T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
SAYISAL YÖNTEMLER BİLİM DALI

VERİ MADENCİLİĞİNDE SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

"BANKACILIK MÜŞTERİ VERİ TABANI ÜZERİNDE BİR UYGULAMA"

Doktora Tezi

ÖZGÜR ÇAKIR

İstanbul, 2008

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
SAYISAL YÖNTEMLER BİLİM DALI

VERİ MADENCİLİĞİNDE SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

"BANKACILIK MÜŞTERİ VERİ TABANI ÜZERİNDE BİR UYGULAMA"

Doktora Tezi

ÖZGÜR ÇAKIR

Danışman : Prof. Dr. İsmail Hakkı ARMUTLULU

İstanbul, 2008

Marmara Üniversitesi
Sosyal Bilimler Enstitüsü Müdürlüğü

Tez Onay Belgesi

İŞLETME Anabilim Dalı SAYISAL YÖNTEMLER Bilim Dalı Doktora öğrencisi ÖZGÜR ÇAKIR nın VERİ MADENCİLİĞİNDE SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI"BANKACILIK MÜŞTERİ VERİ TABANI ÜZERİNDE BİR UYGULAMA" adlı tez çalışması ,Enstitümüz Yönetim Kurulunun 19.06.2008 tarih ve 2008-10/26 sayılı kararıyla oluşturulan jüri tarafından oybirliği/oyçokluğu ile Doktora Tezi olarak kabul edilmiştir.

		Öğretim Üyesi Adı Soyadı	**İmzası**
Tez Savunma Tarihi	:	26.06.2008	
1) Tez Danışmanı	:	**PROF. DR. İSMAİL HAKKI ARMUTLULU**	.....
2) Jüri Üyesi	:	**PROF. DR. EROL ŞERİF YARIZ**	.....
3) Jüri Üyesi	:	**PROF. DR. ŞAHAMET BÜLBÜL**	.....
4) Jüri Üyesi	:	**PROF.DR.DURMUŞ DÜNDAR**	.....
5) Jüri Üyesi	:	**PROF.DR.AHMET GÖKÇEN**	.....

İÇİNDEKİLER

TABLO DİZİNİ

ŞEKİL DİZİNİ

GİRİŞ

Bilgi teknolojilerinde yaşanan hızlı gelişme, karar verme süreçlerinde kullanışlı olabilecek veri miktarının giderek artmasını sağlamış ancak verilerin depolanması ve yönetilmesi konusunda bir takım zorlukları da beraberinde getirmiştir. Bu durum, basit veri tabanı sistemlerini yetersiz kılmış ve yeni bir takım veri tabanı mimarilerinin geliştirilmesine neden olmuştur. Karar destek sistemlerinin bir bileşeni olarak ilişkisel veri tabanlarının ve veri ambarlarının kullanımı giderek yaygınlık kazanmaktadır.

Ancak verilerin elde edilmesi, depolanması ve yönetilmesi de tek başına yeterli olmamaktadır. Veri miktarındaki artış, bu verilerden anlamlı ve kullanışlı bilginin çıkarımı konusunda yeni fırsatlar sunmaktadır. Yığın halindeki verilerin anlamlı ve kullanışlı bilgiye dönüştürülmesi ise günümüz işletme yöneticisi için önemli bir rekabet avantajı yaratmaktadır.

Karar verme süreçlerinde geleneksel veri analizi yöntemlerinin kullanılması halen geçerli olsa da özellikle veri tabanı teknolojileri ve yapay öğrenme alanında yaşanan gelişmelerin de dahil olduğu yeni bilgi keşfi yaklaşımları giderek önem kazanmaktadır. Önceleri *veri tabanlarında bilgi keşfi* olarak adlandırılan bu yaklaşımlara son dönemlerde *veri madenciliği* adı verilmektedir.

Veri madenciliği, verilerin elde edilmesinden anlamlı bilginin çıkarımına ve hatta bu bilginin bir eylem planına dönüştürülmesine kadar geçen tüm bir süreci ifade etmek için kullanılmaktadır. Bu süreç, bilginin çeşitli formlarda elde edildiği ve çok sayıda tekniğin kullanıldığı bir süreçtir.

Veri madenciliği süreci, katı ve açık kuralları olmamakla birlikte işi anlama, verileri anlama, verileri hazırlama, modelleme, değerlendirme ve konuşlandırma gibi temel aşamalarla tanımlanmaktadır.

Veri madenciliği, tanımlayıcı nitelikteki faaliyetleri kapsayabileceği gibi öngörü amaçlı modellerin geliştirilmesini de içerebilir. Veri tanımlama ve görselleştirme, kavram tanımlama, bölümleme, sınıflandırma, öngörü ve bağıntı analizi veri madenciliğinin fonksiyonlarını ifade etmektedir.

Günümüzde çok miktarda verinin ve anlamlı bilgi çıkarımı ihtiyacının olduğu kriminolojiden astronomiye, tıptan meteorolojiye kadar bir çok alanda veri madenciliği uygulamalarından yararlanmak mümkündür. Bu çalışmada ise uygulama alanı olarak bankacılık müşteri veri tabanı kullanılacak ve kapsamı daraltmak amacıyla veri madenciliğinin sınıflandırma fonksiyonuna ilişkin teknikleri üzerinde durulacaktır.

Bu çerçevede, çalışmamın amacı ise veri madenciliğinin sınıflandırma fonksiyonuna ilişkin tekniklerini gerçek yaşam verileri üzerinde uygulamak, veri madenciliği sürecini tüm aşamaları ile gerçekleştirmek, uygulamada elde edilen sonuçlarla tekniklere ilişkin farklılıkları tartışmak ve hangi tekniğin hangi koşullarda uygulanmasının daha etkin olacağına yönelik önerilerde bulunmaktır. Bu nedenle, çalışmamı iki ana bölüm halinde tasarlamış bulunmaktayım.

Birinci bölümde, veri madenciliği kavramı ve gelişimi, veri madenciliği süreci, veri madenciliğinin fonksiyonları ve veri madenciliğinin sınıflandırma fonksiyonu üzerinde duracak, sınıflandırma tekniklerinden lojistik regresyon analizi, yapay sinir ağları ve C5.0 karar kuralı türetme algoritmasını teorik olarak inceleyeceğim.

İkinci bölümde, bankacılık müşteri veri tabanı üzerinde veri madenciliğinin sınıflandırma fonksiyonuna ilişkin teknikleri için veri madenciliği sürecinin tüm aşamalarını kapsayan bir uygulama gerçekleştireceğim. Üç farklı sınıflandırma tekniğinin üç farklı bankacılık ürünü üzerinde uygulanması ile toplam dokuz farklı sınıflandırma modeli elde etmiş olacağım. Bu modelleri değerlendirerek uygulanan teknikleri karşılaştırma yoluna gideceğim.

Çalışmamın sonunda ise bu uygulama sırasında edinilen deneyimlerin ve hem iş hedefleri hem de veri madenciliği amacı doğrultusunda elde edilen bulguların paylaşılması amacıyla bir sonuç bölümü yer alacaktır.

1. VERİ MADENCİLİĞİ VE SINIFLANDIRMA FONKSİYONU

Bu bölümde, veri madenciliği konusunda yapılmış olan çalışmalarla ulaşılan bilgi düzeyi doğrultusunda, ikinci bölümde gerçekleştirilecek uygulama için gereken teorik çerçevenin aktarılması hedeflenmektedir. Bu kapsamda, veri madenciliği kavramından başlayarak veri madenciliği süreci, veri madenciliğinin fonksiyonları ve son olarak veri madenciliğinin sınıflandırma fonksiyonu üzerinde durulacaktır.

1.1. VERİ MADENCİLİĞİ

Karar verme süreçlerinde işe ilişkin uzmanlık ile istatistiksel modelleme araçlarını birleştirme gayreti içerisinde olan geleneksel veri analizi yerine son yıllarda bir takım yeni eğilimlerin ortaya çıktığını görmekteyiz. Bu değişimin oluşmasında veri toplama ve depolama alanında yaşanan teknolojik gelişme, hızla elde edilen ve uygulanan veri tabanlı çözümlemenin yarattığı rekabet avantajı ve çözümleme sonuçlarının karar verici açısından kolayca anlaşılması ve uygulanması ihtiyacı en önemli etkenleri oluşturmaktadır.[1]

Veri toplama ve depolama teknolojilerindeki hızlı gelişme veri tabanı, veri ambarı ve www (world wide web) alanları gibi bir çok veri deposunda her geçen gün daha fazla verinin birikmesine neden olmaktadır. Bu olağanüstü gelişmeyi güçlü veri analizi araçları olmaksızın kavramak insanoğlunun yeteneklerini aşan bir durum haline gelmiş ve bu durum karar vericiyi *veri zengini fakat bilgi* fakiri konumuna sokmuştur.[2]

Aynı dönemde dünya ekonomisinde yaşanan rekabetçi ve pazar payı kazanımına dayanan yapı, müşterinin elde edilmesi ve tutulmasının önemini de aynı oranda arttırmıştır. Artık günümüz işletme yöneticisinin veri zenginliğini bilgi zenginliğine dönüştürmesi kritik başarı faktörü haline gelmiştir.[3]

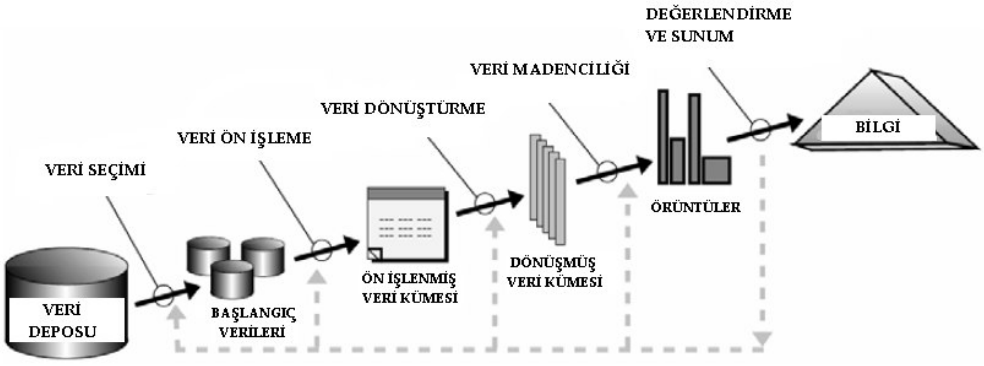
---

[1] Apte C. ve diğerleri, "Business Applications of Data Mining", **Communications of the ACM**, Vol 45, No 8, 49-53, 2002

[2] Han J., Kamber M., **Data Mining : Concepts and Techniques**, Academic Press, 2001, sf.4

[3] Bayraktar R., "Veri Tabanı ve Akılcı Düşünce Üzerine", *http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=127*, [Erişim 14.03.2006]

Günümüzde bir çok işletme, müşteri yaşam çemberinin yeni müşterilerin kazanılması, mevcut müşterilerden kazancın arttırılması ve iyi müşterilerin elde tutulması gibi aşamalarında veri madenciliğini kullanmaktadır. Karlı müşterilerin niteliklerinin belirlenmesi yolu ile benzer nitelikteki müşteri grubu hedeflenmekte, bir ürünü satın alan müşteri tipi yolu ile benzer müşteri grubuna yönelmek mümkün olmakta veya kaybedilen müşteri tipinden hareketle kaybedilme olasılığı güçlü müşterilerin elde tutulması sağlanmaktadır. Çünkü yeni müşterinin kazanılması mevcut müşterinin elde tutulmasından çok daha fazla maliyete neden olmaktadır.



Şekil 1.1. Veri Tabanlarında Bilgi Keşfi Süreci [4]

Teknolojik iletişimin hızla yaygınlaşması ve temelde internet kullanımının ticari uygulamalara getirdiği yeniliklerle ortaya çıkan e-ticaret kavramı dahi biçim değiştirerek artık yerini e-iş kavramına bırakır durumdadır. E-ticaret denildiğinde internet üzerinde gerçekleştirilen ticari işlemler akla gelirken e-iş daha geniş bir anlam ifade etmektedir. E-iş kavramı, eğitim, eğlence, web sayfaları oluşturma ve destekleme, tedarik sağlama, tedarik zinciri yönetimi, müşteri yardım hizmetleri gibi her türlü internet tabanlı ticari faaliyeti kapsamaktadır.[5]

Özellikle 1980'li yıllarda ortaya çıkan ve karar süreçlerinde veri seçimi, veri ön işleme, veri dönüştürme, veri madenciliği, değerlendirme ve sunum basamaklarının

[4] Abonji J., Feil B., **Cluster Analysis for Data Mining and System Identification**, BirkHauser Verlag AG, 2007, sf.X
[5] Clifton C., Thuraisingham B., "Emerging Standards for Data Mining", **Computer Standards & Interfaces**, No.23, 187-193, 2001

tamamını kapsayan süreç "veri tabanlarında bilgi keşfi" adı ile tanımlanmaktaydı. Ancak günümüzde bu sürecin tamamı "veri madenciliği" olarak adlandırılmakta ve "veri tabanlarında bilgi keşfi" adlandırması gittikçe daha az kullanılmaktadır.

Veri madenciliği, tüm dünyada bilinen adı ile *data mining*, sözcük seçimi açısından bakıldığında verilerin içinde saklı bulunan bir şeyler olduğu ya da olabileceği ve bunun verilerin "eşelenmesi" yoluyla ortaya çıkarılabileceği anlamını çağrıştırmaktadır. Konu kapsamı açısından herkesin üzerinde uzlaştığı tek bir ifade bulunmamakla birlikte, tercih edilişleri ve tarihsel sıralaması açısından farklı kaynaklarda en fazla öne çıkan tanımları aşağıdaki gibidir.

Veri madenciliği, veri içerisinde gizli kalmış, önceden bilinemeyen ve potansiyel olarak kullanışlı olan anlamlı bilginin çıkarımıdır.[6]

Veri madenciliği, içerisinde varolan anlamlı örüntü ve kuralları ortaya çıkarmak amacıyla, büyük miktarlardaki verinin otomatik ve yarı otomatik araçlar yardımıyla incelenmesi ve analiz edilmesi sürecidir.[7]

Veri madenciliği, çeşitli mimarilerde depolanmış olan büyük miktarlardaki verilerden ilgi çekici bilginin keşfedilmesi sürecidir.[8]

Veri madenciliği, veri ambarında tutulan çok çeşitli verilere dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarma ve bu bilgileri, karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir.[9]

Veri madenciliği, veriye sahip olan kişi ya da kurum için, kuralları ve ilişkileri keşfederek, önceden bilinemeyen açık ve yararlı sonuçlar elde etmek amacıyla, çok miktardaki verinin seçilmesi, incelenmesi ve modellenmesi sürecidir.[10]

---

[6] Flawley W.J., Piatetsky-Shapiro G., Matheus C.J., "Knowledge Discovery in Databases : An Overview", **AI Magazine**, Vol. 13, No. 3, 57-70, 1992
[7] Berry M.J.A., Linoff G.S., **Mastering Data Mining : The Art and Science of CRM**, John Wiley & Sons, 2000, sf.7
[8] Han J., Kamber M., *a.g.e.*, sf.7
[9] Swift R., **Accelerating Customer Relationship**, Prentice Hall PTR, 2001, sf.93
[10] Giudici P., **Applied Data Mining: Statistical Methods for Business and Industry**, John Wiley & Sons, 2003, sf.2

Yukarıda verilen en eski tanıma zaman içerisinde "otomatik ve yarı otomatik araçlar", "çeşitli mimarilerde depolanmış", "veri ambarında tutulan", "kuralları ve ilişkileri keşfederek" gibi ifadelerin eklendiği görülmektedir. İlerleyen bölümlerde, tanımın tüm bu boyutları üzerinde durulacaktır.

Ancak tüm bu tanımlama çabaları incelendiğinde, "çok fazla miktarda veri" ve "anlamlı bilgi çıkarımı" ifadeleri dikkat çekici bir şekilde ön plana çıkmaktadır.[11]

Yine bu tanımlamalardan hareketle kolayca görülmektedir ki, veri madenciliği belirli bir teknik ya da algoritma olmaktan çok bir süreç ifade etmektedir. Bu süreç, verilerin elde edilmesi ile başlayan ve bulguların gerçek yaşama aktarılmasına kadar süren tüm faaliyetleri kapsamaktadır.[12]

Veri madenciliğinin en önemli üstünlüğü çeşitli disiplinleri bir araya getirerek verilerin analizini gerçekleştirmesi ve işletme stratejileri ya da çeşitli araştırmalara katkıda bulunacak şekilde önemli veri örüntülerini ortaya çıkarabilmesidir. Bu sayede, üst üste biriken ve faydasız bir gömüt halini alan verilerden altın değerindeki bilginin üretilmesi mümkün olmaktadır.

Ancak bu bakış açısı, yaygın olarak bazı yanlış algılamaları da beraberinde getirmiş ve veri madenciliğinin olduğundan daha etkin bir araç olarak görülmesine neden olmuştur. Veri madenciliğini bir yöntem olarak seçen tüm uygulamacıların veri madenciliğini doğru bir şekilde anlaması ve beklentilerini bu doğrultuda belirlemesi son derece önemlidir.

Veri madenciliği uygulamalarının başarılı olabilmesi için uygulamacının mutlaka dikkate alması gereken bir takım yaklaşımlar söz konusudur. Bu yaklaşımları aşağıdaki gibi özetleyebiliriz.[13]

---

[11] Özmen Ş., "İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor", *V. Ulusal Ekonometri ve İstatistik Sempozyumu*, Adana, 2001

[12] Edelstein H., "Data Mining in Depth: Description is not Prediction", Mart 2003, *http://www.dmreview.com/issues/20030301/6388-1.html*, [Erişim 08.05.2005]

[13] Edelstein H., "Data Mining in Depth: Using Data Mining to Find Terrorists", Mayıs 2003, *http://www.dmreview.com/issues/20030501/6655-1.html*, [Erişim 08.05.2005]

Öncelikle, veri madenciliği araçları problemlerimize kendiliğinden çözüm üretmezler. Veri madenciliği her aşamasında özenli çalışma gerektiren ayrıntılı bir süreçtir.

İkinci olarak, veri madenciliğinin her adımında insan zekası ve bilgi birikiminin sürece dahil olması gerekir. İnsan deneyimi ve bilgi birikiminin dahil olmadığı bir veri madenciliği uygulaması düşünülemez. Bu süreçte insan verilerden daha önemlidir.

Üçüncü olarak, veri madenciliği projesinin ürettiği fayda zaman zaman kendi maliyetlerini dahi karşılamayabilir. Bu nedenle, çalışmadan beklenilen fayda ile göze alınan maliyetin önceden ve titizlikle belirlenmesi gerekir.

Ayrıca veri madenciliği yazılımları kullanıcı odaklı tasarlanmış da olsa hem analitik zeka hem de yapılan iş ve uygulanan teknikler açısından bilgi birikimi gerektirir. Veri madenciliğinin kolay uygulanabilir olması herkes tarafından uygulanabileceği anlamına gelmez.

Bununla birlikte, veri madenciliği teknikleri davranış kalıplarını ortaya koysa da nedenleri ile ilgilenmezler. Sadece verilerde gizli bulunan ilişkileri ortaya çıkarırlar. Bu ilişkilerin nedenlerinin ortaya çıkarılması, söz konusu işin uzmanları tarafından gerçekleştirilebilir.

Ve son olarak, veri madenciliğinin en önemli adımı tüm diğer veri analiz tekniklerinde olduğu gibi yine veri hazırlama aşamasıdır. Bu da verilerin madencilikten daha önemli olduğuna işaret eder. Hiçbir veri madenciliği uygulaması kalitesiz verilerle başarılı olarak gerçeklenemez.

Elbette tüm bu yaklaşımları sıralayarak veri madenciliğini kullanmamamız gerektiğini düşünmemekteyim. Aksine veri madenciliğinin araştırmalarda insanın yerini alan değil insanın destekçisi olan doğasını fazlasıyla önemsediğimi belirtmeliyim.

Ancak başarılı veri madenciliği için insanın önemi çok fazla da olsa uygulamanın kendisi için olmazsa olmaz bir diğer koşul ise veri madenciliği araçlarıdır. Veri madenciliği projeleri tamamen otomatik araçlar yoluyla

yürütüldüğünden, belirli bir projede kullanılacak yazılımın seçilmesi de önem kazanmaktadır. Bir çok yazılım, aynı adlandırmaya ancak farklı uygulamaya sahip komutları içermektedir. Bununla birlikte hafıza kullanımı, veri depolama mimarisi, hız ve geçerlilik gibi karakteristikleri açısından da yazılım seçimi son derece önemlidir.[14]

En yaygın kullanımlı veri madenciliği yazılımları arasında SPSS Clementine, SAS Enterprise Miner, IBM Intelligent Miner for Data, DBMiner, Statistica Data Miner ve XLMiner gibi paket yazılımlar ile Weka, Yale ve Orange gibi özgür yazılımlar yer almaktadır.

1.1.1. Veri Madenciliğinin Kökenleri

Veri madenciliğinin yeni bir çalışma alanı olduğu ne kadar doğru ise, yeni bir buluş olduğunu düşünmek de o derecede yanlıştır. Çünkü veri madenciliği sanılanın aksine, çeşitli disiplinleri bir araya getiren bir açılım, yeni bir paradigmadır.[15]

Veri madenciliği, gerçekte çok uzun yıllara dayanan bir evrimin sonucu olmakla birlikte, "veri tabanlarında bilgi keşfi" çalışmalarının devamı olarak bugün kullanılan anlamıyla tam karşılığını 1990'lı yıllarda bulmuştur. Bilimsel çalışmaların paylaşılması ve tartışılmasına olanak tanıyan ilk uluslararası veri madenciliği konferansı* 1995 yılında gerçekleştirilmiş, ilk süreli yayın** 1997 yılında yayımlanmaya başlanmıştır.

Günümüzde Kuzey Amerika'da gerçekleştirilen Knowledge Discovery in Databases, Avrupa'da gerçekleştirilen Principle of Knowledge Discovery in Databases ve Avustrasya ülkelerinde gerçekleştirilen Pasific Asia Knowledge Discovery and Data Mining konferansları veri madenciliği alanında gerçekleştirilen çalışmaların sunulduğu ve tartışıldığı başlıca platformlardır.

Veri madenciliğinin birden çok disiplini bir araya getiren bir çalışma alanı olduğu düşünülürse, köklerini biraz daha genişletmek hatta yüzlerce yıl önceye kadar

---

[14] Two Crows Co., "Introduction to Data Mining and Knowledge Discovery", *http://www.twocrows.com/intro-dm.pdf*, [Erişim 08.05.2005]
[15] Lovell M.C., "Data Mining", **The Review of Economics and Statistics,** Vol.65, No1, 1-12, 1983
* The First International Conference on Knowledge Discovery and Data Mining, Montreal, QC
** Journal of Data Mining and Knowledge Discovery, Springer US

götürmek mümkündür. Çünkü veri madenciliği başta istatistik, yapay öğrenme ve veri tabanı teknolojileri olmak üzere bir çok disiplini içinde barındırmaktadır. Veri madenciliğinde istatistik geçerliliği, veri tabanı teknolojileri verimliliği ve yapay öğrenme ise faydalılığı ön plana çıkarmaktadır.[16]

İstatistiğin kökleri açısından matematiğe, yapay öğrenmenin ise bilgi–işlem teknolojileri ve yapay zeka çalışmalarına dayandığı düşünülürse, veri madenciliğinin tarihsel gelişimi daha iyi anlaşılmış olur.

Bu noktada veri madenciliğinin iki farklı yaklaşımı içerdiğini görmekteyiz. Çünkü istatistik öncelikle teorik çerçeveyi karşılamaya ve daha sonra pratikte uygulamaya vurgu yaparken, yapay öğrenme herhangi bir teorik kanıt aramaksızın öncelikle pratik uygulamadaki verimliliği dikkate almaktadır. Sonuç olarak istatistik modellemeyi, yapay öğrenme ise algoritmayı ön plana çıkarmaktadır.[17]

İstatistiksel veri analizi, üstün teori ve analitik yöntemlerle gerçekleştirilse de bazı noktalarda günümüz ihtiyaçlarına cevap verme konusunda yetersiz kalmaktadır. Bunun başlıca sebebi, standart istatistik yöntemlerin nominal ve yapılandırılmış veri türleri barındıran veri tabanlarında sorun yaşamasıdır. İkincisi ise istatistiğin alan bilgisinin önüne geçecek şekilde tamamen verilere bağımlı olarak gerçeklenmesidir. Ayrıca, istatistiksel analiz sonuçlarının yorumlanması oldukça güç ve göz korkutucu olabilmektedir. Ve son olarak, istatistik yöntemleri verilerin nerede ve ne şekilde analiz edileceği konusunda uygulamacının güçlü rehberliğine ihtiyaç duymaktadırlar.[18]

Bununla birlikte istatistik alanındaki gelişme daha çok mütevazı miktarlardaki veri kümeleri üzerinde gerçekleşmiştir. Hızla gelişen ve maliyetleri azalan veri teknolojileri ile birlikte ortaya çıkan çok daha fazla miktarlardaki verinin analizi ihtiyacı ise yeni çözümlerin keşfedilmesine olanak sağlamıştır. Buradaki kritik nokta, veri madenciliğinin uygulanmada profesyonel istatistikçilere ihtiyaç duymaksızın birikimli çalışanlara da fırsat tanıyan bir yapıya sahip olması ve bu niteliği ile bireylerin üretkenliklerinin artmasına yardımcı olmasıdır.

---

[16] Zhou Z., "Three Perspectives of Data Mining", **Artificial Intelligence**, No.143, 139-146, 2003
[17] Kantardzic M., **Data Mining: Consepts, Models, Methods and Algorithms**, IEEE, 2003, sf.4
[18] Flawley W.J., Piatetsky-Shapiro G., Matheus C.J., *a.g.e.*

Ancak istatistiksel veri analizi teknikleri uzun zaman önce geliştirilmiş olsa da ileri düzey zeki veri analizi teknikleri henüz tam olgunlaşmamış ve evrimini sürdürmektedir. Bu nedenle, günümüzde üretilen veri miktarı ile verilerin anlaşılma düzeyleri arasında hızla büyüyen bir boşluk oluşmaktadır. Aynı zamanda, verilerin işlenmesi ile rekabet avantajı sağlayacak değerli bilginin elde edilebileceği beklentisi ve bu yöndeki uygulamalar da gelişmektedir. Bu açıdan, gelecek dönem bilgi keşfi eğilimlerin ne yönde olacağı önemli bir soru işaretidir. Bu konudaki beklentileri şu şekilde özetleyebiliriz.[19]

Uzmanlığın bilgi keşfinin her aşamasında önem taşımaya devam edeceği konusunda kuşkuya yer yoktur. Ayrıca verimli algoritmaların kullanımının önem kazanacağı, kendi keşfettiği bilgiyi yeniden kullanabilen artımlı (incremental) yöntemlere daha çok ihtiyaç duyulacağı düşünülmektedir. İnsan ve makine birlikteliği için en iyi yaklaşım olan etkileşimli sistemlerin yakın dönemde bilgi keşfi için büyük fırsatlar sunması beklenmektedir. Geleceğin en önemli ihtiyaçlarından biri ise bilgi keşfinin çeşitli seviyelerinde entegrasyonun gerçekleştirilmesi olarak görülmektedir.

Bir başka deyişle, gelecek dönem veri madenciliğinde özdevimli, ölçeklenebilir ve güvenilir tekniklerin geliştirilmesi ile işe yönelik olarak karmaşık tekniklerle elde edilen bilginin kolayca ve hızlı bir şekilde çözümlenmesi ihtiyacı ön plana çıkacaktır.

Kanımca veri madenciliği, bugünkü adlandırması ya da süreç modeli ile olmasa bile, temel hedefleri açısından insanoğlu varoldukça hem varlığını sürdürecek hem de mevcut teknolojik gelişme süreci devam ettikçe önemini giderek artıracaktır. Çünkü veri birikiminin artışı sürdükçe bu birikimin bilgiye dönüştürülmesi ihtiyacının da giderek artacağını söyleyebiliriz.

1.1.2. Veri Tabanı Teknolojilerinin Gelişimi

Verilerin elde edilmesindeki gelişme ile birlikte depo edilmesine yönelik çalışmalar da hız kazanmış ve basit veri dosyalarından gelişmiş veri tabanı mimarilerine kadar geniş bir çeşitlilik ortaya çıkmıştır. Bu açıdan, veri madenciliğinin prensipte farklı veri tabanı mimarileri üzerinde uygulanması söz konusudur.

[19] Flawley W.J., Piatetsky-Shapiro G., Matheus C.J., *a.g.e.*

Basit tanımıyla veri tabanı, verilerin depolanmasını, değiştirilmesini ve çağrılmasını kolaylaştırmak üzere organize edilmiş ve mantıksal olarak bir ya da daha fazla dosya halinde bir araya getirilmiş veri topluluğunu ifade etmektedir.[20]

Günümüzde yaygın olarak kullanılan veri tabanı mimarileri arasında ilişkisel veritabanları, işlemsel veritabanları, veri ambarları, yatay kütükler (flat files) ve www alanları sayılabilmektedir. Ayrıca gelişmiş veritabanı sistemleri olarak adlandırılan nesne merkezli (object – oriented) veritabanları, nesne ilişkili (object – relational) veritabanları ve özel uygulama merkezli (specific application – oriented) veritabanları bu listeye eklenmelidir. Özel uygulama merkezli veritabanları kapsamında uzaysal (spatial) veritabanları, zaman serileri veritabanları, metin veritabanları ve çoklu ortam veritabanları bulunmaktadır.[21]

Aynı zamanda bir veri tabanı üreticisi de olan günümüzün yazılım devi Microsoft, bir veri tabanının sunması gereken temel özellikleri ölçeklenebilir ve esnek olması, güvenilir olması, yüksek performanslı ve yönetilebilir olması olarak sıralanmaktadır.[22]

Ölçeklenebilirlik ve esneklik, veri tabanının gerek bellek gerekse işlemci bazında daha yüksek kapasite ihtiyaçlarını karşılayabilirliği anlamına gelmektedir. Verilerin elde edilmesinde her geçen gün artan bir büyüme söz konusu iken bu büyümeyi karşılayamayan veri tabanı sistemlerinin yaratacağı sorunlarla başa çıkmak hem etkinlik hem de maliyet açısından önemli bir sorun oluşturacaktır.

Güvenilirlik, yüksek teknolojinin her aşamasında önem taşıyan bir sorun olarak görülmektedir. Veri tabanları açısından güvenilirlik, hem ticari hem askeri hem de kamusal alanlarda önem taşımaktadır. Güvenilirlik sorunu, verilerin gizliliğinin korunamaması ya da veri tabanlarının bilgi keşfi sürecinde sağlıklı bir şekilde kullanılamamasına yol açabilmektedir.

---

[20] Flawley W.J., Piatetsky-Shapiro G., Matheus C.J., *a.g.e.*
[21] Han J., Kamber M., *a.g.e.*, sf 10-21
[22] Microsoft Türkiye, "Bir Veri Tabanının Sunması Gereken Temel Özellikler", *http://www.microsoft.com/turkiye/.../.../sqlserver/veritabani.asp*, [Erişim 16.10.2005]

Verilerin sadece depolanmasının yeterli olmayacağı, bilgi keşfi sürecinde verilere erişimin ve veri tabanının yönetilmesinin de gereklilik olduğu açık bir gerçektir. Bu açıdan veri işleme performansı ve yönetilebilirlik, veri tabanları için en önemli başarı kriterleri sayılmaktadır.

Bu noktada, veri madenciliği uygulamalarının başarısı açısından, veri tabanlarının barındırabileceği bir takım sorunları dile getirmekte de fayda görmekteyim. Veri tabanlarında karşılaşılabilecek başlıca sorunlar, dinamik veri yapısı, geçersiz veri alanları, kayıp veri alanları, kayıp veri değerleri, gürültü ve belirsizlik olarak sıralanmaktadır.[23]

Dinamik veri yapısı, verilerin zaman duyarlı olmasına rağmen gözlemlerin zamandan bağımsız olarak elde edilmesinden kaynaklanan bir sorunu ifade eder. Bu sorunun ortadan kaldırılamadığı durumlarda bilgi keşfi sürecinin olumsuz etkilenecektir.

Geçersiz veri alanları, veri kümesinde yer alan veri alanlarının belli bir kısmının bilgi keşfi hedefine uygun düşmemesinden kaynaklanan bir sorundur. Bu nedenle, veri hazırlama sürecinde bilgi keşfi hedefine uygun alanların seçilmesi önem kazanacaktır.

Ayrıca bazı durumlarda bilgi keşfi süreci açısından önem taşıyan veri alanları mevcut veri tabanında yer almayabilir. Bu durum kayıp veri alanları sorunu olarak adlandırılmakta ve ortadan kaldırılması bu veri alanlarına ait verilerin elde edilmesi ile mümkün olmaktadır. Eğer bu alanlara ait verilerin elde edilmesi olanaksız ise bilgi keşfi süreci tam anlamıyla başarılı olamayacaktır.

Bilgi keşfi hedefine uygun, bir başka deyişle geçerli veri alanlarında kayıp veri değerlerinin yer alması da önemli bir sorundur. Bir etkileşimli sistemde kayıp veri değerleri yerine atama yapılabileceği gibi bu değerlerin bilgi keşfi sürecini olumsuz etkilemeyecek şekilde yansız değerler atanarak ortadan kaldırılması da mümkündür. Ancak kayıp veri değerlerinin bilgi keşfi süreci açısından sanki özel bir davranışmış gibi algılanması kesinlikle kaçınılması gereken bir durumdur.

[23] Flawley W.J., Piatetsky-Shapiro G., Matheus C.J., *a.g.e.*

Gürültü ve belirsizlik sorunları ise daha çok verilerin doğasından kaynaklanmaktadır. Geçerli veri alanlarında belirsizlikten kaynaklanan hatanın şiddeti, ilgili alandaki veri türünün izin verilen değerlerine bağlı olarak gerçekleşmektedir. Verilerin sürekli veya tamsayı türünde nicel, isim veya sıralı türde nitel olmasına göre barındıracağı belirsizlik de farklılık göstermektedir. Belirsizliğin farklı bir türü olan gürültü de beklendik bir sorun olarak düşünülmelidir.

Bir veri tabanı sistemi, ya da veri tabanı yönetimi sistemi (DBMS), veri tabanını oluşturan ve karşılıklı ilişki halindeki verilerin bir topluluğu ile, verilere erişimi ve verilerin yönetilmesini sağlayan bir grup yazılımı içermektedir.[24]

Veri tabanı yönetimi sisteminin önemli bir parçası olan veri tabanı yazılımlarına Microsoft Access, dBase, FoxPro, Paradox ve hatta Microsoft Excel gibi kişisel veri tabanı yazılımlarını veya Oracle, DB2, Sysbase, Informix, Progress, MS SQL Server gibi ilişkisel veri tabanı yazılımlarını ve veri ambarları üzerinde karmaşık analizlere imkan veren OLAP (online analytical processing) çözümlerini örnek vermek mümkündür.[25]

Geleneksel sorgulama ve raporlama araçları ile OLAP arasındaki en temel farklılık, OLAP çözümlerinin veri tabanında ortaya çıkan örüntülerin nedenlerini de açıklayabilmesidir. OLAP uzmanı hipoteze dayalı bir dizi örüntü veya ilişki üretmekte ve bunları kanıtlamak veya reddetmek için veri tabanı üzerinde sorguları kullanmaktadır. Bu nedenle OLAP analizi veri madenciliği anlamına gelmemektedir. Çünkü OLAP analizi tümdengelim, veri madenciliği ise tümevarım süreci ifade eder.[26]

Veri ambarları, veri madenciliği söz konusu olduğunda akla gelen en önemli veri tabanı mimarisi olup, yönetim kararları verme sürecini kolaylaştırmak amacıyla çok sayıda heterojen veri kaynağının tekbiçimleştirilerek ortak bir alan altında organize edilmesiyle oluşturulan büyük bir veri deposu olarak tanımlanmaktadır.[27]

---

[24] Han J., Kamber M., *a.g.e.*, sf.10
[25] Bayraktar R., *a.g.e.*
[26] Alpaydın E., "Zeki Veri Madenciliği : Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", *Bilişim 2000 Eğitim Semineri*, İstanbul, 2000
[27] Man D., "Answering Some Common Data Warehousing Questions", **Direct Marketing**, Vol.59, No.8, 12-15, 1996

Bir başka ifade ile veri ambarları, bir kurumda gerçekleşen tüm operasyonel işlemlerin en alt düzeydeki verilerine kadar inebilen, etkili analiz yapılabilmesi için özel olarak modellenen, tarihsel derinliği olan, fiziksel olarak operasyonel verilerden farklı ortamdaki yapılardır.

1.2. VERİ MADENCİLİĞİ SÜRECİ

Veri madenciliğinin sadece belirli tekniklerin uygulanması olarak görülmemesi gerektiğini, gerçekte ham verinin elde edilmesinden anlamlı bilginin çıkarımına kadar bütün bir süreci ifade ettiğini daha önce belirtmiştik. Veri madenciliği, işe ilişkin ihtiyacın ortaya çıkmasından, oluşturulacak modelin gerçek yaşama aktarılmasına kadar yürütülecek tüm faaliyetleri içeren bir süreç olarak düşünülmelidir.[28]

Veri madenciliği sürecinin bir diğer önemli boyutu ise yinelemeli bir süreç olmasıdır. Veri madenciliği uygulamalarında genellikle birden çok model oluşturulmakta ve en uygun modelin elde edilmesi hedeflenmektedir. Eğer elde edilen en uygun model de yeterli görülmez ise süreç yeni kayıtların veri setine dahil edilmesi ya da kayıtlardaki sorunların giderilmesi yoluyla yeniden gerçekleştirilir.[29]

Veri madenciliği süreci, temel bir takım ortak faaliyetleri kapsamakla birlikte, başlangıçta her uygulamacı tarafından farklı şekilde yürütülmekteydi. Bu nedenle, uygulamada ortaya çıkan farklılıkları en aza indirecek bir standart süreç geliştirilmesi ihtiyacı ortaya çıkmıştır. Bu ihtiyaç doğrultusunda, veri madenciliği uygulamalarının dört önemli lideri* ile iki yüzden fazla araştırmacıyı bir araya getiren CRISP–DM** konsorsiyumu, düzenlediği çeşitli çalıştaylarla veri madenciliği uygulamalarının olgunluk kazanması ve tüm uygulamacılara yol gösterecek bir standart süreç modelinin oluşturulması için 1996 yılının sonlarında çalışmalarına başlamıştır.[30]

---

[28] CRISP – DM Konsorsiyumu, "CRISP – DM 1.0 Step – by – Step Data Mining Guide", *http://www.crisp-dm.org/CRISPWP-0800.pdf*, [Erişim 21.01.2005]
[29] Fayyad U., Uthurusamy R., "Data Mining and Knowledge Discovery in Databases", **Communications of the ACM**, Vol.39, No.11, 24-26, 1996
* DaimlerChrysler, SPSS, NCR Systems Engineering, OHRA Verzekeringen en Bank Groep
** CRoss – Industry Standard Process for Data Mining
[30] Shearer C., "The CRISP-DM Model: The New Blueprint for Data Mining", **Journal of Data Warehousing**, Vol.5, No.4, 13-22, 2000

Bu konsorsiyum, bir yıl sonra Avrupa Komisyonu'nun sağladığı fon ile çalışmalarını hızlandırmış ve 1990'lı yılların sonuna gelindiğinde çalışma taslağı büyük oranda oluşturulmuştur. Güncellenmiş şekli ile CRISP–DM 1.0[31] dokümanı ise 2000 yılında yayımlanmıştır. CRISP–DM 1.0, sadece akademik bir çalışma olmayıp veri madenciliği konusunda pratik, gerçek yaşam deneyimleri üzerine kurgulanmış bir süreç modelidir. Bu süreç modeli, temel hedefine ulaşmış olmakla birlikte nihai nitelik taşımamakta ve güncelleştirme ihtiyaçları doğrultusunda CRISP–DM 2.0 üzerindeki çalışmalar yine CRISP–DM SIG tarafından sürdürülmektedir.

Günümüzde veri madenciliği süreci dışında veri madenciliği araçları ve mimarisine dönük standartlaştırma çabaları da tüm güçlüğüne rağmen devam etmektedir.

1.2.1. Veri Madenciliği Standart Süreci

Veri madenciliği standart sürecinin ilk versiyonu olan CRISP – DM 1.0, veri madenciliği uygulamalarında izlenmesi gereken aşamaları standartlaştıran bir kılavuz niteliğindedir. Çalışmamın bu alt bölümünde, veri madenciliği sürecinin aşamalarına ilişkin olarak aktarılacak bilginin temel çatısını da bu dokümanın içeriği oluşturacaktır.

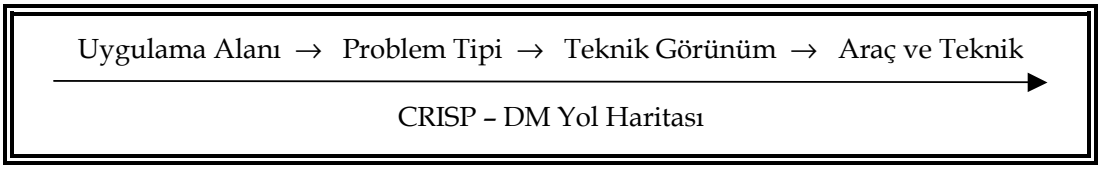
CRISP–DM veri madenciliği metodolojisi (Şekil 1.2), genelden özele olmak üzere dört farklı soyutlama seviyesinden oluşan hiyerarşik bir yapıya sahiptir. Bu hiyerarşide altı farklı aşama, her bir aşamaya özgü genel görevler, tüm genel görevlere ilişkin bir takım özel görevler ve son olarak bu özel görevlerdeki somut faaliyetleri tanımlayan süreç örnekleri yer almaktadır.

Bu hiyerarşinin en üst seviyesinde veri madenciliği sürecinin ana faaliyetlerini ifade eden işi anlama, verileri anlama, verileri hazırlama, model oluşturma, değerlendirme ve konuşlandırma aşamaları yer almaktadır. İkinci seviyede, ilgili aşamada yapılması gereken faaliyetleri sıralayan genel görev, üçüncü seviyede ise genel görevlerin belirli durumlarda nasıl yürütüleceğini tanımlayan özel görevler bulunmaktadır. Son seviyedeki süreç örneği, gerçek bir veri madenciliği uygulamasının sonuçlarını, kararlarını ve eylemlerini ifade etmektedir.

[31] CRISP – DM Konsorsiyumu, *a.g.e*

Şekil 1.2. CRISP–DM Metodolojisi

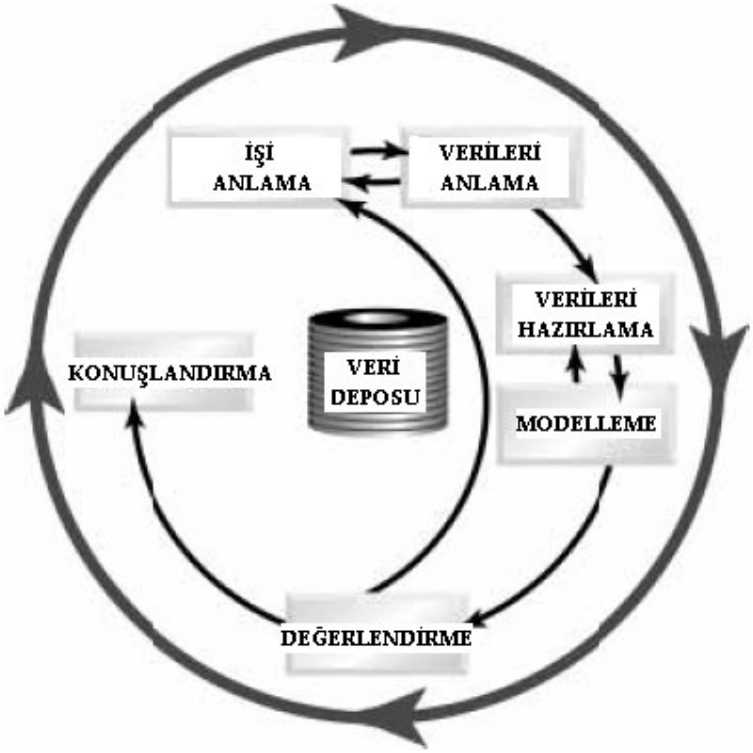
Veri madenciliği çerçevesi (Şekil 1.3) olarak ifade edilen yol haritası ise genelden özele olmak üzere dört farklı seviyede değerlendirilmektedir. En üst seviyede yer alan uygulama alanı veri madenciliği projesinin yürütüleceği alanını, ikinci seviyedeki problem tipi ise projenin ilişkilendirildiği veri madenciliği fonksiyonunu vurgulamaktadır. Üçüncü seviyedeki teknik görünüm, veri madenciliği sırasında genellikle karşılaşılan çeşitli teknik güçlükleri tanımlamaktadır. Son seviyedeki araç ve teknik ise kullanılan yazılımları ve uygulanan teknikleri içerir.



Şekil 1.3. Veri Madenciliği Çerçevesi

CRISP–DM modeli, veri madenciliği sürecini altı temel aşamadan oluşan bir yaşam çemberi (Şekil 1.4) olarak betimlemektedir. Bu betimleme, veri madenciliği projelerinin aşamaları ve sürecin yinelemeli doğasının görsel bir ifadesi gibidir. Model, her bir aşamanın tamamlanması ile yeni bir aşamaya geçilmesini ya da gerektiğinde önceki aşamalara dönülmesini önermektedir. Veri madenciliği projesinin tamamlanması dahi bir son anlamına gelmemekte, süreç boyunca ve süreç sonunda alınan derslerle, yeni ve sıklıkla işe yönelik yeni soruların üretilmesi mümkün olmaktadır.

Görüldüğü gibi, süreç işi anlama aşaması ile başlamakta ve ardışık yinelemeli verileri anlama, verileri hazırlama, modelleme, değerlendirme ve konuşlandırma aşamalarından oluşmaktadır.



Şekil 1.4. Veri Madenciliği Yaşam Çemberi

1.2.2. İşi Anlama

Veri madenciliği projesinin belki de en önemli aşaması proje hedeflerini iş perspektifi ile anlamayı, bu bilgiyi veri madenciliği problem tanımına dönüştürmeyi ve ilgili hedeflere ulaşmak için bir proje planı oluşturmayı kapsayan işi anlama aşamasıdır. İşi anlama, bir başlangıç aşaması olarak, iş perspektifi bakımından projenin hedef ve ihtiyaçlarını anlamaya odaklanmayı ve buradan edinilen bilgi ile veri madenciliğinin problem tanımını oluşturarak, bir önsel plana dönüştürmeyi kapsamaktadır. Bu aşama ile ilgili temel görevler ve ilgili çıktıları aşağıdaki tabloda (Tablo 1.1) görülmektedir.

İş hedeflerinin belirlenmesi görevi, veri madenciliği süreci sonunda tam olarak ne elde edilmek istendiğinin belirlenmesi anlamına gelmektedir. Bu görevdeki başarı, proje sonunda yanlış sorulara doğru cevaplar bulmak gibi bir başarısızlıkla karşılaşmamak açısından son derece önemlidir. Bu görev doğrultusunda, veriye sahip

olan organizasyonun iş durumu hakkında bilgi edinilecek, ilgili organizasyonun bu projeden temel beklentisinin ne olacağının doğru olarak tespit edilmesi mümkün olacak ve projenin başarı kriteri belirlenecektir. Projenin başarı kriteri, genel ve öznel olabileceği gibi, özel ve nesnel de olabilir.

Tablo 1.1. İşi Anlama Aşaması

Aşama	İşi Anlama
Görev	İş Hedeflerini Belirlemek
Çıktı	İş Geçmişi
Çıktı	İş Hedefleri
Çıktı	İş Başarı Kriteri
Görev	Durumu Değerlemek
Çıktı	Kaynakların Envanteri
Çıktı	Gereklilikler, Varsayımlar ve Kısıtlar
Çıktı	Riskler ve Çözüm Alternatifleri
Çıktı	Terminoloji
Çıktı	Maliyetler ve Faydalar
Görev	Veri Madenciliği Başarısını Belirlemek
Çıktı	Veri Madenciliği Amaçları
Çıktı	Veri Madenciliği Başarı Kriteri
Görev	Proje Planını Hazırlamak
Çıktı	Proje Planı
Çıktı	Araç ve Tekniklerin Başlangıç Değerlemesi

Durum değerlemesi görevi, veri madenciliği başarısı ve proje planı oluşturulmasında akla gelebilecek tüm yeterliliklerin, sınırlamaların, varsayımların ayrıntılarının ortaya konmasını sağlamaktadır. Bu görev doğrultusunda, projeye katılabilecek personel olanakları, veri ihtiyaçları, donanım ve yazılım olanakları ortaya çıkmış olacak, bir iş takvimi oluşturulacak, veri güvenliğine ilişkin gereklilikler ile başta teknolojik olmak üzere tüm sınırlamaların ortaya konması sağlanacaktır.

Yine bu görev sırasında, projenin gecikmesine ya da başarısız olmasına neden olabilecek tehditler ile bu tehditlerin gerçekleşmesi halinde yürütülecek alternatif plan da oluşturulacaktır. Ayrıca, projeye ilişkin terimler sözlüğünün oluşturulması önerilmektedir. Ve son olarak bu görev neticesinde, projenin bir fayda maliyet analizi

yapılmış olacak ve üretilecek faydanın maliyeti karşılama oranı mümkün olduğunda net olarak belirlenecektir.

Veri madenciliği başarısını belirleme görevi, iş hedefleri açısından veri madenciliği uygulaması ile neye ulaşılmak istendiğinin ve ulaşılacak sonuçların değerlendirilmesinde öznel ya da nesnel anlamda projenin başarı kriterinin ne olacağının ortaya konmasını ifade eder.

Bu noktada, başarı kriteri kavramını biraz açmakta fayda görmekteyim. Projenin başarısını, iş hedefleri ve veri madenciliği hedefleri açısından ayrı ayrı düşünmek gerekmektedir. İş hedefi, işe sahip olan kişi ya da kurumun bu projeden beklentilerini, veri madenciliği hedefi ise uygulamacının teknik olarak ulaşmak istediği beklentileri ifade etmektedir. Bu iki farklı hedef, proje kapsamında sıkı sıkıya bağlı olsa da bire bir örtüşmeyebilir.

Proje planını hazırlama görevi, hem veri madenciliği hem de iş hedeflerine yönelik olarak sezgisel anlamda projenin aşamalarının bir listesini ve veri madenciliği sürecinde söz konusu olabilecek olası süreç yinelemelerini kapsayan bir proje planının hazırlanmasını ifade etmektedir. Bu plan aynı zamanda, kullanılacak araç ve teknikleri de içermektedir. Proje planı, dinamik bir doküman olduğundan, her tamamlanan aşama ve görevden sonra yenilenmesi gerekecektir.

Başarılı veri madenciliği için anahtar niteliğindeki iki şeyden biri bu aşamada problemin hassas bir tanımlanması, diğeri ise bir sonraki aşamada değineceğimiz gibi doğru verilerin kullanılmasıdır.

1.2.3. Verileri Anlama

Verileri anlama, başlangıç verilerinin elde edilmesini, verileri tanımaya dönük analizleri ve verilerde saklı olabilecek bilgi için ilk izlenimlerin oluşturulmasını kapsayan bir aşamadır.

Bu aşama, verilerin elde edilmesinden başlayarak verinin genel yapısı hakkında fikir geliştirmeye, veri kalitesi problemlerini tespit etmeye ve hatta veri içerisinde ilginç olabilecek alt kümeleri tespit etmeye kadar bir çok çalışmayı kapsamaktadır. Bu

aşama ile ilgili temel görevler ve ilgili çıktıları aşağıdaki tabloda (Tablo 1.2) görülmektedir.

Tablo 1.2. Verileri Anlama Aşaması

Aşama	Verileri Anlama
Görev	Başlangıç Verilerini Elde Etmek
Çıktı	Başlangıç Verilerini Elde Etme Raporu
Görev	Verileri Tanımlamak
Çıktı	Veri Tanımlama Raporu
Görev	Verileri İncelemek
Çıktı	Veri İnceleme Raporu
Görev	Veri Kalitesini Doğrulamak
Çıktı	Veri Kalitesi Raporu

Başlangıç verilerini elde etme görevi, proje yeterlilikleri çerçevesinde tanımlanan verilerin elde edilmesi ve veri kümesinin tanımlayıcı nitelikteki ilk çalışmalarını yapabilmek amacıyla ilgili yazılıma yüklenmesi işlemlerini kapsamaktadır. Bu görev ile veri kümesinin listelenmesi, verinin elde edilmesi ve yüklenmesi sırasında yaşanan zorlukların özetlenmesi, projenin ilerleyen aşamaları açısından varsa bu zorluklara karşı gerçekleştirilmiş başarılı müdahalelerin belirtilmesi gibi bilgileri içeren bir raporun oluşturulması önerilmektedir.

Verileri tanımlama görevi, elde edilen veri kümesinin ilk incelemelerde öne çıkan yüzeysel özelliklerinin gözden geçirilmesini içermektedir. Bu görev sonunda, verilerin genel yapısı, alan ve kayıt sayıları, alanların özellikleri gibi yüzeysel bilgileri içeren ve ayrıca verinin ilgili proje açısından yeterliliğini tartışan bir raporun oluşturulması önerilmektedir.

Verileri inceleme görevinde, veri madenciliğinin veri sorgulama, veri görselleştirme ve raporlama gibi yöntemlerle açıklanabilecek sorularına cevap aranmaktadır. Bu görev, basit ilişki analizleri ve birleştirme işlemleri, temel alt ana kütle karakteristikleri ve basit istatistiksel analiz gibi faaliyetleri kapsamaktadır. Bu faaliyetlerin sonucunda, veri tanımlama ve veri kalitesi raporlarının oluşmasına ya da daha ileri analizleri besleyici çıkarımlara katkıda bulunacak bilginin üretilmesi

beklenmektedir. Bu görevin bir çıktısı olarak, tüm faaliyetleri özetleyen ve alınan dersleri içeren bir raporun oluşturulması önerilmektedir.

Gerçek yaşam verileri genellikle dinamik, eksik, gürültülü ve çok büyük miktarlarda olduğundan veri kalitesinin sağlanması hem uzun zaman hem de yoğun dikkat ve çaba gerektiren oldukça güç bir görevdir.[32]

Veri kalitesini doğrulama görevi ile tüm geçerli alanları kapsayan bir veri kümesine sahip olunup olunmadığı, veri kümesinin hata içerip içermediği, hataların ne derece yaygın olduğu, kayıp veri problemi bulunup bulunmadığı gibi veri kalitesini gözden geçirmeye dayanan sorulara yanıt aranmaktadır. Bu görev sonucunda, veri kalitesi sorunları ve bu sorunlara yönelik uygun çözümlerin bir listesini içerecek raporun oluşturulması önerilmektedir.

1.2.4. Verileri Hazırlama

Bu aşama, modelleme amacıyla kullanılacak olan veri kümesinin oluşturulmasına ilişkin tüm faaliyetleri içermektedir. Projenin tablo, değişken ve kayıt seçimi, birleştirme ve temizleme işlemleri gibi faaliyetleri bu aşamada gerçekleştirilmektedir.

Bu aşama, model geliştirme sürecinde kullanılacak olan veri kümesini oluşturmak için gerekli tüm çalışmaları kapsamaktadır. Bu aşama için önceden belirlenmiş bir kalıp bulunmamakta, mevcut veri kümesi üzerinde tablo, alan ve kayıt seçimlerinin yanı sıra modelleme araçları için veri temizlemesi ve veri dönüşümü gibi işlemler yürütülmektedir. Bu aşama sonucunda, modelleme ya da projenin temel analiz çalışması için kullanılacak veri kümesinin hem oluşturulması hem de tanımlanması mümkün olmaktadır. Bu aşama ile ilgili temel görevler ve ilgili çıktıları aşağıdaki tabloda (Tablo 1.3) görülmektedir.

Verileri seçme görevinde, veri madenciliği hedefi ve veri kalitesi açısından edinilen ilk bilgi ışığında, analiz aşamasında kullanılacak veri kümesinin seçilmesine karar verilir. Bu görev ile, veri kümesine yapılacak ekleme ve çıkarma işlemlerine

[32] Fayyad U., Piatetsky-Shapiro G., Smyth P., " The KDD Process for Extracting Useful Knowledge from Volumes of Data", **Communications of the ACM**, Vol.39, No.11, 27-34, 1996

ilişkin kararların nedenleri açıklanmalıdır. Bu işlemlerde, hem kayıtların hem de alanların seçilmesinin söz konusu olacağı unutulmamalıdır.

Tablo 1.3. Verileri Hazırlama Aşaması

Aşama	Verileri Hazırlama
Çıktı	Veri Kümesi
Çıktı	Veri Kümesi Tanımlaması
Görev	Verileri Seçmek
Çıktı	Veri Seçiminin Mantığı
Görev	Verileri Temizlemek
Çıktı	Veri Temizleme Raporu
Görev	Verileri Yapılandırmak
Çıktı	Türetilmiş Değişkenler
Çıktı	Üretilmiş Kayıtlar
Görev	Verileri Birleştirmek
Çıktı	Birleştirilmiş Veri Seti
Görev	Verileri Biçimlendirmek
Çıktı	Yeniden Biçimlendirilmiş Veri Seti

Verileri temizleme görevi ile veri kalitesinin seçilen analiz teknikleri için gereken seviyeye yükseltilmesi amaçlanmaktadır. Bu görev, verilerin daha yüksek kalitedeki alt kümelerini elde etmeyi ya da modelleme yoluyla kayıp verilerin kestirimi gibi ileri düzey teknikleri içerebilmektedir. Bu görev sonucunda, veri kalitesini doğrulama görevinde raporlanan veri kalitesi sorunlarını çözmeye yönelik ne tür kararlar alındığı ve ne tür eylemler gerçekleştirildiği tanımlanmalı ve raporlanmalıdır.

Verileri yapılandırma görevi, mevcut veri kümesi üzerinde bir ya da daha çok değişken üzerinden yeni değişkenlerin türetilmesi ve tümüyle yeni kayıtların ya da mevcut veri alanları için dönüştürülmüş değerlerin üretilmesi gibi geliştirici veri hazırlama operasyonlarını içermektedir.

Verileri birleştirme görevi, özellikle çoklu veri kaynaklarının bulunması durumunda, aynı obje için bilgi içeren birden çok tablonun bir araya getirilmesi ve tek bir tablo halinde toplanması, verilerin özetlenmesi yolu ile birden çok kaydın bütünleştirilerek yeni kayıtların elde edilmesi gibi faaliyetleri içermektedir.

Ancak her veri tabanı, kayıp veri değerleri ve doğru olmayan veri değerleri gibi olası sorunları içerecektir. Verilerin bir çok kaynaktan birleştirilmesi, aynı terimin farklı anlamlar içermesi, aynı girdi için farklı terimlerin kullanılması, birimlerin ve ölçümlerin farklılaşması, farklı tanımlayıcıların kullanılması gibi yeni veri kalitesi sorunlarının ortaya çıkmasına neden olacaktır.[33]

Verileri biçimlendirme görevi, verilerin sözdizimsel olarak anlamlarını değiştirmeden, modelleme aracı için gerekli olan forma sokulması işlemini içerir. Bu görev sonucunda, veri kümesindeki değişkenlerin ya da kayıtların belli bir düzende ya da rastlantısal olarak sıralanması, karakter uzunluklarının belirlenmesi, ayraç kullanımlarının standartlaştırılması gibi faaliyetlerle oluşturulan biçimleştirilmiş veri kümesi, modelleme aşaması için kullanıma uygun hale getirilmiş olur.

Veri madenciliği başarısında kaliteli verinin ne kadar önemli olduğu dikkate alınırsa veri ön işleme (data preprocessing) faaliyetlerinin veri madenciliği açısından önemi de anlaşılmış olacaktır. Bu açıdan, modelleme öncesinde veri temizleme (data cleaning), veri birleştirme (data integration), veri dönüştürme (data transformation) ve veri indirgeme (data reduction) faaliyetleri anahtar niteliğinde önem taşımaktadır.[34]

Veri ön işleme sürecinin sınıflandırma kesinliği açısından olumlu sonuçlara neden olduğu çeşitli tekniklerin kullanımında deneysel olarak da kanıtlanmıştır.[35]

Veri madenciliği projesinin hem emek hem de zaman açısından büyük bir kısmını verilerin modelleme aşaması için hazır hale getirilmesi oluşturmaktadır. Bu konuda net bir oran vermek doğru olmasa da çalışmaların süre açısından %50 ile %90 aralığında bir kısmını bu faaliyetin oluşturacağı düşünülmektedir.

---

[33] Edelstein H., "Data Mining in Depth: TIAin't", Nisan 2003, *http://www.dmreview.com/issues/20030401/6512-1.html*, [Erişim 08.05.2005]
[34] Oğuzlar A., "Veri Ön İşleme", **Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**, Sayı.21, 67-76, 2003
[35] Crone S.F. ve diğerleri, "The Impact of Preprocessing on Data Mining: An Evaluation of Classifier Sensitivity in Direct Marketing", **European Journal of Operational Research**, No.173, 781-800, 2006

1.2.5. Model Oluşturma

Bu aşamada, modelleme tekniğinin seçilmesi ve bu modellere ilişkin parametrelerin optimal değerlere ayarlanması gibi çalışmaların tümü yer almaktadır. Aynı veri madenciliği problemi için kullanılabilecek birden çok tekniğin bulunması ve bu tekniklerin uygulamada farklı veri formlarına ihtiyaç duyması nedeniyle zaman zaman veri hazırlama aşamasına geri dönmek söz konusu olabilir. Bu aşama ile ilgili temel görevler ve ilgili çıktıları aşağıdaki tabloda (Tablo 1.4) görülmektedir.

Tablo 1.4. Model Oluşturma Aşaması

Aşama	Model Oluşturma
Görev	Modelleme Tekniğini Seçmek
Çıktı	Modelleme Tekniği
Çıktı	Modelleme Varsayımları
Görev	Sınama Tasarımı Üretmek
Çıktı	Sınama Tasarımı
Görev	Model Kurmak
Çıktı	Parametre Ayarları
Çıktı	Modeller
Çıktı	Model Tanımlaması
Görev	Modeli Değerlemek
Çıktı	Model Değerlemesi
Çıktı	Revize Edilmiş Parametre Ayarları

Model kurma süreci denetimli (supervised) ve denetimsiz (unsupervised) öğrenmenin kullanılmasına göre farklılık göstermektedir. Denetimli öğrenme sürecinde veri kümesinde önceden tanımlanmış sınıflara ilişkin özelliklerin belirlenmesi ve bu özelliklerin kural cümleleri ile ifade edilmesi hedeflenmektedir. Süreç tamamlandığında bu kural cümleleri yeni verilere uygulanmakta ve bu verilerin hangi sınıfa ait olduğu öngörülmektedir. Denetimsiz öğrenmede ise verilerin benzerliklerinden ya da uzaklıklardan hareketle ait oldukları sınıfların üretilmesi amaçlanmaktadır.[36]

[36] Hegland M., "Data Mining Techniques", **Acta Numerica**, 313-355, 2001

Denetimli öğrenme sürecinde modelin belirli bir veri kümesi üzerinde oluşturulması ve kalan verilerle modelin geçerliliğinin sınanması yöntemi kullanılmaktadır. Bunun için veri kümesi en az iki alt gruba ayrılmalı ve birinci grupta model parametrelerinin kestirimi, ikinci grupta ise modelin sınanması sağlanmalıdır.

Modelleme tekniğinin seçilmesi görevi, işi anlama aşamasından farklı olarak, kullanılacak olan tekniğin net ve detaylandırılmış tespitini içermektedir. Eğer birden çok teknik uygulanacak ise bu görevin her bir teknik için ayrı ayrı oluşturulması gerekmektedir. Bu görev sonucunda, kullanılacak olan teknik ve ilgili varsayımları belgelenmiş olacaktır.

Model kurma işleminden önce, modelin kaliteli ve geçerli olabilmesi için bir prosedürün geliştirilmesi gerekmektedir. Sınama tasarımı üretme görevinde, verilerin bölünmesi yoluyla, modelin birbirinden tamamen farklı veri kümeleri üzerinde kurulması, sınanması ve değerlendirilmesini sağlayacak bir planın oluşturulması amaçlanmaktadır. Bu görevdeki en önemli nokta ise verilerin bölünmesi işleminin hangi kriterlere göre ve ne şekilde yapılacağının belirlenmesi olacaktır.

Verilerin modelleme öncesinde bölünmesine ihtiyaç duymamızın en önemli nedeni, modelin kesinliğinin belirlenmesinde öğrenme kümesini kullanmanın yanıltıcı ve çoğunlukla iyimser sonuçlar üretmesidir. Çünkü model öğrenme kümesine her zaman aşırı uyum gösterecektir. Burada, modelin öğrenme kümesine ait olup tüm veri setine ait olmayan bazı belirli karakteristiklerden olumsuz etkilenmesini de ölçmek amacıyla bir deneme kümesine ihtiyaç duyulmaktadır.[37]

Model kurma görevi, hazırlanmış veri kümesi üzerinde modelleme tekniğinin uygulanması anlamına gelmektedir. Bu görev sonucunda, ilgili modelleme tekniği için parametrelerin kestirilmesi tamamlanmış ve ortaya gerçek bir model çıkmış olacaktır. İlgili modele ait gerekli açıklamalar ile açıklama konusunda yaşanan güçlükler belgelenmelidir.

---

[37] Fayyad U., Piatetsky-Shapiro G., Smyth P., " From Data Mining to Knowledge Discovery in Databases", **AI Magazine**, Vol.17, No.3, 1996

Veri madenciliği uygulamalarında kullanılabilecek birden çok teknik söz konusu olduğundan hangi tekniğin en uygun olduğuna önceden karar vermek pek mümkün değildir. Bu yüzden veri kümesi üzerinde K adet teknik yolu ile K adet model oluşturulur ve bu modellerin sınanması sonucunda en uygun modele ulaşılır. Hiç bir modelin uygun görülmediği durumlarda olası sorunların tespit edilmesi ile süreç yeniden gerçekleştirilir.

Analizci, oluşturduğu modeli daha çok kendi alanındaki bilgi birikimi, veri madenciliği başarı kriterleri ve arzulanan sınama tasarımı çerçevesinde değerlendirir. Ayrıca analizci, genellikle aynı modeli birden çok kez tekrarlayarak ya da farklı modelleri geliştirerek alternatifler üretir. Bu modelleri değerlendirme kriterlerine göre karşılaştırır ve sıralandırır. Ancak modeli değerleme görevinde, analizcinin yanı sıra, işe ilişkin uzmanlığında modeli değerlendirme sürecine katılması gerekir.

Bu görev sonunda, sonuçlar özetlenir ve geliştirilen modellerin üstünlükleri ortaya konur. Model parametreleri gerekiyorsa revize edilir, en iyi modelin elde edildiğine dair yüksek bir inanç oluşana kadar bu aşama tekrarlanır. Yapılan tüm revizyonların ve değerlemelerin belgelenmesi önerilir.

1.2.6. Değerlendirme

Elde edilen modelin konuşlandırılması öncesinde model oluşum sürecinin dikkatlice gözden geçirilmesini ve modelin iş hedeflerini başarma konusundaki yeterliliğinin değerlendirilmesini içeren değerlendirme aşaması yer almaktadır. Bu aşamada, veri madenciliği sonuçlarının nasıl kullanılacağına da karar vermek gerekecektir.

Bu aşamaya gelindiğinde, veri analizi perspektifi ile yüksek kaliteye sahip olduğu görülen bir veya birden çok model elde edilmiş durumdadır. Ancak modelin geçerli iş hedefine dönük olarak yeterli olup olmadığı ya da herhangi bir iş ödevinin göz ardı edilip edilmediği son kez gözden geçirilir. Bu aşama ile ilgili temel görevler ve ilgili çıktıları aşağıdaki tabloda (Tablo 1.5) görülmektedir.

Tablo 1.5. Değerlendirme Aşaması

Aşama	Değerlendirme
Görev	Sonuçları Değerlendirmek
Çıktı	İş Başarısı Açısından Veri Madenciliği Sonuçlarının Değerlemesi
Çıktı	Onaylanmış Modeller
Görev	Süreci Gözden Geçirmek
Çıktı	Sürecin Gözden Geçirilmesi
Görev	Sonraki Adımları Belirlemek
Çıktı	Olası Eylemlerin Listesi
Çıktı	Karar

Önceki değerlendirme adımları daha çok modelin kesinliği ve genelliği gibi faktörlerle ilgilidir. Bu aşamadaki sonuçları değerlendirme görevi ise modelin iş hedeflerini karşılama derecesini değerlemeye ve herhangi bir nedenle iş açısından yetersiz olmasının söz konusu olup olmadığını belirlemeye yöneliktir. Ve hatta, maliyet ve zaman faktörleri açısından mümkün ise gerçek bir uygulamayla sınanmasını da içerebilir.

Bununla birlikte, iş ile dolaylı ilgili veri madenciliği sonuçları da değerlendirilir. Çünkü veri madenciliği sonuçları, iş hedefiyle doğrudan ilişkili ve iş hedefiyle dolaylı ilişkili ancak ortaya çıkarılması işin tüm boyutları açısından faydalı olabilecek zorlukları, bilgileri ve belirtileri içerebilir. Bu görev sonucunda, projede iş hedefleri doğrultusunda daha önceden belirlenmiş iş başarı kriterleri açısından elde edilen sonuçların değerlendirilmesi özetlenmiş olacak ve bu değerlendirme sonucunda söz konusu kriterleri sağlayan model ya da modeller onaylanacaktır.

Süreci gözden geçirme görevi, iş hedefi açısından tatmin edici olduğu umulan sonuç modelin ve bu modelin elde edilmesi sürecinin çok daha titiz ve incelikli bir şekilde gözden geçirilmesini içerir. Herhangi bir görevde hata oluşup oluşmadığı, gözden kaçan bir unsurun olup olmadığı, verilerin seçilmesinde gözetilen tercihlerin yeniden değerlendirilmesi gibi faaliyetlerle projenin kalitesi garanti altına alınmaya çalışılır. Bu görev sonucunda, gözden kaçmış veya varsa tekrarlanması gereken eylemlerin yürütülmesi ve belgelenmesi sağlanacaktır.

Sonuçların değerlendirilmesi ve sürecin gözden geçirilmesi sonucunda ya projeyi bitirmek üzere bir sonraki aşamaya geçmek ya da daha fazla tekrarlama ve hatta yeni bir proje oluşturmak üzere geri dönmek konusunda bir karar verilmesi gerekecektir. Sonraki adımları belirleme görevi, elde kalan yeterlilikler ve bütçe açısından, bu konuda bir karar vermeyi içerir. Bu görev sonucunda, bir sonraki adımda potansiyel olarak gerçeklenebilecek olan eylemleri ve bu eylemleri destekleyen veya karşı nitelikteki gerekçeleri listelemek mümkün olacak ve ne yapılacağı konusunda mantıksal açıklamaya dayanan bir karar alınacaktır.

1.2.7. Konuşlandırma

Bu aşama, analizciden çok organizasyonun kendisi tarafından anlaşılması gereken bir aşamadır. Geçerliliği sağlanan model iki farklı şekilde kullanılabilir. Birinci yol, analizci tarafından modelin ve sonuçlarının sunularak eylem önerilerinde bulunulmasıdır. İkinci yol ise modelin yeni veri kümeleri üzerinde kullanılmasıdır.

Genellikle bir modelin bilgi kazanımı açısından - çok başarılı da olsa - oluşturulması yeterli olmayıp, kullanıcı için organize edilmesi ve sunulması da gerekmektedir. Bu açıdan konuşlandırma aşaması, gerekliliklere bağlı olarak sadece bir raporlama işlemi kadar basit olabileceği gibi canlı sistemler tarafından tekrarlanabilir bir süreç olarak uygulanması kadar karmaşık da olabilir.

Ancak bu tercih, veri analizcisinden çok uygulamacının karar vermesini gerektiren bir tercih olup, analizci ile uygulamacının farklı olduğu durumlarda uygulamacının modeli ve modelin ihtiyaçlarını yeterli seviyede kavraması gibi bir görevi de beraberinde getirir. Konuşlandırma aşaması ile ilgili temel görevler ve ilgili çıktıları aşağıdaki tabloda (Tablo 1.6) görülmektedir.

Konuşlandırmayı planlama görevi, veri madenciliği sonuçlarının işe uygulanması için, değerlendirme sonuçlarını alarak bir strateji geliştirilmesini ifade eder. Bu görev sonunda, konuşlandırmanın gerekli adımları ve nasıl gerçekleneceğini içeren bir stratejik plan belgelenir.

Tablo 1.6. Konuşlandırma Aşaması

Aşama	Konuşlandırma
Görev	Konuşlandırmayı Planlamak
Çıktı	Konuşlandırma Planı
Görev	İzlemeyi ve Bakımı Planlamak
Çıktı	İzleme ve Bakım Planı
Görev	Sonuç Raporu Oluşturmak
Çıktı	Sonuç Raporu
Çıktı	Sonuç Sunumu
Görev	Projeyi Gözden Geçirmek
Çıktı	Deneyim Dokümantasyonu

Veri madenciliğinin ürünü, canlı bir sistem ve onun çevresinin bir parçası olacak ise izleme ve bakım oldukça önem kazanır. Özenle hazırlanmış bir bakım planı, sonuçların - yanlış bir şekilde - gereğinden daha uzun bir süre boyunca kullanılmasını önler. Veri madenciliğinin sonuçlarını izlemek, ayrıntılı bir plan gerektirir ve bu plan konuşlandırmanın türüne göre şekillendirilir. Bu görev ile, izleme ve bakım için gerekli adımları ve bu adımların nasıl gerçekleneceğini içeren bir stratejik plan belgelenir.

Proje sonunda, proje lideri ve takımı tarafından bir sonuç raporu hazırlanır. Konuşlandırma planına göre bu rapor, sadece projenin ve edinilen deneyimlerin bir özeti olabileceği gibi, veri madenciliği sonuçlarının nihai ve ayrıntılı bir sunumu da olabilir. Ayrıca, sonuç raporu oluşturma görevi doğrultusunda sonuçların uygulamacıya aktarılması için bir sözel sunum da gerçekleştirilebilir.

Projeyi gözden geçirme görevi ile nelerin doğru nelerin yanlış yapıldığı, nelerin yerinde yapılıp nelerin düzeltmeye ihtiyacı olduğu son kez değerlendirilir. Bu görev ve proje sonucunda, proje sırasında edinilen önemli deneyimler özetlenir. İdeal bir projenin belgelenmesinde, proje aşamaları ve görevleri sırasında proje üyelerinin bireysel olarak yazdığı raporlarda yer alır.

1.3. VERİ MADENCİLİĞİNİN FONKSİYONLARI

Veri madenciliği tanımlarında ön plana çıkan öğelerin, çok fazla miktarda veri ve bu veri yığınlarından anlamlı bilginin keşfedilmesi olduğu şeklindeki çıkarımdan önceki bölümde söz etmiştim. Buradan hareketle, veri madenciliğinin veri yığınlarında saklı bulunan bilgiye erişmek için kullanılan bir araç olduğu yargısı oluşmaktadır. Ancak, bilgi söz konusu olduğunda bilginin türü de önem kazanmaktadır.

Veri madenciliğinin fonksiyonları, bir başka deyişle veri madenciliğinin çözüm yöntemi olarak kullanılabileceği problem türleri, ortak bazı noktaları olmakla birlikte hemen her kaynakta farklı şekilde sıralanmıştır. Ancak genel bir kabul olarak, veri madenciliği modelleri öngörü (prediction) modelleri ve tanımlama (description) modelleri olmak üzere iki temel başlık altında incelenmektedir.[38]

Bu çalışmada, veri madenciliği fonksiyonlarına ilişkin olarak CRISP – DM dokümanında yer alan sınıflama esas alınmış ve veri madenciliği problem türleri altı alt başlık altında incelenmiştir. Buna göre, veri madenciliğinin fonksiyonları veri tanımlama ve özetleme (data description and summarization), bölümleme (segmentation), kavram tanımlama (concept description), sınıflandırma (classification), öngörü (prediction) ve bağımlılık analizi (dependency analysis) şeklinde sıralanmıştır. Bu sıralamayı yapmakla birlikte, veri madenciliği uygulamalarında iş hedeflerine ulaşmak için genellikle birden çok problem türü ile ilgilenilmesinin gerekeceğini vurgulamakta ve birden çok tekniğin kullanılmasını önermektedir.

1.3.1. Veri Tanımlama ve Özetleme

Veri tanımlama ve özetleme, verilerin genel yapısını ortaya çıkarmaya ve karakteristiklerini tanımlamaya yönelik faaliyetlerin yürütüldüğü veri madenciliği çalışmalarını ifade etmektedir. Bu faaliyetler, veri madenciliği yelpazesindeki en düşük seviye olarak kabul edilmektedir. Bu açıdan, veri tanımlama ve özetleme faaliyetleri tanımlayıcı veri madenciliği çalışmalarının en belirgin örneğini teşkil etmektedir.

---

[38] Akpınar H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İstanbul Üniversitesi İşletme Fakültesi Dergisi**, Cilt.29, Sayı.1, 1-16, 2000

Bu açıdan, veri tanımlama ve özetleme bir veri madenciliği projesinin nihai hedefi olabileceği gibi, hemen her veri madenciliği projesinin alt hedeflerinden biri olarak da kabul edilmektedir.

Veri madenciliği sürecinin başlangıcında uygulamacı, verilerin yapısını ve veri madenciliği hedefini kesin olarak bilmemekte, veri içerisinde gizlenmiş potansiyel nitelikteki hipotezleri kurgulamak için veriyi anlamaya dönük başlangıç düzeyinde çalışmalara ihtiyaç duymaktadır. Bu nedenle, basit tanımsal istatistik ve görselleştirme teknikleri verinin iç yapısına ilişkin ilk bilgileri sağlayan önemli bir araç olmaktadır.

Veri tanımlama ve özetleme, genellikle diğer veri madenciliği sorunlarıyla birlikte düşünülmekte ve hatta ürettiği sonuçlarla diğer veri madenciliği sorunlarının neler olduğunu da ortaya koymaktadır. Bu nedenle, her veri madenciliği projesi için ilk başta dikkate alınması önerilen faaliyetleri kapsamaktadır.

Yüksek kalitede tanımlama genellikle grafik analizi de içeren keşifsel veri analizi (exploratory data analysis) ile mümkün olmaktadır. Keşifsel veri analizi ile verileri derinlemesine araştırmak, değişkenler arası ilişkileri gözden geçirmek, verilerdeki ilginç alt kümeleri belirlemek ve değişkenlerin kendi aralarında ve bağımlı değişkenle birliktelikleri konusunda düşünce üretmek mümkün olmaktadır.[39]

1.3.2. Bölümleme

Bölümleme, verilerin ortak karakteristikleri paylaşan ilgi çekici ve anlamlı alt gruplara ayrılmasını amaçlayan veri madenciliği çalışmalarını ifade etmektedir. Bu çalışmalarda uygulamacı, kendi bilgi birikimi ya da veri tanımlama ve özetleme çalışmalarında öğrendiklerine dayanarak belirli alt grupları tanımlayabileceği gibi, veri içerisinde saklı bulunan ve öngörülemeyen alt grupların tespit edilmesine dayanan kümeleme yöntemlerini de kullanabilmektedir.

[39] Larose D.T., **Discovering Knowledge in Data: An Introduction to Data Mining**, John Wiley & Sons, 2005, sf.41-66

Kümeleme analizi veri içerisinde saklı bulunan ancak önceden tanımlanmamış sınıfların üretilmesini amaçlamakta ve sınıflar arası benzerliğin en büyükleştirilmesi ya da sınıf içi benzerliğin en küçükleştirilmesi prensiplerini kullanmaktadır.[40]

Kümeleme, kimi zaman yanlış bir şekilde sınıflandırma ile aynı anlamda kullanılmaktadır. Ancak bu iki yöntem temelde alt veri kümelerinin verilerin kendisinden türetilmesi ya da önceden tanımlanmış olması açısından farklılık göstermektedir. Kümelemede ortaya çıkan grupların anlamlılığı ancak konunun uzmanı tarafından açıklanabilir. Bu nedenle hangi değişkenlerin dikkate alınarak kümeleme yapılacağı uygulama alanında uzmanlığı gerektirir.

Bölümleme, bir veri madenciliği projesinin temel sorunu olabileceği gibi diğer veri madenciliği sorunlarının çözümlenmesi için bir adım olarak da düşünülebilir. Bu çalışmalarda amaç, verileri yönetilebilir büyüklükte tutmak ya da analizi kolaylaştırmak açısından daha homojen alt veri kümeleri oluşturmak olabilir. Bölümleme, tanımlayıcı veri madenciliği çalışmaları için çok bilinen bir örnek olup, sonrasında sınıflandırma çalışmaları da uygulanabilmektedir.

1.3.3. Kavram Tanımlama

Kavram tanımlama, verilerde mevcut bulunan kavramların anlaşılır bir tanımını amaçlayan faaliyetleri ifade etmekte ve model oluşturmaktan çok kavrayışı güçlendirmeye dönük olarak gerçekleştirilmektedir. Bu faaliyetleri, tanımlama ya da öngörü modelleri olarak gerçekleştirmek mümkündür.

Kavram tanımlama, hem bölümleme hem de sınıflandırma ile yakın anlam ifade etmektedir. Ancak bölümleme faaliyetinde ulaşılan grupların anlamlandırılması gibi bir hedef bulunmamaktadır. Ayrıca bölümleme, kavram tanımlamadan önce gerçekleşecek bir faaliyettir. Sınıflandırma ile arasındaki en temel fark ise kavram tanımlamanın tüm veri kümesi üzerinde değil sadece ilgili kavram açısından önem arz eden kayıtlarda uygulanmasının yeterli görülmesidir.

[40] Han J., Kamber M., *a.g.e.*, sf.335

Kavram tanımlama, üzerinde durulan kavrama ilişkin verilerle gerçekleştirilecek tanımlayıcı çalışmaları , bu verilerin yapısı ile geriye kalan verilerin yapısının karşılaştırılmasını ya da bu iki faaliyeti birden içerebilir.[41]

1.3.4. Sınıflandırma

Sınıflandırma, veri kümesindeki kayıtları önceden belirlenmiş alt gruplara ait varsayarak tüm kayıtlara doğru sınıf etiketlerinin atanmasını amaçlayan veri madenciliği çalışmalarını ifade etmektedir. Diğer yandan, sınıflandırma modelleri değişkenlikleri ve veri içerisindeki çeşitli problemleri tanımlamak için de kullanılabilmektedirler.

Sınıflandırma modellerinde bağımlı değişken kesikli olmakta ve bu değişkenin her bir değeri birer sınıf etiketi ifade etmektedir. Sınıflandırma modeli, bağımlı değişkenin de değerlerini içeren veri kümesi üzerinden sınıflayıcı kuralları türetmekte ve bu kuralları yeni kayıtlara uygulayarak öngörülerde bulunmaktadır.

Çalışmamın temelini veri madenciliğinin sınıflandırma fonksiyonu oluşturduğundan bu fonksiyonu ileride daha ayrıntılı olarak inceleyeceğim.

1.3.5. Öngörü

Öngörü, hedef değişkenin her bir kayıttaki değerini veri kümesinden edinilen bilgi doğrultusunda doğru olarak belirlemeyi amaçlayan veri madenciliği çalışmalarını ifade etmektedir.

Öngörü ile sınıflandırma arasında bir takım benzerliklerin varlığından söz edilebilse de, iki yaklaşım arasındaki en büyük farklılık hedef değişkenin ölçeği tarafından belirlenmektedir. Çünkü öngörü çalışmasında hedef değişken sınıflandırmada olduğu gibi kesikli değil sürekli ölçektedir. Bu nedenle, öngörü modellerine kimi kaynaklarda regresyon modeli adı da verilmektedir. Ayrıca, öngörü bir zaman serisi ile ilgilenmekte ise tahmin söz konusu olacaktır.

---

[41] Han J., Kamber M., *a.g.e.*, sf.179

1.3.6. Bağımlılık Analizi

Bağımlılık analizi, veri kayıtları veya veri alanları arasında anlamlı bağımlılıkların tespit edilmesine dayanan çalışmaları içermektedir. Bu yöntemlerle, bilgisi verilen kayıtlardan hareketle herhangi bir veri kaydının değerini kestirmek mümkün olmaktadır.

Birliktelik tespiti ve ardışık örüntü tespiti, bağımlılık analizinde kullanılan iki yaygın yaklaşımdır. Birliktelik tespiti, veri kayıtları arasındaki kuralları, yani sıklıkla birlikte ortaya çıkan veri parçalarını tespit etmeye çalışır. Bu birliktelikleri tarayan algoritmalar son derece hızlı olup, çok sayıda birliktelik üretebilirler. Buradaki zorluk, en önemli birlikteliklerin tespit edilmesidir. Benzer şekilde, kuralları zamana bağlı olarak tespit etmeye dayanan analizlere de ardışık örüntü tespiti modelleri adı verilir.[42]

Uygulamada, bağımlılık analizi ile bölümleme sıklıkla birlikte gerçekleşirler. Büyük veri setlerinde, aralarında çok fazla etki yer alması nedeniyle anlamlı bağımlılıklara nadiren rastlanır. Bu tür durumlarda, bağımlılıkların daha homojen alt kümelerde aranması önerilmektedir.

1.4. SINIFLANDIRMA FONKSİYONU

Sınıflandırma problemini diğer birçok veri madenciliği probleminden ayıran en belirleyici özellik, hedef değişkenin iki ya da daha fazla sınıf etiketine (class label) sahip nitelikte yani kategorik ölçekte olmasıdır.

Sınıflandırma görevi, sınıfların açık ve net olarak tanımlanmasını ve önceden sınıflandırılmış örneklerden oluşan bir veri kümesini gerektirir. Burada görevden kasıt, henüz sınıflandırılmamış örneklerin önceden belirlenmiş olan sınıflara atanmasını sağlayacak olan bir model geliştirmektir.

Bir başka deyişle, sınıflandırma çalışması hangi sınıfa ait olduğu bilinen kayıtlar üzerinden bir model oluşturulması ve oluşturulan modelin yeni kayıtların sınıflandırılması amacıyla kullanılması olmak üzere iki basamaklı bir süreçtir.

---

[42] Two Crows Co., *a.g.e.*

Sınıflandırma modeli, hem çeşitli öngörü değişkenlerine hem de hedef değişkene ait değerleri bilinen çok sayıda kaydın gözden geçirilmesi yolu ile oluşturulur. Bu sayede algoritma, hangi değişken kombinasyonlarının hedef değişkende hangi sonucu doğurduğunu öğrenmiş olur.

Ayrıca, sınıflandırma işlemi öncesinde veri kümesinin tekbiçimleştirilmesi de kullanışlı olabilmektedir. Normalleştirme işlemi ile analize dahil edilen değişkenlerin bütün değerleri (0,1) ya da (-1,+1) aralığında ölçeklendirilir. Bu sayede, uzaklık ölçüsünü dikkate alan yöntemlerde yayılma bandı farklılıklarından kaynaklanan aşırı etki önlenmiş olacaktır.

Model oluşturma adımında tüm veri kümesi içerisinden rastlantısal olarak seçilen kayıtların oluşturduğu bir alt veri kümesi kullanılır. Öğrenme kayıtları (training sample) olarak adlandırılan bu kayıtların tümüne öğrenme kümesi (training set) adı verilir. Öğrenme kümesindeki her bir kaydın hedef değişkende aldığı değer de bilindiğinden bu basamak denetimli öğrenme (supervised learning) olarak bilinmektedir. Denetimli öğrenme süreci ile elde edilen model sınıflandırma kuralları, karar ağaçları ya da matematiksel formülasyon formlarında ortaya çıkabilmektedir.

Sürecin ikinci adımında, oluşturulan model hedef değişkendeki değeri henüz bilinmeyen kayıtların ya da yeni girdilerin sınıf etiketini öngörmek amacıyla kullanılır. İlk önce, modelin öngörü kesinliği (predictive accuracy) için kestirimde bulunulur. Modelin kesinliği, öğrenme kümesinden bağımsız ve yine rastlantısal olarak seçilmiş bir deneme kümesi (test set) yoluyla ölçülmelidir. Eğer modelin kesinliği yeterli bulunursa, henüz sınıflandırılmamış kayıtların sınıf etiketinin öngörülmesi için elde edilen model kullanılabilir.

Grafik yöntemi, veri kümesi içerisinde mevcut olan iki ya da üç boyutlu ilişkilerin anlaşılmasına yardımcı olabilmektedir. Ancak daha fazla sayıda öngörü değişkeninin bulunduğu durumlarda grafik yöntemi yetersiz kalmakta ve daha karmaşık yöntemlerin kullanılması kaçınılmaz olmaktadır.[43]

[43] Larose D.T., *a.g.e.*

Bir başka önemli sorun ise çok fazla sayıda veriye rağmen aranılan örneklerin çok az sayıda olmasıdır. Eğer bir değişkenin tespit edilmeye çalışılan değeri toplam veri içerisinde sıfıra yakın bir oranda yer alıyorsa herhangi bir algoritma ilgili değişkenin değeri ile ilgili öngörüsünde herhangi bir girdinin bire yakın oranda aranılan değeri almadığına işaret edecektir. Bu şekilde hem bu davranışın ayırt edilebilmesi zorlaşmakta hem de hatalı öngörü oranı artmaktadır.[44]

Veri madenciliğinin sınıflandırma fonksiyonu için önerilen başlıca yöntemler,

- ✓ Diskriminant Analizi
- ✓ Lojistik Regresyon
- ✓ Kural Türetme Yöntemleri
- ✓ Karar Ağaçları
- ✓ Yapay Sinir Ağları
- ✓ K – En Yakın Komşu Algoritması
- ✓ Genetik Algoritmalar
- ✓ Örnek Tabanlı Çıkarsama
- ✓ Bayesci Sınıflama
- ✓ Katı Küme ve Fuzzy Küme Yaklaşımları

olarak sıralanabilmektedir. Çalışmamın uygulama bölümünde bu yöntemlerden üçüne ilişkin uygulama söz konusu olacağından ilerleyen bölümde istatistik yöntemlerden lojistik regresyon analizi, yapay sinir ağlarından çok katmanlı algılayıcı ve karar türetme algoritmalarından C5.0 algoritması üzerinde detaylı olarak durulacaktır.

Uygulama için bu yöntemlerin seçilmesinde hem veri madenciliği uygulamalarına ilişkin farklı yaklaşımları temsil etmeleri hem de gerçekleştirdiğim kaynak taramalarında bu yöntemlerin veri madenciliğinin öncü yöntemleri konumunda olduğunu tespit etmem etkili olmuştur.

Büyük bir veri kümesi üzerinde sınıflandırma amacıyla kurulacak modele bir takım değişkenlerin yeterli katkıda bulunmaması mümkündür. Hatta bazı değişkenlerin bu model kurma sürecine katılması gereksiz dahi olabilir. Bu nedenle, model kurma aşamasından önce hedef değişkendeki sınıf etiketinin öngörülmesi için hangi değişkenlerin geçerli olduklarının belirlenmesi (feature selection) faydalı

[44] Edelstein H., Nisan 2003, *a.g.e.*

olacaktır. Çünkü geçersiz ve gereksiz değişkenler modeli zayıflatabilir veya hatalı öngörülere neden olabilirler.

Eğer bir değişkenin değerleri belirli bir sınıfı diğerlerinden ayırmak için kullanılabiliyorsa, o değişkenin ilgili sınıf değeri için yüksek derecede geçerli olduğu söylenebilir. Bununla birlikte, aynı değişkenin farklı soyutlama düzeyleri de bir sınıfı diğerlerinden ayırma konusunda beklenmedik şekilde farklı güce sahip olabilirler. Bu yüzden, değişkenlerin geçerlilik analizi (relevance analysis of attributes) farklı soyutlama düzeylerinde gerçeklenmeli ve sadece geçerli değişkenler sürece dahil edilmelidir. Bu kapsamda geçerli değişkenlerin tespiti, sınıf karakteristiğine yönelik olarak yapılıyorsa analitik niteleme, sınıfların karşılaştırılmasına yönelik olarak yapılıyorsa analitik karşılaştırma olarak adlandırılır.[45]

Geçerli değişkenlerin analizinde ölçü olarak kullanılacak tekniğe göre farklılık göstermek üzere bilgi kazanımı (information gain), Gini endeksi, entropi ve korelasyon katsayıları kullanılabilmektedir.

Sınıflandırma yöntemleri denetimli öğrenmeye dayalı olduklarından modelin öğrenme kümesi üzerinde sınanması uygun olmamaktadır. Modelin değerlendirilmesi amacıyla sınama kümesi kullanılmalıdır.

Modelin değerlendirilmesi amacıyla tüm sınıflandırma yöntemlerinde kullanılan en yaygın ve basit araç sınıflandırma tablolarıdır. Bu tablolarda veri kümesinin her bir kategorideki sıklığı ile model öngörülerinin karşılaştırılması mümkün olmakta ve bu şekilde modelin kesinliği hesaplanabilmektedir. Ancak modelin kesinliği ne düzeyde olursa olsun gerçek yaşama uygunluğu garanti edilemez. Bir başka deyişle, her iyi model doğru model olmayabilir. Bu nedenle modelin gerçek yaşamla sınanması gerekmektedir.

[45] Han J., Kamber M. , *a.g.e.*

1.4.1. Lojistik Regresyon Analizi

Lojistik regresyon analizi diskriminant analizi ile birlikte sınıflandırma amacıyla kullanılan istatistik yöntemlerden biridir. Ancak diskriminant analizi tüm öngörü değişkenlerinin normal dağıldığı varsayımı, nominal öngörü değişkenlerinin kullanılamaması ve sınıfları ayıran sınırların doğrusal olması gibi sebeplerle veri madenciliğinde pek yaygın olarak kullanılmamaktadır.[46]

Bir kategorik bağımlı değişkenin öngörülmesinde, bütün bağımsız değişkenler sürekli ve normal dağılıma sahip ise diskriminant analizi, bütün bağımsız değişkenler kategorik ise lojit analizi, bir takım değişkenler sürekli diğerleri kategorik ise ve sürekli değişkenlerin dağılımı hakkında varsayımda bulunulmuyorsa lojistik regresyon analizi kullanılmaktadır.[47]

Basit doğrusal regresyon analizi, bağımlı değişken (Y) ile bağımsız değişken (X) arasındaki fonksiyonel ilişkinin belirlenmesine dayanmaktadır. Bu analizde, bağımlı değişken rastsal bir değişken olurken bağımsız değişken rastsal ya da matematiksel bir değişken olabilmektedir. Bağımsız değişkenin birden çok olması halinde ise çoklu doğrusal regresyon gündeme gelmektedir. Her iki durumda, bağımlı değişken ile bağımsız değişkenler arasındaki ilişki, parametreleri hata kareleri toplamının en küçükleştirilmesine dayanan En Küçük Kareler Yöntemi ile kestirilmekte ve aşağıdaki gibi bir doğrusal denklem ile ifade etmektedir.[48]

$$\mu_Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \qquad (1)$$

Ancak, doğrusal regresyon modelinin uygulanması bir takım varsayımların karşılanmasını gerektirir. Bu varsayımlar aşağıdaki gibi sıralanabilmektedir.[49]

- ✓ Tüm bağımsız değişkenler nitel ya da nicel ölçekte, bağımlı değişken ise sürekli olup aralık ya da oran ölçeğindedir.
- ✓ Bağımlı değişkenin tüm geçerli öngörücüleri analize dahil edilmiş, bağımlı değişken için geçerli olmayan hiçbir öngörü değişkeni analizde yer almamıştır.

---

[46] Two Crows Co., *a.g.e.*
[47] Wuensch K.L., "Binary Logistic Regression with SPSS", *http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.doc*, [Erişim 03.06.2008]
[48] Armutlulu İ.H., **İşletmelerde Uygulamalı İstatistik**, Alfa, 2008, sf.210
[49] Menard S., **Applied Logistic Regression Analysis**, Sage Pub., 2002, sf.4

- ✓ Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusal formdadır.
- ✓ Hata terimlerinin (ε) beklenen değeri sıfırdır.
- ✓ Bağımsız değişkenlerin bütün değerleri için hata terimlerinin varyansı sabittir.
- ✓ Hata terimleri koşullu normal dağılımlara sahiptir.
- ✓ Bağımsız değişkenlerin farklı değerleri tarafından üretilen hata terimleri arasında korelasyon bulunmamaktadır.
- ✓ Hata terimleri ile bağımsız değişkenler arasında korelasyon yoktur.
- ✓ Hiçbir bağımsız değişken diğer bağımsız değişkenlerin kusursuz bir doğrusal kombinasyonu şeklinde ifade edilemez.

Doğrusal regresyonda bağımsız değişken kategorik olabilmekte ancak bağımlı değişkenin kategorik olması halinde en küçük kareler kestirimi ile elde edilen model gerçek durumu ifade etmemektedir. Bu nedenle, Cox (1970) tarafından önerilen ve bağımlı değişkenin öngörülen değeri yerine, bağımlı değişkenin belirli bir kategoride bulunma olasılığının öngörülmesine dayanan lojistik regresyon modeli gündeme gelmektedir.

Bağımlı değişken ikili kategorik {0,1} olduğunda, Bernoulli rastsal değişkenini ifade edeceğinden, beklenen değeri aşağıdaki gibi ifade etmek mümkün olacaktır.

$$E(Y) = 0 \times P(Y = 0) + 1 \times P(Y = 1) \quad (2)$$

$$E(Y) = P(Y = 1) \quad (3)$$

ve benzer şekilde,

$$E(Y^2) = E(Y) = P(Y = 1) \quad (4)$$

Bu olasılık, $\mathbf{X}(X_1, X_2, \ldots\ldots, X_k)$ bağımsız değişken vektörünün fonksiyonu şeklinde ifade edildiğinde ise varyans,

$$V(Y) = \pi(\mathbf{X})[1 - \pi(\mathbf{X})] \quad (5)$$

olarak karşımıza çıkacaktır.

Buna göre, ikili kategorik bağımlı değişken için doğrusal regresyon denklemi,

$$E(Y) = \pi(\mathbf{X}) = \alpha + \boldsymbol{\beta}\mathbf{X} \qquad (6)$$

şeklinde ifade edilir ve bu modele *doğrusal olasılık modeli* adı verilir.[50]

Bir başka deyişle, bağımlı değişken ikili kategorik olduğunda, değişkenin ortalaması söz konusu iki kategoriden daha yüksek olanın olasılığına karşılık gelmektedir. O halde, olasılık teorisinin temel aksiyomlarından hareketle, bu öngörünün [0,1] aralığında gerçekleşmesi gerekir. Oysa, kolayca görüleceği gibi doğrusal model, bağımsız değişkenin görece çok büyük ya da çok küçük değerlerine karşılık (-∞, +∞) aralığında öngörülerde bulunacaktır.

Doğrusal olasılık modelindeki bu türden yapısal sorunlar nedeniyle, yine bağımsız değişkenlerin bir fonksiyonu olarak S tipi bir eğri ifade eden *lojistik regresyon fonksiyonu* kullanışlı olmaktadır. Bu fonksiyon, tek bağımsız değişken söz konusu olduğunda,

$$\pi(X) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \qquad (7)$$

şeklinde olup, $\beta>0$ olduğunda artan ve $\beta<0$ olduğunda azalan forma sahiptir.

Genellikle ikili kategorik bir değişkenin açıklayıcı değişkenler vektörüne bağımlılığını ifade etmek için yaygın olarak kullanılan lojistik regresyon modeli, aynı zamanda bazı parametrik modellerin en büyükleştirilmesi için de kullanılmaktadır.[51]

Lojistik regresyon fonksiyonunda, Y=1 olasılığı yerine Y=1 için olasılık oranı (the odds) dikkate alındığında,

$$\frac{\pi(X)}{1-\pi(X)} = e^{\alpha+\beta X} = e^{\alpha}(e^{\beta})^{x} \qquad (8)$$

şeklinde bir ifadeye ulaşılır. Olasılık oranı, X'te gerçekleşen her birim artışa karşılık e^{β}

[50] Agresti A., **Categorical Data Analysis**, John Wiley & Sons, 1990, sf.84
[51] Strauss D., "The Many Faces of Logistic Regression", **The American Statistician**, Vol.46, No.4, 321-327, 1992

kat artış gösterecektir. Olasılık oranının logaritması alındığında ise,

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right)=\alpha+\beta X \qquad (9)$$

ifadesi elde edilir. İlişkinin doğrusal formda ve bağımlı değişkenin hem sürekli olduğu hem de (-∞,+∞) aralığında değer aldığı bu eşitliğe *logaritmik olasılık oranı* (the logit) adı verilmektedir. Bir başka gösterimle,

$$\text{logit}\,(\pi)=\alpha+\beta X \qquad (10)$$

olduğu söylenebilir.[52]

Bu eşitlikte, $\text{logit}\,(\pi)$ için öngörü (-∞, +∞) aralığında yer alacak ve olasılık öngörüsü π için tanımlı olan yayılma bandının dışına çıkmaksızın, βX çarpımı +∞ yönünde büyüdükçe 1'e ve -∞ yönünde büyüdükçe 0'a yaklaşacaktır.

Yukarıda sözü edilen olasılık, olasılık oranı ve logaritmik olasılık oranı ifadelerinin gerçekte aynı şeyin üç farklı ifadesi olduğunun farkında olmamız gerekir. Nasıl ki, doğrusal regresyon modellerinde doğal logaritma değerleri kullanarak bağımlı ve bağımsız değişkenler arasındaki ilişkiyi doğrusal forma dönüştürüyorsak, kategorik bağımlı değişken için de burada benzer dönüşümü yapmaktayız. Bir başka deyişle, ilişki, değişkenleri açısından doğrusal olmamakla birlikte parametreleri açısından doğrusaldır.

Çalışmamın ikinci bölümünde gerçekleştirilecek lojistik regresyon uygulaması için SPSS Inc. tarafından geliştirilen Clementine yazılımı kullanılacaktır. Söz konusu yazılım, parametrelerin kestiriminde Newton – Raphson yinelemeli kestirim yöntemini kullanmakta ve aşağıda özetlenen çoklu kategorik lojistik regresyon prosedürünü yürütmektedir.[53]

Lojistik regresyona katılan tüm kategorik değişkenler, son kategori hariç olmak üzere, her bir kategoriye karşılık "1" ve diğer kategorilere karşılık "0" değerini alan

[52] Agresti A., *a.g.e.*, sf.85
[53] Integral Solutions Co., *Clementine 8.0 Algorithms Guide*, 2003

yeni kategorik değişkenlerle ifade edilir. Türetilen bu değişkenlere kukla değişken (dummy variable) adı verilmektedir.

Gözlem değerleri, bağımsız değişken kümelerinin çapraz sınıflandırılmasıyla tanımlanan alt anakütlelere göre birleştirilerek sıklık toplamlarına dönüştürülür. Böylece, her bir alt anakütlenin gözlem değeri sayısı,

$$n_i = \sum_{j=1}^{k} n_{ij} \tag{11}$$

şeklinde ifade edilir. Burada, j bağımlı değişkendeki kategori sayısını, i ise alt anakütleyi ifade etmektedir.

Logit modelde, i'nci alt anakütle için P(Y=j) olasılığı,

$$\pi_{ij} = \frac{e^{X_i'\beta_j}}{1 + \sum_{k=1}^{J-1} e^{X_i'\beta_k}} \tag{12}$$

şeklinde ifade edilmektedir. Böylece model,

$$\ln\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = X_i'\beta_j \tag{13}$$

olarak karşımıza çıkar. Kolayca görülebileceği gibi, j indisi en fazla iki değerini aldığında lojistik regresyon modeli ikili kategorik bağımlı değişkene uygun olacaktır.

Lojistik regresyonda hata terimlerinin değişkenliği bağımsız değişkenlerin büyüklüğüne bağlı olarak gerçekleşmekte ve değişken varyanslılık olarak bilinen bu durum, yansız olmalarına rağmen düşük standart hata duyarlılığında en küçük kareler kestiriminin en iyi kestiricileri sağlamadığı anlamına gelmektedir. Bu nedenle, lojistik regresyon modelinde katsayıların kestirimi için En Çok Olabilirlik (maximum likelihood) yönteminin kullanılması uygun bulunmaktadır.

Modelin logaritmik olabilirliği,

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij}\log(\pi_{ij}) \tag{14}$$

veya,

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij}\log\left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}{1+\sum_{k=1}^{J-1} e^{\mathbf{x}_i'\boldsymbol{\beta}_k}}\right) \tag{15}$$

olacaktır. Modele sabit ilave edilecek olursa,

$$c = \sum_{i=1}^{m}\ln\left(n_i!/n_{i1}!\ldots\ldots n_{iJ}!\right) \tag{16}$$

şeklinde hesaplanır.

Logaritmik olabilirliğin birinci türevi,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_s} = \sum_{i=1}^{m} x_{is}\left(n_{is} - n_i\pi_i\right) \tag{17}$$

ve ikinci türevi,

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_s \partial \beta_t} = -\sum_{i=1}^{m} n_i x_{is} x_{it}\pi_i(1-\pi_i) \tag{18}$$

olacaktır.

En küçük kareler kestirimi doğrusal regresyon parametrelerinin elde edilmesinde doğrudan sonuca giderken, lojistik regresyonda kesin olmayan bir çözümle başlanarak modelin gelişmesi sağlanacak şekilde yeniden çözümleme yapılır ve olabilirlik fonksiyonunda ihmal edilebilir bir farklılaşmaya ulaşıncaya kadar bu süreç devam ettirilir. Bir başka deyişle, lojistik regresyonda kestirim, sınama ve yeniden kestirime dayandığından yinelemeli bir süreçtir. Benzerlik fonksiyonunda

ihmal edilebilir bir değişim elde edildiğinde model, yakınsak olarak adlandırılır.

Buradan hareketle, parametre vektörünün (**β**) en çok olabilirlik kestirimi her yinelemede güncellenmek üzere,

$$\boldsymbol{\beta}^{\gamma+1} = \boldsymbol{\beta}^{\gamma} + \left\{\mathbf{X}' \, \mathbf{Diag}[n_i \pi_i^{\gamma}(1-\pi_i^{\gamma})] \, \mathbf{X}\right\}^{-1} \mathbf{X}'(\mathbf{y} - \mathbf{m}^{\gamma}) \qquad (19)$$

şeklinde ifade edilmektedir. Burada, $m_i^{\gamma} = n_i \pi_i^{\gamma}$ ve γ yineleme sayısını göstermektedir. Bu ifade, yineleme sürdükçe bir sonraki $\pi_i^{\gamma+1}$ değerini hesaplamak için kullanılır.

Yineleme işlemi,

$$L(\boldsymbol{\beta}^{\gamma+1}) - L(\boldsymbol{\beta}^{\gamma}) < 0 \qquad (20)$$

olduğu sürece devam ettirilir. Eğer, $\varepsilon_k < 0$ ve $\varepsilon_p < 0$ olmak üzere,

- $\left| L(\boldsymbol{\beta}^{\gamma+1}) - L(\boldsymbol{\beta}^{\gamma}) \right| < \varepsilon_k$ (21)
- $\max \left| \boldsymbol{\beta}_i^{\gamma+1} - \boldsymbol{\beta}_i^{\gamma} \right| < \varepsilon_p$ (22)
- $\max \left(\dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta^{\gamma+1}} \right) < \min(\varepsilon_k, \varepsilon_p)$ (23)

koşullarından en az biri sağlanacak olursa yineleme işlemi yakınsak kabul edilir.

Lojistik regresyon modelinin değerlendirilmesinde, doğrusal regresyonda olduğu gibi ilk olarak modelin bir bütün olarak yeterliliği üzerinde durulmalıdır. Yani, modeli oluşturan bağımsız değişkenlerin bağımlı değişkenle ilişkileri ne kadar güvenilirdir ve analiz edilen veri kümesindeki rastlantısal değişkenliği ne derecede açıklamaktadır.

Doğrusal regresyonda, parametrelerin kestirilmesi açısından hata kareleri toplamı ne ifade ediyor ise lojistik regresyonda logaritmik olabilirlik aynı anlamı ifade etmektedir. Ancak istatistik yazılımları genellikle logaritmik olabilirlik (LL) yerine –2 ile çarpımından elde edilen bir istatistiği (-2LL) hesaplarlar. LL istatistiği negatif

olmakla birlikte –2LL istatistiği pozitif değer alacaktır ve bu istatistiğin değeri büyüdükçe bağımlı değişkenin öngörülerinin kötüye gittiği anlamına gelecektir.

Bu istatistik, kesişimin hariç tutulduğu başlangıç modeli için,

$$-2LL = -2\sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij} \ln(\bar{\pi}_{ij}) \qquad (24)$$

ve kesişimin dahil edildiği başlangıç modeli için,

$$-2LL = -2N \ln\left(\frac{1}{J}\right) \qquad (25)$$

şeklinde hesaplanır. Sonuç modeli için aynı istatistik,

$$-2LL = -2\sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij} \ln(\hat{\pi}_{ij}) \qquad (26)$$

şeklindedir.

Lojistik regresyonda, doğrusal regresyonun açıklanabilen değişkenliğine (SS_R) en yakın karşılık iki logaritmik olabilirliğin farkı ile ifade edilmektedir. Bu fark, modelin ki – kare değeri olarak adlandırılır ve G_M ya da χ^2 ile gösterilir. G_M istatistiği lojistik regresyonda katsayıların anlamlılığını sınamakta ve bu açıdan doğrusal regresyondaki F istatistiğine denk düşmektedir.[54]

Modelin ki – kare istatistiği, başlangıç modeli ile sonuç modelinin –2LL değerlerinin farkı olarak,

$$\{-2LL(\bar{\pi})\} - \{-2LL(\hat{\pi})\} \qquad (27)$$

şeklinde hesaplanmaktadır. Bu istatistik, başlangıç modeli sadece kesişimden oluşuyor ise $p^{nr} - (J-1)$ serbestlik dereceli, model kesişim hariç olduğunda ise p^{nr} serbestlik dereceli ki – kare dağılımına yakınsamaktadır.

[54] Menard S., *a.g.e.*, sf.20

Lojistik modelin R^2 ölçüleri ise, Cox – Snell için,

$$R_{CS}^2 = 1 - \left(\frac{L(\breve{\pi})}{L(\hat{\pi})} \right)^{\frac{2}{n}} \tag{28}$$

Nagelkerke için,

$$R_N^2 = \frac{R_{CS}^2}{1 - L(\breve{\pi})^{2/n}} \tag{29}$$

ve McFadden için,

$$R_M^2 = 1 - \left(\frac{L(\hat{\pi})}{L(\breve{\pi})} \right) \tag{30}$$

şeklinde hesaplanmaktadır.

Yukarıda sıralanan üç istatistik de yaklaşık olarak aynı ölçüleri ifade etmekte ve doğrusal regresyondaki R^2 istatistiğine benzer anlam taşımaktadır. Cox – Snell istatistiği, gözlem başına kareli geometrik ortalamadaki gelişmeyi ifade etmekte, Nagelkerke istatistiği ise bu istatistiğin düzeltilmesine dayanmaktadır. Çünkü lojistik regresyon modeli verilere kusursuz uygunluk gösterse bile Cox – Snell istatistiği "1" değerini almamaktadır. McFadden istatistiği ise –2LL istatistiğindeki ya da logaritmik olabilirlik ölçüsündeki oransal azalmayı ifade etmektedir.[55]

Uyum iyiliği için ise Pearson istatistiği,

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{J} \frac{(n_{ij} - n_i \hat{\pi}_{ij})^2}{n_i \hat{\pi}_{ij}} \tag{31}$$

şeklinde olup, bu istatistiğin dağılımı $m(J-1) - p^{nr}$ serbestlik dereceli ki – kare dağılımına yakınsamaktadır.

[55] Menard S., *a.g.e.*, sf.24

Sonuç modelde, her bir kayıt için bağımlı değişkenin kategorilerine ait logit değeri hesaplanabilmektedir. Buradan hareketle, herhangi bir kaydın bağımlı değişkenin kategorilerinden birine karşılık gelen en yüksek logit değeri bulunarak, söz konusu kategori modelin öngörüsü olarak kabul edilmektedir. Bu değer,

$$r_{ij} = \ln\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = x_i^{'}\beta_j \qquad (32)$$

şeklinde ifade edilebilir.

Doğrusal regresyonda, bağımlı değişkenin aralık ya da oran ölçeğinde varsayılmasından kaynaklı olarak, öngörülerin hatasız olması beklenen bir durum değildir. Logistik regresyonda ise, bağımlı değişkenin sınırlı sayıdaki kategorilerinin ne kadar yakın öngörüldüğünden çok, doğru ya da yanlış olarak öngörülmesi önem kazanır.

Bu nedenle, sonuç model üzerinden hem öğrenme kümesi hem de sınama kümesi için sınıflandırma tablosu oluşturularak, modelin öngörü başarısının belirlenmesi de bir başarı kriteri olmaktadır.

1.4.2. Yapay Sinir Ağları

Yapay sinir ağları (artificial neural networks), canlı organizmaların karşılıklı bağlantılı (interconnected) sinir hücrelerinin (neurons) oluşturduğu karmaşık öğrenme algoritmalarından esinlenilerek geliştirilmiştir.[56]

Kökeni insan beyninin nörofizyolojisi üzerindeki çalışmalara[57] dayanmakla birlikte, sonraki yıllarda yaygın olarak yapay öğrenme[58] ve istatistik[59] çalışmalarının konusu haline gelmiştir.

Yapay sinir ağlarının temel yapı taşları yapay sinir hücreleri (neurons) olarak

---

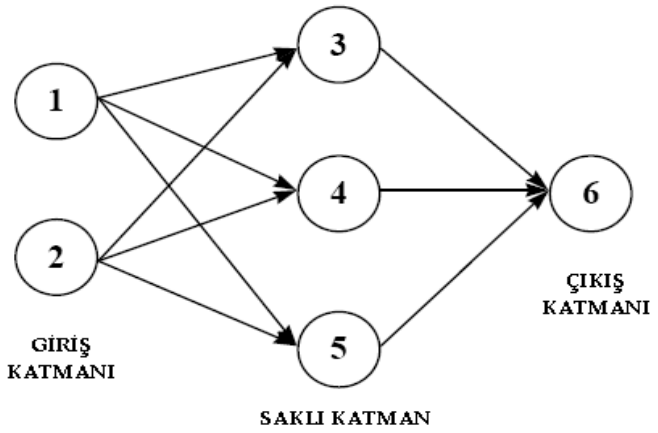
[56] Haykin S., **Neural Networks: A Comprehensive Foundation**, Prentice Hall, 1994, sf.138
[57] Ramon y Cajal, Fizyoloji/Tıp Alanında Nobel Ödülü, 1906
[58] Rosenblatt, F., "The Perceptron: A Probabilistic Model For Information Storage And Organization in the Brain", **Psychological Review**, Vol 65, No 6, 1958
[59] Ripley B.D., **Pattern Recognition and Neural Networks**, Cambridge University Press, 1996

adlandırılmaktadır. Sinir hücreleri, bağlantının kendisi tarafından belirlenen ağırlıklarla birbirlerine bağlanmaktadır. Her ne kadar tek bir sinir hücresi görece basit olsa da karşılıklı bağlantılı sinir hücreleri, son derece karmaşık görevlerin yerine getirilmesinde kullanışlı olmaktadır.

Yapay sinir ağları, veri madenciliği çalışmalarında hem tanımlama hem de öngörü modellerini kapsayan bir çok amaç için kullanılabilmektedir.[60]

Yapay sinir ağları bağımlı değişkenin ve bağımsız değişkenlerle doğrusal olmayan ilişkisini hiçbir istatistiksel varsayım olmaksızın, tamamen verilere dayalı olarak sunmaktadır. Çok katmanlı algılayıcı ise denetimli öğrenme yöntemini kullanan ve genellikle akademik ya da deneysel sınıflandırma amaçları için kullanılan bir yöntemdir. Yapay sinir ağları teorik olarak tüm sürekli değişkenlerle nominal ya da ordinal düzeyde tüm kategorik değişkenlerin kullanılabildiği yöntemlerdir.

Çalışmamın ikinci bölümünde yer alan yapay sinir ağı uygulaması SPSS Inc. tarafından geliştirilmiş olan Clementine 8.0 yazılımı ile gerçekleştirilmiştir. Aşağıda aktarılan bölüm, söz konusu yazılım tarafından kullanılan algoritma esas alınarak hazırlanmıştır.[61]



Şekil 1.5. Çok Katmanlı Algılayıcının Şematik Gösterimi

[60] Giudici P., *a.g.e.*, sf.107
[61] Integral Solutions Limited, Clementine 8.0 Algorithms Guide, 2003

Clementine, çok katmanlı algılayıcı (multilayer perceptron) olarak da bilinen, ileri beslemeli (feedforward) sinir ağlarını (Şekil 1.5) kullanmaktadır. Çok katmanlı algılayıcı, sırasıyla giriş katmanı (input layer), gizli katman (hidden layer) ve çıkış katmanını (output layer) içermektedir. Her bir katmanda, birbirleri ile istisnasız karşılıklı bağlantılı (completely interconnected) sinir hücreleri bulunmaktadır.

Giriş katmanındaki sinir hücrelerinin sayısı genellikle veri kümesindeki değişkenlerin türüne ve sayısına bağlı olmaktadır. Çıkış katmanındaki sinir hücrelerinin sayısı ise üzerinde çalışılan sınıflandırma problemine göre değişkenlik göstermektedir. Bununla birlikte, gizli katmanların sayısı ve gizli katmandaki sinir hücrelerinin sayısı uygulamacı tarafından belirlenebilmektedir.

Yapay sinir ağları, hangi ölçekte olursa olsun bütün değişken yapılarını girdi olarak kabul edebilmektedir. Ancak, değişkenlerin modele dahil edilirken uygun şekilde kodlanması gerekmektedir. Clementine, tüm ölçeklerdeki girdi değişkenleri için farklı olmak üzere, belirli bir kodlama metodolojisi uygulamaktadır.

Yapay sinir ağlarında kullanılacak sürekli değişkenlerin farklı değişkenliklere sahip olması, hem bağlantı ağırlıklarının belirlenmesi sürecini olumsuz etkilemekte hem de bir takım değişkenlerin diğerlerine oranla öngörü değerleri üzerinde daha etkili olmasına neden olmaktadır. Bu nedenle, modele katılacak sürekli ölçekteki değişkenlerin tümü,

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (33)$$

şeklinde bir tekbiçimleştirme işleminden geçmektedir. Bu sayede, tüm sürekli değişkenlerin $[0,1]$ aralığında tanımlanması sağlanmaktadır.

Çoklu kategorik değişkenlerin kodlanmasında, her bir kategoriye karşılık gelecek şekilde $\{0,1\}$ değerlerini alan yeni bir değişken türetilmektedir. Her bir kaydın ait olduğu kategoriye karşılık olarak türetilmiş değişken 1,0 değerini alırken, diğer tüm türetilmiş değişkenler 0,0 değerini almaktadır. Bu şekilde türetilmiş değişkenlere gösterge değişken (indicator variable), bu kodlamaya ise gösterge kodlaması (indicator coding) adı verilmektedir.

Bir değişkendeki kategori sayısının çok fazla olması halinde, sadece bir değişkene karşılık çok fazla sayıda değişken türetilmesi gerekeceğinden ve değişken sayısının artması ile ağın yükü ve hafıza ihtiyacı artacağından, ikili küme kodlaması (binary set encoding) uygulamak daha uygun olmaktadır. Bu yöntemde, türetilen değişkenlerin $\{0,1\}$ şeklindeki değerlerinin kombinasyonu kullanılarak, her bir kategori için ihtiyaç duyulan yeni değişken sayısı azaltılmış olur. Bir başka deyişle, orijinal değişkendeki her bir kategori, türetilen değişkenlerin $\{0,1\}$ şeklindeki değerlerinin benzersiz bir kümesi ile tanımlanmaktadır. Buna göre, k adet kategori için sonuç yukarı yuvarlanarak $\log_2 (k+1)$ adet değişken türetilmesi yeterli olmaktadır.

İkili kategorik değişkenlerin kodlanmasında ise değerleri $\{0,1\}$ olacak şekilde sadece tek bir yeni değişken türetilmesi yeterli olmaktadır. Bu türden değişkenlerde, değeri boş bırakılan kayıt 0,5 olarak kodlanmaktadır.

Çok katmanlı algılayıcının ileri beslemeli doğası, bilginin giriş katmanından çıkış katmanına doğru akmasını ifade etmektedir. Giriş katmanındaki sinir hücreleri gizli katmandaki hücrelere, gizli katmandaki sinir hücreleri ise çıkış katmanındaki hücrelere bağlanarak öngörüleri üretirler.

Çoğu yapay sinir ağı, tek bir gizli katman içerse de gizli katmanların sayısı birden çok olabilmektedir. Gizli katmandaki hücre sayısının arttırılması ise karmaşık yapıların tanımlanmasında ağın gücünü ve esnekliğini arttırmaktadır. Ancak gizli katmanın çok fazla sayıda hücre içermesi de aşırı uyum (overfitting) sorununa neden olabilmektedir. Başka bir deyişle, veri kümesinin ezberlenmesi söz konusu olmakta ve model genellenebilirlikten uzaklaşmaktadır.[62]

Yapay sinir ağları modelinde, veri kümesinden ya da bir önceki katmada yer alan sinir hücrelerinden kabul edilen girdi değerleri, genellikle toplama işlemi barındıran kombinasyon fonksiyonu ile birleştirerek, aktivasyon fonksiyonuna iletilmektedir. Çoğunlukla doğrusal olmayan aktivasyon fonksiyonları, bir çıktı değeri üreterek bir sonraki katmanda yer alan sinir hücrelerinin girdilerini ya da öngörü değerlerini üretirler.

[62] Han J., Kamber M., *a.g.e.*, sf.281

Giriş hücreleri, veri kümesindeki değerleri kabul etmekte ve hiçbir işlem yapmaksızın gizli katmandaki hücrelere iletmektedir. Bu açıdan, giriş katmanındaki hücrelerin yapısı, gizli katman ve çıkış katmanındaki hücrelerin yapısından farklılık göstermektedir. Belirli bir katmandaki hücrelerin çıktıları ve bu hücrelerle bir sonraki katmanda yer alan hücre arasındaki bağlantı ağırlığı, kombinasyon fonksiyonu ile birleştirerek yeni bir değere dönüştürülmektedir. Bu yeni değer *net* olarak adlandırılmakta ve,

$$net_j = \sum_i w_{ij} x_{ij} \qquad (34)$$

şeklinde hesaplanmaktadır. Burada, i'inci hücre ile j'inci hücre arasındaki bağlantı ağırlığı w_{ij} ve i'inci hücreden j'inci hücreye iletilen çıktı değeri x_{ij} olarak gösterilmiştir.

Net değeri, ilgili hücrenin ve dolayısıyla bu hücredeki aktivasyon fonksiyonunun girdisi konumundadır. Aktivasyon fonksiyonu, sinir hücresinin biyolojik yapısında da olduğu gibi genellikle doğrusal olmayıp,

$$y = \frac{1}{1 + e^{-x}} \qquad (35)$$

şeklinde bir sigmoid (S-tipi) fonksiyon ile ifade edilmektedir. Aktivasyon fonksiyonunun sonucu, aynı zamanda ilgili hücrenin çıktısı olmakta ve bir sonraki katmanda yer alan hücrelere belirli bir bağlantı ağırlığı ile iletilmektedir.

Son olarak, bu değerlerin çıkış hücresindeki aktivasyon fonksiyonuna dahil edilmesi ile sinir ağının ilk öngörüsü elde edilmiş olacaktır. Ancak, bu öngörünün en doğru öngörü olmayacağı açık bir gerçek olup iyileştirilmesi gerekmektedir.

Aktivasyon fonksiyonu olarak sigmoid bir fonksiyon seçilmesinin en önemli nedeni, bu fonksiyonun girdi değerlerine bağlı olarak hem doğrusal, hem eğrisel hem de sabit davranış gösterebilme özelliğidir.[63]

[63] Kantardzic M., *a.g.e.*, sf.197

Aslında yapay sinir ağı, doğrusal bir aktivasyon fonksiyonunun uygulanması ve gizli katmanın bulunmaması halinde doğrusal regresyona, doğrusal olmayan bir aktivasyon fonksiyonu söz konusu ise lojistik regresyona karşılık gelmektedir.[64]

Yapay sinir ağları, denetimli öğrenme (supervised learning) yöntemini kullanmaktadır. Bu yönteme göre, hedef değişken de dahil olmak üzere, tüm veri kümesi sürece dahil edilmekte ve veri kümesindeki her kayda karşılık bir öngörü değeri hesaplanmaktadır. Daha sonra, hedef değişkenin ilgili kayda ait gerçek değeriyle hesaplanan öngörü değeri karşılaştırılarak farklılık belirlenir. Bu farklılık, öngörünün iyileştirilmesi için bağlantı ağırlıklarını yeniden hesaplamak üzere geriye doğru yayılır. Bu öğrenme yöntemine hatanın geri yayılımı (backpropagation of error) adı verilmektedir.[65]

Clementine, öğrenme sürecinin başında tüm bağlantı ağırlıklarını –0,5 ile +0,5 aralığında yer alacak şekilde rastlantısal olarak atamaktadır. Ardından, veri kümesindeki n adet kayıt rastlantısal olarak döngülere (cycles ya da epochs) sunulur. Ancak, öğrenme kümesinde n adet kayıt olmasına rağmen seçim rastlantısal olarak gerçekleşeceği için döngülere bazı kayıtların birden çok kez sunulması bazı kayıtların ise hiç sunulmaması söz konusu olabilmektedir.

Bir çok yapay sinir ağı modeli, öngörü hatasını değerlendirmek üzere,

$$SS_E = \sum_i \sum_k (y_i - \hat{y}_{ik})^2 \qquad (36)$$

şeklinde ifade edilen hata kareleri toplamını kullanmaktadır. Burada, kayıt sayısı i ve çıktı sayısı k ile gösterilmektedir.

Bağlantı ağırlıklarının optimal değerinin belirlenmesinde hata kareleri toplamının en küçükleştirilmesi söz konusu olduğundan, regresyon analizindeki katsayılarla bağlantı ağırlıkları arasında benzerlik söz konusudur. Ancak sigmoid fonksiyonun doğrusal olmayan yapısı nedeniyle, en küçük kareler kestirimine karşılık gelecek bir fonksiyonel çözüm (closed-form solution) bulunmamaktadır. Bu yüzden,

[64] Two Crows Co., *a.g.e.*
[65] Haykin S. , *a.g.e.*, sf.281

optimal çözümün sağlanmasından çok, optimal çözüme en yakın çözümün sağlanması söz konusu olmaktadır. Bu amaçla kullanılan başlıca yöntem, eğim düşümü (gradient descent) yöntemidir.[66]

Bu yöntemde amaç, bağlantı ağırlıklarının hata kareleri toplamını en küçükleştiren değerlerini belirlemektir. Sözgelimi, değeri belirlenmesi gereken m adet ağırlık bulunsun ve bunları bir vektör ($\mathbf{w}$) ile ifade edelim. Eğim düşümü yöntemi, hata karelerinin en küçükleştirilmesi için bu vektörün her bir değerinin hangi yönde değişmesi gerektiğini belirleyecektir. Hata kareleri toplamının ağırlık vektörüne göre gradyantı,

$$\nabla SSE(\mathbf{w}) = \left[\frac{\partial SSE}{\partial w_0}, \frac{\partial SSE}{\partial w_1}, \frac{\partial SSE}{\partial w_2}, \ldots\ldots\ldots, \frac{\partial SSE}{\partial w_m} \right] \qquad (37)$$

şeklinde ifade edilir. Bir başka deyişle, hata kareleri toplamının her bir ağırlığa göre kısmi türevlerinin bir vektörüdür.

Basit bir yaklaşımla, değeri belirlenecek tek bir bağlantı ağırlığı olduğunu varsayalım. Hata kareleri toplamının bağlantı ağırlığının değerine göre değişimini de bir parabol ile ifade edelim. Eğer söz konusu ağırlığın mevcut değeri optimal değere göre negatif yönde ise parabolün o noktadaki eğimini gösteren türevi negatif değer alacaktır. Bu sonuç, optimal değere yaklaşmak için ağırlığın mevcut değerinde pozitif yönde bir ayarlama gerektiği anlamına gelecektir. Aksi durumda ise türev değeri pozitif ve gereken ayarlama negatif yönde olacaktır. Sonuç olarak, ilgili ağırlığın mevcut değerinde gerçekleşmesi gereken ayarlamanın yönü, türevin işareti ile ters yönde olacaktır. Bu yüzden, ağırlığın mevcut değerindeki ayarlama,

$$\Delta w = -\left(\frac{\partial SSE}{\partial w} \right) \qquad (38)$$

şeklinde hesaplanacaktır. Eğim, bağlantı ağırlığının optimal değerden uzak olduğu noktalarda yüksek, yakın olduğu noktalarda ise düşük gerçekleşecektir. Ancak, burada hesaplanmış bulunan değişim ayarlamanın yönünü göstermekle birlikte, büyüklüğünü

[66] Larose D.T., *a.g.e.*, sf.135

belirlemede yeterli değildir. Bu yüzden, öğrenme oranı (η) ile çarpılarak,

$$\Delta w = -\eta\left(\frac{\partial SSE}{\partial w}\right) \tag{39}$$

şeklinde ifade edilir. Burada, öğrenme oranı $0 < \eta < 1$ olarak tanımlanır.

Öğrenme oranı, algoritmanın başlangıcında küçük bir değere eşit kabul edilirse, ağın yakınsaması kabul edilemez derecede uzun zaman alabilmektedir. Aksi halde ise ağın optimal sonucu atlaması söz konusu olabilir. Bir çözüm olarak, algoritmanın başlangıcında öğrenme oranı değeri görece yüksek belirlenerek, çözüm yakınsadıkça daha düşük bir değere ayarlanması önerilmektedir.[67]

Bu hesaplamaya dahil edilen bir başka parametre ise momentum terimi (α) olarak bilinmektedir. Momentum teriminin eklenmesi ile geri yayılım algoritması güçlendirilmekte ve hesaplama,

$$\Delta w = -\eta\left(\frac{\partial SSE}{\partial w}\right) + \alpha\,\Delta w' \tag{40}$$

şeklinde olmaktadır. Burada, momentum teriminin değeri $0 \le \alpha < 1$ ve bağlantı ağırlığının bir önceki ayarlaması $\Delta w'$ olmaktadır.

Aslında momentum terimi, ataleti simgelemekte ve aldığı görece büyük değerlerle bağlantı ağırlığındaki mevcut değişimin bir önceki değişimle aynı yönde hareket etmesini sağlamaktadır. Ayrıca geri yayılım algoritmasında momemtum terimi, önceki tüm ağırlık değişimlerinin üstel ortalamasının dikkate alınmasına neden olmaktadır. Öyle ki,

$$\Delta w_i = -\eta \sum_{k=0}^{\infty} \alpha^k \frac{\partial SSE}{\partial w_{i-k}} \tag{41}$$

ifadesinde yer alan α^k terimi mevcut ayarlamanın hesaplanmasında son dönem ayarlamalarının daha fazla dikkate alınmasını sağlar.

[67] Larose D.T., *a.g.e.*, sf.139

Clementine, momentum parametresini öğrenme süreci boyunca sabit kabul ederken, öğrenme parametresi (η) döngülerle birlikte alt sınır değeri ile üst sınır değeri arasında logaritmik olarak,

$$\eta(t) = \eta(t-1)\, e^{\log\left(\frac{\eta_{alt}}{\eta_{üst}}\right)\Big/ d} \qquad (42)$$

şeklinde değişmektedir. Parametrenin başlangıç değeri, alt sınır değeri ve üst sınır değeri uygulamacı tarafından belirlenebilmektedir. Eğer $\eta(t-1) < \eta_{alt}$ olursa, $\eta(t)$ değeri $\eta_{üst}$ değerine ayarlanmakta ve bu değişim öğrenme süreci boyunca devam etmektedir. Bu ifadede, yine uygulamacı tarafından belirlenebilen, öğrenme parametresi döngü sayısı d ile gösterilmiştir.

Clementine, geri yayılmış hata değerini ise bağlantının ağdaki konumunda bağlı olarak farklı ifadelerle hesaplanmaktadır. Çıkış hücrelerine olan bağlantılarda,

$$\delta_{pj} = (t_{pj} - o_{pj}) o_{pj} (1 - o_{pj}) \qquad (43)$$

ifadesi kullanılmaktadır. Burada, p'inci kayıt için j'inci çıkış hücresinin hedef değeri t_{pj} ile gösterilmiştir. Çıkış hücresine bağlı olmayan bağlantılarda ise,

$$\delta_{pj} = o_{pj}(1 - o_{pj}) \sum_{k} \delta_{pk} w_{kj} \qquad (44)$$

ifadesi kullanılmaktadır. Burada, j'inci hücrenin takipçisi konumundaki hücrelerin sayısı k ve k'ıncı hücrenin geri yayılmış hata değeri δ_{pk} şeklinde gösterilmiştir.

Yapay sinir ağları, herhangi bir durma kriteri (stopping criterion) sağlanana kadar öğrenmeyi ve öngörülerini iyileştirmeyi sürdürür. Eğer durma kriteri bir zaman problemi ise algoritma kullanıcının tanımlayacağı parametrelere göre sonlandırılır. Alternatif olarak, hata kareleri toplamının bir alt eşik değere ayarlanması da durma kriteri olarak belirlenebilir. Çoğu yapay sinir ağı uygulamaları aşağıdaki çapraz geçerlilik prosedürünü izlemektedir.[68]

[68] Haykin S. *a.g.e.*, sf.179

- ✓ Veri kümesinin bir kısmının geçerlilik amacıyla ayrılması,
- ✓ Yapay sinir ağının öğrenme kümesi üzerinde yürütülmesi,
- ✓ Öğrenme kümesi üzerinde elde edilen bağlantı ağırlıklarının geçerlilik kümesine uygulanması,
- ✓ Öğrenme kümesi üzerinde elde edilen "son" ağırlıklarla, hata kareleri toplamını en küçükleştirmiş "en iyi" ağırlıkların geçerlilik kümesi üzerinde izlenmesi,
- ✓ "Son" ağırlıkların ürettiği hata kareleri toplamı "en iyi" ağırlıkların hata kareleri toplamından daha büyük olduğunda algoritmanın durdurulması.

Durma kriterinin kullanılması, optimal sonucun elde edildiği anlamına gelmemektedir. Algoritma, hata kareleri toplamı için yerel en küçük değerde sonlanmış da olabilir. Bu durumda, sonuç optimal olmasa da "iyi" olacaktır.

Yapay sinir ağlarında öngörünün belirlenmesi, bağımlı değişkenin türüne göre farklılık göstermektedir. Sürekli bağımlı değişkenlerin öngörüsünde, modelin $[0,1]$ aralığındaki öngörülerinin gerçek değer aralığına dönüştürülmesi gerekecektir. Çoklu kategorik bağımlı değişkenlerde öngörü, çıkış aktivasyonu en yüksek olan kategori olmaktadır. İkili kategorik bağımlı değişken için ise çıktı değeri 0,5 etrafında değerlendirilmektedir. Eğer $o < 0{,}5$ ise öngörü 0.0, $o > 0{,}5$ ise öngörü 1.0 olarak belirlenmektedir.

Yapay sinir ağlarının en büyük dezavantajı, elde edilen modelin anlamlandırılması ve açıklanabilirliğinde yaşanan güçlük olarak gözükmektedir. Genellikle, süreçte kullanılan bağlantı ağırlıklarının ifade ettiği ilişkisel yapıyı anlamak mümkün değildir. Bu durum, sonuçların nasıl ortaya çıktığını kavramamızı imkansız hale getirir. Bu yüzden, sonuçların ortaya çıkışındaki fonksiyonel formdan çok, sonuçların kendisinin önemli olduğu durumlarda uygulaması uygun bulunmaktadır. Yapay sinir ağları, bu yapısından dolayı kara kutu (black box) olarak da anılmaktadır. Ancak, bir çok sinir hücresi barındırmaları nedeniyle, bozuk veri yapıları üzerinde dahi öğrenme kabiliyetine sahiptirler. Bu açıdan, veri madenciliğinin sınıflandırma fonksiyonu için kullanımı giderek yaygınlaşmıştır.

Bununla birlikte, duyarlılık analizi (sensitivity analysis) yolu ile hangi girdi değişkenlerinin öngörüleri hangi ölçüde etkilediğini belirlememiz mümkün olabilmektedir. Duyarlılık analizi, uygulamacıya hangi girdi değişkenlerinin modelin çıktısı üzerinde ne oranda etkili olduğunu anlama imkanı vermektedir. Duyarlılık analizi, aşağıdaki prosedürü izlemektedir.[69]

- ✓ Her bir girdi değişkenin ortalama değerlerinden oluşan yeni bir kayıt üretilmesi,
- ✓ Bu kayıt üzerinden ağın ortalama çıktısının hesaplanması,
- ✓ Her bir girdi değişkenin ortalama değeri yerine, değişkenin en küçük değerini ve en büyük değerini vererek, ortalama çıktıda gerçekleşen değişkenliğin bulunması ve karşılaştırılması.

Elde edilen modelin değerlendirilmesi aşamasında, sınıflandırma amaçlı yapay sinir ağlarında kullanılabilecek en önemli ölçüt kesinlik (accuracy) olarak görülmektedir. Kategorik bağımlı değişkenlerde kesinlik, model tarafından doğru olarak öngörülen kayıt sayısının toplam kayıt sayısına oranını ifade etmektedir. Bu oran, uygulamacı tarafından belirlenen bir seçenekle, öğrenme kümesi ya da sınama kümesi üzerinden hesaplanabilmektedir. Kesinliğin sınama kümesi üzerinden hesaplanması, aşırı uyum (overfitting) sorunundan kaçınmak açısından daha uygun olmaktadır.

Yapay sinir ağları, veri madenciliğinin yapay öğrenme bileşeni açısından en yaygın kullanılan araçları haline gelmektedir. Sınıflandırma probleminin yanı sıra, tahmin ve ilişki analizi gibi amaçlarla da kullanılabilmektedir. Çok katmanlı algılayıcı ise yapay sinir ağı algoritmalarından sadece biridir. Bunun dışında, Radial Basis Function, Recurrent Networks, Self – Organizing Systems gibi bir çok alternatif algoritma da mevcuttur.

[69] Larose D.T., *a.g.e.*, sf.142

1.4.3. C5.0 Algoritması

Ağaç yapısı benzeri akış şemaları olarak bilinen karar ağaçları (decision trees), kök düğüm (root node) ile başlamakta, dallarla (branches) birbirlerine bağlanan iç düğümleri (internal nodes) takip eden yaprak düğümleri (leaf nodes) ile son bulmaktadır. Her bir düğüm, bir değişken üzerinde gerçekleştirilen sınamayı, her bir dal ilgili değişken üzerinde gerçekleştirilen bölümlemeleri ve her bir yaprak da öngörünün sonucunu ifade etmektedir.[70]

Bir başka deyişle karar ağaçları, bir kategori ya da değerle sonuçlanan kural serilerini betimleme yöntemidir. Kategorik bağımlı değişkenin kullanıldığı karar ağaçlarına sınıflandırma ağaçları (classification trees), sürekli bağımlı değişken kullanılanlara ise regresyon ağaçları (regression trees) adı verilmektedir.[71]

Karar ağaçları örüntülerin bir takım kurallarla ifade edildiği ve her bir kuralın kendisinden önceki kurala bağlı olduğu sezgisel yöntemlerdir. Ağaç modelleri, üretilen alt veri kümelerinin saflığını (purity) en büyükleştirmeye dayanan çeşitli kurallarla, aşamalı bölümleme gerçekleştiren bir prosedür olarak tanımlanmaktadır. Bu prosedürün her aşamasında, bölümlemenin gerçekleştirileceği bir sınama değişkeninin belirlenmesi ve bu değişken için bir bölme kuralının türetilmesi söz konusudur. Algoritmaya bağlı olarak, her bir düğüm iki ya da daha fazla dala ayrılmakta ve bu yapılara sırasıyla ikili (binary) ve çok yollu (multiway) ağaç adı verilmektedir.[72]

Karar ağacı uygulamalarının başlangıç aşamasında bir takım ön çalışmaların yapılmasında fayda görülmektedir. İlk olarak, karar ağaçları denetimli öğrenme yöntemini kullandıklarından, yüksek öngörü başarısının elde edilmesi için veri kümesinin yeterli büyüklükte ve değişkenlikte olması gerekmektedir. Bununla birlikte, bağımlı değişkenin dağılımının bilinmesi, asimetri ve aykırı değer (outliers) sorunlarının tespiti, değişkenin şıklarının belirlenmesi gibi tanımlayıcı analizlerin gerçekleştirilmesi hem model seçimi hem de öngörü başarısı konusunda olumlu katkılar doğuracaktır. Ve son olarak, en yüksek başarı için en doğru algoritmanın

[70] Han J., Kamber M., *a.g.e.*, sf.284
[71] Two Crows Co., *a.g.e.*
[72] Guidici P. *a.g.e.*, sf.100

kullanılması gereğinden hareketle, hedefe ve veri yapısına uygun algoritmanın seçimi önem kazanacaktır. En uygun algoritmanın seçilmesinde ise düğümlerde kullanılan bölümleme kriterleri ve ağaç boyutunun küçültülmesine dayanan budama (pruning) yöntemi dikkate alınması gereken iki temel unsur olarak öne çıkmaktadır.[73]

Bölümleme kriterlerinin en çok bilinenleri, bilgi kazanımı (information gain) ya da entropi (entropy), kazanım oranı (gain ratio), gini endeksi (the gini index), ki-kare kontenjans tablosu istatistiği (chi-square contengency table statistic) ve G istatistiği (G-statistic) olarak sıralanmaktadır.

Budama yöntemleri için ise ön budama (prepruning) ve son budama (postpruning) olmak üzere iki yaklaşım söz konusudur. Ön budama yaklaşımı ile bölümleme, ağacın henüz oluşturulması aşamasında durdurulması temeline dayanmaktadır. Bölümlemede kullanılan istatistiksel ölçü için belirlenecek bir eşik değer (threshold) sayesinde bu işlem gerçekleştirilmekte ve bölümlemenin durdurulduğu düğüm bir yaprak düğüm olmaktadır. Ancak en uygun eşik değerin belirlenmesi de oldukça karmaşık ve zor bir karar olacaktır. Son budama yaklaşımı ise karar ağacı oluşturulduktan sonra belirli bir düğümün devamı konumundaki alt ağacın budanması ve budanmaması durumlarında gerçekleşen öngörü hatalarının oranının kullanılmasına dayanmaktadır. Bazı durumlarda her iki yaklaşım birleştirilebilse de, güvenilirlik açısından son budama yaklaşımı önerilmektedir.[74]

Ağaç budama yönteminden farklı olarak, durdurma kuralları (stopping rules) adı verilen yöntemlerle ağacın derinliğini sınırlamak da mümkün olmaktadır. Bu sınırlama, ağaç derinliğinin üst sınırını ya da bir düğümde yer alacak kayıt sayısının alt sınırı belirleyerek gerçekleştirilmektedir.[75]

Başlıca karar ağacı ve karar türetme algoritmaları arasında AID (Automatic Interaction Detector; Morgan J.N., Sonquist J.A. 1963), CHAID (ChiSquared Automatic Interaction Detection, Kass G. 1980), CRT (Classification Regression Trees Breiman L. 1984), ID3 (Quinlan J. 1979) yer almaktadır.

---

[73] Larose D.T., *a.g.e.*, sf.109
[74] Han J., Kamber M., *a.g.e.*, sf.290
[75] Two Crows Co., *a.g.e.*

C4.5 ve C5.0 (Quinlan 1993) ise ID3 algoritmasının geliştirilmiş versiyonları olup, ilk versiyonları kategorik veri türleri ile sınırlı olmakla birlikte, sürekli veri türleri üzerinde de uygulama olanağı sağlamaktadır.

SLIQ* ve SPRINT** gibi son dönem karar ağacı türetme yöntemleri ise hem kategorik hem de sürekli değişkenlerle yürütülebilmekte ve görece daha fazla sayıda veri üzerinde verimli olabilmektedir.

Her karar ağacı algoritmasının bir takım özgünlükleri söz konusudur. Örneğin, bir regresyon ağacı olan CART algoritması ile C5.0 algoritması arasındaki en önemli farklılık; CART algoritmasının ikili bölümlemeye, C5.0 algoritmasının ise çoklu bölümlemeye dayanmalarıdır. Bir diğeri ise bölümleme kriteri olarak kullanılan yöntemdeki farklılıktır. CART ve C5.0 algoritmaları arasında detaylarda mevcut olan farklılıklara rağmen her iki algoritma ile yaklaşık olarak aynı düzeyde bilginin edinilmesi mümkün olmaktadır. Sözgelimi, her iki algoritma da bölümleme için aynı değişkenleri anlamlı bulurken, bu değişkenlerin sadece önem sırasını farklı hesaplamaktadır.[76]

Çalışmamın ikinci bölümünde yer alan C5.0 uygulaması Integral Solutions Limited tarafından geliştirilmiş olan Clementine 8.0 yazılımı ile gerçekleştirilmiştir. Ancak ilgili algoritmanın hakları RuleQuest Reseach[77] tarafından saklı tutulduğu için çalışma esasları konusunda ayrıntılara ulaşmak için ilgili kuruluşun internet adresine ulaşılması önerilmektedir.[78]

Clementine 8.0 üzerinde uygulanabilen Unix tabanlı C5.0 algoritması, Windows tabanlı benzeri olan See5 algoritmasında olduğu gibi, veri kümelerinin önceden belirlenmiş kategorilere atanması amacıyla, kategorileri tanımlayan örüntüleri keşfetme, bu örüntüleri sınıflayıcılara dönüştürme ve bu sınıflayıcıları kullanarak öngörülerde bulunma esasına dayalı olarak kurgulanmıştır. C5.0 algoritmasının çalışma prensiplerine ilişkin olarak aşağıda aktarılan bölümün hazırlanmasında

---

* Supervised Learning in Quest
** Scalable Parallelizable Induction of Decision Trees
[76] Larose D.T., *a.g.e.*, sf.122
[77] RuleQuest Research, *http://www.rulequest.com*
[78] Integral Solutions Limited, Clementine 8.0 Algorithms Guide, 2003

RuleQuest Research tarafından sunulan çevrim içi eğitim dokümanından yararlanılmıştır.[79]

Uygulamaya başlarken bir hedef değişkenin belirlenmesi gerekmektedir. Bu hedef değişken, kayıtların ayrıştırılacağı sınıfların tanımlayıcısı olup sürekli bir değişken de olabilmektedir. Ancak, hedef değişkenin sürekli olması halinde t adet eşik değerin belirlenmesi gerekecektir. Öyle ki, bu eşik değerleri sayesinde sınıfların birbirinden ayrılması sağlanacak ve $t+1$ adet sınıf tanımlanmış olacaktır.

C5.0 uygulamasına konu olan veri kümesindeki değişkenlerin kesikli, sürekli, tarihsel ve etiket niteliğinde açık tanımlı (explicitly-defined) ya da formüllerle türetilmiş şekilde örtülü tanımlı (implicitly-defined) olması mümkündür. Etiket niteliği, her bir kayıt için sıra numarası gibi benzersiz tanımlayıcı değişken tipini ifade etmektedir. Formüllerle değişken türetme işlemi ise mevcut değişkenlerin azaltılmasına ya da arttırılmasına olanak tanımaktadır.

C5.0 algoritması, yüzlerce değişken ile yüz binlerce kayıt barındıran veri depoları üzerinde uygulanabilmektedir. Hem karar ağacı hem de karar kuralları formunda sonuç üretebildiği için yorumlanması, istatistik ya da yapay öğrenme konusunda uzmanlık gerektirmediği için kullanımı son derece kolay bir sınıflandırma aracıdır.

C5.0 karar ağacı türetme algoritması, tek bir düğümle başlamakta ve en uygun sınıflayıcının tespiti için bilgi kazanımı adı verilen buluşsal (heuristic) ve entropi tabanlı bir ölçü kullanmaktadır. Bu ölçüye göre belirlenen değişkenin her bir değeri dallara dönüşmekte ve devam eden süreçte kalan değişkenlerin dikkate alınması ile aynı işlem sürdürülmektedir. Bu süreç, belirli bir düğümdeki tüm örneklerin aynı sınıfa ait olması, bölümlenecek yeni bir değişkenin kalmaması ve veri kümesinde sınama değişkeninin söz konusu değerine sahip kayıt bulunmaması durumlarında son bulmaktadır.

[79] RuleQuest Research, "C5.0 An Informal Tutorial", *http://www.rulequest.com/see5-unix.html*, [Erişim 17.02.2008]

Bilgi kazanımı ve benzerleri, değişken seçimi ölçüsü (attribute selection measure) ya da bölme iyiliği ölçüsü (measure of the goodness of split) olarak da adlandırılmaktadır. En yüksek bilgi kazanımına, bir başka deyişle en büyük entropi azalımına sahip olan değişken ilgili düğüm için sınama değişkeni olarak belirlenmektedir.[80]

Herhangi bir düğümde, m farklı değere ve bu değerlerle tanımlanan $T_1, T_2, \ldots\ldots, T_m$ alt veri kümelerine sahip olası bir sınama değişkeninin (X) seçilmesi sorunu ile karşı karşıya olduğumuzu düşünelim. Herhangi bir örnek kümesi S için, bağımlı değişkenin k adet farklı sınıfı üzerinden C_i sınıfına ait kayıt sayısını $f(C_i, S)$ ve aynı örnek kümesindeki toplam kayıt sayısını $|S|$ ile göstermek üzere, S veri kümesinin bölümlenmesi için ihtiyaç duyulan bilginin beklenen değeri (expected information), yani entropisi,

$$Info(S) = -\sum_{i=1}^{k}\left(\left(f(C_i, S)/|S|\right)\log_2\left(f(C_i, S)/|S|\right)\right) \qquad (45)$$

şeklinde hesaplanmaktadır. Benzer bir yaklaşımla, X değişkeni tarafından gerçekleştirilecek bölümleme sonrası ihtiyaç duyulan bilginin beklenen değeri ise,

$$Info_x(T) = -\sum_{i=1}^{m}\left(\left(|T_i|/|T|\right)Info(T_i)\right) \qquad (46)$$

olmaktadır. Sonuç olarak, X değişkeni tarafından gerçekleştirilen bölümleme ile edinilen bilgi kazanımı,

$$Gain(X) = Info(T) - Info_x(T) \qquad (47)$$

olarak hesaplanmış olacaktır. Bu kritere göre, en yüksek bilgi kazanımını veren değişken bu düğümde sınama değişkeni olarak seçilmiş olur.[81]

Karar ağaçları hakkındaki en yaygın eleştiri, bir düğümdeki bölme kararı verilirken gelecek bölme işlemlerini dikkate almayan "açgözlü" bir algoritma izlendiği

[80] Han J., Kamber M., *a.g.e.*, sf.286
[81] Kantardzic M., *a.g.e.*, sf.144

şeklindedir. Çünkü gelecek tüm bölme işlemleri ilk bölme işlemini koşul kabul ederek gerçekleşecektir. Bununla birlikte, bölümleme işlemi ölçüleri genellikle tek değişkeni dikkate almaktadır. Tek değişkenli bölümleme ile sınırlandırılmamış karar ağaçları ise yine tek değişkenli bölümleme kurallarını kullanmakta ancak birden çok öngörü değişkeninin lineer kombinasyonunu dikkate almaktadır. Bu tür karar ağaçları, dolaylı ağaçlar (oblique trees) olarak bilinmektedir.[82]

C5.0 uygulaması, bir çok sınıflandırma uygulamasında olduğu gibi, veri kümesinin öğrenme ve sınama amaçlı olarak bölümlenmesine olanak tanımaktadır. Bu sayede, hem inşa edilen sınıflayıcının öğrenme kümesi üzerindeki kesinliğini sınamak hem de algoritmanın öğrenme kümesi üzerinde aşırı uyum göstermesini engellemek mümkün olmaktadır.

Algoritmanın gerçekleştirilmesi ile elde edilen çıktıda ilgili öngörü için n ya da n / m şeklinde bir ifade yer almaktadır. Sadece n ifadesi, bu karar kuralına uyan kayıt sayısını ve n/m ifadesi ise bu karar kuralına uyan n adet kayıt içerisinde yanlış sınıflandırılan m adet kayıt bulunduğunu göstermektedir. Eğer hem n hem de m değeri olması gerektiği gibi bir tamsayı ifade etmiyor ise, bu karar kuralına uyan kayıp veri değerlerinin olduğu anlaşılacaktır.

C5.0, hem öğrenme kümesi üzerinde kendisini değerlendirmekte ve yanlış sınıflandırma sayısı üzerinden hata oranını vermekte hem de tanımlanmış bir sınama kümesi söz konusu ise aynı değerlendirmeyi bu veri kümesi üzerinde tekrar etmektedir. Eğer sınıf sayısı yirmiden fazla değilse, çıktı aynı zamanda yanlış sınıflandırmaların ne şekilde gerçekleştiğini gösteren bir hata matrisi (confusion matrix) de üretmektedir. Sınıf sayısı yirmiden fazla olduğunda ise bir özet raporu sunulmaktadır. Bu raporda, öngörülerin hatalı olumlama (false positive) ve hatalı olumsuzlama (false negative) şeklindeki sınıflaması yer almaktadır. Herhangi bir sınıf için hatalı şekilde öngörülmüş olan kayıtlara hatalı olumlama ve bu sınıfta yer alması gerekirken bir başka sınıfta öngörülen kayıtlara hatalı olumsuzlama adı verilmektedir.

[82] Two Crows Co., *a.g.e.*

Bu şekilde sunulan özette, her sınıfın kendisine ait hatalı öngörüleri ile bir başka sınıfa düşen hatalı öngörüleri yer alacağından, gerçek hatalı sınıflama sayısı mevcut sayının yarısı kadar olacaktır. Ayrıca, her bir değişkenin sınıflama işlemine ne oranda katkıda bulunduğunu gösteren bir çıktı almak da mümkündür. Bu çıktıda herhangi bir değişken için verilen oran, öğrenme kümesinin yüzde kaçlık bölümünün sınıflandırılmasında ilgili değişkenin bilinen değerlerinin kullanıldığını ifade etmektedir.

Öğrenme kümesinde sayıca daha fazla bulunan ve bu yüzden öngörülme oranı daha yüksek olan sınıf, sınama kümesinde de kaçınılmaz olarak ağırlık kazanacaktır. Ek olarak, her ne kadar sınama kümesinde yüksek oranda bir öngörü başarısı arzu edilse de, bu başarının öğrenme kümesindekine oranla daha düşük gerçekleşmesi şaşırtıcı olmayacaktır.

C5.0 algoritması budama yönteminin kullanımına da olanak tanımaktadır. Algoritma, öncelikle veri kümesine en uygun sınıflayıcıları üretmekte ve ardından görece yüksek öngörü hatalarına sahip parçaları budamaktadır.

C5.0 algoritması kök düğümden yaprak düğüme uzanan akışla karar kuralları şeklinde de ifade edilebilmektedir. Özellikle çok büyük karar ağaçları yerine karar kurallarını kullanmak insan algısına daha uygun düşmekte ancak veri sayısı arttıkça işlem süresi uzamaktadır. Bir karar ağacında, ağacın alt seviyelerindeki değişken üst seviyelerdeki değişkene oranla daha az kullanılırken, karar kurallarında kuralın sıralamasının önem taşımamaktadır. Ayrıca, karar kuralları öngörü kesinliklerine göre sıralanabilmekte ve öngörü kesinliğini arttırmak amacıyla budanabilmektedir.

Her karar kuralı, sadece kuralı tanımlamayı amaçlayan bir kural numarası ile ifade edilmektedir. Her kural numarasından sonra, karar ağacı formunda olduğu gibi kurala uyan kayıt sayısını gösteren n ve bu kural tarafından hatalı öngörülen kayıt söz konusu ise n/m ifadesi yer alacaktır. Ayrıca kaldıraç (lift) adı verilen bir oran da kuralı değerlendirmek amacıyla kullanılmaktadır. Kuralın kesinliği, $(n-m+1)/(n+2)$ şeklinde hesaplanan Laplace oranı ile belirlenmektedir. Kaldıraç ise kuralın kesinliğinin öngörülen sınıfın öğrenme kümesindeki göreli sıklığına

bölümü şeklinde hesaplanmaktadır. Her kural, bir öngörü üretmekte, her öngörü ifadesinin sonunda ise değeri $[0,1]$ aralığında yer alan güven (confidence) değeri yer almaktadır. Güven değeri, doğru olarak öngörülen kayıt sayısının toplam kayıt sayısına oranını ifade etmektedir. Hiçbir kuralın geçerli kayda karşılık gelmemesi durumunda öngörü, algoritmada varsayılan sınıf olarak belirlenen sınıfın değeri olarak gerçekleşmektedir.

C5.0 algoritmasının son sürümünde, Schapire R. ve Freund Y. tarafından geliştirilen uyarlamalı güçlendirme (adaptive boosting) yöntemi uygulamaya dahil edilmiştir. Bu yöntem, çeşitli sınıflayıcıların kullanılmasına, bu sınıflayıcıların her bir yeni kaydı oylamasına ve öngörünün oylamaların sayılması sonucu üretilmesine dayanmaktadır. Burada, aynı veri kümesinde çeşitli sınıflayıcıların ne şekilde üretileceği önem kazanmaktadır. Öncelikle, veri kümesi üzerinde sınıflayıcıların belirlenmesi gerçekleşmekte ve ardından bu sınıflayıcıların hatalı öngörülerde bulunduğu kayıtlara odaklanan yeni sınıflayıcılar üretilmektedir. Bu süreç, önceden belirlenmiş döngü sayısına ulaşılana kadar devam etmektedir.

C5.0 algoritması, değişkenlerin ayrıştırılmasını sağlamakta ve değişkenlerin göreli önemini de hesaplayabilmektedir. Değişkenlerin göreli önemi, ilgili değişkenin hariç tutulması durumunda hatalı öngörü oranının yüzdesel artışı ile verilmektedir. Değişkenlerin ayrıştırılması ise işlem süresini uzatmakla birlikte çok büyük veri kümeleri için önerilmektedir.

C5.0 algoritmasına dahil edilen sürekli değişkenlerin belirli eşik değerlerle bölümlenmesinden yukarıda bahsedilmişti. Bu bölümleme, uygulamacı tarafından belirlenen eşik değerlere göre olduğu için şüpheli olacaktır. Hedef değişkenin en uygun öngörüleri için bu bölümlemenin kesin değerlerle yapılması yerine farklı eşik değerlerin denenmesi kuşkusuz daha uygun olacaktır. Bu nedenle, C5.0 algoritması, eşik değerler etrafında denemelerde bulunabilmekte ve en uygun eşik değeri belirleyebilmektedir. Bu işleme, eşik değerin yumuşatılması (softening thresholds) adı verilmekte ve bu işlem sadece karar ağaçlarında geçerli olmaktadır.

C5.0 algoritması ile veri kümelerinin çok büyük olduğu durumlarda sınıflayıcıların belirlenmesi sürecini kolaylaştıran ancak öngörü başarısı üzerinde

olumsuz etkisi kaçınılmaz olan örnekleme yöntemini kullanmak da mümkün olmaktadır. Bu yöntemde, sınıflayıcıların rastsal bir alt veri kümesi üzerinde üretilmesi ve bir başka rastsal alt veri kümesi üzerinde sınanması sağlanmaktadır. Örnekleme rastlantısal gerçekleştiği için bu yöntemin ardışık her denemesi farklı bir sonuç üretecektir. Bu durumdan kaçınmak için sabit bir kaynak değeri (seed) belirlenerek her seferinde aynı oranda örneklemenin yapılması yeterli olacaktır.

C5.0 algoritmasında sınıflayıcıların sınanması için bir sınama kümesi kullanmak yerine N katlı çapraz geçerlilik (N-fold cross-validation) yönteminin uygulanmasına da mümkün olmaktadır. Bu yöntem, veri kümesini f adet alt kümeye ayırarak, sınıflayıcıların her bir alt küme üzerinde sınanmasını sağlamaktadır. Gerçek öngörü kesinliği ise her sınama kümesinde gerçekleşen hatalı öngörü sayısının aynı kümedeki kayıt sayısına oranı ile ifade edilen hata oranlarının ortalaması olarak ifade edilecektir. Burada, hatalı öngörü sayılarının standart sapması da hatalı öngörü sayılarında gerçekleşebilecek değişkenliğin ölçüsü olarak dikkate alınmaktadır. Çapraz geçerlilik uygulaması rastsal alt veri kümeleri üzerinde gerçekleştirildiğinden, sabit bir başlangıç noktasının belirlenmesinde fayda olacaktır. Ayrıca, bu uygulamada üretilen sınıflayıcıların veri kümesinin tamamını temsil etmemesinden dolayı kaydedilemediği belirtilmektedir. Eğer sınıflayıcıların daha sonra yeniden kullanılması söz konusu ise çapraz geçerlilik önerilmemektedir.

Buraya kadar, hatalı öngörülerin tamamının aynı öneme sahip olduğu yaklaşımı dikkate alınmıştır. Ancak bazı durumlarda, belirli bir hatalı sınıflamanın göreli önemi diğer hatalı sınıflamalardan daha fazla olabilmektedir. Bu nedenle, C5.0 algoritması, en çok kaçınılan hatalı öngörülerin tanımlanarak sınıflayıcıların bu öngörülerde yoğunlaşmasına olanak tanıyabilmektedir. Bu şekilde yapılan tanımlamaya hatalı sınıflama maliyeti (misclassification cost) adı verilmektedir.

C5.0 ile gerçekleştirilebilen bir başka uygulama ise kayıtlara farklı ağırlıkların atanmasıdır. Bu sayede, belirli nitelikteki kayıtlarda hatalı öngörü oluşma olasılığını düşürmek mümkün olmaktadır. Bu işlem, her bir kaydın ağırlığını tanımlayan yeni bir değişken ile gerçekleştirilmekte ve bu değişken sınıflayıcıların üretilmesi sürecine katılmamaktadır.

1.4.4. Sınıflandırma Yönteminin Seçimi

Birden çok model oluşturulması ve en uygun modelin elde edilmesi veri madenciliği projeleri için önemli başarı göstergelerinden biri olarak kabul edilmektedir. Bu açıdan, farklı yöntemlerin ve bu yöntemler ile geliştirilen modellerin karşılaştırılması sorunu gündeme gelmektedir.

Sınıflandırma tekniğinin seçiminde dikkate alınan kriterler hız, güçlülük, ölçeklenebilirlik, açıklanabilirlik ve öngörü kesinliği şeklinde sıralanabilmektedir.[83]

Hız, modelin üretilmesi ve kullanılması için gereken işlemsel maliyetlere vurgu yapmaktadır. Güçlülük, veri setinde mevcut olan gürültü ve kayıp veri sorunlarında dahi doğru öngörülerde bulunabilme becerisini ifade etmektedir. Ölçeklenebilirlik, büyük miktarlardaki verilerle modelin verimli olarak inşa edilebilmesi olarak değerlendirilebilir. Açıklanabilirlik, modelin anlaşılma ve kavranma düzeyine karşılık gelmektedir. Öngörü kesinliği, modelin yeni girdileri için doğru öngörülerde bulunabilme becerisini ifade etmektedir.

Uygun yöntemin seçiminde değerlendirme aşamasındaki basit geçerlilikten konuşlandırma aşamasındaki yatırımın geri dönüş oranına kadar çok sayıda teknik söz konusudur. Bu teknikleri aşağıdaki gibi sıralayabiliriz. [84]

Bir modelin sınanmasında yaygın olarak basit geçerlilik (simple validation) yöntemi kullanılmaktadır. Bu yöntemle, öğrenme kümesi üzerinde oluşturulan model sınama kümesine uygulanmaktadır. Bu uygulamada, doğru olarak sınıflandırılan kayıtların tüm kayıtlara oranı modelin kesinliği (accuracy) ve yanlış olarak sınıflandırılan kayıtların tüm kayıtlara oranı ise modelin hatası (error) olarak kabul edilir. Basit geçerlilik yönteminde veri kümesinin yaklaşık olarak %5 ile %33 arasındaki kısmının sınama amaçlı olarak ayrılması önerilmektedir.

Veri sayısının sınırlı olması nedeniyle veri kümesinin bir bölümünün sınama amacıyla ayrılması mümkün olmadığında çapraz geçerlilik (cross validation) yönteminin uygulanması söz konusu olabilmektedir. Bu yöntemde, veri kümesi

[83] Han J., Kamber M., *a.g.e.*, sf.283
[84] Akpınar H., *a.g.e.*

öncelikle iki eşit gruba ayrılır. İlk aşamada, model birinci gruptaki verilerle oluşturularak ikinci gruptaki verilerle sınanır. İkinci aşamada ise model diğer veri grubu ile oluşturularak ilk veri grubu üzerinde sınanır. Bu durumda, iki bağımsız hata oranı elde edilmiş olacaktır. Modelin kesinliği iki farklı veri kümesinde elde edilen oranların ortalaması olarak kabul edilir.

Benzer şekilde, veri kümesini ikiden çok alt gruba ayrılarak çapraz geçerlilik uygulaması gerçekleştirmek de mümkündür. Bu yönteme, n katlı çapraz geçerlilik (n fold cross validation) adı verilmekte ve n adet veri kümesi üzerinde elde edilen n adet model, geriye kalan veri kümesi üzerinde sınanmaktadır. Modelin kesinliği ise yine tüm modellerde elde edilen oranların ortalaması olmaktadır.

Veri sayısı kısıtlı olduğunda kullanılan bir başka yöntem ise önyükleme (bootstrapping) olarak bilinmektedir. Önyükleme yönteminde, orijinal veri kümesi içerisinden rastlantısal olarak çok fazla sayıda örnek veri kümesi (bootstrap samples) türetilerek model oluşturma amacıyla kullanılır. Her bir rastlantısal örnek (bazen 1000'den fazla) veri kümesi üzerinden elde edilen oranların ortalaması yolu ile modelin kesinliği hesaplanmış olur.

Kaldıraç (lift) oranı da modeli değerlendirmek için kullanılabilmektedir. Bu oran, konuşlandırma sürecinde gerçekleştiği tespit edilen davranışların model tarafından öngörülen sayısının model dikkate alınmadan rastlantısal olarak belirlenen örneklerde gerçekleşen sayısına oranı şeklinde ifade edilebilir. Ayrıca, modelin kullanılması ile elde edilen kazancın ilgili modelin oluşturulması için katlanılan maliyete oranı şeklinde ifade edilebilecek olan yatırımın geri dönüş oranı da (return on investment) modelin değerlendirilmesi için geçerli bir araçtır.

2. BANKACILIK MÜŞTERİ VERİ TABANI ÜZERİNDE BİR UYGULAMA

Bu bölümde, veri madenciliğinin sınıflandırma fonksiyonuna ilişkin teknikleri ulusal bir bankanın müşteri veri tabanından rastlantısal olarak elde edilen bir örnek veri kümesi üzerinde uygulanacaktır. Bu şekilde, hem veri madenciliği süreci gerçekleştirilmiş ve veri madenciliğinin sınıflandırma problemine ilişkin teknikleri uygulanmış olacak hem de uygulamada ortaya çıkan sorunlar tartışılarak uygulanan tekniklerin karşılaştırılması yoluna gidilecektir.

2.1. AMAÇ, KAPSAM VE YÖNTEM

Çalışmanın amacı, veri madenciliğinin sınıflandırma fonksiyonuna ilişkin tekniklerini gerçek yaşam verileri üzerinde uygulamak, veri madenciliği sürecini tüm aşamaları ile gerçekleştirmek, uygulamada elde edilen sonuçlarla tekniklere ilişkin farklılıkları tartışmak ve hangi tekniğin hangi koşullarda uygulanmasının daha uygun olacağına yönelik önerilerde bulunmaktır.

Çalışma, ulusal bir bankanın müşteri veri tabanından rastlantısal olarak elde edilen örnek veri kümesi üzerinde uygulanacak ve çok geniş bir çalışma alanı olan veri madenciliğinin sınıflandırma fonksiyonuna ilişkin teknikleri ile sınırlandırılacaktır. Bu kapsamda, veri madenciliğinin sınıflandırma teknikleri arasında ölçeklenebilirlikleri ve yaygınlıkları dikkate alınarak ileri beslemeli yapay sinir ağları, C5.0 karar kuralı türetme algoritması ve lojistik regresyon analizi kullanılacaktır.

Uygulamaya konu olan veri kümesi, ulusal bir bankanın veri tabanından rastlantısal olarak elde edilen 17.595 müşterinin 188 farklı değişkene ilişkin değerlerini içermektedir. Uygulamada CRISP–DM konsorsiyumu tarafından önerilen veri madenciliği standart süreci esas alınacaktır. Söz konusu veri kümesi üzerinde Microsoft Access 2000, Microsoft Excel 2000, SPSS for Windows 13.0 ve Clementine 8.0 yazılımları kullanılarak veri madenciliğinin sınıflandırma problemine ilişkin teknikleri için bir örnek uygulama ortaya konacaktır.

2.2. İŞİ ANLAMA

Bu ilk aşamada, uygulamaya başlamadan önce iş perspektifi açısından sürecin ana hatlarıyla tanımlanması için ulaşılmak istenen iş hedeflerinin belirlenmesi, mevcut durumun değerlemesi, veri madenciliği amaçlarının belirlenmesi ve proje planının hazırlanması görevleri üzerinde durulacaktır.

Bilindiği gibi Kasım 2000 ve Şubat 2001 dönemlerinde Türkiye ekonomisi önemli krizler yaşamış ve bu kriz dönemi sonrasında hem ülke ekonomisi hem de bankacılık sektörü önemli sorunlarla karşı karşıya kalmıştır. Kriz sonrası dönemde bazı kuruluşların bankacılık faaliyetleri durdurulmuş, önemli bir kısmı yabancı yatırımcılardan oluşmak üzere çok sayıda hisse satışı ve şirket birleşmesi örnekleri yaşanmıştır.

Geride kalan dönem içerisinde uygulanan ekonomi politikaları neticesinde bankacılık sektörünün olumlu gelişmeler gösterdiği, sektörün bilanço ve özsermaye bazında yaşadığı büyümenin yanı sıra kredi kullanımı, şube sayısı ve personel sayısında yaşanan artışların da bunu desteklediği düşünülmektedir.[85]

Ancak 2007 yılına ilişkin yorumların önceki yıllara göre daha temkinli olduğu görülmektedir. Aşağıda sektörün durumuna ilişkin bir takım istatistikleri görmek mümkündür.[86]

2000 yılında faaliyet gösteren 79 banka varken 2007 yılı sonunda faaliyet gösteren 46 banka bulunmaktadır. Bunlar arasında yer alan 33 banka mevduat bankası, 13 banka ise kalkınma ve yatırım bankası olarak faaliyet göstermektedir. Ayrıca 18 mevduat bankası ile 4 kalkınma ve yatırım bankasının çoğunluk hisseleri yabancı sermayelidir.

2007 sonu itibariyle toplam kredilerde YTL payı %75, yabancı paraların payı %25 olmuştur. Yine toplam krediler içerisinde bireysel kredilerin payı %32 olurken

---

[85] Türkiye Bankalar Birliği, "Bankalarımız 2007", *http://www.tbb.org.tr/turkce/kitap2007/2007.asp*, [Erişim 05.06.2008]

[86] Türkiye Cumhuriyeti Merkez Bankası, "Finansar İstikrar Raporu", *http://www.tcmb.gov.tr/yeni/evds/yayin/finist/finist4.php*, [Erişim 18.07.2007]

bireysel kredilerde konut kredilerinin payı %34, kredi kartlarının payı %29 ve taşıt kredilerinin payı %6 şeklindedir. Kurumsal kredilerde ise en önemli payı imalat, ticaret ve hizmet sektörleri almıştır. İnşaat ve tekstil sektörünün payı ise azalma göstermiştir.

Faaliyet halindeki tüm bankaların toplam şube sayısı 2007 yılında 769 yeni şube eklenerek 7.618 şeklinde gerçekleşmiştir. Aynı dönemde personel sayısı ise benzer bir artışla 158.559 olmuştur. Personel sayısındaki artış daha çok kamu ve yabancı sermayeli bankalarda yaşanmıştır.

Son yıllarda tüketimde yaşanan hızlı artış, nakit kullanımından kredi kartı kullanımına kaymaya başlayan tüketim alışkanlığı ve düşen faizlerle birlikte kredi kullanımında gözlenen dikkate değer hareketlenme, yabancı sermayenin ilgisinden de anlaşılabileceği gibi ülkemizi bankacılık sektörü açısından bir cazibe ve dolayısıyla rekabet merkezi haline getirmiştir.

Bu rekabet ortamı, bankaların mevcut müşterini ve pazar paylarını koruma, yeni müşteri kazanma ve pazar payını arttırma gibi zorunluluklarının önemini de arttırmıştır. Bu rekabet içerisinde başarılı olmak, mevcut müşterileri tanımak, onlara istedikleri düzeyde hizmet sunmak ve mevcut müşterilerden öğrenilen bilgi ışığında potansiyel müşteriler için doğru yaklaşımlar geliştirmek ile mümkün olacaktır.

Bu bakış açısıyla, bankacılık müşteri veri tabanı üzerinde gerçeklenecek veri madenciliği uygulaması ile ulaşılmak istenen iş hedefi, mevcut müşterilerin bankacılık ürünlerine ilişkin davranışlarından hareketle potansiyel müşterilerin ihtiyaçları ve taleplerine yönelik ürünlerin sunulabilmesini sağlamak olarak belirlenmiştir.

Böylesi bir çalışmadan sektörün beklentisini, müşterilerin herhangi bir bankacılık ürününe sahip olma ya da olmama davranışlarına etki eden ilişkilerin belirlenmesi ve bu ilişkilerin modellenmesi olarak kabul etmekteyim.

Ancak bu çalışmanın temel amacı işe dönük bir başarı elde etmek olmadığından çalışmaya ilişkin işe dönük bir başarı kriteri belirlenmemiştir. Buna rağmen çalışmanın sonucunda kullanılan tekniklerle müşterilerin tercih ve davranışlarını temsil kabiliyeti yüksek modellerin elde edilebilmesini ve bu modelin potansiyel müşterin

davranışlarını öngörebilme konusundaki beklentilere cevap verebilecek nitelikte olması beklentisini taşımaktayım.

Bu çalışma herhangi bir ekip çalışması olmayıp kısıtlı olanaklarla gerçekleştirildiğinden ideal bir veri madenciliği projesi olmaktan çok, daha önce de belirtildiği gibi bir örnek uygulama niteliğindedir. Benzer şekilde, üzerinde çalışılan veri kümesi de büyüklük açısından ideal olmamakla birlikte asgari yeterlilikleri sağlamaktadır.

Uygulamada kullanılacak olan veri kümesi ulusal bir bankanın müşterilerine ait olduğundan hem müşterilerin kişisel bilgileri hem de bankanın ticari bir değeri olması itibariyle gizlilik içermektedir. Bu açıdan ilgili banka ve müşterilerine ilişkin bilgi düzeyleri paylaşılmayacaktır.

Her veri madenciliği uygulamasında olduğu gibi başlangıç düzeyinde varsayılan süreç ve iş hedeflerinin gerçeklenmesi bu çalışmada da çeşitli nedenlerle mümkün olmayabilir. Çalışmanın başarısı üzerindeki en önemli tehdit veri kümesinden kaynaklanabilecek sorunlardır. Daha önce de belirtildiği gibi veri kümesi ideal bir veri madenciliği uygulaması için beklenen nicelikte değildir. Bu durum tekniklerin uygulanmasına engel oluşturmasa da sonuçların elde edilmesinde problemlere neden olabilir. Mevcut veri miktarı aranan örüntülerin elde edilmesi için yetersiz kalabilir.

Ayrıca veri kümesinin ilgili bankanın veri tabanından rastlantısal olarak elde edildiği belirtilmişti. Bu işlemde rastlantısallığın arzu edilen seviyede olmaması sonuçların başarısını olumsuz etkileyebilir. Çünkü rastlantısal olarak yapılan seçim ile aranan örüntüleri barındırmayacak yapıda bir veri kümesi oluşmuş olabilir. Veri kümesinden kaynaklanabilecek bu türden sorunların yaşanması durumunda alternatif olarak uygulama daha fazla kaydın elde edilmesi ya da yeni bir veri kümesi edinilmesi yoluyla yenilenecektir.

Yine uygulamanın başarıya ulaşmasını geciktirebilecek unsurlardan biri de teknik bir takım sınırlamalardır. Uygulanmada mevcut donanım ve yazılımların kullanılması yeterli olmaz ise alternatif yazılım ve donanımların kullanılacaktır.

Yukarıda tanımlanmış olan işe dönük hedefler doğrultusunda ulaşılmak istenen veri madenciliği amacı, kullanılacak sınıflandırma teknikleri yolu ile bireysel bankacılık müşterilerinin davranış kalıplarını ortaya çıkaracak nitelikte öngörü modelleri geliştirmek olarak belirlenmiştir.

Veri kümesi üzerinde veri madenciliğinin sınıflandırma tekniklerinin tek tek gerçeklenmesi, geliştirilen öngörü modelleri ışığında bu tekniklerin ilgili amaç doğrultusunda kullanılabilirliğinin sınanması ve tekniklerin karşılaştırılması imkanının yaratılması çalışmanın veri madenciliği açısından diğer amaçlarını oluşturmaktadır.

Veri kümesi üzerinde ilgilenilen tüm tekniklerin başarı ile uygulanması ve sonuçları açısından bu tekniklerin karşılaştırılabilmesini sağlayacak ölçülerin elde edilmesi durumunda arzulanan veri madenciliği başarısı elde edilmiş olacaktır.

Kesin olmamakla birlikte uygulamanın yapılacağı ham veri kümesinden başlayarak uygulamanın sonlandırılmasına kadar olan süreçte yürütüleceği varsayılan görevlerin tümü çalışmanın ekinde sunulan (Ek.1) proje planında listelenmiştir. Veri madenciliği süreci yinelemeli bir süreç olduğundan bu planda değişikliklerin yapılması muhtemeldir. Bu yüzden, ilerleyen basamaklarda duyulacak ihtiyaçlar doğrultusunda proje planının dışına çıkılabilecek, planın yeniden düzenlenmesi ihtiyacı ve yapılacak değişiklikler dile getirilecek ancak her defasında revize edilmiş plan yeniden sunulmayacaktır.

Çalışmanın gerçekleştirilmesinde birden çok yazılımdan yararlanılacak olsa da veri madenciliği açısından temel yazılım olarak Integral Solutions tarafından geliştirilen Clementine 8.0 yazılımının kullanılmasına karar verilmiştir. Bu yazılımın seçilmesinde kullanıcı odaklı olması ve birden çok tekniğin uygulanmasına olanak sağlaması gibi kriterler dikkate alınmıştır.

Çalışma, veri madenciliğinin sınıflandırma problemiyle sınırlandırıldığından, uygulamada kullanılacak tekniklerin Clementine 8.0 tarafından gerçeklenebilen, en yaygın ve etkin veri madenciliği teknikleri olmasına özen gösterilmiştir. Bu çerçevede, veri madenciliğinin üç önemli bileşeni de dikkate alınarak, uygulama için istatistiksel

sınıflandırma tekniklerinden lojistik regresyon analizi, yapay öğrenme algoritmalarından ileri beslemeli yapay sinir ağları ve karar kuralı türetme algoritmalarından C5.0 algoritmasının kullanılması kararlaştırılmıştır.

2.3. VERİLERİ ANLAMA

Bu aşamada, belirlenen iş ve veri madenciliği hedefleri doğrultusunda veri kümesinin incelenmesi, sorunlarından arındırılması ve analize hazır hale getirilmesi için başlangıç verilerini elde etme, verileri tanımlama, verileri inceleme ve veri kalitesini sınama görevleri üzerinde durulacaktır.

Uygulamada kullanılacak olan ham veri kümesi Microsoft Access veri tabanında bulunmaktadır. İlk olarak buradaki veri kümesi uygun sorgularla birleştirilip Microsoft Excel çalışma sayfasına aktarılmıştır. Burada Excel'in seçilmesi, ilerleyen görevlerde ortaya çıkacak veri işleme ihtiyaçları doğrultusunda süzme, sıralama, özetleme ve silme/ekleme gibi pratik fonksiyonları desteklemesinden kaynaklanmaktadır.

Başlangıç verileri 188 değişken ve 17.595 kayıt içermekte, Excel çalışma sayfası biçimindeki dosya büyüklüğü ise yaklaşık olarak 30 Megabyte alana karşılık gelmektedir.

Satırlarda yer alan her bir kayıt, bir müşterinin ilgili sütunlardaki değişkenlere karşılık gelen değerlerinin birleşimini ifade etmektedir. Sütunlarda yer alan her bir değişken ise tüm müşterilerin aynı ölçüme karşılık gelen değerlerinin topluluğunu ifade etmektedir.

Değişkenlerin büyük bir çoğunluğu nominal ve ordinal düzeyde kategorik veriler içerirken, görece az sayıda değişken ise aralık veya oran ölçeğinde sürekli verileri içermektedir.

Veri kümesi hem kolayca inceleme yapılabilmesi hem de ilerleyen süreçte modelleme aşamasına hazır hale getirilmesi açısından bu aşamada sayısal değerlerle kodlanmıştır. Nominal değişkenler ikili, ordinal değişkenler ise sıralı tamsayı değerler ile kodlanmıştır. Aralık ölçeğindeki değişkenlerde bilgi kaybına neden olmamak

açısından bu aşamada herhangi bir işlem yapılmamıştır.

Daha önce belirtildiği üzere müşterilerin kişisel bilgi gizliliği ve veri kümesinin ilgili bankanın ticari değeri oluşu nedeniyle müşterilerin kimlik bilgileri ve değişkenlerin bilgi düzeyleri açıklanmayacaktır.

Bu nedenle veri gizliliği açısından özel olarak kodlanmış başlık etiketleri oluşturulmuş ve ayrıca her bir kayda sıra numarası atamak amacıyla yeni bir değişken (NO) üretilmiştir.

Tablo 2.1. Değişkenlere İlişkin Gruplama Düzeyleri

Grup	Değişken	Alt Grup
D	D01,D02,D03,D04,D05,D06,D07,D08	Demografik Özellikleri
M	M01,M02,M03,M04,M05,M06,M07,M08	Banka Kayıt Bilgileri
H	H01,H02,H03,H04,H05,H06,H23	Hesap - Sahip mi
	H07,H08,H09,H10,H11	Hesap - Aktif mi
	H12,H13,H14,H15,H16,H17	Hesap Sayısı
	H18,H19,H20,H21,H22	Hesap Bakiyesi
L	L01,L02,L03	Kredi - Sahip mi
	L04,L05,L06	Kredi – Aktif mi
	L07,L08,L09	Kredi Sayısı
	L10,L11,L12	Kredi Bakiyesi
C	C01,C02,C03,C04,C05,C06	Kredi Kartı - Sahip mi
	C07,C08,C09,C10,C11,C12	Kredi Kartı – Aktif mi
	C13,C14,C15,C16,C17,C18	Kredi Kartı Sayısı
	C19,C20,C21,C22,C23,C24	Kredi Kartı Bakiyesi
Y	Y01,Y02,Y03,Y04,Y05	Yatırım Ürünü - Sahip mi
	Y06,Y07,Y08,Y09,Y10	Yatırım Ürünü - Aktif mi
	Y11,Y12,Y13,Y14,Y15	Yatırım Ürünü Sayısı
	Y16,Y17,Y18,Y19,Y20	Yatırım Ürünü Bakiyesi
U	U01,U02,U03,U04,U05,U11,U12,U21,U22,U23	Hizmet Ürünü - Sahip mi
	U06,U07,U08,U09,U10	Hizmet Ürünü – Aktif mi
	U13,U14,U15,U16,U17	Hizmet Ürünü Sayı
	U18,U19,U20	Hizmet Ürünü Tutarı
T	T01,T02,T03,T04,T05,.....,T25	Şube İşlemleri Sayısı
A	A01,A02,A03,A04,A05,.....,A17	Telefon İşlemleri Sayısı
N	N01,N02,N03,N04,N05,.....,N22	Internet İşlemleri Sayısı
B	B01,B02,B03,B04,B05,B06	ATM İşlemleri Sayısı

Başlık kodlaması yapılırken değişken grupları oluşturulmasına çaba harcanmıştır. Buna göre veri kümesi müşterilerin demografik özellikleri (D), banka

kayıtları (M), müşteri hesapları (H), kredi kullanımı (L), kredi kartları (C), yatırım ürünleri (Y), çeşitli bankacılık ürünleri (U), şube işlemleri (T), telefon bankacılığı işlemleri (A), internet bankacılığı işlemleri (N) ve ATM işlemleri (B) olmak üzere 11 farklı grupta kodlanmıştır. Yukarıdaki tablo (Tablo 2.1) çalışmanın daha anlaşılır olabilmesi açısından değişken grupları hakkında bazı özet bilgileri içermektedir.

Burada ilk olarak Excel süzme özelliği kullanılarak her bir değişkenin şıkları incelenmiştir. Ardından tüm değişkenlerin sıklık dağılımları ve grafikleri çıkartılarak değişken bazında inceleme yapılmıştır. Veri analize hazır hale geldiğinde bu çıktılar yenileneceğinden bu aşamada sunulmayacaktır.

Bu inceleme ile birlikte sınıflandırma işleminde hedef değişen olarak kullanılabilecek değişkenlerin tespiti de yapılmıştır. Bu değişkenler, çeşitli bankacılık ürünlerine sahip olma bilgisini içeren değişken grupları içerisinden seçilmiş nominal düzeyde kategorik veri yapısına sahip olan değişkenlerdir.

Hedef değişken olarak seçilen ilk değişken belirli bir kredi kartı ürününe sahip olma bilgisini içeren C02 değişkenidir. Bu değişkenin seçilmesinde ilgili kredi kartı ürününün bankanın en yeni kredi kartı ürünü olması, sunduğu olanakların bazı özel tüketim biçimlerini içermesi, görece çok yaygın kullanılmadığı için pazarlama faaliyeti açısından önem taşıması ve veri kalitesi açısından diğer ürünlere göre daha nitelikli olması gibi kriterler etkili olmuştur. C02 değişkeni ilgili kredi kartına sahip olma bilgisini içerdiğinden nominal ölçekte olup, müşteri bu ürüne sahip değil ise "0" sahip ise "1" değerini almaktadır.

Bir diğer hedef değişken olarak belirli bir yatırım ürününe sahip olma bilgisini içeren Y01 değişkeni seçilmiştir. Bu değişkenin seçilmesinde de ilgili yatırım ürününün likit oluşu nedeniyle bireysel müşteri açısından daha cazip bulunması, küçük tutarlardaki birikimlerin değerlendirilmesine elverişli olması, müşterilerin kısa dönemli yatırım ihtiyaçlarına da cevap vermesi, banka açısından ise kısa dönemli ve küçük tutarlardaki likiditenin bünyesine katılmasını sağlaması ve yine veri kalitesi açısından diğer yatırım ürünlerine göre daha nitelikli bulunması gibi kriterlere önem verilmiştir. Y01 değişkeni ilgili yatırım ürününe sahip olma bilgisini içerdiğinden

nominal ölçekte olup, müşteri bu ürüne sahip değil ise "0" sahip ise "1" değerini almaktadır.

Son hedef değişken olarak belirlenen U03 değişkeni ise belirli bir bankacılık hizmet ürününe sahip olma bilgisini içermektedir. Bu değişkenin seçilmesinde de ilgili ürünün görece yeni bir ürün olması, teknoloji ilişkili olup gelecek dönem bankacılığı için büyük önem taşıması, henüz yaygın kullanılmadığı için pazarlama faaliyeti açısından önemi, banka açısından maliyetlere yönelik etkisi ve yine veri kalitesi açısından görece nitelikli bulunması gibi kriterler dikkate alınmıştır. U03 değişkeni ilgili hizmet ürününe sahip olma bilgisini içerdiğinden diğer hedef değişkenlerde olduğu gibi nominal ölçekte olup, müşteri bu ürüne sahip değil ise "0" sahip ise "1" değerini almaktadır.

Bu değişkenler hedef değişken olarak seçilmiş olmakla birlikte sınıflandırma işleminde her biri diğerleri için sınıflayıcı değişken olarak da kullanılabilecektir.

Veri kümesi üzerinde yapılan ilk inceleme sonucunda bir takım veri kalitesi sorunları tespit edilmiştir. Kayıp veri değerleri sorununun hem değişken bazında (örn. bazı değişkenlerin değerlerinin yaklaşık üçte biri oranında bilinmediği) hem de kayıt bazında (örn. bazı müşterilerin demografik değişkenlere ilişkin değerlerinin bilinmediği) yaygın olduğu görülmüştür. Ayrıca bazı çelişkili veri değerleri olduğu (örn. bir kredi kartı ürününe sahip olmadığı görülen müşterinin aynı kredi kartı ürününde bakiye bulunması vb.), bazı veri değerlerinin yanlış kodlandığı (örn. İstanbul ya da Istanbul vb.) ve bazı değişkenlerin uç değer sorunlarına sahip olduğu (örn. bazı müşterilerin onlarca fatura talimatının bulunduğu vb.) tespit edilmiştir.

Bilindiği gibi bir çok veri tabanı güncel olmayan verileri de içermektedir. Özellikle bankacılık sektöründe herhangi bir işlem nedeniyle veri tabanına girmiş bulunan ancak zaman içinde hiçbir işlem gerçekleştirmeyen müşterilerin veri tabanındaki varlığı devam etmektedir. Ayrıca banka el değiştirmeleri, banka birleşmeleri ve veri yönetimi konusunda kullanılan yazılımların güncellenmesi ya da tamamen değiştirilmesi gibi durumlarda da veri kalitesi sorunları oluşabilmektedir. Bu türden sorunların üzerinde çalışılan veri kümesi için de geçerli olduğu görülmüştür.

Veri kalitesini yükseltmek açısından yapılacak ön çalışma ile hem kayıt bazında hem de değişken bazında bazı çözümler öngörülmektedir. Bu çözümlerden biri eleme yöntemine başvurmak olacaktır. Değişken bazında düşünüldüğünde bilgi düzeyi çok düşük ve kayıp veri değerleri sorunu çok yüksek olan değişkenler tespit edilerek elenecektir. Kayıt bazında ise özellikle demografik değişken grubuna ait olmak üzere kayıp veri değerleri sorunu çok yüksek seviyede olan kayıtlar elenecektir.

Eleme işlemi sonucunda ortaya çıkan veri kümesinin kalitesi görece daha yüksek düzeyde olacak ve bu aşamadan sonra veri kalitesi sorunlarının veri kaybına neden olmadan çözülmesine çalışılacaktır. Bu amaçla değişken bazında öngörülen ilk çözüm kayıp veri değerleri yerine kestirimde bulunulması olacaktır. Bir sonraki aşama sıklıkları düşük düzeyde bulunan şıkların birleştirilmesi yoluna gitmek ve ardından uç değer sorunlarını ortadan kaldırılmak olacaktır. Kayıt bazında ise çelişik değerlerin ortadan kaldırılması yoluna gidilecektir.

Ayrıca veri kalitesini yükseltme çalışmalarında hedef değişkenlerin sınıflandırılmasına yönelik önem arz eden değişkenlerin mümkün olduğunca elde tutulmasına gayret edilecektir.

2.4. VERİLERİ HAZIRLAMA

Bu aşamada, üzerinde veri madenciliği tekniklerinin uygulanacağı nihai veri kümesinin oluşturulması amacıyla verileri seçme, verileri temizleme, verileri yapılandırma, verileri birleştirme ve verileri biçimlendirme görevleri üzerinde durulacaktır. Ayrıca bu aşama sonunda, elde edilecek nihai veri kümesine ilişkin tanımlayıcı nitelikte istatistiklerin sunulması ile modelleme aşaması öncesinde eldeki veri kümesinin daha detaylı olarak anlaşılması sağlanacaktır.

İlk olarak hiçbir demografik değişkene ilişkin değerleri bilinmeyen tüm kayıtlar veri kümesinden çıkarılmıştır. Bunun yanı sıra değişkenler arası ilişkilerin incelenmesi sonucunda veri kayıtlarında kolayca fark edilebilecek türden tutarsızlıkların tespit edilmesi yoluna gidilmiş, bu incelenme sonucu tespit edilen ve tutarsızlıklarının düzeltilmesi mümkün görülmeyen kayıtların elenmesi yoluna gidilmiştir. Ayrıca çeşitli kayıtların ticari müşterilere ait olduğu şüphesi oluşmuş, bu kayıtların büyük bir

bölümünde bu durumun gerçekliği değişkenlerle ilişkilendirilerek tespit edilmiş ve bu kayıtlar da veri kümesinden çıkarılmıştır.

Değişken bazında yapılan incelemede ise müşteri profili açısından bilgi içermeyen değişkenlerin belirlenmesi ile bu değişkenlerin veri kümesinden çıkarılması sağlanmıştır. Ardından yine ticari hesaplara yönelik olduğu görünen ve bireysel müşterilere ilişkin bilgi içermeyen değişkenlerin tespiti ve elenmesi gerçekleştirilmiştir. Son aşamada ise oluşturulan sıklık dağılımları yolu ile içerdiği bilgi oranı toplam kayıt sayısı içerisinde çok düşük (%1'den az) olan değişkenlerin belirlenmesi üzerinde çalışılmış ve bu türden değişkenler de veri kümesine dahil edilmemiştir.

Ancak bilgi düzeyi çok düşük olmasına rağmen işe yönelik önemi itibariyle örüntü tanıma sürecinde anlamlı olabileceği düşünülen bazı değişkenlerin (örn. bazı kredi kullanım bilgileri) bu aşamada veri kümesinde tutulması uygun bulunmuştur. Modelleme aşamasında yapılacak geçerli değişken seçimi ile bu değişkenlerin sınıflandırma işleminde yer alıp almayacağı ortaya çıkacaktır.

Verilerin seçilmesi görevi sonucunda veri kümesinden çıkarılan toplam kayıt sayısı 4.321 ve değişken sayısı 86 olarak gerçekleşmiştir. Sonuç olarak sınıflandırma amacıyla kullanılacak veri kümesine 103 değişken ve bu değişkenlere ait 13.274 kayıt dahil edilmiştir. Elde edilen veri kümesi daha detaylı inceleme ve modelleme aşamasına hazırlama amacıyla SPSS ortamına aktarılmıştır.

Verileri seçme görevi ile oluşturulan veri kümesi halen bir takım veri kalitesi sorunlarına sahiptir. İki değişkende kayıp veri değerleri sorunu mevcut iken bazı sürekli değişkenlerde uç değer problemleri bazı kategorik değişkenlerde ise göreli sıklıkların heterojenliği söz konusudur.

Bu nedenle öncelikle uç veri değerleri ve heterojen sıklıklar sorununun çözülmesine çalışılmıştır. Bazı kategorik değişkenlerde kategorilerden bir veya bir kaçına ait sıklığın toplam sıklık içindeki oranının çok düşük (yüzde 1'den az) olduğu, bazı sürekli değişkenlerde ise çok az sayıda kaydın veri kümesinin bütünlüğü içerisinde sıra dışı sayılabilecek (en yüksek ya da en düşük yüzde 1'lik kısım) değerlere sahip olduğu tespit edilmiştir. Bu durum örüntülerin ortaya çıkarılması açısından

sorun yaratacağı için kategorilerin birleştirmesi ya da uç değerlerin budanması yolu ile olabildiğince aşılmaya çalışılmıştır.

Örnek olarak D01 (YAŞ) değişkeni için "–18" kategorisinin göreli sıklığı yüzde 1'den az olduğundan "18-24" kategorisiyle birleştirilmiş ve yeni kategori "–24" olarak adlandırılmıştır. Bir başka örnek olarak ise müşterilerin kredi kartı bakiyesi verilebilir. Kredi kartı bakiyesi negatif ya da 0 (sıfır) değerini almalıdır. Ancak bazı müşterilerin kredi kartı borçlarından fazla ödeme yapmış olmaları nedeniyle bakiye sıfırdan büyük gözükmektedir. Buradaki artı değerlerin tamamı sıfır olarak kabul edilmiştir. Bu kapsamda tam olarak 24 değişkende düzeltme yoluna gidilmiştir.

Kayıp veri değerleri sorunu ise sadece iki değişkende (D04, M04) ancak oldukça yaygın olarak (her iki değişkende de yüzde 10'dan fazla) gözlenmektedir. Bu soruna ilişkin SPSS üzerinde yapılan çalışmalardan sonuç alınamamış ve proje planında değişikliğe gidilerek veriler bu aşamada Clementine 8.0 yazılımına yüklenmiştir. Her iki değişken de kategorik olduğundan kayıp veri değerlerinin kestirilmesine yönelik olarak karar ağaçları yönteminin kullanılması uygun bulunmuştur.

M04 değişkeni sekiz kategoride değer alan bir değişken olup kayıp veri değeri sayısı 1695 (göreli sıklığı yaklaşık yüzde 13) olarak gözlenmektedir. Bu değişkene ait kayıp veri değerlerinin kestirilmesinde çeşitli karar ağaçları yöntemleri denenerek en yüksek kesinlik C5.0 algoritması ile elde edilmiştir. Bu teknik ile tüm kayıp veri değerleri için kestirim yapılmıştır. Model, kayıp veri değeri sorunu bulunmayan 10.190 kayıt üzerinden oluşturulmuştur. Karar ağacı derinliği 28 ve çapraz geçerlilik testi ile kestirimin kesinliği %67,3 olarak hesaplanmıştır.

D04 değişkeni beş kategoride değer almakta ve kayıp veri değerleri sayısı 1385 (göreli sıklığı yaklaşık yüzde 10) olarak görülmektedir. Bu değişkendeki kayıp verilerin kestirilmesinde de C5.0 algoritması en yüksek kesinlikte sonuçları üretmiş, tüm kayıp veri değerleri için kestirim yapılmıştır. Model, kayıp veri değeri sorunu bulunmayan 11.889 kayıt üzerinden oluşturulmuştur. Karar ağacı derinliği 28 ve çapraz geçerlilik testi ile kestirimin kesinliği %64,6 olarak hesaplanmıştır.

Bu işlemlerle birlikte verilerin temizlenmesi tamamlanmış ve veri kümesi veri kalitesi açısından modelleme için hazır hale getirilmiştir. Daha sonra verilerin satır ya da sütun bazında detaylı incelemeleri yapılmış ve değişken türetilmesi ya da yeni kayıtların üretilmesi ihtiyacı görülmemiştir. Ancak modelleme aşamasında ortaya çıkacak bu türden ihtiyaçlar durumunda bu aşamaya geri dönülecektir.

Modelleme aşamasında kullanılacak veri kümesi Access veri tabanından birleştirildiği için tablo bazında yeni bir birleştirme işlemi söz konusu değildir. Ancak değişken bazında birleştirme yapmak mümkündür.

Veri kümesinde yer alan 29 değişken, müşterilerin belirli bir bankacılık ürününe sahip olup olmadıklarını ifade etmekte ve iki şıklı kategorik değer almaktadır. Veri kümesindeki 24 değişken ise müşterilerin söz konusu ürünlerden kaç adet bulundurduğu bilgisini içermektedir. Bu değişkenlerin ürün grupları bazında birleştirilmesinin sınıflandırma işlemine ek katkıda bulunabileceği umulmaktadır.

Bu birleştirme işlemi iki farklı yaklaşımla gerçekleştirilmiştir. Birinci yaklaşımda ürün grupları bazında müşterinin sahip olduğu toplam ürün sayısı için birleştirme işlemi gerçekleştirilmiştir. Bu yaklaşımla, müşterilerin sahip olduğu farklı hesap türleri sayısı (FH), farklı kredi türleri sayısı (FL), farklı kredi kartı ürünleri sayısı (FC), farklı yatırım ürünleri sayısı (FY) ve farklı bankacılık ürünleri sayısı (FU) olmak üzere beş yeni değişken üretilmiştir.

Bir diğer yaklaşımla ise, yine ürün grupları bazında sahip olunan toplam ürün adedi için birleştirme yapılmıştır. Buna göre, müşterilerin sahip olduğu toplam banka hesabı adedi (TH), toplam kullanılan kredi adedi (TL), toplam kredi kartı adedi (TC), toplam yatırım ürünü adedi (TY) ve toplam bankacılık ürünü adedi (TU) olmak üzere beş değişken daha veri kümesine eklenmiştir.

Bu yeni değişkenlere ilişkin veri kalitesi faaliyetlerine geri dönülmüş, tüm değişkenlerle ilgili gerekli kategori birleştirmelerine gidilmiştir. Ancak aynı kredi ürününden birden fazla yararlanan müşteri sayısı çok az olduğundan farklı kredi sayısı (FL) ve toplam kredi sayısı (TL) değişkenleri arasında yüksek bir korelasyon (kendall's tau korelasyonu yaklaşık 1,00) saptanmıştır. Bu durum TL değişkeninin FL

değişkeninden farklı olarak yeni bir bilgi içermediği anlamına geldiğinden veri kümesine dahil edilmemesine neden olmuştur. Böylece veri kümesindeki değişken sayısı 112 olmuş, kayıt sayısı 13.274 olarak kalmıştır.

Modelleme aşamasında sınıflandırma için tek bir hedef değişken kullanılmayacağı ve sınıflandırma işlemi tek bir teknikle yapılmayacağından, veri kümesinin herhangi belirli bir sınıflandırma hedefi doğrultusunda yapılandırılması doğru olmayacaktır. Bu nedenle üç farklı hedef değişken için veri madenciliği tekniklerinin aynı şekilde uygulanacağı üç farklı alt veri kümesi hazırlanmıştır.

Her üç hedef değişken için de veri kümesini biçimlendirmede iki farklı işlem gerçekleştirilmiştir. Birinci işlem ile sınıflandırma sürecini etkileyecek ancak öngörü amaçlı olarak kullanılamayacak değişkenlerin veri kümesinden çıkarılması yoluna gidilmiştir. İkinci olarak, hedef değişkenlerin sıklık dağılımlarındaki orantısızlıkların giderilmesine çalışılmıştır. Sınıflandırma başarısını yükseltmek amacıyla ilgili ürünlere sahip olan müşterilerin tamamı, sahip olmayan müşterilerin ise rastlantısal olarak seçilen belirli bir kısmı (en fazla 1:2 oranında) veri kümelerine dahil edilmiştir.

2.4.1. C02 Değişkeni İçin Verileri Biçimlendirme

C02 hedef değişkeni için veri kümesinin biçimlendirilmesi işleminde C08 (C02 ürününü kullanma aktivitesi), C14 (C02 ürününe sahip olma sayısı) ve C20 (C02 ürünündeki bakiye) değişkenleri veri kümesinden çıkarılmıştır.

Kayıtların seçilmesi açısından ise C02 ürününe sahip olanların tümü ile sahip olmayanlardan rastlantısal 3.370 örnek dikkate alınmıştır. Böylece C02 hedef değişkeni üzerinde gerçekleştirilecek sınıflandırma işlemi için veri kümesine 109 değişken ve 5.055 kayıt dahil edilmiştir.

2.4.2. Y01 Değişkeni İçin Verileri Biçimlendirme

Y01 hedef değişkeni için veri kümesinin biçimlendirilmesi işleminde Y06 (Y01 ürününün aktivitesi), Y11 (Y01 ürününe sahip olma sayısı) ve Y16 (Y01 ürününe ait yatırım tutarı) değişkenleri veri kümesinden çıkarılmıştır.

Kayıt bazında ise Y01 ürününe sahip olanların tümü ile sahip olmayanlardan rastlantısal 2.704 örnek seçilerek, Y01 hedef değişkeni üzerinde gerçekleştirilecek sınıflandırma işlemi için 109 değişken ve 4.056 kayıttan oluşan bir veri kümesi oluşturulmuştur.

2.4.3. U03 Değişkeni İçin Verileri Biçimlendirme

U03 hedef değişkeni için veri kümesinin biçimlendirilmesi işleminde U08 (U03 hizmet ürünü kullanma aktivitesi) ve U15 (U03 hizmet ürününe sahip olma sayısı) değişkenleri veri kümesinden çıkarılmıştır.

Sınıflandırma işleminde kullanılmak üzere U03 ürününe sahip olanların tümü ile sahip olmayanlardan rastlantısal 4.836 örnek seçilmiştir. Böylece U03 hedef değişkeni üzerinde gerçekleştirilecek sınıflandırma işlemi için veri kümesine 110 değişken ve 7.254 kayıt dahil edilmiştir.

2.4.4. Tanımlayıcı İstatistik Çalışması

Veri madenciliği standart sürecinin (CRISP – DM) ikinci aşamasının tamamlanması ile birlikte üç farklı hedef değişken için üç farklı alt veri kümesi elde edilmiştir. Ancak her üç veri kümesinde de farklı sayıda değişken ve kayıt yer almaktadır. Bu nedenle aşağıdaki tablolarda verilen sıklık dağılımları veri kümesinin alt kümelere ayrılmadan önceki yapısını temsil etmektedir. Sadece hedef değişkenlere ait sıklık dağılımları kendileri için hazırlanan alt veri kümeleri üzerinden sunulacaktır.

Veri kalitesi sorunlarından arındırılmış ve henüz hedef değişkenlere uygun olarak bölümlenmemiş veri kümesi 112 değişken ve bunlara ait değerleri içeren 13.274 kayıttan oluşmaktadır.

Yaş değişkenine (D01) ait sıklık dağılımı (Tablo 2.2) incelendiğinde yaklaşık olarak simetrik olduğu ve en yüksek sıklığın "35 – 44" aralığında gerçekleştiği görülmektedir.

Tablo 2.2. Müşterilerin Yaşları Dağılımı

D01		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	- 24	331	2,5	2,5	2,5
	25 - 34	3078	23,2	23,2	25,7
	35 - 44	4381	33,0	33,0	58,7
	45 - 54	3148	23,7	23,7	82,4
	55 -	2336	17,6	17,6	100,0
	Toplam	13274	100,0	100,0	

Cinsiyet değişkeninin (D02) sıklık dağılımından (Tablo 2.3) "erkek" müşterilerin çoğunlukta olduğu, medeni durum değişkeninin (D03) sıklık dağılımından (Tablo 2.4) ise "evli" müşterilerin göreli sıklığının oldukça yüksek olduğu görülmektedir.

Tablo 2.3. Müşterilerin Cinsiyetleri Dağılımı

D02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	Erkek	9752	73,5	73,5	73,5
	Kadın	3522	26,5	26,5	100,0
	Toplam	13274	100,0	100,0	

Tablo 2.4. Müşterilerin Medeni Durumları Dağılımı

D03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	Bekar	3943	29,7	29,7	29,7
	Evli	8980	67,7	67,7	97,4
	Boşanmış	351	2,6	2,6	100,0
	Toplam	13274	100,0	100,0	

Eğitim durumu değişkenine (D04) ait sıklık dağılımında (Tablo 2.5) ise en yüksek göreli sıklığın "lise" düzeyinde olduğu, "lise" ve "üniversite" şıklarının göreli sıklığının %50'den fazla olduğu ve müşterilerin büyük bir çoğunluğunun (yaklaşık %70) "lise ve dengi okul mezunu" olduğu görülmektedir.

Tablo 2.5. Müşterilerin Eğitim Düzeyleri Dağılımı

D04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	İlkokul	2158	16,3	16,3	16,3
	Ortaokul	1887	14,2	14,2	30,5
	Lise	4734	35,7	35,7	66,1
	Üniversite	3434	25,9	25,9	92,0
	Lisans Üstü	1061	8,0	8,0	100,0
	Toplam	13274	100,0	100,0	

Veri kümesinde yer alan tüm değişkenlere ait tanımlayıcı istatistik ve sıklık dağılımları, bu değişkenlere ilişkin bilgi düzeyleri paylaşılamadığı için faydalı olmayacağından çalışmanın ekinde (Ek.2) sunulmuştur.

Bununla birlikte modelleme aşamasında hedef değişken olarak kullanılacak değişkenlerin sıklık dağılımlarının verilmesinde fayda görmekteyim. Daha önce belirtildiği gibi bu değişkenlere ait sıklık dağılımları ilgili değişken için hazırlanan alt veri kümesi dikkate alınarak sunulmuştur.

Aşağıdaki tablolarda sırasıyla C02 kodlu (Tablo 2.6) "kredi kartı ürününe sahip olma" bilgisini içeren hedef değişkenin sıklık dağılımı, Y01 kodlu (Tablo 2.7) "yatırım ürününe sahip olma" bilgisini içeren değişkenin sıklık dağılımı ve değişken kodu U03 olan (Tablo 2.8) "hizmet ürününe sahip olma" bilgisini içeren değişkenin sıklık dağılımı sunulmuştur.

Tablo 2.6. C02 Değişkeni Sıklık Dağılımı

C02 (Kredi Kartı)		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	Yok	3.370	66,7	66,7	66,7
	Var	1.685	33,3	33,3	100,0
	Toplam	5.055	100,0	100,0	

Tablo 2.7. Y01 Değişkeni Sıklık Dağılımı

Y01 (Yatırım Ürünü)		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	Yok	2.704	66,7	66,7	66,7
	Var	1.352	33,3	33,3	100,0
	Toplam	4.056	100,0	100,0	

Tablo 2.8. U03 Değişkeni Sıklık Dağılımı

U03 (Hizmet Ürünü)		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	Yok	4.836	66,7	66,7	66,7
	Var	2.418	33,3	33,3	100,0
	Toplam	7.254	100,0	100,0	

Her üç hedef değişkenin sıklık dağılımları incelendiğinde toplam sıklıkların farklı ancak şıkların göreli sıklığının aynı olduğu fark edilecektir. Bu durum, çalışmanın amaçları arasında uygulanacak tekniklerin karşılaştırılması da yer aldığından, sınıflandırma sürecinde her üç değişken için de aynı koşulları yaratmak açısından önem taşımaktadır.

Ayrıca bu noktada belirtilmeliyim ki, son şekli verilen bu üç veri kümesinin modelleme aşamasında uygulanacak her teknikte aynı şekilde kullanılması gerekmemektedir. Her teknik öncesinde yapılacak geçerli değişken analizi ile o tekniğe ilişkin uygulamada geçersiz ya da gereksiz bulunan değişkenlerin veri kümesinden çıkarılması söz konusu olacak ve her teknik kendisi için geçerli değişkenler dikkate alınarak uygulanacaktır.

2.5. MODEL OLUŞTURMA

Bu aşamada, önceki aşamalarda iş hedefleri ve veri madenciliği amacı doğrultusunda biçimlendirilen bankacılık müşteri verileri üzerinde sınıflandırma modelleri geliştirilecektir. Bu kapsamda modelleme tekniğinin seçimi, sınama tasarımının üretilmesi, modelin kurulması ve model değerlemesi üzerinde ayrıntılarıyla durulacaktır.

Bu çalışmada, daha önce belirtildiği gibi sınıflandırma modellerinin geliştirilmesi amaçlandığından veri madenciliğinin sınıflandırma fonksiyonuna ilişkin yöntemleri arasından seçim yapılmıştır. Bu seçimde gerek literatürdeki yaygınlık gerekse mevcut yazılım ve donanım olanakları önemli rol oynamıştır.

Veri madenciliğinin en önemli bileşenlerinin istatistik, yapay öğrenme ve veri tabanı teknolojileri olduğundan önceki bölümde söz edilmişti. Çalışmada kullanılacak yöntemlerin de bu alanları temsil etmelerine özen gösterilmiştir. Sonuç olarak, yazılım

ve donanım olanakları doğrultusunda kademeli lojistik regresyon analizi, ileri beslemeli yapay sinir ağları ve C5.0 karar kuralı türetme algoritması tekniklerinin uygulanmasına karar verilmiştir.

Seçilen bu tekniklerin hem kategorik hedef değişkenlerin sınıflandırılmasına uygun olmaları hem de her ölçekte değişken tipini bağımsız değişken olarak kabul edebilmeleri nedeniyle hazırlanan veri kümeleri üzerinde uygulanmalarında herhangi bir sakınca bulunmamaktadır.

Yapılacak uygulama sonucu elde edilecek sınıflandırma modelleri için iki temel başarı kriteri söz konusudur. Öncelikle geliştirilen modellerin öğrenme kümesi üzerinde yüksek sınıflandırma başarısı göstermesi gerekir. Ancak veri madenciliği amacı doğrultusunda asıl olan geliştirilen modellerin öngörü amaçlı olarak kullanılabilmesidir. Bu açıdan ikinci başarı kriteri, sınıflandırma modellerinin öğrenme kümesinden tamamen farklı bir veri kümesi üzerinde de yüksek sınıflandırma başarısı göstermesi olacaktır.

Bu nedenle, modelleme sürecinde bu ölçülerin elde edilebilmesi için her üç hedef değişken için hazırlanmış bulunan veri kümeleri %70 oranında öğrenme kümesini ve %30 oranında sınama kümesini oluşturacak şekilde rastlantısal seçimle bölümlenecektir. Sınıflandırma modelleri öğrenme kümesi üzerinde geliştirilecek, sınama kümesi üzerinde öngörü başarıları sınanacaktır.

C02 değişkeni için gerçekleştirilecek uygulamada veri kümesi 5.055 kayıttan oluşmaktaydı. Bu kayıtlardan rastlantısal seçimle 3.491 adedi öğrenme kümesini ve 1.564 adedi sınama kümesini oluşturacak şekilde bölümlenmiştir.

Y01 değişkeni için oluşturulan veri kümesinde 4.056 kayıt yer almaktaydı. Bu kayıtların yine rastlantısal seçimle 2.869 adedi öğrenme kümesi ve 1.187 adedi sınama kümesi için ayrılmıştır.

U03 hedef değişkeni için ise 7.254 kayıttan oluşan bir veri kümesi hazırlanmıştı. Bu veri kümesinde yapılan rastlantısal bölümlemede öğrenme kümesi 5.088 kayıt ve sınama kümesi 2.166 kayıt içerecek şekilde oluşmuştur.

Ayrıca bu bölümlemeyi her defasında aynı şekilde kullanabilmek için öğrenme kümesi ile sınama kümesini birbirinden ayıran ve *örnekleme değişkeni* adı verilen yeni bir değişken türetilmiştir. Bu şekilde aynı öğrenme kümesinde geliştirilen ve aynı sınama kümesi üzerinde test edilen modellerin karşılaştırılması sağlanmış olacaktır.

Sınıflandırma işlemi Clementine 8.0 yazılımı ile gerçekleştirilecek ve her bir hedef değişkenin sınıflandırılmasında daha önce kendisi için hazırlanan veri kümesi kullanılacaktır.

Clementine 8.0 üzerinde çalışmak basit anlamda üç adımlı bir süreçtir. İlk adımda, veri kümesinin Clementine 8.0 tarafından okunması sağlanır. Bir başka deyişle, yazılım veri kümesine tanımlanan köprü üzerinden bağlanır. İkinci aşamada, veri kümesi üzerinde bir dizi yönlendirme işlemi gerçekleştirilir. Değişkenlerin ölçekleri tanımlanır, hedef değişken belirlenir, modele katılmayacak değişkenler elenir. Son aşamada ise veri kümesi belirli bir konuma yönlendirilir. Bu konum bir model olabileceği gibi herhangi bir biçime sahip çıktı da olabilir. Bu işlemlerin tümüne ise verilerin bir noktadan bir başka noktaya hareketi söz konusu olduğundan "veri akışı" (data stream) adı verilir.[87]

Özetlemek gerekirse, her modelleme işlemi için veri kümesi yazılıma yüklenecek, ilgili veri kümesi üzerinde değişken ya da kayıt bazında gereken işlemler yürütülecek ve sürecin sonunda model ya da çıktılar elde edilecektir. Bu çalışmada oluşturulacak olan veri akışları çalışmanın ekinde şematik olarak sunulacaktır.

2.5.1. Lojistik Regresyon Analizi Uygulaması

Bu tekniğin uygulanmasında öncelikle veri kümelerinin yazılıma yüklenmesi gerçekleştirilmiştir. Daha sonra sırasıyla veri kümesinin öğrenme ve sınama kümesi şeklinde ayrılması için örnekleme değişkeninin kullanıldığı *select* düğümü, örnekleme değişkeninin modele katılmasını engellemek için süzme işleminin yapıldığı *filter* düğümü ve hedef değişkenin tanımlandığı *type* düğümü kullanılmıştır. Son olarak, öğrenme kümesi model parametrelerinin kestirildiği *logistic* düğümünden geçirilmiştir.

[87] Integral Solutions Co., *Clementine 8.0 User's Guide*, 2003

Clementine 8.0, *logistic* düğümünde değişken kodlaması gerçekleştirdiğinden her bir bağımsız değişkenden çok sayıda yeni değişkenin türetilmesi söz konusu olmaktadır. Ayrıca nihai modelin oluşturulmasında kademeli değişken seçiminin yapılabilmesi için *stepwise* yöntemi seçilmiş ve değişkenlere ilişkin anlamlılık sınırı ise modele kabul için 0,05 ve modelden çıkarma için 0,10 olarak belirlenmiştir. Diğer model parametreleri için yazılımın varsayılan *(default)* değerleri kullanılmıştır.

2.5.1.1. C02 Değişkeni Üzerinde Uygulama

C02 hedef değişkeni üzerinde uygulanan lojistik regresyon tekniği ile model 15 dakika 01 saniye içerisinde üretilmiştir. Elde edilen lojistik regresyon denklemi aşağıdaki tabloda (Tablo 2.9) sunulmaktadır. Kademeli değişken seçimi kullanılarak 47 bağımsız değişkenli lojistik regresyon denklemi oluşmuştur.

Tablo 2.9. C02 Hedef Değişkenine Ait Lojistik Regresyon Denklemi

Değişken	Katsayı	Değişken	Katsayı	Değişken	Katsayı
Sabit	0,6726	M02 = 8	-0,3938	FH = 3	0,7270
D02 = 2	0,3693	M03 = 2	-1,1870	C03 = 0	1,4670
D04 = 5	-0,0552	M04 = 1	-0,7392	C09 = 0	-1,6010
D04 = 3	0,2847	M04 = 2	0,3943	C12 = 0	0,3561
D04 = 1	0,8483	M04 = 3	1,4020	C18 >= 4	0,2226
D04 = 2	0,6101	M04 = 4	2,3110	C18 = 0	2,1960
M02 >= 15	0,2976	M04 = 5	-0,1973	C18 = 1	0,6334
M02 = 10	0,2980	M04 = 6	2,3260	C18 = 2	0,3495
M02 = 11	0,5285	M04 = 7	-0,0612	U01 = 0	-0,5292
M02 = 12	0,4249	H01 = 0	2,5780	U07 = 0	-0,3650
M02 = 13	0,2780	H03 = 0	-0,8497	U09 = 0	-0,6598
M02 = 14	0,3001	H19	-0,0001	FU >= 5	-0,2025
M02 = 4	-2,3570	FH >= 5	-0,6632	FU = 0	0,9012
M02 = 5	-1,5110	FH = 0	-1,8070	FU = 1	0,9277
M02 = 6	-0,7856	FH = 1	0,8867	FU = 2	0,4116
M02 = 7	-0,6714	FH = 2	0,7068	FU = 3	0,0028

Model iyiliği ölçümünde kullanılan Nagelkerke R – Kare değeri 0,6310 olarak hesaplanmış, modelin öğrenme kümesi üzerindeki sınıflandırma başarısı ise %84,90 olarak elde edilmiştir.

Lojistik regresyon tekniğinin uygulanmasında C02 değişkeni için yazılım üzerinde gerçeklenen veri akışı (Ek.3) ile değişken seçiminde uygulanan süreç tablosu

(Ek.4) çalışmanın ekinde sunulmuştur.

Model öngörü başarısını ölçmek amacıyla sınama kümesi üzerinde uygulanmış ve aşağıda sunulan sınıflandırma tablosu (Tablo 2.10) ortaya çıkmıştır. Buna göre, modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %84,34 olmuştur.

Tablo 2.10. C02 için Lojistik Regresyon Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	415	113	528
Yok	132	904	1.036
Toplam	547	1.017	1.564

2.5.1.2. Y01 Değişkeni Üzerinde Uygulama

Lojistik regresyon tekniğinin Y01 hedef değişkeni üzerinde uygulanması işlemi 15 dakika 18 saniyede sonuç vermiştir. Elde edilen logistik regresyon denklemi aşağıdaki tabloda (Tablo 2.11) sunulmuştur. Kademeli değişken seçimi sonucu elde edilen lojistik regresyon denklemi 49 bağımsız değişkene sahiptir.

Bu uygulama sırasında bir takım sorunlarla karşılaşılmış ve diğer uygulamalardan farklı olarak *filter* düğümünde ek düzenlemelere gidilmiştir. Lojistik regresyon düğümünün ilk çalıştırılmasında nümerik bir hata oluşmuş ve kademeli değişken seçimi başarısız olmuştur. En çok olabilirlik kestiriminin başarısız olmasına neden olan bu sorunun ortadan kaldırılabilmesi için hatanın oluştuğu anda modele en son katılmış bulunan bağımsız değişken modelden çıkarılmış, bir başka deyişle *filter* düğümünde bu değişken elenmiştir. Tekrar edilen uygulamada da aynı sorun yaşanmış ve geçerli model elde edilene kadar aynı yöntem ile devam edilmiştir. Bu süreçte sırasıyla her deneme sonucunda H01, FH, H12 ve TH bağımsız değişkenleri modelin dışında bırakılmıştır.

Bu sorundan arındırılarak elde edilen nihai modelin Nagelkerke R – Kare değeri 0,6640 olarak hesaplanmıştır. Bu modelin öğrenme kümesi üzerindeki sınıflandırma başarısı ise %86,70 olarak elde edilmiştir.

Tablo 2.11. Y01 Hedef Değişkenine Ait Lojistik Regresyon Denklemi

Değişken	Katsayı	Değişken	Katsayı	Değişken	Katsayı
Sabit	-0,2677	H07 = 0	0,7664	M04 = 4	0,8749
C03 = 0	0,5837	H10 = 0	1,7080	M04 = 5	-0,0760
C13 >= 2	-0,8453	H11 = 0	-1,2720	M04 = 6	1,2330
C13 = 0	0,1352	M02 >= 15	1,2300	M04 = 7	0,3138
C24	0,0001	M02 = 10	0,4516	U06 = 0	0,5563
D01 = 2	-0,0015	M02 = 11	0,2950	U10 = 0	-0,8117
D01 = 6	-0,7360	M02 = 12	0,5373	U13 >= 4	-0,6633
D01 = 3	-0,4938	M02 = 13	0,6193	U13 = 0	0,4709
D01 = 4	-0,0950	M02 = 14	0,7604	U13 = 1	0,4149
D02 = 2	-0,4855	M02 = 4	-2,7710	U13 = 2	0,1950
FU >= 5	-0,8224	M02 = 5	-1,9060	U17 >= 2	0,6802
FU = 0	1,8230	M02 = 6	-1,6400	U17 = 0	1,3250
FU = 1	1,4340	M02 = 7	-1,4510	U23 = 0	-1,5450
FU = 2	0,8568	M02 = 8	-0,6998	Y02 = 0	0,7682
FU = 3	0,1161	M04 = 1	-0,3848	Y04 = 0	1,5870
H04 = 0	-1,2140	M04 = 2	0,8567	Y05 = 0	-3,0130
H06 = 0	1,9570	M04 = 3	0,4428		

Lojistik regresyon tekniğinin uygulanmasında Y01 değişkeni için yazılım üzerinde gerçeklenen veri akışı (Ek.5) ile değişken seçiminde uygulanan süreç tablosu (Ek.6) çalışmanın ekinde sunulmuştur.

Modelin öngörü başarısını ölçmek amacıyla sınama kümesi üzerinde uygulanması sonucu aşağıda sunulan sınıflandırma tablosu (Tablo 2.12) ortaya çıkmıştır. Buna göre modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %84,84 olarak gerçekleşmiştir.

Tablo 2.12. Y01 için Lojistik Regresyon Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	285	98	383
Yok	82	722	804
Toplam	367	820	1.187

2.5.1.3. U03 Değişkeni Üzerinde Uygulama

U03 hedef değişkeni üzerindeki uygulama ile son lojistik regresyon modeli üretilmiştir. Bu tekniğin üretilmesi 43 dakika 26 saniye içerisinde sonuç vermiş ve aşağıdaki tabloda sunulan (Tablo 2.13) lojistik regresyon denklemi üretmiştir.

Kademeli değişken seçimi sonucunda modele 53 bağımsız değişken katılmıştır.

Tablo 2.13. U03 Hedef Değişkenine Ait Lojistik Regresyon Denklemi

Değişken	Katsayı	Değişken	Katsayı	Değişken	Katsayı
Sabit	-5,8500	H12 = 1	0,0662	M04 = 3	0,3657
C03 = 0	5,0170	H12 = 2	0,0106	M04 = 4	0,4717
C07 = 0	0,5848	H12 = 3	-0,3221	M04 = 5	0,0152
C08 = 0	0,5605	L02 = 0	-1,9150	M04 = 6	0,8201
C13 >= 2	-0,5034	M02 >= 15	0,1917	M04 = 7	-0,0748
C13 = 0	-0,4178	M02 = 10	-0,4752	TC >= 5	-1,0320
C20	0,0003	M02 = 11	-0,2515	TC = 0	-0,7224
D01 = 2	0,1255	M02 = 12	-0,5178	TC = 1	-0,8194
D01 = 6	0,8955	M02 = 13	-0,2294	TC = 2	-0,9445
D01 = 3	-0,5285	M02 = 14	-0,0712	TC = 3	-0,9912
D01 = 4	-0,2801	M02 = 4	-2,3500	TY >= 3	0,3625
D04 = 5	-0,1398	M02 = 5	-1,3360	TY = 0	1,2070
D04 = 3	0,1152	M02 = 6	-0,9250	TY = 1	0,7429
D04 = 1	0,7596	M02 = 7	-0,7347	U02 = 0	3,2690
D04 = 2	0,8067	M02 = 8	-0,3202	U04 = 0	0,4474
H02 = 0	0,4691	M03 = 1	0,5679	U06 = 0	0,7487
H12 >= 5	-1,0610	M04 = 1	-0,0258	Y16	0,0001
H12 = 0	0,7512	M04 = 2	-0,6876	Y18	0,0002

Bu modelin Nagelkerke R – Kare değeri 0,6640 olarak hesaplanmış, öğrenme kümesi üzerindeki sınıflandırma başarısı ise %90,20 olarak elde edilmiştir.

Lojistik regresyon tekniğinin uygulanmasında U03 değişkeni için yazılım üzerinde gerçeklenen veri akışı (Ek.7) ile değişken seçiminde uygulanan süreç tablosu (Ek.8) çalışmanın ekinde sunulmuştur.

Tablo 2.14. U03 için Lojistik Regresyon Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	611	139	750
Yok	118	1.298	1.416
Toplam	729	1.437	2.166

Modelin öngörü başarısını ölçmek amacıyla sınama kümesi üzerinde uygulanması sonucu yukarıda sunulan sınıflandırma tablosu (Tablo 2.14) elde edilmiştir. Buna göre modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %88,13 şeklinde gerçekleşmiştir.

2.5.2. Yapay Sinir Ağları Uygulaması

Bu uygulanmada öncelikle veri kümeleri yazılıma yüklenmiş ve daha önce belirtildiği gibi örnekleme değişkeni kullanılarak öğrenme ve sınama amaçlı alt veri kümeleri oluşturulmuştur. *Filter* düğümü ile örnekleme değişkeninin modele girmesi engellenmiş, *type* düğümü ile sınıflandırma işleminin gerçekleştirileceği hedef değişken tanımlanmıştır. Daha sonra yapay sinir ağı modelinin oluşturulması amacıyla *Neural Net* düğümü kullanılmıştır. Model, değişken kodlamalarını ve geçerli değişkenlerin seçimini kendiliğinden gerçekleştirmektedir.

Yapay sinir ağları, öğrenme kümesinin bir kısmını bağlantı ağırlıklarının hesaplanmasında geçerlilik amacıyla kullanmaya ihtiyaç duymaktadır. Bu nedenle, daha önce bu uygulama için hazırlanan öğrenme kümeleri yaklaşık olarak 1/3 oranında yeniden bölümlenmiştir. Yapay sinir ağı algoritması, veri kümesinin bir bölümünü (%70) modeli oluşturmak, diğer bölümünü (%30) ise geçerlilik sınaması amacıyla kullanacaktır. Bu bölümleme rastlantısal olarak gerçekleştirildiğinden her defasında aynı kayıtların seçilebilmesi için sabit bir başlangıç değeri (*seed*) belirlenmiştir.

Neural net düğümünde öncelikle olası en küçük ağın oluşturulduğu ve her defasında sınıflandırma başarısını en çok arttıran yeni nöronların ağa katılımının sağlandığı *dynamic* yöntem seçilmiştir. Sonlandırma kriteri olarak yazılımın varsayılan değerleri kabul edilmiştir. Bu şekilde, *dynamic* yönteme uygun olarak arka arkaya beş yeni nöron eklendiği halde sınıflandırma başarısında anlamlı bir değişiklik olmadığı görülürse öğrenme süreci sonlandırılacaktır. Süreç sonlandırıldığında elde edilen en son model değil elde edilmiş en iyi model yazılımın ürettiği nihai model olacaktır.

2.5.2.1. C02 Değişkeni Üzerinde Uygulama

C02 hedef değişkeninin sınıflandırılmasında kullanılan yapay sinir ağı uygulamasında yazılım 22 dakika 50 saniye sonunda sonuca ulaşmıştır. Elde edilen ağın giriş katmanında 203 nöron, ilk saklı katmanda 5 ve ikinci saklı katmanda 5 nöron, çıkış katmanında ise 1 nöron oluşmuştur.

Girdi olarak tanımlanan tüm bağımsız değişkenler modele dahil edilmiş ve modelin geçerlilik kümesindeki sınıflandırma başarısı %83,63 olarak hesaplanmıştır. Ayrıca yazılım üzerinde oluşturulan veri akışı çalışmanın ekinde (Ek.9) sunulmuştur.

Tablo 2.15. C02 için Yapay Sinir Ağı Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	421	107	528
Yok	142	894	1.036
Toplam	563	1.001	1.564

Model öngörü başarısının ölçülmesi amacıyla daha önce öğrenme kümesinden ayrı tutulan veri kümesi üzerinde sınanmış ve yukarıda sunulan sınıflandırma tablosu (Tablo 2.15) ortaya çıkmıştır. Buna göre modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %84,08 şeklinde gerçekleşmiştir.

2.5.2.2. Y01 Değişkeni Üzerinde Uygulama

Y01 hedef değişkeni üzerinde uygulanan yapay sinir ağı 17 dakika 04 saniye sonunda model üretmiştir. Modelin giriş katmanında 206 nöron, ilk saklı katmanda 2 ve ikinci saklı katmanda 4 nöron, çıkış katmanında ise 1 nöron oluşmuştur.

Girdi olarak tanımlanan tüm bağımsız değişkenler modele dahil edilmiş ve geçerlilik kümesi üzerindeki sınıflandırma başarısı %86,12 olarak hesaplanmıştır. Y01 değişkeni üzerinde gerçekleştirilen yapay sinir ağı uygulamasının yazılım üzerinde oluşturulan veri akışı çalışmanın ekinde (Ek.10) yer almaktadır.

Tablo 2.16. Y01 için Yapay Sinir Ağı Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	233	150	383
Yok	49	755	804
Toplam	282	905	1.187

Modelin sınama kümesi üzerinde uygulanması ile elde edilen sınıflandırma tablosu (Tablo 2.16) yukarıda sunulmuştur. Buna göre modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %83,24 şeklinde gerçekleşmiştir.

2.5.2.3. U03 Değişkeni Üzerinde Uygulama

U03 değişkeninin hedef değişken kabul edildiği yapay sinir ağı uygulamasında yazılım 30 dakika 36 saniye sonunda sonuca ulaşmıştır. Elde edilen ağın giriş katmanında 199 nöron, ilk saklı katmanda 7 ve ikinci saklı katmanda 5 nöron, çıkış katmanında 1 nöron oluşmuştur.

Girdi olarak tanımlanan tüm bağımsız değişkenler modele dahil edilmiş ve geçerlilik kümesi üzerindeki sınıflandırma başarısı %88,95 olarak gerçekleşmiştir. U03 değişkeni üzerinde gerçekleştirilen yapay sinir ağı uygulamasının veri akışı çalışmanın ekinde (Ek.11) yer almaktadır.

Sınama kümesi üzerinde uygulanan modelin %87,44 öngörü kesinliğine sahip olduğu görülmüş ve ilgili sınıflandırma tablosu aşağıda (Tablo 2.17) sunulmuştur.

Tablo 2.17. U03 için Yapay Sinir Ağı Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	616	134	750
Yok	138	1.278	1.416
Toplam	754	1.412	2.166

2.5.3. C5.0 Kural Türetme Algoritması Uygulaması

Bu tekniğin uygulanmasında da öncelikle her bir hedef değişken için hazırlanmış olan veri kümeleri yazılıma yüklenmiş, örnekleme değişkeni yolu ile veri kümelerinin öğrenme ve sınama amacıyla ayrılması için *select* düğümü kullanılmıştır. Örnekleme değişkeni *filter* düğümü ile elenmiş, hedef değişken *type* düğümü ile tanımlanmıştır. Ve son olarak *C5.0* düğümü ile model kurma aşamasına gelinmiştir.

C5.0 düğümünde öncelikle beş farklı modelin geliştirilmesi ve bu modellerin kombinasyonu ile tek bir nihai modele ulaşılmasına dayanan *boosting* seçeneği ile sınıflandırma başarısının ölçülebilmesi için *cross-validate* seçeneği tercih edilmiştir. Bu ikinci seçenek ile veri kümeleri beş eşit bölüme ayrılacak ve her bölümde geliştirilen modelin veri kümesinin kalan bölümünde geçerliliği sınanacak yani beş katlı çapraz geçerlilik uygulanacaktır.

Ayrıca gelişmiş seçeneklerde ağaç yapısının her bir koluna en az beş kayıt düşecek şekilde ağaç budama yönteminin kullanılması ve öncelikle geçerli değişken analizi uygulanarak geçersiz değişkenlerin modele alınmaması için *winow attributes* seçeneği tercih edilmiştir. Model çıktısı olarak ise bir karar ağacı elde edilecektir.

2.5.3.1. C02 Değişkeni Üzerinde Uygulama

C02 hedef değişkeni üzerinde gerçekleştirilen C5.0 algoritması 06 saniye içerisinde sonuçlanmış ve ağaç derinliği 17 olarak gerçekleşmiştir. Modele, geçerliliği kabul edilen 32 değişken katılmıştır.

Modelin uyguladığı çapraz geçerlilik işlemi sonucu sınıflandırma başarısı %81,04 ve standart hatası 0,90 olarak hesaplanmıştır. Bu modelin geliştirildiği veri akışı diyagramı (Ek.12) ile elde edilen karar ağacı (Ek.13) çalışmanın ekinde sunulmuştur.

Model öngörü başarısının ölçülmesi için öğrenme kümesinden tamamen farklı olarak hazırlanan veri kümesi ile sınanmış ve aşağıda sunulan sınıflandırma tablosu (Tablo 2.18) ortaya çıkmıştır. Buna göre modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %82,03 olarak gerçekleşmiştir.

Tablo 2.18. C02 için C5.0 Algoritması Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	399	129	528
Yok	152	884	1.036
Toplam	551	1.013	1.564

2.5.3.2. Y01 Değişkeni Üzerinde Uygulama

Y01 hedef değişkeni için yapılan uygulamada C5.0 algoritması 04 saniye içerisinde sonuçlanmış ve ağaç derinliği 8 olarak gerçekleşmiştir. Geçerli değişken analizi sonucunda modele dahil edilen değişken sayısı 15 olmuştur.

Modelin uyguladığı çapraz geçerlilik sonucu sınıflandırma başarısı %82,10 olarak hesaplanmış, standart hata 0,50 olarak gerçekleşmiştir. Uygulama sırasında kullanılan veri akışı diyagramı (Ek.14) ile elde edilen karar ağacı (Ek.15) çalışmanın ekinde sunulmuştur.

Elde edilen model, öngörü başarısını ölçmek için daha önce hazırlanan veri kümesi ile sınanmış ve yukarıda sunulan sınıflandırma tablosu (Tablo 2.19) ortaya çıkmıştır. Buna göre modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %81,21 olarak gerçekleşmiştir.

Tablo 2.19. Y01 için C5.0 Algoritması Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	267	116	383
Yok	107	697	804
Toplam	374	813	1.187

2.5.3.3. U03 Değişkeni Üzerinde Uygulama

C5.0 algoritmasının ve çalışmamın son uygulaması olan U03 hedef değişkeni üzerindeki uygulama 07 saniye içerisinde sonuçlanmış ve ağaç derinliği 18 olarak gerçekleşmiştir. Modele geçerliliği kabul edilen 37 değişken katılmıştır.

Modelin uyguladığı çapraz geçerlilik işlemi sonucu sınıflandırma başarısı %88,30 olarak hesaplanmış, standart hata 0,60 olarak gerçekleşmiştir. Uygulama sırasında kullanılan veri akışı diyagramı (Ek.16) ile elde edilen karar ağacı (Ek.17) çalışmanın ekinde sunulmuştur.

Tablo 2.20. U03 için C5.0 Algoritması Öngörü Başarısı Tablosu

Gerçek / Öngörü	Var	Yok	Toplam
Var	605	145	750
Yok	114	1.302	1.416
Toplam	719	1.447	2.166

Elde edilen model öngörü başarısını ölçmek için sınama kümesi üzerinde uygulanmış ve yukarıda sunulan sınıflandırma tablosu (Tablo 2.20) ortaya çıkmıştır. Buna göre modelin yeni kayıtlar için doğru öngörülerde bulunma kesinliği %88,04 olarak gerçekleşmiştir.

2.5.4. Modellerin Karşılaştırılması

Karşılaştırma sürecinde aynı koşulların sağlanması önem taşıdığından bir önceki adımda elde edilen dokuz farklı modelin veri kümeleri (hedef değişken) bazında karşılaştırılmasının en uygun yaklaşım olacağını düşünmekteyim. Karşılaştırma kriterleri olarak ise hız, ölçeklenebilirlik, sınıflandırma kesinliği ve öngörü kesinliği gibi ölçüleri kullanacağım.

Aynı donanım ve yazılım olanaklarının kullanıldığı dikkate alınırsa her bir tekniğin işlem süresinin karşılaştırılması anlamlı olacaktır. Ancak unutulmamalıdır ki işlem süresi modelin elde edilmesi dışında, ek olarak hesaplatılmak istenen ölçülere ve istatistiklere bağlı olarak da değişkenlik gösterecektir. Aşağıdaki tabloda (Tablo 2.21) her bir modelin üretilmesinde ihtiyaç duyulan işlem süreleri görülmektedir.

Tablo 2.21. Hedef Değişken Bazında Kayıt Sayısı ve İşlem Süreleri

	C02	Y01	U03
	3.491 Kayıt	2.869 Kayıt	5.088 Kayıt
Yapay Sinir Ağı	22 dak. 50 sn.	17 dak. 04 sn.	30 dak. 36 sn.
C5.0 Algoritması	06 sn.	04 sn.	07 sn.
Lojistik Regresyon	15 dak. 01 sn.	15 dak. 18 sn.	43 dak. 26 sn.

Tablodan açıkça görülebileceği gibi işlem süresi açısından en hızlı teknik C5.0 algoritması olmuştur. Modelin verilerdeki güncellemelere bağlı olarak yenilenmesi ya da hızlı karar alma ihtiyacının bulunduğu süreçlerde C5.0 algoritmasının kullanılması önerilebilir. Ancak hız kriterini değerlendirirken kayıt ve değişken sayısına duyarlılık ile birlikte düşünmek gerekir.

Her üç tekniğin de kayıt sayısının göreli fazlalığı nedeniyle U03 hedef değişkeni üzerindeki uygulamalarının daha uzun sürede gerçekleştiği görülmektedir. Bununla birlikte, kayıt sayısına bağlı olarak işlem süresinde en dikkate değer artış lojistik regresyon tekniğinde yaşanmıştır. Bu durum istatistik tekniklerin güçlü teorik temelleri ve karmaşık hesaplama süreçleri açısından güvenilir, sonuçlarının açıklanabilirliği açısından güçlü olmalarına rağmen veri madenciliğinin çokça tercih edilen teknikleri arasında yer almamasını açıklamaktadır.

Veri madenciliği temel olarak çok fazla miktarlardaki verilerin analizine dayandığından kullanılan tekniklerin veri madenciliği açısından değerlendirilmesinde veri sayısına duyarlılıkları da önem taşımaktadır. Ölçeklenebilirlik olarak adlandırılan bu ölçüyü de aynı tablo (Tablo 2.22) yolu ile değerlendirmek mümkündür. Bu açıdan, lojistik regresyon analizi diğer tekniklere oranla ölçeklenebilirliği açısından zayıf bir teknik olarak gözükmektedir.

Şüphesiz ki, tekniklerin işlem süreleri ne kadar kısa, ölçeklenebilirlikleri ne kadar üst düzeyde olursa olsun sınıflandırma kesinliği düşük olan modellere güvenilmesi söz konusu olmayacaktır. Bu nedenle modellerin karşılaştırılmasında sınıflandırma kesinliğini dikkate almak bir zorunluluk sayılabilir.

Ayrıca öğrenme kümesi üzerindeki yüksek sınıflandırma başarısı sınama kümesi üzerinde tekrar etmiyorsa, model güvenilir olmaktan uzak olacaktır. Eğer bir sınıflandırma modeli yeni kayıtlar üzerinde yüksek düzeyde doğru öngörülerde bulunamıyorsa uygulanabilirliği şüphe yaratacaktır. Bu nedenle modelleri karşılaştırmada son ve en önemli kriter olarak öngörü kesinliği dikkate alınmalıdır.

Bu çalışmada elde etmiş olduğum modellerin hem sınıflandırma kesinliği hem de öngörü kesinliği açısından hedef değişkenlere göre karşılaştırılması için aşağıdaki tablolardan faydalanabiliriz.

Tablo 2.22. C02 için Sınıflandırma ve Öngörü Kesinliği

	C02	
	Sınıflandırma Kesinliği	Öngörü Kesinliği
Yapay Sinir Ağı	%83,63	%84,08
C5.0 Algoritması	%81,04	%82,03
Lojistik Regresyon	%84,90	%84,34

İlk tablo (Tablo 2.22) C02 değişkeni üzerinde gerçekleştirilen modellerin karşılaştırılmasını içermektedir. Sınıflandırma kesinlikleri açısından bakıldığında modeller arasında önemli bir farklılık görülmemektedir. Yine de sıralamak gerekirse, en yüksek başarının lojistik regresyon tekniğine ait olduğu ve ardından yapay sinir ağı algoritmasının geldiği görülmektedir.

Modellerin C02 değişkeni üzerindeki öngörü kesinliği değerlendirildiğinde de birbirine yakın oranların gerçekleştiğini görmekteyiz. Sınıflandırma kesinliğindeki sıralama bu ölçüde de değişmemektedir. Ancak lojistik regresyonun öngörülerinde sınıflandırma kesinliğine göre yaşanan azalma beklenir bir durumu ifade ederken, diğer iki teknikteki yükselme dikkat çekici bir sonuç olarak görülmelidir.

Y01 değişkeni üzerinde gerçekleştirilen uygulamaların karşılaştırıldığı ikinci tabloda da (Tablo 2.23) bir önceki değerlendirmeden farklı bir durum gözükmemektedir. Hem sınıflandırma hem de öngörü kesinlikleri birbirine yakın değerlerle ifade edilirken, sıralamada değişiklik olmamıştır. Ancak bu kez, beklendiği gibi tüm modellerin öngörü kesinlikleri sınıflandırma kesinliklerine göre düşüş göstermiştir.

Tablo 2.23. Y01 için Sınıflandırma ve Öngörü Kesinliği

	Y01	
	Sınıflandırma Kesinliği	Öngörü Kesinliği
Yapay Sinir Ağı	%86,12	%83,24
C5.0 Algoritması	%82,10	%81,21
Lojistik Regresyon	%86,70	%84,84

Son tablo (Tablo 2.24) ise U03 değişkeni üzerindeki uygulamalara ait olup, önceki uygulamalarla paralel sonuçların oluştuğuna işaret etmektedir. Her iki kriter açısından da modellerin başarı sıralaması lojistik regresyon, yapay sinir ağı ve C5.0 algoritması şeklinde gerçekleşmiştir.

Tablo 2.24. U03 için Sınıflandırma ve Öngörü Kesinliği

	U03	
	Sınıflandırma Kesinliği	Öngörü Kesinliği
Yapay Sinir Ağı	%88,95	%87,44
C5.0 Algoritması	%88,30	%88,04
Lojistik Regresyon	%90,20	%88,13

Sınıflandırma ve öngörü başarılarının değişkenden değişkene gösterdiği farklılık ise başka bir gerçekliği ortaya koymaktadır. Burada dikkat çekici olan, tüm modellerin en yüksek başarıları U03 değişkeninde göstermeleridir. Bu değişkene ait veri kümesinin diğer veri kümelerinden daha fazla kayıt içerdikleri düşünülürse, veri

madenciliği uygulamalarında veri sayısının önemini doğrulamış oluruz. Veri kümelerindeki örüntülerin tespitinde en önemli etken veri sayısıdır. Veri sayısı ne kadar artarsa, bir örüntü ya da ilişkinin tespit edilebilmesi konusundaki başarı da o oranda artacaktır.

Bu anlamda, bir başka ilginç bulgu ise Y01 değişkenindeki veri sayısının görece C02 değişkeninden daha az olmasına rağmen sınıflandırma ve öngörü başarılarının daha yüksek olmasıdır. Bu durum, veri madenciliğinde veri kalitesinin önemini akla getirmektedir.

Sınıflandırma ve öngörü kesinlikleri açısından en zayıf model C5.0 algoritması olarak görülmektedir. Yapay sinir ağı algoritmasının ise hem hız hem ölçeklenebilirlik hem de sınıflandırma ve öngörü kesinliği açısından en iyi ya da en kötü olmamakla birlikte kararlı bir duruş sergilediği söylenebilir.

2.6. DEĞERLENDİRME

Bu aşamaya kadar yapılan değerlendirmelerde modellerin veri madenciliği açısından başarısı dikkate alınmıştı. Bu aşamada ise elde edilen modellerin iş hedefleri açısından değerlendirmesi yapılacaktır. Bu kapsamda sonuçları değerlendirme, süreci gözden geçirme ve bir sonraki adımı belirleme görevleri üzerinde durulacaktır.

Daha önce belirtildiği gibi, bu çalışmada iş hedeflerine yönelik bir başarı kriteri belirlenmemiştir. Ancak kullanılan tekniklerle müşterilerin tercih ve davranışlarını temsil kabiliyeti yüksek modellerin elde edilebilmesini ve bu modelin potansiyel müşterin davranışlarını öngörebilme konusundaki beklentilere cevap verebilecek nitelikte olması beklentisinden söz edilmişti.

Geliştirilen modellerin tamamı için her iki beklentinin de karşılandığını söylemek mümkündür. Veri madenciliği amacına dönük değerlendirmede belirtildiği gibi her bir model hem öğrenme kümesi üzerinde sınıflandırma görevini hem de sınama kümesi üzerinde öngörü görevini yaklaşık olarak aynı düzeylerde ve başarıyla gerçekleştirmiştir.

Yeni bir ürünün pazarlanması faaliyetlerinde hedef müşteri profilinin yaklaşık olarak %80-90 oranında öngörülebilmesi hem zaman hem de maliyetleri önemli düzeyde azaltabilecek bir bilgiye sahip olunması anlamına gelir. Özellikle öngörü kesinliklerinin %80'den yüksek olması bu modellerle alınacak işe yönelik kararların %20'den daha az risk içereceğini göstermektedir.

İşe yönelik uygulamalarda yapılacak durum değerlendirmesi doğrultusunda bu çalışmada uygulanan üç tekniğin de kullanılması söz konusu olabilir. Elde edilen bulgular doğrultusunda bir değerlendirme yapmak gerekirse, hızlı karar verme gereği olan durumlarda C5.0 algoritmasını, veri sayısının görece fazla olmadığı durumlarda lojistik regresyon tekniğini, çok fazla sayıda veri üzerinde çalışılması gerektiğinde yapay sinir ağları algoritmasını kullanmak akılcı olacaktır.

Yapılan çalışma mümkün olduğunca titizlikle gerçekleştirilmiş, her aşamada ve her görevde alınan kararlar veri madenciliğinin yinelemeli doğası gereği tekrar tekrar değerlendirilmiştir. Başlangıç verilerinden model kurma görevine kadar her adım defalarca tekrarlanarak en uygun sonuçlara ulaşılmasına çaba harcanmıştır.

Gelinen noktada süreç boyunca yapılmış bir hata olmadığı inancı oluşmuştur. Geride bırakılan sürecin başarısının ancak bir sonraki aşama olan konuşlandırma ile gerçek anlamda ortaya çıkarılması mümkün olacaktır.

Bununla birlikte, kontrol edemeyeceğim nedenlerle modellerin ideal düzeyde gerçekleşmiş olmama olasılığı halen mümkündür. Bu anlamda, en büyük sorun başlangıç verilerinden kaynaklanabilir. Hedef değişkenlerin sınıflandırılmasında önem arz edebilecek ancak veri kümesinde mevcut olmayan değişken ya da değişkenlerin olması söz konusu olabilir. Aranılan örüntülerin tespitinde, ilgili örüntüyü temsil eden kayıtların sayısı yetersiz kalmış olabilir. Böylesine sorunları aşmak ancak veri kümesinin güncellenmesi veya yeniden edinilmesi ile aşılabilir. Bunun dışındaki sorunların tümü ise sürecin yenilenmesi ile aşılabilir nitelikte sorunlar olacaktır.

Bu çalışma için bir sonraki adımda yapılması gereken, konuşlandırma aşamasına geçmeden önce elde edilen modellerin sınırlı sürede ve sınırlı sayıda müşteri üzerinde denenmesidir. Bu kısıtlı uygulamadan alınacak derslerle veri

madenciliği sürecinin yenilenmesi ve sorunlardan arındırılması, konuşlandırmanın başarı düzeyini arttıracaktır.

2.7. KONUŞLANDIRMA

Konuşlandırma aşamasının gerçeklenebilmesi için veri madenciliği projesinin veriye sahip olan kurum tarafından ya da bu kurumla işbirliği içerisinde yapılabilmesi gerekmektedir. Bu çalışmanın gerçekleştirilmesi için bu türden bir işbirliği olanağının oluşturulmasına çalışılmış ancak böyle bir olanak yaratılamamıştır. Özellikle ticari bilgi içeren verilerin kullanılmasında işbirliği yapmak bir yana bu verilerin paylaşılmasında dahi önemli engellerle karşılaşılmaktadır.

Bu yüzden, çalışmanın konuşlandırma aşaması bir uygulama içermekten çok öneri ve temennilerden oluşacaktır. Bu kapsamda konuşlandırmayı planlama, izleme ve bakımın planlanması, sonuç raporu oluşturma ve projenin gözden geçirilmesi görevlerinden söz edilecektir.

Konuşlandırmayı planlama, geliştirilen model ya da modellerin iş hedefleri doğrultusunda uygulanmasının planlanmasıdır. Bu uygulamanın nasıl yapılacağına, çalışmanın iş hedefleri perspektifinden bakmak gerekir.

Modelleme sürecinin bağımlı değişkenleri belirli bankacılık ürünlerine sahip olma bilgilerini içerdiğine göre, projenin hedef kitlesi müşteri veri tabanında yer alan ancak ilgili ürünlere henüz sahip olmayan müşterilerin topluluğu olacaktır.

Bu doğrultuda konuşlandırma planı, geliştirilen modellerin ilgili ürünlere sahip olmayan müşterilerin oluşturduğu bir veri kümesi üzerinde uygulanması sonucu, potansiyel müşterilerin tespit edilmesi üzerine kurgulanmalıdır. Modellerin ilgili ürüne sahip olması gerektiğine işaret ettiği müşteriler, bu ürünün pazarlanması faaliyeti açısından potansiyel müşteri kitlesi olarak kabul edilmelidir. Bu müşterilerin ilgili ürünü satın alma olasılıkları, uygulanan modelin öngörü kesinliği ile orantılı olacaktır.

Konuşlandırmanın gerçekleştirilmesi aynı zamanda model ya da modellerin gerçek yaşam üzerinde test edilmesi anlamını taşımaktadır. Bu aşamanın her bir

adımında, veri madenciliğinin yinelemeli doğası gereği, ortaya çıkan sonuçların gözden geçirilmesi ve gerektiğinde sürecin bazı adımlarının yinelenmesi gerekecektir. Projenin izlenmesi ve bakımı, veri madenciliği sürecinin konuşlandırma aşamasında da devam ettiği anlamına gelmektedir.

Bu çalışma, bankacılık müşteri veri tabanından elde edilen rastlantısal veri kümesi üzerinde başlatılmıştır. Veri kümesinin analize hazır hale getirilmesi için değişken seçimi, kayıtların seçimi, verilerin sorunlarından arındırılması, hedef değişkenlerin belirlenmesi ve veri kümesinin hedef değişkenlere uygun olarak yapılandırılması gibi faaliyetler yürütülmüştür.

Veri madenciliği sürecinde yürütülen çalışmaların tümü dikkate alındığında en fazla çabanın ve zamanın beklendiği gibi veri kümesinin anlaşılması ve analize hazır hale getirilmesi aşamasında tüketildiğini söylemek mümkündür. Çok fazla değişken ve kayıt içeren veri kümesini anlamak ve analize hazırlamak, oldukça karmaşık ve özenle çalışmayı gerektiren bir süreç olmuştur.

Modelleme aşamasında ise kullanılan yazılımın öğrenilmesi ve en uygun modelin üretilebilmesi, en fazla zaman ve çaba gerektiren çalışmalar olmuştur. Yazılımın doğru bir şekilde kullanılabilmesi ve en uygun modelin üretilebilmesi için her bir uygulamanın defalarca tekrarlanması gerekmiştir.

Bu çalışmada uygulanan veri madenciliği sürecinin tamamı dikkate alındığında, verilerin analize hazırlanmasının veri madenciliğinde en fazla zamana ihtiyaç duyulan, en fazla çaba harcanan, en karmaşık ve en hassas kararların alındığı işlemleri kapsadığını edindiğim deneyim doğrultusunda kolaylıkla söyleyebilirim. Ayrıca, projenin gerçekleştirilmesinde gelişmiş yazılım ve donanım olanaklarına sahip olmanın ne derece önemli olduğunu defalarca gözlemledim. Ve son olarak, veri madenciliği sürecinin doğru bir şekilde yürütülmesi ve uygulanabilmesi için hem işe yönelik hem de uygulanan tekniklere yönelik güçlü bir bilgi birikiminin gerektiğini uygulama boyunca öğrendim.

SONUÇ

Bu çalışmada, veri madenciliği standart sürecinin tüm aşamaları bankacılık müşteri veri tabanından rastlantısal olarak seçilmiş veri kümesi üzerinde uygulanmış ve çok geniş bir çalışma alanı olan veri madenciliğinin sınıflandırma fonksiyonu üzerinde durulmuştur.

Veri kümesi, başlangıçta 188 değişken ve 17.595 kayıt içerirken, veri hazırlama aşamasında yürütülen faaliyetlerle 103 değişken ve 13.274 kayıt içerecek şekilde biçimlendirilmiştir.

Uygulama, birden çok bağımlı değişken üzerinde birden çok sınıflandırma tekniğini kullanarak bu tekniklerin karşılaştırılması üzerine kurgulanmıştır. Bu nedenle, veri madenciliğinin üç önemli bileşeni olan istatistik, yapay öğrenme ve veri tabanı teknolojilerini temsil edecek şekilde lojistik regresyon analizi, yapay sinir ağları ve C5.0 karar kuralı türetme algoritması uygulamada kullanılacak sınıflandırma teknikleri olarak belirlenmiştir.

Bu tekniklerin çeşitli bankacılık ürünlerine sahip olma bilgisini içeren üç farklı kategorik değişken üzerinde uygulanması ile toplam dokuz farklı model geliştirilmiştir.

Modellerin tarafsız bir şekilde karşılaştırılması için her bağımlı değişkene ilişkin tek bir veri kümesi kullanılmış ve karşılaştırma ölçütleri olarak hız, ölçeklenebilirlik, sınıflandırma kesinliği ve öngörü kesinliği kullanılmıştır.

Hız ölçütü açısından yapılan değerlendirmede, C5.0 algoritmasının tartışmasız bir şekilde avantaj sağladığı görülmüştür. Aynı donanım ve yazılım olanakları çerçevesinde lojistik regresyon analizinin 15 dakika ile 43 dakika arasında, yapay sinir ağlarının 22 dakika ile 30 dakika arasında sonuç verdiği uygulamaları C5.0 algoritması 4 saniye ile 7 saniye arasında gerçekleştirmiştir.

Verilerin sıkça güncellendiği, modelin sürekli yinelenmesi ve karar sürecine hızla dahil edilmesinin gerektiği durumlarda C5.0 algoritmasının önerilebileceği sonucuna varılmıştır.

Ölçeklenebilirlik açısından yapılan değerlendirmede, yapay sinir ağları ve C5.0 algoritmasının veri sayısına daha az duyarlı olduğu, lojistik regresyon tekniğinin ise veri sayısındaki artıştan etkilendiği gözlemlenmiştir. Kayıt sayısı 3.491 iken 15 dakika 01 saniyede sonuca ulaşan lojistik regresyon, kayıt sayısı 5.088 olduğunda 43 dakika 26 saniye gibi bir işlem süresine ihtiyaç duymuştur.

Günümüzün çevrimiçi veri tabanlarında her gün milyonlara varan kaydın biriktiği düşünülürse, lojistik regresyonun görece mütevazı sayılabilecek veri miktarları için daha uygun olacağı görüşü oluşmuştur.

Modellerin, geliştirildikleri veri kümesi üzerinde gösterdikleri sınıflandırma başarısının bir ölçüsü olan, sınıflandırma kesinliği açısından anlamlı bir farklılık göstermedikleri görülmüştür. Tüm modellerin %80 veya daha fazla sınıflandırma başarısı gösterdiği dikkat çekerken, lojistik regresyon her üç bağımlı değişken üzerinde de en yüksek sınıflandırma kesinliğine sahip olmuştur.

Bir diğer karşılaştırma kriteri olarak ise modellerin öğrenme kümesinden tamamen bağımsız bir veri kümesi üzerinde gösterdikleri sınıflandırma başarısını ifade eden öngörü kesinliği kullanılmıştır. Bu kritere göre de modellerin anlamlı bir farklılık göstermedikleri tespit edilmiştir. Ancak öngörü başarısında da görece en yüksek oranların lojistik regresyon uygulamalarında gerçekleştiği dikkat çekmektedir.

Buradan hareketle, zaman sınırlamasının esnek olduğu, yeterli yazılım ve donanım olanaklarının mevcut olduğu ve hata riskinin en küçükleştirilmesinin esas alındığı çalışmalarda lojistik regresyon tekniğinin önerilmesi uygun bulunmuştur.

Ayrıca, veri sayısı arttıkça her üç modelde de sınıflandırma ve öngörü kesinliğinin arttığı görülmüş ve bu durum, veri madenciliğinde veri sayısı arttıkça örüntü ve ilişkilerin tespit edilmesinin kolaylaşacağı şeklindeki yaklaşımın doğrulanması olarak algılanmıştır.

Bununla birlikte, en az sayıda veri ile gerçekleştirilen uygulamaların beklendiği gibi en düşük seviyede sınıflandırma ve öngörü başarısını göstermemiş olması, örüntü

ve ilişkilerin tespitinde veri kalitesinin en önemli rolü oynayacağı gerçeğini hatırlatan nitelikte bir bulgu olarak görülmüştür.

Bir başka beklendik durum ise sınıflandırma ve öngörü kesinliklerinin karşılaştırılmasında yaşanmıştır. Toplam dokuz modelden yedisinde öngörü başarısı sınıflandırma başarısının altında gerçekleşmiştir.

Sonuç olarak, gerçekleştirilen standart veri madenciliği süreci ve uygulanan sınıflandırma teknikleri ışığında, sürecin en zorlu kısmının veri hazırlama aşaması olduğu, sürecin her aşamasında işe ve tekniklere ilişkin güçlü bilgi birikimine ihtiyaç duyulduğu, veri sayısının ve veri kalitesinin uygulamaların başarısında önemli birer faktör olduğu, güncel ve hızlı karar verme ihtiyaçları doğrultusunda en uygun seçimin C5.0 algoritması olacağı, lojistik regresyonun ölçeklenebilirlik açısından zayıflığı ancak güvenilirliğin esas alındığı uygulamalarda üstünlerinin olduğu, hem zaman hem de güvenilirlik açısından optimal çözümün yapay sinir ağlarının kullanılması olacağı görüşleri ağırlık kazanmıştır.

EKLER

EK 1. PROJE PLANI

- Veri kümesinin incelenmesi
- Basit veri kalitesi sorunlarının giderilmesi
- Hedef değişkenlerin belirlenmesi
- İleri düzeyde veri kalitesi çalışmasının yürütülmesi
- Veri kümesinin hedef değişkenler bazında biçimlendirilmesi
- Tanımlayıcı istatistiklerin incelenmesi
- Sınama tasarımının geliştirilmesi
- Lojistik regresyon modellerinin oluşturulması
- Lojistik regresyon modellerinin sınanması
- Lojistik regresyon modellerinin değerlemesi
- Yapay sinir ağı modellerinin oluşturulması
- Yapay sinir ağı modellerinin sınanması
- Yapay sinir ağı modellerinin değerlemesi
- C5.0 modelinin oluşturulması
- C5.0 modelinin sınanması
- C5.0 modelinin değerlemesi
- Modellerin karşılaştırılması
- İş hedefleri açısından değerlendirme
- Konuşlandırmaya ilişkin önerilerin geliştirilmesi

EK 2. TANIMLAYICI İSTATİSTİK VE SIKLIK DAĞILIMLARI

D01		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	2	331	2,5	2,5	2,5
	3	3078	23,2	23,2	25,7
	4	4381	33,0	33,0	58,7
	5	3148	23,7	23,7	82,4
	6	2336	17,6	17,6	100,0
	Toplam	13274	100,0	100,0	

D02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	1	9752	73,5	73,5	73,5
	2	3522	26,5	26,5	100,0
	Toplam	13274	100,0	100,0	

D03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	1	3943	29,7	29,7	29,7
	2	8980	67,7	67,7	97,4
	3	351	2,6	2,6	100,0
	Toplam	13274	100,0	100,0	

D04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	1	2158	16,3	16,3	16,3
	2	1887	14,2	14,2	30,5
	3	4734	35,7	35,7	66,1
	5	3434	25,9	25,9	92,0
	6	1061	8,0	8,0	100,0
	Toplam	13274	100,0	100,0	

M02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	4	1323	10,0	10,0	10,0
	5	4082	30,8	30,8	40,7
	6	2436	18,4	18,4	59,1
	7	1433	10,8	10,8	69,9
	8	1000	7,5	7,5	77,4
	9	806	6,1	6,1	83,5
	> 10	2194	16,5	16,5	100,0
	Toplam	13274	100,0	100,0	

M03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	1	10248	77,2	77,2	77,2
	2	3026	22,8	22,8	100,0
	Toplam	13274	100,0	100,0	

M04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	1	333	2,5	2,5	2,5
	2	790	6,0	6,0	8,5
	3	2642	19,9	19,9	28,4
	4	1183	8,9	8,9	37,3
	5	536	4,0	4,0	41,3
	6	4589	34,6	34,6	75,9
	7	1669	12,6	12,6	88,5
	8	1532	11,5	11,5	100,0
	Toplam	13274	100,0	100,0	

H01		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	1364	10,3	10,3	10,3
	1	11910	89,7	89,7	100,0
	Toplam	13274	100,0	100,0	

H02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	8161	61,5	61,5	61,5
	1	5113	38,5	38,5	100,0
	Toplam	13274	100,0	100,0	

H03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11932	89,9	89,9	89,9
	1	1342	10,1	10,1	100,0
	Toplam	13274	100,0	100,0	

H04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11699	88,1	88,1	88,1
	1	1575	11,9	11,9	100,0
	Toplam	13274	100,0	100,0	

H05		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12417	93,5	93,5	93,5
	1	857	6,5	6,5	100,0
	Toplam	13274	100,0	100,0	

H06		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12459	93,9	93,9	93,9
	1	815	6,1	6,1	100,0
	Toplam	13274	100,0	100,0	

H07		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11512	86,7	86,7	86,7
	1	1762	13,3	13,3	100,0
	Toplam	13274	100,0	100,0	

H08		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11808	89,0	89,0	89,0
	1	1466	11,0	11,0	100,0
	Toplam	13274	100,0	100,0	

H09		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12081	91,0	91,0	91,0
	1	1193	9,0	9,0	100,0
	Toplam	13274	100,0	100,0	

H10		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11796	88,9	88,9	88,9
	1	1478	11,1	11,1	100,0
	Toplam	13274	100,0	100,0	

H11		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12904	97,2	97,2	97,2
	1	370	2,8	2,8	100,0
	Toplam	13274	100,0	100,0	

H12		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	1364	10,3	10,3	10,3
	1	8272	62,3	62,3	72,6
	2	2479	18,7	18,7	91,3
	3	765	5,8	5,8	97,0
	4	244	1,8	1,8	98,9
	5	150	1,1	1,1	100,0
	Toplam	13274	100,0	100,0	

H13		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	8161	61,5	61,5	61,5
	1	2854	21,5	21,5	83,0
	2	1512	11,4	11,4	94,4
	3	436	3,3	3,3	97,7
	4	186	1,4	1,4	99,1
	5	125	,9	,9	100,0
	Toplam	13274	100,0	100,0	

H14		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11932	89,9	89,9	89,9
	1	1105	8,3	8,3	98,2
	2	237	1,8	1,8	100,0
	Toplam	13274	100,0	100,0	

H15		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11699	88,1	88,1	88,1
	1	1126	8,5	8,5	96,6
	2	301	2,3	2,3	98,9
	3	148	1,1	1,1	100,0
	Toplam	13274	100,0	100,0	

H16		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12417	93,5	93,5	93,5
	1	857	6,5	6,5	100,0
	Toplam	13274	100,0	100,0	

H17		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12459	93,9	93,9	93,9
	1	815	6,1	6,1	100,0
	Toplam	13274	100,0	100,0	

H18		
	Geçerli	13274
	Kayıp	0
Ortalama		88,8260
Standart Sapma		799,10045
Çarpıklık		25,712
Basıklık		843,643

H19		
	Geçerli	13274
	Kayıp	0
Ortalama		511,3441
Standart Sapma		4532,74880
Çarpıklık		16,681
Basıklık		352,559

H20		
	Geçerli	13274
	Kayıp	0
Ortalama		819,3543
Standart Sapma		8901,12522
Çarpıklık		36,989
Basıklık		2094,509

H21		
	Geçerli	13274
	Kayıp	0
Ortalama		2946,5244
Standart Sapma		29467,53890
Çarpıklık		29,047
Basıklık		1134,220

H22		
	Geçerli	13274
	Kayıp	0
Ortalama		-7,5732
Standart Sapma		108,06330
Çarpıklık		-35,220
Basıklık		1876,801

H23		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13056	98,4	98,4	98,4
	1	218	1,6	1,6	100,0
	Toplam	13274	100,0	100,0	

FH		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	130	1,0	1,0	1,0
	1	7593	57,2	57,2	58,2
	2	3412	25,7	25,7	83,9
	3	1408	10,6	10,6	94,5
	4	520	3,9	3,9	98,4
	5	211	1,6	1,6	100,0
	Toplam	13274	100,0	100,0	

TH		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	130	1,0	1,0	1,0
	1	6292	47,4	47,4	48,4
	2	2658	20,0	20,0	68,4
	3	1523	11,5	11,5	79,9
	4	989	7,5	7,5	87,3
	5	597	4,5	4,5	91,8
	6	426	3,2	3,2	95,0
	7	263	2,0	2,0	97,0
	8	153	1,2	1,2	98,2
	9	98	,7	,7	98,9
	10	145	1,1	1,1	100,0
	Toplam	13274	100,0	100,0	

L01		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13206	99,5	99,5	99,5
	1	68	,5	,5	100,0
	Toplam	13274	100,0	100,0	

L02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13247	99,8	99,8	99,8
	1	27	,2	,2	100,0
	Toplam	13274	100,0	100,0	

L03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13149	99,1	99,1	99,1
	1	125	,9	,9	100,0
	Toplam	13274	100,0	100,0	

L04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13216	99,6	99,6	99,6
	1	58	,4	,4	100,0
	Toplam	13274	100,0	100,0	

L05		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13247	99,8	99,8	99,8
	1	27	,2	,2	100,0
	Toplam	13274	100,0	100,0	

L06		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13150	99,1	99,1	99,1
	1	124	,9	,9	100,0
	Toplam	13274	100,0	100,0	

L07		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13206	99,5	99,5	99,5
	1	68	,5	,5	100,0
	Toplam	13274	100,0	100,0	

L08		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13247	99,8	99,8	99,8
	1	27	,2	,2	100,0
	Toplam	13274	100,0	100,0	

L09		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13149	99,1	99,1	99,1
	1	125	,9	,9	100,0
	Toplam	13274	100,0	100,0	

L10	Geçerli	13274
	Kayıp	0
Ortalama		-5,8337
Standart Sapma		195,81778
Çarpıklık		-81,035
Basıklık		7885,902

L11	Geçerli	13274
	Kayıp	0
Ortalama		-72,1922
Standart Sapma		3845,63991
Çarpıklık		-73,903
Basıklık		5765,588

L12	Geçerli	13274
	Kayıp	0
Ortalama		-26,9856
Standart Sapma		467,24282
Çarpıklık		-45,147
Basıklık		2964,138

FL		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13057	98,4	98,4	98,4
	1	217	1,6	1,6	100,0
	Toplam	13274	100,0	100,0	

C01		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10806	81,4	81,4	81,4
	1	2468	18,6	18,6	100,0
	Toplam	13274	100,0	100,0	

C02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11589	87,3	87,3	87,3
	1	1685	12,7	12,7	100,0
	Toplam	13274	100,0	100,0	

C03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12950	97,6	97,6	97,6
	1	324	2,4	2,4	100,0
	Toplam	13274	100,0	100,0	

C04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13121	98,8	98,8	98,8
	1	153	1,2	1,2	100,0
	Toplam	13274	100,0	100,0	

C06		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	9013	67,9	67,9	67,9
	1	4261	32,1	32,1	100,0
	Toplam	13274	100,0	100,0	

C07		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11624	87,6	87,6	87,6
	1	1650	12,4	12,4	100,0
	Toplam	13274	100,0	100,0	

C08		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12347	93,0	93,0	93,0
	1	927	7,0	7,0	100,0
	Toplam	13274	100,0	100,0	

C09		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13217	99,6	99,6	99,6
	1	57	,4	,4	100,0
	Toplam	13274	100,0	100,0	

C10		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13206	99,5	99,5	99,5
	1	68	,5	,5	100,0
	Toplam	13274	100,0	100,0	

C12		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10109	76,2	76,2	76,2
	1	3165	23,8	23,8	100,0
	Toplam	13274	100,0	100,0	

C13		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10806	81,4	81,4	81,4
	1	2068	15,6	15,6	97,0
	2	400	3,0	3,0	100,0
	Toplam	13274	100,0	100,0	

C14		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11589	87,3	87,3	87,3
	1	1410	10,6	10,6	97,9
	2	275	2,1	2,1	100,0
	Toplam	13274	100,0	100,0	

C15		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12950	97,6	97,6	97,6
	1	324	2,4	2,4	100,0
	Toplam	13274	100,0	100,0	

C16		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13121	98,8	98,8	98,8
	1	153	1,2	1,2	100,0
	Toplam	13274	100,0	100,0	

C18		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	9013	67,9	67,9	67,9
	1	2443	18,4	18,4	86,3
	2	1240	9,3	9,3	95,6
	3	374	2,8	2,8	98,5
	4	204	1,5	1,5	100,0
	Toplam	13274	100,0	100,0	

C19	Geçerli	13274
	Kayıp	0
Ortalama		-32,0836
Standart Sapma		153,98277
Çarpıklık		-9,168
Basıklık		124,342

C20	Geçerli	13274
	Kayıp	0
Ortalama		-52,7614
Standart Sapma		424,74010
Çarpıklık		-21,361
Basıklık		732,638

C21	Geçerli	13274
	Kayıp	0
Ortalama		-,0662
Standart Sapma		4,05322
Çarpıklık		-95,260
Basıklık		9914,870

C22	Geçerli	13274
	Kayıp	0
Ortalama		-,5936
Standart Sapma		19,48390
Çarpıklık		-72,673
Basıklık		6607,898

C24	Geçerli	13274
	Kayıp	0
Ortalama		-97,4275
Standart Sapma		524,77757
Çarpıklık		-41,246
Basıklık		2989,389

FC		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	7830	59,0	59,0	59,0
	1	2701	20,3	20,3	79,3
	2	2119	16,0	16,0	95,3
	3	624	4,7	4,7	100,0
	Toplam	13274	100,0	100,0	

TC		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	7830	59,0	59,0	59,0
	1	1960	14,8	14,8	73,8
	2	1711	12,9	12,9	86,6
	3	911	6,9	6,9	93,5
	4	464	3,5	3,5	97,0
	5	398	3,0	3,0	100,0
	Toplam	13274	100,0	100,0	

Y01		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11922	89,8	89,8	89,8
	1	1352	10,2	10,2	100,0
	Toplam	13274	100,0	100,0	

Y02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11792	88,8	88,8	88,8
	1	1482	11,2	11,2	100,0
	Toplam	13274	100,0	100,0	

Y03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13153	99,1	99,1	99,1
	1	121	,9	,9	100,0
	Toplam	13274	100,0	100,0	

Y04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12988	97,8	97,8	97,8
	1	286	2,2	2,2	100,0
	Toplam	13274	100,0	100,0	

Y05		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13263	99,9	99,9	99,9
	1	11	,1	,1	100,0
	Toplam	13274	100,0	100,0	

Y06		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11965	90,1	90,1	90,1
	1	1309	9,9	9,9	100,0
	Toplam	13274	100,0	100,0	

Y07		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11792	88,8	88,8	88,8
	1	1482	11,2	11,2	100,0
	Toplam	13274	100,0	100,0	

Y08		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13155	99,1	99,1	99,1
	1	119	,9	,9	100,0
	Toplam	13274	100,0	100,0	

Y09		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12991	97,9	97,9	97,9
	1	283	2,1	2,1	100,0
	Toplam	13274	100,0	100,0	

Y10		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13263	99,9	99,9	99,9
	1	11	,1	,1	100,0
	Toplam	13274	100,0	100,0	

Y11		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11922	89,8	89,8	89,8
	1	1145	8,6	8,6	98,4
	2	207	1,6	1,6	100,0
	Toplam	13274	100,0	100,0	

Y12		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11792	88,8	88,8	88,8
	1	1482	11,2	11,2	100,0
	Toplam	13274	100,0	100,0	

Y13		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13153	99,1	99,1	99,1
	1	121	,9	,9	100,0
	Toplam	13274	100,0	100,0	

Y14		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12988	97,8	97,8	97,8
	1	286	2,2	2,2	100,0
	Toplam	13274	100,0	100,0	

Y15		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13263	99,9	99,9	99,9
	1	11	,1	,1	100,0
	Toplam	13274	100,0	100,0	

Y16	Geçerli	13274
	Kayıp	0
Ortalama		491,6361
Standart Sapma		5943,98787
Çarpıklık		44,597
Basıklık		2627,354

Y17	Geçerli	13274
	Kayıp	0
Ortalama		376,7390
Standart Sapma		6015,12773
Çarpıklık		31,789
Basıklık		1288,686

Y18	Geçerli	13274
	Kayıp	0
Ortalama		19,2039
Standart Sapma		711,94312
Çarpıklık		51,719
Basıklık		3034,716

Y19	Geçerli	13274
	Kayıp	0
Ortalama		536,8930
Standart Sapma		13259,48965
Çarpıklık		71,205
Basıklık		6191,919

Y20	Geçerli	13274
	Kayıp	0
Ortalama		14,5397
Standart Sapma		979,18493
Çarpıklık		68,740
Basıklık		4802,386

FY		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10808	81,4	81,4	81,4
	1	1784	13,4	13,4	94,9
	2	682	5,1	5,1	100,0
	Toplam	13274	100,0	100,0	

TY		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10808	81,4	81,4	81,4
	1	1701	12,8	12,8	94,2
	2	566	4,3	4,3	98,5
	3	199	1,5	1,5	100,0
	Toplam	13274	100,0	100,0	

U01		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	6039	45,5	45,5	45,5
	1	7235	54,5	54,5	100,0
	Toplam	13274	100,0	100,0	

U02		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10002	75,4	75,4	75,4
	1	3272	24,6	24,6	100,0
	Toplam	13274	100,0	100,0	

U03		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10856	81,8	81,8	81,8
	1	2418	18,2	18,2	100,0
	Toplam	13274	100,0	100,0	

U04		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11660	87,8	87,8	87,8
	1	1614	12,2	12,2	100,0
	Toplam	13274	100,0	100,0	

U05		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12497	94,1	94,1	94,1
	1	777	5,9	5,9	100,0
	Toplam	13274	100,0	100,0	

U06		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	7786	58,7	58,7	58,7
	1	5488	41,3	41,3	100,0
	Toplam	13274	100,0	100,0	

U07		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10316	77,7	77,7	77,7
	1	2958	22,3	22,3	100,0
	Toplam	13274	100,0	100,0	

U08		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10972	82,7	82,7	82,7
	1	2302	17,3	17,3	100,0
	Toplam	13274	100,0	100,0	

U09		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12170	91,7	91,7	91,7
	1	1104	8,3	8,3	100,0
	Toplam	13274	100,0	100,0	

U10		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12860	96,9	96,9	96,9
	1	414	3,1	3,1	100,0
	Toplam	13274	100,0	100,0	

U12		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13192	99,4	99,4	99,4
	1	82	,6	,6	100,0
	Toplam	13274	100,0	100,0	

U13		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	6039	45,5	45,5	45,5
	1	5267	39,7	39,7	85,2
	2	1309	9,9	9,9	95,0
	3	442	3,3	3,3	98,4
	4	217	1,6	1,6	100,0
	Toplam	13274	100,0	100,0	

U14		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10002	75,4	75,4	75,4
	1	3272	24,6	24,6	100,0
	Toplam	13274	100,0	100,0	

U15		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	10856	81,8	81,8	81,8
	1	2418	18,2	18,2	100,0
	Toplam	13274	100,0	100,0	

U16		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	11660	87,8	87,8	87,8
	1	367	2,8	2,8	90,6
	2	296	2,2	2,2	92,8
	3	235	1,8	1,8	94,6
	4	213	1,6	1,6	96,2
	5	503	3,8	3,8	100,0
	Toplam	13274	100,0	100,0	

U17		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	12497	94,1	94,1	94,1
	1	530	4,0	4,0	98,1
	2	247	1,9	1,9	100,0
	Toplam	13274	100,0	100,0	

U19	Geçerli	13274
	Kayıp	0
Ortalama		,4180
Standart Sapma		9,35755
Çarpıklık		45,732
Basıklık		2680,557

U20	Geçerli	13274
	Kayıp	0
Ortalama		3,3937
Standart Sapma		31,49986
Çarpıklık		16,427
Basıklık		378,671

U21		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13169	99,2	99,2	99,2
	1	105	,8	,8	100,0
	Toplam	13274	100,0	100,0	

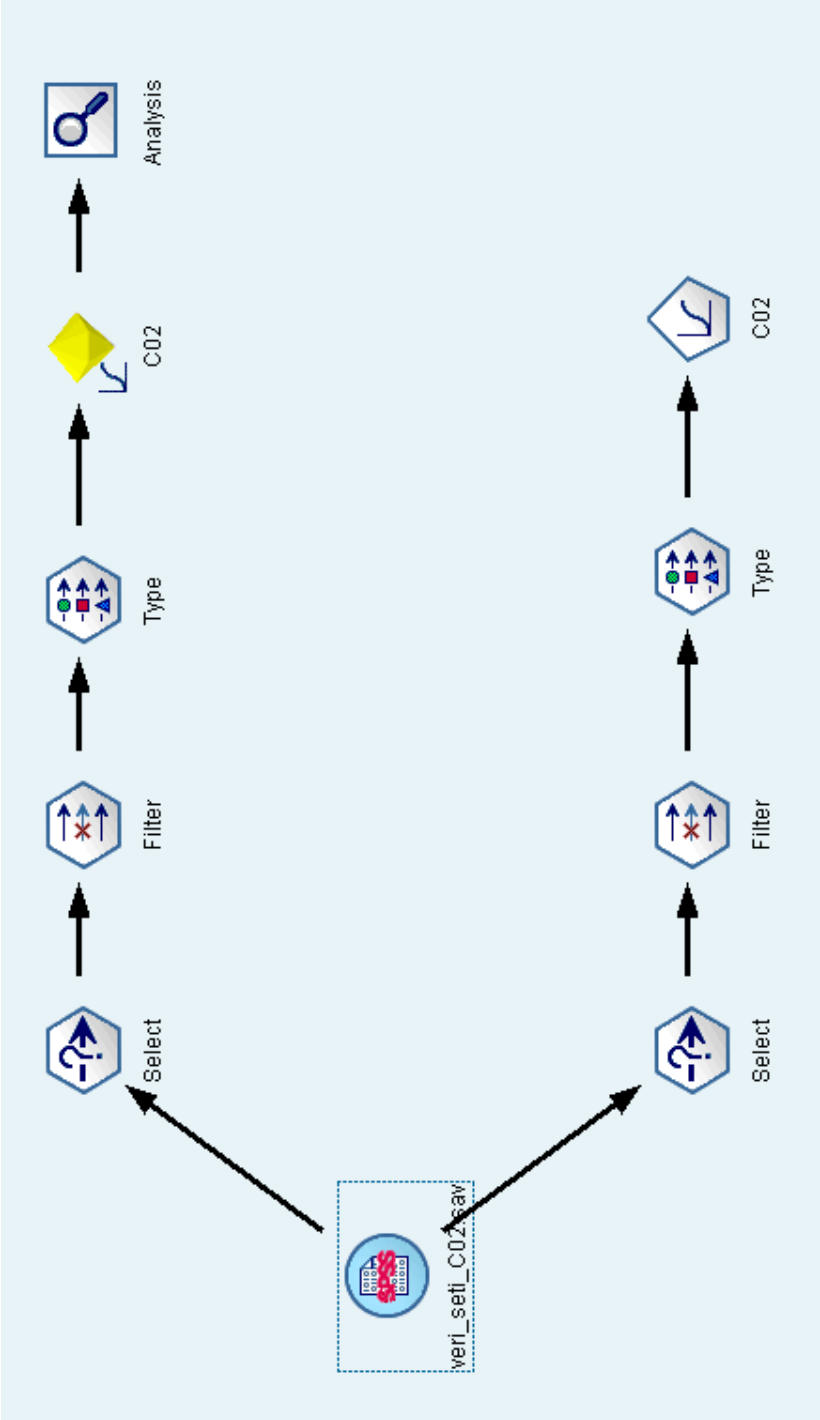
U22		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13182	99,3	99,3	99,3
	1	92	,7	,7	100,0
	Toplam	13274	100,0	100,0	

U23		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	13169	99,2	99,2	99,2
	1	105	,8	,8	100,0
	Toplam	13274	100,0	100,0	

FU		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	5147	38,8	38,8	38,8
	1	4342	32,7	32,7	71,5
	2	1456	11,0	11,0	82,5
	3	1453	10,9	10,9	93,4
	4	677	5,1	5,1	98,5
	5	199	1,5	1,5	100,0
	Toplam	13274	100,0	100,0	

TU		Sıklık	Göreli	Geçerli Göreli	Birikimli Göreli
Geçerli	0	5147	38,8	38,8	38,8
	1	3455	26,0	26,0	64,8
	2	1466	11,0	11,0	75,8
	3	1239	9,3	9,3	85,2
	4	636	4,8	4,8	90,0
	5	398	3,0	3,0	93,0
	6	280	2,1	2,1	95,1
	7	202	1,5	1,5	96,6
	8	187	1,4	1,4	98,0
	9	132	1,0	1,0	99,0
	10	132	1,0	1,0	100,0
	Toplam	13274	100,0	100,0	

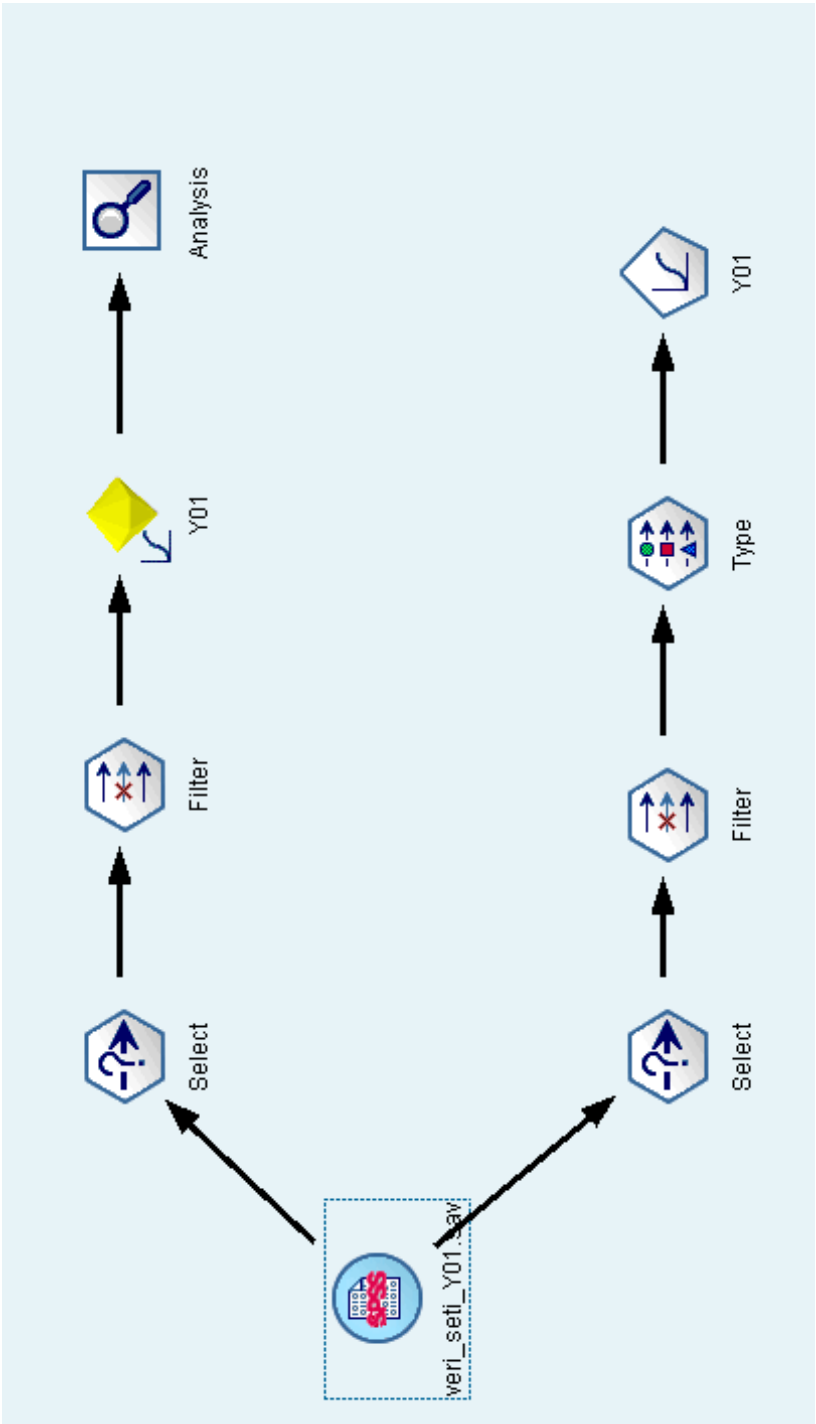
EK 3. C02 İÇİN LOJİSTİK REGRESYON VERİ AKIŞI



EK 4. C02 İÇİN LOJİSTİK REGRESYON DEĞİŞKEN SEÇİMİ

Step Summary							
Model		Action	Effect(s)	-2 Log Likelihood	Chi-Square	df	Sig.
Step 0	0	Entered	Intercept	4432,07			
Step 1	1	Entered	C01	4316,902	115,168	1	0
Step 2	2	Entered	C03	4197,96	118,942	1	0
Step 3	3	Entered	C06	3257,081	940,88	1	0
Step 4	4	Entered	D04	3083,215	173,866	4	0
Step 5	5	Entered	FH	2968,835	114,38	5	0
	6	Removed	C01	2968,918	0,083	1	0,77
Step 6	7	Entered	M02	2721,238	247,68	11	0
Step 7	8	Entered	M03	2663,158	58,079	1	0
Step 8	9	Entered	M04	2424,299	238,859	7	0
Step 9	10	Entered	H03	2404,05	20,25	1	0
Step 10	11	Entered	D02	2391,071	12,979	1	0
Step 11	12	Entered	H01	2380,897	10,174	1	0
Step 12	13	Entered	C18	2366,32	14,577	3	0
	14	Removed	C06	2366,32	0	0	,
Step 13	15	Entered	U01	2359,453	6,867	1	0,01
Step 14	16	Entered	C09	2353,031	6,423	1	0,01
Step 15	17	Entered	C12	2348,178	4,852	1	0,03
Step 16	18	Entered	U09	2343,85	4,328	1	0,04
Step 17	19	Entered	FU	2329,152	14,698	5	0,01
Step 18	20	Entered	H19	2325,075	4,077	1	0,04
Step 19	21	Entered	U07	2320,944	4,132	1	0,04
Stepwise Method: Forward Stepwise							

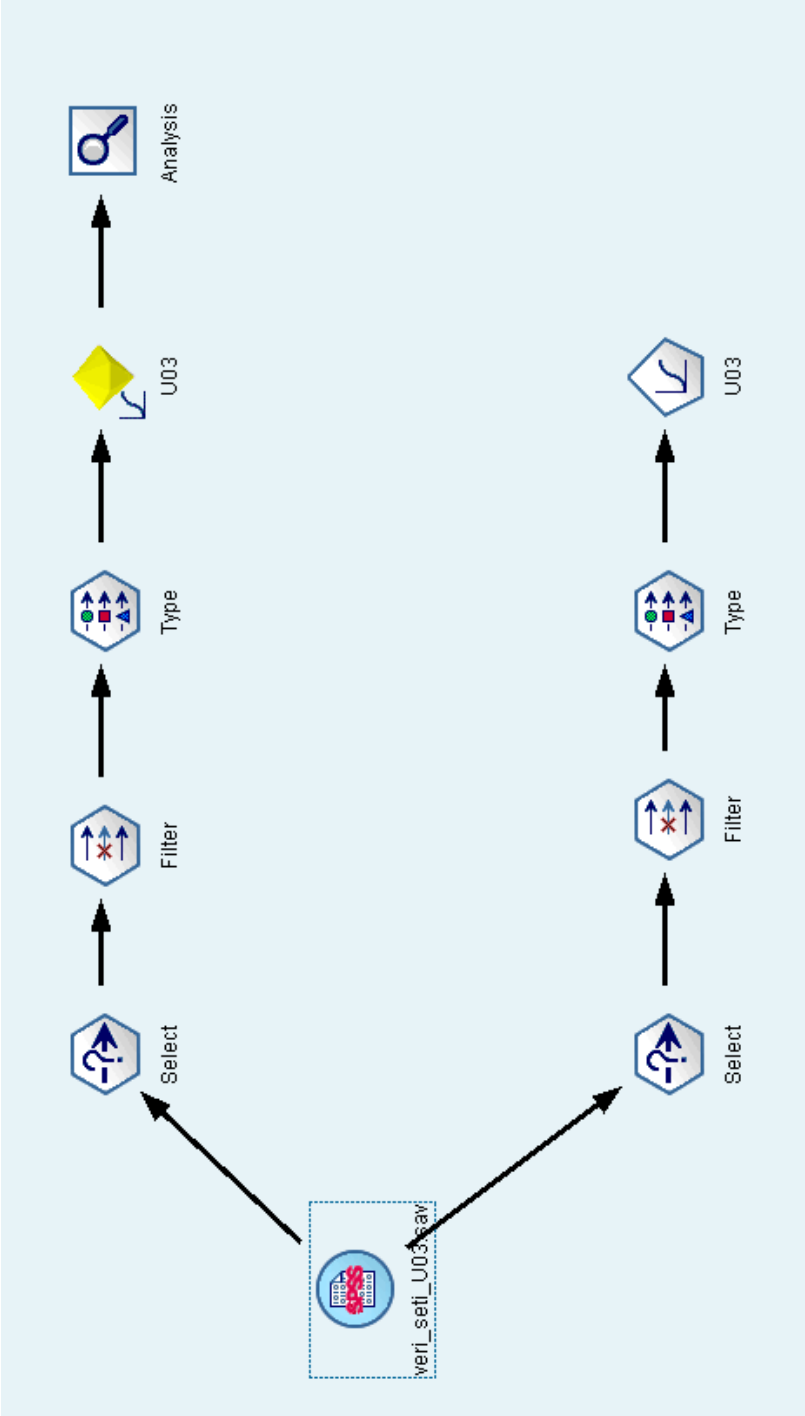
EK 5. Y01 İÇİN LOJİSTİK REGRESYON VERİ AKIŞI



EK 6. Y01 İÇİN LOJİSTİK REGRESYON DEĞİŞKEN SEÇİMİ

Step Summary							
Model		Action	Effect(s)	-2 Log Likelihood	Chi-Square	df	Sig.
Step 0	0	Entered	Intercept	3669,627			
Step 1	1	Entered	C01	3454,812	214,815	1	0
Step 2	2	Entered	C02	3381,762	73,05	1	0
Step 3	3	Entered	C03	3320,41	61,353	1	0
Step 4	4	Entered	D02	3276,272	44,138	1	0
Step 5	5	Entered	D04	3158,24	118,032	4	0
Step 6	6	Entered	FC	3100,907	57,333	3	0
Step 7	7	Entered	FU	2623,98	476,928	5	0
Step 8	8	Entered	H04	2579,656	44,324	1	0
	9	Removed	C02	2581,205	1,549	1	0,213
Step 9	10	Entered	H06	2377,834	203,371	1	0
Step 10	11	Entered	M02	2034,612	343,222	11	0
Step 11	12	Entered	H07	1988,506	46,106	1	0
Step 12	13	Entered	Y04	1947,387	41,12	1	0
Step 13	14	Entered	Y02	1923,293	24,094	1	0
Step 14	15	Entered	M04	1890,963	32,33	7	0
	16	Removed	D04	1896,247	5,284	4	0,259
Step 15	17	Entered	H11	1880,003	16,244	1	0
Step 16	18	Entered	U06	1864,502	15,5	1	0
	19	Removed	FC	1870,161	5,659	3	0,129
Step 17	20	Entered	U17	1857,268	12,894	2	0,002
Step 18	21	Entered	C13	1849,263	8,005	1	0,005
	22	Removed	C01	1849,263	0	0	,
Step 19	23	Entered	U23	1842,032	7,231	1	0,007
Step 20	24	Entered	H10	1834,857	7,175	1	0,007
Step 21	25	Entered	U13	1821,732	13,125	4	0,011
Step 22	26	Entered	C24	1816,423	5,309	1	0,021
Step 23	27	Entered	D01	1806,644	9,78	4	0,044
Step 24	28	Entered	Y05	1802,691	3,953	1	0,047
Step 25	29	Entered	U10	1798,706	3,985	1	0,046
Stepwise Method: Forward Stepwise							

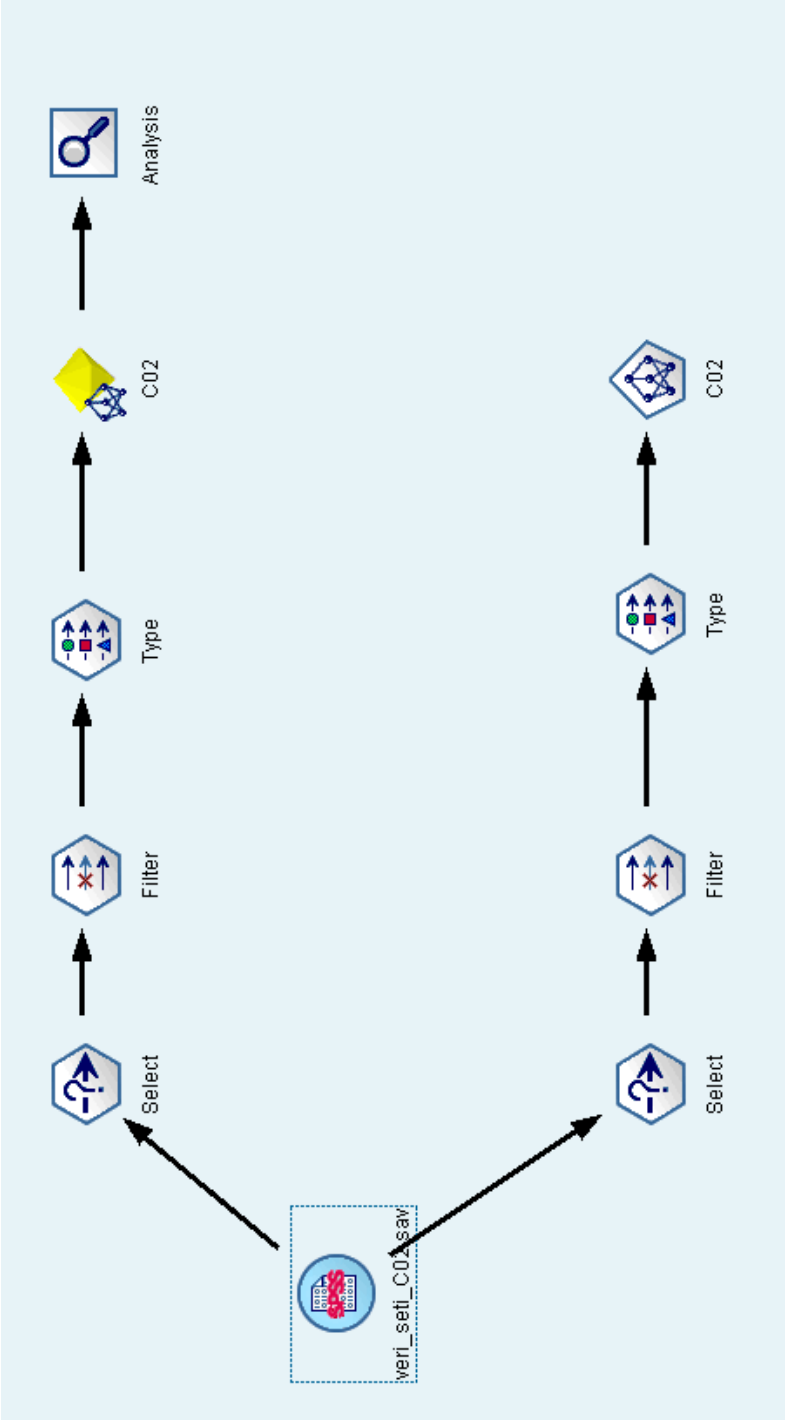
EK 7. U03 İÇİN LOJİSTİK REGRESYON VERİ AKIŞI



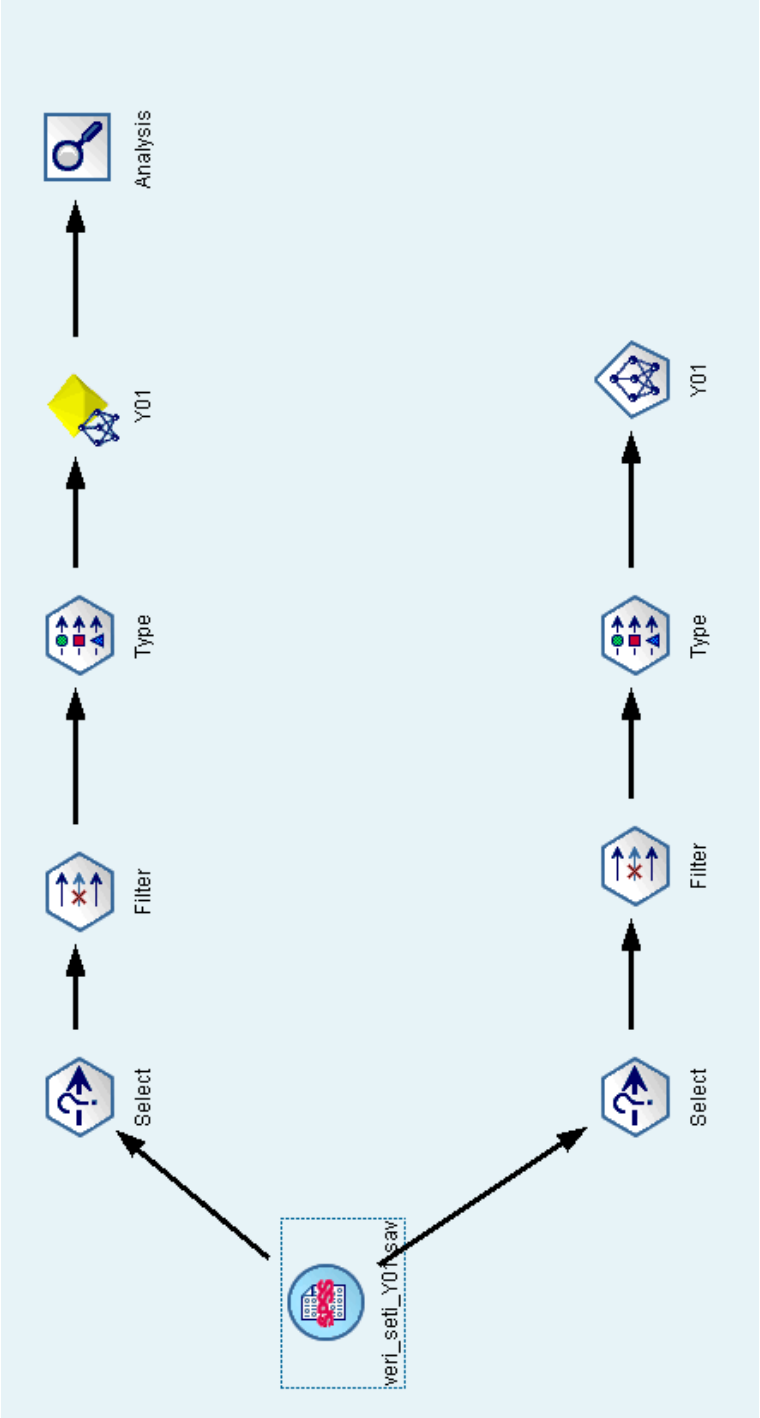
EK 8. U03 İÇİN LOJİSTİK REGRESYON DEĞİŞKEN SEÇİMİ

Step Summary							
Model		Action	Effect(s)	-2 Log Likelihood	Chi-Square	df	Sig.
Step 0	0	Entered	Intercept	6432,687			
Step 1	1	Entered	C01	6026,845	405,842	1	0
Step 2	2	Entered	C02	5776,263	250,582	1	0
Step 3	3	Entered	C03	5388,626	387,637	1	0
Step 4	4	Entered	C06	5309,598	79,028	1	0
Step 5	5	Entered	C07	5258,405	51,193	1	0
Step 6	6	Entered	D01	5049,915	208,49	4	0
Step 7	7	Entered	D04	4852,131	197,784	4	0
Step 8	8	Entered	FC	4797,666	54,465	3	0
Step 9	9	Entered	FH	4532,643	265,024	5	0
Step 10	10	Entered	FY	4335,947	196,696	2	0
Step 11	11	Entered	H01	4280,968	54,979	1	0
Step 12	12	Entered	M02	4075,955	205,013	11	0
Step 13	13	Entered	H12	4016,282	59,673	4	0
	14	Removed	H01	4016,282	0	0	,
Step 14	15	Entered	U01	3840,805	175,477	1	0
Step 15	16	Entered	U02	2688,657	1152,148	1	0
	17	Removed	C02	2688,735	0,078	1	0,78
Step 16	18	Entered	U06	2673,15	15,585	1	0
Step 17	19	Entered	M04	2645,005	28,145	7	0
Step 18	20	Entered	M03	2632,922	12,082	1	0,001
Step 19	21	Entered	U04	2623,449	9,473	1	0,002
	22	Removed	U01	2625,701	2,252	1	0,133
Step 20	23	Entered	C08	2617,643	8,059	1	0,005
	24	Removed	FC	2619,705	2,062	3	0,56
	25	Removed	C06	2619,855	0,15	1	0,698
	26	Removed	C01	2622,533	2,678	1	0,102
Step 21	27	Entered	L02	2616,396	6,137	1	0,013
Step 22	28	Entered	H02	2610,186	6,21	1	0,013
	29	Removed	FH	2617,14	6,954	5	0,224
Step 23	30	Entered	TY	2609,965	7,175	2	0,028
	31	Removed	FY	2611,937	1,972	1	0,16
Step 24	32	Entered	TC	2599,608	12,329	5	0,031
Step 25	33	Entered	Y16	2594,744	4,864	1	0,027
Step 26	34	Entered	Y18	2590,07	4,675	1	0,031
Step 27	35	Entered	C13	2583,253	6,817	2	0,033
Step 28	36	Entered	C20	2578,937	4,316	1	0,038
Stepwise Method: Forward Stepwise							

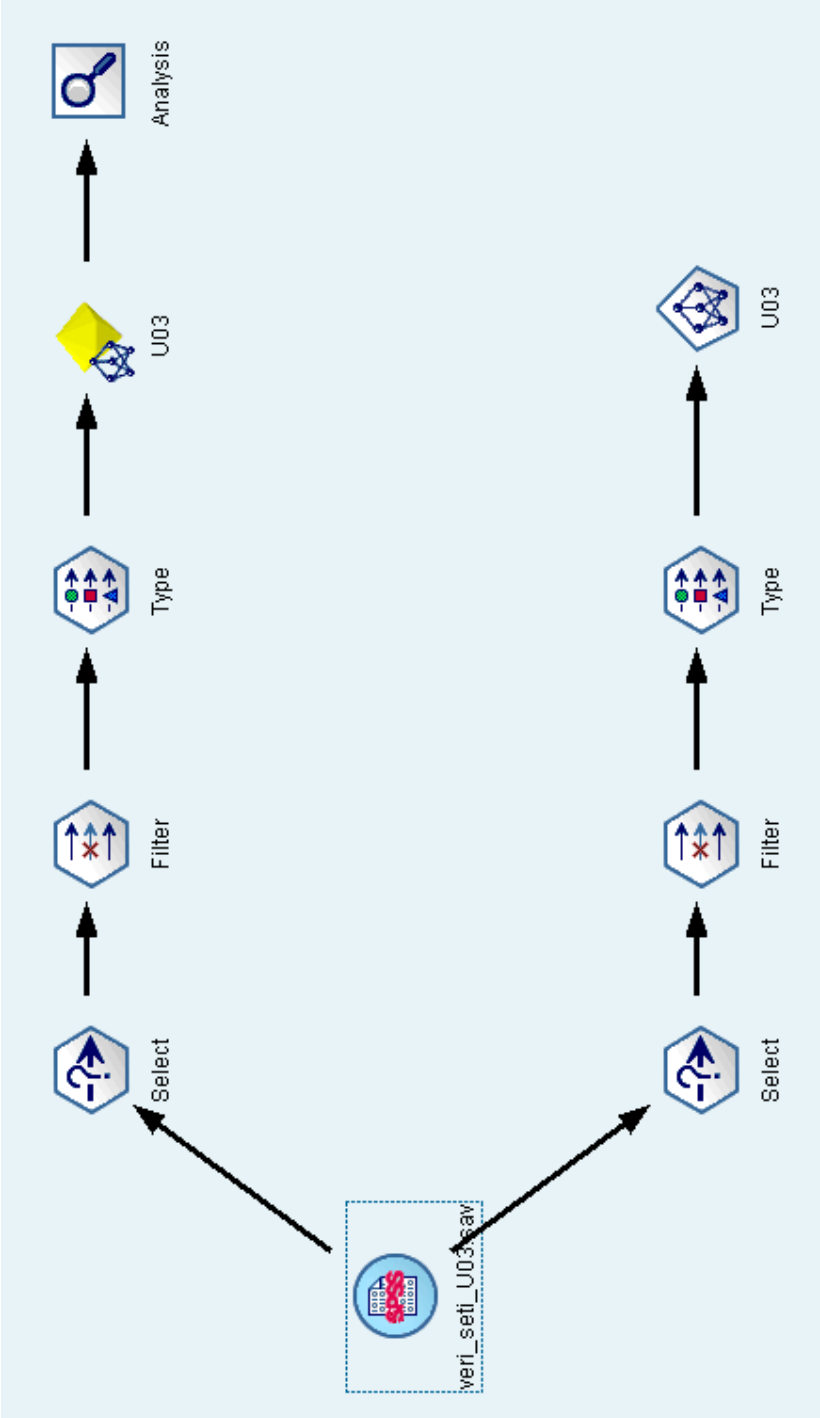
EK 9. C02 İÇİN YAPAY SİNİR AĞI VERİ AKIŞI



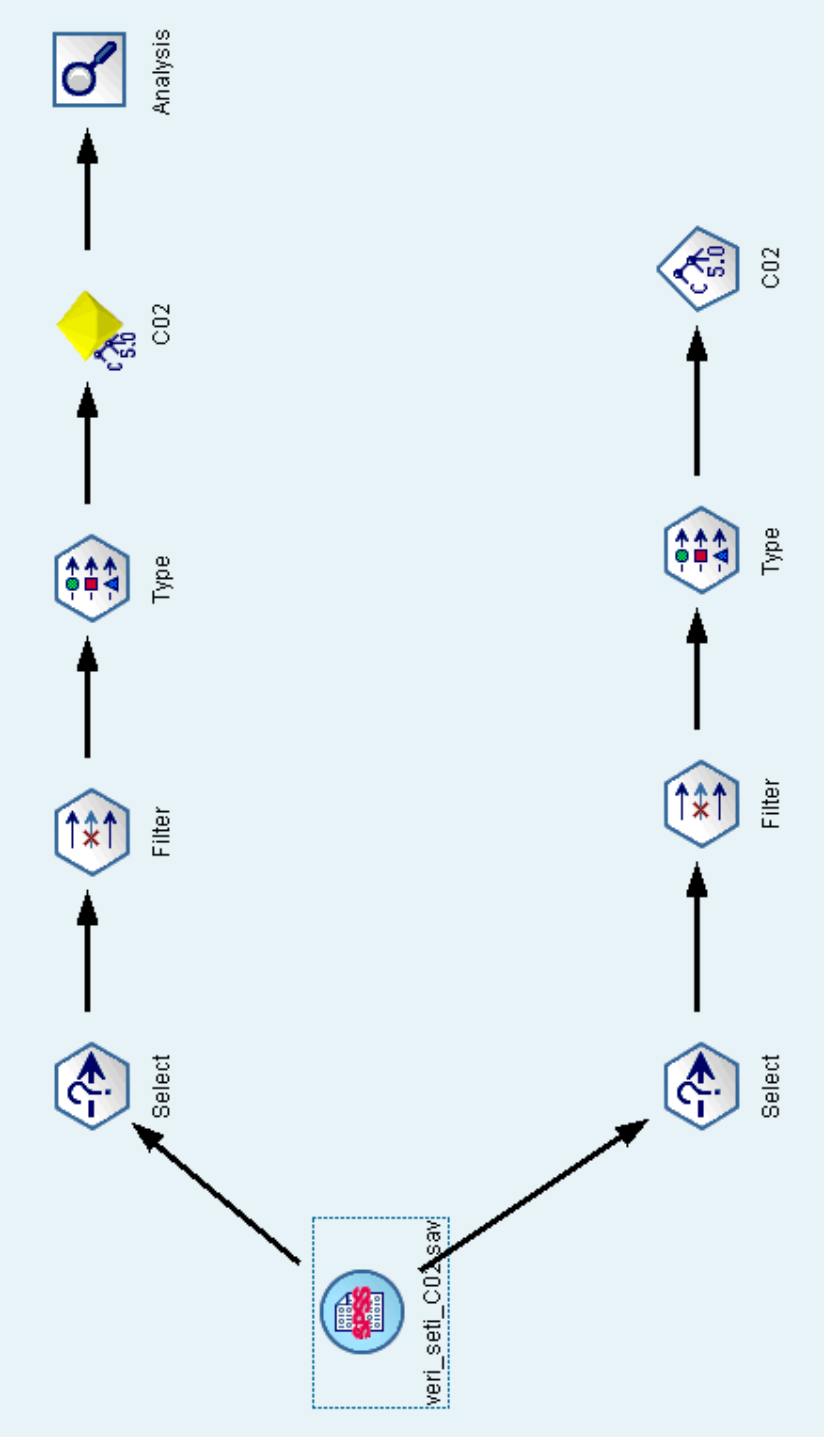
EK 10. Y01 İÇİN YAPAY SİNİR AĞI VERİ AKIŞI



EK 11. U03 İÇİN YAPAY SİNİR AĞI VERİ AKIŞI



EK 12. C02 İÇİN C5.0 ALGORİTMASI VERİ AKIŞI



EK 13. C02 İÇİN C5.0 ALGORİTMASI KARAR KURALLARI

Rule 1 - estimated accuracy 86,74% [boost 84%]
C06 = Yes [Mode: Yes]
M03 = G [Mode: No]
C24 <= -306.23999 [Mode: Yes] => Yes
C24 > -306.23999 [Mode: No] => No
M03 = A [Mode: Yes]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: Yes]
M04 = Elmas [Mode: Yes]
D04 = Graduate [Mode: Yes] => Yes
D04 = High [Mode: Yes] => Yes
D04 = Primary [Mode: Yes] => Yes
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
M04 = Inci [Mode: Yes]
C12 = Yes [Mode: Yes]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: Yes]
D03 = Married [Mode: Yes]
D04 = Graduate [Mode: Yes]
Y06 = Yes [Mode: No] => No
Y06 = No [Mode: Yes] => Yes
D04 = High [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: Yes]
U06 = Yes [Mode: Yes] => Yes
U06 = No [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
D03 = Single [Mode: Yes] => Yes
D03 = Widowed [Mode: Yes] => Yes
C12 = No [Mode: No] => No
M04 = Mercan [Mode: No]
H10 = Yes [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
H10 = No [Mode: No]
U09 = Yes [Mode: No] => No
U09 = No [Mode: No]
H09 = Yes [Mode: Yes] => Yes
H09 = No [Mode: No]
M02 = 10 [Mode: No] => No
M02 = 11 [Mode: No] => No
M02 = 12 [Mode: No] => No
M02 = 13 [Mode: No] => No

M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes] => Yes
M02 = 6 [Mode: No] => No
M02 = 7 [Mode: No] => No
M02 = 8 [Mode: No]
D03 = Married [Mode: No] => No
D03 = Single [Mode: Yes] => Yes
D03 = Widowed [Mode: No] => No
M02 = 9 [Mode: No] => No
M02 = >= 15 [Mode: No] => No
M04 = Opal [Mode: No]
C24 <= -47.310001 [Mode: No] => No
C24 > -47.310001 [Mode: Yes] => Yes
M04 = Safir [Mode: Yes] => Yes
M04 = Turkuaz [Mode: No]
M02 = 10 [Mode: No] => No
M02 = 11 [Mode: No] => No
M02 = 12 [Mode: No] => No
M02 = 13 [Mode: No] => No
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
D04 in ["Graduate" "Upper Graduate"] [Mode: Yes] => Yes
D04 in ["High"] [Mode: Yes] => Yes
D04 in ["Primary"] [Mode: No] => No
D04 in ["Secondary"] [Mode: No] => No
M02 = 6 [Mode: No] => No
M02 = 7 [Mode: No] => No
M02 = 8 [Mode: No] => No
M02 = 9 [Mode: No] => No
M02 = >= 15 [Mode: No] => No
M04 = Yakut [Mode: Yes]
M02 = 10 [Mode: No]
D03 = Married [Mode: No] => No
D03 = Single [Mode: Yes] => Yes
D03 = Widowed [Mode: No] => No
M02 = 11 [Mode: No] => No
M02 = 12 [Mode: Yes]
U04 = Yes [Mode: Yes] => Yes
U04 = No [Mode: No] => No
M02 = 13 [Mode: Yes] => Yes
M02 = 14 [Mode: Yes] => Yes
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes] => Yes
M02 = 6 [Mode: Yes] => Yes
M02 = 7 [Mode: Yes]
U10 = Yes [Mode: Yes] => Yes

U10 = No [Mode: Yes]
H16 = 1 [Mode: No] => No
H16 = 0 [Mode: Yes]
C01 = Yes [Mode: No] => No
C01 = No [Mode: Yes] => Yes
M02 = 8 [Mode: Yes] => Yes
M02 = 9 [Mode: Yes]
H07 = Yes [Mode: No]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: No] => No
H07 = No [Mode: Yes] => Yes
M02 = >= 15 [Mode: No] => No
M04 = Zumrut [Mode: Yes]
C24 <= -483.75 [Mode: Yes] => Yes
C24 > -483.75 [Mode: Yes]
M02 = 10 [Mode: No]
C24 <= -162.00999 [Mode: Yes] => Yes
C24 > -162.00999 [Mode: No] => No
M02 = 11 [Mode: No]
U08 = Yes [Mode: Yes] => Yes
U08 = No [Mode: No] => No
M02 = 12 [Mode: No] => No
M02 = 13 [Mode: No]
D04 = Graduate [Mode: Yes]
C24 <= -62.310001 [Mode: Yes] => Yes
C24 > -62.310001 [Mode: No] => No
D04 = High [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
M02 = 14 [Mode: Yes] => Yes
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes] => Yes
M02 = 6 [Mode: Yes]
U09 = Yes [Mode: No] => No
U09 = No [Mode: Yes] => Yes
M02 = 7 [Mode: Yes]
D04 = Graduate [Mode: Yes]
Y16 <= 295.47 [Mode: Yes] => Yes
Y16 > 295.47 [Mode: No] => No
D04 = High [Mode: No] => No
D04 = Primary [Mode: Yes] => Yes
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
M02 = 8 [Mode: Yes]
U04 = Yes [Mode: No] => No
U04 = No [Mode: Yes] => Yes
M02 = 9 [Mode: Yes]

H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: Yes]
D04 = Graduate [Mode: Yes] => Yes
D04 = High [Mode: No] => No
D04 = Primary [Mode: Yes] => Yes
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
M02 = >= 15 [Mode: Yes] => Yes
C06 = No [Mode: No]
C03 = Yes [Mode: Yes]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No] => No
C03 = No [Mode: No]
Y03 = Yes [Mode: Yes] => Yes
Y03 = No [Mode: No]
C01 = Yes [Mode: No]
M03 = G [Mode: No] => No
M03 = A [Mode: No]
M02 = 10 [Mode: No] => No
M02 = 11 [Mode: No] => No
M02 = 12 [Mode: No] => No
M02 = 13 [Mode: No] => No
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
D04 = Graduate [Mode: Yes] => Yes
D04 = High [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: Yes] => Yes
D04 = Upper Graduate [Mode: Yes] => Yes
M02 = 6 [Mode: Yes]
U06 = Yes [Mode: No] => No
U06 = No [Mode: Yes] => Yes
M02 = 7 [Mode: No] => No
M02 = 8 [Mode: No] => No
M02 = 9 [Mode: No] => No
M02 = >= 15 [Mode: No] => No
C01 = No [Mode: No]
Y16 <= 497.63 [Mode: No]
M04 = Elmas [Mode: No]
U08 = Yes [Mode: Yes] => Yes
U08 = No [Mode: No] => No
M04 = Inci [Mode: No] => No
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: No] => No
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: No]

Y07 = Yes [Mode: Yes] => Yes
Y07 = No [Mode: No] => No
M04 = Zumrut [Mode: No]
U09 = Yes [Mode: No] => No
U09 = No [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
Y16 > 497.63 [Mode: No]
D04 = Graduate [Mode: Yes] => Yes
D04 = High [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
Rule 2 - estimated accuracy 80,38% [boost 84%]
C06 = Yes [Mode: Yes]
M03 = G [Mode: No] => No
M03 = A [Mode: Yes]
C24 <= -460.85001 [Mode: Yes] => Yes
C24 > -460.85001 [Mode: Yes]
H10 = Yes [Mode: Yes]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: Yes]
D04 = Graduate [Mode: Yes] => Yes
D04 = High [Mode: Yes]
U04 = Yes [Mode: Yes] => Yes
U04 = No [Mode: No]
H21 <= 2011.01 [Mode: Yes] => Yes
H21 > 2011.01 [Mode: No]
H21 <= 39099.551 [Mode: No] => No
H21 > 39099.551 [Mode: Yes] => Yes
D04 = Primary [Mode: Yes] => Yes
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
H10 = No [Mode: Yes]
H09 = Yes [Mode: No]
Y06 = Yes [Mode: Yes] => Yes
Y06 = No [Mode: No] => No
H09 = No [Mode: Yes]
M04 = Elmas [Mode: Yes] => Yes
M04 = Inci [Mode: Yes]
Y07 = Yes [Mode: No] => No
Y07 = No [Mode: Yes] => Yes
M04 = Mercan [Mode: No]
U09 = Yes [Mode: No] => No
U09 = No [Mode: No]
U06 = Yes [Mode: No] => No
U06 = No [Mode: No]
H07 = Yes [Mode: No] => No

H07 = No [Mode: Yes]
Y07 = Yes [Mode: Yes] => Yes
Y07 = No [Mode: No]
D04 = Graduate [Mode: No] => No
D04 = High [Mode: Yes] => Yes
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: Yes] => Yes
D04 = Upper Graduate [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: Yes] => Yes
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: Yes]
M02 = 10 [Mode: No] => No
M02 = 11 [Mode: No] => No
M02 = 12 [Mode: Yes] => Yes
M02 = 13 [Mode: No] => No
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes] => Yes
M02 = 6 [Mode: No]
Y16 <= 1352.96 [Mode: No]
FL = >= 1 [Mode: No] => No
FL = 0 [Mode: No]
U06 = Yes [Mode: No] => No
U06 = No [Mode: Yes] => Yes
Y16 > 1352.96 [Mode: Yes] => Yes
M02 = 7 [Mode: Yes] => Yes
M02 = 8 [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
M02 = 9 [Mode: Yes] => Yes
M02 = >= 15 [Mode: No] => No
M04 = Zumrut [Mode: Yes]
FL = >= 1 [Mode: Yes] => Yes
FL = 0 [Mode: Yes]
U10 = Yes [Mode: No] => No
U10 = No [Mode: Yes]
U17 = 0 [Mode: Yes]
U09 = Yes [Mode: No] => No
U09 = No [Mode: Yes]
Y16 <= 2581.51 [Mode: Yes] => Yes
Y16 > 2581.51 [Mode: No] => No
U17 = 1 [Mode: Yes] => Yes
U17 = >= 2 [Mode: No] => No
C06 = No [Mode: No]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]
M04 = Elmas [Mode: Yes]

U06 = Yes [Mode: No] => No
U06 = No [Mode: Yes] => Yes
M04 = Inci [Mode: No] => No
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: No] => No
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: Yes]
H01 = Yes [Mode: Yes]
H09 = Yes [Mode: No] => No
H09 = No [Mode: Yes] => Yes
H01 = No [Mode: No] => No
M04 = Zumrut [Mode: No] => No
Rule 3 - estimated accuracy 78,42% [boost 84%]
M03 = G [Mode: No] => No
M03 = A [Mode: No]
C06 = Yes [Mode: Yes]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: Yes]
M02 = 10 [Mode: No]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: No]
M04 = Elmas [Mode: Yes] => Yes
M04 = Inci [Mode: No] => No
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: Yes] => Yes
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: No] => No
M04 = Zumrut [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
M02 = 11 [Mode: No]
H09 = Yes [Mode: Yes] => Yes
H09 = No [Mode: No]
FL = >= 1 [Mode: Yes] => Yes
FL = 0 [Mode: No] => No
M02 = 12 [Mode: No]
H09 = Yes [Mode: No] => No
H09 = No [Mode: No]
H10 = Yes [Mode: Yes] => Yes
H10 = No [Mode: No]
C01 = Yes [Mode: Yes]
U09 = Yes [Mode: No] => No
U09 = No [Mode: Yes] => Yes
C01 = No [Mode: No]
C12 = Yes [Mode: No] => No
C12 = No [Mode: Yes] => Yes

M02 = 13 [Mode: No]
H09 = Yes [Mode: No] => No
H09 = No [Mode: No]
U09 = Yes [Mode: Yes] => Yes
U09 = No [Mode: No]
H16 = 1 [Mode: No] => No
H16 = 0 [Mode: No]
H21 <= 1181.8199 [Mode: No]
C12 = Yes [Mode: No]
Y07 = Yes [Mode: Yes] => Yes
Y07 = No [Mode: No]
C24 <= -8.46 [Mode: No] => No
C24 > -8.46 [Mode: Yes] => Yes
C12 = No [Mode: Yes] => Yes
H21 > 1181.8199 [Mode: No] => No
M02 = 14 [Mode: No]
Y16 <= 279.92001 [Mode: No] => No
Y16 > 279.92001 [Mode: Yes] => Yes
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
L03 = Yes [Mode: No] => No
L03 = No [Mode: Yes]
C07 = Yes [Mode: Yes]
Y16 <= 8617.0498 [Mode: Yes]
C12 = Yes [Mode: Yes] => Yes
C12 = No [Mode: No] => No
Y16 > 8617.0498 [Mode: No] => No
C07 = No [Mode: Yes] => Yes
M02 = 6 [Mode: Yes] => Yes
M02 = 7 [Mode: Yes]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: Yes]
Y07 = Yes [Mode: Yes] => Yes
Y07 = No [Mode: Yes]
M04 = Elmas [Mode: Yes] => Yes
M04 = Inci [Mode: Yes] => Yes
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: Yes] => Yes
M04 = Safir [Mode: Yes] => Yes
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: No] => No
M04 = Zumrut [Mode: Yes] => Yes
M02 = 8 [Mode: Yes]
FL = >= 1 [Mode: Yes] => Yes
FL = 0 [Mode: No]
H09 = Yes [Mode: Yes] => Yes
H09 = No [Mode: No]
H16 = 1 [Mode: Yes] => Yes

H16 = 0 [Mode: No]
Y06 = Yes [Mode: No] => No
Y06 = No [Mode: No]
Y07 = Yes [Mode: Yes] => Yes
Y07 = No [Mode: No]
C07 = Yes [Mode: No] => No
C07 = No [Mode: No]
U08 = Yes [Mode: Yes] => Yes
U08 = No [Mode: No] => No
M02 = 9 [Mode: No]
U10 = Yes [Mode: Yes] => Yes
U10 = No [Mode: No]
H21 <= 6977.8398 [Mode: No] => No
H21 > 6977.8398 [Mode: Yes] => Yes
M02 = >= 15 [Mode: No] => No
C06 = No [Mode: No]
M04 = Elmas [Mode: No]
H09 = Yes [Mode: No] => No
H09 = No [Mode: Yes] => Yes
M04 = Inci [Mode: No] => No
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: Yes]
U04 = Yes [Mode: No] => No
U04 = No [Mode: Yes] => Yes
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: No]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: No]
D03 = Married [Mode: No]
U17 = 0 [Mode: No]
D04 = Graduate [Mode: No]
Y16 <= 155.50999 [Mode: No]
C07 = Yes [Mode: No] => No
C07 = No [Mode: No]
H21 <= 50 [Mode: No] => No
H21 > 50 [Mode: Yes] => Yes
Y16 > 155.50999 [Mode: Yes] => Yes
D04 = High [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
U17 = 1 [Mode: No] => No
U17 = >= 2 [Mode: Yes] => Yes
D03 = Single [Mode: No] => No
D03 = Widowed [Mode: Yes] => Yes
M04 = Zumrut [Mode: No]
D04 = Graduate [Mode: Yes]

H09 = Yes [Mode: No] => No
H09 = No [Mode: Yes]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: No]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: No] => No
D04 = High [Mode: No]
Y07 = Yes [Mode: No] => No
Y07 = No [Mode: Yes] => Yes
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
Rule 4 - estimated accuracy 77,03% [boost 84%]
M03 = G [Mode: No] => No
M03 = A [Mode: No]
C06 = Yes [Mode: Yes]
U05 = Yes [Mode: Yes]
H10 = Yes [Mode: Yes] => Yes
H10 = No [Mode: Yes]
H09 = Yes [Mode: No] => No
H09 = No [Mode: Yes]
C07 = Yes [Mode: No]
Y16 <= 9159.7197 [Mode: Yes]
U04 = Yes [Mode: Yes] => Yes
U04 = No [Mode: No] => No
Y16 > 9159.7197 [Mode: No] => No
C07 = No [Mode: Yes]
U09 = Yes [Mode: Yes]
U08 = Yes [Mode: Yes] => Yes
U08 = No [Mode: No] => No
U09 = No [Mode: Yes] => Yes
U05 = No [Mode: Yes]
M02 = 10 [Mode: No]
D04 = Graduate [Mode: Yes]
C24 <= -1917.71 [Mode: No] => No
C24 > -1917.71 [Mode: Yes] => Yes
D04 = High [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
M02 = 11 [Mode: No]
D04 = Graduate [Mode: Yes]
U06 = Yes [Mode: No] => No
U06 = No [Mode: Yes] => Yes

D04 = High [Mode: Yes] => Yes
D04 = Primary [Mode: Yes] => Yes
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
M02 = 12 [Mode: No]
Y16 <= 1057.48 [Mode: No]
H09 = Yes [Mode: No] => No
H09 = No [Mode: No]
C03 = Yes [Mode: No] => No
C03 = No [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
Y16 > 1057.48 [Mode: Yes] => Yes
M02 = 13 [Mode: No]
U09 = Yes [Mode: Yes] => Yes
U09 = No [Mode: No]
C12 = Yes [Mode: No] => No
C12 = No [Mode: Yes] => Yes
M02 = 14 [Mode: No]
C12 = Yes [Mode: No]
H09 = Yes [Mode: Yes] => Yes
H09 = No [Mode: No]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: No] => No
C12 = No [Mode: Yes] => Yes
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
Y06 = Yes [Mode: Yes] => Yes
Y06 = No [Mode: Yes]
Y07 = Yes [Mode: Yes] => Yes
Y07 = No [Mode: Yes]
C24 <= -342.17999 [Mode: Yes] => Yes
C24 > -342.17999 [Mode: No]
M04 = Elmas [Mode: Yes] => Yes
M04 = Inci [Mode: No] => No
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: Yes] => Yes
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: Yes]
U06 = Yes [Mode: Yes] => Yes
U06 = No [Mode: No] => No
M04 = Zumrut [Mode: No] => No
M02 = 6 [Mode: Yes]
C24 <= -746.41998 [Mode: Yes] => Yes
C24 > -746.41998 [Mode: Yes]
D04 = Graduate [Mode: Yes]
M04 in ["Elmas"] [Mode: Yes] => Yes

M04 in ["Inci"] [Mode: Yes] => Yes
M04 in ["Mercan"] [Mode: Yes] => Yes
M04 in ["Opal" "Turkuaz"] [Mode: No] => No
M04 in ["Safir"] [Mode: Yes] => Yes
M04 in ["Yakut"] [Mode: No] => No
M04 in ["Zumrut"] [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
D04 = High [Mode: No]
C01 = Yes [Mode: No] => No
C01 = No [Mode: Yes] => Yes
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
M02 = 7 [Mode: Yes]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: Yes]
H09 = Yes [Mode: No] => No
H09 = No [Mode: Yes] => Yes
M02 = 8 [Mode: Yes]
D04 = Graduate [Mode: Yes]
H21 <= 1559.14 [Mode: Yes] => Yes
H21 > 1559.14 [Mode: No] => No
D04 = High [Mode: No]
Y06 = Yes [Mode: No] => No
Y06 = No [Mode: No]
U06 = Yes [Mode: No] => No
U06 = No [Mode: Yes] => Yes
D04 = Primary [Mode: Yes] => Yes
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
M02 = 9 [Mode: Yes]
M04 = Elmas [Mode: Yes] => Yes
M04 = Inci [Mode: Yes] => Yes
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: Yes] => Yes
M04 = Safir [Mode: No] => No
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: Yes] => Yes
M04 = Zumrut [Mode: Yes]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: Yes]
C24 <= -497.57001 [Mode: Yes] => Yes
C24 > -497.57001 [Mode: No] => No
M02 = >= 15 [Mode: Yes] => Yes
C06 = No [Mode: No]
L03 = Yes [Mode: Yes] => Yes
L03 = No [Mode: No]

M04 = Elmas [Mode: No] => No
M04 = Inci [Mode: No] => No
M04 = Mercan [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: Yes]
Y16 <= 1111.4301 [Mode: Yes] => Yes
Y16 > 1111.4301 [Mode: No] => No
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: Yes]
D03 = Married [Mode: No]
D04 = Graduate [Mode: Yes]
Y16 <= 155.50999 [Mode: Yes]
C01 = Yes [Mode: No] => No
C01 = No [Mode: Yes]
H21 <= 50 [Mode: Yes]
U08 = Yes [Mode: No] => No
U08 = No [Mode: Yes] => Yes
H21 > 50 [Mode: Yes] => Yes
Y16 > 155.50999 [Mode: Yes] => Yes
D04 = High [Mode: No]
H21 <= 2411.1899 [Mode: Yes] => Yes
H21 > 2411.1899 [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
D03 = Single [Mode: Yes] => Yes
D03 = Widowed [Mode: Yes] => Yes
M04 = Zumrut [Mode: No]
D04 = Graduate [Mode: No]
U09 = Yes [Mode: No] => No
U09 = No [Mode: Yes]
U06 = Yes [Mode: No]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: No] => No
U06 = No [Mode: Yes] => Yes
D04 = High [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
Rule 5 - estimated accuracy 78,5% [boost 84%]
H01 = Yes [Mode: No]
M03 = G [Mode: No]
C24 <= -37.200001 [Mode: Yes] => Yes
C24 > -37.200001 [Mode: No] => No
M03 = A [Mode: No]
M04 = Elmas [Mode: Yes]
FL = >= 1 [Mode: Yes] => Yes
FL = 0 [Mode: Yes]

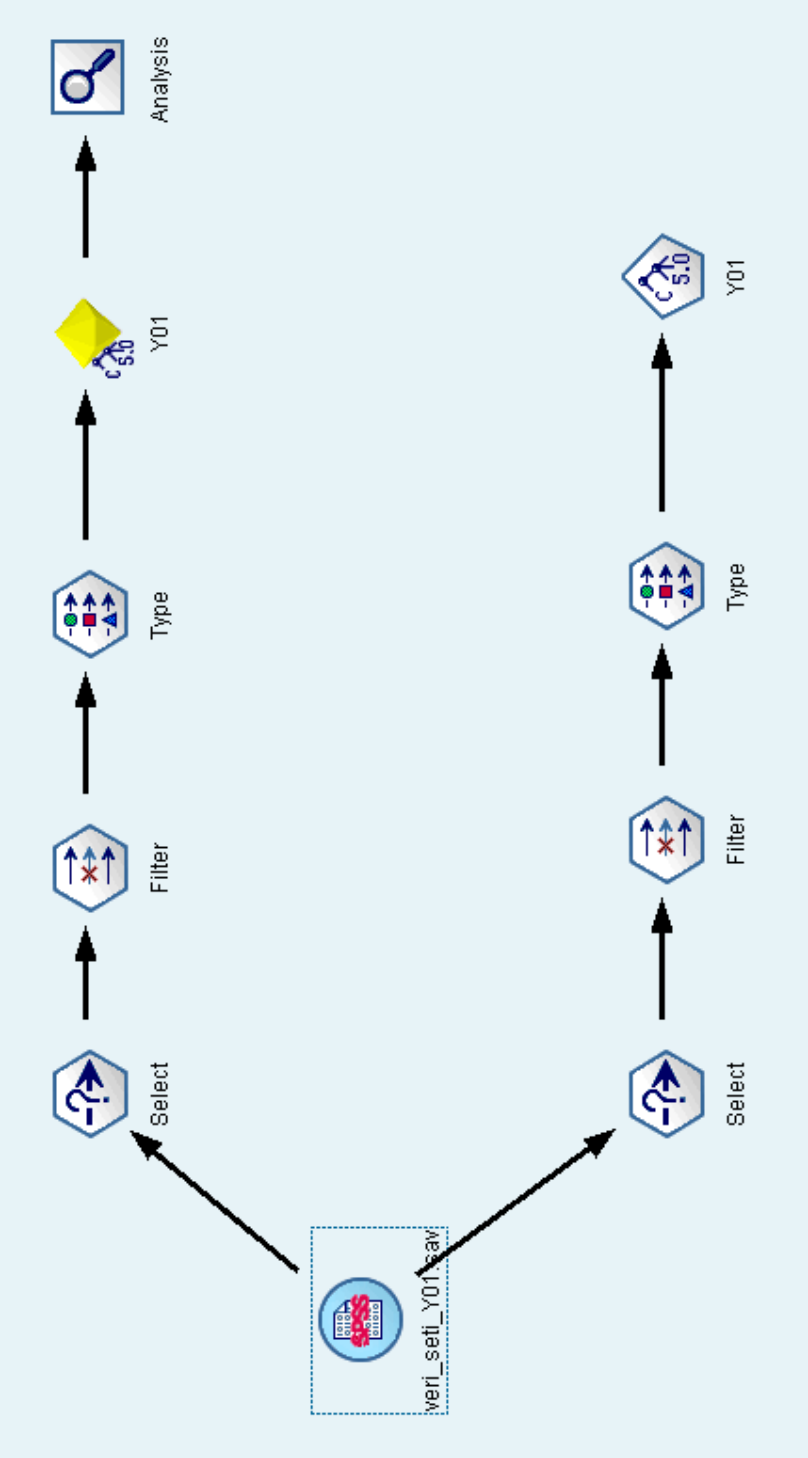
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: Yes]
D04 = Graduate [Mode: Yes]
D03 = Married [Mode: Yes]
Y07 = Yes [Mode: No] => No
Y07 = No [Mode: Yes]
C24 <= -388.45001 [Mode: No] => No
C24 > -388.45001 [Mode: Yes] => Yes
D03 = Single [Mode: Yes] => Yes
D03 = Widowed [Mode: Yes] => Yes
D04 = High [Mode: Yes] => Yes
D04 = Primary [Mode: Yes] => Yes
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
M04 = Inci [Mode: No]
U17 = 0 [Mode: No]
U08 = Yes [Mode: Yes]
H09 = Yes [Mode: Yes] => Yes
H09 = No [Mode: Yes]
C12 = Yes [Mode: Yes] => Yes
C12 = No [Mode: No] => No
U08 = No [Mode: No]
H09 = Yes [Mode: No] => No
H09 = No [Mode: No]
D04 = Graduate [Mode: No]
C12 = Yes [Mode: Yes] => Yes
C12 = No [Mode: No] => No
D04 = High [Mode: No]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: No] => No
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: Yes] => Yes
U17 = 1 [Mode: No] => No
U17 = >= 2 [Mode: Yes] => Yes
M04 = Mercan [Mode: No]
U10 = Yes [Mode: No] => No
U10 = No [Mode: No]
H16 = 1 [Mode: No] => No
H16 = 0 [Mode: No]
C07 = Yes [Mode: Yes]
H21 <= 8.7700005 [Mode: No]
H07 = Yes [Mode: No] => No
H07 = No [Mode: Yes]
C24 <= -549.39001 [Mode: Yes] => Yes
C24 > -549.39001 [Mode: No]
U04 = Yes [Mode: Yes] => Yes
U04 = No [Mode: No]

U06 = Yes [Mode: No] => No
U06 = No [Mode: Yes] => Yes
H21 > 8.7700005 [Mode: Yes] => Yes
C07 = No [Mode: No]
Y03 = Yes [Mode: Yes] => Yes
Y03 = No [Mode: No]
U09 = Yes [Mode: No] => No
U09 = No [Mode: No]
C03 = Yes [Mode: No] => No
C03 = No [Mode: No]
H10 = Yes [Mode: No] => No
H10 = No [Mode: No]
Y07 = Yes [Mode: No] => No
Y07 = No [Mode: No]
C06 = Yes [Mode: No]
U08 = Yes [Mode: Yes] => Yes
U08 = No [Mode: No]
U06 = Yes [Mode: No] => No
U06 = No [Mode: No]
D04 = Graduate [Mode: Yes]
D03 = Married [Mode: Yes] => Yes
D03 = Single [Mode: No] => No
D03 = Widowed [Mode: No] => No
D04 = High [Mode: Yes] => Yes
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: No] => No
D04 = Upper Graduate [Mode: No] => No
C06 = No [Mode: No] => No
M04 = Opal [Mode: No] => No
M04 = Safir [Mode: Yes]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: Yes]
C07 = Yes [Mode: No] => No
C07 = No [Mode: Yes] => Yes
M04 = Turkuaz [Mode: No] => No
M04 = Yakut [Mode: Yes]
L02 = Yes [Mode: Yes] => Yes
L02 = No [Mode: Yes]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: Yes]
FL = >= 1 [Mode: No] => No
FL = 0 [Mode: Yes]
D03 = Married [Mode: Yes]
M02 = 10 [Mode: No] => No
M02 = 11 [Mode: No]
U08 = Yes [Mode: No] => No
U08 = No [Mode: Yes] => Yes
M02 = 12 [Mode: Yes] => Yes

M02 = 13 [Mode: Yes] => Yes
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes] => Yes
M02 = 6 [Mode: Yes]
Y16 <= 1306.3101 [Mode: No]
U04 = Yes [Mode: No] => No
U04 = No [Mode: Yes]
U08 = Yes [Mode: Yes] => Yes
U08 = No [Mode: No]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: No] => No
Y16 > 1306.3101 [Mode: Yes] => Yes
M02 = 7 [Mode: No]
U10 = Yes [Mode: Yes] => Yes
U10 = No [Mode: No]
H16 = 1 [Mode: No] => No
H16 = 0 [Mode: No]
U06 = Yes [Mode: Yes] => Yes
U06 = No [Mode: No] => No
M02 = 8 [Mode: Yes]
U06 = Yes [Mode: No]
H21 <= 6129.6499 [Mode: Yes]
U10 = Yes [Mode: No] => No
U10 = No [Mode: Yes] => Yes
H21 > 6129.6499 [Mode: No] => No
U06 = No [Mode: Yes] => Yes
M02 = 9 [Mode: No]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No] => No
M02 = >= 15 [Mode: No] => No
D03 = Single [Mode: Yes]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: Yes]
C12 = Yes [Mode: Yes]
H07 = Yes [Mode: No] => No
H07 = No [Mode: Yes]
U09 = Yes [Mode: No] => No
U09 = No [Mode: Yes] => Yes
C12 = No [Mode: No]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: No] => No
D03 = Widowed [Mode: Yes] => Yes
M04 = Zumrut [Mode: Yes]
C24 <= -638.23999 [Mode: Yes] => Yes
C24 > -638.23999 [Mode: Yes]
M02 = 10 [Mode: No]
C06 = Yes [Mode: No]

C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No] => No
C06 = No [Mode: No] => No
M02 = 11 [Mode: No]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: No] => No
M02 = 12 [Mode: No]
C12 = Yes [Mode: Yes] => Yes
C12 = No [Mode: No] => No
M02 = 13 [Mode: No] => No
M02 = 14 [Mode: Yes] => Yes
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
U10 = Yes [Mode: Yes] => Yes
U10 = No [Mode: Yes]
C06 = Yes [Mode: Yes]
C01 = Yes [Mode: No] => No
C01 = No [Mode: Yes] => Yes
C06 = No [Mode: Yes] => Yes
M02 = 6 [Mode: Yes]
D03 = Married [Mode: Yes]
H16 = 1 [Mode: Yes] => Yes
H16 = 0 [Mode: Yes]
D04 = Graduate [Mode: No]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: No] => No
D04 = High [Mode: Yes] => Yes
D04 = Primary [Mode: No] => No
D04 = Secondary [Mode: Yes] => Yes
D04 = Upper Graduate [Mode: Yes] => Yes
D03 = Single [Mode: No] => No
D03 = Widowed [Mode: No] => No
M02 = 7 [Mode: Yes]
H21 <= 9952.8496 [Mode: Yes]
U09 = Yes [Mode: No] => No
U09 = No [Mode: Yes] => Yes
H21 > 9952.8496 [Mode: No] => No
M02 = 8 [Mode: Yes]
U05 = Yes [Mode: Yes] => Yes
U05 = No [Mode: Yes]
U09 = Yes [Mode: No] => No
U09 = No [Mode: Yes] => Yes
M02 = 9 [Mode: Yes]
C06 = Yes [Mode: Yes] => Yes
C06 = No [Mode: No] => No
M02 = >= 15 [Mode: Yes] => Yes
H01 = No [Mode: No]
C01 = Yes [Mode: Yes] => Yes

EK 14. Y01 İÇİN C5.0 ALGORİTMASI VERİ AKIŞI

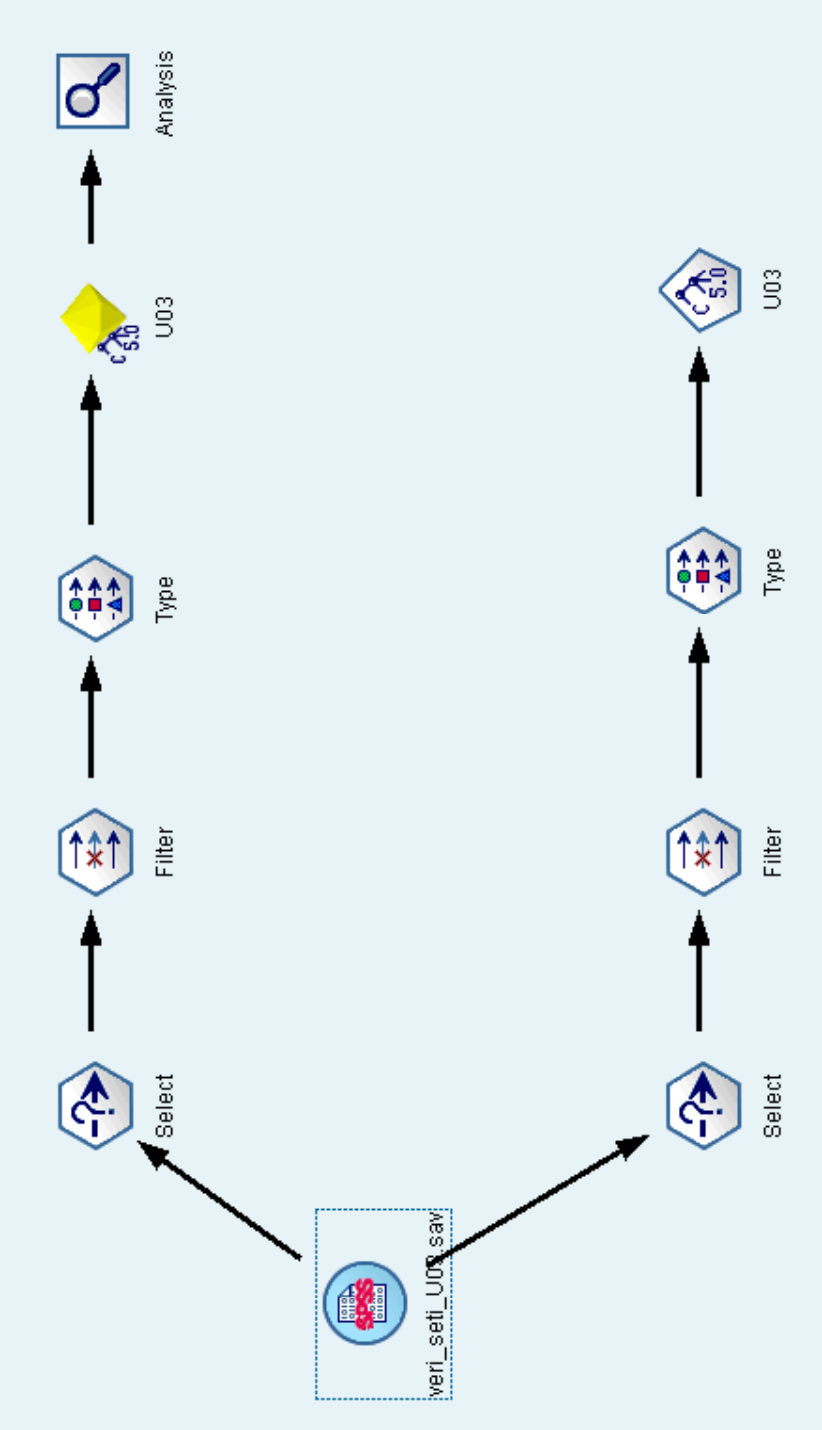


EK 15. Y01 İÇİN C5.0 ALGORİTMASI KARAR KURALLARI

Rule 1 - estimated accuracy 83,03% [boost 81,8%]
H06 = Yes [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No]
U03 = Yes [Mode: Yes] => Yes
U03 = No [Mode: No] => No
H06 = No [Mode: No]
U03 = Yes [Mode: Yes]
U01 = Yes [Mode: Yes]
H22 <= -15.89 [Mode: No] => No
H22 > -15.89 [Mode: Yes] => Yes
U01 = No [Mode: No] => No
U03 = No [Mode: No]
Y04 = Yes [Mode: Yes] => Yes
Y04 = No [Mode: No]
Y12 = >= 1 [Mode: No]
C08 = Yes [Mode: Yes] => Yes
C08 = No [Mode: No]
H05 = Yes [Mode: No] => No
H05 = No [Mode: No]
H10 = Yes [Mode: Yes] => Yes
H10 = No [Mode: No] => No
Y12 = 0 [Mode: No] => No
Rule 2 - estimated accuracy 75,7% [boost 81,8%]
H01 = Yes [Mode: No]
H06 = Yes [Mode: Yes] => Yes
H06 = No [Mode: No]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: No]
Y04 = Yes [Mode: Yes] => Yes
Y04 = No [Mode: No]
Y12 = >= 1 [Mode: Yes] => Yes
Y12 = 0 [Mode: No] => No
H01 = No [Mode: No] => No
Rule 3 - estimated accuracy 73,17% [boost 81,8%]
H06 = Yes [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
H06 = No [Mode: No]
U01 = Yes [Mode: No]
H10 = Yes [Mode: Yes] => Yes
H10 = No [Mode: No]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]
Y04 = Yes [Mode: Yes] => Yes
Y04 = No [Mode: No] => No

U01 = No [Mode: No] => No
Rule 4 - estimated accuracy 71,03% [boost 81,8%]
U01 = Yes [Mode: Yes]
H06 = Yes [Mode: Yes] => Yes
H06 = No [Mode: No]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: No]
H22 <= -59.16 [Mode: No] => No
H22 > -59.16 [Mode: No]
U03 = Yes [Mode: Yes] => Yes
U03 = No [Mode: No] => No
U01 = No [Mode: No] => No
Rule 5 - estimated accuracy 69,29% [boost 81,8%]
U01 = Yes [Mode: Yes]
U23 = Yes [Mode: No] => No
U23 = No [Mode: Yes]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: Yes]
C04 = Yes [Mode: Yes] => Yes
C04 = No [Mode: Yes]
U03 = Yes [Mode: Yes] => Yes
U03 = No [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
U01 = No [Mode: No]
U21 = Yes [Mode: Yes] => Yes
U21 = No [Mode: No]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]
H06 = Yes [Mode: Yes] => Yes
H06 = No [Mode: No] => No

EK 16. U03 İÇİN C5.0 ALGORİTMASI VERİ AKIŞI



EK 17. U03 İÇİN C5.0 ALGORİTMASI KARAR KURALLARI

Rule 1 - estimated accuracy 91,27% [boost 86,7%]
U02 = Yes [Mode: Yes]
H01 = Yes [Mode: Yes]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: Yes]
Y01 = Yes [Mode: Yes] => Yes
Y01 = No [Mode: Yes]
L05 = Yes [Mode: No] => No
L05 = No [Mode: Yes]
L03 = Yes [Mode: Yes] => Yes
L03 = No [Mode: Yes]
M02 = 10 [Mode: Yes]
U16 = 0 [Mode: Yes]
TY in ["0"] [Mode: Yes]
C18 = 0 [Mode: No]
D01 = + 55 [Mode: No] => No
D01 = - 24 [Mode: No] => No
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: Yes] => Yes
D01 = 45 - 54 [Mode: No] => No
C18 = 1 [Mode: Yes] => Yes
C18 = 2 [Mode: Yes] => Yes
C18 = 3 [Mode: No] => No
C18 = >= 4 [Mode: No] => No
TY in ["1"] [Mode: Yes] => Yes
TY in ["2" ">= 3"] [Mode: Yes] => Yes
U16 = 1 [Mode: Yes] => Yes
U16 = 2 [Mode: No] => No
U16 = 3 [Mode: Yes] => Yes
U16 = 4 [Mode: Yes] => Yes
U16 = >= 5 [Mode: Yes] => Yes
M02 = 11 [Mode: Yes]
H22 <= -114.06 [Mode: Yes] => Yes
H22 > -114.06 [Mode: Yes]
U01 = Yes [Mode: Yes]
D01 = + 55 [Mode: Yes] => Yes
D01 = - 24 [Mode: Yes] => Yes
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: Yes] => Yes
D01 = 45 - 54 [Mode: No]
TY in ["0"] [Mode: No] => No
TY in ["1"] [Mode: Yes] => Yes
TY in ["2" ">= 3"] [Mode: No] => No
U01 = No [Mode: No] => No
M02 = 12 [Mode: Yes]
H05 = Yes [Mode: Yes] => Yes

H05 = No [Mode: Yes]
FC = 0 [Mode: No] => No
FC = 1 [Mode: Yes] => Yes
FC = 2 [Mode: Yes]
C01 = Yes [Mode: No] => No
C01 = No [Mode: Yes] => Yes
FC = >= 3 [Mode: Yes] => Yes
M02 = 13 [Mode: Yes]
TC = 0 [Mode: No] => No
TC = 1 [Mode: No] => No
TC = 2 [Mode: Yes] => Yes
TC = 3 [Mode: Yes] => Yes
TC = 4 [Mode: No] => No
TC = >= 5 [Mode: Yes] => Yes
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes] => Yes
M02 = 6 [Mode: Yes]
D01 = + 55 [Mode: No] => No
D01 = - 24 [Mode: Yes] => Yes
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: Yes] => Yes
D01 = 45 - 54 [Mode: Yes]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: Yes]
H15 = 0 [Mode: Yes]
U01 = Yes [Mode: Yes] => Yes
U01 = No [Mode: No] => No
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: No] => No
M02 = 7 [Mode: Yes]
U07 = Yes [Mode: Yes]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: Yes]
C08 = Yes [Mode: Yes] => Yes
C08 = No [Mode: Yes]
C18 = 0 [Mode: Yes]
TY = 0 [Mode: Yes]
D01 = + 55 [Mode: No] => No
D01 = - 24 [Mode: No] => No
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: Yes] => Yes
D01 = 45 - 54 [Mode: No] => No
TY = 1 [Mode: Yes] => Yes
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: Yes] => Yes
C18 = 1 [Mode: No] => No

C18 = 2 [Mode: Yes] => Yes
C18 = 3 [Mode: No] => No
C18 = >= 4 [Mode: No] => No
U07 = No [Mode: Yes] => Yes
M02 = 8 [Mode: Yes]
H15 = 0 [Mode: Yes]
TY in ["0"] [Mode: Yes]
TC = 0 [Mode: Yes] => Yes
TC = 1 [Mode: No] => No
TC = 2 [Mode: No]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: No] => No
TC = 3 [Mode: Yes] => Yes
TC = 4 [Mode: Yes] => Yes
TC = >= 5 [Mode: Yes] => Yes
TY in ["1"] [Mode: Yes] => Yes
TY in ["2" ">= 3"] [Mode: Yes] => Yes
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: No] => No
M02 = 9 [Mode: Yes]
C08 = Yes [Mode: Yes] => Yes
C08 = No [Mode: Yes]
TY = 0 [Mode: No]
U17 = 0 [Mode: No]
TC = 0 [Mode: No]
H19 <= 0.029999999 [Mode: No] => No
H19 > 0.029999999 [Mode: Yes] => Yes
TC = 1 [Mode: No] => No
TC = 2 [Mode: Yes] => Yes
TC = 3 [Mode: Yes] => Yes
TC = 4 [Mode: No] => No
TC = >= 5 [Mode: No] => No
U17 = 1 [Mode: Yes] => Yes
U17 = >= 2 [Mode: No] => No
TY = 1 [Mode: Yes] => Yes
TY = 2 [Mode: No] => No
TY = >= 3 [Mode: Yes] => Yes
M02 = >= 15 [Mode: Yes]
C01 = Yes [Mode: No] => No
C01 = No [Mode: Yes] => Yes
H01 = No [Mode: No]
D01 = + 55 [Mode: Yes] => Yes
D01 = - 24 [Mode: No] => No
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: No] => No
D01 = 45 - 54 [Mode: No] => No
U02 = No [Mode: No]

C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]
H22 <= -36.279999 [Mode: No]
U01 = Yes [Mode: No]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No] => No
U01 = No [Mode: No] => No
H22 > -36.279999 [Mode: No]
C08 = Yes [Mode: No]
H17 = >= 1 [Mode: Yes]
TC = 0 [Mode: Yes] => Yes
TC = 1 [Mode: No] => No
TC = 2 [Mode: Yes] => Yes
TC = 3 [Mode: Yes] => Yes
TC = 4 [Mode: No] => No
TC = >= 5 [Mode: No] => No
H17 = 0 [Mode: No] => No
C08 = No [Mode: No]
Y01 = Yes [Mode: No]
H05 = Yes [Mode: Yes]
H17 = >= 1 [Mode: No] => No
H17 = 0 [Mode: Yes] => Yes
H05 = No [Mode: No]
TY = 0 [Mode: No] => No
TY = 1 [Mode: No] => No
TY = 2 [Mode: No]
C19 <= -37.889999 [Mode: Yes] => Yes
C19 > -37.889999 [Mode: No]
H17 = >= 1 [Mode: No] => No
H17 = 0 [Mode: No]
U05 = Yes [Mode: No] => No
U05 = No [Mode: Yes]
H07 = Yes [Mode: No] => No
H07 = No [Mode: Yes]
H04 = Yes [Mode: No] => No
H04 = No [Mode: Yes] => Yes
TY = >= 3 [Mode: No] => No
Y01 = No [Mode: No] => No

Rule 2 - estimated accuracy 82,94% [boost 86,7%]

C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]
U02 = Yes [Mode: Yes]
H01 = Yes [Mode: Yes]
Y03 = Yes [Mode: Yes] => Yes
Y03 = No [Mode: Yes]
L03 = Yes [Mode: Yes] => Yes
L03 = No [Mode: Yes]
C04 = Yes [Mode: Yes] => Yes

C04 = No [Mode: Yes]
C07 = Yes [Mode: Yes]
H21 <= 2411.1899 [Mode: Yes]
D01 = + 55 [Mode: No] => No
D01 = - 24 [Mode: No] => No
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: Yes]
H21 <= 1202.52 [Mode: Yes] => Yes
H21 > 1202.52 [Mode: No] => No
D01 = 45 - 54 [Mode: Yes]
U20 <= 50.169998 [Mode: Yes] => Yes
U20 > 50.169998 [Mode: No] => No
H21 > 2411.1899 [Mode: Yes] => Yes
C07 = No [Mode: Yes]
M03 = G [Mode: Yes] => Yes
M03 = A [Mode: Yes]
C08 = Yes [Mode: Yes]
U04 = Yes [Mode: Yes] => Yes
U04 = No [Mode: Yes]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: Yes]
H17 = >= 1 [Mode: No] => No
H17 = 0 [Mode: Yes] => Yes
C08 = No [Mode: Yes]
U17 = 0 [Mode: Yes]
U16 = 0 [Mode: Yes]
FC = 0 [Mode: Yes]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: No]
H15 = 0 [Mode: Yes]
TY = 0 [Mode: No]
U07 = Yes [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
U07 = No [Mode: Yes] => Yes
TY = 1 [Mode: Yes] => Yes
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: Yes] => Yes
H15 = 1 [Mode: No] => No
H15 = 2 [Mode: Yes] => Yes
H15 = >= 3 [Mode: No] => No
FC = 1 [Mode: Yes]
H22 <= -6.4499998 [Mode: No] => No
H22 > -6.4499998 [Mode: Yes]
H15 = 0 [Mode: Yes]
H17 = >= 1 [Mode: No] => No

H17 = 0 [Mode: Yes] => Yes
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: No] => No
FC = 2 [Mode: Yes]
U01 = Yes [Mode: Yes] => Yes
U01 = No [Mode: No] => No
FC = >= 3 [Mode: Yes] => Yes
U16 = 1 [Mode: Yes]
H21 <= 22246.199 [Mode: Yes] => Yes
H21 > 22246.199 [Mode: No] => No
U16 = 2 [Mode: Yes] => Yes
U16 = 3 [Mode: Yes] => Yes
U16 = 4 [Mode: Yes] => Yes
U16 = >= 5 [Mode: Yes] => Yes
U17 = 1 [Mode: No]
H19 <= 321.70999 [Mode: No] => No
H19 > 321.70999 [Mode: Yes] => Yes
U17 = >= 2 [Mode: Yes] => Yes
H01 = No [Mode: No] => No
U02 = No [Mode: No]
C08 = Yes [Mode: No]
C01 = Yes [Mode: No] => No
C01 = No [Mode: No]
H22 <= -88 [Mode: Yes] => Yes
H22 > -88 [Mode: No]
H05 = Yes [Mode: No] => No
H05 = No [Mode: No]
H08 = Yes [Mode: Yes] => Yes
H08 = No [Mode: No]
U04 = Yes [Mode: Yes] => Yes
U04 = No [Mode: No]
H19 <= 1.14 [Mode: No]
H19 <= 0.050000001 [Mode: No] => No
H19 > 0.050000001 [Mode: Yes] => Yes
H19 > 1.14 [Mode: No] => No
C08 = No [Mode: No]
Y03 = Yes [Mode: No] => No
Y03 = No [Mode: No]
H05 = Yes [Mode: No]
H04 = Yes [Mode: No] => No
H04 = No [Mode: No]
H19 <= 4.73 [Mode: No] => No
H19 > 4.73 [Mode: Yes] => Yes
H05 = No [Mode: No] => No
Rule 3 - estimated accuracy 81,37% [boost 86,7%]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]

U02 = Yes [Mode: Yes]
L03 = Yes [Mode: Yes] => Yes
L03 = No [Mode: Yes]
M02 = 10 [Mode: No]
H21 <= 8373.3896 [Mode: No]
H15 = 0 [Mode: No]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: No]
H17 = >= 1 [Mode: No] => No
H17 = 0 [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No]
U07 = Yes [Mode: No]
C19 <= -71.709999 [Mode: Yes] => Yes
C19 > -71.709999 [Mode: No] => No
U07 = No [Mode: No] => No
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: Yes] => Yes
H15 = >= 3 [Mode: No] => No
H21 > 8373.3896 [Mode: No] => No
M02 = 11 [Mode: No]
H15 = 0 [Mode: No]
H08 = Yes [Mode: No] => No
H08 = No [Mode: No]
U05 = Yes [Mode: Yes] => Yes
U05 = No [Mode: No] => No
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: Yes] => Yes
H15 = >= 3 [Mode: No] => No
M02 = 12 [Mode: Yes]
M03 = G [Mode: No] => No
M03 = A [Mode: Yes]
H19 <= 905.31 [Mode: Yes]
U01 = Yes [Mode: Yes] => Yes
U01 = No [Mode: No] => No
H19 > 905.31 [Mode: Yes] => Yes
M02 = 13 [Mode: Yes] => Yes
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
U01 = Yes [Mode: Yes] => Yes
U01 = No [Mode: No]
U05 = Yes [Mode: Yes] => Yes
U05 = No [Mode: No]
C18 = 0 [Mode: No] => No
C18 = 1 [Mode: Yes] => Yes
C18 = 2 [Mode: No] => No
C18 = 3 [Mode: No] => No

C18 = >= 4 [Mode: No] => No
M02 = 6 [Mode: Yes]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: Yes]
H19 <= 0.02 [Mode: No]
U05 = Yes [Mode: No] => No
U05 = No [Mode: No]
H21 <= 3344.1799 [Mode: No]
U01 = Yes [Mode: Yes]
FC = 0 [Mode: Yes] => Yes
FC = 1 [Mode: Yes] => Yes
FC = 2 [Mode: No] => No
FC = >= 3 [Mode: No] => No
U01 = No [Mode: No] => No
H21 > 3344.1799 [Mode: Yes] => Yes
H19 > 0.02 [Mode: Yes] => Yes
M02 = 7 [Mode: Yes]
TY = 0 [Mode: Yes]
U07 = Yes [Mode: Yes]
U17 = 0 [Mode: Yes]
U16 = 0 [Mode: No]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No]
H04 = Yes [Mode: Yes] => Yes
H04 = No [Mode: No]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: No] => No
U16 = 1 [Mode: Yes] => Yes
U16 = 2 [Mode: Yes] => Yes
U16 = 3 [Mode: No] => No
U16 = 4 [Mode: No] => No
U16 = >= 5 [Mode: Yes] => Yes
U17 = 1 [Mode: No] => No
U17 = >= 2 [Mode: Yes] => Yes
U07 = No [Mode: Yes] => Yes
TY = 1 [Mode: Yes] => Yes
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: Yes] => Yes
M02 = 8 [Mode: Yes]
C08 = Yes [Mode: Yes] => Yes
C08 = No [Mode: No]
U17 = 0 [Mode: No]
D01 = + 55 [Mode: Yes] => Yes
D01 = - 24 [Mode: Yes] => Yes
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: No] => No
D01 = 45 - 54 [Mode: Yes] => Yes
U17 = 1 [Mode: No] => No

U17 = >= 2 [Mode: No] => No
M02 = 9 [Mode: No]
C08 = Yes [Mode: Yes] => Yes
C08 = No [Mode: No]
FC = 0 [Mode: Yes] => Yes
FC = 1 [Mode: No]
C19 <= -252.78999 [Mode: Yes] => Yes
C19 > -252.78999 [Mode: No]
U04 = Yes [Mode: Yes] => Yes
U04 = No [Mode: No] => No
FC = 2 [Mode: No] => No
FC = >= 3 [Mode: Yes] => Yes
M02 = >= 15 [Mode: Yes] => Yes
U02 = No [Mode: No]
H01 = Yes [Mode: No]
C08 = Yes [Mode: No]
U16 = 0 [Mode: No]
H21 <= 27240.67 [Mode: No]
H15 = 0 [Mode: No]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: No]
TY = 0 [Mode: No] => No
TY = 1 [Mode: No] => No
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: No] => No
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: Yes] => Yes
H15 = >= 3 [Mode: Yes] => Yes
H21 > 27240.67 [Mode: No] => No
U16 = 1 [Mode: Yes] => Yes
U16 = 2 [Mode: Yes] => Yes
U16 = 3 [Mode: No] => No
U16 = 4 [Mode: No] => No
U16 = >= 5 [Mode: Yes] => Yes
C08 = No [Mode: No]
Y03 = Yes [Mode: No] => No
Y03 = No [Mode: No]
U20 <= 28.01 [Mode: No]
H19 <= 0.31 [Mode: No]
U05 = Yes [Mode: No] => No
U05 = No [Mode: No]
M02 = 10 [Mode: No] => No
M02 = 11 [Mode: No] => No
M02 = 12 [Mode: No] => No
M02 = 13 [Mode: No] => No
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: No]
U04 = Yes [Mode: Yes] => Yes

U04 = No [Mode: No]
C18 in ["0"] [Mode: No] => No
C18 in ["1"] [Mode: No] => No
C18 in ["2"] [Mode: Yes] => Yes
C18 in ["3" ">= 4"] [Mode: Yes] => Yes
M02 = 5 [Mode: No] => No
M02 = 6 [Mode: No]
C19 <= -50.540001 [Mode: Yes] => Yes
C19 > -50.540001 [Mode: No]
H21 <= 9469.1504 [Mode: No]
H17 = >= 1 [Mode: No] => No
H17 = 0 [Mode: No]
TY = 0 [Mode: No] => No
TY = 1 [Mode: No] => No
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: No] => No
H21 > 9469.1504 [Mode: Yes] => Yes
M02 = 7 [Mode: No] => No
M02 = 8 [Mode: No] => No
M02 = 9 [Mode: No] => No
M02 = >= 15 [Mode: No] => No
H19 > 0.31 [Mode: No]
M03 = G [Mode: No] => No
M03 = A [Mode: No]
U16 = 0 [Mode: No]
U17 = 0 [Mode: No]
H19 <= 8.5900002 [Mode: No]
C01 = Yes [Mode: Yes] => Yes
C01 = No [Mode: No] => No
H19 > 8.5900002 [Mode: No] => No
U17 = 1 [Mode: No] => No
U17 = >= 2 [Mode: Yes] => Yes
U16 = 1 [Mode: Yes] => Yes
U16 = 2 [Mode: No] => No
U16 = 3 [Mode: Yes] => Yes
U16 = 4 [Mode: Yes] => Yes
U16 = >= 5 [Mode: No] => No
U20 > 28.01 [Mode: No] => No
H01 = No [Mode: No] => No
Rule 4 - estimated accuracy 79,57% [boost 86,7%]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]
Y03 = Yes [Mode: Yes] => Yes
Y03 = No [Mode: No]
U02 = Yes [Mode: Yes]
D01 = + 55 [Mode: No]
M03 = G [Mode: Yes] => Yes
M03 = A [Mode: No]

H19 <= 1580.84 [Mode: No]
C19 <= -146.67 [Mode: Yes] => Yes
C19 > -146.67 [Mode: No]
Y09 = Yes [Mode: No] => No
Y09 = No [Mode: No]
U07 = Yes [Mode: No] => No
U07 = No [Mode: Yes] => Yes
H19 > 1580.84 [Mode: Yes] => Yes
D01 = - 24 [Mode: Yes] => Yes
D01 = 25 - 34 [Mode: Yes]
Y09 = Yes [Mode: Yes] => Yes
Y09 = No [Mode: Yes]
Y01 = Yes [Mode: Yes]
C08 = Yes [Mode: Yes] => Yes
C08 = No [Mode: Yes]
H19 <= 5734.5298 [Mode: Yes] => Yes
H19 > 5734.5298 [Mode: No] => No
Y01 = No [Mode: Yes]
U16 = 0 [Mode: Yes]
U17 = 0 [Mode: Yes]
H17 = >= 1 [Mode: No] => No
H17 = 0 [Mode: Yes]
H15 = 0 [Mode: Yes]
TC = 0 [Mode: Yes]
H07 = Yes [Mode: No] => No
H07 = No [Mode: Yes]
H19 <= 4.73 [Mode: Yes] => Yes
H19 > 4.73 [Mode: No] => No
TC = 1 [Mode: Yes] => Yes
TC = 2 [Mode: Yes] => Yes
TC = 3 [Mode: Yes] => Yes
TC = 4 [Mode: No] => No
TC = >= 5 [Mode: No] => No
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: No] => No
U17 = 1 [Mode: Yes] => Yes
U17 = >= 2 [Mode: Yes] => Yes
U16 = 1 [Mode: Yes] => Yes
U16 = 2 [Mode: Yes] => Yes
U16 = 3 [Mode: Yes] => Yes
U16 = 4 [Mode: Yes] => Yes
U16 = >= 5 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: Yes]
U16 = 0 [Mode: Yes]
FC = 0 [Mode: Yes]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: Yes]

Y01 = Yes [Mode: No] => No
Y01 = No [Mode: Yes]
TY = 0 [Mode: Yes]
H19 <= 0.81 [Mode: No]
M03 = G [Mode: No] => No
M03 = A [Mode: Yes]
U01 = Yes [Mode: Yes]
H19 <= 0.25999999 [Mode: Yes] => Yes
H19 > 0.25999999 [Mode: No] => No
U01 = No [Mode: No] => No
H19 > 0.81 [Mode: Yes] => Yes
TY = 1 [Mode: Yes] => Yes
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: Yes] => Yes
FC = 1 [Mode: Yes]
H19 <= 1142.17 [Mode: Yes]
H19 <= 950.85999 [Mode: Yes]
TY = 0 [Mode: No]
H07 = Yes [Mode: Yes] => Yes
H07 = No [Mode: No] => No
TY = 1 [Mode: Yes] => Yes
TY = 2 [Mode: No] => No
TY = >= 3 [Mode: Yes] => Yes
H19 > 950.85999 [Mode: No] => No
H19 > 1142.17 [Mode: Yes] => Yes
FC = 2 [Mode: Yes]
H21 <= 4334.2798 [Mode: Yes]
H04 = Yes [Mode: Yes] => Yes
H04 = No [Mode: Yes]
C01 = Yes [Mode: Yes]
H19 <= 5.2399998 [Mode: Yes] => Yes
H19 > 5.2399998 [Mode: No] => No
C01 = No [Mode: Yes]
Y09 = Yes [Mode: No] => No
Y09 = No [Mode: Yes] => Yes
H21 > 4334.2798 [Mode: No] => No
FC = >= 3 [Mode: Yes] => Yes
U16 = 1 [Mode: Yes] => Yes
U16 = 2 [Mode: Yes] => Yes
U16 = 3 [Mode: Yes] => Yes
U16 = 4 [Mode: Yes] => Yes
U16 = >= 5 [Mode: Yes]
H22 <= -1417.49 [Mode: No] => No
H22 > -1417.49 [Mode: Yes] => Yes
D01 = 45 - 54 [Mode: Yes]
H19 <= 3.4000001 [Mode: No]
Y01 = Yes [Mode: Yes] => Yes
Y01 = No [Mode: No]

U16 = 0 [Mode: No]
H17 = >= 1 [Mode: Yes] => Yes
H17 = 0 [Mode: No]
H15 = 0 [Mode: No]
TY = 0 [Mode: No]
FC = 0 [Mode: No]
H19 <= 0 [Mode: No] => No
H19 > 0 [Mode: Yes] => Yes
FC = 1 [Mode: No] => No
FC = 2 [Mode: No] => No
FC = >= 3 [Mode: Yes] => Yes
TY = 1 [Mode: Yes] => Yes
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: No] => No
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: No] => No
U16 = 1 [Mode: No] => No
U16 = 2 [Mode: Yes] => Yes
U16 = 3 [Mode: No] => No
U16 = 4 [Mode: No] => No
U16 = >= 5 [Mode: Yes] => Yes
H19 > 3.4000001 [Mode: Yes]
C19 <= -562.29999 [Mode: No] => No
C19 > -562.29999 [Mode: Yes] => Yes
U02 = No [Mode: No]
M03 = G [Mode: No]
H17 = >= 1 [Mode: Yes] => Yes
H17 = 0 [Mode: No] => No
M03 = A [Mode: No]
Y01 = Yes [Mode: No]
U01 = Yes [Mode: Yes]
D01 = + 55 [Mode: No] => No
D01 = - 24 [Mode: Yes] => Yes
D01 = 25 - 34 [Mode: Yes] => Yes
D01 = 35 - 44 [Mode: Yes] => Yes
D01 = 45 - 54 [Mode: No] => No
U01 = No [Mode: No] => No
Y01 = No [Mode: No]
Y09 = Yes [Mode: No] => No
Y09 = No [Mode: No]
U16 = 0 [Mode: No]
L03 = Yes [Mode: No] => No
L03 = No [Mode: No]
U17 = 0 [Mode: No]
C08 = Yes [Mode: No]
H17 = >= 1 [Mode: Yes] => Yes
H17 = 0 [Mode: No]

H22 <= -94.650002 [Mode: Yes] => Yes
H22 > -94.650002 [Mode: No] => No
C08 = No [Mode: No] => No
U17 = 1 [Mode: No] => No
U17 = >= 2 [Mode: Yes] => Yes
U16 = 1 [Mode: No] => No
U16 = 2 [Mode: No] => No
U16 = 3 [Mode: No] => No
U16 = 4 [Mode: No] => No
U16 = >= 5 [Mode: No] => No
Rule 5 - estimated accuracy 76,17% [boost 86,7%]
C03 = Yes [Mode: Yes] => Yes
C03 = No [Mode: No]
H01 = Yes [Mode: No]
U01 = Yes [Mode: No]
TY = 0 [Mode: No]
L05 = Yes [Mode: No] => No
L05 = No [Mode: No]
M02 = 10 [Mode: No]
U16 = 0 [Mode: No]
H15 = 0 [Mode: No]
C01 = Yes [Mode: No] => No
C01 = No [Mode: No]
U05 = Yes [Mode: Yes] => Yes
U05 = No [Mode: No]
M03 = G [Mode: No] => No
M03 = A [Mode: No]
H19 <= 0.5 [Mode: No]
H22 <= -3.73 [Mode: No] => No
H22 > -3.73 [Mode: No]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: No] => No
H19 > 0.5 [Mode: Yes] => Yes
H15 = 1 [Mode: No] => No
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: Yes] => Yes
U16 = 1 [Mode: No] => No
U16 = 2 [Mode: No] => No
U16 = 3 [Mode: No] => No
U16 = 4 [Mode: Yes] => Yes
U16 = >= 5 [Mode: Yes] => Yes
M02 = 11 [Mode: No]
H15 = 0 [Mode: No]
C19 <= -14 [Mode: No] => No
C19 > -14 [Mode: No]
U07 = Yes [Mode: Yes] => Yes
U07 = No [Mode: No] => No
H15 = 1 [Mode: Yes] => Yes

H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: Yes] => Yes
M02 = 12 [Mode: No]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: No]
M03 = G [Mode: No] => No
M03 = A [Mode: No]
H19 <= 1153.41 [Mode: No] => No
H19 > 1153.41 [Mode: Yes] => Yes
M02 = 13 [Mode: No] => No
M02 = 14 [Mode: No] => No
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
C04 = Yes [Mode: Yes] => Yes
C04 = No [Mode: No]
C19 <= -262.48001 [Mode: No] => No
C19 > -262.48001 [Mode: Yes]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No]
C01 = Yes [Mode: No] => No
C01 = No [Mode: No]
H19 <= 0.0099999998 [Mode: No] => No
H19 > 0.0099999998 [Mode: Yes] => Yes
M02 = 6 [Mode: No]
C19 <= -392.41 [Mode: Yes] => Yes
C19 > -392.41 [Mode: No]
FL = >= 1 [Mode: Yes] => Yes
FL = 0 [Mode: No]
C08 = Yes [Mode: Yes] => Yes
C08 = No [Mode: No]
U16 = 0 [Mode: No]
U05 = Yes [Mode: No] => No
U05 = No [Mode: No]
H19 <= 0.44999999 [Mode: No]
C07 = Yes [Mode: Yes] => Yes
C07 = No [Mode: No]
H21 <= 8221.2402 [Mode: No]
U02 = Yes [Mode: No]
M03 = G [Mode: Yes] => Yes
M03 = A [Mode: No] => No
U02 = No [Mode: No] => No
H21 > 8221.2402 [Mode: Yes] => Yes
H19 > 0.44999999 [Mode: Yes] => Yes
U16 = 1 [Mode: No] => No
U16 = 2 [Mode: Yes] => Yes
U16 = 3 [Mode: No] => No
U16 = 4 [Mode: No] => No
U16 = >= 5 [Mode: No] => No

M02 = 7 [Mode: No]
H15 = 0 [Mode: No]
D01 = + 55 [Mode: No] => No
D01 = - 24 [Mode: No] => No
D01 = 25 - 34 [Mode: No]
U05 = Yes [Mode: Yes] => Yes
U05 = No [Mode: No] => No
D01 = 35 - 44 [Mode: Yes]
H05 = Yes [Mode: No] => No
H05 = No [Mode: Yes]
C18 = 0 [Mode: Yes] => Yes
C18 = 1 [Mode: Yes] => Yes
C18 = 2 [Mode: No] => No
C18 = 3 [Mode: Yes] => Yes
C18 = >= 4 [Mode: No] => No
D01 = 45 - 54 [Mode: No] => No
H15 = 1 [Mode: No] => No
H15 = 2 [Mode: Yes] => Yes
H15 = >= 3 [Mode: Yes] => Yes
M02 = 8 [Mode: No]
U05 = Yes [Mode: No] => No
U05 = No [Mode: No]
H19 <= 302.5 [Mode: No]
H19 <= 0.52999997 [Mode: No]
H19 <= 0.07 [Mode: No]
U07 = Yes [Mode: Yes] => Yes
U07 = No [Mode: No] => No
H19 > 0.07 [Mode: Yes] => Yes
H19 > 0.52999997 [Mode: No] => No
H19 > 302.5 [Mode: Yes] => Yes
M02 = 9 [Mode: No]
C18 = 0 [Mode: No]
C19 <= -254.61 [Mode: Yes] => Yes
C19 > -254.61 [Mode: No] => No
C18 = 1 [Mode: No] => No
C18 = 2 [Mode: Yes] => Yes
C18 = 3 [Mode: No] => No
C18 = >= 4 [Mode: No] => No
M02 = >= 15 [Mode: No] => No
TY = 1 [Mode: Yes]
L05 = Yes [Mode: Yes] => Yes
L05 = No [Mode: Yes]
M03 = G [Mode: Yes] => Yes
M03 = A [Mode: Yes]
M02 = 10 [Mode: No]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: No] => No
M02 = 11 [Mode: No] => No

M02 = 12 [Mode: Yes] => Yes
M02 = 13 [Mode: No] => No
M02 = 14 [Mode: Yes] => Yes
M02 = 4 [Mode: Yes] => Yes
M02 = 5 [Mode: Yes]
H05 = Yes [Mode: Yes] => Yes
H05 = No [Mode: Yes]
H21 <= 15930.89 [Mode: Yes]
H08 = Yes [Mode: No] => No
H08 = No [Mode: Yes] => Yes
H21 > 15930.89 [Mode: No] => No
M02 = 6 [Mode: Yes] => Yes
M02 = 7 [Mode: Yes]
U20 <= 55.360001 [Mode: Yes]
H15 = 0 [Mode: Yes] => Yes
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: Yes] => Yes
U20 > 55.360001 [Mode: Yes] => Yes
M02 = 8 [Mode: No] => No
M02 = 9 [Mode: No] => No
M02 = >= 15 [Mode: No] => No
TY = 2 [Mode: Yes]
FL = >= 1 [Mode: Yes] => Yes
FL = 0 [Mode: Yes]
TC = 0 [Mode: Yes] => Yes
TC = 1 [Mode: Yes] => Yes
TC = 2 [Mode: Yes]
C19 <= -119.08 [Mode: No] => No
C19 > -119.08 [Mode: Yes] => Yes
TC = 3 [Mode: Yes] => Yes
TC = 4 [Mode: No] => No
TC = >= 5 [Mode: Yes] => Yes
TY = >= 3 [Mode: Yes] => Yes
U01 = No [Mode: No]
Y03 = Yes [Mode: No] => No
Y03 = No [Mode: No]
M03 = G [Mode: No]
H17 = >= 1 [Mode: Yes] => Yes
H17 = 0 [Mode: No]
U07 = Yes [Mode: Yes] => Yes
U07 = No [Mode: No] => No
M03 = A [Mode: No]
C08 = Yes [Mode: No]
U05 = Yes [Mode: No] => No
U05 = No [Mode: Yes]
H15 = 0 [Mode: Yes]
TY = 0 [Mode: No]

H19 <= 372.03 [Mode: Yes]
H19 <= 173.52 [Mode: No]
H19 <= 0.72000003 [Mode: Yes] => Yes
H19 > 0.72000003 [Mode: No] => No
H19 > 173.52 [Mode: Yes] => Yes
H19 > 372.03 [Mode: No] => No
TY = 1 [Mode: No] => No
TY = 2 [Mode: Yes] => Yes
TY = >= 3 [Mode: Yes] => Yes
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: Yes] => Yes
H15 = >= 3 [Mode: No] => No
C08 = No [Mode: No]
U02 = Yes [Mode: Yes]
U17 = 0 [Mode: Yes]
H17 = >= 1 [Mode: No] => No
H17 = 0 [Mode: Yes]
Y01 = Yes [Mode: Yes] => Yes
Y01 = No [Mode: Yes]
H15 = 0 [Mode: Yes]
D01 = + 55 [Mode: Yes] => Yes
D01 = - 24 [Mode: Yes] => Yes
D01 = 25 - 34 [Mode: No] => No
D01 = 35 - 44 [Mode: No] => No
D01 = 45 - 54 [Mode: Yes] => Yes
H15 = 1 [Mode: Yes] => Yes
H15 = 2 [Mode: No] => No
H15 = >= 3 [Mode: Yes] => Yes
U17 = 1 [Mode: Yes] => Yes
U17 = >= 2 [Mode: Yes] => Yes
U02 = No [Mode: No] => No
H01 = No [Mode: No] => No

KAYNAKÇA

KİTAPLAR

Abonji J., Feil B., **Cluster Analysis for Data Mining and System Identification**, BirkHauser Verlag AG, 2007

Agresti A., **Categorical Data Analysis**, John Wiley & Sons, 1990

Andersen E.B., **Introduction to the Statistical Analysis of Categorical Data**, Springer, 1997

Armutlulu İ.H., **İşletmelerde Uygulamalı İstatistik**, Alfa, 2000

Berry M.J.A., Linoff G.S., **Mastering Data Mining : The Art and Science of CRM**, John Wiley & Sons, 2000

Cabena P., Stadler R., Zanasi A., **Discovering Data Mining : From Consept to Implementation**, Prentice Hall, 1998

Flach P., Lavrac N., **Intelligent Data Analysis**, Springer, 1999

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

Han J., Kamber M., **Data Mining : Concepts and Techniques**, Academic Press, 2001

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

Haykin S., **Neural Networks: A Comprehensive Foundation**, Prentice Hall, 1994

Hosmer D.W., Lemeshow S., **Applied Logistic Regression**, John Wiley & Sons, 2000

Kantardzic M., **Data Mining: Consepts, Models, Methods and Algorithms**, IEEE, 2003

Larose D.T., **Discovering Knowledge in Data: An Introduction to Data Mining**, John Wiley & Sons, 2005

McLachlan G.J., **Discriminant Analysis and Statistical Pattern Recognition**, John Wiley & Sons, 2004

Menard S., **Applied Logistic Regression Analysis**, Sage Publication, 2002

Michie D., **Machine Learning, Neural and Statistical Classification**, Ellis Horwood, 1995

Ripley B.D., **Pattern Recognition and Neural Networks**, Cambridge University Press, 1996

Swift R., **Accelerating Customer Relationship**, Prentice Hall PTR, 2001

SÜRELİ YAYINLAR

Akpınar H., "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İstanbul Üniversitesi İşletme Fakültesi Dergisi**, Cilt.29, Sayı.1, 2000

Apte C. ve diğerleri, "Business Applications of Data Mining", **Communications of the ACM**, Vol.45, No.8, 2002

Bender R., Grouven U., "Using Binary Logistic Regression Models for Ordinal Data with Non-proportional Odds", **Journal of Clinical Epidemiol**, Vol.51, No.10, 1998

Bircan H., "Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama", **Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi**, Sayı.2, 2004

Bose I., Mahapatra R.K., "Business Data Mining: A Machine Learning Perspective", **Information & Management**, No.39, 2001

Brachman R.J. ve diğerleri, "Mining Business Databases", **Communications of the ACM**, Vol.39, No.11, 1996

Brito P., Malerba D., "Mining Official Data", **Intelligent Data Analysis,** No.7, 2003

Caudill S.B., "The Necessity of Mining Data", **Atlantic Economic Journal**, Vol.16, No.3, 1988

Chou D.C., Chou A.Y., " A Manager's Guide to Data Mining", **Information Systems Management**, Fall 1999

Clifton C., Thuraisingham B., "Emerging Standards for Data Mining", **Computer Standards & Interfaces**, No.23, 2001

Crone S.F. ve diğerleri, "The Impact of Preprocessing on Data Mining: An Evaluation of Classifier Sensitivity in Direct Marketing", **European Journal of Operational Research**, No.173, 2006

Denton F.T., "Data Mining as An Industry", **The Review of Economics and Statistics**, Vol.65, No.1, 1985

Dhar V., "Data Mining in Finance: Using Counterfactuals to Generate Knowledge from Organizational Information Systems", **Information Systems**, Vol.23, No.7, 1998

Dreiseitl S., Ohno-Machado L., "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review", **Journal of Biomedical Informatics,** No.35, 2002

Fayyad U., Haussler D., Stolorz P., "Mining Scientific Data", **Communications of the ACM**, Vol.39, No.11, 1996

Fayyad U., Piatetsky-Shapiro G., Smyth P., " From Data Mining to Knowledge Discovery in Databases", **AI Magazine**, Vol.17, No.3, 1996

Fayyad U., Piatetsky-Shapiro G., Smyth P., " The KDD Process for Extracting Useful Knowledge from Volumes of Data", **Communications of the ACM**, Vol.39, No.11, 1996

Fayyad U., Uthurusamy R., "Data Mining and Knowledge Discovery in Databases", **Communications of the ACM**, Vol.39, No.11, 1996

Flawley W.J., Piatetsky-Shapiro G., Matheus C.J., "Knowledge Discovery in Databases : An Overview", **AI Magazine**, Vol. 13, No. 3, 1992

Fogler H.R., "Investment Analysis and New Quantitative Tools", **Journal of Portfolio Management**, Vol.21, No.4, 1995

Forcht K.A., Cochran K., "Using Data Mining and Data Warehousing Techniques", **Industrial Management & Data Systems**, Vol.99, No.5, 1999

Giraud-Carrier C., Povel O., "Characterising Data Mining Software", **Intelligent Data Analysis,** No.7, 2003

Glymour C. ve diğerleri, "Statistical Inference and Data Mining", **Communications of the ACM**, Vol.39, No.11, 1996

Glymour C. ve diğerleri, "Statistical Themes and Lessons for Data Mining", **Data Mining and Knowledge Discovery**, Vol.1, 1997

Goil S., Choudhary A., "High Performance OLAP and Data Mining on Parallel Computers", **Data Mining and Knowledge Discovery**, No.1, 1997

Hand D.J., "Pattern Recognition", **Handbook of Statistics,** Vol.24, 2005

Hand D.J., "Statistics and Data Mining: Intersecting Disciplines", **SIGKDD Explorations**, Vol.1, No.1, 1999

Hegland M., "Data Mining Techniques", **Acta Numerica**, 2001

Hormozi A.M., Giles S., "Data Mining: A Competitive Weapon for Banking and Retail Industries", **Information Systems Management**, Spring 2004

Imielinski T., Mannila H., "A Database Perspective on Knowledge Discovery", **Communications of the ACM**, Vol.39, No.11, 1996

Kaastra I., Boyd M., "Designing a Neural Network for Forecasting Financial and Economic Time Series", **Neurocomputing**, No.10, 1996

Kaufman K.A., Michalski R.S., "From Data Mining to Knowledge Mining", **Handbook of Statistics**, Vol.24, 2005

Leinweber D.J., Arnott R.D., "Quantitative and Computational Innovation in Investment Management", **Journal of Portfolio Management**, Vol.21, No.2, 1995

Liew P.L. ve diğerleri, "Comparison of Artificial Nueral Networks with Logistic Regression in Prediction of Gallbladder Disease Amoung Obese Patients", **Digestive and Liver Diseases**, No.39, 2007

Liu Y., Schumann M., "Data Mining Feature Selection for Credit Scoring Models", **Journal of the Operational Research Society**, 2005

Lovell M.C., "Data Mining", **The Review of Economics and Statistics,** Vol.65, No.1, 1983

Man D., "Answering Some Common Data Warehousing Questions", **Direct Marketing**, Vol.59, No.8, 1996

McCarty J.A., Hastak M., "Segmentation Approaches in Data Mining: A Comparison of RFM, CHAID and Logistic Regression", **Journal of Business Research**, No.60, 2007

Newing R., "Data Mining", **Financial Management**, Vol.74, No.9, 1996

Oğuzlar A., "Veri Ön İşleme", **Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**, Sayı.21, 2003

Pan S-L., Tan C., Lim E.T.K., "Customer Relationship Management in e-Government: A Relational Perspective", **Decision Support Systems**, No.42, 2006

Quinlan J.R., " Improved Use of Continuous Attributes in C4.5", **Journal of Artificial Intelligence Research**, Vol.4, 1996

Raghavan N.R.S., "Data Mining in e-commerce: A Survey", **Sadhana**, Vol.30, 2005

Rosenblatt, F., "The Perceptron: A Probabilistic Model For Information Storage And Organization in the Brain", **Psychological Review**, Vol.65, No.6, 1958

Rygielski C., Wang J., Yen D.C., "Data Mining Techniques for Customer Relationship Management", **Technology in Society**, No.24, 2002

Shearer C., "The CRISP-DM Model: The New Blueprint for Data Mining", **Journal of Data Warehousing**, Vol.5, No.4, 2000

Spangler W.E., May J.H., Vargas L.G., "Choosing Data Mining Methods for Multiple Classification: Representational and Performance Measurement Implications for Decision Support", **Journal of Management Information Systems**, Vol.16, No.1, 1999

Strauss D., "The Many Faces of Logistic Regression", **The American Statistician**, Vol.46, No.4, 1992

Wegman E.J., Solka J.L., " Statistical Data Mining", **Handbook of Statistics**, No.24, 2005

Zhang D., Zhou L., "Discovering Golden Nuggets: Data Mining in Financial Application", **IEEE Transactions on System, Man, and Cybernetics**, Vol.34, No.4, 2004

Zhou Z., "Three Perspectives of Data Mining", **Artificial Intelligence**, No.143, 2003

TEBLİĞLER, RAPORLAR, KITAPÇIKLAR, INTERNET KAYNAKLARI

Alpaydın E., "Zeki Veri Madenciliği : Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", *Bilişim 2000 Eğitim Semineri*, İstanbul, 2000

Bayraktar R., "Veri Tabanı ve Akılcı Düşünce Üzerine", *http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=127*, [Erişim 14.03.2006]

CRISP – DM Konsorsiyumu, "CRISP – DM 1.0 Step – by – Step Data Mining Guide", *http://www.crisp-dm.org/CRISPWP-0800.pdf*, [Erişim 21.01.2005]

Edelstein H., "Data Mining in Depth: Description is not Prediction", Mart 2003, *http://www.dmreview.com/issues/20030301/6388-1.html*, [Erişim 08.05.2005]

Edelstein H., "Data Mining in Depth: TIAin't", Nisan 2003, *http://www.dmreview.com/issues/20030401/6512-1.html*, [Erişim 08.05.2005]

Edelstein H., "Data Mining in Depth: Using Data Mining to Find Terrorists", May.2003, *http://www.dmreview.com/issues/20030501/6655-1.html*, [Erişim 08.05.2005]

Integral Solutions Co., *Clementine 8.0 Algorithms Guide*, 2003

Integral Solutions Co., *Clementine 8.0 User's Guide*, 2003

Koyuncugil A.S., "Veri Madenciliği ve Sermaye Piyasalarına Uygulanması", *SPK Araştırma Raporu*, 2007

Microsoft Türkiye, "Bir Veri Tabanının Sunması Gereken Temel Özellikler", *http://www.microsoft.com/turkiye/.../.../sqlserver/veritabani.asp*, [Erişim 16.10.2005]

Özmen Ş., "İş Hayatı Veri Madenciliği ile İstatistik Uygulamalarını Yeniden Keşfediyor", *V. Ulusal Ekonometri ve İstatistik Sempozyumu*, Adana, 2001

RuleQuest Research, "C5.0 An Informal Tutorial", *http://www.rulequest.com/see5-unix.html*, [Erişim 17.02.2008]

Türkiye Bankalar Birliği, "Bankalarımız 2007", *http://www.tbb.org.tr/turkce/kitap2007/2007.asp*, [Erişim 05.06.2008]

Türkiye Cumhuriyeti Merkez Bankası, "Finansar İstikrar Raporu", *http://www.tcmb.gov.tr/yeni/evds/yayin/finist/finist4.php*, [Erişim 18.07.2007]

Two Crows Co., "Introduction to Data Mining and Knowledge Discovery", *http://www.twocrows.com/intro-dm.pdf*, [Erişim 08.05.2005]

Widom J., "Research Problems in Data Warehousing", *4th International Conference on Information and Management*, 1995

Wuensch K.L., "Binary Logistic Regression with SPSS", [Erişim 03.06.2008] *http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.doc*