,GRUP_NO) AS BMI,
		dbo.GRUP_GETIR('DM',DM,CINS,GRUP_NO) AS DM,
		dbo.GRUP_GETIR('HGB',HGB,CINS,GRUP_NO) AS HGB,
		dbo.GRUP_GETIR('HDL',HDL,CINS,GRUP_NO) AS HDL,
		dbo.GRUP_GETIR('HL',HL,CINS,GRUP_NO) AS HL,
		dbo.GRUP_GETIR('HT',HT,CINS,GRUP_NO) AS HT,
```

```
                dbo.GRUP_GETIR('PDW',PDW,CINS,GRUP_NO) AS PDW,
                dbo.GRUP_GETIR('PLT',PLT,CINS,GRUP_NO) AS PLT,
                dbo.GRUP_GETIR('RDW',RDW,CINS,GRUP_NO) AS RDW,
                dbo.GRUP_GETIR('WBC',WBC,CINS,GRUP_NO) AS WBC,
               dbo.GRUP_GETIR('TRIGLISERID',TRIGLISERID,CINS,GRUP_NO)
AS TRIGLISERID,
                dbo.GRUP_GETIR('KREATININ',KREATININ,CINS,GRUP_NO)
AS KREATININ,
                dbo.GRUP_GETIR('LDL',LDL,CINS,GRUP_NO) AS LDL,
                dbo.GRUP_GETIR('BUN',BUN,CINS,GRUP_NO) AS BUN,
                dbo.GRUP_GETIR('TK',TK,CINS,GRUP_NO) AS TK,
                dbo.GRUP_GETIR('UA',UA,CINS,GRUP_NO) AS UA,
                dbo.GRUP_GETIR('HASTA',HASTA,CINS,GRUP_NO) AS CLASS
        INTO DATA
        FROM ##DATA
        CROSS JOIN ##GRUPLAR
END
```

EK 2: GRUP_GETIR FONKSİYONU

```
CREATE FUNCTION [dbo].[GRUP_GETIR]
(
	@PARAMETRE NVARCHAR(50),
	@DEGER FLOAT,
	@CINS NVARCHAR(1),
	@GRUP_NO INT
)
RETURNS NVARCHAR(50)
AS
BEGIN
	DECLARE @GRUP NVARCHAR(50)
	SELECT @GRUP = GRUP
	FROM REFERANS_DEGERLER
	WHERE PARAMETRE = @PARAMETRE
	AND @DEGER BETWEEN ALT_SINIR AND UST_SINIR
	AND CINSIYET LIKE '%'+@CINS+'%'
	AND GRUP_NO = @GRUP_NO
	RETURN @GRUP
END
```

EK 3: DOSYA_ICERIGI_OLUSTUR PROSEDÜRÜ

```
CREATE PROCEDURE [dbo].[DOSYA_ICERIGI_OLUSTUR]
        @GRUP_NO INT
AS
BEGIN
DECLARE @ALAN VARCHAR(50),
            @DOSYA_ICERIGI VARCHAR(8000)
 DECLARE CURSOR_ALAN CURSOR FOR
        select C.NAME
        FROM SYS.columns C
        INNER JOIN SYS.objects O
        ON O.object_id = C.object_id
        WHERE O.name = 'DATA'
        AND C.NAME NOT IN ('ID','GRUP_NO','AD_SOYAD')
OPEN CURSOR_ALAN
FETCH NEXT FROM CURSOR_ALAN INTO @ALAN
        WHILE @@FETCH_STATUS = 0
   BEGIN
                CREATE TABLE #ALAN_ICERIGI(ICERIK VARCHAR(8000))
                SET @DOSYA_ICERIGI = ISNULL(@DOSYA_ICERIGI +
CHAR(10) ,'@relation hasta' + CHAR(10) + char(10) ) +'@attribute ' + @ALAN + ' {'
                DECLARE @SQL VARCHAR(500) = 'DECLARE @X
VARCHAR(8000) SELECT @X = ISNULL(@X+", ","")+ICERIK FROM (SELECT
DISTINCT '+@ALAN+' AS ICERIK FROM DATA WHERE GRUP_NO =  ' +
CAST(@GRUP_NO AS VARCHAR(1)) + ') A SELECT @X'
                PRINT @SQL
                INSERT INTO #ALAN_ICERIGI
                EXEC(@SQL)
                SELECT @DOSYA_ICERIGI = @DOSYA_ICERIGI + ICERIK + '}'
                FROM #ALAN_ICERIGI
                DROP TABLE #ALAN_ICERIGI
   FETCH NEXT FROM CURSOR_ALAN INTO @ALAN
   END
 CLOSE CURSOR_ALAN DEALLOCATE CURSOR_ALAN
SELECT @DOSYA_ICERIGI+ CHAR(10)+CHAR(10)+'@data' + CHAR(10) AS
ICERIK
END
```

EK 4: DOSYA_OLUSTUR KOMUT DİZİSİ

```
using System;
using System.Data;
using Microsoft.SqlServer.Dts.Runtime;
using System.Windows.Forms;

namespace ST_628d86d2b5bb40e68397914d12db6722.csproj
{
   [System.AddIn.AddIn("ScriptMain", Version = "1.0", Publisher = "", Description =
"")]
   public partial class ScriptMain :
Microsoft.SqlServer.Dts.Tasks.ScriptTask.VSTARTScriptObjectModelBase
   {

      #region VSTA generated code
      enum ScriptResults
      {
         Success = Microsoft.SqlServer.Dts.Runtime.DTSExecResult.Success,
         Failure = Microsoft.SqlServer.Dts.Runtime.DTSExecResult.Failure
      };
      #endregion


      public void Main()
      {
         string bugun = DateTime.Now.Year.ToString() +
DateTime.Now.Month.ToString() + DateTime.Now.Day.ToString() +
DateTime.Now.Hour.ToString() +
DateTime.Now.Minute.ToString()+DateTime.Now.Second.ToString() ;
         string dosyaadi = @"d:\tez\arff\Data_Grup1_" + bugun + ".arff";
         Dts.Variables["DosyaAdi"].Value = dosyaadi;
         System.IO.FileStream dosyaislemi = System.IO.File.Create(dosyaadi);
         dosyaislemi.Close();
         System.IO.File.WriteAllText(dosyaadi,
Dts.Variables["DOSYA_ICERIGI"].Value.ToString());
         Dts.TaskResult = (int)ScriptResults.Success;
      }
   }
}
```

ÖZGEÇMİŞ

Adı Soyadı : Serap ERKUŞ

Sürekli Adresi : Eğitim mh. Gelincik sk. Anıt Sitesi D Blok No:3/16 Kadıköy/İstanbul

Doğum Yeri ve Yılı : Silivri 1983

Yabancı Dili : İngilizce

İlk Öğretim : Piri Mehmet Paşa İlkokulu 1994, Nurullah Baldöktü İlköğretim Okulu 1997

Orta Öğretim : Silivri Teknik Lisesi Bilgisayar Yazılım 2001

Ön Lisans : Trakya Üniversitesi Bilgi Teknolojileri ve Programlama 2004

Lisans : Anadolu Üniversitesi İşletme 2008

Yüksek Lisans : Bahçeşehir Üniversitesi

Enstitü Adı : Fen Bilimleri Enstitüsü

Program Adı : Bilgi Teknolojileri

Çalışma Hayatı :
ING Emeklilik-Veri Ambarı ve İş Zekası Kıdemli Uzmanı 2013.12 - halen
Yapı Kredi Emeklilik- Veri Ambarı ve İş Zekası Uzmanı 2012.06 – 2013.11
Sompo Japan Sigorta- Veri Ambarı ve İş Zekası Uzmanı 2012.01 - 2012.05
MBEM Teknoloji-Kıdemli Yazılım Uzmanı 2008.10 – 2011.12
IDB Yazılım Şirketi-Yazılım Uzmanı 2004.10 – 2008.10