T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
İŞLETME YÖNETİMİ
YÜKSEK LİSANS PROGRAMI

VERİ MADENCİLİĞİ: SINIFLANDIRMA ve TAHMİN YÖNTEMLERİNİ KULLANARAK BİR UYGULAMA

YÜKSEK LİSANS TEZİ

PELİN BİÇEN

Tez Danışmanı: Prof. Dr. S. ÜMİT OKTAY FIRAT

İSTANBUL, 2002

TEŞEKKÜR

"Veri madenciliği: Sınıflandırma ve Tahmin Yöntemlerini Kullanarak Bir Uygulama" başlıklı yüksek lisans tezimin hazırlanmasında bana her anlamda yardımcı olan ve akademik dünyada kalmam için manevi desteğini benden esirgemeyen Sayın Prof. Dr. S. Ümit Oktay Fırat'a ve aileme,

Tezin içerik ve teknik anlamda hazırlanmasında bana her konuda yardımcı olan, desteğini her zaman hissettiğim, yaşadığım problemleri çözerken günün her saati yanımda olmaya çalışan Fatih Köksal'a,

Çalışmamın uygulama aşamasında gerekli olan verinin ve ilgili veri madenciliği paket programının sağlanmasında bana yardımcı olan ve bunun yanında manevi anlamda beni yalnız bırakmayan SAS Ins. Türkiye çalışanlarına,

Tezimin uygulama aşamasında yaşadığım zorlukları yenmemde akademi ve sektor tecrübeleri ile bana yardımcı olmaya çalışan Yonca Aslanbay Karapazar'a, Zehra Üstdiken'e, İpek Özel'e ve Mehtap Öztürk'e,

Tezimi yazdığım dönem boyunca her türlü ruh halimi kaldıran Timuçin Kurt, Aslı Telli, Füsun Akdoğan, Özlem Hesapçı ve Aybike Hatemi'ye teşekkür etmeyi bir borç bilirim.

Pelin Biçen

İÇİNDEKİLER

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

İşletme ve bilimsel içerikli veri tabanlarının gün geçtikçe büyümesi, veri tabanlarında bulunan verinin analiz edilmesini ve yorumlanmasını zorlaştırdı. Bu noktada, veri tabanı analiz sürecini otomatikleştirecek yeni nesil tekniklere ve araçlara ihtiyaç duyulmaya başlandı. Bu anlamda, bu teknikler ve araçlar veri tabanlarında bilgi keşfi ve veri madenciliği teknikleri olarak bilinen ve çok hızlı gelişen bir alana konu oldular.

Veri madenciliği teknolojisi her geçen gün gelişmektedir. Tıp, finans, sağlık, pazarlama, sigorta ve diğer işletme sektörlerinde bu gelişen teknolojiye artan derecede ihtiyaç duyulmaktadır. İstatistik paket programlar konusunda uzman uluslararası yazılım firmaları, veri madenciliği pazarındaki rekabette lider şirket olabilmek için sürekli yeni yazılımlar geliştirmektedir. Bu anlamda veri madenciliği alanı istatistik yazılım programlarının bir uzantısı olarak görülmektedir. SAS şirketinin felsefesine göre, "SAS kütüphanesinde istatistik herşeydir, veri madeciliği ise geriye kalan herşeydir."

Bu çalışmada, ilk olarak veri tabanlarında bilgi keşfi ve veri madenciliği kavramları detaylı olarak açıklanmıştır. Veri madenciliği teknikleri ve uygulama alanları geniş bir çerçevede incelenmiştir. Son bölümde, günümüz işletme dünyasında çok sık karşılaşılan, müşterilerin kredi taleplerinin değerlendirilmesi ve karlılık durumlarına göre müşterilerin bölümlendirilmesi problemi, veri madenciliği sınıflandırma ve tahmin modelleri uygulanarak çözümlenmiştir. Çözüm sürecinde SAS Enterprise Miner 4.2 veri madenciliği paketi kullanılmıştır.

SUMMARY

The explosive growth of many business and scientific databases has far exceeded the ability to interpret the data. At this point, there was a creating need for a new generation of tools and techniques for automated database analysis. The tools and techniques are the subject of the rapidly emerging field of knowledge discovery in databases (KDD) and datamining techniques.

Data mining technology is rapidly evolved. There is a creating need of this emerging technology in medical, finance, health, marketing, insurance and other related sector. Expert multinational firms that produce statistical application packages has continuously developed new statistics software in order to be a leader in data mining market. The philosophy of SAS Inst. is " Statistics is everything in SAS Library and data mining is everything else". By this definition, it is easy to understand that data mining has now ceased to exist.

In this research, first of all data mining and knowledge discovery in databases (KDD) concepts were explained. Data mining techniques and their application areas were tried to be examined in extense form. In the last chapter, customer credit scoring and customer segmentation problem that was steadily encountered in current business world was solved with predictive and classification data mining modeling techniques. In the solution period, SAS Enterprise Miner 4.2 data mining package was used.

GİRİŞ

1990'lı yıllarda çok miktarda verinin ortaya çıkmasıyla beraber teknik olarak ve işletmecilik anlamında bazı problemler baş gösterdi. Zamanla artan verinin toplanması ve saklanması dışında analiz edilmesi, özetlenmesi ve mevcut veriden anlamlı bilgi çıkarımı insanların yapabileceğinden çok daha fazlasını gerektiriyordu. Geleneksel anlamda insan odaklı veri analizi, çok ciddi miktarlardaki veri ile çalışmayı olanaksız bir duruma getirmişti. Veri tabanı teknolojisinin gelişmesiyle beraber verinin toplanması ve saklanması olanaklı duruma getirildi. Geriye sadece çok miktarda verinin nasıl analiz edileceği ve veriden nasıl anlamlı bilgi çıkarımı sağlanacağı problemi kalıyordu. Bu noktada, işletme dünyası veri tabanlarında bilgi keşfi (KDD- knowledge discovery in databases) ve veri madenciliği (data mining) kavramları ile tanıştı. Yapay zeka (artificial intelligence), istatistik, matematik, makine öğrenmesi (machine learning), örüntü tanıma (pattern recognition) ve veri görselleştirmesi (data visualisation) kavramlarını birarada kullanan veri tabanlarında bilgi keşfi ve veri madenciliği süreci, zamanla artan veriden anlamlı bilgiler çıkarmada ve işletme problemlerini çözmede işletmelere yardımcı olmaya başladı.

Bu çalışmada, veri madenciliği kavramı ve teknikleri temel alınmıştır. Bu anlamda, ilk bölümde genel olarak veri tabanlarında bilgi keşfi ve veri madenciliği kavramları, veri madenciliğinin hayati döngüsü, veri madenciliğinin alanları ve veri madenciliği modelleme sürecinin aşamaları ile veri ambarı ve online analitik işlemenin veri madenciliği kavramı ile ilişkisi incelenmiştir. İkinci bölümde, veri madenciliği tekniklerinden sepet analizi (market basket analysis), bellek tabanlı yöntemler (memory based reasoning), karar ağaçları (decision trees), yapay sinir ağları (artificial neural networks) ve kümeleme analizi (cluster analysis) detaylı olarak incelenmiştir. Veri madenciliği tekniklerinin olumlu ve olumsuz yönleri ile bu tekniklerin hangi işletme problemlerine çözüm aradığı anlatılmıştır. Son bölümde, ilk olarak veri madenciliğinin uygulama alanları incelenmiş daha sonra teorik çerçevede değerlendirilen modellerden tahmin ve sınıflandırma modellerinin uygulandığı bir çalışma gerçekleştirilmiştir. Bu çalışmada veri madenciliği tekniklerinden lojistik regresyon, karar ağacı analizi ve kümeleme analizi söz konusu probleme çözüm sunmak amacıyla uygulanmıştır.

BÖLÜM 1
VERİ MADENCİLİĞİ KAVRAMI

Karmaşık, dinamik ve kaotik bir ekonomi ortamında işletmelerin karar verme süreçlerindeki etkinlikleri, çok değişkenli büyük miktarlardaki veri kümelerinde saklı bulunan bilginin elde edilmesi ve işlenmesine bağlıdır. Bu noktada, doğru yanıtlandığında işletmelerin isabetli karar almalarında belirleyici rol oynayacak stratejik önemde sorular vardır. Örneğin;

i. En iyi müşterilerinizi nasıl koruyabilirsiniz?

ii. Hangi müşterilerinizin sizi bırakıp rakiplerinize gitme olasılığı daha yüksektir?

iii. Müşterilerinizin gelecekte nasıl davranacağını doğru bir şekilde nasıl öngörebilirsiniz?

Yukarıdaki soruların doğru yanıtları çok büyük miktardaki veri yığınlarının altında gizlidir. Veri madenciliği (data mining) bu veri yığınlarının içinden işletme yöneticileri için en gerekli olanlarının seçilmesi, düzenlenmesi ve modellenmesi süreçlerini içerir. Bu noktada veri madenciliğini, karar verme mekanizmaları için yeni bilgiler üreten teknikler ve kavramlar bütünü olarak tanımlamak mümkündür.[1]

Bu bölümde ilk olarak veri tabanlarında bilgi keşfi süreci ve veri madenciliği kavramları açıklanacak ve veri madenciliğinde kullanılan modeller ile bu modellerin işlevlerine göre ayrımları incelenecektir. Daha sonra, veri madenciliğinin işletme problemlerini çözmede ve karar verme süreçlerinde işletmelerde oluşturduğu katma değer incelecektir. Veri madenciliğinin uygulama sürecinde hangi aşamalardan geçtiğinin ayrıntılı biçimde incelenmesi ve veri madenciliği tekniklerinin performanslarının nasıl değerlendirildiği aşamasından sonra veri madenciliğinde önemli bir yeri olan veri ambarı (dataware house) ve on-line analitik işleme (OLAP- online analytical processing)

[1] "SAS VERİ MADENCİLİĞİ".http/www.sas.com/offices/Europe/Turkey/cozveri.com, s:1 (09.09.2001)

kavramları genel hatlarıyla ele alınacaktır. Veri ambarı ve on-line analitik işlemeninin genel tanımları verildikten sonra veri madenciliği ile aralarındaki ilişki incelenecektir.

1.1. Veri Tabanlarında Bilgi Keşfi Süreci (Knowledge Discovery in Databases)

Dijital dünyadaki teknolojik gelişmeler, kullanılan ve saklanması gereken veri miktarını her geçen gün arttırmaktadır. Boyutları hızla artan veriden anlamlı bilgiler çıkarmak için bilgisayar hızlarının ve güçlerinin artmasını sağlayacak yeni teoriler ve araçlar geliştirilmektedir. Bu teoriler ve araçlar veri tabanlarında bilgi keşfi (knowledge discovery) süreçlerinin konusunu oluşturmaktadır.

Veriden anlamlı ilişkiler ve örüntüler çıkarma sürecine literatürde[2], veri madenciliği, bilgi çıkarımı (knowledge extraction), bilgi keşfi, veri arkeolojisi ve veri örüntü işleme (data pattern processing) gibi isimler verilmektedir. Veri madenciliği tanımını daha çok istatistikçiler, veri analizcileri (data analyzer) ve yönetim bilişim sistemleri (management information systems-MIS) kullanıcıları kullanmaktadırlar. İlk olarak 1989 yılında yapılan bir atölyede (workshop), veri işleme sürecinde bilginin son ürün olduğunu vurgulamak için "veri tabanlarında bilgi keşfi" tanımlaması kullanılmıştır.

Veri tabanlarında bilgi keşfi, veriden gerçekten anlamlı ve yararlı bilginin çıkarıldığı süreç olarak tanımlanmaktadır. Bu anlamda veri madenciliği bu sürecin sadece bir kısmını oluşturmaktadır. Bilgi keşfi süreci; makine öğrenmesi (machine learning), örüntü tanıma (pattern recognition), istatistik, yapay zeka (artificial intelligence-AI) ve veri görselleştirmesi (data visualisation) kavramlarının bir araya gelmiş şeklidir. Veri madenciliği; bilgi keşfi sürecinde makine öğrenmesi, örüntü tanıma ve istatistik tanımlamalarını birarada kullanan bir aşamadır. Bilgi keşfi sürecinde amaç, büyük veri kümelerindeki düşük seviyedeki veriden yüksek seviyede bilgi çıkarımını sağlamaktır.[3]

---

[2] Piatesky- Shapiro, "Knowledge Discovery in Real Databases: A report on the IJCAI-89 Workshop", *AI Magazine*, 1999, cilt:11 , sayı:5, s: 68-70

[3] Usama Fayyad, Gregory Piatesky- Shapiro, Padhric Smith and Ramasamy Uthunamy. *Advances in Knowledge Discovery and Data Mining*.USA: MIT Press,1996 s:7

Veri tabanlarında bilgi keşfi, verinin nasıl saklanması ve algoritmaların büyük veri kümelerine nasıl uyarlanması gerektiği, sonuçların nasıl yorumlanacağı ve görselleştirileceği, insan-makine etkileşiminin nasıl modelleneceği sorularının cevabını aramaktadır. Bilgi keşfi çok disiplinli bir aktivitedir.[4]

Veri tabanlarında bilgi keşfi interaktif ve tekrarlı (iterative) bir süreçtir. Bu sürecin interaktif yapısının aşamaları aşağıda incelenmektedir.[5]

i. Veri tabanı bilgi keşfi sürecinin amacının net olarak belirlenmesi ve uygulama sahasının geliştirilmesi

ii. Hedef veri kümesinin yaratılması, veri kümesinin ve keşfin yapılacağı değişkenlerin seçimi

iii. Verinin ön-işlemesi (pre-processing)

iv. Veri projeksiyonu ve boyut sayısının indirgenmesi (data dimension reduction)

v. Veri tabanı bilgi keşfi sürecinin amaçları ile veri madenciliği yöntemlerinin (regresyon, sınıflandırma, tahmin, bağıntı kurma, kümeleme) karşılaştırılması

vi. Keşifsel analiz, model ve hipotez seçimi

vii. Veri madenciliği

viii. Veri madenciliği sonuçlarının yorumlanması, örüntülerin tanımlanması

ix. Keşfedilen bilgilerin kullanılması, sonuçların kontrol edilmesi.

1.2. Veri Madenciliği (Data Mining)

Bilgisayar sistemlerinin her geçen gün güçlerinin artması, işlemcilerinin gittikçe hızlanması çok büyük miktardaki verinin saklanabilmesine imkan vermektedir. Saklanan milyonlarca bilgiden hareketle her malın zaman içindeki hareketini izlemek ve bu malı tüketen müşterilerin verilerine ulaşmak mümkündür.

---

[4] J. Sharager ve P. Langley, *Computational Models of Scientific Discovery and Theory Formation*, California: Kaufmann Press, 1990, s: 37

[5] R. Brachmen ve T. Anand, *The Process of Knowledge Discovery in Databases. A human Centered Approach. In Advances in Knowledge Disvovery and Data Mining*, California: AAAI Press, 1996, s:44

Veri kendi başına değersiz olduğundan verinin amacımız doğrultusunda bilgiye çevrilmesine veri analizi (data analysis) denmektedir. Veri analizi yaparak bir mal için bir sonraki ayın satış tahminlerini çıkarabilir, müşterileri satın aldıkları mallara göre gruplayabilir, yeni çıkacak bir ürün için potansiyel müşterileri belirleyebilir, müşterilerin hareketlerini izleyerek ve inceleyerek onların davranışları ile ilgili tahminler yapabiliriz. Milyonlarca malın ve müşterinin olabileceği düşünülürse bu analizin otomatik olarak yapılmasının zorunluluğu ortaya çıkmaktadır. Bu noktada veri madenciliği devreye girmektedir. Veri madenciliği, büyük miktardaki veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanılarak aranmasıdır.[6]

Beery ve Linoff'un "Data Mining Techniques"[7] kitabında verilen daha geniş bir tanıma göre ise veri madenciliği; büyük miktardaki verinin, bu veriden anlamlı ve yorumlanabilir modeller ve kurallar çıkarabilmek amacıyla analiz edilmesidir. Bu tanımdan yola çıkarak veri madenciliğinin amacını, işletmelerin müşterilerini daha iyi anlayabilmesi ve bu mantıkla pazarlama, satış ve müşteri ilişkileri yönetimini (CRM- customer relationship management) geliştirebilmesi olarak tanımlayabiliriz.

Veri madenciliği uzun süredir üzerinde çalışılan bir konu olmasına rağmen, son zamanlarda iş dünyasında daha etkin bir maliyet kontrolü ve daha yüksek kârlılık elde etme konusunda sağladığı katkılar ile ilgi kazanmıştır. Gartner Group Araştırma Şirketi, gelecek on yıl içinde, hedef pazarlarda veri madenciliği kullanımının yüzde 80'lere ulaşacağı tahmininde bulunmaktadır. META Group ise veri madenciliği pazarının 2002 yılı içerisinde 800 milyon dolara yükseleceğini öngörmektedir.[8]

Veri madenciliği uygulamaları, pek çok sektör ve iş fonksiyonlarında kullanılmaktadır. Telekomünikasyon, hisse senedi işlemleri, kredi kartı ve sigorta

---

[6] Ethem Alpaydın, "Zeki Veri Madenciliği: Ham veriden altın bilgiye ulaşma yöntemleri", *Bilişim 2000 Eğitim Semineri*, İstanbul (2000),s:1.
[7] Michael J.A. Berry ve Gordon Linoff. *Data Mining Techniques: ForMarketing, Sales and Customer Support.* USA: John Wiley&Sons, Inc., 1997, s:5
[8] "SAS VERİ MADENCİLİĞİ",age, s:2, (09.09.2001)

şirketleri veri madenciliğini hizmetlerinin istismar edilmesini önlemek için uygulamaktadır. Tıp endüstrisi ameliyat prosedürlerinin, tıbbi testlerin ve ilaçla tedavinin etkinliğini tahmin etmekte, perakende sektörü de özel uygulamaların etkinliğini değerlendirmektedir.

Veri madenciliği'nin tüm uygulama alanlarının içinde en çok kullananı ise veritabanı pazarlamacılığı ve müşteri ilişkileri yönetimidir. Pazarlamacılar, bu yolla hedefledikleri kampanyalar için uygun müşteri adaylarını belirlemekte ve müşterilerin firmanın rakiplerini tercih nedenlerini saptamaktadır. Böylelikle maliyetler düşürülmekte ve kârlılık arttırılmaktadır. Bu anlamda tablo 1.1'de görüleceği gibi veri madenciliği tipik operasyonel sistemlerden farklı çalışmaktadır.

Tablo 1.1. Veri madenciliğinin tipik operasyonel sistemlerden farkları

Tipik Operasyonel Sistemler	Veri madenciliği sistemleri
Geçmiş verilerin raporlanması durumu sözkonusudur	Geçmiş ve günümüz verileri kullanılarak ve analiz edilerek gelecekteki hareketler belirlenir
Belirli bir takvim sistemine bağlı iş akışı söz konudur	İşletme ve pazarlama ihtiyaçlarına paralel bir iş akışı söz konudur
Sınırlı veri kullanımı söz konusudur	Ne kadar çok veri olursa, sonuçlarda o denli sağlıklı olur
İşletme odaklılık söz konusudur	Müşteri, ürün, satış odaklılık söz konusudur
Cevap süresi interaktif sistemler için dakikalarla, raporlar için hafta/aylarla ölçülür.	Cevaplama süresi her durumda dakika ve saatlerle ölçülür
Tanımlayıcıdır. (descriptive)	Yaratıcıdır. (creative)

Kaynak: Michael J.A. Berry ve Gordon Linoff. *Data Mining Techniques: ForMarketing, Sales and Customer Support.* USA: John Wiley&Sons, Inc., 1997, s:33

1.3. Veri Madenciliğinde Model ve Algoritma Kavramları

Veri madenciliğinde kullanılan modelleri ve bu modelleri işlevlerine göre ise ana başlıklar halinde incelemeye başlamadan önce tanımlanması gereken iki kavram bulunmaktadır. Bunlardan ilki model, ikincisi ise algoritma kavramlarıdır.

1.3.1. Model Kavramı

Gerçek dünyadaki bir olayın, sürecin veya birimlerden oluşan ve birimleri arasındaki iç ilişkiler yanında çevre ile dış ilişkilere göre işleyen bir sistemin belli bir anlatımına model denir. Anlatım, sözle ve çizimle belli bir ölçekte fiziki benzer oluşturmak veya başka bir şekilde yapılmakla birlikte en geçerli anlatım, bilimin ortak dili olan matematik ile yapılmaktadır.

Bir modelin var olması, en iyi ve en doğru sonuçlara ulaşacağımız anlamına gelmemelidir. İyi ve kötü modeller mevcuttur. Önemli olan modeli kurarken ve geliştirirken sonuçlarını değerlendirebilmektir. Modeller veri madenciliği hakkında konuşurken kullandığımız ortak dildir.

Bir modeli oluştururken karşılaşacağımız bazı sorunlar söz konusudur. Az uygunluk (underfitting) ve aşırı uygunluk (overfitting) problemleri bu sorunların başında gelmektedir.

Veri madenciliği tekniklerini kullanırken bu problemlerle oldukça sık karşılaşılır. Örneğin, belirgin bir yeri aramak için çok detaylı bir haritadan yararlanıyorsak, çok fazla bilginin içinde kaybolma olasılığımız yüksektir. Bu duruma aşırı uygunluk (overfitting) problemi diyoruz. Özellikle bir bölgeyi aradığımızda ise kullandığımız harita eğer çok ana hatlarıyla şehri ele alan türden ise sorunumuzun çözülme ihtimali çok düşüktür. Bu duruma da az uygunluk (underfitting) problemi diyoruz. Aşırı uygunluk problemi, veri sayısının yetersiz olmasından kaynaklanır. Model, eğitim kümesinde bulunan verinin azlığından dolayı mevcut olan bir grup verinin tüm özelliklerini ezberler ve genelleştirme yapamaz.

Az uygunluk durumu ise veride tahmin gücünü yükseltebilecek yapıda olan değişkenlerin analiz edilecek veri kümesinden çıkarılması halinde karşımıza çıkar. Genelde bu durumla, istatistiksel metodların kullanımı sırasında çok sık karşılaşılır. Örnek vermek gerekirse, bir şirket belirli bir marka bebek maması kullanımını arttırmak için bir postalama yapmak istemektedir. Bebeğin erkek veya kız olmasının analiz aşamasında dikkate alınmadığını ve bu bilginin de önemsenmediğini varsayacak olursak, kız çocuğu ailelerinin fatura ödemelerine daha sadık olduğu bilgisini ihmal etmiş oluruz ve yaptığımız postalamanın hedefine ulaşma imkanını sınırlandırırız.[9]

1.3.1.1. Veri Madenciliği Modelleri ve İşlevleri

Akpınar'a göre veri madenciliğinde kullanılan modeller, tahmin edici (predictive) ve tanımlayıcı (descriptive) olmak üzere iki genel başlık altında incelenmektedir.[10] Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. Yazar, veri madenciliği modellerini gördükleri işlevlere göre,

i. Sınıflandırma (classification) ve regresyon,
ii. Kümeleme (clustering),
iii. Birliktelik kuralları (association rules) ve ardışık zamanlı örüntüler (sequential patterns) olmak üzere üç ana başlık altında incelemektedir.

Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir.

---

[9] Berry&Linoff, *Data Mining Techniques*, s: 117
[10] Haldun Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", www.isletme.istanbul.edu.tr/dergi/nisan2000 (24.01.2002)

Berry ve Linoff'a göre[11] veri madenciliği modellerini dört ana başlık altında toplamak mümkündür. Bunlar; sınıflandırma modeli, tahmin modeli, kümeleme modeli ve zaman serileri modelidir.

Modelleri işlevlere göre incelediğimizde ise sınıflandırma, tahmin (prediction), öngörme (estimation), bağıntı oluşturma (affinity grouping-market basket analysis), kümeleme ve tanımlayıcı olmak üzere altı ana başlıkta toplandığını görebiliriz.

SAS Enstitüsü'nün "Data Mining Primer"[12] adlı kitapçığına göre, veri madenciliği modelleri, tahmin edici model, kümeleme modeli ve birliktelik kuralları olmak üzere üç ana başlık altında toplanmıştır.

Bu çalışmada veri madenciliği modelleri genel olarak, tahmin edici model, kümeleme modeli ve tanımlayıcı model olmak üzere üç ana başlık altında incelenecektir. Kümeleme modeli bir tür tanımlayıcı model olmasına rağmen, çalışmamızda kümeleme modelinin aynı zamanda tahmin süreçlerinde bir başlangıç aşaması konumunda olması bu modelin ayrı konumlandırılması gereğini ortaya çıkarmıştır. Bu modelleri gördükleri işlevlere göre ise,

i. Denetimli sınıflandırma (supervised classification) ve regresyon
ii. Kümeleme (denetimsiz sınıflandırma- unsupervised classification)
iii. Birliktelik kuralları ya da bağıntı oluşturma (affinity grouping)

olarak üç genel başlık altında toplamaya çalışacağız.

Bu modellerden denetimli sınıflandırma ve regresyon tahmin edici, kümeleme adından anlaşılacağı üzere kümeleme modeli, birliktelik kuralları ise tanımlayıcı modeller olarak tanımlanmıştır.

Denetimli sınıflandırma, veri madenciliğinin en sık kullanılan modellerindendir. Yaşadığımız dünyayı daha iyi anlayabilmek için nesneleri, konuları ve problemleri

---

[11] Berry&Linoff, *Data Mining Techniques*, s: 116-117
[12] William J. Potts, *Data Mining Primer: Overview of Applications and Methods*, USA: SAS Ins. ,1998, s:23

sınıflandırma ya da kategorize etme yoluna gideriz.[13] Sınıflandırma, yeni sunulan verilerin özelliklerini inceleyerek bu verileri daha önceden belirlenmiş (predefined) sınıflardan birine dahil etme işlemidir. Sınıflandırmanın amacı belirli bir sınıfa ait olmayan verileri sınıflandırmak amacıyla bir model oluşturmaktır. Sınıflandırma ve regresyon modelleri arasındaki en temel fark, tahmin edilecek olan bağımlı değişkenin kategorik veya sürekli değer olması ile bağlantılıdır.

Daha sonraki bölümlerde de inceleyeceğimiz üzere veri madenciliği tekniklerinden karar ağaçları (decision trees), bellek tabanlı yöntemler (memory based reasoning), yapay sinir ağları (artificial neural networks) ve lojistik regresyon denetimli sınıflandırma ve regresyon modellerinde kullanılan başlıca veri madenciliği tekniklerindendir.

Kümeleme modeli, heterojen bir popülasyonu homojen alt gruplara ayırma görevini yerine getiren bir veri madenciliği modelidir. Kümeleme bir tür sınıflandırma olmasına rağmen, kümeleme modelini denetimli sınıflandırmadan ayıran en belirgin özellik önceden belirlenmiş sınıfların olmamasıdır. Bu nedenle kümeleme modelinin diğer bir adı denetimsiz sınıflandırmadır.[14] Denetimli sınıflandırmada, önceden belirlenmiş sınıflara özelliklerine göre birbirinden ayırt ettiğimiz verileri yerleştirme işlemini gerçekleştiriyorduk. Kümelemede ise önceden belirlenmiş kümeler ya da sınıflar bulunmamaktadır. Kümelemede kaç küme oluşturulacağı analizin amacına göre değişebilir. Denetimsiz sınıflandırma, denetimli sınıflandırma ve regresyon problemlerini çözmede kullanılan etkili bir basamaktır. Örneğin pazar bölümlendirmesi yaparken ya da " müşteriler ne tür bir promosyon aktivitesine en iyi şekilde cevap verir?" sorusuna yanıt ararken ilk basamak müşterileri alım davranışları ve alışkanlıklarına göre kümelemek olacaktır.

Birliktelik Kuralları ya da bağıntı oluşturmada, bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından

[13] Berry&Linoff, *Data Mining Techniques*, s: 52

[14] Potts, age, s:23

biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları, pazarlama amaçlı olarak pazar sepet analizi (market basket analysis) adı altında veri madenciliğinde yaygın olarak kullanılmaktadır.[15]

1.3.2. Algoritma Kavramı

Veri madenciliği algoritması, veri madenciliği tekniğinden farklı kullanılmaktadır. Teknik, elimizdeki veriden anlamlı bilgiler çıkarmada kullanılan bir kavramdır. Algoritma ise ilgili tekniği oluşturma aşamasında yapılan adım-adım detaylandırmadır. Algortima kavramı, bilgisayar bilimleri için temel olan bir kavramdır. Algoritma, belli bir görevi yerine getirmek amacıyla takip edilen açıklamalardan oluşan sonlu sayıda elemena sahip olan kümedir. Örneğin kümeleme bir tekniktir. Gauss K-ortalamalar (Gaussian k-means), basit K-ortalamalar (simple k-means) ve kendini düzenleyen haritalar (self-organizing maps -SOMs) kümeleme tekniğinde kullanılan algoritmalardır.

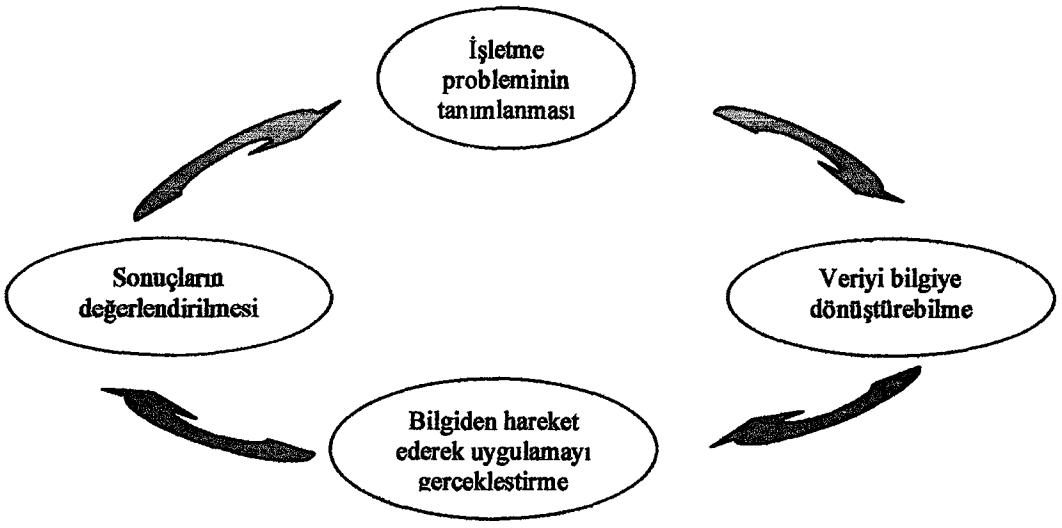
1.3. Veri Madenciliğinin Hayati Döngüsü (Virtuous Cycle of Data Mining)

Veri madenciliğin ana amacı milyonlarca baytlık veri içerisinde saklanmış olan ilginç örüntüleri bulmaktır. Fakat yalnızca ilginç örüntüleri bulmak yeterli olmayabilir. Şekil 1.1'de görüldüğü gibi, bu örüntüler üzerinden hareket ederek, veriyi bilgiye dönüştürmeliyiz, bu bilgiden hareketle bazı işletme kararları alıp bu kararlardan işletme problemlerini çözecek sonuçlara ulaşmalıyız.[16] Bu duruma özetle veri madenciliğinin hayati döngüsü diyoruz. Veri madenciliğinin bu amacına ulaşabilmesi için pazarlama, satış, müşteri destek hizmetleri, ürün dizaynı ve envanter kontrolü gibi süreçlerle işbirliği içine girmesi gerekmektedir.

---

[15] Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", age, s:5

[16] Berry&Linoff, *Data Mining Techniques*, s: 23



Şekil 1.1 Veri Madenciliğinin Hayati Döngüsü

Veri madenciliğinin hayati döngüsü, veri madenciliğini diğer işletme süreçleri ile bütünleştiren bir yapıdadır. Şekil 1.1'de ayrıntılı olarak görüldüğü gibi,

i. İşletme problemini tanımlama,

ii. Veriyi bilgiye dönüştürme (model oluşturma süreci) ve modelin değerlendirilmesi,

iii. İşletme uygulamasını gerçekleştirme,

iv. Uygulama sonuçların değerlendirilmesi , başarılı bir veri madenciliği sürecinde mutlaka izlenmesi gereken temel aşamalardır. Bu adımları kısaca inceleyelim.

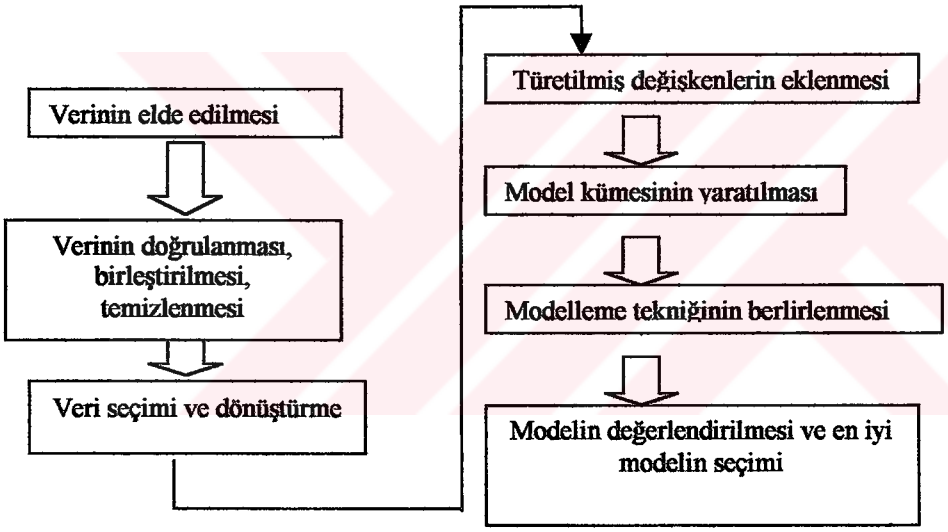
1.4.1. İşletme Probleminin Tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçülebileceği tanımlanmalıdır. Ayrıca yanlı

tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.[17]

1.4.2.Veriyi Bilgiye Dönüştürme ve Modelin Değerlendirilmesi

Veriyi bilgiye dönüştürme ve modelin değerlendirilmesi veri madenciliğinin hayati döngüsündeki en önemli aşamadır. Bu kısımda bir veri madenciliği modelinin nasıl oluşturulduğuna dair genel bir taslak verilecektir. Şekil 1.2, veriyi bilgiye dönüştürme sürecinin (veri madenciliği modeli oluşturma süreci) aşamalarını göstermektedir.



Şekil 1.2. Veriyi Bilgiye Dönüştürme Süreci (Veri madenciliği modeli oluşturma süreci)

1.4.2.1. Verinin Elde Edilmesi

Veri madenciliği modeli oluşturma sürecinde ilk adım verinin elde edilmesidir. Verinin işletme problemlerini çözmek için gerekli olan bilgiyi barındırması gerekir. Dolayısıyla verinin toplanacağı kaynakların önceden belirlenmesi ve bu kaynakların

[17] Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", age, s:6

güvenilirlikleri konusunda emin olunması ileriki safhalarda problemle karşılaşılma riskini önemli ölçüde azaltır.

1.4.2.2. Verinin Doğrulanması

Bir sonraki aşama verinin doğrulanması, birleştirilmesi ve temizlenmesidir. Veri madenciliği çalışmasının sonuçları büyük ölçüde verinin ne kadar sağlıklı olduğuna bağlıdır. Verinin toplanması aşamasında birden fazla kaynaktan elde edilen veride uyumsuzlukların olması olası bir durumdur. Uyumsuzlukların çoğu, faklı kodlama ve farklı ölçü birimlerinin kullanılmasından kaynaklanır. Bunun yanı sıra, verideki eksik değerler de (missing values) gözden geçirilmelidir. Veri girme aşamasında yanlış girilen değerler olabileceğinden bu değerlerin göz ardı edilmesi analiz sonuçlarını fazlasıyla değiştirebilir. Bu nedenle, aykırı değerlerin (outlier) gözden geçirilip, atılmasıyla beraber analizde yapabileceği değişiklikler hesaplandıktan sonra uygun görüldüğü takdirde veri tabanından silinmesi gerekir. Bütün bu geçerlemeler yapıldıktan sonra tüm veri tek bir veri tabanında düzenli bir şekilde tutulmalıdır.

1.4.2.3. Veri Seçimi ve Dönüştürme

Veri seçimi ve dönüştürme aşamasında, temizlenen veride gereksiz görülen kolonlar elenir. Veri madenciliğinde, çok fazla sayıda verinin analizi söz konusu olduğundan gereksiz kolonların varlığı, analizi yavaşlatmaktadır. Gereksiz görülen kolonların elenmesi ve bu eleme işleminin analizci tarafından yapılması, sürecin daha sağlıklı işlemesini sağlar. Veri madenciliği algoritmaları veri üzerinde belirli satırlar üzerinde çalışırlar. Bu nedenle, modellenecek verinin seviyesinin doğru belirlenmesi gerekir. Literatürde[18] bu duruma verinin özetlenmesi (granularity) diyoruz. Verinin özetlenmesi oldukça zaman alıcı bir işlemdir. Bu işlemi kolaylaştıran programlar SAS, SPSS, Ab Initio ve PERL gibi yazılımlarda mevcuttur.

---

[18] Berry&Linoff, *Mastering Data Mining: The Art and Science of Customer Relationship Management*, USA: John Wiley&Sons, Inc., 2000, s:51

1.4.2.4. Türetilmiş Değişkenlerin Eklenmesi

Türetilmiş değişkenlerin eklenmesi ise, verideki bilgiler kullanılarak faydalı yeni bilgilerin elde edilmesi ve bunların veriye eklenmesidir. Yaş ve cinsiyet bilgilerinin geçmiş karlılık değerleriyle beraber kullanılması karlılığın demografik özelliklere göre belirlenmesini sağlar. Analiz aşamasında üç bilgi yerine tek bir bilginin kullanılması analizin hızını arttırabilir.

1.4.2.5. Model Kümesinin Yaratılması

Model kümesinin yaratılması veri madenciliği modelinin oluşturulmasında çok önemli bir aşamadır. Model kuruluş süreci denetimli (supervised) ve denetimsiz (unsupervised) öğrenimin kullanıldığı modellere göre faklılık göstermektedir. Örnekten öğrenme olarak olarak da isimlendirilen denetimli öğrenimde bir denetçi tarafından ilgili sınıflar önceden belirlenen kritere göre ayrılarak, her sınıflayıcı için çeşitli örnekler verilir.

Sistemin amacı verilen örneklerden hareket ederek herbir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir. Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir. Denetimsiz öğrenimde, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır. Denetimli öğrenimde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi diğer bir kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenimi eğitim kümesi kullanılarak gerçekleştirilir.[19] Sürecin bu kısmında karşımıza bazı problemler çıkabilir. Bunlardan biri aşırı uygunluk problemidir. Oluşturdugumuz model eğitim kümesini ezberlemiş olursa genelleştirme yapmakta zorlanırız. Bu da yapacağımız tahminlerin

[19] Akpınar, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", age, s:7

dogru çıkmaması anlamına gelir. Oluşturdugumuz modeli en iyileştirebilmek (optimisation) için test kümesini kullanmalıyız. Test kümesini kullanmakla, egitim kümesinin yapısal özelliklerine tamamen bağımlı olan kuralları ve ilişkileri ortadan kaldırmış oluruz. Oluşturduğumuz modelin performansı konusunda emin olmak için, daha önceden sınıflandırdığımız veri üzerinde bir test yaparız ve bunun sonucunda oluşan kümeye de değerlendirme kümesi deriz.

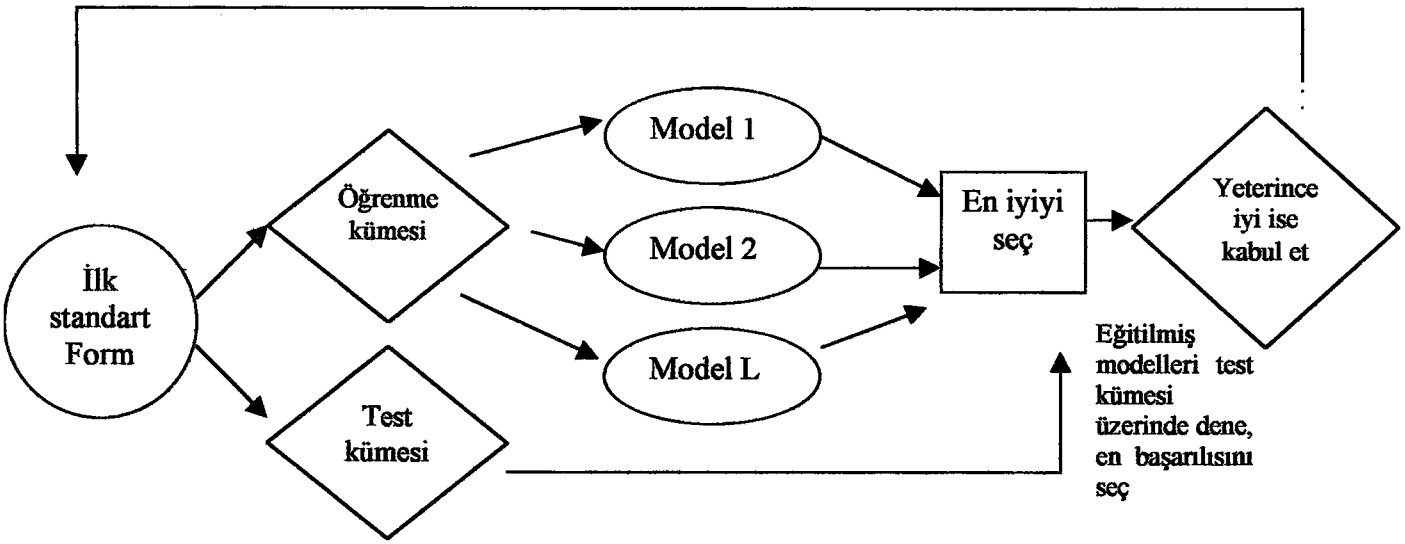
Denetimsiz öğrenmeyi verideki ilişkiyi *tanımlayabilmek* (description) için, denetimli öğrenmeyi ise bu ilişkiyi *açıklayabilmek* (explanation) için kullanabileceğimizi söyleyebiliriz.

1.4.2.6. Modelleme Tekniğinin Belirlenmesi

Her uygulamada kullanılanabilecek birden çok modelleme tekniği vardır ve önceden hangisinin en başarılı olacağını kestirmek olası değildir. Bu yüzden eğitim kümesi üzerinde L değişik modelleme tekniği kullanılarak L tane model oluşturulur. Sonra bu L model test kümesi üzerinde denenerek en başarılı olanı, yani test kümesi üzerindeki tahmin başarısı en yüksek olanı seçilir. Veri madenciliği çalışmalarında kullanılan metodoloji şekil 1.3'te verilmiştir.

Eğer bu en iyi model yeterince başarılıysa kullanılır, aksi takdirde başa dönerek çalışma tekrarlanır. Tekrar sırasında başarısız olan örnekler incelenerek bunlar üzerindeki başarının nasıl arttırılabileceği araştırılır. Örneğin standart forma yeni alanlar ekleyerek programa verilen bilgi arttırılabilir veya olan bilgi değişik bir şekilde kodlanabilir veya amaç daha değişik bir şekilde tanımlanabilir.[20]

[20] Alpaydın, age, s:5



Şekil 1.3 Veri madenciliği çalışmasında kullanılan metodoloji.

1.4.2.7. Modelin Değerlenmesi ve En iyi Modelin Seçimi

Veri madenciliği; verinin toplanması, analize hazırlanması, yazılım altyapısının oluşturulması, problemin formülasyonu, modelin kurulması ve analizi gibi oldukça ağır adımlardan geçen işlemlerden oluştuğundan işletmeler için pahalı bir süreçtir. Analize başlamadan önce, sonuçların harcanan paraya, zamana ve çabaya değecek kadar etkin olup olmadığından emin olmamız gerekir.

Asansör (lift) oranı ve grafiği, bir modelin sağladığı faydanın değerlendirilmesinde ve modellerin performanslarının karşılaştırılmasında tercih edilen bir kavramdır. Asansörün matematiksel olarak ifadesi aşağıdaki gibidir.[21]

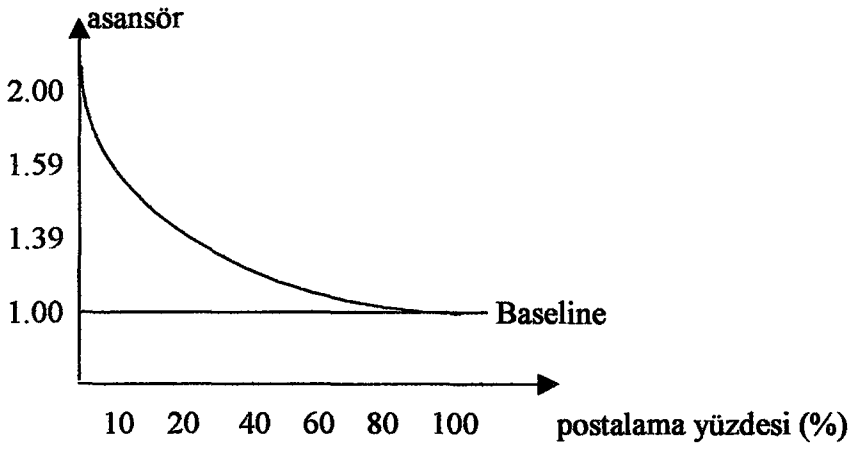
$$\text{Asansör} = \frac{P(\text{sınıf}(t) / \text{örneklem})}{P(\text{sınıf}(t) / \text{anakütle})}$$

[21] Berry&Linoff, *Data Mining Techniques*, age, s:107

Bu kavramı daha iyi anlayabilmek için, konuyu bir örnekle açabiliriz. Asansör terimi doğrudan pazarlama sektöründen gelen bir kavram olduğundan, bu alanda çok kullanılan bir uygulamayla konuya açıklık kazandırılabilir. Yapılacak doğrudan pazarlama kampanyasında, postalamaya en çok yanıt verme ihtimali olan müşterileri tahmin edebileceğimiz bir model kuracağımızı varsayalım. Başlangıçta, önceden sınıflandırılmış (preclassified) ve eğitilmiş veri kümesini ve test kümesini kullanarak bir model inşa etmemiz gerekir. Daha sonra da değerlendirme veri kümesini kullanarak modelin asansörü hesaplanır. Sınıflandırmada veri "cevaplama ihtimali var" ve "cevaplama ihtimali yok" şeklinde işaretlenir.

Bu işlem her zaman tam olarak hedefine ulaşamayabilir. Fakat modeli ne kadar iyi oluşturabilirsek cevap verme ihtimali yüksek olan müşterileri tahmin etme olasılığımız da o kadar yüksek olur. Genel müşteri listesinden "cevap verme ihtimali var" şeklinde işaretlediğimiz müşterilerin kümesi tahmin modeli kullanarak oluşturduğumuz örneklemi oluşturmaktadır. Eğer değerlendirme kümemizde gerçekte postalamaya cevap verecek olan insanların % 10'u bulunmaktaysa ve örneklemimizde postalamaya yanıt verecek insanların %20'sini içermekteyse modelimizin asansörü 2'dir (20/10=2).

Bir modelin asansörünün ne kadar yüksek olduğu o modelin performansının da çok iyi olabileceği anlamına gelmez. Şekil 1.4' de bir asansör grafiği örneği verilmektedir. Postalama yapacağımız müşterilerin listesi kabardıkça örneklem hacmimiz artacağından oluşturacağımız modelin asansörü de azalacaktır.



Şekil 1.4. Asansör grafiği

Kurulan modelin değerinin belirlenmesinde kullanılan diğer bir ölçü, model tarafından önerilen uygulamadan elde edilecek olan kazancın bu uygulamanın gerçekleştirilmesi için katlanılacak maliyete bölünmesi ile elde edilecek olan yatırımın geri dönüşüdür (ROI- return of investment).

1.4.3. İşletme Uygulamasını Gerçekleştirme

İşletmeler, kurulan ve geçerliliği kabul edilen modelden anlamlı uygulamalar oluşturabilir. Cevap verme ihtimali yüksek müşterileri hedef alan bir doğrudan pazarlama kampanyası, ek kart alma ihtimali yüksek olan müşterilere yapılacak olan promosyon, ait oldukları risk grupları belirlenen müşterilerin yeni kredi kart limitlerinin tespiti, şirketlerin siparişlerini önceden tahmin edecek olan bir uygulamanın veya sigorta poliçelerinde dolandırıcılık yapmış olanların önceden tespiti işletmeler tarafından yapılabilecek uygulamalar arasındadır.

1.4.4. Uygulama Sonuçlarının Değerlendirilmesi

Oluşturulan modele bağlı olarak işletme tarafından belirlenen uygulamanın sonuçlarının değerlendirilmesi, veri madenciliğinin hayati döngüsünün son kısmıdır. Sonuçları değerlendirmenin en etkili yolu, beklenen sonuç ile gerçekleşen sonucun

birbirlerine oranının tespit edilmesidir. Yapılacak kampanya sonunda 1000 müşterinin ankete cevap vermesi beklenirken 800 kişinin vermesi, %80 doğruluk payı ile kampanyanın gerçekleştirilmiş olduğunun bir göstergesidir.

Müşteri ilişkileri yönetiminde kullanılan, ömür boyu müşteri değerinin (life time customer value) tespit edilmesi ise uygulama sonuçlarının değerlendirilmesinde kullanılan bir başka yöntemdir. Bu yöntemde, bir işletmenin bir müşterisinin bu işletme için taşıdığı değer zamansal ve maddi açıdan hesap edilmektedir. Yapılan uygulamalar sonucunda müşterinin değerinin zamansal ve maddi açıdan değişimi, uygulamaların etkinliğinin bir göstergesidir.

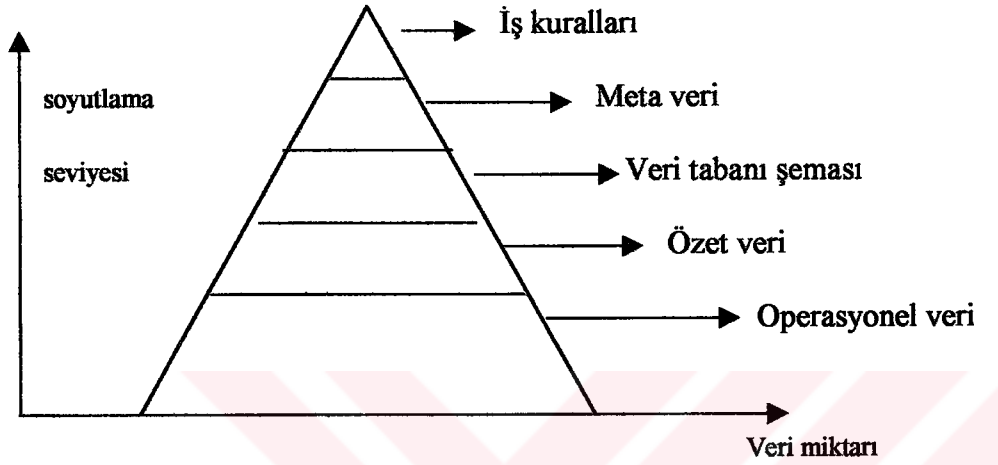
1.5. Veri Ambarı (Dataware House)

Teknolojinin hızlı bir şekilde ilerliyor olması ve iş hayatının bir anlamda otomatikleştirilmesi, birçok verinin bir arada tutulmasını kolaylaştırmıştır. Çok yüksek miktarlarda müşteri bilgilerinin bir arada tutulması "veri ambarı" (dataware house) teriminin hayatımıza girmesini sağlamıştır. Veri ambarı, rapor çıkarmak için gerekli dökümü yapabilecek kadar kolay, potansiyel müşterileri yapay sinir ağları tekniği uygulayarak bulacak kadar da sofistike bir sistemdir.

Veri ambarının asıl amacı, karar destek organlarından gelen verileri biraraya toplamaktır. İyi dizayn edilmiş ve tasarlanmış bir veri ambarı, veri madenciliği sürecini çok kolaylaştırabilir. Verinin temizlenmiş ve eksik kısımlarının tamamlanmış olması veri madenciliği sürecinin daha sağlıklı olmasını sağlar. Bu çerçevede iyi bir veri ambarı süreci, iyi bir veri madenciliği sürecini beraberinde getirir. İyi bir veri ambarı veri madenciliği için katalizör görevi görmektedir. İlişkisel veri tabanı yönetim sistemi (relational database management system- RDBMS) veri ambarının kalbidir. Daha sonra inceleyecek olduğumuz on-line analitik işleme teknolojisi ilişkisel veri tabanı yönetim sisteminin önemli bir parçasıdır.[22]

[22] Berry&Linoff, *Data Mining Techniques*, s:361

Veri ambarı ve veri madenciliğinde esas olan veridir. Verinin kalitesi ve miktarı bu iki sürecin performansını da yakından etkilemektedir. Performansı arttırmak için süreç içinde işlenen verinin kendi içinde bir hiyererşisi vardır. Şekil 1.5'te de görebileceğiniz gibi bu soyutlama süreç içinde çok önemlidir.



Şekil 1.5 Veri Hiyerarşisi.

Şekil 1.5' de görülebileceği gibi veri miktarı azaldıkça veriden çıkarılan bilgi de orantılı olarak artmaktadır. Adımlara kısa bir göz atmak gerekirse;

i. **Operasyonel veri:** Verinin en ham olduğu kısımdır. Bütün karar destek organlarından gelen verinin bir araya toplandığı ve hiçbir soyutlama işlevinin gerçekleştirilmediği ortamdır. Bu kısımdaki veri, envanter ve faturalama gibi temel işlemler için kullanılmaktadır.

ii. **Özet veri**: Özet veri kısmı, soyutlama sürecinin ilk basamağıdır. Operasyonel veriye bağlı bir soyutlama yapıldığından, sürekli güncellenmelidir. Çünkü operasyonel veri hergün değişmektedir. Kullanıcıya veriye yönelik statik bir görüş açı kazandırır.

iii. **Veri tabanı şeması:** Bir sonraki adımda, veriye yönelik bir alt-yapı şemasının hazırlandığı süreç vardır. Veri tipleri, şemalar, tablolar ve indekslerin bulunduğu kısımdır. Verinin fiziksel alt-yapısı oldukça

önemlidir. Ne tür bir veriye sahip olduğumuzu ve bu veriden ne tür bir bilgi sağlayacağımızı bu kısımda daha rahat anlayabiliriz.

iv. **Meta veri:** Daha sonraki adım olan meta-veri ise fiziksel alt yapıdan farklı olarak işletme terimleriyle açıklanmaya çalışılmaktadır. Ayrıca verinin kendi içinde ilişkilendirilmesi de bu adımda yapılmaktadır. Bu yüzden, verideki değişimlere çok duyarlı bir adımdır.

v. **İş kuralları:** İş kuralları ise veri madenciliği sürecine geçilmeye başlandığı adımdır. Bu adımda ilişkilerin nedenlerinin ve sonuçlarının çıkarımı yapılır ve anlamlı kurallar üretilir. Veri madenciliği tekniklerinden sepet analizi ve karar ağaçları bu adımın amacına benzer bir süreç yapısı göstermektedir.

1.5.1. Veri Ambarının Yapısı

Birçok kaynakta bulunan veriyi biraraya toplayarak veri madenciliği için katalizör görevi gören veri ambarında, verinin akışını ve son kullanıcıya kadar ulaşması sürecini yakından inceleyecek olursak, veri ambarının çok katmanlı yapısıyla karşılaşırız. Bu çok katmanlı yapıyı kısaca özetleyecek olursak:[23]

i. **Kaynak sistemler**: Verinin ilk elden toplandığı ve soyutlama seviyesinin en düşük olduğu kısımdır. Operasyonel anlamda yararlanılan verinin karar destek için kullanılması söz konusu değildir.

ii. **Veri nakli ve temizlenmesi:** Bu kısımda veriyi kaynak sistemlerden çıkararak veri ambarı ve analiz ortamına nakleden yazılımlar kullanılır.

[23] Berry&Linoff, *Data Mining Techniques*, s:369

iii. **Merkezi depo:** Veri ambarının teknik olarak en gelişmiş kısmıdır. Veriyi içinde bulunduran çok büyük bir veri tabanıdır.

iv. **Meta veri:** Verinin soyutlanması konusunda incelediğimiz gibi, meta veri verinin fiziksel alt yapısını hazırlar. Analiz için gerekli olan kısımların ön plana çıkarılmasına ve indeks, tablo, alan sayılarının belirlenmesine çalışılır. Veriye kendisi hakkında bilgi sağlar. Çoğu zaman veri ambarı çerçevesinde göz ardı edilen bir konudur.

v. **Datamartlar:** Bir işletmede aynı anda farklı bilgilere ihtiyaç duyan insanlar olacaktır. Verinin tümü veri ambarında olduğuna göre aynı anda bu veri ambarından farklı bilgiler sağlamak mümkün mü? Bu sorunun cevabı datamart'dadır. Datamartlar bir departman için gerekli olan bilgiyi merkezileştirme özelliğine sahip bir sistemdir.

vi. **Operasyonel geri besleme:** Bu noktaya kadar olan, veri işleme sonuçlarının geri besleme olarak operasyonel sisteme verilmesi sürecidir. Veri madenciliğinin hayati döngüsünü tamamlama yeteneğine sahip bir süreçtir. Bu yüzden oldukça önemlidir.

vii. **Son kullanıcı:** Veri ambarının yapısı içindeki en önemli kısımdır. Son kullanıcıdan amaç analizciler, uygulama geliştiriciler ve işletmecilerdir.

1.5.2. Veri Madenciliği ve Veri Ambarı:

Son yıllarda hemen hemen her alanda veri ambarları kullanılmaya başlanmış bulunmaktadır. Günümüzde hipermarket satışlarından bankacılığa, astronomiden fiziğe birçok alanda büyük veri tabanları kullanılmaktadır. Veri madenciliğinin kaynak olarak

değerlendirdiği alan (maden) veri ambarlarıdır. Veri madenciliği bilginin veri ambarlarından çekilip çıkarılacağı araçlar kümesini sağlar.[24]

Verinin bir arada toplu bir şekilde bulunduğu veri ambarında hareket edilebilir bilgi üretmek oldukça zordur. Raporlama ve faturalama gibi faaliyetleri kolaylaştıran veri ambarı, bilgi çıkarımı konusunda çok etkili değildir. Fakat bu konuda, veri madenciliğine yardımcı olmaktadır. Verinin bir arada, temizlenmiş olarak bulunması ve veri madenciliğinin hayati döngüsünü tamamlayıcı özelliği, veri ambarının veri madenciliği için ne kadar önemli olduğunu kısaca özetlemektedir.

Bilindiği gibi veri madenciliğini standart istatistiksel yöntemlere üstün kılan özelliği, çok fazla miktarda veriyle çalışılabilir olmasıdır. Standart istatistikte, anakütleden seçilen bir örneklem üzerinde çalışarak genelleştirme yapılmaya çalışılır. Fakat bu durumun gelecekteki işletme ihiyaçlarını tam olarak tahmin edememe, iş çevresindeki gelişmelere ve değişimlere cevap verememe gibi olumsuz yönleri vardır.

Bu amaçla pahalı da olsa veri madenciliği tekniklerini uygulamak daha isabetli karar verilmesini sağlar. Tüm veriyle çalışılan veri madenciliğinde, bütün veriyi sağlayan organ veri ambarıdır. Bu amaçla veri madenciliğinin, veri ambarına göründüğünden daha fazla ihtiyacı vardır.

Bunun dışında, veri ambarı veri madenciliğinde kullanılacak veriyi temizler. Veri ambarının olmaması durumunda veri madenciliği süreci gereğinden fazla uzar. Ayrıca, veri ambarı çok basit ve cevabı kullanışlı olacak soruların cevabını hızlı bir şekilde alarak veri madenciliğinin işini kolaylaştırmaktadır. Yapılan bir kampanyanın sonuçlarının başarısını belirlemek gibi geri beslemesi yüksek olan noktaları belirlemede ise veri ambarı oldukça etkilir.

---

[24] N. Gürsakal, F. Acar, "İstatistik, Veri Analizi ve Veri Madenciliği", IV. Ulusal Ekonometri ve İstatistik Sempozyumu Bildirileri, Mayıs 1999, s: 5-6

1.6. On-line Analitik İşleme (OLAP)

On-line analitik işleme (On-line Analytical Processing-OLAP) kavramı ilk olarak veri tabanı teknolojisinin babası sayılan Dr. E.F. Codd tarafından ortaya atılmıştır. On-line analitik işleme, son kullanıcıların hergün ihtiyaç duydukları rapor ve analizleri karşılayan özel bir teknoloji olarak tanımlanabilir. On-line analitik işleme, operasyonel sistemlerin ve ilişkisel veri tabanlarının tamamlanmasını sağlar. Bu teknoloji sayesinde veriler, sadece ihtiyaç duyuldukları zamanlarda değil, sürekli olarak tanımlanır.[25]

Veri ambarında veri oluşturulduktan sonra bu verinin elle veya gözle analizi yapılabilir. Bunun için on-line analitik işleme programları kullanılır. On-line analitik işleme, yöneticilerin, araştırmacıların, analizcilerin ve diğer tüm karar alma organlarının veriden hızlı, etkili, tutarlı ve interaktif bilgi çıkarımını kolaylaştıran, kullanıcıya ham veriden bilgi üretme yeteneğini kazandıran çok boyutlu bir yazılım teknolojisi kategorisidir.[26] Bu özellikleri, on-line analitik işlemeye gelecekteki eylemlere yönelik karar verme işlemlerini gerçekleştirme imkanı verir.

California'daki Symmetri Corp'den Sarah Forsman'a göre[27], on-line analitik işleme, analiz yapan kişilere ve yöneticilere, bilginin olabilecek tüm görünüşlerine, hızlı, uyumlu ve etkileşimli erişimizi sağlayarak veri hakkında görüş kazanmasına olanak tanımaktadır. On-line analitik işleme ham veriyi, kullanıcının kavrayabildiği, kurumun gerçek boyutunu yansıtan bir şekle dönüştürür. On-line analitik işleme, tek başına yararlı bilgi oluşturmaz ve bir dönüştürücü etmen gibi davranarak, ham veriyi karar verme desteğinde kullanılmak üzere yönetimsel bilgiye veya işle ilgili istihbarat bilgilerine çevirir. On-line analitik işleme'nin ortaya çıkardığı sonuçların daha da analiz edilmesi ve korelasyonunun alınmasıyla kazanılan bazı görüşler sonucunda, gerçek yararlı bilgiye dönüşüm sağlanmış olur. On-line analitik işleme, bütçe oluşturma, eylemlere dayalı masraf çıkarma, finansal performans analizleri ve finansal modelleme uygulamaları gibi

---

[25] P. Dean, "Implementing the ORACLE OLAP Applications Products- What is OLAP?", Temmuz, 1997
[26] Catherine Ma, David Chou, David Yen, "Datawarehousing, technology assessment and management", *Industrial Management & Data Systems*, cilt:100, sayı:3 (2000), s:128
[27] OLAP Council. www.on-line analitik işlemecouncil.org/research/whtpapco.html (28.01.2002) s:1

konularda kuruluşların finans bölümlerinde yaygın bir şekilde kullanılmaktadır. Ayrıca, satış analizi ve tahmin yürütme, pazar araştırması, müşteri bölümlendirmesi (customer segmentation), müşteri analizi, üretim planlaması ve hatalı ürün analizinde çokça kullanılmaktadır.

1.6.1. On-line Analitik İşleme ve Veri Madenciliği

Çoğu kişi on-line analitik işleme ve veri madenciliği kavramlarını birbirine karıştırmaktadır. Her ikisi de veri ambarı üzerinde yürütülen iki önemli fonksiyondur. Özbilginin yönetimi açısından her iki fonksiyonun da amacı, ham veri içinde gizli duran işle ilgili yararlı bilgileri ortaya çıkarmaktır. On-line analitik işleme ve veri madenciliği birbirini tamamlayan öğeler olmasına rağmen, on-line analitik işleme veri madenciliğinden farklıdır.

Veri madenciliğinde amaç, kullanıcının bilgi çıkarma sürecinde katkısının olabildiğince az tutulması, işin olabildiğince otomatik olarak yapılabilmesidir. Çünkü on-line analitik işleme programlarını kullanırken bulunabilecek sonuçlar kullanıcının sormayı düşündüğü sorgularla sınırlıdır. Ama veri içinde çocuk bezi ile bira örneğindeki bağıntı gibi kullanıcının hiç aklına gelemeyecek bilgiler de olabilir. Zaten veri madenciliğinde amaç bu tip bilgileri bulabilmektir.[28] Veri madenciliğinde, veriye bağlı olarak bilgi büyük veritabanlarından çekip çıkarılır. Veri madenciliğinde, veri içinde örtülü olarak bulunan belirli bir örüntü açığa çıkarılır. Kullanıcı, ortaya çıkarılan bu olgulara bakarak, olayın önemini anlar. Bu işlemde insan etkeni oldukça önemlidir. Veri madenciliği süreci genelde özbilginin yaratılmasıyla son bulur.

Veri madenciliğinde; istatistik, matematik, makine öğrenmesi ve yapay zeka disiplinlerinden oldukça yararlanılır. IBM, veri madenciliği sürecini on-line analitik işleme sürecine üstün kılan özelliğini şu şekilde açıklamaktadır: Veri madenciliği tekniği,

[28] Alpaydın, age, s:3

"ne?" sorusunun arkasında yatan "niçin?" sorusundan başka, "başka ne?" ve "neden o?" gibi soruları yanıtlayarak on-line analitik işleme tekniğini geçmeye çalışır.[29]

Veriye birden fazla perspektifden bakma imkanı sağladığından on-line analitik işlemeyi bir küp şeklinde hayal etmek gerekir. Standart sorgulama dilinin (SQL-standart query language) saatle veya gün ile ölçülen cevaplama süresi on-line analitik işlemede dakikalarla ölçülmektedir. Bunun dışında elle analiz edebilme imkanı da sağlamaktadır. On-line analitik işleme verilen diğer bir isim çok boyutlu veri tabanıdır (multi dimensional database-MDD). İyi dizayn edilmiş bir on-line analitik işleme küpünde her kayıt sadece bir alt-kümeye girmelidir. Bu, çok boyutlu veri tabanı olmanın en önemli kuralıdır.[30]

Bunun yanında çok boyutluluk on-line analitik işlemenin en temel özelliğidir. Çok boyutlu görünümler, verinin üretildiği veya yakalandığı biçimde görülmek istenmeyip, bilginin bir iş kullanıcısı gözüyle algılanmak istenmesi çabasıdır. Örneğin kullanıcı sadece satış verisini görmek istemeyecek, aynı zamanda belirli bir ürüne veya belirli bir zaman periyoduna yönelik satış bilgilerini görmek isteyecektir. Ürün, zaman ve periyodun herbiri, satış verisinin boyutlarıdır. Kullanıcı çoğunlukla, verinin kendisine, aynı anda çeşitli boyutlarca düzenlenerek sunulmasını ister. Örneğin bir kullanıcı, geçen yıla ilişkin satışları, ürünlere, müşteriye, satış temsilcisine, dağıtım kanalına ve bölgeye göre görmek isteyebilir. On-line analitik işleme sistemleri kullanıcılara verinin çok boyutlu görünümlerini doğal olarak sunmakta ve onları karmaşık sorgu sentaksından yalıtmaktadır.[31] Son olarak, gerçek işletme problemlerini modelleyebilme yeteneği ve kullanıcıların kaynakları daha verimli kullanma imkanını sağlaması on-line analitik işlemeyi günümüzde işletmeler için vazgeçilmez kılmaktadır. Pazar taleplerinin daha hızlı cevaplanmasını sağlayan on-line analitik işleme bu sayede işletmelerin yatırımlarının geri dönüş sürelerini kısaltmaktadır.

---

[29] Türker Cambazoğlu, "Kurumlarda Yararlı BilGinin (Kowledge) Yönetimi ve İlintili Teknolojiler-18", www.bilişimrehber.com.tr/araştirma (28.01.2002)s:2-3

[30] Berry&Linoff, *Data Mining Techniques*, s: 398

[31] Cambazoğlu,age, s:2

BÖLÜM 2

VERİ MADENCİLİĞİ TEKNİKLERİ

Günümüz dünyasında uygulamalı matematiğin ve ölçüm yapmayı kolaylaştıran bilgisayar teknolojisinin ilerlemesi, zamanla artan çok miktarda verinin saklanmasını ve bu verilerden elde edilen bilgi çıkarımını kolaylaştırmıştır. Şirketlerin, iş dünyasının yoğun rekabet ortamında başarılı olabilmeleri için veri ambarlarında saklı bulunan verilerinin anlamlı bilgilere dönüşmesini sağlayan veri madenciliği tekniklerini iyi anlamaları ve ilgili problemlerin hangi tekniklerle çözüm bulacağı hakkında fikir sahibi olmaları gerekmektedir.

Veri madenciliği teknikleri genel olarak istatistiksel ve matematiksel tekniklerle, örüntü tanıma teknolojilerini beraber kullanan süreçlerden oluşmaktadır. Veri madenciliğinin ortaya çıkış sürecinde, örüntü tanıma ve sınıflandırma problemleri üzerine yoğunlaşan yapay zeka ve istatistik disiplinlerindeki gelişmelerin veri madenciliği tekniklerinin temelini oluşturduğunu görmekteyiz.[32]

Bu bölümde, veri madenciliği tekniklerinden sepet analizi tekniği (market basket analysis), bellek tabanlı yöntemler (memory based reasoning), kümeleme analizi (cluster analysis), yapay sinir ağları (artificial neural networks) ve karar ağaçlarını (decision trees) detaylı bir şekilde incelemeye ve bu incelemede aşağıda bulunan noktaların üzerinden geçerek veri madenciliği teknikleri için bir çerçeve oluşturmaya çalışacağız. Bu bölümde özetle;[33]

i. Tekniklerin avantajları ve dezavantajları
ii. Hangi tekniğin teorik ve pratik anlamda en iyi olduğunun araştırılması
iii. En çok anlamlı bilgi verebilen tekniğin seçimi
iv. Tekniklerin performanslarının karşılaştırılması, konuları incelenecektir.

---

[32] Akpınar, age, s:1-2
[33] Berry&Linoff, *Data Mining Techniques*, age, s:112

2.1. Sepet Analizi Tekniği (Market Basket Analysis)

Sepet analizinde amaç değişkenler arasındaki ilişkiyi bulmaktır. Bu ilişkilerin bilinmesi şirketin karını arttırmak için kullanılabilir. Eğer X malını alanların Y malını da çok yüksek olasılıkla aldıklarını biliyorsanız veya bir müşteri X malını alıyor ama Y malını almıyorsa o potansiyel bir Y müşterisidir.[34] Sepet analizi günlük işlemler sonucu elde edilen verilerden anlamlı bağıntılar (association) çıkarmada kullanılır. "Eğer A malını alıyorlarsa, % x ihtimalle B malını almaya da meyillidirler" şeklinde bir sonuç A malını satan bir mağaza için çok faydalı bir bilgi olabilir. Sepet analizi uygulamaları; çapraz satış (cross-selling), mağaza raflarının düzenlenmesi (layout), katalog dizaynı ve fiyatlandırma (pricing) gibi alanlarda kullanılmaktadır.

Sepet analizinde mallar arasındaki bağıntı, destek ve güven kriterleri ile hesaplanır. Destek ve güven kriterlerinin tanımlarını aşağıdaki gibi özetleyebiliriz:[35]

$$\text{Destek (support): } P(X \text{ ve} Y) = \frac{\text{X ve Y mallarını satın alan müşterilerin sayısı}}{\text{Toplam müşteri sayısı}}$$

$$\text{Güven (confidence): } P(X/Y) = \frac{P(X \text{ ve } Y)\ \{X \text{ ve} Y \text{ mallarını alan müşterilerin sayısı}\}}{P(Y)\ \{Y \text{ malını satın alan müşterilerin sayısı}\}}$$

Destek kriteri, veride mallar arasındaki bağıntının ne kadar sık olduğunu, güven kriteri ise Y malını almış olan bir kişinin hangi olasılıkla X malını alacağını söyler. İki ürünün satın alınmasındaki bağıntının önemli olması için hem destek kriterinin hem de güven kriterinin olabildiğince yüksek olması gerekir.

Sepet analizini yaparken karşımıza bazı problemler çıkabilir. Bunlardan en önemlisi, bu tekniğin analiz sonunda bulunan bağıntının rastlantısal olup olmadığını

[34] Akpınar, age, s:1-2
[35] Alpaydın, age, s: 9

anlayamamasıdır.[36] Bir örnekle duruma açıklık kazandırmak gerekirse, bir markette tuz alan müşterilerin %40'ının şeker de almakta olduğunu düşünelim. Bu bilgi ilk bakışta ilginç görünse de çok anlamlı olmayabilir. Çünkü, marketten alışveriş yapan insanların belki de %40'ı zaten şeker alıyordur, dolayısıyla tuz ve şeker arasındaki bağıntı sadece rastlantısaldır. Alışveriş yapanların sadece %15'i şeker alırken, tuz alanların %40'ının şeker almasının bilgisi analizin sonucunu çok daha farklı etkileyecektir, bu duruma ürünlerin birbirini çekmesi (attract) diyoruz. Başka bir durumda şeker alanlar tüm alışveriş verisinin %65'ini oluştururken tuz alanların sadece %40'ının şeker almasının bilgisi bu ürünlerin birbirini ittiğini (repel) gösterir. Karşılaşılan problemlerin çoğu sepet analizi tekniğinin bağıntılar arasındaki rastlantısallığın derecesini anlamıyor olmasından kaynaklanmaktadır.Özetle, sepet analizi tekniğinin güçlü ve zayıf yönlerini sıralamak gerekirse;

Güçlü yönleri;

i. Denetimsiz veri madenciliği yöntemini başarı ile uygular
ii. Anlaşılabilir sonuçlar verir.

Zayıf yönleri;

i. İleri derecede bilgisayar performansına bağımlıdır
ii. Arasında bağıntı oluşturulacak ürünlerin doğru olarak seçilmesi analiz süreci için oldukça önemlidir.

2.2. Bellek Tabanlı Yöntemler (Memory-Based Reasoning)

Bellek tabanlı yöntemler denetimli öğrenmenin kullanıldığı veri madenciliği tekniklerindendir. Bu tekniğin temel özelliği, daha önceki deneyimlerimizden faydalanarak elimizdeki problemlere benzer durumları tanımlayıp geçmiş benzer problemlere getirdiğimiz uygun çözümleri mevcut problemimize uygulamaya çalışmaktır.

---

[36] Robert Groth. *Data Mining: Building Competitive Advantages.*USA: Printice Hall, 2000, s:29-30

Bellek tabanlı yöntem tekniği (BTY tekniği), en yakın komşu algoritması (k nearest neighboorhood) olarak da adlandırılmaktadır.[37] Bunun nedeni elimizdeki kayıda (record) en yakın komşu kayıtları bularak bu komşuları sınıflandırma ve tahmin için kullanmasından kaynaklanmaktadır. BTY tekniğinin performansını belirleyen iki fonksiyon vardır. Bunlar, uzaklık ve kombinasyon fonksiyonlarıdır.

Uzaklık fonksiyonu iki kayıt arasındaki uzaklığı bulmamıza, kombinasyon fonksiyonu ise sonuçları anlamlı çözüm sunacak şekilde birleştirmemize olanak sağlar. BTY tekniğinin sözünü ettiğimiz fonksiyonları kullanmasının bir faydası her türlü veri tipi için geçerliliğinin olmasıdır. Bu tekniğin uygulamada çok avantajlı yanlarının olmasının (her türlü veriye uygulanabilirliği, adaptasyon kolaylığı) dışında, geçmiş tarihi verileri saklama maliyeti bu yöntemi oldukça pahalı bir teknik haline getirmektedir. Yeni kayıtların sınıflandırılması bu kayıtlara en yakın komşu kayıtların sistemde taranacağı anlamına geldiğinden, bu teknik yapay sinir ağları ya da karar ağaçları tekniklerinden çok daha fazla zaman alıcı olmaktadır.

Uzaklık fonksiyonu, kombinasyon fonksiyonu ve en yakın komşu sayısının belirlenmesi BTY tekniğinin iyi bir sonuç vermesinde oldukça önemlidir. Bu bilgiler ışığında BTY tekniğinin çalışma düzeneği ; eğitim kümesinin belirlenmesi, uzaklık fonksiyonunun belirlenmesi ve kombinasyon fonksiyonunun belirlenmesi aşamalarından oluşmaktadır.

2.2.1. Eğitim Kümesinin Belirlenmesi

Eğitim kümesinin belirlenmesi, genel olarak veri madenciliği sürecinin başında gerçekleştirilen bir aşamadır. Verinin bir kısmı eğitilmek üzere, bir kısmı doğrulanmak üzere kalan kısmı ise test edilmek üzere ayrılmaktadır.

---

[37] Estelle Brand ve Rob Gerritsen, "Naïve- Bayes and Nearest Neighbor", *DDMS- Data Mining Solutions Supplement*, (www. dbmsmag.com), s: 5, (20.03.2002)

Az sayıda verinin eğitilmesi oluşacak modelin veriyi ezberlemesini sağlayacağından eğitilecek veri sayısının kabul edilebilir bir seviyede belirlenmesi gerekmektedir. BTY tekniğinin verimliliği kullandığımız eğitim kümesinin performansının ne kadar yüksek olduğuyla çok yakından ilintilidir.[38]

2.2.2. Uzaklık Fonksiyonunun Belirlenmesi

Uzaklık fonksiyonu, BTY tekniğinin benzerliği ölçmek için kullandığı bir yoldur. Matematiksel anlamda A ve B arasındaki uzaklık kavramının dört özelliği mevcuttur:[39]

i. İyi tanımlanmış olmak (well defined) A ve B arasındaki uzaklık her zaman pozitif olmak zorundadır. $d(A, B) \geq 0$

ii. Birim özelliği: Bir noktanın kendine uzaklığı her zaman 0'dır. $d(A,A) = 0$

iii. Komütatiflik (değişme özelliği): Yön uzaklık belirlemede etkili değildir. A'dan B'ye uzaklık B'den A'ya uzaklıkla aynıdır. $d(A, B)=d(B,A)$

iv. Üçgensel eşitsizlik: A ve B arasındaki bir C noktasından geçmek uzaklığı azaltmaz. $d(A,B) \leq d(A,C)+d(C,B)$

İlk özellik olan iyi tanımlanmış olmak, mutlaka her kayıdın veri tabanında bir komşusunun bulunduğunu söyler. Birim özelliği ise elimizdeki bir kayıda en fazla benzeyen kayıdın orjinal kayıdın kendisi olduğunu anlatmaktadır. Son özellik ise veri tabanına yeni bir kayıdın eklenmesinin var olan kayıdın daha da yakınlaşacağı anlamını taşımadığını ifade etmektedir.

Sayısal alanlarda en çok kullanılan üç tür uzaklık fonksiyonu bulunmaktadır.[40]

i. Mutlak farklılık : $|A-B|$

ii. Farkların karesi: $(A-B)^2$

iii. Normalize mutlak farlılık: $|A-B| /$ (maksimum uzaklık)

---

[38] Tom Mitchell, Machine Learning, Singapore: Mcgrow Hill, 1997, s: 231
[39] Berry & Linoff, *Data Mining Techniques*, age, s: 171
[40] Berry & Linoff, *Data Mining Techniques*, age, s: 173

Müşterilerin yaş ve ücret bilgileri sayısal büyüklük olarak birbirlerinden çok farklı olduğu için (yaş: 20 ücret: 400 milyon gibi) normalize mutlak farlılık kullanmak daha doğru olur.[41] Eğer elimizde müşterilere ait birden fazla değişken varsa ve bu müşterilerin birbirlerine olan uzaklıklarını ölçmek istiyorsak, müşterilerin benzer bilgileri arasındaki (yaş-yaş, ücret-ücret gibi) uzaklığı ölçtükten sonra bu uzaklıkları birleştirmemiz (merge) gerekmektedir. Birleştirme işlemi üç farklı yolla yapılmaktadır.[42]

i. Toplam: d sum(A,B)= dx (A,B)+dy (A,B)
ii. Normalize toplam: dnorm (A,B)= dsum (A,B) / max dsum (A,B)
iii. Öklid uzaklığı: deuclid= [dx (A,B) ^{2} + dy (A,B) ^{2}]½

2.2.3. Kombinasyon Fonksiyonlarının Belirlenmesi

Kombinasyon fonksiyonları içinde en temel olanı k-yakın komşu algoritmasıdır.Bu algoritma ilk olarak komşu sayısını belirlemeyle sürece başlar. Komşu sayısını belirlemede en klasik yol, veri tabanındaki müşterilerin kategori sayılarına bir eklemektir. Veri tabanında bulunan kategori sayısı arttıkça k yakın komşu algoritmasının çalışması çok daha verimli olur. BTY tekniğinin güçlü ve zayıf yanlarını sıralamak gerekirse: Güçlü yanları;

i. Kolay anlaşılır sonuçlar üretmesi
iii. Her türlü veri tipine uyarlanabilir olması
iv. Eğitim kümesinin büyümesinin peformansını düşürmemesidir.

Zayıf yanları;

i. Sınıflandırma ve tahmin görevlerini yerine getirme aşamasında diğer tekniklere nazaran oldukça maliyetli olması
ii. Sonuçların performansının uzaklık ve kombinasyon fonksiyonları ile komşu sayısının seçimine oldukça sıkı derecede bağlı olmasıdır.

---

[41] Bu duruma literatürde verinin standardize edilmesi (data standardization) diyoruz.

[42] Berry & Linoff, *Data Mining Techniques*, age, s: 174

2.3. Kümeleme Analizi (Clustering)

Kümeleme analizi ilk olarak Tryon tarafından kullanılmıştır.[43] Denetimsiz öğrenmenin kullanıldığı veri madenciliği ve çok değişkenli sınıflandırma tekniklerindendir. Daha önceki tekniklerde olduğu gibi, önceden sınıflandırılmış ve eğitilmiş bir veri kümesi yoktur, buna ek olarak bağımsız ve bağımlı değişken gibi bir ayrımda söz konusu değildir. Kümeleme analizinde yapılan birbirine çok benzeyen nesneleri veya bireyleri aynı gruba yerleştirmektir. Diğer bir deyişle, kümeleme analizi bilinen istatistik testlerinden farklı olarak "nesneleri uygun kümelere koyma" işlevini yerine getirecek algoritmaların toplanmış halidir.

Kümeleme analizinde amaç birbirine en çok benzeyen nesneleri aynı grupta toplamaktır. Benzemekten amaç geometrik anlamda uzaklık olarak birbirine en yakın nesnelerin seçilmesidir. Bu yüzden nesnelerin sayısal değerler (metric data- quantitative data) olması gerekir. Bu noktada değişkenlerin dörde ayrıldığını eklemekte fayda vardır. Bunlar kısaca:[44]

i. Kategorik değişkenler (nominal- categorical variables): Bu değişkenler arasında sadece birbirine benzemez durumu söz konusudur. Sıralama mümkün değildir. (örneğin siyah>beyaz durumu söz konusu değildir.)

ii. Sıralama değişkenleri: (ordinals) X>Y şeklinde bir sıralama yapmak mümkündür. Ama büyüklüğün ne kadar olduğu belli değildir. (X-Y bulunamaz)

iii. Aralık ölçekli (interval) değişkenler: İki nokta arasındaki uzaklığı bulabiliriz. Fakat bu tür değişkenlerde gerçek "0" değeri yoktur.

iv. Oran ölçekli (ratio) değişkenler: Anlamlı "0" noktasının bulunduğu, her türlü dört işleme açık değişkenler toplıluğudur. (örneğin; 20 yaşında biri 10 yaşında olan birinin iki katı yaştadır denilebilir.)

---

[43] STAT Softinc "Cluster Analysis", www.statsoftinc.com/textbook/stcluan.html (28.08.2001), s:1

[44] Berry & Linoff, *Data Mining Techniques*, age, s: 196-197

Kategorik ve sıralama değişkenlerinin, matematiksel hesaplamaların yapılabileceği sayısal değerlere dönüştürülmesi şarttır.

Kümeleme analizinde, diğer çok değişkenli istatistik analizlerinde önemli olan verilerin normalliği varsayımı prensipte kalmakta, uzaklık değerlerinin normalliği yeterli görülmektedir.[45] Yukarıda da bahsettiğimiz gibi benzerlik ölçüsü olarak uzaklık kriteri esas alınmaktadır. İki nokta arasında uzaklığın ölçülmesinin en çok kullanılan yolu öklid uzaklığıdır (deuclid= [dx (A,B) ² + dy (A,B) ²]½).[46] Bu methodun avantajlı olduğu bir nokta, öklid uzaklığı iki nesne arasındaki uzaklığı ölçümlediğinden başka bir nesne tarafından (outlier-aykırı değer) sonucun etkilenmesi söz konusu değildir. Bunun dışında öklid uzaklığı ölçeklerden çok etkilenir. Birim olarak dakika olan bir değişkenin birimi saniye olarak değiştirilirse sonuç da bu duruma bağlı olarak çok değişecektir. Öklid uzaklığı dışında kulanılan uzaklık türleri; manhanttan uzaklığı, chebychev uzaklığı, percent disaggrement, city block, kareli öklid, mahalanobis uzaklığı olarak sıralanabilir.[47]

Kümeleme analizinde oldukça önemli olan iki nokta daha vardır: Ölçekleme (scaling) ve ağırlıklandırma (weighting). Normalde geometrik anlamda iki nokta arasındaki uzaklığı ölçerken A ve B gibi iki nesnenin aynı birimlere sahip olduğunu düşünürüz. Fakat A nesnesi $, B nesnesi de yıl cinsinden olursa ne tür bir işlem yapmamız gerekir? Ortak bir ölçü aracı olmadığına göre bütün birimlerin belli bir aralığa gönderme (mapping) yapılması gerecektir. Bu sayede oranlama yapmak kolaylaşır. Birimlerin bu şekilde belli bir aralığa ([-1,1] gibi) gönderilmesine ölçekleme diyoruz. Genelde istaistikte z değeri (z-score) olarak bilinen değer, bu oranlamanın yapılmasında kullanılan bir araçtır.[48]

Kümeleme analizinde her değişken bizim için aynı anlamı ifade etmeyecektir. Bazı değişkenler analizimizde daha önemli bir noktada bulunacağından, değişkenler arası bir ağırlıklandırmaya gitmek anlamlı olacaktır. Örneğin bir ailenin geliri ailenin genel yaş

---

[45] H. Tatlıdil, *Uygulamalı Çok Değişkenli İstatistik Analiz*, İstanbul: Engin Yayınları
[46] Berry & Linoff, *Data Mining Techniques*, age, s: 200
[47] STAT Softinc. "Cluster Analysis", age, s:3-4

[48] Hair, Anderson, Tatham ve Black, *Multivariate Data Analysis*, New Jersey: Prentice Hall, 1998, s:486

ortalamasından çok daha önemli bir bilgiyse, ilk bilgiye daha fazla ağırlık vererek analize devam etmek daha faydalı olacaktır.[49]

Kümeleme analizinde çok önemli bir konu olan benzer nesneleri gruplandırırken ya da kümelerken hangi prosedürü kullanacağımız, seçeceğimiz kümeleme algoritmasına göre farklılık gösterecektir. Veri madenciliği ve çok değişkenli istatistik analizi yapan birçok paket program farklı kümeleme algoritması kullanmaktadır. Bu algoritmaların ortak özelliği, kümeler arası uzaklığın küme içi uzaklığa oranını maksimize etmeye çalışmasıdır (Bu oran varyans analizindeki F oranı ile yapılmaktadır). Bu çalışmada en çok kullanılan kümeleme algoritmaları iki kategoride incelenecektir: Hiyerarşik yöntemler ve hiyerarşik olmayan yöntemler.

2.3.1. Hiyerarşik Yöntemler

Hiyerarşik kümeleme yöntemleri, kümeleri ardısıra birleştirme sürecidir ve bir grup, diğeri ile bir kere birleştirilir ise daha sonraki adımlarda bir daha kesinlikle ayrılamaz.[50] Süreçte ilk başta bütün noktalar kendi kümesini oluşturur. Süreç boyunca kümeler anlamlı bir şekilde biraraya gelerek en sonda tek büyük bir küme oluştururlar. Bu nedenle bu yöntem aglomeratif-yığışma algoritmaları olarak da bilinmektedir.Bu algoritmaya göre:[51]

i. İlk basamakta benzerlik matrisi yaratılır (similarity matrix). Bu matriste bütün noktaların birbirlerine uzaklık durumları gösterilir.

ii. Matristeki en küçük değerler bulunur ve ait oldukları kümeler birleştirilir. Benzerlik matrisi bu işlemler sonucu yeniden güncellenir.

iii. Sonuçta tek bir küme kalıncaya kadar sürece devam edilir. Birleştirme anlarında hangi kümelerin birleştirildiği ve birbirlerinden ne kadar uzak oldukları not edilir. Büyüklük olarak hangi uzaklıklarda sıçramanın (jump) olduğu tespit edildikten sonra kaç küme oluşturulacağı belirlenmiş olur.

---

[49] Berry & Linoff, *Data Mining Techniques*, age, s: 205

[50] Ümit Fırat, "Kümeleme Analizi: İstihdam Sektörel Yapısı Açısından Avrupa Ülkelerinin Karşılaştırılması", Sosyal Bilimler Dergisi, Cilt 3, sayı:2 s:54

Kümelerin birleştirilmesi iki grupta incelenebilir.[52]

i. Bağlantı (linkage): Bağlatı yöntemi içinde en çok bilinenler; tek bağlantı (single linkage), tam bağlantı (complete linkage), ortalama bağlantı (average linkage) yöntemleridir.

ii. Merkezi (centroid)

2.3.1.1. Tek Bağlantı Yöntemi

Tek bağlantı yöntemi, en yakın komşuluk (nearest neigboorhood) kavramı olarak da adlandırılmaktadır. Bu yöntemde, mesafe olarak birbirine en yakın iki nesne seçildikten sonra birbirlerine bağlanarak ilk kümeyi oluşturur. Daha sonra ya oluşturulan ilk kümeye en yakın mesafedeki nesne seçilerek kümeye eklenir ya da daha yakın mesafede olan iki nesne belirlenerek başka bir küme oluşturmaları sağlanır. Bu işlem bütün kümeler birleşerek tek bir kümeye dönüşünceye kadar devam eder.[53] Tek bağlantı kümeleme yönteminin sonuçları ağaç diyagramı veya bir dendogram şeklinde grafik ile gösterilebilmektedir. Ağaçlardaki dallar, kümeleri gösterir. Dallar boğumlarda biraraya gelir.[54]

2.3.1.2. Tam Bağlantı Yöntemi

Tam bağlantı yöntemi, tek bağlantı yöntemine kümeleme kriteri dışında- en uzak mesafenin baz alınması- işleyiş açısından çok benzemektedir. Bu yüzden en uzak komşuluk (farthest neigboorhood) yaklaşımı olarak da anılmaktadır. Bu yöntem sonucunda kümeler, birbirlerinden maksimum uzaklıkta olan dolayısıyla birbirine çok az benzeyen nesnelerin biraraya gelmesiyle oluşmaktadır.

---

[51] Hair, Anderson, Tatham ve Black, age, s: 474-475
[52] Fırat, age, s:54
[53] Hair, Anderson, Tatham ve Black, age, s: 494
[54] Fırat, age, s:7

2.3.1.3. Ortalama Bağlantı Yöntemi

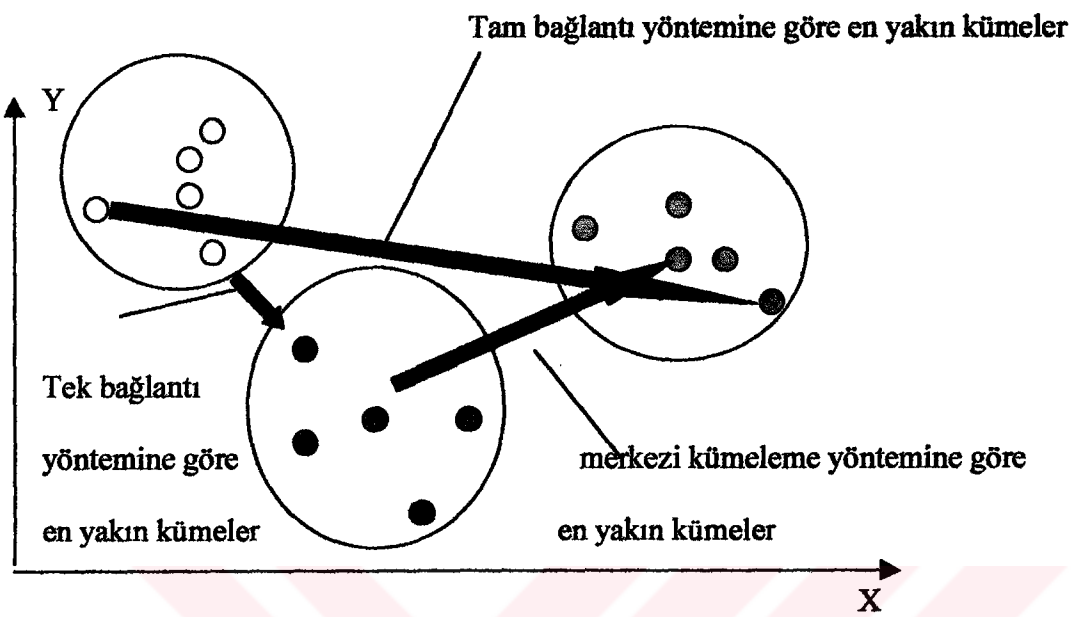
Ortalama bağlantı yöntemi işleyişe, tek ve tam bağlantı yönteminde olduğu gibi başlar. Fakat kümeleme kriteri farklılık göstermektedir. Kümeler arası bağlantı oluşturmada, bir kümedeki nesnelerin diğer kümedeki nesnelere ortalama uzaklıkları baz alınmaktadır. Bu yöntemde, aşırı değerlerden (tek ve tam bağlantı yönteminde olduğu gibi) ziyade küme içindeki bütün nesnelerin uzaklık belirlemede fonksiyonları bulunmaktadır. Bu yüzden, süreç sonunda oluşan kümelerin küme-içi varyansları çok düşüktür. Ayrıca, birbirine benzer varyansta olan kümelerin oluşması da doğal bir gelişmedir.[55] Genellikle tam bağlantı ve ortalama bağlantı konfigürasyonları benzer dendogramlar üretmektedir. Bununla birlikte herbir yöntemde uzaklık farklı tanımlanmıştır, dolayısıyla birleştirmeler farklı seviyelerde ortaya çıkabilmektedir.[56]

2.3.1.4. Merkezi Kümeleme Yöntemi

Hiyerarşik kümeleme yönteminde en son ele alacağımız yöntem, merkezi kümeleme yöntemidir. Bu yöntemde iki küme arasındaki uzaklık, kümelerin merkezleri arasındaki uzaklığa eşittir. Uzaklık ölçüsü olarak öklid ya da kareli öklid kullanılır.[57] Kümelerin merkezlerinden amaç, küme elemanlarının aritmetik ortalamasıdır (mean value). Dolayısıyla, her defasında kümelere yeni bir nesne eklendiğinde kümelerin aritmetik ortlaması değişmektedir. Hiyerarşik kümeleme yöntemleri aykırı değerlerden (outlier) en az etkilenen yöntemdir.

---

[55] Hair, Anderson, Tatham ve Black, age, s: 496
[56] Fırat, age, s:8
[57] Hair, Anderson, Tatham ve Black, age, s: 496



Şekil 2.1. Kümeler arası uzaklığı ölçen üç hiyerarşik yöntem

Yukarıdaki şekil 2.1 hiyerarşik yöntemlerden tek bağlantı, tam bağlantı ve merkezi kümeleme yöntemlerini göstermeye çalışmaktadır[58].

2.3.2. Hiyerarşik Olmayan Yöntemler (K- Ortalamalar Yöntemi)

Hiyerarşik yöntemlerin tersine, hiyerarşik olmayan yöntemler ağaç diyagramları ya da diğer bir deyişle dendogramlar kullanmazlar. Bunun yerine küme sayısı ya başta özellikle saptanır ya da yöntemin bir aşaması olarak farklı küme büyüklükleri denenerek belirlenebilir. Literatürde K-ortalamalar yöntemi (k means algorithm) en bilinen hiyerarşik olmayan kümeleme yöntemidir. Araştırmamızda, bu yöntem teorik ve pratik anlamda incelenmeye çalışılacaktır. K ortalamalar yönteminin nasıl işlediğini özetlemeye çalışırsak;[59]

[58] Berry & Linoff, *Data Mining Techniques*, age, s:209

[59] P.E. Green, *Analyzing Multivariate Data*.Hındsdale: Holt, Rinehart & Winston,1978, s:428

i. K tane kümenin merkezini (cluster seeds) oluşturacak nokta seçilir.

ii. Seçilen merkezi noktalara, uygun olan uzaklık ölçüsü kullanılarak, en yakın olan noktalar işaretlenir ve oluşturulan kümelerin sınırları belirlenir.

iii. Mevcut olan kümelerin merkezleri yeniden tayin edilir ve ikinci aşamadan itibaren süreç tekrarlanır. Kümelerin merkezlerindeki sapmalar minimum oluncaya ve kümelerin sınırları değişmeyinceye kadar sürece devam edilir.

Hiyerarşik olmayan yöntemlerde en fazla karşılaşılan problem, ilk aşamada yer alan küme merkezlerinin nasıl seçileceğine ilişkindir. K-ortalamalar algoritmasının daha sağlıklı sonuçlar vermesi için bazı metodlar geliştirilmiştir.[60]

i. Başlangıç küme merkezlerini bulmak için geliştirilen metodlar

ii. Bir sonraki merkezi hesaplamamızı kolaylaştıracak metodlar

iii. Nesneleri kümelemede kullanacağımız uzaklık ölçümü üzerine olasılık yoğunluğu (probability density) kullanmak.

Nesneleri kümelemede olasılık yoğunluğu kullanmak literatürde[61] "gaussian mixture model" olarak geçmektedir. Bu modelin kullanılmasının nedeni, K ortalamalar algoritmasının bazı dezavantajlarının olmasından kaynaklanmaktadır. Bunları özetleyecek olursak;

i. Kümeleri üst üste bindirmede (over lapping) yeteri kadar iyi bir algoritma değildir.

ii. Kümelerin merkezleri çok rahat bir şekilde aykırı değerler tarafından çekilmektedir.

iii. Her nesne bir kümenin ya içinde ya da dışında kalmaktadır. Nesnelerin gerçekten olmaları gereken kümelerin içinde olması açısından bir yenilik getirmemektedir.

[60] Berry&Linoff, *Data Mining Techniques*, s:205
[61] Berry&Linoff, *Data Mining Techniques*, s:205-206

Hiyerarşik yöntemlerin birbirlerine göre avantajlı veya dezavantajlı oldukları noktalar vardır. Fakat çoğu zaman her iki yöntemi beraber uygulamak da çok uygun bir fikir olabilir. Hiyerarşik yöntem kullanılarak küme sayısı, kümelerin merkezleri ve aykırı değerler tanımlanır ve aykırı değerler elimine edildikten sonra hiyerarşik olmayan yöntemler kullanılarak optimal küme sayısına ulaşılır. Kaç tane küme oluşturulacağı hala araştırmacılar tarafından cevabı tam olarak verilmemiş sorulardandır (stopping rule problem). Bu soruya verilebilecek bir cevap kümeleme sürecinin her aşamasında kümeler arası uzaklığı ölçerek, ilk sıçramada (first jumping) kümeleme sürecini durdurmak ve mevcut küme sayısını elde etmek olabilir. Bunun dışında bu tür sorulara cevap arayan istatistiksel testler de vardır. Point-biserial ya da likelihood ratio olarak bilinen bu testler istatistik paketlerinde (örneğin; SAS- Cubic Clustering Criterion) bulunmaktadır.[62]

Kümeler içindeki benzerliğin ölçüsü varyanstır. Küme içi varyansın düşük olması iyi bir küme olduğunun kanıtıdır. Fakat bu durum hiyerarşik kümelemede çok anlamlı olmayacaktır, bunun yerine çok güçlü kümeler çok uzun zaman da oluşur kriteri söylenebilir. Kümeleme analizinde K-ortalamalar methodu ve aglomeratif metodlar dışında kullanılan iki alternatif method daha vardır: Divisive methods (karar ağaçları tekniğinde kullanılmaktadır) ve kendini düzenleyen haritalar (self organizing maps). Kümeleme tekniğinin güçlü ve zayıf yönlerini özetleyecek olursak[63];

Güçlü yönleri;

i. Başarılı bir denetimsiz öğrenimin kullanıldığı veri madenciliği tekniği olması
ii. Her türlü veri tipinde rahatlıkla çalışabilir olması ve uygulamada sağladığı rahatlık.

Zayıf yönleri;

i. Sağlıklı sonuçların çıkmasını sağlayacak olan uygun uzaklık ölçüsünü seçme güçlüğünün bulunması

[62] Hair, Anderson, Tatham ve Black, age, s: 497
[63] Berry&Linoff, *Data Mining Techniques*, s:213-214

ii. Sonucun başlangıç parametrelerine çok bağımlı olması, süreç sonunda oluşan kümeleri yorumlama güçlüğü şeklinde sıralanabilir.

2.4. Karar Ağaçları (Decision Trees)

Karar ağaçları tekniği, çok güçlü bir sınıflandırma ve tahmin aracıdır. Diğer veri madenciliği tekniklerine nazaran çok daha anlaşılır bir dile sahiptir. Örneğin; kredi kartı başvurusunda bulunan bir müşteri için başvurusunun red edilmesinin sebebinin gelir<1 milyar ve kredi kartı sayısı<3 olması yeteri kadar açıklayıcı olacaktır. Bunun yanısıra modelin başarısı kadar, başarılı ve başarısız bir modelin nasıl çalıştığını araştırması bu tekniği diğer tekniklere göre farklı kılmaktadır. Karar ağaçları daha sonra açıklayacağımız bazı kriterlere gore farklı algoritmalar yardımıyla oluşturulabilir. Bunlardan en çok kullanılanları; CART, CHAID, C4.5 ve ID3' tür.

Karar ağaçları denetimli öğrenimin kullanıldığı veri madenciliği tekniklerindendir. Bu anlamda tahmin edilmesi gereken bir hedef değişken vardır. Hedef, kesikli (discrete) veya sürekli (continous) değerlerden oluşabilir. Kesikli değerlerden oluşuyorsa sınıflandırma ağacı (classification tree), sürekli değerlerden oluşuyorsa regresyon ağacı (regression tree) olarak adlandırılır.[64]

Karar ağaçları soru serilerinden oluşmaktadır. Genelde karar ağaçları bir kök düğümünden (root node) başlamak suretiyle süreç sonunda yaprakların (leaves) oluşumuyla sona erer.

Kök düğümünde, bir sonraki çocuk düğümlerin (child nodes) ne olacağını belirleyen başlangıç testleri bulunmaktadır. Bu testleri seçmede yardımcı olan birçok farklı algoritma vardır. Fakat bu algoritmaların amaçları genelde aynıdır: En iyi ayrımı yapacak olan testi seçmek. Bir sonraki düğümü belirleyecek testlerin uygulanması yaprak düğümlere ulaşıncaya kadar tekrarlı bir şekilde devam eder. Aynı yaprakta son bulan kayıtlar aynı şekilde sınıflandırılmış olanlardır. Kökten herbir yaprağa sadece bir tek yol

[64] Potts, *age.* s: 50

vardır. Bu yolu belirleyen kurallar aynı zamanda o yaprakta son bulan kayıtların da sınıflandırılma kuralı olmaktadır. Birçok farklı yaprak aynı sınıflandırmayı yapabilir. Fakat herbir yaprak bu sınıflandırmayı farklı nedenlerle yapmaktadır.[65]

Daha önce bir karar ağacında herbir yolun bir kural ifade ettiğini belirtmiştik. Bu kurallardan bazıları diğerlerine nazaran daha iyi olacaktır. Karar ağaçlarının performansının iyileştirilmesi açısından kötü kuralları ifade eden dalların geriye doğru budanması (prunning back) isabetli bir karar olacaktır.

Daha yakından incelediğimizde, karar ağaçları benzer kayıtların biraraya toplandığı kümeler olarak da düşünülebilir. Aynı kutuda olan kayıtlar o kutunun kuralına uyan kayıtlar olacağından aynı sınıflandırma kriterine de uymuş olacaktır. Bu noktada istatistiksel yöntemlerle karar ağaçları tekniği arasında bir fark bulunmaktadır. İstatistiksel yöntemlerde kayıtların dahil olduğu sınıfları birbirinden ayırmak için düz bir çizgi ya da bir eğri kullanılır, karar ağaçlarında ise kutucuklar bulunur. Böylelikle sınıflandırma daha verimli olacaktır.[66]

2.4.1. Karar Ağaçlarının Oluşumu

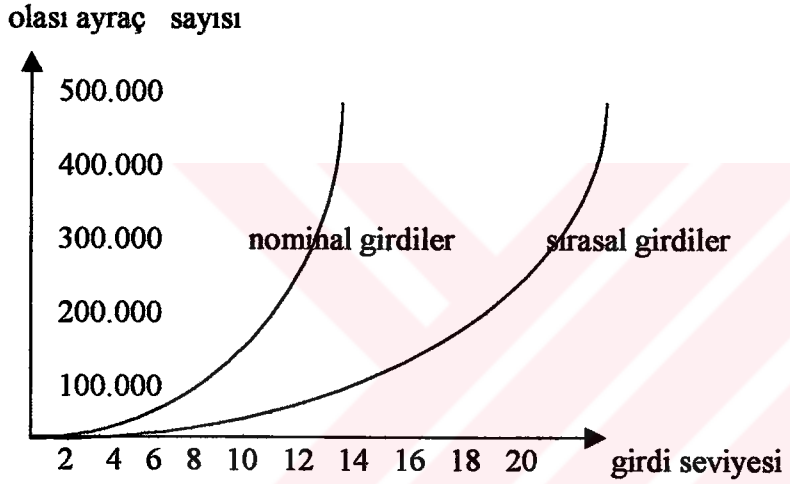
En iyi sonucu verecek karar ağacının oluşturulması için aşağıda bulunan sorulara cevap aramalıyız:

i. Hangi ayraçlar (splitter) dikkate alınmalıdır (ayraç araştırması- splitter search) ?
ii. Hangi ayraç en iyisidir (ayırma kriteri-splitting criterion) ?
iii. Ayırma işlemi nerede sona ermelidir (durma kuralı- stopping rules) ve bazı dallar budanmalı mıdır (budama kuralları-pruning rules) ?

[65] Berry&Linoff, *Data Mining Techniques*, s:244-245
[66] Berry&Linoff, *Data Mining Techniques*, s:251

2.4.1.1. Ayraç Araştırması

Şekil 2.2'de de görüldüğü gibi girdi sayısı arttıkça olası ayraç sayısı da buna bağlı olarak artmaktadır. Hiçbir ayraç algoritması bütün bölümleri derinlemesine inceleyemez. Bunun yerine olası ayraçların sayısını kısıtlayacak bazı sınırlandırmalar getirir. Bunlar ; ikili ayraç (binary split), sürekli girdilerin gruplandırılması (binning continous inputs), adım adım araştırma algoritması (stepwise search algorithm) ve örneklem (sampling) oluşturmadır.[67]



Şekil 2.2. Girdi seviyesine gore olası ayraç sayısı

Kaynak: William J. Potts, "Data Mining Primer: Overview of Applications and Methods, SAS Institute Inc., USA, 1998, s:52

2.4.1.2. Ayırma Kriteri

Hangi ayraçları inceleyeceğimizi belirledikten sonra bunlardan hangisinin en iyi ayraç olacağına karar vermeliyiz. Bu aşamada kullanılan bazı ayırma kriterleri bulunmaktadır. Sınıflandırma ağacı için kullanılan üç kriter: Ki-kare testi, Gini endeksi ve Entropidir. Genelde bu üç method da benzer sonuçlar vermektedir.

[67] Potts, age , s:52

2.4.1.3. Durma ve Budama Kuralları

Kök düğümünden başlayan bölümlendirme kendini tekrarlar bir biçimde herbir yaprakta bir kayıt oluncaya kadar devam eder. Bu noktada oluşturulan karar ağacı genelleştirme yeteneğini kaybetmiş demektir. Bu amaçla, bölümlendirme işlemi sırasında bir noktada durmak gerekir. Fakat durmamız gereken noktayı nasıl belirleyeceğiz? Bu soruyu cevaplamamızda yardımcı olacak iki yaklaşım bulunmaktadır.

i. Ağacın büyümesini engelleyecek olan ön budama kuralı (pre-prunning):

Eğer düğüm daha fazla bölümlendirilemiyorsa diğer bir deyişle saf düğüm (pure node) ise ön-budama (pre-prunning) kuralı kullanılır. Bunun dışında eğer bir düğümde önceden belirlediğimiz sayıdan daha az sayıda kayıt varsa ya da ayraç istatistiksel anlamda belirlenen seviyede değilse bölümlendirme işlemini durdururuz.

ii. Geriye doğru budama (post-pruning):

Bu yaklaşımda tahmin gücü azalmış olup, tüm ağacın performansını düşüren dallar budanır. Bu zayıf dalları tanımamızda yardımcı olacak bir kavram vardır; uyarlanmış hata oranı (adjusted error rate). Uyarlanmış hata oranı (AE) kullanılarak, tüm ağacın uyarlanmış hata oranından düşük olan dallar belirlenir ve budanır. Geriye doğru budama yaklaşımı ön-budama yaklaşımına göre daha çok tercih edilir.

2.4.2. Karar Ağaçları Algoritmaları:

Konumuzun başında da belirttiğimiz gibi karar ağaçlarının oluşumunu belirleyen yüzlerce algoritma vardır. Bunlardan en çok kullanılanları; CART, CHAID, C4.5 ve ID3' tür. Bu algoritmalara kısaca bir göz atmak gerekirse;

CART, sınıflandırma ve regresyon ağaçlarında kullanılan bir algoritmadır. İlk olarak 1984 yılında Breiman tarafından ortaya atılmıştır. Standart CART yaklaşımı,

ayraçları sınıflandırma kriteri olarak ikili ayraç yöntemini ve büyümeyi durdurma kuralı olarak da ön budama methodunu kullanmaktadır. Olası tüm ikili ayraçları hesaba katar. Ayırma kriteri (en iyi ayracı belirleme kriteri) olarak da sınıflandırma ağaçlarından Gini endeksini, regresyon ağaçlarından ise varyans azaltma (variance reduction) yöntemini kullanmaktadır. [68]

CHAID adından da anlaşılacağı gibi, ayırma kriteri olarak Ki-kare' yi kullanır. Ki-kare testi 1900 yılında Karl Pearson tarafından geliştirilmiştir. Eğer deneyimiz belirli sayıda (k sayıda) sonuca sahipse Ki-kare analizi için bu sonuçları k sayıda ayrı hanede göstermek durumundayız. Deney n kere tekrarlanarak sonuçlar hanelerde gösterildiğinde, gözlemlerle elde edilmiş olan bu sonuçların belirli bir hipoteze gore ümit edilen teorik frekanslara uygunluk derecesinin incelenmesi ile bir karara varmak mümkün olmaktadır. Bu sebeple Ki-kare bölünmesi araştırma sonuçlarının frekans bakımından kategorilere ayrılabildiği ve sürekli değişkenlerle anlatımın mümkün olmadığı hallerde kullanılır. [69]

CHAID algoritması ilk defa J. A Hartigan tarafından 1975 yılında ortaya atılmıştır. 1963 yılında Morgan ve Songuist tarafından geliştirilen AID algoritmasının güncelleştirilmiş halidir. SPSS ve SAS gibi güçlü veri madenciliği yazılımlarının kullandığı güçlü bir algoritmadır. Ayraçların sayısını kısıtlamak üzere adım adım araştırma algoritması kullanarak çokyollu ayraç yöntemini (multiway split), durdurma kuralı olarak da ön-budama methodunu kullanır. Genelde sürekli değişkenlerden çok, kategorik değişkenlerle çalışır.[70]

C4.5 algoritması ise en yeni karar ağacı algoritmasıdır. Kategorik değişkenlerde çokyollu ayraç, sürekli değişkenlerde ise ikili ayraç yöntemini ayraçları sınırlandırma kriteri olarak kullanmaktadır. En iyi ayracı belirlemede ise Entropi kullanır. Durdurma kuralı olarak da geriye doğru budama kuralını kullanmaktadır.[71]

---

[68] Potts, age , s:54
[69] Bilge Aloba Köksal, *İstatistik: Analiz Methodları*.5. Basım,İstanbul: Çağlayan Kitapevi s: 294
[70]SPSS. "Data Mining Techniques: Decision Trees", www.spss.com/datamine/tress (09.09.2001), s:1
[71] Brand ve Gerritsen, "Decision Trees", DBMS, www.dbmsmag.com,(20.03.2002), s: 8-9

ID3 algoritması sadece sınıflandırma ağaçlarında kullanılmaktadır. 1993 yılında Quinland tarafından makine öğrenmesi literatürüne katılmıştır.[72]

Ağaç diyagramı hangi değişkenlerin önemli olduğunu ve birbirleriyle nasıl etkileştiklerini incelemede çok faydalıdır. Sonuçları herkesin anlayacağı türden olmaktadır. Bunun yanında eksik değerlerin tespit edilmesinde ve iyileştirilmesinde kendini tekrarlayan bölümlendirme oldukça etkilidir. Her türlü değişken (nominal, ordinal, aralık, oran) tipi ile rahatlıkla çalışılabilir. Karar ağaçları olumlu yanlarının dışında, olumsuz birkaç özelliğe sahiptir. Standart istatistiksel metodlara üstün olan dikdörtgensel kutu yaratma özelliği, bazen basit bir doğru veya eğri ile ifade edilebilecek türden sınıflandırmalarda gereksiz bölümlendirme yapabilir. Bu da oluşturulan ağacın topolojisinde büyük etkiler yaratabilir.[73]

Karar ağaçları tekniği özellikle şirketlerin müşteri kayıplarını anlamada (aşınma-attrition/churn), çapraz satış fırsatları yaratmada, kredi risklerini ve hileli (fraud) durumları belirlemede kullanılır.

2.5. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay sinir ağları (YSA), en fazla bilinen bunun yanında sonuçları en az anlaşılan; tahmin, sınıflandırma/regresyon ve kümeleme gibi genel amaçlı kullanılan çok güçlü bir veri madenciliği aracıdır. Günlük hayatta hileli (fraudulent) durumların ve kredi risklerinin tahmin edilmesi, elyazısı ve görüntü tanıma (handwriting and image recognition) alanlarında oldukça sık kullanılmaktadır. YSA'ların sonuçlarının anlaşılması güçlüğü, bu tekniğin sonuçlarının tamamen tahmine dayalı olmasından kaynaklanmaktadır. Bu nedenle YSA tekniği kara kutu (black box) teknolojisi olarak adlandırılır.[74]

---

[72] Hair, Andersen, Tatham, Black, age, s:682
[73] Potts , age, s:56
[74] Brand& Gerritsen, "Neural Networks", www.dbmsmag.com

Yapay sinir ağlarını çekici kılan özelliği insan beynini modellemesinden kaynaklanmaktadır. Bu tarz bir modelleme bu tekniğe bazı dezavantajlar getirmektedir. Bunlardan en önemlisi, eğitilmiş sinir ağlarının sonuçları bu sinir ağları içerisinde dağıtılmış ağırlıklardır ve bu ağırlıklar sonucun neden doğru veya geçerli olduğu hakkında pek fazla ipucu verememektedir.[75] Uzmanların tek bildiği analiz sonuçlarının doğru olduğudur. Neden doğru olduğu sorusu ise hala bu alanda ciddi bir araştırma konusudur. Sinir ağları çözümlerinde matematiksel denklemleri, eksponansiyel fonksiyonları ve birçok parametreyi kullanır. Fakat bu denklemler insanların gözünde hala anlaşılmazdır. Sinir ağları, çok fazla girdi olduğu anlarda çok iyi çalışamaz. Birçok özellik örüntü tanımayı zorlaştıracağından sinir ağının performansını düşürür. Bu yapı, karar ağaçları ile beraber çok etkili çalışır. En önemli değişkenleri bulmada karar ağaçları etkili olurken, bulunan değişkenler sinir ağını eğitmede kullanılırlar.[76]

Yapay sinir ağları ve diğer veri madenciliği arasındaki en temel fark, yapay sinir ağlarının sadece sayısal değerler ile çalışıyor olmasıdır. Bunun sonucu olarak herhangi bir sayısal olmayan değer, bağımlı ya da bağımsız değişken, mutlaka sayısal bir değere dönüştürülmelidir. YSA'ların birçok değişik modeli vardır. Bu çalışmada, YSA modelleri arasında en fazla kullanılan " çok katmanlı ileriye beslemeli geriye yayılım ağı" (multilayer feed forward backpropagation) incelenecektir.

2.5.1. Yapay Sinir Ağlarının Tarihçesi

İlk olarak Wallen McCulloch tarafından 1930'lı ve 1940'lı yıllarda nöronların nasıl çalıştığına dair bilgiler edinilmeye başlandı. O dönemde McCulloch ile birlikte çalışan Walter Pitts insan beyninin matematiksel anlamda modellenebileceğini iddia etti. 1949'da Donald O. Hebb tarafından yazılan "Organization of Behaviour" adlı kitapta sinir ağlarının nasıl öğrenebileceğine dair bilgiler sunuldu ve geriye yayınım ağı methodunun atası sayılan Hebbs kuralı tanıtıldı. 1956 yılında Marvin Minsky, John McCarthy, Nathaniel Rochester ve Claude Shannon tarafından organize edilen ilk yapay sinir ağları

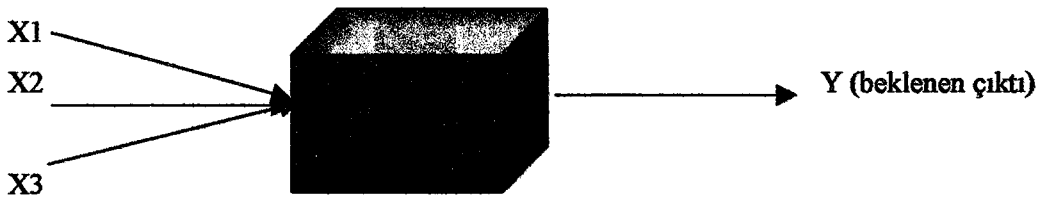
[75] Berry&Linoff, *Data Mining Techniques*, s:286-287
[76] Berry&Linoff, *Mastering Data Mining*, s:128

konferansında varsayılan ve savunulan düşünce, her öğrenme ifadesinin bir makinaya uygulanabilirliğiydi.[77]

Bilgisayarların olmadığı döneme denk gelen bu çalışmalar daha sonra nörobiyoloji dışındaki problemleri de çözmeye çalışan bir olgu haline geldi. 1968'de Minski ve Papert, yapay sinir ağları çalışmalarının yavaş gitmesinin nedeni olarak bilgisayarların yavaş çalışmasını ve bazı teorik eksikliklerin olduğunu gösterdi. Bu teorik zayıflıkların etkisini azaltmak için 1982 yılında John Hopfield, yapay sinir ağlarını eğitmede kullanılacak olan geriye yayınım ağı (backpropagation) öğrenme methodunu buldu.[78] 1988 yılında Teuvo Kohenen denetimli yapay sinir ağlarına alternatif olarak denetimsiz yapay sinir ağları tekniği olarak "Kohonen haritaları" ve "kendini düzenleyen haritalar (SOMs)" olarak adlandırılan yöntemleri buldu.[79] 1980'yılların ikinci yarısında bu gelişmeler labaratuar ortamından çıkarılarak ticari hayata girmeye ve gerçek iş dünyası problemlerine çözüm bulmaya başladı. Bu yıllarda bilgisayarların hızlarının çok artması, analizcilerin sinir ağları kavramına alışarak bu konuda kendilerini geliştirmeleri ve çoğu firmanın operasyonel sistemlerinin otomatikleştirilmesi, yapay sinir ağları konusunun hızlı bir şekilde ilerlemesine ve gelişmesine yol açtı.

2.5.2. Yapay Sinir Ağlarının Çalışma Mekanizması

Yapay sinir ağları, insanların tecrübelerle öğrenmesi gibi örnekler vasıtasıyla öğrenir. Şekil 2.3'de görüldüğü üzere kapalı bir kutu gibi çalışan yapay sinir ağı, uygun değişkenleri girdi olarak kullanıp beklenen çıktıyı üretir.



Şekil 2.3. Yapay Sinir Ağı Modeli

[77] G. David Garson. *Neural Networks for Social Sciences,*London: SAGE Publ., 1998, s: 2
[78] Mitchell, age, s.81

[79] Christopher Bishop. *Neural Networks for Pattern Recognition,* Great Britain: Oxford, 1995, s:188

Sinir ağları girdi değerleri 0 ile 1 aralığında iken çok iyi çalışır. Bu yüzden sürekli ve sırasal değişkenleri bu iki değer aralığına getirerek iyileştirmek gerekir. Bu iyileştirme yapıldıktan sonra, sinir ağını eğitmek daha mantıklı olacaktır. Sinir ağları, sonuçları bilinen eski verileri kullanarak tahmin ettiği çıktıları gerçek çıktılar ile karşılaştırarak, tahmin gücünü arttırmak amacıyla iç ağırlıkları (internal weights) yeniden düzenler. Bütün örneklerin üzerinden birkaç kez geçildikten sonra ağ ağırlıkları hesaplanmış olur. Bundan sonra oluşturduğumuz sinir ağına daha önce hiç görmediği bir test kümesi girdi olarak verilir. Tahmin sonuçları tatmin eder nitelikte ise, sinir ağı modeli oluşturulmuş demektir. Yalnız, elde ettiğimiz çıktı 0 ile 1 aralığında çıkacağından, sonucun anlam ifade etmesi için bu değeri gerçek değerine getirmeyi unutmamalıyız. Yukarıda anlattığımız süreci maddeler halinde özetlemek gerekirse:[80]

i. Girdi ve çıktılarının belirleyici özelliklerini tanımlamak
ii. Girdi ve çıktı değerlerini 0 ile 1 aralığına getirmek
iii. Uygun topolojide bir ağ inşa etmek
iv. Eğitim kümesindeki örnekleri kullanarak sinir ağını eğitmek
v. Eğitim kümesindeki örneklerden bağımsız olarak seçilen test kümesinin üzerinde, oluşturduğumuz sinir ağını denemek.
vi. Değerlendirme kümesini kullanarak modelin performansını belirlemek
vii. Bilinmeyen girdiler kullanarak modeli uygulamak.

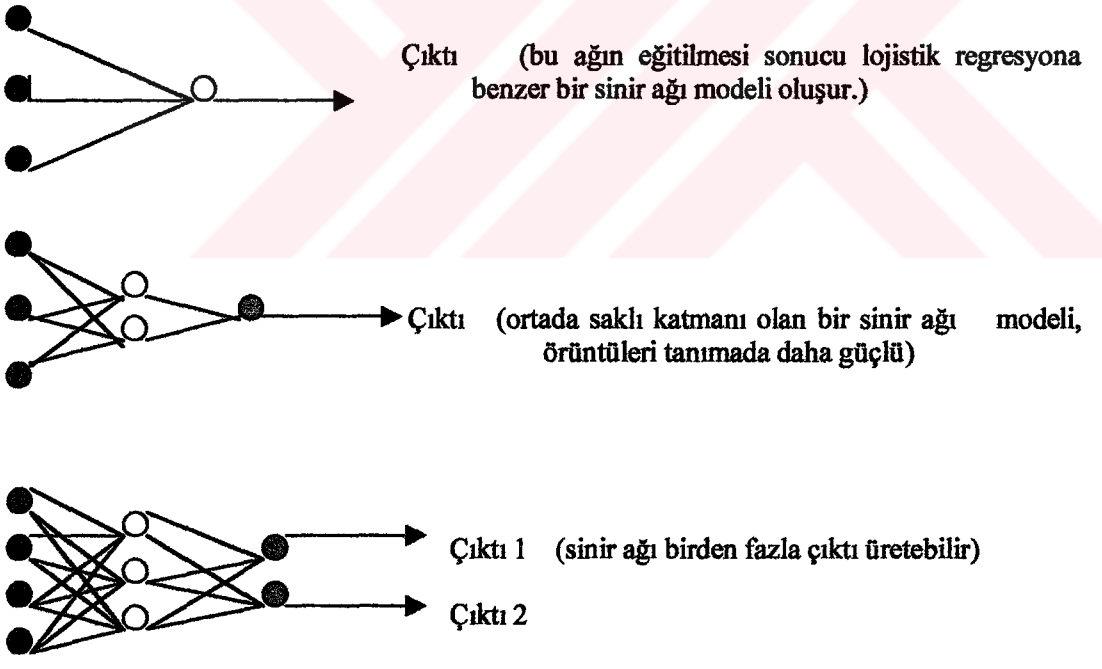
2.5.3. Sinir Ağlarının Yapısı

Genel olarak bir yapay sinir ağı modeli n adet katman (layer), her katmanda biyolojik sinir hücrelerine benzer işlevi yerine getiren ve değişik sayılarda olabilen hesaplama elemanları arasındaki yoğun bağlantılardan meydana gelmektedir. Farklı yapay sinir ağı modellerinde kullanılan hesaplama elemanları, yapay sinir hücresi

[80] Berry&Linoff, *Data Mining Techniques*, s:294

(artificial neuron), düğüm (node), birim (unit) ve işlem elemanı (processing element) olarak isimlendirilmektedir.[81]

Sinir ağları biyolojik nöronlar üzerinde modellenmiş temel birimlerden oluşur. Her birim basit bir çıktı üreten girdilere sahiptir. Bazı birimlerin çıktıları diğer birimler için girdi niteliğinde olabilir. Bu durumda saklı bir orta katman (hidden layer) var demektir. Bu saklı katmandan çok fazla olmaması sinir ağının performansı açısından önemlidir. Çok fazla saklı katman olması, modelin veriyi ezberleyeceği ve genelleştirme yapamayacağı anlamına gelmektedir. Uygulamanın amacının genelleşirme olması bu durumun önemini daha da iyi açıklar. Şekil 2.4'de de görebileceğiniz türden sinir ağları ileri besleme sinir ağları şeklinde adlandırılmaktadır.[82] Bu ağ içinde döngü olmadığı, girdiden çıktıya tek bir yol olduğu anlamına gelmektedir.



Şekil 2.4. İleriye besleme yapay sinir ağlarına üç adet örnek

Kaynak: Michael J.A. Berry ve Gordon Linoff. *Data Mining Techniques: ForMarketing, Sales and Customer Support.* USA: John Wiley&Sons, Inc., 1997, s:296

[81] Akpınar, "Yapay Sinir Ağları ve Kredi Taleplerininin Değerlendirilmesinde Bir Uygulama Önerisi", Araştırma Raporu, Mayıs 1993, s:7

[82] Potts, age, s: 49

Şekildeki ilk örnekte görüldüğü gibi, genelleştirilmiş lineer modeller ileriye besleme ve saklı katmanı olmayan yapıda sunulurlar. Standart lineer regresyon ve lojistik regresyon bu duruma örnek olarak verilebilir. Bu basit yapılılık (saklı katmanı olmama durumu) bu uygulamaları basit ve uygulanabilir hale getirmektedir. Lineer-lojistik modellerin basitliği bu uygulamaları çekici hale getirmenin dışında esnekliklerini fazlasıyla sınırlandırmaktadır.[83]

Daha önceden de belirtildiği gibi, biyolojik bir sinir hücresine benzer şekilde faaliyet gösteren bir düğümün birden fazla girdisi, fakat ağda yer alan diğer düğümlere gönderebileceği tek bir çıktısı bulunmaktadır. Önceki düğümlerden gönderilen bu girdilerin her biri, mevcut düğümün bu girdiler için belirlediği ağırlık değerleri (weights) ile çarpılmakta, bu çarpımlar sonucunda elde edilen toplam, mevcut düğümün net girdi (net inputs) değerini oluşturmaktadır. Düğümler arasında oluşturulan engelleyici bağlantılar negatif ağırlıklarla, uyarıcı bağlantılar ise pozitif ağırlıklarla çarpılmaktadır. Bu işlem, n uyarı gönderen düğüm sayısını, Xk k. düğümden gelen girdiyi, Wik i. düğümün k. düğüme uyguladığı ağırlığı ve net(i) i. düğümün net girdisini göstermek üzere,

$$\text{Net}(i) = \sum_{k=1}^{n} X_k * W_{ik}$$

eşitliği ile genellenebilir. Bu aşamayı faaliyet fonksiyonunun (activation function) hesaplanması izleyecektir. Eşitlikte Ai (t), t. zamanda i. düğümün faaliyet değerini göstermektedir. Eşitlikte görüldüğü gibi, t. zamandaki faaliyet değeri, t-1. zamandaki faaliyet değeri ve t. zamandaki net girdi değerinin bir fonksiyonu olarak hesaplanmaktadır. Ancak genel yapı bu olmakla beraber, birçok yapay sinir ağı modelinde faaliyet ve net girdi değeri eşdeğer olarak kullanılmaktadır.

$$A_i(t) = F_i(A_i(t-1), \text{net}_i(t))$$

[83] Potts, age, s:49

Faaliyet fonksiyonunun hesaplanmasını çıktı değerinin belirlenmesi izler. Çıktı değerinin hesaplanmasında kullanılan eşitliği, Xi= Fi (net i) şeklinde yazabilmek mümkündür.[84]

Aktivasyon fonksiyonu iki kısma ayrılmaktadır. Ilki kombinasyon ikincisi ise transfer fonksiyonudur.[85] Bazı kitaplarda bu ayrım toplam (summation) ve aktivasyon fonksiyonu olarak gerçekleşmektedir.[86] Kombinasyon fonksiyonu ağırlıklandırılmış girdileri birleştirerek tek bir değer haline getirir. Ençok kullanılan kombinasyon fonksiyonu ağırlıklandırılmış toplamdır (weighted sum). Transfer fonksiyonu ise kombinasyon fonksiyonundan elde edilen değeri çıktı değerine transfer eder. Üç tip transfer fonksiyonu vardır: Sigmaud, lineer ve hiperbolik tanjant. Lineer transfer fonksiyonu sadece lineer regresyon yapar, sigmaud ve hiperbolik tanjant ise lineer olmayan fonksiyonlarda çalışır. Sigmaud ve hiperbolik tanjant fonksiyonları arasındaki tek fark ise sigmaudun çıktı değerlerinin 0 ile 1 arasında, diğerinin ise –1 ile 1 arasında olmasıdır. Bu fonksiyonlardan en çok kullanılanı sigmaud fonksiyonudur. [87]

Sigmaud (x) $= 1/(1+e^{-x})$ x: kombinasyon fonksiyonunun sonucu

Bir düğüm birden fazla düğüm tarafından etkilendiği için, etkileyici düğümlerdeki yetersizlik veya bozukluk, sistemin tüm performansını etkilemeyecektir. Bu durum yapay sinir ağlarının klasik bilgi işlem uygulamalarına göre, yetersiz ve ya bozuk verilerle çalışma sırasında ortaya çıkabilecek hatalara karşı ne kadar esnek olduğunun bir göstergesidir.[88]

---

[84] Akpınar, "Yapay Sinir Ağları ve Kredi Taleplerininin Değerlendirilmesinde Bir Uygulama Önerisi", age, s :9-10
[85] Berry&Linoff, *Mastering Data Mining,* s:124
[86] Garson, age, s:42
[87] Bishop, age, s:83
[88] Akpınar,"Yapay Sinir Ağları ve Kredi Taleplerininin Değerlendirilmesinde Uygulama Önerisi",age,s :11

2.5.4. Sinir Ağlarının Öğrenme Methodları

Bir sinir ağını eğitme, herbir birimin girdilerinin en iyi ağırlıklarını seçme sürecidir. Bu noktada eğitim kümesinin kullanılmasının amacı olası çıktı değerini gerçek çıktı değerine yakınlaştırabilmek için gerekli ağırlıkları üretmesidir. Bu işlemi yapan ve en çok uygulanan öğrenme methodu Hopfiel tarafından geliştirilen geriye yayılım methodudur.[89]

Yayınma ve uyum gösterme olmak üzere iki aşamada işlemleri gerçekleştirilen geriye yayılım methodunda, katmanlar arasında tam bir bağlantının bulunduğu çok katmanlı, ileri beslemeli ve denetimli olarak eğitilen bir yapay sinir ağı modelidir. Bu model içerisinde girdi, gizli ve çıktı olmak üzere üç katman bulunmakla beraber problemin özelliklerine bağlı olarak gizli katman sayısını arttırabilmek mümkündür.[90] Bu methodda uygulanan adımları gözden geçirmek gerekirse:[91]

i. Sinir ağı eğitim kümesindeki örnekleri kullanarak ve varolan ağırlıkları da hesaba katarak olası çıktılar hesaplanır

ii. Gerçek çıktı ile bulunan çıktı değerleri arasındaki farkı bularak hata hesaplanır

iii. Hatadan yola çıkılarak ağırlıklar yeniden düzenlenir ve hata minimize edilmeye çalışılır.

Bu adımlar arasında en önemli olanı üçüncü adımdır. Hata bilinerek birimler ağırlıklarını nasıl baştan düzenleyebilirler? Bu soruyu cevaplamak için ilk olarak çıktının herbir girdiye ne derece duyarlı olduğu hesaplanır. Eğitim kümesindeki yeni düzenlemeler ağırlıkları optimal değerlerine yaklaştırır. Amaç eğitim kümesindeki örüntüleri tanımaktır. Hata oranı daha fazla düşmüyorsa eğitim sürecini durdurmak gerekir. Sinir ağı girdi örüntülerini öğrenmiş demektir.

---

[89] Neil Gershenfield, "The Nature of Mathematical Modeling", UK: Cambrige, 1999, s: 150

[90] Akpınar,"Yapay Sinir Ağları ve Kredi Taleplerininin Değerlendirilmesinde Uygulama Önerisi",age,s:35

[91] Berry&Linoff, *Data Mining Techniques,* s:304

Genelleştirilmiş delta kuralı (generalized delta rule) ağırlıkların düzenlenmesini sağlayan bir tekniktir. Bu kuralda momentum ve öğrenme oranı (learning rate) şeklinde iki parametre bulunmaktadır. Momentum herbir birim için ağırlıkların ne yönde değiştiğini hesaplar. Araştırmacılar, sinir ağlarının öğrenme hızını arttıracak bir teknik aramaktadırlar. Bazı sinir ağı yazılımları birden çok eğitim methodu sunarak kullanıcıların problemlerine uygun çözümleri sunmaktadır.[92]

Eğitim tekniklerini kullanırken karşılaşacağımız bir problem lokal optimum problemidir.[93] Verilen eğitim kümesi için en iyi sonucu üreten ve hesaplanmış ağırlıkların daha fazla değişemediği durumlarda bu ağırlıkların en iyisi olduğuna karar verilir. Halbuki, çözümü daha iyi bir sonuca ulaştıracak başka bir ağırlıklar kombinasyonu olabilir. Dolayısıyla, öğrenme oranı ve momentum kavramlarını kullanarak en iyi local çözümü bulmaktansa en iyi global çözümü bulmaya çalışmalıyız.

Saklı katmanların ne büyüklükte olması gerektiği de sinir ağının öğrenme performansını etkileyen bir sorundur. Bu konuda bir görüş, saklı katmanın büyüklüğünün giriş katmanındaki (input layer) büyüklüğün iki katından fazla olmayacağı, başka bir görüş ise girdi katmanındaki büyüklüğe eşit sayıda olması gerekitiği yönündedir.

Eğitim kümesinin büyüklüğünün seçimi de sinir ağlarının performansını arttırabilecek yönde alınacak kararlar arasındadır. Eğitim kümesi girdilerin bütün özelliklerini kapsayacak kadar büyük olmalıdır. Sinir ağındaki herbir ağırlık için seçilmiş birden fazla eğitim örneği olmalıdır. Örneğin, x girdi birimi, y saklı birimi ve bir sonucu olan bir sinir ağı modelinde n*(x+1)+y+1 tane ağırlık olmalıdır. (n, eğitim kümesi için gerekli örnek sayısı). Son olarak sinir ağının daha iyi sonuç üretebilmesi için öğrenme oranı ve momentum parametrelerinin seçimi de etkili olmaktadır diyebiliriz.

---

[92] Michael Berthold and David J.Hand. *Intelligent Data Analysis*, Italy: Springer, 1999, s:234
[93] Bishop, age, s:255

Yapay sinir ağlarını genel anlamda bu şekilde özetledikten sonra, olumlu ve olumsuz yönlerini de anlatarak veri madenciliği teknikleri arasındaki yerini daha net belirleyebiliriz.

Olumlu yönleri:

i. Çok farklı yapıdaki problemlere çözüm sağlaması,
ii. Kategorik ve sürekli değişkenleri içeren problemlere uygun çözümler üretebilmesi (değişken probleminin olmaması) şeklinde sıralanabilir.

Olumsuz yönleri ise :

i. Girdilerin sayısal değerlerinin mutlaka 0 ile 1 aralığında olma zorunluluğu, bulduğu sonuçları nasıl bulduğuna ilişkin ipucu vermemesi (Kara kutu özelliği).
ii. Sonuçlarının anlaşılamaz olması. (Kurallar ile tanımlanamaz olması)

2.5.5. Kendini Düzenleyen Haritalar (SOMs)

Şu ana kadar sinir ağı modellerinin denetimli veri madenciliğinde kullanıldığı alanlara değindik. Sinir ağları aynı zamanda denetimsiz veri madenciliğinde de kullanılan bir araçtır. Dr. Tuevo Kohenen tarafından keşfedilen kendini düzenleyen haritalar (self-organizing maps) ya da "kohonen feature maps" kavramı verideki kümeleri tanımlamada, görüntü ve sesleri tanımada kullanılmaktadır. Eğitim methodu olarak geriye yayılım methodunundan farklı bir method kullanılmaktadır. Çünkü sinir ağı yapısı olarak farklı topolojidedir. SOMs, verideki bilinmeyen örüntüleri tanımaya çalışan bir sinir ağı modelidir.

SOMs'un amacı verideki boyut sayısını azaltmaktır (dimension reduction). Genelde bu methodla ağdaki boyut sayısı ikiye indirgenir. Birbirine benzeyen üyelerden oluşan kümelerin ortaya çıkarılması söz konusudur. Veri madenciliği araçlarından kümeleme (clustering) tekniğinde de kullanılan bir sinir ağı modelidir. Oluşturulan sinir ağı modelinde genelleştirmeyi engelleyen boyut sayısındaki fazlalık, SOMs'lar kullanılarak

ikiye indirgenir, böylelikle sinir ağının sonuçları ezberlemekten ziyade genelleştirmesi sağlanır.

Yapay sinir ağları kapsamında yer alan, denetimsiz ve rekabetçi (competitive) öğrenimin önemli bir örneği olan kendini düzenleyen haritalar tekniği, projeksiyon ve kümeleme problemlerine etkin bir çözüm sunmaktadır. Düşünme, konuşma, görme, işitme ve motor fonksiyonları gibi farklı faaliyetlerin merkezlerinin bulunduğu beyin zarının işlem ve ilişki yapılarından esinlenen bu teknik, görüntü ve ses tanıma uygulamaları için geliştirlmiş olmakla birlikte çeşitli projeksiyon ve kümeleme analizlerinde de başarı ile kullanılmaktadır.[94]

[94] Teuvo Kohonen, *Selg Organizing Maps*, Springer Verlag, ,1995, s:27-28

BÖLÜM 3

SINIFLANDIRMA ve TAHMİN YÖNTEMLERİ KULLANARAK BANKA MÜŞTERİLERİ BÖLÜMLENDİRMESİ VE KREDİ SKORLAMA MODELİ

Her geçen gün artan rekabet ortamında, işletmeler mevcut müşterilerini kaybetmemek, bunun yanında potansiyel müşterilerinin kim olduğunu ve genel anlamda müşterilerinin isteklerinin neler olduğunu anlama konusunda teknolojiden yararlanma yolunu seçmişlerdir. Veri madenciliği teknolojisi, milyonlarca müşteri verisini analiz ederek işletmelerin temel sorularına cevap vermede oldukça etkili olmaktadır.

Çalışmanın bu kısmında veri madenciliğinin işletmelerde hangi amaçlarla kullanıldığı ve başlıca uygulama alanları verilecektir. Daha sonra gerçek bir veri kümesine veri madenciliği modellerinden sınıflandırma ve tahmin yöntemlerini baz alan veri madenciliği teknikleri (karar ağaçları, regresyon, yapay sinir ağları ve kümeleme analizi) uygulanarak bu tekniklerin sonuçlarının işletmeler için önemi gösterilecek ve tekniklerin performanslarının karşılaştırılması yapılacaktır.

3.1. Veri Madenciliğinin Başlıca Uygulama Alanları

Veri madenciliği astronomi, biyoloji, finans, pazarlama, sigorta, tıp ve birçok başka dalda uygulanmaktadır. Son 20 yıldır ABD' de çeşitli veri madenciliği algoritmalarının gizli dinlemeden, vergi kaçakçılarının ortaya çıkarılmasına kadar çeşitli uygulamalarda kullanıldığı bilinmektedir.[95] Günümüzde veri madenciliği teknikleri özellikle işletmelerde çeşitli alanlarda başarı ile kullanılmaktadır. Bu uygulamaların başlıcaları ilgili alanlara göre aşağıda özetlenmiştir:

[95] Akpınar,"Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği",age, s:2

i. Perakende /Pazarlama:

Günümüz serbest rekabet ortamında zaman ve piyasa verilerinin karar destek amaçlı bilgi haline dönüşmüş şekli, satış ve pazarlama faaliyetleri açısından kritik öneme haiz iki unsurdur. Doğru karar alma ve bu kararların gecikmeden hayata geçirilmesi işletmelerin varlığını devam ettirebilmeleri açısından çok önemlidir. Her müşteri adayı aynı zamanda potansiyel bir müşteri adayıdır ve günümüzün yoğun rekabet ortamında şirketlerin başarıya ulaşabilmesi müşterilerini iyi tanıması, var olan müşterilerine yeni satışlar yapabilmesi ve onları memnun edebilmesi, müşterilerinin aynı sektörde hizmet veren diğer şirketlerle çalışmaya başlamasının engelleme amaçlı pazarlama politikaları geliştirebilmesi, yaptığı promosyonlardan hangilerinin hangi özellikli müşterileri tarafından ilgi göreceğini önceden bilmesi, şirkete kalıcı müşteriler kazandıracak promosyon paketlerinin hangileri olduğunu belirleyebilmesi gerekir.[96]

Bu noktada veri madenciliği yöntemlerinden özellikle müşterilerin satın alma davranışlarının tanımlanması, müşterilerin demografik özellikleri arasındaki ilişkilerin anlaşılması, postalama kampanyalarına cevap oranının tahmin edilmesi, müşteri sadakati belirleme, sepet analizi ve müşteri ilişkileri yönetimi konusunda faydalanılmaktadır.

ii. Bankacılık:

Bankacılık sektöründe veri madenciliği teknikleri genel olarak altı amaç dahilinde kullanılmaktadır. Bunlar; müşteri segmentasyonu ve profilleme, müşteri aşınmalarının tahmin edilmesi, kredi skorlama, kredi kartı dolandırıcılıklarının (fraud detection) tespiti, şirket analizi ve risk yönetimi ve son olarak kapasite planlama tahminidir.[97]

Bankalar için ne tür müşteri kitlesine sahip olduğunun bilgisi oldukça önemlidir. Bu bilgi dahilinde, bankalar müşterilerinin özelliklerini kullanarak müşteriye özel satış promosyonları, çapraz satış ve postalama kampanyaları düzenleyebilirler. Bunun yanında

[96] SPSS Inc., "Clementine", www.spss.com.tr/clementine, (14.03.2002), s:2-3

[97] SAS Ins Inc. "How Can Data Mining Help in Banking", www.sas.com (25.03.2002), s:1

bankayı bırakmak üzere olan müşterileri tespit ederek bu müşterilere özel teklifler sunabilirler.

Veri madenciliği tekniklerinin şirkete ait analizlerinde ve özellikle risk yönetimi konusundaki kullanımı ise oldukça yaygındır. Bu yönde yapılan bazı çalışmalar finansal planlama ve aktif varlıkların değerlendirilmesi kapsamında nakit akışlarının analizi ve tahmini, aktif varlıkların değerlendirilmesi için şüpheli alacaklarının tespiti, zaman serileri ve cross-sectional analizi, işletmelerin performansına yönelik geleceğe dair tahminler, müşterilerin sınıflara ayrılması ve sınıflara göre fiyat politikası tayinidir.[98] Veri madenciliği teknikleriyle geçmişe ait veriler kullanılarak, geçmişte dolandırıcılık yapmış kişilere ait bilgiler incelenebilir ve bunlara ait bir model kurulabilir.

Müşteri bölümlendirmesi (segmentation) ve müşteri aşınması tahmini, farklı sektörlerde faaliyet gösteren işletmelerin pazarlama departmanlarında kullanılan veri madenciliği faaliyetleridir. Kredi skorlama ise özellikle bankacılık sektöründe kullanılmaktadır. Kredi skorlama yönteminde müşterilerin demografik ve psikografik özelliklerine göre borç geri ödeme durumları skorlanır. Yapılacak tahminlere göre yeni gelen müşteri verileri incelenerek kredi başvurusunun kabul veya red durumuna karar verilir.

iii. Diğer Sektörler

Veri madenciliği tekniklerinin başlıca kullanım alanlarının perkendecilik ve bankacılık sektörü olmasının yanısıra özellikle sağlık, sigorta, ulaşım (transportation) ve ilaç (medical) sektöründe de kullanım alanları mevcuttur.

Yeni poliçe satın alma ihtimali olan müşterilerin tahmin edilmesi, sigorta poliçelerindeki dolandırıcılıkların/kötü amaçlı kullanımın tahmin edilmesi, riskli müşterilerin davranış örüntülerinin tanımlanması, farklı hastalıklara uygulanabilecek tıbbi tedavilerin bulunması ve son olarak hasta davranışlarının karakterize edilerek ofis

[98] SPSS Inc. , "Clementine",age, s:2-3

ziyaretlerinin tahmini bu sektörlerde veri madenciliğinin uygulama alanlarını daha net göstermektedir.

3.2. Araştırmanın Tanımlanması

Çalışmamızda bir bankanın bireysel bankacılık departmanına gelen müşteri kredi taleplerinin değerlendirilme süreci incelenecektir. Bu bölümde sırasıyla araştırmanın amacı, araştırmanın varsayımları ve çerçevesi, ana kütle ve değişkenlerin tanımı verilecektir. Daha sonra uygulama için öngörülen tahmin modelleri oluşturulacaktır. Tahmin modellerinin performans karşılaştırılmaları yapıldıktan sonra son olarak müşteri bölümlendirme analizi olarak bilinen kümeleme analizi uygulanacaktır.

3.2.1. Araştırmanın Amacı

Çalışmada, bir bankanın bireysel bankacılık departmanına borç konsolidasyonu ve konut yenileme nedeniyle gelen kredi taleplerinin, veri madenciliği teknikleri uygulanarak değerlendirilmesi ve kredi taleplerinin kabul veya red kararının bu doğrultuda verilmesi amaçlanmaktadır.

Çalışmada ilk olarak bankaya gelen kredi başvurularının kabul veya red kararının otomatik olarak verileceği tahmin ve denetimli sınıflandırma modeli oluşturulacaktır. Bu amaçla tahmin modellerinden regresyon analizi ve denetimli sınıflandırma modellerinden karar ağacı analizi modele eklenecektir. Modellerin performans karşılaştırılmaları yapıldıktan sonra bankanın mevcut problemini çözmede etkin olan model değerlendirmeye alınacaktır.

Bankanın kredi başvuru taleplerini değerlendirmede etkin olabilecek bir başka teknik olan kümeleme analizi de uygulamanın diğer bir aşamasını oluşturacaktır. Kümeleme analizinde müşterilere ait değişkenler detaylı şekilde incelendikten sonra

kümeleme analizinin hiyerarşik olmayan yöntemlerinden K-ortalamalar algoritması uygulanacaktır. Kümeler analitik ve işletme filtrelerinden geçirildikten sonra profillendirilecektir. Kümelerin profillerini belirleyen değişkenlerin ne oranda doğru tahmin edildiğini test edilmek için kümeleme analizi sonunda tahmini multinomial regresyon analizi yapılacaktır.

3.2.2. Araştırmanın Varsayımları ve Çerçevesi

Çalışamızda amaç bir bankanın bireysel bankacılık departmanına gelen kredi başvurularının kabul veya red kararının otomatik olarak verilmesi ve bankanın müşteri portföyünün anlamlı bölümlere ayrılmasını sağlamaktır. Bu amaçla bankanın kararını ve uygulamalarını hızlandıracak veri madenciliği tekniklerinin kullanacağı verinin güvenilir bir kaynaktan elde edildiği, güncel olduğu ve verinin önceden anlamsız bilgilerden temizlenmiş ve doğrulanmış olduğu varsayılmaktadır. Uygulamalardaki testler aşağıdaki çerçevelerde ele alınmıştır;

Çerçeve 1: Veri madenciliği tekniklerinden regresyon ve karar ağacı analizleri geleceğe yönelik anlamlı tahminler oluşturarak ve çok miktarda veri kümesinden anlamlı bilgiler çıkararak işletmelerin karar alma süreçlerini hızlandırmaktadır.

Çerçeve 2: Kümeleme analizi, müşterileri anlamlı bölümlere ayırarak işletmelerin müşterilerini tanımasını sağlamaktadır. Bunun yanında işletmelerin herbir bölüm için farklı pazarlama stratejileri belirleyerek maksimum kar minumum zarar etmesine de katkıda bulunmaktadır.

3.2.3 Anakütle ve Değişkenlerin Tanımı

Uygulamada söz konusu veri kümesi ve veri madenciliği paketi SAS Institute Türkiye tarafından sağlanmıştır. Uygulama SAS Enterprise Miner 4.2 veri madenciliği paket programı ile gerçekleştirilmiştir.

Uygulamada kulllanılan veri kümesi tüm ana kütle olarak belirlenmiştir. Veri kümesinde 5960 müşteriye ait 13 adet değişken bulunmaktadır. Bu değişkenlerin isimleri, değişken tipleri (ikili, nominal, sırasal, aralık ölçekli, oran ölçekli), veri madenciliği modelindeki rolleri ve kod açılımları Tablo 3.1'de ayrıntılı olarak verilmektedir.

Tablo 3.1. Değişkenlerin tanımlanması

Değişken İsmi	Değişkenin Modeldeki Rolü	Değişken Tipi	Değişkenin Tanımı
BAD	girdi	ikili	1=Borcunu ödememiş 0=Borcunu ödemiş
REASON	girdi	ikili	HomeImp:ev yenileme Debcon: Borç konsolidasyonu
JOB	girdi	Nominal	6 meslek kategorisi
LOAN	girdi	aralık ölçekli	Talep edilen kredi miktarı
MORTDUE	girdi	aralık ölçekli	Konut ipotek değeri
VALUE	girdi	aralık ölçekli	Mevcut mal varlığının bugünkü değeri
DEBTINC	girdi	aralık ölçekli	Borç gelir oranı
YOJ	girdi	aralık ölçekli	Müşterinin mevcut mesleğinde geçirdiği toplam sene
DEROG	girdi	aralık ölçekli	Borç ihbar belgesi sayısı
CLNO	girdi	aralık ölçekli	Kredi başvuru sayısı
DELINQ	girdi	aralık ölçekli	Ödenmeyen kredi sayısı
CLAGE	girdi	aralık ölçekli	İlk yapılan kredi başvuru süresinden itibaren ay bazında geçen toplam süre
NINQ	girdi	aralık ölçekli	Kredi soruşturma sayısı

Kaynak: Wielenga Doug , Lucas Bob & George Jim, "Applying Data Mining Techniques- Course Notes", USA, SAS Ins., 1999, s:106

3.2.4 Tahmin ve Denetimli Sınıflandırma Tekniklerinin Uygulaması

Uygulamada öncelikli olarak tahmin ve denetimli sınıflandırma teknikleri için bir hedef değişken belirlenmesi gerekmektedir. Bu amaçla, BAD değişkeni hedef değişken olarak konumlandırılmıştır. BAD değişkenine ait 1189 (%20) ödenmemiş ve 4771 (%80) ödenmiş konut kredisi bilgisi kullanılarak oluşturulacak kredi skorlama modeli, gelecekte başvuruda bulunacak potansiyel müşterilerin konut kredisi başvurularını otomatik olarak kabul veya red edecektir. Bu amaçla çalışmamızda, denetimli sınıflandırma modellerinden karar ağaçları tekniği ile performans karşılaştırması yapmak amacıyla tahmin modellerinden lojistik regresyon teknikleri kullanılacaktır.

Çalışmamızda ilk adım verinin elde edilmesidir . Analizi yapılacak olan veri kümesi CRSSAMP. HMEQ adıyla kodlanmıştır. Analizimize başlamadan önce veri kümesindeki değişkenlerin modeldeki rolleri ve değişken tipleri belirlenmiştir (bknz tablo 3.1.).Veri kümesindeki tüm değişkenler model oluşturma sürecinde aktif rol oynayacağından bütün değişkenler kullanılır statüdedir. Verinin SAS Enterprise Miner'a yerleştirilmesinden sonra veri madenciliği modelinde eğitim ve değerlendirme kümesi olarak kullanacağımız veri kümelerinin oranlarının belirlenmesi gerekmektedir. Eğitim ve değerlendirme kümelerinin oranları sırasıyla %70 ve %30 olarak belirlenmiştir. Ana kitlede "borcunu ödemiş" ve "borcunu ödememiş" müşteri sayısının tüm ana kitleye oranı sırasıyla %80 ve %20 olduğundan bu oranların oluşturulan eğitim ve değerlendirme kümesinde korunması sınıflandırma ve tahmin modelinin performansını olumlu yönde etkileyecektir. Bu amaçla eğitim ve değerlendirme kümelerindeki müşteriler katmanlı örneklem yöntemine (stratified sampling method) göre seçilmiştir.

Verinin yerleştirilmesi ve eğitim/test kümelerinin oranlarının belirlenmesinden sonra verideki değişkenlerin dağılımının (skewness, kurtosis, variance ve frekans dağılımı), değişkenlerin birbirleriyle ilişkilerinin ve aykırı değerlerin (outlier) tespitinin yapıldığı verinin betimsel istatistikleri (descriptive statistics) yer almaktadır.

3.2.4.1. Betimsel İstatistikler

Veri kümesimizde bulunan 10 aralık ölçekli (interval) değişkenin betimsel istatistiği (minimum ve maksimum değerleri, aritmetik ortalama, standart sapma, skewness, kurtosis, eksik değerler yüzdesel oran) incelenerek normal dağılıma uymayan ve aykırı değerlere sahip değişkenler tablo 3.2'de ayrıntılı biçimde verilmiştir. Veride betimsel istatikler incelenerek normal dağılıma uymayan değişkenler için uygun transformasyon yöntemleri, eksik değerler için ise uygun tamamlama yöntemleri belirlenecektir.

Tablo 3.2. Aralık Ölçekli Değişkenlerin Betimsel İstatistiği

Değişken ismi	Min. değer	Maks. değer	Aritmetik ort.	Standart sapma	Eksik değer(%)	Skewness	kurtosis
CLAGE	0	1168.2	179.77	85.81	%5	1.34	7.60
CLNO	0	71	21.296	10.14	%4	0.78	1.16
DEBTINC	0.52	203.31	33.78	8.60	%21	2.85	50.5
DELINQ	0	15	0.45	1.13	%10	4.02	23.36
DEROG	0	10	0.25	0.85	%12	5.32	36.87
LOAN	1100	89900	18608	11207	%0	2.02	6.93
MORTDUE	2063	399550	73761	44458	%9	1.81	6.48
NINQ	0	17	1.1861	1.73	%9	2.62	9.78
VALUE	8000	855909	101776	57386	%2	3.05	24.36
YOJ	0	41	8.92	7.574	%9	0.98	0.37

i. Tamamlama Yöntemleri

Tablo 3.2' de görüldüğü gibi LOAN değişkeni dışında bütün değişkenler belli bir oranda eksik değere sahiptir. Özellikle DEBTINC değişkeni %21 ile en fazla eksik değere sahip değişkendir. Modelin performansının arttırılabilmesi için veride mevcut

eksik değerlerin tamamlanması (imputation) gerekmektedir. Bu işlem veri madenciliği tekniklerinden özellikle regresyon ve yapay sinir ağları teknikleri için çok önemlidir. Çünkü bu tekniklerde eksik değerlere sahip müşteri bilgileri analizden çıkarılır ve sadece tam bilgiye sahip değerler ile analize devam edilir. Sonuç olarak analiz çok az veri ile çalışmak zorunda kalabilir. Bu durum karar ağaçları tekniği için geçerli değildir. Tamamlama yöntemleri kullanmadan da karar ağaçları başarı ile çalışabilir.

Uygulamada, tamamlama yöntemlerinden aralık ölçekli, nominal ve ikili değişkenler için ağaç tamamlanması (tree imputation) en uygun yöntem olarak belirlenmiştir. Ağaç tamamlanması, tamamlama değerlerini PROC SPLIT kullanarak tahmin etmektedir.Herbir değişkende eksik değerler, kalan diğer değişkenler kullanılarak tamamlanmaktadır. Hedef değişken olarak belirlenen değişken ise veri tamamlamasında kullanılmamaktadır.Tamamlamada sadece girdi mahiyetindeki değişkenler kullanılmaktadır. Ağaç tamamlanması diğer tamamlama yöntemlerine (aritmetik ortalama, medyan, Tukey's Biweight, Hubert, sabir değer) göre daha etkilidir.

Veri kümemizde değişkenlerin minimum ve maksimum değerleri ve herbir değişkenin kutu-bıyık (box-whisker) grafikleri gözlendiğinde herhangi bir aykırı değere rastlanmamıştır. Bu amaçla herhangi bir aykırı değer filtreleme işlemine gerek duyulmamıştır.

ii. Transformasyon Yöntemleri

Oluşturulacak modelin performansının iyi olabilmesi için değişken dağılımlarının normal dağılım olması gerekir. Tablo 3.2'yi incelediğimizde DEBTINC, DELINQ, DEROG, NINQ ve VALUE değişkenlerinin dağılımında pozitif bir skewness değeri gözlenmektedir. Ayrıca CLNO ve YOJ değişkenlerinin dışında tüm değişkenlerin kurtosis değerleri de oldukça yüksek çıkmıştır. Bu değerler göz önünde bulundurularak analizin performansının arttırılabilmesi için bu değişkenlerin dağılımının normalize edilmesi gerekmektedir. Bu amaçla değişkenlerin transformasyonu söz konusudur. Model kurma aşamasında SAS Enterprise Miner 4.2'de değişken dağılımlarının normalizasyonunu hedefleyen transformasyon yöntemlerinden değişken dağılımlarının

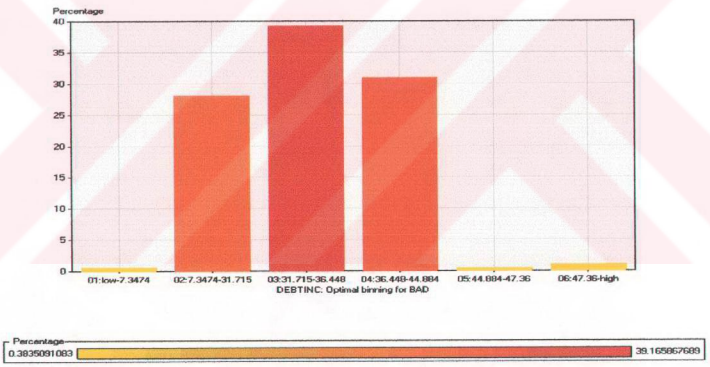
normal dağılıma uygun duruma getirilmesi transformasyonu (maximize normality transformation) ve girdi değişkenlerinin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu (optimal binning for relationships to target tranformation) iki alternatif seçenek olarak modele eklenmiştir. Bu yöntemlerden değişken dağılımlarının normal dağılıma uygun duruma getirilmesi transformasyonu, değişkenlerin betimsel istatistik değerlerini göz önünde bulundurarak, bu değişkenlerin dağılımlarının normal olabilmesi için değişik transformasyonlar (logaritmik, ters fonksiyon, karekök, eksponansiyel) denemektedir. Yapılan değişken dağılımlarının normal dağılıma uygun duruma getirilmesi transformasyonu sonucu değişkenlerin yeni betimsel istatistik sonuçları tablo 3.3'te verilmiştir.

Tablo 3.3. Değişkenlerin dağılımlarının normal dağılıma uygun duruma getirilmesi transformasyonu sonucu oluşan betimsel istatistik sonuçları

Değişken ismi	Formül	Aritmetik ort.	Standart sapma	skewness	kurtosis
CLAGE	Sqrt	13.03	3.10	0.24	1.37
CLNO	Sqrt	4.60	1.08	-0.2	0.93
DEBTINC	Sqrt	5.80	0.68	-0.47	17.2
DELINQ	log	0.27	0.45	1.72	2.51
DEROG	log	0.20	0.40	2.07	4.28
LOAN	log	9.67	0.57	-0.35	0.57
MORTDUE	sqrt	258	77	0.64	1.58
NINQ	log	0.61	0.56	-0.43	0.98
VALUE	log	0.50	-0.08	1.10	0.04
YOJ	sqrt	2.94	1.15	0.22	-0.55

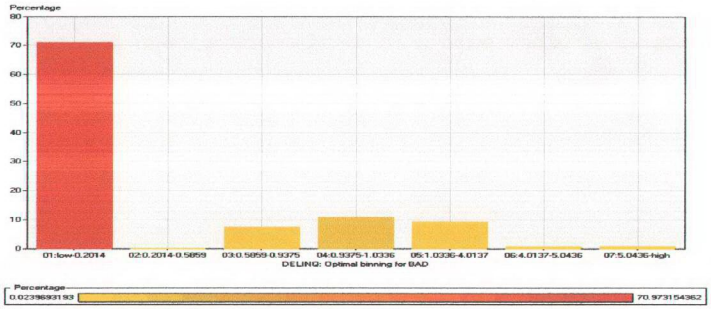
İkinci opsiyon olarak belirlenen girdi değişkenlerinin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu, bağımsız değişkenleri hedef değişkene bağlı şekilde optimal olarak n eşit gruba ayırır. Bu

transformasyon yöntemi genelde bağımsız değişkenlerle bağımlı değişkenler arasında lineer olmayan bir ilişki söz konusu olduğunda uygulanır. İlk gruplamayı oluşturabilmek için transformasyon nodu söz konusu değişken değerlerini 64 eşit parçaya ayırır. Bu 64 grup daha sonra Ki-kare değerlerinin maksimize edildiği iki gruba ayrılır. Ayırma işlemi, Ki-kare değeri eşik değer olan 3.84'ü aştığında uygulanır. Aksi takdirde girdi değeri transform edilmez. İlk ayırma işlemi bittikten sonra, aynı süreç iteratif bir şekilde 4'lü gruba ayırmada uygulanır. Enterprise Miner'ın bu nodunda belirtilen maksimum ayraç sayısına (maximum split numbers) ve bağımsız değişkenler ile hedef değişken arasındaki lineer olmayan ilişkinin Ki-kare test sonuçlarına göre ayırma işlemi devam eder. Çalışmamızda maksimum ayraç sayısı 8 olarak belirlenmiştir. Yapılan transformasyon sonucu çıkan sonuçlar önemli görülen bazı değişkenler (VALUE, NINQ, DEBTINC, DEROG ve DELINQ) bazında grafik halinde aşağıda verilmiştir.



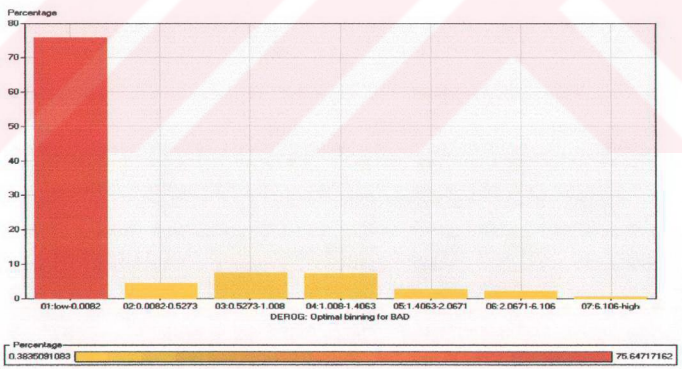
Şekil 3.1 DEBTINC değişkeninin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu sonrası dağılımı

Şekil 3.1'de görüldüğü gibi DEBTINC değişkeni transformasyon sonucu 6 gruba ayrılmıştır. Modelin performansını artırmada üçüncü grubun etkisinin daha fazla olduğu gözlemlenmektedir. Birinci, beşinci ve altıncı grupların etkisi oldukça azdır.



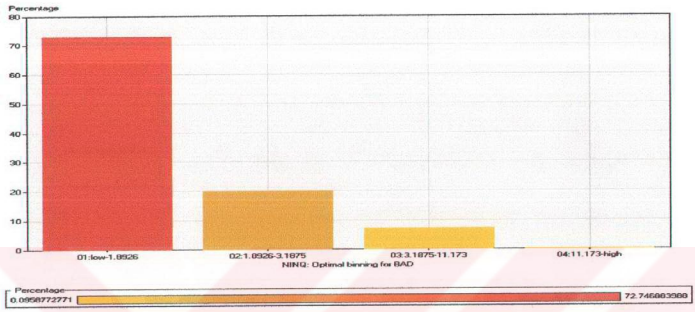
Şekil 3.2 DELINQ değişkeninin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu sonrası dağılımı

DELINQ değişkeni transformasyon sonucu 7 gruba ayrılmıştır. Modelin performansını artırmada birinci grubun etkisinin daha fazla olduğu gözlemlenmektedir. İkinci, altıncı ve yedinci grupların etkisi ihmal edilebilecek kadar azdır.



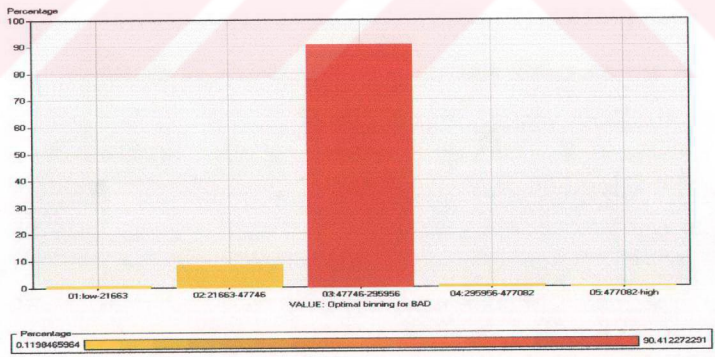
Şekil 3.3 DEROG değişkeninin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu sonrası dağılımı

DEROG değişkeninin transformasyonu sonucu hedef değişkeni tahmin etmede modelin performansını arttırmada birinci grubun etkisinin daha fazla olduğu göze çarpmaktadır.



Şekil 3.4 NINQ değişkeninin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu sonrası dağılımı

NINQ değişkeninin transformasyonu sonucu dört grup oluşmuştur. Modelin performansı açısından grupların etkisi azalarak devam etmektedir.



Şekil 3.5. VALUE değişkeninin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu sonrası dağılımı

Transform edilen VALUE değişkeninin dağılımında üçüncü grubun hedef değişkenin tahmini konusunda diğer gruplara nazaran daha etkili olduğunu görmekteyiz.

Sonuç olarak bağımsız değişkenlerin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu noduna göre transformasyon sonuçları tahmin modelimizin performansını arttırmada değişken dağılımlarının normal dağılıma uygun duruma getirilmesi transformasyonu noduna göre daha başarılı olmuştur. Bu amaçla oluşturulan tahmin modelinde en uygun transformasyon yöntemi bağımsız değişkenlerin hedef değişken ile aralarındaki ilişkinin optimal gruplandırma yöntemiyle transformasyonu olarak belirlenmiştir.

Tahmin ve denetimli sınıflandırma için veride gerekli olan tüm değişiklikler (eksik değerlerin tamamlanması (replacement), aykırı değerlerin tespiti, değişkenlerin dağılımının normalize edilmesi (tranformation)) yapıldıktan sonra veri madenciliği modelleme sürecine başlanabilir. Veri madenciliğinin başlıca görevlerinden biri gelecekte olabilecek hareketleri önceden tahmin edebilmektir. Bu amaçla çalışmamızda veri madenciliği tekniklerinden regresyon ve karar ağacı analizi tahmin modeli oluşturmada kullanılacaktır.

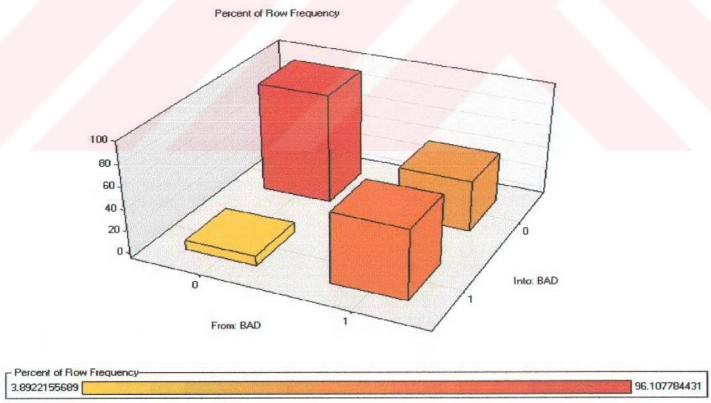
3.2.4.2. Regresyon Analizi

Veri madenciliği modellerinden regresyon analizi bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi inceleyen bir tahmin modelidir. Tahmin edilecek bağımlı değişkenin model rolü ikili değişken (1/0) olduğundan uygulamada lojistik regresyon uygun regresyon tekniği olarak belirlenmiştir. Regresyon methodlarından geri adım (backward) ve adım adım (stepwise) methodları alternatif olarak analize eklenmiştir. Geri adım methodunda analize tüm bağımsız değişkenler ile başlanmaktadır. Bağımlı değişkenin tahmin edilmesi sürecinde belirlenen güven aralığında[99] etkili olmayan değişkenler analizden çıkarılmaktadır. Adım adım methodunda ise bağımlı değişkenin tahmin edilmesi sürecinde etkili olan değişken ile analize başlanır. Herbir değişkenin

[99] Uygulamada güven aralığı %95 olarak belirlenmiştir.

analize eklenmesi ile mevcut durum incelenir ve bağımlı değişkenin tahmin edilmesinde etkisiz olan değişkenler modelden çıkarılır. Değişkenlerin modelden çıkarılması kısmi korelesyon katsayısılarının (partial correlation coefficient) belirlenen güven aralığının altına düşmesiyle mümkün olmaktadır. Güven aralığının üstünde olan değerler ile analize devam edilmektedir. Geri adım ve ileri adım metodlarında olduğu gibi sadece geriye veya ileriye bir hareket söz konusu değildir. Bağımlı değişkenin tahmin edilmesindeki yeteneklerine göre değişkenlerin eklenmesi veya elenmesi söz konudur. Hareket hem ileriye hem geriye doğrudur. Bu amaçla istatistikçiler tarafından tercih edilen bir regresyon methodudur. Genelde bağımlı değişkenin rolü ikili veya sırasal olduğunda geri adım methodu tercih edilmemektedir.[100]

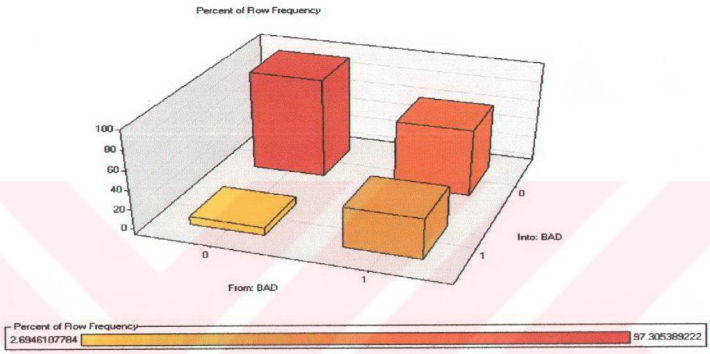
Uygulamada yapılan analiz sonucu da bu doğrultuda gerçekleşmiştir. Adım adım ve geri adım yöntemine göre kurulan lojistik regresyon modelinin sonuçları sırasıyla şekil 3.6 ve 3.7'de karşılaştırma matrisi (confusion matrix) olarak verilmiştir. Bu matrislere göre adım adım metodunu kullanan lojistik regresyon modeli daha başarılı olmuştur.



Şekil 3.6 Adım adım lojistik regresyon modeline göre karşılaştırma matrisi

[100] Doug Wielenga, Bob Lucas ve Jim Georges, SAS Enterprise Miner : Applying Data Mining Techniques Course Notes, USA: SAS Ins. Inc. , 1999 , s:64

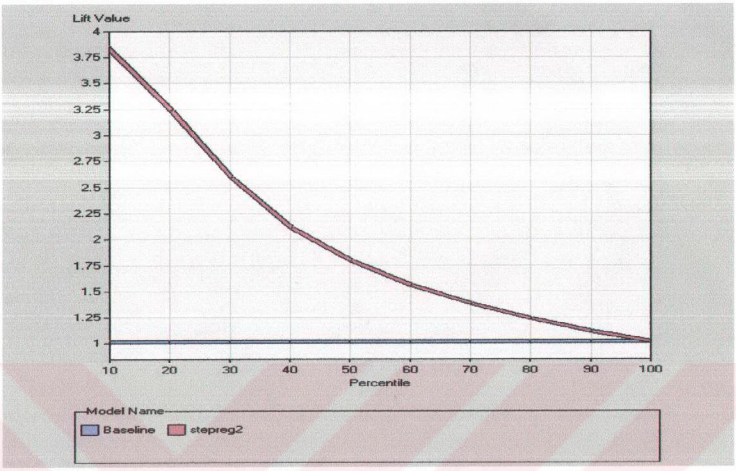
Adım adım lojistik regresyon modeli karşılaştırma matrisine göre borcunu ödeyecek olan müşterilerin tahmin başarı yüzdesi %96.1, borcunu ödemeyecek olan müşterilerin tahmin başarı yüzdesi ise %58.6' dır. Şekil 3.6'daki grafikte ana köşegen üzerindeki barların uzunluğu modelin başarısını ifade etmektedir.



Şekil 3.7 Geri adım lojistik regresyon modeline göre karşılaştırma matrisi

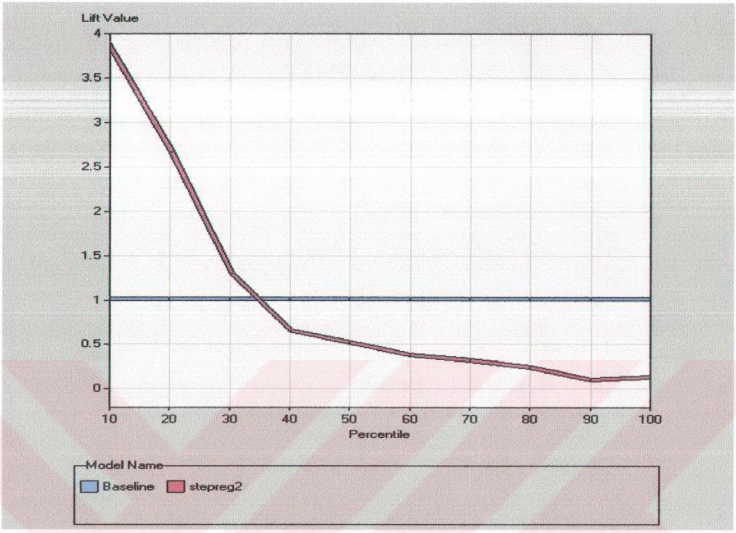
Geri adım lojistik regresyon modeli karşılaştırma matrisine göre borcunu ödeyecek olan müşterilerin tahmin başarı yüzdesi %97.3, borcunu ödemeyecek olan müşterilerin tahmin başarı yüzdesi ise %36.3' tür. Borcunu ödeyecek müşterilerin tahmini yüksek olmasına rağmen ödemeyecek olan müşterilerin tahmin başarı yüzdesinin düşük olması bu modelin etkinliğini azaltmaktadır. Sonuç olarak adım adım lojistik regresyon modeli daha başarılı bulunmuştur.

Oluşturulan regresyon modelinin değerlendirilmesi aşamasına gelindiğinde modelin asansör grafiğinin yorumlanması gerekmektedir. Şekil 3.8 ve 3.9'da seçilen adım adım lojistik regresyon modelinin sırasıyla kümülatif ve kümülatif olmayan asansör grafikleri incelenmektedir.



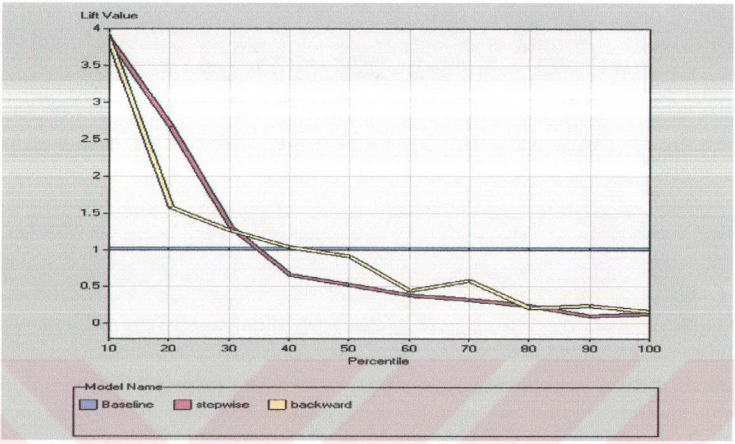
Şekil 3. 8 Adım adım lojistik regresyon modelinin kümülatif asansör grafiği

Şekil 3.8'de bulunan asansör grafiği incelendiğinde asansör değerinin oldukça yüksek çıktığını görmekteyiz. Bu durum modelin performansının da başarılı olduğunu ifade etmektedir. Veri madenciliği modeli uygulanarak yapılan analizin sonucu, rastlantısal olarak seçilen örneklem üzerinde veri madenciliği modeli uygulanmadan yapılan analiz sonucuna göre 3.8 kat daha iyi sonuç vermiştir. Bu durum ana kitlenin ilk %20'lik diliminde borcunu ödemeyecek müşterilerin %76'sına ulaşıldığı bilgisini vermektedir. Modelin gerçek anlamda başarılı olabilmesi için başvurusu red edilecek olan müşterilerin tüm ana kitle içindeki oranlarını belirlemek gerekecektir. Bu bilgiye şekil 3.9'da kümülatif olmayan asansör grafiğinden ulaşılabilir.



Şekil 3.9 Adım adım lojistik regresyon modelinin kümülatif olmayan asansör grafiği

Şekil 3.9'da da açıkça görüldüğü gibi %35'lik dilimden sonraki müşterilerin kredi başvurularının reddi bankanın zarar etmesine yol açacaktır. Bu amaçla modelin başarısının amacına ulaşabilmesi için başvuruda bulunan potansiyel müşterilerin %35'inin başvurusunun reddi söz konusudur.



Şekil 3.10. Adım adım ve geri adım regresyon modellerinin kümülatif olmayan asansör grafiklerinin karşılaştırılması

Şekil 3.10'daki kümülatif olmayan asansör grafiği incelendiğinde daha önce tespit ettğimiz adım adım regresyon modelinin başarı grafiği daha net gözlenmektedir. İlk %10'luk dilimdeki asansör oranları aynı olsa da geri adım methodunu kullanan regresyon modeli %20'lik dilimde ciddi bir düşüş gerçekleştirmiştir. Ayrıca geri adım regresyon modelinin kümülatif olmayan asansör grafiğinin iniş çıkışlar göstermesi başarısız bir model olduğunun başka bir göstergesidir. [101]

3.2.4.3. Karar Ağaçları Analizi

Karar ağaçları, denetimli öğrenimin kullanıldığı güçlü bir tahmin modelidir. Sonuçlarının kolay yorumlanabilir ve kolay inşa edilebilir bir model olması karar ağaçlarını diğer tahmin modellerine kıyasla avantajlı bir noktaya getirmektedir.

[101] Berry&Linoff, *Mastering Data Mining*, s:189

Karar ağacı modeli eksik değerler ile çalışabildiğinden oluşturulan tahmin modelinde analiz edilecek veri kümesinin eksik değerlerinin tamamlanması ve transformasyonu yapılmadan önceki durumu incelenmiştir.

Aday ayraçları belirlemede çeşitli ayraç arama stratejileri (split search strategy) kullanılmaktadır. Adaylar belli olduktan sonra en iyi ayracı belirlemek için ayraç kriteri (splitting criterion) kullanılır. Seçilen ayraç kriteri ana düğümlerle (parent nodes) karşılaştırıldığında çocuk düğümlerdeki (child nodes) hedef değişken dağılımının değişkenliğinin azalırlığını ölçmektedir. Sınıflandırma ağaçlarında üç tür ayraç kriteri kullanılmaktadır: Entropy, Gini ve Ki-kare testi. Analiz sürecimizde üç ayraç kriterini de performans karşılaştırması yapmak amacıyla inceleyeceğiz.

i. Gini ayraç kriteri :Düğümün saflığını (impurity) ölçen bir endekstir. Gini kriterinin formülü;

$$i(P_1, P_2, P_3 \ldots\ldots P_r) = 1 - \sum_{i=1}^{r} P_i^2,$$

(Pi: düğümdeki hedef değişkenin relatif oranları şeklinde ifade edilmektedir.)

Saf bir düğümün Gini endeksi 0'dır. Model rolü ikili olan hedef değişkenler için formül $2\ P_1 * (1 - P_2)$ şeklinde ifade edilir ve alabileceği en büyük değer ½'dir. Düğüm içindeki saflık derecesi arttıkça Gini endeks değeri 0'a yaklaşmaktadır.[102]

ii. Entropi ayraç kriteri : Düğümün saflığını ölçen başka bir endeks ise Entropi kriteridir. Enformasyon teorisinden gelen bu kavrama göre r adet birbirinden bağımsız olayda belirli bir sonucun azlık derecesi ;

$-\log_2(P_i)$ değeri ile hesaplanmaktadır.

Entropi azlık derecelerinin ortalama değeridir ve bu yüzden oluşacak sonucun belirsizliğini ölçer. Entropi kriterinin formülü;

[102] Potts, "Decision Tree Modeling Course Notes", USA: SAS Ins. Inc., 1999, s: 16

$$H(P_1, P_2, P_3 \ldots\ldots P_r) = -\sum_{i=1}^{r} P_i \log_2 (P_i),$$ ile ifade edilmektedir.

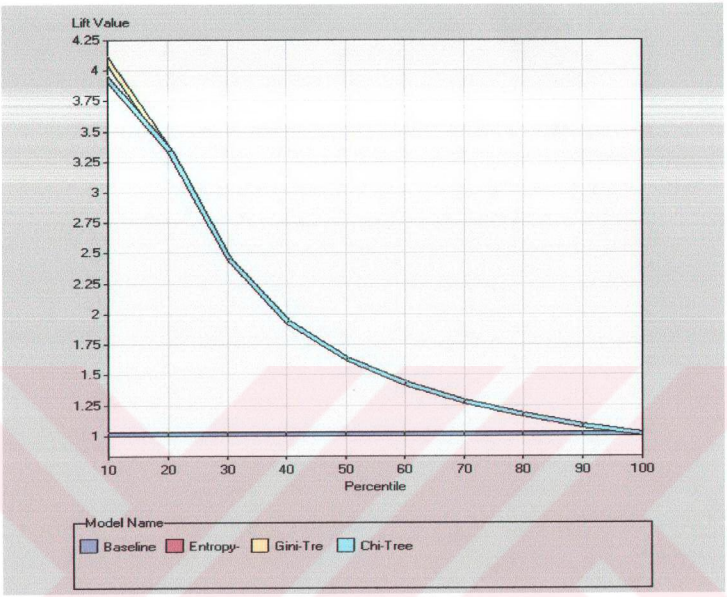
İkili ayraç kullanan sınıflandırma ağaçlarında Gini kriteri bir dalda (branch) en fazla hedef sınıfı yaratmaya eğilimli iken Entropi kriteri ayraç dengesi kurmaya eğilimlidir. Gini ve Entropi kriterleri dal sayısı arttıkça değerleri artan kriterlerdir. Çoklu ayraçlardan çok ikili ayraçlar için uygundurlar.

iii. Ki-kare ayraç kriteri : Pearson Ki-kare testi de karar ağacında oluşan düğümlerin saflığını ölçer. Ki-kare değeri gözlenen değerler ile olması gereken değerler arasındaki farkların karelerinin toplamından oluşur. Testin serbestlik derecesi (degree of freedom);

$$v=(r-1)^* (B-1)$$

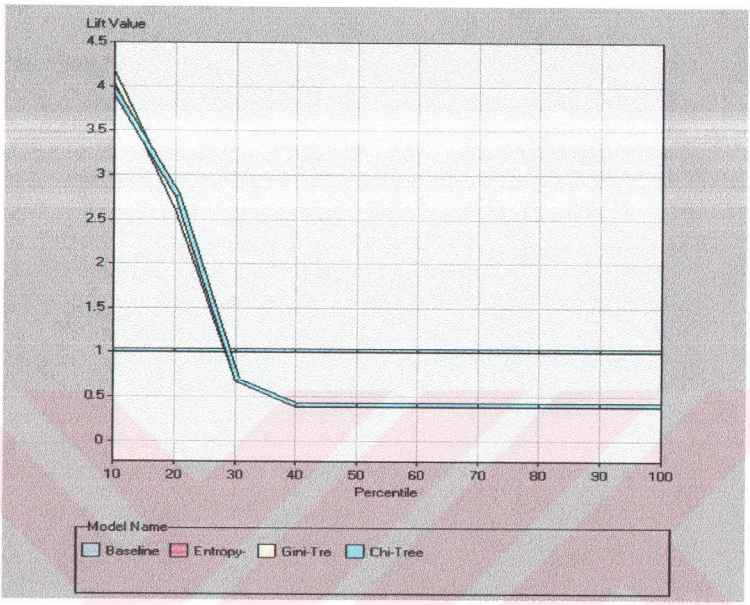
(r:hedef seviyesi (örneğimizde bu değer 2'dir. 0 ve 1 den oluşmaktadır.) B: dal sayısı şeklinde ifade edilmektedir. Formüldeki r ve B tablonun boyutunu temsil etmektedirler. Ki-kare testi sonucu ortaya çıkan değerin p-değeri (p value) incelenir. P değerinin belirlenen güven aralığında sıfıra yaklaşması düğümün saflığının arttığının bir göstergesidir.)

Çalışmada üç ayraç kriteri de değerlendirmeye alınmıştır. Her bir ayraç kriteri oluşturduğu ağaç diyagramı içerisinde en iyi ayraçları belirlemiştir. En iyi karar-ağacı analiz sonucunu veren modeli belirlemek için üç modelinde asansör grafiğini incelemek gerekmektedir. Şekil 3.11'de Gini, Entropi ve Ki-kare testi ayraç kriterlerine göre oluşturulmuş karar ağacı modellerinin asansör grafikleri karşılaştırılmalı olarak verilmektedir.



Şekil 3.11 Gini, Entropi ve Ki-kare testi ayraç kriterlerine göre oluşturulmuş karar ağacı modellerinin karşılaştırmalı kümülatif asansör grafikleri

Şekil 3.11'deki asansör grafiklerinde Entropi ve Ki-kare testi kriterlerine göre oluşturulan karar ağacı modelleri aynı sonucu vermektedir. Gini kriterine göre oluşturulan karar ağacı modeli ise diğer iki modele göre daha iyi bir sonuç üretmektedir. Gini kriterine göre oluşturulan karar ağacı modeli tahmini, ilk %10'luk dilimde veri madenciliği modeli oluşturmadan raslantısal olarak seçilen örneklem üzerinde yapılan kredi başvurusu reddi tahminine göre 4.25 kat daha iyi sonuç vermiştir. Bu oran Entropi ve Ki-kare karar ağaçlarında yaklaşık olarak 3.85'dir. %20'lik dilimden sonra her üç model de benzer tahminler üretmektedir. Başvuru reddinin nerede kesilmesi (cut off point) gerektiği ise şekil 3.12'de kümülatif olmayan karşılaştırmalı asansör grafiğinde açıkça gözlenmektedir.



Şekil 3.12 Gini, Entropi ve Ki-kare testi ayraç kriterlerine göre oluşturulmuş karar ağacı modellerinin karşılaştırmalı kümülatif olmayan asansör grafikleri

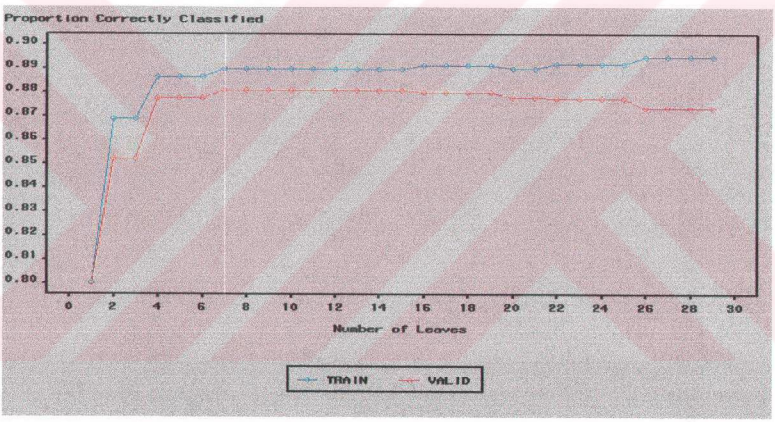
Şekil 3.12'deki grafiğe göre %29'luk dilimden sonraki müşterilerin kredi başvurularının reddi bankanın zarar etmesine neden olacaktır.

Karar ağaçları modelleri arasında en iyi model olarak belirlenen Gini karar ağacının değerlendirme veri kümesine göre oluşan karşılaştırma matrisi (confusion matrix) tablo 3.4'te verilmiştir. Koyu ve altı çizili olan değerler modelin performansının yüksek olduğunu ifade etmektedir.

Tablo 3.4. Gini endeksini kullanan karar ağacı karşılaştırma matrisi

	1	0	Toplam
1	**212 (%59)**	145(%41)	357
0	68 (%5)	**1363 (%95)**	1431

Şekil 3.13'de Gini kriterine göre oluşturulan karar ağacı modelinde yaprak oluşumunu eğitim ve değerlendirme kümeleri çerçevesinde ifade eden bir grafik bulunmaktadır. Bu grafiğe göre Gini karar ağacı 7 yapraktan oluşmaktadır.



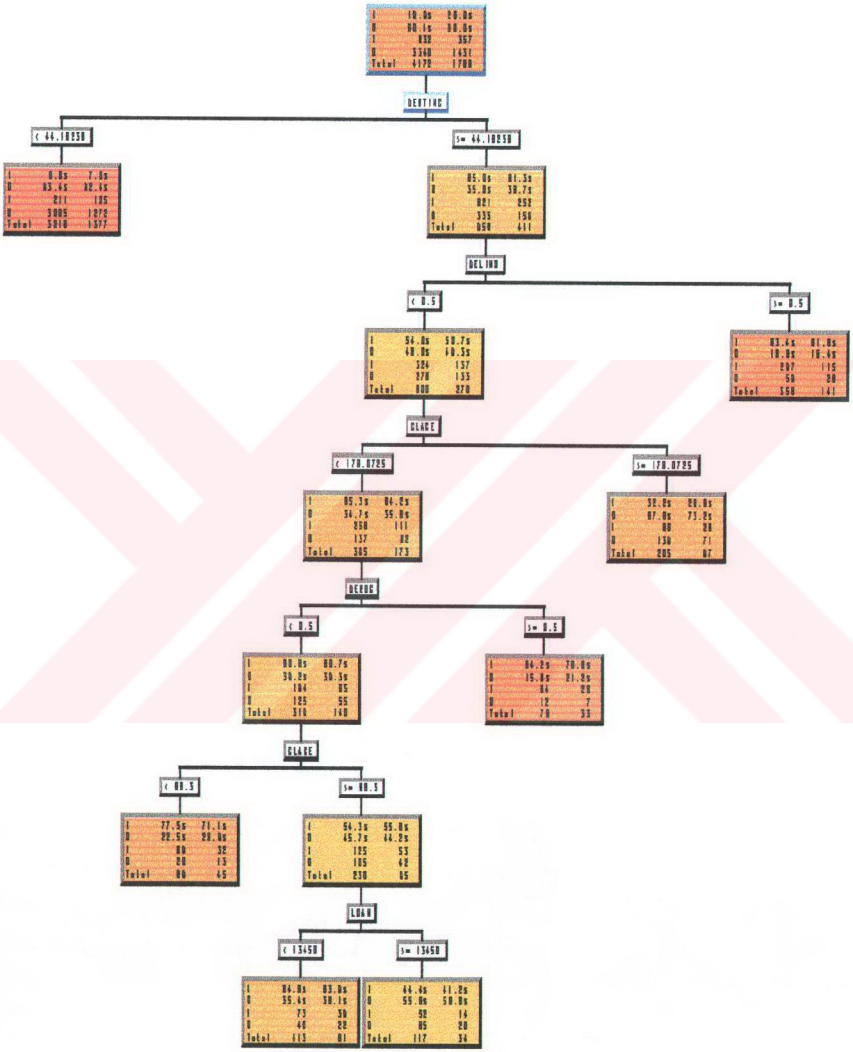
Şekil 3.13. Yaprak sayısının belirlenmesi

Yaprakların belirlenmesinde bazı değişkenler diğer değişkenlere göre daha fazla belirleyici olmaktadır. Tablo 3.5'te oluşturulan Gini karar ağacı modelinde yaprakları belirleyici değişkenler önem sırasına göre verilmektedir.

Tablo 3.5. Yaprak oluşumunu belirleyen önemli değişkenler

Değişken ismi	Önem değeri	Modeldeki rolü
DEBTINC	1.0000	girdi
DELINQ	0.4773	girdi
CLAGE	0.4765	girdi
LOAN	0.2578	girdi
DEROG	0.2373	girdi

Tablo 3.5'de sonuca göre karar ağacın oluşmasında en belirleyici değişken müşterinin borç-gelir oranıdır. Ikinci en önemli değişken müşterinin ödemediği kredi sayısıdır. Üçüncü diğer bir değişken ilk yapılan kredi talebinden sonra ay bazında geçen toplam süredir. LOAN olarak adlandırılan önem sırasında dördüncü olan değişken ise müşterinin bankadan talep ettiği kredi miktarıdır. Önem sırasında son sırada müşteriye gönderilen borç ihbar belgesi sayısı bulunmaktadır. Tabloda olmayan değişkenler ise karar ağaç diyagramının oluşmasında rol oynamamıştır. Şekil 3.14' de oluşturulan Gini karar ağacı diyagramı bulunmaktadır.



Şekil 3.14 Gini karar ağacı diyagramı

Veri madenciliği tahmin modellerinden karar ağacı tekniğinin tercih edilmesinin en önemli nedeni çıkacak olan sonuçların anlaşılır ve kolay yorumlanabilir olmasından kaynaklanmaktadır. Şekil 3.14'deki Gini karar ağacı diyagramını bir kurallar dizisi olarak incelediğimizde sonuçlar üzerine yorum yapmak daha kolay olacaktır. Oluşan kuralları maddeler halinde yazmak gerekirse;

i. **Eğer DEBTINC < 44.18 ise** müşterinin borcunu ödeme ihtimali %93.4, **ödememe ihtimali %6.6' dır.**

ii. **Eğer DEBTINC değişkeni >=44.18 ve DELINQ>=0.5** ise müşterininin borcunu ödeme ihtimali % 16.6, **ödememe ihtimali ise %83.4'tür.**

iii. **Eğer DEBTINC değişkeni >=44.18 ve DELINQ< 0.5 ve CLAGE>=178.0725** ise müşterininin borcunu ödeme ihtimali % 67.8, **ödememe ihtimali ise %32.2'dir.**

iv. **Eğer DEBTINC değişkeni >=44.18 ve DELINQ< 0.5 ve CLAGE<178.0725 ve DEROG>=0.5** ise müşterininin borcunu ödeme ihtimali % 15.8, **ödememe ihtimali ise %84.2'dir.**

v. **Eğer DEBTINC değişkeni >=44.18 ve DELINQ< 0.5 ve CLAGE<178.0725 ve DEROG<0.5 ve CLAGE<66.3** ise müşterininin borcunu ödeme ihtimali % 22.5, **ödememe ihtimali ise %77.5'dir.**

vi. **Eğer DEBTINC değişkeni >=44.18 ve DELINQ< 0.5 ve CLAGE<178.0725 ve DEROG<0.5 ve CLAGE>=66.3 ve LOAN<13450** ise müşterininin borcunu ödeme ihtimali % 35.4, **ödememe ihtimali ise %64.6'dır.**

vii. **Eğer DEBTINC değişkeni >=44.18 ve DELINQ< 0.5 ve CLAGE<178.0725 ve DEROG<0.5 ve CLAGE>=66.3 ve LOAN>=13450** ise müşterininin borcunu ödeme ihtimali % 55.6, **ödememe ihtimali ise %44.4'dür.**

Kuralları inceleğimizde potansiyel müşterilerin mevcut durumlar karşısında borçlarını ödememe olasılık değerleri açıkça görülmektedir. Başvuruda bulunan müşterilerin karar ağacını oluşturan DEBTINC, DELINQ, DEROG, CLAGE ve LOAN değerleri hesaplandıktan sonra borcunu ödememe olasılığı kolayca hesaplanabilir.

Fakat müşterilerin başvurularının reddine karar verirken belli bir karar eşik değeri (decision threshold) seçmek gerekmektedir. Bu değeri aşan müşterilerin başvurularının reddine, altında kalan müşterilerin ise başvurularının kabulune karar verilir. Bu eşik değeri teorik ve pratik anlamda elde edilebilir değerlerdir. Her iki yaklaşımda da eşik değer hesaplamasında yanlış sınıflandırma maliyeti kullanılmaktadır. Teorik yaklaşımda Bayes kuralı kullanılır. Karar teorisine göre optimal eşik değeri ;[103]

$$\theta = 1 / \left(1 + (\text{yanlış negatif maliyeti} / \text{yanlış pozitif maliyeti})\right)$$

Yanlış negatif maliyeti: borcunu ödeyemeyecek durumdaki müşterileri ödeyebilir statüsünde değerlendirme maliyeti

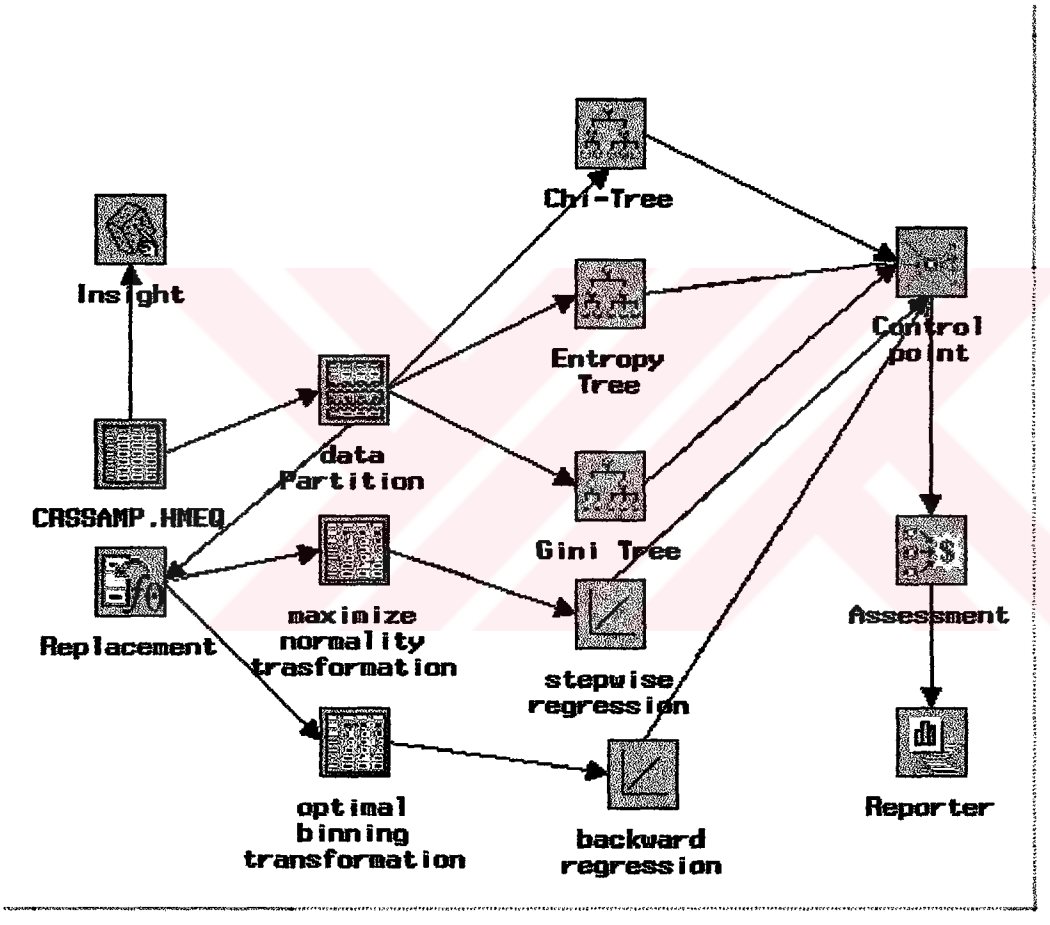
Yanlış pozitif maliyeti: borcunu ödeyecek durumdaki müşterileri ödeyemeyecek statüsünde değerlendirme maliyeti

Bu değerin tespiti bankaların kredilere uyguladıkları faizlere göre değişiklik göstermektedir. Başvuruda bulunan müşterinin, bayes kuralına göre hesaplanan optimal eşik değerinin altında veya üstünde olması durumuna göre başvurusunun reddine veya kabulune karar verilir.

[103] Doug Wielenga, Bob Lucas ve Jim Georges, s:125

3.2.4.4. Tahmin ve Denetimli Sınıflandırma Modeli Uygulamasının Yorumu

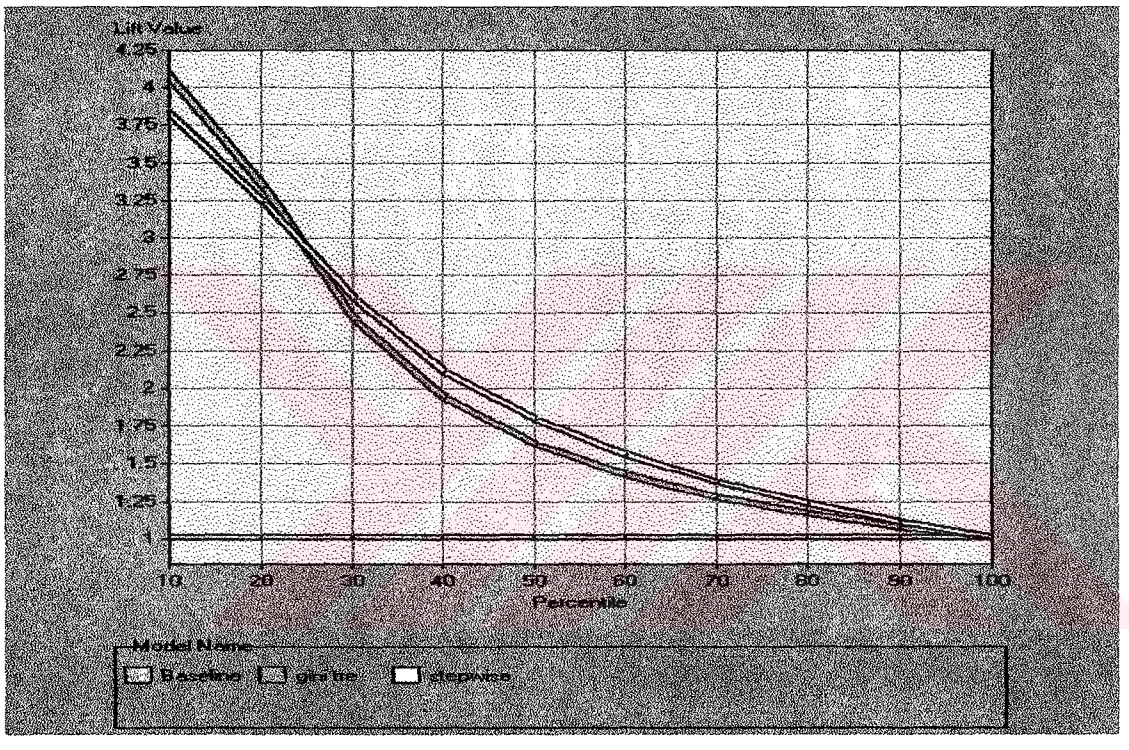
Çalışmamızda SAS Enterprise Miner 4.2 kullanılarak oluşturulan tahmin modeli veri madenciliği diyagramı şekil 3.15'de verilmiştir.



Şekil 3.15. SAS Tahmin Modeli Veri Madenciliği Diyagramı

Yapılan regresyon ve karar ağacı analizi sonucunda tahmin modelinin performansını arttıran veri madenciliği tekniğini bulabilmek için tekniklerin karşılaştırılmalı asansör grafiklerine bakmak gerekecektir. Bu amaçla kümülatif ve kümülatif olmayan karşılaştırmalı asansör grafiklerine, toplam geri dönüş yüzde (captured response) grafiklerine ve tekniklerin hedef değişkeni doğru sınıflandırma

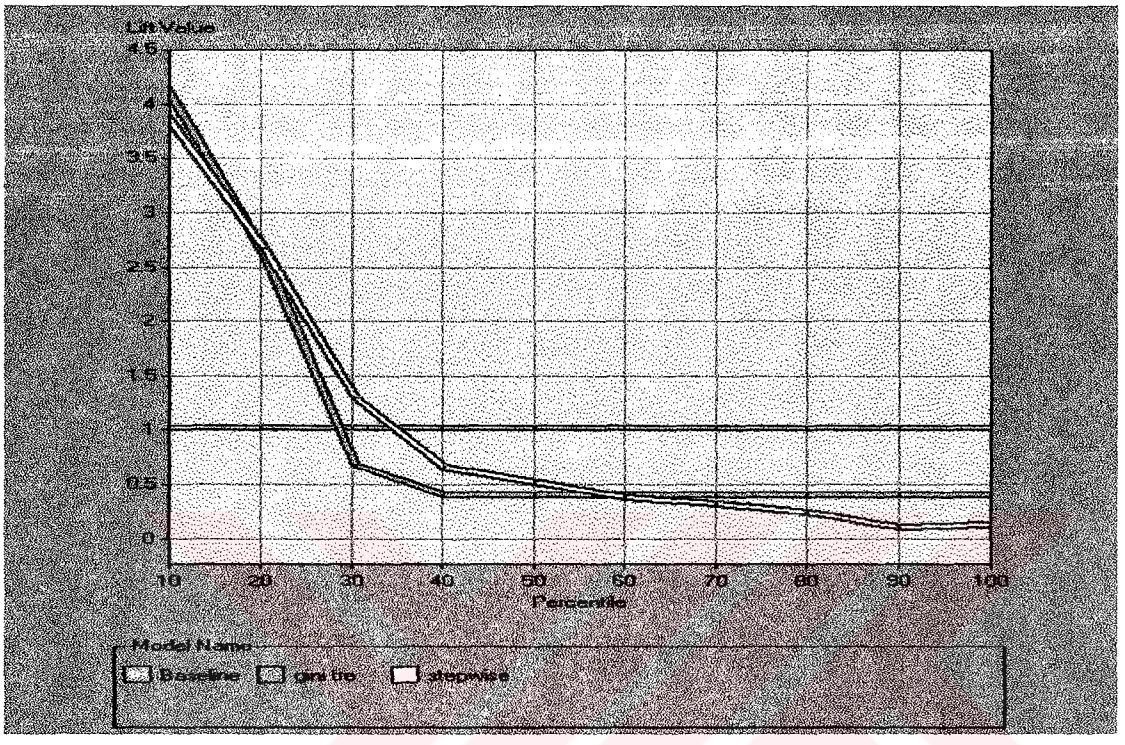
yüzdelerine bakılacaktır. Daha önce tekniklerin kendi içlerinde performans analizleri yapıldığında regresyon analizinde adım adım regresyon analizi ile karar ağacı analizinde Gini karar ağacı analizinin sonuçları başarılı bulunmuştu. Bu sonuçlardan yola çıkarak mevcut probleme dair en iyi tahmin modelini sunan analizi bulabilmek için Gini karar ağacı analizi ve adım adım regresyon analizi karşılaştırılacaktır.



Şekil 3.16. Gini karar ağacı ve adım adım regresyon analizi karşılaştırmalı kümülatif asansör grafiği

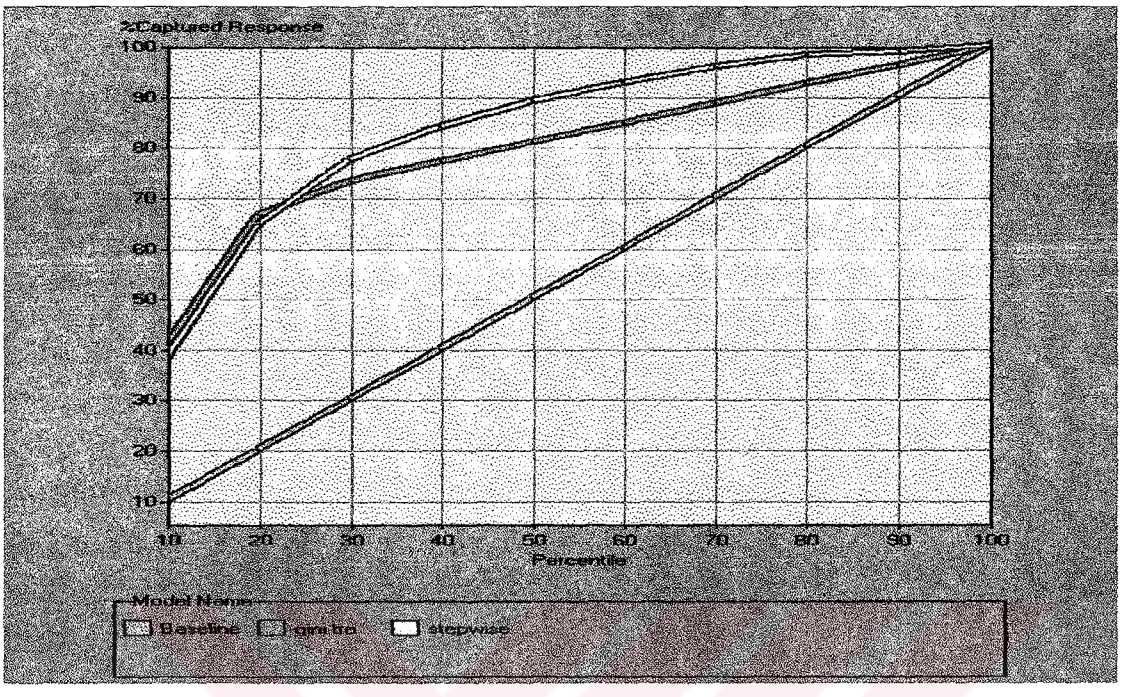
Şekil 3.16'daki karşılaştırılmalı kümülatif asansör grafiğine göre %20'lik dilimde Gini karar ağacı analizi adım adım regresyon analizine göre daha iyi bir performans sağlamıştır. %10'luk dilimde Gini karar ağacı analizinin asansör değeri 4.1 iken adım adım regresyon modelinin asansör değeri 3.8'de kalmıştır. %20'nci dilimde ise analiz sonuçları birbirine yakınlaşmıştır. Gini karar ağacı analizinde asansör değeri 3.25 olurken adım adım regresyon analizinin asansör değeri 3.3 sonucu vermiştir. %25'lik dilimden sonra adım adım regresyon analizinin performansı Gini karar ağacı analizine göre iyileşme göstermektedir. Tekniklerin performanslarına dair daha net bir açıklama

yapabilmek için tekniklerin karşılaştırmalı kümülatif olmayan asansör grafiklerine ve toplam geri dönüş yüzdelerine bakmak gerekecektir.



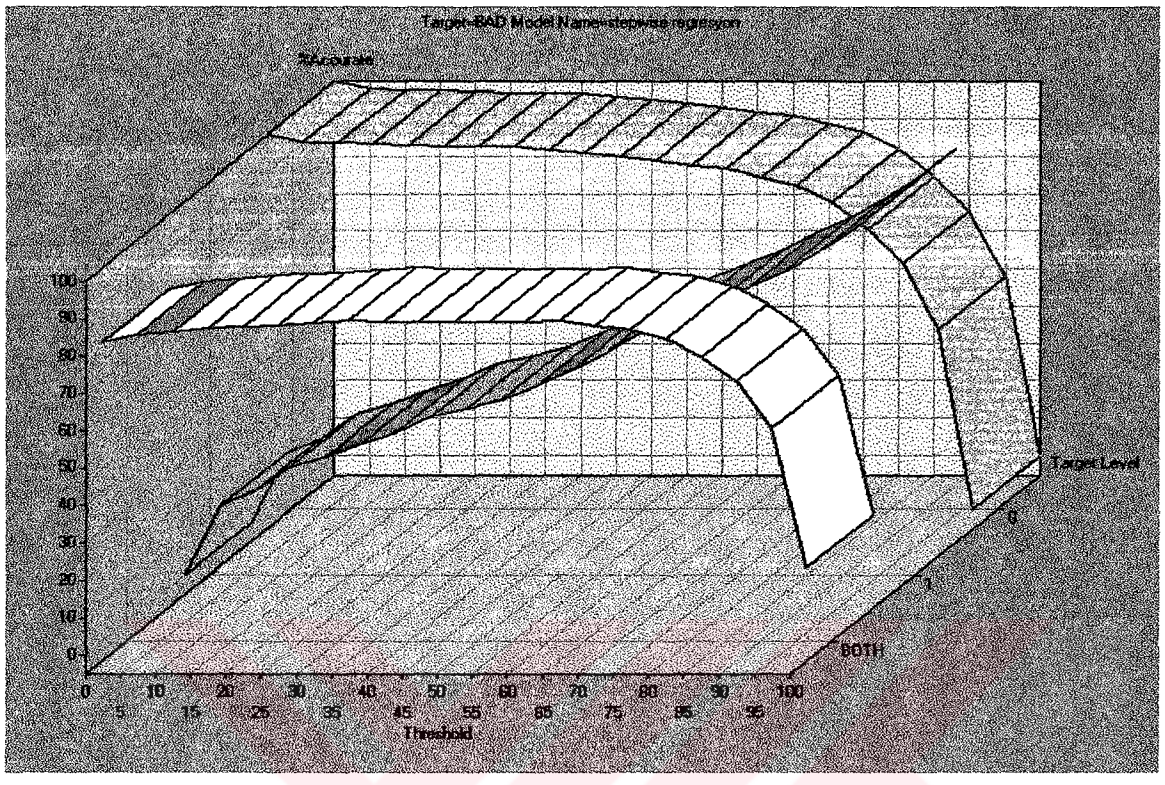
Şekil 3.17. Gini karar ağacı ve adım adım regresyon analizi karşılaştırmalı kümülatif olmayan asansör grafiği

Şekil 3.17'deki grafiği incelediğimizde, Gini karar ağacının kümülatif olmayan asansör grafiğine göre yaklaşık %30'luk dilimden sonraki müşterilerin başvurularının reddi bankanın zarar etmesine yol açacaktır. Bu amaçla Gini karar ağacında optimal çözüm, baştaki (skorlama sonucu oluşan sıralama) ilk %30'luk dilimdeki müşterilerin başvurularının reddine karar vermektir. Adım adım regresyon analizine göre ise optimal çözüm ilk %35'lik dilimdeki müşterilerin başvurularının reddine karar vermektir. Kümülatif olmayan asansör grafiklerinden de hangi tekniğin performansının daha iyi olduğu çok iyi anlaşılamamaktadır. Şekil 3.18'deki geri dönüş yüzdelerini veren grafiği incelediğimizde ise adım adım regresyon modelinin performansının daha fazla pozitif geri dönüşü verdiğini görmekteyiz.

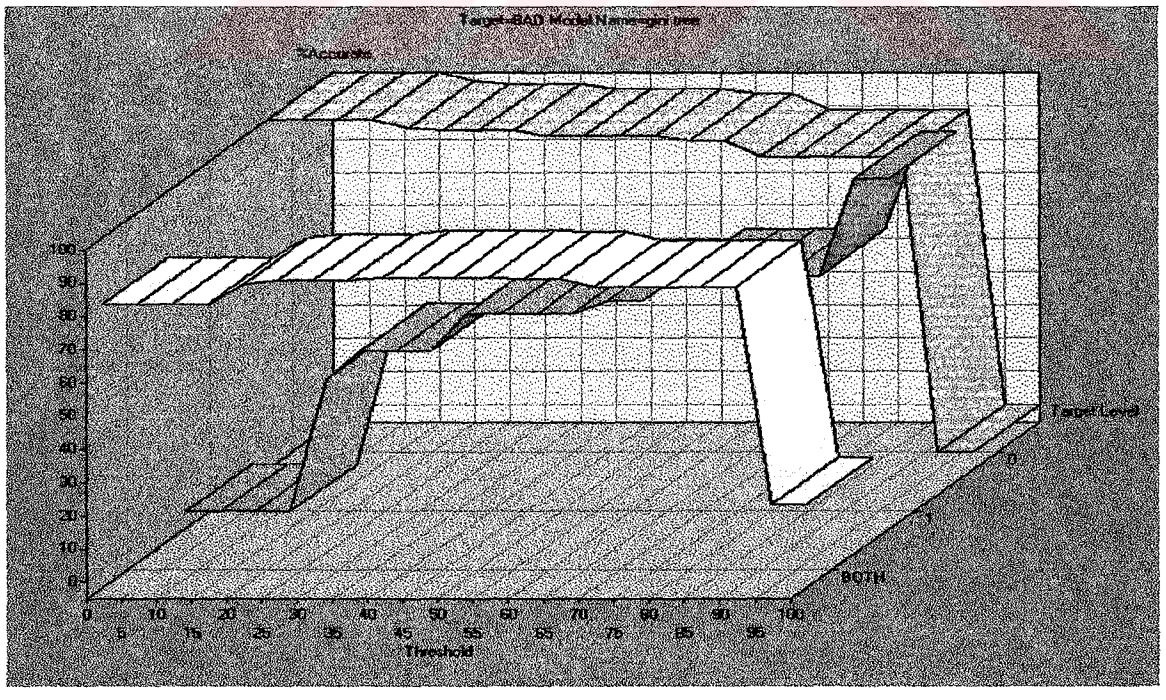


Şekil 3.18. Gini karar ağacı ve adım adım regresyon analizi karşılaştırmalı toplam geri dönüş yüzde grafiği

Şekil 3.18'de açıkça görüldüğü gibi genel olarak adım adım regresyon modelinin performansı Gini karar ağacı analizine göre daha iyidir. Eşik değer olan ilk %30'luk dilimde müşterinin başvurularının reddine karar verildiğinde veri madenciliği tekniği uygulamadan başvuruda bulanacak ve gelecekte borcunu ödemeyecek müşterilerin %30'una, Gini karar ağacı analizinde %75'ine, adım adım regresyon analizinde ise %79'una ulaşılacaktır. Son olarak eşik değer göz önüne alınarak analizlerin doğru sınıflandırma yapma performanslarına göre karşılaştırma grafiği şekil 3.19 ve 3.20'de verilmiştir.



Şekil 3.19 Adım adım regresyon analizinde eşik değere bağlı doğru sınıflandırma oranı



Şekil 3.20 Gini karar ağacı analizinde eşik değere bağlı doğru sınıflandırma oranı

Şekil 3.19 ve 3.20'deki grafikleri incelediğimizde, adım adım regresyon modelinin kötü (borcunu ödemeyecek) ve iyi müşteri (borcunu ödeyecek) sınıflandırma oranlarının Gini karar ağacı analizine göre daha iyi sonuç verdiğini görmekteyiz. Yapılan analizler ve incelen grafikler sonucu oluşturulan veri madenciliği tahmin modelinde genel anlamda adım adım regresyon analizinin performansı Gini karar ağacı analizine göre daha iyi sonuç vermiştir.

3.2.5. Kümeleme Analizi Uygulaması

Bankacılık sektöründe çoğu veri madenciliği projesi bölümlendirme analizi (segmentation analysis) ile başlamaktadır. Segmentasyon analizi ile müşteri yapısının anlaşılması ve ortak özellikte olan müşterilerin gruplandırılması sonucu oluşacak müşteri segmentlerine farklı pazarlama taktikleri ile yaklaşılması amaçlanmaktadır. Bu sayede uzun vadede amaç yüksek karlılık düzeyindeki müşterilere daha iyi hizmet vermek, daha verimli pazarlama araçlarıyla karlılığı arttırmak ve müşteri memnuniyeti ve bağımlılığı (customer satisfaction and loyalty) oluşturmaktır.

Segmentasyon analizine başlamadan önce bu analizin sürekli bir süreç olduğunu anlamak gerekmektedir. Bir başka önemli nokta ise müşteri segmentasyonunun doğru ve yanlış bir yolu olmadığının, asıl amacın işletme değeri (business value) olan segmentler yaratmak olduğunun anlaşılmasıdır.

Kümeleme analizi denetimsiz segmentasyon analizidir (unsupervised segmentation analysis). Denetimsiz olarak adlandırılmasının nedeni, bölüm 1.3.3.1'de de belirtildiği gibi tahmin edilmesi gereken bir hedef değişkenin olmamasından kaynaklanmaktadır.

3.2.5.1. Kümeleme Analizi Uygulamasında Değişkenlerin Tanımlanması

Kümeleme analizine başlamadan önce ilk yapılması gereken, değişkenlerin iyi tanımlanmasıdır. Değişkenlerden bazıları segmentlerin oluşturulmasında önemli rol oynayacağından bu değişkenler aktif (active) olarak tanımlanmıştır. Diğer değişkenler ise segmentler oluşturulduktan sonra, segmentlerin işletme değeri olduğunun anlaşılmasında rol oynayacağından tanımlayıcı (descriptive) olarak adlandırılmıştır. Tanımlayıcı değişkenler kümeleme analizi sürecinde kullanılmayacaktır, bunun yerine segmentlerin profillendirilmesi ve değerlendirilmesi sürecinde rol oynayacaklardır.

Aktif ve tanımlayıcı değişkenlerin belirlenmesi sürecinde değişkenlerin birbirleriyle ne kadar ilişkili olduğunu anlayabilmek için korelasyon analizi yapmak gerekmektedir. Bu analiz sonucu yüksek korelasyona sahip değişkenlerin birarada analize alınmaması kümeleme analiz sürecini çok farklı etkileyecektir. Korelasyon analizinde korelasyon değeri 0.5'ten fazla çıkan değişkenler analiz sürecinde beraber değerlendirilmeyecektir. Tablo 3.6'da değişkenlerin korelasyon değerleri bulunmaktadır.

Tablo 3.6 Değişkenlerin Korelasyon Değerleri

	Clage	Clno	Debtinc	Delinq	Derog	Loan	Mortdue	Ninq	Value	Yoj
Clage	1.00	0.24	-0.05	0.22	-0.08	0.09	0.14	0.11	0.17	0.20
Clno	0.24	1.00	0.18	0.16	0.06	0.07	0.32	0.08	0.26	0.02
Debtinc	-0.05	0.18	1.00	0.05	0.02	0.08	0.15	0.14	0.13	-0.05
Delinq	0.22	0.16	0.05	1.00	0.21	-0.03	-0.001	0.07	-0.01	0.04
Derog	-0.08	0.06	0.02	0.21	1.00	-0.001	-0.04	0.17	-0.04	-0.06
Loan	0.09	0.07	0.08	-0.03	-0.001	1.00	0.23	0.04	0.33	0.10
Mortdue	0.14	0.32	0.15	-0.001	-0.04	0.23	1.00	0.03	**0.87**	-0.08
Ninq	0.11	0.08	0.14	0.07	0.17	0.04	0.03	1.00	-0.004	-0.07
Value	0.17	0.26	0.13	-0.01	-0.04	0.33	**0.87**	-0.004	1.00	0.007
Yoj	0.20	0.02	-0.05	0.04	-0.06	0.10	-0.08	-0.07	0.007	1.00

Tablo 3.6'yı incelediğimizde sadece MORTDUE ve VALUE değişkenlerinin korelasyon değerlerinin çok yüksek çıktığını görmekteyiz. Bu nedenle kümeleme analizinde bu iki değişkenin beraber analiz edilmemesi gerekmektedir. Değişkenlerin bankacılık sektöründeki tanımları gereği ve korelasyon analizi sonucu kümeleme analizinde değişkenlerin rolü tablo 3.7'de ayrıntılı biçimde verilmiştir.

Tablo 3.7. Değişkenlerin Kümeleme Analizindeki Rolleri

Değişken İsmi	Değişken Tanımı	Değişken Rolü
BAD	1=Borcunu ödememiş 0=Borcunu ödemiş	Tanımlayıcı
REASON	HomeImp:ev yenileme Debcon: Borç konsolidasyonu	Tanımlayıcı
JOB	6 meslek kategorisi	Tanımlayıcı
LOAN	Talep edilen kredi miktarı	Aktif
MORTDUE	Konut ipotek değeri	Tanımlayıcı
VALUE	Mevcut mal varlığının bugünkü değeri	Aktif
DEBTINC	Borç gelir oranı	Aktif
YOJ	Müşterinin mevcut mesleğinde geçirdiği toplam sene	Tanımlayıcı
DEROG	Borç ihbar belgesi sayısı	Aktif
CLNO	Kredi başvuru sayısı	Aktif
DELINQ	Ödenmeyen kredi sayısı	Aktif
CLAGE	İlk yapılan kredi başvuru süresinden itibaren ay bazında geçen toplam süre	Tanımlayıcı
NINQ	Kredi soruşturma sayısı	Tanımlayıcı

Kümeleme analizinde değişkenlerin rolleri belirlendikten sonra veride aktif değişkenlerde aykırı ve eksik değerlerin tespiti yapılmalıdır. Aktif değişkenlerde olası tüm aykırı değerler (dağılımda %1'den az ya da %99'dan fazla dilimlerde oluşan değerler) tespit edildikten sonra bunların eliminasyonu gerekmektedir. Aktif değişkenlerin dağılımının normalizasyonu amacıyla SAS Enterprise Miner 4.2'de değişken dağılımlarının normalizasyonunu hedefleyen transformasyon yöntemlerinden gruplandırma (quantile) transformasyonunun seçilmesi aykırı değerleri elimine etmektedir. Bu amaçla ayrıca bir aykırı değer filtreleme aracına gerek duyulmamıştır.

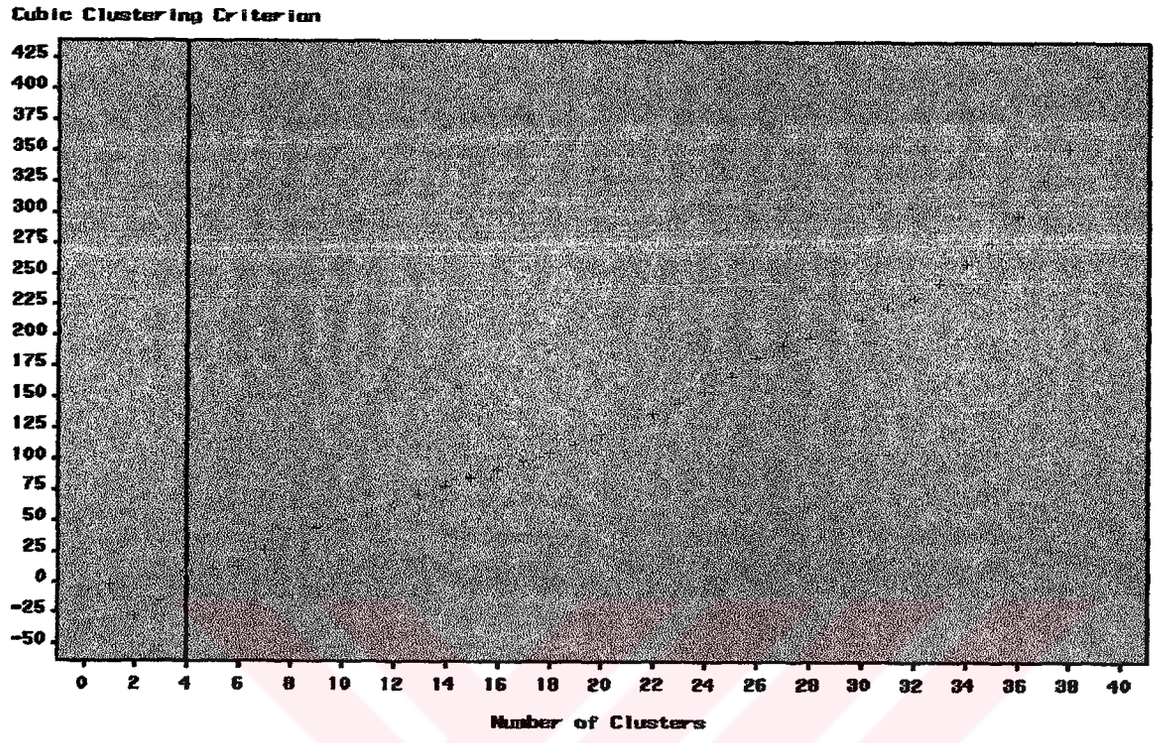
Kümeleme analizinde değişken dağılımlarının normal dağılıma uygun hale getirilmesi amacıyla yapılan transformasyon ve aykırı değerlerin elenmesinden sonra eksik değerlerin tespiti ve tamamlanması (replacement) süreci gelmektedir. Sınıflayıcı ve aralık ölçekli değişkenlerde eksik değerlerin tamamlanmasında diğer yöntemlere oranla en uygun araç ağaç yöntemi olarak olarak belirlenmiştir.

3.2.5.2 Kümeleme Analizinin Analitik Kontrolü

Verideki değişkenlerin aktif ve tanımlayıcı olarak tanımlanması, aykırı değerlerin elenmesi ve eksik değerlerin tamamlanmasından sonra kümeleme analizi sürecine başlayabiliriz. Kümeleme analizinde hiyerarşik olmayan yöntemlerden K-ortalamalar yöntemi kullanılmıştır.

Kümeleme analizi sürecinde ilk yapılması gereken değişkenlerin aynı ölçeğe indirgenmeleridir SAS Enterprise Miner 4.2'de değişkenleri aynı ölçeğe indirgeyebilmek için iki türlü standardizasyon aracı bulunmaktadır: Değişim aralığı (range) ve standart sapma (standart deviation). Uygulamada her iki standardizasyon aracı kullanılmış ve değişim aralığı standardizasyonu daha iyi bir sonuç vermiştir.

K-ortalamalar algoritmasında ilk yapılması gereken küme merkezlerinin seçilmesidir Küme merkezlerinin seçilmesinde en küçük kareler (least square) yöntemi en uygun yöntem olarak belirlenmiştir. K-ortalamalar algoritmasında merkez seçimi tekrarlı (iterative) olacağından tekrar sayısı 100 ile sınırlı tutulmuştur. SAS Enterprise Miner'da kümeleme analizi yapan CCC (Cubic Clustering Criterion) aracına göre ilk olarak küme sayısı otomatik olarak belirlenir. Şekil 3.20'de CCC grafiği görülmektedir.



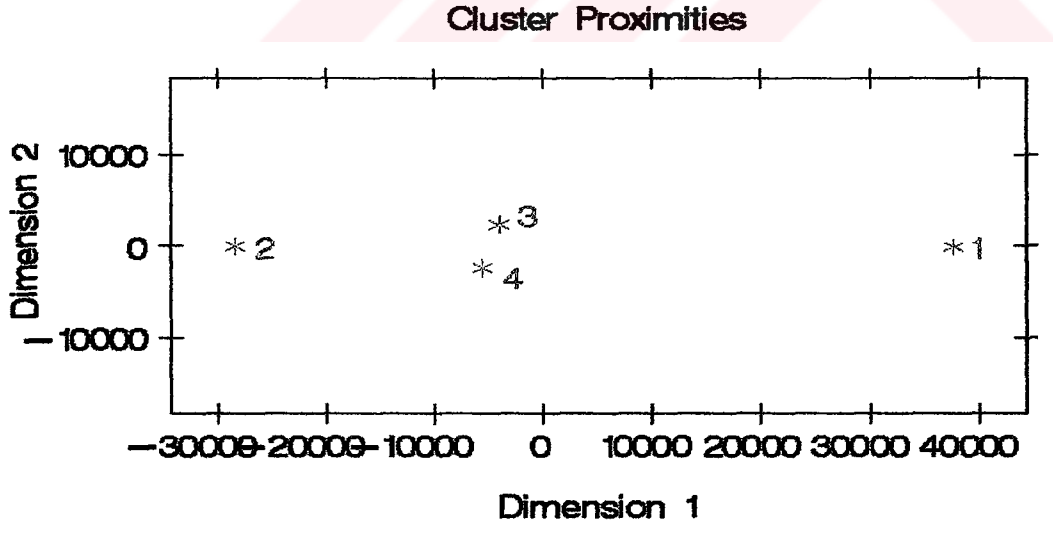
Şekil 3.21 Cubic Clustering Criterion grafiği

Şekil 3.21'deki CCC grafiğine göre otomatik olarak 4 küme belirlenmiştir. Normalde grafikte sıçramanın olduğu olası noktalar alternatif küme sayılarını ifade etmektedir. Şekildeki grafikte otomatik olarak belirlenen küme sayısından daha iyi bir alternatif olmadığı gözlenmektedir. Oluşturulan kümelerin istatistikleri tablo 3.8'de incelenmektedir.

Tablo 3.8. K-ortalamalar algoritması sonucu oluşan kümelerin istatistik değerleri

Kümeler	Küme Frekansları	Küme merkezinden maksimum uzaklık	En yakın küme	En yakın kümeye uzaklık
1	2340	1.147	2	0.781
2	2387	1.079	1	0.781
3	584	1.315	2	1.096
4	649	1.385	1	0.997

Tablo 3.8'de de gözlendiği gibi 1. ve 2. kümeler diğer kümelere oranla biraz daha büyüktür. 3. ve 4. kümelerin frekanslarının diğer iki kümeye oranının çok düşük olmamasından dolayı kümelerin frekans dağılımları kabul edilebilir niteliktedir. Bu nedenden dolayı en yakın küme kolonunda büyük kümelerin olduğu gözlenmektedir. Kümelerin birbirlerinden uzaklığı şekil 3.22'de daha net olarak incelenebilir.



Şekil 3.22. Kümelerin birbirine uzaklığı

Kümeleme analizinin analitik olarak geçerlenmesinde son olarak kümelerin oluşmasında etkili olan aktif değişkenlerin incelenmesi gerekmektedir. Tablo 3.9'da değişkenler önem sırasına göre verilmiştir.

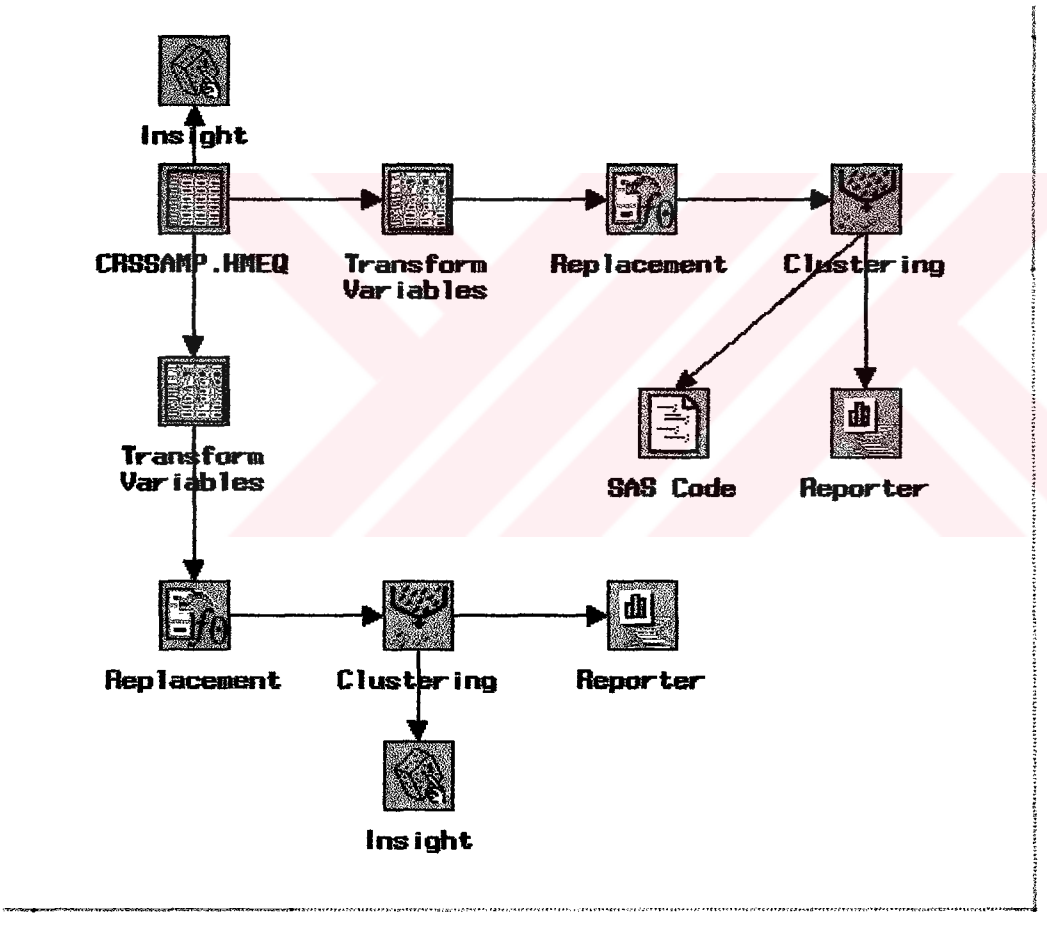
Tablo 3.9 Kümelerin oluşmasında etkili olan değişkenlerin önem sırası

Değişken ismi	Önem sırası	Değişken rolü	Değişken kodu
VALUE	1.000	Aralık ölçekli	
CLNO	0.8636	Sırasal	Quantile (CLNO)
DEROG	0.8319	İkili	Quantile (DEROG)
DELINQ	0.8130	Sırasal	Quantile (DELINQ)
LOAN	0.5433	Aralık ölçekli	
LOAN	0.3784	Sırasal	Quantile (LOAN)
VALUE	0.3617	Sırasal	Quantile (VALUE)
DEBTINC	0.0794	Aralık ölçekli	

Tablo 3.9'da görüldüğü gibi kümelerin oluşmasında etkin olan 8 adet değişken bulunmaktadır. Bu değişkenlerden 5 tanesi tranformasyon sonucu oluşan yeni değişkenlerdir. Kümelerin oluşmasında etkin olan değişkenler SAS Enterprise Miner 4.2 veri madenciliği programının kümeleme aracının altında çalışan karar ağacı mekanizması tarafından belirlenmektedir.

Kümeleme analizinin analitik olarak inşa edilmesinden sonra oluşan SAS grafiği şekil 3.23'de incelenmektedir. Şekil 3.23'den de görüleceği gibi kümeleme modelinin oluşturulması tahmin modeline oranla daha kolaydır. Kümeleme modelinde zor olan, oluşan kümelerin yorumlanmasıdır. Kümeleme analizi sonucunda işletme değeri olmayan kümelerin oluşması oluşturulan kümeleme modelinin başarısız olacağı anlamına gelmektedir. Tahmin modelinde ise oluşturulan modelin analitik yapısı ile yorumlanması doğru orantılıdır. İyi bir tahmin modelinin sonuçları da başarılı olacaktır.

Sonuç olarak, yapılan kümeleme analizi analitik olarak değerlendirildiğinde (küme merkezlerinin birbirinden maksimum uzaklığı, birbirine en yakın kümelerin incelenmesi, küme içi frekans dağılımı, kümelerin oluşmasında etken olan değişkenlerin incelenmesi) K-ortalamalar algoritmasının başarılı olduğu gözlenmektedir. Analitik kontrolden geçen kümeleme analizinin sonuçlarının genelleştirilebilmesi için oluşturulan kümelerin işletme değeri (business value) taşıması gerekmektedir.



Şekil 3.23 Kümeleme modeli SAS grafiği

3.2.5.3. Kümeleme Analizinin İşletme Değeri Kontrolü

K-ortalamalar algoritması sonucu oluşan kümelerin analitik olarak başarılı olması kümeleme analizini sonlandırmak için yeterli değildir. Başarılı bir kümeleme analizi süreci için, oluşturulan kümeler üzerinden işletme için efektif kararların alınabiliyor olması gerekmektedir. Bölüm 3.2.5.2'de analitik olarak değerlendirilen 4 kümenin işletme değeri açısından anlamlı sonuçlarının kontolü aşamasına kümeleme analizinde profilleme ve geçerleme süreci (profiling and validation process) denmektedir. Bu aşamada aktif değişkenlerin yanında tanımlayıcı değişkenlerde rol oynamaktadır. Kümelerin işletme değerini tam olarak anlayabilmek için sınıflayıcı ve aralık ölçekli değişkenlerin ayrı ayrı küme-içi istatistik değerlerinin incelenmesi gerekmektedir. İlk olarak sınıflayıcı değişkenlerinin küme içi yüzdesel değerleri tablolar halinde incelenecektir.

Tablo 3.10 BAD (Ödeme Durumu) değişkeninin küme içi yüzdesel değerleri

BAD	1.küme (%)	2.küme(%)	3.küme(%)	4.küme(%)	GENEL(%)
0	**92.31**	83.54	56.35	51.63	80.1
1	7.69	16.46	43.65	**48.37**	19.9

Tablo 3.10'a göre borcunu ödeyen müşteriler 1. kümede yüzdesel olarak ağırlıktadır. Genel ortalamayla karşılaştırıldığında borcunu ödeyenlerin yüzdesi 1. ve 2. kümelerde ortalamanın üstünde değerler almıştır. Borcunu ödemeyenler ise özellikler 4. kümede gruplanmıştır. 3. ve 4. kümelerin borcunu ödemeyenler cinsinden yüzdesel değerleri ortalamanın üstünde kalmıştır.

Tablo 3.11 JOB (İş Kategorisi) değişkeninin küme içi yüzdesel değerleri

JOB	1.küme(%)	2.küme (%)	3.küme(%)	4.küme(%)	GENEL
Müdür	15.26	8.29	16.24	14.88	13.5
Ofis İşi	18.21	[illegible]	8.12	16.74	16.7
Diğer	**31.92**	56.06	**49.75**	**41.40**	**42**
Müdür Yardımcısı	23.50	13.61	15.23	21.86	22.05
Satış	1.92	1.24	3.05	0.93	1.9
Kendi işi	4.10	2.23	[illegible]	4.19	3.4

Tablo 3.11'i incelediğimizde ve genel dağılımla karşılaştırdığımızda mesleği müdür olan müşterilerin 3. kümede, ofis işi olan ve diğer işlerle ilgilenen müşterilerin 2. kümede, müdür yardımcısı olan müşterilerin 1. kümede, satış ve kendi işi ile ilgilenen müşterilerin ise 3. kümede gruplandığını görebiliriz. Bunun yanında kümesel olarak incelediğimizde ise 1. ,2., 3. ve 4. kümelerde diğer işlerle ilgilenen müşterilerin ağırlıkta olduğunu görmekteyiz. Bunun nedeni popülasyonun genelinde diğer işlerle ilgilenen müşterilerin oranında oluşan fazlalıktan kaynaklanmaktadır.

Tablo 3.12 REASON (borç alma nedeni) değişkeninin küme içi yüzdesel değerleri

REASON	1.küme(%)	2.küme(%)	3.küme(%)	4.küme(%)	GENEL(%)
Borç konsolidasyonu	**75.38**	66.34	**75.63**	63.72	68.8
Ev Yenileme	24.62	33.66	24.37	**36.28**	31.2

Tablo 3.12'yi incelendiğinde kredi talebini özellikle borç konsolidasyonu nedeniyle yapanların 1. ve 3. kümelerde gruplandığı gözlenmektedir. Ev yenileme nedeniyle kredi talebinde bulunan müşteriler ise 4. kümede toplanmışlardır.

Sınıflayıcı değişkenlerinin küme içi değerlendirilmelerinden sonra, aralık ölçekli değişkenlerin küme içi değerlendirilmesi süreci aktif ve tanımlayıcı değişkenlere göre, tablo 3.13 ve tablo 3.14'te toplu olarak değerlendirilmiştir.

Tablo 3.13 Aktif-Aralık ölçekli değişkenlerin küme içi aritmetik ortalamaları

Değişken ismi	1. küme	2.küme	3.küme	4. küme	GENEL
CLNO	25.53	14.82	21.07	**27.06**	21.30
DEBTINC	**35.34**	32.35	34.31	**35.41**	33.78
DELINQ	0.10	0.07	0.37	**2.94**	0.45
DEROG	0.00	0.00	**1.80**	0.34	0.25
LOAN	**24489**	14149	20957	15267	18607
VALUE	**136538**	50933	99960	97110	101776

Tablo 3.13 incelendiğinde CLNO değişkeninin 4. kümede, DEBTINC değişkeninin küçük bir farkla 1. ve 4. kümelerde, DELINQ değişkeninin 4. kümede, DEROG değişkeninin 3. kümede, LOAN ve VALUE değişkeninin ise 1. kümede aritmetik ortalama değerlerinin yüksek çıktığını görmekteyiz.

Tablo 3.14 Tanımlayıcı-aralık ölçekli değişkenlerin küme içi aritmetik ortalamaları

Değişken ismi	1. küme	2.küme	3.küme	4. küme	GENEL
CLAGE	**201.93**	164.41	155.40	186.24	179.77
MORTDUE	**92382**	50933	72584	70827	73760
NINQ	1.13	0.78	**1.73**	1.23	1.19
YOJ	8.94	7.94	6.48	**9.86**	8.92

Tablo 3.14 incelendiğinde CLAGE ve MORTDUE değişkenlerinin 1. kümede, NINQ değişkeninin 3. kümede, YOJ değişkeninin ise 4. kümede aritmetik ortalamala değerlerinin yüksek çıktığını görmekteyiz.

Sınıflayıcı, aktif ve tanımlayıcı aralık ölçekli değişkenlerin genele göre dağılımlarını incelediğimizde oluşan 4 kümenin işletme değerlerinin olduğunu görmekteyiz. Analitik olarak geçerlenen kümeler işletme değeri açısından da geçerlenmektedir.

3.2.5.4. Kümelerin Profilleri

Analitik olarak ve işletme değeri açısından geçerlenen kümeleme analizi sonucu Tablo 3.10 ve tablo 3.14 arasındaki tablolardaki bilgileri derlediğimizde her kümenin kendine ait bir profili olduğu ortaya çıkmaktadır.

1. küme: Kümenin sınıf, aktif ve tanımlayıcı değişkenlerini incelediğimizde birinci kümenin en belirgin özelliklerini aşağıda bulunduğu gibi özetleyebiliriz.

i) Borcuna en sadık müşteri kümesi .
ii) Genelde müdür yardımcısı statüsünde olan müşterilerden oluşuyor. Diğer işlerle ilgilen müşterilerin sayısı da oldukça fazla.
iii) Kredi talebinde bulunmalarının sebebi borç konsolidasyonu.
iv) Borç gelir oranı en yüksek olan müşteri kümesi.
v) Talepte bulunulan kredi miktarı en yüksek olan müşteri kümesi.
vi) Mevcut mal varlığının bugünkü değeri oldukça yüksek ve dolayısıyla maddi durumları iyi olan müşterilerden oluşan bir küme.
vii) Kredi talebinde uzun zamandır bulunan müşterilerden oluşan bir küme.
viii) Hakkında borç ihbar belgesi sayısı hiç olmayan müşterilerden oluşan bir küme
ix) Kredi soruşturma sayısı ve ödenmemiş kredi sayısı çok az olan müşterilerden oluşan bir küme.

Birinci kümenin özelliklerini biraraya getirdiğimizde ortaya çıkan müşteri profili bankanın sadık olabilecek müşteri grubunu oluşturmaktadır. Uzun zamandır borç konsolidasyonu nedeniyle kredi talebinde bulunan, maddi durumu iyi, maddi durumuna kıyasla yüksek borç talebinde bulunan fakat çoğunlukla borcunu zamanında ödeyen, hakkında kötü raporu olmayan müşteriler, birinci kümenin müşteri profilini oluşturmaktadır. Müşteri ilişkileri yönetimi uygulanarak (CRM) bu segmentteki müşterilerin bankaya sadakat düzeyleri arttırılabilir. Bankanın adının bir marka olduğunu düşünürsek marka bağımlılığı oluşturmak için gerekli aksiyonlar alınabilir. Birinci küme banka için karlılığı arttırabilecek müşteri grubunu oluşturmaktadır.

2. küme : Kümenin sınıf, aktif ve tanımlayıcı değişkenlerini incelediğimizde ikinci kümenin en belirgin özelliklerini aşağıda bulunduğu gibi özetleyebiliriz.

i) Borcuna sadık olan bir müşteri kümesi.
ii) Diğer müşteri kümelerine oranla ofis işi yapanların çoğunlukta olduğu, diğer işlerle ilgilenen müşterilerin de ağırlıkta olduğu bir müşteri kümesi.
iii) Kredi talebinde bulunma sebebleri çoğunlukla borç konsolidasyonu olan bir müşteri kümesi.
iv) Borç gelir oranı en düşük olan müşteri kümesi.
v) Kredi talebi olarak belirtilen kredi miktarının en düşük olduğu müşteri kümesi.
vi) Mevcut mal varlığının bugünkü değerinin en düşük olduğu ve dolayısıyla maddi durumları çok iyi olmayan müşterilerden oluşan bir küme.
vii) Hakkında borç ihbar belgesi sayısı hiç olmayan müşterilerden oluşan bir küme
viii) Kredi soruşturma sayısı en az olan müşteri kümesi.

İkinci kümenin özellikleri biraraya getirdiğimizde ortaya çıkan müşteri profili; maddi durumu çok iyi olmadığı halde bankadan aldığı krediyi zamanında ödeyen, hakkında kötü raporu olmayan, kredi soruşturma sayısı en az olan, gelirine göre talep ettiği kredi tutarını en azda tutan ve çoğunlukla borç konsolidasyonu amacıyla kredi talebinde bulunan müşteri kümesi şeklindedir. Birinci küme ile kıyasladığımızda ikinci kümenin banka için yüksek karlılık düzeyi oluşturabilecek bir müşteri grubu olmadığını görmekteyiz. Fakat

uygulanacak verimli bir müşteri ilişkileri yönetimi ile bu segmentteki müşterilerin uzun vadede birinci segmentteki müşterilerin segmentine dahil edilmesi süreci olası görünmektedir. Bu amaçla bu segmentteki müşteri kümesinde uzun vadede marka bağımlılığı yaratmak amacıyla çeşitli promosyonel pazarlama aktivitelerinde bulunmak banka için olumlu sonuçlar üretebilecektir.

3. küme: Kümenin sınıf, aktif ve tanımlayıcı değişkenlerini incelediğimizde üçüncü kümenin en belirgin özelliklerini aşağıda bulunduğu gibi özetleyebiliriz.

i) Borcunu genel olarak ödeyen müşterilerden oluşan bir küme. Bunun yanında birinci ve ikinci kümeye oranla borcunu ödemeyen müşterilerin sayısında ciddi bir yükseliş gözleniyor.
ii) Meslek olarak müdürlük, satış ve kendi işi ile uğraşan müşterilerin ağırlıkta olduğu bir müşteri segmenti. Bu anlamda müşteri meslek profili çeşitlilik arzeden bir küme.
iii) Kredi talebi genelde borç konsolidasyonu amacıyla yapılıyor.
iv) Borç ihbar belgesi sayısı en fazla olan müşteri kümesi.
v) Birinci ve ikinci kümelere oranla ödenmeyen kredi sayısı yüksek olan bir müşteri kümesi.
vi) Borç gelir oranı genel ortalamanın üstünde olan bir küme.
vii) Bankadan talep edilen kredi miktarı genel ortalamanın üstünde olan bir müşteri kümesi.
viii) Mevcut mal varlığının bugünkü değerinin ortalamadan düşük olduğu ve dolayısıyla maddi durumları çok iyi olmayan müşterilerden oluşan bir küme.
ix) Hakkında kredi soruşturma sayısı en fazla olan müşteri kümesi.

Üçüncü kümenin özelliklerini birarada değerlendirdiğimizde, bu kümenin banka için zarar nedeni arzeden müşterilerden oluştuğu gözlenmektedir. Maddi imkanı çok iyi olmadığı halde bankadan çoğunlukla borç konsolidasyonu nedeniyle yüksek kredi talebinde bulunan ve borcuna sadık olmayan müşterilerin oluşturduğu bu kümede ayrıca müşteriler hakkında oldukça fazla sayıda kredi soruşturması da bulunmaktadır. Bu durum müşteriler hakkında adli takibin olduğu ve maddi durumlarının kötülüğünden

kaynaklanan bir geri ödemenin olmadığı anlamına gelmektedir. Borç gelir oranının yüksekliği ve ödenmemiş kredi sayısının çokluğu bu müşteri grubunun banka için olumsuz etkileri olduğu anlamına gelmektedir. Bu amaçla, bu segmentteki müşterilerin banka müşteri portföyünden çıkarılması gerekmektedir.

4. küme: Kümenin sınıf, aktif ve tanımlayıcı değişkenlerini incelediğimizde dördüncü kümenin en belirgin özelliklerini aşağıda bulunduğu gibi özetleyebiliriz.

i) Borcunu ödemeyen müşterilerin diğer kümelere oranla en fazla olduğu küme. Bunun yanında, kendi içinde borcunu ödeyenler fazlalıkta.

ii) Diğer iş kategorileri ile ilgilenen müşterilerin ağırlıkta olduğu bir küme.

iii) Kendi içinde borç konsolidasyonu nedeniyle kredi talebinde bulunan müşterilerin ağırlıkta olduğu, fakat diğer kümelere oranla ev yenileme nedeniyle kredi talebinde bulunanların ağırlıkta olduğu bir müşteri kümesi.

iv) Diğer kümelere oranla, bankadan kredi talebi sayısı daha yüksek olan bir müşteri kümesi.

v) Borç gelir oranı en yüksek olan müşteri kümesi.

vi) Müşterilerinin borç ihbar sayıları ortalamanın üstünde olan bir müşteri kümesi.

vii) Bankadan talep edilen tutarın ortalamanın altında olduğu bir müşteri kümesi.

viii)Mevcut mal varlığının bugünkü değerinin ortalamadan düşük olduğu ve dolayısıyla maddi durumları çok iyi olmayan müşterilerden oluşan bir küme.

ix) Kredi soruşturma sayısının ortalamanın üstünde olduğu müşterilerden oluştuğu bir küme.

x) Uzun zamandır kredi talebinde bulunan müşterilerden oluşan bir küme.

Son küme olan dördüncü kümenin özelliklerini biraraya getirdiğimizde ortaya çıkan müşteri profili; maddi durumları çok iyi olmadığı gibi bankadan talep edilen kredi tutarının diğer kümelere oranla düşük olduğu, fakat bunun yanında borç gelir oranı en fazla olan, uzun zamandır bankadan borç konsolidasyonu ve ev yenileme nedeniyle kredi talebinde bulunan, kredi talep sayısının diğer müştere oranla daha fazla olduğu, haklarında çıkarılan kredi soruşturma sayısının ve borç ihbar belgesi sayısının genel

ortalamanın üstünde olduğu şeklindedir. Bu bilgileri biraraya topladığımızda uzun vadede banka için zarar nedeni olabilecek müşterilerden oluştuğunu görmekteyiz. Yapılan profilleme sonucu bu müşteri kümesine yapılan yatırımın geri dönmeyeceği (ROI'nin çok düşük olacağı) sonucuna varılmaktadır. Bankadan talep edilen borç miktarının fazla olmaması nedeniyle yakın zamanda fark edilemeyecek olan zararın uzun vadede bankaya zarar vereceği göz önünde bulundurulursa bu müşterilerin banka müşteri portföyünden çıkarılması banka için olumlu sonuçlar getirecektir.

Küme-içi profillendirmeler sonunda kümelerin birbirleriyle karşılaştırılmaları yapıldığında birinci ve ikinci kümenin banka için olumlu sonuçlar ürettiğini görmekteyiz. Bunun yanında ikinci kümeye özel yapılabilecek pazarlama aktiviteleri ile bu gruptaki müşterilerin uzun vadede birinci kümeye dahil olabileceği anlaşılmaktadır. Üçüncü ve dördüncü kümeler borç ödeme durumları kötü olan müşterilerden oluşmaktadır. Gelir durumlarına oranla talep ettikleri borç müktarı da diğer iki kümeye oranla fazla olan üçüncü ve dördüncü kümenin banka müşteri portföyünden çıkarılması banka için olumlu sonuçlar üretecektir.

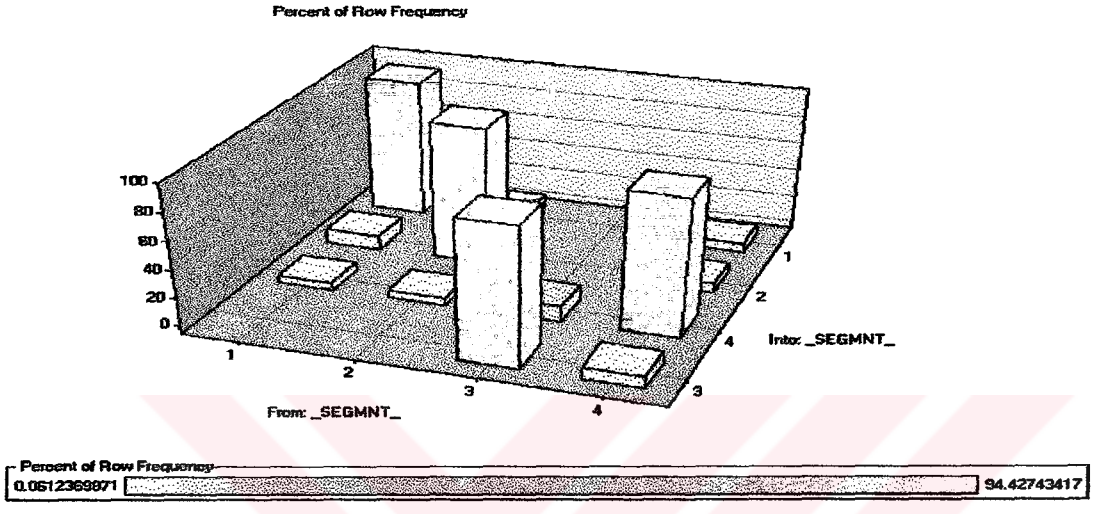
3.2.5.4 Kümeleme Analizinin Doğrulanması Süreci

Kümeleme analizi, çoğu veri madenciliği projesinde ilk yapılan modellemedir. Veri kümesinde benzer verileri gruplandıran kümeleme analizinin ardından kurulan tahmin modelleri oluşan kümelerin etkinliğini ölçmektedir.[104] Uygulamamızda, kümeleme analizi sonucu oluşan 4 adet kümeyi birbirinden ayırt eden değişkenleri belirlemek ve kümelerin etkinliğini ölçmek için, tahmin modellerinden adım adım regresyon tekniği kullanılmıştır. Kümeleme analizi sonucu oluşan küme sayısı ikiden fazla olduğundan dolayı en uygun regresyon tekniği multinomial regresyon olarak belirlenmiştir. Modelde, kümeleme analizi sonucu oluşan SEGMNT değişkeni tahmin edilmesi gereken hedef değişken olarak konumlandırılmıştır.

---

[104] Berry & Linoff, *Data Mining Techniques*, s:215

Şekil 3.24'te kümelerin etkinliğini belirlemede kullanılan multinomial adım adım regresyon analizinin karşılaştırma matrisi bulunmaktadır.



Şekil 3.24 Multinomial Adım adım Regresyon Tekniğinin Karşılaştırma Matrisi

Şekil 3.24'te incelenen regresyon karşılaştırma matrisinde, verinin değerlendirme kümesi baz alınmıştır. Şekilde bar grafiklerinin yüksekliğine denk gelen rakamlar tablo 3.15'te ayrıntılarıyla incelenmektedir.

Tablo 3.15 Multinomial regresyonun doğru sınıflandırma oranları

Kümeler/ayrışma yüzdeleri	1. küme	2. küme	3. küme	4. küme	Toplam
1. küme	**1542 (%94.43)**	90(%5.51)	0	1(%0.06)	**1633**
2.küme	88 (%5.28)	1574 (%94.36)	0	6(0.36	1668
3. küme	0	0	**384(%93.89)**	25(%6.11)	**409**
4. küme	7(%1.52)	9(%1.95)	16(3.46)	**430(93.07)**	462

Şekil 3.24 ve tablo 3.15'de açıkça görüldüğü gibi kümelerin birbirlerinden ayrışmalarının başarılı olduğu söylenebilir. Bunun nedeni, tablo 3.15'de doğru sınıflandırma oranlarının en düşük %93.07 olmak üzere çok yüksek oranda gerçekleştiği görülmektedir. Bu durum kümeleme analizinin analitik olarak başarılı olduğu bilgisini doğrular niteliktedir. Bunun yanında multinomial regresyon analizi, kümelerin oluşmasında etkin olarak belirlenen değişkenlerin doğruluk derecesini de ölçmektedir. Bu sayede kümelerin profillerini belirleyen değişkenlerin ne oranda doğru tahmin edildiği de test edilmektedir. Multinomial regresyon analizi sonucu kümelerin profillerini belirleyen değişkenler tablo3.16'da ayrıntılarıyla verilmektedir.

Tablo 3.16. Multinomial Regresyon Analizi Sonucu Kümeleri Belirleyen Değişkenler

Değişken İsmi	Parametre Tahmini	Mutlak parametre değeri	t Skoru	Mutlak t skoru
CLNO: _SEGMNT_=2	-0,4590326	0,4590326	-19,110469	19,110469
CLNO: _SEGMNT_=3	-0,01833857	0,01833857	-0,55411672	0,55411672
CLNO: _SEGMNT_=4	0,09300617	0,09300617	3,011747241	3,01174724
DEBTINC: _SEGMNT_=2	-0,16237066	0,16237066	-10,1000426	10,1000426
DEBTINC: _SEGMNT_=3	-0,02335538	0,02335537	-0,49771657	0,49771657
DEBTINC: _SEGMNT_=4	0,004271916	0,00427192	0,092903991	0,09290399
DELINQ: _SEGMNT_=2	-1,01159799	1,01159799	-2,49703529	2,49703529
DELINQ: _SEGMNT_=3	22,99222738	22,9922274	1,967076243	1,96707624
DELINQ: _SEGMNT_=4	25,74592154	25,7459215	2,203059662	2,20305966
DEROG: _SEGMNT_=2	-2,94784014	2,94784014	-0,0000782	7,82E-05
DEROG: _SEGMNT_=4	54,411618	54,411618	325,2786456	325,278646

JOB MGR: _SEGMNT_=2	-0,14874654	0,14874654	-0,50045272	0,50045272
JOB MGR: _SEGMNT_=3	-0,54989065	0,54989065	-0,25207056	0,25207056
JOB MGR: _SEGMNT_=4	-0,80716891	0,80716891	-0,37355204	0,37355204
JOB OFFICE: _SEGMNT_=2	-0,75468312	0,75468312	-3,37568721	3,37568721
JOB OFFICE: _SEGMNT_=3	0,950807478	0,95080748	0,444102501	0,4441025
JOB OFFICE: _SEGMNT_=4	0,895752106	0,89575211	0,419974125	0,41997412
JOB OTHER: _SEGMNT_=2	-0,15971744	0,15971744	-0,80532284	0,80532284
JOB OTHER: _SEGMNT_=3	0,939980885	0,93998088	0,444091674	0,44409167
JOB OTHER: _SEGMNT_=4	0,936482232	0,93648223	0,443686313	0,44368631
JOB PROFEXE: _SEGMNT_=2	-0,39379631	0,39379631	-1,86509571	1,86509571
JOB PROFEXE: _SEGMNT_=3	0,362290843	0,36229084	0,169737599	0,1697376
JOB PROFEXE: _SEGMNT_=4	0,265961872	0,26596187	0,125353384	0,12535338
JOB SALES: _SEGMNT_=2	-0,49251147	0,49251147	-1,02475766	1,02475766
JOB SALES: _SEGMNT_=3	-3,67643826	3,67643826	-0,35630306	0,35630306
JOB SALES: _SEGMNT_=4	-4,17834703	4,17834703	-0,40448983	0,40448983
LOAN: _SEGMNT_=2	-0,00029986	0,00029986	-17,5197653	17,5197653
LOAN: _SEGMNT_=3	-0,00024723	0,00024723	-5,35558897	5,35558897
LOAN: _SEGMNT_=4	-0,00036741	0,00036741	-8,22765995	8,22765995
MORTDUE: _SEGMNT_=2	2,07E-06	2,07E-06	0,375744258	0,37574426

MORTDUE: _SEGMNT_=3	-1,9151E-05	1,9151E-05	-1,16642558	1,16642558
MORTDUE: _SEGMNT_=4	-3,4118E-05	3,4118E-05	-2,12115848	2,12115848
REASON DEBTCON: _SEGMNT_=2	-0,34682228	0,34682228	-3,47989739	3,47989739
REASON DEBTCON: _SEGMNT_=3	-0,1349249	0,1349249	-0,46679122	0,46679122
REASON DEBTCON: _SEGMNT_=4	-0,36241658	0,36241658	-1,41391569	1,41391569
VALUE: _SEGMNT_=2	-0,00010556	0,00010556	-15,1579626	15,1579626
VALUE: _SEGMNT_=3	-2,6476E-05	2,6476E-05	-1,81511675	1,81511675
VALUE: _SEGMNT_=4	-2,6151E-05	2,6151E-05	-1,82125508	1,82125508

Tablo 3.16'da her değişkene ilişkin parametre için analizdeki anlamlılıkları t testleri ile birlikte verilmektedir. T skoru 0'a yaklaşan değişkenler için anlamsız olmaktadır.

Tablo 3.16'da bulunan değişkenlerin paratmetre tahminleri ve t test skorları, tablo 3.10 ve tablo 3.14 arasındaki değişken istatistikleri ile karşılaştırıldıklarında JOB MGR: _SEGMNT_=2, JOB OFFICE: _SEGMNT_=2, JOB OFFICE: _SEGMNT_=3, JOB OTHER: _SEGMNT_=2, JOB PROFEXE: _SEGMNT_=3, JOB PROFEXE: _SEGMNT_=4, JOB SALES: _SEGMNT_=2, JOB SALES: _SEGMNT_=3, değişkenlerinin ilgili segmentlerindeki tahminlerinde sapmalar olduğu gözlenmektedir. Sapma payının özellikle JOB değişkenine ait olması bu değişkenin nominal olmasından ve iş dağılımının özellikle diğer iş statüsünde yoğunlaşmasından kaynaklanmaktadır. Diğer değişken istatistiklerinde herhangi bir sapma olmaması kümelerin işletme değerlerininin ve profillerinin başarılı konumlandırıldığı bilgisini vermektedir.Sonuç olarak, multinomial regresyon analizi kümeleme analizinin analitik ve profillendirilme açısından başarılı olduğunu göstermektedir.

SONUÇ ve ÖNERİLER

Günümüz dünyasında, insanların alışverişlerinde, bankacılık işlemlerinde, işletmelerin günlük işlemlerinde transfer edilen verilerin sayısı her geçen gün artmaktadır. 1995 yılında birincisi düzenlenen "Knowledge Discovery in Databases" adlı konferansta günlük işlemler sonucu oluşan veri yığınının taşıdığı önem şu şekilde açıklanmıştır:[105]

"Dünyadaki bilgi miktarı her 20 ayda bir yaklaşık olarak ikiye katlanmaktadır. Bu kadar veri ile alınması gereken kararlar ne olmalıdır?"

Veri sayısının inanılmaz boyutlarda artış problemi, bilgisayar sistemlerinin her geçen gün güçlerinin artırmasıyla aşılmaktadır. İşlemcilerin gittikçe hızlanması, bilgisayarların daha büyük miktarlardaki veriyi saklamasına olanak tanımaktadır. Bunun yanında bilgisayar ağlarındaki ilerleme ve bu veriye başka bilgisayarlardan da hızla ulaşabilmek olanaklı duruma gelmiştir.[106]

Veri sayısının büyük miktarlara ulaşması, bu verilerden anlamlı bilgi çıkarabilmek amacıyla kullanılan istatistik yöntemeleri yetersiz duruma getirmiştir. Bu amaçla veri madenciliği ve bilgi keşfi (data mining and knowledge discovery), büyük miktarlardaki veri içerisinden anlamlı sonuçlar çıkarmada kullanılmaktadırlar. Bilgi keşfi veri madenciliği, makine öğrenmesi, örüntü tanıma, yapay zeka gibi kavramları içine alan geniş bir kavramdır.

Veri madenciliğinde amaç istatistikten farklı olarak kolaylıkla mantıksal kurallara ve görsel sunumlara çevrilebilecek modellerin çıkarılmasıdır. Bu anlamda, veri madenciliği insan merkezlidir. Veri madenciliği, verilerin içindeki örüntüyü, ilişkiyi ve kuralları yarı otomatik olarak keşfetmektedir.

---

[105] SPSS Inc. Data Mining: An Introduction. www.spss.com (24.05.2002)
[106] Alpaydın, age, s:1

Çalışmamızda veri madenciliği modelleri üç ana başlıkta incelenmiştir: tahmin, tanımlayıcı ve kümeleme modeli. Tahmin modelinde tahmin edilmesi gereken bir hedef değişkenk bulunmaktadır. Bu amaçla daha önceden hedef değişkenin değerinin bilindiği bir veri kümesinde model eğitilmektedir. Bu nedenden dolayı tahmin modelinde denetimli öğrenmenin kullanıldığı teknikler bulunmaktadır. Eğitilen model, daha sonra hedef kolonun değerinin bilinmediği bir veri kümesinde değerlendirilmektedir. Model başarılı bulunduğu takdirde, geleceğin geçmişten çok farklı olmayacağı varsayılırsa geçmiş veriden çıkarılmış olan kurallar gelecekte de geçerli olacak ve ilerisi için doğru tahmin yapmamızı sağlayacaktır.[107] Çalışmada tahmin modellerinden karar ağaçları, yapay sinir ağları ve bellek tabanlı yöntemler tanıtılmaktadır.

Tanımlayıcı modelde, karar vermeye ilişkin örüntü tanıma işlemi gerçekleştirilmektedir. Bu anlamda veri içinde bağıntı kurma işlevi yerine getirilmektedir. Çalışmada tanımlayıcı modellerden sepet analizi kısaca tanıtılmıştır.

Kümeleme modelinde amaç, birbirine benzer özellikte olan verilerin bir arada gruplanmasını, birbirlerinden faklı özellikte olan verilerin ise farklı gruplarda yer almasını sağlamaktır. Kümeleme modelini tahmin modelinden ayrıran en önemli kriter hedef değişkenin bulunmamasıdır. Kümeleme analizine bu nedenden dolayı denetimsiz öğrenmenin kullanıldığı veri madenciliği tekniği de denilmektedir.

Veri madenciliğinin işletmelerde oluşturduğu katma değeri anlayabilmek için çalışmamızda veri madenciliğinin hayati döngüsü de (virtuous cycle of data mining) ele alınmıştır. Veri madenciliğinin hayati döngüsünde önemli olan dört aşama vardır:

i. İşletme problemlerinin tanımlanması
ii. Veriyi bilgiye dönüştürme
iii. Bilgiden hareketle uygulamayı gerçekleştirme
iv. Sonuçların değerlendirilmesi

[107] Alpaydın, age, s:1

Veri madenciliğinin hayati döngüsündeki aşamaları geçen bir veri madenciliği süreci, işletmelerin karlılıklarının artmasına ve yatırımların geri dönüşlerinin hızlı alınmasına olumlu anlamda katkıda bulunacaktır.

Veri madenciliği modelleri oluşturulduktan sonra modelin etkinlik derecesini ölçmek amacıyla asansör grafikleri oluşturulmaktadır. Uygulama aşamasında kurulan modellerin performanslarını değerlendirmede kümülatif ve kümülatif olmayan asansör grafiklerinden yararlanılmıştır.

Veri madenciliği büyük miktarlarda veriyi inceleme amacı üzerine kurulmuş olduğu için veri tabanları ile yakından ilişkilidir. Kulanıcılar için verinin amaca uygun bir şekilde saklanması ve gerektiğinde hızla ulaşılması önemlidir. Günümüzde yaygın olarak kullanılmaya başlanan veri ambarları günlük kullanılan veri tabanlarının birleştirilmiş şeklidir. Veri ambarında veri oluşturulduktan sonra bu verinin elle veya gözle analizi yapılabilir. Bunun için on-line analitik işleme programları kullanılmaktadır. Bu programlar veriye her boyutu veride bir alana karşılık gelen çok boyutlu bir küp olarak bakmayı ve incelemeyi sağlar. Bu şekilde boyut bazında gruplama, boyutlar arasındaki korelasyonları inceleme ve sonuçları grafik veya rapor olarak sunma olanağı sağlar.[108]

Çalışmamızın uygulama aşamasında kullanılan veri kümesi SAS Inst. Türkiye tarafından sağlanmıştır. Uygulamada kullanılan veri madenciliği paketi SAS Enterprise Miner 4.2 olarak belirlenmiştir.

Uygulamada, bir bankanın bireysel bankacılık departmanına gelen konut kredisi başvurularının kabul red kararının otomatik olarak verilmesi ve başvuruda bulunan müşterilerin ortak özelliklerinin belirlenerek anlamlı müşteri segmentleri yaratma süreci incelenmiştir.

[108] Alpaydın, age, s:3

Kredi talebinin otomatik olarak incelenmesi ve cevaplanması süreci, veri madenciliği tahmin ve sınıflandırma modeli oluşturularak değerlendirilmiştir. Performans karşılaştırması yapmak ve en iyi modeli seçmek amacıyla çalışmamızda tahmin modellerinden lojistik regresyon ve sınıflandırma modellerinden karar ağacı uygulanmıştır. Her iki model için, 5960 müşteriye ait 13 değişkenden BAD (müşterilerin borç ödeme durumu) değişkeni hedef değişken olarak belirlenmiştir. Verinin %70'i eğitim kümesi, %30'u da değerlendirme kümesi olarak bölümlendirilmiştir. Regresyon ve karar ağacı analiz sonuçları değerlendirme veri kümesi (validation data set) üzerinden gerçekleştirilmiştir.

Değişkenlerin dağılımının normal dağılıma uygun duruma getirilmesi gerekmiştir. Bu amaçla. girdi değişkenlerinin, hedef değişken ile aralarındaki ilişkinin optimal gruplandırma (optimal binning for relationships to target) yöntemiyle transformasyonu kullanılmıştır. Verideki eksik değerlerin tamamlanması işlemi regresyon ve karar ağacı modellerinde farklılık göstermiştir. Regresyon analizinde eksik değerler boşluk oluşturduğundan ve analiz esnasında eksik değer olarak esas alındığından bu boşlukların doldurulması gerekmiştir. Bu amaçla tamamlama yöntemlerinden ağaç tamamlanması uygun yöntem olarak modele eklenmiştir. Karar ağacı analizinde eksik değerler boşluk olarak algılanmadığından bu değerlerin yerlerinin doldurulmasına gerek duyulmamıştır.

Verinin temizlenmesi ve geçerlenmesi aşamalarından sonra regresyon analizinde adım adım ve geri adım yöntemlerinin performasları değerlendirilmiş, herbir yöntemin asansör grafikleri incelenmiş ve adım adım regresyon modelinin daha başarılı olduğu sonucuna varılmıştır. Genel olarak asansör grafiklerinde ve karşılaştırma matrisilerinde değerlerin yüksek çıkması regresyon modelinin başarılı olduğu bilgisini vermektedir.

Karar ağacı analizinde sınıflandırmanın yapılabilmesi amacıyla ayraç kriterlerinden Gini, Entropi ve Ki-kare ayraç kriterlerinin bulunduğu alternetif karar ağacı diyagramları oluşturulmuştur. Modelde üç ayraç kriterinin değerlendirildiği karar ağacı analizlerinin asansör grafikleri ve karşılaştırma matrisleri değerlendirildiğinde Gini ayraç kriterini esas

alan karar ağacı analizinin performansının daha başarılı olduğu sonucuna varılmıştır. Gini karar ağacı analizi sonucu yedi adet yaprak oluşmuş ve bu yaprakların oluşmasında etkili olan değişkenler önem sırasına göre verilmiştir. Gini karar ağacında yaprakların oluşmasında etkili olan kurallar maddeler halinde verilmiştir. Kurallar incelendiğinde en belirgin olan ifadeler:

i. Başvuruda bulunan müşterilerin borç gelir oranı 44.18'den küçükse, bu müşterilerin borçlarını ödeme ihtimali %93.4'tür.
ii. Başvuruda bulunan müşterilerin borç gelir oranı 44.18'den büyükse ve ödenmemiş kredi sayısı 0.5'den büyükse, bu müşterilerin borçlarını ödeme ihtimali %16.6'dır.
iii. Başvuruda bulunan müşterilerin borç gelir oranı 44.18'den büyükse, ödenmemiş kredi sayısı 0.5'den küçükse fakat haklarında oluşan borç ihbar sayısı 0.5'den büyükse, bu müşterilerin borçlarını ödeme ihtimalleri 15.8'dir.

Tahmin modeli olarak lojistik regresyon ve denetimli sınıflandırma tekniği olarak karar ağacı analizlerinin kendi içlerinde değerlendirme süreçlerinden sonra en başarılı modeli belirleyebilmek için lojistik regresyon ve Gini karar ağacı analizlerinin asansör grafikleri ve eşik değere bağlı doğru sınıflandırma grafikleri karşılaştırılmış ve adım adım lojistik regresyon modelinin performansı daha başarılı bulunmuştur.

Tahmin ve denetimli sınıflandırma modellerinden sonra hedef değişkenin bulunmadığı, müşterilerin 13 değişkene göre ortak özelliklerinin incelenerek segmentlere ayrıldığı denetimsiz sınıflandırma tekniklerinden kümeleme analizi yapılmıştır. Çalışmamızda kümeleme analizinin hiyerarşik olmayan yöntemlerinden K-ortalamalar algoritması incelenmiştir.

Kümeleme analizinde ilk adım, değişkenlerin aktif ve tanımlayıcı olarak sınıflandırılması olmuştur.Aktif değişkenler analizin analitik olarak değerlendirilmesinde, tanımlayıcı değişkenler ise analiz sonucu oluşacak kümelerin profillendirilmesinde kullanılmıştır.

Kümelerin analitik olarak değerlendirilmesinde kümelerin birbirlerinden uzaklığı ve küme içi frekanslar incelenmiştir. Analiz sonucu oluşan kümelerden birinci ve ikinci kümelerin, üçüncü ve dördüncü kümelere kıyasla daha büyük olmasına rağmen küme frekanslarının dağılımının çok fark oluşturmaması ve kümelerin birbirine uzaklığının ideal olması kümeleme analizinin analitik olarak başarılı olduğunu göstermiştir. Kümeleme analizi öncesi belirlenen aktif ve tanımlayıcı değişkenlerin beraber değerlendirilerek kümelerin işletme değerlerinin incelenmesi aşamasında, değişkenlerin küme içi dağılımlarında anlamlı sonuçlar oluştuğu gözlenmiştir.

Kümeleme analizinin analitik ve işletme değeri açısından başarılı olarak geçerlenmesi ve doğrulanmasından sonra ortaya çıkarılan kümelerin profillendirilmesi kümelerin anlamlı olduğu sonucunu vermektedir. Bu anlamda, banka için en değerli müşterilerin birinci kümede olduğu ve bankanın bu müşteri segmenti için her türlü yatırımı yapması gerektiği ortaya çıkmıştır. İkinci kümedeki müşterilerin banka için bir zarar nedeni olmadığı bunun yanında uzun vadede kar getirebilecek müşterilerden oluştuğu görülmektedir. İkinci kümedeki müşteri segmentine uygulanacak müşteri ilişkileri yönetimi (CRM) ile uzun vadede bu müşterilerin birinci kümeye dahil edilmesi mümkün görünmektedir. Üçüncü ve dördüncü kümedeki müşteriler banka için bir zarar kaynağıdır. Bu amaçla üçüncü kümedeki müşterilerin ve dördüncü kümedeki müşterilerin banka müşteri portföyünden çıkarılması bankanın boş yere para harcamasını engelleyecektir. Bu müşterilere ayrılan sermayenin ilk olarak birinci küme, ikincil olarak da ikinci küme için ayrılması sermayenin doğru yöne kanalize olamasını sağlayacaktır. Bu sayede banka, yatırımlarının geri dönüşünü daha hızlı alabilecektir. Küme profilleri tablo 3.17'de özetlenmektedir.

Tablo 3.17. Küme Profilleri

	1.küme	2.küme	3.küme	4.küme
Borç ödeme durumu	EN SADIK	SADIK	SADIK DEĞİL	SADAKAT DERECESİ EN DÜŞÜK
Kredi talep nedeni	Borç konsolidasyonu	Borç konsolidasyonu	Borç konsolidasyonu	Borç konsolidasyonu ve ev yenileme
Borç gelir oranı	EN YÜKSEK	EN DÜŞÜK	YÜKSEK	YÜKSEK
Talep edilen kredi tutarı	EN YÜKSEK	EN DÜŞÜK	YÜKSEK	YÜKSEK
Mevcut mal varlık değeri	EN YÜKSEK	EN DÜŞÜK	DÜŞÜK	ÇOK DÜŞÜK
Kredi soruşturma sayısı	YOK	YOK	EN YÜKSEK	YÜKSEK

Kümelerin analitik ve işletme değeri açısından değerlendirilmesi ve küme profillerinin oluşturulması aşamalarından sonra kümeleme analizinin geçerlenmesi ve doğrulanması süreci bulunmaktadır. Bu aşamada denetimli sınıflandırma tekniklerinden multinomial regresyon kullanılmıştır. Kümeleme analizi sonucu oluşan SEGMNT değişkeni, analizin geçerlenmesi aşamasında regresyon için hedef değişken olarak belirlenmiştir. Regresyon analiz sonucu, analitik ve işletme değeri açısından başarılı bulunan kümeleme analizi geçerlenmiştir. Regresyon karşılaştırma matrisi kullanılarak kümelerin biribirlerinden ne oranla ayrıştığı incelenmiş ve yaklaşık %94 tahmin oranı ile kümeler birbirlerinden başarı ile ayrılmıştır.

Konu araştırılırken yapılan literatür taramasında, yurt dışı yayınlarda gerek kitap gerekse makale anlamında yeterli sayıda kaynağa ulaşılmıştır. "Journal of Knowledge Discovery in Databases (KDD)" adlı dergide veri madenciliği konusunda en güncellenmiş yayınlara ulaşılabilmektedir. "KDD Nuggets" adlı web sitesinde ise veri madenciliği ve veri tabanlarında bilgi keşfi süreci konuları ile ilgilenen ve bu konular üzerine akademik çalışma yapan kişilerin yayınlarına ulaşmak mümkündür. "Data Mining Techniques" ve "Mastering Data Mining: The Art and Science of Customer Relationship Management" adlı kitapların yazarı Michael Berry ve Gordon Linoff'un elektronik ortamda bu konuda çalışan insanları buluşturan dataminer. com adında bir web siteleri bulunmaktadır. Bu site üzerinden, veri madeciliği ve veritabanlarında bilgi keşfi konuları ile ilgili kitaplara ve makalelere ulaşabilmek mümkündür. Aynı sitede veri madeciliği ile ilgili konferansların, seminer ve eğitimlerin de duyurusu yapılmaktadır.

Teknolojinin hızla ilerliyor olması veri madenciliği pazarının da dinamik olmasını gerekmiştir. Bu konuda çalışan uzman şirketler veri madenciliği ile ilgili ürünlerini değişen ve gelişen talebe göre güncellemektedirler. En çok bilinen veri madenciliği yazılım programları ; Clementine (SPSS), Enterprise Miner (SAS), Intelligent Miner (IBM), Mineset (SGI), Model 1 (Unicacorp), Model Quest (Abtech), PRW, Neuro Shell (Wardsystem), OLPARS (Partech) ve son olarak S-Plus olarak sıralanabilir.

Bilgi teknolojilerinin hızlı gelişiminin, çok fazla miktarda bilginin saklanmasına imkan vermesi günümüz iş dünyasında birçok alanda çalışmayı kolaylaştırmıştır. Tıp, finans, pazarlama, sigorta, araştırma ve ölçüm hizmetleri ve güvenlik sektörlerinde amaç çok fazla sayıda veriden bilgi çıkarımı olduğundan veri madenciliği ve veritabanlarında bilgi keşfi, bu alanda çalışan şirketlere vizyonlarını belirlemede yardımcı olmaktadır.

Türkiye'de yeni tanınmaya başlayan veri madenciliği kavramında, veri madenciliği tekniklerinin ve uygulama alanlarınındaki eksiklikleri gidermek amacıyla yapılan bu çalışmanın araştırmacılara yardımcı olmasını ve gelecekte bu alanda yapılacak çalışmalara bir zemin oluşturmasını dilerim.

KAYNAKLAR

Akael, Al- Attar. " White Paper: Data Mining- Beyond Algorithms". www.attar.com/tutor/mining (21.05.2002)

Akpınar, Haldun. "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği". www.isletme.istanbul.edu.tr/dergi/nisan2000. (24.01.2002)

Akpınar, Haldun. "Kendini Düzenleyen Haritalar, Avrupa Birliği'ne Üye ve Aday Ülkelerin Karşılaştırılması". www.isletme.istanbul.edu.tr/akpinar. (10.05.2002)

Akpınar, "Yapay Sinir Ağları ve Kredi Taleplerininin Değerlendirilmesinde Bir Uygulama Önerisi", Araştırma Raporu, Mayıs 1993

Alpaydın, Etjem. "Zeki Veri Madenciliği: HamVeriden Altın Bilgiye Ulaşma Yöntemleri". Bilişim 2000 Veri Madenciliği Eğitim Semineri.

Berry, Michael and Linoff, Gordon. *Mastering Data Mining: The Art and Science of Customer Relationship Management.* USA: John Wiley and Sons Inc, 2000

Berry, Michael and Linoff, Gordon. *Data Mining Techniques.* USA: John Wiley and Sons Inc, 1997

Berthold, Michael and Hand, David J., *Intelligent Data Analysis,* Italy: Springer, 1999

Bishop, Christopher M. *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press, 1995

Brachmen, R ve Anand, T. *The Process of Knowledge Discovery in Databases: A Human Centered Approach. In Advances in Knowledge Disvovery and Data Mining.* California: AAAI Press, 1996

Brand, Estelle. And Gerritsen Rob. "Decision Trees". DBMS- Data Mining Solutions Supplement. www. dbmsmag.com. (20.03.2002)

Brand, Estelle. And Gerritsen Rob. "Naïve Bayes and Nearest Neighbor". DBMS- Data Mining Solutions Supplement. www. dbmsmag.com. (20.03.2002)

Brand, Estelle. And Gerritsen Rob. "Association and Sequencing". DBMS- Data Mining Solutions Supplement. www. dbmsmag.com. (20.03.2002)

Brand, Estelle. And Gerritsen Rob. "Data Mining and Knowledge Discovery". DBMS- Data Mining Solutions Supplement. www. dbmsmag.com. (20.03.2002)

Brand, Estelle. And Gerritsen Rob. "Neural Networks". DBMS- Data Mining Solutions Supplement. www. dbmsmag.com. (20.03.2002)

Brand, Estelle. And Gerritsen Rob. "Classification and Regression". DBMS- Data Mining Solutions Supplement. www. dbmsmag.com. (20.03.2002)

Cambazoğlu, Türker. "Kurumlarda Yararlı Bilginin (Knowledge) Yönetimi ve İlintili Teknolojiler-18". www.bilisimrehber.com.tr/arastirma. (28.01.2002)

CDG Consulting. " Data Mining and Modeling". www.cdgconsulting.com/datamininghome. (19.03.2002)

Cheng, Bing and Tittrington D.M. "Neural Networks: A Review from a Statistical Perspective". Statistical Sciences. Vol. 9. no: 1, 1994. p:2-54

Coppock, S. David. "Data Mining and Data Mining Modeling: When is a Black Box Not Enough". www.dmreview.com/editorial. (18.04.2002)

Dean, P. "Implementing the ORACLE OLAP Applications Products- What is OLAP?", Temmuz, 1997

Fayyad, Usama, Piatesky-Shapiro Gregory, Padhric, Symith and Uthurasamy, Ramasomy. *Advances in Knowledge Discovery and Data Mining.* USA: MIT Press,1996

Fayyad, Usama. "Data Mining". *Technology Review.* vol. 104, no:1, January/February 2001, p: 101-105

Forcht, Kren A. and Cochran Kevin. "Using Data Mining and Datawarehousing Techniques". www.zaccaria.emeraldinsight.com. (01.10.2001)

Gershenfield, Neil. "The Nature of Mathematical Modeling", UK: Cambrige, 1999

Groth, Robert. *Data Mining: Building Competitive Advantages.* USA: Printice Hall, 2000

Green, P.E. *Analyzing Multivariate Data.* Hındsdale: Holt, Rinehart & Winston,1978

Gürsakal, Necmi ve Acar, F. "İstatistik, Veri Analizi ve Veri Madenciliği", IV. Ulusal Ekonometri ve İstatistik Sempozyumu Bildirileri. Antalya. Mayıs 1999

Hair, Joesph F., Anderson E. Ralph , Tatham L. Ronald, and William C.Black. *Multivariate Data Analysis.* USA: Printice Hall, 1998

Hand, David J., "Statistics and Data Mining: Intersecting Disciplines". *SIGKDD Explorations.* Vol.1. no:1. June 1999. p:16-17

Kohonen, Teuvo. *Selg Organizing Maps.* Springer Verlag, 1995.

Köksal, Bilge Aloba *İstatistik: Analiz Methodları.* 5. Basım, İstanbul: Çağlayan Kitapevi, 1998

Lee, Sang Jun and Siau, Keng. "A Review of Data Mining Techniques". Industrial Management & Data Systems. Vol.101. no:1. p:1-7

Ma, Catherine, Chou, David C. and Yen, David c. "Data warehousing, Technology Assesment and Management". *Industrial Management & Data Systems.* Vol. 100. no:3, 2000.p:125-135.

Mannila, Heikki. "Theoratical Frameworks for Data Mining". *SIGKDD Explorations.* Vol.1. no:2. January 2000. p:30-32.

Mitchell, Tom M., *Machine Learning.* USA: McGraw Hill, 1997

Next Action Technology. "Set Based Segmentation- A New Way to View Your Customer".March 1998

Oğuzlar, Ayşe. "Çok Boyutlu Ölçekleme ve Kümeleme Analizi Arasındaki İlişkiler". www. iktisat.uludağ.edu.tr/dergi. (24.01.2002)

Oktay, S. Ümit Fırat. "Kümeleme Analizi: İstihdamın Sektörel Yapısı Açısından Avrupa Ülkelerinin Karşılaştırılması". *Sosyal Bilimler Dergisi.* Cilt 3. Sayı 2.Temmuz 1993, s:50-59

OLAP Council White Paper. www.olapcouncil.org/research. (28.01.2002)

Parr Rud, Olivia. *Data Mining Cookbook.* USA: John Wiley and Sons Inc., 2001

Piatesky-Shapiro, G., and Frawley, W.*Konowledge Discovery in Databases.*California: AAAI Press, 1991

Piatesky- Shapiro, G. "Knowledge Discovery in Real Databases: A report on the IJCAI-89 Workshop", *AI Magazine*, 1999, cilt:11 , sayı:5, s: 68-70

Pitta, Dennis A., "Marketing One-to One and Its Dependence on Discovery in Databases",*Journal of Consumer Marketing*. Vol 15. no:5. 1998, p:468-480

Potts, William J.E. *Decision Tree Modeling Course Notes*. USA: SAS Inst., 1999

Potts, William J.E. *Data Mining Primer:Overview of Applications and Methods*. USA: SAS Inst., 1998

Pyle, Dorian. *Data Preparation for Data Mining*. California: Academic Press, 1999

StatSoft Inc. "Cluster Analysis". www. statsoftinc.com/textbook (28.08.2001)

Ripley, B.D. "Neural Networks and Related Methods for Classification".*Journal of the Royal Statistical Society*. vol 56. no:3.1194. p: 409-437

SAS Analytic Intelligence. "Turn Raw Data into Business Gold with Data Mining". www.sas.com/technologies/data_mining. (14.03.2002)

Schrager, J. and Langley, P. "Computational Models of Scientific Discovery and Theory Formation". California: Morgan Kaufmann, 1990.

Sharma, Subbash. *Applied Multivariate Techniques*. USA: John Wiley and Sons Inc., 1998

SAS. "How Can Data Mining help in Banking". www.sas.com. (01.10.2001)

SAS. "SAS VERİ MADENCİLİĞİ".
http/www.sas.com/offices/Europe/Turkey/cozveri.com, s:1 (09.09.2001)

SPSS. "Data Mining Techniques: Decision Trees". www.spss.com/datamine/trees. (09.09.2001)

SPSS. "Clementine".www.spss.com.tr/clementine2. (14.03.2002)

Westpall, Christopher and Blaxton, Teresa. *Data Mining Solutions: Methods and Tools for Solving Real World Problems.* USA: John Wiley and Sons Inc., 1998

SPSS. "Data Mining with Clementine for Smarter Retailing". www.spss.com./whitepaper (21.02.2002)

SPSS. " Data Mining and Statistics: Gain a Competitive Advantage". www.spss.com/whitepaper. (12.04.2002)

SPSS. "More on What Data Mining is and isn't". www.spss.com/datamine. (28.01.2002)

SPSS. "Data Mining Techniques". www.spss.com/datamine/techniques. (28.01.2002)

Tatlıdil,H. *Uygulamalı Çok Değişkenli İstatistik Analiz,* İstanbul: Engin Yayınları, 1996

Thearling, Kurt. "Scoring Your Customer". www3.shore.net. (24.11.2001)

Thearling, Kurt. "An Introduction to Data Mining". www3.shore.net. (05.11.2001)

Thearling, Kurt. "Data Mining and Privacy: A Conflict in The Making ?". www3.shore.net. (24.11.2001)

Thearling, Kurt. "Understanding Data Mining: Its All in The Interaction". www3.shore.net. (24.11.2001)

Thearling, Kurt and Frawley Andrew. "Data Mining Can Bring Pinpoint Accuracy to Sales". www3.shore.net. (24.11.2001)

Thearling, Kurt. "Data Mining and CRM: Zeroing in on Your Best Customer". www.dmreview.com/editorial/dmdirect. (24.11.2001)

WhiteCross White Paper. "Mining Very Large Databaes to Support Knowledge Exploration". Version1. January 5, 2001. p: 1-22

Wielenga, Doug, Lucas Bob ve Georges Jim. *Enterprise Miner: Applying Data Mining Techniques Course Notes.* USA: SAS Inst., 1999

ÖZGEÇMİŞ

Doğum Tarihi:	**18 Kasım 1976**	
Doğum Yeri:	**Burdur**	
Lise:	**1990-1993**	**Bursa Fen Lisesi**
Lisans:	**1993-1999**	**Boğaziçi Üniversitesi** **Fen-Edebiyat Fakültesi** **Matematik Bölümü**
Yüksek Lisans:	**1999-2002**	**Yıldız Teknik Üniversitesi** **Sosyal Bilimler Fakültesi** **İşletme Yönetimi Programı**
Çalıştığı Kurumlar	**1999-2000**	**ALICE/BBDO Reklam Ajansı** **(İstatistik- Veri Tabanı Yöneticisi)**
	2000-	**İstanbul Bilgi Üniversitesi** **Reklamcılık Bölümü** **Araştırma Görevlisi**