AKDENİZ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

Sezgin IRMAK

VERİ MADENCİLİĞİ YÖNTEMLERİ İLE SAĞLIK SEKTÖRÜ VERİTABANLARINDA BİLGİ KEŞFİ: TANIMLAYICI VE KESTİRİMCİ MODEL UYGULAMALARI

Danışman
Doç. Dr. Can Deniz KÖKSAL

İşletme Anabilim Dalı
Doktora Tezi

Antalya, 2009

Akdeniz Üniversitesi
Sosyal Bilimler Enstitüsü Müdürlüğüne,

Sezgin IRMAK'ın, bu çalışması jürimiz tarafından İşletme Anabilim Dalı Doktora Programı tezi olarak kabul edilmiştir.

Başkan : Prof. Dr. Ayşe Kuruüzüm

Üye (Danışmanı) : Doç.Dr.Can Deniz KÖKSAL

Üye : Prof. Dr. Gabil Adilov

Üye : Doç.Dr. Levent Dönmez

Üye : Yrd. Doç. Dr. Murat Tuhan

Onay : Yukarıdaki imzaların, adı geçen öğretim üyelerine ait olduğunu onaylarım.

Tez Savunma Tarihi : 13/11./2009

Mezuniyet Tarihi : 23/11./2009

Prof.Dr. Burhan VARKIVANÇ
Müdür

.....

İ Ç İ N D E K İ L E R

ŞEKİLLER LİSTESİ

Sayfa

TABLOLAR LİSTESİ

Sayfa

KISALTMALAR LİSTESİ

A.B.D.	Amerika Birleşik Devletleri
ACF	AutoCorrelation Function
AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
ARIMA	AutoregRessive Integrated Moving Averages
CART (C&RT)	Classification and Regression Trees
CHAID	Chi-squared Automatic Interaction Detector
CRISP-DM	Cross Industry Standart Process for Data Mining
CRM	Customer Relationships Management
dbo	Database Object
dL	Desilitre
ERP	Enterprise Resource Planing
FCM	Fuzzy C-Means
g	Gram
HgbA1c	Hemoglobin A1c
ICD	International Classification of Diseases
K.B.B.	Kulak, Burun ve Boğaz
KDD	Knowledge Discovery in Databases
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MaxAE	Maximum Absolute Error
MaxAPE	Maximum Absolute Percentage Error
MSDN	Microsoft Developer Network
Norm. BIC	Normalised Bayesian Information Criterion
ODBC	Open Database Connectivity
OLAP	OnLine Analytical Processing
PACF	Partial AutoCorrelation Function
RBFN	Radial Basis Function Network
RMSE	Root Mean Squared Error
S.B.	Sağlık Bakanlığı
SEMMA	Sample, Explore, Modify, Model, Assess
SQL	Structured Query Language

SSMA	SQL Server Migration Assistant
SSK	Sosyal Sigortalar Kurumu
SSE	Sum of Squared Errors
TID	Transaction Identification Number
TL	Türk Lirası
TOAD	Tool for Application Developers
TSDM	Time Series Data Mining
UNOS	United Network for Organ Sharing

ÖZET

Veri madenciliği yöntem ve tekniklerinin sağlık sektörü veritabanlarında kullanımıyla özellikle hastane veya sağlık kurumları yöneticilerinin veya bu alanda politika yapıcılarının öngörü edinmesine yardımcı olabilecek bilgilerin elde edilmesi mümkündür. Bu çalışmada sağlık alanındaki veri madenciliği uygulamalarının hangi adımlardan oluşacağı ortaya konulmuş, bu alanda var olan literatürden teorik ve uygulama bilgisi aktarılmıştır. Uygulama aşamasında ise hali hazırda işleyen bir hastane veritabanında bazı önemli veri madenciliği teknikleri uygulanmış ve sonuçları aktarılmıştır.

Veri madenciliği kavramının yalnızca problemlere uygulanan bilgisayar tabanlı araç ve yöntemler topluluğu değil, probleme özgü tasarlanmış, ilgili yöntem, teknik ve uygulamaları da içine alan, sonuçları itibariyle probleme özgü olmak üzere ilişkileri, kuralları, örüntüleri veya eğilimleri modelleyen ve gösteren bir süreç olduğu ele alınmıştır.

Uygulama aşamasında kullanılan hastane veritabanında veri transferi, filtreleme ve veri ön-işleme faaliyetleri gerçekleştirilmiş sonrasında da kümeleme, birliktelik kuralları, zaman serileri ve yapay sinir ağları teknikleri kullanılarak birçok veri madenciliği modeli üretilmiştir. Bu modeller ile hastanede konsültasyon hizmetleri örüntülerinin belirlenmesi ve hastanenin gelecekteki hasta yoğunluğunun tahmin edilmesi uygulamaları gerçekleştirilmiştir.

Özellikle birliktelik kuralları tekniği kullanılarak birimler arası konsültasyon hizmetleri örüntülerinin belirlenmesi uygulaması bu tekniğin böyle bir amaçla ilk defa kullanımı ve anlamlı sonuçlar üretmesi bakımından önemli bulunmaktadır. Birimler arası konsültasyon hizmetleri, istem ve hizmeti verme ilişkileri ve bu ilişkilerin yoğunlukları kurallar ile ifade edilmiş ve grafiklerle de görselleştirilmiştir.

Gelecekteki hasta yoğunluklarının tahmin edilmesi uygulamalarında üstel düzgünleştirme, ARIMA ve yapay sinir ağları yöntemleri önce kendi içlerindeki farklı modellerle kıyaslanmış sonra da her yöntemin en kestirimci modelleri birbirleriyle kıyaslanarak bu konuda en iyi sonucu veren model tespit edilmeye çalışılmıştır. Üstel düzgünleştirme yöntemlerinden Winters Additive modeli, ARIMA yöntemlerinden ARIMA$(3,1,0)(1,0,0)_{12}$ modeli ve yapay sinir ağları yöntemlerinden Prune yöntemi ile elde edilen model en iyi sonuçları vermiştir. Winters Additive üstel düzgünleştirme modeli ise bunlar arasında en kestirimci model olarak öne çıkmış ve gerçekleşen değerlere en yakın tahminleri üretmiştir.

SUMMARY

KNOWLEDGE DISCOVERY IN HEALTH SECTOR DATABASES BY USING DATA MINING METHODS: APPLICATIONS OF DESCRIPTIVE AND PREDICTIVE MODELS

The utilization of data mining methodologies and techniques in health sector databases enable discovering knowledge which is useful for health institution managers or policy makers of this domain. In this study, the steps of data mining processes in health sector were exhibited, also theoretical and practical information from the literature were presented. In the application phase, some techniques of data mining were used in a currently active database of a hospital and the results were presented.

Data mining was emphazised as a process rather than only a collection of computer-based tools and techniques. This process includes problem specific methodologies and techniques, and it models relationships, rules, patterns or trends, and it summarizes and visualizes the results in a meaningful way.

Data transfer, filtering and data preprocessing activities were performed in the hospital database. Many data mining models were generated by using clustering, association rules, time series and artificial neural network techniques. The applications of determining the patterns of medical consultation services among the clinics and policlinics of the hospital, and predicting the patient volume of the hospital were executed by using these models.

Particularly the application of determining the patterns of medical consultation services among the clinics and policlinics by using association rules was regarded as important, because this technique was used for this kind of purpose for the first time and the generated results were meaningful.

In order to predict the future volumes of patients, different models of exponential smoothing, ARIMA and neural network techniques were evaluated. The best models of each technique then evaluted again to determine the best predictive model. Winters Additive exponential smoothing model, ARIMA$(3,1,0)(1,0,0)_{12}$ model, and ANN model which was trained by using Prune method have yield beter results. The results of comparison showed that Winters Additive exponential smoothing model was the best predictive model for this data and this model generated the closest predictions to actual values.

ÖNSÖZ

Bir insan için en büyük zenginlik yanındaki ve yakınındaki insanlardır. Şu ana kadar geçen akademik çalışma hayatım boyunca bu yönden kendimi çok şanslı hissettim. Bundan dolayı kendisinden çok şey öğrendiğim Anabilim Dalı başkanım değerli hocam Prof. Dr. Ayşe Kuruüzüm'e ve hem yüksek lisans tezimde hem de doktora tezimde danışmanım olan, bu süreçlerde her zaman desteğini hissettiğim değerli hocam Doç. Dr. Can Deniz Köksal'a öncelikle teşekkür etmek isterim.

Doktora tezimin izleme komitelerinde yer alan, özellikle uygulamalar ve sonuçlarının yorumlanmasıyla ilgili konulardaki görüşleriyle tezimin uygulama bölümünün şekillenmesine önemli katkılar sağlayan Tıp Fakültesi Öğretim Üyesi sayın hocam Doç. Dr. Levent Dönmez'e katkılarından dolayı teşekkür ederim.

Uzun süre birlikte asistanlık yaptığım ve tezimin son dönemlerinde yaptığı son okumalarıyla da tezime katkı sağlayan arkadaşım Yrd. Doç. Dr. Emre İpekçi Çetin'e birlikte yaptığımız çalışmalar ve en önemlisi dostluğu için teşekkür ederim. Bunun yanında akademik çalışmayı keyifli hale getiren diğer tüm asistan arkadaşlarıma ve hocalarıma da teşekkür etmek isterim.

Doktora tez çalışmamı maddi olarak destekleyen Akdeniz Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi'ne katkılarından dolayı teşekkür ederim. Doktora eğitimim süresinde her ne zaman Enstitü'ye uğradıysam beni güleryüzleriye karşılayan ve yardımcı olan Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü'nün tüm çalışanlarına teşekkür ederim.

Son olarak, tüm süreçlerde beni destekleyen, hatta bazı durumlarda katlanmak zorunda kalan, aileme, kardeşime ve arkadaşlarıma teşekkür eder, bu tez çalışması sonucunda elde ettiğim doktora derecesinin mutluluğunu paylaşırım.

Sezgin Irmak

Kasım 2009, Antalya

.

GİRİŞ

Veri madenciliği kavramının günümüzde giderek yaygınlaştığı görülmektedir. İş dünyasında işletmeler daha fazla kar sağlamak amacıyla birçok farklı uygulama için veri madenciliği yöntemlerini kullanmaktadır. Bu yöntemlerle işletmelerin veritabanlarından, veri ambarlarından veya farklı biçimlerdeki veri setlerinden elde edilen bilgiler karar destek sistemlerinde ve stratejik planlama çalışmalarında girdi olarak kullanabilmektedirler. Ayrıca veri madenciliği yöntem ve tekniklerinin geliştirilmesi çalışmalarına ek olarak, özellikle yığın veri içeren sağlık, bankacılık, perakendecilik vb. gibi farklı alanlardaki akademik çalışmalarda da veri madenciliği yöntemleri giderek daha sık kullanılmaya başlanmıştır.

İşletmelerde veri madenciliği yöntem ve teknikleri müşteri sınıflarının belirlenmesi, toplamda, bölgeler temelinde ve hatta şubelere ve ürün kategorilerine göre satış tahminlerinin yapılmasında, maliyetlerin tahmininde, gelecekte uygulanacak kampanya stratejilerinin belirlenmesinde ve benzeri birçok uygulamada kullanılabilmektedir. Ayrıca, bu çalışmanın uygulamalarında olduğu gibi kurum içi faaliyetlerin belirli yapılarının analizinde ve gelecek tahminlemesinde kullanılabilmekte ve veri madenciliği yöntemleri ile stratejik düzeyde karar destek bilgileri elde edilebilmektedir.

Veri madenciliği günümüzün yükselen analiz yöntemlerinden biridir. Hali hazırda insanlar günlük hayatlarında veri madenciliği uygulamaları ile karşılaşmaktadırlar. Örneğin bir süpermarketten üye kartıyla alışveriş yapıldığında bu bilgiler veri madenciliği amaçlarıyla veritabanlarında saklanmakta ve indirim kampanyalarının belirlenmesinde kullanılmaktadır. Internet'ten yapılan alışverişlerde ise bunun sonuçları ile daha hızlı karşılaşma imkanı ortaya çıkmaktadır. İncelenmekte olan bir ürüne göre, önceki satışlardan elde edilen bilgiler kullanılarak, ilgilenilebilecek başka bir ürün sayfanın bir bölümünde öneri olarak karşımıza çıkmaktadır. Önümüzdeki yıllarda örneğin kişisel sağlık bilgi kartları gibi akıllı kart uygulamalarının yardımıyla sağlık verileri ve veri madenciliği yöntemleri kullanılarak yeni ve yaygın önleyici hekimlik uygulamalarının gerçekleştirilebilmesi mümkün gözükmektedir. Ancak bu uygulamaların işin henüz başlangıç adımları olduğunu söylemek gerekmektedir.

Veri madenciliğinin kökleri yapay zeka, makine öğrenimi, matematik, istatistik ve bilgisayar uygulamaları alanlarına dayanmaktadır. Veriden bilgi çıkarımı insanoğlunun yüzyıllar boyunca yaptığı bir iştir. Ancak günümüzde bilgisayar, bilgisayar ağları ve

depolama sistemleri teknolojilerindeki gelişmelerle birlikte hayatı büyük miktarlarda veri kuşatmaktadır ve bu verinin büyük bir kısmı organizasyonel, ticari, bilimsel, stratejik ve benzeri amaçlarla büyük veritabanları veya veri ambarlarında saklanmaktadır.

Büyük, karmaşık ve bilgi zengini veri setlerinin anlaşılabilmesi ihtiyacı, işletme, bilim ve mühendisliğin bütün alanları için ortaktır. İşletmecilik dünyasında, şirket ve müşteri verisi şirketin stratejik varlıkları olarak algılanmaya başlamıştır ve bu verilerde saklı olan faydalı bilgilerin çıkarılması ve bu bilgiye dayalı hareket edebilme yeteneği, bugünün rekabetçi dünyasında giderek artan bir öneme sahip olmaktadır (Kantardzic, 2003, s.1-2). Dolayısıyla işletmelerin gelecekte veri madenciliği ve bununla bağlantılı konularda daha fazla yatırım yapmalarının bir zorunluluk haline geldiği görülmektedir. Bu zorunluluğa paralel olarak ise akademik alanda veri madenciliği alanıyla ilgili eğitim ve araştırma faaliyetlerinin yaygınlaşması bir ihtiyaç olarak ortaya çıkmaktadır.

Sağlık alanı yığın veri tutulan alanlardan biridir ve bu alan veri madenciliği faaliyetleri sonucunda elde edilecek faydalı bilgilerden marjinal olarak en çok yararlanabilecek alanlardan biridir. Bu çalışmada bir sağlık alanı veritabanı kullanılarak veri madenciliği uygulamaları gerçekleştirilmiştir. Veri madenciliği kavramı, tarihsel gelişimi, ilişkili olduğu alanlar ve süreçleri hakkında bilgiler verilmiş daha sonrasında ise veri madenciliği teknikleri ayrıntılı olarak incelenmiştir. Ayrıca sağlık alanında yapılmış önceki veri madenciliği uygulamalarından örnekler verilmiştir.

Çalışmanın birinci bölümünde veri madenciliği kavramı farklı kaynaklardan aktarılmış ve sonuçta kapsayıcı bir tanımlama yapılmıştır. Aynı zamanda bu bölümde veri madenciliğinin bir süreç olarak algılanması gerektiği vurgulanmış ve veri madenciliği süreci ile ilgili farklı yaklaşımlara yer verilerek bunlar tartışılmıştır. Birinci bölümün sonunda veri madenciliği konusunda sağlık alanında yapılmış uygulamalı çalışmalar bilimsel yayın veritabanları kullanılarak araştırılmış ve bu çalışmalar kullandıkları veri setleri, teknikler ile elde edilen bulgular doğrultusunda özetlenmiştir.

İkinci bölümde veri madenciliğinde yaygın olarak kullanılan önemli teknikler ayrıntılı olarak anlatılmış ve bu tekniklerin işleyişi açıklanmaya çalışılmıştır. Burada yer verilen teknikler; Bayes sınıflandırıcılar, karar ağaçları, kümeleme, birliktelik kuralları, yapay sinir ağları ve zaman serileridir. Özellikle çalışmanın uygulama bölümünde kullanılan yöntemlerden birliktelik kuralları ve Apriori algoritması, yapay sinir ağları, zaman serilerinde

üstel düzgünleştirme ve ARIMA yöntemleri ile yapay sinir ağlarının zaman serilerinde kullanımı burada açıklanmıştır.

Son bölümde üç tip bulguya yer verilmiştir. Bunlardan birincisinde hastanenin hasta profili ve verilen hizmetlerle ilgili tanımlayıcı bulgular verilmiştir. İkinci tip bulgu olarak ise birliktelik kuralları ile yapılan modelleme sonucunda konsültasyon hizmetleri istem ve hizmetin verilmesi konularında birimler arasındaki örüntüler istemi yapan ve hizmeti veren birimler merkezli olmak üzere iki farklı biçimde ortaya konulmuştur. Üçüncü tip bulgular ise yapılan hastane yoğunluk tahmini analizlerine ilişkin olmuştur. Burada üstel düzgünleştirme, ARIMA ve yapay sinir ağları teknikleri önce kendi içinde farklı modellerle karşılaştırılmış sonrasında da her tekniğe ait en iyi modeller kendi aralarında kıyaslanmıştır. Böylece en kestirimci model belirlenmeye çalışılmıştır.

BİRİNCİ BÖLÜM
VERİ MADENCİLİĞİ

1.1. Veri Madenciliği Tanımı

Veri madenciliği kavramını anlayabilmek için işin en başında kelimelerin yalın anlamlarından yola çıkılabilir. Madencilik yeryüzünün gizli ve kıymetli kaynaklarının açığa çıkarılması süreci olup, bu kelimenin veri kelimesi ile ilişkilendirilmesi ise veri yığınları içerisinde ilk bakışta fark edilemeyen kıymetli bilgilerin bulunması ve çıkartılması fikrini uyandırmaktadır (Giudici, 2003, s.1).

Bir bilim disiplininin tanımlanması çoğu zaman tartışmalı bir iştir; araştırmacılar kendi çalışma alanlarının kesin aralığı ve sınırları konusunda genellikle aynı fikirde değildirler (Hand vd., 2001, s.1). Aşağıda, veri madenciliği konusunda literatürde yer alan bazı tanımlar verilmektedir:

- Veri madenciliği, geniş veritabanlarından bilgi çıkarımını hedeflemek için makine öğrenimi, örüntü tanıma, istatistik, veritabanı ve görselleştirme tekniklerini bir araya getiren disiplinler arası bir alandır (Cabena, vd., 1998, s. 12).

- Veri madenciliği, (genellikle büyük) gözlemsel veri setlerinin veri sahibi için anlaşılabilir ve faydalı olması amacıyla, tahmin edilemeyen ilişkilerin bulunması için analiz edilmesi ve bunların sözel yollarla özetlenmesidir (Hand vd., 2001, s.1).

- Veri madenciliği, otomatik veya yarı-otomatik biçimlerde verinin analiz edilerek gizli örüntülerin bulunmasıdır (Tang ve MacLennan, 2005, s.2).

- Veri madenciliği veride var olan örüntüleri keşfetme sürecidir. Süreç otomatik veya (daha çok) yarı otomatiktir. Keşfedilen örüntüler anlamlı olmalıdır ve genellikle ekonomik avantaj olmak üzere fayda sağlamalıdır (Witten ve Frank, 2005, s.5).

- Veri madenciliği, büyük veri depolarında faydalı bilgilerin otomatik olarak keşfedilmesi sürecidir (Tan, vd., 2006, s.2).

- Veri madenciliği, veri ambarlarında depolanan büyük miktarlardaki verinin istatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerinin de kullanılarak incelenmesi yoluyla anlamlı yeni ilişkiler, örüntüler ve eğilimler bulunması sürecidir (Gartner Group, 2007).

Literatürde yer alan tanımlardan ve veri madenciliği süreçlerinde yaşanılan deneyimlerden yola çıkılarak şu kapsayıcı tanımlama yapılabilir;

- ✓ Veri madenciliği, büyük veri setlerinde, veritabanlarında veya veri ambarlarında bulunan veriler arasında var olan, bilinmeyen, klasik yöntemlerle görülemeyen ve sıradan olmayan ilişkileri, örüntüleri, belirli yapıları veya eğilimleri ortaya çıkarmak amacıyla istatistik, matematik, makine öğrenimi ve bilgisayar uygulamaları alanlarının birleşimi tekniklerin kullanılarak analiz edilmesi ve sonuçların anlamlı bir şekilde özetlenmesi ve görselleştirilmesi sürecidir.

Bununla birlikte veri madenciliğinin kendine özgü birtakım karakteristik özellikleri olmalıdır. Bir sistemin veri madenciliği sistemi olabilmesi için büyük miktarlarda veri ile çalışabilmesi, birleşik sorgulara cevap verebilir bir yapıda veri ve bilgi geri alma işlemlerini gerçekleştirebilmesi gerekmektedir (Han ve Kamber, 2006, s.9).

1.2. Veri Madenciliğinin Gelişimi

Veri madenciliği, henüz nispeten yeni ve gelişmesini sürdüren bir alan olarak tanımlanabilir. Veri madenciliği ile ilgili ilk kitap 1991 yılında çıkartılmıştır (Piatetsky-Shapiro ve Frawley, 1991) ve veritabanlarında bilgi keşfi konulu bir çalıştayda sunulan çalışmalardan oluşmaktadır (Witten ve Frank, 2005). Uluslararası Veri Madenciliği ve Bilgi Keşfi Kongresi'nin ilk olarak 1995 yılında gerçekleştirildiği de (Shmueli, vd., 2007, s.1) düşünüldüğünde bu yaklaşım doğru sayılabilir. Veri madenciliği, genellikle veritabanlarında bilgi keşfi (Knowledge Discovery in Databases, KDD) bağlamında yer bulmaktadır ve bu terim orijinal olarak yapay zeka (Artificial Intelligence, AI) araştırma alanında çıkmıştır (Hand, vd., 2001, s.3).

Veri madenciliği yöneticiler, karar vericiler ve sonuçların plana göre uygulanmasına dahil olan kişilerle yakından bağlantılı iyi-yapılandırılmış bir standart süreçtir (Larose, 2005, s.4). Veri madenciliğindeki amaç, toplanmış olan bilgilerin, bir takım sayısal yöntemlerle incelenip ilgili kurum ve yönetim destek dizgelerinde kullanılmak üzere değerlendirilmesidir. Veri madenciliğinde geleneksel yöntemlerde olduğunun aksine başlangıçta kesin bir amaç ya da varmak istenilen net bir sonuç yoktur.

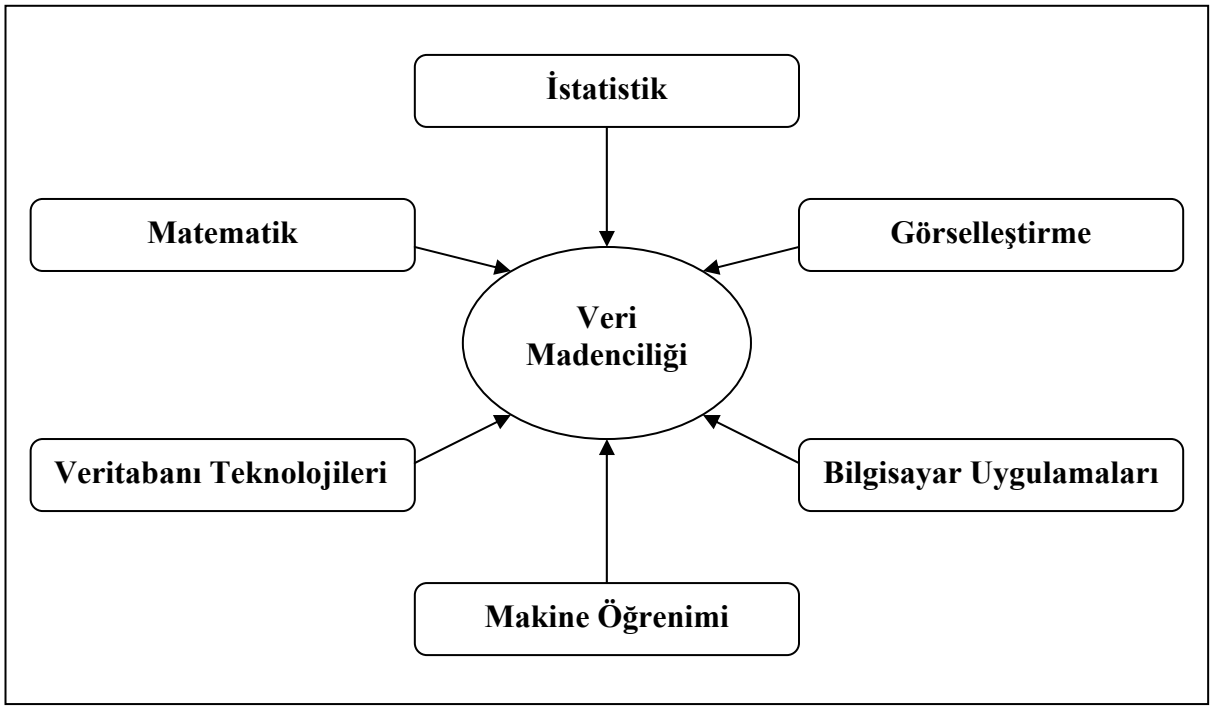
Veri madenciliği nispeten yeni ortaya çıkmış bir alan olmasına rağmen, bu yaklaşımın kökleri yaklaşık otuz yıllık bir geçmişe sahip araştırma ve uygulama geleneğine dayanmaktadır. Bu dönem boyunca istatistiksel analizin öncüleri SAS, SPSS ve IBM gibi firmalar olmuştur. Günümüzde de bu firmalar veri madenciliği alanında oldukça aktiftirler ve bu alanda, yıllara dayanan deneyimleri sonucunda geliştirilmiş yüksek kabul gören ürünleri bulunmaktadır (Marakas, 2003, s.18).

Günümüzün elektronik iş ortamında, veri madenciliği daha fazla dikkat toplamaya başlamıştır. Çünkü veri madenciliği analiz ve keşfetme üzerinedir. Otomatik veya yarı-otomatik biçimlerle, çok miktarda veri anlamlı örüntü veya kuralların ortaya çıkarılmasına yardımcı olabilir. Bu örüntüler ve kurallar şirketlere, müşterilerini daha iyi anlamaları için, pazarlama, satış ve müşteri desteği operasyonlarını geliştirmede yardımcı olabilir. Yıllar boyu şirketler, kurumsal kaynak planlama (Enterprise Resource Planning, ERP), müşteri ilişkileri yönetimi (Customer Relationships Management, CRM) veya diğer operasyonel sistemler gibi uygulamalardan çok büyük veritabanları oluşturmuşlardır (Soni vd., 2008).

Veri madenciliği için evrensel bir tanım üzerinde uzlaşılamadığı gibi, bu alanın köklerini dayandırdığı alanlar veya hangi disiplinlerin kesişim kümesi içerisinde yer aldığı konusunda da bir ortak düşünce oluşmamıştır. Ancak bu konuda nispeten herkesin hemfikir olabileceği gerçekler vardır.

Birçok veri madenciliği problemi ve bu problemlere ilişkin çözümlerin klasik veri analizinde kökleri bulunmaktadır. Veri madenciliğinin birçok disiplinde kökenleri olduğu gibi, bunlardan en önemli ikisi istatistik ve makine öğrenimidir. İstatistiğin kökenleri matematiğe dayanmaktadır, bunun için veri madenciliğinde matematiksel katılığa, yani bir şeyin uygulanmasından önce teorik temelinde hassas olarak kurgulanması isteğine bir vurgu vardır. Buna tezat olarak, makine öğrenimi topluluğu köklerini büyük oranda bilgisayar uygulamalarından almaktadır. Bu da, uygulamacı bir yönelime, yani etkinliğinin biçimsel bir

kanıtı için beklemeksizin, ne kadar iyi performans gösterdiğini görmek amacıyla bir şeyi test etme isteğine neden olmaktadır (Kantardzic, 2003).



Şekil 1.1. Veri Madenciliğinin Dayandığı Alanlar

Buna ek olarak, veri madenciliğinin temellerini oluşturan disiplinler arasında insan zekasına benzetim yapan yapay zeka alanı da gösterilmektedir. Makine öğrenimi ise istatistik ve yapay zekanın birleşimi olarak tanımlanmaktadır, ancak yapay zeka ticari bir başarı elde edemezken teknikleri büyük oranda makine öğrenimi tarafından alınmış ve uyarlanmıştır. Şekil 1.1'de veri madenciliğinin dayandığı alanlar gösterilmiştir. Buradan yola çıkarak, veri madenciliği konusunda istatistik ve matematiğe kökleri dayanan, makine öğrenimi alanından teknikleri barındıran, veritabanı teknolojileri ve üzerinde çalışan bilgisayar uygulamaları ile süreci uygulamaya koyan ve görselleştirme teknikleri ile de sonuçları karar destek sistemlerine anlamlı ve anlaşılır bir biçimde çıktı olarak veren çok disiplinli bir alan diyebiliriz.

1.3. Veri Madenciliği Süreci

Her ne kadar bazı araştırmacılar ve uygulamacılar tarafından veri madenciliği, problemlere uygulanan bilgisayar tabanlı araçlar ve yöntemler topluluğu olarak algılansa da bu algı gerçek yaşam uygulamalarında bir eksiklik ve yanlışlık oluşturmaktadır. Veri madenciliği yalnızca yöntemlerinin herhangi bir problemin çözümüne yönelik anlık

uygulaması olarak değil, aşağıdaki nedenlerden dolayı bir süreç olarak algılanmalıdır (Kantardzic, 2003):

- Birinci neden, veri madenciliğinin sadece her biri diğerinden tamamıyla farklı olan ve problemle eşleşmeyi bekleyen, birbirinden izole araçlar koleksiyonu olmamasıdır.

- İkinci neden, problemin yöntem ile eşleşmesi nosyonudur. Çok nadiren, bir araştırma problemi tam olarak bir yöntemin uygulanmasıyla etkin sonuç verir. Doğrusu, uygulamada veri madenciliğinin iteratif bir süreç olduğudur. Bir yöntem veriyi analiz eder, bazı analitik tekniklerle sınar ve başka bir yönden bakmaya veya belki değiştirmeye karar verir; sonra başa dönerek daha iyi veya farklı sonuçlar elde etmek amacıyla başka bir veri analiz tekniği uygular. Bu her defasında veriye faklı bir soru sormak amacıyla farklı açılardan sonda yapılarak birçok kere tekrarlanabilir.

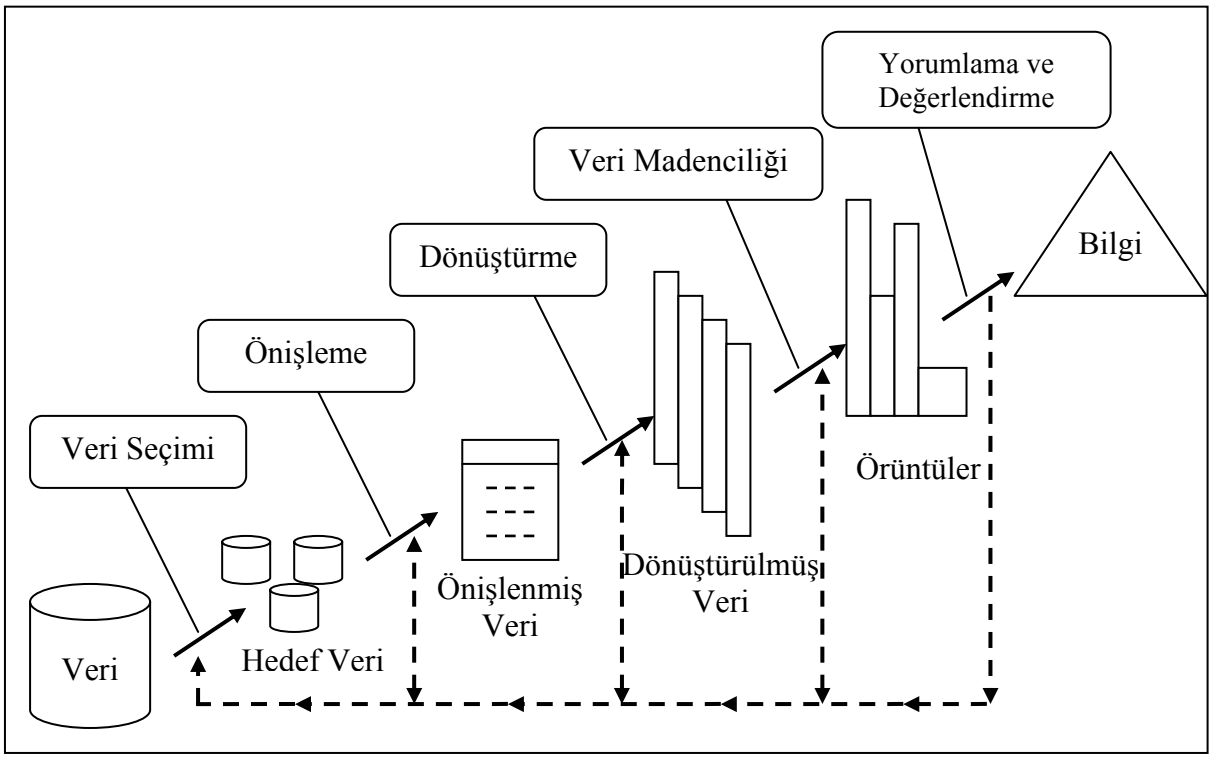
- Üçüncü neden, veri madenciliği istatistik, makine öğrenimi ve diğer yöntem ve araçların tesadüfi uygulanması değildir; bilakis bu süreç analitik teknikler uzayında bir rassal yürüyüş olmadığı gibi, neyin en kullanışlı, sonuç verici ve açıklayıcı olduğuna yönelik özenli planlanmış ve düşünülmüş bir karar verme sürecidir.

Bu durumda veri madenciliği sadece bir yöntem ve teknikler topluluğu olarak değil, probleme özgü tasarlanmış, ilgili yöntem, teknik ve uygulamaları da içine alan, sonuçları itibariyle probleme özgü olmak üzere ilişkileri, kuralları, örüntüleri, eğilimleri, vb. modelleyen ve gösteren bir süreç olarak algılanacaktır.

1.3.1. Literatürde Veri Madenciliği Süreci

Literatürde bu sürecin tanımlanması konusunda henüz bir ortak fikir oluşmadığı görülmektedir. Veri madenciliği bazı kaynaklarda veritabanlarında bilgi keşfi kavramı ile birbirinin yerine geçebilir şekilde kullanılmaktadır ve süreç veri madenciliği süreci veya veritabanlarında bilgi keşfi süreci olarak adlandırılmaktadır (Han ve Kamber, 2006, s.7; Tsai ve Tsai, 2005; Kendall ve Kendall, 2008, s.516). Bazı kaynaklarda ise süreç veri madenciliği süreci (Olaru ve Wehenkel, 1999; Brohman, 2006; Delavari, 2005) ve bu işlevi gerçekleştiren

bir sistem de veri madenciliği sistemi (Li ve Khosla, 2005) olarak adlandırılmaktadır. Ancak bazı kaynaklar veri madenciliğini veritabanlarında bilgi keşfi sürecinde bir aşama olarak konumlandırmaktadırlar (Tan, vd., 2006, s.3; Fu, 1997; Dunham, 2003; Fayyad, vd., 1996) ve sürecin bütününü veritabanlarında bilgi keşfi kavramı ifade etmektedir. Bazı kaynaklarda ise bu kavram karmaşası içerisinde iki kavram birleştirilerek süreç bilgi keşfi ve veri madenciliği süreci (Microsoft Research, 2006; Fayyad, 1996) olarak isimlendirilmektedir.



Şekil 1.2. Veritabanlarında Bilgi Keşfi Sürecini Oluşturan Aşamalar
(Fayyad vd., 1996, s.41)

Burada iki farklı yaklaşım olduğunu görmek mümkündür. Birincisinde, araştırmacılar veri madenciliğini, veritabanlarında bilgi keşfi sürecinde veri madenciliği tekniğinin uygulandığı ve modelin geliştirildiği aşama olarak tanımlamaktadırlar. Fayyad vd., (1996) bu yaklaşım doğrultusunda veritabanlarında bilgi keşfi sürecini oluşturan aşamaları Şekil 1.2'de görüldüğü gibi tanımlamışlardır. İkinci yaklaşımda ise, araştırmacılar süreci veri madenciliği süreci olarak adlandırmakta ve veritabanlarında bilgi keşfi sürecini bunun eş anlamlısı olarak görmektedirler. Son yıllarda kullanılan kavramın genellikle veri madenciliği süreci olması ve bu süreci metodolojik olarak standardize etmek amacıyla geliştirilen endüstri standartlarının (SEMMA, CRISP-DM, vb.) süreci veri madenciliği olarak tanımlıyor olmaları ve endüstride de genel kabul görmüş olması nedeniyle bu çalışmada, veri ön işlemeden modelin uygulanmasına kadar olan aşamaların tümü veri madenciliği süreci olarak ele alınacaktır.

Veri madenciliği süreci farklı kaynaklarda birbirine benzemekle birlikte sayısı üç ile on bir arasında değişen farklı aşamalarda tanımlanmıştır. Bu bölümde literatürde yer alan bazı veri madenciliği süreçleri kısaca özetlenerek ayrı ayrı verilmiştir. Buradaki amaç, veri madenciliği sürecine farklı bakış açılarını görmek ve her bir yaklaşımda süreçle ilgili bir bakış geliştirebilmektir.

Tan vd. (2006, s.3) süreci veri ön işleme, veri madenciliği ve son işleme olarak üç aşamada ifade etmiştir. Veri ön işleme aşamasında özellik (değişken) seçimi, boyutsal indirgeme, normalizasyon ve veri alt kümeleme işlemleri yer almaktadır. Burada veri ön işlemenin amacı ham girdi verisini bir sonraki analiz için uygun formata dönüştürmektir. Son işleme kısmında ise örüntülerin filtrelenmesi, görselleştirme ve örüntülerin yorumlanması işlemleri yer almaktadır. Son işleme aşamasının amacı ise sadece geçerli ve kullanışlı sonuçların karar destek sisteminin bünyesine dahil edilmesini garanti etmektir.

Berry ve Linoff (2004, s.54) veri madenciliği sürecini on bir aşamada ayrıntılı olarak tanımlamıştır ve birbirini takip eden düz bir çizgi yerine iç içe döngüler kümesi olarak algılanması gerektiğini vurgulamışlardır. Bu on bir aşama aşağıda verilmiştir:

1. İş probleminin veri madenciliği problemine dönüştürülmesi
2. Uygun verinin seçimi
3. Verinin öğrenilmesi
4. Bir model setinin oluşturulması
5. Verideki problemlerin giderilmesi
6. Verinin dönüştürülmesi
7. Modellerin yapılandırılması
8. Modellerin değerlendirilmesi
9. Modellerin uygulanması
10. Sonuçların değerlendirilmesi
11. Gerekli adımdan yeniden başlanması

Burada adımlar doğal bir sıraya sahiptirler, fakat bir sonrakine geçebilmek için önceki adımın bitirilmesi veya tamamlanması bir zorunluluk değildir. Ayrıca sonraki adımlarda öğrenilen durumlar önceki adımların yeniden gözden geçirilmesini gerektirebilir (Berry ve Linoff, 2004, s.55).

Bir diğer yaklaşımda Giudici (2003, s.6-10) veri madenciliği sürecini amaçların tanımlanmasından sonuçların değerlendirilmesine kadar yedi aşama olarak tanımlamıştır. Bu aşamalar Giudici'nin (2003) çalışmasından derlenerek aşağıda özetlenmiştir.

Amaçların Tanımlanması: Bu aşama analizin amaçlarını tanımlamayı içerir ve amaçlanan şirket hedefleri genellikle nettir. Bununla birlikte problemlerin, analiz edilmesi gereken detaylı amaçlara dönüştürülmesi daha karmaşıktır. Dolayısıyla bu aşamada amaçlar iş hedefleri doğrultusunda tanımlanabilir.

Verinin Düzenlenmesi: Analizin hedefleri tanımlandıktan sonra analiz için gerekli veriyi seçmek önemlidir. Öncelikle veri kaynaklarının seçilmesi gerekir. Genellikle veri daha güvenli olan iç kaynaklardan alınır. Bunlar da şirket veritabanı, veri ambarı veya küçük veri depoları olabilir. Bu aşamada ayrıca değişkenlerin içeriği incelenmeli, kayıp ve doğru olmayan veri için gerekli çalışmalar yapılmalıdır. Son olarak da uygun verinin tamamı veya uygun bir örneklemi seçilerek analiz için hazır hale getirilmelidir.

Verinin Keşfedici Analizi: Bu aşama verinin hazırlık niteliğinde olan Çevrimiçi Analitik Süreçler (OnLine Analytical Processing, OLAP) tekniklerine benzer şekilde keşfedici analizlerini içerir. Keşfedici analiz, örneğin veri içerisindeki anomali içeren değişken ve kayıtların ön tespiti için kullanılabilir. Anormal veri önemli bilgiler içeriyor olabilir, bu yüzden analizin amaçları göz önünde alınarak hemen elenmemelidir. Verinin keşfedici analizi aşaması, analizcinin bir sonraki aşamada kullanılacak yöntemlerin iyi kestirimine olanak sağlar.

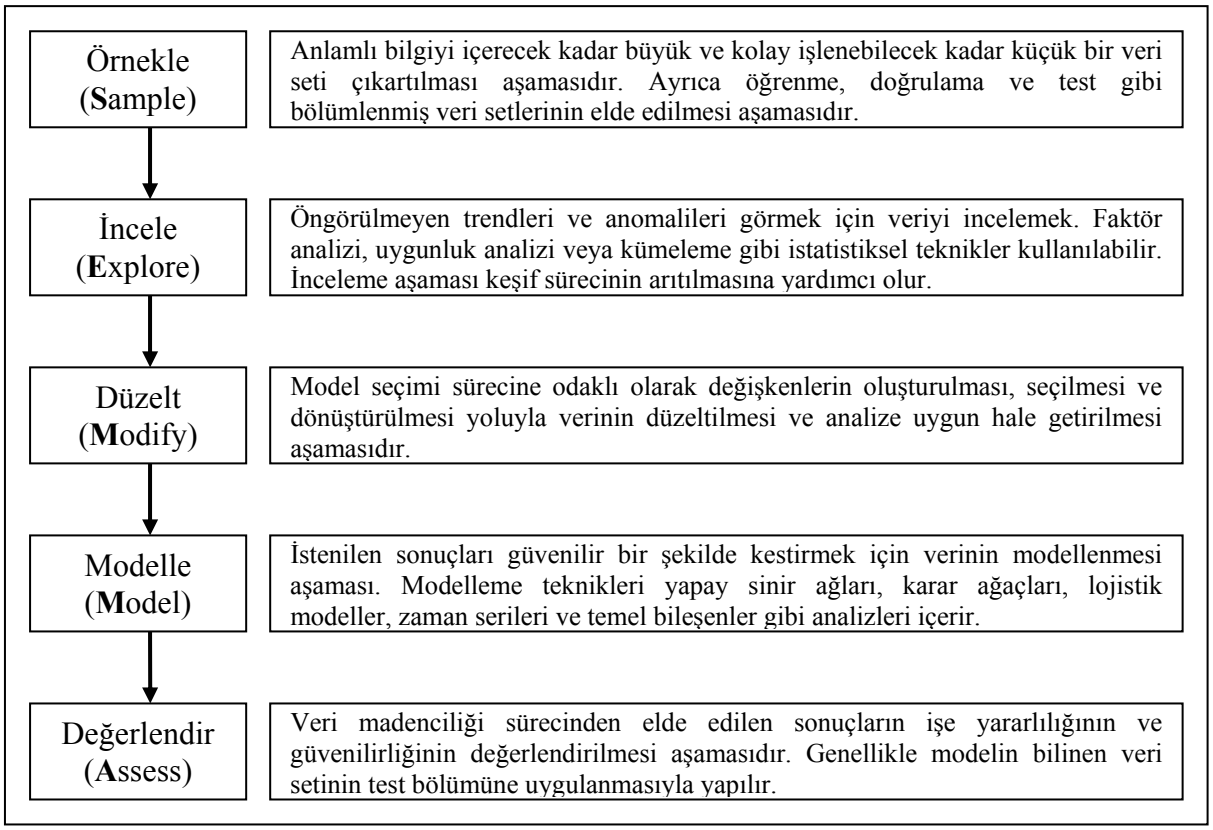
İstatistiksel Yöntemlerin Belirlenmesi: Yöntemin seçilmesi, üzerinde çalışılan probleme veya mevcut uygun verinin yapısına bağlıdır. Veri madenciliği süreci uygulamalar tarafından yönlendirilmektedir ve bu nedenle yöntemler analizin amacına göre sınıflandırılabilir. Tanımlayıcı yöntemler daha özet bir biçimde veri gruplarını tanımlamayı amaçlar. Bunlar ayrıca simetrik, denetimsiz veya dolaylı yöntemler olarak adlandırılırlar. Kestirimci yöntemler, bir veya daha fazla değişkeni tüm diğer değişkenlerle ilişkili olarak tanımlamayı amaçlar. Bunlar ayrıca asimetrik, denetimli veya doğrudan yöntemler olarak da adlandırılırlar. Birliktelik kuralları gibi algoritmalar da global olmaktan çok yerel yöntemler olarak tanımlanmaktadır ve veritabanında bir alt küme ile ilgili belirli karakteristikleri tanımlamayı amaçlar.

Veri Analizi: İstatistiksel yöntemler belirlendikten sonra bilgisayar uygulamalarında çalıştırılmak üzere uygun algoritmalara dönüştürülmelidir. Uygulanan algoritmalar, üzerinde çalışılan veritabanından istenen sonuçların elde edilmesini sağlar. Veri madenciliği için kapsamlı ve çok çeşitli yazılımlar vardır. Bunun için amaca özel (ad hoc) algoritmalar geliştirilmesine gerek olmadan yazılımla gelen algoritmaların yeterli olması beklenir.

İstatistiksel Yöntemlerin Değerlendirilmesi: Bir son kararın üretilebilmesi için uygun olan istatistiksel yöntemler arasından veri analizinin en iyi modeli seçilmelidir. Model seçimi farklı yöntemlerle elde edilen sonuçların belirli kriterler doğrultusunda karşılaştırılmasına dayalıdır. Veri madenciliğinde veriyi analiz etmek için nadiren tek bir istatistiksel yöntem kullanılır. Farklı yöntemlerin diğer türlü gözden kaçabilecek farklı bakış açılarını ortaya çıkarma potansiyeli vardır.

Yöntemlerin Uygulanması: Veri madenciliği sadece verinin analizi değil, aynı zamanda sonuçlarının kurumun karar aşamalarına yansıtıldığı bir süreçtir. Model seçildikten ve veri seti ile test edildikten sonra kurallar bütün referans popülasyona uygulanabilir. Veri madenciliği sürecinin kuruma dahil edilmesi aşamalı olarak yapılmalıdır. Önce gerçekçi amaçlar tanımlanmalı ve sonuçlarına bakılmalıdır. Son olarak ise veri madenciliği diğer faaliyetlerle tamamen entegre edilmeli ve karar süreçlerini desteklemek amacıyla kullanılmalıdır.

Veri madenciliği sürecini mantıksal olarak organize eden metodolojilerden biri de SAS firmasının önerdiği SEMMA'dır. SAS firması SEMMA'yı bir veri madenciliği metodolojisi olmaktan çok veri madenciliğinin temel görevlerini yerine getirmek için bir mantıksal düzenleme olarak tanımlamaktadır. SEMMA süreci ismini, aşamalarının baş harflerinin birleşmesinden almaktadır. Şekil 1.3'de SAS firmasının SEMMA dokümanından derlenen süreçler ve açıklamaları görülmektedir (SAS, 2007).

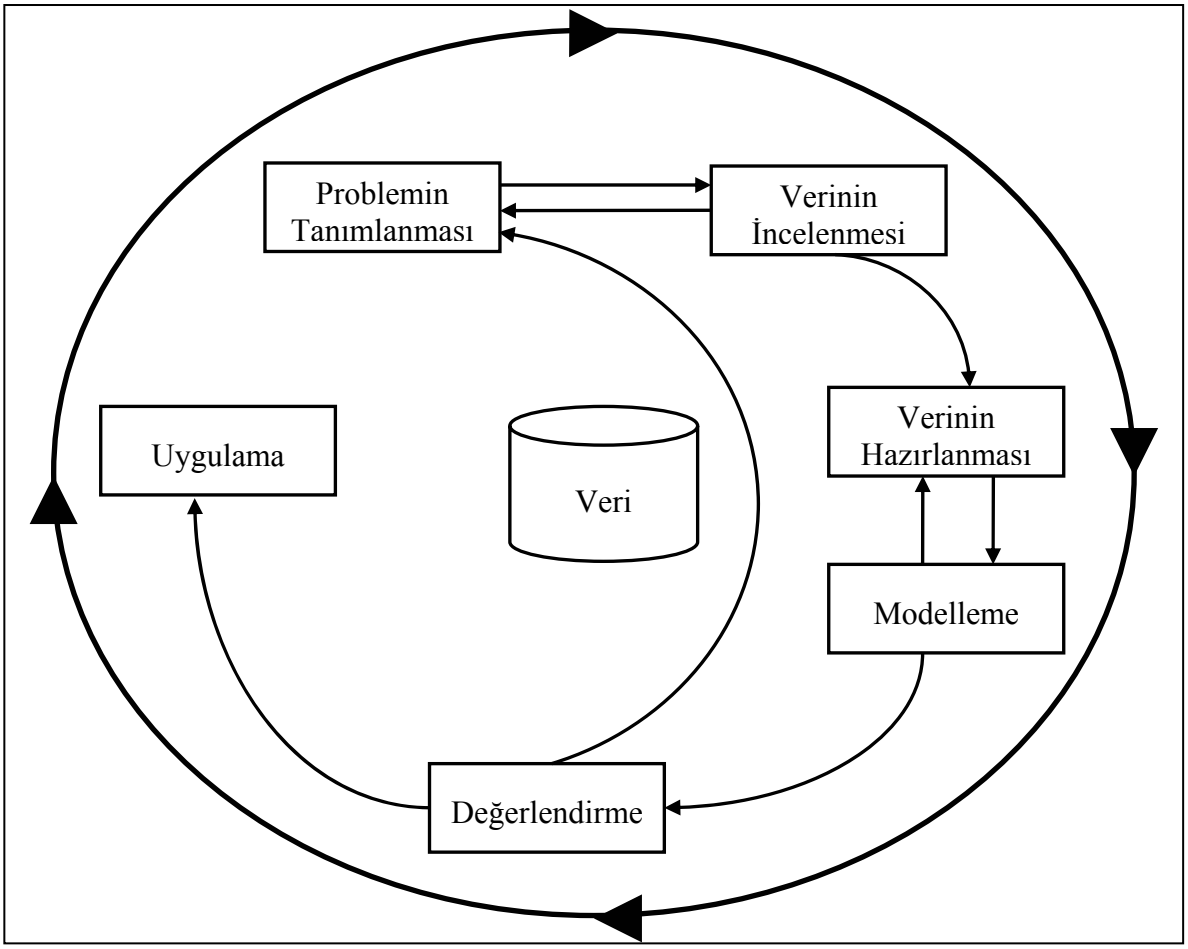


Şekil 1.3. SEMMA Aşamaları (SAS, 2007).

1.3.2. CRISP-DM

Veri Madenciliği için Çapraz Endüstri Standart Süreci (Cross-Industry Standard Process for Data Mining – CRISP-DM) projesi endüstri ve araçlardan bağımsız bir veri madenciliği süreci modeli geliştirmek amacıyla ortaya çıkmıştır. CRISP-DM, veri madenciliği projelerinin belirli standart süreçler içerisinde yürütülebilmesi için SPSS, Daimler Chrysler ve NCR firmaları tarafından oluşturulan konsorsiyum tarafından geliştirilen bir endüstri standardıdır. Halen standardın 1.0 versiyonu kullanımda olup, 2.0 versiyonunun geliştirilmesi için özel ilgi grubu (Special Interest Group – SIG) oluşturma çalışmaları devam etmektedir (CRISP-DM, 2007).

CRISP-DM Referans Modeli, veri madenciliği projesi için genel bir bakış sağlamaktadır. Şekil 1.4'de görüldüğü gibi veri madenciliği projesinin hayat döngüsü altı aşamadan oluşmaktadır. Aşamaların sırası kesin değildir ve aşamalar arasında ileri veya geri doğru hareket her zaman gereklidir. Şekilde yer alan oklar, aşamalar arasındaki en önemli ve en sık bağımlılıkları temsil etmektedir (Chapman, vd., 2000, s.13).



Şekil 1.4. CRISP-DM Metodolojisine Göre Veri Madenciliği Süreci (IBM, 2007)

1.3.2.1. Problemin Tanımlanması

Bir veri madenciliği projesi iş probleminin anlaşılması ile başlar. Veri madenciliği uzmanları, işletme uzmanları ve alan uzmanlarının birlikte çalışmalarını gerektiren bir aşamadır. Proje hedeflerinin ve işletme perspektifinden gereksinimlerin tanımlanması bu aşamada gerçekleştirilir. Proje hedefleri daha sonra veri madenciliği problem tanımlamalarına dönüştürülür. Problemin tanımlanması aşamasında veri madenciliği araçları henüz gerekli değildir (IBM, 2007).

Problemin tanımlanması veya işin kavranması (business understanding) aşaması, iş hedeflerinin belirlenmesi, mevcut durumun değerlendirilmesi, veri madenciliği hedeflerinin yapılandırılması ve bir proje planı geliştirilmesi faaliyetlerini içerir. Bir veri madenciliği çalışması için anahtar bileşen çalışmanın ne için olduğunun bilinmesidir. Bu da yeni bir bilgiye duyulan yönetimsel ihtiyaçla ve yapılacak çalışmanın iş hedeflerinin tanımlanmasıyla başlar. Amaçların, "Ürünlerimizden her biri için hangi tip müşteriler ilgi gösteriyor?" veya

"Müşterilerimizin tipik profilleri nelerdir ve her biri bize ne kadar değer sağlıyor?" gibi ifadelerle tanımlanmasına ihtiyaç vardır. Sonrasında bu tür bilgi ihtiyaçlarının karşılanması için verinin toplanması, analiz edilmesi ve sonuçların raporlanması faaliyetlerinin ve ilgili sorumlulukların tanımlandığı bir plan geliştirilmelidir. Bu aşamada çalışmayı destekleyecek bir bütçe en azından başlangıç düzeyinde hazırlanmalıdır (Olson ve Shi, 2007, s. 20-21).

Tablo 1.1'de Problemin Tanımlanması aşamasının altında yer alan görevler, bu görevlere ait çıktılar, ilgili görev ve çıktılara ait açıklamalar CRISP-DM konsorsiyumu tarafından yayımlanan CRISP-DM 1.0 Veri Madenciliği Rehberi'nden (Chapman vd., 2000, s.10-17) derlenerek verilmiştir.

Tablo 1.1. Problemin Tanımlanması Aşamasında Tanımlanan Görevler ve Çıktılar

Görevler ve *Çıktıları*	**Açıklama**
İş Hedeflerinin Belirlenmesi	İş hedeflerinin belirlenmesi görevinde işletmenin bakış açısıyla veri madenciliği projesinden istenilenin ne olduğu tam olarak anlaşılmalıdır. Bu aşamanın atlanması veya burada yanlışlık yapılmasının muhtemel sonuçları, yanlış sorulara doğru cevaplar üretmek ve bunun için oldukça büyük bir insan ve para kaynağı harcamak olabilir.
Altyapı Bilgisi	Projenin başlangıcında işletmenin durumu hakkındaki bilgilerin kayıt altına alınmasıdır.
İşletme Hedefleri	İşletmenin birincil hedefleri ile bunlara bağlı yan hedefler ve iş sorularıdır.
Başarı Kriterleri	İşletme bakış açısıyla proje çıktılarının başarılı veya kullanışlı olup olmadığını belirleyecek olan kriterlerdir.
Mevcut Durumun Değerlendirilmesi	Bu görev veri analizi hedefleri ve proje planının belirlenmesinde göz önüne alınacak kaynakların, kısıtların, varsayımların ve benzeri faktörlerin tümü hakkında daha detaylı bilgi toplama faaliyetlerini içerir.
Kaynaklar Envanteri	Proje için uygun olan kaynakların listesidir. Bunlar arasında personel (işletme uzmanları, veri uzmanları, teknik destek, veri madenciliği uzmanları), veri (sabit çıktılar, veri ambarı verisi, işlemsel veritabanı verisi), bilgisayar (donanım platformları) ve yazılım (veri madenciliği araçları, diğer ilgili yazılım) kaynakları bulunur.
Gereksinimler, Varsayımlar ve Kısıtlar Listesi	Projeye ait tüm gereksinimlerin listesidir ve bitiş çizelgesi, sonuçların yorumlanabilirliği ve kalitesi, güvenlik ve yasal konuları içerir. Aynı zamanda veri veya iş ile ilgili varsayımlar ve kısıtları da listeler.
Riskler ve Olası Arızalar Listesi	Projede gecikmeye veya başarısızlığa neden olabilecek riskleri veya olayları listeler. Beklenmedik durum planlarını içerir ve belirlenen risk veya olaylar oluştuğunda nelerin yapılacağını listeler.

Tablo 1.1. devamı

Görevler ve *Çıktıları*	Açıklama
Mevcut Durumun Değerlendirilmesi	
Terminoloji Sözlüğü	Proje ile ilgili terminolojilerin ve açıklamalarının derlendiği bir sözlüktür. İlgili sektör terminolojisi ve veri madenciliği terminolojisi olmak üzere iki unsuru olabilir.
Fayda-Maliyet Analizi Raporu	Proje için bir fayda-maliyet analizi yapılandırır. Projenin başarılı olması durumunda potansiyel faydaları ile maliyetlerinin karşılaştırmasıdır.
Veri Madenciliği Hedeflerinin Belirlenmesi	Bir iş hedefi ilgili sektör terminolojisi ile amaçların tanımlanmasıdır. Veri madenciliği hedefi ise proje amaçlarının teknik terimlerle ifade edilmesidir. Örneğin bir iş hedefi "Mevcut müşterilere katalog satışlarının arttırılması" olabilirken, veri madenciliği hedefi "Müşterilerin geçmiş 3 yıllık alışveriş kayıtları, demografik bilgileri ve alınan parçaların fiyatları bilgilerine dayalı olarak bir müşterinin kaç parça malzeme alacağının kestirilmesi" olabilir.
Veri Madenciliği Hedefleri	İş hedeflerinin başarılmasına katkı sağlamak amacıyla planlanan proje hedeflerinin tanımlanmasıdır.
Veri Madenciliği Başarı Kriterleri	Projenin çıktılarının başarılı olduğunun belirlenmesi için kriterlerin, örneğin kesin bir kestirim doğruluğu seviyesi gibi teknik terimlerle tanımlanmasıdır.
Proje Planının Üretilmesi	Veri madenciliği hedeflerine ulaşabilmek ve dolayısıyla iş hedeflerinin gerçekleştirilmesi için amaçlanan planın tanımlanması görevidir. Plan, araç ve tekniklerin başlangıç seçimini de içermek üzere projenin geri kalanında gerçekleştirilecek adımları tanımlamalıdır.
Proje Planı	Projede gerçekleştirilecek aşamaların, süreleri, gerekli kaynakları, girdileri, çıktıları ve bağımlılıkları ile birlikte tanımlandığı listedir. Proje planı her aşama için detaylı planları içerir ve bu aşamada değerlendirme aşamasında hangi değerlendirme stratejisinin izleneceğine karar verilmelidir. Proje planı dinamik bir dokümandır ve her aşamanın sonunda bir gözden geçirme sürecine ihtiyaç vardır. Güncellemeler için gözden geçirme noktaları oluşturulması proje planının bir parçasıdır.
Araç ve Tekniklerin Başlangıç Değerlendirmesi	İlk aşamanın sonunda veri madenciliği araç ve tekniklerinin bir başlangıç değerlendirmesi yapılmalıdır. Sürecin farklı aşamaları için çeşitli yöntemleri destekleyen bir veri madenciliği aracı seçilmelidir. Araç ve tekniklerin seçimi tüm projeyi etkileyebileceği için, sürecin başında araç ve tekniklerin değerlendirilmesi önem taşımaktadır.

1.3.2.2. Verinin İncelenmesi

Verinin incelenmesi aşamasında veri madenciliği araçları veya klasik istatistiksel veri analizi araçları kullanılabilir. Tablo 1.2'de verinin incelenmesi aşamasında yer alan görevler ve bu görevlere ait çıktılar CRISP-DM konsorsiyumu tarafından hazırlanan rehberden yararlanılarak özetlenmiştir (Chapman vd., 2000, s.18-19).

Tablo 1.2. Verinin İncelenmesi Aşamasında Tanımlanan Görevler ve Çıktılar

Görevler ve *Çıktıları*	**Açıklama**
Verinin Toplanması	Proje kaynaklarında listelenen verinin alınması veya bu veriye erişilmesi görevidir. Eğer verinin incelenmesi aşaması için özel araçlar kullanılıyorsa veri bu araçlara yüklenebilir. Bu görev veri önişleme adımlarının gerçekleştirilmesini sağlar. Eğer çoklu veri kaynağı varsa veri birleştirme de gerekli bir işlemdir.
Veri Toplama Raporu	Elde edilen veri setlerinin konumları, alma yöntemleri ve varsa problemleri ile birlikte listesidir. Karşılaşılan kayıt problemleri ve uygulanan çözümleri de içerebilir.
Verinin Tanımlanması	Elde edilen verinin tüm veya yüzeysel özelliklerini inceleme ve sonuçları raporlama görevidir.
Veri Tanımlama Raporu	Elde edilen verinin formatı, miktarı (örneğin her tabloda kaç adet alan ve kayıt bulunduğu) ve alanların özelliklerinin tanımlandığı rapordur.
Verinin İncelenmesi	Bu görevde sorgulama, görselleştirme ve raporlama gibi teknikler kullanılır. Bu teknikler anahtar özniteliklerin dağılımı, öznitelikler arasındaki ilişkiler, belirli alt-popülasyonların özellikleri ve bazı istatistiksel analizleri içerebilir. Bu analizler direkt olarak veri madenciliği hedeflerine yönelik olabileceği gibi, veri tanımlama ve veri kalitesi raporlarını iyileştirmeye de katkıda bulunabilir.
Veri İnceleme Raporu	Veri inceleme sonuçlarını tanımlar. Bu raporda ilk bulgular, başlangıç hipotezleri ve bunların projenin geri kalanına etkileri tanımlanır. Eğer uygunsa, grafikler ve veri karakteristiklerini yansıtacak çizimler kullanılabilir.
Veri Kalitesinin Doğrulanması	Burada, "Veri tam mı?", "Veri doğru mu?", "Eğer hatalar içeriyorsa oranı nedir?", "Kayıp veri içeriyor mu, içeriyorsa bu sorun nasıl çözülüyor?" gibi sorulara yanıt verecek şekilde verinin kalitesi incelenir.
Veri Kalitesi Raporu	Veri kalitesi doğrulama işlemleri ve sonuçlarının listesidir. Eğer kalite problemleri varsa, olası çözümler listelenir. Veri kalitesi problemlerinin çözümü genellikle hem veri, hem de iş bilgisine dayanır.

İşletme hedefleri ve proje planı oluşturulduktan sonra, verinin incelenmesi aşaması veri ihtiyaçlarının belirlenmesi, veri toplanması, verinin tanımlanması ve veri kalitesinin doğrulanması gibi işlemleri kapsayan bir adımdır. Bu aşamanın sonunda özet istatistikler gibi sonuçlar üretilebilir. Kümeleme analizi gibi modeller de ayrıca bu aşamada veri içerisindeki örüntülerin tanımlanması için kullanılabilir. Veri seçimi farklı veri kaynaklarından ve farklı veri tiplerinden yapılabilir. Genellikle işletme uygulamaları için veri tipleri, demografik veriler (gelir, eğitim düzeyi, yaş, vb.), sosyografik veriler (hobiler, klüp üyelikleri vb.), işlemsel veriler (satış kayıtları, kredi kartı harcamaları, düzenlenmiş çekler, vb.) gibi veriler olabilir. Bunun yanında veri tipleri nicel (quantitative) ve nitel (qualitative) olarak da ayrılmaktadır. Nicel veriler kesikli (tamsayı) olabileceği gibi sürekli (reel sayı) de olabilir. Nitel veya kategorik veriler ise nominal veya ordinal veri içerebilir. Nominal veri cinsiyet değişkenindeki erkek ve kadın gibi sonlu ve sıralı olmayan değerler içerir. Ordinal veri ise müşteri kredi kartı derecelendirme değişkenindeki iyi, vasat ve kötü gibi sonlu ve belirli bir sıraya (veya dereceye) sahip değerler içerir. Nicel veriler herhangi bir olasılık dağılımı yoluyla kolaylıkla incelenebilir. Nitel veriler ise sayısal olarak kodlanarak frekans dağılımları ile incelenebilir (Olson ve Shi, 2007, s.20-22).

1.3.2.3. Verinin Hazırlanması

Kullanılabilir veri kaynakları tanımlandıktan sonra, verinin seçilmesi, temizlenmesi, istenilen formda yapılandırılması ve biçimlendirilmesi gerekmektedir. Verinin ön-işlenmesi aşaması olarak da tanımlanabilen bu aşamada veri temizleme ve veri dönüştürme işlemleri gerçekleştirilmektedir. Veri ön-işlemenin temel amacı seçilen veriyi en iyi veri kalitesini sağlayacak şekle getirmektir. Seçilen bazı veriler farklı veri kaynaklarından elde edildiği için farklı biçimlerde (format) olabilir. Eğer seçilen veri, metin dosyaları (flat files), web metinleri veya farklı veritabanları gibi farklı kaynaklardan alınıyorsa, bu veri tutarlı bir biçime dönüştürülmelidir. Seçilen verinin ön-işlenmesi için birçok istatistiksel yöntem veya görselleştirme teknikleri kullanılabilir. Maksimum, minimum, ortalama ve mod gibi yaygın istatistik yöntemleri veriyi birleştirmek veya düzgünleştirmek için kullanılabilir. Serpilme diyagramları ve kutu grafikleri ise genellikle aykırı değerleri filtrelemek için kullanılırlar. Veri kalitesi için gerekli ihtiyaçlara dayalı olarak, regresyon analizi, kümeleme analizi ve karar ağaçları gibi gelişmiş teknikler de kullanılabilir (Olson ve Shi, 2007, s.20-23).

Tablo 1.3'de verinin hazırlanması aşamasında yer alan görevler ve çıktıları CRISP-DM Veri Madenciliği Rehberinden (Chapman vd., 2000, s.21-23) tablolaştırılarak özetlenmiştir.

Tablo 1.3. Verinin Hazırlanması Aşamasında Tanımlanan Görevler ve Çıktıları

Görevler ve *Çıktıları*	**Açıklama**
Verinin Seçilmesi	Analizler için kullanılacak olan veriye karar verme görevidir. Burada kriterler, veri madenciliği hedeflerine uygunluk, kalite ve veri miktarı ve veri tipiyle ilgili limitler gibi teknik kısıtlardır. Verinin seçimi bir veritabanı tablosundaki kayıtlar ile birlikte alanları (değişkenleri) de kapsamaktadır.
Dahil Etme / Hariç Tutma Gerekçeleri	Dahil edilecek ve hariç tutulacak verinin bir listesidir ve aynı zamanda bu kararların gerekçeleri de burada yer almalıdır.
Verinin Temizlenmesi	Verinin, seçilen analiz tekniklerinin gerektirdiği veri kalitesi düzeyine eriştirilmesi görevidir. Bu görev içerisinde, verinin temiz alt-setlerinin seçilmesi, uygun varsayılan değer (default) ataması veya kayıp verinin modelleme yoluyla tahmini gibi daha iddialı teknikler yer alabilir.
Veri Temizleme Raporu	Verinin İncelenmesi aşamasındaki Veri Kalitesinin Doğrulanması görevinde raporlanan veri kalitesi problemleri ile ilgili ne gibi kararların alındığını ve hangi faaliyetlerin gerçekleştirildiğini açıklayan bir rapordur.
Verinin Yapılandırılması	Bu görev, yeni alanlar (değişkenler) veya yeni kayıtlar elde edilmesi, veya hali hazırdaki alanların değerlerinin dönüştürülmesi gibi yapılandırıcı veri hazırlama faaliyetlerini içermektedir.
Elde Edilen Alanlar	Yeni alanlar, bir veya daha fazla eski alanın değerleri kullanılarak aynı kayıt için hesaplanan yeni değerleri içermektedir.
Üretilen Kayıtlar	Tamamen yeni bir kaydın üretilmesini tanımlar. Örneğin bir yıl içinde hiç alışveriş yapmamış bir müşterinin ham veride kaydının olamaması normaldir, ancak analiz amacı için veritabanında yer alan bir müşteri için sıfır alışverişi temsil eden bir kaydın eklenmesi gerekebilir.
Verinin Birleştirilmesi	Burada yer alan yöntemler, birden fazla tablodan ve kayıttan gelen veriyi kullanarak, yeni kayıt veya değerler oluşturulması süreçlerini kapsamaktadır.
Birleştirilmiş Veri	Tabloların birleştirilmesi, iki veya daha fazla tabloda yer alan aynı nesneye ait farklı bilgilerin bir araya getirilmesi işlemidir. Birleştirilmiş veri aynı zamanda birçok kayıt ve tablodan gelen bilgilerin özetlenerek yeni değerler, kayıtlar ve tablolar elde edilmesini de kapsar.
Verinin Biçimlendirilmesi	Bu aşamada, modelleme aracı tarafından ihtiyaç duyulabilecek olan, verinin anlamını değiştirmemekle birlikte biçimsel veya sözdizimsel (syntactic) düzenlemeler yapılabilir.
Yeniden Biçimlendirilmiş Veri	Veri setindeki kayıtların sırasının değiştirilmesi önemli olabilir. Örneğin yapay sinir ağları kullanımında veri sunumunun rasgele sıra ile yapılması önemlidir. Burada örneğin bir alan için tüm değerlerin 32 karakter ile sınırlandırılması gibi işlemler de yapılabilir.

1.3.2.4. Modelleme

Modelleme aşaması, veri madenciliği yazılımı yardımıyla uygun teknikler kullanılarak farklı durumlar için sonuçlar üretilmesi aşamasıdır. Genellikle ilk olarak kümeleme analizi ve verinin görselleştirilmesi teknikleri kullanılmaktadır. Verinin tipine göre daha sonra çeşitli modeller uygulanabilir. Eğer amaç verinin gruplandırılması ise ve gruplar belirli ise, diskriminant analizi uygun olabilir. Eğer amaç tahmin ise ve veri sürekli ise regresyon analizi, eğer veri sürekli değilse lojistik regresyon uygun olabilir. Her iki amaç için de yapay sinir ağları kullanılabilir. Verinin sınıflandırılması için karar ağaçları da başka bir teknik olarak kullanılabilmektedir (Olson ve Shi, 2007, s. 24).

Veri madenciliği modelleme tekniklerinin seçiminde bazı kriterler sırasıyla takip edilmelidir. Öncelikle analize uygun teknikler belirlenmeli ve mevcut veri madenciliği araçlarının hangi teknikleri barındırdığı da dikkate alınmalıdır. Sonrasında organizasyonun hedefleri ve yönetimin istekleri göz önünde bulundurulmalıdır. Son olarak ise zaman, personelin eğitim ve bilgi durumu gibi kısıtlar değerlendirilmelidir.

CRISP-DM metodolojisinde yer alan Modelleme aşaması görevleri ve bunlara ait çıktılar Tablo 1.4'de özetlenmiştir (Chapman vd., 2000, s. 23-25). Bununla birlikte, yaygın olarak kullanılan veri madenciliği tekniklerine ayrıntılı olarak bu çalışmanın Veri Madenciliği Teknikleri bölümünde yer verilmiştir.

Tablo 1.4. Modelleme Aşamasında Tanımlanan Görevler ve Çıktıları

Görevler ve *Çıktıları*	**Açıklama**
Modelleme Tekniğinin Seçilmesi	Modellemenin birinci adımı olarak, kullanılacak olan kesin modelleme tekniğinin seçilmesi görevidir. Problemin Tanımlanması aşamasında, Proje Planının Üretilmesi görevinde bir araç belirlenmiş olmalıdır, ancak burada örneğin C5.0 algoritması ile karar ağacı yapılandırılması veya geri yayılım algoritması kullanan yapay sinir ağı modellemesi gibi spesifik modelleme tekniğinin seçilmesi gerekmektedir.
Modelleme Tekniği	Kullanılacak spesifik modelleme tekniğinin dokümanlaştırılması.
Modelleme Varsayımları	Birçok modelleme tekniği veri ile ilgili varsayımlar yapmaktadır. İlgili varsayımlar burada kayıt altına alınmalıdır.
Test Tasarımının Üretilmesi	Modelin yapılandırılmasından önce, modelin kalite ve geçerliliğini test etmek amacıyla bir prosedür veya mekanizmanın oluşturulması gerekmektedir. Örneğin, ortalama mutlak yüzde hata gibi hata oranları veri madenciliği modellerinin kalitesi için ölçü olarak kullanılabilir.
Test Tasarımı	Modellerin eğitilmesi, test edilmesi ve değerlendirilmesi için planların tanımlanması. Bu planda birincil bileşen eldeki veri setinin eğitim, test ve doğrulama setleri olarak nasıl ayrılacağının belirlenmesidir.
Modelin Yapılandırılması	Hazırlanan veri seti üzerinde bir veya daha fazla model üretmek üzere modelleme aracının koşturulması görevidir.
Parametre Ayarları	Bütün modelleme araçlarında genellikle ayarlanması gereken birçok parametre bulunmaktadır. Burada ilgili parametreler ve ayarları, değerlerinin seçilme nedenleriyle birlikte listelenmelidir.
Modeller	Burada modelleme aracı tarafından üretilen gerçek modeller yer almalıdır.
Model Tanımlamaları	Sonuçlanan modellerin tanımları. Burada modellerin yorumları ve anlamlarıyla ilgili karşılaşılan güçlükler raporlanmalıdır.
Modelin Değerlendirilmesi	Veri madenciliği analisti modelleri, alan bilgisi, veri madenciliği başarı kriterleri ve amaçlanan test tasarımına göre değerlendirir. Bu görev yalnızca modellerin değerlendirilmesini amaçlamaktadır. Bir sonraki Değerlendirme aşaması ise projenin tamamıyla ilgilenmektedir.
Model Değerlendirme Raporu	Değerlendirme görevinin sonuçlarının özetinin yer aldığı bir rapordur. Burada üretilen modellerin kaliteleri ve birbirlerine göre sıralamaları listelenmelidir.
Revize Edilen Parametre Ayarları	Model değerlendirmelerine göre parametre ayarları revize edilerek bir sonraki Modelin Yapılandırılması görevi için ayarlanmalıdır. Modelin Yapılandırılması ve Modelin Değerlendirilmesi görevleri en iyi model veya modellerin elde edildiğine inanılana kadar devam ettirilmelidir. İlgili revizyonlar ve değerlendirmeler burada raporlanmalıdır.

1.3.2.5. Değerlendirme

Model sonuçları, birinci aşamanın birinci görevinde belirlenen iş hedefleri doğrultusunda değerlendirilmelidir. İşin kavranması, veri madenciliğinde iteratif bir süreçtir. Çeşitli görselleştirme, istatistik ve yapay zeka araçları analiste yeni ilişkiler gösterebilir ve bu yolla işletme faaliyetleri konusunda başlangıçtakinden daha derin bir anlayışa sahip olunabilir. Yorumlama ve değerlendirme aşaması önemlidir çünkü, veri madenciliği sonuçlarının özümsenmesi bu aşamada gerçekleşmektedir. Burada iki önemli konu vardır. Birincisi, veri madenciliği aşamasında elde edilen bilgilerin, iş değeri olarak nasıl algılanacağıdır. İkincisi ise, veri madenciliği sonuçlarının gösterilmesi için hangi görselleştirme araçlarının ve tekniklerinin kullanılacağıdır (Olson ve Shi, 2007, s. 20-26). Tablo 1.5'de bu aşama için CRISP-DM veri madenciliği rehberinde (Chapman vd., 2000, s. 26-27) tanımlanmış görevler ve çıktıları tablolaştırılarak özetlenmiştir.

Tablo 1.5. Değerlendirme Aşamasında Tanımlanan Görevler ve Çıktıları

Görevler ve *Çıktıları*	**Açıklama**
Sonuçların Değerlendirilmesi	Bu aşama modelin iş hedeflerini karşılama derecesinin ve eğer varsa hangi işletme nedenleriyle modelin eksik olduğunun belirlenmesiyle ilgilenmektedir.
Veri Madenciliği Sonuçlarının Derecelendirilmesi	Burada değerlendirme sonuçlarının iş hedefleri kriterleri yönünden özetlenmesi ve projenin başlangıçtaki iş hedeflerini karşılayıp karşılamadığının tanımlanması yapılmalıdır.
Onaylanmış Modeller	Modellerin başarı kriterlerine dayalı olarak değerlendirilmesinden sonra, seçilen kriterleri sağlayan modeller onaylanmış modeller olarak belirlenir.
Gözden Geçirme Süreci	Bu görevde, kalite güvence konuları açısından önemli faktörler ve görevler gözden geçirilmelidir ve örneğin modelin doğruluğu veya kullanılan değişkenler kontrol edilmelidir.
Gözden Geçirme Raporu	Gözden geçirme işlemleri raporlanmalı ve ihmal edilen veya tekrarlanması gereken görevler raporlanmalıdır.
Gelecek Aşamaların Belirlenmesi	Bu aşamada proje ekibi, projenin tamamlanması ve uygulamaya geçilmesine, daha fazla iterasyon yapılmasına veya yeni bir veri madenciliği projesine başlanmasına karar verebilir.
Muhtemel Faaliyetlerin Listesi	Gerçekleştirilmesi muhtemel faaliyetlerin, her seçenek için olumlu ve olumsuz nedenleri ile birlikte listelenmesidir.
Karar	Kararlar ve nasıl gerçekleştirilecekleri, dayanakları ile birlikte tanımlanmalıdır.

1.3.2.6. Uygulama

Veri madenciliği çalışması yeni bilgileri ortaya çıkarır ve bu bilgiler veri madenciliği proje hedefleri ile bağlantılı olmalıdır. Bu sayede yönetim iş ortamı ile ilgili bu yeni anlayışı uygulamak için bir pozisyon alabilir. Belirli bir veri madenciliği çalışmasından elde edilen bilginin değişim için izlenmesi önemlidir. Çünkü veri toplandığı zaman doğru olan bir durum hali hazırda değişmiş olabilir. Eğer temel bir değişiklik gerçekleşmişse elde edilen bilgiler artık doğru olmayabilir. Bundan dolayı, ilgi alanı uygulama süreci boyunca izlenmelidir (Olson ve Shi, 2007, s. 27). Tablo 1.6'da CRISP-DM metodolojisi için konsorsiyum tarafından hazırlanan rehberden (Chapman vd., 2000, s. 28-29) yararlanarak özetlenen Uygulama aşaması görevleri ve bunlara ait çıktılar verilmiştir.

Tablo 1.6. Uygulama Aşamasında Tanımlanan Görevler ve Çıktıları

Görevler ve *Çıktıları*	**Açıklama**
Uygulamanın Planlanması	Bu görevde değerlendirme sonuçları alınarak, uygulama için bir strateji belirlenir.
Uygulama Planı	Bu plan, uygulama stratejisini özetler ve içeriğinde uygulamada gerekli adımlar ve bunların nasıl gerçekleştirileceği yer alır.
İzleme ve Bakımın Planlanması	İzleme ve bakım özellikle veri madenciliği sonuçları işletmenin günlük faaliyetlerinin bir parçası haline geldiyse önemli konulardır. Bakım stratejisinin dikkatli hazırlanması veri madenciliği sonuçlarının uzun süreler yanlış kullanımını engelleyecektir.
İzleme ve Bakım Planı	İzleme ve bakım stratejisini, gerekli adımlar ve bunların nasıl gerçekleştirileceği konularını da içerecek şekilde özetleyen plandır.
Sonuç Raporunun Üretilmesi	Projenin sonunda, proje ekibi bir sonuç raporu hazırlamalıdır. Uygulama planına da bağlı olarak, bu rapor yalnızca projenin ve projede elde edilen deneyimlerin bir özeti olabilir veya veri madenciliği sonuçlarının son ve kapsamlı bir sunumu olabilir.
Sonuç Raporu	Bu doküman, veri madenciliği yükümlülüğünün son yazılı raporudur. Önceki aşamalardaki çıktıları, sonuçların özet ve organizasyonlarını içerir.
Son Sunum	Veri madenciliği projesinin sahibi olan tarafa projenin sonuçlarının sunumu için ayrıca bir toplantı da düzenlenmelidir.
Projenin Gözden Geçirilmesi	Bu görev, nelerin iyi gittiği nelerin kötü gittiği, nelerin iyi yapıldığı ve nerelerde iyileştirmeler olabileceğinin değerlendirilmesidir.
Deneyim Dokümantasyonu	Proje boyunca elde edilen önemli deneyimlerin özetlendiği bir dokümandır. İdeal bir projede deneyim dokümantasyonu, projenin herhangi bir aşamasında takım üyeleri tarafından yazılmış bireysel raporların tümünü kapsamalıdır.

1.4. Veri Madenciliği Uygulama Örnekleri

Veri madenciliği farklı işletme ve kurumlarda birçok alanda ve çeşitli amaçlarla uygulanmaya başlanmıştır. Bu çalışmada sağlık alanı veritabanlarında bilgi keşfine odaklanıldığı için bu bölümde sağlık alanında yapılmış çalışmaları ve uygulamaları içeren yayınlar, aralarında tıp alanı veritabanları da olmak üzere önemli bilimsel veritabanları kullanılarak elde edilmiş ve incelenmiştir. Sağlık alanındaki çalışmaların amaçları sağlık kuruluşlarının yönetiminde karar desteği sağlamak olabileceği gibi tıbbi konularda da veri madenciliği teknolojileri yoluyla etkinlik sağlamak olabilmektedir. Tablo 1.7'de incelenen çalışmaların yazarları ve yayın yılı, kullanılan veri madenciliği teknikleri ve çalışmada ne yapıldığının kısa açıklaması olmak üzere üç tip bilgi verilmiştir.

Tablo 1.7. Sağlık Sektöründe Yapılmış Veri Madenciliği Uygulamaları

Yazarları ve Yılı	**Kullanılan Teknikler**	**Çalışmanın Amacı, Yöntemi ve Sonuçları**
Oztekin vd. (2009)	Yapay Sinir Ağları, Karar Ağaçları, Lojistik Regresyon	Çalışma, birleştirilmiş kalp ve akciğer organ naklinde izleyen dönemlerdeki sonuçların kestirimini geliştirmek için tümleşik bir veri madenciliği yöntemi önermektedir. Organ Paylaşımı için Birleşik Ağ (United Network for Organ Sharing, UNOS) isimli tıbbi, bilimsel ve eğitimle ilgili bir kuruma ait 16.604 vaka kaydı ve 283 değişken içeren geniş bir veri seti kullanılarak, makine öğrenimi temelli kestirimci modeller geliştirilmesi ve en önemli kestirimci faktörlerin ortaya çıkarılması amaçlanmıştır. Sonuçlara göre yapay sinir ağları modelinin doğruluk oranları farklı veri setlerinde %79 ile %86 arasında, lojistik regresyon için %78 ile %86 ve karar ağaçları için de %71 ile %79 arasında değişmiştir.
Delen vd. (2009)	Yapay Sinir Ağları, Karar Ağaçları	A.B.D.'de birçok bireyin sağlık hizmetleri kapsamına sahip olmadığı ve bu olguya neden olan faktörlerin araştırılması gerektiği vurgulanmaktadır. Çalışmada A.B.D.'de bireylerin sağlık hizmetleri kapsamlarını çok çeşitli kestirimci faktörler üzerinde popüler makine öğrenimi tekniklerinin kullanılarak incelenmesi amaçlanmaktadır. 23 değişken ve 193.373 kayıttan oluşan bir risk faktörleri denetim sistemi verisi kullanılmıştır. Yapay sinir ağarı ve karar ağaçları modelleri geliştirilmiş ve kestirimci yetenekleri bakımından birbiri ile karşılaştırılmıştır. Sonuçlara göre bu olgu için en iyi sınıflandırıcının %78,45 doğru sınıflandırma oranı ile ileri beslemeli yapay sinir ağı modeli olduğu görülmüştür.

Tablo 1.7. devamı

Yazarları ve Yılı	Kullanılan Teknikler	Çalışmanın Amacı, Yöntemi ve Sonuçları
Zhuang vd. (2009)	Vaka Tabanlı Muhakeme (Case-Based Reasoning), Kümeleme, Kohonen Ağları	Avustralya'da pratisyen hekimlerin patoloji istemleri için karar desteği sağlayacak Veri Madenciliği ve Vaka Tabanlı Muhakeme yöntemlerini entegre eden bir metodoloji önerilmektedir. Veri seti olarak Avustralya'da pratisyen hekimler tarafından istenen 1,5 milyonun üzerindeki patoloji istem kayıtları kullanılmıştır. Kohonen Ağları (veya Kendini Düzenleyen Haritalar) hastaların demografik özellikleri ve patoloji hizmeti alma örüntülerine dayalı olarak homojen hasta gruplarını belirlemek amacıyla kullanılmıştır.
Chen vd. (2009)	Yapay Sinir Ağları, Destek Vektör Makineleri, Lojistik Regresyon	Çalışma hasta güvenliğini sağlamada önemli bir konu olan gecikmiş tanılama problemini incelemeyi amaçlamaktadır. Sonuçlar arasında çıkan ilginç bir bulgu eğer hasta normal bir şekilde nefes alabiliyor fakat kan basıncı veya nabzında anormallik varsa yüksek olasılıkla gecikmiş tanılamanın var olduğu görülmüştür.
Razali ve Ali (2009)	Karar Ağaçları	Çalışma, sağlık hizmetleri sağlayan kurumlar için tedavi süreçlerindeki karmaşıklığı gidermek ve hataları yönetmek için bir tedavi planı oluşturmayı amaçlamaktadır. Ayakta tedavi gören poliklinik hastalarına odaklanılmıştır ve Malezya'daki farklı sağlık merkezlerinden toplanan veri kullanılmıştır. Araştırma sürecinde CRISP-DM metodolojisi kullanılmıştır. Karar ağaçları (C5 algoritması) ile geliştirilen modelin doğruluk derecesi %94,73 olmuştur ve araştırma bulguları sonuçlarına göre kullanılacak teknolojilerin sağlık hizmeti sağlayan kurumlara sağlayacağı fayda ile birlikte tedavi sürecinde oluşacak hataları azaltabileceği vurgulanmıştır.
Batyrshin ve Sheremetov (2008)	Zaman Serileri	Zaman Serileri Veri Madenciliği (Time Series Data Mining, TSDM), sınırlı belirlilik altında karar vermeye rehberlik edebilecek bilgi yapılarını ortaya çıkarmak amacıyla kullanılan bilgi yönetimi aracı olarak giderek önemi artan bir yöntemdir. Çalışmada zaman serileri veri madenciliği konusunda birçok etkin algoritma bulunmasına rağmen bunların insanların karar verme süreçlerine entegre edilmesinin halen açık bir problem olduğu vurgulanmış ve zaman serileri veritabanlarında TDSM yöntemleri ile kelime ve algılarla programlama ve uzman bilgileri entegre edilerek bir algı-temelli karar verme sistemi mimarisi önerilmektedir.

Tablo 1.7. devamı

Yazarları ve Yılı	Kullanılan Teknikler	Çalışmanın Amacı, Yöntemi ve Sonuçları
Silva vd. (2008)	Yapay Sinir Ağları, Lojistik Regresyon	Yoğun bakım ünitelerinde yatan hastaların organ yetmezliği riskini tahmin etmek için, olumsuz vakalara dayalı olarak bir veriden yola çıkan bir yaklaşım sunulmuştur. Oldukça geniş bir veritabanı kullanılmıştır. Avrupa'daki 42 yoğun bakım ünitesinden 4.425 hastaya ait 25.215 günlük kayıt içermektedir. Girdi değişkenleri yaş, tanı, başvuru tipi, sevk eden yer gibi vaka için tanımlayıcı değişkenler ve sistolik kan basıncı, kalp hızı, puls oksimetre oksijen doygunluğu, idrar tahlili sonuçları gibi fizyolojik değişkenlerden oluşmaktadır. Çıktı değişkeni ise normal, işlev bozukluğu ve organ yetmezliği olmak üzere üç kategoriye sahip organ durumu değişkenidir. Bu analizler solunum, dolaşım, karaciğer (hepatik), kardiyovasküler, sinir (nörolojik), ve boşaltım (böbrek, renal) olmak üzere altı organ sistemi için yapılmıştır. R yazılımı kullanılarak yapılan iki veri madenciliği yönteminin bu amaçla karşılaştırılmasında, yapay sinir ağları altı sistem için ortalama %74 tahmin doğruluğu ile oldukça farklı olarak daha iyi sonuçlar vermiştir.
Phillips-Wren vd. (2008)	Karar Ağaçları, Lojistik Regresyon, Yapay Sinir Ağları	Çalışmada; 1) Sağlık hizmetleri kaynaklarından akciğer kanseri hastalarının yararlanma durumlarını demografik karakteristiklerine, sosyo-ekonomik göstergelerine, etnik kökenlerine, tıbbi geçmişlerine ve sağlık hizmetleri kaynaklarına erişim durumlarına göre değerlendirilmesi; 2) Geleneksel istatistik yöntemlerle birlikte kullanıldığında, veri madenciliği yöntemlerinin akciğer kanseri hastaların sağlık hizmetlerinden yararlanmaları için değerli yeni bilgiler ve yeni anlayışlar ortaya koyabileceğinin gösterilmesi; 3) Veri madenciliğinin büyük ve halka açık Medicare kullanım verisine uygulanabilirliğinin incelenmesi; ve 4) Eğilim dereceleme yöntemlerinin veri madenciliği ve istatistik yöntemlerin performanslarını değerlendirmede kullanılabileceğinin gösterilmesi amaçlanmıştır. Yazılım olarak SAS Enterprise Miner ve veri madenciliği metodolojisi olarak SEMMA kullanılmıştır. Sonuçlara göre karar ağaçları (CART ve CHAID) ve yapay sinir ağlarının (ileri beslemeli ve tek gizli katmanlı) özellikle birlikte kullanıldığında sağlık hizmetleri kararlarına destek sağlayabilecek kestirimci ve tanımlayıcı modelleri lojistik regresyona göre daha iyi ürettiği görülmüştür.

Tablo 1.7. devamı

Yazarları ve Yılı	Kullanılan Teknikler	Çalışmanın Amacı, Yöntemi ve Sonuçları
Jonsdottir vd. (2008)	Naïve Bayes Sınıflandırıcı, Karar Ağaçları	Klinik uygulamalarda göğüs kanseri tanısı konulan hastalar risk gruplarına göre sınıflandırılmaktadırlar. Çalışmada göğüs kanseri için bir kestirimci sonuç modeli geliştirilmiş ve kanser vakasının 5 yıllık sonuçlarını doğru olarak tahmin etme yetenekleri değerlendirilmiştir. İzlanda'da göğüs kanseri tanısı konulmuş 257 kadına ait 100 farklı veri seti kullanılmıştır. Temel sonuç olarak kullanılan algoritmadan bağımsız olarak benzer sonuçlar gözlenmiştir.
Ramon vd. (2007)	Karar Ağaçları, Bayes Ağları	Yoğun bakım ünitesinde yatan hastaların gelişimlerini tahmin etmek için veri madenciliği yöntemlerinin uygulanmasını tanımlanmaktadır. İlgili metotların sağlık hizmetlerindeki önemi tartışılmıştır. Üç ana başlıkta toplanabilen veri kullanılmıştır. Bunlar: Hastaların demografik bilgileri, geçmiş hastalık, tanı ve tedavi bilgileri ve yoğun bakım ünitesinde kalış süresince tutulan verilerdir. Son başlıktaki veriler de klinik parametre ölçümleri (ateş, tansiyon, vb.), laboratuar verisi, bakteriyolojik veri (enfeksiyonlara ait), doktor ve hemşirelerin gözlemleri, verilen ilaçlar, tedaviler (fizik tedavi, ameliyat, vb.), hastaların beslenme verisi ve tedavi planı kararları olarak sıralanmaktadır.
Kuo ve Shih (2007)	Birliktelik Kuralları, Karınca Kolonisi Sistemi	Birliktelik kurallarının oluşturulmasında önemli bir aşama sık öğe setlerinin keşfedilmesidir. Çalışma veri madenciliği analisti tarafından tanımlanan kısıtları göz önüne alarak kısıt-temelli veri madenciliği yoluyla analist tarafından önem verilen öğe setleri üzerine odaklanmaya ve dolayısıyla veri madenciliği görevlerinin etkinliğinin arttırılmasına yönelik bir çok-boyutlu kısıtlar problemi çözümü amaçlamaktadır. Sağlık Sistemi Veritabanı kullanılarak yapılan çalışmada, tümleşik optimizasyon problemleri için yeni bir meta-sezgisel teknik olan Karınca Kolonisi Sistemi ile Birliktelik Kuralları tekniğini (Apriori) birlikte kullanarak çok-boyutlu kısıtları da göz önüne alan ve birliktelik kurallarını daha etkin bir biçimde bulabilecek bir yöntem ve bir algoritma önerilmektedir.
Lavrač vd. (2007)	Aykırı Değer Tespiti, Görselleştirme	Çalışmada Slovenya'da Celje Bölgesinde yer alan 11 Kamu Sağlık Merkezinin verileri kullanılarak bu bölgeye ait temelde tanımlayıcı istatistikler ve bir matematiksel model kullanılarak Kamu Sağlık Merkezinin uygunluğu değerleri elde edilmiş ve görselleştirme teknikleri ile analiz edilmiştir.

Tablo 1.7. devamı

Yazarları ve Yılı	Kullanılan Teknikler	Çalışmanın Amacı, Yöntemi ve Sonuçları
Mullins vd. (2006)	Birliktelik Kuralları	Çalışma bir akademik medikal sistemde 667.000 adet yatan ve ayakta tedavi gören hastanın verisi kullanılarak yapılan çalışma klinik veri depolarından kestirimci analiz ve örüntü keşfi yapmayı amaçlamaktadır. Sonuçları itibariyle klinik hastalık birlikteliklerinin tanımlanması yoluyla önerilen yaklaşımların araştırma yeteneklerini geliştirebileceği belirtilmektedir.
Yang ve Hwang (2006)	Birliktelik Kuralları	Çalışma sağlık hizmetlerinde dolandırıcılık ve kötüye kullanma olaylarının ortaya çıkarılması için bir veri madenciliği sistemi önermektedir. Tayvan Ulusal Sağlık Sigortası Programı verileri kullanılan çalışma sağlık hizmetleri hizmet sağlayıcılarının hileli ve yolsuzluk barındıran hareketlerini tespit edecek modelin geliştirilmesini kolaylaştırmak için izlenen klinik yolları kullanmıştır.
Poynton ve McDaniel (2006)	Yapay Sinir Ağları	Çalışma geri yayılım algoritması (backpropagation) kullanan yapay sinir ağı sınıflandırıcısının hali hazırda sigara içenleri ve geçmişte sigara içenleri sınıflandırabilme yeteneğini incelemektedir. Amerikan Ulusal Sağlık Araştırması verisi kullanılan çalışmada analizler 14.416 yetişkin kaydı ve 1.429 değişken kullanılarak gerçekleştirilmiştir. Sigara bırakma danışmanlığı için karar destek sistemlerinde geri yayılım algoritması kullanan yapay sinir ağlarının faydalı olabileceği belirtilmektedir.
Delen vd. (2005)	Karar Ağaçları, Yapay Sinir Ağları, Lojistik Regresyon	Çalışmanın amacı göğüs kanseri vakalarında hayatta kalabilirliği tahminlemek için üç veri madenciliği yöntemini karşılaştırmaktadır. 200.000'den fazla kayıt içeren büyük bir veri seti kullanılmıştır. Sonuçlara göre karar ağaçları (C5 algoritması) %93,6 tahmin doğruluğu ile en iyi kestirimci olmuştur. Yapay sinir ağları %91,2 ve lojistik regresyon ise %89,2 tahmin doğruluğu göstermiştir.
Gren vd. (2004)	Kümeleme Analizi, Sınıflandırma	Michigan Üniversitesi Çok-disiplinli Ağrı Merkezi'nin ikincil veritabanı kullanılarak kronik ağrı problemi yaşayan Afrika kökenli ve Beyaz Amerikalılar arasında etnik köken, yaş ve cinsiyet etkilerini incelemeyi amaçlamaktadır. Benzer fiziksel, duygusal ve ağrı karakteristiklerine sahip hastalar arasında önemli etnik ve yaş bağlantılı değişkenlikler bulunmuştur.

Tablo 1.7. devamı

Yazarları ve Yılı	Kullanılan Teknikler	Çalışmanın Amacı, Yöntemi ve Sonuçları
Neumann vd. (2004)	Lojistik Regresyon, Karar Ağaçları	Fransa'da 106 Yoğun Bakım Ünitesinde yatan 87.099 hasta veri seti ile yapılan çalışmada yoğun bakımda ölüm oranı ve potansiyel olarak önlenebilir hastane yeniden kabulleri olmak üzere iki önemli performans göstergesini kullanılmıştır. Lojistik regresyon ve karar ağaçları birleştirilmiş olarak kullanılarak sağlık hizmetleri performans değerlendirmesi yapılmıştır.
Semenova (2004)	Birliktelik Kuralları	Büyük yönetimsel sağlık veritabanlarında tıbbi uygulamalarla ilgili örüntülerin keşfedilmesi konusunda karmaşıklığı azaltıcı alternatif bir teknik önerilmektedir. Kullanılan veritabanında 3.617.556 hasta kaydı, 2.145.864 vaka ve 13.192.395 işlem kaydı bulunmaktadır
Lucas (2004)	Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees – CART), Yapay Sinir Ağları, Destek Vektör Makineleri, Bayes Sınıflandırıcı	Çalışma veri madenciliği tekniklerinin ve Bayesyan yöntemlerin biyomedikal araştırmalar ve sağlık hizmetlerindeki hali hazırdaki rolünü tartışmayı amaçlamaktadır. Bayes ağlarının ve diğer olasılıksal grafik modellerin biyomedikal veri içerisinde örüntülerin keşfedilmesine yönelik olarak ve klinik karar almada altta yatan belirsizliklerin gösterilmesinde temel olarak yükselen yöntemler olduğu belirtilmektedir. Aynı zamanda makine öğrenimi tekniklerinin de biyomedikal ve sağlık hizmetleri ile ilgili problemlerin çözümünde kullanıldığı belirtilmektedir.
Jerez-Aragonés vd. (2003)	Yapay Sinir Ağları, Karar Ağaçları	Çalışmada İspanya Malaga Üniversitesi Hastanesi Medikal Onkoloji Servisinde 1990 ile 2000 yılları arasındaki 10 yıllık periyotta yatan göğüs kanseri hastaları verisi kullanılarak hastalığın sonucunu tahmin etmek (prognoz) için bir karar destek yöntemi önerilmektedir. 85 değişken arasından 14 değişken çalışmanın bağımsız değişkenleri olarak belirlenmiş ve sonucunda da kural sayısını azaltan ve anlamlı prognoz faktörleri olan değişkenleri seçmeyi etkinleştiren bir karar ağacı algoritması önerilmiştir.

Tablo 1.7. devamı

Yazarları ve Yılı	Kullanılan Teknikler	Çalışmanın Amacı, Yöntemi ve Sonuçları
Ma vd. (2003)	Birliktelik Kuralları	Hastane enfeksiyonu kontrol programının özellikle önemli fonksiyonlarından biri olan antibiyotik direnci ve nozokomiyal enfeksiyon gözetimi konusunda yeni, beklenmeyen ve potansiyel olarak önemli olabilecek örüntüleri keşfetmek amacıyla Birliktelik Kuralları yöntemi ve Apriori Algoritması uygulanmıştır. Çalışmada 10 adet hastanenin verisi kullanılmış ve özellikle düşük destek ve güven değerlerine sahip kurallar aranarak az sayıda vakada karşılaşılan beklenmeyen hastane enfeksiyonu salgınlarının tespit edilmesi amaçlanmıştır.
Chen vd. (2003)	Birliktelik Kuralları	Çalışmanın amacı Tayvan'da Ulusal Sağlık Sigortası Programı çerçevesinde geri ödemesi yapılan antasit ilaçların reçete edilmesinin ölçeğini tahmin etmek ve bu ilaçlarla birlikte yazılan ilaçların örüntülerini ortaya çıkarmaktır. Tayvan Ulusal Sağlık Sigortası Araştırma Veritabanında yer alan veriler kullanılmış ve Birliktelik Kuralları antasitlerle birlikte reçete edilen ilaçları belirlemek amacıyla kullanılmıştır. Birliktelik Kuralları analizinde minimum destek seviyesi %1,0 olarak belirlenmiştir.
Chae vd. (2003)	Karar Ağaçları	Kalite yükseltme stratejileri geliştirmek için veri madenciliği kullanarak sağlık hizmetleri kalite göstergeleri analizi ortaya konmaktadır. Yatan hasta ölüm oranını etkileyen önemli faktörleri belirlemek amacıyla Karar Ağaçları (CHAID algoritması) yöntemi 8405 hasta kaydı üzerinde uygulanmıştır. Yatan hasta ölüm oranı üzerine etkili olan önemli faktörler yatış süresi, hastalık sınıfı, taburcu edilen departman ve yaş grupları olarak çıkmıştır. Çalışmada ek olarak kalite gösterge trendlerini analiz etmek ve izlemek için bir karar destek sistemi geliştirilmiştir.
Breault vd. (2002)	Sınıflandırma ve Regresyon Ağaçları (CART)	Bir diyabet veritabanında HgbA1c >9,5 olmak üzere ikili (binary) tipteki bağımlı değişken ve 10 adet bağımsız değişken (yaş, cinsiyet, kardiyovasküler rahatsızlık, hipertansiyon, son safha göz rahatsızlığı, vb.) kullanılarak bir sınıflandırma uygulaması gerçekleştirilmiştir. Kötü glisemik kontrol için en önemli değişkenin beklenmeyen bir şekilde genç yaş çıktığı vurgulanmaktadır.

Tablo 1.7. devamı

Yazarları ve Yılı	Kullanılan Teknikler	Çalışmanın Amacı, Yöntemi ve Sonuçları
Nguyen vd. (2002)	Yapay Sinir Ağarı, Lojistik Regresyon	Pediyatrik meningokok hastalığının kötü sonuçlanmasının tahmin modelini oluşturmak amacıyla yapay sinir ağları ve çok değişkenli lojistik regresyon analizi karşılaştırılmıştır. Meningokok hastalığı konusunda önceki çalışmalarda kötü sonuçlanma ile ilgili sekiz değişken kullanılmış ve her iki modelin performans parametreleri arasında anlamlı bir fark görülmemiştir.
Liao vd. (2002)	Diskriminant Analizi, Yapay Sinir Ağları, Genetik Algoritmalar, Karar Ağaçları, Bayes Sınıflandırıcılar	Kalp hastalıkları veritabanı kullanılarak, dört tipteki veri (sürekli, ikili, nominal ve ordinal veri) için yan sütunda sıralanan beş farklı sınıflandırma tekniği uygulanmıştır. Kategorik verinin, kalp hastalığı olan ve olmayan gruplar gibi veri madenciliği sınıflandırma tekniklerinin bir çoğunda kullanışlı olduğu ve tıbbi bilgi çıkarımı için nispeten kolay olduğu sonucu bildirilmektedir.
Goodwin vd. (2001)	Lojistik Regresyon, Yapay Sinir Ağları, Sınıflandırma ve Regresyon Ağaçları (CART)	Erken doğumların erken kestirimcilerini belirlemek amacıyla geleneksel istatistik yöntemler ile veri madenciliği yöntemlerinin karşılaştırılması amaçlanmıştır. Çalışmada Duke Üniversitesi Perinatal Veritabanında bulunan yaklaşık 72.000 kayıt arasından veri temizleme ve filtreleme prosedürlerinden sonra 19.970 hamile kadına ait 1622 değişkene sahip veri seti kullanılmıştır. Veri madenciliği yöntemleri arasında en iyi sonuçları CART algoritmasının verdiği belirtilmiştir.
Richards vd. (2001)	Birliktelik Kuralları	İngiltere'de 21.000 hastanın klinik kayıtlarının tutulduğu bir diyabet hastaları veritabanında yapılan çalışmanın amacı hastaların hastaneye ilk yatışlarında kaydedilen gözlemler ile erken ölüm sonuçları arasındaki ilişkileri (associations) tanımlamaktır. Çıkarılan en önemli birliktelik kuralları tıp topluluğu tarafından genellikle kabul edilmeyen ilişkiler olmakla birlikte yeni bağımsız çalışmaların ilgili birlikteliklerin geçerliğini doğruladığı belirtilmektedir.
Lee vd. (2000)	Görselleştirme, Diskriminant Analizi, Yapay Sinir Ağları	Yüksek riskli kalp hastalığı olan hastaları ve kalp hastalığına neden olan en önemli faktörleri belirlemek amacıyla hali hazırdaki tıbbi bilgiyi temsil eden bir model oluşturulmuş ve yüksek riskli hastaları sınıflandırmak amacıyla iki parametrik olmayan teknik kullanılmıştır. Bunlardan yapay sinir ağları, diskriminant analizine göre daha iyi sınıflandırma başarısı göstermiştir.

İKİNCİ BÖLÜM
VERİ MADENCİLİĞİ TEKNİKLERİ

2.1. Bayes Sınıflandırıcılar

Bayes sınıflandırıcılar istatistiksel sınıflandırıcılardır ve belirli bir değişkenler grubunun belirli bir sınıfa ait sınıf üyeliği olasılıklarını tahmin ederler. Bu sınıflandırma Thomas Bayes'in teoremine (1763) dayanmaktadır (Han ve Kamber, 2006, s.310). Bayes teoremi iki rassal olayın koşullu ve marjinal olasılıklarını ilişkilendirir. Bayes sınıflandırıcılar temel olarak her özniteliğin verilen sınıf etiketine göre sınıfsal koşullu olasılıklarını öğrenir (Hsu vd., 2008, s.1081).

Bir sınıflandırma probleminde amaç belirli kestirimci değişkenler kümesi verildiği durumda, her sınıfa ilişkin üyelik olasılıklarını tahmin etmektir. Bu tür bir olasılık koşullu olasılık olarak adlandırılır. Y olayının verilen X olayına göre koşullu olasılığı $P(Y \mid X)$ sadece X olayının gerçekleştiği durumlarda Y olayının da gerçekleşmesi olasılığını temsil eder. (Shmueli vd., 2007, s.94). X ve Y'nin bağımsız rassal değişkenler çifti olduğu düşünülürse, bunların bileşik olasılıkları $P(X = x, Y = y)$'dir. Bunların koşullu olasılığı ise bir rassal değişkenin diğer rassal değişkenin değerinin bilindiği durumda vereceği sonuç olarak tanımlanabilir. Örneğin $P(Y = y \mid X = x)$ koşullu olasılığı Y değişkeninin, X'in x değerini aldığının gözlendiği bir durumda, y değerini alacağı durumdur. X ve Y'nin bileşik ve koşullu olasılıkları aşağıdaki biçimde ilişkilendirilebilir (Tan vd., 2006, s.228):

$$P(X,Y) = P(Y \mid X) \times P(X) = P(X \mid Y) \times P(Y)$$

Yukarıdaki formülde son iki ifadenin yeniden düzenlenmesi ile Bayes teoremi olarak bilinen aşağıdaki formül elde edilir (Bolstad, 2004, s.63-64):

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Naïve Bayes sınıflandırıcılar Bayes teoremine dayalı daha sade ve bazı veri madenciliği tekniklerine oranla daha hızlı ve yüksek performanslı bir veri madenciliği yöntemidir. Naïve Bayes sınıflandırıcılar belirli bir sınıfta yer alan değişkenin değerinin diğer değişkenlerinin değerlerinden bağımsız olduğunu varsayar. Bu varsayım koşullu bağımsızlık kavramı ile ifade edilir (Han ve Kamber, 2006, s.310).

Koşullu bağımsızlık X, Y ve Z rassal değişkenler setinde X değişkeninin belirli bir Z değerinde Y'den tamamen bağımsız olması durumudur ve aşağıdaki şekilde ifade edilebilir (Tan vd., 2006, s.231):

$$P(X,Y \mid Z) = P(X \mid Z) \times P(Y \mid Z)$$

Koşullu bağımsızlık varsayımı ile X'in her kombinasyonu için koşullu olasılıkları hesaplamak yerine, sadece istenilen Y durumu için her X_i'nin koşullu olasılığını hesaplamak gereklidir. İkinci yaklaşım uygulamaya daha uygundur, çünkü iyi bir olasılık tahmini elde etmek için büyük bir eğitim setine (training set) gerek olmayacaktır (Tan vd., 2006, s.232).

Naïve Bayes sınıflandırıcı aşağıdaki şekilde çalışmaktadır (Han ve Kamber, 2006, s. 311):

1. D değişken grupları ve ilgili sınıf etiketlerinden oluşan bir eğitim seti olsun. Her değişkenler grubu bir n-boyutlu öznitelik vektörü $X = (x_1, x_2, \ldots, x_n)$ ile temsil edilmektedir ve $A_1, A_2, \ldots, A_n$ n adet öznitelikten oluşan bir değişkenler grubunda yapılan n adet ölçümü belirtmektedir.

2. $C_1, C_2, \ldots, C_m$ olmak üzere m adet sınıf olduğu kabul edilirse, verilen bir X değişkenler grubu için sınıflandırıcı X'in, X'e koşullu olmak üzere ($P(C_i \mid X)$), en yüksek ardıl olasılığa (posterior probability) sahip olan sınıfa ait olduğunu kestirecektir. Bu aşağıdaki durum sağlandığında Naïve Bayes sınıflandırıcının X değişkenler grubunun C_i sınıfına ait olduğunu kestirmesi anlamına gelmektedir.

$$P(C_i \mid X) > P(C_j \mid X) \quad ; \quad 1 \le j \le m \, , \, j \ne i$$

Dolayısıyla $P(C_i \mid X)$ maksimize edilmiş olur. $P(C_i \mid X)$'in maksimize edildiği C_i sınıfı maksimum ardıl hipotez olarak adlandırılır. Bunun Bayes teoremi ile ifadesi aşağıdaki şekildedir.

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$$

3. $P(X)$ bütün sınıflar için sabit olduğuna göre, burada sadece $P(X \mid C_i)P(C_i)$'in maksimize edilmesi gerekmektedir. Eğer sınıfların öncül olasılıkları (prior probabilities) bilinmiyorsa, yaygın olarak kabul gören şekilde bütün sınıflar eşit olasılığa sahiptir yani, $P(C_1) = P(C_2) = \cdots = P(C_m)$. Bu durumda $P(X \mid C_i)$'yi maksimize ederiz, aksi durumda ise $P(X \mid C_i)P(C_i)$ maksimize edilir. Sınıf öncül olasılıkları $P(Ci) = |C_{i,D}| / |D|$ ile tahmin edilebilir. Burada $|C_{i,D}|$ C_i sınıfının D eğitim seti (training set) içerisindeki eğitim değişkenler grubu (training tuples) sayısıdır.

4. Birçok öznitelik (attribute) içeren veri setlerinde $P(X \mid C_i)$'nin hesaplanması bilgisayar kaynakları kullanımı açısından oldukça maliyetli olabilir. $P(X \mid C_i)$'nin değerlendirilmesinde hesaplamaların azaltılabilmesi için, Naïve sınıf koşullu bağımsızlık varsayımı yapılır. Bu varsayım, özniteliklerin değerlerinin birbirinden koşullu bağımsızlığını öngörür. Yani öznitelikler arasında bir bağımlılık ilişkisi yoktur. Dolayısıyla,

$$\begin{aligned} P(X \mid C_i) &= \prod_{k=1}^{n} P(x_k \mid C_i) \\ &= P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \cdots \times P(x_n \mid C_i) \end{aligned}$$

$P(x_1 \mid C_i), P(x_2 \mid C_i), \ldots, P(x_n \mid C_i)$ olasılıkları eğitim değişken gruplarından kolaylıkla hesaplanabilir. Burada x_k, X değişkenler grubu için A_k özniteliğinin aldığı değeri temsil etmektedir. Her öznitelik için, özniteliğin kategorik veya sürekli bir değişken olup olmadığına bakılmalıdır. $P(X \mid C_i)$'yi hesaplamak için aşağıdaki durumlar göz önüne alınmalıdır.

a. Eğer A_k kategorik ise, bu durumda $P(X_k \mid C_i)$, C_i sınıfının D içindeki A_k'nın x_k değerini aldığı değişken grubu sayısının, C_i sınıfının D içindeki toplam değişken grubu sayısına ($|C_{i,D}|$) oranıdır.

b. Eğer A_k sürekli bir değişken ise, biraz daha fazla ancak daha kolay bir hesaplama yapmak gereklidir. Sürekli değer alan bir değişkenin tipik olarak aşağıda tanımlandığı şekilde μ ortalaması ve σ standart sapmasına sahip Gauss dağılımına sahip olduğu varsayılır.

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

böylece

$$P(x_k \mid C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

Burada μ_{C_i} ve σ_{C_i}'yi, yani C_i sınıfının eğitim değişken grupları için A_k özniteliğinin değerlerine ait ortalama ve standart sapmayı hesaplamak gerekmektedir. Sonra bu iki değer Gauss dağılımı denkleminde $P(x_k \mid C_i)$'yi tahmin etmek üzere x_k ile birlikte yerine konulur.

5. X'in sınıf etiketini kestirmek için her sınıf için $P(X \mid C_i)P(C_i)$ değerlendirilir. Sınıflandırıcı aşağıdaki durum gerçekleştiğinde X değişkenler grubunun sınıf etiketini C_i olarak kestirir.

$$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j) \quad ; \quad 1 \le j \le m,\ j \ne i$$

Diğer bir deyişle, $P(X \mid C_i)P(C_i)$'nin maksimum olduğu durumda kestirilen sınıf etiketi C_i'dir.

Naïve Bayes sınıflandırıcıların iyi tarafı basitliğine, bilgisayar kaynakları ve hesaplamalar açısından etkinliğine ve iyi sınıflandırma performansına dayandırılabilir. Amacın veri setindeki kayıtların belirli bir sınıfa aidiyetlerinin olasılık değerlerine göre sınıflandırılması veya derecelendirilmesi olduğu durumlarda Naïve Bayes sınıflandırıcılar iyi performans göstermektedir (Shmueli vd., 2007, s.100-103). Ancak amacın asıl olarak sınıf üyelikleri değerlerini tahmin etmek olduğu durumlarda, bu metod yanlı sonuçlar üretebilmektedir.

Bunun için Naïve Bayes, örneğin kredi derecelendirme gibi uygulamalarda nadiren kullanılmaktadır (Larsen, 2005, s. 76).

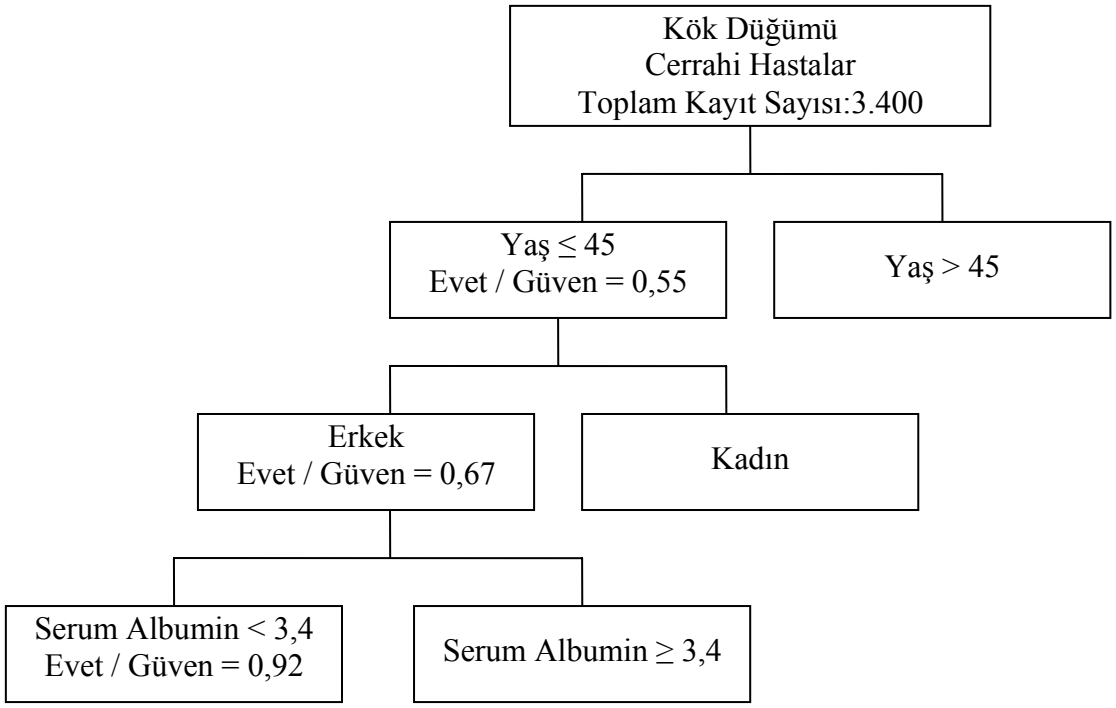
2.2. Karar Ağaçları

Karar ağacı geniş bir kayıtlar koleksiyonunu bir dizi karar kuralları uygulayarak daha küçük kayıt kümelerine ayırmaya yarayan bir yapıdır. Her ardışık bölünme ile birlikte oluşan kümelerin elemanları gittikçe birbirine daha çok benzerler (Berry ve Linoff, 2004, s.166). Karar ağaçları en yaygın veri madenciliği yöntemlerinden biridir. Karar ağaçları ile örneğin, bir bankanın her müşteri için kredi riskini belirlemesi veya Internet şubesi de olan bir perakendeci için hangi müşterilerin çevrimiçi satın alma yapabileceğini bulma gibi işlemler yapılabilir. Karar ağaçlarında temel fikir, verinin, tekrarlayan bir şekilde alt kümelere bölünmesidir ve böylece her alt küme hedef değişkenin (kestirilebilir öznitelik) az veya çok homojen durumlarını içerir. Ağaçtaki her düğümde, tüm girdi değişkenlerinin kestirilebilir öznitelik üzerindeki etkileri değerlendirilir. Bu özyinelemeli (recursive) süreç tamamlandığında karar ağacı yapılandırılmış olur (Tang ve MacLennan, 2005, s. 145-146).

Karar ağaçları ve karar kuralları birçok gerçek yaşam uygulamasında sınıflandırma problemlerine güçlü çözümler üreten veri madenciliği yöntemleridir. Veriden sınıflandırıcılar üretmek için en etkin yöntemlerden biri karar ağaçlarıdır ve karar ağacı gösterimleri yaygın olarak kullanılan mantıksal yöntemlerdir. Başlıcaları makine öğrenimi ve uygulamalı istatistik alanları olmak üzere literatürde tanımlanmış çok sayıda karar ağacı endüksiyon algoritması vardır. Bu algoritmalar bir girdi-çıktı örneklem setinden karar ağaçları yapılandıran denetimli öğrenme metotlarıdır. Tipik bir karar ağacı öğrenme sistemi, arama uzayının bir bölümünde bir çözüm arayan yukarıdan-aşağıya bir strateji kullanır. Karar ağacı, özniteliklerin test edildiği düğümlerden (node) oluşur. Bir düğümün altında kalan dallar, düğümde test edilen tüm olası sonuçlara karşılık gelir (Kantardzic, 2003, s. 139-140).

Şekil 2.1'de bir karar ağacı örneği görülmektedir. Bu karar ağacı ameliyat edilmiş hastaların hangi özelliklere göre ve hangi olasılıklarla belirli bir cerrahi komplikasyon geliştirebileceğini bulmak için oluşturulmuştur. Tanımlanan problem için bağımlı değişken ikili yapıdadır (binary) ve bu yapıda 1 komplikasyon geliştirmeyi yani Evet seçeneğini gösterirken 0 da Hayır'ı göstermektedir. Modelin en tepesinde ağacın kök düğümü bulunmaktadır ve 3400 gözlemin tamamını barındırmaktadır. Karar ağacı yapılandırıldıkça benzer davranış gösteren homojen hastalar grubunun daha rafine alt bölümlerini

(subsegments) oluşturmaktadır. Herhangi bir terminal düğümden (son düğüm) yukarı doğru bakıldığında o hastalar grubunu tanımlayan bir eğer-ise (if-then) kurallar zinciri elde edilir. Karar ağacında yukarıdan aşağı doğru inildikçe ise kurala yeni karakteristikler eklenir ve tahminin doğruluğu konusundaki güven değeri de artar. Örnekte verilen sol alt tarafta yer alan terminal düğümünün kuralı şu şekilde tanımlanabilir: 45 yaşından küçük, erkek ve serum albumin değerleri 3,4 g/dL'den daha aşağı olan hastaların belirli bir cerrahi komplikasyon geliştirme riski güven seviyesi %92'dir (Berger ve Berger, 2004).



Şekil 2.1. Cerrahi Komplikasyon Riski Taşıyan Hastalar için Karar Ağacı Örneği

Karar ağaçlarında her iç düğüm bir değer kümeleri listesi ile etiketlenir ve her değer kümesi bir asta giden yol ile ilişkilendirilir. Eğer bir nesnenin öznitelik değeri ilgili değerler kümesi aralığına düşerse, arama ağaçtan aşağı doğru uygun düşen yolu izleyerek ilerler. Her son düğüm bir sınıf etiketi ile ilişkilendirilir ve bu etiket de son düğüme düşen nesnelere atanır. Karar ağacı yapılandırması boyunca en ayırt edici özniteliklerin üst seviyelerde seçilmesi istenilen bir durumdur. Böylece bir karar ağacı, sınıfları mümkün olduğunca erken ayırabilir ve ağaç yapılandırmasının etkinliğini arttırabilir (Bajcsy, 2005, s.25).

2.2.1. Karar Ağaçlarının Yapılandırılması

Karar ağacı modellerinin temel prensibi, girdi değişkenleri tarafından genişletilen uzayı, sınıf saflık derecesini maksimize etmek üzere bölümlere ayırmaktır. Örneğin x, y ve z olmak

üzere üç girdi değişkeninin oluşturduğu bir uzayda x bölündüğünde, girdi uzayı ikiye bölünmüş olur. Daha sonra bu hücrelerin her biri ikiye bölünebilir ve x, y veya z için tanımlanmış bir eşik değere kadar bu böyle devam eder. Bu süreç bütün düğümler tanımlanıncaya kadar gerektiği kadar tekrarlanır. Girdi değişkenlerinin değeri bilinen, yeni bir gözlemin (veya kayıtın) sınıf değerini kestirmek için, yukarıdan aşağıya bir yöntem izlenir ve her düğümde, ilgili düğüme ait girdi değişkeninin eşik değeri ile yeni gözlemin değeri karşılaştırılarak uygun alt dal seçilir (Hand vd., 2001, s. 344).

Prensipte, verilen öznitelikler kümesinden karar ağaçları yapılandırmak için birçok algoritma vardır. Bazı karar ağaçları diğerlerinden daha tutarlı olabilirken, optimal karar ağacını bulmak arama uzayının büyüklüğünün aşırı olması nedeniyle bilgisayar kaynakları kullanımı açısından yapılabilir (feasible) değildir. Bununla birlikte, optimalliğin altında fakat yeterince tutarlı ve makul sürelerde karar ağaçları yapılandırmak için, etkin algoritmalar geliştirilmiştir (Tan vd., 2006, s. 151).

Karar ağacı modellerini yapılandırmak için temel strateji, girdi değişkenleri uzayının hücrelerini tekrarlayan bir süreçte (recursively) bölmektir. Herhangi bir hücreyi bölmek için (veya düğümü bölecek değişkeni ve değişkene ait eşik değerini seçmek için) tanımlanmış skor fonksiyonunda en büyük iyileştirmeyi sağlayan bölünme eşik değerini bulmak için bütün olası değişkenler ve eşik değerleri aranır. Skor eğitim veri seti (training data set) elemanları temelinde tayin edilir. Eğer amaç bir nesnenin, iki sınıftan hangisine ait olduğunu kestirmek ise, lokal skorda (iki alt düğümün ortalaması) en büyük ortalama iyileştirmeyi sağlayan değişken ve eşik değeri seçilir. Bir düğümün bölünmesi eğitim seti verisinin skor fonksiyonunda yozlaşmaya (deterioration) neden olmamalıdır. Sınıflandırma için, sınıflandırma hatasının (classification error) doğrudan kullanılması bölünme için değişken seçiminde kullanışlı bir skor fonksiyonu değildir. Entropi gibi daha dolaylı ölçüler daha kullanışlıdır. Ordinal değişkenler için ikili (binary) ayrım, değişken değerleri üzerinde tek bir eşik değerine karşılık gelir. Nominal değişkenler için ayrım ise, değişken değerlerinin iki ayrı değerler alt kümesine bölünmesine karşılık gelir (Hand, vd. 2001, s. 344).

Karar ağaçlarını yapılandırmak için geliştirilen bazı algoritmalardan en yaygın olarak kullanılanları ID3 (Quinlan, 1986) ve bu algoritmanın geliştirilmiş versiyonu olan C4.5 (Quinlan, 1993) algoritmalarıdır. Bunların yanında Breiman vd. (1984) tarafından geliştirilen CART (C&RT – Classification and Regression Trees) algoritması ve Kass (1980) tarafından geliştirilen CHAID (Chi-Squared Automatic Interaction Detector) algoritması da yine yaygın

olarak kullanılan algoritmalardır. Karar ağacı algoritmaları tanımlanabilir birçok ortak bileşen içermektedir ve bu bileşenler aşağıdaki şekilde tanımlanabilirler (Jenhani vd., 2008, s. 786-787):

1. Öznitelik seçim ölçüsü (Attribute selection measure): Bilişim teorisine (information theory) dayalı olarak, her karar düğümünde aday öznitelikler listesinden seçim yapmak için kullanılan bir kriterdir. Seçilen öznitelik, nesnelerin daha az rassal olarak dağıldığı bölümler oluşturur. Burada amaç veri ile uyumlu en küçük ve tutarlı ağaç yapısını oluşturmaktır. Bu konuda en bilinen ölçü C4.5 algoritmasında kullanılan Kazanç Oranı (Gain Ratio)'dur.

   Herhangi bir A_k özniteliği için, A_k'ya göre bilgi kazancı aşağıdaki biçimde tanımlanır:

$$Gain(T, A_k) = E(T) - E_{A_k}(T)$$

   burada

$$E(T) = -\sum_{i=1}^{n} \frac{n(C_i, T)}{|T|} \log_2 \frac{n(C_i, T)}{|T|}$$

   ve

$$E_{A_k}(T) = \sum_{v \in D(A_k)} \frac{|T_v^{A_k}|}{|T|} E(T_v^{A_k})$$

   Burada $n(C_i, T)$, T eğitim seti içerisindeki C_i sınıfına ait nesnelerin sayısını temsil etmektedir. $D(A_k)$ ise A_k özniteliğinin sonlu tanım kümesini, ve $|T_v^{A_k}|$ ise A_k özniteliği v değerini aldığında nesneler kümesinin niceliğini temsil eder. $\frac{n(C_i, T)}{|T|}$, C_i sınıfının T içerisindeki olasılık değerine karşılık gelir. Böylece, $E(T)$ de T setinin entropisine karşılık gelir. Kazanç oranı (Gain ratio) aşağıdaki şekilde ifade edilir:

$$Gr(T, A_k) = \frac{Gain(T, A_k)}{SplitInfo(T, A_k)}$$

   Burada $SplitInfo(T, A_k)$ eğitim seti T'nin n alt kümeye bölünmesiyle üretilebilen potansiyel bilgiyi temsil etmektedir ve şu şekilde ifade edilir:

$$SplitInfo(T, A_k) = -\sum_{v \in D(A_k)} \frac{|T_v^{A_k}|}{|T|} \log_2 \frac{|T_v^{A_k}|}{|T|}$$

2. Bölümleme stratejisi (Partitioning strategy): Seçilen bir özniteliğe ait bütün olası değerler için, bir bölüm üretilmesine neden olan, tüm olası öznitelik değerleri (kesikli veri için) uyarınca eğitim setinin bölümlenmesini içerir. Sürekli veri için bir kesikli hale getirme sürecine ihtiyaç vardır. Literatürde farklı kesikli hale getirme stratejileri vardır (Fayyad ve Irani, 1992, s.87-102). Kesikli hale getirme işleminden sonra, her kesikli aralığa (dicretized interval) yeni bir ordinal değer atanır (Tan vd., 2006, s. 157).

3. Durma kriteri (Stopping criteria): Bölümleme sürecinin durdurulması için gerekli kriterdir. Genellikle bölümleme işlemi tüm nesneler sadece bir sınıfa ait olduğunda durdurulur ve ilgili düğüm belirli bir sınıf değeri ile etiketlenerek son düğüm (leaf node) olarak tanımlanır. Ayrıca karar ağacının yapılandırması test edilecek öznitelik kalmadığında da durdurulur. Bu durumda baskın olan sınıf değeri son düğüm etiketi olarak tanımlanır.

2.2.2. Karar Ağaçlarının Sadeleştirilmesi (Pruning)

Bir karar ağacı yapılandırıldıktan sonra, bazı dallar eğitim veri seti içerisinde yer alan gürültü veya aykırı değerlerden kaynaklanan anomalileri yansıtır. Karar ağacı sadeleştirme yöntemleri, bu veriye aşırı uyum problemine çözüm olarak kullanılır. Bu yöntemler en az güvenilir olan dalları elemek için istatistiksel yöntemler ve ölçümler kullanır. Sadeleştirilmiş karar ağaçları daha küçük ve daha az karmaşık olmaya eğilimlidir ve bu nedenle de anlaşılmaları ve yorumlanmaları kolaydır (Han ve Kamber, 2006, s. 304).

Genellikle kullanılan sadeleştirme yöntemlerinde herhangi bir düğüm için, elde edilen alt karar ağacının yerine bir son düğüm yerleştirilir. Bu işlem kestirim hata yüzdesi, belirli bir tolerans altında ve alt karar ağacı için son düğümden yüksek olduğu durumda yapılır (Pomorski ve Perche, 2001, s. 157).

Karar ağaçlarının sadeleştirilmesinde temel fikir, modelin eğitilmesinde kullanılmamış test örnekleminin sınıflandırma doğruluğu oranına katkı sağlamayan karar ağacı bölümlerinin (alt karar ağacı) çıkarılmasıdır. Karar ağaçlarının sadeleştirilmesinde izlenen iki yol vardır (Kantardzic, 2003, s. 153):

1. Tanımlanmış şartlar altında örneklem setlerinin daha fazla bölünmemesine karar vermek. Durdurma kriteri genellikle χ^2 gibi bazı istatistiksel testlere dayalıdır. Eğer sınıflandırma doğruluğunda (classification accuracy) bölünmeden önce ve sonra anlamlı farklılıklar yoksa, mevcut düğüm son düğüm olarak atanır. Burada karar bölünme işleminden önce verilir ve dolayısıyla bu yaklaşım ön sadeleştirme (prepruning) olarak adlandırılır.

2. Belirlenmiş bir doğruluk (accuracy) kriteri kullanılarak, oluşturulmuş karar ağacı yapısının bir bölümünü, geriye dönük olarak çıkarmak. Bu süreçteki karar, karar ağacı yapılandırıldıktan sonra verildiği için, son sadeleştirme (postpruning) olarak adlandırılır.

2.2.3. Karar Ağaçlarının Etkinliğinin Değerlendirilmesi

Bütün olarak düşünüldüğünde bir karar ağacının etkinliği, karar ağacına bir test kümesinin (karar ağacını oluştururken kullanılmamış olan kayıtlar koleksiyonu) uygulanmasıyla ve doğru sınıflandırma yüzdelerinin gözlenmesiyle belirlenir. Bu işlem karar ağacının bütünü için bir sınıflandırma hata yüzdesi sağlar ve bu ayrıca karar ağacının bağımsız dallarının kalitesi hakkında da bilgi sağlaması açısından önemlidir. Karar ağacında yer alan her yol bir kuralı temsil eder ve bazı kurallar diğerlerinden daha anlamlıdır. Son düğüm veya dallanma düğümlerinin her birinde aşağıdaki unsurlar ölçülebilir (Berry ve Linoff, 2004, s. 176):

1. Düğüme giren kayıt sayısı
2. Her sınıfta yer alan kayıtların yüzdesi
3. Mevcut düğümün son düğüm olması durumunda buradaki kayıtların nasıl sınıflandırılacağı
4. Mevcut düğümde sınıflandırılan kayıtların sınıflama doğruluğu yüzdesi
5. Eğitim seti (training set) ve kontrol seti (test set) arasındaki varyans dağılımı

Karar ağacı temelli modeller nispeten basit, yorumlanabilir ve hızlı üretilebilir modellerdir. Birçok istatistiksel yaklaşımdan farklı olarak, mantıksal yaklaşım öznitelik değerlerinin dağılımı veya özniteliklerin bağımsızlığı varsayımlarına dayanmaz. Bunun yanında bu yöntem görevler arasında birçok diğer istatistiksel yönteme göre daha tutarlı olmaya eğilimlidir. Ancak bu yaklaşımın bazı dezavantajları ve kısıtlılıkları da vardır. Veri

madenciliği analisti bu konularda dikkatli olmalıdır, çünkü uygun metodolojinin seçimi bir veri madenciliği sürecinin başarısında anahtar bir adımdır (Kantardzic, 2003, s. 157).

2.3. Kümeleme

Kümeleme, küçük öznitelikler seti söz konusu olduğunda, insanların günlük hayatlarında da yaptıkları basit, doğal ve hatta otomatik bir işlemdir. Ancak öznitelik sayısı arttıkça kümeleme problemi artan bir şekilde zorlaşır ve insan zihninin baş edemeyeceği seviyeye gelir. Günümüzün veri setleri tipik olarak düzinelerce boyut içerir ve öznitelikler arasındaki muhtemel ilişkileri anlamayı ve gruplar oluşturmayı zorlaştırır (Tang ve MacLennan, 2005, s.187-188).

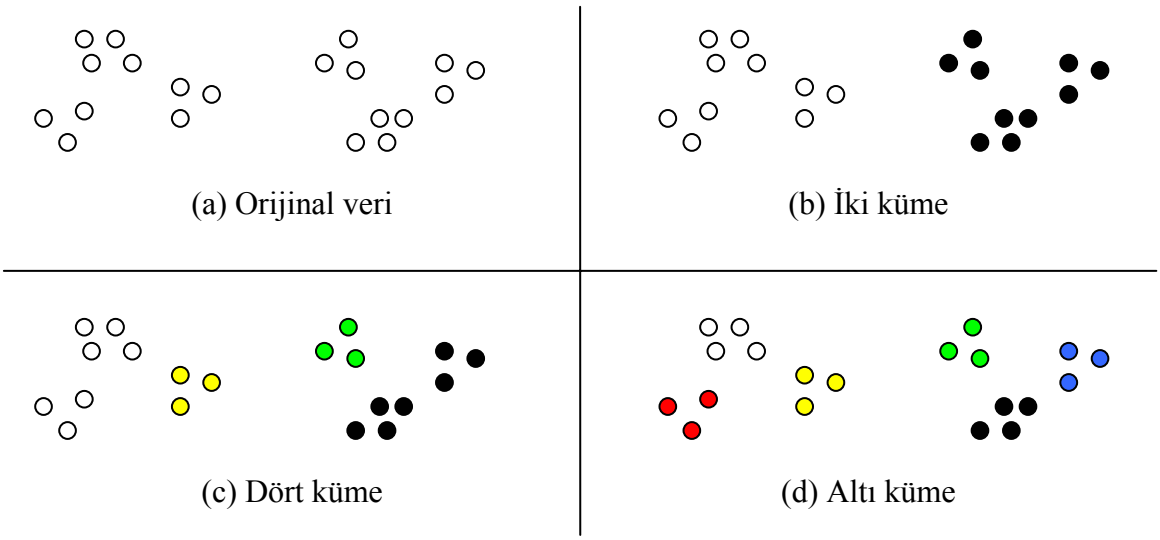
Bir nesneler setinin, benzer nesnelerden oluşan gruplara ayrılması sürecine kümeleme adı verilir. Bir küme, aynı küme içerisinde birbirlerine benzeyen, fakat diğer kümelerdeki nesnelerden farklı olan veri nesneleri koleksiyonudur. Veri nesneleri kümesi kolektif olarak tek bir grup şeklinde algılanabilir ve bu yönüyle de bir anlamda veri kısaltma biçimi olarak düşünülebilir (Han ve Kamber, 2006, s.383). Kümeleme teknikleri, ortada tahmin edilecek bir sınıf etiketi olmadığı ve bunun yerine gözlemlerin doğal gruplara ayrılacağı durumlarda uygulanır. Bu kümeler, gözlemlerin çekildiği etki alanındaki (domain) belirli mekanizmaları yansıtır. Bu mekanizmalar bazı gözlemlerin birbirleriyle, geri kalanlara göre daha güçlü bir benzerlik taşıdığını ifade eder (Witten ve Frank, 2005, s. 136).

Kümeleme analizi en yaygın ve bilinen, tanımlayıcı veri madenciliği yöntemlerinden biridir. n adet gözlem ve p adet değişkenden oluşan bir veri matrisinde, kümeleme analizinin amacı gözlemleri kendi içinde homojen (iç tutarlılık) ve aralarında heterojen (dışsal ayrışım) olan gruplara ayırmaktır (Guidici, 2003, s.75-76). n adet kayıt içeren bir veri seti için, hiyerarşik ve hiyerarşik olmayan (amaç fonksiyonu temelli) şeklinde, iki genel tip kümeleme algoritması tanımlanabilir (Shmueli vd., 2007, s. 220-222):

1. Hiyerarşik yöntemler yığmacı (agglomerative) veya ayırmacı (divisive) olabilir. Yığmacı yöntemler n adet küme ile başlarlar ve benzer kümeleri bir tek küme elde edilene kadar sırayla birleştirirler. Ayırmacı yöntemler ise, bütün gözlemleri içeren bir tek küme ile başlamak üzere, tam tersi bir şekilde çalışırlar. Hiyerarşik yöntemler özellikle amacın, kümeleri doğal hiyerarşisi içerisinde düzenlemek olduğu durumlarda kullanışlıdırlar.

2. Hiyerarşik olmayan, örneğin k-ortalamalar gibi yöntemler önceden belirlenmiş sayıda küme kullanarak gözlemleri her bir kümeye atarlar. Bu yöntemler genellikle bilgisayar kaynakları ve hesaplamalar açısından daha az yoğundur. Bu yüzden çok büyük veri setlerinde tercih edilirler.

Birçok uygulamada kümeyi ve kümeyi oluşturan yapıyı tam olarak tanımlamak zordur. Şekil 2.2'de yirmi noktadan oluşan veriyi kümelere ayırmanın üç farklı yolu görülmektedir. İnsan gözü ilk bakışta iki (Şekil 2.2.b) veya altı (Şekil 2.2.d) kümeli yapının uygun olduğunu görebilir, ancak aynı veri seti için dört kümeli (Şekil 2.2.c) bir yapının olması da muhtemeldir. Burada bir kümenin tanımının kesin olmadığı ve en iyi tanımın verinin doğasına ve amaçlanan sonuçlara bağlı olduğu söylenilebilir (Tan vd., 2006, s. 490).



Şekil 2.2. Aynı Veri Setini Kümelemenin Farklı Biçimleri

2.3.1. Benzerlik ve Uzaklık Ölçüleri

Uzaklık (veya benzemezlik) ve bunun duali olan benzerlik kavramları kümelemenin bütün formları için önemli bileşenlerdir. Bu ölçüler veri uzayında ve küme formları arasında hareket etme imkanını verirler. Uzaklığı hesaplayarak, iki örüntünün (pattern) birbirlerine ne kadar yakın oldukları algılanabilir ve ifade edilebilir. Buna dayalı olarak da bu iki örüntü aynı kümeye dahil edilebilir (Pedrycz, 2005, s. 2). Uzaklıklar birçok yol ile tanımlanabilir ancak genellikle aşağıdaki koşullar sağlanmalıdır (Shmueli vd., 2007, s. 222-223):

Negatif olmama özelliği: $d(x,y) \geq 0$

Kendine yakınlık özelliği: $d(x,x) = 0$ (Her kaydın kendine uzaklığı sıfırdır.)

Simetri özelliği: $d(x,y) = d(y,x)$

Üçgen eşitsizliği özelliği: $d(x,y) \leq d(x,z) + d(z,y)$ (Üç noktanın herhangi bir çifti arasındaki uzaklık, diğer iki uzaklığın toplamından fazla olamaz.)

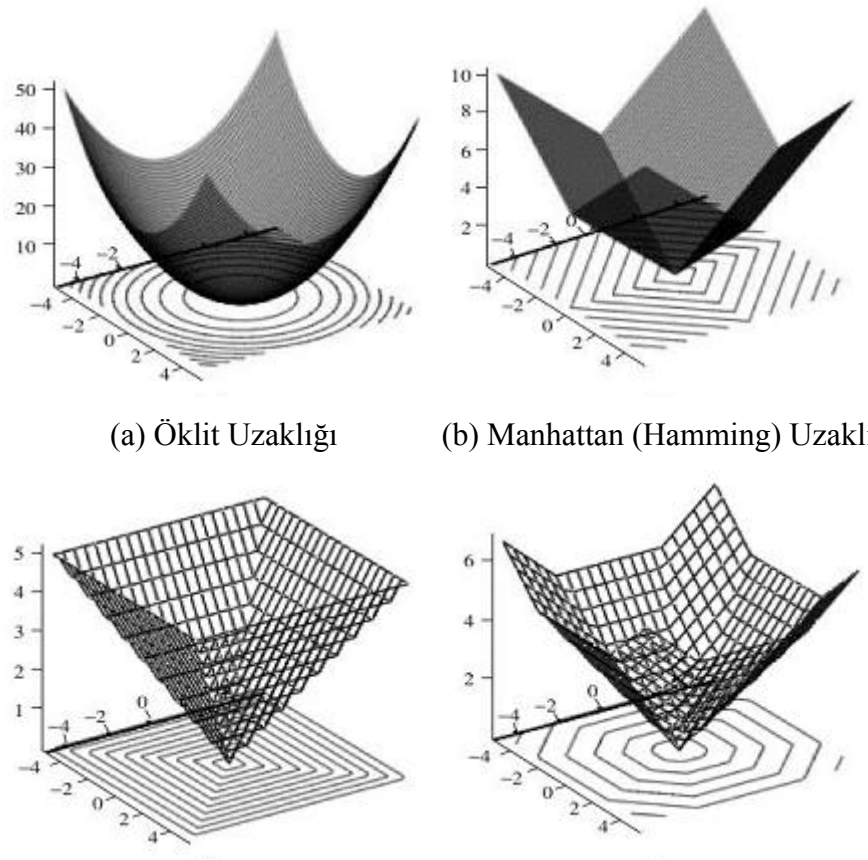
Tablo 2.1. **x** ve **y** Örüntüleri Arasındaki Uzaklık Fonksiyonları

Uzaklık Fonksiyonu	**Formülü**
Öklit Uzaklığı	$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Manhattan (Hamming veya City Block) Uzaklığı	$d(x,y) = \sum_{i=1}^{n}\lvert x_i - y_i\rvert$
Tchebyshev Uzaklığı	$d(x,y) = \max_{i=1,2,\ldots,n}\lvert x_i - y_i\rvert$
Minowski Uzaklığı	$d(x,y) = \sqrt[p]{\sum_{i=1}^{n}\lvert x_i - y_i\rvert p} \quad , \quad p > 0$
Canberra Uzaklığı	$d(x,y) = \sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert}{x_i + y_i} \quad , \quad x_i, y_i > 0$
Korelasyon Uzaklığı	$d(x,y) = 1 - r_{(x,y)}^2$
Mahalanobis Uzaklığı	$d(\vec{x},\vec{y}) = \sqrt{\lvert\vec{x}_i - \vec{y}_i\rvert^T \sum^{-1} \lvert\vec{x}_i - \vec{y}_i\rvert}$
Açısal Ayrım (Bu ölçü, x ve y doğrultusunda birim vektörler arasındaki açıyı tanımlayan bir benzerlik ölçüsüdür.)	$d(x,y) = \frac{\sum_{i=1}^{n} x_i\, y_i}{\left[\sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} y_i^{\,2}\right]^{\frac{1}{2}}}$

(Kaynaklar: Pedrycz, 2005, s.3; Shmueli vd., 2007, s.225-226)

Değişkenlerin sürekli olduğu durumda uygulanabilir birçok uzaklık fonksiyonu vardır. Bu fonksiyonlar ve formülleri Tablo 2.1'de verilmiştir. Uzaklık fonksiyonlarından her biri geometrileri nedeniyle veriye farklı bir bakış sağlar. Bu geometri sadece iki öznitelik $\mathbf{x}=[x_1, x_2]^T$ ele alındığında kolaylıkla görselleştirilebilir ve **x**'in orijinden olan uzaklığı

hesaplanabilir. Şekil 2.3'de farklı uzaklık fonksiyonları için ne tür bir geometrik yapının oluştuğunu gösteren sabit uzaklık eğrileri görülmektedir (Pedrycz, 2005, s. 4).



(a) Öklit Uzaklığı (b) Manhattan (Hamming) Uzaklığı

(c) Tchebyschev Uzaklığı (d) Manhattan ve Tchebyshev Karma

Şekil 2.3. Üç Boyutlu Gösterim ve Eş Uzaklık Eğrileri ile Uzaklık Fonksiyonu Örnekleri

Öklit uzaklığı en yaygın kullanılan uzaklık ölçüsüdür. Ancak öklit uzaklığının üç temel özelliği kullanılırken dikkate alınmalıdır. Birincisi, yüksek oranda ölçeğe bağımlı olmasıdır. Bir değişkenin biriminin değiştirilmesinin sonuçlar üzerinde büyük etkisi olabilir. Standardizasyon bu sorun için bir çözüm olabilir. Öklit uzaklığının dikkat edilmesi gereken ikinci özelliği de ölçümler arasındaki ilişkileri dikkate almamasıdır. Bu yüzden eğer ölçümler güçlü bir şekilde ilişkili (ör. yüksek korelasyon) ise Mahalanobis uzaklığı gibi başka uzaklık fonksiyonları kullanılmalıdır. Üçüncü olarak ise eğer veri aykırı değerler (outliers) içeriyor ve bunların giderilmesi mümkün değilse, Manhattan uzaklığı gibi daha sağlam (robust) uzaklık fonksiyonları tercih edilebilir (Shmueli vd., 2007, s. 223-225).

2.3.2. Bağıntı Yöntemleri

Bir küme, bir veya daha fazla kayıttan oluşan bir kayıtlar setidir. Kümeler arasındaki uzaklıkları ölçmek için temel fikir ise gözlemler arasındaki uzaklık ölçülerinin genişletilerek

kümeler arasındaki uzaklıkları tanımlamaktır. m adet gözlem içeren bir A kümesi ($A_1, A_2, \ldots, A_m$) ve n adet gözlem içeren bir B kümesi ($B_1, B_2, \ldots, B_n$) arasındaki uzaklığı ölçmek için en yaygın kullanılan kümeler arası uzaklık ölçüleri şunlardır (Shmueli, vd., 2007, s. 227):

Minimum Uzaklık (Tekil Bağıntı – Single Linkage): A_i ve B_j kayıt çiftleri arasındaki uzaklığın birbirine en yakın olduğu ölçüdür. Diğer bir deyişle iki kümenin birbirine en yakın iki gözlemi arasındaki uzaklıktır.

$$\min\left(dist\left(A_i, B_j\right)\right), \quad i = 1,2,\ldots,m\,; \quad j = 1,2,\ldots,n$$

Maksimum Uzaklık (Tam Bağıntı – Complete Linkage): A_i ve B_j kayıt çiftleri arasındaki uzaklığın birbirine en uzak olduğu ölçüdür. Diğer bir deyişle iki kümenin birbirine en uzak iki gözlemi arasındaki uzaklıktır.

$$\max\left(dist\left(A_i, B_j\right)\right), \quad i = 1,2,\ldots,m\,; \quad j = 1,2,\ldots,n$$

Ortalama Uzaklık (Ortalama Bağıntı – Average Linkage): Bir kümedeki gözlemler ile diğer kümedeki gözlemler arasındaki tüm olası uzaklıkların ortalamasıdır.

$$average\left(dist\left(A_i, B_j\right)\right) = \frac{1}{m\,n}\sum_{i=1}^{m}\sum_{j=1}^{n} d_{i,j}$$

Merkezi Uzaklık (Centroid Distance): İki küme merkezi arasındaki uzaklıktır. Bir küme merkezi, o kümede yer alan tüm gözlemlerin ölçüm ortalamaları vektörüdür. p adet kayıt içeren A kümesi ile k adet kayıt içeren B kümesi arasındaki uzaklık aşağıdaki şekilde ifade edilebilir.

$$\vec{x}_A = \left[\left(\frac{1}{m}\sum_{i=1}^{m} x_{1i}, \ldots, \frac{1}{m}\sum_{i=1}^{m} x_{pi}\right)\right]$$

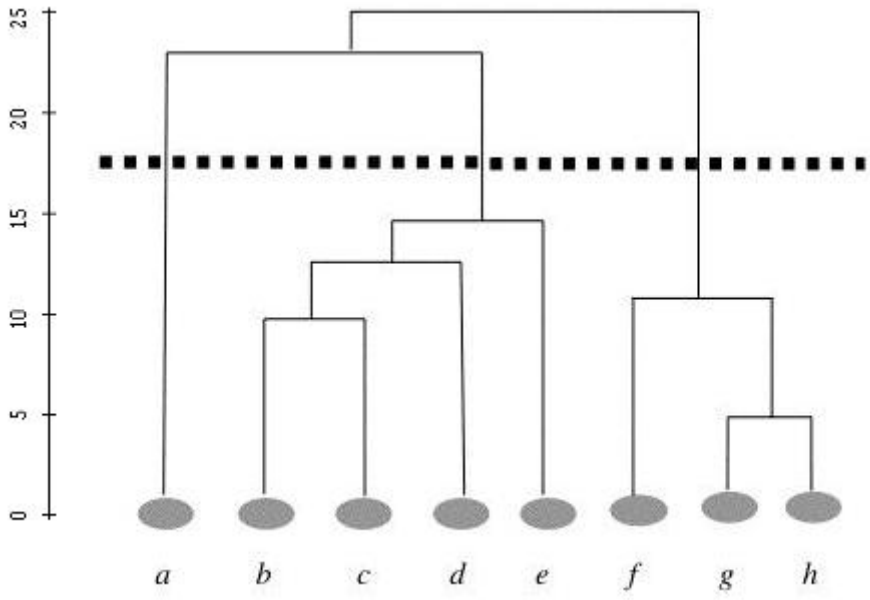
$$\vec{x}_B = \left[\left(\frac{1}{n}\sum_{j=1}^{n} x_{1j}, \ldots, \frac{1}{n}\sum_{j=1}^{n} x_{kj}\right)\right]$$

$$dist\left(A, B\right) = \left|\vec{x}_A - \vec{x}_B\right|$$

2.3.3. Hiyerarşik Kümeleme

Hiyerarşik kümeleme algoritması iki bileşene dayalıdır: nesneler arasındaki benzerlik veya uzaklık ölçüsü ve nesne sınıfları arasında bir birleştirme (amalgamation) veya bağıntı (linkage) kuralı (Kojadinovic, 2004, s. 270). Hiyerarşik kümeleme analizi teknikleri dendogram adı verilen bir yapı ile verinin grafiksel olarak sunumunu üretir (Duda vd., 2001). Grafiklerin oluşturulmasında yukarıdan-aşağıya ve aşağıdan-yukarıya olmak üzere iki yol izlenir. Aşağıdan-yukarıya izlenen yol yığmacı (agglomerative) yaklaşım olarak adlandırılır. Bu yaklaşımda başlangıçta her bir eleman ayrı bir kümedir ve algoritmanın her ilerlemesinde, kullanılan bağıntı yöntemine bağlı olarak birbirlerine en yakın olan örüntüler aynı kümede birleştirilmektedir. Bu süreç tek küme elde edilene kadar veya önceden belirlenmiş bir eşik değere (treshold value) ulaşılana kadar ilerlemektedir.Yukarıdan-aşağıya izlenen yol ayırmacı (divisive) yaklaşım olarak adlandırılır ve bütün set tek bir küme olarak algılanarak küçük kümelere ayırma süreci izlenir (Pedrycz vd., 2005, s. 6).

Hiyerarşik kümeleme sonuçları genellikle dendogramlar ile gösterilir. Şekil 2.4'de örnek bir dendogram sunumu gösterilmiştir. Şeklin sol tarafında bulunan uzaklık ölçeği, kümeler arasındaki uzaklıkların ölçülmesi ve değerlendirilmesine yardımcı olur. Uzaklık ölçeği basit bir durdurma kriterini temsil eder: belirlenmiş bir uzaklık eşik değerinde kümelerin birleştirilmesi durdurulur. Kümeler arasındaki uzaklığın eşik değerini aşması, bu iki ayrı yapının birleştirilmesinin sonuca anlamlı bir katkısı olmadığı anlamına gelir (Pedrycz, 2005, s. 7).



Şekil 2.4. Örüntü Yapısının Görselleştirilmesi için Dendogram Örneği

Şekil 2.4'de gösterilen dendogramın analizinde kesikli çizgi, kümelerin belirlenmesi için kümeler arası uzaklık eşik değerini temsil etmektedir. Burada $\{a\}$, $\{b,c,d,e\}$ ve $\{f,g,h\}$ şeklinde üç kümenin oluştuğu görülmektedir.

2.3.3.1. Yığmacı (Agglomerative) Yöntemler

Yığmacı yöntemler, kümeler arasındaki uzaklıklara dayalıdır. Temelde, birbirine en yakın iki kümeyi birleştirerek daha az sayıda küme elde etmeyi hedefler. Bu işlem her defasında birbirine en yakın iki kümeyi birleştirerek, bütün veri tek bir kümede toplanana kadar devam eder. Sürecin başlangıç noktası her bir kümenin sadece bir veri noktasını içerdiği başlangıç kümelemesidir (initial clustering). Bundan sonra prosedür, kümelenmesi gereken n adet nokta ile başlar (Hand vd., 2001, s. 311).

Birçok yığmacı hiyerarşik kümeleme algoritması tekil-bağıntı veya tam-bağıntı algoritmalarının değişik biçimleridir. Bu iki temel algoritma küme çiftleri arasındaki benzerlikleri karakterize etme yollarında (uzaklık hesaplama yöntemleri) farklılaşırlar (Kantardzic, 2003, s. 125-126). Hiyerarşik kümeleme algoritmalarında her iterasyonda sadece tek küme çifti birleştirildiği için bu algoritmalar büyük hiyerarşiler oluşturur. Bu nedenle, bu algoritmalar hiyerarşinin her seviyesinde, yeni küme ve diğer kümeler arasındaki uzaklıkları veya benzerlikleri yeniden hesaplamak için çok zaman ve kaynak harcarlar (Gil-García vd., 2006, s. 1).

2.3.3.2. Ayırmacı (Divisive) Yöntemler

Ayırmacı yöntemler bütün veri noktalarından oluşan tek bir küme ile süreci başlatır ve bu kümeyi alt bileşenleri ayırma işlemlerini yapar. En sonunda, süreç her kümenin tek bir veri noktası içerdiği seviyede sona erer. Tekli (monothetic) ayırmacı yöntemler her seferinde tek değişken kullanarak kümeleri ayırır. Bu yöntemler, sonuçları dendogram tarafından kolayca gösterilebildiği ve her düğümdeki ayrımın ilgili değişken ile tanımlanabildiği için kullanışlı olabilir. Çoklu (polythetic) ayırmacı yöntemler ayrımları değişkenlerin tümünü temel alarak yaparlar. Herhangi bir kümeler arası uzaklık ölçüsü kullanılabilir. Genel olarak ise ayırmacı yöntemler bilgisayar kaynakları kullanımı açısından yoğun ve hesaplama zamanları uzun olduğu için yığmacı yöntemlere göre daha az yaygındır (Hand vd., 2001, s. 314-315).

2.3.4. Amaç Fonksiyonu Temelli Kümeleme

Kümelemenin ikinci genel kategorisi olan amaç fonksiyonu temelli kümeleme yöntemleri, veri setlerinden kümeler oluşturmayı, bazı performans endekslerine veya başka bir deyişle amaç fonksiyonlarına dayalı olarak gerçekleştirirler. Esas itibariyle, N adet örüntüyü c adet kümeye bölme işlemi çözülmesi zor olan bir problemdir (Pedrycz, 2005, s. 8-9). Amaç fonksiyonu temelli kümeleme, bazı kaynaklarda bölümsel (partitional) kümeleme (Kantardzic, 2003, s. 129; Tan vd., 2006, s. 491) veya hiyerarşik olmayan kümeleme (Guidici, 2003, s. 83; Shmueli vd., 2007, s. 233) olarak da tanımlanır.

Amaç fonksiyonu temelli yöntemler büyük veri setlerinden oluşan uygulamalarda üstünlük sağlarlar. Bu yöntemler tanımlanmış amaç fonksiyonunu optimize ederek kümeleri oluştururlar. Amaç fonksiyonu lokal (örneklem alt kümesinde tanımlı) veya global (bütün örneklem üzerinde tanımlı) olarak tanımlanmış olabilir. Bir global kriter olan öklidyen hata-kareleri ölçüsü, her kümeyi bir prototip veya küme merkezi (centroid) ile temsil eder ve gözlemleri en yakın prototiplere göre kümelere atar. Bir lokal kriter olan en küçük ortak komşu uzaklığı (minimal mutual neighbor distance) ise veri içerisindeki lokal yapıyı kullanarak kümeleri biçimlendirir. Dolayısıyla, veri uzayındaki yüksek yoğunluklu bölgelerin tanımlanması kümelerin oluşturulmasında temel kriterdir (Kantardzic, 2003, s. 129).

2.3.4.1. K-Ortalamalar (K-Means) Kümeleme

K-ortalamalar, prototip temelli kümeleme yöntemlerinden biridir. Prototip temelli yöntemler veri nesnelerinin tek-seviye bölümlemesini üretirler ve bunlardan en önde gelen ikisi K-ortalamalar (K-means) ve K-ortancalar (K-medoids) teknikleridir. K-ortalamalar prototipi küme merkezi (centroid) üzerinden tanımlar. Küme merkezi, genellikle veri noktalarının ortalamasıdır ve tipik olarak n-boyutlu sürekli değişkenler uzayındaki nesnelere uygulanır. K-ortancalar ise prototipi ortanca (medyan) üzerinden tanımlar. Ortanca, bir veri noktaları grubu için en temsil edici veri noktasıdır. Centroid hemen hiçbir zaman bir gerçek veri noktasına karşılık gelmezken, ortanca, tanımı gereği, mutlaka bir gerçek veri noktası olmalıdır (Tan vd., 2006, s.497).

K-ortalamalar algoritması gürültülü veriye ve aykırı değerlere aşırı duyarlıdır. Bunun nedeni, bu türdeki az sayıda verinin bile ortalama değeri (mean) büyük ölçüde etkileyebilir olmasıdır. K-ortalamalar yönteminin aksine K-ortancalar yöntemi küme temsilcisi olarak

gözlemlerin ortalama değerini almak yerine, bir küme içerisinde en merkezi olarak yerleşmiş nesneyi (medyan) kullandığı için, gürültülü veriye ve aykırı değerlere daha az duyarlıdır (Kantardzic, 2003, s. 131-132).

K-ortalamalar yönteminde k oluşturulacak grup sayısını belirtir ve öncül olarak tanımlanır. K-ortalamalar algoritması n adet başlangıç elemanını k adet gruba kümeleme işlemini gerçekleştirir. Bunu aşağıdaki işlem akışına göre yapar (Guidici, 2003, s. 84):

1. Başlangıç Durumuna Getirme (Initialization): Grup sayısının ve başlangıç küme merkezlerinin (seed) belirlenmesi. Başlangıç bölümlemesinde küme merkezlerini (centroid), başlangıç küme merkezleri (seed) teşkil ederler. Başlangıç küme merkezleri arasında, algoritmanın yakınsamasını geliştirmek için yeterli uzaklık bulunmalıdır. Bunun için örneğin SAS yazılımı Fastclust olarak adlandırılan bir prosedür kullanır ve başlangıç küme merkezlerini etkin belirlemek için verinin ön analizini gerçekleştirir. Başlangıç küme merkezleri belirlendikten sonra gözlemlerin başlangıç bölümlemesi gerçekleştirilir ve her gözlem kendisine en yakın küme merkezine atanır.

2. Transfer Değerlendirmesi (Transfer Evaluation): Her gözlemin, k adet kümenin merkezine (centroid) olan uzaklığı hesaplanır. Her gözlem ile, kendisinin atandığı kümenin merkezinin uzaklığı minimum olmalıdır. Eğer bu uzaklık minimum değilse gözlem kendisine daha yakın olan küme merkezinin sahibi olan kümeye atanır. Eski küme ve yeni kümenin küme merkezleri yeniden hesaplanır.

3. Tekrar (Repetition): Kümelerin uygun bir şekilde durağanlığa erişmesine kadar (örneğin hata kareleri toplamının minimize edilmesine kadar) 2. adım tekrarlanır.

K-ortalamalar algoritmasında kullanılan matematiksel formüller tek boyutlu veri ve amaç fonksiyonunun hata kareleri toplamının minimizasyonu olduğu durum için aşağıda verilmiştir (Tan vd., 2006, s. 513-514):

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} \left(c_i - x\right)^2$$

Burada SSE hata kareleri toplamı (sum of the squared error), C_i, i numaralı kümedir. x, C_i kümesi içerisinde yer alan bir nokta, c_i ise i numaralı kümenin ortalamasıdır (centroid). k numaralı centroid c_k için SSE'nin türevini alıp sıfıra eşitleyerek yukarıdaki eşitliği minimize eden çözüm aşağıdaki gibidir.

$$\begin{aligned}\frac{\partial}{\partial c_k} SSE &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \\ &= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} (c_i - x)^2 \\ &= \sum_{x \in C_k} 2(c_k - x_k) = 0\end{aligned}$$

$$\sum_{x \in C_k} 2(c_k - x_k) = 0 \quad \Rightarrow \quad m_k c_k = \sum_{x \in C_k} x_k \quad \Rightarrow \quad c_k = \frac{1}{m_k} \sum_{x \in C_k} x_k$$

2.3.4.2. Bulanık C-Ortalamalar (Fuzzy C-Means) Kümeleme

Katı kümeleme yaklaşımında veri, ayrık kümelere bölünürler ve her veri noktası kesin olarak bir kümeye dahil olur. Bulanık kümeleme (fuzzy clustering) yaklaşımında ise veri noktaları birden fazla kümeye ait olabilir ve her veri noktası bir üyelik seviyeleri (membership level) seti ile ilişkilendirilir. Bu bulanık setler (fuzzy sets), ilgili veri noktası ve belirli bir küme arasındaki bağlantının gücünü temsil eder. Bulanık kümeleme, üyelik seviyelerinin belirlenmesi ve bunların veri noktalarının bir veya daha fazla kümeye atanması için kullanılması sürecidir.

Bulanık C-Ortalamalar (Fuzzy C-Means – FCM) algoritması kısmi üyeliğe izin veren bir algoritmadır ve bunun ilk olarak ortaya atılması Dunn (1973) tarafından olmakla birlikte genelleştirilmesi Bezdek (1981) tarafından yapılmıştır. FCM algoritmasının prensipleri aşağıdaki biçimde ifade edilebilir (Pedrycz, 2005, s. 12-15):

Veri uzayında aramayı yönlendiren performans endeksi veya amaç fonksiyonu,

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \left\| x_k - v_i \right\|^2$$

biçiminde ifade edilir, burada u bulanık bölümleme matrisidir, üyelik derecelerinin toplamı 1'e eşittir. Dolayısıyla aitlik dağılımları 1'e eşittir. Yukarıdaki amaç fonksiyonu için uzaklığı quadratik formda şu şekilde yazabiliriz,

$$\|x_k - v_i\|^2 = (x_k - v_i)^T (x_k - v_i) = x_k{}^T x_k - 2x_k{}^T v_i + v_i{}^T v_i$$

Minimizasyon, bölümleme matrisi ve prototiplere göre tamamlanır. Bulanıklaştırma faktörü (fuzzification factor, m), $m > 1$ olmak üzere, küme biçimlerini kontrol etmeye yardımcı olur ve üyelik dereceleri arasında bir denge oluşturur. Çözümün elde edilmesi iki adımda tamamlanır. Birincisi bölümleme matrisinde yerine konulan kısıtları içerir ve bu kısıtlar Lagrange çarpanları yardımıyla dahil edilir. Burada her örüntü için $t = 1, 2, \ldots, N$, genişletilmiş fonksiyon aşağıdaki şekilde formüle edilir,

$$V = \sum_{i=1}^{c} u_{it}^{m} d_{it}^{2} - \lambda \left(\sum_{i=1}^{c} u_{it} - 1 \right)$$

Burada λ Lagrange çarpanını gösterir. V'nin u_{st}'ye göre türevini alarak ve bunu sıfıra eşitleyerek aşağıdaki eşitliği elde ederiz,

$$\frac{\partial V}{\partial u_{st}} = m u_{st}^{m-1} d_{st}^{2} - \lambda = 0$$

ve buradan,

$$u_{st} = \left(\frac{\lambda}{m} \right)^{1/m-1} \frac{1}{\left(d_{st} \right)^{\frac{2}{m-1}}}$$

Aitlik dağılımlarının 1'e eşitliği kısıtını, $\sum_{j=1}^{c} u_{jt} = 1$, göz önüne alırsak,

$$\left(\frac{\lambda}{m} \right)^{1/m-1} \sum_{j=1}^{c} \frac{1}{\left(d_{jt} \right)^{\frac{2}{m-1}}} = 1$$

Bu, Lagrange çarpanının belirlenmesini sağlar (λ):

$$\left(\frac{\lambda}{m}\right)^{1/m-1} = \frac{1}{\sum_{j=1}^{c} \frac{1}{\left(d_{jt}\right)^{\frac{2}{m-1}}}}$$

İkinci adımda ise son elde ettiğimiz ifadeyi bulanık bölümleme matrisi (u_{st}) eşitliğinde yerine koyarsak,

$$u_{st} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{st}}{d_{jt}}\right)^{\frac{2}{m-1}}}$$

elde edilir. Prototiplerin hesaplanması ileriye doğrudur. Minimum Q'nun v_s'ye göre hesaplanmasında $\nabla_{v_s} Q = 0$ elde edilir. Detaylı çözüm uzaklık fonksiyonuna dayalıdır. Öklit uzaklığı kullanıldığı durumda,

$$2\sum_{k=1}^{N} u_{sk}^{m}\left(x_k - v_s\right) = 0$$

Buradan aşağıdaki çözümü elde ederiz,

$$v_s = \frac{\sum_{k=1}^{N} u_{sk}^{m} x_k}{\sum_{k=1}^{N} u_{sk}^{m}}$$

Uzaklık fonksiyonlarının Manhattan veya Tchebyshev gibi diğer formlarında, uzaklıklar hızlı çözüme götürmezler ve burada daha fazla bir optimizasyon çabası gerektirirler.

Kısaca FCM algoritması, prototiplerin ve bölümleme matrisinin ardışık hesaplamalarını içeren yinelemeli (iterative) bir süreçtir. Parametrelerin değerleri önceden belirlenir. Parametreler şu bileşenleri içerir: küme sayısı (c), uzaklık fonksiyonu ($\|x_k, v_i\|$),

bulanıklaştırma faktörü (fuzzification factor) (m) ve sonlandırma kriteri (ε) (Pedrycz vd., 2005, s. 15).

2.4. Birliktelik Kuralları (Association Rules)

Veritabanı işlemlerindeki detaylı bilgi elverişliliği, veritabanında tutulan öğeler arasındaki bağlantıları (associations) otomatik yöntemlerle araştıran tekniklerin geliştirilmesine neden olmuştur. Birliktelik kuralları algoritmaları en çok kullanılan ve en yaygın veri madenciliği uygulamalarından biridir. Genellikle en çok pazar sepeti analizi (market basket analysis) kavramı ile ilişkilendirilse de, Agrawal vd. (1993) tarafından ortaya atıldığından bu yana birliktelik kuralları akademisyenler ve uygulamacılar tarafından müşteri ilişkileri yönetimi, üretim süreçleri, finans, meteoroloji, veri güvenliği, tanılama (diagnosis) ve tıp bilişimi gibi alanlarda da kullanılmaktadır (Feng vd., 2001; He vd., 2004; Chen vd., 2005; Ezziane, 2006; Chen vd., 2007; Yi ve Zhang, 2007; Kuo vd., 2007; Kianmehr ve Alhajj, 2008).

Birliktelik kurallarının ilgi çekici yanlarından biri sonuçlarının açıklığı ve uygulanabilirliğidir. Bu sonuçlar, öğeler veya öğe grupları arasındaki kurallar biçimindedir (Berry ve Linoff, 2004, s. 296). Bir birliktelik kuralı, veritabanı içerisinde belirli nesnelerin birlikte görülme olasılığı ifadesidir (Hand vd., 2001, s. 158).

Birliktelik kuralları yöntemlerinin uygulanmasında dikkat edilmesi gereken iki konu vardır. Birincisi, büyük bir işlemsel veri setinden örüntülerin çıkartılmasının bilgisayar kaynakları ve hesaplamalar açısından oldukça maliyetli ve zaman alıcı olmasıdır. İkincisi ise bulunan bazı örüntülerin şans eseri ortaya çıkma ve sahte olma ihtimalidir. Bunun için keşfedilen örüntüler ayrıca değerlendirme süreçlerinden geçirilmelidir (Tan vd., 2006, s. 328).

Denetimsiz öğrenme algoritmaları, bir çıktı değişkeni veya kestirilecek veya sınıflandırılacak bir değişken olmadığı durumlarda kullanılan algoritmalardır (Shmueli vd., 2007, s. 11). Birliktelik kuralları, veri madenciliği teknikleri arasında denetimsiz öğrenme sistemlerinde yerel örüntü keşfinin (local pattern discovery) en yaygın biçimidir (Kantardzic, 2003, s. 165).

Birliktelik kuralları tekniklerinin çıktısı olarak çok sayıda kural üretildiği için, burada amaç, öncül ve ardıl öğe setleri arasındaki güçlü bağlılıkları belirten kuralların tespit edilmesi olacaktır. Bir kural tarafından ifade edilen bağlılığın gücünü ölçmek için destek (support),

güven (confidence) ve kaldırma oranı (lift ratio) gibi ölçüler kullanılır. Buradaki güven ölçüsü istatistikte yaygın olarak kullanılan güven aralığı ve güven düzeyi kavramları ile ilgisiz ve bunlardan bağımsızdır. Aynı şekilde, kaldırma oranı kavramı da finans alanındaki kaldıraç oranı kavramından farklı ve bağımsızdır. Birliktelik kurallarında kullanılan bu üç ölçüye ait açıklamalar aşağıda verilmiştir.

Burada $I = \{i_1, i_2, \ldots i_n\}$ öğe seti (itemset), $D = \{t_1, t_2, \ldots t_m\}$ işlemleri (transactions) içeren bir veritabanı, her işlem öğelerin alt kümesini içermek, $t \subseteq I$ ve her işleme ait bir benzersiz (unique) tanımlayıcı TID olmak üzere, t işlemi A öğe setini eğer ve sadece $A \subseteq t$ durumunda içermektedir. Bir birliktelik kuralı ise, $A \subset I$, $B \subset I$ ve $A \bigcap B = \Phi$ olduğu durumda $A \Rightarrow B$ formunun bir dolaylı anlatımıdır. Bu kuralın destek, güven ve kaldırma oranı aşağıdaki biçimlerde ölçülür (Han ve Kamber, 2006, s.230; Shmueli vd., 2007, s. 206-207):

$$support(A \Rightarrow B) = P(A \cup B)$$

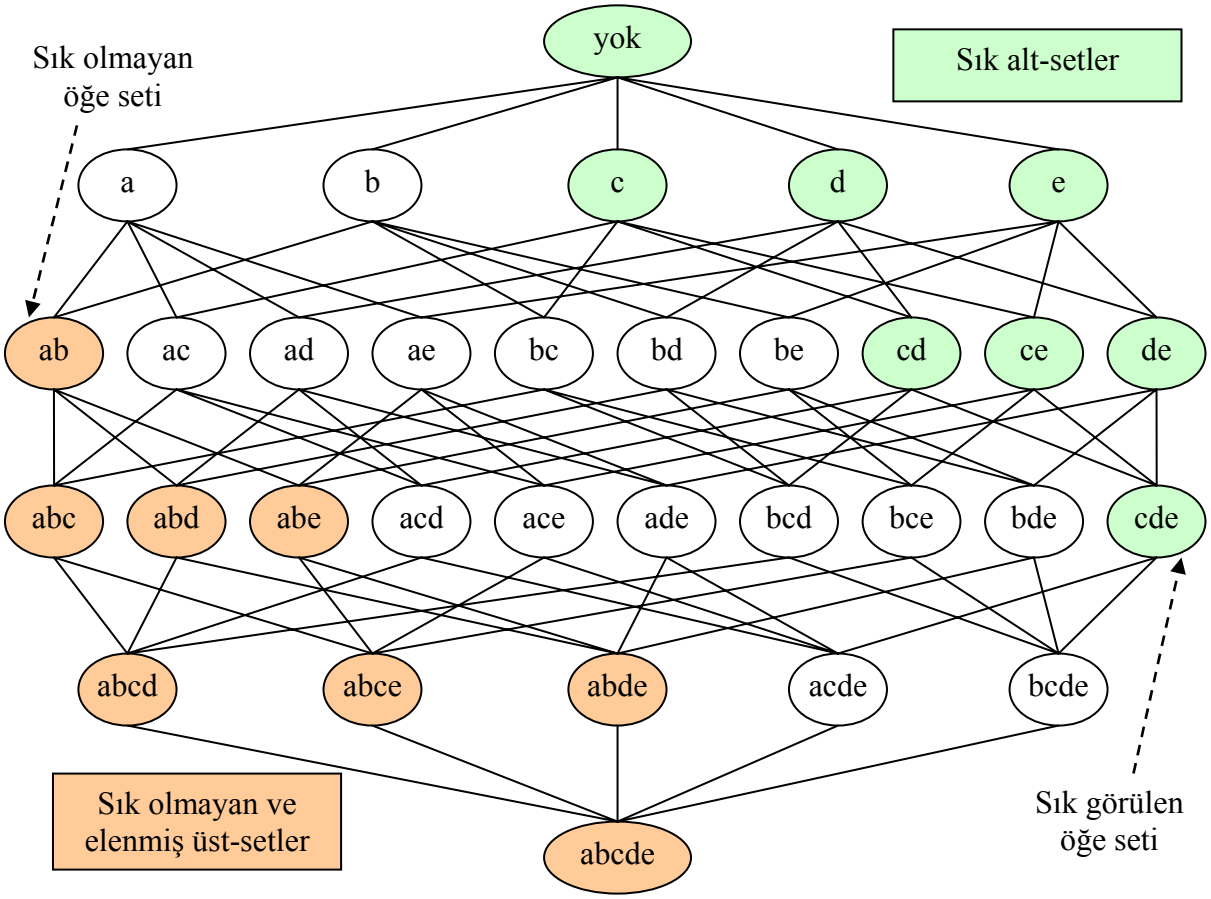
$$confidence(A \Rightarrow B) = P(B \mid A) = \frac{P(A \cup B)}{P(A)}$$

$$liftratio(A \Rightarrow B) = \frac{confidence(A \Rightarrow B)}{expected\ confidence(A \Rightarrow B)} = \frac{P(B \mid A)}{P(B)}$$

Burada A öncül (antecedent), B ise ardıl (consequent) öğe setidir. Destek (support) ölçüsü, işlemlerin (t) öncül ve ardıl öğe setlerini birlikte içerme olasılığıdır. Bir başka deyişle destek, öncül ve ardıl öğe setlerini birlikte içeren işlem sayısının veritabanındaki toplam işlem sayısına oranıdır. Güven (confidence) ölçüsü, ardıl öğe setinin öncül öğe setine koşullu olasılığıdır. Bir başka deyişle, işlemlerin öncül ve ardıl öğe setlerini birlikte içerme olasılığının, öncül öğe setini içerme olasılığına oranıdır. Kaldırma oranı (lift ratio) ölçüsü ise, kuralın güven değerinin beklenen güven değerine oranıdır. Beklenen güven değeri (expected confidence) ardıl öğe setini içeren işlem sayısının veritabanındaki toplam işlem sayısına oranıdır.

2.4.1. Apriori Algoritması

Apriori, Agrawal ve Srikant (1994) tarafından önerilen ve Boolean birliktelik kuralları için sık öğe setleri madenciliğinde kullanılan bir algoritmadır. Algoritmanın ismi, sık öğe setlerinin özelliklerine ait öncül bilgi kullanması prensibine dayanmaktadır (Han ve Kamber, 2006, s. 234-235). Apriori prensibinde destek (support) ölçüsü, sık öğe seti üretimi sırasında ortaya çıkan aday öğe seti sayısının azaltılması için kullanılmaktadır. Bu prensibe göre eğer bir öğe seti sık görülüyorsa, bu öğe setinin tüm alt setleri de sık görülür (Tan vd., 2006, s.333-334). Şekil 2.5'de Apriori prensibini açıklamak için örnek öğe setleri verilmiştir.



Şekil 2.5. Apriori Algoritmasında Sık Görülen ve Sık Olmayan Öğe Setleri

Şekil 2.5'de görülen $\{c,d,e\}$ öğe setinin sık görülen bir öğe seti olduğu düşünülürse, bu öğe setini içeren tüm işlemlerin, bunun alt-setlerini de $\{c,d\}$, $\{c,e\}$, $\{d,e\}$, $\{c\}$, $\{d\}$ ve $\{e\}$ içereceği açıktır. Dolayısıyla, eğer $\{c,d,e\}$ sık görülen bir öğe seti ise, bütün alt-setleri de sık görülen öğe setleridir. Bunun tersine eğer $\{a,b\}$ öğe seti sık olmayan bir öğe seti ise bunun tüm üst-setleri de $\{a,b,c\}$, $\{a,b,d\}$, $\{a,b,e\}$, $\{a,b,c,d\}$, $\{a,b,c,e\}$, $\{a,b,d,e\}$ ve $\{a,b,c,d,e\}$

sık olmayan öğe setleridir. Bu sayede, Apriori stratejisinde $\{a,b\}$ öğe setinin sık olmayan (infrequent) olarak bulunması durumunda bunun tüm üst-setleri de hemen elenir. Üstel arama uzayının bu şekilde destek ölçüsüne dayalı olarak azaltılması stratejisi destek temelli sadeleştirme (support based pruning) olarak adlandırılır (Tan vd., 2006, s. 333-335).

Apriori algoritmasına ve alt prosedürlerine ait algoritma aşağıda verilmiştir (Han ve Kamber, 2006, s. 239).

Algorithm Apriori:
Girdiler:
D, İşlem veritabanı
$\min\_\sup$, minimum destek eşik değeri
Çıktılar:
L, D içerisinde yer alan sık öğe setleri
Yöntem:

(1) $L_1 = find\_frequent\_1-itemsets(D)$;
(2) **for** $(k = 2; L_{k-1} \neq \Phi; k++)$ {
(3) $C_k =$ **apriori_gen**(L_{k-1});
(4) **for each** transaction $t \in D$ { // adaylar için D'nin taranması
(5) $C_t = subset(C_k, t)$; // t'nin aday olan alt setlerinin elde edilmesi
(6) **for each** candidate $c \in C_t$
(7) $c.count++$;
(8) }
(9) $L_k = \{c \in C_k \mid c.count \geq \min\_\sup\}$
(10) }
(11) **return** $L = \bigcup_k L_k$;

procedure apriori_gen$(L_{k-1} : frequent(k-1)-itemsets)$
(1) **for each** itemset $l_1 \in L_{k-1}$
(2) **for each** itemset $l_2 \in L_{k-1}$
(3) **if** $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[k-1] = l_2[k-1])$ **then** {
(4) $c = l_1 \otimes l_2$; // birleştirme adımı: adayların üretilmesi
(5) **if has_infrequent_subset**(c, L_{k-1}) **then**
(6) **delete** c; // sadeleştirme adımı: başarısız adayların çıkarılması
(7) **else add** c **to** C_k;
(8) }
(9) **return** C_k;

procedure
has_infrequent_subset$(c : candidate\ k-itemset; L_{k-1} : frequent(k-1)-itemsets)$;
(1) **for each** $(k-1)-subset\ \ s$ **of** c
(2) **if** $s \notin L_{k-1}$ **then**
(3) **return** TRUE;
(4) **return** FALSE;

Apriori algoritması örüntüleri analiz etmek için kullanılmaz, bunun yerine aday öğe setlerini (itemsets) üretir ve bunların sayısını bulur. Bir öğe, analiz edilen verinin türüne göre bir olayı, bir ürünü veya bir özniteliğin değerini temsil edebilir (MSDN, 2008).

2.5. Yapay Sinir Ağları

Yapay sinir ağları (Artificial Neural Networks – ANN) insan beyninin bilişsel öğrenme sürecinin benzetimini yapmak amacıyla ilk olarak 1940'ların başında ortaya atılmıştır. Yapay sinir ağları, birincil fonksiyonları deneme yanılmaya dayalı olarak problem uzayının bir modelini yapılandırmak olan bilgisayar-tabanlı yöntemlerdir. Modelin yapılandırılması süreci kavramsal olarak bir kısım veri setinin yapay sinir ağına gönderilmesiyle başlar ve yapay sinir ağı bir çıktı değeri tahmin eder. Bu tahmin değeri, gerçek (veya doğru) değer ile bir geribildirim biçimi ile karşılaştırılır. Eğer tahmin doğru ise ağ başka bir faaliyet göstermez. Eğer tahmin yanlış ise yapay sinir ağı, kestirimin kalitesini arttırmak için hangi iç parametrelerin ne şekilde düzeltileceğini belirlemek için kendi kendisini analiz eder. Bu parametre ayarlamaları yapıldıktan sonra yapay sinir ağı veri setinin başka bir bölümünü alır ve süreci tekrar eder. Bu süreç içerisinde zamanla yapay sinir ağı büyük oranda doğru bir modele yakınsamaya başlar (Marakas, 2003, s. 130-131).

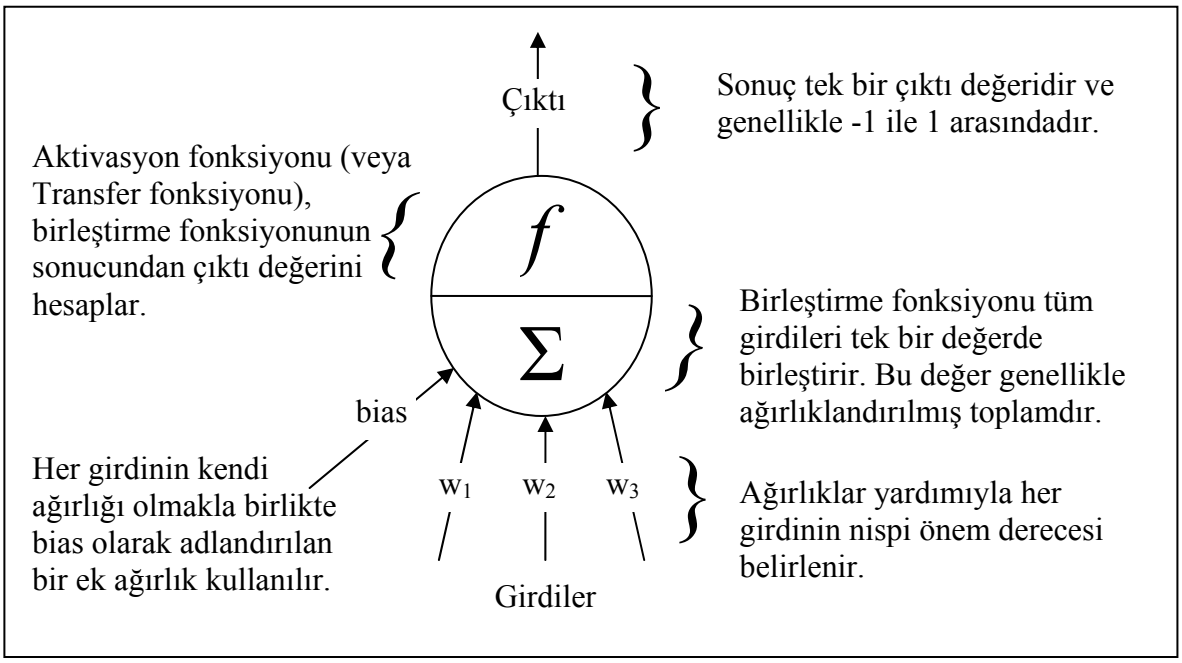
Yapay sinir ağları yüksek oranda parametrelerle ifade edilen istatistiksel model sınıflarından biridir ve son yıllarda oldukça çok ilgi çekmiştir. Yüksek oranda parametrelerle ifade edildiğinden dolayı yapay sinir ağları oldukça esnektirler ve bu yüzden fonksiyonlardaki nispeten küçük aykırılıkları doğru bir şekilde modelleyebilirler. Diğer yandan böylesi bir esneklik ciddi bir aşırı-uyum (overfitting) tehlikesini beraberinde getirir (Hand vd., 2001, s.391).

Yapay sinir ağları güçlü ve genel amaçlı veri madenciliği yöntemlerinden biridir ve kestirim, sınıflandırma ve kümeleme problemlerinde uygulanabilir. Bu alanda yapılan güncel çalışmalar yapay sinir ağlarının güçlü örüntü sınıflandırma ve kestirim kapasitelerini kanıtlamışlardır (Zhang, 2004, s. 1). Yapay sinir ağları geniş bir yelpazede ve birçok alanda problemlerin çözümünde kullanılabilen çok güçlü ve esnek sayısal araçlardan biridir (Pham vd., 2006, s. 307). Yapay sinir ağlarının kullanılma amaçları arasında sınıflandırma, örüntü tanıma ve eşleştirme, örüntü tamamlama, optimizasyon ve kontrol sayılabilir (Tsetsekas vd., 2006, s. 2987).

Yapay sinir ağlarının esas gücü yüksek kestirimci performansından kaynaklanmaktadır. Bu ağların yapısı kestirimci değişkenler ve bağımlı değişken arasındaki çok karmaşık ilişkileri yakalamayı desteklemektedir. Yapay sinir ağlarının bu gücü diğer sınıflandırıcı yöntemlerde genellikle mümkün olmamaktadır (Shmueli vd., 2007, s. 167).

2.5.1. Nöronun Yapısı

Bir yapay sinir ağı, nöronlar olarak adlandırılan temel sayısal birimler kümesinden oluşur ve bu yapay nöronlar biyolojik nöronlardan yola çıkılarak modellenmişlerdir. Şekil 2.6'da bir yapay sinir ağı nöronunun bileşenleri gösterilmektedir. Burada nöronun çıktısı, girdilerinin doğrusal olmayan (nonlinear) kombinasyonudur.



Şekil 2.6. Yapay Sinir Ağı Nöronunun Bileşenleri (Berry ve Linoff, 2004, s. 223'den uyarlanmıştır.)

Nöronlar ağırlıklandırılmış bağlantılar yoluyla birbirlerine bağlanırlar. Bu birimler katmanlar (layer) içerisinde organize edilirler ve bir katmandaki her bir nöron yalnızca kendinden bir önceki ve kendinden bir sonraki katmanlarda yer alan nöronlara bağlanır. Her nöron otonom bir sayısal birimi temsil eder ve girdileri, aktivasyonunu belirleyen bir sinyal serisi olarak alırlar. Aktivasyonu takiben her nöron bir çıktı sinyali üretir. Girdi sinyalleri nörona aynı anda ulaşır ve nöron birden fazla girdi sinyali alır, ancak tek bir çıktı sinyali üretir. Her girdi sinyali bir ağırlıkla ilişkilendirilir ve bu ağırlıklar girdi sinyalinin nöron tarafından iletilen sonuç sinyalini üretmedeki nispi önemini belirler. Bias olarak adlandırılan

bir eşik değeri de genellikle girdi olarak kullanılır. Bias, regresyon modelindeki sabit (intercept) ile benzer bir yapıdadır (Giudici, 2003, s. 107).

2.5.2. Birleştirme ve Aktivasyon Fonksiyonları

Birleştirme fonksiyonu (combination function) tüm girdileri tek bir değerde birleştiren fonksiyondur. En yaygın kullanılan birleştirme fonksiyonu ağırlıklandırılmış toplam fonksiyonudur (Berry ve Linoff, 2004, s. 222). Burada her girdi ilgili ağırlığı ile çarpılır ve sonuçlar birbirine eklenir. Ağırlıklar, doğal nöronun biyolojik sinaptik gücüne benzetim yapmaktadır. Yapay sinir ağları literatüründe ağırlıklandırılmış toplam fonksiyonunun sonucu genellikle net girdiyi ifade eden net ile gösterilir ve birden çok girdi değişkeni $x_i, i = 1,\ldots,m$ olmak üzere net girdi aşağıdaki şekilde ifade edilmektedir (Kantardzic, 2003, s. 197):

$$net_k = bias_k + x_1 w_{k1} + x_2 w_{k2} + \cdots + x_m w_{km}$$

Eğer bias ifadesini $w_{k0} = bias_k$ olarak ifade ederek ve sabit girdi değerini $x_0 = 1$ olarak tanımlarsak, net girdi toplamı ifadesini şu şekilde yazabiliriz:

$$net_k = \sum_{i=0}^{m} x_i w_{ki}$$

Bazı diğer birleştirme fonksiyonları da zaman zaman kullanışlıdır ve bunlara örnek olarak ağırlıklandırılmış girdilerin maksimumu, minimumu, değerlerin mantıksal (logical) VE (AND), VEYA (OR) veya ÖZEL VEYA (Exclusive OR – XOR) işlemleri gösterilebilir. Birleştirme fonksiyonunun seçiminde esneklik olmakla birlikte birçok durumda ağırlıklandırılmış toplamlar iyi sonuç vermektedir (Berry ve Linoff, 2004, s. 222).

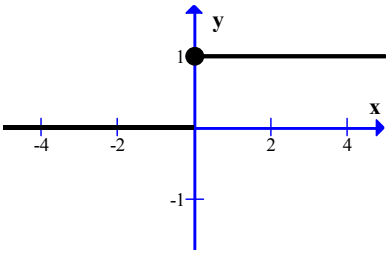
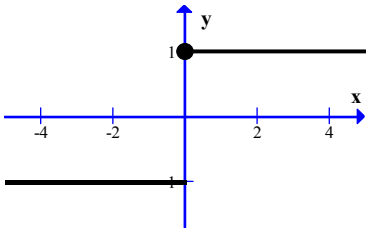
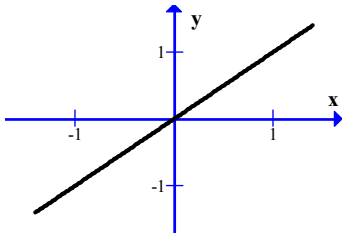
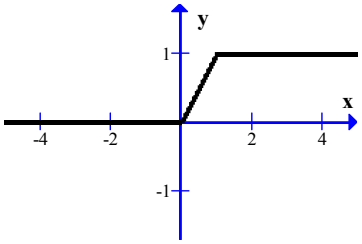
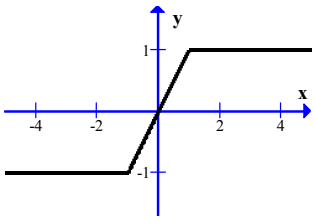
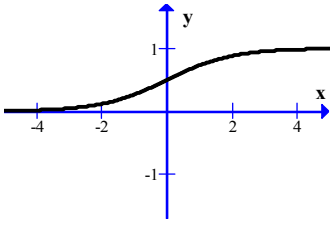
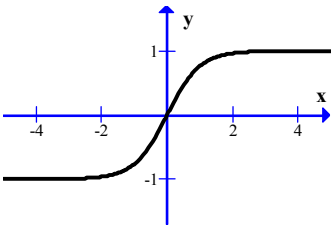
Aktivasyon fonksiyonu kavramı yerine literatürde bazı kaynaklarda (örn. Katayama vd., 2003; Kros vd., 2006; Qu vd., 2008; Gougam vd., 2008) transfer fonksiyonu kavramı kullanılmaktadır. Ancak yaygın olarak aktivasyon fonksiyonu kavramının kullanıldığı görülmektedir (Daqi ve Genxing, 2003; Solazzi ve Uncini, 2004; Tang ve MacLennan, 2005; Han ve Kamber, 2006; Chen ve Wang, 2007; Huang ve Cao, 2008; Amin ve Murase, 2009; Skubalska-Rafajlowicz, 2009). Bazı çalışmalarda da aktivasyon fonksiyonunu, birleştirme fonksiyonu ve transfer fonksiyonunun birlikte oluşturduğu fonksiyonun tümü olarak

tanımlamıştır (Berry ve Linoff, 2004; Kros vd., 2006). Bu çalışmada ise en çok kabul gören şekliyle net girdiyi (net input) hesaplayan fonksiyon birleştirme fonksiyonu, net girdiyi kullanarak nöronun çıktı değerini belirleyen fonksiyon ise aktivasyon fonksiyonu olarak kabul edilmiştir.

Aktivasyon fonksiyonu, bir nöronun aktivasyonunu nöronun çıktısına çevirir ve bunu gerçekleştirmek için doğrusal, sigmoid ve gaussian gibi farklı tiplerde aktivasyon fonksiyonları kullanılabilir (Kros vd., 2006, s. 3137). Aktivasyon fonksiyonlarının tiplerinin, yapay sinir ağlarının öğrenme hızları, doğru sınıflandırma yüzdeleri ve doğrusal olmayan eşleştirme kestirimleri gibi konular üzerinde önemli etkileri vardır (Daqi ve Genxing, 2003, s.870).

k numaralı nöronun çıktısını net_k net girdi değerinin belirli bir fonksiyonu olarak hesaplayan aktivasyon fonksiyonunu genel biçimde $y_k = f(net_k)$ olarak ifade edebiliriz. Tablo 2.2'de bazı yaygın kullanılan aktivasyon fonksiyonları ve bunların grafikleri gösterilmiştir (Kantardzic, 2003, s. 198).

Tablo 2.2. Yapay Sinir Ağlarında Kullanılan Aktivasyon Fonksiyonları

Aktivasyon Fonksiyonu	**Girdi/Çıktı Bağımlılığı**	**Grafiği**
Katı Limit (Hard Limit)	$y = \begin{cases} 1 & if\ net \geq 0 \\ 0 & if\ net < 0 \end{cases}$	
Simetrik Katı Limit (Symmetrical Hard Limit)	$y = \begin{cases} 1 & if\ net \geq 0 \\ -1 & if\ net < 0 \end{cases}$	
Doğrusal (Linear)	$y = net$	
Doygun Doğrusal (Saturating Linear)	$y = \begin{cases} 1 & if\ net > 1 \\ net & if\ 0 \leq net \leq 1 \\ 0 & if\ net < 0 \end{cases}$	
Simetrik Doygun Doğrusal (Symmetric Saturating Linear)	$y = \begin{cases} 1 & if\ net > 1 \\ net & if\ -1 \leq net \leq 1 \\ -1 & if\ net < -1 \end{cases}$	
Sigmoid veya Log-Sigmoid	$y = \frac{1}{1+e^{-net}}$	
Hiperbolik Tanjant Sigmoid (Hyperbolic Tangent Sigmoid)	$y = \frac{e^{net} - e^{-net}}{e^{net} + e^{-net}}$	

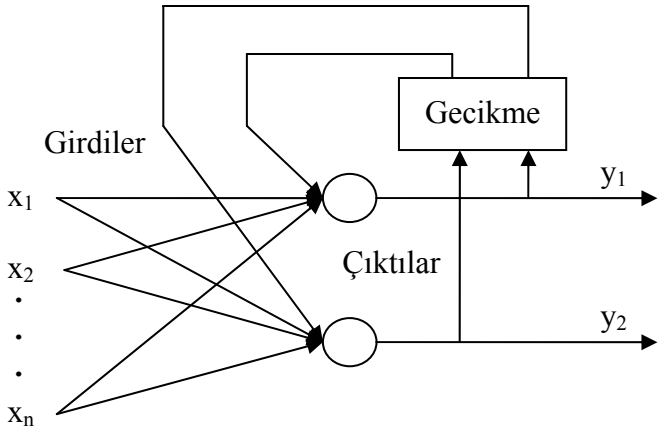
Kaynak: Kantardzic, 2003, s.198

2.5.3. Yapay Sinir Ağları Mimarisi

Yapay sinir ağlarının arkasında yatan fikir, girdi bilgilerini, bunların kendi aralarında ve bağımlı değişkenle olan karmaşık ilişkilerini yakalamak ve bunu çok esnek bir yolla birleştirmektir. Örneğin, doğrusal regresyon modellerinde bağımlı değişken ve kestirimci değişkenler arasındaki ilişkinin biçimi doğrusal olarak varsayılır. Oysa birçok durumda ilişkilerin gerçek biçimi daha karmaşıktır. Yapay sinir ağı, kestirimcilerin kendi aralarındaki ve bağımlı değişkenle olan ilişkilerini veriden öğrenmeye çalışır. Gerçekte, doğrusal regresyon ve lojistik regresyon, yapay sinir ağlarının sadece girdi ve çıktı katmanları olan ve gizli katmanı olmayan, daha basit ve özel bir durumu olarak düşünülebilir (Shmueli vd., 2007, s. 168).

Gizli katmana sahip olmayan ve girdi ve çıktı katmanları olmak üzere iki tip katmana sahip yapıya yapay sinir ağları literatüründe perceptron adı verilmektedir. Bir yapay sinir ağı ise perceptron modelinden daha karmaşık bir yapıya sahiptir (Tan vd., 2006, s. 247-251). Yapay sinir ağları, katmanları arasındaki bağlantıların tiplerine göre genel olarak, ileri beslemeli (feedforward) ve yinelemeli (recurrent) olmak üzere iki kategoride sınıflandırılırlar.

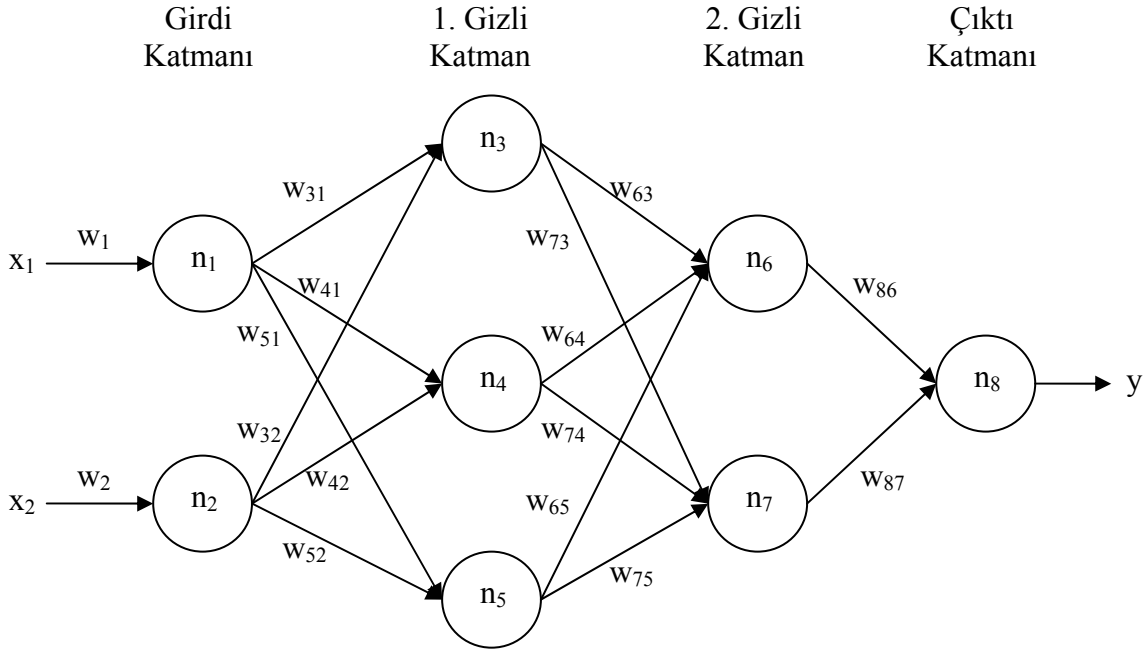
Yapay sinir ağında döngüsel yol oluşturan bir geri besleme (feedback) hattı varsa, bu ağ yinelemeli ağ olarak adlandırılır. Bu yapıda genellikle senkronizasyonu sağlamak amacıyla bir gecikme (delay) bileşeni de kullanılmaktadır. Şekil 2.7'de bir yinelemeli yapay sinir ağı mimarisi görülmektedir (Katardzic, 2003, s. 200).



Şekil 2.7. Yinelemeli (Recurrent) Yapay Sinir Ağı Mimarisi

Yinelemeli yapay sinir ağları dinamik yapılı ağlardır ve mimarileri statik olanlara göre geri besleme içermelerinden dolayı temel olarak farklıdır. Yinelemeli ağlar, ileri beslemeli olanlara göre daha küçük bir mimariye sahip olmaktadırlar (Goh ve Mandic, 2003, s. 1095).

Yapay sinir ağlarında en yaygın kullanılan mimari, yapılarındaki esneklik, iyi temsil yetenekleri ve geniş sayıdaki öğrenme algoritmaları ile ileri beslemeli ağlardır (Ma ve Khorasani, 2003, s. 361). İleri beslemeli yapay sinir ağları (Feedforward Neural Networks) bir girdi katmanı, bir veya daha fazla gizli katmanı ve bir de çıktı katmanı olan ağlardır. Girdi katmanı, eğitim seti girdilerini alarak ağırlıklandırır ve çıktılarını kendisinden sonra gelen gizli katmana eş zamanlı olarak gönderir. Bir gizli katmanın çıktıları, kullanılan gizli katman sayısına bağlı olarak bir sonraki gizli katmanın girdileri olabilir. Gizli katman sayısı analizcinin takdirine göre seçilebilir, ancak uygulamada genellikle tek gizli katman kullanılmaktadır. Son gizli katmanın ağırlıklandırılmış çıktıları, çıktı katmanının girdileridir ve bu katman çıktı olarak yapay sinir ağının verilen eğitim seti için yaptığı kestirimi verir (Han ve Kamber, 2006, s. 328).



Şekil 2.8. İleri Beslemeli Yapay Sinir Ağı

Şekil 2.8'de iki gizli katmanı olan bir ileri beslemeli yapay sinir ağı modeli görülmektedir. Bazı kaynaklarda ileri beslemeli yapay sinir ağına, çok katmanlı perceptron (multilayer perceptron) adı verilmektedir (Giudici, 2003, s. 111; Wang ve Fu, 2005, s. 25). Bishop (2005, s. 116-117) çok katmanlı perceptron kavramı ile ileri beslemeli yapay sinir ağı kavramlarının her ikisini birden birbirinin yerine geçebilir şekilde kullanmıştır. Bazı kaynaklarda bu yapıya

çok katmanlı ileri baslemeli yapay sinir ağı (multilayer feedforward neural network) adı verilmektedir (Han ve Kamber, 2006, s. 328; Tan vd., 2006, s. 251). Bazı kaynaklarda ise sadece ileri beslemeli yapay sinir ağı (feedforward neural network) adı verilmektedir (Arulampalam ve Bouzerdoum, 2003, s. 561; Kantardzic, 2003, s. 200; Tang ve MacLennan, 2005, s. 248; Benardos ve Vosniakos, 2007, s. 365). Bu tip ağ topolojisine sahip bir modelin adlandırılması bu çalışmada iki nedenle ileri beslemeli yapay sinir ağı olarak benimsenmektedir. Birincisi, çok katmanlı perceptron'un bir yapay sinir ağı biçimi olmasıdır. İkincisi ise, bu mimarinin perceptron değil de yapay sinir ağı olması için en az bir gizli katmana sahip olması gerekmektedir ve dolayısıyla zaten çok katmanlı olmak durumundadır.

2.5.4. Yapay Sinir Ağlarında Modelin Öğrenmesi ve Geri Yayılım Algoritması

Optimizasyon algoritması, hata fonksiyonunu minimize etmek yoluyla ağırlık tahminlerini elde edebilecek sayısal olarak etkin bir yöntemdir (Guidici, 2003, s.113). Maliyet fonksiyonu (cost function) olarak da adlandırılan hata fonksiyonu, hedef değişken (bağımlı değişken) ile gerçekleşen çıktı vektörü arasındaki uzaklıkla temsil edilir (Solazzi ve Uncini, 2004, s. 252). Yapay sinir ağları öğrenme algoritmasının temel amacı ağırlıklar kümesinin değerlerinin (**w**), hata fonksiyonunu minimize etmek üzere belirlenmesidir. Hata fonksiyonu, yapay sinir ağının tahmin değeri ile gerçek hedef değişken değeri arasındaki hata kareleri toplamının yarısı olarak ifade edilebilir (Tan vd., 2006, s. 253):

$$E(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Öğrenme olarak adlandırılabilecek olan hataların kullanılarak ağırlıkların düzeltilmesi için en yaygın yöntem geri yayılım (backpropagation) algoritmasıdır (Shmueli vd., 2007, s. 172). Geri yayılım şeması ileri beslemeli yapay sinir ağlarının öğrenmesi için kullanılan temel bir denetimli öğrenen algoritmadır (Kathirvalavakumar ve Thangavel, 2006, s. 111).

Geri yayılım algoritması kavramsal olarak, tekrarlamalı (iterative) bir kademeli azalma algoritmasıdır (gradient descent algorithm). Öğrenme aşaması süresince yapay sinir ağının çıktıları, gerçek hedef değerler (bağımlı değişken değerleri) ile karşılaştırılır. Eğer bu değerler birbirlerinden farklı ise bir hata fonksiyonu üretilir. Bu hata ağ boyunca geri yayılır ve ağırlıklar buna göre düzeltilir (Detienne vd., 2003, s. 245-246).

Yapay sinir ağının tahmin değerini veren çıktı katmanı da dahil olmak üzere, tüm gizli katmanlar için çıktı değerleri (O_j) hesaplanır. Uygulamada her birimin ara çıktı değerlerini kaydetmek veya belleğe almak yararlı olmaktadır. Çünkü bu değerlere daha sonra hatanın geri yayılımı sırasında ihtiyaç duyulmaktadır. Bu durum hesaplama miktarını ve zamanını düşürmektedir. Hesaplanan hata, ağın tahmin hatasını yansıtmak üzere geriye doğru ağırlıkları ve bias değerini düzelterek yayılır. Geri yayılım algoritmasında parametreler aşağıdaki biçimde hesaplanmaktadır (Han ve Kamber, 2006, s. 332-333):

Aktivasyon fonksiyonu olarak sigmoid (veya lojistik) fonksiyonu kullanıldığı durumda çıktı katmanındaki j nöronunun hatası Err_j aşağıdaki biçimde hesaplanır.

$$Err_j = O_j\left(1 - O_j\right)\left(T_j - O_j\right)$$

Burada O_j, j nöronunun çıktı değeridir, T_j ise hedef değer yani bağımlı değişkenin gerçek değeridir. $O_j\left(1 - O_j\right)$ ise sigmoid fonksiyonun türevidir. Gizli katmandaki bir j nöronunun hatasını hesaplamak için, bu nörona bağlı bir sonraki katmandaki tüm nöronların hatalarının ağırlıklandırılmış toplamlarını hesaba katmak gerekir. Dolayısıyla bir gizli katman nöronunu hatası:

$$Err_j = O_j\left(1 - O_j\right)\sum_k Err_k w_{jk}$$

ile formüle edilir. Burada w_{jk}, j nöronunun bir sonraki katmanda yer alan k nöronu ile bağlantısının ağırlığıdır. Err_k ise üst katmandaki k nöronunun hatasıdır. Böylece bir sonraki katmanda yer alan j nöronuna bağlı tüm nöronların ağırlıklandırılmış hata toplamları hesaba katılmış olur. Ağırlıklar ve bias değerleri geri yayılan hataları yansıtmak üzere düzeltilirler:

$$\Delta w_{ij} = (l)Err_j O_i \qquad \Delta\theta_j = (l)Err_j$$
$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad \theta_j = \theta_j + \Delta\theta_j$$

Burada Δw_{ij}, w_{ij} ağırlığındaki değişimdir, $\Delta\theta_j$ ise θ_j bias değerindeki değişimdir. l ise öğrenme oranıdır (learning rate) ve genellikle 0 ile 1 arasında değer alır. Öğrenme oranı,

karar uzayında yerel minimuma takılmaktan kurtulmaya ve global minimumu bulmaya yardımcı olur. Eğer öğrenme oranı çok küçük olursa öğrenme çok yavaş gerçekleşir, eğer çok büyük olursa da yetersiz çözümler arasında gidip gelme durumu gerçekleşebilir.

Yapay sinir ağının ilk yapılandırılması sürecinde tüm nöron ağırlıklarının rassal olarak ayarlanması gerekmektedir. Bu yapay sinir ağının öğrenmesinin önemli bir aşamasıdır. Eğer nöron ağırlıkları birbirleriyle aynı şekilde ayarlanırsa, her nöron örneğin, girdi sinyalinin aynı kısmını öğrenmeyi deneyecektir. Bunun sonrasında geri yayılım algoritmasıyla her bir nörona birbiriyle aynı hatalar geri gelecektir ve hepsi de girdinin aynı küçük bölümünde takılıp kalacaklardır. Ağırlıkların rassal olarak ayarlanması ile, nöronların hepsi de girdinin aynı kısmına yakınsamaya çalışsalar da geri yayılımdan gelen hatalar farklı olacağından, baskın olan nöron dışındaki diğer nöronlar sinyalin diğer kısımlarına yakınsamaya çalışacaklardır (Pyle, 1999, s. 368).

2.6. Zaman Serileri

Bir zaman serisi veritabanı, zaman içerisinde tekrarlayan ölçümlerden elde edilen değerleri içerir. Bu değerler tipik olarak eşit zaman aralıklarında ölçülürler (günlük, haftalık ve yıllık gibi). Zaman serisi veritabanları, hisse senedi piyasası analizleri, ekonomik tahminler, bütçe analizi, süreç ve kalite kontrolü ve iş yükü tahminleri gibi uygulamalarda kullanılmaktadır (Han ve Kamber, 2006, s. 489). Zaman serisi analizi, zamansal veri içerisindeki örüntü ve trendleri keşfetmek için güçlü bir tekniktir. Ancak bu veri madenciliği tekniği için kavramsal model eksikliği analizcinin yapılandırılmamış veri ile uğraşmasına neden olur. Bu türdeki veri düşük seviyede bir düzenlenmişliğe sahiptir ve yönetimi zordur (Zubcoff vd., 2009, s. 977).

Zaman serileri, değişkenlerin gün, hafta, ay, mevsim veya yıl gibi herhangi bir zaman dilimine göre dağılımını gösteren serilerdir. Çeşitli değişkenler için düzenlenmiş zaman serileri için özel tahmin teknikleri geliştirilmiştir. Zaman serisi analizlerinin kullanıldığı en önemli alanlar ekonomi ve işletme alanlarıdır. Tahmin edilecek değişkenlerin geçmiş değerlerinin çeşitli yöntemlerle incelenmesine dayanan zaman serisi analizleri altı grupta toplanmaktadır (Orhunbilge, 1999, s. 3-4):

1. Zaman serilerinin bileşenlerine ayrılması yöntemi (Decomposition methods)
2. Üstel düzgünleştirme yöntemleri (Exponential smoothing methods)

3. Otoregresif yöntemler (Autoregressive methods)
4. Hareketli ortalama yöntemleri (Moving average methods)
5. Bileşik otoregresif hareketli ortalama yöntemleri (Autoregresive integrated moving average methods)
6. Zaman serilerinde regresyon yöntemleri (Regression methods in time series)

Mevcut çalışmada zaman serisi analizlerinde üstel düzgünleştirme yöntemleri ve bileşik otoregresif hareketli ortalama (Box-Jenkins) yöntemleri kullanıldığından bu kısımda bu iki yönteme ve ayrıca zaman serilerinde yapay sinir ağlarının kullanılmasına değinilecektir.

2.6.1. Üstel Düzgünleştirme Yöntemleri

İleriye dönük kestirim sistemlerinde, zaman serilerinin kestirimleri ilerleyen periyotlar için her periyotta yapılmaktadır. Dolayısıyla kestirim denklemi ve zaman serisi parametrelerinin tahminleri her periyodun sonunda en yeni gözlemi de hesaba katarak düzeltilmelidir. Bu düzeltme işlemi parametrelerdeki zaman içerisinde olabilecek değişimleri hesaba katacak şekilde yapılmalıdır. Buna ek olarak, bu değişiklikler, parametre tahminleri düzeltildiğinde zaman serisi gözlemlerine eşit olmayan ağırlıklar uygulanması gerekliliğini gösterebilir. Üstel düzgünleştirme, gözlenen zaman serisi değerlerini eşit olmayan bir biçimde ağırlıklandıran bir kestirim yöntemidir (Bowerman ve O'Connell, 1993, s. 379).

Üstel düzgünleştirme yöntemleri özellikle verinin gürültülü olduğu zamanlarda veriyi düzgünleştirmek için kullanışlı olabilmektedir (Kirkup, 2002, s. 379). Üstel düzgünleştirme, zaman serisinin trend ve mevsimsel faktör gibi bileşenleri zaman içerisinde değişken olduğunda etkin bir tahmin yöntemidir. Bu yöntem, gözlenen zaman serisi değerlerini eşit olmayan bir şekilde ağırlıklandırmaktadır. Daha yakın gözlemler daha yüksek ağırlıklandırılırken, daha uzak gözlemler daha düşük ağırlıklandırılmaktadır. Eşit olmayan ağırlıklandırma bir veya daha fazla düzgünleştirme sabiti (smoothing constant) kullanılarak gerçekleştirilir ve bu sabit her gözleme ne kadar ağırlık atanacağını belirlemektedir (Bowerman vd., 2005, s. 345).

Mevsimsel zaman serileri için en yaygın bilinen ve kullanılan yöntemler Winters (1960) tarafından öne sürülen yöntemlerdir. Bunlardan biri toplamsal mevsimsellik için olan Toplamsal (Additive) Holt-Winters yöntemi, diğeri ise çarpımsal mevsimsellik için önerilen Çarpımsal (Multiplicative) Holt-Winters yöntemidir (Koehler vd., 2001, s. 269).

2.6.1.1. Toplamsal (Additive) Holt-Winters Yöntemi

Eğer zaman serisi belirli bir büyüme oranında (β_1, growth rate) doğrusal bir trende ve sabit bir toplamsal (additive) değişkenlikte bir mevsimsel örüntüye (pattern) sahipse, aşağıdaki modelle ifade edilebilir (Bowerman vd., 2005, s. 367):

$$y_t = (\beta_0 + \beta_1 t) + SN_t + \varepsilon_t$$

Toplamsal Holt-Winters yönteminin uygulanması aşağıdaki şekilde özetlenebilir (Bowerman vd., 2005, s. 367-369):

1. $y_1, y_2, \ldots, y_n$ gibi bir zaman serisinin doğrusal bir trend gösterdiği ve sabit bir mevsimsel değişkenliği olan bir mevsimsel örüntüye sahip olduğu varsayımıyla değerlerin, büyüme oranının ve mevsimsel örüntünün değişken olduğu düşünülürse, zaman serisinin T dönemi için değer tahmini ℓ_T, büyüme oranı tahmini b_T ve mevsimsel faktör tahmini sn_T düzgünleştirme denklemleri aşağıdaki gibi yazılabilir:

$$\ell_T = \alpha(y_T - sn_{T-L}) + (1-\alpha)(\ell_{T-1} + b_{T-1})$$
$$b_T = \gamma(\ell_T - \ell_{T-1}) + (1-\gamma)b_{T-1}$$
$$sn_T = \delta(y_T - \ell_T) + (1-\delta)sn_{T-L}$$

   Burada α, γ ve δ, 0 ile 1 arasında değişen düzgünleştirme sabitleridir. ℓ_{T-1} ve b_{T-1}, $T-1$ zamanında değer (seviye) ve büyüme oranı için tahminlerdir, sn_{T-L} ise $T-L$ dönemi için mevsimsel faktör tahminidir.

2. T zaman periyodunda $y_{T+\tau}$ noktası için yapılan tahmin:

$$\hat{y}_{T+\tau}(T) = \ell_T + \tau b_T + sn_{T+\tau-L} \qquad (\tau = 1, 2, \ldots)$$

   biçiminde yazılabilir. Burada $sn_{T+\tau-L}$, $T+\tau$ zaman periyoduna ilişkin mevsim için "en yakın" mevsimsel faktör tahminidir.

3. $y_{T+\tau}$ için T zaman periyodunda %95 güven aralığında hesaplanan tahmin aşağıdaki biçimdedir:

$$\left[\hat{y}_{T+\tau}(T) \pm z_{[,025]} s\sqrt{c_\tau}\right]$$

$$Eger \quad \tau = 1 \quad ise \quad c_1 = 1$$

$$Eger \quad 2 \le \tau \le L \quad ise \quad c_\tau = \left[1 + \sum_{j=1}^{\tau-1} \alpha^2 (1 + j\gamma)^2\right]$$

$$Eger \quad L \le \tau \quad ise \quad c_\tau = 1 + \sum_{j=1}^{\tau-1} \left[\alpha(1 + j\gamma) + d_{j,L}(1 - \alpha)\delta\right]^2$$

Burada, eğer j, L'nin tam sayı katı ise $d_{j,L} = 1$'dir. Eğer değilse $d_{j,L} = 0$'dır. T zaman periyodu için hesaplanan standart hata s aşağıdaki gibidir:

$$s = \sqrt{\frac{SSE}{T-3}} = \sqrt{\frac{\sum_{t=1}^{T}\left[y_t - \hat{y}_t(t-1)\right]^2}{T-3}} = \sqrt{\frac{\sum_{t=1}^{T}\left[y_t - (\ell_{t-1} + b_{t-1} + sn_{t-L})\right]^2}{T-3}}$$

4. Toplamsal Holt-Winters yönteminde düzgünleştirme denklemleri için hata düzeltme biçimi aşağıdaki şekildedir:

$$\ell_T = \ell_{T-1} + b_{T-1} + \alpha\left[y_T - (\ell_{T-1} + b_{T-1} + sn_{T-L})\right]$$

$$b_T = b_{T-1} + \alpha\gamma\left[y_T - (\ell_{T-1} + b_{T-1} + sn_{T-L})\right]$$

$$sn_T = sn_{T-L} + (1 - \alpha)\delta\left[y_T - (\ell_{T-1} + b_{T-1} + sn_{T-L})\right]$$

2.6.1.2. Çarpımsal (Multiplicative) Holt-Winters Yöntemi

Eğer mevsimsel faktör sabit bir değer değil, fakat sabit bir oran ise çarpımsal yöntem uygulanabilir. Bu durumda düzeltilmiş trend ile düzeltilmiş mevsimsel faktör toplanmak yerine çarpılmaktadır. Verinin mevsim etkisinden arındırılması için ise, uygun mevsimsel faktörün veriden çıkarılması yerine bu faktöre bölünmektedir (Gaynor ve Kirkpatrick, 1994, s. 387). Çarpımsal Holt-Winters yöntemi ile bir seri aşağıdaki biçimde ifade edilebilir.

$$y_t = (\beta_0 + \beta_1 t) \times SN_t \times IR_t$$

Burada IR_t düzensiz (irregular) bileşeni, SN_t ise mevsimsel faktörü ifade etmektedir. $\beta_0 + \beta_1 T$, T zamanındaki seviyeyi gösterirken, β_1, bu seviye için artış oranını (growth rate) göstermektedir. Çarpımsal Holt-Winters yöntemi aşağıda özetlenmiştir (Bowerman vd., 2005, s. 374-377):

1. $y_1, y_2, \ldots, y_n$ gibi bir zaman serisinin doğrusal bir trend gösterdiği ve artan (çarpımsal) bir mevsimsel değişkenliği olan bir mevsimsel yapıya sahip olduğu varsayımıyla değerlerin, büyüme oranının ve mevsimsel yapının değişken olduğu düşünülürse, zaman serisinin T dönemi için değer tahmini ℓ_T, büyüme oranı tahmini b_T ve mevsimsel faktör tahmini sn_T düzgünleştirme denklemleri aşağıdaki gibi yazılabilir:

$$\ell_T = \alpha\left(\frac{y_T}{sn_{T-L}}\right) + (1-\alpha)(\ell_{T-1} + b_{T-1})$$

$$b_T = \gamma(\ell_T - \ell_{T-1}) + (1-\gamma)b_{T-1}$$

$$sn_T = \delta\left(\frac{y_T}{\ell_T}\right) + (1-\delta)sn_{T-L}$$

Burada α, γ ve δ, 0 ile 1 arasında değişen düzgünleştirme sabitleridir. ℓ_{T-1} ve b_{T-1}, $T-1$ zamanında değer (seviye) ve büyüme oranı için tahminlerdir, sn_{T-L} ise $T-L$ dönemi için mevsimsel faktör tahminidir.

2. T zaman periyodunda $y_{T+\tau}$ noktası için yapılan tahmin:

$$\hat{y}_{T+\tau}(T) = (\ell_T + \tau b_T)sn_{T+\tau-L} \qquad (\tau = 1, 2, \ldots)$$

olarak yazılabilir. Burada $sn_{T+\tau-L}$, $T+\tau$ zaman periyoduna ilişkin mevsim için "en yakın" mevsimsel faktör tahminidir.

3. $y_{T+\tau}$ için T zaman periyodunda %95 güven aralığında hesaplanan tahmin aşağıdaki biçimdedir:

$$\left[\hat{y}_{T+\tau}(T) \pm z_{[,025]} s_r \left(\sqrt{c_\tau}\right)\left(sn_{T+\tau-L}\right)\right]$$

$$Eger \quad \tau = 1 \quad ise \quad c_1 = \left(\ell_T + b_T\right)^2$$

$$Eger \quad \tau = 2 \quad ise \quad c_2 = \alpha^2 (1+\gamma)^2 \left(\ell_T + b_T\right)^2 + \left(\ell_T + 2b_T\right)^2$$

$$Eger \quad \tau = 3 \quad ise \quad c_3 = \alpha^2 (1+2\gamma)^2 \left(\ell_T + b_T\right)^2 + \alpha^2 (1+\gamma)^2 \left(\ell_T + 2b_T\right)^2 + \left(\ell_T + 3b_T\right)^2$$

$$Eger \quad 2 \le \tau \le L \quad ise \quad c_\tau = \sum_{j=1}^{\tau-1} \alpha^2 \left(1 + [\tau - j]\gamma\right)^2 \left(\ell_T + b_T\right)^2 + \left(\ell_T + \tau b_T\right)^2$$

T zaman periyodu için hesaplanan standart hata s aşağıdaki gibidir:

$$s_r = \sqrt{\frac{\sum_{t=1}^{T} \left[\frac{y_t - \hat{y}_t(t-1)}{\hat{y}_t(t-1)}\right]^2}{T-3}} = \sqrt{\frac{\sum_{t=1}^{T} \left[\frac{y_t - \left(\ell_{t-1} + b_{t-1}\right) sn_{t-L}}{\left(\ell_{t-1} + b_{t-1}\right) sn_{t-L}}\right]^2}{T-3}}$$

4. Çarpımsal Holt-Winters yönteminde düzgünleştirme denklemleri için hata düzeltme biçimi aşağıdaki şekildedir:

$$\ell_T = \ell_{T-1} + b_{T-1} + \alpha \left[\frac{y_T - \left(\ell_{T-1} + b_{T-1}\right) sn_{T-L}}{sn_{T-L}}\right]$$

$$b_T = b_{T-1} + \alpha\gamma \left[\frac{y_T - \left(\ell_{T-1} + b_{T-1}\right) sn_{T-L}}{sn_{T-L}}\right]$$

$$sn_T = sn_{T-L} + (1-\alpha)\delta \left[\frac{y_T - \left(\ell_{T-1} + b_{T-1}\right) sn_{T-L}}{\ell_T}\right]$$

2.6.2. Bileşik Otoregresif Hareketli Ortalama (Box-Jenkins) Yöntemi

George E. P. Box ve Gwilym M. Jenkins 1970 yılında otoregresif modeller ile hareketli ortalama yöntemlerinin bileşimi ARIMA (Autoregressive Integrated Moving Average Model) yöntemini geliştirmişlerdir. Literatürde Box-Jenkins yöntemi adıyla da geçen bu teknik tanımlama, tahmin ve testlerden oluşan üç aşamalı bir yöntemdir. Bileşik otoregresif hareketli ortalama yöntemi, otoregresif ve hareketli ortalama modellerinin bileşimi olduğu için bu

modeller hakkındaki tüm bilgilerin bir arada kullanılmasına dayanmaktadır (Orhunbilge, 1999, s. 189).

Box-Jenkins modelleri, mevsimsel ve mevsimsel olmayan modeller şeklinde ikiye ayrılmaktadır. Mevsimsel olmayan Box-Jenkins modelleri genel olarak ARIMA(p,d,q) şeklinde gösterilmektedir. Burada p, otoregresyon (AR) modelinin derecesi, d, fark alma işlemi sayısı ve q, hareketli ortalama (MA) modelinin derecesi olmaktadır. Mevsimsel Box-Jenkins modelleri ise genellikle ARIMA(p,d,q)$(P,D,Q)_s$ biçiminde ifade edilmektedir. Burada P, mevsimsel otoregresyon (SAR) modelinin derecesi, D, mevsimsel fark alma işlemi sayısı, Q, mevsimsel hareketli ortalama (SMA) modelinin derecesi ve s, periyot olmaktadır (Kadılar, 2005, s. 185).

Bir bileşik otoregresif hareketli ortalama modelinde, bir değişkenin gelecekteki değeri, geçmiş gözlemlerin ve rassal hataların doğrusal bir fonksiyonu olarak varsayılır. Dolayısıyla, zaman serilerini üreten temel sürecin biçimi aşağıdaki şekilde ifade edilebilir (Zhang, 2003, s.162):

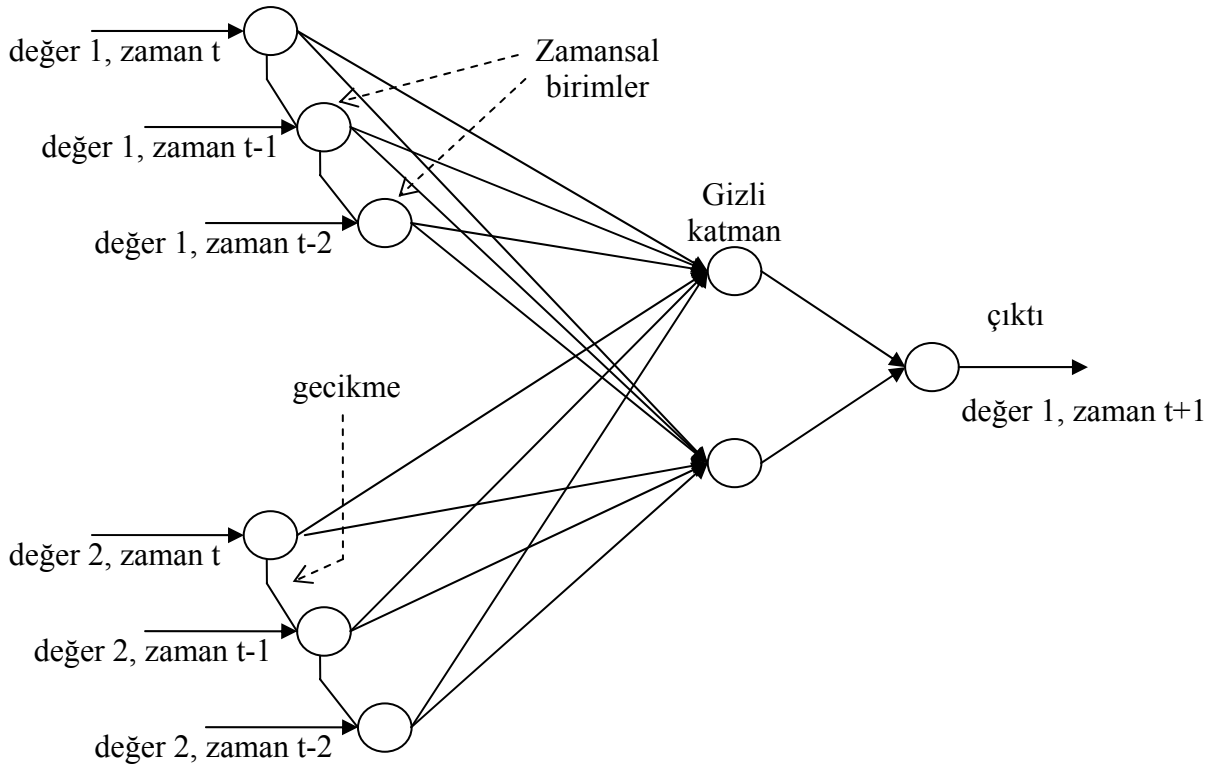
$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

Burada y_t ve ε_t, sırasıyla t periyodundaki gerçek değer ve rassal hatadır. $\phi_i \ (i = 1,2,\ldots,p)$ ve $\theta_j \ (j = 0,1,2,\ldots,q)$ model parametreleri olmak üzere, p ve q genellikle modelin üstel dereceleri olarak ifade edilen tam sayılardır. Rassal hataların (ε_t) ise ortalamaları sıfır edecek şekilde ve σ^2 sabit varyansla bağımsız ve özdeş olarak dağıldığı varsayılmaktadır. Bu eşitlik aynı zamanda ARIMA ailesi modellerinin özel durumlarını da içermektedir. Eğer $q = 0$ olursa, model p'inci dereceden AR modeli haline gelir. Eğer $p = 0$ olursa model q'uncu dereceden MA modeline indirgenir. ARIMA modelini yapılandırmanın ana görevlerinden birisi de (p,q) uygun model derecelerinin belirlenmesidir (Zhang, 2003, s. 162).

2.6.3. Zaman Serilerinde Yapay Sinir Ağları Kullanımı

Yapay sinir ağları, zaman serisi analizleri için kolayca adapte edilebilmektedir. Şekil 2.9'da zaman serileri için adapte edilmiş bir yapay sinir ağı modeli görülmektedir. Yapay sinir ağı, zaman serisi verisi üzerinde en eski veri noktasından başlayarak öğrenir. Daha sonra

öğrenme ikinci en eski veri noktası ve girdi setindeki bir sonraki zaman noktası şeklinde ilerleyerek devam eder. Ağın öğrenmesi ileri beslemeli yapay sinir ağı biçimindedir ve geri yayılım algoritması her adımda seride yer alan bir sonraki değeri kestirmeye çalışır (Berry ve Linoff, 2004, s. 245).



Şekil 2.9. Zaman Gecikmeli Yapay Sinir Ağı

Geleneksel zaman serisi yöntemleri olasılıksal istatistik kavramlarına dayanmaktadır, ancak son yıllarda yapay sinir ağları kavramı zaman serisi tahmin yöntemleri arasına entegre olmaya başlamıştır. Yapay sinir ağları ve geleneksel zaman serisi yöntemleri karşılaştırmalarının birçok çalışmada yapıldığı görülmektedir (Tseng vd., 2002, s. 72). Bazı çalışmalarda kullanılan test setlerinin tahminlerinde yapay sinir ağlarının en iyi sonuçları verdiği gösterilmiştir (Kohzadi vd., 1996; Hill vd., 1996; Prybutok vd., 2000; Ho vd., 2002; Zou vd., 2007). Maier ve Dandy (1996) ise, ARIMA modellerinin kısa-dönem tahminlerde, yapay sinir ağlarının ise uzun-dönem tahminlerde kullanımını önermektedir.

Yapay sinir ağlarının, içerisinde gürültü barındıran veya barındırmayan, doğrusal-olmayan (nonlinear) zaman serilerini modellemede ve tahminlemede etkin olduğu gösterilmiştir. Ancak yapay sinir ağlarının doğrusal olmayan bir yapıya sahip problemler için uygun olduğu beklense bile, uygulamada üzerinde çalışılan problemin doğrusal olup olmadığını belirlemek genellikle güçtür (Zhang, 2001, s. 1185).

ÜÇÜNCÜ BÖLÜM
SAĞLIK SEKTÖRÜ VERİTABANI UYGULAMASI

3.1. Araştırmanın Amacı ve Kapsamı

Bu araştırmanın amacı, veri madenciliği yöntem ve teknikleri ile sağlık sektörü ve özellikle hastane veritabanlarında bulunan işlemsel verileri kullanarak yönetim karar desteği sağlayabilecek bilgilerin çıkartılması süreçlerini gerçekleştirmek, bu süreçlerde geçilen aşamaları ortaya koymak ve kullanılan yöntem ve tekniklerin etkinliğini araştırmaktır. Bu amaç birçok uygulama alternatifini kapsayabilir ve ilgili hedefler doğrultusunda birden fazla veri madenciliği yöntem ve tekniği kullanılabilir. Bu çalışmada, hastaların ve hastane hizmetlerinin mevcut durumlarının analizi, birimler arası konsültasyon hizmetlerinin analizi ve hastaneye yapılan başvuru sayılarının gelecek dönemlere ilişkin tahminlerinin yapılması olmak üzere üç tip uygulama gerçekleştirilmiştir.

Bu bölümde öncelikle uygulama platformu tanıtılmıştır. Daha sonra hastane veritabanından çıkartılan, hastalara ve hastanede verilen hizmetlere ait tanımlayıcı istatistikler verilmiştir. Sonrasında ise veri madenciliği modelleme teknikleri kullanılarak, birliktelik kuralları tekniği ile birimler arası konsültasyon hizmetlerinin analizi ve zaman serisi yöntemleri ve yapay sinir ağları uygulamaları ile hasta başvurularının ileriye dönük tahmin edilmesi uygulamaları gerçekleştirilmiştir.

3.2. Uygulama Platformu

Uygulama çalışması, hali hazırda kullanımda olan, oldukça büyük boyuta sahip ve oldukça karmaşık bir veri kaynağında yapılan bir veri madenciliği uygulamasıdır. Bu yüzden tezin uygulama platformunu oluşturan yazılım sayısı ve çeşidi fazlalaşmıştır. Yazılımların ne şekilde kullanıldığı, Hastane Veritabanı, Veri Aktarımı ve Dönüştürme ve Veri Madenciliği Yazılımı olarak üç aşamalı başlıklar altında anlatılacaktır.

3.2.1. Hastane Veritabanı

Çalışmada kullanılan veri kaynağı, Akdeniz Üniversitesi Hastanesinin 1997 yılından bu yana kullanılan ve hastane otomasyonu sisteminin verilerinin tutulduğu hastane veritabanıdır.

Bu veritabanı işlemsel veritabanı özelliğindedir ve mevcut durumda ORACLE 9i veritabanı yönetim sistemi kullanılmaktadır. Veritabanı zamanla gelişen otomasyon ihtiyaçları doğrultusunda dönüştürülerek geliştirilmiştir. Mevcut durumda 93 tabloda yaklaşık 210 milyon kayıt vardır. Ancak veri madenciliği amaçları doğrultusunda Tablo 3.1'de bilgileri verilen sekiz tablo seçilmiştir. Ayrıca bazı tablolar ve değişkenler zaman içerisinde ihtiyaca göre eklendiğinden, veri tutarlılığı ve kayıt sayısının fazlalığı göz önüne alınarak zaman serisi analizleri hariç, 2005 öncesi kayıtlar elenerek yalnızca 01 Ocak 2005 ve sonrası kayıtlar alınmıştır.

Tablo 3.1. Veritabanından Seçilen Tablolara Ait Bilgiler

TABLOLAR	**Toplam Değişken Sayısı**	**Toplam Kayıt Sayısı**	**2005 Sonrası Kayıt Sayısı**
HASTA_HAREKET	34	66.799.175	35.665.845
HASTA_ISTEM	20	18.235.884	9.897.628
HASTA_KABUL	38	4.528.759	2.272.492
HASTA_TANILAR	6	2.379.299	2.379.096
HASTALAR	36	1.018.127	1.018.127
ICD10	4	13.566	13.566
BIRIM	12	871	871
HIZMETLER	25	28.956	28.956

Oracle veritabanı sunucusu tarafındaki işlemler TOAD for Oracle 8.5.1 aracı kullanılarak gerçekleştirlmiştir. Bu araç, sektördeki Oracle veritabanı profesyonelleri tarafından yaygın olarak kullanıldığı ve veritabanı yönetim sistemi işlemlerini kolaylaştıran bir arayüz olduğu için tercih edilmiştir. Burada veri önişleme faaliyetleri doğrultusunda verinin incelenmesi, düzenlenmesi ve yeni değişkenler elde edilmesi gibi bazı işlemler gerçekleştirilmiştir.

Yeni değişkenler elde edilmesi işlemine örnek olarak, herhangi bir hastanın hastaneye başvurusunda Hasta Kabul numarası olarak atanan KHASTAKABUL değişkenini kullanarak, başvuru tarihini Tarih/Saat değişken tipine sahip yeni bir değişken olarak elde etme ve bu kabul tarihinin haftanın hangi gününde olduğunu belirleyen başka bir değişken elde edilmesi işlemlerini verebiliriz. Yeni değişken elde etme işlemi için TOAD SQL Editöründe aşağıdaki örnek SQL cümlesi biçimi kullanılmıştır:

```
UPDATE MEDISYS.HASTA_KABUL
SET BASTARIH = TO_DATE(SUBSTR(KHASTAKABUL,1,8),'YYYY.MM.DD
HH:MI:SS')
```

Yukarıdaki SQL cümlesi, değişken tipi Tarih/Saat olan BASTARIH isimli bir değişken oluşturmaktadır. Başvuru tarihinin haftanın hangi günü olduğunun elde edilmesi için ise aşağıdaki SQL cümlesi kullanılmıştır:

```
UPDATE MEDISYS.HASTA_KABUL
SET WEEKDAY = TO_CHAR(BASTARIH,'D')
```

Bu sorgu `NLS TERRITORY='TURKEY'` olmak üzere haftanın günlerini Pazartesi = 1 olacak şekilde numaralandırmaktadır. Eğer örneğin `NLS TERRITORY='AMERICA'` olursa Cumartesi = 1 olmaktadır.

Hastanede birimlere göre inceleme yapılacağında, bazı birimlerde yatış sürelerinin kısa olup başvuru sayılarının fazla olduğu, bazı birimlerde ise başvuru sayılarının nispeten az olduğu ancak yatış sürelerinin uzun olduğu göz önünde bulundurulmalıdır. Dolayısıyla birim yoğunlukları incelenirken hastaların yatış sürelerinin mutlaka dikkate alınması gerekmektedir. Yatan hastaların hastanede toplam kaç gün yattığının elde edilmesi için başvuru kapanış tarihinden (KAPATAR) hasta kabul numarasından elde edilen kabul tarihi çıkarılmakta ve bu değer sayı biçimine dönüştürülerek GUN değişkeni elde edilmektedir. Bunun için kullanılan SQL örneği aşağıda verilmiştir.

```
UPDATE MEDISYS.HASTA_KABUL
SET GUN = TO_NUMBER(KAPATAR-
(TO_DATE(SUBSTR(KHASTAKABUL,1,8),'YYYY.MM.DD HH:MI:SS')))
WHERE YATISONAY='E'
```

3.2.2. Veri Aktarımı ve Dönüştürme

Analiz için veriler hastane veritabanından, çalışmanın yapıldığı bilgisayarda bulunan SQL Server 2005 veritabanına alınmıştır. Bu işlem sırasında filtreleme yapılarak analize uygun olmayan ve özel bilgi niteliğinde olan Ad, Soyad, T.C. Kimlik No., Kurum Karne No., vb. gibi bilgiler aktarılmamıştır. Bunun yanı sıra analizde ve farklı tablolardaki verilerin birleştirilmesi süreçlerinde gerekli olabilecek Yılın Ayı (örneğin 200501, 200502, vb.), Yaş (başvuru tarihinden doğum tarihi çıkarılarak) ve Hasta İstem numarasından Hasta Kabul numarası üretmek gibi bazı değişkenler mevcut veriden üretilmiştir.

Verinin Oracle'dan SQL Server 2005'e aktarılması (data migration) için, SQL Server Migration Assistant (SSMA) for Oracle v3.1 kullanılmıştır (Microsoft, 2007). Bu araç kullanılırken her iki veritabanı sunucusuna da bağlantı gerçekleştirilmektedir.

Oracle veritabanı bağlantısının gerçekleştirilmesinden sonra Oracle tarafından sunucudan tüm şemalar SSMA arayüzüne getirilmektedir. SQL Server veritabanı sunucusuna bağlantı yapılırken eğer SQL Server üzerinde henüz MEDISYS veritabanı yoksa bunun oluşturulmasını istemektedir. İlgili veritabanı, hem SSMA kullanılarak, hem de SQL Server Management Studio kullanılarak oluşturulabilir. Ancak SSMA kullanılırsa başlangıç veritabanı dosya boyutları (initial size) çok küçük olduğundan (3MB gibi) ve Autogrowth oranı da çok küçük atandığından, büyük miktarlardaki veri transferinde sorun oluşturacaktır. Bunun için veritabanı ya SQL Server Management Studio kullanılarak oluşturulmalı veya SSMA'da oluşturulduktan sonra Veritabanı Yönetim Sistemi üzerinden Database Properties seçilmeli ve Files kısmında Initial Size değeri, Autogrowth oranı ve veritabanı log dosyası ile ilgili parametreler değiştirilmelidir.

Mevcut projede MEDISYS veritabanı SSMA'da oluşturulmuş olup daha sonra SQL Server Management Studio'da MEDISYS Database Properties kısmında dosya özellikleri değiştirilmiştir. Oracle tarafında, TOAD Database Browser'da MEDISYS_DATA dosyası kullanılan alanı (TABLESPACE) 12.883 MB olarak görüldüğünden, başlangıç dosya boyutu 13 GB (13.312 MB) ve 512 MB Autogrowth boyutu tanımlanmıştır.

Her iki veritabanı sunucusuna da bağlantı yapıldıktan sonra sonra kaynak şema (source schema) MEDISYS ve hedef şema (target schema) MEDISYS.dbo olmak üzere şema dönüştürme işlemi SSMA'da Convert Schema kullanılarak gerçekleştirilmiştir. Burada veritabanındaki aktarılacak (migrate edilecek) olan tüm tablolar ve indekslere ait üst-veriler (metadata) dönüştürülmektedir. Mevcut çalışmada MEDISYS için dönüştürme istatistikleri Tablo 3.2'de verilmiştir:

Tablo 3.2. Veritabanı Şema Dönüştürme İstatistikleri

Cümle Tipi	Toplam	Dönüştürülen	Dönüştürülemeyen
Tüm	1.319	% 100	% 0
Sütun	1.231	% 100	% 0
İndeks	88	% 100	% 0

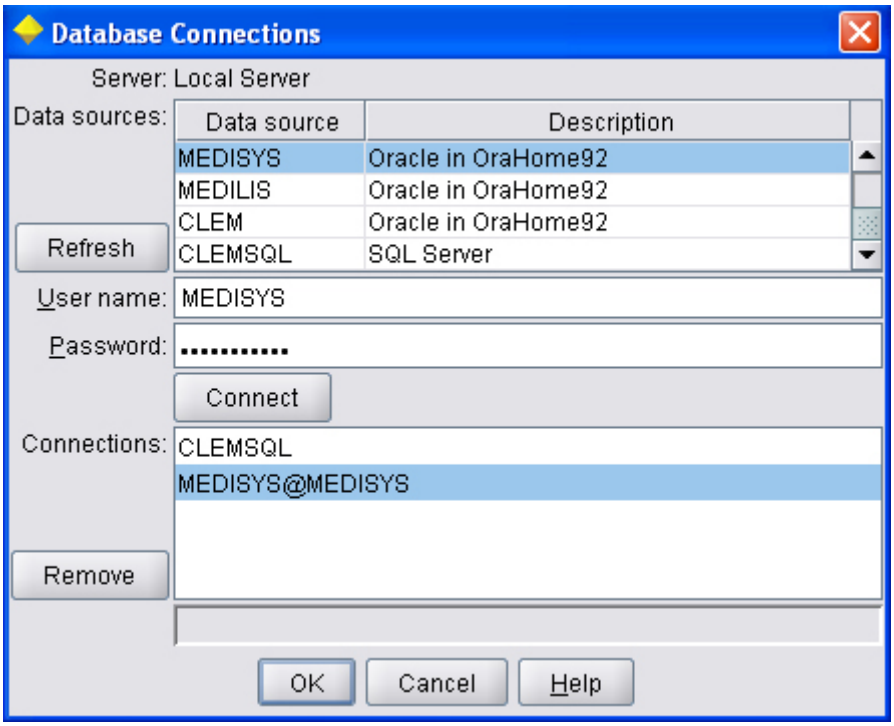
Bu aşamada SQL Server veritabanı senkronizasyonu yapılmalı ve dönüştürülen tablolar veritabanında oluşturulmalıdır. Daha sonra SSMA'da Migrate Data seçeneği ile verinin aktarılmasına başlanmıştır. Bu işlemler sonucunda seçilmiş tablolarda yer alan tablo tipine göre bütün veriler veya 01/01/2005 tarihinden sonraki veriler SQL Server 2005 veritabanı sunucusuna aktarılmıştır.

3.2.3. Veri Madenciliği Yazılımı

Tez çalışmasında veri madenciliği yazılımı olarak SPSS firmasının Clementine 11.1 paketi seçilmiştir. Clementine, CRISP-DM endüstri standart süreç modeli çerçevesinde tasarlanmış bir veri madenciliği çalışma tezgahı (workbench) olarak tanımlanmaktadır (SPSS, 2007a, s.1). Clementine, veri madenciliği algoritmalarını uygulamanın yanı sıra, veri seçme, veri birleştirme, veri dönüştürme gibi veri önişleme aşamalarını ve veri madenciliği modellerinin elde edilmesinden sonra model değerlendirme ve görselleştirme gibi işlemleri gerçekleştirme imkanı sağladığı için tercih edilmiştir. Bütün bu aşamaları da CRISP-DM metodolojisi çerçevesinde bir bütünlük ve mantıksal akış düzeni içerisinde gerçekleştirmeye imkan vermektedir.

Mevcut çalışma için akademik versiyon kullanılmıştır. Bu versiyon, Temel (Base) modülüne ek olarak Bölümleme (Segmentation), Sınıflandırma (Classification) ve Birliktelik (Association) modüllerini içermektedir. Bu modüller içerisinde Karar Ağaçları, Kümeleme, Birliktelik Kuralları, Yapay Sinir Ağları, Zaman Serileri, Kohonen Ağları gibi teknikler için çeşitli işlem düğümleri (nodes) bulunmaktadır (SPSS, 2007a, s.2-7).

Clementine çeşitli biçimlerde olan ve farklı platformlarda bulunan veriye erişim imkanı sağlamaktadır. SPSS, SAS ve Excel veri dosyalarına, ayrıca csv, dat ve txt uzantılı dosya tiplerine erişerek bunlardan veri alabilmektedir. Bununla birlikte tüm ilişkisel veritabanı sunucularına ODBC yoluyla bağlanarak veriye erişilebilmektedir. Tezin uygulamasında Oracle MEDISYS ve MEDILIS veritabanları ve SQL Server 2005 veritabanı için ODBC veri kaynakları (data source) oluşturulmuştur. Şekil 3.1'de Clementine veritabanı bağlantıları arayüzü görülmektedir. Burada veri kaynakları kısmında ODBC veri kaynaklarının hepsi listelenmektedir ve gerekli veri kaynağı veya kaynakları seçilerek bağlantılar yapılmaktadır.



Şekil 3.1. Veritabanı Bağlantıları Arayüzü

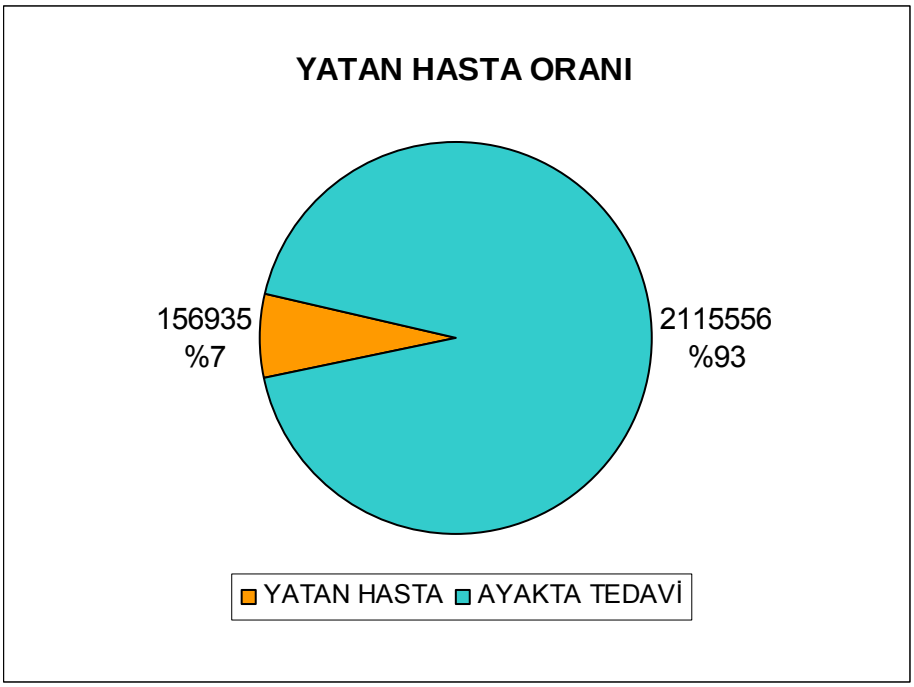
Bu çalışmada Oracle 9i ve SQL Server 2005 veritabanları ve buralarda yer alan tabloların yanı sıra, tablolardan elde edilen verilerin dönüştürülmesi sonucu hazırlanan verinin kaydedildiği csv, txt ve dat uzantılı veri dosyaları ile Microsoft Excel uygulaması da kullanılmıştır.

3.3. Tanımlayıcı Bulgular

3.3.1. Başvuran Hasta İstatistikleri

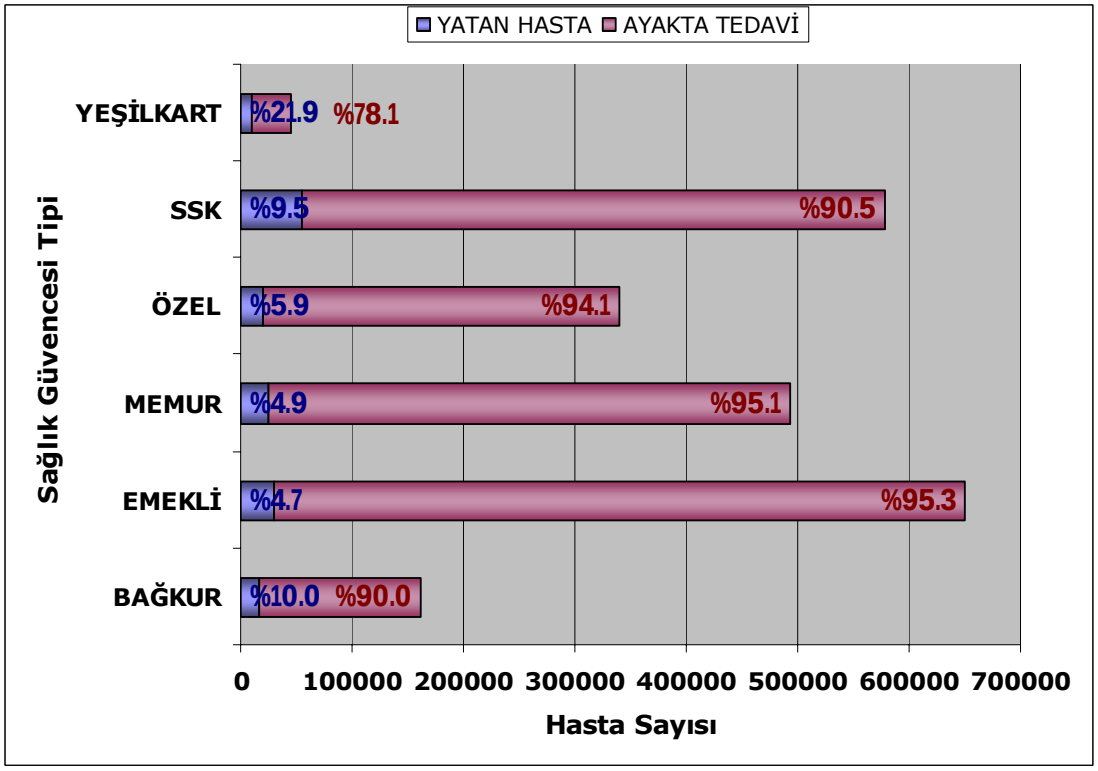
Bu bölümde frekans alma yöntemine göre ve veri madenciliğinde görselleştirme tekniği kullanılarak veritabanında kayıtlı bulunan hastalara ait bulgulara yer verilmiştir. Sonuçları elde edebilmek için ilgili değişkenlerin parametrelerine göre SQL cümleleri yazılarak veritabanından frekans değerleri elde edilmiştir.

Şekil 3.2'de Ocak 2005 ile Aralık 2008 dönemlerini kapsayan dört yıllık dönemde hastaneye başvuru yapan hastaların yatış yapma oranları verilmiştir. Hastanenin hasta profiline bu açıdan bakıldığında %7'sinin yatan hasta olduğu ve %93'ünün ayakta tedavi gören hastalar olduğu görülmektedir.



Şekil 3.2. Yatan Hasta ve Ayakta Tedavi Oranı

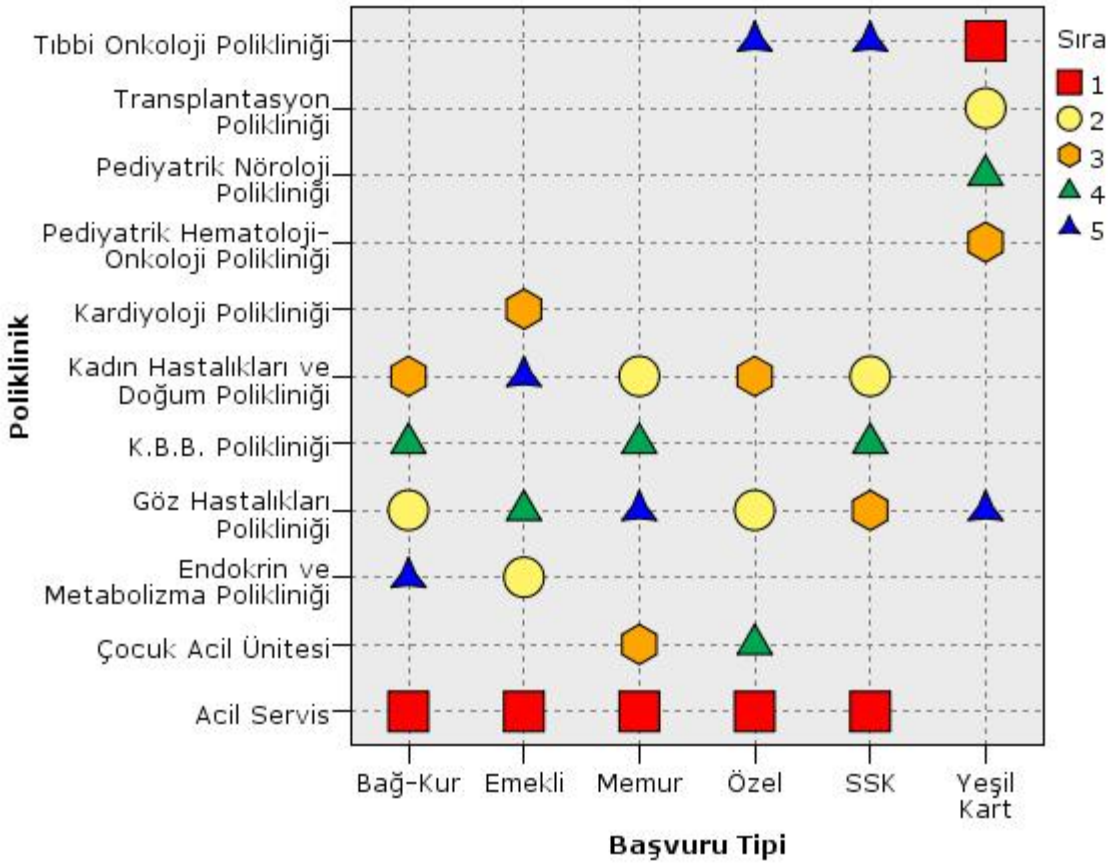
Şekil 3.3'de hastaneye yatış yapan hasta sayılarının sağlık güvencesi tiplerine göre dağılımları görülmektedir. Grafiğin üzerinde yer alan yüzdelik oranlar her sağlık güvencesi tipine göre kendi içerisinde yatan hasta ve ayakta tedavi oranlarını vermektedir.



Şekil 3.3. Hasta Başvurularının Sağlık Güvencesi Tiplerine Göre Dağılımı

Hastaneye en çok başvuru yapan grubun Emekliler (649.386 başvuru) ve SSK'lılar (578.112 başvuru) olduğu görülmektedir. Ancak yatış yapma oranlarına bakıldığında Yeşil Kart sahibi hastaların kendi grubu içinde yatış yapma oranlarının (%21,9) diğerlerine kıyasla oldukça yüksek olduğu görülmektedir. Yeşil Kart sahibi hastalardan sonra Bağ-Kur'lu ve SSK'lı hastaların yatış yapma oranlarının yüksekliği göze çarpmaktadır (sırasıyla %10,0 ve %9,5). Dolayısıyla Yeşil Kartlı hastaların diğerlerine göre daha fazla yataklı tedaviler için bu hastaneye geldikleri, Emekli ve Memurların ise oldukça yüksek oranda (sırasıyla %95,3 ve %95,1) ayakta yapılan tedaviler amaçlı başvurdukları söylenebilir. Sağlık güvencesi tiplerine göre başvuru sayısı tahminleri Hastane Yoğunluk Tahmini bölümünde yapılmıştır.

Hastaların sağlık güvencesi tiplerine göre başvuru sayısı olarak ilk beş sırayı oluşturan poliklinikler ve yatış yaptıkları klinikler araştırılmıştır. En çok başvuru yapılan poliklinikler ile ilgili sonuçlar Şekil 3.4'de gösterilmiştir.



Şekil 3.4. Sağlık Güvencesi Tiplerine Göre En Çok Başvuru Yapılan Beş Poliklinik

Şekil 3.4'de Bağ-Kur, Emekli, Memur, Özel ve SSK'lıların en çok Acil Servis'e başvurdukları görülmektedir (sırasıyla 11.981, 57.307, 42.206, 39.060 ve 44.634 başvuru).

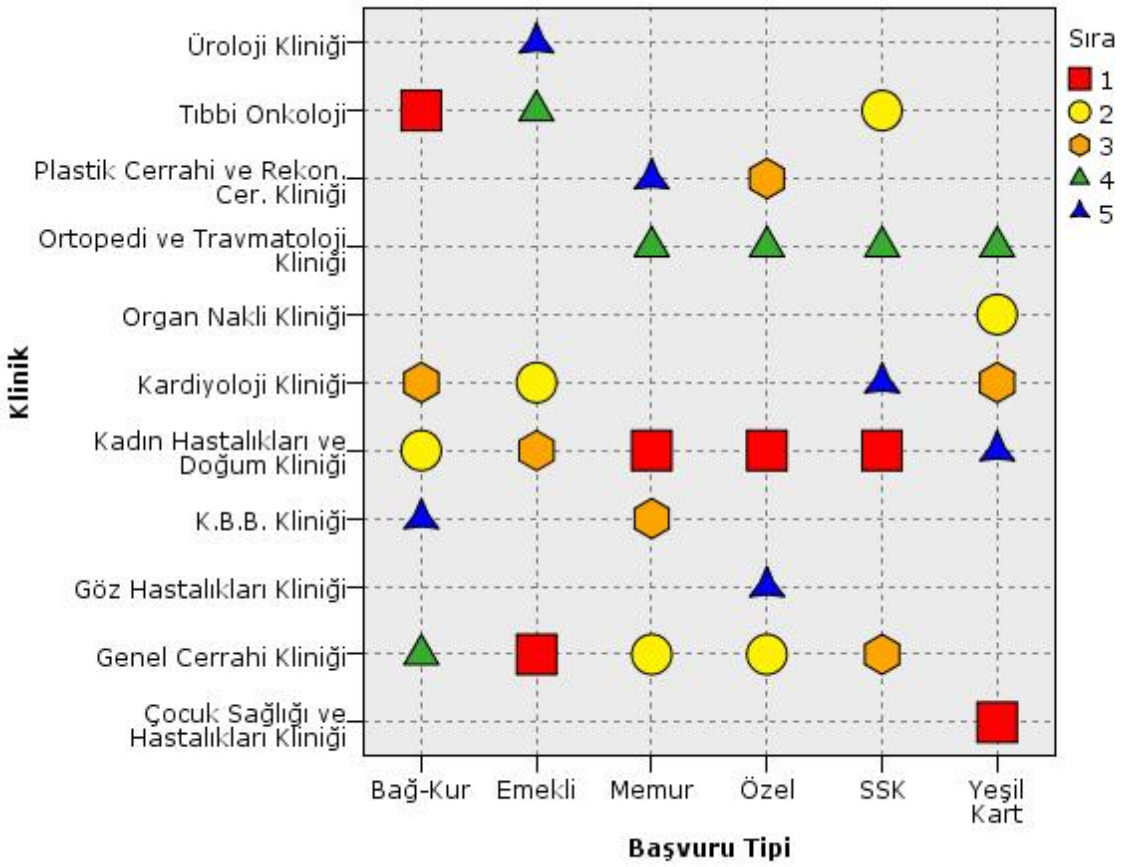
Ancak Yeşil Kartlı hastaların diğerlerinden farklı olarak en çok Tıbbi Onkoloji Polikliniğine başvuru yaptıkları görülmektedir (3.230 başvuru).

Hastaların ikinci ve üçüncü sırada en çok başvuru yaptıkları polikliniklere bakıldığında başvuru tipine göre ayrım daha belirginleşmektedir. Emekliler, Endokrin ve Metabolizma (42.887 başvuru) ile Kardiyoloji (42.368 başvuru) polikliniklerine ikinci ve üçüncü sırada başvuru yapmaktadırlar. Bağ-Kurlu ve Özel hastalar ise sırasıyla Göz Hastalıkları ile Kadın Hastalıkları ve Doğum Polikliniklerine başvuru yapmaktadırlar. SSK'lı ve Memur hastalar ise ikinci sırada Kadın Hastalıkları ve Doğum Polikliniğine, SSK'lılar üçüncü sırada Göz Hastalıkları Polikliniğine, Memurlar ise üçüncü sırada Çocuk Acil Ünitesine başvuru yapmaktadırlar. Yeşil Kartlı hastaların ise yine diğerlerinden farklı bir yapıda oldukları görülmektedir. Yeşil Kartlılar ikinci sırada Transplantasyon Polikliniğine başvuru yaparken üçüncü sırada Pediyatrik Hematoloji-Onkoloji Polikliniğine başvuru yaptıkları görülmektedir.

Şekil 3.4'e bir de poliklinikler açısından baktığımızda ise Göz Hastalıkları Polikliniğinin tüm başvuru tipleri için ilk beş sıra içerisinde olduğu görülmektedir. Kardiyoloji Polikliniği yalnız Emekli hastalar için ilk beş sıra içerisinde yer alırken, Transplantasyon, Pediyatrik Hematoloji-Onkoloji ve Pediyatrik Nöroloji Polikliniklerinin sadece Yeşil Katlı hasta başvurularında ilk beş sıra içerisinde yer aldığı görülmektedir.

Benzer bir durumu yatan hastalar açısından incelediğimizde, hastaların sağlık güvencesi tipine göre en çok yatış yaptığı ilk beş klinik Şekil 3.6'da gösterilmektedir. Burada poliklinik başvurularına nazaran daha fazla bir farklılaşma göze çarpmaktadır.

Şekil 3.5 incelendiğinde Memur, Özel ve SSK'lı hastaların en çok Kadın Hastalıkları ve Doğum Kliniğine yatış yaptıkları görülmektedir. Bunlar arasında Memur ve Özel tipte başvuran hastaların ikinci sırada Genel Cerrahi Kliniğinde yatış yaptıkları görülmektedir. SSK'lılar için ise Genel Cerrahi Kliniği üçüncü sırada gelirken, ikinci sırada Tıbbi Onkoloji yer almaktadır. Tıbbi Onkoloji Bağ-Kur'lu hastalar için ise birinci sırada gelmektedir. Yeşil Kart sahibi hastaların ise birinci sırada Çocuk Sağlığı ve Hastalıkları Kliniği'nin geldiği görülmektedir. Yeşil Kartlılar için ikinci sırada Organ Nakli Kliniği, üçüncü sırada ise Kardiyoloji Kliniği gelmektedir. Emeklilerin Kardiyoloji Kliniğine ikinci sırada Bağ-Kurluların ise üçüncü sırada yatış yaptıkları görülmektedir.



Şekil 3.5. Sağlık Güvencesi Tiplerine Göre En Çok Yatış Yapılan Beş Klinik

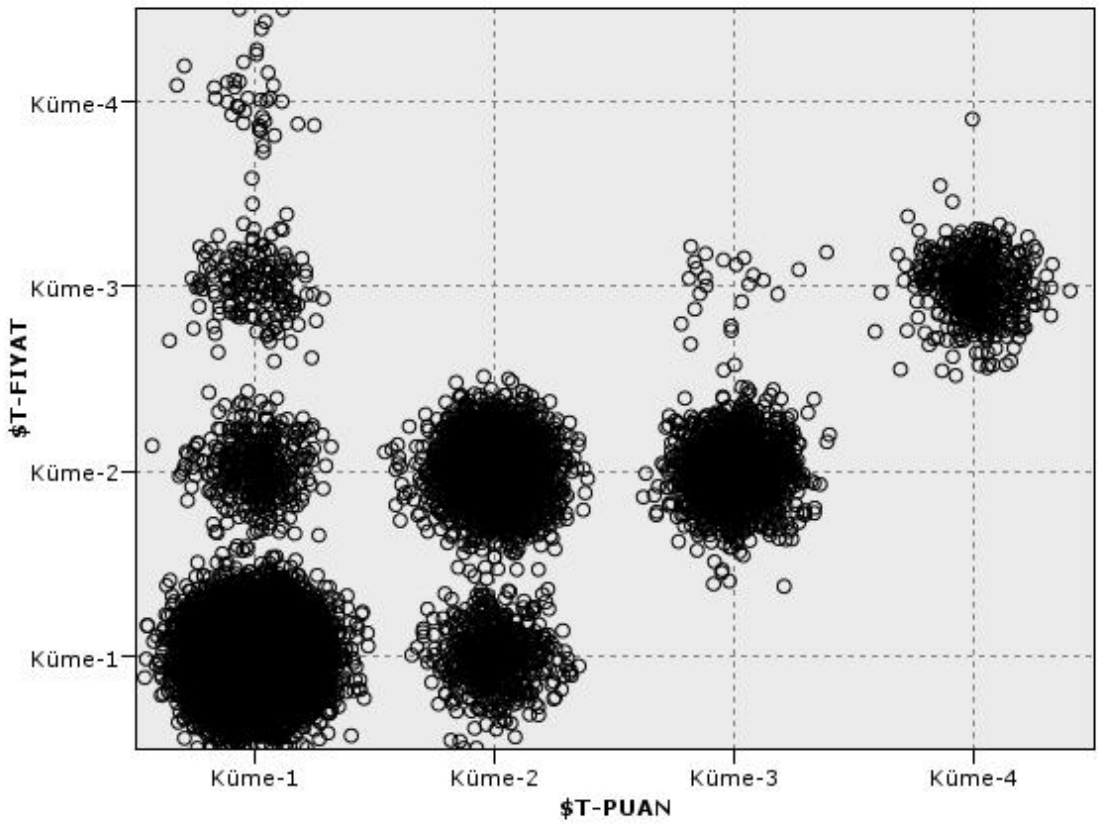
Şekil 3.5'e klinikler açısından baktığımızda ise Çocuk Sağlığı ve Hastalıkları Kliniği ile Organ Nakli Kliniğinin yalnızca Yeşil Kartlı hastalar açısından ilk beş sıra içerisinde yer aldığı dikkat çekmekte ve bu hastalar için yatış yapma sayısı açısından sırasıyla birinci ve ikinci sırada yer aldığı görülmektedir. Üroloji Kliniğinin sadece Emekliler için yatış yapmada ilk beş sırada yer aldığı, Göz Hastalıkları Kliniğinin ise sadece Özel tipteki başvurularda ilk beş klinik içerisinde yer aldığı görülmektedir. Kadın Hastalıkları ve Doğum Kliniğinin ise tüm başvuru tipleri için ilk beş sırada yer aldığı görülmektedir. Plastik Cerrahi ve Rekonstrüktif Cerrahi Kliniğinin, Memur ve Özel başvuru tipindeki hastalar için ilk beş içerisinde yer aldığı ve Özel tipteki başvurular için üçüncü sırada yer aldığı dikkat çekmektedir.

Polikliniklere yapılan başvurular ve kliniklerde yatan hastalara ait yatış sayıları, birimlere ve başvuru tiplerine göre incelenmiştir. Toplam hasta başvuru sayıları ilk otuz poliklinik için ve yatışı yapılan toplam hasta sayısı ilk otuz klinik için Ek-1(a) ve Ek-1(b)'de verilen grafiklerde sunulmuştur.

3.3.2. Hastanede Verilen Hizmetlerin Kümelenmesi

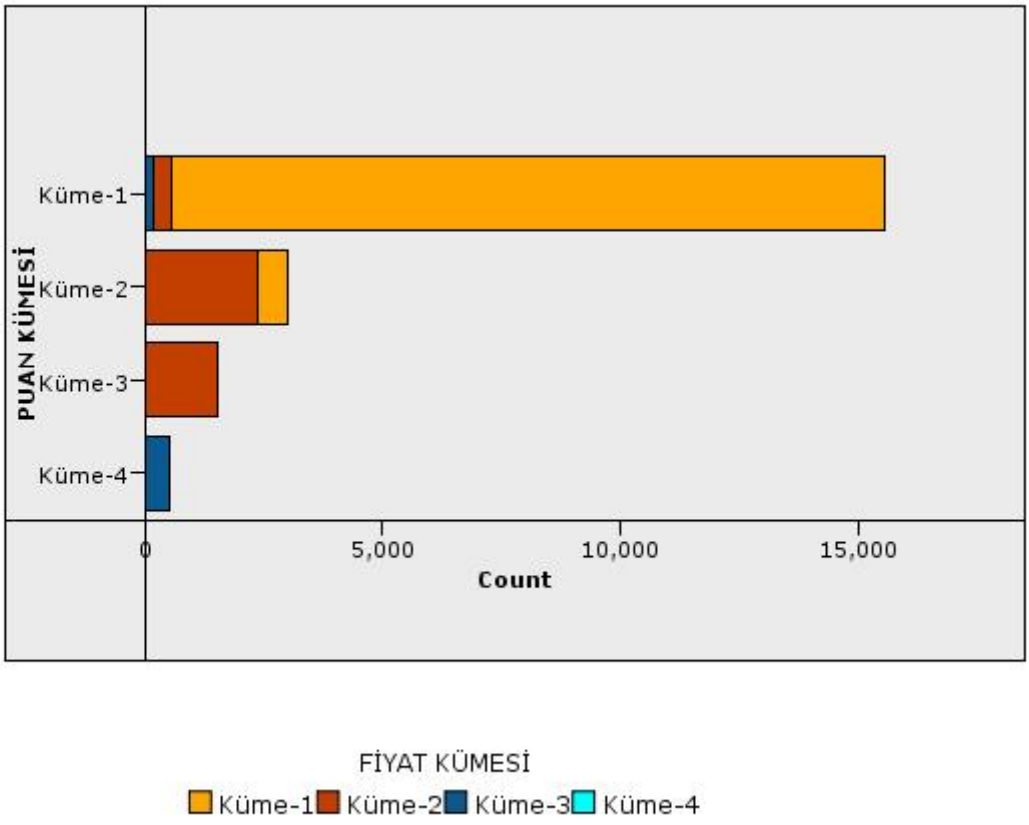
Hastanede verilen tüm hizmetler HIZMET tablosunda tutulmaktadır ve bu hizmetler Sağlık Bakanlığı kodlarına dayalı olarak detaylı tanımlanmıştır. Bu tabloda 28.956 adet farklı hizmet yer almaktadır. Bunlara örnek olarak Radyoloji'de çekilen grafiler, laboratuarda yapılan testler, kan merkezinde yapılan işlemler, Eczane biriminden kullanılan ilaçlar, yatan hastalar için çıkarılan yemekler, hemşirelik hizmetleri, konsültasyon hizmetleri ve daha birçok örnek verilebilir. Kısaca bu tabloda hastanede, hasta için yapılan küçük veya büyük tüm işler detaylı olarak yer almaktadır. Her hizmet, bir hizmet kodu (KHIZMET) ile tanımlanmıştır. Bunun yanı sıra bu tablodan hizmeti yapan birim (KBIRIM), sosyal güvenlik kuruluşlarınca ödenen resmi fiyatı (RESMI_FIYAT) ve Sağlık Bakanlığı tarafından belirlenen performans puanı (SB_PUAN) alanları kullanılmıştır.

Şekil 3.6'da hastanede verilen hizmetlerin resmi fiyatlarına ve performans puanlarına göre yapılan kümeleme analizi sonuçları gösterilmektedir. Bu analiz yapılırken iki aşamalı (Two-Step) kümeleme tekniği kullanılmıştır.



Şekil 3.6. Hastanede Verilen Hizmetlerin Sağlık Bakanlığı Performans Puanları ve Fiyatlarına Göre Kümelenmesi

Şekil 3.6'da hizmetlerin kümelere dağılımları görselleştirilmiştir. Burada yatay eksende performans puanına göre kümeler yer alırken, dikey eksende resmi fiyata göre kümeler yer almaktadır. Grafikte yer alan küme numaralarında 1. küme en düşük puan veya en düşük fiyatı göstermektedir. 4. küme ise en yüksek puan veya en yüksek fiyatı göstermektedir. Bekleneceği gibi hastane hizmetlerinin sayı olarak büyük bir kısmı en düşük puanlı ve en düşük fiyatlı kümelerde yer almıştır. Hizmetlerin sayı olarak hangi fiyat ve puan kümelerinde yer aldıkları Şekil 3.7'de daha açık olarak gösterilmiştir.



Şekil 3.7. Hizmetlerin Kümelere Dağılımı Grafiği

Şekil 3.7'de yatay eksende, hizmet sayıları, dikey eksende S.B. performans puanlarına göre kümeler gösterilmiştir. Renklere göre ise fiyat kümeleri gösterilmiştir. Görüldüğü gibi hastanede verilen hizmetlerin 15.000'den fazlası en düşük puan kümesinde (Küme-1) yer almaktadır. S.B. performans puanlarının yüksek olması verilen hizmetin daha yüksek nitelikli olduğunu göstermektedir. En yüksek puan kümesinde (Küme-4) yer alan hizmetlerin genellikle 3. fiyat kümesinde yer aldığı görülmektedir. 1. puan kümesi ve 3. fiyat kümesinde yer alan hizmetler ise genellikle yüksek fiyatlı malzeme veya ilaçları kapsamaktadır. Resmi fiyata ve S.B. performans puanına göre oluşan kümelere ait küme merkezleri, standart sapmaları ve her kümede yer alan hizmet sayıları Tablo 3.3'de verilmiştir.

Tablo 3.3. Kümeleme Sonuç Değerleri

Küme No.	Resmi Fiyat (TL)			S.B. Performans Puanı		
	Küme Merkezi	Standart Sapma	Hizmet Sayısı	Küme Merkezi	Standart Sapma	Hizmet Sayısı
Küme-1	40,18	53,49	21.817	57,08	71,81	16.592
Küme-2	393,48	147,41	5.945	442,73	93,25	3.883
Küme-3	1121,18	324,38	1.092	916,40	177,31	2.165
Küme-4	5866,24	4871,02	102	1808,63	431,39	746

Bir sonraki analizde incelenen konsültasyon hizmetlerinin tümü Fiyat ve Puan bakımından birinci kümede yer almaktadırlar. Konsültasyon hizmetlerinin fiyat ortalaması 6,28 TL'dir (En Düşük = 3,30 TL, En Yüksek = 17,80 TL, Standart Sapma = 1,89). Bu hizmetlerin Sağlık Bakanlığı Performans Puanı ortalaması ise 10,47'dir (En Düşük = 5, En Yüksek = 30, Standart Sapma = 3,21).

3.4. Konsültasyon Hizmetlerinin Birliktelik Kuralları ile Analizi

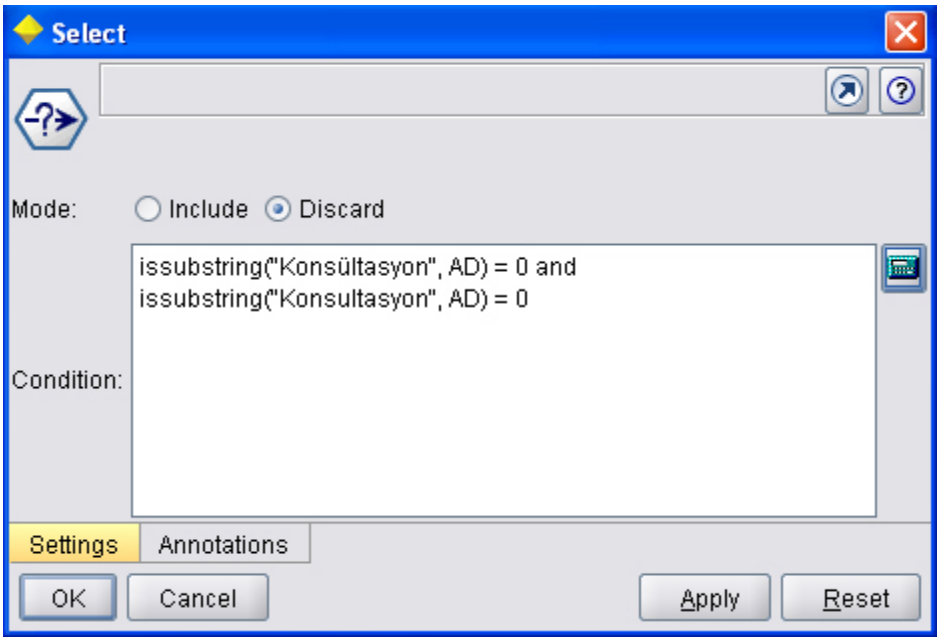
Birimler arasındaki konsültasyon hizmetlerinin analizi, Birliktelik Kuralları tekniği ve bu teknik için en yaygın olan Apriori algoritması kullanılarak, konsültasyonu talep eden birim (ISTEMBIRIM) merkezli ve konsültasyon hizmetini veren birim (YAPBIRIM) merkezli olmak üzere iki farklı biçimde yapılmıştır.

3.4.1. Verinin Hazırlanması

Analizlerin gerçekleştirilmesi için tüm hizmetlerin tanımlı olduğu HIZMET tablosu ve hasta başvurularında, hasta için verilen tüm hizmetlerin kaydedildiği işlemsel bir tablo olan HASTA_HAREKET tablosu kullanılmıştır. HASTA_HAREKET tablosunda hastaya verilen hizmetlerin istemini yapan birim (ISTEMBIRIM) ve bu hizmeti sağlayan birimin (YAPBIRIM) kayıtları tutulmaktadır. Dolayısıyla burada yalnızca konsültasyon hizmetlerine odaklanılarak birliktelik kuralları çıkartılması çalışması gerçekleştirilmiştir. Verinin hazırlanması ve analizinde aşağıdaki adımlar takip edilmiştir.

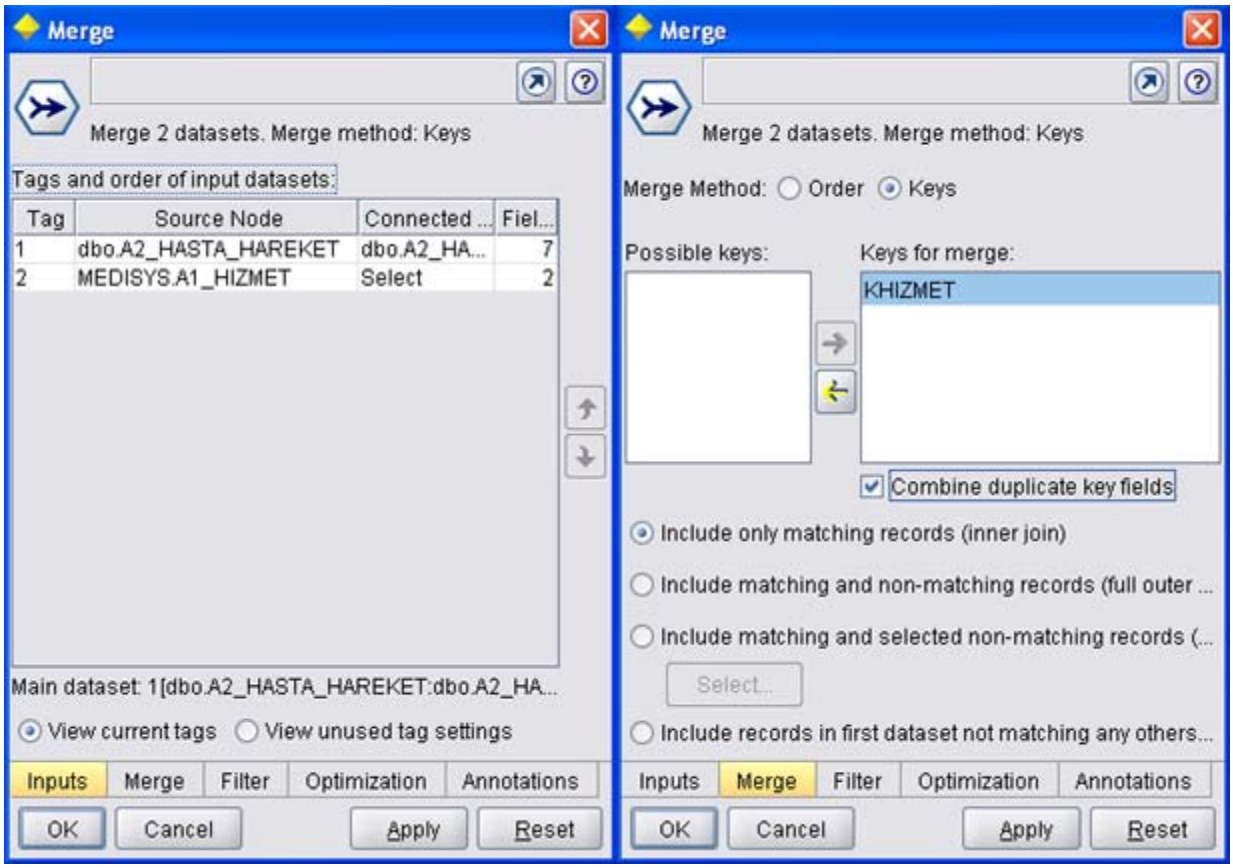
1. HIZMET tablosundan AD alanı içerisinde geçen "konsultasyon" veya "konsültasyon" kelimelerine göre seçim yapılarak, verilen tüm hizmetler arasından sadece konsültasyon hizmetleri belirlenmiştir. Bu işlem Clementine'de Select işlemcisi kullanılarak yapılmıştır.

Select işlemcisinin kullanımı Şekil 3.8'de gösterilmiştir. Burada issubstring fonksiyonu kullanılmıştır. Bu fonksiyon issubstring (SUBSTRING, STRING) şeklinde kullanılmakta ve belirtilen metin tipindeki değişken içerisinde alt metni taramaktadır. Eğer ilgili değişkende alt metin bulunursa sonuç olarak bu alt metni döndürmekte, eğer bulunamazsa değer olarak sıfır döndürülmektedir. Veri seti incelendiğinde konsültasyon hizmetlerinin "Konsultasyon" veya "Konsültasyon" şeklinde iki biçimde girildiği görülmüştür. Select işlemcisinde Şekil 3.8'de görüleceği gibi koşul AD değişkeni içerisinde "Konsultasyon" veya "Konsültasyon" terimleri bulunmayanlar olarak tanımlanmış (sıfıra eşitlenerek) ve Discard seçeneği seçilerek bunların dışarıda bırakılması sağlanmıştır.



Şekil 3.8. Konsültasyon Hizmetlerin Seçilmesi

2. Veritabanında her hasta başvurusunda verilen tüm hizmetlerin kaydedildiği HASTA_HAREKET tablosu ile sadece Konsültasyon hizmetlerinin seçildiği HIZMET tablosu Merge işlemcisi ile birbirine bağlanmıştır. Burada anahtar alan (Key) olarak KHIZMET kullanılmış ve iç birleştirme (iner join) yapılarak yeni oluşan kayıt setinde sadece eşleşen kayıtların, yani konsültasyon hizmetleri ile ilgili hareket kayıtlarının içerilmesi sağlanmıştır. Şekil 3.9'da bu işlemin Clementine'de Merge işlemcisi kullanılarak nasıl yapıldığı görülmektedir. Clementine'de Kaynak düğümleri (Source Nodes), tablolar aynı veritabanından olsa bile her bir tablo için ayrı ayrı oluşturulur. Şekilde sol tarafta yer alan Inputs sekmesinde bağlanacak tablolar listelenmektedir, sağ tarafta ise Merge sekmesi ve burada birleştirme işleminin nasıl yapılacağı görülmektedir.

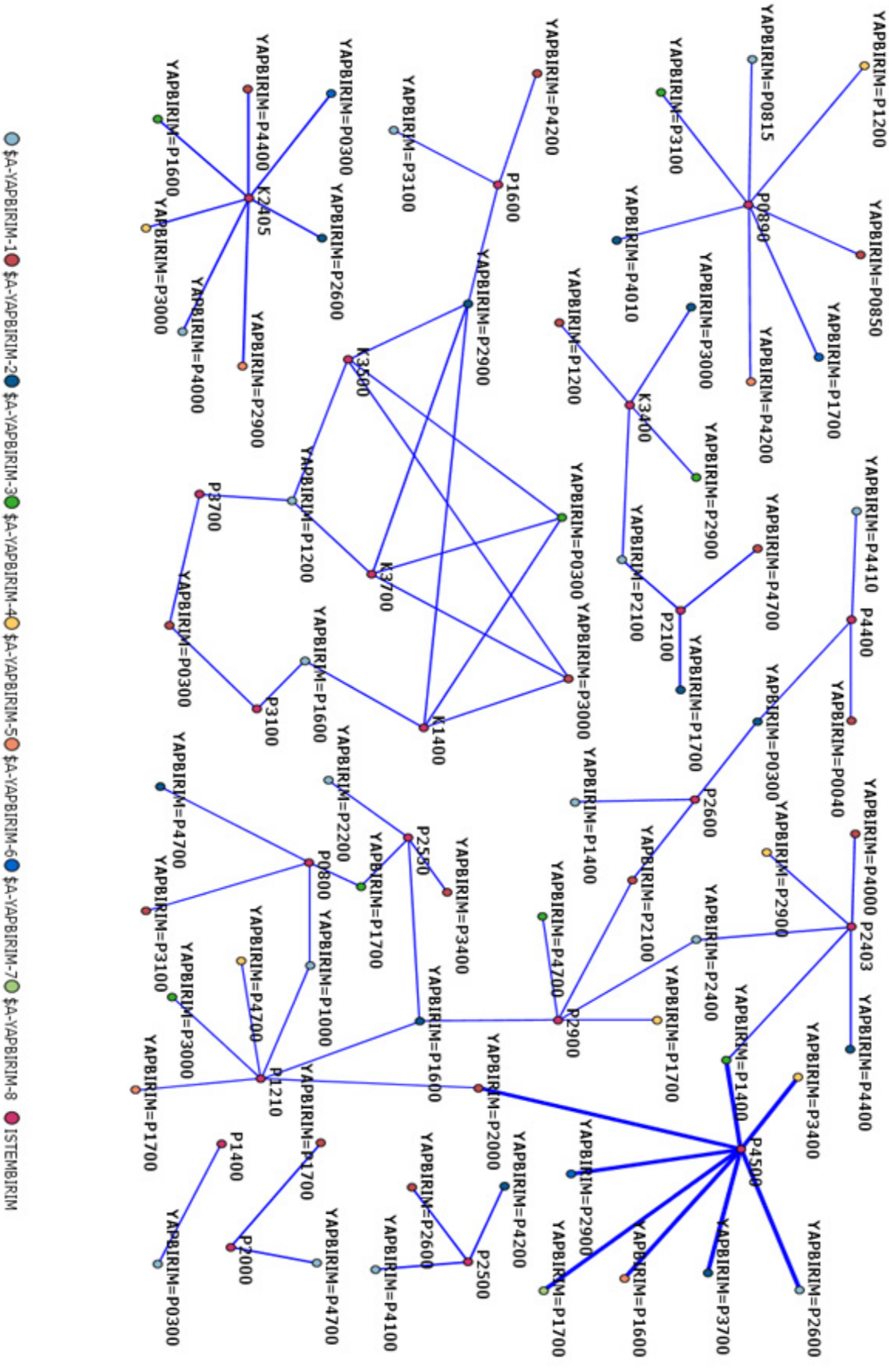


Şekil 3.9. HASTA_HAREKET ve HIZMET Tablolarının Birleştirilmesi

3. Veri seçimi ve düzenlenmesi işlemlerinin ardından Clementine modelleme işlemcileri arasında yer alan Apriori işlemcisi bağlanarak Apriori algoritması çalıştırılmıştır. Burada analizin durumuna göre öncül (antecedent) ve ardıl (consequent) öğeler HASTA_HAREKET tablosunda bulunan ISTEMBIRIM veya YAPBIRIM değişkenleri olarak seçilmiştir.

3.4.2. Konsültasyon Hizmetini İsteyen Birime Göre Birimler Arası İlişkiler

Birimler arası konsültasyon hizmetlerinin analizinde ilk olarak, istemi yapan birimlerin en çok hangi birimler ile ilişkide olduğu ve hangi birimlerden konsültasyon talep ettiklerinin belirlenmesi için birliktelik kuralları tekniği uygulanmıştır. Burada öncül öğe ISTEMBIRIM, ardıl öğe ise YAPBIRIM olarak tanımlanmıştır. Böylece hizmet hareketi kayıtlarında konsültasyonu isteyen birim merkezde olmak üzere hizmeti alma konusunda hangi birimlerle yoğun ilişkide olduğunun çıkarılması hedeflenmiştir. Şekil 3.10'da birliktelik kuralları analizi sonuçları bir ağ şeması şeklinde gösterilmiştir. Bu şema, elde edilen modelin Clementine grafik modülünde yer alan Web (ağ) grafiği işlemcisi kullanılarak görselleştirilmiştir.



Şekil 3.10. Birimler Arası Konsültasyon İstekleri Ağ Grafiği (İsteyen Birime Göre)

Apriori algoritmasının parametreleri olarak minimum destek %0,1, minimum güven %5,0 ve maksimum öncül öğe sayısı da 1 olarak belirlenmiştir. HIZMET tablosunda 112 adet tanımlanmış farklı konsültasyon hizmeti, kullanılan veri setinde 101 farklı YAPBIRIM ve 151 farklı ISTEMBIRIM olduğu için destek ve güven seviyeleri düşük tutulmuştur. Kullanılan veri seti 310.463 kayıttan oluşmaktadır. Destek düzeyi tüm işlemler içerisinde öncül ve ardıl öğelerin (ilgili ISTEMBIRIM ve YAPBIRIM öznitelik değerlerinin) birlikte bulunma olasılığı iken, tüm işlemler içerisinde %0,1'lik seviye, normalde örneğin 30 adet ana ürün kategorisinin incelendiği bir pazar sepeti analizinde düşük kalabilmekte, ancak bu veritabanı için uygun bir seviye olmaktadır. Bir kurala ait %5'lik minimum güven seviyesi ise, hizmet isteminde bulunan ve hizmeti veren iki birimi de içeren işlemlerin, ardıl öğe durumundaki konsültasyon hizmetini veren birimi içeren işlemler içerisinde %5 veya daha fazlası olduğu anlamına gelmektedir ki bunun da mevcut çalışma için uygun bir seviye olduğu düşünülmektedir. Maksimum öncül öğe (antecedent) sayısı ise 1 olarak belirlenmiştir, çünkü veritabanında tanımlanmış herhangi bir konsültasyon hizmeti için istemi yapan ve de hizmeti veren birim sayısı 1 adettir.

Şekil 3.10'da konsültasyon isteyen birimlerin yoğunlukla hangi birimlerden istemde bulundukları grafik olarak gösterilmiştir. Bu grafikte kırmızı renkte olan ve analizin öncül öğesi olan ISTEMBIRIM'ler sadece birim kodları ile görülmektedir. İlgili birim kodlarının hangi birimler olduğu Tablo 3.4'den görülebilir. Grafikte bir ISTEMBIRIM'in yoğun istemde bulunduğu en fazla 8 adet YAPBIRIM gösterilmiştir. Birimler arasındaki ilişkileri gösteren çizgilerin kalınlığı ise bağlı olduğu ISTEMBIRIM tarafından yapılan konsültasyon istemlerinin miktarının fazlalığını belirtmektedir.

Grafikte P4500 birim kodu ile görülen Acil Servis'in en yoğun konsültasyon talep eden birim olduğu çok açık bir şekilde görülebilmektedir. Bu Acil Servis birimi için beklenen bir durum olarak karşımıza çıkmaktadır. Acil Servisin arkasından İç Hastalıkları Polikliniği (P2000), Organ Nakli Kliniği (K2405), Tıbbi Onkoloji Polikliniği (P2500) ve Endokrin ve Metabolizma Polikliniği (P2100) gelmektedir. Bu birimlerin hepsinin konsültasyon talep sayıları birbirine çok yakın olmakla birlikte, Acil Servisin istem sayısının ¼'leri dolayındadır.

Tablo 3.4'de, belirtilen parametrelere göre elde edilen birliktelik kuralları birincil olarak en çok istemde bulunan öncül öğeye göre ikincil olarak ise güven değerine göre sıralanmış halde verilmiştir.

Tablo 3.4. İstem Yapan Birime Göre Birliktelik Kuralları

Öncül Öğe (Antecedent) (ISTEMBIRIM)	Ardıl Öğe (Consequent) (YAPBIRIM)	P(A) %	Güven (Confidence) %	Destek (Support) %	Kaldırma Oranı (Lift Ratio)
Acil Servis (P4500)	İç Hastalıkları Polikliniği (P2000)	13,074	14,028	1,834	4,637
	Ortopedi ve Travmatoloji Polikliniği (P3700)	13,074	12,904	1,687	4,123
	Kardiyoloji Polikliniği (P2900)	13,074	10,357	1,354	1,705
	Kadın Hastalıkları ve Doğum Polikliniği (P2600)	13,074	8,862	1,159	3,093
	Genel Cerrahi Polikliniği (P1400)	13,074	7,854	1,027	2,449
	Nöroloji Polikliniği (P3400)	13,074	6,989	0,914	1,968
	Göz Hastalıkları Polikliniği (P1700)	13,074	5,378	0,703	0,654
	Göğüs Hastalıkları Polikliniği (P1600)	13,074	5,134	0,671	1,146
İç Hastalıkları Polikliniği (P2000)	Beslenme ve Diyet Polikliniği (P4700)	3,805	31,801	1,210	5,447
	Göz Hastalıkları Polikliniği (P1700)	3,805	27,594	1,050	3,355
Organ Nakli Kliniği (K2405)	Üroloji Polikliniği (P4400)	3,566	16,602	0,592	6,406
	Psikiyatri Polikliniği (P4000)	3,566	15,012	0,535	5,881
	Kardiyoloji Polikliniği (P2900)	3,566	14,976	0,534	2,466
	Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	3,566	13,594	0,485	1,258
	Enfeksiyon Hastalıkları Polikliniği (P3000)	3,566	10,315	0,368	1,889
	Kadın Hastalıkları ve Doğum Polikliniği (P2600)	3,566	6,106	0,218	2,131
	Göğüs Hastalıkları Polikliniği (P1600)	3,566	5,429	0,194	1,211
Tıbbi Onkoloji Polikliniği (P2500)	Radyoloji (P4200)	3,506	43,891	1,539	11,132
	Radyasyon Onkolojisi Polikliniği (P4100)	3,506	16,893	0,592	14,175
	Kadın Hastalıkları ve Doğum Polikliniği (P2600)	3,506	6,807	0,239	2,376

Tablo 3.4. devamı

Öncül Öğe (Antecedent) (ISTEMBIRIM)	Ardıl Öğe (Consequent) (YAPBIRIM)	P(A) %	Güven (Confidence) %	Destek (Support) %	Kaldırma Oranı (Lift Ratio)
Endokrin ve Metabolizma Polikliniği (P2100)	Göz Hastalıkları Polikliniği (P1700)	3,385	39,848	1,349	4,845
	Beslenme ve Diyet Polikliniği (P4700)	3,385	18,849	0,638	3,229
	Endokrin ve Metabolizma Polikliniği (P2100)	3,385	17,783	0,602	5,516
Genel Cerrahi Kliniği (K1400)	Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	2,517	32,536	0,819	3,011
	Enfeksiyon Hastalıkları Polikliniği (P3000)	2,517	19,698	0,496	3,606
	Kardiyoloji Polikliniği (P2900)	2,517	9,484	0,239	1,561
	Göğüs Hastalıkları Polikliniği (P1600)	2,517	5,542	0,139	1,237
Ortopedi ve Travmatoloji Kliniği (K3700)	Fiziksel Tıp ve Rehabilitasyon Polikliniği (P1200)	2,461	25,834	0,636	6,834
	Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	2,461	23,466	0,578	2,172
	Enfeksiyon Hastalıkları Polikliniği (P3000)	2,461	7,263	0,179	1,330
	Kardiyoloji Polikliniği (P2900)	2,461	7,146	0,176	1,176
Transplantasyon Polikliniği (P2403)	Nefroloji Polikliniği (P2400)	2,405	23,246	0,559	15,323
	Genel Cerrahi Polikliniği (P1400)	2,405	15,252	0,367	4,756
	Üroloji Polikliniği (P4400)	2,405	9,989	0,240	3,854
	Kardiyoloji Polikliniği (P2900)	2,405	6,253	0,150	1,030
	Psikiyatri Polikliniği (P4000)	2,405	6,106	0,147	2,392
Kadın Hastalıkları ve Doğum Polikliniği (P2600)	Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	2,183	45,809	1,000	4,240
	Endokrin ve Metabolizma Polikliniği (P2100)	2,183	9,091	0,198	2,820
	Genel Cerrahi Polikliniği (P1400)	2,183	5,962	0,130	1,859
Kulak, Burun ve Boğaz Hastalıkları Polikliniği (P3100)	Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	2,177	55,674	1,212	5,153
	Göğüs Hastalıkları Polikliniği (P1600)	2,177	5,903	0,129	1,317

Tablo 3.4'de birimlerin yoğun konsültasyon isteminde bulundukları birimler güven, destek ve kaldırma oranı değerleriyle birlikte verilmiştir. Tabloda yer alan P(A) değeri ise ilgili konsültasyon isteminde bulunan birimin istemlerinin tüm konsültasyon istemleri içinde yüzde kaç olduğunu göstermektedir. Burada Acil Servis'in %13,074 ile en çok konsültasyon isteyen birim olduğu görülmektedir. Acil Servis'in en çok istemde bulunduğu birim ise İç Hastalıkları Polikliniği olarak görülmektedir. Bu birliktelik kuralının güven değeri %14,028'dir ve güven değerinin formülünü de göz önüne aldığımızda bu Acil Servis'in konsültasyon istemlerinin %14,028'ini İç Hastalıkları Polikliniği'nden istediği şeklinde yorumlanmaktadır. Bu kuralın destek değeri ise %1,834'tür. Destek değeri de Acil Servis'in, İç Hastalıları Polikliniği'nden talep ettiği konsültasyon hizmetlerinin hastanede yapılan tüm konsültasyon istemlerinin %1,834'ünü oluşturduğu anlamına gelmektedir. İlgili birliktelik kuralının kaldırma oranı (lift ratio) değeri 3,687'dir. Bu değer öncül öğe'nin olduğu durumların ardıl öğe'nin gerçekleşme olasılığını ne kadar arttırdığının göstergesi olarak kabul edilmektedir. Bu anlamda kaldırma oranı bir birliktelik kuralının önem derecesinin göstergelerinden biri olarak kabul edilebilir. Değerin 1'e yakın olduğu durumlarda kuralın önem derecesi düşük, 1'den küçük olduğu durumlarda ardıl öğenin gerçekleşme olasılığına negatif katkısı olduğu, 1'den büyük olduğu durumlarda ise bu olasılığa pozitif katkısı olduğu söylenebilir. Dolayısıyla "ISTEMBIRIM = Acil Servis $\Rightarrow$ YAPBIRIM = İç Hastalıkları Polikliniği" kuralı için 3,687 olan kaldırma oranı değeri bu kuralın önemini gösterirken aynı zamanda İç Hastalıkları Polikliniği'nin konsültasyon hizmeti verme olasılığını isteyen birimin Acil Servis olması durumunun yükselttiğini göstermektedir.

Acil Servis'in konsültasyon isteminde bulunduğu diğer birimleri, güven değerleri doğrultusunda yorumlayacak olursak, ikinci sırada en çok istemde bulunduğu birim Ortopedi ve Travmatoloji Polikliniği (%12,908), üçüncü sırada en fazla istemde bulunduğu birim Kardiyoloji Polikliniği (%10,357) ve birliktelik kuralları analizinde minimum güven değerini sağlayan son sırada en çok istemde bulunduğu birim ise Göğüs Hastalıkları Polikliniği (%5,134) olarak görülmektedir.

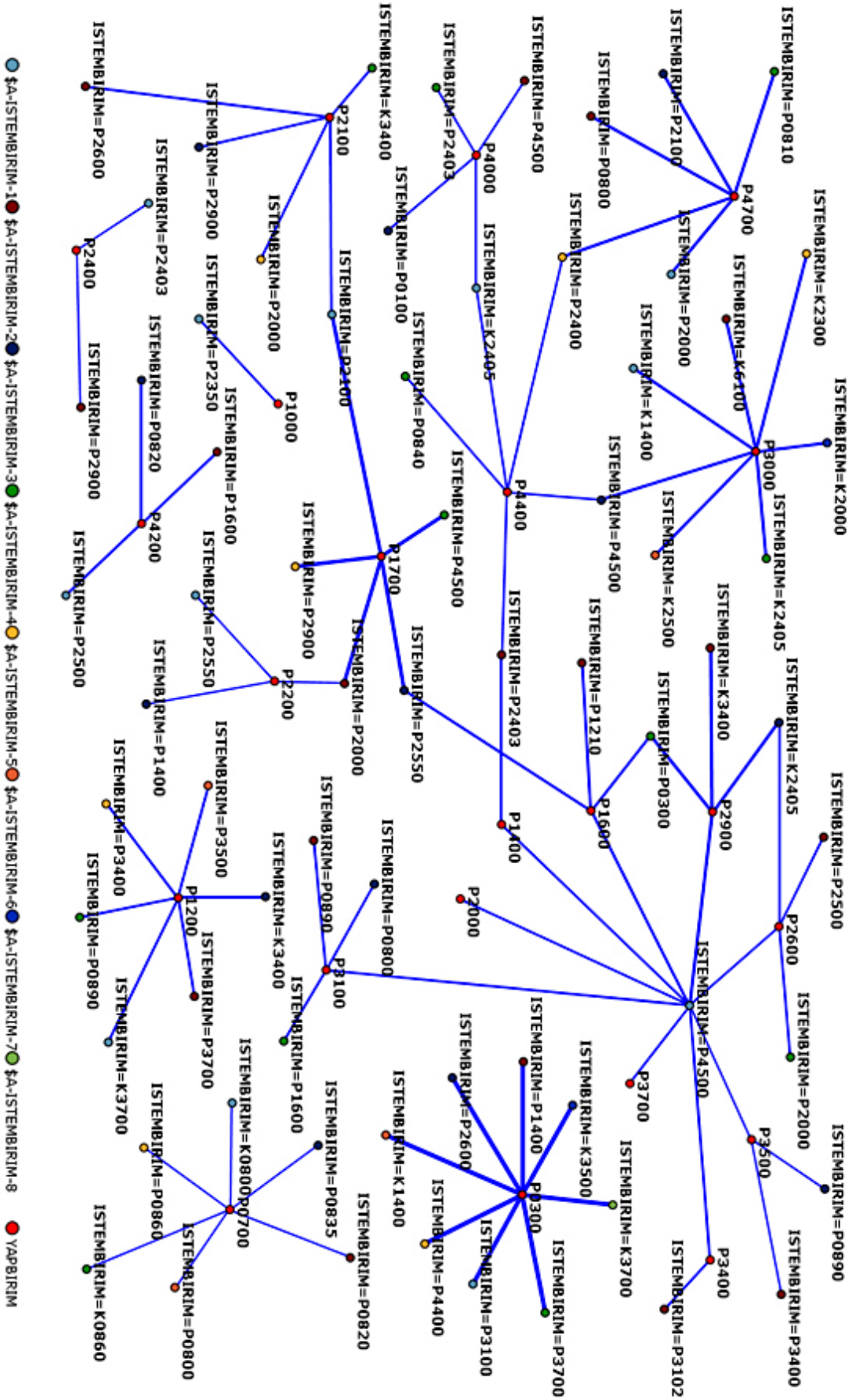
Hastanede ikinci sırada en çok konsültasyon isteminde bulunan birim olarak karşımıza çıkan İç Hastalıkları Polikliniği'ni incelediğimizde, bu birimin konsültasyon hizmetleri konusunda sadece iki birimde yoğunlaştığı görülmektedir. Bunlar, ilk sırada Beslenme ve Diyet Polikliniği (%31,801), ikinci sırada ise Göz Hastalıkları Polikliniği'dir (%27,594). Bu iki birimden sonra gelen birimlerin ise güven değerinin %5'ten küçük olduğu anlaşılmaktadır. Beslenme ve Diyet Polikliniği ile Göz Hastalıkları Polikliniği, İç Hastalıkları Polikliniği'nin

tüm konsültasyon istemlerinin %59,395'ine yanıt vermektedirler. Birliktelik kuralları analizi sonucu ortaya çıkan bu yoğun ilişki ve benzeri sonuçların tıbbi açıdan yorumlanmasının ise bu çalışmanın kapsamı dışında olması gerektiği düşünülmektedir. Ancak buradan elde edilen sonuçlara göre hastane yönetimi konsültasyon hizmetlerinin daha hızlı gerçekleştirilebilmesi için birbiri ile yoğun konsültasyon talep ve arzı içerisinde bulunan birimleri dikkate alarak gerekli tedbirleri alma yoluna gidebilir.

3.4.3. Konsültasyon Hizmetini Veren Birime Göre Birimler Arası İlişkiler

Birimler arası konsültasyon hizmetlerinin analizinde ikinci olarak, konsültasyon hizmetini veren birimlerin en çok hangi birimler ile ilişkide olduğu ve hangi birimlere konsültasyon hizmeti verdiklerinin belirlenmesi için birliktelik kuralları tekniği uygulanmıştır. Burada analizin konsültasyonu isteyen ve yapan birimlere göre ayrı ayrı yapılmasının nedeni farklı birimlerin hizmet alma ve hizmet verme konusunda yoğunluklarının farklı olmasıdır. Örneğin bir birimin konsültasyon istemi yoğun iken ve bu istemleri belirli birimlerden farklı yoğunluklarda alırken, başka bir birim konsültasyon hizmetini verme konusunda yoğunluk yaşıyor olabilir ve bu hizmetleri de yine belirli birimlere farklı yoğunluklarda veriyor olabilir. Dolayısıyla birimler arası konsültasyon hizmetleri birliktelik kuralları analizinin konsültasyonu hizmetini isteyen ve yapan birimlere göre iki farklı bakış açısıyla incelenmesinde yarar vardır.

Bu aşamada öncül öğe YAPBIRIM, ardıl öğe ise ISTEMBIRIM olarak tanımlanmıştır. Böylece hizmet hareketi kayıtlarında konsültasyon hizmetini veren birim merkezde olmak üzere hizmeti verme konusunda hangi birimlerle yoğun ilişkide olduğunun çıkarılması hedeflenmiştir. Şekil 3.11'de konsültasyon yapan birime göre birliktelik kuralları analizi sonuçları bir ağ şeması şeklinde gösterilmiştir.



Şekil 3.11. Birimler Arası Konsültasyon İstekleri Ağ Grafiği (Yapan Birime Göre)

Apriori algoritmasının parametreleri bu analizde de yine minimum güven değeri %5, minimum destek değeri %0,1 ve maksimum öncül öğe sayısı da 1 olarak belirlenmiştir. Şekil 3.11'de birimler arası konsültasyon istemlerinin ağ grafiği gösterilmiştir. Burada merkezde konsültasyon hizmetini veren birimler olmak üzere, bu birimlerden yoğun şekilde konsültasyon hizmeti isteyen birimler en fazla sekiz adet olacak şekilde bu birimlere bağlı şekilde gösterilmektedir.

Grafiği incelediğimizde Anestezi ve Reanimasyon Bilim Dalı Polikliniği'nin diğer birimlere en kalın çizgilerle bağlı olduğu görülmektedir. Dolayısıyla Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300) hastanede en fazla konsültasyon hizmeti veren birim olarak ortaya çıkmaktadır. Bunun nedeni ise Anestezi ve Reanimasyon Bilim Dalı Polikliniği'nde ameliyat olacak hastaların preoperatif dönemde değerlendirmeleri ve operasyona anestezi açısından hazırlanması süreçlerinde bu birimden konsültasyon alınmasıdır. Grafikte P0300 noktasına bağlanan birimler incelendiğinde bunların Genel Cerrahi Polikliniği (P1400), Kulak, Burun ve Boğaz Hastalıkları Polikliniği (P3100), Kadın Hastalıkları ve Doğum Polikliniği (P2600), Nöroşirurji Kliniği (K3500), vb. gibi cerrahi bilim dallarına ait poliklinik ve klinikler olduğu görülmektedir.

Grafikte ikinci en kalın çizgilerle diğer birimlere bağlı olan birimin Göz Hastalıkları Polikliniği (P1700) olduğu görülmektedir. Dolayısıyla Göz Hastalıkları Polikliniği hastanede en yoğun konsültasyon hizmeti veren ikinci birim olarak dikkat çekmektedir. Aynı zamanda grafikteki bağlantı sayıları da birimlerin konsültasyon hizmeti verme konusunda belirli az sayıda birimlere yoğunlaşmaları veya yoğunluklarının birçok farklı birime dağılmış olması konusunda da fikir vermektedir. Örneğin Anestezi ve Reanimasyon Bilim Dalı Polikliniği'nin sekiz birime bağlantısı görülmekte iken Göz Hastalıkları Polikliniği'nin beş birime bağlantısı görülmektedir. İç Hastalıkları Polikliniği (P2000) ve Ortopedi ve Travmatoloji Polikliniği (P3700) gibi birimlerin ise sadece bir birime yoğunluklu olarak konsültasyon hizmeti verdikleri görülmektedir. Bu birimlerin konsültasyon hizmeti alma konusunda ise birden fazla birimle yoğunluklu ilişkide olduğunu bir önceki analize göre söyleyebiliriz.

Grafikte bir başka dikkat çeken nokta ise analizin öncül öğesi olmamasına rağmen Acil Servis'e (P4500) birçok bağlantının olmasıdır. Bu durum Acil Servis'e çok sayıda birimin konsültasyon hizmeti verdiğini ortaya koymaktadır. Grafikte eşik değeri aşan dokuz adet birimin Acil Servis'e konsültasyon verdiği görülmektedir.

Tablo 3.5. Konsültasyon Hizmetini Yapan Birime Göre Birliktelik Kuralları

Öncül Öğe (Antecedent) (YAPBIRIM)	**Ardıl Öğe (Consequent) (ISTEMBIRIM)**	**P(A) %**	**Güven % (Confidence)**	**Destek % (Support)**	**Kaldırma Oranı (Lift Ratio)**
Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	Kulak, Burun ve Boğaz Hastalıkları Polikliniği (P3100)	10,804	11,218	1,212	5,153
	Genel Cerrahi Polikliniği (P1400)	10,804	9,907	1,070	4,561
	Kadın Hastalıkları ve Doğum Polikliniği (P2600)	10,804	9,254	1,000	4,240
	Ortopedi ve Travmatoloji Polikliniği (P3700)	10,804	8,166	0,882	3,840
	Üroloji Polikliniği (P4400)	10,804	7,912	0,855	5,248
	Genel Cerrahi Kliniği (K1400)	10,804	7,578	0,819	3,011
	Nöroşirurji Kliniği (K3500)	10,804	6,881	0,743	3,370
	Ortopedi ve Travmatoloji Kliniği (K3700)	10,804	5,345	0,578	2,172
Göz Hastalıkları Polikliniği (P1700)	Endokrin ve Metabolizma Polikliniği (P2100)	8,224	16,402	1,349	4,845
	İç Hastalıkları Polikliniği (P2000)	8,224	12,768	1,050	3,355
	Romatoloji Polikliniği (İç Hastalıkları) (P2550)	8,224	8,902	0,732	4,540
	Acil Servis (P4500)	8,224	8,550	0,703	0,654
	Kardiyoloji Polikliniği (P2900)	8,224	6,791	0,559	4,372
Kardiyoloji Polikliniği (P2900)	Acil Servis (P4500)	6,074	22,293	1,354	1,705
	Nöroloji Kliniği (K3400)	6,074	8,951	0,544	4,729
	Organ Nakli Kliniği (K2405)	6,074	8,792	0,534	2,466
	Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	6,074	8,066	0,490	8,629
Beslenme ve Diyet Polikliniği (P4700)	İç Hastalıkları Polikliniği (P2000)	5,838	20,728	1,210	5,447
	Çocuk Sağlığı ve Hastalıkları Polikliniği (P0800)	5,838	13,523	0,789	7,540
	Endokrin ve Metabolizma Polikliniği (P2100)	5,838	10,930	0,638	3,229
	Pediyatrik Endokrinoloji Polikliniği (P0810)	5,838	7,068	0,413	6,990
	Nefroloji Polikliniği (P2400)	5,838	5,065	0,296	3,995

Tablo 3.5. devamı

Öncül Öğe (Antecedent) (ISTEMBIRIM)	Ardıl Öğe (Consequent) (YAPBIRIM)	P(A) %	Güven % (Confidence)	Destek % (Support)	Kaldırma Oranı (Lift Ratio)
Enfeksiyon Hastalıkları Polikliniği (P3000)	Genel Cerrahi Kliniği (K1400)	5,462	9,076	0,496	3,606
	Anestezi Yoğun Bakım Kliniği (B Blok) (K6100)	5,462	8,905	0,486	8,337
	Acil Servis (P4500)	5,462	8,769	0,479	0,671
	Organ Nakli Kliniği (K2405)	5,462	6,735	0,368	1,889
	Hematoloji Kliniği (K2300)	5,462	6,505	0,355	7,419
	Tıbbi Onkoloji Kliniği (K2500)	5,462	5,496	0,300	5,902
	İç Hastalıkları Kliniği (K2000)	5,462	5,237	0,286	4,399
Göğüs Hastalıkları Polikliniği (P1600)	Acil Servis (P4500)	4,481	14,979	0,671	1,146
	Romatoloji Polikliniği (FTR) (P1210)	4,481	8,496	0,381	5,224
	Romatoloji Polikliniği (İç Hastalıkları) (P2550)	4,481	6,886	0,309	3,511
	Anestezi ve Reanimasyon Bilim Dalı Polikliniği (P0300)	4,481	5,757	0,258	6,159
Radyoloji (P4200)	Tıbbi Onkoloji Polikliniği (P2500)	3,943	39,033	1,539	11,132
	Göğüs Hastalıkları Polikliniği (P1600)	3,943	20,080	0,792	12,401
	Pediyatrik Hematoloji-Onkoloji Polikliniği (P0820)	3,943	5,653	0,223	4,237
Fiziksel Tıp ve Rehabilitasyon Polikliniği (P1200)	Ortopedi ve Travmatoloji Kliniği (K3700)	3,780	16,820	0,636	6,834
	Ortopedi ve Travmatoloji Polikliniği (P3700)	3,780	16,036	0,606	7,541
	Nöroloji Kliniği (K3400)	3,780	7,524	0,284	3,975
	Pediyatrik Nöroloji Polikliniği (P0890)	3,780	6,288	0,238	3,385
	Nöroloji Polikliniği (P3400)	3,780	6,178	0,234	4,241
	Nöroşirurji Polikliniği (P3500)	3,780	5,496	0,208	11,857
Nöroloji Polikliniği (P3400)	Acil Servis (P4500)	3,551	25,730	0,914	1,968
	KBB Vertigo Polikliniği (P3102)	3,551	5,931	0,211	10,614

Tablo 3.5'de konsültasyon hizmetini yapan birime göre elde edilen birliktelik kuralları, birincil olarak en çok istemde bulunan öncül öğeye göre ikincil olarak ise güven değerine göre sıralanmış halde verilmiştir. Öncül öğe için P(A) değerlerine baktığımızda hastanede yapılan konsültasyon hizmetlerinin %10,804'ünü Anestezi ve Reanimasyon Bilim Dalı Polikliniği'nin verdiğini, ondan sonra da sırasıyla Göz Hastalıkları Polikliniği (%8,224), Kardiyoloji Polikliniği (%6,074), Beslenme ve Diyet Polikliniği (%5,838) ve Enfeksiyon Hastalıkları Polikliniklerinin (%5,462) geldiği görülmektedir. Bir önceki konsültasyon isteyen birimlere göre yapılan analiz ile kıyasladığımızda, burada bir birimin aşırı baskın olduğu bir durumun göze çarpmadığı görülmektedir.

Tablo 3.5'de yer alan kuralları güven değerlerine göre yorumlayacak olursak Anestezi ve Reanimasyon Bilim Dalı Polikliniği'nin en çok konsültasyon hizmeti verdiği birim Kulak, Burun ve Boğaz Hastalıkları Polikliniği olarak görülmektedir (%11,218). Bundan sonra da sırasıyla Genel Cerrahi Polikliniği (%9,907) ve Kadın Hastalıkları ve Doğum Polikliniği (%9,254) gelmektedir. Kardiyoloji Polikliniği'nin yoğun olarak Acil Servis'e (%22,293) konsültasyon hizmeti verdiği görülürken, Beslenme ve Diyet Polikliniği ise yoğun olarak İç Hastalıkları Polikliniği'ne (%20,728) konsültasyon hizmeti verdiği görülmektedir. Benzer bir biçimde Göğüs Hastalıları Polikliniği ve Nöroloji Polikliniği de en yoğun olarak Acil Servis'e konsültasyon vermektedirler (sırasıyla %14,979 ve %25,730).

Radyoloji Polikliniği'nin konsültasyon hizmetleri konusunda iki birime yoğun hizmet verdiği ve bunların toplamın %59,113'ünü oluşturduğu görülmektedir. Bu birimler sırasıyla Tıbbi Onkoloji Polikliniği (%39,033) ve Göğüs Hastalıkları Polikliniği'dir (%20,080). Fiziksel Tıp ve Rehabilitasyon Polikliniği ise yoğun olarak Ortopedi ve Travmatoloji Kliniği (%16,820) ve Ortopedi ve Travmatoloji Polikliniği (%16,036) birimlerine konsültasyon hizmeti vermektedir. Bu iki birimin toplamı ise Fiziksel Tıp ve Rehabilitasyon Polikliniği'nin verdiği konsültasyonların %32,836'lık kısmını oluşturmaktadır.

Burada dikkat çekebilecek bir başka konu ise oldukça yoğun konsültasyon istemi bulunan Acil Servis birimine yoğun konsültasyon veren birimlerden bazılarının oldukça az (%5 güven değerinin altında) konsültasyon hizmeti veriyor olmasıdır. Bu birimler Anestezi ve Reanimasyon Bilim Dalı Polikliniği, Beslenme ve Diyet Polikliniği, Radyoloji ve Fiziksel Tıp ve Rehabilitasyon Polikliniği olarak sıralanabilir. Bu dört birim hastanede verilen konsültasyon hizmetleri toplamının %24,365'ini vermektedirler.

3.5. İleriye Yönelik Hastane Yoğunluk Tahmini Analizleri

Yoğunluk tahmini analizleri hastaneye yapılan toplam aylık başvuru sayıları dikkate alınarak gelecek aylarda gerçekleşecek olan başvuru yoğunluklarının tahmin edilmesi amacıyla Zaman Serileri (Üstel Düzgünleştirme ve ARIMA) ve Yapay Sinir Ağları (Geri Yayılım Algoritması ile İleri Beslemeli Yapay Sinir Ağı) yöntemleri kullanılarak gerçekleştirilmiştir.

3.5.1. Verinin Hazırlanması

Veri setini hazırlamak amacıyla HASTA_KABUL tablosunda yer alan 1 Ocak 2001 ve sonrasında gerçekleşmiş tüm kayıtlar ait oldukları yıl ve aya göre hesaplanmıştır. Zaman serisi analizleri burada toplam hasta sayısı yoğunluğunu tahmin etmek üzere gerçekleştirilmiştir, ancak bu analizler ayrıca sosyal güvence tiplerine ve birimlere göre hasta yoğunluklarının tahmin edilmesi şeklinde de gerçekleştirilebilmektedir.

2001 yılından başlamak üzere aylara göre toplam hasta sayısını elde etmek için TOAD for ORACLE SQL Editöründe aşağıdaki SQL ifadesiyle veri seti elde edilmiştir:

```
SELECT SUBSTR(KHASTAKABUL,1,6), COUNT
FROM HASTA_KABUL
WHERE KHASTAKABUL > '200100'
GROUP BY SUBSTR(KHASTAKABUL,1,6);
```

Hastaların sosyal güvence tiplerine ve aylara göre hasta sayılarının elde edilmesi için ise aşağıdaki SQL ifadesi kullanılmıştır. Bu ifade sonucunda aynı satırda kayıtlar yılın ayı, sosyal güvence tipi ve hasta sayısı olarak gelmektedir. Bu veri seti daha sonra Clementine'de sosyal güvence tipleri sütunları oluşturacak şekilde düzenlenmiştir.

```
SELECT SUBSTR(KHASTAKABUL,1,6), KABULTIPI, COUNT
 FROM HASTA_KABUL
 WHERE KHASTAKABUL > '200100'
 GROUP BY SUBSTR(KHASTAKABUL,1,6), KABULTIPI;
```

Birimlere göre başvuru sayılarını elde etmek için ise aşağıdaki SQL ifadesi kullanılmış ve birden çok birim için gelen kayıtlar Clementine'de düzenlenmiştir. Aşağıdaki örnekte Acil Servis ve K.B.B. Hastalıkları Kliniği için veri setleri elde edilmiştir.

```
SELECT SUBSTR(KHASTAKABUL,1,6), KKABULBIRIM,
 GET_SERVICE_NAME(KKABULBIRIM), COUNT
 FROM HASTA_KABUL
 WHERE KHASTAKABUL > '200100' AND KKABULBIRIM IN ('P4500',
 'K3100')
 GROUP BY SUBSTR(KHASTAKABUL,1,6), KKABULBIRIM;
```

Bu işlemlerden sonra 2001 ile 2008 yılları arasında bu yılları da kapsayan 96 aylık eğitim veri seti elde edilmiştir. Zaman serisi analizlerinde genellikle veri noktası sayısının %10'u kadar ileriye dönük kestirim gerçekleştirilmektedir. Ocak 2009 ile Eylül 2009 arasında 9 aylık veri seti ise test veri seti olarak kullanılmak üzere veritabanından çekilmiştir.

3.5.2. Toplam Hasta Sayısına Göre Yoğunluk Tahmini

Bu kısımda gelecek aylardaki toplam hasta sayısını tahmin etmek amacıyla 96 aylık veri kullanılarak (1) üstel düzgünleştirme yöntemleri ile geliştirilen zaman serisi modelleri, (2) ARIMA yöntemleri ile geliştirilen zaman serisi modelleri ve (3) yapay sinir ağları ile geliştirilen tahmin modelleri incelenmiştir. Her yöntemin modelleri önce kendi içinde kıyaslanmış, daha sonra da her yöntemin en iyi modelleri birbirleriyle kıyaslanmıştır.

3.5.2.1. Üstel Düzgünleştirme Modelleri

Bu aşamada Clementine Zaman Serileri Modelleme aracı kullanılarak; Winters Additive, Winters Multiplicative, Simple Seasonal, Damped Trend, Holt's Linear Trend ve Brown's Linear Trend üstel düzgünleştirme yöntemleriyle hasta sayısı yoğunluğu gelecek 9 ay için tahmin edilmiş ve aynı zamanda üstel düzgünleştirme yöntemleri ile eğitilmiş zaman serisi veri madenciliği modelleri birbirleriyle kıyaslanmıştır.

Tablo 3.6'da 2009 yılı Eylül ayına kadar farklı üstel düzgünleştirme yöntemleri ile yapılmış olan 9 aylık toplam hasta sayısı tahminleri yer almaktadır. Tabloda ayrıca bu aylar için gerçekleşen hasta sayıları da verilmiştir. Böylece her yöntemin tahmin rakamları ile gerçek değerler karşılaştırılmaktadır. Özellikle 9. aya (Eylül) gelindiğinde dahi bazı modellerin tahminlerinin gerçekleşen değerlere oldukça yakın olduğu görülmektedir. Bu tabloda yer alan değerlerin grafikleri Ek-2'de her model için gerçek değerlerle karşılaştırmalı bir şekilde verilmiştir.

Tablo 3.6. Üstel Düzgünleştirme Yöntemleri 2009 Yılı 9 Aylık Hasta Sayısı Tahminleri

2009 Yılı (Aylar)	Gerçekleşen Toplam Hasta Sayısı	Üstel Düzgünleştirme Model Tahminleri (Toplam Hasta Sayısı)					
		Winters Additive	Winters Multiplicative	Simple Seasonal	Damped Trend	Holt's Linear Trend	Brown's Linear Trend
Ocak	64.615	61.737	65.002	63.305	57.432	57.495	56.414
Şubat	58.330	58.548	59.632	57.622	57.718	57.857	56.106
Mart	66.385	63.719	63.551	61.084	58.004	58.220	55.799
Nisan	64.946	62.003	62.012	59.036	58.289	58.582	55.492
Mayıs	62.632	63.937	62.886	59.463	58.574	58.944	55.185
Haziran	67.427	61.685	61.319	57.597	58.858	59.307	54.877
Temmuz	64.100	59.881	60.456	56.649	59.142	59.669	54.570
Ağustos	59.534	57.779	55.938	52.761	59.426	60.031	54.263
Eylül	59.310	59.826	56.620	53.760	59.709	60.393	53.955

Üstel düzgünleştirme yöntemleri ile geliştirilen modeller arasında hangi modelin en iyi kestirimci olduğunu anlamak için, Tablo 3.7'de geliştirilen modellerin Uyum İyiliği (Goodness of Fit) Kriterleri verilmiştir.

Burada modellerin uyum iyiliği kriterleri birbirleri ile karşılaştırmalı bir şekilde değerlendirilmektedir. R^2, yaygın olarak bilinen bir ölçüdür ve doğrusal modelin uyum iyiliği ölçütüdür, çoğunlukla determinasyon katsayısı olarak da isimlendirilmektedir. 0 ile 1 arasında değişmekte ve küçük değerler modelin dataya uyumunun iyi olmadığını göstermektedir. Durağan R^2 ise modelin durağan kısmınıyla temel modeli karşılaştıran bir ölçüdür. Bir eğilim (trend) veya mevsimsel bir örüntü olduğu durumda tercih edilmektedir. RMSE, hata karelerinin ortalamasının kareköküdür. Bağımlı serilerin model tarafından kestirilen seviyeden ne kadar farklı olduğunu ifade etmek için kullanılmaktadır. Küçük değerleri model kestirimlerinin daha iyi olduğunu göstermektedir. MAPE, ortalama mutlak yüzde hatayı göstermektedir. Serilerin birimlerinden bağımsızdır, dolayısıyla farklı serilerin karşılaştırılmasında da kullanılabilmektedir. MAE, ortalama mutlak hatayı göstermektedir ve serilerin kendi birimleriyle ifade edilmektedir. MaxAPE, en yüksek mutlak yüzde hata ölçüsüdür. Tahmin edilen değerler arasında gerçekleşen en yüksek hatayı gösterir, yüzde

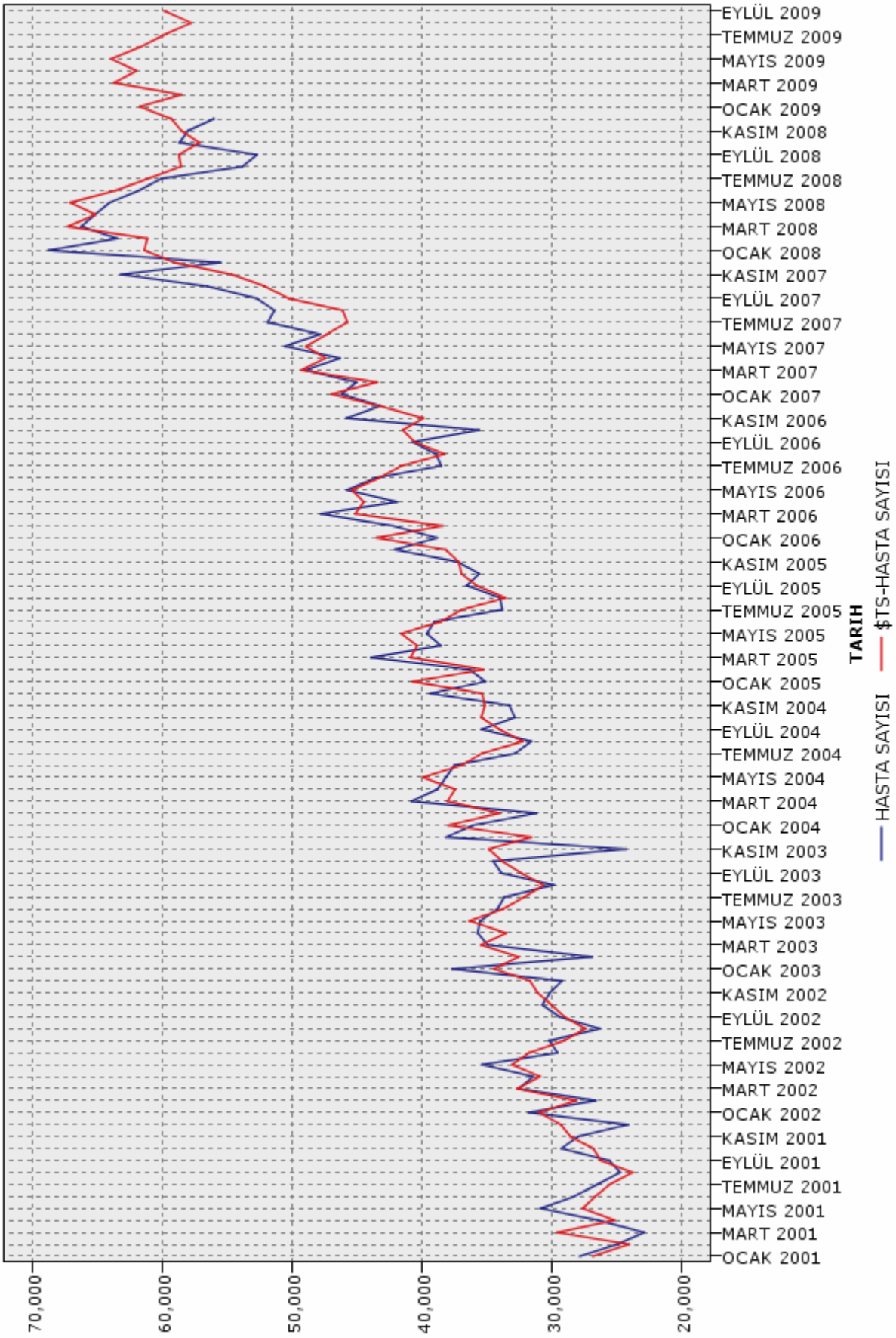
olarak ifade edilir dolayısıyla birimden bağımsızdır. Tahminler arasında en kötü senaryo uygulamaları için kullanılabilen bir ölçüdür. MaxAE, en yüksek mutlak hatayı göstermektedir ve bağımlı seri ile aynı birimde ifade edilmektedir. Norm. BIC, normalize Bayesyan bilgi kriteri, modelin toplam uyumunun genel ölçüsüdür. Bu ölçü, aynı seriler söz konusu olduğunda farklı modeller arasında karşılaştırma yapmak için kullanılmaktadır ve düşük değerler daha iyi bir modeli göstermektedir (SPSS, 2007b, s. 675-676).

Tablo 3.7. Üstel Düzgünleştirme Modellerinin Karşılaştırılması

	Uyum İyiliği Kriterleri							
Model	Durağan R^2	R^2	RMSE	MAPE	MAE	MaxAPE	MaxAE	Norm. BIC
Winters Additive	0,517	0,924	3.150,914	6,116	2.299,916	43,569	10.585,578	16,254
Winters Multiplicative	0,387	0,913	3.373,920	5,980	2.355,255	42,277	10.271,709	16,390
Simple Seasonal	0,416	0,915	3.311,182	6,047	2.353,968	39,322	10.934,058	16,305
Damped Trend	0,277	0,880	3.948,081	8,032	2.998,154	41,339	11.169,115	16,705
Holt's Linear Trend	0,762	0,880	3.930,564	8,123	3.019,423	41,804	10.945,612	16,648
Brown's Linear Trend	0,738	0,868	4.095,312	8,386	3.167,584	40,294	10.288,938	16,683

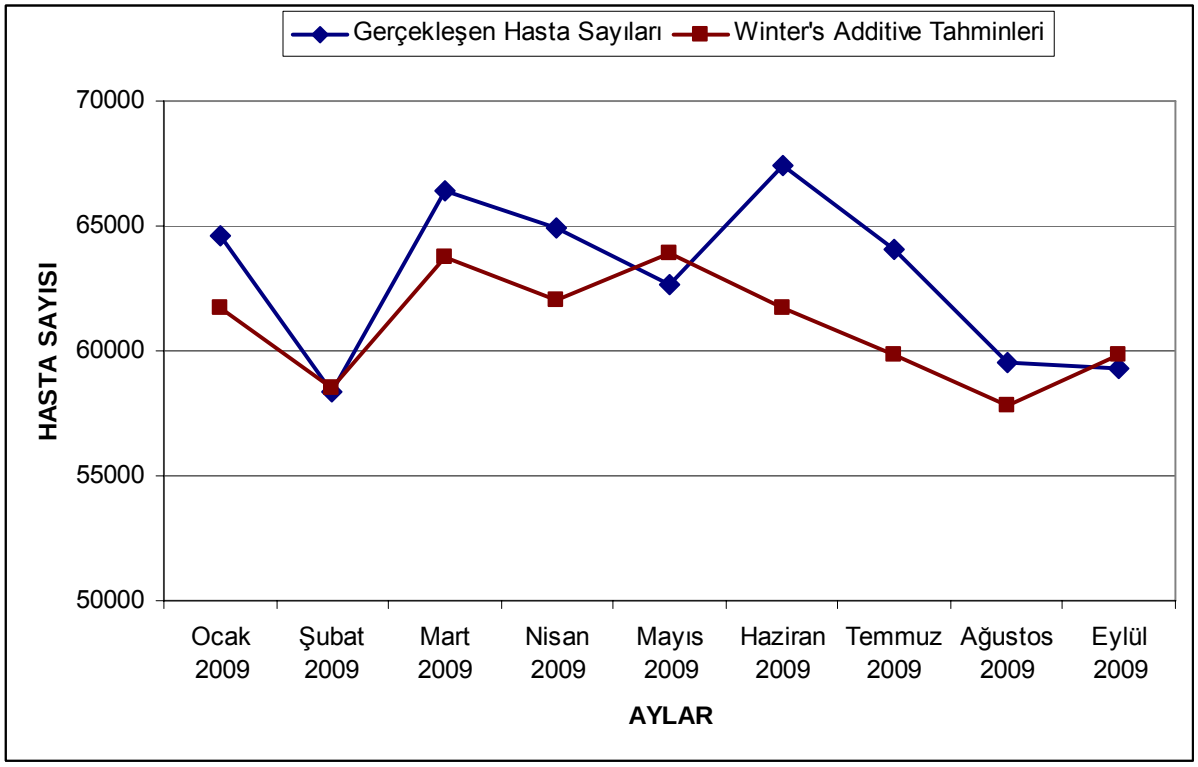
Tablo 3.7'den R^2, RMSE (Root Mean Square Error, Hata Kareleri Ortalamasının Karekökü), MAE (Mean Absolute Error, Ortalama Mutlak Hata), ve Norm. BIC (Normalized Bayesian Information Criterion, Normalize Bayesyan Bilgi Kriteri) değerlerine göre en iyi kestirimci modelin Winters Additive modeli olduğu görülmektedir.

Winters Additive modeli ile hasta sayısı tahmin serisi ve hedef değişken değerleri serisi Şekil 3.12'de verilmiştir. Şekildeki sonuçlar Clementine yazılımında eğitilen zaman serisi modeli kullanılarak elde edilmiştir. Diğer üstel düzgünleştirme modellerinin sonuç grafikleri ve üstel düzgünleştirme modelleri kazanç (gain) grafiği Ek-2'de verilmiştir. Kazanç grafiği incelendiğinde de Winters Additive modeli en iyi model olarak gözükmektedir. Şekil 3.12 incelendiğinde özellikle her yılın Mart ve Nisan aylarında grafiğin bir tepe yaptığı, Temmuz ve Ağustos aylarında ise bir dip yaptığı dikkati çekmektedir.



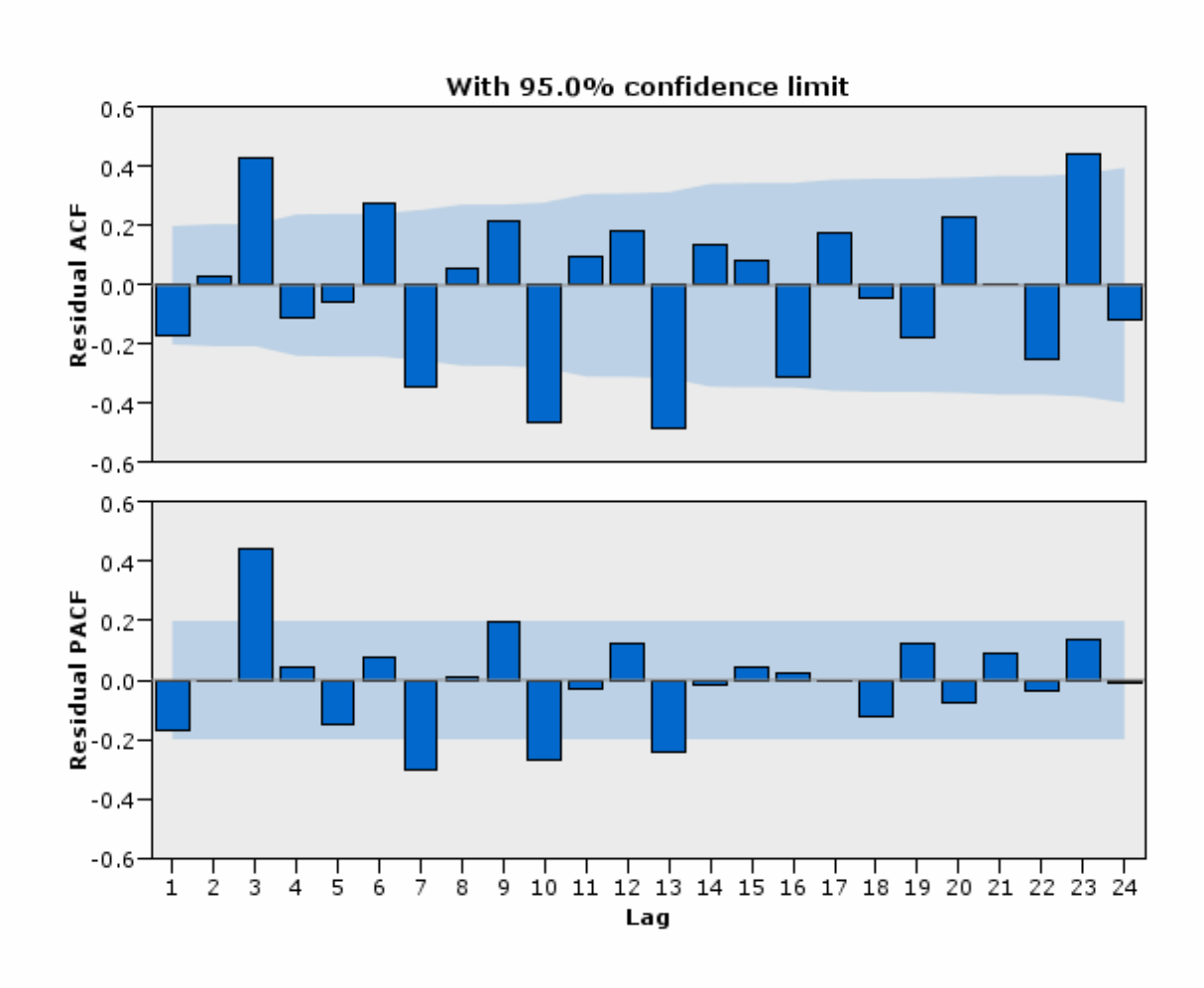
Şekil 3.12. Winters Additive Modeli ile Hasta Sayısı Tahmini

Şekil 3.13'de 2009 yılı ilk 9 aylık dönemi için Winters Additive modelinin tahmin değerleri ve bu dönemde gerçekleşen hasta sayıları grafik olarak karşılaştırılmıştır. Bu seri için en iyi üstel düzgünleştirme modeli olan Winters Additive tahminlerinin çoğu aylarda gerçekleşen değerlere çok yakın olduğu, en büyük sapmanın Haziran ayında 5.742 hasta ile gerçekleştiği görülmektedir. Bu da yaklaşık %8,5'lik bir hata anlamına gelmektedir. Ancak diğer aylarda ve özellikle 8. ve 9. aylarda elde edilen yakın tahminler modelin kestirimci gücünün yüksekliği konusunda bir fikir vermektedir.



Şekil 3.13. 2009 Yılı ilk 9 Ayı için Hasta Sayısı Tahminleri ve Gerçekleşen Değerler (Winters Additive Modeli Tahminleri)

Şekil 3.14'de Winters Additive modeli için artık terimlerin otokorelasyon fonksiyonu ve kısmi otokorelasyon fonksiyonu grafikleri verilmiştir. Otokorelasyon için genellikle analizdeki toplam gözlem sayısının ¼'i kadar gecikmede (lag) test yapılmaktadır, daha uzun gecikmelerdeki tahminler ise istatistiksel olarak güvenilir bulunmamıştır (Box ve Jenkins, 1970). Bu çalışmada aylık veri içeren sekiz yıllık bir zaman serisi girdi olarak kullanılmıştır ve dolayısıyla 96 gözlem noktasına sahiptir. Artık terim analizi de noktalar arasında en fazla 24 noktalık bir gecikme için ve %95 güven aralığında yapılmıştır.



Şekil 3.14. Winters Additive Modelinde Artık Terimlerin Analizi

Şekil 3.14'de yer alan otokorelasyon fonksiyonu grafiği incelediğinde aralarında 3, 6, 7, 10 ve 23 gecikme bulunan zaman serisi değerleri arasında %95 güven aralığında anlamlı korelasyon bulunduğu görülmektedir. Kısmi otokorelasyon fonksiyonu incelendiğinde ise aralarında 3, 7, 10 ve 13 gecikme bulunan zaman serisi değerleri arasında korelasyon bulunduğu ve korelasyonların, terimler arasında 3 gecikme için pozitif, 7, 10 ve 13 gecikme için negatif yönde etkili olduğu görülmektedir. Ayrıca kısmi otokorelasyon fonksiyonunun geometrik olarak azaldığı dikkati çekmektedir.

3.5.2.2. ARIMA Modelleri

Bileşik Otoregresif Hareketli Ortalamalar (Auto Regressive Integrated Moving Averages, ARIMA) yöntemleri ile gelecek aylardaki hasta yoğunluğunun tahmin edilmesi analizlerinde ARIMA $(p,d,q)(P,D,Q)_s$ parametreleri kullanılarak mevsimsel ve mevsimsel olmayan AR(1), MA(1), IMA(1,1) ve ARMA(1,1) süreçleri incelenmiş mevsimsel modellerin mevsimsel olmayanlara göre daha kestirimci olduğu görülmüştür. Bundan sonra AR(2), MA(2),

ARMA(2,1), ARMA(1,2), ARMA(2,2), AR(3), ARMA(3,1), ARMA(1,3), MA(3), ve ARMA(3,3) süreçleri incelenmiştir.

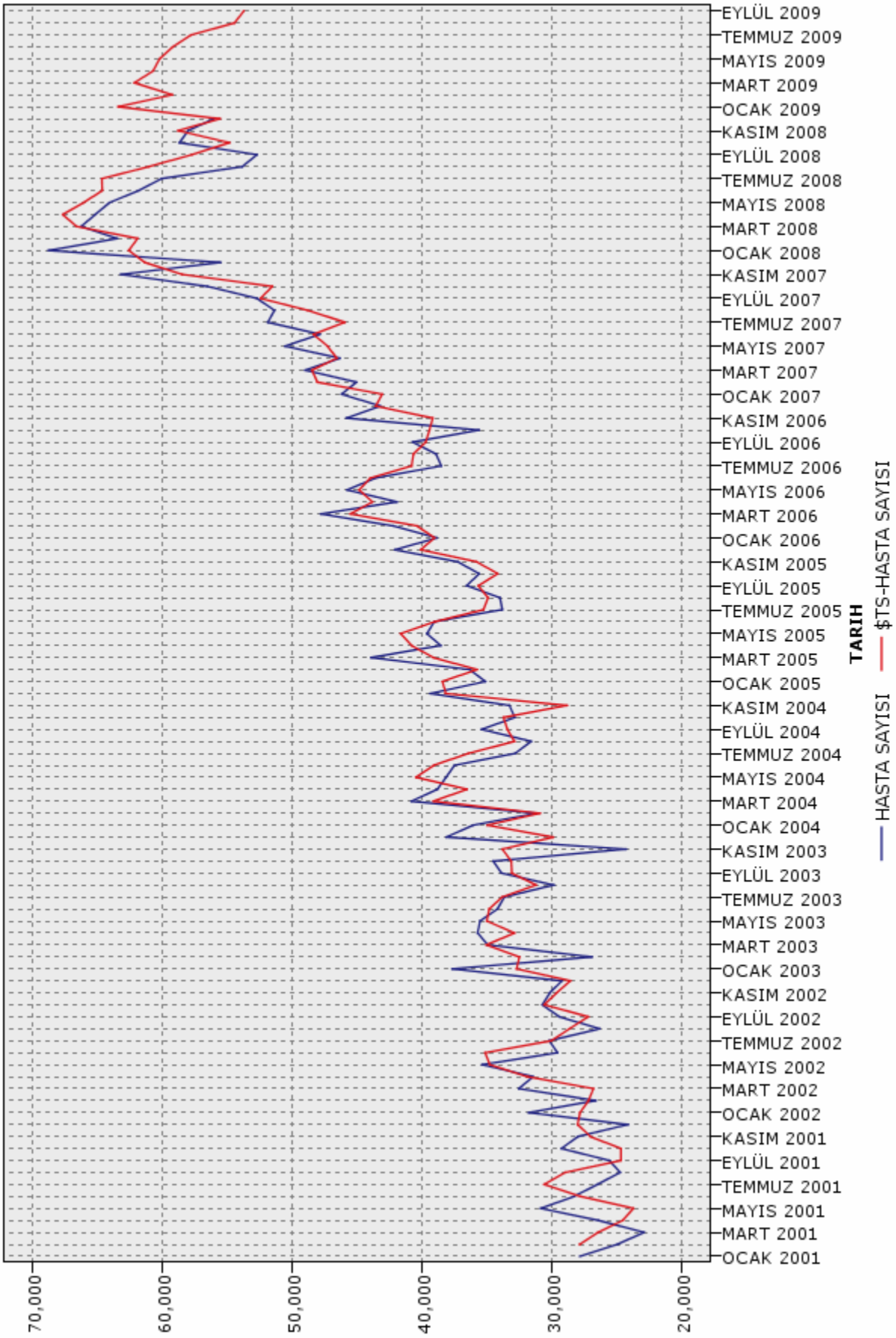
Tablo 3.8. ARIMA Modellerinin Karşılaştırılması

Model	Uyum İyiliği Kriterleri								
ARIMA $(p,d,q)(P,D,Q)_s$	Drğn. R^2	R^2	RMSE	MAPE	MAE	Max APE	Max AE	Norm. BIC	Ljung-Box Q
$(1,0,0)(0,0,0)$	0,867	0,867	4.155,114	8,454	3.170,557	43,669	14.162,726	16,807	94,312
$(1,0,0)(1,0,0)_{12}$	0,897	0,897	3.679,666	7,284	2.727,212	42,635	13.308,541	16,611	76,263
$(0,0,1)(0,0,0)$	0,840	0,840	4.568,613	9,015	3.463,164	42,697	16.095,681	16,997	92,506
$(0,0,1)(0,0,1)_{12}$	0,866	0,866	4.196,148	8,125	3.160,519	42,020	15.137,585	16,874	71,601
$(1,0,1)(0,0,0)$	0,885	0,885	3.883,086	8,035	2.985,599	42,451	12.108,389	16,719	77,827
$(1,0,1)(1,0,1)_{12}$	0,913	0,913	3.423,285	6,676	2.488,048	41,615	10.928,684	16,562	53,109
$(0,1,1)(0,0,0)$	0,276	0,879	3.968,003	8,054	3.010,915	41,145	10.901,085	16,716	79,850
$(0,1,1)(0,1,1)_{12}$	0,306	0,876	3.878,137	6,810	2.746,343	40,882	10.079,179	16,739	86,666
$(1,1,1)(0,0,0)$	0,299	0,883	3.926,681	8,238	3.061,153	40,564	9.855,467	16,743	81,090
$(1,1,1)(1,1,1)_{12}$	0,416	0,896	3.602,229	6,366	2.545,552	41,807	10.157,431	16,698	67,670
$(2,0,0)(1,0,0)_{12}$	0,915	0,915	3.356,685	6,634	2.461,653	42,501	10.325,942	16,475	68,334
$(2,0,0)(2,0,0)_{12}$	0,917	0,917	3.344,853	6,509	2.409,809	42,031	10.211,878	16,516	49,353
$(0,0,2)(0,0,1)_{12}$	0,889	0,889	3.837,951	7,404	2.834,506	44,695	11.738,252	16,743	59,804
$(0,0,2)(0,0,2)_{12}$	0,895	0,895	3.760,512	7,092	2.724,944	42,753	11.279,174	16,750	61,918
$(2,0,1)(2,0,1)_{12}$	0,917	0,917	3.377,305	6,556	2.443,445	41,762	10.146,529	16,630	55,448
$(1,0,2)(1,0,2)_{12}$	0,927	0,927	3.179,124	6,393	2.334,745	39,237	9.532,924	16,509	25,720
$(2,0,2)(1,0,1)_{12}$	0,925	0,925	3.202,446	6,547	2.373,299	38,525	9.360,005	16,524	27,575
$(2,0,2)(2,0,2)_{12}$	0,928	0,928	3.181,942	6,428	2.330,537	38,662	9.393,233	16,606	20,174
$(3,0,0)(1,0,0)_{12}$	0,919	0,919	3.294,821	6,442	2.407,417	40,490	9.837,549	16,485	36,220
$(3,1,0)(1,0,0)_{12}$	0,507	0,918	3.274,266	6,601	2.451,548	39,223	9.529,668	16,332	26,634
$(3,0,1)(1,0,0)_{12}$	0,921	0,921	3.286,861	6,359	2.374,321	40,307	9.792,886	16,528	30,708
$(1,0,3)(0,0,1)_{12}$	0,920	0,920	3.291,599	7,028	2.544,469	37,674	9.153,292	16,531	22,154
$(0,1,3)(0,0,1)_{12}$	0,469	0,911	3.454,565	7,281	2.664,887	37,698	9.159,145	16,583	28,093
$(0,0,3)(0,0,1)_{12}$	0,911	0,911	3.469,406	6,965	2.596,557	41,392	10.056,642	16,589	29,676
$(3,0,3)(1,0,1)_{12}$	0,927	0,927	3.208,422	6,423	2.334,686	38,331	9.312,874	16,623	22,632

Tablo 3.8'de karşılaştırma yapmak amacıyla incelenen ARIMA modelleri ve bunlara ait uyum iyiliği kriterleri verilmiştir. Burada da en kestirimci modeli seçmek için Normalize edilmiş Bayesyan Bilgi Kriteri (Normalized Bayesian Information Criterion, Norm.BIC) kullanılmıştır. Bu kriter Akaike Bilgi Kriteri'ne (Akaike Information Criterion, AIC) benzer bir kriter ve bu kriterle ilişkili olmakla birlikte, karşılaştırılan model parametrelerine koyduğu sınırlandırma daha güçlü olduğu için aşırı uyum (overfitting) problemi konusunda daha hassas bir ölçüdür. Bu ölçüye göre ARIMA modelleri arasında en iyi kestirimci model ARIMA$(3,1,0)(1,0,0)_{12}$ olarak görülmektedir. Bu model için Norm.BIC = 16,332 R^2 = 0,918 olmuştur.

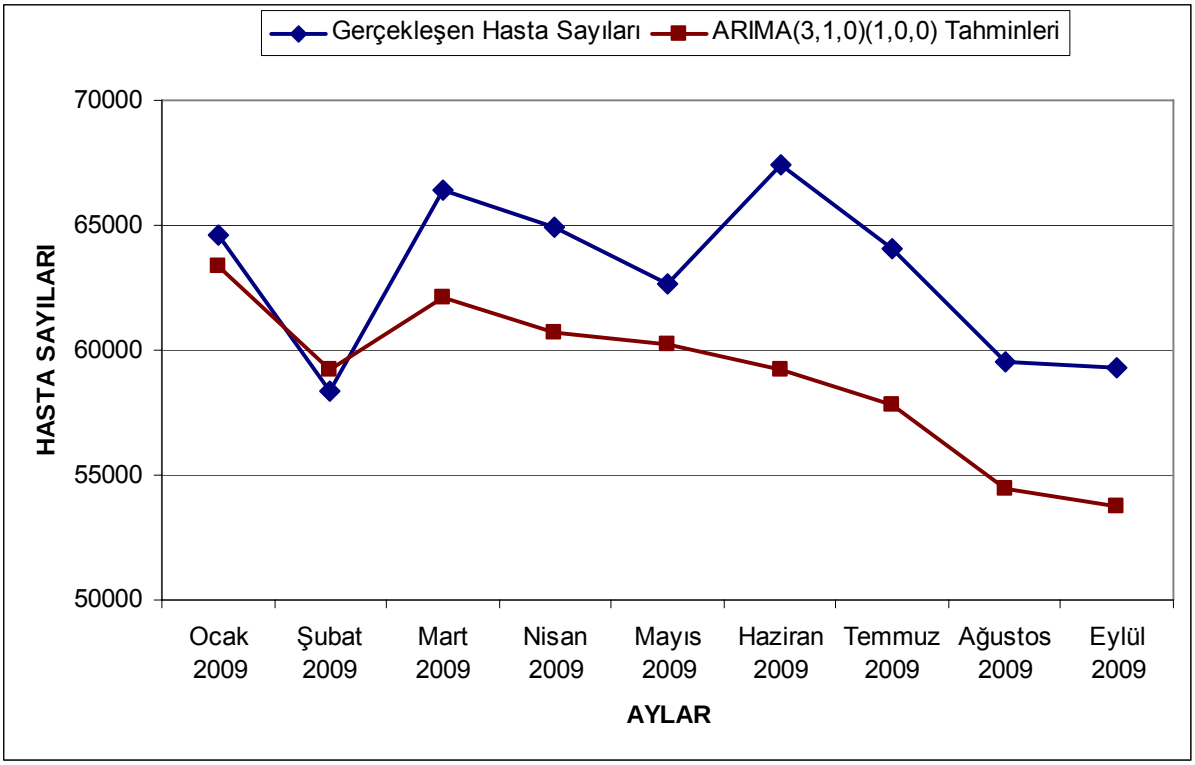
Clementine programında yer alan Expert Modeler aracı tüm alternatifleri deneyerek ARIMA modelleri içinde en düşük Norm.BIC değerine sahip modeli çıkartmaktadır. Bu araç kullanıldığında da ARIMA$(3,1,0)(1,0,0)_{12}$ en kestirimci ARIMA modeli olarak bulunmuştur. Ancak Expert Modeler birden fazla alternatif model getirmemektedir. Bunun için farklı ARIMA süreçleri ile elde edilen modelleri kendi aralarında kıyaslamak için Tablo 3.8'de yer alan modeller tek tek oluşturularak denenmiş ve sonuçlarına kıyaslama amacıyla burada yer verilmiştir. En iyi model dışında Norm. BIC. değerlerine göre seçilen beş adet örnek ARIMA modeline ait grafik ve sonuçlara Ek-3'de yer verilmiştir.

Şekil 3.15'de ARIMA$(3,1,0)(1,0,0)_{12}$ modelin hasta sayısı tahmin grafiği ve kullanılan veri serisindeki gerçek hasta sayıları grafiği verilmiştir. Grafik incelendiğinde ARIMA$(3,1,0)(1,0,0)_{12}$ modelinin tahminlerinin gerçek veriye oldukça uyum sağladığı ancak 1 fark alındığı için aradaki 1 aylık gecikme dikkati çekmektedir. Ayrıca gelecek değerlerin tahmininde de aşağıya doğru bir yönelimin olduğu görülmektedir.



Şekil 3.15. ARIMA(3,1,0)$(1,0,0)_{12}$ Modeli ile Hasta Sayısı Tahmini

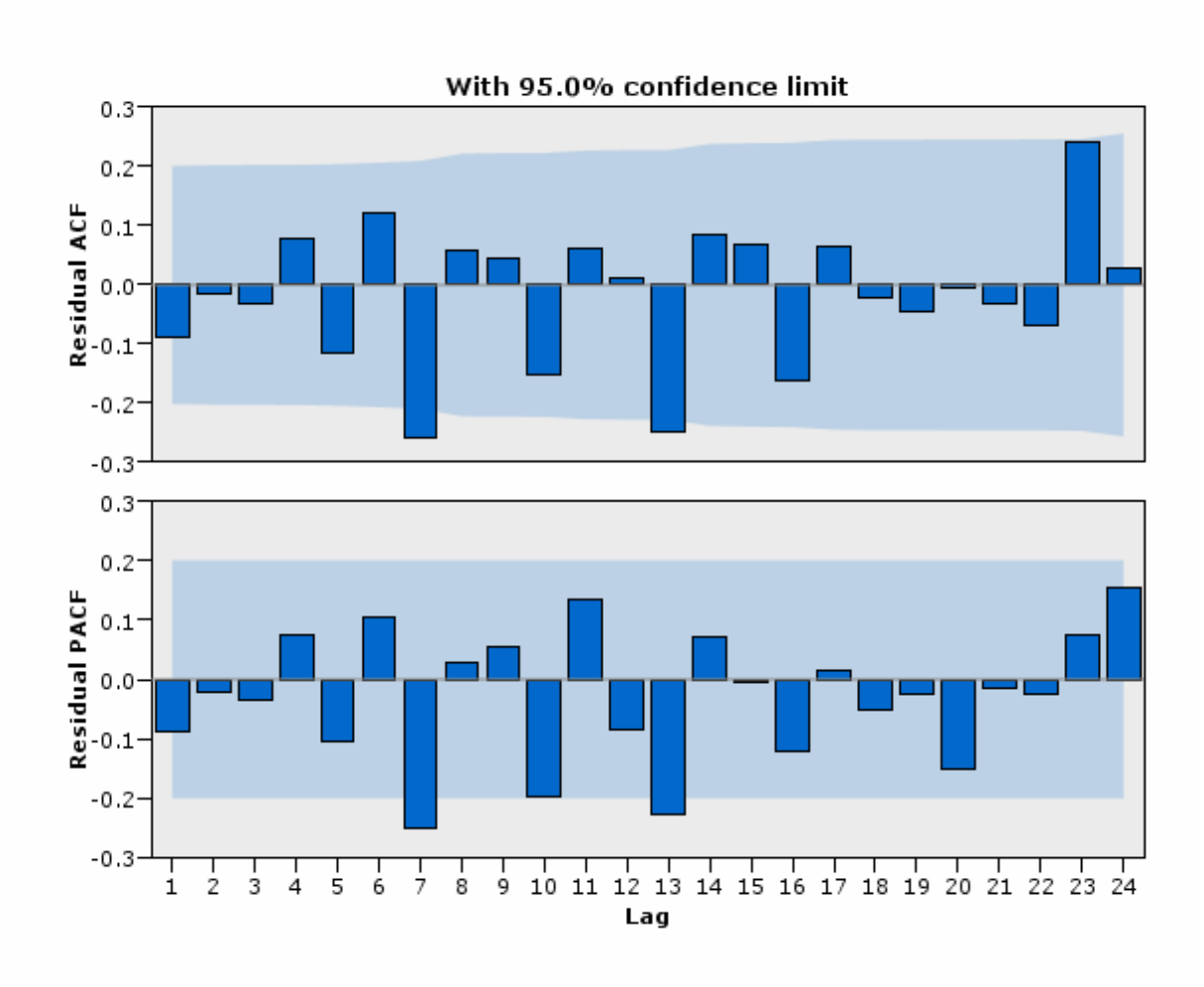
Farklı ARIMA modellerine ait 2009 yılı Ocak ve Eylül ayları arası tahmin rakamları tablosu Ek-3'de verilmiştir. Şekil 3.16'de ARIMA(3,1,0)(1,0,0)$_{12}$ modelinin 2009 yılı ilk 9 ayına ait tahmin değerleri ve bu aylarda gerçekleşen toplam hasta yoğunlukları verilmiştir.



Şekil 3.16. 2009 Yılı ilk 9 Ayı için Hasta Sayısı Tahminleri ve Gerçekleşen Değerler (ARIMA(3,1,0)(1,0,0)$_{12}$ Modeli Tahminleri)

Grafik incelendiğinde ilk aylarda daha yakın tahminler görülürken süre arttıkça tahmin değerleri ve gerçekleşen değerler arasında daha fazla fark olduğu görülmektedir. Yine ARIMA modelinde de Üstel Düzgünleştirmede olduğu gibi en fazla tahmin hatasının Haziran ayında 8.230 hasta (%12,2) ile gerçekleştiği görülmektedir.

Şekil 3.17'de (ARIMA(3,1,0)(1,0,0)$_{12}$ modelinin artık terimlerine ait otokorelasyon fonksiyonu ve kısmi otokorelasyon fonksiyonu grafikleri verilmiştir. Artık terim analizi en fazla 24 gecikme için ve %95 güven aralığında yapılmıştır. Otokorelasyon fonksiyonu incelendiğinde 7 ve 13 gecikmeli terimler arasında anlamlı korelasyon bulunduğu görülmektedir. Aynı şekilde kısmı otokorelasyon fonksiyonu grafiği incelendiğinde de 7 ve 13 gecikme için yüksek korelasyon görülmektedir.



Şekil 3.17. (ARIMA(3,1,0)(1,0,0)$_{12}$ Modelinde Artık Terimlerin Analizi

Kısmi otokorelasyon fonksiyonu grafiğinde göze çarpan bir durum ise gecikme sayısı arttıkça korelasyon değerlerinin yüksek olmasıdır. Aynı zamanda otokorelasyon fonksiyonunda da giderek azalan bir eğilimin görülmemesi serinin durağan olmadığı konusunda bir fikir vermektedir.

3.5.2.3. Yapay Sinir Ağları Modelleri

Yapay sinir ağlarının zaman serilerinde kullanımı giderek artmaktadır. Yapay sinir ağları ile yapılan tahminler bazen zaman serileri ile yapılan tahminlerden daha iyi sonuçlar verebilmektedir ancak bu her zaman olan bir durum değildir. Bu nedenle gelecek kestirimi konusunda yapay sinir ağlarının, zaman serileri modelleri ile kıyaslamaları yapılmaktadır. Bu çalışmadaki uygulamada gelecekteki hasta yoğunluklarının yapay sinir ağları ile tahmin edilmesi uygulaması için Clementine'de bulunan yapay sinir ağları modelleme aracı kullanılmıştır.

Bu aşamada gerçekleştirilen analizde geri yayılım algoritması ile ağırlıkları düzelten ileri beslemeli yapay sinir ağı ve girdi değişkeni olarak da önceki zaman serisi yöntemlerinde olduğu gibi sadece zaman endeksi (TimeIndex) değeri kullanılmıştır.

Yapay sinir ağı modellerinin eğitilmesi Exhaustive Prune, Prune, RBFN, Multiple ve Dynamic olmak üzere beş farklı yöntemle gerçekleştirilmiştir. Dynamic yönteminde bir başlangıç topolojisi oluşturulmakta ve modelin eğitimi sürecinde gizli birimler (katman veya düğüm) eklenerek veya çıkartılarak topoloji iyileştirilmektedir. Multiple yönteminde ise başlangıçta farklı topolojilere sahip birden fazla yapay sinir ağı üretilmekte ve ağlar paralel yöntemle eğitilmektedir. Bunun sonucunda en düşük ortalama hata karelerinin karekökü (RMSE) değerine sahip ağ son model olarak seçilmektedir. Prune yönteminde başlangıçta geniş bir yapay sinir ağı ile başlanmakta ve modelin eğitilmesi sürecinde gizli veya girdi katmanlarındaki en zayıf birimler elimine edilerek en iyi model bulunmaya çalışılmaktadır. RBFN (Radial Basis Function Network, Radyal Tabanlı Fonksiyon Ağları) yöntemi çok boyutlu uzayda hedef değişkenin değerlerine bağlı olarak eğri uydurma yaklaşımıdır. Modelin eğitimi daha az zaman gerektirir ancak iyi sonuçların alınabilmesi için fazla miktarda veriye ihtiyaç duymaktadır. Exhaustive Prune yönteminde ise Prune yöntemine benzer bir yaklaşım söz konusudur ancak model eğitimi parametreleri mümkün modeller uzayının tamamının taranmasından emin olunacak şekilde seçilmektedir. Bu modelin eğitimi en yavaştır ancak genellikle en iyi sonuçları üretmesi beklenir. Bununla birlikte Exhasutive Prune yaklaşımında aşırı öğrenme probleminin ortaya çıkması da beklenebilir bir durumdur.

Tablo 3.9. Eğitilen Yapay Sinir Ağı Modelleri ve Sonuçları

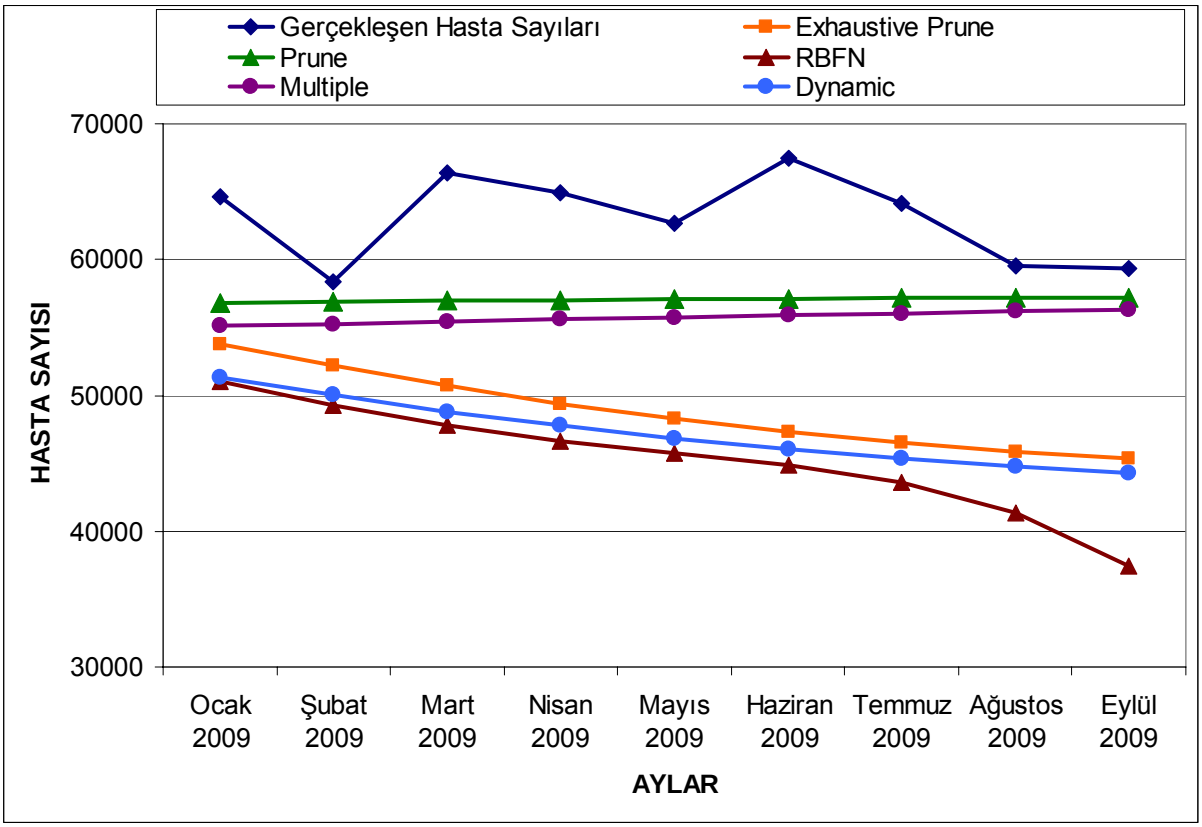
Model Eğitim Yöntemleri	Exhaustive Prune	Prune	RBFN	Multiple	Dynamic
Model Topolojisi	1:9:4:1	1:2:1	1:20:1	1:4:2:1	1:12:6:1
En Düşük Hata (Min.Error)	-9.666	-11.221	-9.226	-11.932	-9.606
En Yüksek Hata (Max. Error)	7.039	13.805	6.424	16.269	7.786
Ortalama Hata (Mean Error)	-308,438	-149,885	-73,49	-115,729	133,188
Ortalama Mutlak Hata (MAE)	2.687,708	3.259,260	2.366,698	3.778	2.794
Standart Sapma	3.359,185	4.240,077	2.977,093	4.982,346	3.370,143
Doğrusal Korelasyon	0,955	0,93	0,965	0,903	0,954

Tablo 3.9'da beş farklı yapay sinir ağı eğitim yöntemi ile elde edilen modellere ait bilgiler verilmiştir. Model topolojilerine bakıldığında Prune ve RBFN'de tek gizli katman yer aldığı

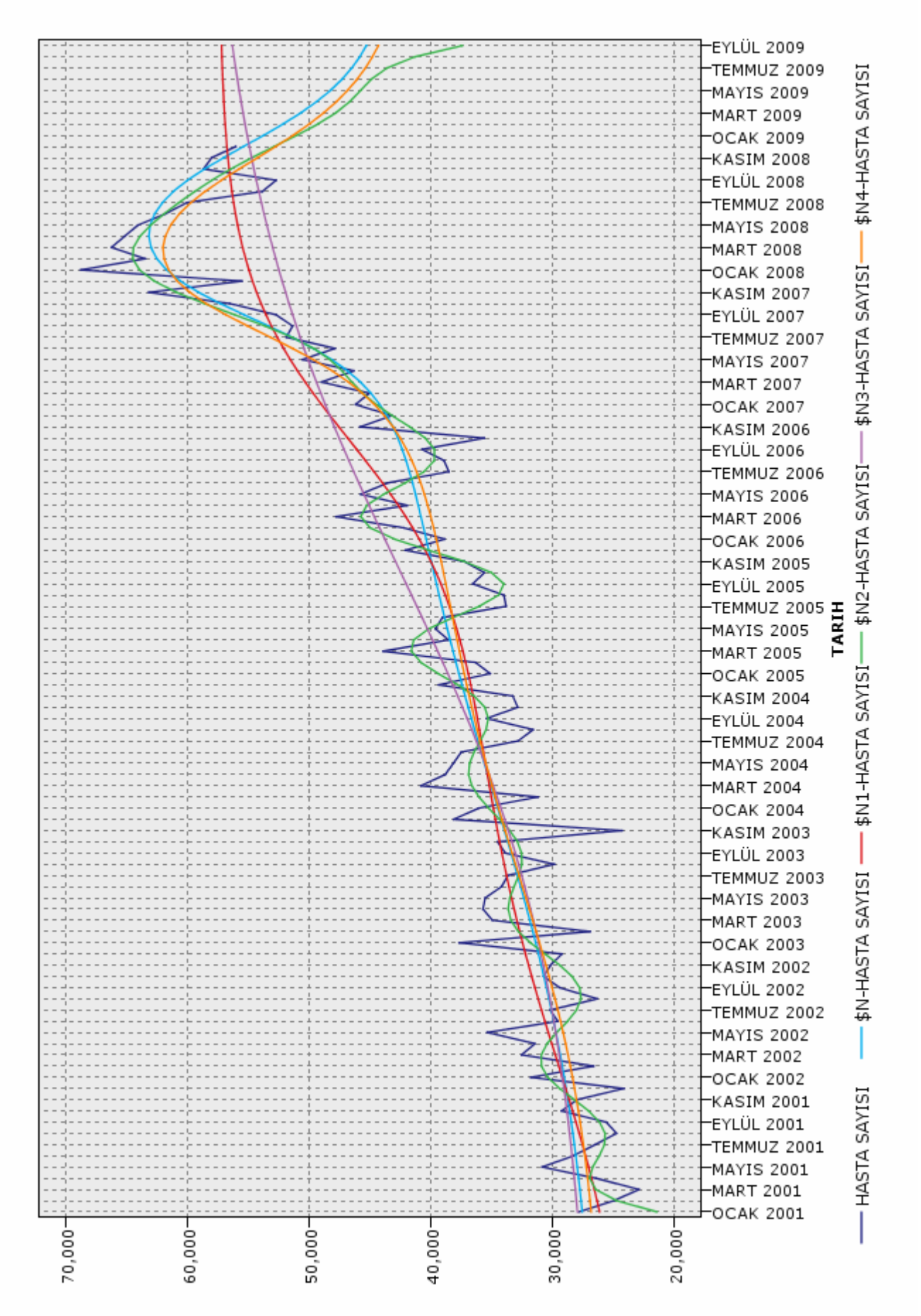
diğerlerinin ise ikişer gizli katmana sahip olduğu görülmektedir. Ancak RBFN modelinde gizli katmanda 20 gizli birim (düğüm) kullanılırken, Prune yöntemiyle elde edilen modelin gizli katmanında sadece 2 nöron vardır. Modeller arasında en küçük yapıya sahip olan da budur. Yine Multiple yöntemle elde edilen model iki gizli katmanda toplam 6 nöron içerirken Dynamic ve Exhaustive Prune ile elde edilen modeller hem ikişer gizli katmana hem de daha fazla nöron sayısına sahiptirler.

Tablo 3.9'da yer alan sonuçlar incelendiğinde veriye en fazla uyum sağlayan modelin RBFN olduğu görülmektedir. RBFN modeline çok yakın sonuçları olan diğer iki yöntem de Exhaustive Prune ve Dynamic modeleri olmuştur. Bu üç model ile Prune ve Multiple yöntemleri ile elde edilen modeller arasında veriye uyum açısından belirgin bir fark bulunmaktadır. Nispeten veriye daha düşük uyum sağlayan bu iki model arasında ise sonuçları itibariyle Prune modeli daha iyi bir model olarak gözükmektedir.

RBFN modeli veriye oldukça yüksek uyum göstermektedir. Ancak yapay sinir ağları modellerinin iyi bir kestirimci olup olmadığının ölçülmesi için gelecek 9 aydaki tahmin değerleri ve gerçekleşen hasta sayıları dikkate alınmıştır. Bu değerler Ek-4'de verilmiştir. Şekil 3.18'de modellerin tahmin değerleri ve gerçekleşen hasta sayısı grafikleri verilmiştir.



Şekil 3.18. Yapay Sinir Ağları Modelleri Tahminleri ve Gerçekleşen Hasta Sayıları



Şekil 3.19. Beş Farklı Yapay Sinir Ağı Model Sonuçları

Şekil 3.18 ve Şekil 3.19 birlikte incelendiğinde yapay sinir ağlarının veriye aşırı uyum problemleri ve en iyi kestirimci modeller daha iyi görülmektedir. Şekil 3.18'deki tahmin değerleri grafikleri incelendiğinde gerçekleşen hasta sayısına en yakın tahmin değerlerinin Prune yöntemiyle elde edilen model ile ürettiği görülmektedir. Ondan sonra da en yakın tahmin sonuçlarını Multiple yöntemi modelinin ürettiği görülmektedir. Yapay sinir ağları modelleri arasındaki farklılık üstel düzgünleştirme veya ARIMA yöntemleriyle üretilen modellerin kendi aralarındaki farklılıklardan çok daha belirgin olmuştur. Şekil 3.19'da beş farklı yöntemle üretilmiş olan yapay sinir ağı modellerinin grafikleri ile gerçek verinin zaman serisi grafiği görülmektedir. $N, $N2 ve $N4 ile belirtilen (Exhaustive Prune, Dynamic ve RBFN) yapay sinir ağı modellerinin veriye aşırı uyum göstermekle birlikte gelecek tahminlerinde oldukça başarısız oldukları görülmektedir. $N1 ve $N3 ile belirtilen (Prune ve Multiple) yapay sinir ağı modellerinin ise gerçek veriye daha az uyum sağlamakla birlikte gelecek aylardaki değerleri çok daha iyi kestirdiği görülmektedir. Bu sonuçlara göre hasta sayısının tahmini için en iyi kestirimci yapay sinir ağı modelinin Prune yöntemiyle elde edilen 1 girdi katmanı ve 2 nöronlu 1 gizli katmanı bulunan model olduğu söylenebilir.

3.5.3. Üstel Düzgünleştirme, ARIMA ve Yapay Sinir Ağları Model Sonuçların Karşılaştırılması

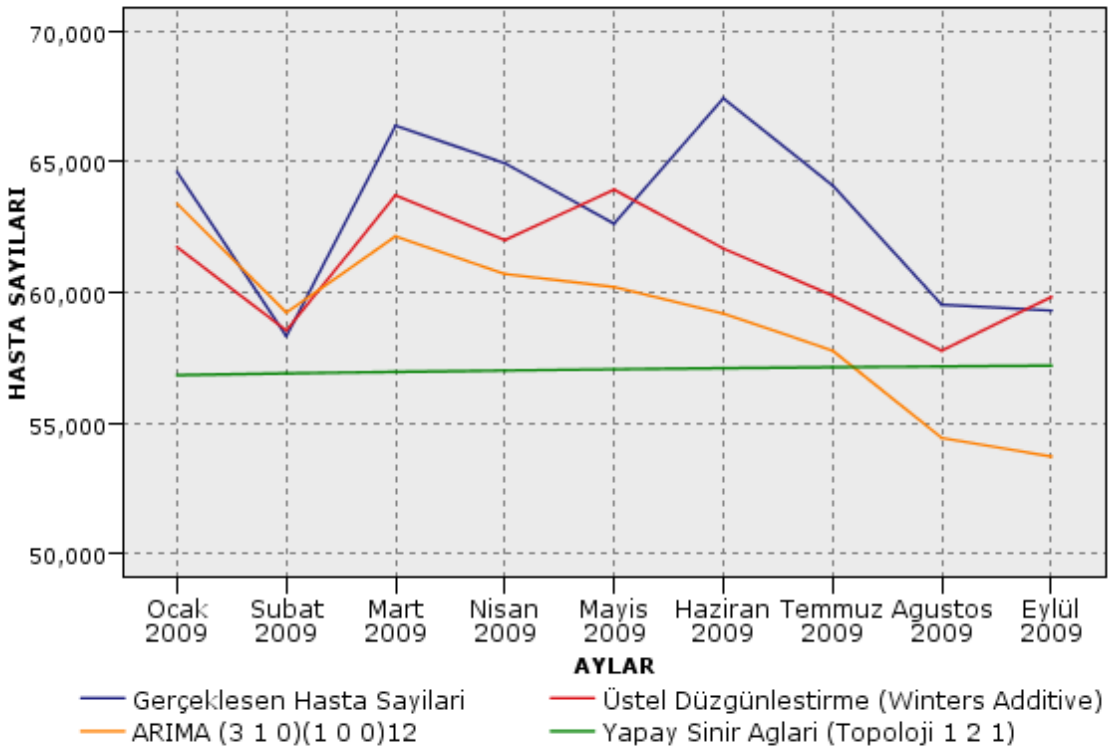
Üstel düzgünleştirme, ARIMA ve yapay sinir ağları modelleri hastanenin gelecek dokuz aylık dönemdeki hasta sayılarının tahmin edilmesi amacıyla kendi içinde karşılaştırılmış ve en iyi tahminleri yapan modeller belirlenmiştir. Burada ise her üç yöntemin en iyi modelinin hangisinin daha iyi kestirimci olduğu incelenecektir.

Tablo 3.10'da ilgili modellerin uyum iyiliği kriterleri özetlenmiştir. Bu kriterlere göre bakıldığında en büyük mutlak hata (MaxAE) ve en büyük mutlak yüzde hata (MaxAPE) kriterleri dışında Winters Additive üstel düzgünleştirme yöntemi modelinin en iyi değerlere sahip olduğu, bu kriterlere göre ise en düşük hata değerlerine ARIMA modelinin sahip olduğu görülmektedir. Ancak ortalama hatalara bakıldığında; ortalama mutlak yüzde hata (MAPE) ve ortalama mutlak hata (MAE) bakımından üstel düzgünleştirme modelinin hatasının daha düşük olduğu görülmektedir. Ortalama hatalar bakımından ise yapay sinir ağı modelinin hata değerleri diğerlerinden yüksektir. Hata kareleri ortalamalarının karekökü (RMSE) ve R^2 değerine göre ise en iyi model Winters Additive üstel düzgünleştirme modelidir. Bundan sonra sırasıyla ARIMA$(3,1,0)(1,0,0)_{12}$ modeli ve Prune yöntemiyle elde edilen, bir gizli katmanı ve iki gizli katman nöronu bulunan yapay sinir ağı modeli gelmektedir.

Tablo 3.10. Üstel Düzgünleştirme, ARIMA ve Yapay Sinir Ağı Modelleri Uyum İyiliği Kriterleri

Model	R^2	RMSE	MAPE	MAE	MaxAPE	MaxAE
Üstel Düzgünleştirme (Winters Additive)	0,924	3.150,914	6,116	2.299,916	43,569	10.585,578
ARIMA(3,1,0)(1,0,0)$_{12}$	0,918	3.274,266	6,601	2.451,548	39,223	9.529,668
Yapay Sinir Ağları (Topoloji 1:2:1)	0,859	4.220,598	8,382	3.259,260	42,126	1.3805

Şekil 3.20'de Winters Additive üstel düzgünleştirme modeli, ARIMA(3,1,0)(1,0,0)$_{12}$ modeli ve yapay sinir ağı modeline ait gelecek dokuz aylık hasta sayısı tahminleri ve gerçekleşen hasta sayılarının grafikleri verilmiştir.



Şekil 3.20. Üstel Düzgünleştirme, ARIMA ve Yapay Sinir Ağı Modelleri Tahminleri ve Gerçekleşen Hasta Sayıları

Grafik incelendiğinde Winters Additive üstel düzgünleştirme modelinin tahminlerinin gerçekleşen hasta sayısı değerlerine diğerlerinden daha yakın olduğu görülmektedir. Ağustos ve Eylül aylarında ise yapay sinir ağı modelinin tahminlerinin ARIMA(3,1,0)(1,0,0)$_{12}$ modeli tahminlerinden daha iyi olduğu görülmektedir.

SONUÇ

Veri madenciliği giderek günümüzün önemi ve yaygınlığı artan teknolojilerinden biri olmaktadır. Özellikle yığın halde verinin üretildiği alanlarda üretilen bilgi stratejik varlık olarak görülmektedir ve bu alanlarda veri madenciliğinin gelecekte çok daha yoğun olarak kullanılacağı görülebilmektedir. Yığın veri üretilen alanlardan önemli bir tanesi de sağlık hizmetleri alanıdır. Dolayısıyla sağlık alanında veri madenciliği yöntemlerinden faydalanılarak gerek tıbbi amaçlarla gerekse hastane yönetimlerine veya sağlık hizmetleri konusunda politika yapıcılara karar desteği sağlamak amacıyla veriden bilgi üretimi bu alandaki gelişmeye önemli ölçüde katkı sağlayabilecektir. Bu çalışmada özellikle hastane yönetimleri için stratejik varlık olarak nitelendirilebilecek bilginin kurumun kendi veritabanlarından nasıl çıkarılabileceği, ne tür bilgilerin hangi teknikler kullanılarak üretilebileceği ve güvenilir bilgiyi üreten en iyi modellerin seçilmesi konuları üzerinde odaklanılmıştır.

Çalışmada hastanenin veritabanında yer alan bilgiler özetlenerek tanımlayıcı istatistikler elde edilmiş, bunlardan bazıları veri görselleştirme teknikleriyle sunulmuştur. Hastanede verilen hizmetler kümeleme analizi yoluyla incelenmiştir. Bu hizmetler arasında konsültasyon hizmetlerine yoğunlaşılarak birimler arasındaki konsültasyon istemleri ve bu konuda hizmet verme ilişkileri birliktelik kuralları ile modellenmiş ve ilişkili birimlerle birlikte yoğunlukları ve olasılıkları da ortaya konulmuştur. Ayrıca kestirimci analiz olarak gelecekteki hasta yoğunluklarının tahmin edilmesi amaçlanmış ve üç farklı veri madenciliği tekniği ve bunların da kendi içinde farklı modelleri üretilerek gelecekteki hasta yoğunluklarının tahmin edilmesi ve bu konuda en iyi modellerin belirlenmesi gerçekleştirilmiştir.

Veri madenciliği kavramı ve veri madenciliği süreci ile ilgili bilgi ve tartışmalara birinci bölümde yer verilmiştir. Veri madenciliği kavramı ile veritabanlarında bilgi keşfi kavramının kullanım biçimleri ve bu konudaki farklılaşmalara değinilmiş ve veri madenciliği kavramının veri ön-işlemeden, modelleme de dahil olmak üzere, sonuçların uygulamaya geçirilmesine kadar olan sürecin tamamı için kullanılan biçiminin bu çalışmada benimsendiği dayanaklarıyla vurgulanmıştır. Burada veritabanlarında bilgi keşfi kavramı süreç olarak veri madenciliği ile birbirinin yerine geçebilir şekilde kabul edilmiştir. Ayrıca farklı endüstri standardı süreçlere de yer verilmiş ve bir veri madenciliği projesinin geliştirilmesi, gerçekleştirilmesi ve sonuçlarının uygulanmasının hangi aşamalarla tanımlandığı incelenmiştir. Bu çalışmada benimsenen Veri Madenciliği için Çapraz Endüstri Standart

Süreci (CRISP-DM) ayrıntılı bir şekilde tanımlanmış, bu süreci esas alarak yapılacak bir veri madenciliği projesinde hangi aşamalarda nelerin gerçekleştirileceği, bu adımların çıktılarının neler olacağı ve yapılan işlemler ile çıktıların dokümantasyonlarının nasıl yapılacağı her aşama için hazırlanan tablolarda ayrıntılı olarak verilmiştir. CRISP-DM sürecinin aşamalarının isimlendirilmesinde CRISP-DM konsorsiyumu üyelerinden IBM'in dokümanlarında yer alan şekli esas alınmıştır.

Sağlık alanında daha önce yapılmış çalışmalar literatür taranarak incelenmiş ve birinci bölümün son kısmında tablo olarak özetlenmiştir. Son yıllarda sağlık alanında yapılan veri madenciliği uygulamalarının arttığı görülmüştür. Çalışmalarda kullanılan teknikler bakımından yapay sinir ağları, karar ağaçları ve birliktelik kuralları tekniklerinin yoğun kullanıldığı görülmektedir. Özellikle karar ağaçları ve yapay sinir ağlarının sınıflandırma amaçlı kullanımlarında lojistik regresyon analizinin de bir karşılaştırma referans analizi olarak yoğun kullanıldığı ve yöntemlerin birbirleriyle kıyaslanması uygulamalarının gerçekleştirildiği görülmektedir. Bununla birlikte veri görselleştirme, kümeleme, zaman serileri, vaka tabanlı muhakeme, Bayes sınıflandırıcılar, destek vektör makineleri ve diskriminant analizi tekniklerinin de kullanıldığı görülmektedir.

Yaygın olarak kullanılan veri madenciliği yöntemleri ikinci bölümde ayrıntılı olarak işlenmiştir. Bu bölümde özellikle yöntemlerin matematiksel yapısı ve nasıl çalıştıkları tanımlanmıştır. Bu yaklaşım ilgili tekniklerin kullanılması sırasında analizcinin neyi niçin yaptığını iyi kavramasına ve sonuçları doğru yorumlamasına yardımcı olmaktadır. Bu bölümde Bayes sınıflandırıcılar, karar ağaçları, kümeleme, birliktelik kuralları, yapay sinir ağları ve zaman serileri yöntemlerine yer verilmiştir.

Çalışmanın uygulaması Akdeniz Üniversitesi Hastanesi veritabanı kullanılarak yapılmıştır. Hastane veritabanı olarak Oracle 9i kullanmaktadır ve bu veritabanında 1997 yılından bu yana tutulan veriler bulunmaktadır. Hastane otomasyonu yazılımları günün şartlarına göre geliştirildikçe veritabanının yapısında da değişiklikler yapılmış, özellikle bazı alanlara veri girişi daha tutarlı hale getirilmiştir. Ancak kullanılan veri bir gerçek yaşam uygulama verisi olduğu için tutarsız, dönüştürülmüş, hatalı veya eksik veri ile karşılaşılmıştır. Bu tip problemler veri ön-işleme süreçleri ile çözülmüştür. Bazı durumlarda ise planlanan uygulamadan vazgeçilmiştir. Bu durum literatürde yer alan veri madenciliği projesinin %70 çaba ve zamanının veri ön-işleme süreçlerine gittiği bilgisiyle de örtüşmektedir. Aylık başvuru sayıları zaman serileri dışında diğer uygulamalar için 2005 yılı ve sonrası veriler

kullanılmıştır. Hastane veritabanında tüm tablolarda 210 milyondan fazla kayıt bulunmaktadır. 2005 yılından sonrasında ise tablolarda yüz milyondan fazla kayıt bulunmaktadır. Veri aktarımı, verilerin filtrelenmesi, veri dönüştürme ve veri ön-işleme süreçleri veri madenciliği çalışmaları için öncelikle gerçekleştirilmiştir.

Bu çalışmada yer verilen uygulamalar üç temel kısma ayrılabilir. Birinci kısım hastanenin hasta profiline ve hastanede verilen hizmetlere ait tanımlayıcı sonuçlardır. Hastaneye başvuru yapan hastaların %93'ü ayakta tedavi süreçlerinden geçmektedir. Hastaların %7'sinin ise yatarak tedavi gören hastalar olduğu görülmüştür. Hastaneye dört yıllık dönemde toplam 2.272.491 başvuru yapıldığı ve en çok başvuru yapan sağlık güvencesi tipinin 649.386 başvuru ile emekliler olduğu görülmüştür. Emekli, Memur ve Özel tipteki başvuruların yatarak tedavi oranlarının %5'ler düzeyinde olduğu, Bağ-Kur ve SSK'lıların yatarak tedavi oranlarının %10'lar düzeyinde olduğu, Yeşil Kartlıların ise yaklaşık %22'sinin yatarak tedavi olduğu görülmüştür.

Sağlık güvencesi tiplerine göre yapılan detaylı incelemede hastaların başvuru yaptıkları ilk 5 poliklinik ve yatış yaptıkları ilk 5 klinik araştırılmıştır. Burada da özellikle Yeşil Kartlıların diğerlerinden farklılaştığı görülmüştür. Diğer gruplarda ilk 5'e hiç girmeyen ve sadece Yeşil Kartlı başvurularında ilk 5'te yer alan poliklinikler; Transplantasyon Polikliniği, Pediyatrik Hematoloji-Onkoloji Polikliniği ve Pediyatrik Nöroloji Polikliniği olurken, yatış yapılan klinikler; Çocuk Sağlığı ve Hastalıkları Kliniği ile Organ Nakli Kliniği olmuştur. Bu ayrışma özellikle gelir durumu düşüklüğü belirgin olan Yeşil Kart grubu için ileri araştırmaya değer bir bulgu olarak karşımıza çıkmaktadır.

Hastanede verilen konsültasyon hizmetlerinin istem yapan birime ve hizmeti veren birime göre birimler arası ilişkilerinin ve bu ilişkilerin yoğunluklarının belirlenmesi amacıyla birliktelik kuralları tekniğiyle iki farklı modellemesi yapılmıştır. İlk modelde istemi yapan birim öncül öğe olarak tanımlanmış ve böylece en çok istem yapan birimler ve bu birimlerin yoğun ilişkide olduğu birimler belirlenmiştir. Bu analizin sonucunda elde edilen birliktelik kuralları güven ve destek değerlerine göre yorumlanmış ve sonuçlar bir ağ (network) grafiği ile görselleştirilmiştir. Birliktelik kuralları tekniği literatürde ilk defa böyle bir uygulama amacıyla kullanılmış ve sonuçları itibariyle oldukça kullanışlı bilgiler üretilmiştir. En çok konsültasyon isteminde bulunan birim beklenebileceği gibi Acil Servis olurken, Acil Servis'in 7. sırada yoğun konsültasyon isteminde bulunduğu Göz Hastalıkları Polikliniği, 2. yoğun konsültasyon isteminde bulunan İç Hastalıkları Polikliniğinin yoğun konsültasyon isteminde

bulunduğu iki birimden biri olmuştur. Bu ve benzeri şekilde ağ yapısı içerisinde yorumlanabilecek 41 adet kural güven, destek ve kaldırma oranı değerleriyle sunulmuştur.

İkinci tip modellemede konsültasyon hizmetini veren birim öncül öğe olarak tanımlanarak en çok hizmet veren birimler ve bu birimlerin yoğun hizmet verdiği birimler belirlenmiştir. Burada elde edilen örüntüler de yine bir ağ grafiği ile gösterilmiştir. En çok konsültasyon hizmeti veren birim Anestezi ve Reanimasyon Bilim Dalı Polikliniği olarak çıkarken, bu birimin yoğun hizmet verdiği birimler beklenen şekilde cerrahi branşlar olmuştur. En yoğun konsültasyon verilen birim K.B.B. Polikliniği olurken, ikinci sırada Genel Cerrahi Polikliniği çıkmıştır. Konsültasyon hizmetini veren birime göre de elde edilen 44 adet kural güven, destek ve kaldırma oranı değerleriyle verilmiştir. Burada elde edilen her iki türdeki kurallar, birimler arasındaki ilişkilerin ve yoğunluklarının belirlenmesine ve hastane yönetimlerinin bu konuda etkinlik sağlayıcı önlemler alma yoluna gitmesine yardımcı olabilir niteliktedir.

Hastanenin gelecek aylardaki hasta yoğunluğunun tahmin edilmesi için 96 aylık hasta başvuru sayıları veritabanından derlenerek bir zaman serisi elde edilmiş ve bu seri kullanılarak 3 farklı yöntem ile zaman serisi analizleri yapılarak 2009 yılı ilk 9 ayı için hasta başvuru sayıları tahmin edilmiştir. Üstel düzgünleştirme yöntemleri, bileşik otoregresif hareketli ortalama (ARIMA) yöntemleri ve yapay sinir ağları yöntemleri önce kendi içlerinde kıyaslanmış ve en kestirimci modeller belirlenmiş, sonrasında da bu modeller birbirleriyle kıyaslanarak en iyi tahminleri yapan kestirimci model belirlenmiştir.

Bu aşamada bütün modellerin eğitilmesi ayrı ayrı yapılmış ve her modele ait değerler kıyaslama için elde edilmiştir. Ancak özellikle ARIMA süreçlerinde ortaya çıkabilecek model sayısı çok aşırı olduğu için muhtemel olasılıklar üzerinden gidilerek ve belirli modellere odaklanılarak model oluşturma süreci izlenmiştir. Daha sonrasında Expert Modeler kullanılarak en iyi Norm. BIC. değerine sahip modelin kıyaslanan modeller arasına dahil olduğunun sağlaması gerçekleştirilmiştir. Ayrıca modellere ilişkin karşılaştırma ve kazanç (gain) grafikleri Ek-2, Ek-3 ve Ek-4'te verilmiştir. Üstel düzgünleştirme yöntemleri arasında en kestirimci model Winters Additive modeli olmuştur. ARIMA süreçleri içinde en kestirimci model ARIMA$(3,1,0)(1,0,0)_{12}$ modeli olmuştur. Yapay sinir ağları yöntemleri arasında ise en kestirimci model Prune yöntemiyle elde edilen model olmuştur. Yapay sinir ağları modelleri arasında veriye daha fazla uyum gösteren modeller olmasına rağmen bunlarda aşırı-öğrenme probleminin gerçekleştiği görülmüştür.

Her yöntemin en kestirimci modellerinin birbirleriyle kıyaslanması, uyum iyiliği kriterleri ve modellerin tahmin değerleri ile hastane veritabanından elde edilen gerçekleşen hasta sayısı değerlerinin karşılaştırılması yöntemleri kullanılarak gerçekleştirilmiştir. Her iki konuda da Winters Additive üstel düzgünleştirme modeli en kestirimci model olmuştur. Uyum iyiliği kriterleri bakımından ve ilk 7 aydaki tahminlerin gerçeğe yakınlığı bakımından ARIMA$(3,1,0)(1,0,0)_{12}$ modeli ikinci en iyi olsa da 8. ve 9. aylardaki tahmin değerleri kötüye gitmiş ve yapay sinir ağları modeli bu aylarda Winters Additive üstel düzgünleştirme modelinden sonra ikinci en iyi tahminleri gerçekleştirmiştir. Gerçekleşen sayıların tahminlere oldukça yakın olması bu tekniklerin hastanenin yoğunluk tahminleri için kullanılabileceğini göstermektedir.

Sonuç olarak, sağlık alanında üretilen yığın bilgi hastanelerin veya sağlık kurumlarının veritabanlarında veya veri ambarlarında tutulmaktadır. Bu yığın veriden birçok amaç doğrultusunda faydalı bilgiler elde edilmesi sağlık kurumları, yöneticileri, çalışanları ve hastalar açısından olumlu sonuçları beraberinde getirecektir. Veri madenciliği teknikleri kullanılarak bahsedilen veritabanları veya veri ambarlarından faydalı bilgilerin elde edilmesi mümkündür. Bu çalışmada bu işin hangi süreçlerin takip edilerek ve bu süreçlerde hangi işlemlerin gerçekleştirilerek yapılabileceği tanımlanmaya çalışılmış, buna ilişkin literatürde ve endüstrideki uygulamalara ait bilgilere yer verilmiş ve bir gerçek yaşam uygulamasının içine girilerek tanımlayıcı ve kestirimci analizlerle bir veri madenciliği uygulamasının en başından en sonuna kadar gerçekleştirilmesi sunulmuştur.

KAYNAKÇA

Agrawal, R. ve Srikant, R., "Fast Algorithms for Mining Association Rules", Proceedings of International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile, (1994), 487-499.

Agrawal, R., Imieliński, T. ve Swami, A., "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, (1993), 207-216.

Amin, M. F. ve Murase, K., "Single-Layered Complex-Valued Neural Network for Real-Valued Classification Problems", Neurocomputing, 72, (2009), 945-955.

Arulampalam, G. ve Bouzerdoum, A., "A Generalized Feedforward Neural Network Architecture for Classification and Regression", Neural Networks, 16, (2003), 561-568.

Bajcsy, P., Han, J., Liu, L. ve Yang, J., "Survey of Biodata Analysis from a Data Mining Perspective" (in Eds. Wang, J. T. L., Zaki, M. J., Toivonen, H. T. T. ve Shasha, D.) Data Mining in Bioinformatics, Springer-Verlag London Limited, London, UK, (2005), 9-39.

Batyrshin, I. Z. ve Sheremetov, L. B., "Perception-Based Approach to Time Series Data Mining", Applied Soft Computing, 8, (2008), 1211-1221.

Bernardos, P. G. ve Vosniakos, G. C., "Optimizing Feedforward Artificial Neural Network Architecture", Engineering Applications of Artificial Intelligence, 20, (2007), 365-382.

Berger, A. M. ve Berger, C. R., "Data Mining as a Tool for Research and Knowledge Development in Nursing", Computers, Informatics, Nursing, 22(3), (2004), 123-131.

Berry, M. J. A. ve Linoff, G. S., Data Mining Techniques for Marketing, Sales, and Customer Relationship Management (Second Edition), Wiley Publishing Inc., Indianapolis, Indiana, 2004.

Bezdek, J. C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

Bishop, C. M., Neural Networks for Pattern Recognition, Oxford University Press, Great Clarendon Street, Oxford, UK, 2005.

Bolstad, W. M., Introduction to Bayesian Statistics, John Wiley & Sons, Hoboken, New Jersey, 2004.

Bowerman, B. L., O'Connell, R. T. ve Koehler, A. B., Forecasting, Time Series, and Regression: An Applied Approach, Fourth Edition, Thomson Brooks/Cole, Belmont, CA, USA, 2005.

Bowerman B. L. ve O'Connell, R. T., Forecasting and Time Series: An Applied Approach, Third Edition, Duxbury Thomson Learning, Pacific Grove, CA, USA, 1993.

Box, G. E. P. ve Jenkins, G. M., Time Series Analysis: Forecasting and Cotrol, Holden-Day, San Francisco, 1970.

Breault, J. L., Goodall, C. R. ve Fos, P. J., "Data Mining a Diabetic Data Warehouse", Artificial Intelligence in Medicine, 26, (2002), 37-54.

Breiman, L., Friedman, J. H., Olshen, R. A. ve Stone C. J., Classification and Regression Trees, Wadsworth & Brooks, Monterey, CA, 1984.

Brohman, M. K., "Knowledge Creation Opportunities in the Data Mining Process", Proceedings of the 39th Hawaii International Conference on System Sciences, Vol.8, (2006), 1-10.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. ve Zanasi, A., Discovering Data Mining: From Concept to Implementation, Prentice Hall PTR, Upper Saddle River, NJ, 1998.

Chae, Y. M., Kim, H. S., Tark, K. C., Park, H. J. ve Ho, S. H., "Analysis of Healthcare Quality Indicator Uning Data Mining and Decision Support System", Expert Systems with Applications, 24, (2003), 167-172.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. ve Wirth, R., CRISP-DM 1.0: Step by Step Data Mining Guide, SPSS, Chicago, IL, USA, 2000.

Chen, B. ve Wang, J., "Global Exponential Periodicity and Global Exponential Stability of a Class of Recurrent Neural Networks with Various Activation Functions and Time-Varying Delays", Neural Networks, 20, (2007), 1067-1080.

Chen, C.-M., Hsieh, Y.-L. ve Hsu, S.-H., "Mining Learner Profile Utilizing Association Rule for Web-based Learning Diagnosis", Expert Systems with Applications, 33(1), (2007), 6-22.

Chen, T.-J., Chou, L.-F. ve Hwang, S.-J., "Application of a Data-Mining Technique to Analyze Coprescription Patterns for Antacids in Taiwan", Clinical Therapeutics, 25, (2003), 2453-2463.

Chen, W.-C., Tseng, S.-S. ve Wang, C.-Y., "A Novel Manufacturing Defect Detection Method Using Association Rule Mining Techniques", Expert Systems with Applications, 29(4), (2005), 807-815.

Chen, Y.-W., Larbani, M., Hsieh, C.-Y. ve Chen, C.-W., "Introduction of Affinity Set and Its Application in Data-Mining Example of Delayed Diagnosis", Expert Systems with Applications, 36, (2009), 10883-10889.

Daqi, G. ve Genxing, Y., "Influences of Variable Scales and Activation Functions on the Performances of Multilayer Feedforward Neural Networks", Pattern Recognition, 36, (2003), 869-878.

Delavari, N., Beikzadeh, M. R. ve Phon-Amnuaisuk, S., "Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System", IEEE ITHET 6th Annual International Conference, Juan Dolio, Dominican Republic, (2005), F4B/1-6.

Delen, D., Fuller, C., McCann, C. ve Ray, D., "Analysis of Healthcare Coverage: A Data Mining Approach", Expert Systems with Applications, 36, (2009), 995-1003.

Delen, D., Walker, G. ve Kadam, A., "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods", Artificial Intelligence in Medicine, 34, (2005), 113-127.

Detienne, K. B., Detienne, D. H. ve Joshi, S. A., "Neural Networks as Statistical Tools for Business Researchers", Organizational Research Methods, 6(2), (2003), 236-265.

Duda, R. O., Hart, P. E. ve Stork, D. G., Pattern Classification (Second Edition), John Wiley, New York, 2001.

Dunham, M. H., Data Mining: Introductory and Advanced Topics, Prentice-Hall, Upper Saddle River, NJ, USA, 2003.

Dunn, J. C., "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Cybernetics and Systems, 3(3), (1973), 32-57.

Ezziane, Z., "Applications of Artificial Intelligence in Bioinformatics: A Review", Expert Systems with Applications, 30, (2006), 2-10.

Fayyad, U. M., "Data Mining and Knowledge Discovery: Making Sense Out of Data", IEEE Intelligent Systems, 11(5), (1996), 20-25.

Fayyad, U. M., Piatetsky-Shapiro, G. ve Smyth, P., "From Data Mining to Knowledge Discovery in Databases", Artificial Intelligence Magazine, Fall, (1996), 37-54.

Fayyad, U. M. ve Irani, K. B., "On the Handling of Continuous-Valued Attributes in Decision Tree Generation", Machine Learning, 8, (1992), 87-102.

Feng, L., Dillon, T. ve Liu, J., "Inter-transactional Association Rules for Multi-dimensional Contexts for Prediction and Their Application to Studying Meteorological Data", Data & Knowledge Engineering, 37(1), (2001), 85-115.

Fu, Y., "Data Mining: Tasks, Techniques and Applications", IEEE Potentials, 16(4), (1997), 18-20.

Gaynor, P. E. ve Kirkpatrick, R. C., Introduction to Time-Series Modeling and Forecasting in Business and Economics, McGraw-Hill International Editions, Singapore, 1994.

Gil-García, R. J., Badía-Contelles, J. M. ve Pons-Porrata, A., "A General Framework for Agglomerative Hierarchical Clustering Algorithms", The 18th International Conference on Pattern Recognition (ICPR'06) IEEE Computer Society, Hong Kong, 20-24 August 2006.

Giudici, P., Applied Data Mining: Statistical Methods for Business and Industry, John Wiley & Sons, West Sussex, England, 2003.

Goh, S. L. ve Mandic, D. P., "Recurrent Neural Networks with Trainable Amplitude of Activation Functions", Neural Networks, 16, (2003), 1095-1100.

Goodwin, L. K., Iannacchione, M. A., Hammond, W. E., Crockett, P., Maher, S. ve Schlitz, K., "Data Mining Methods Find Demographic Predictors of Preterm Birth", Nursing Research, 50(6), (2001), 340-345.

Gougam, L. A., Tribeche, M. ve Mekideche-Chafa, F., "A Systematic Investigation of a Neural Network for Function Approximation", Neural Networks, 21, (2008), 1311-1317.

Gren, C. R., Ndao-Brumblay, S. K., Nagrant, A. M., Baker, T. A. ve Rothman, E., "Race, Age, and Gender Influences among Clusters of African American and White Patients with Chronic Pain", The Journal of Pain, 5(3), (2004), 171-182.

Han, J. ve Kamber, M., Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann Publishers, San Francisco, CA, USA, 2006.

Hand, D., Mannila, H. ve Smyth, P., Principles of Data Mining, The MIT Press, Cambridge, Massachusetts, USA, 2001.

He, Z., Xu, X., Huang, J. Z. ve Deng, S., "Mining Class Outliers: Concepts, Algorithms and Applications in CRM", Expert Systems with Applications, 27(4), (2004), 681-697.

Hill, T., O'Connor, M. ve Remus, W., "Neural Network Models for Time Series Forecasts", Management Science, 42, (1996), 1082-1092.

Ho, S. L., Xie, M. ve Goh, T. N., "A Comperative Study of Neural Network and Box-Jenkins ARIMA Modeling in Time Series Prediction", Computers & Industrial Engineering, 42, (2002), 371-375.

Hsu, C.-C., Huang, Y.-P., Chang, K.-W., "Extended Naïve Bayes Classifier for Mixed Data", Expert Systems with Applications, 35, (2008), 1080-1083.

Huang, G. ve Cao, J., "Multistability of Neural Networks with Discontinuous Activation Function", Communications in Nonlinear Science and Numerical Simulation, 13, (2008), 2279-2289.

Jenhani, I., Ben Amor, N., ve Elouedi, Z., "Decision Trees as Possibilistic Classifiers", International Journal of Approximate Reasoning, 48, (2008), 784-807.

Jerez-Aragonéz, J. M., Gómez-Ruiz, J. A., Ramos-Jiménez, G., Muñoz-Pérez, J. ve Alba-Conejo, E., "A Combined Neural Network and Decision Trees Model for Prognosis of Breast Cancer Relapse", Artificial Intelligence in Medicine, 27, (2003), 45-63.

Jonsdottir, T., Hvannberg, E. T., Sigurdsson, H. ve Sigurdsson, S., "The Feasibility of Constructing a Predictive Outcome Model for Breast Cancer Using the Tools of Data Mining", Expert Systems with Applications, 34, (2008), 108-118.

Kadılar, C., SPSS Uygulamalı Zaman Serileri Analizine Giriş, Bizim Büro Basımevi, Ankara, 2005.

Kantardzic, M., Data Mining: Concepts, Models, Methods, and Algorithms, IEEE Press, Hoes Lane, Piscataway, NJ, USA, 2003.

Kass, G. V., "An Exploratory Technique for Investigating Large Quantities of Categorical Data", Journal of Applied Statistics, 29(2), (1980), 119-127.

Katayama, K., Sakata, Y. ve Horiguchi, T., "Layered Neural Networks with Non-monotonic Transfer Functions", Physica A, 317, (2003), 270-298.

Kathirvalavakumar, T. ve Thangavel, P., "A Modified Backpropagation Training Algorithm for Feedforward Neural Networks", Neural Processing Letters, 23, (2006), 111-119.

Kendall, K. E. ve Kendall, J. E., Systems Analysis and Design, 7/E, Prentice Hall, Upper Saddle River, NJ, USA, 2008.

Kianmehr, K. ve Alhajj, R., "CARSVM: A Class Association Rule-Based Classification Framework and Its Application to Gene Expression Data", Artificial Intelligence in Medicine, 44(1), (2008), 7-25.

Kirkup, L., Data Analysis with Excel: An Introduction for Physical Scientists, Cambridge University Press, Cambridge, UK, 2002.

Koehler, A. B., Snyder, R. D. ve Ord, J. K., "Forecasting Models and Prediction Intervals for the Multiplicative Holt-Winters Method", International Journal of Forecasting, 17, (2001), 269-286.

Kohzadi, N., Boyd, M. S., Kermanshahi, B. ve Kaastra, I., "A Comparison of Artificial Neural Network and Time Series Model for Forecasting Commodity Prices", Neurocomputing, 10, (1996), 169-181.

Kojadinovic, I., "Agglomerative Hierarchical Clustering of Continuous Variables based on Mutual Information", Computational Statistics & Data Analysis, 46, (2004), 269-294.

Kros, J. F., Lin, M. ve Brown, M. L., "Effects of the Neural Network s-Sigmoid Function on the KDD in the Presence of Imprecise Data", Computers & Operations Research, 33, (2006), 3136-3149.

Kuo, R. J. ve Shih, C. W., "Association Rule Mining through the Ant Colony System for National Health Insurance Research Database in Taiwan", Computers and Mathematics with Applications, 54, (2007), 1303-1318.

Kuo, R. J., Lin, S. Y. ve Shih C. W., "Mining Association Rules through Integration of Clustering Analysis and Ant Colony System for Health Insurance Database in Taiwan", Expert Systems with Applications, 33(3), (2007), 794-808.

Larose, D. T., Discovering Knowledge in Data. An Introduction to Data Mining, John Wiley & Sons, Hoboken, NJ, USA, 2005.

Larsen, K., "Generalized Naïve Bayes Classifiers", SIGKDD Explorations, 7(1), (2005), 76-81.

Lavrač, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. ve Kobler, A., "Data Mining and Visualization for Decision Support and Modeling of Public Health-Care Resources", Journal of Biomedical Informatics, 40, (2007), 438-447

Lee, I.-N., Liao, S.-C. ve Embrechts, M., "Data Mining Techniques Applied to Medical Information", Medical Informatics, 25(2), (2000), 81-102.

Li, Q. ve Khosla, R., "Performance Optimization of Data Mining Applications Using a Multi-layered Multi-agent Data Mining Architecture", CIMSA 2005 – IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Giardini Naxos, Italy, July (2005), 227-231.

Liao S.-C. ve Lee, I.-N., "Appropriate Medical Data Categorization for Data Mining Classification Techniques", Medical Informatics, 27(1), (2002), 59-67.

Lucas, P. "Bayesian Analysis, Pattern Analysis, and Data Mining in Health-Care", Current Opinion in Critical Care, 10, (2004), 399-403.

Ma, L. ve Khorasani, K., "A New Strategy for Adaptively Constructing Multilayer Feedforward Neural Networks", Neurocomputing, 51, (2003), 361-385.

Ma, L., Tsui, F.-C., Hogan, W. R. Wagner, M. M. ve Ma, H., "A Framework for Infection Control Surveillance Using Association Rules", AMIA 2003 Symposium Proceedings, (2003), 410-414.

Maier, H. R. ve Dandy, G. C., "Neural Network Models for Forecasting Univariate Time Series", Neural Networks World, 6, (1996), 747-772.

Marakas, G. M., Modern Data Warehousing, Mining, and Visualization: Core Concepts, Prentice Hall, Upper Saddle River, New Jersey, USA, 2003.

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S. ve Knaus, W. A.,

"Data Mining and Clinical Data Repositories: Insights from a 667,000 Patient Data Set", Computers in Biology and Medicine, 36, (2006), 1351-1377.

Neumann, A., Holstein, J., Gall, J.-R. L. ve Lepage, E., "Measuring Performance in Health Care: Case-Mix Adjustment by Boosted Decision Trees", Artificial Intelligence in Medicine, 32, (2004), 97-113.

Nguyen, T., Malley, R., Inkelis, S. H. ve Kuppermann, N., "Comparison of Prediction Models for Adverse Outcome in Pediatric Meningococcal Disease Using Artificial Neural Network and Logistic Regression Analyses", Journal of Clinical Epidemiology, 55, (2002), 687-695.

Olaru, C. ve Wehenkel, L., "Data Mining", IEEE Computer Applications in Power, 12(3), (1999), 19-25.

Olson D. L. ve Shi, Y., Introduction to Business Data Mining, McGraw-Hill/Irwin, New York, NY, USA, 2007.

Orhunbilge, N., Zaman Serileri Analizi Tahmin ve Fiyat İndeksleri, Avcıol Basım Yayın, İstanbul, 1999.

Oztekin, A., Delen, D. ve Kong Z. J., "Predicting the Graft Survival for Heart-Lung Transplantation Patients: An Integrated Data Mining Methodology", International Journal of Medical Informatics, (article in pres), (2009), doi:10.1016/j.ijmedinf.2009.04.007

Pedrycz, W., Knowledge-Based Clustering: From Data to Information Granules, John Wiley & Sons, Hoboken, New Jersey, 2005.

Piatetsky-Shapiro, G. ve Frawley, W., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA, USA, 1991.

Pham, D. T., Packianather, M. S. ve Charles, E. Y. A., "A Novel Self-Organised Learning Model with Temporal Coding for Spiking Neural Networks", (in Eds. Pham, D. T., Eldukhri, E. E. ve Soroka, A. J.) Intelligent Production Machines and Systems, Cardiff University, Manufacturing Engineering Centre, Cardiff, UK., (2006), 307-312.

Phillips-Wren, G., Sharkey, P. ve Dy, S. M., "Mining Lung Cancer Patient Data to Assess Healthcare Resource Utilization", Expert Systems with Applications, 35, (2008), 1611-1619.

Pomorski, D. ve Perche, P. B., "Inductive Learning of Decision Trees: Application to Fault Isolation of an Induction Motor", Engineering Applications of Artificial Intelligence 14, (2001), 155-166.

Poynton, M. R. ve McDaniel, A. M., "Classification of Smoking Cessation Status with a Backpropagation Neural Network", Journal of Biomedical Informatics, 39, (2006), 680-686.

Prybutok, V. R., Yi, J. ve Mitchell, D., "Comparison of Neural Network Models with ARIMA and Regression Models for Prediction of Houston's Daily Maximum Ozone Concentrations", European Journal of Operational Research, 122, (2000), 31-40.

Pyle, D., Data Preparation for Data Mining, Academic Press, San Diego, CA, 1999.

Qu, H., Yi, Z. ve Wang, X. B., "Switching Analysis of 2-D Neural Networks with Nonsaturating Linear Treshold Transfer Functions", Neurocomputing, 72, (2008), 413-419.

Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, CA, 1993.

Quinlan, J. R., "Induction of Decision Trees", Machine Learning, 1, (1986), 81-106.

Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M. ve Berghe, G. V. D., "Mining Data from Intensive Care Patients", Advanced Engineering Informatics, 21, (2007), 243-256.

Razali, A. M. ve Ali, S., "Generating Treatment Plan in Medicine: A Data Mining Approach", American Journal of Applied Sciences, 6(2), (2009), 345-351.

Richards, G., Rayward-Smith, V. J., Sönksen, P. H., Carey, S. ve Weng, C., "Data Mining for Indicators of Early Mortality in a Database of Clinical Records", Artificial Intelligence in Medicine, 22, (2001), 215-231.

Semenova, T., "Discovering Patterns of Medical Practice in Large Administrative Health Databases", Data and Knowledge Engineering, 51, (2004), 149-160.

Shmueli, G, Patel, N. R. ve Bruce, P. C., Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, John Wiley & Sons, Hoboken, NJ, USA, 2007.

Silva, A., Cortez, P., Santos, M. F., Gomes, L. ve Neves, J., "Rating Organ Failure via Adverse Events Using Data Mining in the Intensive Care Unit", Artificial Intelligence in Medicine, 43, (2008), 179-193.

Skubalska-Rafajlowicz, E., "Neural Networks with Sigmoidal Activation Functions-Dimension Reduction Using Normal Random Projection", Nonlinear Analysis, (in press), (2009), doi:10.1016/j.na.2009.01.124

Solazzi, M. ve Uncini, A., "Regularising Neural Networks Using Flexible Multivariate Activation Function", Neural Networks, 17, (2004), 247-260.

Tan, P.-N., Steinbach, M. ve Kumar, V., Introduction to Data Mining, Pearson, Addison-Wesley, Boston, MA, USA, 2006.

Tang, Z. H. ve MacLennan, J., Data Mining with SQL Server 2005, Wiley Publishing Inc., Indianapolis, IN, USA, 2005.

Tsai, C. Y. ve Tsai, M. H., "A Dynamic Web Service based Data Mining Process System", Proceedings of The Fifth International Conference on Computer and Information Technology (CIT'05), Washington, DC, USA, IEEE Computer Society, (2005), 1033-1039.

Tseng, F. M., Yu, H. C. ve Tzeng, G. H., "Combining Neural Network Model with Seasonal Time Series ARIMA Model", Technological Forecasting & Social Change, 69, (2002), 71-87.

Tsetsekas, C. A., Fertis, A. G. ve Venieris, I. S., "Dynamic Application Profiles using Neural Networks for Adaptive Quality of Service Support in the Internet", Computer Communications, 29, (2006), 2985-2995.

Wang, L. ve Fu, X., Data Mining with Computational Intelligence, Springer-Verlag Berlin Heidelberg, Germany, 2005.

Winters, P. R., "Forecasting Sales by Exponentially Weighted Moving Averages", Management Science, 6, (1960), 324-342.

Witten, I. H. ve Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann Publishers, San Francisco, CA, USA, 2005.

Yang, W.-S. ve Hwang, S.-Y., "A Process-Mining Framework for the Detection of Healthcare Fraud and Abuse", Expert Systems with Applications, 31, (2006), 56-68.

Yi, X. ve Zhang, Y., "Privacy-preserving Distributed Association Rule Mining via Semi-trusted Mixer", Data & Knowledge Engineering, 63(2), (2007), 550-567.

Zhang, G. P., Neural Networks in Business Forecasting, Idea Group Publishing, Hershey, PA, 2004.

Zhang, G. P., "Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model", Neurocomputing, 50, (2003), 159-175.

Zhang, G. P., "An Investigation of Neural Networks for Linear Time-Series Forecasting", Computers & Operations Research, 28, (2001), 1183-1202.

Zhuang, Z. Y., Churilov, L., Burstein, F. ve Sikaris, K., "Combining Data Mining and Case-Based Reasoning for Intelligent Decision Support for Pathology Ordering by General Practitioners", European Jounal of Operational Research, 195, (2009), 662-675.

Zou, H. F., Xia, G. P., Yang, F. T. ve Wang, H. Y., "An Investigation and Comparison of Artificial Neural Network and Time Series Models for Chinese Food Grain Price Forecasting", Neurocomputing, 70, (2007), 2913-2923.

Zubcoff, J., Pardillo, J. ve Trujillo, J., "A UML Profile for the Conceptual Modeling of Data-Mining with Time-Series in Data Warehouse", Information and Software Technology, 51(6), (2009), 977-992.

-----, SPSS, Clementine11.1 User's Guide, Integral Solutions Limited, Chicago, IL., 2007a.

-----, SPSS, Clementine11.1 Node Reference, Integral Solutions Limited, Chicago, IL, 2007b.

Çevrimiçi Kaynaklar:

Soni, S., Tang, Z. ve Yang, J., "Performance Study of Microsoft Data Mining Algorithms", Microsoft Technet, http://technet.microsoft.com/en-us/library/cc917687.aspx, (Erişim Tarihi: 16.12.2008).

-----, CRISP-DM, http://www.crisp-dm.org, (Erişim Tarihi: 08.03.2007).

-----, Gartner Group, http://www.gartner.com/6_help/glossary/GlossaryD.jsp, (Erişim Tarihi: 16.09.2007).

-----, IBM, The Data Mining Process, DB2 Universal Database,http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.im.easy.doc/c_dm_process.html, (Erişim Tarihi: 08.03.2007).

-----, Microsoft Research, Data Mining: Efficient Data Exploration and Modeling, http://research.microsoft.com/dmx/DataMining, (Erişim Tarihi: 12.05.2006).

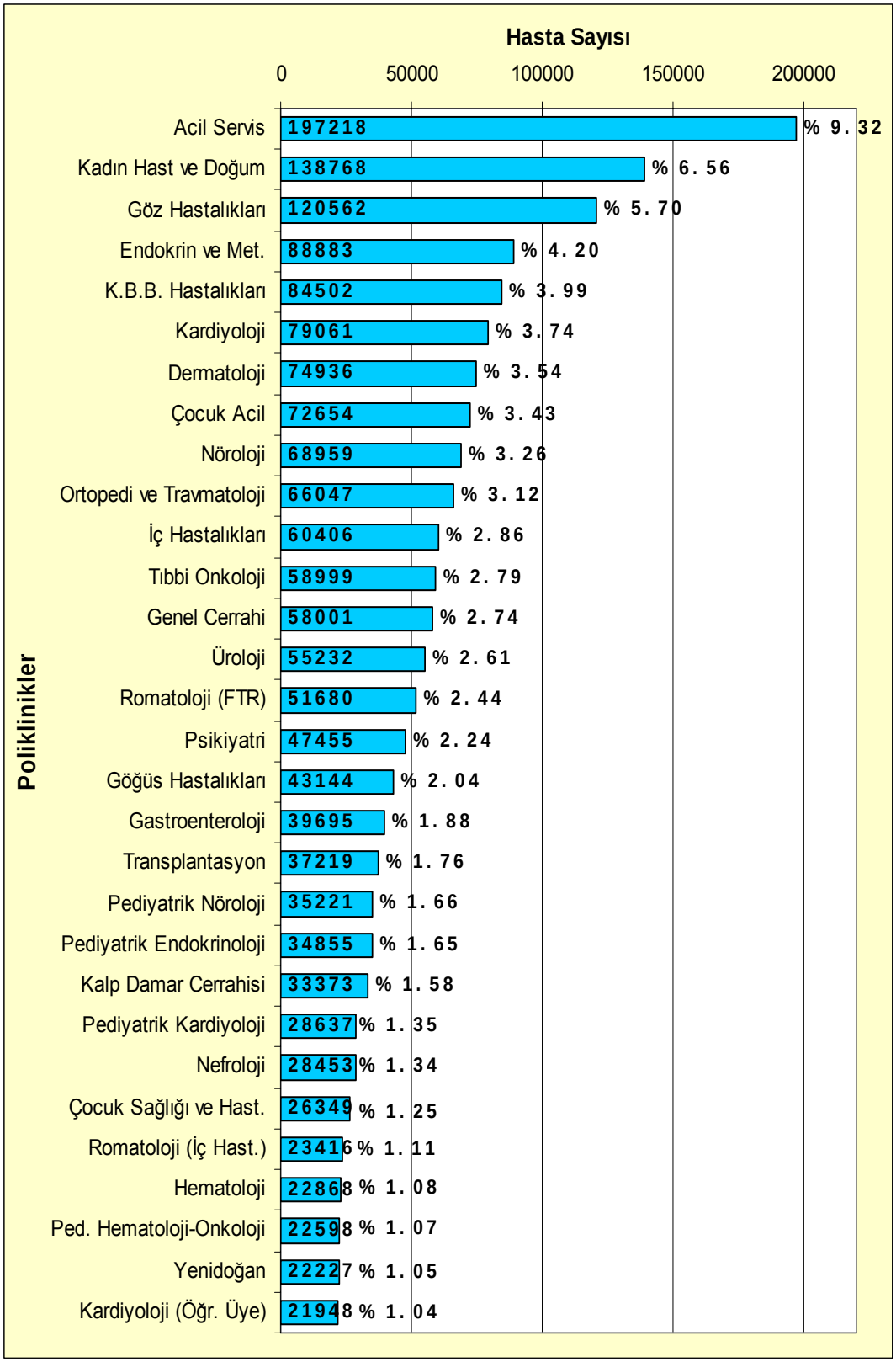
-----, Microsoft, Migration for Oracle, http://www.microsoft.com/sqlserver/2005/en/us/migration-oracle.aspx, (Erişim Tarihi: 25.05.2007).

-----, MSDN, Microsoft Association Algorithm Technical Reference, SQL Server Developer Center, http://msdn.microsoft.com/en-us/library/cc280428.aspx, (Erişim Tarihi:16.12.2008).

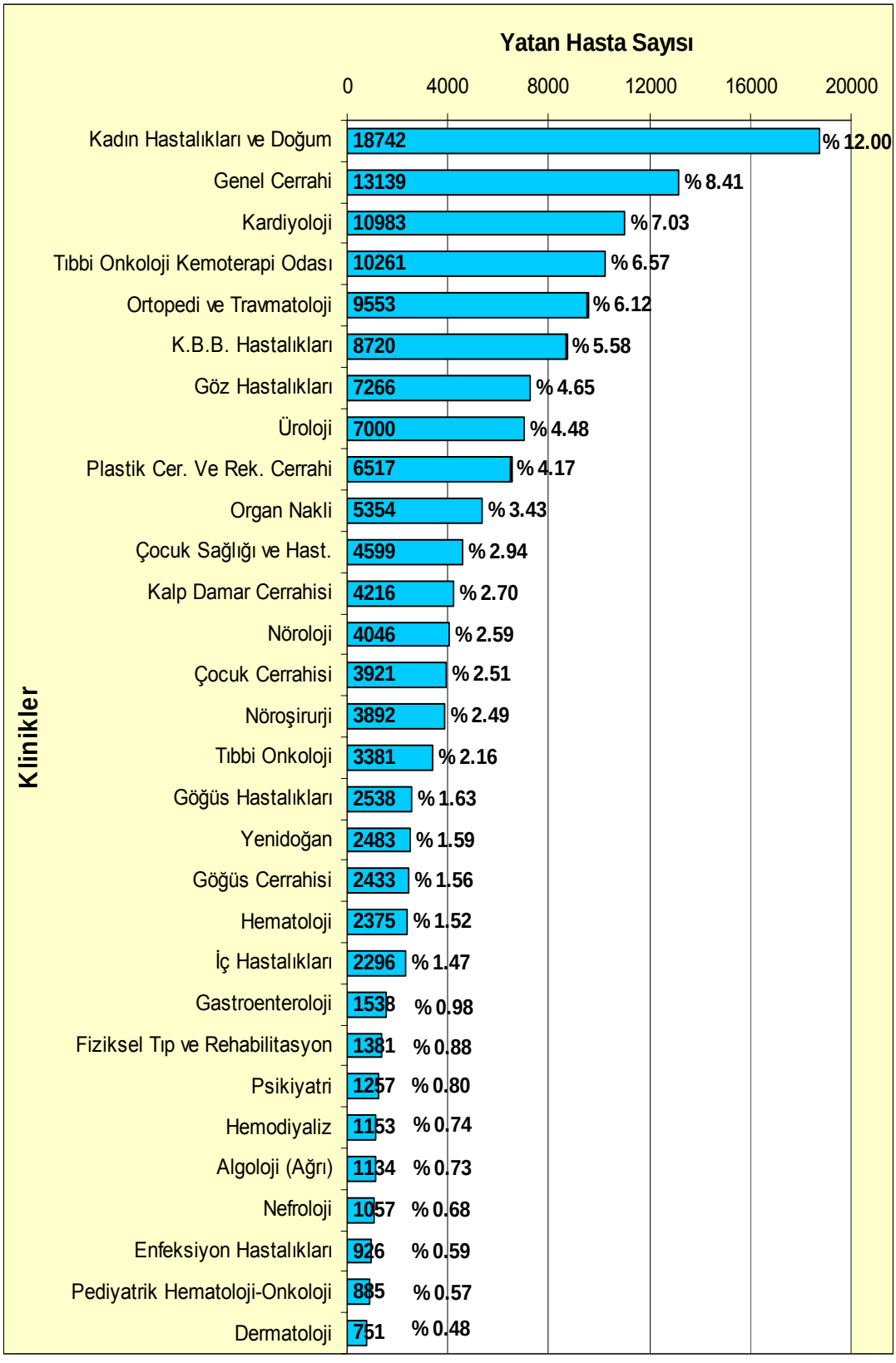
-----, SAS, SAS Enterprise Miner, http://www.sas.com/offices/europe/uk/technologies/analytics/ datamining/miner/semma.html, (Erişim Tarihi: 08.03.2007).

EKLER

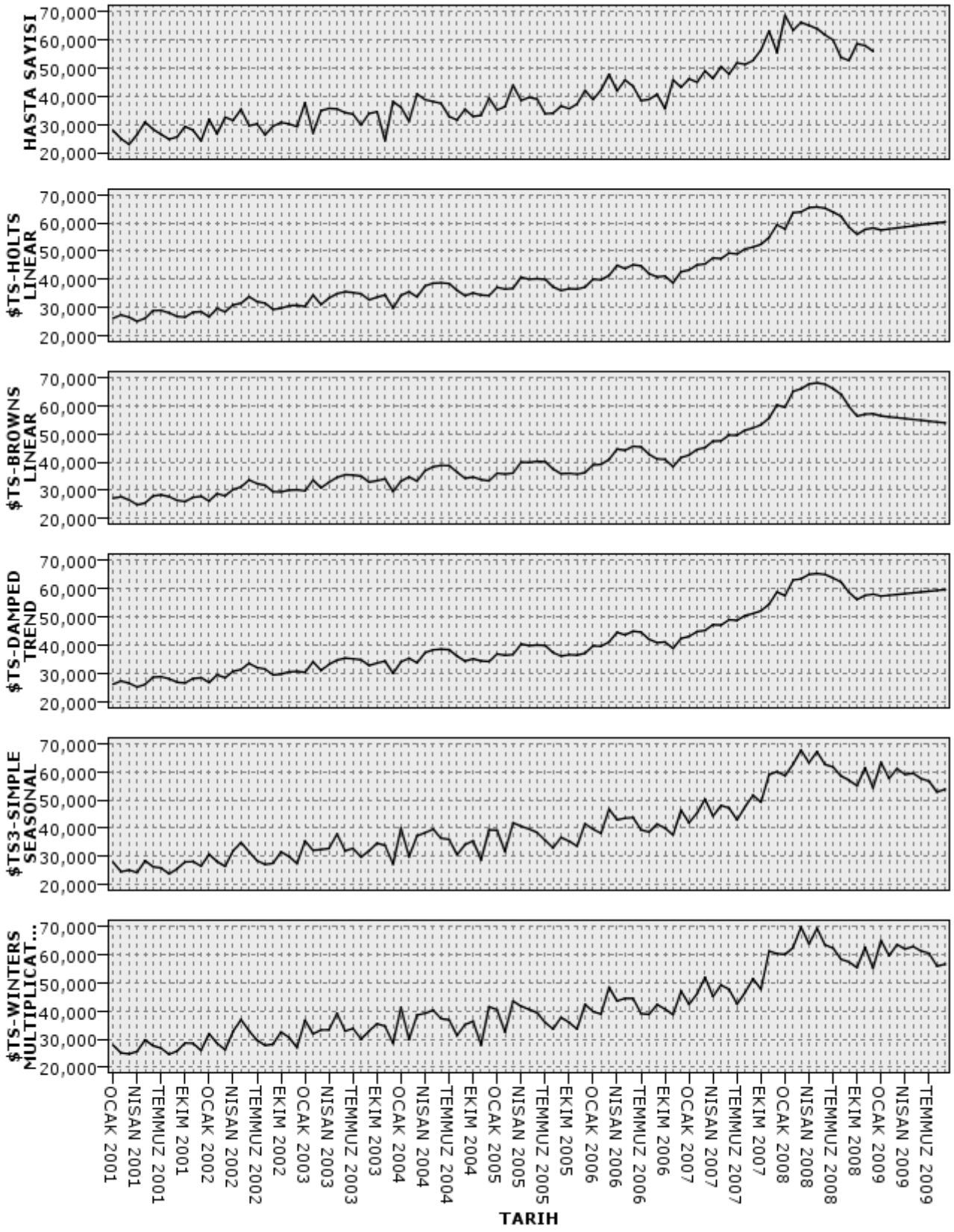
Ek-1(a): Poliklinik Hasta Başvuru Sayıları (Ocak 2005 - Aralık 2008 Arası)

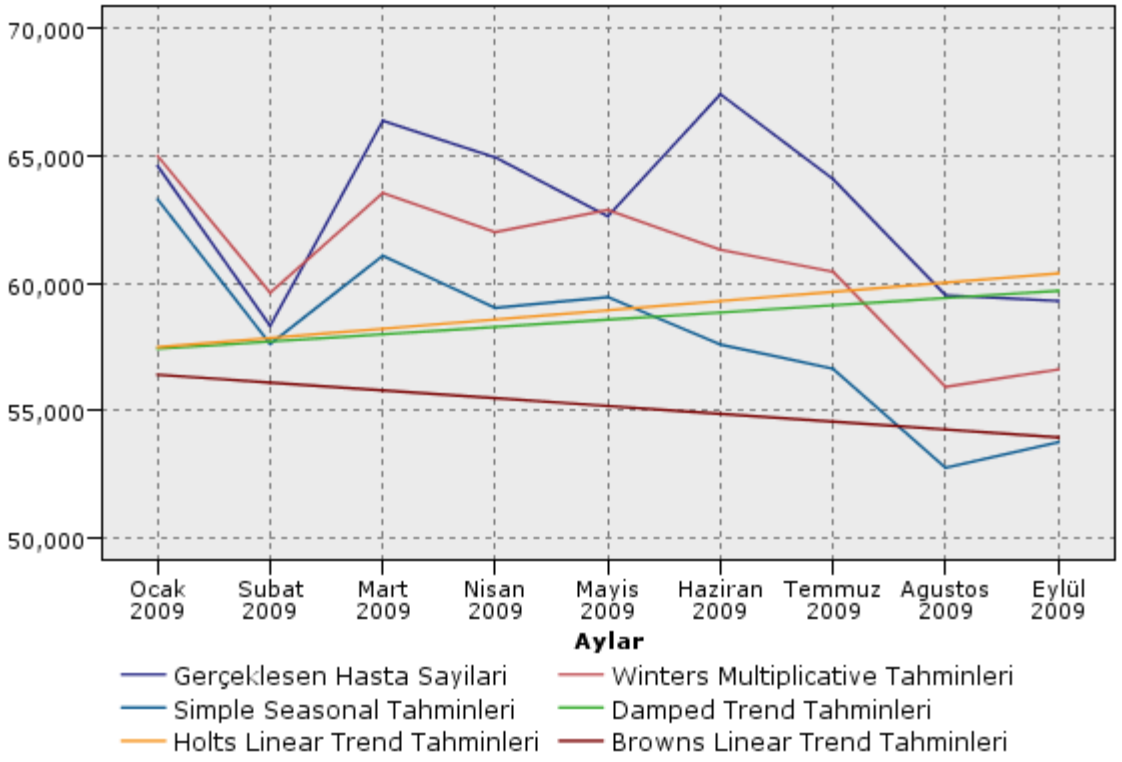


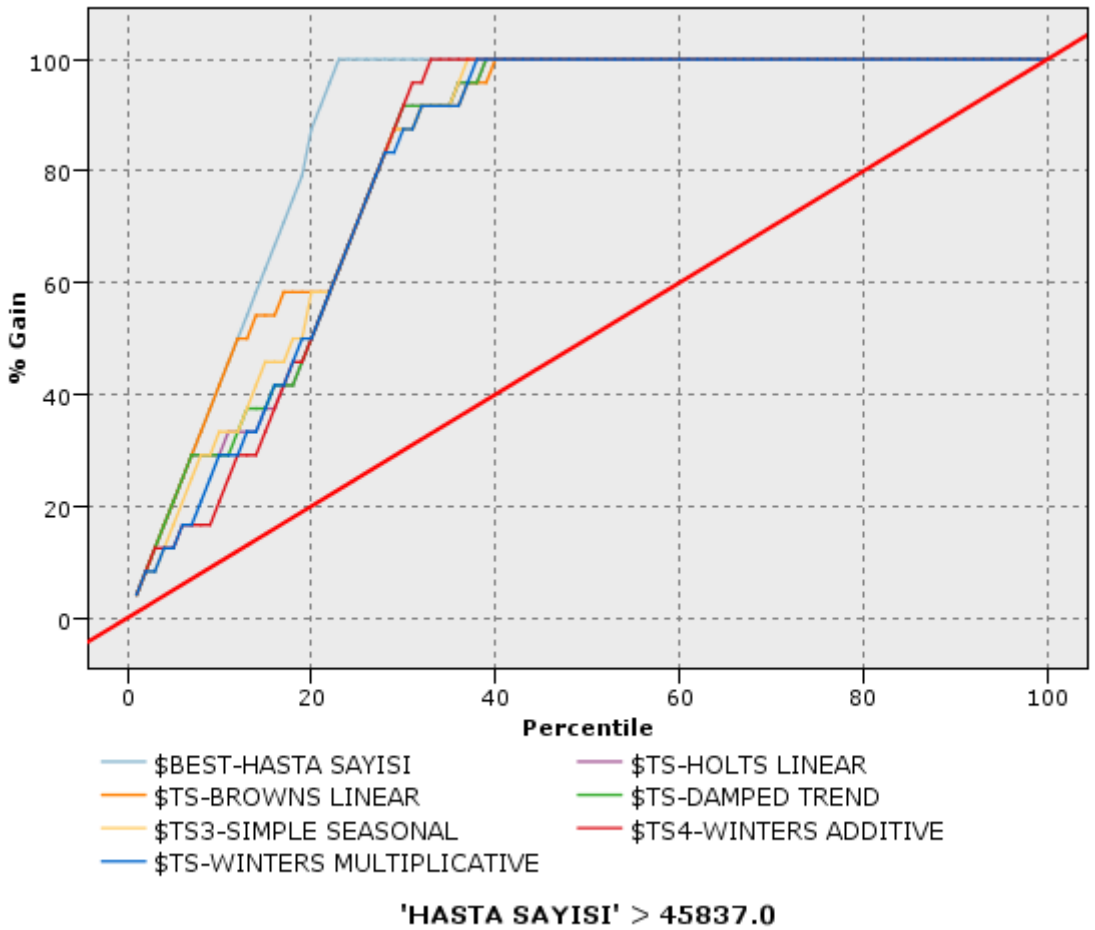
Ek-1(b): Kliniklerde Yatan Hasta Sayıları (Ocak 2005 - Aralık 2008 Arası)



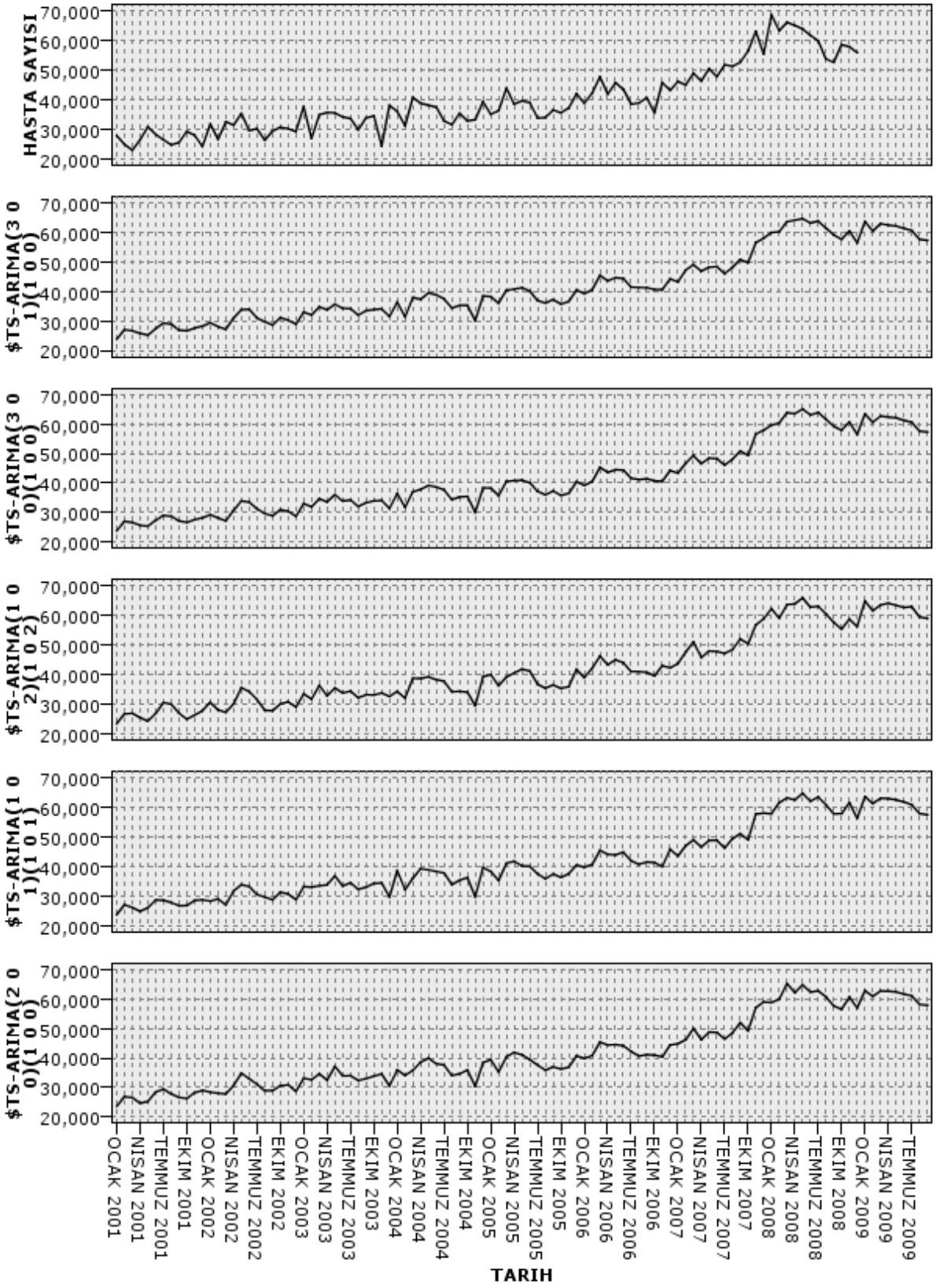
Ek-2: Toplam Hasta Sayısı Tahmini Üstel Düzgünleştirme Modelleri Sonuçları

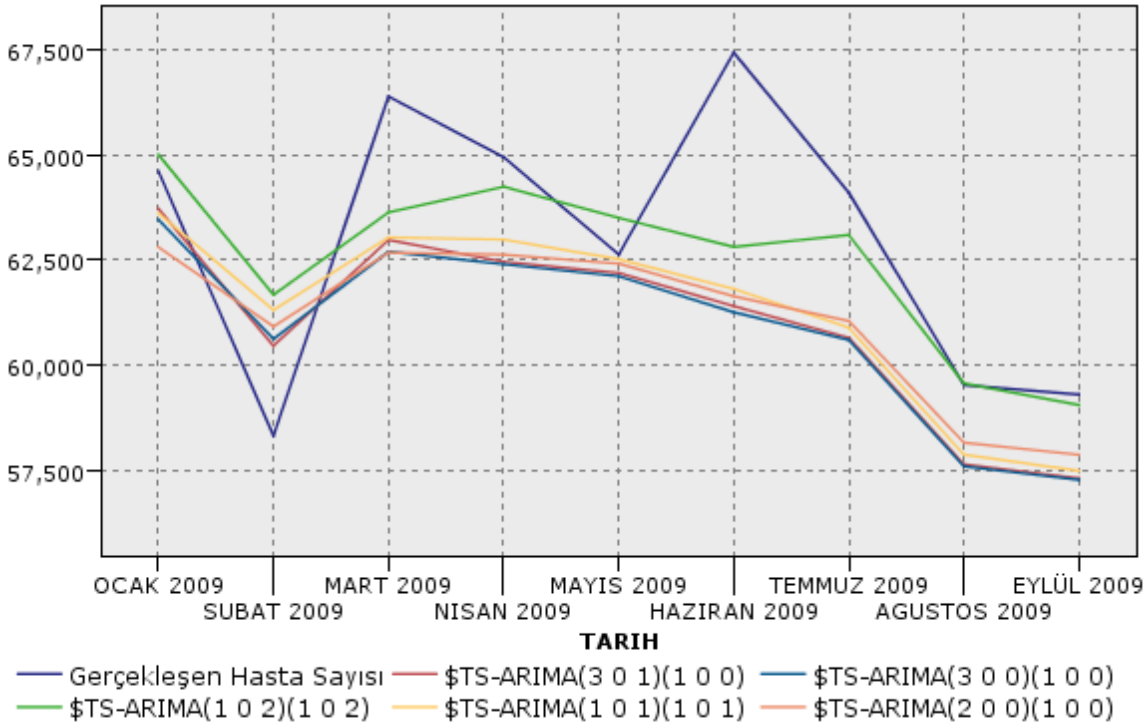


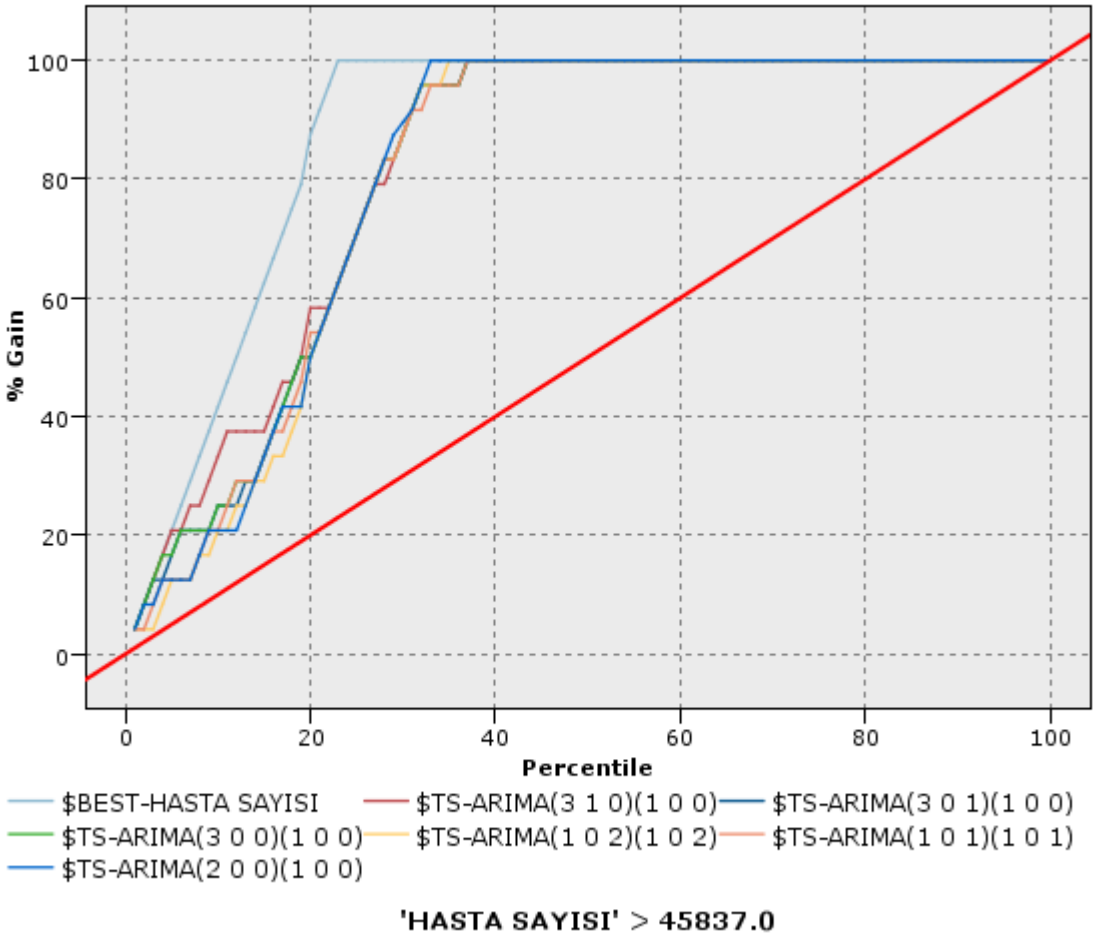

70,000
65,000
60,000
55,000
50,000
Ocak 2009
Subat 2009
Mart 2009
Nisan 2009
Mayis 2009
Haziran 2009
Temmuz 2009
Agustos 2009
Eylül 2009
Aylar
Gerçeklesen Hasta Sayilari
Simple Seasonal Tahminleri
Holts Linear Trend Tahminleri
Winters Multiplicative Tahminleri
Damped Trend Tahminleri
Browns Linear Trend Tahminleri


% Gain
100
80
60
40
20
0
0
20
40
60
80
100
Percentile
$BEST-HASTA SAYISI
$TS-BROWNS LINEAR
$TS3-SIMPLE SEASONAL
$TS-WINTERS MULTIPLICATIVE
$TS-HOLTS LINEAR
$TS-DAMPED TREND
$TS4-WINTERS ADDITIVE
'HASTA SAYISI' > 45837.0

Ek-3: Toplam Hasta Sayısı Tahmini ARIMA Modelleri Sonuçları



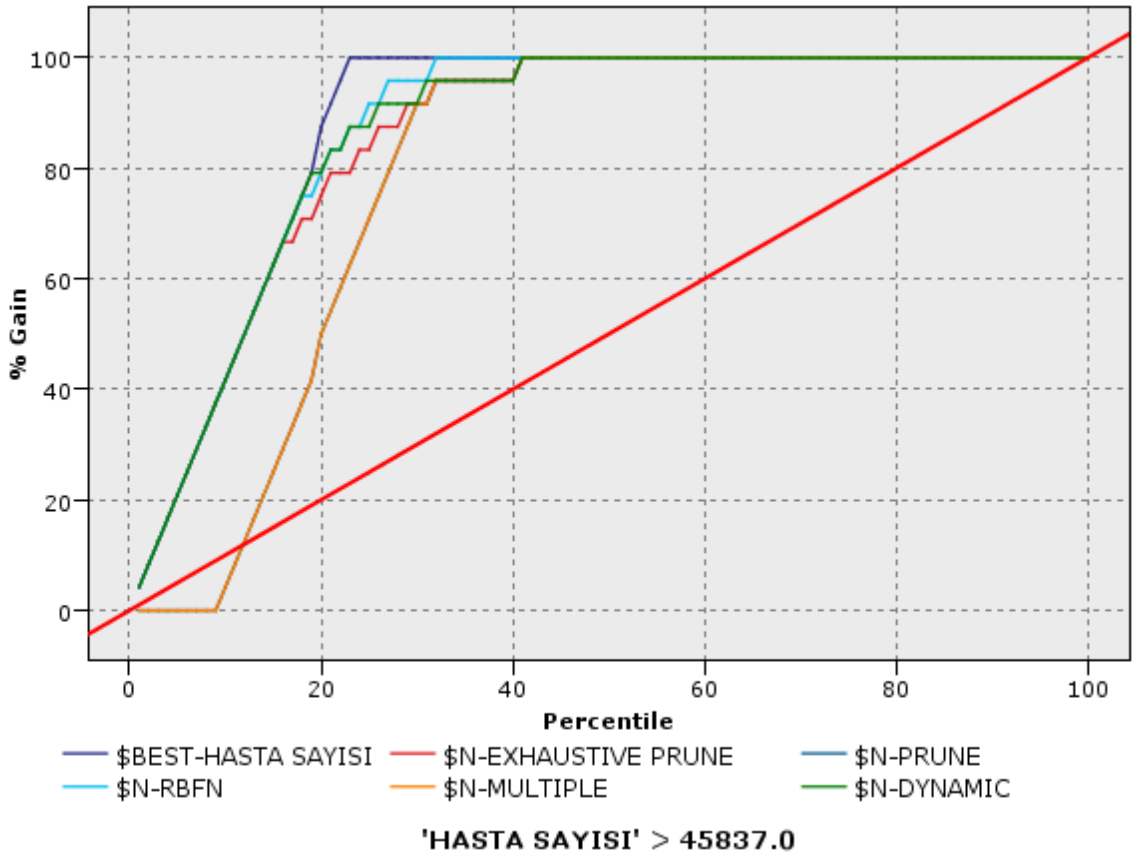

67,500
65,000
62,500
60,000
57,500
OCAK 2009
SUBAT 2009
MART 2009
NISAN 2009
MAYIS 2009
HAZIRAN 2009
TEMMUZ 2009
AGUSTOS 2009
EYLÜL 2009
TARIH
Gerçekleşen Hasta Sayısı
$TS-ARIMA(3 0 1)(1 0 0)
$TS-ARIMA(3 0 0)(1 0 0)
$TS-ARIMA(1 0 2)(1 0 2)
$TS-ARIMA(1 0 1)(1 0 1)
$TS-ARIMA(2 0 0)(1 0 0)


100
80
60
40
20
0
% Gain
0
20
40
60
80
100
Percentile
$BEST-HASTA SAYISI
$TS-ARIMA(3 1 0)(1 0 0)
$TS-ARIMA(3 0 1)(1 0 0)
$TS-ARIMA(3 0 0)(1 0 0)
$TS-ARIMA(1 0 2)(1 0 2)
$TS-ARIMA(1 0 1)(1 0 1)
$TS-ARIMA(2 0 0)(1 0 0)
'HASTA SAYISI' > 45837.0

ARIMA Modelleri 2009 Yılı Hasta Sayısı Tahminleri

	Aylar								
Modeller	Ocak 2009	Şubat 2009	Mart 2009	Nisan 2009	Mayıs 2009	Haziran 2009	Temmuz 2009	Ağustos 2009	Eylül 2009
ARIMA(1,0,0)(0,0,0)	56677	57195	57647	58059	58448	58824	59192	59555	59916
ARIMA(1,0,0)(1,0,0)	63880	62138	64148	64085	63905	63096	62479	59634	59297
ARIMA(0,0,1)(0,0,0)	56867	57576	57936	58297	58658	59019	59380	59741	60102
ARIMA(0,0,1)(0,0,1)	63972	63204	64228	64185	63435	63045	61475	59144	58755
ARIMA(1,0,1)(0,0,0)	57276	57581	57892	58208	58528	58851	59178	59508	59840
ARIMA(1,0,1)(1,0,1)	63616	61317	63040	62989	62524	61815	60897	57881	57497
ARIMA(0,1,1)(0,0,0)	57837	58303	58772	59244	59719	60197	60677	61160	61647
ARIMA(0,1,1)(0,1,1)	65783	61715	64802	62897	62970	60486	59426	54635	54048
ARIMA(1,1,1)(0,0,0)	57963	57950	58562	58977	59459	59923	60396	60869	61346
ARIMA(1,1,1)(1,1,1)	64312	59954	63035	61723	61753	59213	57135	52388	52240
ARIMA(2,0,0)(1,0,0)	62811	60923	62677	62630	62417	61641	61054	58165	57879
ARIMA(2,0,0)(2,0,0)	63573	61512	63030	63346	62513	61986	60739	57446	56919
ARIMA(0,0,2)(0,0,1)	63341	61705	64683	64586	63692	63236	61319	58728	58252
ARIMA(0,0,2)(0,0,2)	64391	62588	65746	65822	64867	64048	63800	60440	59857
ARIMA(2,0,1)(2,0,1)	62951	60713	62394	62317	62104	61329	60868	57959	57657
ARIMA(1,0,2)(1,0,2)	65005	61677	63634	64244	63502	62814	63096	59575	59060
ARIMA(2,0,2)(1,0,1)	64980	62746	64957	64939	64460	63563	62519	58964	58441
ARIMA(2,0,2)(2,0,2)	66464	62415	64428	66088	64708	64469	65167	61451	60920
ARIMA(3,0,0)(1,0,0)	63461	60629	62701	62401	62117	61262	60605	57603	57283
ARIMA(3,1,0)(1,0,0)	63398	59234	62148	60708	60212	59197	57782	54439	53728
ARIMA(3,0,1)(1,0,0)	63722	60464	62979	62462	62194	61409	60656	57650	57318
ARIMA(1,0,3)(0,0,1)	64150	62785	64085	64188	63373	62990	60928	58237	57749
ARIMA(0,1,3)(0,0,1)	62454	60899	61959	61810	60843	60332	58290	55599	55066
ARIMA(0,0,3)(0,0,1)	66006	63070	63595	64497	63645	63207	61394	58878	58426
ARIMA(3,0,3)(1,0,1)	65617	62355	64580	64484	64060	63082	62254	58785	58336
Gerçekleşen Hasta Sayıları	64615	58330	66385	64946	62632	67427	64100	59534	59310

Ek-4: Toplam Hasta Sayısı Tahmini Yapay Sinir Ağı Modelleri Sonuçları



Yapay Sinir Ağı Modelleri 2009 Yılı Hasta Sayısı Tahminleri

		Hasta Sayısı Tahminleri				
Aylar	Gerçekleşen Hasta Sayıları	Exhaustive Prune	Prune	RBFN	Multiple	Dynamic
Ocak 2009	64615	53744	56839	51033	55101	51323
Şubat 2009	58330	52156	56906	49295	55273	50004
Mart 2009	66385	50694	56966	47818	55438	48816
Nisan 2009	64946	49392	57018	46670	55598	47766
Mayıs 2009	62632	48263	57065	45783	55752	46852
Haziran 2009	67427	47306	57106	44910	55900	46063
Temmuz 2009	64100	46506	57142	43617	56043	45387
Ağustos 2009	59534	45847	57174	41333	56181	44810
Eylül 2009	59310	45309	57202	37437	56314	44317

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı ve Soyadı : Sezgin IRMAK

Doğum Tarihi ve Yeri: 04.01.1978 – Serik/Antalya

Adres : Yeni Mah. 2453. Sk. No:21 07060 ANTALYA

Telefon : 0.242.3101845

E-Posta : sezgin@akdeniz.edu.tr
sezginirmak@yahoo.com

Yabancı Dil(ler) : İngilizce

Eğitim Durumu

Doktora : Akdeniz Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı `2009
Tez Konusu: "Veri Madenciliği Yöntemleri ile Sağlık Sektörü Veritabanlarında Bilgi Keşfi: Tanımlayıcı ve Kestirimci Model Uygulamaları"

Yüksek Lisans: Akdeniz Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı `2004
Tez Konusu: "İlişkisel Veritabanı Yönetim Sistemler ve Akdeniz Üniversitesi S.B.E. Öğrenci İşleri Otomasyonu Uygulaması"

Lisans : Marmara Üniversitesi, Elektronik ve Bilgisayar Eğitimi (İngilizce) `2000

Lise : Antalya Teknik Lisesi, Bilgisayar Bölümü `1995

İş Durumu

2001 - ~ : Akdeniz Üniversitesi, Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Araştırma Görevlisi