KOCAELİ ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

CRISP-DM YÖNTEMBİLİMİ KULLANILARAK DENİZ KUVVETLERİ VERİSİ ÜZERİNDE VERİ MADENCİLİĞİ SINIFLANDIRMA TEKNİKLERİNİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS

Erkan KIYAK

Anabilim Dalı: Bilgisayar Mühendisliği

Danışman: Yrd. Doç. Dr. Nevcihan DURU

KOCAELİ, 2006

KOCAELİ ÜNİVERSİTESİ * FEN BİLİMLERİ ENSTİTÜSÜ

CRISP-DM YÖNTEMBİLİMİ KULLANILARAK DENİZ KUVVETLERİ VERİSİ ÜZERİNDE VERİ MADENCİLİĞİ SINIFLANDIRMA TEKNİKLERİNİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

Erkan KIYAK

Tezin Enstitüye Verildiği Tarih: 26 Mayıs 2006
Tezin Savunulduğu Tarih: 06 Temmuz 2006

Tez Danışmanı	Üye	Üye
Yrd.Doç.Dr. Nevcihan DURU	Prof.Dr.A.Coşkun SÖNMEZ	Doç.Dr.Yaşar BECERİKLİ
(.....)	(.....)	(.....)

KOCAELİ, 2006

ÖNSÖZ ve TEŞEKKÜR

Günümüzde gerek kişisel, gerekse kurumsal alanda bilgisayar kullanımının artması ve internetin yaygınlaşması sonucu olarak toplanan veri miktarı, devasa boyutlara ulaşmıştır. Dünyada oluşturulan veri miktarının her yirmi ayda bir ikiye katlandığı da hesaba katılırsa, bu olağanüstü büyük veriden ihtiyaç duyulan bilgiler elde edilmediği sürece, toplanan verinin hiç bir anlamı kalmayacaktır. İnsanoğlunun, bu kadar büyük veriyi gözle inceleyip bilgiler oluşturmasına da imkan yoktur. Bunu sağlayabilecek olan teknoloji veri madenciliğidir.

Naïve Bayes yönteminin kökleri 1760'lı yıllara kadar uzanmaktadır. Sınıflandırma alanında kullanılan bu yönteme, hem basit çalışma prensibi olması ve hem de karar ağacı ve yapay sinir ağları gibi yöntemlerle veri madenciliği çalışmalarının ne kadar ilerlediğini göstermesi bakımından bu tez çalışmasında yer verilmiştir.

Bu tez çalışmasında, Deniz Kuvvetleri Komutanlığı'ndaki giyecek siparişleri verisi, Naïve Bayes, karar ağacı ve yapay sinir ağı yöntemleriyle modellenerek, mevcut veri için en uygun sınıflandırma yönteminin bulunması amaçlanmıştır.

Bana veri madenciliği konusunda çalışma fikrini veren, çalışmam süresince fikir ve yapıcı eleştirileriyle desteğini esirgemeyen değerli hocam Sayın Yrd.Doç.Dr.Nevcihan DURU'ya teşekkür ederim.

Ayrıca çalışmam ve tüm hayatım boyunca hep yanımda olan değerli eşim Ayşen KIYAK başta olmak üzere tüm aileme teşekkür ederim.

Son söz olarak, bu tez çalışması esnasında dünyaya gelen biricik evladım Ege KIYAK'tan, tez çalışmaları nedeniyle kendisiyle ilgilenmem gereken zamandan çaldığım için özür dilerim.

Mayıs 2006, KOCAELİ Erkan KIYAK

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER DİZİNİ

Semboller

Pr	: Olasılık
E	: Durum
H	: Olay
S	: Destek (Support)
C	: Güven (Confidence)
MSE	: Ortalama Kareler Hatası (Mean Squared Error)
MAE	: Ortalama Mutlak Hata (Mean Absolute Error)

Kısaltmalar

A.B.D.	: Amerika Birleşik Devletler
OLAP	: Online Analitical Process
V.T.B.K.	: Veri Tabanlarında Bilgi Keşfi
K.D.D.	: Knowledge Discovery in Databases
SQL	: Structered Query Language
RDBMS	: Relational Database Management Systems
CRISP-DM	: Cross Industry Standart Process for Data Mining
SEMMA	: Sample, Explore, Modify, Model, Assess
CHAID	: Chi-squared Automatic Interaction Detection
CART	: Classification and Regression Trees
JDM	: Java Data Mining
JSR	: Java Specification Request
JSP	: Java Server Pages
ARFF	: Attribute Relation File Format
JVM	: Java Virtual Machine
WEKA	: Waikato Environment for Knowledge Analysis
MLP	: Multilayer Perceptron

CRISP-DM YÖNTEMBİLİMİ KULLANILARAK DENİZ KUVVETLERİ VERİSİ ÜZERİNDE VERİ MADENCİLİĞİ SINIFLANDIRMA TEKNİKLERİNİN KARŞILAŞTIRILMASI

Erkan KIYAK

Anahtar Kelimeler: Veri Madenciliği, Sınıflandırma, Naive Bayes, Karar Ağacı, Yapay Sinir Ağı, CRISP-DM.

Özet: Bu çalışmada, Deniz Kuvvetleri Komutanlığı giyecek sipariş sisteminin iyileştirilmesi amaçlanmıştır. Veri madenciliğinin tüm sürecini belirli bir disiplin altına alan CRISP-DM yöntembilimi kullanılarak, işin anlaşılması, verinin anlaşılması, verinin temizlenmesi, modelleme, değerlendirme ve gerçekleme adımları hazırlanmıştır. Modelleme adımında, veri madenciliği sınıflandırma yöntemlerinden olan Naive Bayes, karar ağacı ve yapay sinir ağları modelleme yöntemleri karşılaştırılarak, Deniz Kuvvetleri verisi için en uygun yöntemin belirlenmesi hedeflenmiştir. Kredili sistemlerde veri madenciliği yapılmasına, herhangi bir veri madenciliği literatüründe karşılaşılmadığından, bu tez çalışmasının, kredili sistemler üzerine yapılacak veri madenciliği çalışmalarında yol gösterici olduğu değerlendirilmektedir. Diğer yandan, CRISP-DM yöntembiliminin adımlarının izlenmesi, bir veri madenciliği çalışmasının, sadece modellemeden oluşmadığını göstermesi açısından önemlidir.

COMPARISON OF DATA MINING CLASSIFICATION ALGORTIHMS ON TURKISH NAVY DATA BY USING CRISP-DM METHODOLOGY

Erkan KIYAK

Keywords: Data Mining, Classification, Naïve Bayes, Decision Tree, Artificial Neural Network, CRISP-DM.

Abstract: The objective of this thesis work is to improve the "Clothing Ordering System" of Turkish Navy. The phases of business understanding, data understanding, data preparation, modeling, evaluation and deployment have been prepared by using CRISP-DM Methodology which holds the process of data mining under a certain dicipline. In modeling phase, data mining classification algorithms such as Naive Bayes, Decision Tree and Artificial Neural Network modeling techniques have been compared in order to find the best technique for Turkish Navy data. It is evaluated that this thesis work will be a guide for data mining studies on credit based systems, since no data mining implement in credit based systems has been encountered in data mining literature. On the other hand, following the phases of CRISP-DM Methodology is important for demonstrating that a data mining study is not only composed of modeling.

1. GİRİŞ

A.B.D.'nin bir önceki başkanı olan Bill Clinton, 6 Kasım 2002 tarihli konuşmasında; FBI ajanlarının 11 Eylül 2001 saldırılarının hemen sonrasında tüketici verisi üzerinde yaptıkları analizler sonucunda, 11 Eylül saldırılarına katılan teröristlerden birinin toplam bakiyesi 250,000 $ olan 30 farklı kredi kartına sahip olduğunu ve A.B.D.'de 2 yıldan az bir zamandır yaşadığını, ayrıca, teröristlerin lideri olan Muhammed Atta'nın A.B.D.'de 12 farklı adresi olduğunu ve bunlardan 2 tanesinin kendisine ve kalan 10 tanesinin güvenilir yandaşlarına ait olduğunu tespit etiklerini belirtmiştir. Ayrıca, Bill Clinton, konuşmasının sonunda, bu tarz verinin, felaketler olmadan önce iyi analiz edilip, önlemler alınması gerektiğini belirtmiş ve;

"Eğer bir insan bir kaç yıldır bu ülkede oturuyorsa ve 12 farklı adresi varsa, ya gerçekten çok zengindir ya da kötü niyetli bir kişidir. Bunlardan hangisi olduğunu bulmak, o kadar da zor olmasa gerek"

demiştir. Aslında bu konuşmalarda anlatılan ve tarif edilen veri madenciliğidir (Larose, 2006).

Veri madenciliği, "ZDNET News" teknoloji dergisi tarafından önümüzdeki on yılın en devrimci gelişmelerinden birisi olarak gösterilmiştir. "MIT Technology Review" dergisi ise veri madenciliğini dünyayı değiştirecek ilk on yeni teknolojiden biri olarak göstermiştir (Konrad, 2001). Veri madenciliği üzerine son 10 yıl içerisinde çok fazla miktarda çalışmalar yapılmış ve yayınlanmış olması bunun açık bir kanıtıdır.

İlk olarak veri analizi ile başlayan çalışmalar sonucunda veri madenciliği teknolojisine ulaşılmıştır. Aslında veri madenciliği, 1960'lı yıllarda IBM ve CDC gibi firmalar tarafından o günün teknolojisi olan devasa boyutlu ve bugünkü işlemcilerle kıyaslanamayacak kadar yavaş çalışan bilgisayarlar aracılığıyla, kasetler ve diskler üzerinde yazılmış verinin analizi ile başlamıştır. Veri analizinin tarihsel

gelişimi ve sonuçta nasıl veri madenciliği teknolojisine ulaşıldığı, Tablo 1.1'de verilmiştir.

Tablo 1.1: Veri Analizinin Tarihsel Gelişimi (Squier, 2001)

Gelişim Adımı	**Cevap Aranan Soru Örneği**	**Teknoloji**	**Ürün Sağlayıcılar**	**Karakteristiği**
Veri toplama (1960'lı yıllar)	Son 5 yıllık toplam gelirim ne kadar?	Bilgisayar, kaset, disk	IBM,CDC	Geriye dönük, statik veri dağıtımı
Veri Erişimi (1980'li yıllar)	Geçen mart ayında New England'a yapılan birim satışlar nelerdi?	RDBMS, S.Q.L., ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Geriye dönük, kayıt seviyesinde dinamik veri dağıtımı
Veri ambarı ve karar destek (1990'lı yıllar)	Geçen mart ayında New England'a yapılan birim satışlar nelerdi? Boston'a yapılan satışlara göre değerlendir.	OLAP, veri ambarları	SPSS, Comshare, Arbor, Cognos, Microstartegy, NCR,	Geriye dönük, çoklu kayıt seviyesinde dinamik veri dağıtımı
Veri Madenciliği (2000'li yıllar)	Önümüzdeki ay Boston'daki birim satış miktarı ne olur?Nedeni?	Gelişmiş algoritmalar, çok işlemcili bilgisayarlar, devasa veri tabanları	SPSS, Lockhead, IBM, SGI,SAS,NCR	İleriye dönük, proaktif bilgi dağıtımı

Veri madenciliği üzerine basılan yayınlara bakılacak olursa; basılan ilk kitap Piatetsky-Shapiro ve Frawley (1991) tarafından yazılmış olup, 1989 yılında gerçekleştirilen bir seminerdeki makalelerin bir araya getirilmesi ile oluşturulmuştur. 1994 yılında yapılan bir seminere ait makalelerden yola çıkarak hazırlanan bir diğer kitap Fayyad ve diğ. (1996) tarafından yazılmıştır.

Veri madenciliği ile ilgili olarak yayınlanan sonraki kitaplar, işin teorisinden çok pratiğiyle ilgili ve doğrudan iş odaklı olarak hazırlanmıştır. Bunların arasında

Syllogic firmasından Adriaans ve Zantige veri madenciliği hakkındaki ilk çalışmalardan birini hazırlamışlardır. IBM firmasında çalışan Cabena ve diğ. (1998) tarafından yapılan çalışmada ise gerçek yaşamdaki uygulama örnekleri ile birlikte veri madenciliği süreç ve yöntemleri incelenmiştir. Dhar ve Stein (1997) ise, veri madenciliği yöntemlerini iş hayatı bakış açısı ile irdelemişlerdir.

Veri madenciliği üzerine yazılım geliştiren bir firmada çalışan Groth (1998), veri madenciliği üzerine yapılan yazılım ürünlerinin detaylı bir incelemesini yapmıştır. Weiss ve Indurkhya (1998), büyük veriden hareketle ileriye dönük tahminlerde bulunmak için kullanılması gereken istatistiksel yöntemlerler konulu geniş bir çalışma yapmışlardır. Han ve Kamber (2001), büyük ve birleştirilmiş veri tabanlarında bilgi keşfi konusuna odaklı olarak veri madenciliği konusunu, veri tabanı bakış açısı ile incelemişlerdir. Han ve diğ. (2001), veri madenciliği konusunda söz sahibi yazarların çalışmalarından disiplinler arası bir kitap oluşturmuşlardır.

Mohammadian (2004); yazmış olduğu kitabında, hem internet, hem de veri tabanlarının olağanüstü büyümesi nedeniyle iyice karmaşıklaşan anlamlı bilgilerin elde edilmesi süreci için akıllı sistemlere ihtiyaç duyulduğunu belirtmiştir. Veri tabanlarında ve internette anlamlı bilgilerin araştırılması ve elde edilmesi için akıllı ajanların (intelligent agent) kullanılması konusunu irdelemiştir. Bu amaçla, dünya üzerinde akıllı ajanlar konusunda çalışmalar yapan uluslararası araştırmacıların çalışmalarını kitabında toplamıştır.

Soukup ve Davidson (2002); ham verinin, işletmelerin yararlanabileceği veri kümeleri haline dönüştürülmesini ve sonrasında bu veri kümelerinin görsel veri madenciliği yöntemleri kullanılarak analiz edilmesini incelemişlerdir. Kitabın yazarlarının görüşüne göre, görselleştirme, diğer işletme zekası (business intelligence) yöntem ve araçlar ile kıyaslandığında, veri içindeki bilinmeyen örüntü ve kural dışılıkları bulma süresini son derece azaltmaktadır. Sonuçta, resim sanatçıları, binlerce kelimeyle anlatılabilecek duyguları, bir tek resime sığdırabilmektedir. Kitaplarını üç bölüm halinde hazırlamışlardır. Birinci bölümde;

tanıtım ve proje planının çıkartılması, ikinci bölümde; verinin hazırlanması ve son bölümde; verinin analiz edilmesini anlatmışlardır.

Keim (2004); hazırlamış olduğu sunumda, özellikle görsel veri keşfi sürecinde kullanılan yöntemleri bir araya getirmiştir. İncelediği görsel veri keşfi teknikleri arasında; geometrik teknikler, ikona tabanlı teknikler, piksel tabanlı teknikler, hiyerarşik teknikler, grafik tabanlı teknikler ve hibrid teknikler vardır.

Venkayala (2005); Java geliştiricileri dergisinde yayınlanan makalesinde, Java Data Mining (JDM)1.0 standardını açıklamıştır. Kendisi, JSR-73 altında geliştirilen JDM standardı uzman geliştiricilerindendir. Venkayala, makalesinde, Java yazılım dili ile veri madenciliği yapabilmenin standardı olan JDM'in pratikte nasıl kullanılabileceği konusunu detaylı olarak incelemiştir.

Wang (2003); 2001 yılında konusunda uzmanlaşmış kişilere yaptığı çağrısı sonucunda, veri madenciliği ile ilgili yeni teorilerden uygulamalara kadar çok geniş bir yelpazede topladığı makaleler üzerinde yaptığı bir buçuk yıllık titiz bir çalışma sonucunda bir kitap oluşturmuştur. Kitabın ana amacı, yeni yöntemler ve uygulama alanları ile ilgili yayınlanan bir çok eseri bir araya getirmektir. Böylelikle konu hakkında araştırma veya uygulama yapan bilim adamları, işletmeler, öğrenciler ve yöneticiler gibi çok geniş bir kesime yol gösterebilmektir. Kitapta 7 ülkeden toplam 44 uzmana ait makaleler mevcuttur.

Wang (2006); 2003 yılında hazırlamış olduğu ve çeşitli makaleleri topladığı çalışmasından sonra, 2006 yılında hazırlamış olduğu geniş içerikli kitabında, veri madenciliği ve veri ambarı konusunda uzmanlaşmış, toplam 358 uluslararası araştırmacının makalelerine yer vermiştir. 2 ciltten oluşan kitabın, 2006 basımı olması, özellikle veri madenciliği ve veri ambarı konusunda yapılan en son araştırma ve teknikleri içermesi açısından önemlidir.

Freeman ve Skapura (2001); yapay sinir ağları (artificial neural networks) konusunda detaylı bir kitap hazırlamışlardır. Çalışmalarında işledikleri konular özetlenecek

olursa; yapay sinir ağlarının genel tanıtımı, sayısal sinyal işleme (Digital Signal Processing, DSP)'de yazılımsal filtreleme metodu olan "Adeline" filtresi ve bunun çoklu kullanımı olan "Madaline" filtresi ve geri beslemeli (back propogation) yapay sinir ağları başta olmak üzere yapay sinir ağları ile ilgili tüm yöntemler incelenmiştir.

Tang ve MacLennan (2005); SQL sunucu yüklü ortamlarda veri madenciliğinin nasıl yapılabileceği konusunu incelemişlerdir. Microsoft firmasının bir ürünü olan ve dünya üzerinde yoğun olarak kullanılan Microsoft SQL sunucu veri tabanlarında, veri madenciliği tekniklerinden olan Naive Bayes, karar ağaçları, zaman serileri, kümeleme, birliktelik kuralları, yapay sinir ağları ile veri madenciliği yapılması anlatılmıştır. Kitaplarında, veri madenciliği yöntemlerinin yanında, OLAP küpleri ve veri madenciliği yazılımları hakkında bilgiler verilmiştir.

Kasabov (1998)'e göre, insan zekasının bilgisayarlara uyarlanması için bilim adamlarınca çeşitli yöntemler denenmiştir. Yapay zeka, sembolleri işleyerek mantıksal çıkarımlar yapar. Bulanık (Fuzzy) sistemler ise akıllı ve etkili çıkarımlar yapmak için örneksel (analog) girdiler kullanır. Her ikisi de insan zekasının semboller ve kurallar seviyesinde anlaşılabilmesi için büyük ölçekte yöntemlerdir. Yapay sinir ağları ise nöronların etkileşimlerinden yola çıkılarak tasarlanan küçük ölçekte bir yöntemdir. Tüm bu yaklaşımların hepsi, insan beynini temsil etmekte, kısmen başarılıdırlar. Yapay zeka, matematiği kullanıyor olsa da değişik şartlara uyum problemi olan ve gerçek hayata uyarlanması zor bir yöntemdir. Bulanık sistemler ise, her türlü ortama kolay uyum sağlayacak nitelikte çıkarımlar yapsa da, çıkarımların tam olarak kesinliği ve ayrıntısı konularında zayıftırlar. Yapay sinir ağları öğrenme ve kendi başlarına hareket etme yeteneklerine sahiptir ama, diğer yandan sembolik çıkarımların çözümünde başarısızdır. Önemli olan nokta, insan zekasının mümkün olduğa kadar iyi bir şekilde bilgisayar ortamına nasıl uyarlanabileceğidir ve Kasabov kitabında, bu üç metodun zayıf yönlerini azaltacak ve güçlü yönlerini artıracak şekilde nasıl birleştirilebileceği konusuna odaklanmıştır. Kitapta anlatılan konular, birçok gerçek dünya örneğiyle, biraz daha pekiştirilmiştir.
Weiss ve diğ. (2005); veri madenciliği üzerine yapılan en önemli çalışma atölyelerinden biri olan K.D.D. 2005'te sunulan çalışmaları, eserlerinde

toplamışlardır. Atölyede, hem veri madenciliği, hem de makine öğrenmesi üzerine bir çok uygulamalar sunulmuştur. Atölyeler, uygulama tabanlı olduğu için işletmelerin veri madenciliğinden maksimum fayda sağlamaları için son derece faydalı olmaktadırlar.

Mattison (1997); son yıllarda gelişen telli iletişim alanında veri ambarlama ve veri madenciliği araştırmaları yapmıştır. Kitabında, veri ambarı ve veri madenciliği kullanılarak, iletişim alanında faaliyet gösteren işletmelerdeki değerlerin tanımlanması ve yaratılması konularına odaklanmıştır. Veri yöntemlerinden olan yapay sinir ağları ve coğrafi veri madenciliği ile telli iletişim alanında çalışan işletmelerin değer ve bilgi elde etme yollarını çeşitli uygulamalarla açıklamıştır.

Keogh ve diğ. (2004); değiştirgesiz (parametresiz) veri madenciliği konusunu ortaya atmışlardır. Çoğunluk veri madenciliği algoritmaları, başlangıçta bir çok değiştirgenin düzgün bir şekilde ayarlanmasına ihtiyaç duymaktadır. Bu değiştirgelerin yanlış şekilde ayarlanması, iki büyük hataya neden olabilir. Bunlardan birincisi, algoritmanın doğru örüntüyü bulamaması, ikincisi ve belki de daha kritik olanı, algoritmaların, gerçekte varolmayan örüntüler bulması veya mevcut örüntülerin önemini, olduğundan çok daha fazla kuvvetliymiş gibi göstermesidir. Veri madenciliğinde değiştirgeler mümkün olduğunca az olmalıdır. Değiştirgeler kullanılmadan yapılan veri madenciliği çalışmaları, önyargıların, beklentilerin ve tahminlerin ortadan kalkmasını sağlayarak, sadece verinin kendisinin konuşmasını sağlayacaktır. Keogh ve diğ. makalelerinde değiştirgesiz veri madenciliğinin nasıl yapılabileceğini göstermişlerdir.

Mitra ve Acharya (2003); veri madenciliğinin sınıflandırma, kümeleme ve benzer gruplama gibi geleneksel kavram ve fonksiyonlarının yanında, özellikle çoklu ortam (multimedia) ve bilgisayar destekli biyoloji (bioinformatics) alanlarında veri madenciliği yapılması konularına odaklanmışlardır. İnternet kullanımının giderek yaygınlaşması ve internet ortamında çoklu ortam uygulamalarının yoğun bir şekilde kullanılıyor olması, veri madenciliği açısından bir çok yeni araştırma konuları çıkaracaktır. Bu kadar büyük verinin çoğunlukla sıkıştırılarak kullanılması nedeniyle,

kitabın bir bölümünde sıkıştırılmış uygulamalarda veri madenciliği yapılması konusu incelenmiştir. Kitap; metin, imge ve internet ortamında veri madenciliği yapma yöntemlerini detaylı olarak irdelemiştir.

Berry ve Linoff (2004), ilki 1997 yılında yayımlanan kitaplarının ikinci basımında veri madenciliği konusunu genel olarak üç ayrı kısımda incelemişlerdir. Birinci kısımda; veri madenciliğini tanıtan ve niçin gerekli olduğunu vurgulayan bölümü de içerecek şekilde işletmeler açısından veri madenciliğinin anlamı anlatılmıştır. İkinci bölümde; verinin bilgi haline getirilmesi için hangi durumlarda, hangi veri madenciliği tekniklerinin kullanılması gerektiği detaylandırılmıştır. Üçüncü ve son bölümde ise; veri madenciliği yöntemleri ile ilgili en iyi uygulama alanları, örneklendirilerek anlatılmıştır. Veri madenciliği konusu son yıllarda, akademik ortamlarda olduğu kadar işletmelerde de tartışılır ve uygulanır olmuştur. İşletmelerin veri madenciliğine başlaması için gerekli başucu kitaplardan bir tanesidir.

Pyle (2003); veri madenciliği konusunu işletmeler açısından ele alan bir başka araştırmacıdır. Işletmelerin, iyi bir veri madenciliği yapabilmek için nereden başlamaları ve sonrasında neler yapmaları konusunu incelemiştir. İşletmelerin elinde bol miktarda veri ve yine bol miktarda problem sahaları vardır. Diğer tarafta ise, ellerindeki veriyi işleyerek bilgiler oluşturacak ve problemleri çözecek veri madenciliği teknikleri ve araçları vardır. Öyleyse sorun, hangi tür veri ve problemler için, hangi tür teknik ve araçların kullanılması ile ilgilidir. Pyle, bu konulara 4 bölümde açıklık getirmeye çalışmıştır. Birinci bölümde; mevcut durumun ve çevre şartlarının bir haritası çıkartılmıştır. Ikinci bölümde; işletme modelini, üçüncü bölümde; veri madenciliği ve yöntemlerini, dördüncü bölümde ise; kendisinin hazırlamış olduğu bir veri madenciliği yöntembilimini anlatmıştır.

Witten ve Frank (2005), Yeni Zellanda'nın Waikato Üniversitesi bilgisayar bilimleri bölümünde çalışan iki öğretim görevlisidir. Yazmış oldukları kitap iki açıdan çok önemlidir. Birincisi, veri madenciliği konusuna uygulanabilirlik açısından baktıklarından bir başucu kitabı olmasıdır. İkincisi ise, bu kitabın yazarlarının, veri madenciliği arenasında ve özellikle akademik ortamlarda çok sıklıkla kullanılan

WEKA aracını oluşturmuş olmalarıdır. Bu tez çalışmasının son bölümündeki uygulama, açık kaynak kodlu WEKA veri madenciliği aracı ile yapılmıştır.
Bu tez çalışmasında, CRISP-DM (Cross Industry Standart Process for Data Mining) yöntembilimi kullanılarak Deniz Kuvvetleri verisinde, veri madenciliği sınıflandırma yöntemlerinin karşılaştırılması yapılmıştır. Uygulamada, Deniz Kuvvetleri Komutanlığı'nda görev yapan personelin, kendilerine verilen kredi doğrultusunda, siparişini vermiş oldukları giyecekler, veri medenciliği yöntemleri ile analiz edilmiştir. Karar ağaçları, yapay sinir ağları ve Naive Bayes veri madenciliği yöntemleriyle yapılan modellemeler sonucunda, kredi karşılığı alınan malzemenin sınıflandırılması için en uygun olan veri madenciliği yönteminin belirlenmesi amaçlanmıştır. Sınıflandırma işlemi sonucunda, personel tarafından siparişi edilen malzemenin, diğer öznitelikler cinsinden fonksiyonu belirlenmeye çalışılmıştır. Daha açık bir ifadeyle; miktar, dönem, personel tipi, kredi yılı gibi girdi verisinin bilindiği durumlarda, çıktı verisi olan malzeme adının belirlenmesi amaçlanmıştır.

Bu tez beş bölümden oluşmaktadır. Birinci bölümde; veri madenciliği hakkında genel bilgiler verilmiş, literatür taraması yapılmış ve bu tezde ele alınan uygulamanın amacı hakkında genel bilgi verilmiştir.

İkinci bölümde; veri madenciliği ile ilgili detaylı bilgiler verilerek, veri madenciliği yöntemleri hakkında bilgi verilmiştir.

Üçüncü bölümde; en çok kullanılan veri madenciliği yöntembilimi olan CRISP-DM yöntembilimi detaylı olarak anlatılmıştır. Ayrıca, bu bölümde, bu tez çalışmasında karşılaştırılan veri madenciliği yöntemleri olan karar ağacı, Naive Bayes ve yapay sinir ağı yöntemlerinin detaylı incelemesi yapılmıştır.

Dördüncü bölümde, Deniz Kuvvetleri Komutanlığı kredili giyecek sistemi yazılımı aracılığıyla toplanan, sipariş bilgileri verisi, CRISP-DM yöntembilimi ışığında, karar ağaçları, yapay sinir ağları ve Naive Bayes veri madenciliği sınıflandırma yöntemleri kullanılarak analiz edilmiş, yapılan çalışmalarda elde edilen sonuçlara yer verilmiştir.

Beşinci ve son bölümde, bu tez çalışmasında elde edilen sonuçlara ve önerilere yer verilmiştir.

2. VERİ MADENCİLİĞİNE GİRİŞ

2.1. Veri madenciliği ile Veri Tabanında Bilgi Keşfinin Karşılaştırılması

Veri madenciliği teriminin ne anlama geldiğinin detaylarına girmeden önce, veri madenciliği ile veri tabanında bilgi keşfi (V.T.B.K.) terimlerine açıklık getirmek yerinde olacaktır.

V.T.B.K., terim olarak, veri arasından yararlı bilgiler keşfetme sürecidir. Veri madenciliği ise V.T.B.K. sürecinin sadece belirli bir bölümü olup, veri içinde örüntüler bulmak için çeşitli algoritmaların kullanılması (modelleme) işlemidir. V.T.B.K. sürecinde, veri madenciliği anlamına gelen modelleme basamağından önce; yapılacak işin anlaşılması, verinin analiz edilmesi ve anlaşılması, verinin veri madenciliği için hazırlanması işlemleri, modelleme basamağından sonra ise modelleme sonuçlarının değerlendirilmesi ve sonuç olumlu ise model gerçek hayata uyarlanması işlemleri vardır. (Bradlet ve diğ., 1998)'in bu konuda bir de uyarısı vardır;

"Sadece veri madenciliği (modelleme) yapmak, anlamsız örüntüler elde edilmesini sağlayan tehlikeli bir araç haline gelebilir."

Veri madenciliği terimi ilk başlarda, istatistikçiler, veri analistleri ve veri tabanı ile ilgilenenler tarafından kullanılan bir terimdir. 1990'lı yıllarda, yani veri madenciliğinin ilk emekleme yıllarında, V.T.B.K.-Veri Madenciliği ayrımı vardır. (Waiganjo, 2002)'nin bu ayrımı gösterir şekilde oluşturduğu V.T.B.K. denklemi şöyledir;

V.T.B.K. = veri hazırlama + veri madenciliği + keşfedilen örüntü veya ilişkilerin yorumlanması ve gerçek hayata uyarlanması.

İkibinli yıllardan itibaren, gerek V.T.B.K., gerekse veri madenciliği için genel olarak veri madenciliği terimi kullanılmaya başlanmıştır. Günümüzdeki veri madenciliği terimi, işi anlama, veriyi anlama, veriyi hazırlama, modelleme, değerlendirme ve uygulama işlemlerinin bütünü için kullanılmaktadır. Bu tez çalışmasında da, veri madenciliği terimi bu anlamda ele alınmıştır.

2.2. Veri Madenciliği Nedir?

Veri madenciliği ile ilgili olarak yapılan çalışma sayısı ile doğru orantılı olarak bir çok tanımlama mevcuttur. Bu tanımlamalar her ne kadar aynı anlama geliyor olsalar da, bu tanımlamaları toplu halde bir arada görmek, veri madenciliğini kavramak açısından kolaylık sağlayacaktır;

Büyük miktardaki veri içinde, mantıklı, şaşırtıcı, potansiyel olarak yararlı ve anlaşılır örüntüler bulmak için gerekli olan işlemler bütünüdür (Fayyad ve diğ., 1996).

İstatistik ve matematik tekniklerle birlikte örüntü tanıma (pattern recognition) teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilmesi sürecidir (Witten ve Frank, 2005).

Büyük miktarda verinin, öngörülmeyen ilişkiler bulmak amacıyla analiz edilmesi ve sonrasında verinin sahibinin işine yarayacak, anlaşılabilir şekilde, yeni ve farklı bir biçimde özetlenmesidir (Hand ve diğ., 2001).

Büyük miktarda veri içinden, gelecekle ilgili tahmin yapılmasını sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır (Alpaydın, 2000).

Önceden öngörülemeyen ve saklı durumdaki değerli bilgi ya da bilgilerin, eldeki veriden, matematiksel yöntemler ile süzülerek, anlamlı bir şekilde ortaya çıkarılması sürecidir (Alkan, 2003).

Büyük miktardaki verinin, anlamlı örüntüler ve kurallar bulabilmek için, otomatik olarak incelenmesi ve analiz edilmesidir (Berry ve Linoff, 2004).

2.3. Veri, Enformasyon ve Bilgi

Veri, enformasyon ve bilgi terimleri sıklıkla birbirleriyle karıştırılmaktadır. Veri madenciliğini ilgilendiren alanlarda sıkça adı geçen bu terimlerin açıklanmasında yarar vardır;

Veri (Data): Bilgisayarlar tarafından işlem gören herhangi bir olgu, rakam veya metindir. Wong ve Leung (2002), veri için, çok değerli bir hazine benzetmesi yapmıştır.

Enformasyon (information): Veri içindeki örüntüler, birliktelikler veya ilişkiler enformasyon sağlayabilirler. Örneğin, bir işletmenin satış işlemlerine ait verinin analiz edilmesi ile, hangi ürünün ne zaman satıldığı enformasyonu elde edilebilir.

Bilgi (Knowledge): Enformasyonlar, geriye yönelik örüntüler veya ileriye yönelik eğilimler hakkında bilgilere dönüştürülebilirler. Yukarıdaki örnekte elde edilen hangi ürünün ne zaman satıldığı enformasyonu, müşterilere yapılan promosyonlar ışığında analiz edilerek, müşterilerin satın alma davranışları bilgisi elde edilebilir. Böylelikle üretici ya da satıcılar, hangi ürünün promosyona en yatkın ürün olduğunu belirleyebilirler. Özetle enformasyon bizim için bir anlam ifade ediyor ve biz o enformasyonu kullanıyorsak, enformasyon bizim için bilgi olur, aksi taktirde haberden öteye geçemez.

Bu terimler, birbirlerinin yerine kolaylıkla kullanılabilmektedir. Diğer bir problem ise, kavramlar arasındaki bu karmaşa ortadayken, bu terimlerin karşılığı Türkçe terim olmaması veya bulunamayışıdır. Bu konuda Aktaş (2004)'e hak vermemek elde değildir:

"İngilizce (data/information/knowledge) üçlüsünü biz Türkler yıllar önce kestirmeden halledip, 'bilgi' deyip çıktık. Hatırlayın 'data processing' için 'bilgi işlem' dedik. Sonra

‘data’ya ‘veri’, ‘information’a bilgi dedik , ‘management information systems’ için ‘yönetim bilgi sistemleri’ olduğu gibi. Son yıllarda da ‘knowledge‘ için ‘bilgi’, ‘information’ için de ‘haber’ veya ‘enformasyon’ demeye başladık. Aslına bakarsanız o kadar da dert değil bence. Zira, Amerikalılar ve İngilizler, nasıl ‘information’ sözcüğünü (data/information/knowledge) üçlüsü için genel olarak kullanıyorlarsa , biz de ‘bilgi’ sözcüğünü (veri/haber/bilgi) üçlüsü için kullanabiliriz. Zira bu kaygan ve kaypak terimlerin kişiye, ortama ve zamana göre değiştiğini artık biliyoruz. Benim için ‘bilgi’ olan bir mesaj sizin için pekala ‘veri’ olabileceği gibi, aynı mesaj bir süre sonra benim için de bir ‘haber’, hâttâ , ‘veri’ olabilir.”

2.4. Veri Madenciliğine Niçin İhtiyaç Duyulmuştur?

Veri madenciliğinin son yıllarda popüler olmasının bir çok nedeni vardır, bunlardan en önemlileri; firmaların yaptıkları işlerle ilgili olarak daha fazla veri toplamaya başlamış olmaları, verilerin toplanma ve saklanma maliyetlerinin çok hızlı bir şekilde düşmesi ve küreselleşmenin firmalar üzerine olan baskısıdır (Wang, 2003).

Veri madenciliğinin ne kadar yaygın olarak kullanıldığını işaret eden bir örnek olarak; Amerikan profesyonel basketbol ligi (NBA)’de mücadele eden 29 takımdan 16’sının IBM firmasının NBA için özel olarak hazırladığı “Advanced Scout” adlı veri madenciliği yazılımını kullanmaları verilebilir. Hikayenin başlangıcı ise, IBM firmasında veri madenciliği ve veri analizi uygulamaları üzerine çalışan Inderpal Bhandari’nin büyük basketbol tutkusu nedeniyle New York Nicks takımıyla anlaşması ve kendisinin liderliğinde bir yazılım ortaya koymasıdır. Uygulama; basket, şut, pas, rebound, asist sayıları gibi NBA tarafından tutulan veriye dayanmaktadır. Yapılan veri madenciliği sonucunda, takım koçlarının bile fark edemediği örüntüler bulunmuş ve bu durumlara karşı alınan önlemler neticesinde, takım çok daha iyi yerlere gelmiştir. New York Nicks takımının bu başarısının ardından, diğer NBA takımlarının çoğunluğu, aynı veri madenciliği yazılımını kullanmaya başlamışlardır (Larose, 2006).

Bugün işletmeler terabayt (1,000 megabayt) büyüklükte veri tabanlarına sahiptir ve bu muazzam bilgi artışı nedeniyle bu veri tabanlarının önümüzdeki bir kaç yıl içinde petabayt (1,000 terabayt) seviyelerinde veri tabanlarına gereksinim duyulacağı bir gerçektir (Whiting, 2002). Gartner Grubu, 2004 yılında yaratılan verinin, 1999

yılında yaratılanın 30 katı olduğunu ve son 30 yılda yaratılan verinin, ondan önceki 5000 yılda yaratılan veriden daha fazla olduğunu tahmin etmiştir (Wurman, 1989). Kurumsal veri tabanlarındaki bu olağanüstü büyümeye rağmen, IBM firmasındaki bir araştırmacı olan Brown (2002), işletmelerin analiz için ellerindeki verinin %1'inden bile daha azını kullandıklarını açıklamıştır. İçinde yaşadığımız bilgi çağının temel ironisi de budur: işletmeler olağanüstü büyük miktarlarda veriye sahip olmalarına karşın, faydalandıkları gerçek bilgi miktarı bir o kadar küçüktür. 450 üst düzey yönetici arasında yapılan bir araştırma, yöneticilerden %90'ının, ihtiyaç duydukları anda gerekli bilgiye sahip olamadıklarından, içgüdüleriyle hareket ettiklerini göstermiştir (Brown, 2002).

Berkeley Üniversitesi 2000 yılında çok önemli bir araştırma yapmış ve sonuçlarını yayınlamıştır. Aynı araştırma, 2003 yılında tekrarlanmıştır. Araştırmaların amacı, dünya üzerinde yaratılan bilgi miktarını yaklaşık olarak hesaplamaktır. Her iki araştırma sonucunda, bilgi miktarı tahminleri yaratılma ortamlarına göre Tablo 2.1'de verilmiştir. Araştırma sonuçlarına göre, 2002 yılında yaratılan toplam veri miktarı 5 eksabayt (5 milyon terabayt) olarak tahmin edilmiştir. Oluşturulan bu verinin ortalama olarak yarısı, stratejik iş uygulamaları veya karar destek sistemleri ihtiyacı kaynak bilgilerden oluşmaktadır (Kestelyn, 2002).

Berkeley üniversitesinin araştırmasına göre bilgi; kağıt, film, optik ve manyetik olmak üzere dört farklı fiziksel ortamda kaydedilmekte, saklanmakta ve dağıtılmaktadır. Veri madenciliği, genel olarak manyetik ortam altındaki sabit disklerde saklanan veri ile ilgilidir. Aynı araştırmanın sabit disklerle ilgili olarak yayınladığı miktar ve kapasite olarak yıllık üretilen sabit disk miktarları Tablo 2.1'de verilmiştir (Lyman ve diğ., 2003). Tablo 2.2'de de görüldüğü üzere kapasite açısından bakıldığı zaman sabit disk üretimi ve satışının ortalama olarak 18 ayda bir ikiye katlandığı görülmektedir.

Benzer şekilde, intel firmasının kurucularından (Moore, 1965), bir mikroçipin içinde bulunan transistor sayısının her 18 ayda bir ikiye katlandığını belirtmiştir. Yarı iletken endüstrisinden gelen veri, bu öngörünün gerçekliğini kanıtlamıştır. Bu

öngörü, Moore yasası olarak bilinmektedir. Berkeley üniversitesinin yapmış olduğu araştırma sonucu ile Moore yasası arasındaki ilişki dikkat çarpıcıdır. Dünya üzerinde toplanan tüm verinin miktarı yaklaşık olarak 18 ayda bir ikiye katlanmaktadır. Bu kadar büyük verinin toplanması ve depolanması için, veri tabanları ve depolama ortamları da en az aynı oranda büyümektedir.

Tablo 2.1**:** 2002 Yılında Yaratılan Bilgi Miktarı

Yedekleme Ortamı	**2002 Yılı Üst Sınır (terabayt cinsinden)**	**2002 Yılı Alt Sınır (terabayt cinsinden)**	**1999 Yılı Üst Sınır (terabayt cinsinden)**	**1999 yılı Alt Sınır (terabayt cinsinden)**	**Üst Sınırlar Arası Değişim Oranı**
Kağıt Ortamı	1,634	327	1,200	240	36 %
Film Endüstrisi	420,254	76,690	431,690	58,209	-3 %
Manyetik	4,999,230	3,416,230	2,779,760	2,073,760	80 %
Optik	103	51	81	29	28 %
TOPLAM	5,421,221	3,433,298	3,212,731	2,132,238	69 %

Verideki bu artışın oluşmasında internetin rolü yadsınamaz. Çünkü coğrafi olarak uzak yerleri bile çok yakın eden, dünya üzerindeki herkesin veri oluşmasına kolayca katkı sağlamasını sağlayan internet devrimi olmuştur. Mevcut internet teknolojisi ve bu teknolojinin muazzam büyümesi göz önüne alındığında, dünyanın her noktasından girilen verinin yorumlanıp bilgi haline getirilmesi için çok daha gelişmiş veri madenciliği tekniklerine ihtiyaç duyulacaktır. 1998 yılında A.B.D. başkanına sunulan 21 nci yüzyıl için enformasyon teknolojileri (Information Technology for 21st Century) konulu raporda, gelişen internet ve çoklu ortam (multimedia) uygulamalarının; bilginin görselleştirilmesini, yorumlanmasını, işlenmesini ve analiz edilmesini zorunlu kılmakta olduğu belirtilmiştir (Mitra ve Acharya, 2003). Tüm bu nedenlerden dolayı, veri madenciliği tekniklerinin iyileştirilmesi ve geliştirilmesi önümüzdeki yıllarda da önemli bir çalışma alanı olacaktır.

Yaratılan veri miktarı o kadar fazladır ki, insanoğlu, yaratılan verinin yaklaşık %20'sini gözle inceleyebilmektedir. Geri kalan %80'lik veriden gerekli bilgiler oluşturulamamaktadır.

Tablo 2.2**:** Yıllık Sabit Disk Üretim Tablosu

Yıl	**Satılan Sabit Disk Miktarı (X1000)**	**Depolama Kapasitesi (Petabyt)**
1992	42.000	bilinmiyor
1995	89.054	104,80
1996	105.686	183,90
1997	129.281	343,63
1998	143.649	724,36
1999	165.857	1.394,60
2000	200.000	4.630,50
2001	196.000	7.279,14
2002	213.000	10.849,56
2003	235.000	15.892,24
TOPLAM	1.519.527	41.402,73

İşletmeler ve organizasyonlar tarafından her gün çok büyük miktarlarda veri oluşturulmaktadır. En basit örnek olarak Türkiye'nin önde gelen GSM operatörlerinden Türkcell'in 2006 Mart ayı sonu itibari ile abone sayısı yaklaşık 28,7 milyon kişidir. Bu kadar insanın yapmış olduğu tüm işlemlere ilişkin her türlü ayrıntıların tutulduğu işlembilgi (transaction) düşünüldüğünde, oluşan verinin boyutu daha iyi anlaşılabilir. Bunun yanına, tarifeler, servisler ve kampanyalar eklendiğinde iş biraz daha içinden çıkılmaz olacaktır.

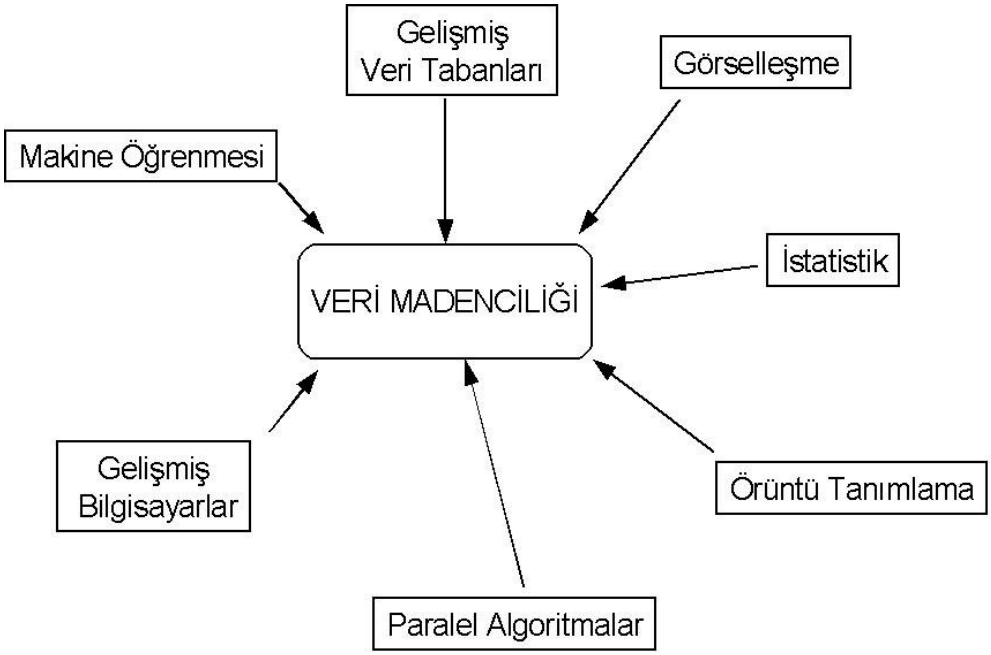
Veri, iç veya dış kaynaklardan toplanabilmektedir. Bu kaynaklar, genel olarak; mevcut kurulu sistemler, müşteri ilişkileri yönetimi (CRM), kurumsal kaynak planlama (ERP) uygulamaları, e-ticaret sistemleri, devlet organizasyonları ile ürün / servis sağlayıcılar ve ortaklıklarından oluşmaktadır (Nemati ve Barko, 2004).

Winter Corporation firmasının 2003 yılı araştırmasına göre, Fransa Telekomu 30 Terabayt ve AT&T 26 terabayt ile en büyük karar destek veri tabanına sahip

firmalardır. Internet açısından bakılırsa, Alexa internet arşivinde 7 yıllık veri tutulmaktadır ve toplam büyüklüğü 500 terabayttır. Google arama motorunun üzerinde 2003 yılı rakamları ile 4 milyarın üzerinde sayfa ve yüzlerce terabaytlık veri mevcuttur.

Şekil 2.1'de gösterildiği üzere, diğer bilim alanlarındaki gelişmeler, veri madenciliğini tetiklemiştir. Veri madenciliği, Şekil 2.1'de gösterilen diğer bilim dallarının hepsinin bir arada olmasının doğal bir sonucudur. Bunlardan sadece bir veya birkaç tanesinin olması, veri madenciliği açısından çok fazla bir şey ifade etmezken, hepsinin bir araya gelmesi tetiklemiştir.

Bu noktada, makine öğrenmesi ile istatistik arasındaki ana farkın belirtilmesinde fayda vardır. İstatistik, genel anlamda bir varsayımın sınanması ile ilgilenirken, makine öğrenmesi daha çok olası varsayımları tarayarak bir genelleme yapılabilmesinin denklem haline getirilmesi ile ilgilenmektedir.



Şekil 2.1: Veri Madenciliği Yapılmasını Tetikleyen Gelişmeler

Şekil 2.1'deki veri madenciliğini tetikleyen bilim dallarının dışında, veri madenciliğini en çok etkileyen gelişmeler, Şekil 2.2'de gösterilmiştir.

Şekil 2.2**:** Veri Madenciliğinin Dayanak Noktaları (Thearling, 2002)

2.5. Veri Madenciliği Hakkındaki Yanlış İnanışlar ve Gerçekler

Veri madenciliği sihirli bir değnek değildir. Ama, veri madenciliğinin sihirli bir değnek olduğu, veri içinde saklı olan önemli bilgileri insana gereksinim duymadan, kendi başına bulduğunu düşünen kişi sayısı azımsanmayacak miktardadır. Aşağıda, veri madenciliği hakkındaki yanlış inanışlar ve onların doğruları verilmiştir.

Yanlış İnanış 1: Veri üzerinde otomatik olarak çalışan ve problemlere otomatik olarak çözüm bulan veri madenciliği yazılımları vardır.

Gerçek 1: Kullanıcı tarafından hiç bir işlem yapılmadan, otomatik olarak problemlere yanıt bulan herhangi bir veri madenciliği aracı bulunmamaktadır. Veri madenciliği bir süreçtir.

Yanlış İnanış 2: Veri madenciliği insan gözetiminin hiç gerek duyulmadığı ya da çok az duyulduğu kendi başına çalışan bir süreçtir.
Gerçek 2: Veri madenciliği, her aşamasında insan gözetimine ihtiyaç duyan bir süreçtir. Veri madenciliği modeli oluşturulduktan sonra bile, yeni veri için sıklıkla modelin güncellenmesi gerekmektedir. Analistler tarafından, veri madenciliği

kalitesinin sürekli olarak takip edilmesi ve değerlendirme ölçümlerinin yapılması gerekmektedir.

Yanlış İnanış 3: Veri madenciliği çalışmaları, kendisine harcanan kaynakları çok kısa bir sürede geri kazandırır.

Gerçek 3: Veri madenciliğine harcanan kaynakların veri madenciliğinin işletmeye kazandırdıkları ile kendisini amorti etmesi, ilk başlangıç için harcanan kaynakların büyüklüğüne, analiz eden personelin ücretlerine ve veri ambarı oluşturma masraflarına göre değişir.

Yanlış İnanış 4: Veri madenciliği yazılımları genellikle kullanımı kolay olan ve sezgilerle bile kullanılabilen yazılımlardır.
Gerçek 4: Kullanım kolaylığı yazılımdan yazılıma değişmektedir. Bununla birlikte, veri analisti, ihtiyaç duyulan bilgiyi, analitik düşünce ve işletmenin genel amaçları ile araştırma modeline olan aşinalığıyla oluşturması gerekir.

Yanlış İnanış 5: Veri madenciliği, işletme ile ilgili problemleri, kendi başına ortaya çıkarmaktadır.

Gerçek 5: Veri madenciliği süreci, örüntülerin bulunması konusunda yardımcıdır. İşletme ile ilgili problem sahalarını belirlemek, yine insanların görevidir.

Yanlış İnanış 6: Veri madenciliği, dağınık bir veri tabanını düzenler ve temizler.

Gerçek 6: Bunu otomatik olarak yapamaz. Veri madenciliğinin ilk adımlarından biri olan verinin hazırlanması aşamasında, genellikle uzun süre el atılmamış ve incelenmemiş veri ele alınır. Bu nedenle, ilk defa veri madenciliği yapan bir işletme uzun zamandır elden geçirilmemiş, eski ve güncellenmeye ihtiyaç duyan veri problemleriyle yüz yüze gelmektedir.

2.6. Veri Madenciliğinin Uygulama Alanları

Veri madenciliğinin bir çok uygulama alanı vardır. Veri madenciliği uygulamalarına konu olmuş alanların en önemlileri şunlardır;

• Pazarlama: Pazar dağılımı, müşteri değerleme ve çapraz satış analizleri.

• Bankacılık: Risk yönetimi, usulsüzlük tespiti, müşteri kazanma ve mevcut müşterileri elde tutma analizleri, kredi işlemleri, firma derecelendirme, faiz oranlarının tahmini, borçlanma ve iflas tahminleri.

• Sigortacılık: Müşteri kaybı sebeplerinin belirlenmesi, usulsüzlüklerin önlenmesi, ana giderlerin azaltılması, poliçe fiyatlarının belirlenmesi.

• Perakendecilik: Satış noktası veri analizleri, alış veriş sepeti analizleri, tedarik ve mağaza yerleşim iyileştirmeleri.

• Borsa: Hisse senedi fiyat tahmini, genel piyasa analizleri, alım satım stratejilerinin iyileştirmeleri.

• Telekomünikasyon: Kalite iyileştirme, hile tespiti, hatların yoğunluk tahminleri, müşteri kazanma ve elde tutma analizleri.

• İlaç: Test sonuçlarının tahmini, ürün geliştirme.

• Sağlık: Tıbbi teşhis, uygun tedavi sürecinin belirlenmesi.

• Endüstri: Kalite kontrol, lojistik, üretim süreçlerinin iyileştirmesi.

• Bilim ve mühendislik: Ampirik veri üzerinde modeller kurularak bilimsel ve teknik problemlerin çözümlenmesi.

• Internet: Arama motorları.

• Devlet: Vergi hırsızlıklarının belirlenmesi, terörü önleme.

2.7. Veri Madenciliğinin Diğer Adları

Özellikle veri madenciliği kavramlarının yeni filizlendiği 1980'li yıllarda, konu ile ilgilenen bilim adamları tarafından çeşitli adlandırmalar yapılmıştır. Veri madenciliği adlandırması, özellikle iki binli yıllardan sonra standart olarak kullanılmaya başlanmıştır. Ama hala, çok az da olsa, bazı bilim çevrelerince farklı adlandırmalar kullanılmaktadır. Bunlardan bazıları ve kullanıma başlanma yılları Tablo 2.3'te verilmiştir.

Tablo 2.3**:** Veri Madenciliği Adlandırmaları

Adlandırma	**Kullanım Yılı Aralığı**	**"Google" Arama Sonucu**
Veri Tarama (Data Dredging)	1960 - ...	27.800
Veri tabanında bilgi keşfi (Knowledge Discovery in Databases)	1989 - ...	3.170.000
Veri Madenciliği (Data Mining)	1990 - ...	49.300.000

Her bir adlandırma, 2006 yılı mayıs ayında google arama motorunda yazılmış ve google arama motorunun bulduğu sonuçların toplam miktarları, adlandırmaların yanında verilmiştir. Bu rakamlardan da kolayca anlaşılacağı üzere, veri madenciliği adlandırması standart hale gelmiştir. Bilgi keşfi adlandırması ise daha çok yapay zeka ve makine öğrenmesi üzerinde çalışanlar arasında kullanılmaya devam etmektedir. Yukarıdaki adlandırmaların dışında, bazı kesimlerce kullanılan ama fazla rağbet görmeyen adlandırmalar ise şunlardır:

- Veri Arkeolojisi (Data Archaeology)
- Veri Avcılığı (Data Fishing)

- Bilgi Hasadı (Information Harvesting)
- Bilgi Keşfi (Information Discovery)
- Bilgi Çıkarımı (Knowledge Extraction)

2.8. Veri Madenciliği ve Etik

Özellikle bireylere ait verinin, veri madenciliği amacıyla kullanılması ciddi etik tartışmaları beraberinde getirmiştir. Veri madenciliği ile ilgilenen kişiler, bu etik kaygıları hesaba katarak, dikkatli olmalıdır.

Veri madenciliği, doğrudan insanlara ait veriye uygulandığında, kimlere borç verilmeli veya verilmemeli, kimlere özel indirimler uygulanmalı veya uygulanmamalı gibi doğası gereği ayırımlar yapar. Bu ayırımlar, etik açısından herhangi bir problem teşkil etmemektedir. Öbür taraftan, özellikle ırk ve din gibi bazı ayırımlar etik olmayabilir. Bu tarz ayırımlar, etik olmadığının yanında, kanunlara da uygun değillerdir.

Veri madenciliğindeki ayırım konusu oldukça karmaşık bir durumdur. Cinsiyet veya ırk gibi bir verinin, bir hastalığın iyileştirilmesi amacıyla kullanılması etik iken, aynı bilginin, kimlere borç verilmemeli sorusunun cevabını bulurken kullanılması etik değildir. Hatta bazen ırk veya din gibi hassas veri kullanılmasa da, veri madenciliği sonucunda elde edilen bilgi bu durumu işaret edebilmaktedir. Doğal olarak, bu durum da etik değildir. Yurdumuzdan örnek vermek gerekirse, Türkler, memleketleri dışındayken, genellikle hemşehrileri ile birlikte iskan etmek isterler. İstanbul örneği ele alınırsa, her mahalle veya semtte, genellikle aynı yörenin insanı yaşamaktadır. Ordulular, Sivaslılar, Malatyalılar gibi. Veri madenciliği yapılmak istenilen verinin içinden memleket / yöre bilgisi çıkartılsa bile, o mahallede veya semtte oturan insanların kredi başvurusunun geri çevrilmesi gerekir şeklinde bir sonuçla karşılaşıldığında, o yörenin tüm insanları için bir ayırım yapılmış olunur ki, bu durum etik olmaz.

Peki, etik sorunu nasıl aşılır? Bu konudaki en yaygın çözüm, veri madenciliği yapmak isteyen kişinin, kişilere ait veriyi, hangi amaçla kullanacağını, bilginin gizliliğini korumak için hangi önlemleri alacağını ve hatalarını nasıl düzeltebileceğini çok iyi bilmesi gerektiğidir (Witten ve Frank, 2005).

Avrupa birliği veya A.B.D. gibi bazı devlet ve oluşumlar, insanlara ait verinin ayrımcılığa neden olabilecek şekilde kullanılmamasını kanunlar ile korumuşlardır.

2.9. Veri Madenciliği Yöntembilimleri

Veri madenciliği yöntembilimleri, veri madenciliği yapılırken baştan sona neler yapılacağını sistematik bir şekilde sunan yapılardır. Veri madenciliği yöntembilimlerinin amacı, veri madenciliği çalışmalarını belirli bir disiplin altına almaktır.
Veri madenciliğinin nasıl yapılması gerektiği konusunda organizasyonların farklı yaklaşımları olabilmektedir. Uygulamalardan, veri madenciliği yazılımlarından ve endüstrilerden bağımsız bir yöntembilime ihtiyaç duyulduğu açıktır (Larose, 2006).

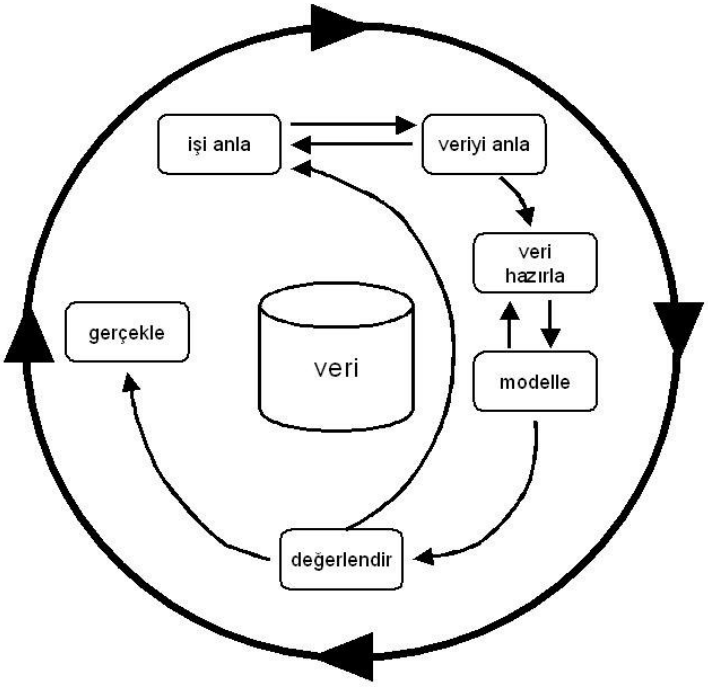
Uygulamalardan, yazılımlardan ve endüstriden bağımsız iki ana yöntembilim bulunmaktadır. Bunlardan en çok kullanılanı, NCR Systems Engineering Copenhagen (A.B.D. ve Danimarka), Daimler Chrysler AG (Almanya), SPSS Inc.(A.B.D.) ve OHRA Verzekeringen en Bank Groep B.V.(Hollanda) firmalarının bir araya gelerek oluşturduğu CRISP-DM (CRoss Industry Standart Process for Data Mining) yöntembilimidir. Diğeri ise IBM firmasının hazırlamış olduğu SEMMA (Sample, Explore, Modify, Model, Assess)'dır. Bu iki yöntembilim haricinde, veri madenciliği yapan insanlar, kendilerine ait bir yöntembilim geliştirebilmektedir.

Kdnuggets (2004)'ten alınan anket sonuçlarına göre, veri madenciliği ile ilgilenen kişi ve kurumların kullandıkları veri madenciliği yöntembilimleri Tablo 2.4'te verilmiştir. Tablo 2.4'teki anket çalışması, veri madenciliği üzerine çalışan uzmanlar arasında yapıldığından, genel anlamda gerçek sonuçlar olduğu şeklinde değerlendirilebilir.

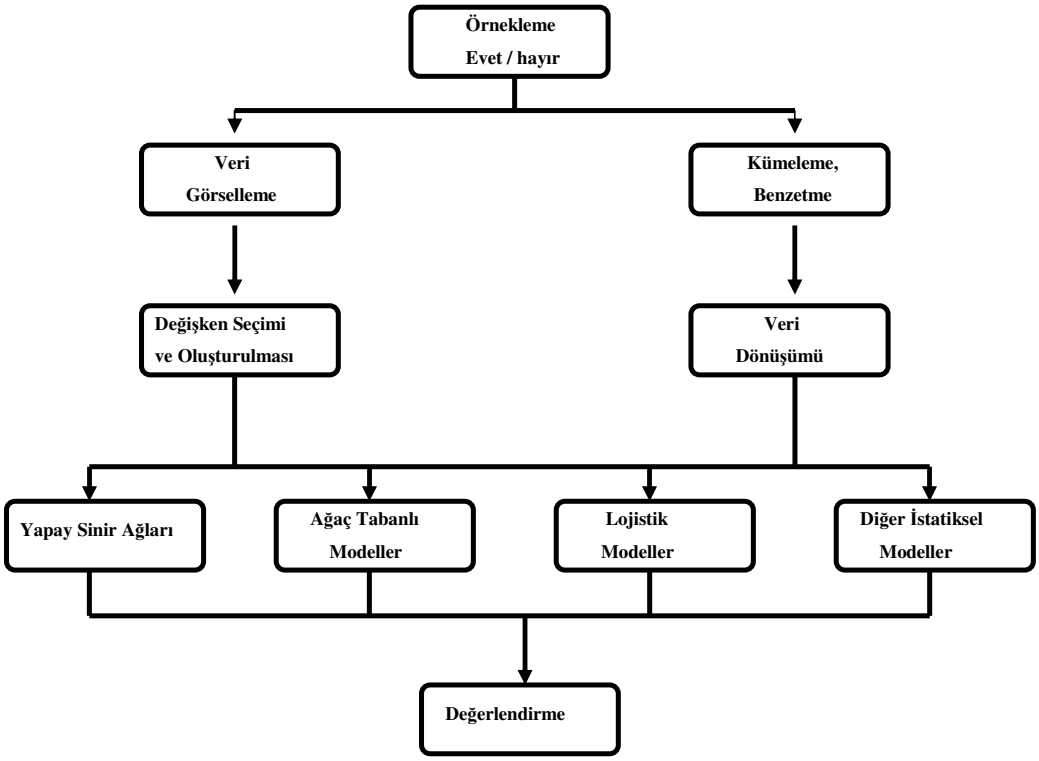
CRISP-DM yöntembiliminin ana yapısı Şekil 2.3'te, SEMMA yöntembiliminin ana yapısı Şekil 2.4'te verilmiştir. Bu tez çalışmasındaki tüm veri madenciliği aşamaları, CRISP-DM yöntembilimi rehberliğinde hazırlanmış olup, bundan sonraki bölümlerde bu yöntembilimin adımları görülmektedir.

Tablo 2.4**:** Yöntembilimlerin Kullanılma Oranları

Kullanılan Yöntembilim	**Oranı**
CRISP-DM yöntembilimi	42 %
SEMMA yöntembilimi	10 %
İşyerine özel yöntembilimler	6 %
Bireylerin kendi yöntembilimleri	28 %
Diğer	6 %
Hiçbiri	7 %



Şekil 2.3**:** CRISP-DM Yöntembilimi



Şekil 2.4 : SEMMA Yöntembilimi (Firestone, 1997)

2.10. Modelleme Teknikleri

2.10.1. Sınıflandırma (Classification)

İnsanoğlu doğası gereği sınıflandırır. Dünyayı daha iyi anlamak ve tanımlamak için nesneleri sınıflandırır veya derecelendir. Örneğin, insanları; Türk, İngiliz, zenci gibi, köpekleri; terrier, buldog, kaniş gibi, hava durumunu; bulutlu, yağmurlu, karlı, güneşli, soğuk gibi, kredi talebinde bulunan müşterileri; yüksek, orta veya düşük riskli gibi sınıflandırır.

Sınıflandırmanın özünde, yeni durumun özelliklerini inceleyerek, mevcut durumlardan hangisine ait olduğunu belirlemek yatmaktadır. Sınıflandırılacak nesneler veri tabanında veya dosyada bir kayıt iken, sınıflandırma işleminin kendisi, veri tabanına sınıf kodu içeren yeni bir sütun eklemektir. Sınıflandırma, sınıfsal değerler ile ilgilenir.

2.10.2. Kestirim (Estimation)

Aslında, kestirimin kendisi de bir sınıflandırmadır. Tek fark, sınıflandırma sınıfsal değerler ile ilgilenirken, kestirim rakamsal değerler ile ilgilenir. Verilen girdi değerleri için, maaş, boy veya kredi kart bakiyesi gibi bilinmeyen değerleri getirmek, kestirimin işidir.

Örneğin, bir bankanın, mevcut iki milyon kredi kartı kullanıcısından sadece yüz bin kişiye yeni bir teklif sunacağı düşünüldüğünde, banka, bu yüz bin kişiyi diğerlerinin arasından nasıl seçecektir? Çözüm olarak, müşterinin mevcut borç ödeme alışkanlığı, harcama miktarları, yeniliklere açıklığı gibi, bankanın sunmak istediği teklif ile ilgili olabilecek bir çok niteliğe bakılarak rakamsal bir puanlama yapılabilir. Bu puanlamadan, yani kestirimden alınan sonuca göre, en yüksek puanlı ilk yüz bin müşteriye teklifte bulunulabilir.

2.10.3. Tahmin (Prediction)

Tahmin, sınıflandırma veya kestirim ile çok benzerdir. Diğerlerinden farkı, tahminin geleceğe yönelik olmasıdır. Tahmin işleminde, sınıflandırmanın doğruluğunu kontrol etmenin tek yolu bekleyip görmektir.

Sınıflandırma ve kestirim için kullanılan tüm teknikler, tahmin için de kullanılabilir. Bu durumda, mevcut durum; daha önce olmuş olayları, yeni durum ise gelecekte olacak olan olayları tanımlar. Daha önce olmuş olaylardan bir model oluşturularak, ilerisi için bir tahmin yaratılır.

2.10.4. Benzer gruplama (Affinity grouping)

Amaç, hangi nesnelerin beraber olduğunu ya da birbirlerinin varlıklarını etkilediğini bulmaktır. Benzer gruplamaya en iyi örnek, alışveriş sepeti örneğidir. Alış veriş sepeti analizi, insanların hangi iki veya daha fazla malzemeyi aynı anda almak istediklerini araştırmaktadır. Mağazalar, bu bilgiyi kullanarak, hangi malzemenin

hangi sıra veya rafa konulması gerektiği ya da tanıtım kataloglarının nasıl hazırlanması gerektiği konusunda fikir sahibi olabilmektedirler.

2.10.5. Kümeleme (Clustering)

Kümeleme, heterojen olan bir büyük gurubun, mümkün olduğu kadar homojen olan alt gruplara veya kümelere bölünmesi işlemidir. Kümelemenin sınıflandırmadan farkı, daha önceden tanımlı sınıfları kullanmayıp, kendi sınıfını oluşturmasıdır.

Kümelemede, önceden tanımlı sınıf ya da örnek yoktur. Kayıtlar benzerliklerine göre sınıflandırılır. Kümeleme, başka bir modelleme tekniği kullanmadan önce yapılması gereken bir ilk iş olarak sık kullanılan bir yöntemdir. Örneğin, bir mağaza için müşterilerin hangi tür promosyonlara olumlu tepki verdiğini ölçmeden önce, müşteriler öncelikle, satın alma alışkanlıklarına göre kümelenir ve daha sonra hangi küme gurubunun ne tür promosyonlara olumlu tepki verdiği tespit edilir.

2.10.6. Tanımlama ve belgileme (Description and profiling)

Veri madenciliği, verinin üretmiş olduğu bireylere ait kayıtlar, ürünler veya işlemler gibi veri tabanında neler olduğunu anlamamıza yardımcı olmak amacıyla da kullanılabilir. Yeteri kadar iyi bir tanımlama, beraberinde iyi bir açıklamayı da getirecektir. En kötü ihtimalde bile, iyi bir tanımlama, makul bir açıklama için nereden başlanması gerektiği konusunda veri madeni analistine bilgi verecektir.

2.11. Veri Madenciliği Model Tipleri

Veri madenciliği işlevleri genel olarak iki farklı ulamda (kategoride) sınıflandırılmaktadır: tanımlayıcı veri madenciliği ve tahmin edici veri madenciliği. Tanımlayıcı veri madenciliği, veri kümesini kısa ve özet bir biçimde tanımlamaktadır ve verinin ilginç özelliklerini göstererek, mevcut verinin tanınmasında yardımcı olmaktadır. Tahmin edici veri madenciliği ise, mevcut veri kümesi üzerinde bir

model ya da modeller kümesi oluşturarak, yeni veri kümelerinin davranışlarının nasıl olacağını tahmin etmeye çalışmaktadır.

2.11.1. Tahmin edici (Predictive) Modeller

Tahmin edici modeller, bir özniteliğin bilinmeyen veya gelecekteki olası değerini bulmak için kullanılır. Bu tür modellemelerde, girdi verisinin yanında en az bir tane çıktı verisi olur. Mevcut durum için bir model oluşturulur ve oluşturulan bu model, bilinmeyen veya gelecekteki değer tahmini için kullanılır. Tahmin edici model tipleri şunlardır;

- Sınıflandırma (Classification)
- Kestirim (Estimation)
- Tahmin (Prediction)

2.11.2. Tanımlayıcı (Descriptive) Modeller

Tanımlayıcı modeller, veriyi tanımlayan ve insanların yorumlayabileceği örüntüleri bulur. Girdi verisi bulunmak zorunda olmasına karşın çıktı verisi bulunmak zorunda değildir. Mevcut durumu tanımlayan bir model oluşturur. Tanımlayıcı model tipleri şunlardır;

- Sınıflandırma (Classification)
- Benzer Gruplama (Affinity Grouping)
- Kümeleme (Clustering)
- Tanımlama (Description)

2.12. Önemli Veri Madenciliği Teknikleri

2.12.1. Naive Bayes

Naive Bayes, tek taramalı bir algoritmadır, bu yüzden hızlıdır. Hızlılığının yanında çok basit yapıya sahip olması bu modellemenin en büyük avantajıdır. Diğer yandan, tüm öznitelikler eşit derecede öneme sahiptir ve bu yüzden istatistiksel olarak bağımsızdır. Bu sebeple, bir özniteliğin değerini biliyor olmak, başka bir özniteliğin değeri hakkında hiçbir bilgi vermemektedir. Bu da çok önemli bir dezavantajdır.

Naïve Bayes veri madenciliği yöntemi, bu tez çalışmasında karşılaştırılan yöntemlerden biri olduğu için 3 ncü bölümde detaylı olarak incelenmiştir.

2.12.2. Karar ağaçları (Decision trees)

Veri madenciliği konusunda uzman bir çok araştırmacı (German ve diğ., 1999; Pal ve Mather, 2001), karar ağaçları ile maksimum olabilirlik sınıflandırması (Maximum Likelihood Classification, MLC) veya yapay sinir ağları gibi diğer sınıflandırıcıların karşılaştırmasını yapmışlar ve karar ağaçlarının en iyi sınıflandırıcı olduğu konusunda hemfikir olmuşlardır. Gahegan ve West (1998), karar ağaçlarının, yapay sinir ağlarından farklı olarak, kolaylıkla eğitilebildiğini, çok hızlı bir şekilde doğru sonuçlar verdiğini ve her adımının kolaylıkla izlenilip, anlaşılabildiğini belirtmişlerdir. Karar ağaçlarının geleneksel istatistiksel sınıflayıcılardan farkı, kayıp ve gürültülü veriyle uyumlu çalışabilmesi ve değiştirgesiz bir sınıflayıcı olmasıdır. Değiştirgesiz sınıflayıcılar, istatistiği temel almazlar, bu nedenle, verinin özelliklerinden bağımsızdırlar ve eğitici kümenin dağılımını hesaba katmazlar. En çok bilinen değiştirgesiz sınıflayıcılar; karar ağaçları ve yapay sinir ağlarıdır.

Karar ağaçları metodu, böl ve yönet stratejisiyle çalışmaktadır (Bharti, 2004). Veri madenciliği teknikleri arasında en çok kullanılanıdır.

Karar ağaçları veri madenciliği yöntemi, bu tez çalışmasında karşılaştırılan yöntemlerden biri olduğu için 3 ncü bölümde detaylı olarak incelenmiştir.

2.12.3. Yapay sinir ağları (Artificial Neural networks)

Yapay sinir ağları yöntemi, çok güçlü bir tahmin modelleme tekniğidir. Bu gücün bir kısmını, hemen her alanda uygulanabilir olmasından almaktadır. Veri madenciliği ve karar destek sistemlerindeki bir çok başarılı uygulamalarla geçerliliğini ispatlamış bir yöntemdir.

10^{11} adet nöron içeren insan beyni, karmaşık hesaplamaların bile üstesinden gelebilecek yetenektedir. İnsan beyni, bilgi üretebilmek için neden sonuç ilişkilerinden, hislerine kadar bir çok metodu kullanır (Kasabov, 1998). Yapay sinir ağları, insan beyninin işleyişi taklit edilerek oluşturulmuştur.

Yapay sinir ağları, uzmanları tarafından bile her zaman tam olarak anlaşılamayan çok karmaşık modeller üretmektedir. Bu durum, yapay sinir ağlarının en büyük dezavantajıdır. Modelin kendisi rakamsal değerlerin karmaşık hesaplamaları ile uğraştığı için, girdilerinin de rakamsal olmasını bekler. Girdinin veya çıktının rakamsal olmaması durumunda, verinin uygun şekilde dönüştürülmesi gerekmektedir. Bir çok veri madenciliği yazılımı, bu dönüşümü otomatik olarak yapabilecek yeteneğe sahiptir.Yapay sinir ağlarında eğitilme işlemi, çok zaman alır, dönüştürme yapıldığı sürece tüm veri tipleri ile çalışır, ama yapay sinir ağlarından maksimum fayda sağlamak için veri hazırlamanın çok iyi yapılması gerekmektedir. Aksi taktirde, elde edilen sonuçlar, işletmeleri yanlış yönlendirebilir. Yapay sinir ağı yöntemi, doğru veriden ne kadar mükemmel doğru sonuçlar üretiyorsa, hatalı veya normal dışı değerlerden de bir o kadar hatalı sonuçlar üretme durumundadır.

Yapay sinir ağları veri madenciliği yöntemi, bu tez çalışmasında karşılaştırılan yöntemlerden biri olduğu için 3 ncü bölümde detaylı olarak incelenmiştir.

2.12.4. Bellek tabanlı akıl yürütme (Memory based reasoning)

İnsanoğlu, yeni bir durumla karşılaştığı zaman, geçmişte deneyim kazandığı benzer durumlar rehberinde hareket tarzı geliştirir (Berry ve Linoff, 2005).

Bellek tabanlı akıl yürütmenin temeli de budur. Benzerlik, algoritmanın tam merkezinde yer alır. Yeni bir veri geldiği zaman, veri tabanında bulunan benzer oluşumlar veya komşular aranır. Bu komşular sınıflandırma ve kestirim için kullanılır.

Örneğin, konuşması duyulan bir kişinin şivesinden, bu kişinin nereli olduğu tahmin edilebilir. Ya da, ortak film zevkleri olan iki kişiden biri diğerine, beğendiği bir filme gitmesini tavsiye ediyorsa, diğer kişi, o filmi beğeneceğini düşünür. Çünkü, arkadaşının daha önceden önerdiği filmleri beğenmiştir.
Bellek tabanlı akıl yürütmenin benzerlik ölçüsü için mesafe bilgisi kullanılır. Dolayısı ile mesafenin minimum olduğu nokta benzerlik için gerekli ve yeterlidir. Bu mesafelerin gerçek veri üzerinde nasıl hesaplandığını göstermek üzere; Şekil 2.5'te pazarlama veri tabanından beş müşteriye ait kayıtlar incelenmiştir.

Şekil 2.5'teki tablonun öznitelikleri cinsiyet, yaş ve maaştır. Bu özniteliklerin her biri için ayrı mesafe matrislerinin hesaplanması gerekmektedir. Şekil 2.6'da görüldüğü üzere, cinsiyetler arası geçişler daha büyük mesafe değerine sahiptir. Bayan – bayan ve erkek – erkek arası mesafe 0 iken, bayan – erkek veya erkek – bayan geçişleri 1 değerini alır.

Yaş için mesafe matrisi hazırlanırken, matrisin satır ve sütununa yaş özniteliğinin tüm değerleri girilir. Satır ve sütunlarda yaş değerlerinin kesiştiği noktalarda, iki yaş arasındaki fark bulunur. Sonrasında bu fark değeri, olabilecek en büyük fark değerine bölünür. Şekil 2.5'teki örnek veri için en büyük değer 52 ve en küçük değer 27'dir. Aralarındaki fark ise 25'tir. Her bir değer 25 rakamına bölünmek suretiyle [0,1] değer aralığında normalleştirilmiş bir mesafe elde edilir. Aynı işlem maaş için de

hesaplanır. Maaş matrisi için fark değerleri, en büyük fark olan 86K (105K-19K) değerine bölünür.

SıraNo	Cinsiyet	Yaş	Maaş
1	Bayan	27	19.000 $
2	Erkek	51	64.000 $
3	Erkek	52	105.000 $
4	Bayan	33	55.000 $
5	Erkek	45	45.000 $

Şekil 2.5**:** Pazarlama Veri Tabanından 5 Müşteriye ait Kayıt

Mesafe matrisleri hazırlandıktan sonra, veri tabanına yeni bir kayıt geldiği zaman, bu yeni gelen kayıtın daha önceki kayıtlardan hangisine benzediğini bulmak için, yeni kayıt ile mevcut kayıtlar arasındaki tüm mesafelerin bulunarak, bu mesafelerden en küçüğünün seçilmesi gerekir. Örneğin, Şekil 2.7'deki gibi bir yeni kayıt geldiğinde, bu yeni kayıt en çok hangi kayıta benzemekte sorusunun cevabı bulunmak istensin.

$d_{cinsiyet}(bayan, bayan) = 0$
$d_{cinsiyet}(bayan, erkek) = 1$
$d_{cinsiyet}(erkek, bayan) = 1$
$d_{cinsiyet}(erkek, erkek) = 0$

	27	51	52	33	45
27	0.00	0.96	1.00	0.24	0.72
51	0.96	0.00	0.04	0.72	0.24
52	1.00	0.04	0.00	0.76	0.28
33	0.24	0.72	0.76	0.00	0.48
45	0.72	0.24	0.28	0.48	0.00

	19K	64K	105K	55K	45K
19K	0.00	0.52	1.00	0.41	0.30
64K	0.52	0.00	0.47	0.10	0.22
105K	1.00	0.47	0.00	0.58	0.69
55K	0.41	0.10	0.58	0.00	0.11
45K	0.30	0.22	0.69	0.11	0.00

Şekil 2.6**:** Cinsiyet, Yaş ve Maaşa Göre Mesafe Matrisleri

SıraNo	Cinsiyet	Yaş	Maaş
1	Bayan	27	19.000 $
2	Erkek	51	64.000 $
3	Erkek	52	105.000 $
4	Bayan	33	55.000 $
5	Erkek	45	45.000 $
yeni	Bayan	45	100.000 $

Şekil 2.7**:** Yeni bir Kayıt Eklendiğinde Kendisine En Yakın Grup Hangisidir?

Yeni kayıt ile 4 sıra numaralı kayıt arasındaki mesafe aşağıdaki yolla hesaplanabilir.

$$d_{toplam}(A,B) = d_{cinsiyet}(A,B) + d_{yaş}(A,B) + d_{maas}(A,B) \quad (2.1)$$

Bu mesafe denklemine Manhattan mesafesi denmektedir. Cinsiyet, yaş ve maaşa ait mesafeler;

$$d_{cinsiyet}(bayan, bayan) = 0 \quad (2.2)$$

$$d_{yas}(45,33) = \frac{45-33}{52-27} = \frac{12}{25} = 0.480 \quad (2.3)$$

$$d_{maas}(100.000,55.000) = \frac{100.000-55.000}{105.000-19.000} = \frac{45.000}{86.000} = 0.523 \quad (2.4)$$

bulunur. Buradan, toplam mesafe;

$$d_{toplam} = 0 + 0.480 + 0.523 = 1.003 \quad (2.5)$$

elde edilir. Normalleştirilmiş toplamı bulmak için 1.003 değerinin 3'e bölünmesi gerekir, çünkü cinsiyet, yaş ve maaşın maksimum mesafeleri 1'dir. Üçünün toplamından 3 elde edilir.

$$d_{normalize} = \frac{1.003}{3} = 0.334 \text{ değeri elde edilir.} \quad (2.6)$$

Bu değer yeni kayıt ile dört sıra numaralı kayıt arasındaki mesafeyi göstermekte olup, yeni kayıt ile mevcut tüm kayıtlar arası mesafelerin toplamı ve normalleştirilmiş değerleri Şekil 2.8'de görülebilir. Şekil 2.8'den de anlaşılacağı üzere, yeni kayıt ile en küçük mesafe değerine sahip mevcut kayıt, 4 sıra numaralı kayıttır. Buradan sonuçla, yeni kayıt, mevcut kayıtlardan 4 numaralı kayıta benziyor denilebilir.

	1	2	3	4	5
d_{toplam}	1.662	1.659	1.338	1.003	1.640
d_{norm}	0.554	0.553	0.446	0.334	0.547

Şekil 2.8**:** Yeni Kayıt, 4 Sıra Numaralı Kayıta Yakındır.

2.12.5. K-Ortalama (K-means)

K-ortalama metodu, kümeleme amaçlı olarak kullanılan bir yöntemdir. Kümelemenin, sınıflandırmadan en büyük farkı, kontrolsüz öğrenme olması, yani daha önceden sınıflandırılmış bir duruma göre tahminde bulunmamasıdır. K-ortalama algoritması, sınıflandırılmamış veri üzerinde oluşan doğal gruplaşmaları bulur.

Gerçek veri arasındaki mesafeyi bulmak için öklid mesafesi kullanılır. Öklid mesafesinin denklemi (2.7)'de verilmiştir.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.7}$$

Sınıfsal veri arasındaki mesafe bulunurken; değerler birbirinden farklı ise 1, aynı ise 0 değeri alınır.

K-ortalama metodu bir örnekle açıklanırsa; eldeki veriye ait iki boyutlu bir dağılım şeması olduğu farz edilsin. Bu dağılımın, k-ortalama metoduyla değerlendirilmesi için gerekli adımlar şunlardır;

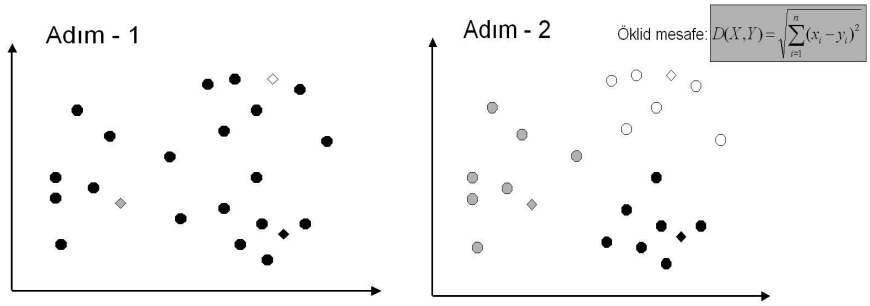
- Adım 1: Rasgele k sayıda küme merkezi seçilir. Örnekte k=3 adet seçilmiştir (Şekil 2.9).

- Adım 2: Her bir eleman, kendisine en yakın küme merkezine atanır. (koyu renkli daireler şeklinde gösterilen her bir veri için adım-1'de seçilen küme merkezlerine olan mesafe öklid mesafe kullanılarak hesaplanır) (Şekil 2.9).

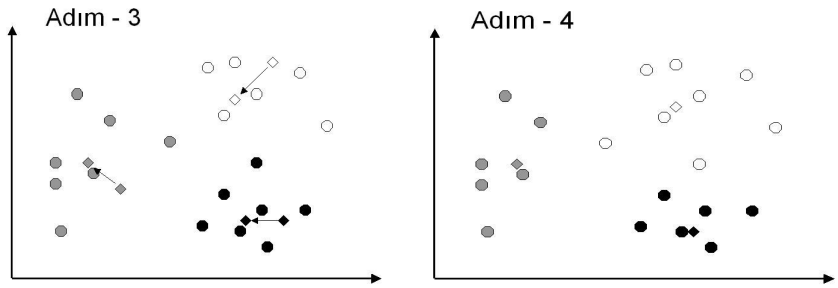
- Adım 3: Her bir küme merkezi, kendisine atanan elemanların merkezine kaydırılır. Bu durumda, küme merkezlerinin yerleri değişecektir (Şekil 2.10).

- Adım 4: Adım 2 ve 3, küme merkezi yer değişimleri belirli bir eşik altına düşünceye kadar yinelenir (Şekil 2.10).

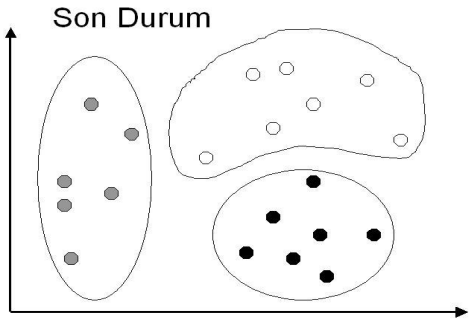
Hesaplanan küme merkezi yer değiştirmeleri belli bir eşik değerinin altına düştükten sonra, yani yer değiştirme oranı ihmal edilebilir derecede küçük olunca, kümelemenin son hali elde edilmiş olur. Örneğe ait son durum, Şekil 2.11'de verilmiştir.



Şekil 2.9**:** K-ortalama Metodu Adım-1 ve Adım-2



Şekil 2.10**:** K-ortalama Metodu Adım-3 ve Adım-4



Şekil 2.11**:** K-ortalama Metodu Son Durum

2.12.6. Apriori algoritması (Apriori algorithm)

Apriori algoritmasında amaç, ilişkisel veri tabanlarında veya veri ambarlarında bulunan malzeme ya da nesnelerin birbirleri arasındaki ilişkileri ya da korelasyonu (birlikte olma veya birinin olmasının diğerinin oluşumunu etkilemesi gibi) bulmaktır (Han ve Kamber, 2001).

Diğer bir deyişle, veri tabanı kayıtları arasında nedensellik ilişkisi bulmayı amaçlayan yaklaşımdır. Örneğin, X ürününü satın alan müşterilerin %90'ı aynı zamanda Y ürününü de satın almaktadır kuralı, pazarlama kitlesinin veya mağazalarda ürün yerleşim planlarının belirlenmesinde son derece yararlı olması nedeniyle özellikle pazarlama yöneticileri ve mağazacılık sektörü ile ilgilenenler arasında son derece yaygın olarak kullanılmaktadır.

Apriori algoritması ve diğer birliktelik kuralı algoritmaları için destek ve güven terimlerinin bilinmesi gerekmektedir. Destek ve güven terimleri bir örnekle açıklanırsa;

$$A \,\&\, B \Rightarrow C \tag{2.8}$$

için, %50 minimum destek ve %50 güvene sahip tüm kurallar bulunmak istensin.

Destek : Bir işlemin {A,B,C} içerme olasılığıdır.
Güven : {A,B} içeren işlemin aynı zamanda {C} içerme şartlı olasılığıdır.

İşlem No	Alınan Malzeme
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Şekil 2.12**:** Satın Alınan Malzemeler Listesi

Şekil 2.12'de satın alınan malzemelere ait liste verilmiştir. Şekildeki veri incelendiğinde, %50 minimum destek ve %50 minimum güvene sahip kurallar şunlar olacaktır.

$$A \Rightarrow C \ (50\%,66.6\%) \tag{2.9}$$

$$C \Rightarrow A \ (50\%,100\%) \tag{2.10}$$

Parantez içindeki ilk rakam destek, ikinci rakam güvendir. Şekil 2.12'deki 4 kayıtlık malzemeler listesinde hem A, hem de C malzemesini içeren 2 adet kayıt vardır.
Yani, hem $A \Rightarrow C$ ve hem de $C \Rightarrow A$ için;

$$\text{destek oranı} = \frac{2}{4} = 50\%\text{'dir.} \tag{2.11}$$

A malzemesini alanların aynı zamanda C malzemesini satın almaları, $A \Rightarrow C$ kuralı için güven oranını vermektedir. Listede A alan kişi sayısı 3'tür. A malzemesini alanlardan aynı zamanda C malzemesini de alanların sayısı 2'dir. Bu durumda;

$$A \Rightarrow C \text{ için güven} = \frac{2}{3} = 66.6\ \%\text{'dır.} \tag{2.12}$$

C malzemesini alanların, aynı zamanda A malzemesini satın almaları, $C \Rightarrow A$ kuralı için güven oranını verir. Listede C alan kişi sayısı 2'dir. C malzemesini alanlardan aynı zamanda A malzemesini de alanların sayısı 2'dir. Bu durumda;

$C \Rightarrow A$ için güven$= \frac{2}{2} = 100$ %'dür. (2.13)

Apriori algoritmasının mantığına gelince;

- Sık karşılaşılan malzeme kümesinin tüm alt kümeleri de sık karşılaşılan olur. Örneğin {Süt, çocuk bezi, bira} sık karşılaşılan ise, bu durumda {Süt, çocuk bezi} 'de sık karşılaşılan olmalıdır. Diğer bir deyişle, nadir karşılaşılan bir malzemenin üst kümeleri sık karşılaşılan olamaz.Malzemenin sık karşılaşılan olması için maksimum destek oranına sahip olması gerekir. Destek oranı düşük olan elemanlar nadir karşılaşılan olduğu için listeden silinir.

- Tek malzemeli küme, iki malzemeli küme oluşturmak için kullanılır, iki malzemeli küme, üç malzemeli küme oluşturmak için kullanılır...Yani (k-1) malzemeli kümeler, (k) malzemeli kümeler oluşturmak için kullanılır.

Şekil 2.13'de apriori algoritması örneği verilmiştir. Örnekte de görüleceği üzere, ilk önce malzeme alım kayıtlarını tutan D veri tabanı taranmıştır. Bu tarama sonucunda;

- C_1 adımıyla tek malzemeli tüm malzemeler çıkarılmıştır.

- Daha sonra L_1 adımıyla en küçük destek oranına sahip olanlar, ki burada 4 sıra no'lu malzeme, listeden çıkarılmıştır.

- C_2 adımıyla, tek malzemeli kümeden olası tüm iki malzemeli kümeler oluşturulmuş ve destek değerleri hesaplanmıştır.

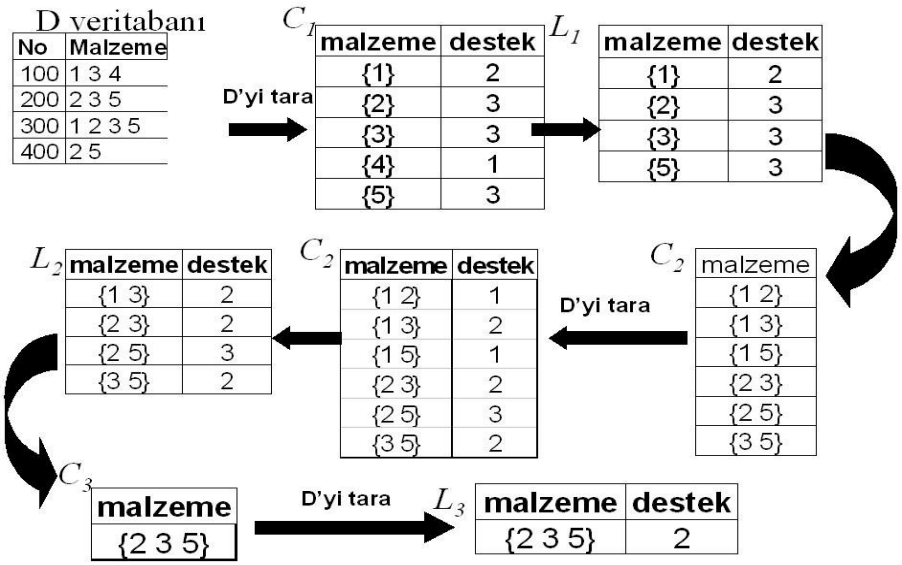
- L_2 adımıyla en küçük destek oranına sahip olanlar, ki burada {1 2} ve {1 5} kümeleri, listeden çıkarılmıştır.

- C_3 adımıyla, iki malzemeli kümelerden, üç malzemeli küme oluşturulmuş ve sonuçta tek küme {2 3 5} kalmıştır.

- L_3 adımıyla {2 3 5} kümesinin desteği hesaplanmıştır.

Apriori algoritmasına göre en sık karşılaşılan üçlü küme {2 3 5} kümesidir. D veri tabanında toplam 4 kayıt olduğuna ve {2 3 5} kümesi 2 kayıtta bulunduğuna göre;

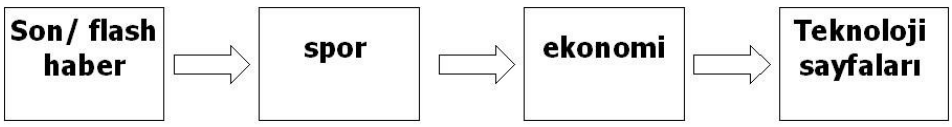
$$\{2\ 3\ 5\} \text{ kümesinin destek oranı} = \frac{2}{4} = 50\% \text{'dir} \qquad (2.14)$$



Şekil 2.13**:** Apriori Algoritması Örneği (Han ve Kamber, 2001)

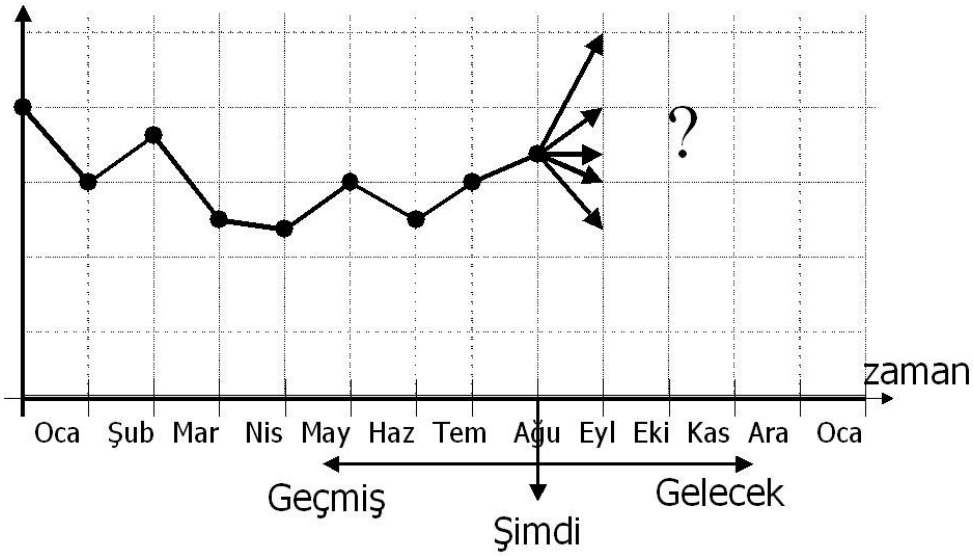
2.12.7. Zaman serileri (Time series)

Zaman serileri, zaman içinde değişen veri ile uğraşır. Örneğin, www.ntvmsnbc.com internet sitesinde gezinen bir ziyaretçinin, hangi ana konulara hangi sırayla baktığı tespit edilerek, gerekirse ileriye dönük olarak, sayfa ve konu tasarımında nelerin çıkartılıp, nelerin eklenmesi gerektiği veya konuların sayfa içerisinde nereye yerleştirilmesi gerektiği konusunda fikir edinilebilir. (Şekil 2.14)



Şekil 2.14**:** www.ntvmsnbc.com Sitesindeki Gezinme Sırası

Borsa zaman serilerine iyi bir örnektir. Borsadaki kağıtların veya borsa endeksinin gelecekte ne olacağının tahmin edilmesi için kullanılabilir. Tahmin amaçlı kullanılabildiği gibi, tanımlayıcı anlamda da kullanılabilir. Bu tarz yöntemler daha çok link analizi olarak tanımlanır (Şekil 2.15).



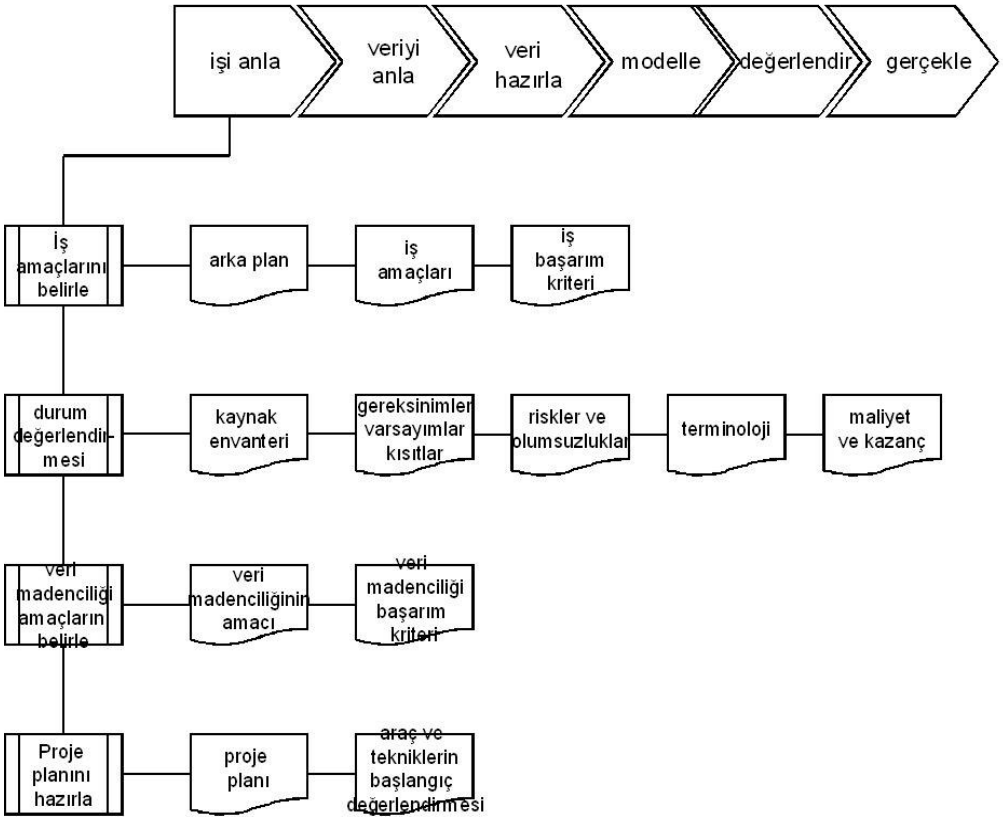
Şekil 2.15**:** Borsa Endeksinin Dünü, Bugünü ve Geleceği

3. CRISP-DM YÖNTEMBİLİMİ

CRISP-DM yöntembilimi, veri madenciliği ile ilgilenen kişi veya organizasyonlar tarafından en çok kullanılan yöntembilimdir. Bir veri madenciliği sürecinde atılması gereken adımların tümünün detayını açıklamaktadır.

Genel olarak 6 ana adımdan oluşmaktadır; işin anlaşılması, verinin anlaşılması, verinin temizlenmesi, modelleme, değerlendirme ve gerçekleme. Tez çalışmasının bu bölümünde, CRISP-DM yöntembilimi anlatılmıştır.

3.1. İşi Anlama



Şekil 3.1**:** CRISP-DM Basamak 1: İşi Anlama

CRISP-DM yöntembiliminin bu ilk adımı, projenin amaç ve gereksinimlerini, iş bakış açısı ile anlamaya ve sonrasında bu bilgiyi, veri madenciliği problemini tanımıya ve amaçları gerçekleştirmeye yönelik bir ilk plan haline dönüştürmeye yoğunlaşmıştır (Chapman ve diğ., 2000).

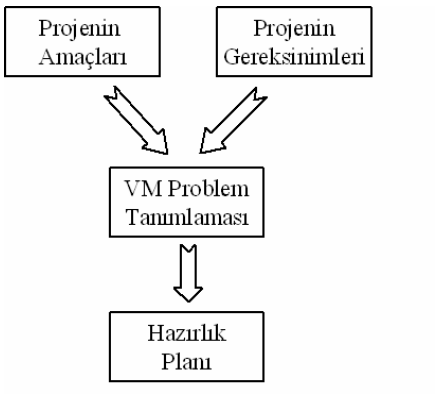
Veri madenciliği yapacak analistlerin ilk amacı, müşterinin gerçekten neyi başarmak istediğini tam anlamıyla öğrenmektir. Müşterilerin, genellikle, uygun bir şekilde dengede tutulması gereken birden fazla amacı vardır. Analistin asıl amacı ise, sonucu etkileyebilecek önemli faktörleri daha işin başlangıç aşamasında bulmaktır. İşi anlama basamağına gerektiği önemin verilmemesi, işyeri açısından, yanlış sorulara doğru cevaplar üretmekten başka bir anlam ifade etmeyecektir.

Şekil 3.1'te görüldüğü üzere bu basamakta, iş amaçlarının ve başarım kriterlerinin tam olarak belirlenmesi, mevcut durumun değerlendirilmesi, veri madenciliği yapmaktaki amacın belirlenmesi ve proje planının hazırlanması gibi işlemler yapılmalıdır. Diğer bir deyişle kim, neyi, ne zaman, nerede, niçin ve nasıl yapılmalı sorularının cevaplarını bulunmasıdır.

Veri Madenciliğinin ilk adımı olan işi anlama basamağı doğru bir şekilde yapılmadığı sürece, ne kadar mükemmel veri temizleme veya ne kadar verimli modelleme yapılırsa yapılsın, sonuç güvenilemez olacaktır. İşin tam olarak anlaşılmadan, çözülmek istenen problemin tanımlanamayacağı ve sonuçların doğru bir şekilde açıklanamayacağı açık bir gerçektir.

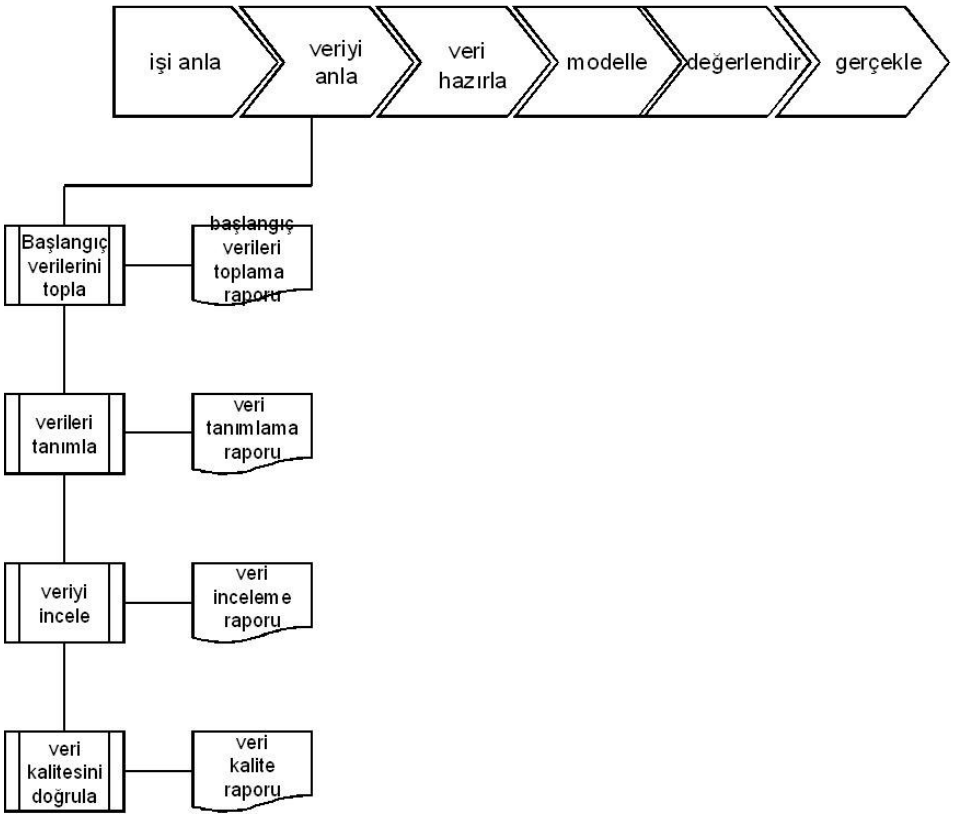
İşi anlama basamağının tam yapılmadığı durumlarda, veri madenciliği sonuçlarına belli bir oranda güvenilebilir, ama veri madenciliğinden maksimum fayda sağlamak için bu basamağın layıkıyla yapılması gerekmektedir.

Problemin açık bir şekilde betimlenmesi; daha sonraki aşamalarda hangi yöntemin kullanılması gerektiği ya da hangi yolun izlenmesi gerektiği konularında gerekli ipuçlarını verecektir. İşi anlama basamağını özetleyen yapı, Şekil 3.2'de gösterilmiştir.



Şekil 3.2**:** İşi anlama

3.2. Veriyi Anlama

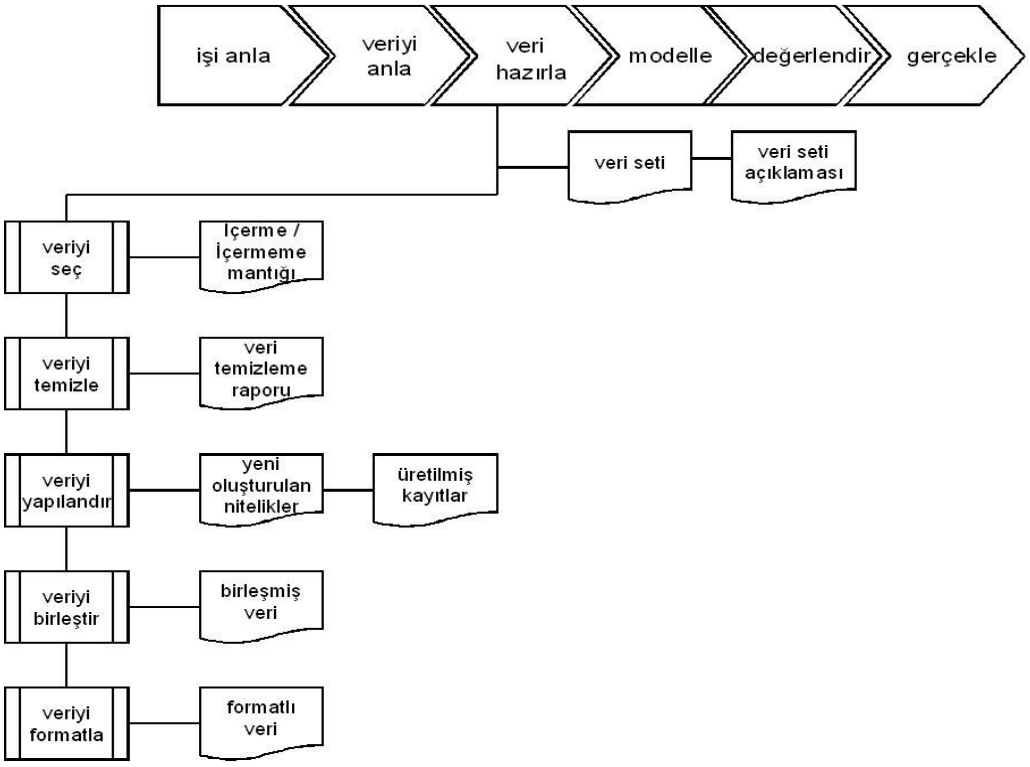


Şekil 3.3: CRISP-DM Basamak 2: Veriyi Anlama

Veriyi anlama basamağında ilk yapılması gereken işlem, verinin toplanmasıdır. Birden fazla kaynakta ve değişik biçimlerde olabilen veri tek bir tablo haline getirilmelidir (Şekil 3.3).

Veri toplandıktan sonra yapılması gereken işlem, verinin genel olarak incelenmesidir. Veriyi anlama basamağının veriyi hazırlama basamağından farkı, verinin hiçbir şekilde değiştirilmemesi, sadece genel olarak incelenmesidir. Veri anlama basamağında incelenen veri, rapor haline getirilebilir. Böylelikle verinin bir sonraki basamakta hangi açıdan temizlenmesi veya hazırlanması gerektiği ortaya çıkartılmış olur.

3.3. Veriyi Hazırlama

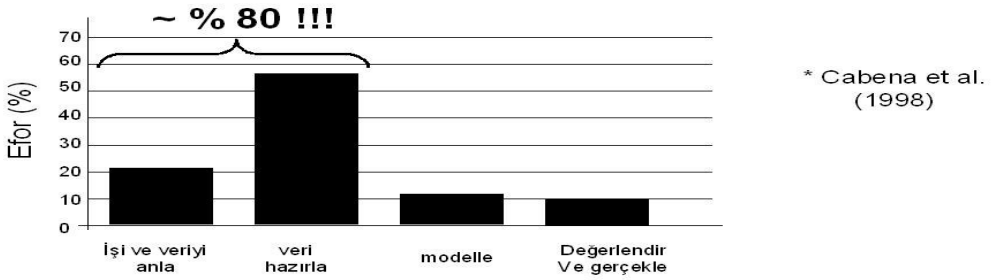


Şekil 3.4: CRISP-DM Basamak 3: Veriyi Hazırlama

Şekil 3.4'te veri hazırlama adımları belirtilmiştir. Bir önceki basamakta, veri incelenmiş ve verinin hazırlanması için neler yapılması gerektiği kaba taslak çıkartılmıştır. Veriyi hazırlama basamağında ise, modelleme yapılabilmesi için gerekli olan her türlü veri hazırlığı işlemleri yapılır.

Günümüzdeki veri madenciliği çalışmaları, genellikle modelleme üzerine odaklanmıştır. Bu yaklaşım teorik açıdan doğru da olsa, gerçek hayatta hiçbir veri

hazır halde değildir. Veri hazırlama işlemine kesinlikle ihtiyaç duyulur. İşi anlama, veriyi anlama ve veriyi hazırlama çalışmaları için harcanan efor, tüm veri madenciliği süreci için harcanan eforun yaklaşık %80'ini oluşturur (Cabena ve diğ., 1998). Yapılan araştırmalarda bu tez doğrulanmıştır. Şekil 3.5'te veri madenciliği sürecinde harcanan eforun dağılımı görülmektedir.



Şekil 3.5**:** Veri Madenciliği Esnasında Harcanan Eforun Dağılımı

3.3.1. Veri seçimi

Veriyi anlama basamağında ilgili veri toplanır. Toplanan bu verinin, doğal olarak bir çok özniteliği olacaktır. Veri seçimi ile yapılması gereken, veri madenciliği için hangi özniteliklerin gerekli olduğuna karar vermektir. Sonuca etkisinin olmadığı değerlendirilen öznitelikler, veri madenciliği kapsamından çıkartılır. Çok fazla öznitelik ile çalışmak, hem zaman açısından, hem de veri madenciliği sonuçlarının güvenilirliği açısından sorun yaratacaktır. Dolayısıyla, modelleme yapılırken, sadece sonuca etki edeceği değerlendirilen öznitelikler göz önüne alınmalıdır. Bu noktada, CRISP-DM yöntembiliminin ilk iki basamağının önemi bir kez daha ispatlanmaktadır. Çünkü, hangi özniteliğin sonuca ne derecede etkisi olacağının tahmini için, öncelikle sahip olunan veri ve iş yerinin amaçları konusunda gerçek bilgilere ihtiyaç duyulmaktadır.

Veri madenciliği için kullanılan özniteliklerin arasında, birbirleriyle bire bir ilişki içerisinde olan, birbirini tamamlayan öznitelikler bulunmamalıdır. Aralarında bire bir ilişki bulunan öznitelikler, veri madenciliği sonuçlarının; ya olması gerekenden çok fazla abartılmış, ya da tamamen yanlış sonuçlar veriyor olmasına neden olabilir. Böyle bir durumda veri madenciliği çalışmaları sonucunda önemli kararlar alacak

veya hareket yöntemleri geliştirecek olan işletme ve organizasyonlar için veri madenciliği, oldukça tehlikeli bir araç haline gelebilir.

3.3.2. Veriyi temizleme

Gerçek anlamda veri temizlemenin ilk şartı, veriyi görselleştirmektir. Çeşitli grafikler kullanılarak, hangi verinin normal dışı olduğunu tespit etmek gibi kirli verinin tespit edilmesi işlemleri çok kolaylaşmaktadır.

Veri temizlemeden kasıt; hatalı, normal dışı, eksik, tutarsız ve tekrar eden verinin uygun şekilde düzeltilmesidir.

Veri hatalı olabilir, örneğin maaş değeri için eksi bir değer girilmesi gibi. Hatalı olan verinin mutlaka düzeltilmesi gerekir. Hatalı kayıt, mantık dışı kayıt anlamında kullanılmıştır.

Veri tutarsız olabilir, örneğin, yaş özniteliği için 35 girildiği halde, doğum tarihi olarak 23.05.1996 girilmesi gibi. Birbirleriyle ilişkisi olan özniteliklerin değerleri arasında da benzer bir ilişki olması gerekir. Diğer taraftan, birbirleriyle lineer bir ilişkisi olan öznitelikler, veri madenciliği sonuçları açısından zararlıdır. Bu özniteliklerden birinin veri seçimi esnasında çıkartılması gerekmektedir. Örneğin, bir tablonun özniteliklerinden birinin maaşı, diğer bir özniteliğin de zam oranını verdiği ele alınsın. Tüm personele maaşının %10'u kadar bir zam yapılacaksa, bu zam oranı özniteliğini de modelleme için hesaba katmak, hatalı sonuçlara neden olabilir.

Veri eksik olabilir. Örneğin, cep telefonu özniteliği için bazı personelin kayıtları boş olabilir. Bu doğaldır, çünkü bazı personel cep telefonu sahibi olmayabileceği gibi, bazı personel cep telefonu bilgisini vermek istemeyebilir. Eksik olan kayıtları temizlemenin genel anlamda iki yolu vardır. Bunlardan birincisi, diğer örneklere bakılarak olabilecek en uygun değerin verilmesi, ikincisi ise o kayıtın tamamen silinmesidir. İkinci metod olan kayıtın silinmesi çözümü için, eksik kayıt oranın gerçekte çok düşük, yani ihmal edilebilir düzeyde olması gerekir.

Veri, birden fazla kaynaktan toplanabileceği için, aynı bilgiler değişik kaynaklarda tekrar tekrar yazılmış olabilir. Bunun neticesi olarak, toplanmış olan veride, aynı kayıt birden fazla oluşabilir. Bu durum, veri madenciliği sonuçlarını olumsuz olarak etkileyebilmektedir.

Tüm veri madenciliği çalışmaları içerisinde modelleme ve veri temizlemenin ayrı bir yeri vardır. Her ikisinin de sonuç üzerine etkisi diğerlerinden daha büyüktür.

3.3.3. Veriyi yapılandırma

Veri madenciliği çalışması yapılırken, bazı durumlarda, mevcut öznitelikler yeterli olmayabilir. Bu gibi durumlarda yeni öznitelikler yaratılmasına ihtiyaç duyulur.

Örneğin, bir hastanede tedavi görmüş olan kalp hastalarının kayıtlarının tutulduğu bir tablo ele alınsın. Bu tabloda hastaların kolesterol, tansiyon, kanda bulunan sodyum ve potasyum miktarı bilgileri ile hastanın tedavisinde kullanılan ilaçların tutulduğu farzedilsin. Bu tablonun görselleştirilmesi, yani 2 boyutlu hale getirilmesi durumunda, kandaki sodyum ile potasyum oranı arasında sıkı bir ilişki olduğu görülür. Bu durumda yeni bir değişken yaratılır ve değer olarak sodyumun potasyuma oranı alınır. Bu yeni öznitelik, veri madenciliği açısından, çok daha kesin sonuçlar verecektir. Bu nedenle, yeni bir öznitelik yaratılabiliyor ve bu yeni özniteliğin, veri madenciliği sonuçlarını olumlu olarak etkileyeceği değerlendiriliyorsa, bu yapılmalıdır.

3.3.4. Veriyi birleştirme

Aynı özniteliğe ait verinin, birden fazla tablo ya da kayıtta olması durumunda, bu veri, uygun bir biçimde birleştirilmelidir.
Veri birleştirmenin amacı, aynı özniteliğe işaret eden verinin, farklı ortamlarda ve farklı biçimlerde olması durumunda, bu verinin tek bir öznitelik altında birleştirilmesidir. Örneğin, bir süpermatketler zincirinde 3 farklı tablo tutuluyor

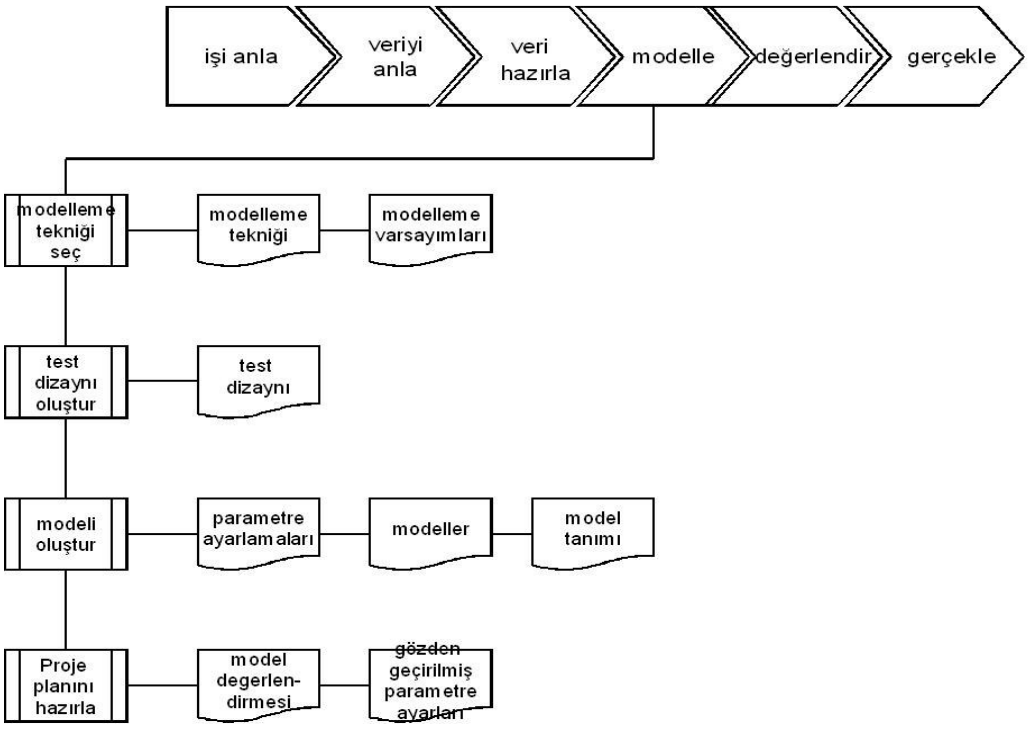
olsun. Birinci tabloda; her bir süpermarketin toplam alanı, tipi gibi genel karakteristik verisi, ikinci tabloda; geçen yıla göre satış rakamı değişiklikleri, kar oranı gibi özet satış verisi, üçüncü tabloda ise; süpermarketin bulunduğu bölgenin demografik yapısı hakkında veri bulunsun. Bu üç tablodaki veri, her bir süpermarket için bir kayıt olacak ve ilgili özniteliklerin karşısına gelecek şekilde tek bir tablo altında birleştirilmelidir.

3.3.5. Veriyi biçimleme

Farklı kaynaklardan gelen veri, aynı özniteliğe ait olsalar bile farklı biçimlerde olabilirler. Örneğin, öğrencilerin notlarını tutan bir öznitelik yaratılmak istensin. A kaynağından alınan veri, öğrencilerin aldığı A, B+, B, C+, C, D+, D, F gibi dereceleri verebilir. B kaynağından alınan veri ise, öğrencilerin aldığı 68, 37, 92 gibi gerçek notları verebilir. Bu durumda ne yapılmalıdır? Her ikisi de aynı özniteliğe ait olduklarına göre, bu verinin iki ayrı öznitelikte tutulmaması gerekir. Bu farklı biçimdeki verinin, uygun bir şekilde, tek bir biçime indirgenmesi şarttır. En çok kullanılan çözüm yolu kullanılarak, öğrencilerin aldığı gerçek notların derece karşılığı bulunup, tüm öğrencilerin notları derece cinsinden yazılmalıdır.

Bazı modelleme tipleri kesin olarak rakamsal (numeric) veya sınıfsal (nominal) değerlere ihtiyaç duyabilir. Örneğin, yapay sinir ağları yöntemi rakamsal değerler ile çalışır. Öğrenci notları özniteliğini de kapsayan bir yapay sinir ağı modellemesi için, derece cinsinden olan tüm notların, rakamsal notlara çevrilmesi gerekecektir. Son yıllarda piyasaya sunulan akıllı veri madenciliği araçları, bu biçim değişimlerini otomatik olarak yapabilmektedir.

3.4. Modelleme



Şekil 3.6**:** CRISP-DM Basamak 4: Modelleme

Modelleme basamağı, veri madenciliğinin en önemli basamaklarındandır. Şekil 3.6'da görüleceği üzere, asıl model bu basamakta oluşturulur. Bu tez çalışmasında, Deniz Kuvvetleri Komutanlığı giyecek sipariş verisi üzerinde, Naive Bayes, karar ağacı ve yapay sinir ağı veri madenciliği yöntemlerinin karşılaştırılması yapılmıştır. Bu üç yönteme ait detaylı bilgi, aşağıdaki maddelerde sunulmuştur.

3.4.1. Naive Bayes

Thomas Bayes, kendisinin ölümünden sonra yayınlanan çalışmasında bir olasılık denklemi ortaya koymuştur (Bayes, 1763). Bu denklem, Bayes denklemi olarak anılır.

Bayes denklemine göre, durumu (E) verilen olayın (H) olasılığının denklemi;

$$\Pr[H \mid E] = \frac{\Pr[E \mid H]\Pr[H]}{\Pr[E]} \quad \text{'dir} \tag{3.1}$$

(3.1) denkleminde, Pr[H]; olayın, durum belirtilmeden önceki olasılığını, Pr[H|E] ise; olayın, durum belirtildikten sonraki olasılığını göstermektedir.

Şekil 3.7'de tenis oynama ile ilgili bir veri tabanı ve öznitelik değerleri verilen yeni bir durum için tenis oynanıp oynanamayacağı sorusu sorulmuştur.

Gün	Hava Durumu	Sıcaklık	Nem	Rüzgar	Tenis Oynanır
1	Güneşli	Sıcak	Yüksek	Zayıf	Hayır
2	Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
3	Bulutlu	Sıcak	Yüksek	Zayıf	Evet
4	Yağmurlu	Ilık	Yüksek	Zayıf	Evet
5	Yağmurlu	Soğuk	Normal	Zayıf	Evet
6	Yağmurlu	Soğuk	Normal	Kuvvetli	Hayır
7	Bulutlu	Soğuk	Normal	Kuvvetli	Evet
8	Güneşli	Ilık	Yüksek	Zayıf	Hayır
9	Güneşli	Soğuk	Normal	Zayıf	Evet
10	Yağmurlu	Ilık	Normal	Zayıf	Evet
11	Güneşli	Ilık	Normal	Kuvvetli	Evet
12	Bulutlu	Ilık	Yüksek	Kuvvetli	Evet
13	Bulutlu	Sıcak	Normal	Zayıf	Evet
14	Yağmurlu	Ilık	Yüksek	Kuvvetli	Hayır

Hava Durumu	Sıcaklık	Nem	Rüzgar	Tenis Oynanır
Güneşli	Soğuk	Yüksek	Kuvvetli	?

Şekil 3.7: Tenis Oynama Veri Tabanı ve Yeni Kayıt Örneği

Naive Bayes denklemi, birden fazla öznitelik için güncellenirse;

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]...\Pr[E_n \mid H]\Pr[H]}{\Pr[E]} \tag{3.2}$$

elde edilir. Örnekteki sorunun cevabını bulmak için, (3.2) denklemi düzenlenirse;

$$\begin{aligned}\Pr[evet \mid E] = {} & \Pr[havadurumu = gunesli \mid evet] \\ & \times \Pr[sicaklik = soguk \mid evet] \\ & \times \Pr[nem = yuksek \mid evet] \times \Pr[ruzgar = kuvvetli \mid evet] \\ & \times \frac{\Pr[evet]}{\Pr[E]}\end{aligned} \tag{3.3}$$

elde edilir. Şekil 3.7'deki veriyi, (3.3)'te verilen denklemde kullanacak şekilde yeni bir şekil oluşturulabilir. Şekil 3.8'de, veri tabanının özniteliklerindeki değerlerin, Naive Bayes yöntemi için özel bir biçimde hazırlanmış hali bulunmaktadır.

hava durumu			sıcaklık			nem			rüzgar			oynanır	
	evet	hayır		evet	hayır		evet	hayır		evet	hayır	evet	hayır
güneşli	2	3	sıcak	2	2	yüksek	3	4	zayıf	6	2	9	5
bulutlu	4	0	ılık	4	2	normal	6	1	kuvvetli	3	3		
yağmurlu	3	2	soğuk	3	1								
güneşli	2/9	3/5	sıcak	2/9	2/5	yüksek	3/9	4/5	zayıf	6/9	2/5	9/14	5/14
bulutlu	4/9	0/5	ılık	4/9	2/5	normal	6/9	1/5	kuvvetli	3/9	3/5		
yağmurlu	3/9	2/5	soğuk	3/9	1/5								

Şekil 3.8: Tenis Oynama Veri Tabanının Naive Bayes Yöntemi İçin Biçimli Hali

Şekil 3.8'de her bir özniteliğin alabileceği değerlerin evet veya hayır olma olasılıkları verilmiştir. Değerler denklemde yerine konulduğunda, yeni kayıt için tenis oynanabilirlik değerinin evet olma olasılığı;

$$\begin{aligned}\Pr[evet \mid E] &= \Pr[havadurumu = gunesli \mid evet] \times \Pr[sicaklik = soguk \mid evet] \\ &\times \Pr[nem = yuksek \mid evet] \times \Pr[ruzgar = kuvvetli \mid evet] \times \frac{\Pr[evet]}{\Pr[E]} \\ &= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0052\end{aligned} \tag{3.4}$$

bulunur. Yeni kayıt için tenis oynanabilirlik değerinin hayır olması olasılığı;

$$\begin{aligned}\Pr[hayir \mid E] &= \Pr[havadurumu = gunesli \mid hayir] \times \Pr[sicaklik = soguk \mid hayir] \\ &\times \Pr[nem = yuksek \mid hayir] \times \Pr[ruzgar = kuvvetli \mid hayir] \times \frac{\Pr[hayir]}{\Pr[E]} \\ &= \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0205\end{aligned} \tag{3.5}$$

olur. Değerler normalleştirilirse;

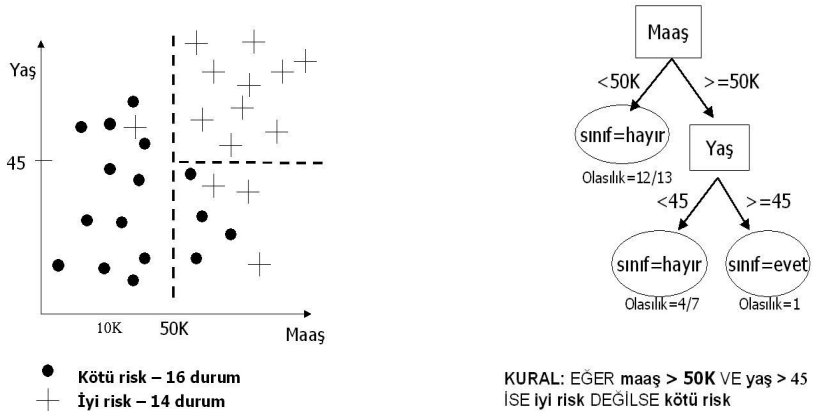
$$P("evet") = \frac{0.0052}{0.0052 + 0.0205} \cong \%20 \tag{3.6}$$

$$P("hayir") = \frac{0.0205}{0.0052 + 0.0205} \cong \%80 \tag{3.7}$$

sonuçları elde edilir. Buradan sonuçla debilebilir ki; Şekil 3.7'da verilen yeni kayıt için tenis oynanır özniteliğinin evet olma olasılığı % 20, hayır olma olasılığı % 80'dir.

3.4.2. Karar ağaçları

Karar ağaçları, büyük miktardaki kayıtların bir dizi basit kurallar uygulanarak, daha küçük kayıtlara bölünmesi için kullanılan bir yapıdır. Başarılı olan her bölünme sonucunda, elemanlarının birbirleriyle olan benzerliğinin arttığı küçük kümeler oluşturulur. Basit karar kuralları ve bunun sonucunda oluşan karar ağacı Şekil 3.9'te gösterilmiştir.

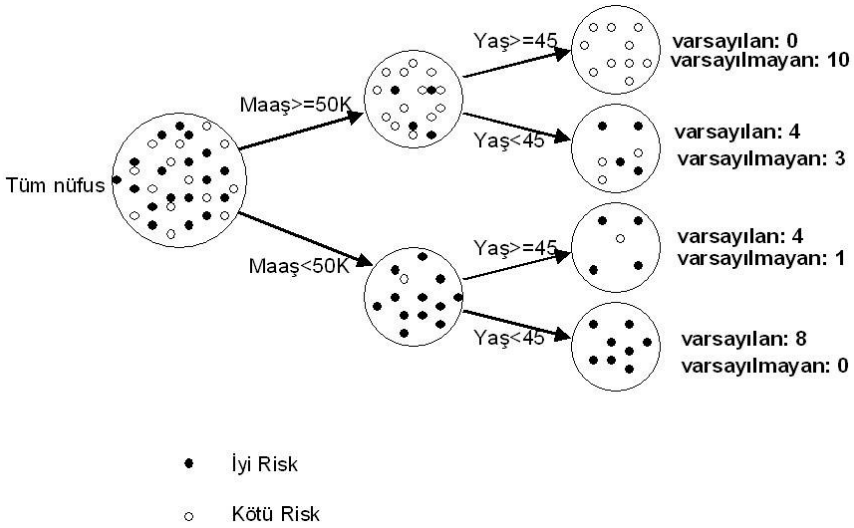


Şekil 3.9**:** Karar Ağacı Oluşturma

Karar ağacı yöntemi, büyük ve heterojen bir kayıdın, daha küçük ve homojen alt gruplara tek bir sonuç değişkenine bağlı olarak bölünmesini sağlayan kurallar kümesidir. Şekil 3.10'da verilen örnekte müşterilerin kredi riski, sonuç değişkenidir.

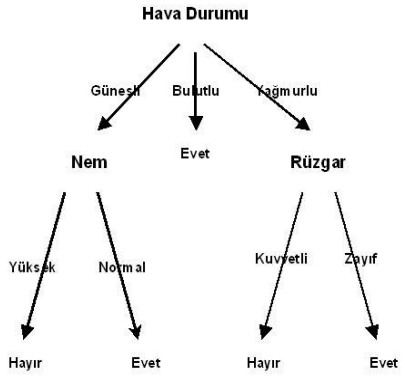
Karar ağaçlarında alt kümelere bölünmeler tüm seviyelerde ikili (binary) olabileceği gibi, bazı seviyelerde veya tüm seviyelerde ikiden fazla bölünme de olabilir. Şekil 3.11'deki örnekte de görüldüğü üzere, hava durumu özniteliğinin alabileceği değerler kümesi üçtür, bu yüzden bu öznitelik üçe bölünmüştür. Hava olaylarına göre tenis

oynanabilirlik veri örneği Quinlan (1996)'dan alınmıştır ve veri madenciliği veya makine öğrenmesi gibi alanlarda sıklıkla kullanılır.



Şekil 3.10**:** Heterojen Gruptan Homojen Gruplar Oluşturma

Gün	Hava Durumu	Sıcaklık	Nem	Rüzgar	Tenis Oynanır
1	Güneşli	Sıcak	Yüksek	Zayıf	Hayır
2	Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
3	Bulutlu	Sıcak	Yüksek	Zayıf	Evet
4	Yağmurlu	Ilık	Yüksek	Zayıf	Evet
5	Yağmurlu	Soğuk	Normal	Zayıf	Evet
6	Yağmurlu	Soğuk	Normal	Kuvvetli	Hayır
7	Bulutlu	Soğuk	Normal	Kuvvetli	Evet
8	Güneşli	Ilık	Yüksek	Zayıf	Hayır
9	Güneşli	Soğuk	Normal	Zayıf	Evet
10	Yağmurlu	Ilık	Normal	Zayıf	Evet
11	Güneşli	Ilık	Normal	Kuvvetli	Evet
12	Bulutlu	Ilık	Yüksek	Kuvvetli	Evet
13	Bulutlu	Sıcak	Normal	Zayıf	Evet
14	Yağmurlu	Ilık	Yüksek	Kuvvetli	Hayır

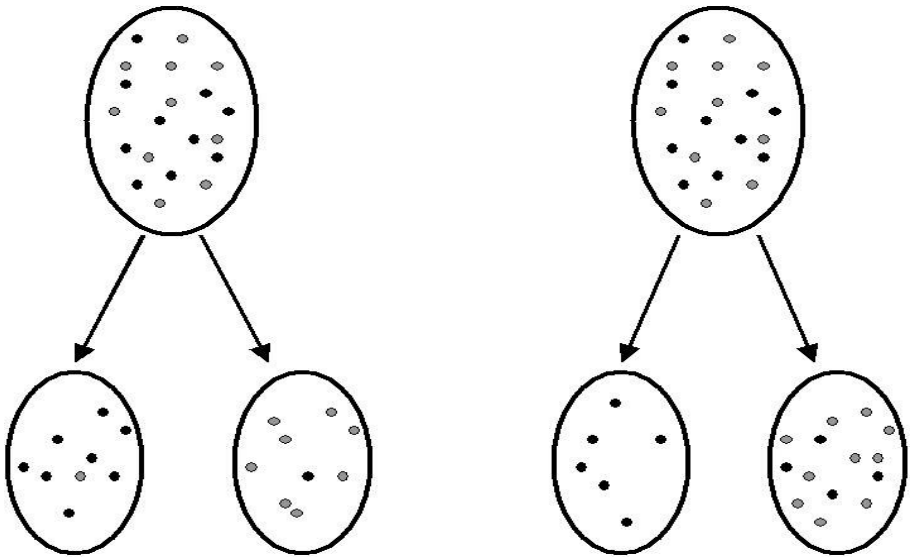


Şekil 3.11**:** Veriden Karar Ağacı Oluşturulması (Quinlan, 1996)

Bir çok karar ağacı algoritması mevcuttur, bunlardan bazıları şunlardır;

- Hunt Algoritması
- CART,CHAID
- ID3, C4.5, C5.0
- SLIQ, SPRINT...

Karar ağaçlarında amaç; bölünme sayısını minimum tutacak şekilde mümkün olduğu kadar saf alt gruplar oluşturmaktır. Hangi özniteliğin bölünmesi gerektiği kararını verebilmek için bilgi kazancı (information gain), entropi (entropy), kirlilik (impurity) ve Gini indeksi (Gini index) değerleri hesaplanır. Örneğin, Şekil 3.12'de sorulan sorunun cevabını bulmak için;



Şekil 3.12**:** Hangisinin Saflığı Daha Yüksektir?

Gini indeksi denklemi şöyledir;

$$Gini(T) = 1 - \sum_{j-1}^{n} p_j{}^2 \tag{3.8}$$

Burada p_j ifadesi, j sınıfının T dallanması içindeki görece frekansını vermektedir. Birinci ağacın sol dallanmasının Gini indeksi;

$$\text{Gini\_1\_sol} = \left(\frac{1}{10}\right)^2 + \left(\frac{9}{10}\right)^2 = 0.01 + 0.81 = 0.82 \tag{3.9}$$

Birinci ağacın sağ dallanmasının Gini indeksi;

$$\text{Gini\_1\_sağ} = \left(\frac{9}{10}\right)^2 + \left(\frac{1}{10}\right)^2 = 0.81 + 0.01 = 0.82 \tag{3.10}$$

Birinci ağacın genel Gini indeksi ;

$$\text{Gini\_1\_genel} = \left(\frac{10}{20} \times Gini_{sol}\right) + \left(\frac{10}{20} \times Gini_{sag}\right) = \left(\frac{10}{20} \times 0.82\right) + \left(\frac{10}{20} \times 0.82\right) = 0.82 \tag{3.11}$$

İkinci ağacın sol dallanmasının Gini indeksi;

Gini_2_sol = 1'dir. (3.12)

İkinci ağacın sağ dallanmasının Gini indeksi;

$$\text{Gini\_2\_sağ} = \left(\frac{4}{14}\right)^2 + \left(\frac{10}{14}\right)^2 = 0.082 + 0.510 = 0.592 \tag{3.13}$$

İkinci ağacın genel Gini indeksi;

$$\text{Gini\_2\_genel} = \left(\frac{6}{20} \times Gini_{sol}\right) + \left(\frac{14}{20} \times Gini_{sag}\right) = \left(\frac{6}{20} \times 1\right) + \left(\frac{14}{20} \times 0.592\right) = 0.714 \tag{3.14}$$

Birinci karar ağacının Gini indeksi olan 0.82 değeri, ikinci karar ağacının Gini indeksi olan 0.714 değerinden daha büyük olduğu için, birinci ağacın saflığı daha yüksektir.

Entropinin hesaplanma denklemi şöyledir;

$$Entropi(p_1, p_2, ..., p_n) = -(p_1 \log p_1 + p_2 \log p_2 + ... + p_n \log p_n) \tag{3.15}$$

Burada p değerleri olasılığı vermektedir. İşlem entropi kanalıyla hesaplanırsa;

$$\text{Entropi}=(-1)\times\left(P(koyurenk).\log_2 P(koyurenk)+P(acikrenk).\log_2 P(acikrenk)\right) \quad (3.16)$$

bulunur. Örnekte, hem koyu renkli için hem de açık renkli için olasılık %50, yani 0.5'tir. Buradan; birinci ağacın genel entropisi;

$$\begin{aligned} entropi &= -\left(0.5\times\log_2(0.5)+0.5\times\log_2(0.5)\right)=-\left(0.1\times\log_2(0.1)+0.5\times\log_2(0.1)\right) \\ &= 0.33+0.14=0.47 \end{aligned} \quad (3.17)$$

değeri elde edilir. Birinci ağacın bilgi kazancı;

$$\text{Bilgi\_kazancı\_1 = 1 -entropi=1-0.47=0.53} \quad (3.18)$$

bulunur. İkinci ağacın sol tarafı için entropi = 0'dır.

İkinci ağacın sağ tarafı için entropi;

$$\text{entropi} = -\left(\left(\frac{4}{14}\right)\log_2\left(\frac{4}{14}\right)+\left(\frac{10}{14}\right)\log_2\left(\frac{10}{14}\right)\right)=0.516+0.347=0.863 \quad (3.19)$$

hesaplanır. İkinci ağacın genel entropisi ;

$$\text{entropi} =\left(\frac{6}{20}\times Entropi_{sbl}\right)+\left(\frac{14}{20}\times Entropi_{sag}\right)=\left(\frac{6}{20}\times 0\right)+\left(\frac{14}{20}\times 0.863\right)=0.604 \quad (3.20)$$

bulunur. İkinci ağacın bilgi kazancı;

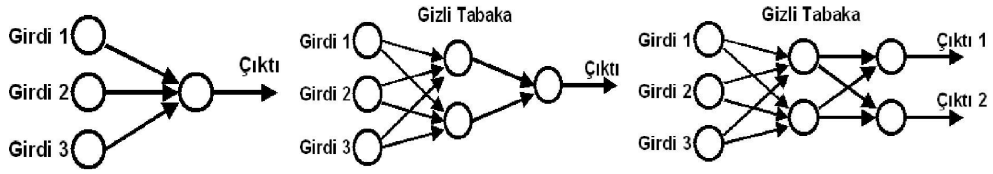
$$\text{Bilgi\_kazancı\_2 = 1 -entropi=1-0.604=0.396} \quad (3.21)$$

değeri bulunur. Birinci karar ağacının bilgi kazancı (0.53), ikinci karar ağacının bilgi kazancından (0.396) daha büyük olduğu için Gini indeksinde olduğu gibi, entropi ve bilgi kazancı hesaplamasına göre de birinci ağacın saflığı daha yüksektir.

3.4.3. Yapay sinir ağları

Yapay sinir ağları, tanım olarak, basit işlemci birimlerden oluşan büyük ve paralel dağılımlı bir işlemcidir. Birimler arası bağlanıtlatıların kuvvetlerinden deneysel bir şekilde bilgi oluşturma ve bu bilginin kulanılmasını sağlama yeteneğine sahiptir (Kantardzic, 2003).

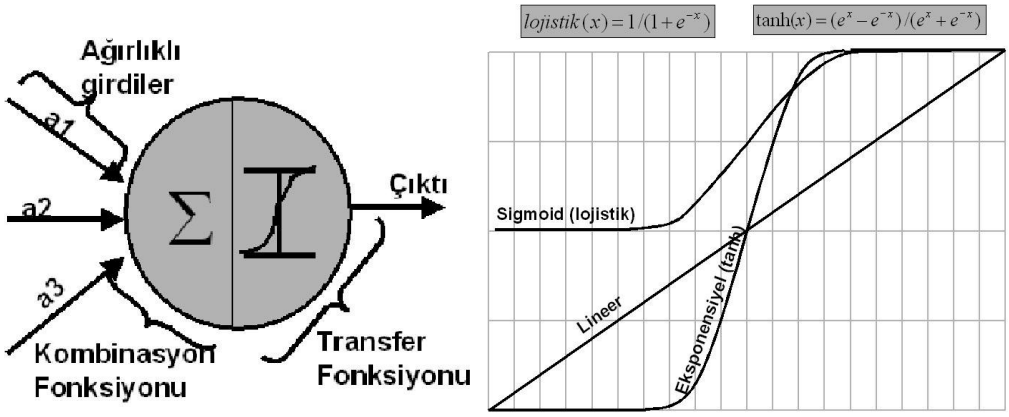
Yapay sinir ağlarının genel yapısı Şekil 3.13'te gösterilmiştir. Veri madenciliğinde kullanılan yapay sinir ağı yöntemleri, genellikle bir veya birden fazla gizli tabakaya sahiptir.



Şekil 3.13**:** Yapay Sinir Ağı Yapıları

Yapay sinir ağları yöntemi, gerçek sinir ağı yapısından esinlenilerek ortaya çıkarılmıştır. Her bir girdinin bir ağırlığı vardır ve çıktı esnasında, bu ağırlıklı girdilerin lineer olmayan kombinasyonu alınarak, bir transfer fonksiyonundan geçirilir. Şekil 3.14'te yapay sinir ağının yapısı ve kullanılan transfer fonksiyonları görülebilir. Bu transfer fonksiyonlarından en çok kullanılanı sigmoid fonksiyondur.

Yapay sinir ağları bağlantı tiplerine göre iki sınıfa ayrılır: ileri beslemeli (feedforward) ve yinelemeli (recurrent). İleri beslemeli yapay sinir ağının özelliği, ağ içindeki akışın, tek yönlü ve girdiden çıktıya doğru olması ve ağ içerisinde bulunan aynı katman içerisinde hiç bir döngü olmamasıdır. Yinelemeli yapay sinir ağlarında ise, aynı katmanda bulunan noktalar arasında döngü vardır. Bunu yapabilmek için bir gecikme zamanı eklenir.



Şekil 3.14**:** Yapay Sinir Ağının Çıktı Yapısı ve Transfer Fonksiyonu

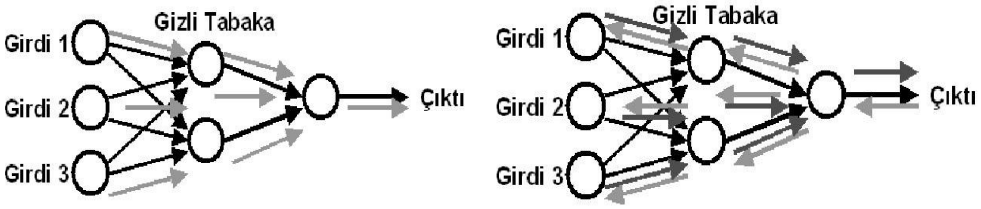
Yapay sinir ağları, daha iyi öğrenme yeteneğine sahip olabilmek için, miktarını kendisinin belirlediği bir veya daha fazla gizli katman kullanmaktadır. Bu tip yapay sinir ağlarına çok katmanlı yapay sinir ağı denir.

Endüstride ve işletmelerde kullanılan yapay sinir ağı uygulamalarının yaklaşık % 90'ı ileri beslemeli çok katmanlı yapay sinir ağıdır (Kantardzic, 2003). İleri beslemeli çok katmanlı yapay sinir ağları, genel olarak çok katmanlı algılayıcı (MLP-Multilayer Perceptron) olarak bilinmektedir.

Bu tez çalışmasının uygulama bölümünde yöntemler karşılaştırılırken, WEKA'nın çok katmanlı algılayıcı yöntemi kullanılmıştır.

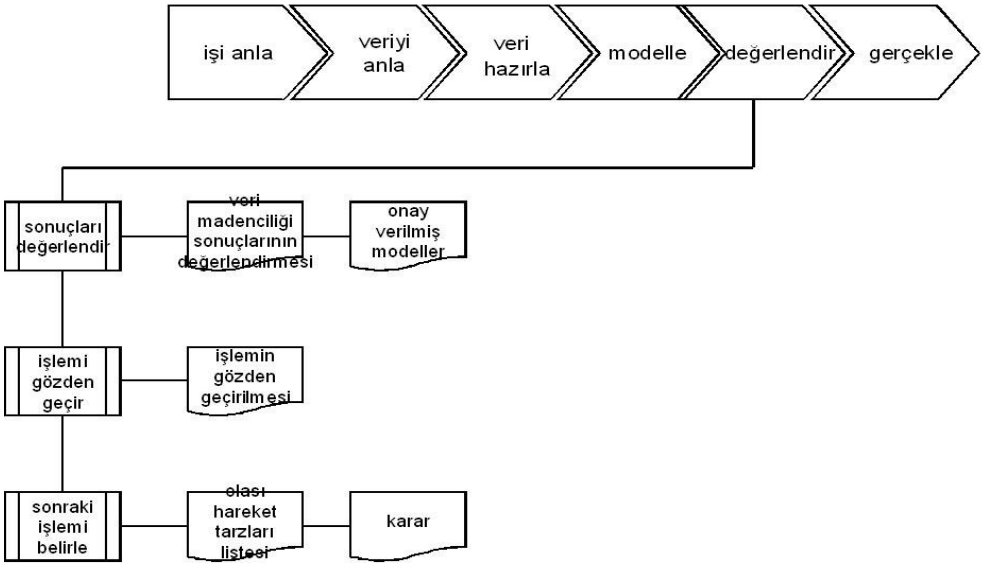
Çok katmanlı algılayıcının en zor problemlerin bile üstesinden gelebilmesini sağlayan, hatanın geri yayılması algoritmasıdır (error back propogation algorithm). Geri yayılmalı yapay sinir ağının özelliği, elde edilen tahmini çıktı değeri ile gerçek çıktı değerinin karşılaştırılması ve aradaki farkın azaltılması için, girdilerin ağırlıklarının değiştirilmek üzere, ağ akış yönünün geriye doğru yönlendirilmesidir. Dolayısıyle, geri yayılma algortimasının temeli, çıktı değerinde bir hata olduğu varsayımıyla başlar.

Şekil 3.15'te hem ileri beslemeli, hem de ileri beslemeli geri yayılmalı yapay sinir ağının ağ akış yönleri gösterilmiştir.



Şekil 3.15**:** İleri Beslemeli ve Geri Yayılmalı Yapay Sinir Ağı Yapısı

3.5.Değerlendirme



Şekil 3.16**:** CRISP-DM Basamak 5: Değerlendirme

3.5.1. Genel değerlendirme esasları

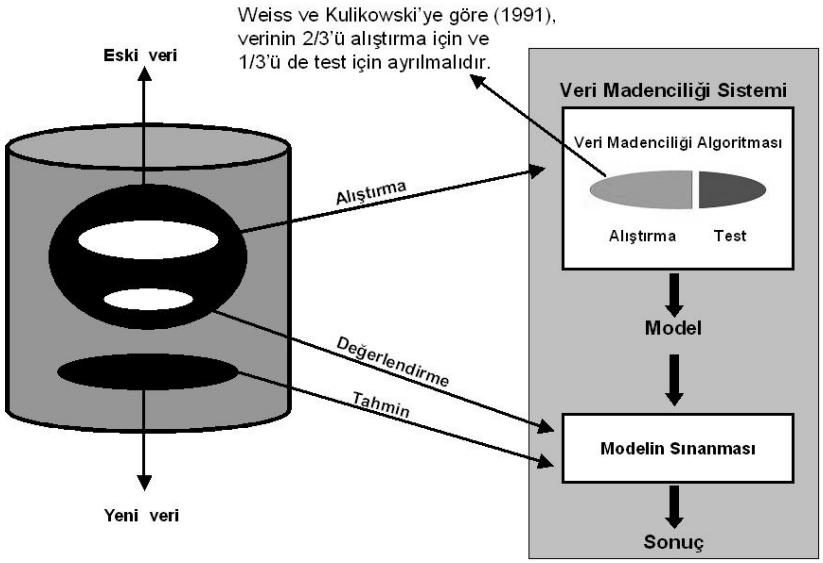
Değerlendirme adımında, önceki adımlarda yapılmış olan tüm işlemler tekrar gözden geçirilmeli, veri madenciliği sonucunun, en başta belirlenen iş problemine tam olarak cevap verip vermediğine bakılmalıdır. Bunların yanında, kullanılan veri madenciliği yönteminin, mevcut veriye uygunluğu ve oluşturulan modelin, gerçekten doğru bir model olup olmadığı değerlendirilmelidir.

3.5.2. Verinin modelleme ve değerlendirme için kullanılma yöntemleri

Veri madenciliği sonuçlarının değerlendirilebilmesi için, öncelikle oluşturulan modelin gerçekten iyi bir model olduğundan emin olmak gerekmektedir. Büyük boyutlu veri için, genellikle Weiss ve Kulikowski (1991)'in önerdiği sistem uygulanır. Bu iki araştırmacıya göre, büyük boyutlu verinin 2/3'ü modeli oluşturmak ve 1/3'ü de modeli test etmek için kullanılmalıdır. Alıştırma ve testler neticesinde en uygun model oluşturulduktan sonra bu model, değerlendirme için ayrılan kısım ile veya veri tabanına yeni giren veri ile sınanır. Bu sınamalar sonucunda, veri madenciliğinin sonucuna ulaşılır. Şekil 3.17, verinin, veri madenciliği sonucu üretmek için nasıl kullanılması gerektiğini göstermektedir.

Eğer, mevcut verinin boyutu çok fazla değilse, mevcut veriden en iyi model oluşturma yöntemi çapraz doğrulamadır. Çapraz doğrulamada uygulanması gereken üç basamak vardır;

- Birinci adımda; tüm veri küçük eşit parçalara bölünür. Bunun ideali on eşit parçaya bölmektir. Verinin boyutu ne kadar küçük ise, veriyi bir o kadar çok eşit parçaya bölmektir.

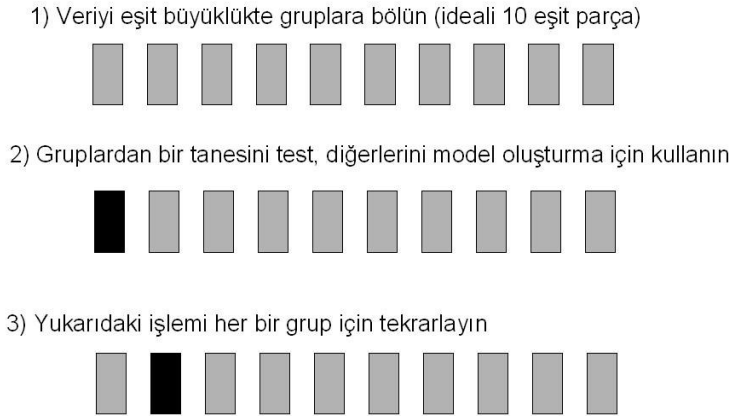


Şekil 3.17**:** Büyük Veri Kümeleri İçin Verinin Kullanılması

- İkinci adımda; eşit parçalara bölünmüş olan verinin ilk parçası test, diğerleri model oluşturma için kullanılır. Yani, on eşit parçaya bölünmüş bir veri topluluğu için, dokuz eşit parçadaki veri kullanılarak modelleme yapılır ve onuncu parça oluşturulan parçanın test edilmesi için kullanılır.

- Üçüncü adım; ikinci adımın tekrarlamalı (iterasyonlu) halidir. Yani, daha önce test işlemi için ayrılmamış tüm parçalar, birer birer test için ayrılır ve kalan dokuz parça modelleme için kullanılır.

Tüm parçaların bir kere test ve dokuz kere modelleme için kullanılma işlemi tamamlandıktan sonra, aralarında en iyi sonucu veren model, sistemin modeli olarak kullanılır. Çok büyük veri için bu yöntemin kullanılması durumunda, çok fazla zaman alacağı açıktır. Ayrıca, büyük boyutlu veriden, doğru sonuçlar üreten bir model oluşturmak daha kolay olduğu için, çapraz doğrulama metoduna gerek yoktur. Çapraz doğrulamaya ilişkin grafik, Şekil 3.18'da görülebilir.



Şekil 3.18**:** Çapraz Doğrulama

3.5.3. Değerlendirme analizleri

3.5.3.1. Doğruluk oranı

Veri madenciliği sonuçlarının değerlendirilmesi için çeşitli yöntemler mevcuttur. Çeşitli tablolar ve hesaplamalar ile veri madenciliği modellemesinin sonuçları değerlendirilebilir. Bunların ilki, doğruluk oranı değerlendirmesidir;

$$dogruluk\_orani = \frac{S}{N} \tag{3.22}$$

Burada,

S : Doğru tahminlerin sayısı,

N : Tüm kayıtların sayısıdır.

Örneğin doğru tahminlerin sayısı 18.000, tüm kayıt sayısı da 20.000 olması durumunda;

$$dogruluk\_orani = \frac{S}{N} = \frac{18.000}{20.000} = 90\% \tag{3.23}$$

bulunur. Model, %90'lık bir doğruluk oranına sahiptir. Doğruluk oranı ne kadar büyükse, modelleme o kadar başarılıdır. Doğruluk oranı ile modellemenin başarısı arasında doğru orantı vardır.

Doğruluk oranı değerlendirmesi için tablo yapısı da kullanılabilir. Şekil 3.19'daki tabloda mevcut özniteliklerin yanında gerçek çıktı, modelin tahmini ve tahminin doğru / yanlış olması gibi sütunlar eklenmiştir. Böylelikle, modelin tahmin ettiği değer ile gerçek değer karşılaştırılarak, doğru tahminlerin tüm kayıtlara oranı bulunabilir.

Şekil 3.20'de iki ayrı durum için doğruluk oranlarını gösteren bir başka tablo yapısı verilmiştir. Bu tablo türü, yapıyı genel anlamda göstermesi açısından oldukça sık

kullanılmaktadır. WEKA veri madenciliği aracı bu doğruluk oranı tablosunu kullanmaktadır.

Girdiler				Gerçek çıktı	Modelin Tahmini	Doğru / Yanlış Tahmin
Hava Durumu	Sıcaklık	nem	Rüzgar	Tenis oynanır	Tenis oynanır	
Güneşli	Ilık	Yüksek	Zayıf	1	1	√
Bulutlu	Sıcak	Normal	Kuvvetli	0	0	√
Güneşli	Soğuk	Normal	Zayıf	0	1	X
Yağmurlu	Ilık	Yüksek	Kuvvetli	1	1	√
...	...	...	...	...	...	...

Şekil 3.19: Doğruluk Oranı Tablosu

Durum 1

Tahmin Edilen Sonuç	Gerçek Sonuç	
	İyi Risk	Kötü Risk
İyi Risk	60	10
Kötü Risk	20	10

S/N: 70 %

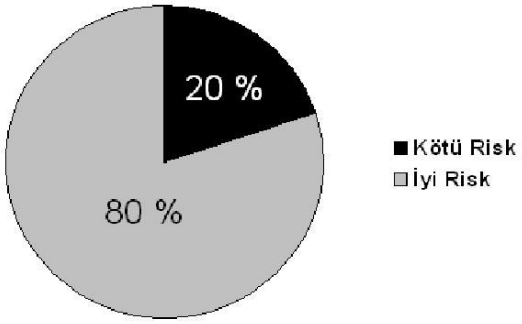
Durum 2

Tahmin Edilen Sonuç	Gerçek Sonuç	
	İyi Risk	Kötü Risk
İyi Risk	70	5
Kötü Risk	10	15

S/N: 85 %

Şekil 3.20: Doğruluk Oranı İçin Karşılaştırmalı Tablo

İyi bir veri madenciliği çalışması için minimum doğruluk oranı ne olmalı sorusunun cevabı, rasgele bir seçime göre makul bir kazanç sağlaması olmalıdır. Bankanın müşterilerini kredi verme durumuna göre iyi riskli veya kötü riskli olarak sınıflandırma örneğine bakılırsa (Şekil 3.21); bankanın tüm müşterileri iyi riskli olarak sınıflandırılmış olunursa, modelin doğruluk oranı zaten %80 olacaktır. Dolayısı ile Şekil 3.20'deki örneğe geri dönüldüğünde; birinci durum, rasgele bir seçime göre kazanç değil, kayıp sunmaktadır. Bu yüzden kabul edilemez. İkinci durum ise, rasgele seçime göre fazladan %5'lik bir kazanç sunmaktadır ki, bir çok durumda bu oran, veri madenciliği sonucu olarak yeterlidir denilebilir



Şekil 3.21: Tüm Müşterilerin İyi Riskli Olarak Kabul Edilmesi Durumu

3.5.3.2. Ortalama kareler hatası (Mean squared error - MSE) ve ortalama mutlak hata (mean absolute error-MAE)

Veri madenciliği elde edilen sonucun hata miktarını bulmak için çeşitli denklemler vardır. Bunların arasından en çok kullanılanı, ortalama kareler hatası ve ortalama mutlak hatadır. Hata miktarının deklemi (3.24)'te verilmiştir.

$$Hata_i = (p_i - a_i) \tag{3.24}$$

olacaktır. Burada;

$a_1, a_2, ..., a_n$: gerçek miktarlar

$p_1, p_2, ..., p_n$: tahmin edilen miktarlardır. Bu durumda, ortalam kareler hatası;

Ortalama Kareler Hatası: $MSE = \frac{\left[(p_1 - a_1)^2 + (p_2 - a_2)^2 + ... + (p_n - a_n)^2\right]}{n}$ (3.25)

denklemiyle gösterilir.

Şekil 3.22'de verilen örnek tablo için ortalama kareler hatası, öncelikle 1 sıra no'lu kayıtın sipariş miktarı özniteliğinin 80 olduğu duruma göre hesaplanırsa;

$$MSE = \frac{\left[(83-80)^2 + (131.3-140)^2 + (178-175)^2 + (166-168)^2 + (117-120)^2 + (198-189)^2\right]}{6} = 31.86 \tag{3.26}$$

elde edilir. Ortalama kareler hatası, 1 sıra no'lu kayıtın sipariş miktarı özniteliğinin 400 gibi normal dışı bir değer olduğu durumuna göre hesaplanırsa;

$$MSE=\frac{\left[(83-400)^2+(131.3-140)^2+(178-175)^2+(166-168)^2+(117-120)^2+(198-189)^2\right]}{6}=16778.5 \tag{3.27}$$

değeri elde edilir. Yukarıda verilen iki örneğin sonuçları karşılaştırıldığında, ortalama kareler hesabının zayıf noktası ortaya çıkmaktadır. Görüldüğü üzere, özniteliklerden herhangi birinin normal dışı bir değer alması durumunda, sonuç olağan üstü değişikliğe uğramaktadır. Ortalama kareler hatasının bu zayıf noktasını ortadan kaldırmak üzere, ortalama mutlak hata denklemi oluşturulmuştur.

Sıra	Gelir	Yaş	Sipariş Miktarı	Tahmin edilen Sip. Mik.
1	23.000	30	80 / 400	83
2	51.100	40	140	131,1
3	68.000	55	175	178
4	74.000	46	168	166
5	23.000	47	120	117
6	100.000	49	189	198

Şekil 3.22: Ortalama Kareler Hatası ve Ortalama Mutlak Hata İçin Örnek Tablo

Ortalama Mutlak Hata (MAE), farkların kareleri yerine mutlak değerlerini aldığı için, normal dışı değerlerin sonuca olan etkisi azaltılmıştır. Ortalama mutlak hatayı bulan denklem (3.28)'de verilmiştir.

$$MAE=\frac{\left[|p_1-a_1|+|p_2-a_2|+\ldots+|p_n-a_n|\right]}{n} \tag{3.28}$$

Şekil 3.22'de verilen örnek tablo için ortalama mutlak hata, öncelikle 1 sıra no'lu kayıtın sipariş miktarı özniteliğinin 80 olduğu duruma göre hesaplanırsa;

$$MSE=\frac{\left[|83-80|+|131.3-140|+|178-175|+|166-168|+|117-120|+|198-189|\right]}{6}=4.8 \tag{3.29}$$

değeri elde edilir. Ortalama kareler hatası, 1 sıra no'lu kayıtın sipariş miktarı özniteliğinin 400 gibi normal dışı bir değer olduğu duruma göre hesaplanırsa;

$$MSE = \frac{\left[|83-400| + |131.3-140| + |178-175| + |166-168| + |117-120| + |198-189|\right]}{6} = 57.15 \quad (3.30)$$

değeri elde edilir. Yukarıda verilen örneklerde görüldüğü üzere, ortalama mutlak hata, normal dışı değerlere karşı çok daha verimli sonuçlar üretebilmektedir.

3.5.3.3. Maliyet duyarlı değerlendirme (Cost sensitive evaluation)

Gerçekte kötü risk tanımlı birisine, iyi risk tahminiyle kredi vermenin maliyetinin kişi başına 20 YTL, gerçekte iyi risk tanımlı birisine, kötü risk tahminiyle kredi vermenin maliyetinin kişi başına 5 YTL varsayımından hareketle, Şekil 3.23'teki veri kullanılarak hatanın maliyeti hesaplanırsa;

Durum 1

Tahmin Edilen Sonuç	Gerçek Sonuç	
	İyi Risk	Kötü Risk
İyi Risk	60	0
Kötü Risk	20	20

S/N: 80 %

Durum 2

Tahmin Edilen Sonuç	Gerçek Sonuç	
	İyi Risk	Kötü Risk
İyi Risk	70	5
Kötü Risk	10	15

S/N: 85 %

Şekil 3.23**:** Maliyet Duyarlı Değerlendirme İçin Örnek Sonuçlar

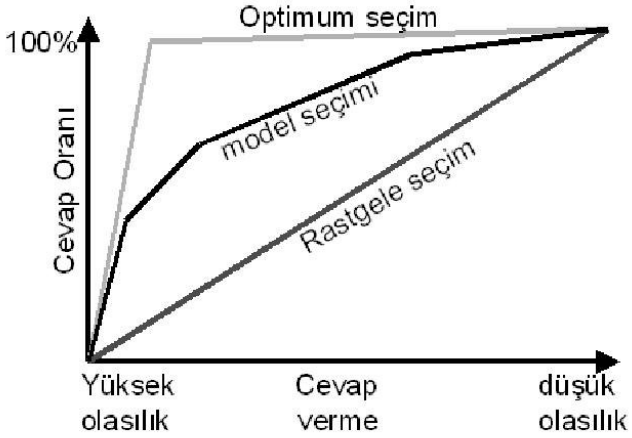
Durum-1 toplam hata maliyeti = $(0\times 20)+(20\times 5)=100YTL$

Durum-2 toplam hata maliyeti = $(5\times 20)+(10\times 5)=150YTL$

değerleri bulunur. Buradan çıkarılması gereken, Şekil 3.23'ün durum-1'inde verilen sonucun S/N oranı daha düşük olduğu halde, durum-2'de verilen sonuçtan daha iyi bir seçenek olduğudur.

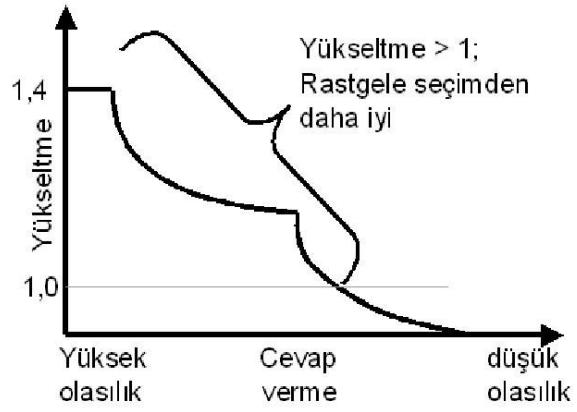
3.5.3.4. Değerlendirme eğrileri

- Cevap eğrisi (Response curve): Genel olarak, veri madenciliğinin sonucu ile rasgele seçimin sonucunun karşılaştırılmasını gösterir. Şekil 3.24'deki örnekte müşterilerine mektup / katalog ile ulaşan bir firmanın müşterilerinden aldığı cevap oranı gösterilmiştir. Rasgele seçimde yükselen bir doğru elde edilirken, veri madenciliği ile optimuma yakın seçimler oluşturulabilir. Bu sayede firma, mektubu / katalogu, cevap verme oranı en yüksek belli sayıdaki müşterisine göndererek, masraflarından büyük tasarruf sağlayabilir.



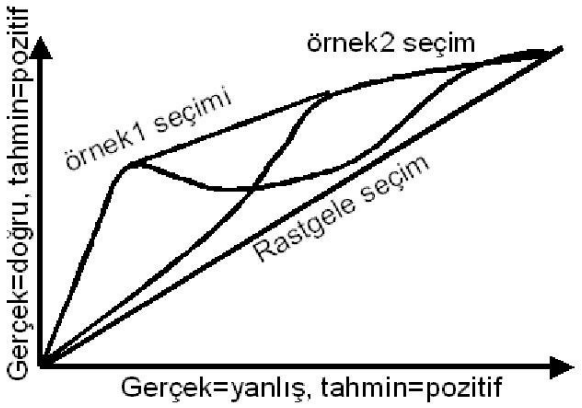
Şekil 3.24**:** Cevap Eğrisi (Response Curve)

- Yükseltme eğrisi (Lift curve) : Bu eğrinin cevap eğrisinden en önemli farkı, modellemenin sisteme olan katkısını, oran olarak ve direk vermesidir. Yükseltme miktarının 1 değerinden büyük olduğu yerler, modellemenin rasgele seçimden daha iyi olduğu durumları gösterir. Firmanın müşterilerine dağıtacağı promosyonlar bu eğriye göre hesaplanabilir. Firma genel olarak, en yüksek yükseltme oranına sahip azınlık müşteri için, en değerli ürünü promosyon olarak verebilir. Promosyon malzemesinin değeri, yükseltme oranına göre göreceli olarak düşer. Bu pazarlama stratejisi, günümüzde oldukça sık kullanılmaktadır (Şekil 3.25).



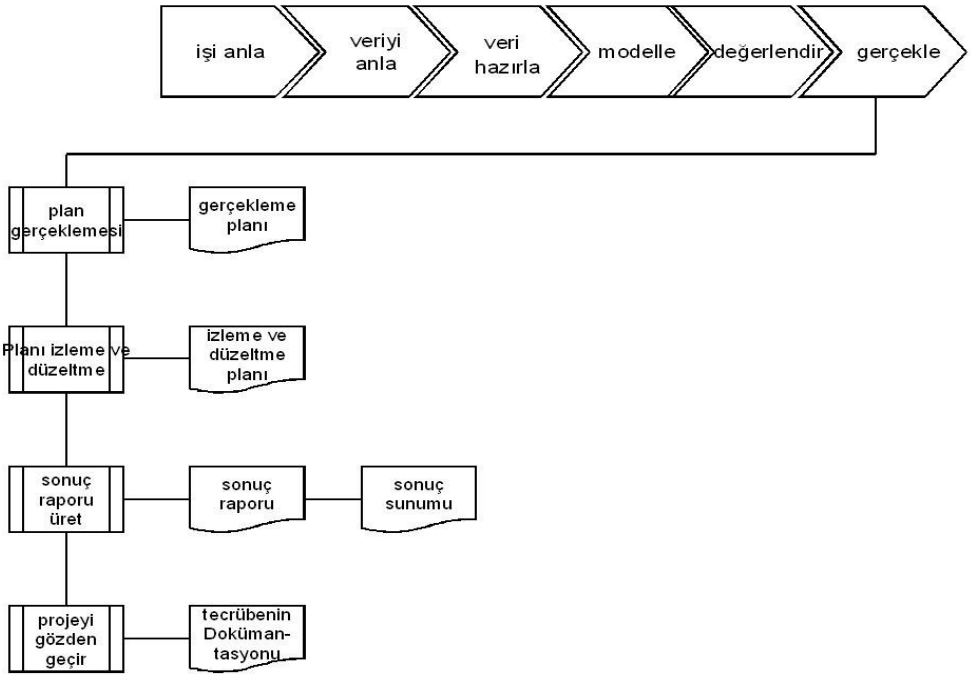
Şekil 3.25**:** Yükseltme Eğrisi (Lift Curve)

- Alıcı işletim eğrisi (Receiver operating curve - ROI): Şekil 3.26'deki eğrileri oluşturmak için, birbirinden bağımsız örnekleme sonuçları tabloya işlenir. Son işlem olarak, bağımsız örneklemelerin tepe noktaları birleştirilir. Bu yaklaşımın verimi, en az cevap eğrisi düzeyindedir. Bağımsız örneklemelerin tepe noktaları daha yüksek değerlere ulaştığı için, genel anlamda sonuç, cevap eğrisine göre çok daha verimlidir.



Şekil 3.26**:** Alıcı İşletim Eğrisi (Receiver Operating Curve - ROI)

3.6. Gerçekleme



Şekil 3.27**:** CRISP-DM Basamak 6: Gerçekleme

Sadece model yaratılması ve modelin değerlendirilmesi genel anlamda yeterli değildir. Modellemenin amacı, verinin bilgiye dönüştürülmesi olsa da, elde edilen bilgi, işletmenin faydalanacağı şekilde organize edilmeli ve sunulmalıdır.

Şekil 3.27'de görüldüğü üzere veri madenciliği yöntembiliminin bu son adımında, gerçekleme, izleme ve düzeltme planları yapılmalıdır. Ayrıca sonuç raporu ve dokümantasyon yapılması da firma için son derece faydalı olacaktır. Gerçekleme işlemi, ihtiyacın durumuna göre, sadece bir rapor oluşturulması gibi basit olabileceğinin yanında, tüm şirketi kapsayan ve sürekli tekrarlanan bir veri madenciliği işlemi kadar da karmaşık olabilir.

Veri madenciliği sonuçları, başka bir yazılıma girdi olarak aktarılabilir. Böyle durumlarda, veri madenciliği işlemlerinin hangi sıklıkta tekrarlanacağı bir plana bağlanmalıdır.

Gerçekleme işleminde genel olarak cevaplanması gereken sorular şunlardır;

- Veri madenciliği sonuçları nasıl kullanılacak?
- Kimler kullanacak?
- Hangi sıklıkta kullanacak?

Bu sorulara verilen yanıtlara göre, işletme veya organizasyonların veri madenciliği sonuçlarını en etkin şekilde kullanabilmesi için yapılması gereken işlemlere gerçekleme denir.

4. CRISP-DM KULLANILARAK DENİZ KUVVETLERİ VERİSİ ÜZERİNDE VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Bu tez çalışmasının uygulama bölümünde, CRISP-DM yöntembilimi kullanılarak Deniz Kuvvetleri Komutanlığı verisinde, veri madenciliği sınıflandırma yöntemlerinin karşılaştırılması yapılmıştır. Uygulamada, Deniz Kuvvetleri Komutanlığı'nda görev yapan personelin, kendilerine verilen kredi doğrultusunda bir yazılım üzerinden vermiş oldukları siparişlere ait veri, üç farklı veri madenciliği sınıflandırma yöntemi ile analiz edilmiştir. Karar ağaçları, yapay sinir ağları ve Naive Bayes veri madenciliği yöntemleriyle ayrı ayrı modellemesi yapılan veri madenciliği çalışması sonucunda; personel tipi, miktar, dönem, kredi yılı gibi girdi verisinden yararlanılarak, siparişi verilen malzeme çıktı verisinin örüntüsü belirlenmeye çalışılmıştır.

Önceki bölümlerde de anlatıldığı üzere, veri madenciliği, sadece modelleme yapılması anlamına gelmemektedir. Günümüzde veri madenciliği terimi, bir işletme ya da organizasyonun veri madenciliği yapabilmek için yapmış olduğu tüm adımları içine alan genel bir şüreçtir. Tezin uygulama bölümünde, tam anlamıyla bir veri madenciliği uygulaması yapılabilmesi gerekli olan adımları içeren yöntembilimlerden, dünya üzerinde en çok kabul göreni olan CRISP-DM referans alınmıştır.

Bu tez çalışmasında, Deniz Kuvvetleri Komutanlığı personelinin giyecek siparişi vermek için kullandığı Kredili Giyecek Sistemi yazılımının veri tabanından alınan veri kullanılmıştır.

Tezin bundan sonraki bölümünde, CRISP-DM yöntembilimi doğrultusunda, uygulamaya ait tüm veri madenciliği süreci anlatılmıştır.

4.1. İşi Anlamak

4.1.1. İş amaçlarının belirlenmesi

Deniz Kuvvetleri Komutanlığı'nda görev yapan subay ve astsubay personel, diğer rütbeli personel gibi mesai saatleri içerisinde üniforma giymek zorundadır. Görev başında giyilmesi gereken bu üniformaları, devlet vermektedir. 2002 yılı öncesinde, her giyeceğin, belirli bir kullanım ömrü vardır mantığıyla bir sistem oluşturulmuştur. Personele verilen giyeceklerin, daha önceden belirlenen kullanım ömürleri bittiğinde, yenisi verilmektedir. Fakat, personelin giymesi gereken üniformalar, görev yerine göre değişmektedir. Örneğin gemilerde görev yapan personel, işbaşı denilen giyeceği çoğunlukla giyerken, kara birliklerinde görev yapan personelin işbaşı giymeleri yasaklanmıştır. Buna benzer örnekler nedeniyle, giyeceklere kullanım ömrü uygulamasından 2002 yılında vazgeçilerek, kredili bir sisteme geçilmiştir. Bu amaç için, kredili giyecek sistemi yazılımı oluşturulmuştur. Personel, bu yazılımı kullanarak, kendisine her yıl verilen kredi doğrultusunda ihtiyaç duyduğu giyeceklerin siparişini verebilmektedir.

Sistemden girilen siparişler, ilgili birimler tarafından değerlendirilerek, dikimevi müdürlükleri için hammadde ve üretim planlaması yapılmaktadır. Kumaş, düğme, iplik gibi ihtiyaç duyulan hammaddeler, 4734 ve 4735 sayılı Kamu İhale Kanunu hükümleri uyarınca satın alınmaktadır. Satınalma sürecinin ortalama olarak altı ay aldığı ve hammaddeler satın alındıktan sonra, sipariş edilen malzemelerin üretilmesi ve istek sahibi personele elden teslim edilmesi sürelerinin de hesaba katılmasıyla, ortalama olarak bir yıl öncesinden siparişlerin verilmesi gerekmektedir.

Personel, kredili giyecek sistemi üzerinden siparişini verdiği malzemeleri, en erken bir yıl sonra alabilmektedir. Bir yıllık teslim alma süresi içerisinde personelin başka bir birliğe ataması olabilmektedir. Bu durumda, ihtiyaç duyacağı giyecekler değişmesine rağmen, sadece belirli zaman aralıklarında sipariş girilebilen kredili giyecek sistemi üzerinden sipariş güncellemesi yapamamaktadır. Bunun sonucunda ataması olan personel, mağdur olmaktadır.

Deniz Kuvvetleri Komutanlığının giyecek sisteminde iyileştirme yapılabilecek alanların bilinemiyor olması, problemlerin en büyüğünü teşkil etmektedir. Çünkü, personelin, giyecek sisteminden tam olarak memnun olmadığı bilinmektedir. Ama sipariş verisinin gerçek anlamda analizi yapılamadığı için, problem sahaları ve çözümleri bilinememektedir. Veri madenciliğinin sisteme olan katkısı, sadece bilinen problemlere çözümler bulmaya çalışmak değil, aynı zamanda problem sahalarını belirlemesi ve çözüm önerilerini sunmasıdır.

4.1.2. Durum değerlendirmesi

Kullanılmakta olan "kredili giyecek sistemi" uygulaması ile, tüm personel istekleri doğrultusunda giyecek aldığından, giyecek israfının, dolayısı ile kaynak israfının önüne geçilmiştir. Personel, kendisine verilen krediyi, o yıl içerisinde kullanmadığı zaman, kalan kredinin bir sonraki yıla devretmesi nedeniyle, kullanmayacağı giyecek siparişi vermemektedir. Bu nedenle, uygulamadaki kredili giyecek sisteminin yerine, başka bir sistem koymaya çalışmak gereksizdir. Bunun yerine, mevcut sistemin eksikliklerinin giderilmesi daha mantıklı bir seçim olarak öne çıkmaktadır.

Mevcut sistemin eksiklerinin ortaya çıkartılması için, öncelikle personelin sipariş verme örüntüsünün belirlenmesi gerekmektedir. Bu tez çalışmasında amaçlanan bu olmuştur.

4.1.3. Veri madenciliği amaçlarının belirlenmesi

Bu tez çalışmasında, Deniz Kuvvetlerinde görev yapan personelin kredili giyecek sistemi üzerinden vermiş olduğu sipariş verisi üzerinde veri madenciliği yapılarak, personelin sipariş verme örüntülerinin ortaya çıkartılması amaçlanmıştır. Ortaya çıkartılan bu örüntüler kullanılarak, Deniz Kuvvetleri Komutanlığı giyecek sisteminin eksiklikleri ortaya çıkartılabilir ve bu eksikliklerin giderilmesi sağlanabilir.

Sipariş verisi, veri madenciliği sınıflandırma yöntemleri ile analiz edilerek, personel tipi, kredi yılı, dönem, miktar gibi girdi verisi biliniyorken, sipariş edilen malzeme çıktı verisinin belirlenmesi örüntüsü bulunmaya çalışılmıştır. Bu örüntü, özet olarak, hangi durumlarda hangi malzemenin sipariş edildiği bilgisini verecektir.

Bu tez çalışmasında yapılan veri madenciliği, tahmin edici değil, tanımlayıcı veri madenciliğidir. Veri madenciliği sınıflandırma yöntemlerinde, en az iki veya daha fazla öznitelik kullanılır. Bu özniteliklerden sadece bir tanesi hedef öznitelik olarak belirlenebilir. Bu tez çalışmasında; sipariş edilen malzeme verisini içeren ALTGRUP_ALTGRUPKODU, siparişin yaz dağıtım dönemi için mi, kış dağıtım dönemi için mi olduğu verisini içeren DONEM, sipariş edildiği yıl verisini içeren KREDIYILI, sipariş miktarı verisini içeren MIKTAR, siparişi veren personelin rütbesi verisini içeren RUTBE ve siparişi verenin subay mı, astsubay mı olduğu verisini içeren PERSONEL_PERSTIP özniteliklerinin arasından, sipariş edilen malzeme verisini içeren ALTGRUP_ALTGRUPKODU özniteliği hedef öznitelik olarak seçilmiştir. Bu tezde ele alınan veri düşünüldüğünde, veri madenciliğinin, bir sonraki yılda hangi malzemelerden ne kadar sipariş edileceğini tahmin etmek için kullanılamayacağı sonucu ortaya çıkar. Bunun nedeni ise, hedef öznitelik olarak hem ALTGRUP_ALTGRUPKODU özniteliğinin, hem de MIKTAR özniteliğinin beraber hedef öznitelik olarak seçilememesidir.

Deniz Kuvvetleri giyecek siparişi verisinin tanımlanması en önemli amaçtır. Veri madenciliği sınıflandırma yöntemleri sonucunda oluşturulan kurallar, daha önceden bilinmeyen örüntüler bulunmasını sağlayacaktır. Bu örüntüler içinde yararlı bir çok kurallar olacaktır. Kurallar, ilgili kişiler tarafından yorumlanarak, bilgiye dönüştürülebilir. Örneğin;

- Yüzbaşı rütbesindeki bir personel, 2002-2005 yılları arasında, yazlık giyecek siparişi vermişse, sipariş ettiği malzeme kısa kollu yazlık elbisedir,

- Tüm subaylar ve astsubaylar pardesü malzemesi sipariş edeceklerse, sadece 1 tane sipariş etmektedirler,

- Personele Mont dağıtılmaya başlandıktan sonra, yağmurluğa olan talep çok aza inmiştir gibi bilgiler oluşturulabilir. Bu bilgiler ışığında, Deniz Kuvvetleri Komutanlığı giyecek sistemi üzerinde bir çok olumlu gelişimler yaratılabilir. Örneğin, sipariş edilen malzemelerin hammaddelerinin temini, üretiminin yapılması ve dağıtılması süreçlerinde iyileştirmeler yapılarak, en az 1 yıl olan sipariş verme ile malzemeyi teslim alma arasındaki süre azaltılabilir. Yeni bir malzeme dağıtılmadan önce, personelin bu malzemeye olan tepkisi tahmin edilebilir. Bazı malzemelerin dağıtılmasına son verilebilir.

4.1.4. Proje planının hazırlanması

Öncelikle mevcut kredili giyecek sistemi detaylı olarak incelenmiştir. Sistem, Oracle veri tabanını kullanmaktadır ve JSP (Java Server Pages) yazılım dili ile yazılmıştır. Personel sisteme, kendisine ait kullanıcı adı ve şifre ile girdikten sonra, sistem üzerinde kalan kredisini, içinde bulunulan yıl ve daha önceki yıllarda vermiş olduğu malzeme siparişlerini görebilmektedir.

Sisteme, yılın belirli zamanlarında, personele kalan kredisi doğrultusunda yeni sipariş girme yetkisi tanınmaktadır. Personel kendisine ait beden / drop bilgilerini istediği zaman güncelleme hakkına sahiptir. Sistemin Oracle veri tabanı üzerine yazmış olduğu veri genel olarak incelenmiştir.

Veri madenciliği uygulaması için, Yeni Zellanda'daki Waikato Üniversitesi bilgisayar bilimlerinde görev yapan öğretim üyelerince hazırlanan ve özellikle akademik çevrelerde çok yaygın olarak kullanılan, açık kaynak kodlu WEKA veri madenciliği yazılımı seçilmiştir. Bu seçimin yapılmasının en önemli nedenleri; akademik çevrelerde çok fazla kullanıldığından, güvenilirliğinin bugüne kadar defalarca test edilmiş olması, tamamen açık kaynak kodlu olduğundan herhangi bir telif hakkı sorunun olmaması ve içeriğinin çok zengin olmasıdır. WEKA uygulamasını hayata geçirmiş olan Witten ve Frank (2005), kitaplarının bir bölümünü bu yazılımın tanıtımına ayırmışlardır.

Oracle üzerinde bulunan verinin, Weka uygulamasında kullanılabilmesi için, ARFF (Attribute Relation File Format) biçimine çevrilmesi gerekmektedir. Bu amaçla, Ayhan Alkan, veri tabanlarından topladigi veriyi, WEKA yazılımının ihtiyaç duyduğu ARFF biçimine çeviren bir araç geliştirmiştir. Sayin Alkan, geliştirdiği bu araca, büyük Türk matematikçisi Cahit Arf'in anısına Cahit Arf ismini vermiştir.

4.2. Veriyi Anlamak

4.2.1. Başlangıç verisinin toplanması

Veriyi anlama basamağında veri genel olarak incelenir, veri üzerinde hiç bir işlem yapılmaz. Modelleme öncesi yapılması gereken, verinin hazırlanmasına yönelik işlemler bir sonraki veriyi hazırlama başlığı altında verilmiştir.

Veriyi anlamak için ilk yapılması gereken verinin toplanmasıdır. Veri madenciliği için verinin tek bir tabloda toplanması en uygunu olacaktır.

Kredili giyecek sistemi uygulamasına ait veri, kendi içerisinde büyük bir uygulama olduğu için, birden fazla tabloda tutulmaktadır. Bu tablolardan hangi verinin tek bir tablo altında toplanacağının tespiti, yapılması gereken ilk iştir. Kredili giyecek sistemi geliştiricileri, üretim planlamasını yapan ilgililer ve personel ile yapılan görüşmeler sonucunda, personelin hangi malzemeyi sipariş edebileceğini etkilediği düşünülen öznitelikler öncelikli olarak belirlenmiştir. Bu öznitelikler ve içinde bulunduğu tablolar, Tablo 4.1'de verilmiştir.

Tablo 4.1'de de görüldüğü üzere beş farklı tablodan toplam 11 öznitelik bir araya getirilmiştir. Bu 11 öznitelik ve değerleri tek bir tablo altında toplanmıştır.

Oluşturulan yeni tablodaki toplam kayıt sayısı 182.612'dir.

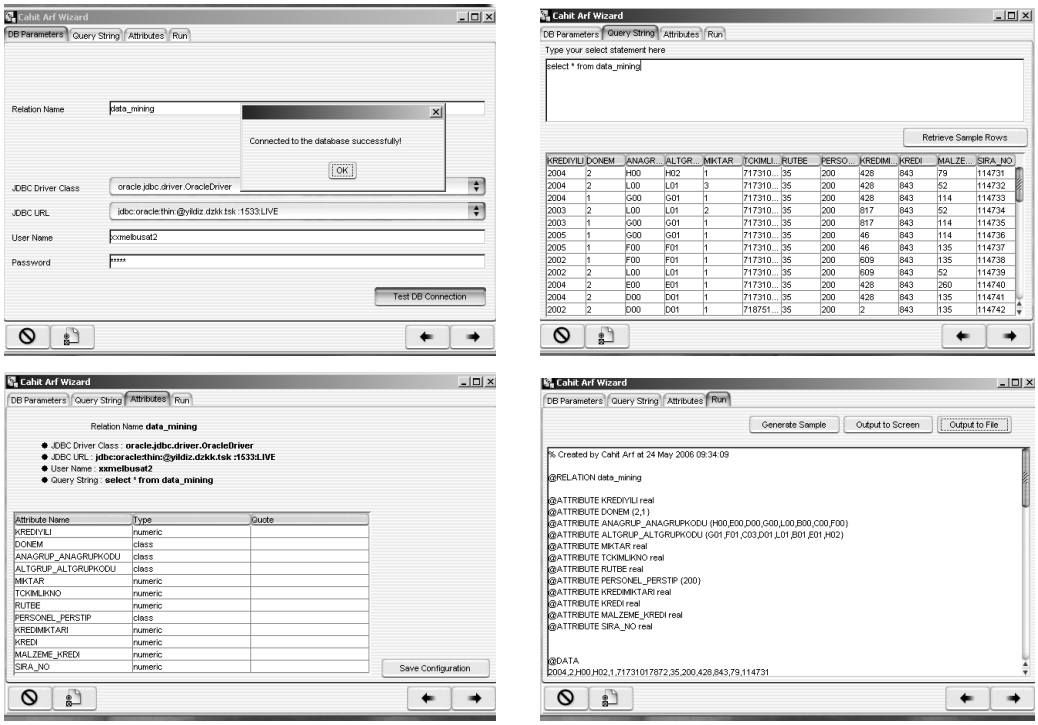
Tablo 4.1**:** Uygulamada Kullanılan Öznitelikler ve Bunların Bulunduğu Tablolar

Tablonun Adı	**Öznitelikler**
melbusat_istekleri	ANAGRUP_ANAGRUPKODU, ALTGRUP_ALTGRUPKODU, DONEM, KREDIYILI, MIKTAR
personel	TCKIMLIKNO, RUTBE, PERSONEL_PERSTIP
personel_kredi	KREDIMIKTARI
rütbe_kredi	RUTBE_KREDI
alt_grup_kodlari	MALZEME_KREDI

4.2.2. Verinin tanımlanması

Verinin tanımlanması noktasında, WEKA uygulaması zengin bir içerik sunmaktadır. Bu nedenle, Cahit Arf v1.0 uygulaması ile ARFF biçimine dönüştürülen malzeme sipariş verisi WEKA ortamında açılabilir. Cahit Arf v1.0 uygulaması ile Oracle veri tabanından verilerin alınması ve ARFF biçimine dönüştürülmesine ait 4 adım, Şekil 4.1'de verilmiştir.

Java ortamında çalışan WEKA uygulamasını çalıştırmak için, öncelikle java sanal makinenin (JVM-Java Virtual Machine) kurulması gerekmektedir. Bu nedenle, java sanal makine v1.5.0 kurulmuştur. Kayıt sayısının fazla olması nedeniyle, WEKA uygulamasını çalıştırırken, JVM üzerinde, ayrılan bellek miktarının artırılması gibi bazı değiştirgelerin ayarları yapılmıştır. WEKA aracı ile açılan Deniz Kuvvetleri verisi Şekil 4.2'de gösterilmiştir.

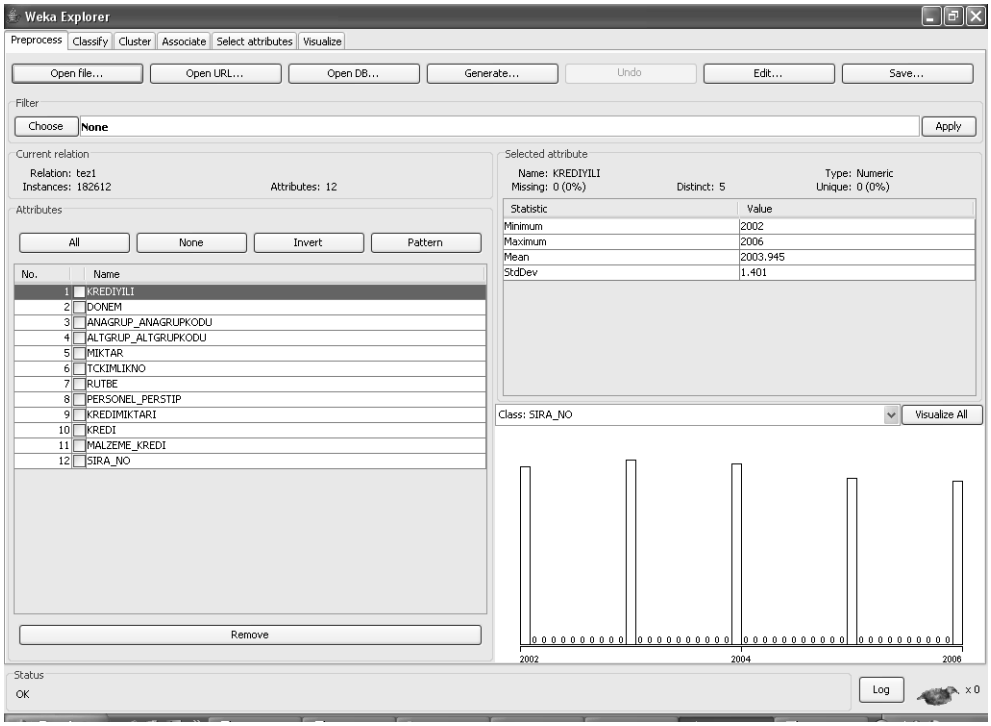


Şekil 4.1: Cahit Arf Uygulaması ile 4 Adımda Verinin Dönüştürülmesi

Tüm veriye ait genel dağılımlar, WEKA aracının sunduğu 2-boyutlu grafiklerle incelenmiştir. WEKA aracının tüm özniteliklerin değer dağılımını gösteren toplu grafik ekranı Şekil 4.3'de verilmiştir.

Şekil 4.3'de görüldüğü üzere, beş ayrı tablodaki veri, veri madenciliği için tek bir tablo altında toplanırken, 11 özniteliğin yanında bir de SIRA_NO özniteliği eklenmiştir. SIRA_NO özniteliğinin, veri madenciliğine herhangi bir katkısı olmadığından, tablo 4.2'de verilen söz konusu özniteliklerin açıklamaları arasında yer verilmemiştir.

Tablo 4.2'de, her bir özniteliğin, uygulaması yapılan giyecek verisi içindeki açıklamalarına yer verilmiştir. Bu öznitelikler, veri işleme adımında, tekrardan gözden geçirilecek ve aralarında veri madenciliği uygulamasına katkı sağlamayacağı değerlendirilenler, WEKA aracının sunduğu filtreler sayesinde uygulama dışında bırakılacaklardır. SIRA_NO özniteliği de, benzer şekilde, veri işleme aşamasında uygulama dışında bırakılacaktır.



Şekil 4.2**:** Giyecek Verisinin WEKA Uygulamasında Görünümü

Şekil 4.3**:** Giyecek Verisinin Genel Dağılımı

Tablo 4.2**:** Özniteliklerin Açıklamaları

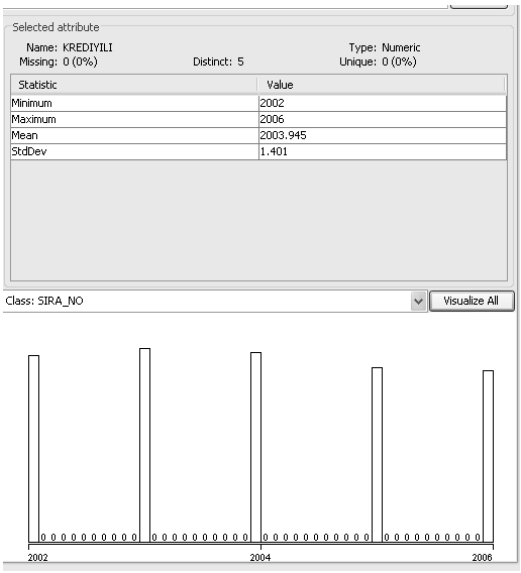
Tablonun Adı	**Açıklaması**
ANAGRUP_ANAGRUPKODU	Her giyecek çeşidi için bir alt grup vardır ve alt gruplar, bir ana grup altında yer alır.
ALTGRUP_ALTGRUPKODU	Her bir giyecek çeşidi için verilen koddur.
DONEM	Giyecekler, yaz ve kış dönemi olmak üzere 2 dönem halinde sipariş edilir ve teslim edilir.
KREDIYILI	Hangi yıl için sipariş girildiğini gösterir.
MIKTAR	Sipariş edilen malzemenin miktarını gösterir.
TCKIMLIKNO	Siparişi veren personelin kimliğini verir.
RUTBE	Siparişi veren personelin rütbesini gösterir.
PERSONEL_PERSTIP	Siparişi veren personelin subay ya da astsubay olduğunu gösterir.
KREDIMIKTARI	Siparişi veren personelin, ne kadar kredisi olduğunu gösterir.
KREDI	Her rütbe için farklı kredi verilir. Rütbe karşılığı verilen kredi miktarını gösterir.
MALZEME_KREDI	Her bir malzemenin kredi cinsinden fiyatını gösterir.

4.2.3. Verinin incelenmesi

Weka aracı kullanarak, tüm özniteliklere ait veri incelenmiştir. İnceleme yapılırken, her bir özniteliğe ait verinin minimum, maksimum, ortalama ve standart sapma istatistiksel bilgilerine ve verinin genel dağılımına bakılmıştır. Bu tez çalışmasında, çok fazla yer kaplamaması için, tüm özniteliklere ait dağılımlar ve istatistiksel bilgilerin verilmesi yerine, örnek olması açısından WEKA aracından alınan bilgilerden, sadece KREDI_YILI özniteliğine ait bilgiler Şekil 4.4'te verilmiştir.

Bu tez çalışmasında, veri madenciliğinin sınıflandırma yöntemleri karşılaştırılarak, giyecek sipariş verisi için en uygun olan yöntem tespit edilmeye çalışılmıştır. Sınıflandırma işlemi yapılırken, öncelikle hedef öznitelik belirlenir. Sonrasında, bu özniteliğin, diğer öznitelikler cinsinden formüle edilmesi sağlanır. Bu nedenle, hedef öznitelik olarak seçilen ve sipariş edilen giyecek malzemelerini tanımlayan

ALTGRUP_ALTGRUPKODU özniteliği değerlerinin ne anlama geldiğinin bilinmesinde yarar vardır (Tablo 4.3).



Şekil 4.4**:** KREDI_YILI Özniteliğine ait Veri İncelemesi

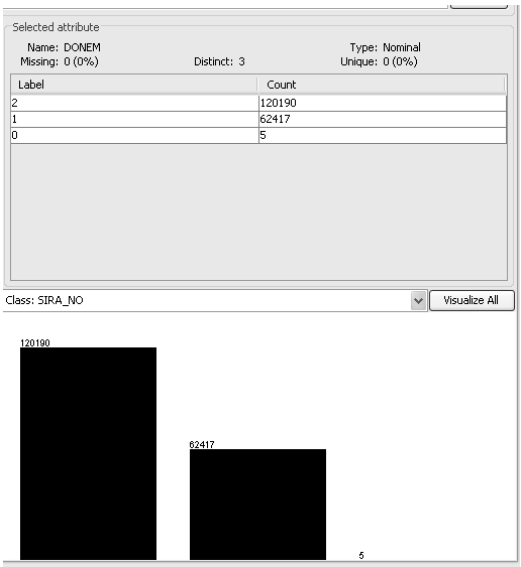
Tablo 4.3: ALTGRUP_ALTGRUPKODU Özniteliğinin Değerleri ve Anlamları

Öznitelik Değeri	**Anlamı**
B01	Pardesü
C01	Yağmurluk Tip 1
C03	Yağmurluk Tip 2
D01	Mont
E01	Elbise Kışlık Siyah
F01	Elbise Yazlık Uzun Kollu Beyaz
G01	Elbise Yazlık Kısa Kollu Beyaz
H01	Kazak Tip 1
H01	Kazak Tip 2

4.2.4. Veri kalitesinin doğrulanması

Kredili Giyecek Sistemi uygulaması, son kullanıcının hatalı veri girmesini engelleyecek şekilde tasarlanmıştır. Ama yine de bazı durumlarda, veri tabanında kirli veri olabilmektedir. Gerçek hayat verisinde de kirli veri ile uğraşmak zorunda olmak kaçınılmazdır.

Uygulama verisinin incelenmesi yapılırken, benzer bir veri kirliliği ile karşılaşılmıştır. Kredili giyecek sisteminde giyecekler, yazlık ve kışlık olmak üzere iki ayrı dönem için sipariş edilmektedir. Sipariş edilen malzemeler, yaklaşık bir yıl sonra, yaz dönemi ve kış dönemi öncesinde olmak üzere iki seferde dağıtılmaktadır. Şekil 4.5'te de görüleceği üzere, giyecek verisinin içerisinde, normalde sadece dönem-1 ve dönem-2 olması gerekirken, 5 kayıtlı bir dönem-0 bilgisi ile karşılaşılmıştır. Bu verinin kirli olduğu, böyle bir verinin olmaması gerektiği kararı konunun uzmanlarına danışılarak verilmiştir. Diğer veri içerisinde herhangi bir veri kirliliğine rastlanmamıştır.



Şekil 4.5**:** DONEM Özniteliğinde Kirli Veri Tespiti

4.3. Veriyi Hazırlamak

4.3.1. Verinin seçilmesi

Verinin seçilmesi için, öncelikle hangi özniteliklerin veri madenciliği için seçileceğine karar verilmesi gerekmektdir. Uygulamada kullanılan veriye bakıldığında, her bir ANAGRUP_ANAGRUPKODU özniteliğinin altında bir veya daha fazla ALTGRUP_ALTGRUPKODU olduğu tespit edilmiştir. Buradan, iki öznitelik arasında doğru orantılı bir ilişki olduğu anlaşılmaktadır. İster doğru, ister ters orantılı olsun, tüm bire bir ilişkiler, veri madenciliği çalışmasının doğru sonuç üretmesine engel olacaktır.

Larose (2005)'in de belirttiği gibi, veri madenciliği veya istatistiksel modelin içerisinde birbirleriyle bire bir ilişkili özniteliklere yer verilmemelidir. Çünkü, ilişkili öznitelikler kullanıldığında, elde edilen veri madenciliği sonucu en iyi ihtimalle abartılı olacaktır. Daha kötü ihtimaller arasında, tutarsız veya tamamen yanlış bir sonuç çıkartılması vardır. Yanlış sonuca göre işlem yapan işletme veya organizasyonlar, bu durumdan büyük zarar görebilirler.

Benzer şekilde, uygulama verisi içinde 10 farklı malzeme, yani 10 farklı ALTGRUP_ALTGRUPKODU vardır. Her bir malzemenin kredisi / fiyatı farklı olduğuna göre, ALTGRUP_ALTGRUPKODU ile MALZEME_KREDI arasında da bire bir ilişki vardır.

Aralarında bire bir ilişki olan diğer bir öznitelik çifti ise RUTBE ve KREDI öznitelikleri arasında vardır. RUTBE özniteliği, sipariş veren personelin rütbesi bilgisini tutarken, KREDI özniteliği, her bir rütbenin yıllık ne kadar kredisi olduğu bilgisini tutmaktadır.

TCKIMLIKNO ve SIRA_NO öznitelikleri, her bir kayıt için farklıdır. Veri madenciliğinin amacı, veri içinde anlamlı örüntüler bulmak olduğuna göre, bu özniteliklerin de çıkartılması gerekmektedir.

KREDIMIKTARI özniteliği, siparişi veren personelin, ne kadar kredisi olduğunu göstermektedir. Hedef öznitelik olarak sipariş edilen malzeme seçilmiştir. Yani, malzeme bakış açısı ile bakılmıştır. KREDIMIKTARI özniteliği ise sipariş veren kişi ile ilgili bir özniteliktir. Eğer, bu çalışmanın amacı, kim hangi malzemeyi sipariş ediyor bilgisine ulaşmak olsaydı, KREDIMIKTARI özniteliğinin, o durumda, veri madenciliği sonucuna olumlu bir etkisi olabilirdi. Ama, mevcut durum itibari ile, sipariş edilen malzemeler üzerine herhangi bir etkisi olmadığından, sonucun yanlış üretilmesine neden olabilir. Bu nedenle, KREDIMIKTARI özniteliğinin de çıkartılması gerekmektedir.

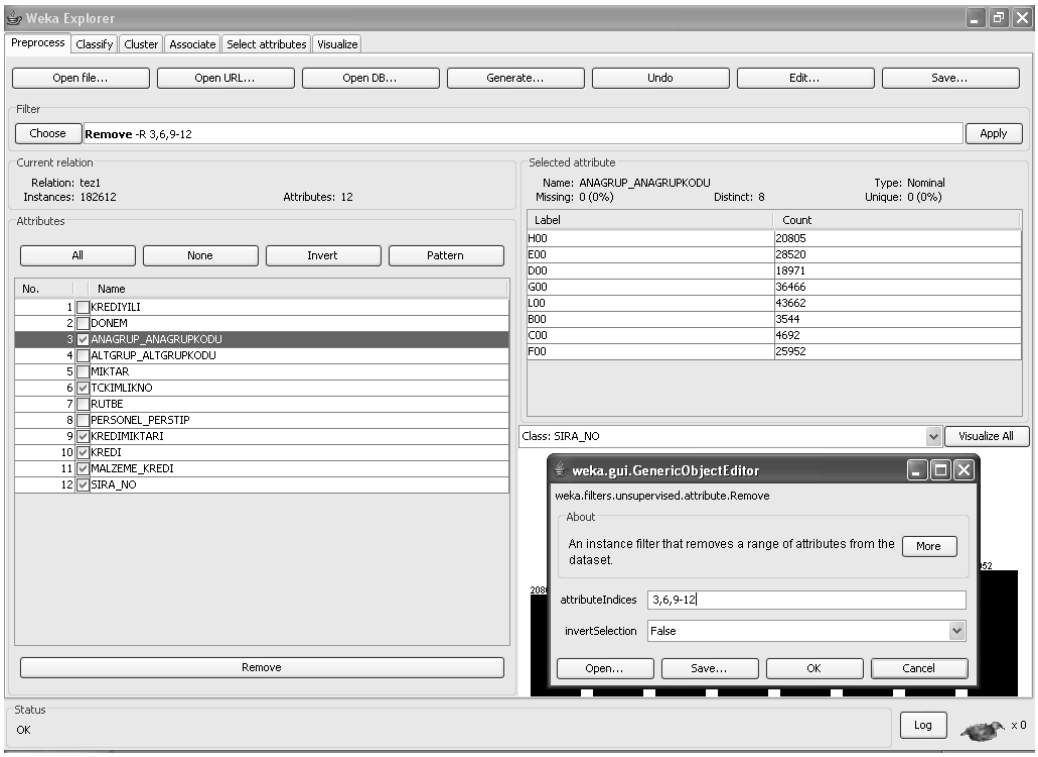
Sonuç olarak, bu tez çalışmasından sağlıklı ve tutarlı sonuçlar alabilmek için, ANAGRUP_ANAGRUPKODU, MALZEME_KREDI, KREDI, TCKIMLIKNO, SIRA_NO ve KREDIMIKTARI öznitelikleri, veri madenciliği için gerekli olan öznitelikler arasından çıkartılmalıdır.
WEKA aracının "Remove" filtresi kullanılarak ANAGRUP_ANAGRUPKODU, MALZEME_KREDI, KREDI, TCKIMLIKNO, SIRA_NO ve KREDIMIKTARI öznitelikleri, veri madenciliği için gerekli olan öznitelikler arasından çıkartılmıştır (Şekil 4.6).

4.3.2. Verinin temizlenmesi

Verinin temizlemesi aşamasında; hatalı, normal dışı, eksik, tutarsız ve tekrar eden verinin uygun şekilde düzeltilmesi yapılır.

Kredili Giyecek Sistemine ait veri, verinin incelenmesi ve veri kalitesinin doğrulanması aşamalarında analiz edilmişti. Kredili giyecek sistemi gibi yazılımlar, aynı zamanda veri bütünlüğünü ve güvenirliğini maksimum seviyede sağlayan yazılımlardır. Bu nedenle, veri temizlemesi gerektirecek herhangi bir normal dışı, eksik veya tekrar eden veriye rastlanılmamıştır. Fakat, veri kalitesinin doğrulanması aşamasında yapılan analizlerin sonuç bölümünde de belirtildiği üzere, yazılımın bir şekilde hataya düşmesi nedeniyle DONEM özniteliği için 5 adet kayıt, 0 değeri

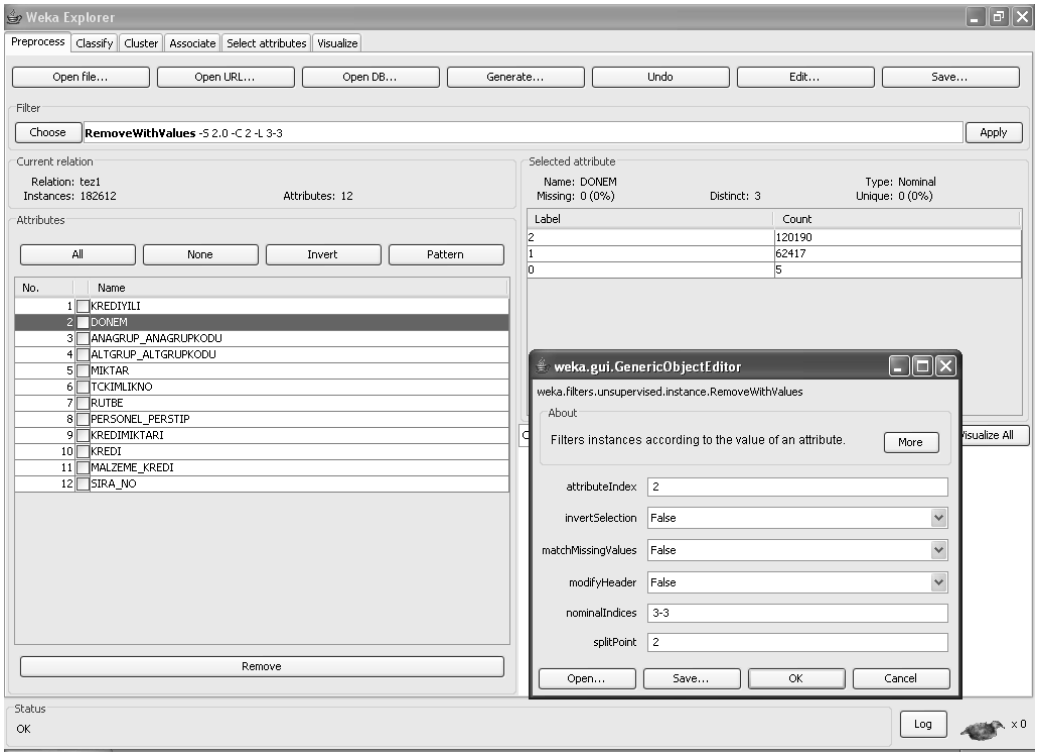
almıştır. Bu hatalı bir veridir, çünkü kredili giyecek sisteminde sadece 2 dönem vardır, dönem-1 ve dönem-2.



Şekil 4.6**:** Özniteliklerin Seçilmesi

Hatalı verinin düzeltilmesi için farklı yöntemler vardır. Eğer ki, hatanın nedeni bulunabiliyorsa, hatalı veri, olması gereken doğru veri ile değiştirilir. Hatanın nedeni bilinmiyorsa; hatalı veri, hatalı olmayan veri üzerine dağıtılır veya sayı olarak sonucu etkilemeyecek kadar az miktardaysa, bu değerler silinebilir. Kredili giyecek sistemi verisinin veri madenciliğinin yapıldığı bu uygulamada, 182.612 kayıttan sadece 5 tanesi, dönem-0 olduğu için, bu kayıtları silmek en doğru yaklaşım olacaktır. Sonuç olarak, veri madenciliği işlemleri, 182.607 kayıt üzerinde yapılacaktır.

WEKA aracı, daha önce de belirtildiği üzere, her türlü veri madenciliği ön işlemleri için çeşitli filtrelere sahiptir. Bir özniteliğin, belirli değerlere sahip kayıtlarının silinmesi için "RemoveWithValues" filtresi kullanılmaktadır. Şekil 4.7'da gösterildiği gibi, gerekli değiştirge ayarlamaları yapılarak, hatalı kayıtlar silinmiştir.



Şekil 4.7**:** Bir Özniteliğin, Belirli Değerlere Sahip Kayıtlarının Silinmesi

4.3.3. Verinin yapılandırılması

Verinin yapılandırılması aşamasında, eğer veri madenciliği süreci içerisinde, yeni bir öznitelik oluşturma ihtiyacı olursa, bu işlemler yapılır ve rapor haline getirilir. Bu tez çalışmasına konu olan uygulamada, böyle bir ihtiyaç görülmediği için, yeni bir öznitelik oluşturulmamıştır.

4.3.4. Verinin birleştirilmesi

Yeni bir öznitelik oluşturmaya gerek olmadığı gibi, mevcut verinin birleştirilmesine de ihtiyaç bulunmamaktadır.

Gerçek dünyada veri, kredili giyecek sisteminde olduğu gibi düzgün ve hatasız bir şekilde gelmeyebilir. Bu gibi durumlarda, hatalı verinin düzeltilmesi, yeni öznitelikler yaratılması, verinin birleşitirilmesi, olağan dışı olanların düzeltilmesi

veya sistemden atılması, eksik verinin düzeltilmesi veya sistemden atılması işlemleri gerekebilmektedir.

4.3.5. Verinin biçimlenmesi

Veri, genel anlamda üç çeşittir.

- Rakam (Numeric) - 25, 6.34, 45632,… gibi
- Sınıf (Nominal) - dönem1, dönem2,… gibi
- Karakter Dizisi (String) - Tc Kimlik No, Ad, Soyad,… gibi

Her veri madenciliği yöntemi, aşağıdaki veri tiplerinden her biriyle çalışmayabilir. Örneğin, yapay sinir ağları, işlem yapabilmek için tüm verinin rakam olmasını ister. Aksi taktirde işlem yapamaz. Son zamanlarda piyasaya sürülen veri madenciliği araçlarının çoğunluğu, sınıfsal veriyle de çalışmaktadır. Aslında, veri madenciliği aracının tek yaptığı, kendi içinde otomatik olarak, sınıfsal verinin rakamsal veri haline dönüştürülmesidir. WEKA aracı da bu dönüştürme işlemini kendi içinde yapmaktadır.

Bu uygulamada kullanılan öznitelikler ve bu özniteliklerin içindeki verinin tipleri Tablo 4.4'te verilmiştir. Bu veri tiplerinin değiştirilmesine ihtiyaç bulunmamaktadır.

Tablo 4.4**:** Özniteliklere Ait Veri Tipleri

Öz Nitelik Adı	**Veri Tipi**
ALTGRUP_ALTGRUPKODU	Sınıf
DONEM	Sınıf
KREDIYILI	Rakam
MIKTAR	Rakam
RUTBE	Rakam
PERSONEL_PERSTIP	Sınıf

4.4. Modelleme

4.4.1. Modelleme tekniğinin seçilmesi

Modelleme tekniğinin seçilmesi adımı, veri madenciliği sürecinin, belki de en önemli aşamasıdır. Elde edilen sonuçlar, seçilen veri madenciliği yöntemine göre çok büyük oranlarda değişebilmektedir. Aslında, herhangi bir yöntemin mutlak üstünlüğünden bahsedilemez. Her yöntem, farklı veri üzerinde en iyi yöntem olabilir. Buna rağmen, veri madenciliği arenasında, verdiği doğru sonuçlar nedeniyle görece tercih edilen yöntemler vardır.

Bu tez çalışmasında, hangi malzemelerin istendiği sorusuna cevap bulunmak istendiğinden, yapılacak iş, bir sınıflandırma işlemidir. Sınıflandırma konusunda, en çok tercih edilen veri madenciliği yöntemleri Naive Bayes, karar ağaçları ve yapay sinir ağları olduğu için, modelleme yöntemi seçilirken, bu üç yöntem karşılaştırılarak, Kredili Giyecek Sistemi verisi için en uygun yöntem bulunmaya çalışılacaktır.

4.4.2. Test tasarımı

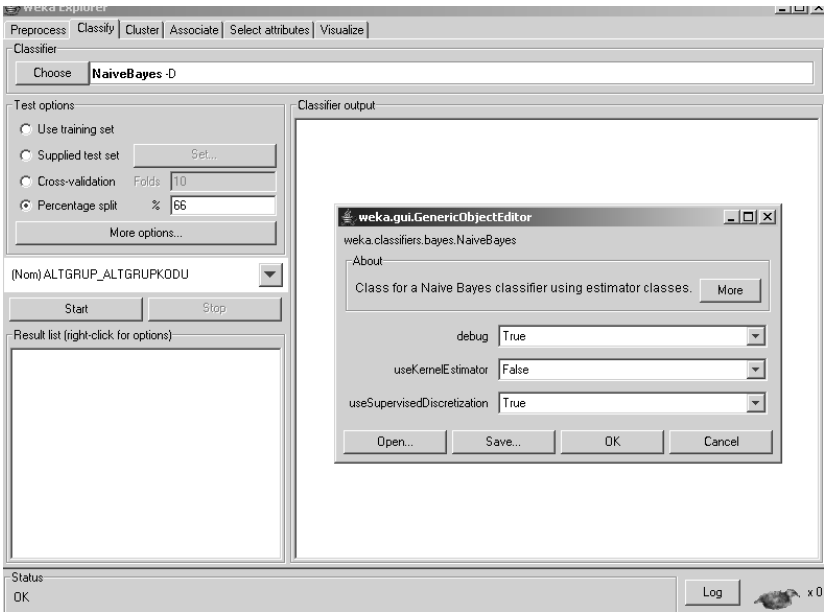
WEKA veri madenciliği aracında veri madenciliği yapabilmek için öncelikle alıştırma ve test verisinin belirlenmesi gerekmektedir. Bu tez çalışmasında kullanılan veri miktarı büyük olduğu için, Weiss ve Kulikowski (1991)'in önerdiği şekilde, tüm verinin 2/3'ü modeli eğitim, ya da diğer bir deyişle alıştırma için, geri kalan 1/3'lük kısım ise modeli test etmek için ayrılmıştır.

4.4.3. Modelin oluşturulması

4.4.3.1. Naive Bayes

WEKA veri madenciliği aracında Naive Bayes yöntemini uygulamak için öncelikle sınıflandırma (classify) sekmesi seçilmiştir. Gelen ekran Şekil 4.8'da gösterilmiştir.

Şekil 4.8‘de görüldüğü üzere; ilk önce, sınıflandırıcı yöntemlerden Naive Bayes seçilmiştir. Daha sonra, Naive Bayes yazısı üzerine tıklanıldığı zaman gelen değiştirge ayarları girilmiştir. Naive Bayes yöntemi, son derece yalın bir yöntem olduğu için az sayıda değiştirgesi vardır. Değiştirgelerden, sınıflandırıcının ilave sonuçları da ekrana dökmesi ve rakamsal verileri otomatik olarak sınıfsal veri haline getirmesi işaretlenmiştir. Test seçenekleri bölümünde, tüm verinin %66’sının eğitim, geri kalan %34’lük kısımın test amaçlı olarak kullanılacağı belirtilmiştir. Hedef öznitelik olarak ALTGRUP_ALTGRUPKODU özniteliği seçildikten sonra yöntem çalıştırılmışır. Toplam 182.612 kayıta sahip ve 5 özniteliği bulunan Kredili Giyecek Sistemi verisi için WEKA aracının Naive Bayes yöntemi ile ürettiği sonuçlar Şekil ’da verilmiştir.



Şekil 4.8: WEKA Naive Bayes Değiştirge Ayarları

```
=== Run information ===
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     tez1-weka.filters.unsupervised.attribute.Remove-R3,6,10-12-
weka.filters.unsupervised.instance.RemoveWithValues-S2.0-C2-L3-3-H-
weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Remove-R6

Instances:    182607

Attributes:   5
              KREDIYILI, DONEM,  ALTGRUP_ALTGRUPKODU,  MIKTAR, RUTBE
Test mode:    split 66% train, remainder test

=== Classifier model (full training set) ===
Naive Bayes Classifier


Class G01: Prior probability = 0.24
KREDIYILI:  Normal Distribution. Mean = 2004.0142 StandardDev = 1.5722 WeightSum = 43326 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  17088 26240  (Total = 43328)
MIKTAR:  Normal Distribution. Mean = 1.4997 StandardDev = 0.6441 WeightSum = 43326 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 24.2569 StandardDev = 10.5274 WeightSum = 43326 Precision =
2.4166666666666665

Class F01: Prior probability = 0.08
KREDIYILI:  Normal Distribution. Mean = 2003.8919 StandardDev = 1.0385 WeightSum = 14409 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  1 14410  (Total = 14411)
MIKTAR:  Normal Distribution. Mean = 1.3813 StandardDev = 0.4972 WeightSum = 14409 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 18.9377 StandardDev = 7.0946 WeightSum = 14409 Precision =
2.4166666666666665

Class H01: Prior probability = 0.04
KREDIYILI:  Normal Distribution. Mean = 2004.6131 StandardDev = 0.6458 WeightSum = 6651 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  6120 533  (Total = 6653)
MIKTAR:  Normal Distribution. Mean = 1.9057 StandardDev = 0.8617 WeightSum = 6651 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 16.3869 StandardDev = 2.4136 WeightSum = 6651 Precision =
2.4166666666666665

Class C01: Prior probability = 0.11
KREDIYILI:  Normal Distribution. Mean = 2004.1706 StandardDev = 1.5535 WeightSum = 20129 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  10529 9602  (Total = 20131)
MIKTAR:  Normal Distribution. Mean = 1.9868 StandardDev = 0.8855 WeightSum = 20129 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 28.9209 StandardDev = 10.2955 WeightSum = 20129 Precision =
2.4166666666666665

Class L01: Prior probability = 0.03
KREDIYILI:  Normal Distribution. Mean = 2003.8415 StandardDev = 1.4652 WeightSum = 6371 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  6124 249  (Total = 6373)
MIKTAR:  Normal Distribution. Mean = 2.9913 StandardDev = 1.647 WeightSum = 6371 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 26.1228 StandardDev = 10.6655 WeightSum = 6371 Precision =
2.4166666666666665
```

Şekil 4.9: WEKA Naive Bayes Yöntemi Sonuç Raporu

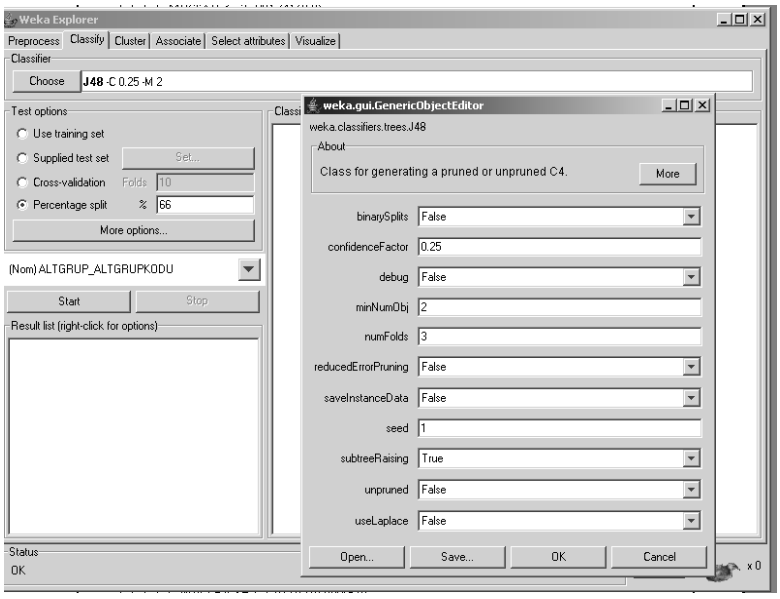
```
Class C03: Prior probability = 0.13
KREDIYILI:  Normal Distribution. Mean = 2003.2355 StandardDev = 0.9696 WeightSum = 23566 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  13978 9590  (Total = 23568)
MIKTAR:  Normal Distribution. Mean = 1.5011 StandardDev = 0.6828 WeightSum = 23566 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 34.4736 StandardDev = 8.8666 WeightSum = 23566 Precision =
2.4166666666666667

Class D01: Prior probability = 0.04
KREDIYILI:  Normal Distribution. Mean = 2003.0981 StandardDev = 1.1849 WeightSum = 6954 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  6479 477  (Total = 6956)
MIKTAR:  Normal Distribution. Mean = 2.3281 StandardDev = 0.9521 WeightSum = 6954 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 32.4603 StandardDev = 8.0735 WeightSum = 6954 Precision =
2.4166666666666665

Class B01: Prior probability = 0.08
KREDIYILI:  Normal Distribution. Mean = 2003.6387 StandardDev = 1.7615 WeightSum = 14575 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  14576 1  (Total = 14577)
MIKTAR:  Normal Distribution. Mean = 1.6045 StandardDev = 0.7509 WeightSum = 14575 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 28.3091 StandardDev = 11.3118 WeightSum = 14575 Precision =
2.4166666666666665

Class E01: Prior probability = 0.17
KREDIYILI:  Normal Distribution. Mean = 2004.2695 StandardDev = 1.19 WeightSum = 31546 Precision = 1.0
DONEM:  Discrete Estimator. Counts =  30224 1324  (Total = 31548)
MIKTAR:  Normal Distribution. Mean = 1.2212 StandardDev = 0.2595 WeightSum = 31546 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 24.3753 StandardDev = 10.0107 WeightSum = 31546 Precision =
2.4166666666666665

Class H02: Prior probability = 0.08
KREDIYILI:  Normal Distribution. Mean = 2004.3609 StandardDev = 1.162 WeightSum = 15080 Precision =
1.0
DONEM:  Discrete Estimator. Counts =  15081 1  (Total = 15082)
MIKTAR:  Normal Distribution. Mean = 1.1781 StandardDev = 0.1944 WeightSum = 15080 Precision =
1.1666666666666667
RUTBE:  Normal Distribution. Mean = 37.9968 StandardDev = 2.3196 WeightSum = 15080 Precision =
2.4166666666666665

=== Summary ===
Correctly Classified Instances       40081               64.5541 %
Incorrectly Classified Instances     22008               35.4459 %
Mean absolute error                  0.118
Root mean squared error              0.241
=== Confusion Matrix ===
    a    b    c    d    e    f    g    h    i    j   <-- classified as
10235  883   69  189    1 1521  258  616  612  237 |    a = G01
  633 3777    0  456    0    0    0    0    0    0 |    b = F01
  198    0 1397    0    0    0    0    0  697    0 |    c = H01
 2315    0  345 1987  236 1191  251  608   30    0 |    d = C01
   54    0   29   25 1004   55  373  181  254  184 |    e = L01
   85    0    0  261    0 5938  675    5 1028   39 |    f = C03
    0    0   27   60  195  210  873  515  345  164 |    g = D01
  571    0  891    0    0    0    0 2617  809   61 |    h = B01
  315    0   66   47    0  167  171    2 7549 2382 |    i = E01
    0    0    0    0    0   82  201   24  109 4704 |    j = H02
```

Şekil 4.9 (Devam): WEKA Naive Bayes Yöntemi Sonuç Raporu

4.4.3.2. Karar ağaçları

WEKA veri madenciliği aracında karar ağaçları yöntemi J48 adı ile geçmektedir. J48, C4.5 algoritmasının WEKA'ya uyarlanmış halidir. C.4.5 algoritması Quinlan (1993) tarafından oluşturulmuştur. WEKA veri madenciliği aracında karar ağaçları yöntemini uygulamak için; öncelikle sınıflandırma (classify) sekmesi seçilmiştir. Gelen ekran Şekil 4.10'da gösterilmiştir. Şekil 4.10'da görüldüğü üzere, sınıflandırıcı yöntemlerden J48 seçilmiştir. J48 üzerine tıklandığı zaman gelen değiştirge ayarlamaları sayfasında; güven değeri olarak 0.25 değeri girilmiştir. Güven değeri ne kadar küçük olursa, oluşacak ağaç şeklinin dallanması da o oranda küçük olacaktır. Dallanmaların minimum sayısının 2 olması ve alt dallanmaların olması değiştirge değerleri girilmiştir.

Test seçenekleri bölümünde, tüm verinin %66'sının eğitim, geri kalan %34'lük kısımın test amaçlı olarak kullanılacağı belirtilmiştir. Hedef öznitelik olarak ALTGRUP_ALTGRUPKODU özniteliği seçildikten sonra yöntem çalıştırılmışır. Toplam 182.612 kayıta sahip ve 5 özniteliği bulunan Kredili Giyecek Sistemi verisi için WEKA aracının J48 karar ağaçları yöntemi ile ürettiği sonuçlar Şekil'de verilmiştir.



Şekil 4.10: WEKA Karar Ağacı Değiştirge Ayarları

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:                                   tez1-weka.filters.unsupervised.attribute.Remove-R3,6,10-12-
weka.filters.unsupervised.instance.RemoveWithValues-S2.0-C2-L3-3-H-
weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Remove-R6
Instances:    182607
Attributes:   5
              KREDIYILI, DONEM,  ALTGRUP_ALTGRUPKODU,  MIKTAR, RUTBE
Test mode:    split 66% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

DONEM = 2
|  MIKTAR <= 1
|  |  KREDIYILI <= 2002
|  |  |  RUTBE <= 35: G01 (10248.0/1909.0)
|  |  |  RUTBE > 35: B01 (7631.0)
|  |  KREDIYILI > 2002
|  |  |  RUTBE <= 35
|  |  |  |  KREDIYILI <= 2004
|  |  |  |  |  RUTBE <= 20: E01 (14837.0)
|  |  |  |  |  RUTBE > 20
|  |  |  |  |  |  KREDIYILI <= 2003: C03 (2516.0/511.0)
|  |  |  |  |  |  KREDIYILI > 2003: G01 (2463.0/1834.0)
|  |  |  |  KREDIYILI > 2004
|  |  |  |  |  RUTBE <= 20
|  |  |  |  |  |  RUTBE <= 14: C03 (3932.0/1186.0)
|  |  |  |  |  |  RUTBE > 14
|  |  |  |  |  |  |  KREDIYILI <= 2005
|  |  |  |  |  |  |  |  RUTBE <= 18: H01 (2590.0/864.0)
|  |  |  |  |  |  |  |  RUTBE > 18: B01 (2164.0/1082.0)
|  |  |  |  |  |  |  KREDIYILI > 2005
|  |  |  |  |  |  |  |  RUTBE <= 18: E01 (2572.0/1204.0)
|  |  |  |  |  |  |  |  RUTBE > 18: B01 (2122.0/1044.0)
|  |  |  |  |  RUTBE > 20: E01 (4341.0)
|  |  |  RUTBE > 35
|  |  |  |  KREDIYILI <= 2003
|  |  |  |  |  RUTBE <= 37: H02 (3287.0)
|  |  |  |  |  RUTBE > 37: C03 (4251.0/697.0)
|  |  |  |  KREDIYILI > 2003
|  |  |  |  |  RUTBE <= 37
|  |  |  |  |  |  KREDIYILI <= 2004: H02 (3195.0)
|  |  |  |  |  |  KREDIYILI > 2004: E01 (5365.0)
|  |  |  |  |  RUTBE > 37
|  |  |  |  |  |  KREDIYILI <= 2004: E01 (4168.0/2512.0)
|  |  |  |  |  |  KREDIYILI > 2004: H02 (6930.0)
|  MIKTAR > 1
|  |  KREDIYILI <= 2004
|  |  |  RUTBE <= 20
|  |  |  |  MIKTAR <= 3
|  |  |  |  |  KREDIYILI <= 2003: G01 (6116.0)
|  |  |  |  |  KREDIYILI > 2003: H01 (3171.0)
```

Şekil 4.11: WEKA Karar Ağacı Yöntemi Sonuç Raporu

```
|  |  |  |  MIKTAR > 3
|  |  |  |  |  KREDIYILI <= 2003: L01 (1066.0)
|  |  |  |  |  KREDIYILI > 2003
|  |  |  |  |  |  MIKTAR <= 4: D01 (572.0)
|  |  |  |  |  |  MIKTAR > 4: L01 (224.0)
|  |  |  RUTBE > 20
|  |  |  |  KREDIYILI <= 2002
|  |  |  |  |  MIKTAR <= 2: D01 (2259.0/878.0)
|  |  |  |  |  MIKTAR > 2
|  |  |  |  |  |  MIKTAR <= 4: D01 (926.0)
|  |  |  |  |  |  MIKTAR > 4: L01 (48.0)
|  |  |  |  KREDIYILI > 2002
|  |  |  |  |  RUTBE <= 35
|  |  |  |  |  |  MIKTAR <= 2
|  |  |  |  |  |  |  KREDIYILI <= 2003: D01 (747.0/285.0)
|  |  |  |  |  |  |  KREDIYILI > 2003: C01 (798.0/319.0)
|  |  |  |  |  |  MIKTAR > 2
|  |  |  |  |  |  |  MIKTAR <= 4: D01 (782.0)
|  |  |  |  |  |  |  MIKTAR > 4: L01 (105.0)
|  |  |  |  |  RUTBE > 35
|  |  |  |  |  |  MIKTAR <= 4: C03 (5123.0)
|  |  |  |  |  |  MIKTAR > 4: L01 (98.0)
|  |  KREDIYILI > 2004
|  |  |  RUTBE <= 18
|  |  |  |  MIKTAR <= 3: B01 (4160.0)
|  |  |  |  MIKTAR > 3: L01 (988.0)
|  |  |  RUTBE > 18
|  |  |  |  MIKTAR <= 4: C01 (9947.0)
|  |  |  |  MIKTAR > 4: L01 (448.0)
DONEM = 1
|  KREDIYILI <= 2004
|  |  RUTBE <= 20
|  |  |  MIKTAR <= 1
|  |  |  |  KREDIYILI <= 2002: C01 (5896.0)
|  |  |  |  KREDIYILI > 2002: F01 (10934.0)
|  |  |  MIKTAR > 1
|  |  |  |  KREDIYILI <= 2002
|  |  |  |  |  MIKTAR <= 2: E01 (898.0)
|  |  |  |  |  MIKTAR > 2
|  |  |  |  |  |  MIKTAR <= 3: E01 (54.0)
|  |  |  |  |  |  MIKTAR > 3: L01 (6.0)
|  |  |  |  KREDIYILI > 2002
|  |  |  |  |  MIKTAR <= 3: G01 (2218.0)
|  |  |  |  |  MIKTAR > 3
|  |  |  |  |  |  MIKTAR <= 4: G01 (22.0)
|  |  |  |  |  |  MIKTAR > 4: L01 (4.0)
|  |  RUTBE > 20
|  |  |  KREDIYILI <= 2003
|  |  |  |  MIKTAR <= 1
|  |  |  |  |  RUTBE <= 35
|  |  |  |  |  |  KREDIYILI <= 2002: G01 (1338.0/382.0)
|  |  |  |  |  |  KREDIYILI > 2002: C03 (1530.0/305.0)
|  |  |  |  |  RUTBE > 35: C03 (9102.0/978.0)
|  |  |  |  MIKTAR > 1
|  |  |  |  |  KREDIYILI <= 2002: D01 (314.0/136.0)
|  |  |  |  |  KREDIYILI > 2002
```

Şekil 4.11 (Devam): WEKA Karar Ağacı Yöntemi Sonuç Raporu

```
|  |  |  |  |  MIKTAR <= 2: D01 (496.0/208.0)
|  |  |  |  |  MIKTAR > 2
|  |  |  |  |  |  MIKTAR <= 3: E01 (26.0/10.0)
|  |  |  |  |  |  MIKTAR > 3: L01 (3.0)
|  |  |  KREDIYILI > 2003
|  |  |  |  RUTBE <= 35: G01 (1492.0/275.0)
|  |  |  |  RUTBE > 35
|  |  |  |  |  MIKTAR <= 1: C01 (4320.0/889.0)
|  |  |  |  |  MIKTAR > 1: G01 (382.0/162.0)
|  KREDIYILI > 2004
|  |  MIKTAR <= 1: G01 (19427.0/1440.0)
|  |  MIKTAR > 1
|  |  |  RUTBE <= 17
|  |  |  |  MIKTAR <= 2: G01 (1324.0)
|  |  |  |  MIKTAR > 2: L01 (204.0)
|  |  |  RUTBE > 17
|  |  |  |  MIKTAR <= 3: F01 (2398.0)
|  |  |  |  MIKTAR > 3: L01 (29.0)
Number of Leaves  :         58
Size of the tree :  115

=== Summary ===

Correctly Classified Instances       55555               89.4793 %
Incorrectly Classified Instances      6532               10.5207 %
Mean absolute error                   0.0313
Root mean squared error               0.1254

=== Confusion Matrix ===
    a     b     c     d     e     f     g     h     i     j   <-- classified as
 13255     0    59   295     0   571    43   110   391     0 |     a = G01
   334  4484     0     1     0     0     7     0     0     0 |     b = F01
   174     0  1648     0     0    78     0   230    79     0 |     c = H01
    89     0     8  6725     0     7     0     3    17     0 |     d = C01
   230     0    28    83  1083    96   327    42   278     0 |     e = L01
   171     0     0     1     0  7852     0    81    43     0 |     f = C03
   361     0    39     7     0   185  1548    35   237     0 |     g = D01
    18     0    51     0     0     5     0  4727   119     0 |     h = B01
   399     0   107     3     0   219    75   207  9667     0 |     i = E01
   299     0     0    14     0   102    30     0   144  4566 |     j = H02
```

Şekil 4.11 (Devam): WEKA Karar Ağacı Yöntemi Sonuç Raporu

4.4.3.3. Yapay sinir ağları

WEKA veri madenciliği aracında, ileri beslemeli geri yayılmalı yapay sinir ağı yöntemi olan çok katmanlı algılayıcı (MultilayerPerceptron) kullanılmaktadır. WEKA veri madenciliği aracında yapay sinir ağları yöntemini uygulamak için; öncelikle sınıflandırma (classify) sekmesi seçilmiştir. Gelen ekran Şekil 4.13'da gösterilmiştir. Şekil 4.13'de görüldüğü üzere; sınıflandırıcı yöntemlerden çok katmanlı algılayıcı seçilmiştir. Çok katmanlı algılayıcılar ile tek katmanlı

algılayıcılar arasındaki farkı anlamak için XOR işlemi uygulanması yeterlidir. Çünkü, tek katmanlı algılayıcılarda AND, OR ve NOT işlemleri yapılabilirken, XOR işlemi yapılamaz. Çok katmanlı algılayıcılar ise XOR işlemini abrayabilmektedir.

Çok katmanlı algılayıcının nasıl öğrendiği en önemli konulardan bir tanesidir. Bu sorunun çözümü için iki yöntemi vardır. Bunlar; ağ yapısının öğrenilmesi ve bağlantı ağırlıklarının öğrenilmesidir. Bilinen bir ağ yapısı için ağırlıkları belirleyebilen görece basit bir algoritma vardır, bu algoritma geri yayılma (backpropogation) algoritmasıdır. Ağ yapısını öğrenmeye çalışan çeşitli algoritma denemeleri mevcut olsa da, bu sorun genellikle alıştırma yöntemi ile çözülmektedir. Tahmin oranını maksimum düzeyde tutacak sayıda birimlerden oluşan sadece bir adet gizli katman çoğunluk problemin çözümü için yeterlidir (Witten ve Frank, 2005). Algılayıcının öğrenme kuralı Şekil 4.12'de gösterilmiştir.

- Tüm ağırlıkları sıfıra eşitle.
- Öğrenme için ayrılan verideki tüm kayıtlar doğru olarak sınıflandırılıncaya kadar işleme devam et
  - Öğrenme için ayrılan verideki her bir I kayıtı için tekrarla.
    - Eğer I kayıtı, algılayıcı tarafından yanlış sınıflandırılmış ise;
      - Eğer I kayıtı ilk sınıfa aitse, I kayıtının değerini ağırlık vektörüne ekle.
      - Diğer durumlarda, I kayıtının değerini ağırlık vektöründen çıkart.

Şekil 4.12: Algılayıcı Öğrenme Kuralı (Witten ve Frank, 2005).

Geri yayılma, her bir birimin son tahmin değerine olan katkısı oranında gizli birimleri besleyen bağlantıların ağırlıklarının değiştirilmesidir. Bunu yapmak için kademeli düşme (gradient descent) standart matematik algoritması kullanılır. Kademeli düşme algoritması, fonksiyonun türevini kullandığından, basamak fonksiyonunun, türevi alınabilir bir fonksiyon olmasına ihtiyaç duyar. Bu nedenle, basamak fonksiyonu olarak genellikle sigmoid fonksiyon kullanılır.

Kademeli düşme algoritması, minimize edilecek olan hata fonksiyonun türevinden elde edilen bilgiyi kullanır. Algoritma, bu bilgiyi, bir fonksiyonun değiştirgelerinin

ayarlanmasında kullanan bir yinelemeli eniyileme (optimizasyon) yordamıdır. Hata fonksiyonunun türevini alır, öğrenme oranı (learning rate) denilen küçük bir sabit değer ile çarpar ve sonucu değiştirgenin o andaki değerinden çıkartır. Bu işlem, değiştirgenin yeni değerleri için de bir minimum değere ulaşılıncaya kadar teker teker hesaplanarak yinelenir. Hata fonksiyonu birden fazla minimum noktaya sahip olabilir, bu durumda kademeli düşme algoritması, en düşük minimum noktayı bulamayabilir. Çok katmanlı algılayıcılar, bu nedenle, destek vektör makinaları gibi yeni bazı yöntemlerle kıyaslandığında daha az beceriye sahiptirler.

Bu tez çalışmasındaki uygulamada, çok katmanlı algılayıcı üzerine tıklandığı zaman gelen değiştirge ayarlamaları sayfasındaki değiştirgeler ve girilen değerler aşağıda sunulmuştur;

- Grafiksel Kullanıcı Arayüzü (GUI) : Ağ, gizli katmanları ile birlikte grafiksel kullanıcı arayüzünde gösterilir. Bu tez uygulamasında, daha detaylı rapor vermesi nedeniyle grafiksel kullanıcı arayüzü yerine normal arayüz kullanılmıştır.

- Otomatik Katman Oluşturma (autoBuild) : İhtiyaç durumunda gizli katman eklenmesi gerektiğinde, sistem ağa otomatik olarak gizli katman veya katmanlar ekler. Uygulamada bu değer pozitif olarak işaretlenmiştir.

- Doğrulama (debug) : Bu değer işaretliğinde, çıktı olarak ek bilgiler verebilir. Sistem performansını kötü anlamda etkileyeceğinden negatif olarak işaretlenmişir.

- Kademeli Azaltma (decay) : Öğrenme oranının düşmesine neden olur. Yeni öğrenme oranını belirlemek için, başlangıçtaki öğrenme oranını döngü sayısına böler. Sistem performansını kötü anlamda etkileyeceğinden negatif olarak işaretlenmişir.

- Gizli Katmanlar (hiddenLayers) : Yapay sinir ağının gizli katmanlarını tanımlar. İstenilen gizli katman sayısı yazılabileceği gibi, sistem tarafından özel olarak oluşturulmuş kodlar da kullanılabilir. Bu kodlardan ‘a’ kodu (öznitelik sayısı + sınıf

sayısı) / 2, ‘i’ kodu öznitelik sayısı ,’o’ kodu sınıf sayısı ve’t’ kodu öznitelik sayısı + sınıf sayısını belirtir. Uygulamada, ‘a’ kodu işaretlenmiştir.

- Öğrenme Oranı (learningRate) : Güncellenen ağırlıkların büyüklüğüdür. Öğrenme oranı , basamakların büyüklüğünü belirler. Öğrenme oranı çok büyük ve hata fonksiyonun birden fazla minimum noktası olduğunda, minimumlardan bir veya birkaçının bulunamaması olasılığı oluşabilir. Öğrenme oranın çok küçük olması durumunda ise, minimum nokta bulma işlemi çok fazla zaman alabilir. Uygulama için ‘0,3’ değeri girilmiştir.

- İvmelenme (momentum) : Yeni ağırlık değişimine, bir önceki iterasyondan gelen değişim miktarının küçük bir oranı eklenerek performans artışı sağlanabilir. Uygulamada, binde iki, yani ‘0,2’ değeri girilmiştir.

- Sınıfsaldan İkiliye Filtre(NominalToBinaryFilter) : Veri içerisinde sınıfsal değerlerin bulunması durumunda, bu değerler, bu filtre sayesinde işlemler öncesinde ikiliye çevrilerek performans artışı sağlanabilir. Uygulama verisi içinde sınıfsal veri bulunduğundan pozitif olarak işaretlenmiştir.

- Özniteliklerin Normalizasyonu (normalizeAttributes) : Öznitelikleri normalize eder. Hem sınıfsal, hem de sayısal değerleri normalize ederek performans artışı sağlayabilir. Uygulama verisi içinde hem sınıfsal, hem de sayısal veri bulunduğundan pozitif olarak işaretlenmiştir.

- Sayısal Sınıfların Normalizasyonu (normalizeNumericClass) : Sınıfın değerleri sayısal ise, bunları normalize eder. Sınıf değerlerini –1 ile +1 arasında olacak şekilde normalize ederek performans artışı sağlayabilir. Uygulama verisi içinde sayısal sınıflar bulunduğundan pozitif olarak işaretlenmiştir.

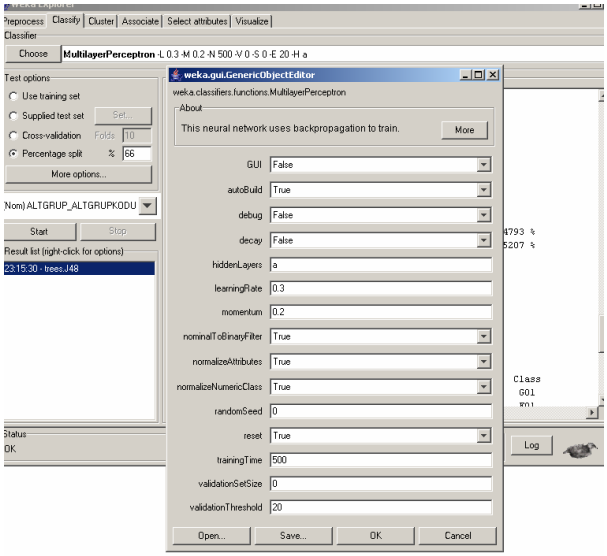
- Rastgele Kök (randomSeed) : Kök, rastgele sayı üreticisinin başlatılması için kullanılır. Rastgele sayılar, nodlar arası bağlantıların ilk ağırlık değerlerinin belirlenmesinde kullanılır. ‘0’ değeri girilmiştir.

- Yinele (reset) : Ağın, daha düşük bir öğrenme oranı ile tekrar çalışmasını sağlar. Eğer, sonuçtan uzaklaşılacak şekilde bir ilerleyiş varsa, sistem otomatik olarak daha düşük bir öğrenme oranı ile yinelenir. Yineleme pozitif olarak işaretlenmiştir.

- Alıştırma Zamanı (trainingTime) : Ağ üzerinde kaç yineleme yapılacağını gösterir. Uygulama için 500 yineleme yapılması işaretlenmiştir.

- Değerlendirme Kümesinin Büyüklüğü (validationSetSize) : Değerlendirme kümesinin orantısal büyüklüğüdür. Alıştırma işlemi, değerlendirme kümesindendeki hatanın sürekli olarak büyüme eğilimi gözlenmesine veya yineleme sayısına ulaşılmasına kadar devam eder. Tüm verinin değerlendirmede kullanılması için '0' değeri girilmiştir.

- Değerlendirme Eşik Değeri (validationThreshold) : Değerlendirme işlemini sonlandırmak için kullanılır. Buradaki değer, alıştırma işleminin, değerlendirme kümesinin bir kayıtında kaç tane hata ile karşılaşıldığında sonlandırılacağını göstermektedir. Uygulama için, 20 hata gözlendiğinde alıştırma işleminin sonlandırılması değeri girilmiştir.



Şekil 4.13: WEKA Yapay Sinir Ağı Değiştirge Ayarları

Test seçenekleri bölümünde, tüm verinin %66'sının eğitim, geri kalan %34'lük kısımın test amaçlı olarak kullanılacağı belirtilmiştir. Hedef öznitelik olarak ALTGRUP_ALTGRUPKODU özniteliği seçildikten sonra yöntem çalıştırılmışır. Toplam 182.612 kayıta sahip ve 5 özniteliği bulunan Kredili Giyecek Sistemi verisi için WEKA aracının çok katmanlı algılayıcı yapay sinir ağı yöntemi ile ürettiği sonuçlar Şekil 'te verilmiştir.

```
=== Run information ===
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H
Relation:                                          tez1-weka.filters.unsupervised.attribute.Remove-R3,6,10-12-
weka.filters.unsupervised.instance.RemoveWithValues-S2.0-C2-L3-3-H-
weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Remove-R6
Instances:    182607
Attributes:   5
              KREDIYILI, DONEM,  ALTGRUP_ALTGRUPKODU, MIKTAR, RUTBE
Test mode:    split 66% train, remainder test

=== Classifier model (full training set) ===
Sigmoid Node 0
    Inputs    Weights
    Threshold    -23.439780095921822
    Node 10    75.03656130604568
    Node 11    26.534476386010315
    Node 12    -4.021810474266425
    Node 13    -78.14591194448383
    Node 14    26.49019034285591
    Node 15    14.094855153503662
    Node 16    -77.18543437454439
Sigmoid Node 1
    Inputs    Weights
    Threshold    -25.35528230616546
    Node 10    12.524735119538027
    Node 11    12.081954610230468
    Node 12    14.496006061289126
    Node 13    -4.758984751967913
    Node 14    -62.67284852440831
    Node 15    -7.69718880370697
    Node 16    2.5933561180198055
Sigmoid Node 2
    Inputs    Weights
    Threshold    -55.285032275510346
    Node 10    -28.84630818569475
    Node 11    -47.166209668542066
    Node 12    51.94647454150852
    Node 13    -1.5612876006062206
    Node 14    -49.04137182926891
    Node 15    51.29299549694294
    Node 16    -8.26585079611836
```

Şekil 4.14 : WEKA Yapay Sinir Ağı Yöntemi Raporu

```
Sigmoid Node 3
    Inputs    Weights
    Threshold    -25.05559271916344
    Node 10    -0.2834927358015941
    Node 11    -115.6013848460862
    Node 12    118.38984735201296
    Node 13    20.51969648323724
    Node 14    -114.72452651187047
    Node 15    -116.94044486834473
    Node 16    -3.888400173630891
Sigmoid Node 4
    Inputs    Weights
    Threshold    -29.80323705017128
    Node 10    -1.217863841316322
    Node 11    3.251029576781274
    Node 12    21.474047762009665
    Node 13    2.043554471273471
    Node 14    3.047628981904765
    Node 15    2.0888863624744305
    Node 16    0.3438473352054314
Sigmoid Node 5
    Inputs    Weights
    Threshold    -23.114792183070026
    Node 10    -2.5789551547225886
    Node 11    -35.368990774476146
    Node 12    4.276399934089863
    Node 13    -3.4390361106783915
    Node 14    20.14733580174717
    Node 15    -13.914993022591005
    Node 16    6.024913676741946
Sigmoid Node 6
    Inputs    Weights
    Threshold    -57.59034380870406
    Node 10    -4.532934497955773
    Node 11    -5.406732872282652
    Node 12    11.661835445856166
    Node 13    -0.3103112357909848
    Node 14    37.37867378520206
    Node 15    10.383516568923298
    Node 16    4.488806303393089
Sigmoid Node 7
    Inputs    Weights
    Threshold    -67.52719973487889
    Node 10    -26.198981907133295
    Node 11    56.810965441921425
    Node 12    2.8201520280233763
    Node 13    8.965171287206452
    Node 14    -32.800272463819816
    Node 15    51.89161165639149
    Node 16    53.28602202262221
Sigmoid Node 8
    Inputs    Weights
    Threshold    -5.906589256153146
    Node 10    5.942849048089077
    Node 11    -54.07379452147904
    Node 12    -51.539051664207456
```

Şekil 4.14 (Devam): WEKA Yapay Sinir Ağı Yöntemi Raporu

```
Node 13    14.72730114348907
    Node 14    0.4643633473198882
    Node 15    -4.166849493682477
    Node 16    -11.817578377153138
Sigmoid Node 9
    Inputs    Weights
    Threshold    15.889531153821759
    Node 10    -3.7997580097727663
    Node 11    -23.92493643987477
    Node 12    -62.1324335396935
    Node 13    -47.38306762095864
    Node 14    -3.9935661995645346
    Node 15    -90.8243046834562
    Node 16    34.35712899011583
Sigmoid Node 10
    Inputs    Weights
    Threshold    -2.3923333446766137
    Attrib KREDIYILI    -46.86984018206309
    Attrib DONEM    6.7634755892900635
    Attrib MIKTAR    21.80103372115957
    Attrib RUTBE    -20.6577848152882
Sigmoid Node 11
    Inputs    Weights
    Threshold    -34.343957772335976
    Attrib KREDIYILI    133.8894134828896
    Attrib DONEM    74.61381670642763
    Attrib MIKTAR    -14.621196872533242
    Attrib RUTBE    -72.48376346069846
Sigmoid Node 12
    Inputs    Weights
    Threshold    219.26686192088647
    Attrib KREDIYILI    -12.354740052586598
    Attrib DONEM    7.813873599179282
    Attrib MIKTAR    225.8568858234359
    Attrib RUTBE    -5.1330343429558045
Sigmoid Node 13
    Inputs    Weights
    Threshold    15.194301868387962
    Attrib KREDIYILI    -27.781894895191567
    Attrib DONEM    -30.196535784164986
    Attrib MIKTAR    -0.3678314191963723
    Attrib RUTBE    -16.913850512982794
Sigmoid Node 14
    Inputs    Weights
    Threshold    25.551071582483416
    Attrib KREDIYILI    -12.8690876021949
    Attrib DONEM    -1.4289659627980216
    Attrib MIKTAR    31.072005250824574
    Attrib RUTBE    6.258728126150747
Sigmoid Node 15
    Inputs    Weights
    Threshold    -89.91724572037437
    Attrib KREDIYILI    -135.01332584828646
    Attrib DONEM    -100.73527650169021
    Attrib MIKTAR    30.576007816225182
    Attrib RUTBE    -79.77742434949265
```

Şekil 4.14 (Devam): WEKA Yapay Sinir Ağı Yöntemi Raporu

```
Sigmoid Node 16
    Inputs    Weights
    Threshold    -118.80753618709869
    Attrib KREDIYILI    -26.052951940096108
    Attrib DONEM    -18.0293363463466
    Attrib MIKTAR    -55.76325790657239
    Attrib RUTBE    72.26017182892983
Class G01
    Input
    Node 0
Class F01
    Input
    Node 1
Class H01
    Input
    Node 2
Class C01
    Input
    Node 3
Class L01
    Input
    Node 4
Class C03
    Input
    Node 5
Class D01
    Input
    Node 6
Class B01
    Input
    Node 7
Class E01
    Input
    Node 8
Class H02
    Input
    Node 9

=== Summary ===
Correctly Classified Instances       49952               80.4548 %
Incorrectly Classified Instances     12135               19.5452 %
Mean absolute error                  0.0482
Root mean squared error              0.1841

=== Confusion Matrix ===
    a     b     c     d     e     f     g     h     i     j   <-- classified as
 13134    96   100   326     1   480     0    71    35   481 |     a = G01
   557  4268     0     1     0     0     0     0     0     0 |     b = F01
   302     0  1843     0     0     0     0    53    11     0 |     c = H01
   210     8     8  5826   188     0     0   430     2   177 |     d = C01
   230    76    40     0   998    54   258    35    11   465 |     e = L01
   768     0   235     0     0  6893     0   168     0    84 |     f = C03
   361    39   169    36    62   351   852    59     7   476 |     g = D01
   221     0   409     0     0     0     0  4019   188    83 |     h = B01
   903    27   176   188    12    95     3    84  8413   776 |     i = E01
   177     0     0     0     0  1242    30     0     0  3706 |     j = H02
```

Şekil 4.14 (Devam): WEKA Yapay Sinir Ağı Yöntemi Raporu

4.4.4. Modelleme sonuçlarının yorumlanması

Bir önceki bölümde, WEKA veri madenciliği aracı kullanılarak, Deniz Kuvvetleri giyecek siparişi verisi üzerinde Naïve Bayes, karar ağacı ve yapay sinir ağı yöntemleri için ayrı ayrı modelleme yapılmıştır. WEKA aracında, Naïve Bayes için yine Naïve Bayes; karar ağacı için J48 ve yapay sinir ağı için çok katmanlı algılayıcı (MLP-MultilayerPerceptron) kullanılmaktadır.

WEKA aracında, Naïve Bayes yöntemi ile modelleme yapılması için Şekil 4.8'de ayrıntıları verilen gerekli değiştirge ayarlamaları yapıldıktan sonra model çalıştırılmış ve Şekil 'da verilen sonuç raporu alınmıştır.

Sonuç raporu incelenirse; raporun başında, kullanılan sınıflandırıcının adı, veri hazırlama işlemlerinde veri üzerinde yapılan değişiklikler, kayıt sayısı, öznitelik sayısı ve adları ile eğitim ve test verisinin oranları gibi genel bilgiler görülebilir. Her üç yöntem için başlık kısmı aynı olup; sadece kullanılan sınıflayıcı adı değişmektedir.

Naïve Bayes yöntemi sonuç raporunun sonraki bölümünde, hedef özniteliği olan giyecek malzemelerinin (ALTGRUP_ALTGRUPKODU) her biri için genel hesaplamalara yer verilmiştir. Örneğin, ALTGRUP_ALTGRUPKODU hedef özniteliğinin G01, yani, elbise yazlık kısa kollu beyaz, değeri için istatistiksel bilgiler verilmiştir. Bu istatistiksel bilgilerin içinde ilk olarak, tüm kayıtlar içinde G01 olanların oranını veren ilk olasılık değeri verilmiştir (%24). Daha sonra, hedef özniteliğindeki değer G01 iken, diğer tüm özniteliklerin istatistiksel değerleri verilmiştir. Bu istatistiksel hesaplamalar arasında; normal dağılım, ortalama değer, standart sapma, ağırlıklı toplam gibi girdi öznitelik değerlerinin bilgileri verilmiştir. Hedef özniteliğine ait tüm değerler için yukarıdaki işlemler tekrarlandıktan sonra, yöntemin oluşturduğu sonucun özet bilgileri verilmiştir. Buna göre, Naïve Bayes modellemesinin doğru olarak sınıflandırdığı kayıt oranı %64,55'tir. Eğer modelleme yapılmadan tahminde bulunulmak istenseydi, ALTGRUP_ALTGRUPKODU hedef

özniteliğinde 10 farklı değer olduğuna göre, G01 değeri için olasılık değerinin yüzdesi

$$\Pr('G01') = \frac{100}{10} = 10 \qquad (4.1)$$

olacaktır. Diğer yandan, raporun ilk bölümünde verilen ve G01 değeri alan kayıtların, tüm kayıtlara oranından hesaplanan değer olarak %24 verilmişti. Bu değer %10 değerinden çok daha iyi bir sonuçtur. Ama Naïve Bayes sınıflandırma yönteminin bulduğu olasılık değeri olan ~%65'lik değer, bu değerlerin çok üzerindedir. Bu nedenle, Naïve Bayes yöntemi bu veri üzerinde başarılı olmuştur denilebilir.

Her üç raporun en son kısmında dağılım matrisi verilmiştir. Bu matriste, yöntemin eğitim sonucunda bulduğu değer ile gerçek değerin karşılaştırması verilmiştir. Örneğin, eğitim sonucu G01 olarak sınıflandırılan kayıtlar test sonucunda değerlendirildiğinde, bunlardan 10235 tanesinin doğru olarak G01, 633 tanesinin yanlış olarak F01, 198 tanesinin yanlış olarak H01 değerlendirildiği görülmektedir.

Naive Bayes yönteminde, her bir kayıtın değerinin ne olacağı tahmin edilirken, (3.2) denklemi kullanılmıştır.

Raporda, modelin doğruluğunun değerlendirmesi, doğru olarak sınıflandırılan kayıtların oranları bilgisi yanında, ortalama mutlak hata ve ortalama kareler hatası gibi bilgiler de verilmiştir. Naive Bayes yönteminin oluşturduğu modelin ortalama mutlak hatası 0,118 ve ortalama kareler hatası 0,241 olarak bulunmuştur. Bu hata oranları, ne kadar düşük ise, modelin doğruluk oranı bir o kadar yüksektir.

WEKA aracının karar ağaçları için ürettiği sonuç raporu Şekil'de verilmiştir. Bu sonuç raporunun başlık kısmı, kullanılan sınıflandırıcı adının J48 olduğu bilgisinin dışında, Naive Bayes için üretilen raporun başlık kısmı ile aynıdır. Daha sonra ağacın

dallanmalarını temsil edecek şekilde, yöntemin oluşturduğu kurallar kümesi verilmiştir. Örneğin, ilk kural olan ;

DONEM = 2
| MIKTAR <= 1
| | KREDIYILI <= 2002
| | | RUTBE <=35 : G01 (10248.0/1909.0)

kuralına bakılacak olursa, bu kuralın anlamı şu şekildedir;

Eğer malzeme, 2001 ve 2002 yıllarında sipariş edilmişse ve bu malzeme yazlık bir giyecek ise, siparişi veren subay veya başçavuş ise, bu durumda sipariş edilen malzeme kısa kollu yazlık elbisedir.

Rapor, oluşturulan ağacın 58 adet yaprağının olduğunu belirtmektedir. Bunun anlamı 58 farklı kural yazılabilir demektir.

Özet bölümünde; doğru olarak sınıflandırılan kayıtların oranının % 89,48 olduğu belirtilmiştir. Bu rakam ciddi anlamda yüksek bir rakamdır. Karar ağacı yöntemi ile, girdi öznitelikleri bilinen veri için yaklaşık %90 olasılıkla, sipariş edilen malzeme doğru olarak tahmin edilmiştir. Modelin ortalama mutlaka hatası 0,0313, ortalama kareler hatası 0,1254 olarak hesaplanmıştır. Naive Bayes modeli ile karşılaştırıldığı zaman, hata oranların çok daha düşük olduğu görülmektedir. Doğruluk oranı ile tespit edilen karar ağacının daha iyi bir model oluşturduğu sonucu, hata oranlarının karşılaştırılması ile bir kez daha doğrulanmıştır.

Weka aracı, karar ağaçları için metin şeklindeki rapor yanında, modelin ağaç görünümünü de vermektedir. Bu ağaç görünüm Şekil 4.15'te verilmiştir.

WEKA aracının yapay sinir ağı yöntemi kullanarak ürettiği sonuç raporu Şekil 'te verilmiştir. Bu sonuç raporunun başlık kısmı, kullanılan sınıflandırıcı adının çok katmanlı algılayıcı (multilayer perceptron) olduğu bilgisinin dışında, Naive Bayes ve karar ağaçları için üretilen raporun başlık kısmı ile aynıdır.

Rapor detaylı olarak incelendiği zaman, öncelikle transfer fonksiyonu olarak sigmoid fonksiyonun kullanıldığı görülmektedir. Ayrıca, çok katmanlı algılayıcı yönteminin 2 tane gizli katman belirlediği görülmektedir. Birinci katmanda 10,11,12,13,14,15,16 numaralı toplam 7 tane düğüm vardır. Girdi öznitelikleri bu 7 tane düğüme bağlanmıştır. Her bir düğüme, tüm girdi özniteliklerinden modelin belirlediği ağırlıkta bir girdi olmaktadır. Bu düğümler, kendilerine gelen girdilerin ağırlıklarının toplamından oluşmaktadır. Örneğin, 10 numaralı düğüme gelen girdiler ve ağırlıkları Tablo 4.5'te verilmiştir.

Tablo 4.5: 10 Numaralı Düğüme Gelen Girdiler ve Ağırlıkları

Girdi	**Ağırlığı**
KREDIYILI	-46.8698401820630900
DONEM	6.7634755892900635
MIKTAR	21.8010337211595700
RUTBE	-20.6577848152882000

Numarası 10 ile16 arasında olan düğümler; birinci gizli katmanı, numarası 0 ile 9 arası olan toplam 10 düğüm ise ikinci gizli katmanı oluşturmuştur. Yani, 0-9 arasında numarası olan düğümler için girdiler, sıra numarası 10-16 olan düğümler olmuştur. Son olarak, numarası 0-9 arası olan düğümler, çıktı özniteliği olan ALTGRUP_ALTGRUPKODU özniteliğinin her bir değeri için girdi olmuştur.

Çok katmanlı algılayıcının çalışma yöntemine göre, ilk etapta çıktı düğümlerinde hata miktarı bulunmuş, daha sonra düğümlerin ağırlıkları değiştirilerek, hata miktarının en az olduğu durum bulunmaya çalışılmıştır.

Özet bölümünde; doğru olarak sınıflandırılan kayıtlar oranının % 80,45 olduğu belirtilmiştir. Bu rakam da yüksek bir rakamdır. Naive Bayes yöntemi ile karşılaştırıldığında çok daha iyi bir doğruluk oranına sahip olduğu görülürken, karar ağacı yöntemi ile yapılan modellemenin sonucu ile karşılaştırıldığında, doğruluk oranı biraz düşük kalmaktadır. Girdi öznitelikleri bilinen veri için sipariş edilen malzeme, yaklaşık %80 olaslıkla doğru olarak tahmin edilmiştir. Modelin ortalama mutlaka hatası 0,0482 ve ortalama kareler hatası 0,1841 olarak hesaplanmıştır.

Doğruluk oranı ile tespit edilen Naive Bayes'ten daha iyi, karar ağacından daha kötü bir model oluşturduğu sonucu, hata oranlarının karşılaştırılması ile de doğrulanmaktadır.

4.5. Değerlendirme

4.5.1. Sonuçların değerlendirilmesi

Tüm verinin 2/3'ünün model oluşturma ve geri kalan 1/3'ünün test amaçlı olarak kullanıldığı önceki bölümlerde belirtilmişti. Naïve Bayes yönteminin doğruluk tablosu, Tablo 4.6'da, karar ağacı yönteminin doğruluk tablosu Tablo 4.7'de ve yapay sinir ağı yönteminin doğruluk tablosu Tablo 4.8'de verilmiştir. Tabloların hem satır, hem de sütun başlığında ALTGRUP_ALTGRUPKODU özniteliğine ait değerler verilmiştir. Satır hanesi, modelin tahmin ettiği, sütun hanesi ise test verisinde karşılaşılan değeri göstermektedir. Örneğin, ilk satırdaki G01 ile son sütundaki H02'yi çakıştırıldığında 481 sayısı görülür. Bunun anlamı; modelin G01 olarak tahmin ettiği, ama test verisinde F01 olduğu görülen kayıt sayısı 633'tür. Aynı şekilde ilk satır ve ilk sütuna bakılırsa; modelin G01 olarak tahmin ettiği ve test verisinde de G01 olduğu görülen kayıt sayısı 10235'tir.

Model oluşturulduktan sonra, test verisi üzerinde modelin doğruluğunun test edilmesi sonucunda Tablo 4.9'daki sonuçlar elde edilmiştir. Tablonun incelenmesinden anlaşılacağı üzere, Kredili giyecek sistemi verisi üzerinde en iyi sınıflandırma yapan yöntem % 89.48 oranıyla, karar ağacı yöntemidir. Karar ağacı yöntemiyle alıştırma verisi üzerinde oluşturulan model, test verisi üzerinde % 89.48'lik bir doğruluk payı ile tahminde bulunmuştur.

Yapay sinir ağı yöntemi, bu tez çalışmasında incelen veri üzerinde % 80.45'lik doğruluk payı ile karar ağacı yöntemi kadar başarılı olmasa da, iyi bir alternatif olduğunu kanıtlamıştır.

Tablo 4.6**:** Naïve Bayes Yönteminin Doğruluk Tablosu

	G01	**F01**	**H01**	**C01**	**L01**	**C03**	**D01**	**B01**	**E01**	**H02**
G01	10235	883	69	189	1	1521	258	616	612	237
F01	633	3777	0	456	0	0	0	0	0	0
H01	198	0	1397	0	0	0	0	0	697	0
C01	2315	0	345	1987	236	1191	251	608	30	0
L01	54	0	29	25	1004	55	373	181	254	184
C03	85	0	0	261	0	5938	675	5	1028	39
D01	0	0	27	60	195	210	873	515	345	164
B01	571	0	891	0	0	0	0	2617	809	61
E01	315	0	66	47	0	167	4	155	7549	2382
H02	0	0	0	0	0	82	201	24	109	4704

Tablo 4.7**:** Karar Ağaçları Yönteminin Doğruluk Tablosu

	G01	**F01**	**H01**	**C01**	**L01**	**C03**	**D01**	**B01**	**E01**	**H02**
G01	13255	0	59	295	0	571	43	110	391	0
F01	334	4484	0	1	0	0	7	0	0	0
H01	174	0	1648	0	0	78	0	230	79	0
C01	89	0	8	6725	0	7	0	3	17	0
L01	231	0	28	83	1082	95	327	42	278	0
C03	171	0	0	1	0	7852	0	81	43	0
D01	361	0	39	7	0	185	1548	35	237	0
B01	18	0	51	0	0	5	0	4727	119	0
E01	399	0	107	3	0	219	75	207	9667	0
H02	299	0	0	14	0	102	30	0	144	4566

Naïve Bayes yönteminin % 64.55'lık doğruluk oranı ise, diğer iki yöntemin becerisinin oldukça gerisinde kalmıştır.

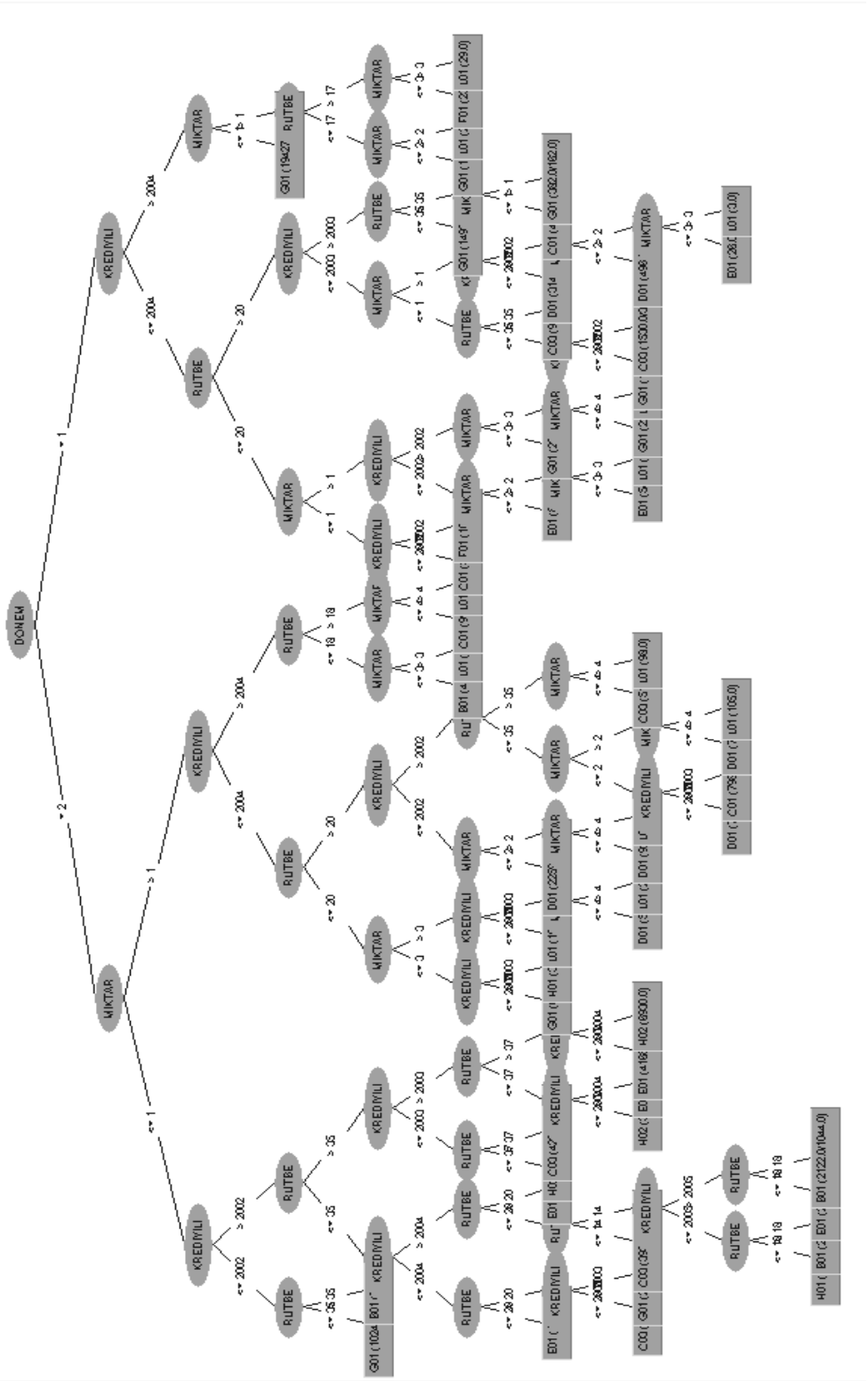
En yüksek doğruluk oranını veren karar ağacı yöntemine ait ağaç görünümü Şekil 4.15'de gösterilmiştir.

Tablo 4.8**:** Yapay Sinir Ağları Yöneminin Doğruluk Tablosu

	G01	**F01**	**H01**	**C01**	**L01**	**C03**	**D01**	**B01**	**E01**	**H02**
G01	13134	96	100	326	1	480	0	71	35	481
F01	557	4268	0	1	0	0	0	0	0	0
H01	302	0	1843	0	0	0	0	53	11	0
C01	210	8	8	5826	188	0	0	430	2	177
L01	230	76	40	0	998	54	258	35	11	465
C03	768	0	235	0	0	6893	0	168	0	84
D01	361	39	169	36	62	351	852	59	7	476
B01	221	0	409	0	0	0	0	4019	188	83
E01	903	27	176	188	12	95	3	84	8413	776
H02	177	0	0	0	0	1242	30	0	0	3706

Tablo 4.9**:** Yöntemlerin Doğruluk Oranları ve Hataları

Yöntem Adı	**Doğruluk Oranı**	**Ortalama Mutlak Hata (MAE)**	**Ortalama Kareler Hatası (MSE)**
Naïve Bayes	64.55 %	0.1180	0.2410
Karar Ağacı	89.48 %	0.0313	0.1254
Yapay Sinir Ağı	80.45 %	0.0482	0.1841



Şekil 4.15**:** Karar Ağacı Yöntemi Sonucunun Ağaç Görünümü

4.5.2. İşlemlerin gözden geçirilmesi

Elde edilen sonuçların kararlı ve doğru olduğunu anlamak için, aynı veri her üç yöntem ile tekrar analiz edilmiş ve sonuçlar aynı çıktığı için modellemelerin doğru ve kararlı olduğu tespit edilmiştir. Yapılan ikinci doğrulama analizi esnasında, tüm değiştirgeler tekrar incelenmiş ve onlarda da herhangi bir sorun olmadığı saptanmıştır.

4.5.3. Sonraki işlemin belirlenmesi

İşlem gözden geçirme işleminde herhangi bir sorunla karşılaşılmadığından, elde edilen sonuçların doğru olduğuna karar verilmiştir. Karar ağacı yöntemi, en yüksek doğruluk oranına sahip olduğundan, bu tez konusu uygulama verisi için en uygun yöntemdir.

CRISP-DM yöntembiliminin gerçekleme adımında, kredili giyecek sistemi verisi için karar ağacı yöntemi kullanılarak, KREDIYILI, DONEM, MIKTAR ve RUTBE özniteliklerinin değerlerinin bilinmesi durumunda, hangi giyecek malzemesinin seçileceği bilgisini yaklaşık %90 olasılıkla doğru tahmin edilebilmektedir.

4.6. Gerçekleme

4.6.1. Planın gerçeklemesi

CRISP-DM yöntembiliminin modelleme adımında karşılaştırılan üç yöntemden, bu tez çalışmasında incelenen veri için en uygun olanının, karar ağacı yöntemi olduğu belirlenmiştir.

Karar ağacı yöntemi ile modelleme sonucunda, diğer dört özniteliğin bilindiği durumlarda hangi malzemenin sipariş edileceği bilgisine yaklaşık olarak % 90 olasılıkla bilinmektedir.

Sipariş edilecek olan malzemelerin önceden bilinmesi için gerekli olan 4 adet öznitelik; KREDIYILI, DONEM, MIKTAR ve RUTBE öznitelikleridir. Bu tez çalışmasında yapılan, ALTGRUP_ALTGRUPKODU bilgisinin, KREDIYILI, DONEM, MIKTAR ve RUTBE cinsinden denklem haline getirilmesidir.

Deniz Kuvvetleri Komutanlığı'nın giyecek sistemi ile ilgilenen birimleri, bu veri madenciliği çalışması sonuçlarını kullanarak; giyecek sistemininin sorunlarını belirleyebilir ve bu sorunlara çözümler üreterek, giyecek sisteminin hem devlet, hem de personel yararına olacak şekilde iyileştirilmesini sağlayabilir.

4.6.2. Planın izlenmesi ve düzeltilmesi

Kredili Giyecek Sistemi verisi üzerinde yapılan veri madenciliği çalışmasına ait adımlar tekrar gözden geçirilmiş ve herhangi bir problemin olmadığı görülmüştür. Bu nedenle, başlangıçtaki plan üzerinde yapılması gereken herhangi bir düzeltme ihtiyacı bulunmamaktadır.

Yapılan bu veri madenciliği çalışması, her yılın sonunda oluşturulan veri ile tekrarlanarak, personelin sipariş verme eğilimlerindeki değişiklikler bulunabilir.

4.6.3. Sonuç raporunun düzenlenmesi

Kredili Giyecek Sistemi verisi üzerinde yapılan veri madenciliği çalışmaları kapsamında, verinin 2/3'ü model oluşturmak ve geri kalan 1/3'lük kısım ise oluşturulan modeli test etmek amacıyla kuşllanılmıştır.

Veri madenciliği sürecinin modelleme adımında, Naïve Bayes, karar ağaçları ve yapay sinir ağları yöntemleri denenerek hangisinin bu tez çalışmasında ele alınan veri için en doğru sonuçları ürettiği tesipit edilmiştir. Sonuç olarak, yaklaşık % 90'lık bir doğruluk oranı ile karar ağacının, bu çalışmada kullanılan veri için en uygun yöntem olduğu anlaşılmıştır.

4.6.4. Projenin gözden geçirilmesi

Tüm proje tekrar gözden geçirilerek, projenin adımlarında herhangi bir sorun olmadığı tekrar doğrulanmıştır.

5. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasının amacı; veri madenciliğinin genel hatlarıyla tanıtılması, yöntemlerinin incelenmesi, CRISP-DM yöntembilimin tanıtılması ve uygulanması ile Deniz Kuvvetleri Komutanlığı personelinin giyecek siparişi verme sürecinin nasıl iyileştirilebileceğinin belirlenemsi konuları yer almıştır. Bu amaçla, sınıflandırma ile ilgili yöntemlerden Naive Bayes, karar ağaçları ve yapay sinir ağları karşılaştırılmıştır. Deniz Kuvvetleri giyecek sipariş verisi üzerinde veri madenciliği sınıflandırma yöntemleri karşılaştırılarak, giyecek sisteminin iyileştirilmesi için en uygun yöntemin, CRISP-DM yöntembilimi kullanılarak belirlenmesi amaçlanmıştır.

CRISP-DM yöntembilimi kullanarak Deniz Kuvvetleri verisinde, veri madenciliği yöntemlerinin karşılaştırılması uygulamasının yapıldığı bu tez çalışması, beş ana bölüm halinde hazırlanmıştır.

Birinci bölümde veri madenciliği hakkında genel bilgiler verilmiş, literatür taraması yapılmış ve bu tez çalışmasında ele alınan uygulamanın amacı hakkında genel bilgi verilmiştir.

İkinci bölümde, veri madenciliği ve veri madenciliği yöntemleri ile ilgili detaylı bilgiler verilmiştir.

Üçüncü bölümde, en çok kullanılan veri madenciliği yöntembilimi olan CRISP-DM yöntembilimi detaylı olarak anlatılmıştır. Bu tez çalışmasında karşılaştırılan Naive Bayes, karar ağacı ve yapay sinir ağı sınıflandırma yöntemleri detaylı olarak ele alınmıştır.

Dördüncü bölümde, veri madenciliği arenasında en çok kullanılan yöntembilim olan CRISP-DM yöntembilimi kullanılarak, veri madenciliği çalışmaları belirli bir disiplin altına alınmıştır. Bu bölümde, CRISP-DM yöntembiliminin adımları takip

edilerek, Deniz Kuvvetlerinde kullanılan veri üzerinde veri madenciliği yöntemlerinin karşılaştırılması amaçlanmıştır. Bu amaç için sırasıyla;

- Yapılan iş ile ilgili detaylara yer verilmiştir: Deniz Kuvvetleri Komutanlığı personeline görev gereği verilen üniformalara ait sipariş süreci incelenmiştir. Bu süreç içerisinde, personelin, Kredili Giyecek Sisteminden genel olarak memnun olduğu, ama çeşitli problemlerin de olduğu tespit edilmiştir. Veri madenciliğinin, sorunların neler olduğunun belirlenmesi ve bu sorunlara cevap üretilmesi konusunda yardımcı olabileceği değerlendirilmiştir.

- Veri incelenmiştir: Siparişlerin Kredili Giyecek Sistemi aracılığıyla girildiği görülmüştür. Oracle veri tabanını kullanan bu uygulama incelenmiştir. Daha sonra, oracle veri tabanındaki ilgili tablolar incelenerek, personelin siparişini verdiği malzemeyi etkileyebilecek öznitelikler belirlenmiştir. Bu öznitelikler Cahit Arf v1.0 uygulaması kullanılarak, ARFF biçimine dönüştürülmüştür. ARFF biçimine dönüştürülmüş veriyi WEKA veri madenciliği aracı ile açtıktan sonra, WEKA'nın özelliklerinden faydalanılarak, uygulamada kullanılan tüm verinin genel karakteristikleri incelenmiştir.

- Veri temizlenmiştir: Öncelikle, veri madenciliği sonucunu etkilemeyeceği değerlendirilen öznitelikler, WEKA aracının "Remove" filtresi kullanılarak veri madenciliği çalışması dışında tutulmuştur. Eksik veri olmamasına rağmen, hatalı veri ile karşılaşıldığından, WEKA aracının "RemoveWithValues" filtresinden faydalanılarak 5 hatalı kayıt silinmiştir. Bunun haricinde veri temizlemesi gerektirecek bir durum ile karşılaşılmamıştır.

- Modelleme yapılmıştır: Kredili Giyecek Sistemi verisi için en uygun sınıflandırma yönteminin bulunması amacıyla Naive Bayes, karar ağacı ve yapay sinir ağı yöntemleri ile modelleme yapılmıştır. Yapılan modelleme çalışması sonucunda, karar ağacı yöntemi %89.48, yapay sinir ağı yöntemi % 80.45 ve Naive Bayes yöntemi % 64.55 doğruluk oranı ile Deniz Kuvvetleri Komutanlığı personeli tarafından sipariş edilen malzemeyi tahmin etmişlerdir.

- Değerlendirme yapılmıştır: Modelleme adımında karşılaştırılan yöntemlerin detaylı değerlendirilmesi yapılmıştır. Bunun için Naive Bayes, karar ağacı ve yapay sininr ağı yöntemlerinden her biri için doğruluk oranı tabloları hazırlanmıştır. Bu bölümde, her bir yöntem için, bir diğer değerlendirme ölçütü olan ortalama mutlak hata ve ortalama kareler hatası ile yöntemlerin değerlendirmeleri yapılmıştır. Değerlendirme sonucunda, doğruluk oranlarından da anlaşılacağı üzere, bu tez çalışmasında ele alınan veri için en uygun veri madenciliği yöntemi olarak karar ağacı belirlenmiştir.

- Gerçekleme yapılmıştır: Bu tez çalışmasında yapılan veri madenciliği çalışmasından, Deniz Kuvvetleri Komutanlığı'nın nasıl faydalanabileceği konusu incelenmiştir. Yapılan veri madenciliği çalışmasının amacı olan, sipariş edilen malzemelerin, diğer öznitelikler cinsinden denklem haline getirilmesinin ne anlam taşıdığı belirtilmiştir.

Bu tez çalışmasının konusu gereğince ele alınan konu, sipariş edilen giyecek malzemelerinin diğer öznitelikler yardımıyla tahmin edilmesidir. Bu tez çalışmasında ele alınan konunun daha ileri götürülmesi için giyecek malzemlerinin neler olduğunun bulunması yanında, hangi malzemelerden ne miktarda sipariş edilebileceğinin tahmin edilmesi konusu bir başka tez konusu olabilir. İleriye dönük olarak miktarlarıyla birlikte malzemeleri tahmin etmek için zaman serilerinden faydalanılabilir. Zaman serileri ile çalışırken, model oluşturma ve test verisi, bu tezde kullanıldığından farklı olmak zorundadır. Mevcut veri incelendiğinde, 2002, 2003, 2004, 2005 ve 2006 yılları için veri vardır. Bunlardan, 2002, 2003, 2004 ve 2005 verisi model oluşturmak ve 2006 yılı verisi modeli test etmek için kullanılabilir.

Bu tez çalışması esnasında, ticari veri madenciliği yazılımları da incelenmiştir. Özellikle grafiksel oluşturma ortamı ile oldukça kullanışlı ve akıllı veri madenciliği araçları geliştirilmiştir. Diğer yandan, veri madenciliği araçları ne kadar yetenekli olursa olsun, veri madenciliği kousunda uzmanlaşmış personel olmadan, veri madenciliğinden kazanç sağlamak imkansızdır. Bu noktada, Waikato Üniversitesinin önayak olması ile WEKA aracını geliştiren tüm gelişitirici personelin hakkını

vermek gerekir. Açık kaynak kodlu olmasına rağmen, oldukça zengin bir içerikle hazırlanan yazılım, sadece akademik ortamlarda değil, ticari ortamlarda da, diğer ücretli ticari yazılımlarla baş edebilecek kalitededir.

KAYNAKLAR

Adriaans, P., and Zantige, D., "Data Mining", ***Addison Wesley Publishing***, (1996).

Aktaş, Z., 2002, *Bilgi ve Bilgi Toplumu Üstüne*, Çankaya Üniversitesi, http://www.cankaya.edu.tr/eng/publications/h1g1.php, **(Ziyaret tarihi: 10 Mayıs 2006).**

Alkan, A., "Veri Madenciliğine Kısa Bir Giriş", Simya Danışmanlık, Ekim (2003).

Alpaydın, E., 2000, *Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*, Bilişim 2000 Veri Madenciliği Eğitim Semineri, http://www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden-2k-notlar.doc, (**Ziyaret tarihi: 16 Ocak 2006)**

Bayes, T., "An essay towards solving a problem in the doctrine of chances", ***Philosophical Transactions of the Royal Society***, 53, 370-418, (1763).

Berry, M.J.A. and Linoff, G.S., "Data Mining Techniques For Marketing, Sales and Customer Relationship Management", Second edition, ***Wiley Publishing***, (2004).

Bharti, A, "A Decision Tree Approach to Extract Knowledge for Improving Satellite Image Classification", Doktora Tezi, ***International Institute for Geo-information Science and Earth Observation***,2-4, (2004)

Brown, E., "Analyze this", ***Forbes***, 169, 96–98, (2002).

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A., "Discovering Data Mining: From concept to implementation", ***Prentice Hall Publishing***, (1998).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., "Step by step data mining guide", ***SPSS inc.,*** (2000).

Fayyad, U.M., Piatestsky-Shapiro, G., Smith, P., "Advances in Knowledge Discovery and Data Mining", (1996).

Firestone, J.M., "Data Mining and KDD: A Shifting Mosaic", 12 Mart (1997).

Freeman, J.A.,and Skapura, D.M., "Neural Networks Algorithms, Applications and Programming Techniques", ***Addison-Wesley Publishing***, (2001).

Gahegan, M. ve West, G., The classification of Complex Data Sets: An operational Comparison of Artificial Neural Networks and Decision Tree Classifiers, Eylül (1998).

Han, J. and Kamber M.,"Data Mining: Concepts and Techniques", Morgan Kaufmann Publishing, (2001).

Hand, D., Mannila, H., Smyth, P., "Principles of Data Mining", ***MIT Press***, (2001).

German, G.W.H., West, G., Gahegan M., "Statistical and AI Techniques in GIS Classification: A Comparison", ***Eleventh Annual Colloquium of the Spatial Information Research Centre***, University of Otago, New Zeland, 13-15 December (1999).

Kantardzic, M., "Data Mining: Concepts, Models, Methods, and Algortihms", ***John Wiley & Sons Publishing***, (2003).

Kasabov, N.K., "Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering", ***The MIT Press***, (1998).

Kdnuggets, 2004, www.kdnuggets.com/polls/2004/data_mining_methodology.htm **(Ziyaret tarihi: 19 Mart 2006)**.

Keim, D.A., "Visual Techniques for Exploring Databases", ***University of Halle-Wittenberg***, (2004).

Keogh, E., Lonardi, S., Ratanamahatana, C.A., "Towards Parameter-Free Data Mining", ***ACM Publishing***, (2004).

Kestelyn, J., "No longer an afterthought", ***Intelligent Enterprise***, 12 Ağustos (2002).

Konrad, R., "Data mining: Digging user info for gold", ***ZDNET News***

Larose, D.T., "Data Mining Methods and Models", ***Wiley Computer Publishing***, (2006).

Larose, D.T., "Discovering Knowledge in Data: An Introduction to Data Mining", ***Wiley Computer Publishing***, (2005).

Mateyaschuk, J., 2000, *The 1999 National IT Salary Survey: Pay up*, Information Week, http://www.informationweek.com/731/salsurvey.htm **(Ziyaret tarihi: 10 Mayıs 2006)**.

Mattison, R, "Data Warehousing and Data Mining for Telecommunication", ***Artech House Publishing***, (1997).

Mitra, S., and Acharya T., "Data Mining: Multimedia, Soft Computing and Bioinformatics", ***Wiley Computer Publishing***, (2003).

Moore, G.E., "Cramming More Components Onto Integrated Circuits", ***Electronics Magazine***, 38, (1965).

Mohammadian, M., "Intelligent Agents for Data Mining and Information Retrieval", ***Idea Group Publishing***, (2004).

Nemati, N.R. and Barko, C.D., "Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance", ***Idea Group Publishing***, (2004)

Pal, M. and Mather, P.M., Decision Tree Based Classification of Remotely Sensed data, (2001).

Quinlan, J.R., "C4.5: Programs for Machine Learning", ***Morgan Kaufman Publishing***, (1993).

Quinlan, J.R., "Bagging, boosting, and C4.5", ***13th AAAI Conference on Artificial Intelligence***, AAAI Press, (1996).

Romeu, J.L., "Operations Research / Statistics Techniques: A Key to Quantitative Data Mining"

Rud, O.P., "Data Mining Cookbook", ***Wiley Computer Publishing***, (2001).

Soukup, T. and Davidson, I., "Visual Data Mining: Techniques and Tools for Data Visualization and Mining", ***Wiley Computer Publishing***, (2006).

Squier, L, What is Data Mining, 13 Kasım (2001).

Srivastava, A, Han, E.H., Kumar V., Singh, V., "Parallel Formulations of Decision Tree Classification Algorithms", (1999).

Tang, Z.and MacLennan, J., "Data Mining with SQL Server 2005", **Wiley Computer Publishing**, (2005).

Thearling, K.,2002, *An Introduction to Data Mining*, www.thearling.com **(Ziyaret tarihi: 10 Mayıs 2006)**.

Venkayala, S., Using Java Data Mining to Develop Advanced Analytics Applications, Java Developers Journal, Nisan (2005).

Waiganjo, P., Data Management 2002 Seminar, 5 Haziran (2002).

Wang, J., "Data Mining : Opportunities and Challenges", ***Idea Group Publishing***, (2003).

Wang, J., "Encyclopedia of Data Warehousing and Mining", ***Idea Group Publishing***, (2006).

Weiss, G., Saar-Tsechansky, M., Zadrozny, B., First International Workshop on Utility-Based Data Mining, ***Association for Computing Machinery Publishing***, (2005).

Weiss, S.M., Kulikowski, C.A., "Computer Systems that Learn: Classification and Prediction Methods fom statistics, Neural Nets, Machine Learning and Expert Systems", ***Morgan Kaufmann Publishing***, (1991).

Whiting, R., Tower of power, ***InformationWeek***, 875, 40–43, (2002).

Wilson, R., Advances in Instance-Based Learning Doctoral Dissertation, (1997).

Witten, I.H. and Frank, E., "Data Mining, Practical Machine Learning Tools and Techniques", Second Edition***, Elsevier Press***, (2005).

Wong, M.L. and Leung, K.S., "Data Mining Using Grammar Based Genetic Programming And Applications", ***Kluwer Academic Publishers***, (2002).

Wurman, R., "Information anxiety is produced by the ever-widening gap between what we understand and what we think we should understand", NewYork, (1989).

ÖZGEÇMİŞ

1972 yılında Menemen / İzmir'de doğdu. İlk ve orta öğrenimini Denizli'de, lise öğrenimini İstanbul Deniz Lisesi'nde tamamladı. 1990 yılında girdiği Deniz Harp Okulu Kontrol ve Bilgisayar Mühendisliği Bölümünden 1994 yılında lojistik subayı olarak mezun oldu. 1998 yılında Boğaziçi Üniversitesi Bilgisayar Mühendisliği Bölümü'nde 1 yıl süreli Otomatik Bilgi İşlem eğitimi aldı. Gölcük Envanter Kontrol Merkezi Komutanlığı'nda Yazılım Test ve Eğitim Kısım Amirliği görevini yürütmekte olup, evli ve bir çocuk babasıdır.