YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

kredi skorlamada kullanılan yönte..

Yüksek Lisans Tezi

nurşahver turangil

FULL

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KREDİ SKORLAMADA KULLANILAN YÖNTEMLER VE UYGULAMALARI

İstatistikçi Nurşahver TURANGİL

F.B.E. İstatistik Anabilim Dalı'nda
Hazırlanan

YÜKSEK LİSANS TEZİ

Tez Danışmanı: Yrd. Doç. Dr. Doğan YILDIZ

Prof. Dr. Taygan LİNAL

Yrd. Doç. Dr. Doğan Yıldız

Yrd. Doç. Dr. Atıf Evren

İSTANBUL, 2006

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

$A(x)$	$p \times p$ simetrik sonlu pozitif bir matrisi,
A_G	İyi olarak cevaplanan sorular
A_B	Kötü olarak cevaplanan sorular
a_i	Muhtemel hatalar için pozitif veya sıfır değeri, dışsal sapma
B, b	Toplam kötü sayısı
b_i	i niteliğini sergileyen kötü müşteri sayısı
$C(i,j)$	Grup j'nin başvurusunu yanlış olarak grup i'ye atamanın maliyeti
c	Bir kesme değeri (cutoff)
D	Concordance istatistiği, uzaklık matrisi
d_{ij}	Uzaklık matrisi elemanları
D^2	Karesel uzaklık
e_i	İçsel sapma
$F(u)$	Lojistik fonksiyon
F_1	Girdi katmanından sonraki ilk katman
F_2	Çıktı katmanındaki aktivasyon fonksiyonunu
G, g	Toplam iyi sayısı, grup adı
g_i	i niteliğini sergileyen iyi müşteri sayısı
$i(v)$	Basit safsızlık indeksi, Gini indeksi, Entropi İndeksi
K	İyiyi reddetmenin ortalama kayıp karı
K_{vk}	Gizli katmandaki nöron k ve çıktı katmanındaki nöron v'yi birleştiren y_k katmanına uygulanan ağırlık değeri
k	Karakteristiklerin sınıf sayısı, başvuru sayısı
L	Faktör yükleri matrisi, logaritma – logit
l	Sol düğüm
ℓ_{ij}	i. değişkenin j. faktör üzerindeki yükü
M	Ortalama varolan maliyet, bir kötüyü iyi olarak yanlış sınıflamanın maliyeti
m	Değer ortalamaları arasındaki orta nokta
N, n	Gözlem sayısı, gerçekleşen olay sayısı
$o_v(t)$	Nöron v deki olay t için gözlenen gerçek sonuçları
R	Korelasyon matrisi
r	Sağ düğüm
S	Kovaryans matrisi, benzerlik matrisi
s	Kabul–ret skoru, çıktı katmanındaki nöronların sayısı
T	Sınıflandırma ağacı düğüm seti
T^2	Hotelling Uzaklığı
v	Düğüm
w	Ağırlıklı skorlar
y_k	Birinci gizli katmandan elde edilen çıktı değerleri
$y_v(t)$	Tahmin edilen sonuçlar
X, x	Veri matrisi, değişken
x_B	Kötü özellikliler
x_G	İyi özellikliler
z_v	Çıktı katmanındaki nöron v'nin çıktı değeri
χ^2	Ki-kare istatistiği
ψ_i	i. değişkenin spesifik varyansı

η	sabit deneme oranı katsayısı

KISALTMA LİSTESİ

ADA	Adımsal diskriminant analizi
HSO	Yüksekken geçersiz kılmak (High-side overrides)
DS	Davranış skorlaması
KS	Kredi Skorlama
LSO	Alçakken geçersiz kılmak (Low-side overrides)
MMD	Maksimum sapmanın minimizasyonu
MSD	Sapmaların mutlak değerler toplamını
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖNSÖZ

Bana çok değişkenli istatistiksel teknikleri tanıtan ve sevdiren, tez çalışmamda orijinal fikirleriyle yol gösterici olan, anlayışıyla çalışmalar sırasında cesaretimi kırmayan değerli hocam Yrd. Doç. Dr. Doğan YILDIZ'a

Hayatım boyunca bana verdikleri güven ve destekten dolayı babam Rıfat TURANGİL ve annem Sevim TURANGİL'e, her anlamda çok emeğini gördüğüm abim A. Hamdi TURANGİL'e, yirmi dört yıldır bana dostluk eden kardeşim Kübra TURANGİL'e ve M.Akif TURANGİL'e

Master eğitimim boyunca kahrımı çok çeken ve hep destekçim olan Selman USLU'ya

Master yapmam konusunda her türlü imkanı sağlayan eski yöneticilerim Sn. Mehmet Atilla KURAMA ve Sn. Osman BAYRAKTAR'a çok teşekkür ederim.

ÖZET

Gelişmekte olan bankacılık sektöründe kredi kartları önemli bir ürün olarak yerini almıştır. Müşteriye kredi kartı verilmesi-verilmemesi ise banka açısından dikkatle incelenerek alınması gereken bir karardır ve kredi kartı talebi arttıkça başvuruların değerlendirilmesi daha da karmaşık bir hal almaktadır. Değerlendirmeyi yapan kişiler farklı kriterleri dikkate alabilecekleri için alınan kararlar subjektif olabilir. Bu durumda, hem artan başvuru sayısına doğru zamanda cevap vermek hem de subjektif kriterlerden arınıp objektif kararlar alabilmek için çeşitli istatistiksel ve istatistiksel olmayan teknikler kullanılmaktadır.

Bu çalışmada, bir bankanın kredi kartı müşterilerine ait on üç değişken ile iyi müşteri-kötü müşteri ayırımı yapılmaya çalışılmıştır ve uygulama sonuçları birbirleriyle kıyaslanmıştır.

Uygulamada kullanılan bazı istatistiksel teknikler; Diskriminant Analizi, Lojistik Regresyon, Kümeleme Analizi (Cluster Analysis), Sınıflandırma Ağaçları (Classification Tree)'dır ve istatistiksel olmayan tek teknik ise Yapay Sinir Ağları'dır. Çalışmada bu tekniklerin uygulamaları ele alınmış, teorik ayrıntılara değinmekten kaçınılmıştır. Uygulamalarına yer verilmeden teorik bilgi olarak sunulan teknikler ise Lineer Programlama ve Tamsayı Programlama konularıdır. Bu tekniklerle ilgili uygulamalar ileride yapılacak çalışmalara bırakılmıştır.

Modellerin uygulamasında çeşitli avantajlar-dezavantajlar mevcut olmasıyla birlikte kullanılan veri setine göre tahmin başarısı en yüksek olan modelin lojistik regresyon olduğu söylenebilir. İlgili veri setine göre tüm modellere giren AYLIK NET GELİR, ÖĞRENİM DURUMU, MÜŞTERİ TİPİ değişkenleri ise bu çalışma için kredi skorlamada etkili değişkenler olarak gözlemlenmiştir.

Anahtar kelimeler: Kredi Skorlama, Lojistik Regresyon, Diskriminant Analizi, Kümeleme Analizi, En Yakın Komşu, Sınıflandırma Ağaçları, Yapay Sinir Ağları

ABSTRACT

Nowadays, credit cards become the most important product in developing finance sector. On behalf of banks, to issue or not to issue credit cards to clients should be decided after carefully investigating their backgrounds. By increasing credit card demands, the evaluation of credit card applications become more complex system. Because credit cards specialist consider different criteria, the decisions can be subjective. So, the statistical and non-statistical methods are used to both respond increasing credit card applications at right time and make objective decisions by getting rid of subjectivity.

In this work, good-bad clients are tried to be separated by regarding thirteen variables belongs to credit card clients of a specific bank and the results of application are compared between each other.

The statistical methods used to in this work are Discriminant Analysis, Logistic Regression, Cluster Analysis and Classification Tree, and the only non-statistical method is Neural Networks. In this study, the applications of those methods are focused on without given so much details in theory.

Whereas there are some advantages and disadvantages of implementing those methods, the estimation rate of logistic regression is the highest among them in accordance with used data sets. Because MONTHLY NET INCOME is common variables in all models regarding to related data set, those are observed as the most effective varibles in the credit scoring in this work.

Keywords: Credit Scoring, Logistic Regression, Discriminant Analysis, Cluster Analysis, Nearest Neighbour, Classification Trees, Neural Networks.

1. GİRİŞ

İlk geliştirilen finansal risk yönetimi konularından biri olan kredi skorlama finans ve bankacılıkta kullanılan istatistiksel ve operasyonel modellerin en başarılılarından birisidir ve geçen zaman içerisinde kredi skorlama analistlerine daha fazla ihtiyaç duyulmaktadır. Kredi kartlarındaki artıştan da etkilenen kredi skorlama otomatik olarak riskin hesaplanmasını sağlamaktadır ve bu hesabı yapan modeller kredi kartı veren bankaların kart hacimlerini ellerindeki verilere dayanarak daha kolay genişletebilmelerine imkan vermektedirler.

Bu çalışmada ilk bölümde kredi skorlamanın tanımı, tarihçesi ve skor kart geliştirilmesiyle ilgili bilgilere yer verilmiştir. Skor kart oluştururken örnekleme seçimi, veri kaynakları, müşterilerin iyi-kötü olarak ayrılmasından kredi kartı başvuru formunda istenilen verilerin sınıflandırılmasına kadar modelin temel noktaları bu bölümde ele alınmıştır. Uygulama aşamasında kullanılan veri seti bir bankanın kredi kartı müşterilerine ait bilgileri içermektedir ve hazır verilerle çalışıldığı için anlatılan teorik altyapının daha önce verilerin hazırlanmasında kullanıldığı varsayılarak veriler üzerinde değişiklik yapılmamıştır.

Çalışmanın ikinci bölümünde ise uygulamada kullanılan bazı istatistiksel ve istatistiksel olmayan tekniklerin teorik altyapılarıyla ilgili bilgiler verilmiştir. Çalışma uygulamayı ön plana çıkardığı ve bazı tekniklerin kıyaslanması amaçlı olduğu için teorik altyapıda detaya inilmemiş, uygulamada yer alan Diskriminant Analizi, Lojistik Regresyon, Kümeleme Analizi, Sınıflandırma Ağaçları gibi istatistiksel ve Yapay Sinir Ağları gibi istatistiksel olmayan tekniklerin teorilerine kısaca değinilmiştir. Kullanılabilecek yöntemlerin bir arada toplanması açısından uygulamada yer almayan Lineer Programlama ve Tamsayı Programlama tekniklerinin teorik altyapılarına da değinilmiştir.

Yukarıda bahsedilen tekniklerin uygulamalarına ise üçüncü bölümde yer verilmiştir. Her tekniğin uygulamasının sonunda uygulamayla ilgili genel değerlendirmenin yapıldığı sonuç kısmı yer almaktadır.

Dördüncü bölümde ise tüm uygulama tekniklerinin sonuçları bir arada değerlendirilerek teknikler arası kıyaslamalara yer verilmiştir.

2. KREDİ SKORLAMA'NIN TANIMI TARİHÇESİ ve OLUŞTURULMASI

2.1 Kredi Skorlama Nedir

Bir kredi başvurusunda müşterinin krediyi geri ödeyememesi olasılığını (probability of default) hesaplamaya kredi skorlama denir.

KS (kredi skorlama), borç verene tüketici kredisi vermesinde yardımcı olan karar modelleri ve teknikleridir. Bu teknikler kime, ne kadar kredinin verileceğine, ne tür operasyonel stratejilerin borç alanın karlılığını artıracağına yönelik kararlar alınmasını sağlar.

KS yaparak yüksek riskli müşterilere kredi vermeyi reddetmek finansal kurumun olası zararını azaltacak, düşük riskli müşterilere kredi vererek karını arttıracak, üstelik müşterilerin ödeyemeyecekleri kredilerden dolayı rahatsızlığını azaltacaktır.

KS teknikleri belirli bir müşteriye verilen borcun riskini dağıtır. Kredi değerliliği ağırlık, uzunluk veya gelir gibi kişisel bir özellik değildir. Borçlunun borç veren ile ilişkisini göstermekte ve her iki tarafın şartlarını yansıtmakta olup borç veren açısından gelecekteki muhtemel ekonomik senaryoları göstermektedir. Böylece borç verenler, bir bireyi krediye layık veya layık olmamasına göre sınıflarlar. KS'nin uzun dönemli en büyük tehlikesi bu işlemin durması ve bazı müşteriler bütün kredi verenlerden borç alıyorken bazı müşterilerin hiç alamamasıdır. Bir müşteriyi krediye uygun değil diye tanımlamak tepkiye yol açar. Kredi verenlere gerçeği göstermek en iyisidir. Her zaman alınan borcun geri ödenmeme riski mevcuttur, kredi verenler bunu hiç unutmamalıdırlar.

Bir borç veren iki çeşit karar vermelidir: yeni bir başvuruya kredi verilip-verilmeyeceğine karar vermek ve kredi limitlerini artırmak isteyen mevcut müşterilere karşı nasıl davranılacağını tespit etmek. İlk tip soruyu açıklayan teknikler kredi skorlama olarak adlandırılırken, ikinci tip karar davranış skorlama olarak adlandırılır.

Hangi teknik kullanılırsa kullanılsın, her iki karar tipinde de var olan hayati nokta; önceki müşterilere ait oldukça fazla, detaylı bilgi ve kredi geçmişi bilgisi içeren örnekleme ihtiyacıdır. Bütün teknikler örneklemeyi müşterilerin özellikleri arasındaki ilişkileri tanımlamak ve sahip oldukları geçmişe dayanarak iyi-kötü ayrımını yapmak için kullanır. Tekniklerin çoğu skor kart oluşturur, bu kartta özelliklere bir skor verilmiştir ve bu skorların toplamı bir kişiye kredi vermenin kötü sonuç doğurup doğurmadığını belirlemeyi sağlar. Bazı teknikler skor kart üretmeseler de, müşteriye kredi vermenin iyi olup olamayacağını doğrudan anlarlar ve bu teknikler kredi ve davranış skorlama ile birlikte paralel hareket ederler.

Genel olarak skorlama kredi verilmesinde kullanılmasına karşın, özellikle son zamanlarda bir çok farklı alanda da kullanılmaktadır. Özellikle, doğrudan ve diğer pazarlama tekniklerinde hedef müşteri kitlesinin tespitinde çok kullanışlıdır. Finans ve perakende sektörlerinde bir çok şirket veri depolamak için skorlama tekniklerini uygulamaya gereksinim duymaktadır. Benzer şekilde, veri madenciliği ve çok gelişmiş karmaşık bilgi sistemlerinde çok başarılı bir şekilde skorlama uygulamaları gerçekleştirilmektedir.

2.2 Kredinin Tarihçesi

İnsanlar iletişime başladığından beri, ödünç para alma ve ödeme işlemleri de başlamıştır. İlk kullanılan kredi antik çağda Babil'de gerçekleşmiştir. O zamanlarda, çiftçiler nakit akışı problemlerini, mahsulü borç alıp hasat mevsimi geldiğinde faizi ile birlikte geri ödeyerek çözerlerdi.

Yunan ve Roma İmparatorluğu zamanına geldiğimizde, gelişmiş kredi enstitüleri kurulmuştur. Takip eden 1000 yıllık zamanda bu konuda çok az gelişme olmuştur, 13. yy.daki Haçlı seferleri sırasında borç veren mağazalar kurulmuştur. Bu tür mağazalar ilk kurulduklarında faiz uygulamazken, 1350 yılından sonra bütün Avrupa'da faiz uygulamaya başlamışlardır. Bu tip işletmeler hala çoğu Avrupa ve Güney Amerika ülkelerinde görülmektedirler (üçlü top işareti olan mağazalar). Orta çağda, verilen bağışlar ya da borçlar üzerinde faiz işletmek ahlaki açıdan uygun görülmüyordu. Günümüzde bu durum İslam Ülkelerinde geçerlidir. Ortaçağ Avrupa'sında eğer bir kişi parası üzerine çok düşük bir faiz uyguluyorsa buna pek bir şey denilmezken çok yüksek faiz işletiliyorsa kötü olarak algılanıyordu. Bu çağda, krallar ve hükümdarlar yaptığı savaşları finanse etmek için borç para alıyorlardı. Bu tür borç almalar, iş ilişkisinden ziyade politik ilişkilere dayanıyordu ve az borç almak gücün zayıfladığına işaret olarak algılanıyordu.

1800'lü yıllarda orta sınıfın gelişmesi, bir kaç tane özel bankanın kurulmasına yol açmıştır. Amaç, iş kurmak ve masraflarını karşılamak isteyen kişileri finanse etmekti. Ancak, tüketici kredisinin bu ilk başlangıcı nüfusun çok küçük bir kesimine hitap ediyordu. Bir malı almak için bir kaç kişi bir araya gelip eşit bir pay ödüyor ve daha sonra kendi aralarında çekiliş yaparak mal sahibini belirliyorlardı. Bu davranış ilk kredi birliklerinin kurulmasına yol açmıştır.

Gerçek devrim, tüketicilerin motorlu araçları satın almaya başlamasıyla 1920'lerde gerçekleşmiştir. İkinci Dünya savaşından önce, Finans şirketleri bu ihtiyacı karşılamak için hızla kurulmaya başlamıştır. Bu dönemde, posta yoluyla satış yapan şirketler müşterilerine

taksit imkanı tanımaya başlamıştır.

20. yy.ın son yarısında, borç verenler çok hızlı artmıştır. Tüketici kredileri bütün sektörlerde en yüksek büyümeye sahip olmuştur. 1960'larda kredi katının ortaya çıkması bu büyümenin en büyük görsel delili olup, günümüzde kart sahibi olmamak imkansız hale gelmiştir. Artık her şey kredi kartı ile satın alınabilmektedir.

2.3 Kredi Skorlamanın Tarihçesi

Kredinin tarihçesi 5000 yıla dayanmasına karşın, kredi skorlama yalnızca 50 yıldır kullanılmaktadır. KS, birbirleriyle ilişkili özelliklere dayanarak bir ana kütledeki farklı grupları tanımlamanın en önemli yoludur. Fisher 1936'da bu tür problemlerin çözümü için bir istatistiksel yaklaşımı ilk defa ortaya atmıştır. İris aslı bir çiçeğin iki türünü, fiziksel büyüklüğüne ve yapıtaşına göre ayırmaya çalışmıştır. 1941'de Durand, aynı teknikleri kullanarak ilk defa iyi ve kötü borçları tasnif etmeye çalışmıştır. Yaptığı çalışmalar Amerikan Ekonomik Araştırmalar Bürosu için yaptığı bir projede kullanılmıştır ve tahmin amaçlı değildir.

1930'larda, posta ile satış yapan şirketler kredi kararlarındaki uyumsuzlukları ortadan kaldırmak için sayısal bir skor sistemi geliştirmişlerdir. İkinci Dünya Savaşı'nın başlamasıyla birlikte, bütün kredi veren ve posta ile satış yapan şirketler kredi yönetiminde zorluklar yaşamıştır. Kredi analistlerinin askere alınmasıyla, bu sektörde uzman kişilerin sayısı büyük oranda azalmıştır. Böylece, şirketler analistlerinden kime kredi verip verilmeyeceğinin kararında uyguladıkları kuralları yazmalarını istemişlerdir (Johnson, 1992). Bunlardan bazıları sayısal skorlama sistemlerinin kurulmasına neden olurken, bazıları da ihtiyaçların tatminini oluşturan şartları oluşturmuştur. Bu kurallar böylece uzman olmayan kişilerin de kredi verme kararı almalarına yol açmıştır (Thomas vd., 2002).

Savaşın bitmesinden çok kısa bir süre sonra, otomatik kredilendirme sistemleri, istatistiksel sınıflandırma modelleri borç verme kararlarında kullanılmaya başlanmıştır. Bu konudaki ilk danışmanlık şirketi San Francisco'da 1950'lerin başlarında Bill Fair ve Earl Isaac tarafından kurulmuştur ve müşterileri finans evleri, perakendeciler ve posta ile satış yapan şirketler olmuştur (Thomas vd., 2002).

1960'ların sonunda kredi kartlarının kullanılmaya başlanmasıyla birlikte kredi kartı verenler için KS çok kullanışlı hale gelmiştir. Bilgisayarların kullanılmasıyla birlikte bu teknik her gün çok sayıda kişinin başvurusunu değerlendirebilmektedir. Böylece şirketler KS'yi çok iyi bir

tahmin edici ve karar verme aracı olarak görmüşlerdir. KS, 1975 ve 1976'da Amerika'da çıkarılan Eşit Kredi Fırsatları Kanunu (Equal Credit Oppırtunity Acts) ile birlikte kredilendirmede kullanılan yasal bir teknik olmuştur. Böylece, sonraki 25 yılda KS analizleri hızla büyüyen bir meslek alanı olmuştur. Özellikle Amerika'da ve İngiltere'de çok yaygınlaşmıştır (Thomas vd., 2002).

1980'lerde, KS'nin kredi kartlarındaki başarısıyla birlikte bankalar kişisel krediler, ev kredileri ve küçük yatırımcı kredileri gibi diğer ürünlerinde de bu tekniği kullanmaya başlamışlardır. 1990'larda, doğrudan pazarlamada skor kartlarının kullanılması reklam kampanyalarına geri dönüşlerin artmasına yol açmıştır. Bilgisayardaki gelişmeler diğer tekniklerin de skor kartlarının oluşturulmasında kullanılmasına izin vermiştir. 1980'lerde, bugün kullanılan en önemli iki teknik olan lojistik regresyon ve lineer programlama tekniği kullanılmaya başlanmıştır. Son zamanlarda, yapay zeka ve sinir ağları teknikleri deneme mahiyetinde kullanılmaktadır.

Günümüzde amaç fonksiyonu, müşterilerin borç ödememelerini minimize etmenin yerine şirketlerin bu tür müşterilerden daha fazla nasıl gelir elde edebilecekleri üzerine kurulmaya başlanmıştır. Skor kartlar ile borç ödemeyen müşterilerin risk tahmininde büyük gelişmeler kaydedilmiştir. Skor kartlar; "Müşteriler yeni bir ürüne ve doğrudan satışa nasıl tepki verecekler?", "Müşteriler ne kadar sıklıkla bir ürünü kullanacaklar?", "Yeni bir ürünün ortaya çıkmasıyla eski ürünü daha ne kadar kullanacaklar?" "Müşteriler başka bir kredi verene gidecekler mi?", "Müşterilerin borçlarını ödememeye başlamasıyla birlikte bunlara karşı nasıl bir tutum sergilenecek?" ve "Başvurulardaki sahtekarlıklar nasıl engellenecek?" gibi soruların cevabını bulmaya yardımcı olur.

2.4 Kredi Skorlama ve Veri Madenciliği

Veri madenciliği, verilerde anlamlı ilişkileri ve yapıları tespit etmek için veri analizi ve araştırma tekniğidir. Madenciliğe benzer şekilde, bu teknikte de gerekli olan verinin nereden ve nasıl bulunacağı belirlenmeye çalışılır. Son yıllarda, şirketler özellikle de bankalar ve perakendeciler, müşterileri hakkında sahip oldukları bilginin tanımlanmasının değerini kavramışlardır. Elektronik fon transferinin ve bağlılık kartlarının yaygın olarak kullanılmasıyla birlikte, bu tür şirketler müşterileri hakkında bilgileri kolaylıkla toplayabilmektedirler. Bilgisayar teknolojisi de toplanan büyük miktarlardaki verilerin analizinde kolaylık sağlamaktadır. Rekabetin artması, ikame ürünler ve internet gibi kolay iletişim kanalları müşterilerin kolaylıkla yer değiştirmesine neden olmaktadır. Böylece,

müşteri davranışlarını anlamak ve analiz etmek büyük bir önem arz etmektedir. Bu nedenle şirketler veri ambarları oluşturmak ve veri madenciliği gibi teknikleri kullanmak için çok büyük miktarda paralar harcamaktadırlar.

Veri madenciliğinin ana tekniklerine göz atıldığında, bu tekniğin kredi skorlamada çok başarılı neticeler alınmasını sağladığı görülmüştür. Temel veri madenciliği teknikleri arasında veri özeti (data summary), değişken düşürme (variable reduction), gözlem kümeleme (observation clustering), tahmin ve açıklama (prediction and explanation) yer almaktadır. Frekans, ortlama, varyans ve çapraz tablolama gibi standart tasvir edici istatistikler verilerin özetlenmesi için kullanılır. Aynı zamanda, kesikli sınıflar için sürekli değişkenlerin kategorize edilmesi için de çok kullanışlıdır. Betimsel istatistikler, KS'de çok kullanılan kaba sınıflama tekniğidir. Hangi değişkenlerin en önemli olduğunu tespit etmek, gereksiz olanları analizden çıkarmak KS uygulamalarında çok sık kullanılan teknikler olduğu gibi veri madenciliği uygulamalarında da kullanılmaktadır. Müşterileri satın aldığı farklı ürünlere göre veya diğer özelliklere göre gruplara ayırmak diğer bir veri madenciliği aracıdır. KS aynı zamanda müşterilerin davranışlarına göre farklı gruplar oluşturur ve her bir grup için ayrı bir skor kart düzenlenir. Bu fikir alt ana kütlelerin segmentasyonu anlamına gelir ve böylece her bir alt ana kütle için bir skor kart profili oluşturulur.

Esasında KS'de kullanmak için geliştirilen, örneğin gelecek yıl hangi müşterinin hangi finansal aracı kullanacağını tahmin eden teknikler veri madenciliği için de çok önemlidir. Gerçekte, veri madenciliğinin kullandığı segmentasyon analizi belirli tip davranışlara sahip segmentleri göstermek için kullanılır.

Böylece, veri madenciliği KS için çok önemli olmazsa olmaz bir teknik ve teknoloji olup, daha geniş bir alanda uygulanması gerekir. KS'yi uygulamadaki hatalardan, eksikliklerden korunmak ve diğer alanlarda uygulayabilmek için, veri madenciliğini KS ile birlikte kullananlar çalışmalarında çok daha fazla başarıya ve gelişmeye ulaşacaklardır.

2.5 Örnekleme Seçimi

Kredi Skorlama (KS) ve Davranış Skorlaması (DS) ile ilgili tüm metotlar skorlama (scoring) sistemini geliştirmek için müşterilere ait geçmiş verilere ve onların hikayelerine ihtiyaç duymaktadır. Örnekleme seçiminde dikkat edilmesi gereken iki husus mevcuttur. Birincisi; örnekleme gelecekte başvurusu mümkün olan kişileri temsil etmelidir. İkincisi; örnekleme ödeme alışkanlığının iyi yada kötü olduğunu yansıtabilecek kadar yeterli bilgiyi içine almalıdır. Bunu en iyi açıklayacak örnekleme, mümkün olan en yakın geçmişte borç almış

kişilerin bilgilerinin yer aldığı veri tabanıdır. Başvuru skorlamasında (Application Scoring) bu son 12 ay, davranış skorlamasında ise son 18-24 aydır (Thomas vd., 2002).

Örnekleme seçiminde dikkat edilmesi gereken diğer bir nokta; örnekleme büyüklüğünün ne kadar olacağı ve iyi-kötü kredilerin hangi unsurlara göre ayrılacağıdır. Örneklemedeki iyi müşteri-kötü müşteri oranı eşit mi olmalıdır, yoksa ana kütlede olduğu şekliyle mi temsil edilmelidir? Ana kütledeki oranına göre iyi-kötü oranı belirlendiğinde kötü kredilerin açıklanmasına yetecek kadar veri örneklemede bulunmayacağından dolayı genelde bu oran 50:50 olarak kabul edilir. Örneklemede iyi-kötü değişkenlerinin dağılımları aynı değilse buna izin verecek örneklemeyi elde etmek için sonuçların düzeltilmesi gerekir.

Lewis, örnekleme büyüklüğü ve iyi-kötü kredi oranı için; 1500 iyi ve 1500 kötü sayısını yeterli bulmuştur (Lewis, 1992). Pratikte ise, çok daha büyük örneklemeler kullanılır.

Örnekleme mevcut ana kütleden rasgele seçilmişse bunun gerçekten rasgele olduğuna emin olmak gerekir. Eğer ana kütle listesindeki iyilerden her 10 tanede bir tanesi seçilirse örneklemede bunun temsili %10 olacaktır. Bununla birlikte, şubelere gidilmesi zorunlu olduğunda ilk olarak yapılması gereken iyilerin kırsal ve kentsel olarak rasgele seçildiğinden emin olmaktır. Ya da örneklemenin seçildiği tarihte belirli ürünlerin pazarlaması yapılıp-yapılmadığı veya belirli kitlelere hitap edilip-edilmediğine bakılmalıdır. Örneğin belirli bir ay örnekleme olarak alındığında o ayda bir ürünün pazarlaması yapılıyor olabilir ve bu ürün gençlere hitap ediyorsa örneklemedeki genç oranı daha yüksek çıkması kaçınılmazdır.

2.6 İyi Müşteri-Kötü Müşteri Tanımlaması

Skor kart geliştirilmesinde aşamalardan biri iyi müşteri-kötü müşteri sınıflamasının ne şekilde yapılacağıdır. Bazı müşterileri kötü olarak tanımlama diğer tüm müşterilerin iyi olduğu anlamına gelmez. Müşterilerin iyi-kötü olması dışında en az 2 seçenek daha vardır. Birincisi "tanımlanamayanlar", ikincisi "yeterli gözlem değeri olmayanlar".

Genellikle KS geliştirmede ödemeler arasında 3 dönem sorun yaşananlar sorunlu kredi olarak nitelendirilir. Tanımlanamayanlar 2 dönem problem yaşayıp 3. döneme düşmeyenlerdir.

İyi-kötü ayrımı ne şekilde yapılırsa yapılsın KS tekniği etkilenmez. Örneklemeden "tanımlanamayanlar" ve "yeterli gözlem değeri olmayanlar" ı çıkarıp sadece iyi-kötü sınıflaması için skor kart geliştirmek gerekir. Tabii ki iyi-kötünün farklı sınıflandırılması farklı skor kart sonuçları geliştirecektir. Diğer bir sorun da kötü kredinin ekstrem şekilde tanımlanmasından kaynaklanmaktadır. Bu durumda çok az olan kötü kredi nedeniyle modelin

güvenilirliği sarsılabilir (Thomas vd., 2002).

2.7 Veri Kaynakları

İyi-kötü kredilerin ayırt edilebilmesi için karakteristikler gereklidir ve bu karakteristikler farklı yollarla elde edilebilir. Bunlardan en önemlisi kredi kartı başvuru formudur. Yapılacak analiz çeşidine uygun karakteristiklerin elde edilebilmesi için çeşitli formatlarda kredi kartı başvuru formu düzenlenebilir. Bu şekilde verilen cevaplar şıklara bağlı olarak aynı formatta olacaktır ve analiz aşamasına kolaylıkla geçilebilecektir.

Başvuru formundaki karakteristiklerin ne olacağına karar verilirken meşru cevaplar alınabilecek şekilde format oluşturulmalıdır. Örneğin, ev durumu karakteristiğine bakıldığında "size ait", "kira", "aile fertlerinden birine ait", lojman" seçenekleri olduğu varsayılsın. Tabi bu daha geniş seçenekli de olabilirdi. Sorun belirlenen seçeneklere "hayır" cevabı verildiğinde ne yapılacağıdır. Genel yaklaşım her karakteristik için "diğer" seçeneği eklenmesidir. Fakat cevabı gerçekten "hayır" olan karakteristikler de vardır. Bu da çelişki yaratabilir.

İş karakteristiği yüzlerce iş seçeneği olduğu durumlarda kodlanması zor bir alandır. Unvanlarla ilgili bilgiler çok açıklayıcı olamamaktadır. Örneğin başvuru yapan kişi yönetici seçeneğini işaretlediğinde; çok büyük bir fabrikada yönetici olduğu anlamı çıkabileceği gibi, küçük bir işletmede yönetici olduğu anlamı da çıkabilir. Daha genel yaklaşımlarla bu aşılmak istenmektedir.

Gelir karakteristiği de net bir şekilde ifade edilmediği takdirde karışıklığa neden olabilir. Aylık gelir mi, toplam gelir mi, yada hane halkının toplam geliri mi? Genellikle başvuru formlarında bunların her biri ayrı ayrı istenilmektedir ve analize farklı değişkenler olarak girmektedirler.

Sonuç olarak, verilerin geçerliliğini denetlemek gerekir. Mesela 18 yaş altı yada 100 yaş üstü değerler işaretlenebilir. Farklı karakteristikler için karakteristiğin dağılımı incelenerek uç olan durumlar gözlemlenebilir. Bu şekilde veri doğrulama hilekarlık tespitinde denenen bir yoldur. Verilen cevaplardaki uyuşmazlık çok sık gözlemlenebilmektedir.

Diğer bir problem de kredi kartı başvurusu yapan kişinin formu şube yetkilisiyle birlikte doldurup doldurmayacağıdır. Şube yetkilisi müşterinin başvurusunun kabul edilebilmesi için ne tür cevaplar vermesi gerektiği şeklinde müşteriyi yönlendiriyor olabilir. Bu durumda müşterinin gerçek bilgilerinden ziyade oluşturulan skor kartta başvurusunun onaylanması için

gerekli olan yanlış bilgiler girilmiş olacaktır ve bu da müşteri hakkında onay aşamasında yanlış karar verilmesi anlamına gelir.

Çizelge 2.1'de Finans evleri (Finance house), Amerika kredi kartları (U.S. credit card), İngiltere kredi kartları (U.K. credit card) kredi kartı formlarında istenilen bilgiler yer almaktadır.

Çizelge 2.1 Üç farklı başvuru formundaki karakteristikler (Thomas vd.,2002)

Karakteristik	Finans Evleri	Amerika Kredi Kartları	İngiltere Kredi Kartları
Posta kodu	X	X	X
İkamet süresi	X	X	X
Ev durumu	X	X	X
Mesleği	X	X	X
Çalışma süresi	X	X	X
Maaş	X	X	X
Diğer gelir	X	X	
Çocuk sayısı	X	X	X
Vadesiz hesap	X	X	X
Vadeli hesap	X	X	
Kredi kartları	X	X	X
Store kartları	X	X	X
Yaşı			X
Telefon numarası		X	X
Aylık giderleri	X		
Toplam varlıkları	X		
Araba yaşı	X		

2.8 Alt Ana kütle Belirlemek (Determining Subpopulation)

Bundan sonraki iki alt bölüm skorlama sisteminde hangi değişkenlerin ne şekilde kullanılması gerektiğiyle ilgilidir. Değişkenlerin belirlenmesinde önemli bir teknik de ana kütlenin alt ana kütlelere ayrılarak her bir sınıf için ayrı skor kart geliştirilmesidir. Bu teknik istatistiksel nedenlerle olduğu kadar kredi politikası açısından da uygulanır. Örneğin DS'de ilgili banka ile son dönemlerde çalışan müşteriler olduğu gibi uzun süredir çalışan müşteriler de mevcuttur. Örneğin 6 aylık ortalama bakiyelere bakıldığında ikinci tip müşteriler için ulaşılabilir bir veridir fakat birinci tip müşteriler için ulaşılabilir bir veri değildir. Bankanın uygulayacağı bir başka politika da genç müşterileri yaşlı müşterilerden farklı olarak değerlendirmek olabilir. Bu durumda da 25 yaş altı ve üstü için ayrı skor kart oluşturma

yoluna gidilebilir.

Ana kütlenin alt ana kütlelere ayrılmasının istatistiksel nedeni ise bir karakteristik ve diğer karakteristikler arasında çok fazla etkileşim olmasıdır. Bundan dolayı karakteristiklerin farklı davranışları için farklı skor kart geliştirilmektedir. Uygulamacılar diğer karakteristiklerle en çok etkileşim halinde olan karakteristik için Sınıflandırma Ağacı (Classification Tree) kullanarak skor kart oluşturabilirler. Sınıflandırma ağacındaki en iyi ayrılma noktası farklı skor kart oluşturulması için uygun alt ana kütleleri göstermektedir.

Alt ana kütlelere bölmede politik nedenler istatistiksel nedenlere göre daha fazla kullanılmaktadırlar. Banasik'in belirttiği gibi alt ana kütlelere bölme her zaman tahmini geliştirmez. Alt ana kütleler farklı değilse, daha küçük örneklemeyle çalışmak skor kartın performansını, alt ana kütlelere bölmenin sağladığı avantaja göre daha çok etkileyecektir. Bu durumda alt ana kütleye bölmenin bir anlamı yoktur (Banasik vd., 1996).

2.9 Karakteristiklerin Genel Olarak Sınıflanması

Her bir karakteristik verilecek cevapların birbirleriyle ilişkilerine göre alt ana kütlelere ayrılmalıdır. Alt ana kütlelere bölme işlemi cevapların kategorik-sürekli olmasına bağlı olarak iki farklı şekilde yapılabilir. Kategorik veriler için; çok fazla farklı cevap şıkkı olabilir ve her bir cevapla ilgili yeterli çoğunluk sağlanamadığı için güvenli bir analiz yapılamaz.

2.9.1 Ki- Kare İstatistiği

Sorulara verilen cevaplarda i niteliğini sergileyen iyi-kötü müşteri sayısı g_i ve b_i şeklinde gösterilir. g ve b ise toplam iyi-kötü sayısıdır.

$$\hat{g}_i = \frac{(g_i + b_i)\,g}{g+b} \tag{2.1}$$

$$\hat{b}_i = \frac{(g_i + b_i)\,b}{g+b} \tag{2.2}$$

$$s^2 = \sum_{i=1} \left(\frac{g_i - \hat{g}_i}{\hat{g}_i} \right)^2 + \left(\frac{b_i - \hat{b}_i}{\hat{b}} \right)^2 \tag{2.3}$$

χ^2 (ki-kare) istatistiğidir. Bu istatistik mümkün olduğunca farklı sınıflarda iyi:kötü oranı aynı olacak şekilde sınıflar oluşturmaya yarar. Ve $k-1$ (k karakteristiklerin sınıf sayısı) serbestlik

derecesinde ki-kare istatistiğiyle kıyaslanır. Ayrıca farklı sınıflardaki iyi:kötü oranlarının ne kadar farklı olduğunu ölçmek için de kullanılır.

Örneğin "ev durumu" değişkeni için; ev sahibi olma, kiracı olma, ailesi ile yaşama, diğerleri şeklinde dört seçenek olduğu varsayılsın ve bu seçeneklerin birbirlerinden en iyi şekilde ayrılarak şıkların belirlenmesi istenilsin. Burada cevabı aranan soru; ev sahibi/kiracı ayrımının mı yoksa ev sahibi/ailesiyle yaşama ayrımının mı daha belirleyici olduğudur. Karar verilmesi için şu adımlar izlenir;

- Ev sahibi/kiracı ayrımı için ki-kare istatistiği ev sahibi, kiracı, diğerleri nitelikleri için hesaplanır.
- Ev sahibi/ailesiyle yaşama ayrımı için ki-kare istatistiği ev sahibi, ailesiyle yaşama, diğerleri nitelikleri için hesaplanır.
- İki ki-kare değeri kıyaslanarak büyük olan ki-kare değerinin iyi bir ayrımı gösterdiği yönünde karar alınır.

2.9.2 Somer's D Concordance İstatistiği

Somer's D Concordance İstatistiği karakteristiklerin sınıflarının en düşük iyi oranından en yüksek iyi oranına doğru sıralanması varsayımını test eder. Concordance istatistiği kötü özellikli müşteriler (x_B) iyi özellikli müşterilerden (x_G) daha düşük sınıfta olduğunda iyiler arasından rasgele bir iyi seçilmesi ve kötüler arasından rasgele bir kötü seçilmesi ihtimalini tanımlar. Bu olasılık ne kadar yüksek olursa iyi-kötü ayrımı ana kütleyi o kadar iyi yansıtır. D istatistiğinin tam tanımı değişkenlerin beklenen sonucudur. Sıralamada kötüler iyilerden aşağıdaysa 1, kötüler iyilerden yukarıdaysa -1, ikisi de aynı seviyedeyse 0 değerlerini alır.

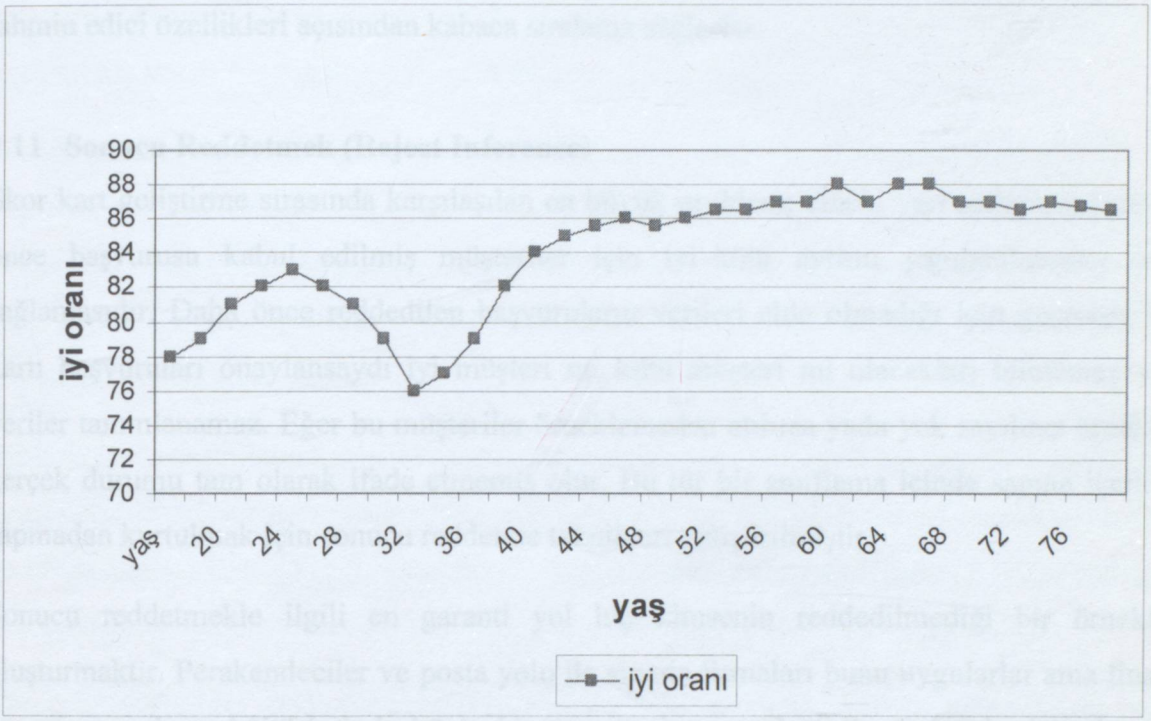
$$D = 1.P\{x_B < x_G\} - 1.P\{x_B > x_G\} + 0.P\{x_B = x_G\}$$

$$= \sum_i \frac{\left(\sum_{j<i} b_j\right) g_i - \left(\sum_{j<i} g_j\right) b_i}{b_g} \qquad (2.4)$$

Uygulamada ki-kare istatistiği ile aynı adımlar takip edilir.

Kategorik verilerle işlem yapıldığı gibi sürekli verilerle de işlem yapılması gerekebilir. Genellikle regresyonda, sürekli bir değişkenle çalışıldığında değişken kendi içinde ayrılır ve sadece ağırlıklandırılması gereken katsayılar hesaplanır. Çünkü istenilen belirli ilişkilerin ortaya çıkarılmasıdır. Eğer riski tahmin etmeye çalışmak yerine, sürekli değişkenler ayrılırsa

yeni oluşturulan değişkenlerde riskin monoton olacağı garanti edilir. Örneğin, Şekil 2.1 kredi kart sahiplerinin tüm yaş gruplarında iyi müşteri olma oranlarını göstermektedir. Burada dikkati çeken şeklin monoton olmadığıdır. İyi oranı önce artmış, sonra azalmış, sonra tekrar artmıştır. 30-40 yaş arasında ve gençlik dönemlerinde bu oranın daha düşük olduğu görülmektedir. Bundan dolayı yaşları belirleyici alt gruplara ayırmak gerekmektedir. 18-21, 21-28, 29-36, 37-59, 60+.



Şekil 2.1 İyi oranı-yaş grafiği (Thomas vd., 2002)

Bu tür sürekli veriler ne şekilde sınıflandırılmalıdır? Ana kütlenin %10'unu alacak şekilde 10 grup oluşturulabilir ya da %5'ini alacak şekilde 20 grup ya da %1'ini alacak şekilde 100 grup. Önemli olan en yakın komşular bir araya gelecek şekilde bu sınıflamanın yapılmasıdır.

2.10 Karakteristik Seçimi (Choosing Characteristics)

Karakteristikler kabaca sınıflandığında, çok sayıda durum (seçenek) ile karşılaşabilir. Başvuru skor kartıyla ilgili olarak 30–40 karakteristik için 200'den fazla durum söz konusudur. DS'de ise bazen yüzlerce karakteristik ve bunlarla ilişkili olarak 1000'den fazla durum olabilir. Her bir durum ikili karakteristik olarak alındığında da, lojistik regresyon ya da sınıflandırma

ağacında başa çıkabilmek zor olacaktır. Daha da önemlisi skor kart çözümlense de sonuçları yöneticilerin anlayamayacağı karmaşalıkta olacaktır. Bundan dolayı karakteristik sayısı 20'den fazla olmamalıdır. Peki bu karakteristikler nasıl seçilmelidir?

Bazı karakteristikler zayıf tahmin ediciler oldukları için göz ardı edilebilirler. Bunun için de daha önce hesaplanan istatistiklerden yola çıkılır. χ^2 Ki-kare İstatistiği, Information İstatistiği F, Concordance İstatistik D her bir değişken için hesaplanırsa, bu istatistikler değişkenlerin tahmin edici özellikleri açısından kabaca sıralama sağlarlar.

2.11 Sonucu Reddetmek (Reject Inference)

Skor kart geliştirme sırasında karşılaşılan en büyük problem; eldeki veri setinin sadece daha önce başvurusu kabul edilmiş müşteriler için iyi–kötü ayrımı yapılabilmesine imkan sağlamasıdır. Daha önce reddedilen başvuruların verileri elde olmadığı için geçmişte kredi kartı başvuruları onaylansaydı iyi müşteri mi kötü müşteri mi olacakları bilinemez ve bu veriler tanımlanamaz. Eğer bu müşteriler örneklemeden atılırsa yada yok sayılırsa örnekleme gerçek durumu tam olarak ifade etmemiş olur. Bu tür bir sınıflama içinde sapma içerir. Bu sapmadan kurtulmak için sonucu reddetme teknikleri geliştirilmiştir.

Sonucu reddetmekle ilgili en garanti yol hiç kimsenin reddedilmediği bir örnekleme oluşturmaktır. Perakendeciler ve posta yolu ile sipariş firmaları bunu uygularlar ama finansal organizasyonların kültürlerinde böyle bir şey söz konusu değildir. Geleneksel bankacılıkta borç ödeme gecikme oranı kriter olarak alınır çünkü bu gecikmelerdeki kayıpların çok büyük bir miktar olmadığı varsayılır. Bu durumda, herkes alınmayarak fakat borcun gecikme riski x olan $p(x)$ oranı alınarak kayıplar azaltılabilir. Bu oran x'e göre değişir; x 1'e yakınken oldukça küçüktür ve x küçüldükçe 1'e yaklaşır (Thomas vd., 2002).

Sonucu reddetmekle ilgili teknikler Kötü Olarak Tanımlama (Define as Bad), Dış Değerleme (Extrapolation), Geliştirme (Augmentation), Dağılımların Karışımı (Mixture of Dstribution), Üç Grup Yaklaşımı (Three-group Approach) olarak bilinmektedir. Tekniklerin ayrıntılarına bu çalışmada değinilmemiştir.

2.12 Sonucu Geçersiz Kılmak Ve Skor Karta Etkisi (Overrides and their effect in the scorecards)

Geçersiz kılmak kredi veren kişinin skorlama sistemi sonucuna ters bir karar vererek hareket etmesi durumudur. Yüksekken geçersiz kılmak (High-side overrides-HSOs) kesim noktası

(cut off) üstünde çıkmasına rağmen başvuruya kredi verilmemesidir. Alçakken geçersiz kılmak (Low-side overrides-LSOs) ise kesim noktası altında çıkmasına rağmen başvuruya kredi verilmesidir. Bunun nedeni kredi veren kişinin o başvuruyla ilgili skor karttan daha çok şey biliyor olması, yada şirketin mevcut politikası olabilir.

Bilgiye dayalı geçersiz kılma genellikle çok nadir olur fakat şubeden kredi başvurusuyla ilgili gelen bilgiler başvuru yapanın tüm hikayesini anlatıyor nitelikte olmaz. Örneğin başvuru yapanın maaşındaki artış onaylansa da henüz zamlı maaşını hiç almamış olabilir. Bu tip durumlarda kararı tam tersine çevirecek bilgiler faydalı olabilir.

Politikadan kaynaklanan geçersiz kılma ise kredi veren kişi belli karakteristikleri kredi kapsamı dışında tutma kararı almış olabilir. Örneğin, öğrencilerin uzun dönemde iyi beklentileri olduğu ve hesaplarını aşan harcamaları mesele etmediklerinden dolayı onlarla ilgili farklı bir kredi politikası izlenebilir. Bu tip durumlarda geçersiz kılma doğru karar noktasında yardımcı olacaktır.

Bir çok LSO doğrulamaktadır ki başka ürünler için de çalışılabilecek birçok müşteri banka tarafından ilk başvuruları reddedildiği için tamamen kaybedilebilmektedirler. Bundan dolayı ne kadar kazanç kaybedildiği yada ileride borcun ödenmeme ihtimalinin ne kadar olduğu iyi analiz edilmelidir.

Kredi başvurularını değerlendiren kişiler skorlama sistemini çalıştırdıktan sonra kişisel kararlarla geçersiz kılma yapıyorlarsa bu çelişkidir. Çünkü ya skorlama sistemi hatalıdır, yada kendi görüşleri. Eğer kredi değerlendiren farklı bir kriterle karar alıyorsa o zaman o kriter de skorlama modeline eklenerek model yeniden yapılandırılmalıdır. Bu durum kredi değerlendirenler açısından kabul edilemez olabilir. Uzmanlar kendi pozisyonlarında kendilerini ön plana çıkaracak kararlar alıyor olmak isterler ama bir sistem sayesinde bu otomatik olarak yapıldığında kendilerinin ön plana çıkması mümkün değildir. Bundan dolayı sistemin çalışmasını engellemek isteyebilirler.

2.13 Kesim Noktasının Belirlenmesi (Setting the Cutoff)

Yeni bir skor kart oluştururken kesme değerinin seçimi için çeşitli yollar vardır.

En basit yaklaşım, varolan skor kartla aynı kabul oranını veren bir kesme değerini seçmektir. Skor kartta, onun programlamasından ve diğer uygulamalarından yeterince emin olana kadar bu seçim işlemi devam edebilir. Yeni skor kartın amacı kabul oranını arttırmak ve kötülerin seviyesini korumak olsa bile, en azından birkaç haftalığına da olsa hali hazırdaki kabul oranı

kesme değeri olarak korunabilir, kullanılabilir. Böylece çok şey değişmeden sistemde meydana gelebilecek herhangi bir anormallik ortadan kaldırılmış olunur. Anormallikler karşılıklı değişen setlerde ortaya çıkar, örneğin önceden reddedilip şimdi kabul edilenler veya tam tersi.

Skor kart geliştirme safhasından sonra doğrulama aşamasında analiz, verilen bir kesme değeri için karşılıklı değişen setlere her bir bireyin özelliği dikkate alınarak uygulanır.

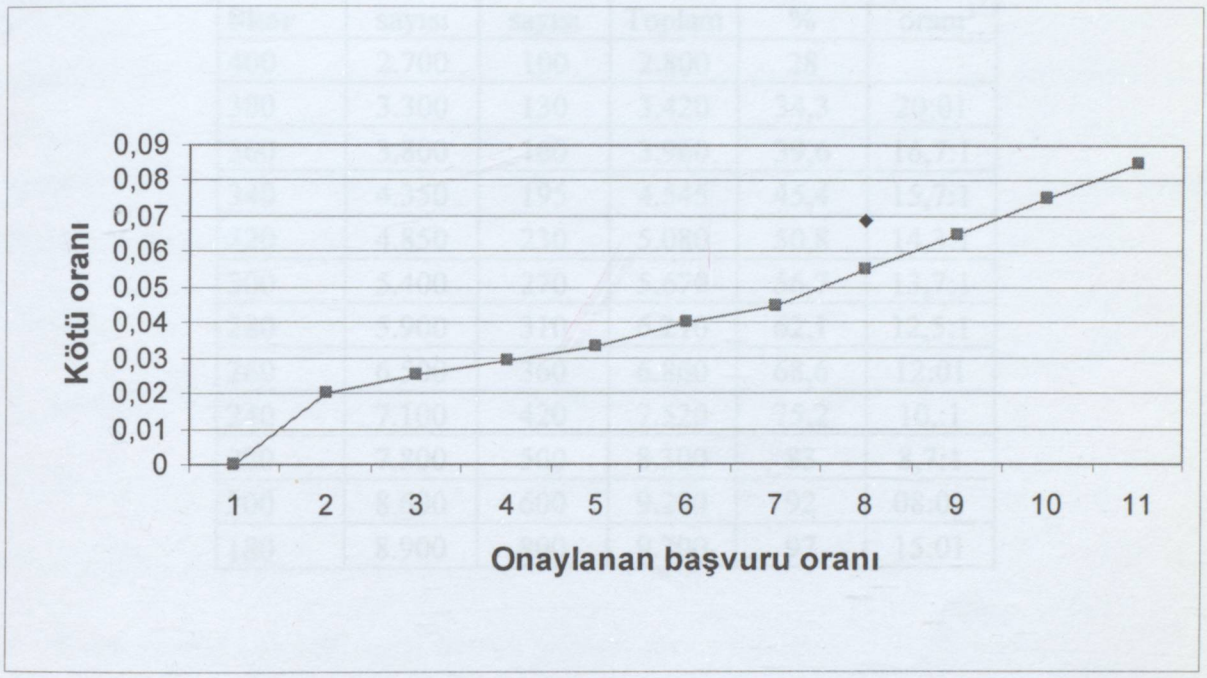
Aynı kabul oranını uygulamak sadece birkaç hafta için tavsiye edilir. Aynı kabul oranını korumak kesinlikle kötü oranının düşmesini sağlayacaktır. Buna karşın, bu fayda ancak bir kaç ay sonra gözlenebilir. Bir çok firmada, kabul oranını yükseltmeye yönelik bir baskı vardır.

Eğer yeni skor kartın eskisinin yerini alacağı göz ardı edilirse, bu durumdaki yaklaşım kesme değerini başa baş noktası değerine eşitlemektir. Grupların geri ödeme, oturduğu yer, ödenmemiş borç, en son ödeme bilgisi ve bunların zamanları hakkında gelecek performansları ile ilgili kusursuz bir bilgiye sahip olunduğu kabul edilsin. Bu durumda, eğer bu başvurular kabul edilirse şirkete ne kadar para getireceğinin bilinmesi gerekir. Her bir olası skor için gelir ve gideri dikkate alarak bu veriler üretilebilir. Temel bir görüşe göre, bütün başvuruları kabul etmek küçükte olsa bir kar getirecektir. Başka bir görüşe göre de karın sıfır olacağı kabul edilir. Bu teknik basit fakat yanlış değil, sadece bir başlangıç noktasıdır.

Kar, net gelir veya net şimdiki değerden ziyade, yatırımların geri dönüşü dikkate alınmalıdır. Skorlamanın bakış açısına göre önemli olan nokta, kesme değerinin kardan ziyade geri dönüşe göre ayarlanması gerektiğidir. Böylece, kesme değeri ayarlanan noktada başvuru yapanlar gerekli minimum geri dönüş eşiğini karşılamak zorunda kalacaklardır. Belirlenen eşik değerinden yüksek skorlara sahip olanların bu minimum geri dönüşü karşıladığı bilinir.

Kar ve geri dönüşün yanı sıra, diğer bir finansal terim de dikkate alınmalıdır. Bu terim ise sabit maliyetlerin dağıtılmasıdır. Kar, geri dönüş ve sabit maliyetler kombinasyonunu tatmin edebilecek bir kesme değeri için net bir şey söylenemez. Şirketler bunun cevabını kendi içlerinde bulurlar. Bir başvuru için kar ve geri dönüşün tespitinde başka bir problem ile karşılaşılır. Geleceğin geçmiş ile aynı olduğu kabul edilsin. Borcun peşin veya erken ödenip ödenmeyeceği bilinemez. Yalnızca, benzer borçların geçmiş performanslarına dayalı kabullerde bulunulabilir. Gelecekte karşılaşılabilecek problemler için bir iyileştirme performansına yönelik kabuller yapılmalıdır. Böylece, kesme değerinin belirlenmesinde, birbiriyle ilişkili çeşitli kredi ve finans konuları dikkate alınmalıdır.

Geliştirilen skor karttan , Şekil 2.2'ye benzer bir grafik elde edilebilmelidir. Örneğin, %50 kabul oranında (2), kötüler %2; %70 kabulde (6) bu oran %4'e çıkıyor; kabul oranı %90'da (10) kötülerin oranı yaklaşık %7.5 oluyor. Belirteç, yeni skor kart uygulamadan önce şuan ki durumu göstermektedir. Açıkça görülüyor ki, eğer bu pozisyon sağa kaydırılırsa kötü oranının sabit kaldığı buna karşın kabul oranının yükseldiği gözlemlenir. Gerçekte, karenin eğri üzerideki aşağı yukarı-sağa sola hareketleri şu anki durum üzerinde yapılan geliştirmeleri göstermektedir. Fakat, ortak görüşe göre, sadece grafik üzerinde uygulanan pozisyonlar gerçeği yansıtır.



Şekil 2.2 Strateji eğrisi (Thomas vd., 2002)

Genellikle üzerinde çalışılan örnekleme ait skor kartın sonuçlarını içeren veriler bir tablo halinde gösterilir. Çizelge 2.2 her bir kesme skorunun bütün ana kütle üzerindeki etkisini göstermektedir.

Kar ve geri dönüş bir skor kartın kesme değerinin kararında itici güç olmasına karşın, çeşitli operasyonel faktörlerde bulunmaktadır. Örneğin, eğer kabul oranı yükseltilmek isteniyorsa "Toplam para değerini yükseltecek kadar yeterli fon var mı?" ayrıca, "Çok sayıda vakayı inceleyip tamamlayacak, fonları geri çekebilecek kadar operasyonel kapasite mevcut mu?", "Belirlenen zamandan önce para çekilmesini karşılanabilecek mi?" sorularına cevap aranır.

Eğer davranışsal bir skor kart uygulanırsa, benzer konular ortaya çıkar; "Telefon ve mektup sayısında çok fazla yükselme var mı?", "Yeterli kapasite mevcut mu?", "Finansal açıdan, kredi limitleri arttırılabilir mi, bunun için onaya ihtiyaç duyuluyor mu?" yada "Ödenmiş borçlar için provizyonlar yükseltilebilir mi?"

Çizelge 2.2 Run-book örneği (Thomas vd., 2002)

Skor	Skor altındaki iyi sayısı	Skor altındaki kötü sayısı	Toplam	Toplam %	iyi-kötü oranı
400	2.700	100	2.800	28	
380	3.300	130	3.420	34,3	20:01
360	3.800	160	3.960	39,6	16,7:1
340	4.350	195	4.545	45,4	15,7:1
320	4.850	230	5.080	50,8	14,3:1
300	5.400	270	5.670	56,7	13,7:1
280	5.900	310	6.210	62,1	12,5:1
260	6.500	360	6.860	68,6	12:01
240	7.100	420	7.520	75,2	10,:1
220	7.800	500	8.300	83	8,7:1
200	8.600	600	9.200	92	08:01
180	8.900	800	9.700	97	15:01

3. KREDİ SKORLAMADA KULLANILAN YÖNTEMLER

3.1 Diskriminant analizi

Diskriminant analizi belirli aralıklarda tanımlanmış kategorik olarak belirlenen bağımlı (kriter) değişken ile bağımsız (kestirimci) değişkenler arasında fonksiyonel bir ilişki kurarak gruplar arası ayırımı sağlayan bir data analiz tekniği olarak tanımlanabilir (Malhotra ve Naresh, 1993). Diskriminant analizi, değişkenlerin değerlerinin gruplar içinde olabildiğince az, gruplar arasında olabildiğince fazla dağıldığı lineer bileşimler bulmaya çalışır (Yıldız, 1995).

Diskriminant Analizi, iki ve daha fazla sayıda grubun çok sayıda değişkene göre karşılaştırılmasını sağlayan bir tekniktir. Analizin amacı, grupların hangi değişkenler açısından birbirinden farklılaştığının ortaya çıkarılmasıdır. Diğer bir ifadeyle, grupları ayırıcı özelliklerinin belirlenmesidir.

Diskriminant analizinin temel görevleri:

- Grupların birbirlerinden ayrılmasını sağlayan bir model kurmak
- Yeni gözlemleri uygun sınıflara yerleştirmek.

Diskriminant analizi varsayımları;

- X veri matrisi çok değişkenli normal dağılım göstermelidir.
- Varyans kovaryans matrisleri homojen olmalıdır.
- Değişkenlerin ortalamaları ve varyansları arasında anlamlı korelasyon bulunmamalıdır.
- Değişkenler arasında çoklu bağımlılık (multicollinearity) bulunmamalıdır.
- X matrisi grupların birbirlerinden ayrılmasında rol oynamayacak gereksiz değişkenler içermemelidir (redundancy).

Araştırmacının amacı doğrultusunda farklı çözüm teknikleri geliştirilmiştir. Bunlardan bazılarının aşağıdaki gibi özetlenmesi mümkündür.

- Bireylerden ya da birimlerden oluşan iki veya daha çok sayıda grubun ortalama özellikleri açısından ,istatistiksel olarak anlamlı bir ayrım gösterip göstermediklerinin belirlenmesi
- Grupların ayrımını yapmada her bir kestirimci değişkenin etkisinin belirlenmesi

- Herhangi bir birimin hangi gruba dahil olabileceğinin kestirilmesi
- Grup içi değişime oranla gruplar arası ayırımı maksimize eden kestirimci değişkenlerin bunları en iyi yansıtan doğrusal (veya doğrusal olmayan) bir fonksiyona dönüştürülmesi

Burada söz konusu olan bireyleri hatalı sınıflandırma olasılığını minimum yapacak biçimde bir fonksiyon elde etmektir (Yıldız, 1995).

Diskriminant analizi, grup kovaryans matrislerinin eşit olup olmamasına göre farklı biçimlerde uygulanmaktadır. Grup kovaryans matrisleri benzer olduğunda Doğrusal Diskriminant Analizi, benzer olmadığında da Karesel Diskriminant Analizi kullanılır (Özdamar, 2002b).

3.2 İki Grup Karesel Diskriminant Analizi

İki gruplu diskriminant analizinin temeli bireyler hakkında toplanan, çoklu ve korelasyonlu verilerden hareketle, iki grup öğeleri arasında maksimum ayırımı sağlayan tek bir karesel bileşim modeline geçmeye dayanır. Böylece çok değişkenli profiller tek değişkenli veriler kümesine dönüştürülür.

Kestirimci değişkenlerin dağılımlarına ilişkin belirli varsayımların yapılması halinde iki grubun ortalamalarından oluşan vektöre ilişkin ortalamalara karşılık gelen odak noktaların eşitliği hipotezi sınanabilir. Eğer farklı oldukları sonucuna varılırsa, söz konusu iki gruba yeni çok değişkenli profiller ekleyerek, bunların hangi gruba dahil olacaklarını diskriminant analizi ile belirlemek mümkün olur. Bu da her profil için bir diskriminant puanı (skoru) hesaplayarak ve bu değeri örneklemeden elde edilen karesel bileşen değeriyle kıyaslayarak yapılabilir. Eğer söz konusu değer sınır değerinin bir yakasındaysa 1. gruba , diğer yakasındaysa 2. gruba sokulur. Sınır değerine eşitse herhangi bir gruba sokulabilir (Yıldız, 1995).

Diskriminant analiz fonksiyonlarını oluştururken ele alınan yapının varsayımlarını belirlemek genel yapının formüle edilmesi açısından önem kazanmaktadır. Her şeyden önce kriter (bağımlı-nitel) Y değişkeninin, gösterge değişkenle kodlanacağı (0,1) belirtilir. Örnekleme büyüklükleri (gruplardaki gözlem sayıları) eşit olarak alındığında, ortalamalar vektörünü oluşturan elemanların toplamı sıfır olacaktır. Burada grup değerleri genel ortalamadan sapmalar olarak alındığında, birinci grubun değerleri negatif, ikinci grubun pozitif sapmalardan oluşacak ve toplamları sıfır olacaktır (Yıldız, 1995).

A ve B isimli iki toplum düşünüldüğünde bu toplumlarda n birimlik p tane birbirleri ile ilişkili

gözlemler var ise X veri matrisinde gözlemler toplumdan topluma farklılıklar sergileyecektir. X matrisi A toplumundan alınan X_1 gözlem matrisi ve B toplumundan alınan X_2 gözlem matrisinden oluşmaktadır.

Fisher, çok değişkenli X gözlem matrisini A ve B toplumlarına ilişkin tek değişkenli y değerlerine dönüştürmeyi ve bu y değerlerinin toplumlara göre birbirlerinden olabildiğince farklı olmalarını sağlamayı amaçlamıştır. y'leri ise X gözlem matrisinin karesel bileşenleriyle elde etmiştir.

X_1 ve X_2 gözlem matrislerine ait kovaryans matrisleri S_1 ve S_2'dir

Bazen katsayılar ölçeklendirilerek (scaled) ya da normalize edilerek kullanılmaktadır. Normalizasyonun temel amacı katsayıları genel katsayılar içinde ağırlıklandırarak elemanların kolay yorumlanmasını sağlamaktır.

İki tür normalizasyon yaklaşımı vardır;

$$b^* = \frac{b}{\sqrt{b'b}} \tag{3.1}$$

$$b^* = \frac{b}{b_1} \tag{3.2}$$

burada b_1 en büyük karesel bileşendir.

Gruplar arasındaki farklılığı maksimize edecek bir diskriminant fonksiyonu aracılığı ile grupları birbirinden ayırmak mümkün olacaktır. Bu nedenle ortak bir diskriminant fonksiyonu belirlenir. i ve j grupları arasındaki diskriminant fonksiyonu;

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p \tag{3.3}$$

şeklinde yazılır.

Bu fonksiyondaki b_i karesel bileşenleri de ortalama fark vektörü ile aşağıdaki gibi bulunur.

$$b = (S_1^{-1} - S_2^{-1})(\bar{x}_1 - \bar{x}_2)' \tag{3.4}$$

$A = (S_1^{-1} - S_2^{-1})$ olmak üzere her bir grubun b_i katsayılar vektörü;

$$b_{ij} = A^*(\bar{x}_1 - \bar{x}_2)' \quad i=1, 2,..., g \quad j=1, 2,..., p \tag{3.5}$$

biçiminde hesaplanır.

Sabit değer b_0 katsayısı ise;

$$b_0 = -(1/2)\bar{x}'(S_1^{-1} - S_2^{-1})\bar{x} \tag{3.6}$$

Diskriminant analizi her bir grup için birer diskriminant fonksiyonu hesaplar.

$$Y_i = b_{0i} + b_{1i}X_1 + b_{2i}X_2 + ... + b_{Pi}X_P \quad i=1, 2, ...,p \tag{3.7}$$

Bu fonksiyonda b_{0i} sabit değeri, b_{ij} ise karesel bileşenleri (kanonik değişkenler) belirtmektedir.

S_1 ve S_2 kovaryans matrisleri ve $\bar{x}_i$ *i.* grup ortalama vektörü olmak üzere her bir grubun b_i katsayılar vektörü;

$$b_{ij} = (S^{-1} - S_2^{-1})(\bar{x}_i) \quad i=1, 2,..., g \quad j=1, 2,..., p \tag{3.8}$$

biçiminde hesaplanır.

Sabit değer ise;

$$b_{i0} = -(1/2)\bar{x}_i'(S_1^{-1} - S_2^{-1})\bar{x}_i \tag{3.9}$$

biçiminde hesaplanır.

Gruplara göre belirlenen sabit ve kanonik katsayılar değişken değerleri ile çarpılarak karesel diskriminant fonksiyonları hesaplanır.

1. grup için diskriminant fonksiyonu;

$$Y_1 = b_{10} + b_{11}X_{11} + b_{12}X_{12} + ... + b_{1p}X_{1p} \tag{3.10}$$

eşitliğinde değerler yerine konularak hesaplanır.

2.grup için diskriminant fonksiyonu;

$$Y_2 = b_{20} + b_{21}X_{21} + b_{22}X_{22} + ... + b_{2p}X_{2p} \tag{3.11}$$

eşitliğinde değerler yerine konularak hesaplanır.

p değişkenli g grup arasındaki karesel uzaklığı veren ve Mahalanobis tarafından ileri sürülen D^2 uzaklığı aşağıdaki gibi hesaplanır.

$$D_{ij}^2 = \left(\bar{x}_i - \bar{x}_j\right)' S^{-1} \left(\bar{x}_i - \bar{x}_j\right) \tag{3.12}$$

D^2 uzaklığının i ve j gruplarını birbirinden ayırmada etkin rol oynayıp oynamadığı test edilebilir. Bu amaçla Hotelling T^2 yaklaşımı ile D^2 'nin önemliliği test edilebilir.

$$T^2 = \left(\frac{n_1 * n_2}{n_1 + n_2}\right) D^2 \tag{3.13}$$

T^2 'nin önemliliğinin test edilmesi için F yaklaşımından yararlanılır.

$$F = \frac{(n_1 + n_2 - p - 1)}{p(n_1 + n_2 - 2)} T^2 \tag{3.14}$$

F'in önemliliği; $p, (n_1 + n_2 - p - 1)$ serbestlik dereceli F dağılımının kritik değerleri kullanılarak belirlenir $\left[F(\alpha, p, (n_1 + n_2 - p - 1)\right]$.

3.2.1 Kovaryans Matrislerinin Eşitliğinin Sınanması

Eğer diskriminant fonksiyonu hesaplamadan yada Hotelling T^2 testini uygulamadan önce kovaryans matrisleri homojenliği test edilmek istenirse, başka bir test gerekli olacaktır. Barlett, iki yada daha fazla kovaryans matrisinin eşitliğini test etmek için bir χ^2 yaklaşımı geliştirmiştir. Daha sonra Box, F testini temel alan daha gelişmiş ve karmaşık bir yöntem ortaya atmıştır.

İki toplumdan oluşan ve p değişkenin kullanıldığı bir örnekleme düşünülsün. Ayrıca G kovaryans matrisi S_g (g = 1,2,.....G) ve toplam gözlem sayısı $\sum_{g=1}^{G} n_g = n$ olan toplanmış

gruplar içi kovaryans matrisi S göz önünde tutulsun. Bu durumda aşağıdaki sıfır hipotezi yazılabilir.

$$H_0: \Sigma_1 = \Sigma_2 = \ldots\ldots = \Sigma_g = \ldots = \Sigma_G$$
$$H_1: \Sigma_1 \neq \Sigma_2 = \ldots\ldots \neq \Sigma_g \neq \ldots\ldots \neq \Sigma_G$$

Burada Σ_g, g grup için ana kütle kovaryans matrisini göstermektedir. Barlett tarafından geliştirilen kovaryans matrislerinin eşitliğini testinde kullanılacak, iki ve daha çok grup için genel formül aşağıdaki gibidir.

$$B = (n-G)\ln|S| - \sum_{g=1}^{G}(n_g - 1)\ln|S| \qquad (3.15)$$

Bulunan sonuç $s.d = \frac{1}{2}[(G-1)(n)(n+1)]$'li χ^2 değerleriyle karşılaştırıldığında eşitlik reddedilirse lineer diskriminant fonksiyonu uygulanamaz. Eğer kovaryans matrisleri eşit olmazsa odak noktaların eşitliğinin sınanmasında sistematik bir hata ortaya çıkar. İki gruplu durumda kovaryans matrisleri eşit olmaması halinde Hotelling T^2 testindeki sıfır hipotezinin daha sık kabul edildiği görülecektir.

Dahası, sınıflandırma hataları terimleriyle birbirine eşit olmayan kovaryans matrisleri halinde lineer bir diskriminant fonksiyonunun kullanımı, bu matris en çok toplanmış gruplar içi kovaryans matrisine katkıda bulunduğu için, büyük kovaryans matrisine sahip gruba pek çok noktanın atanmasına yol açacaktır.

Cooley ve Lohnes'in değindiği gibi dikkate değer ölçüdeki büyük örneklemelerde, kovaryans matrislerinin eşitliği sıfır hipotezinin, reddedilmesinin mümkün olduğu varsayımı genel olarak kabul görmüştür. Böylece odak noktalarının eşitliğinin testinde Hotelling T^2 testinin yeterince kuvvetli olduğu varsayımı ile çoğu araştırmacı kovaryans matrislerinin eşitliği üzerinde durmaz. Sonuçta kuvvetlilik (robustness) konusunda yeni gelişmeler olmadan bu ilksel testlerin kabul edilmesi önerilir (Yıldız, 1995).

Buraya kadar diskriminant analizi için en iyi ayırıcı değişkenlerin bilindiği varsayımıyla modeller incelenmiştir. Tabii ki potansiyel ayırıcı değişkenlerin olduğu fakat diskriminant analizi için en iyi ayırımı sağlayan değişken setinin bilinmediği durumlar da sözkonusu olabilir. Adımsal Diskriminant Analizi (ADA) diskriminant fonksiyonu için en iyi değişken setini seçmede kullanılan faydalı bir tekniktir (Sharma, 1996).

ADA ileri doğru seçme (forward selection), geriye doğru seçme (backward selection ve adımsal seçme (stepwise selection) olmak üzere üç başlık altında incelenmektedir.

Adımsal seçme tekniği ileri doğru seçme ve geri doğru seçme tekniklerinin kombinasyonu şeklinde çalışmaktadır. Diskriminant fonksiyonunda hiçbir değişken yokken analiz başlar ve her adımda yeni değişkenler eklenir ya da elenir. Diskriminant fonksiyonundaki değişkenler istatistiksel kriter olarak belirlenen ayırma gücünden (discriminating power) daha düşükse modelden çıkarılır ya da tam tersi şekildeyse modelden elenir. Süreç hiçbir değişkenin modele girmediği ve çıkmadığı anda durdurulur (Sharma, 1996).

Değişkenler arasında ilişki yoksa üç teknik de aynı sonucu verecektir. Bununla birlikte verilerde ciddi bir çoklu doğrusal bağlantı (multicollinearity) problemi varsa sonuçlar birbirlerinden oldukça farklı çıkacaktır. Araştırmacılar veri setlerinde çoklu doğrusal bağlantı olduğuna inanırlarsa genellikle adımsal diskriminant analizini seçerler.

Genellikle kullanılan seçme kriteri Wilks' Λ, Roo's V, Mahalanobis Kareli Uzaklıklar (Squared Distance), Gruplararası F Oranıdır (Between Group F Ratio). Bunlardan uygulamalarda kriter olarak daha sık seçilen Wilks' Λ açıklanacaktır.

Wilks' Λ

$$W=\frac{SS_w}{SS_b+SS_w} \tag{3.16}$$

Wilks' Λ grup içi kareler toplamının kareler toplamına oranıdır. Her adımda diskriminant modelindeki mevcut değişkenlerden sonra en küçük Wilks' Λ değerine sahip olan değişken modelden çıkarılır. Wilks' Λ değeri gruplar arası ayırma ve grup içi homojenliğe dayanmaktadır (Sharma, 1996).

Wilks' Λ değeri değişkenin en iyi ayırım yapacağını gösterse de adımsal seçme tekniği değişkenler arasında çoklu doğrusal bağlantı var ise bu değişkenlerden birini modele almaz. Çünkü iki değişkenin de modelde olması gereksizdir (Sharma, 1996).

3.2.2 Yeni Gözlemlerin Sınıflandırılması

Fisher, X veri matrisinde, A toplumundan genel örneklerin gözlemlerinden hesaplanan Y'lerin ortalamasını μ_{1Y} ve B toplumundan gelen örneklerin gözlemlerinden hesaplanan Y'lerin ortalamasını da μ_{2Y} olarak nitelemiş ve μ_{1Y} ile μ_{2Y} arasındaki karesel uzaklığı

maksimum yapacak bir karesel bileşen bulmayı amaçlamıştır (Özdamar, 2002b).

Her iki toplum aynı kovaryans matrisine sahip toplumlar olmak üzere;

$\mu_1 = E(X/A)$: 1. toplum çok değişkenli gözlemlerin beklenen değeri

$\mu_2 = E(X/B)$: 2. toplum çok değişkenli gözlemlerin beklenen değeri

olarak bulunur.

Karesel dönüştürme;

$$Y = (\overline{x}_1 - \overline{x}_2)' Ax \tag{3.17}$$

biçiminde hesaplanır.

Y değerlerinin ortalamaları;

$$\mu_{1Y} = E(Y/A) = E(b'X/A) = b'\mu_1 \tag{3.18}$$

$$\mu_{2Y} = E(Y/B) = E(b'X/B) = b'\mu_2 \tag{3.19}$$

biçiminde hesaplanır.

Karesel diskriminant fonksiyonu, *A ve* B toplumlarından alınan çok değişkenli gözlemleri ortak toplum varyansına göre birbirlerinden olabildiğince farklı olan tek değişkenli ortalamalar biçiminde gösterecek bir forma dönüştürür.

İki toplumda Y dönüştürülmüş değerler ortalamaları arasındaki orta nokta;

$$k = \left(\frac{1}{2}\right)\ln\left(\frac{|S_1|}{|S_2|}\right) + \frac{1}{2}\left(\overline{x}_1' S_1^{-1} \overline{x}_1 + \overline{x}_2' S_2^{-1} \overline{x}_2\right) \tag{3.20}$$

biçiminde hesaplanır.

k değeri yeni gözlemin hangi sınıfta yer alabileceğini belirlemeye yarayan bir kriterdir.

x_0 gözlem vektörüne sahip bir yeni gözlemin 1. ya da 2. toplumdan hangisine ait bir birim olduğuna karar vermek için önce y_0 gibi tek değişkenli bir değerle ifade edilmesi gerekir.

$$y_0 = -\frac{1}{2} x_0' A x_0 + (x_1' S_1^{-1} - x_2' S_2^{-1}) x_0 - k \qquad (3.21)$$

y_0 değerinin hangi topluma ait olduğu aşağıdaki gibi belirlenir.

Eğer $y_0 - k \geq 0$ ise x_0 1. topluma ait bir gözlemdir.

Eğer $y_0 - k \leq 0$ ise x_0 2. topluma ait bir gözlemdir.

Genelde toplumlara ilişkin kovaryans matrisleri $\sum$ ve ortalama vektörleri μ_i bilinmez. Bu durumda karesel diskriminant analizi veri matrisinden elde edilen ortalama vektörlerine ve kovaryans matrisine göre yürütülür.

3.3 Lojistik Regresyon

Lojistik regresyon; cevap değişkenin kategorik, ikili, üçlü ve çoklu kategorilerde gözlendiği durumlarda açıklayıcı değişkenlerle-sonuç ilişkisini belirlemede yararlanılan bir teknikdir. Lojistik regresyon ikili (binary) değişkenlerin sürekli değişkenlerle bağımlılığını tanımlamak için ilk defa Cox tarafından geliştirilmiştir (Andersen, 1994).

Bağımsız değişken kategorik olduğunda değişken ortalaması durumun hangi kategoriye daha yakın olduğunu gösteren olasılık fonksiyonudur (Menard, 1995). Değişkenleri 0 ve 1 olarak kodlama değişken ortalaması değişkenin kategorilerinden hangisine ait olacağı olasılığını göstermektedir ve y'nin tahmin değeri iki kategoriden hangisine ait olduğunun tahmin olasılığıdır (predicted probability) (Menard, 1995).

Basit ve çoklu regresyon analizleri bağımlı değişken ile açıklayıcı değişken yada değişkenler arasındaki matematiksel bağıntıyı analiz etmekte kullanılmaktadır. Bu tekniklerin uygulanabileceği veri setlerinde bağımlı değişkenin normal dağılım göstermesi, bağımsız değişkenlerin normal dağılım gösteren toplum yada toplumlardan çekilmiş olması ve hata varyansının $\varepsilon \cong N(0, \sigma^2)$ parametreli normal dağılım göstermesi gerekmektedir. Bu ve benzeri koşulların yerine getirilemediği veri setlerine basit yada çoklu regresyon analizi uygulanamaz (Özdamar, 2002a).

Bağımlı değişken ikili gözlemler içeriyorsa normal olmayan hata ve sabit olmayan varyans gibi problemlere sahibtir (Kunter vd., 1996).

Sınıflama ve regresyon modelleri arasındaki temel fark tahmin edilen bağımlı değişkenin

kategorik ya da süreklilik gösteren bir değere sahip olmasıdır. Ancak çok terimli lojistik regresyon (Multinomial Logistic Regression) gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirlerine yaklaşmakta ve bunun sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır (Akpınar, 2000).

Lojistik regresyon analizi, temelde bir regresyon çözümlemesi olmakla birlikte bir ayırıcı çözümleme tekniği olma özelliği göstermektedir (Akkuş vd., 2005). Normal dağılım varsayımı, süreklilik varsayımı ön koşulu yoktur. Varsayım kısıtlaması olmaması, bu tekniğe olan ilgiyi arttırmaktadır. Bununla birlikte değişkenler arasında çok değişkenli normal dağılım mevcutsa çözüm daha uygun olacaktır. Ayrıca diğer regresyon tekniklerinde olduğu gibi değişkenler arasında çoklu doğrusal bağıntı olması yanlı tahmincilere ve standart hatada artışa neden olabilmektedir. Bundan dolayı tahmin edici değişkenler kategorik değişkenler olduğu takdirde süreç daha etkili olacaktır (S.P.S.S., 2001).

Bağımlı değişken üzerinde açıklayıcı değişkenlerin etkileri olasılık olarak elde edilerek risk faktörlerinin olasılık olarak belirlenmesi sağlanır.

Diskriminant analizi de verilerin belirli olasılıklara göre belirli sınıflara atanmasını sağlar ve çok değişkenli normal dağılım varsayımını kabul etmektedir. Lojistik regresyonun böyle bir varsayım gerektirmemesi ayırıcı bir özelliğidir.

Lojistik regresyon, lojistik modellere göre parametre tahminleri yapar. Bağımlı değişkenin tahmini değerlerini olasılık olarak hesaplayarak, olasılık kurallarına uygun sınıflama yapma imkanı veren bir istatistiksel tekniktir (Özdamar, 2002a).

$$P_i = E(Y = 1|X_i) = \beta_1 + \beta_2 X_i \tag{3.22}$$

$$P_i = \frac{1}{1+e^{-Z_i}} \tag{3.23}$$

Burada $Z = \beta_1 + \beta_2 X_i$ 'dir. Bu fonksiyon lojistik dağılım fonksiyonu olarak adlandırılır. P_i 0 ile 1 arasında değerler alır ve Z_i ile ve dolayısıyla X_i ile ilişkisi doğrusal değildir. P_i yalnız X ile değil, β 'larla da doğrusal ilişki içerisinde değildir. Ama bu görüntüsel bir durumdur (Gujarati, 1995). Model doğrusallaştırılabilir.

Şöyle ki;

$$1-P_i=\frac{1}{1+e^{Z_i}} \tag{3.24}$$

$$\frac{P_i}{1-P_i}=\frac{1+e^{Z_i}}{1+e^{-Z_i}}=e^{Z_i} \tag{3.25}$$

Bu durumda $P_i/1-P_i$ bir olayın gerçekleşmesinin bahis (odds) oranıdır ve doğal logaritması alındığında;

$$L_i=\ln\left(\frac{P_i}{1-P_i}\right)=Z_i \tag{3.26}$$

$$=\beta_1+\beta_2 X_i$$

Yani bahis oranı logaritması L, yalnız X'e göre değil, ana kütle katsayılarına göre de doğrusaldır. L'ye logit, bu modellere de logit model denir.

Logit modeli şu özelliklere sahiptir.

- P, 0'dan 1'e giderken (Yani Z_i , -∞'dan +∞'a doğru değişirken), logit L de -∞'dan +∞'a doğru değişir. Olasılıklar 0 ile 1 arasında yer alırken, logitler için sınırlama söz konusu değildir.
- L, X'e göre doğrusal olmakla birlikte olasılıkların kendileri böyle değildir.
- Yorumu şöyle yapılır; β_2, eğim, X'teki bir birim değişmeye karşılık L'deki değişmeyi ölçer. Sabit terim β_1 ise, X 0 olursa bağımlı değişkenin log-bahis oranıdır. Regresyonda sabit terimlerde her zaman görüldüğü gibi bu da fiziksel bir anlam taşımayabilir.
- Logit modeli log-bahis oranının X ile doğrusal ilişki içinde olduğunu varsayar.
- Belli bir X^* veriyken, bağımlı değişkenin bahis oranı değil de kendi olasılığı tahmin edilmek istenirse, β_1 ile β_2 tahminleri bir kez elde edildikten sonra doğrudan bulunabilir. β_1 ile β_2 nin nasıl tahmin edileceği ise aşağıda anlatılmaktadır.

3.3.1 Logit Modelinin Tahmin Edilmesi

Tahmin modeli;

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i + u_i \qquad (3.27)$$

Ana kütle katsayılarını tahmin edebilmek için en yüksek olabilirlik (maximum likelihood) tekniğine başvurulabilir. X_i düzeyindeki N_i gözlemden n_i tanesinde olayın gerçekleştiği varsayılırsa ($n_i \leq N_i$);

$$\hat{P}_i = \frac{n_i}{N_i} \qquad (3.28)$$

Hesaplandığında, her X_i düzeyine karşılık gelen P_i tahmininde bu kullanılabilir. Eğer N_i hayli büyükse, $\hat{P}_i$, P_i'nin iyi bir tahmini olacaktır. Tahmin edilmiş P_i ile tahmin edilmiş logit modeli;

$$\hat{L}_i = \ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad (3.29)$$

Bu da, her X_i'deki gözlem sayısı N_i hayli büyükse, gerçek logit L_i'nin iyi bir tahmincisi olacaktır.

Bilinmeyen P_i'yi $\hat{P}_i$ ile değiştirip σ^2'nin bir tahmin edicisi olan şu ifade kullanılabilir;

$$\sigma^2 = \frac{1}{N_i \hat{P}_i\left(1-\hat{P}_i\right)} \qquad (3.30)$$

3.3.2 Adımsal Süreç (Stepwise Procedure)

Adımsal süreç ileri doğru seçme (forward selection) ve geriye doğru eleme (backward elimination) olmak üzere ikiye ayrılmaktadır. Bu çalışmada lojistik regresyon uygulamasında geriye doğru eleme tekniği kullanıldığı için kısaca teknik hakkında bilgi verilecektir.

Geriye doğru eleme tekniği kompleks modellerle birlikte kullanılmaya başlanmıştır. Her adımda, modelden çıkarılması modelde daha az bozulmaya neden olan değişkenler seçilir (en geniş p değeri). Yeni bir değişkenin silinmesi uyum iyiliğinde zayıf bir etki yarattığında süreç sona erer. Çoğu istatistikçi geriye doğru elemeyi ileri doğru seçme tekniğine tercih etmektedirler. Kompleks bir modelden bir değişken silmek, basit bir modele değişken

eklemekten daha kolaydır (Agresti, 2001).

3.4 Kümeleme Analizi

Kümeleme analizi, doğal gruplamaları veya kümeleri kesin olarak bilinmeyen birim ve değişkenlerin sınıflandırılması için kullanılan teknikler topluluğudur. Birimleri ve değişkenleri birbirlerine benzerliklerine ve farklılıklarına göre uygun kümelere ayırmaktadır. Bu ayırma işlemi sırasında da yakınlık yada uzaklık ölçülerinden (similarity yada dissimilarity measures) faydalanır. Kümeleme analizi "sınıflandırma analizi", "kümeleme çözümlemesi" yada "sayısal taksonomi" olarak ta adlandırılır. Bu kümelerde;

- Her bir grup veya küme belirli bir özelliğe göre homojendir. Yani, her bir gruptaki gözlemler bir diğerine benzerdir.
- Her bir grup aynı özelliklere göre diğer gruplardan farklı olmalıdır. Yani, bir grubun gözlemleri diğer grupların gözlemlerinden farklı olmalıdır (Aytaç ve Bayram, 1999).

Çok değişkenli istatistiksel analiz, çok sayıda değişken arasındaki ilişkileri ölçme ve açıklamada kullanılan teknikler topluluğunu ifade eder ve bu analiz ile ilgili geliştirilmiş teknikler, bağımlılık analizinde kullanılan teknikler ve karşılıklı bağımlılık analizinde kullanılan teknikler olmak üzere iki grupta toplanabilir.

Bağımlılık analizinde, bir değişken veya değişken grubu, diğer değişkenler tarafından tahmin edilmekte veya açıklanmaktadır. Burada, bir değişken diğerlerine bağlı olup onlarla tahmin edilmektedir. Karşılıklı bağımlılık analizinde ise değişkenler arasındaki bağımlılık yerine karşılıklı bağıntılar söz konusudur. Bu tip analizin en iyi örneklerinden biri Kümeleme Analizi'dir.

Kümeleme analizi, kümelerin sayısına veya küme yapılarına ilişkin herhangi bir varsayımda bulunmaz. Diğer çok değişkenli istatistiksel analiz tekniklerinde önemli bir yer tutan normallik, doğrusallık ve homojenlik varsayımları bu teknikte prensipte kalmakta ve uzaklık değerlerinin normalliği yeterli görülmektedir (Çelik vd., 2005). Ayrıca kümeleme analizinde kovaryans matrisine ilişkin herhangi bir varsayım bulunmamaktadır (Tatlıdil, 1996). Kümeleme analizi, istatistiksel anlamda, birbirlerinden farklılık gösteren gruplar yaratır. Gruplara daha sonra katılacakların, hangi kıstaslara göre sınıflandırılacağına dair bir bilgi içermez.

Kümeleme analizi şu amaçlara yönelik kullanılmaktadır.

- Dağınık halde bulunan *n* sayıda birimi kendi içinde homojen ve diğerleriyle heterojen alt ana kütlelere ayırarak işlenebilir hale getirmek.
- *p* sayıda değişkeni birimlerde gözlemlenen değerler baz alınarak faktör yapılarını açıklayan alt ana kütlelere ayırmak.
- Hem birimleri hem de değişkenleri birlikte ele alarak ortak *n* birimi *p* değişkene göre ortak özellikli alt kümelere ayırmak.

Kümeleme analizi, genel amacının yanı sıra model uydurmanın kolaylaştırılması, gruplar için ön tahmin, hipotezlerin testi, veri yapısının netleştirilmesi, indirgenmesi ve aykırı değerlerin bulunmasında da kullanılır (Akpınar ve Gürcan, 2002).

Kümeleme analizinin uygulanması sırasında izlenecek adımlar:

- Problemin formüle edilmesi
- Uygun kümeleme tekniği (algoritma) belirlenmesi
- Veri matrisinin belirlenmesi
- Mesafe ölçüsünün belirlenmesi (Benzerlik ya da farklılık matrisinin belirlenmesi)
- Kümeleme Süreçünün belirlenmesi
- Küme sayısına karar verilmesi
- Kümelerin yorumlanması

Kümeleme analizi ile değişkenler ve kümeler arasında ilişki kurulur. Kümeleme işleminin sonucunda elde edilen kümeler gözlenen değişkenler üzerine oturtulmuş anlamlı hipotezlerdir. Kümeler belirlenmiş çalışma alanında önemli etkileri ve objektif karşılıklı etkileşimleri kapsamaktadır. Bundan dolayı analiz, önceden belirlenmiş çalışma alanıyla sınırlıdır. Çalışma alanını açıklayan hipotezler, kümeler aracılığıyla açıklanmaktadır.

Bu yüzden kümeleme analizi ile yeni tezler geliştirirken değişkenlerin seçiminde son derece dikkatli davranılmalıdır (Koç, 1997).

Kümeleme analizi birbirleri ile ilgili çok sayıda değişkenin kullanılmasına uygun bir analiz tekniğidir.

Kümeleme analizinde veriler kümelemeye uygun şekilde girildikten sonra uzaklık ölçülerinden yararlanılarak uzaklıklar matrisi elde edilir. Kümeleme analizi çok sayıda değişik

işlevi yerine getirmektedir. Bu nedenle farklı amaçlar için farklı teknikler (prosedürler) kullanılır. Ayrıca değişkenlerin ölçü birimlerinin ve ölçümleme tekniklerinin farklı olmasından dolayı birimlerin benzerliklerinin ortaya konmasında da değişik ölçüler kullanılır. Eğer analize girecek veriler aralıklı veya oransal ölçek düzeyinde ölçülmüş ise, en çok kullanılan uzaklık ölçüleri; Korelasyon Uzaklığı, Öklit, Kare Öklit, Minkowski ve Manhattan City- Blok'dur. Eğer veriler sınıflayıcı veya sıralayıcı ölçek düzeyinde ölçülmüş ise kullanılan uzaklık ölçüleri; Ki-kare ve Normalleştirilmiş Ki-Kare olarak bilinen Phi-Kare'dir (Aytaç ve Bayram, 1999). Eğer ikili (binary) gözlemlere göre ölçümler yapılmış ise birimler arasındaki benzerlikleri belirlemede Öklit, Kare Öklit, Size Difference, Pattern Difference, Lance and Williams Difference, Shape Difference gibi benzerlik yada farklılık ölçülerinden yararlanılmaktadır (Özdamar, 2002b).

Kümeleme teknikleri izledikleri yaklaşımlara göre hiyerarşik kümeleme tekniği ve hiyerarşik olmayan kümeleme tekniği olmak üzere iki ana gruba ayrılırlar.

Hiyerarşik kümeleme tekniklerinde kümeler ardı ardına birleştirilir ve bir grup diğeri ile bir kez birleştirildikten sonra, devam eden adımlarda bir daha ayrılmaz. Bu teknikler ele alınan değişkenler için hiyerarşik bir yapı oluştururlar. Hiyerarşik kümeleme tekniklerinde küme sayısına görsel olarak karar verilir. Bu durumda genellikle dendogram olarak bilinen ağaç diyagramı kullanılır (Aytaç ve Bayram, 1999).

Hiyerarşik kümeleme teknikleri kendi içinde ikiye ayrılır;

- Birleştirici aşamalı kümeleme teknikleri (Agglomerative Hierarchical Clustering Procedures)
- Ayırıcı aşamalı kümeleme teknikleri (Divisive Hierarchical Clustering Procedures)

Birleştirici aşamalı kümeleme teknikleri: Başlangıçta tüm birimlerin ayrı birer küme oluşturduğunu kabul ederek, n birimi aşamalı olarak sırasıyla n, n-1, n-2, ... , ...3, 2, 1 kümeye yerleştirmeyi amaçlayan bir yaklaşımdır.

Ayırıcı aşamalı kümeleme teknikleri: Başlangıçta tüm birimlerin bir küme oluşturduğunu kabul ederek birimleri aşamalı olarak n birimi sırasıyla 1, 2, 3, ... , n-r, n-3, n-2, n-1, n kümeye ayırmayı amaçlayan bir yaklaşımdır (Özdamar, 2002b).

Birleştirici aşamalı kümeleme tekniklerinden sıklıkla kullanılan ve genel kabul görmüş olanları aşağıdaki gibi sayılabilir.

- Tek Bağlantı Kümeleme Tekniği (TEKBY, SINGLE LINKAGE, NEAREST NEIGHBOUR METHOD)
- Ortalama Bağlantı Kümeleme Tekniği (ORTBKY, AVERAGE LINKAGE METHOD)
- Tam Bağlantı Kümeleme Tekniği (TAMBKY, COMPLETE LINKAGE METHOD)
- McQUITTY Bağlantı Kümeleme Tekniği (MCQUITTY LINKAGE METHOD)
- Küresel Ortalama Bağlantı Kümeleme Tekniği (KOBKY, CENTROID LINKAGE METHOD)
- Ortanca Bağlantı Kümeleme Tekniği (OBKY, MEDIAN LINKAGE METHOD)
- Ward Bağlantı Kümeleme Tekniği (WBKY, WARD LINKAGE METHOD)

Tekniklerin kriterleri şu şekildedir;

Tek Bağlantı Kümeleme Tekniği:

$$d_{mj} = \min\left(d_{kj}, d_{lj}\right) \tag{3.31}$$

Ortalama Bağlantı Kümeleme Tekniği:

$$d_{mj} = \left(N_k d_{kj} + N_l d_{lj}\right) / N_m \tag{3.32}$$

Tam Bağlantı Kümeleme Tekniği:

$$d_{mj} = \max\left(d_{kj}, d_{lj}\right) \tag{3.33}$$

McQUITTY Bağlantı Kümeleme Tekniği:

$$d_{mj} = \left(d_{kj} + d_{lj}\right) / 2 \tag{3.34}$$

Küresel Ortalama Bağlantı Kümeleme Tekniği:

$$d_{mj} = \left(N_k d_{kj} + N_l d_{lj}\right) / N_m - N_k N_l d_k l / N_M^2 \tag{3.35}$$

Ortanca Bağlantı Kümeleme Tekniği:

$$d_{mj} = \left(d_{kj} + d_{lj}\right)/2 - d_{kl}/4 \tag{3.36}$$

Ward Bağlantı Kümeleme Tekniği:

$$d_{mj} = \left(\left(N_j + N_k\right)d_{kj} + \left(N_j + N_l\right)d_{lj} - N_j d_{kl}\right)/\left(N_j + N_m\right) \tag{3.37}$$

Kümeleme Analizi'nde, küme sayısının belirlenmesi konusunda son yıllarda yoğun çalışmalar yapılmaktadır. Halen küme sayısının belirlenmesinde kullanılan en pratik yol;

$$k = \sqrt{\frac{n}{2}} \tag{3.38}$$

biçiminde belirtilmektedir (Doğan, 2002).

Veri matrisinde yer alan n birimin p değişkene göre uzaklıkları, uzaklık matrisi adı verilen D matrisi ile gösterilir. D matrisinin elemanları d_{ij} yada $d(i,j)$ biçiminde gösterilir.

D uzaklık matrisinin uzaklık fonksiyonu olabilmesi için X, Y ve Z öklit uzayında üç nokta olmak üzere;

1. Negatif olmama, $D(X,Y) \geq 0$;
2. Simetri, $D(X,Y) = D(Y,X)$;
3. Teşhis işareti, $D(X,X) = 0$;
4. Kesinlik, $D(X,Y) = 0$ ancak ve ancak $X = Y$;
5. Üçgen eşitsizliği, $D(X,Y) \leq D(X,Z) + D(Z,Y)$

Koşulları geçerli ise D, bir uzaklık olarak adlandırılır. Birimlerin birbirleriyle olan benzerlik düzeyleri benzerlik matrisi S ile gösterilir. Benzerlik matrisi elemanları D matrisi elemanlarından elde edilir (Doğan İ., 2002).

Kümelerin birim yada değişkenlerin benzerlik veya farklılıkları dikkate alınarak belirlendiği daha önce belirtilmişti. Bu benzerlik ve farklılıkların ölçümünde kullanılan bazı uzaklık ölçüleri;

- Minkowski Uzaklığı;

$$d(x,y)=\left[\sum_{i=1}^{p}\left|x_i-y_i\right|^m\right]^{1/m} \tag{3.39}$$

- Öklit (Euclide) Uzaklığı;

$$d(i,j)=\sqrt{\sum_{k=1}^{p}\left(X_{ik}-X_{jk}\right)^2} \tag{3.40}$$

- Pearson Uzaklığı;

$$d_p(i,j)=\sqrt{\sum_{k=1}^{p}\left(X_{ik}-X_{jk}\right)^2/S_k^2}\quad i,j=1,2...,n;k=1,2...,p \tag{3.41}$$

- Manhattan (City-Block) Uzaklığı;

$$d_M(i,j)=\sum_{k=1}^{p}\left(\left|X_{ik}-X_{jk}\right|\right)\ldots i,j=1,2...,n;k=1,2...,p \tag{3.42}$$

Kümeleme analizinde veri girişinden sonra, hesaplanan uzaklık değerlerinden yararlanarak birey yada nesnelerin kümelere atanması işlemi yapılmaktadır.

3.4.1 En Yakın Komşu Yaklaşımı (Nearest Neighbour Approach)

En yakın komşu tekniği, ilk defa Fix ve Hudges tarafından geliştirilen sınıflandırma probleminde standart, parametrik olmayan bir yaklaşımdır. Bu teknik KS'de ilk defa Chatterjee ve Barcun ve daha sonra Henley ve Hand tarafından uygulanmıştır. Bu tekniğin dayandığı mantık, herhangi iki başvurunun birbirinden ne kadar uzak olduğunu ölçmek için başvuru veri uzayında bir mesafeyi seçmeye dayanmaktadır. Geçmiş başvuruların bir örnekleme temsili standart alınır, ve yeni bir başvuru temsili örneklemede k kadar yakın başvurular arasındaki (yeni başvurunun en yakın komşuları) iyi-kötü oranına dayanarak iyi veya kötü olarak sınıflanır (Thomas vd., 2002).

En yakın komşu yaklaşımı eşit iyi-kötü oranı olan örneklemelerde optimum performans gösterir (http://www.fractalanalytic.com).

Bu yaklaşımı uygulamak için üç tane parametreye ihtiyaç vardır: mesafe, en yakın komşular setini oluşturan kaç tane başvuru sayısı olduğu (k), ve bir başvurunun iyi olarak sınıflanabilmesi için başvuru iyi oranının ne olması gerektiği. Normal olarak, eğer komşuların çoğunluğu iyi ise, başvuru iyi olarak sınıflandırılır, aksi taktirde başvuru kötü olarak

sınıflanır. Ortalama varolan maliyet M, ve iyiyi reddetmenin ortalama kayıp karı K olarak tanımlansın. Eğer en yakın komşuların en azından $M/M+K$ tanesi iyi ise, yeni bir başvuru iyi olarak sınıflandırılır. Eğer yeni bir başvurunun iyi olma olasılığı iyi olan komşuların oranı ise, bu kriter beklenen kaybı minimize edecektir.

Mesafenin seçimi oldukça önemlidir. Fukanaga ve Flick genel bir mesafe tanımı yapmıştır:

$$d(x_1,x_2)=(x_1-x_2)A(x_1)\left((x_1-x_2)^T\right)^{1/2} \tag{3.43}$$

$A(x)$, $p\times p$ simetrik sonlu pozitif bir matristir. Eğer x'e bağlı ise, $A(x)$ yerel mesafe olarak adlandırılır, eğer x'ten bağımsız ise global (genel) mesafe olarak adlandırılır. Yerel mesafenin eksikliği, genelde uygun olmayan deneme setinin özelliklerini dikkate alır. Bu nedenle bir çok araştırmacı global mesafeye odaklanır. KS'de en yakın komşu yaklaşımının en detaylı uygulaması Henley ve Hand tarafından yapılmıştır. Bu teknik, öklit uzunluğu ve iyi ile kötüyü en iyi ayıran yönün uzunluğu karışımına odaklanılır. Eğer w ; p boyutlu yön vektörü ise, Henley ve Hand'in mesafe ifadesi şöyledir (Thomas vd., 2002);

$$d(x_1,x_2)=\left\{(x_1-x_2)^T\left(I+D_w.w^T\right)(x_1-x_2)\right\}^{1/2} \tag{3.44}$$

KS'de lineer ve lojistik regresyon yaklaşımları kadar çok sık kullanılmamasına karşın, en yakın komşu tekniği gerçek uygulamalar için bazı önemli özelliklere sahiptir. Dinamik olarak yeni olaylar eklemeyerek deneme setini güncellemek çok kolaydır, ve eklenenin iyi veya kötü olduğu bilindiğinde kolayca o olay örneklemeden çıkarılabilir İlk seferde iyi bir mesafe bulmak, skor kart oluşturmada regresyon tekniğiyle hemen hemen eşdeğer bir netice verir. Böylece çoğu uygulayıcı bu noktada ilerlemeyi durdurup geleneksel bir skor kart kullanmayı tercih ederler. Sınıflandırma ağacı yaklaşımı ile kıyasladığımızda, en yakın komşu yaklaşımı her bir başvuranın özelliği için bir skor üretmez, uygulayıcılar için bir denge noktası belirler ve onların gerçekte sistemin ne yaptığını anlamasını sağlar.

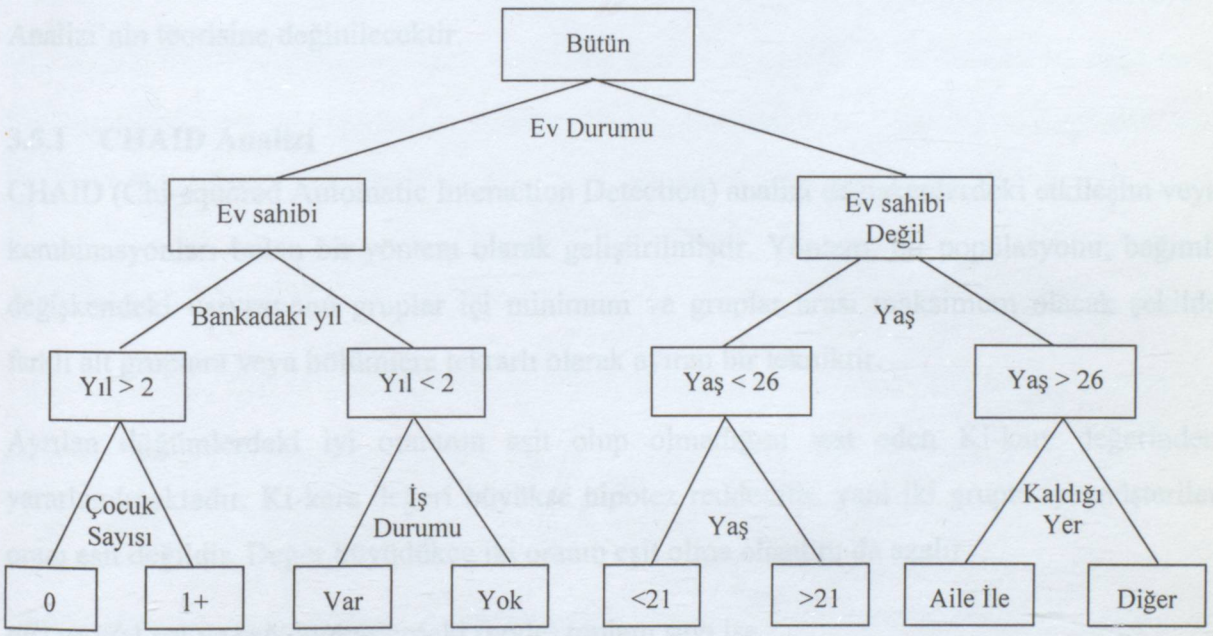
3.5 Sınıflandırma Ağaçları (Classification Tree)

Sınıflandırma Ağaçları sıralı yada sırasız kategorik bağımlı değişkenleri kategorik yada sürekli bağımsız değişkenlerle analiz eden bir nonparametrik tekniktir (Arminger vd., 1997).

Sınıflandırma ve ayırma için tamamıyla farkı bir yaklaşım, Sınıflandırma Ağaçları yada Tekrarlanan Kısımlara Ayırma Algoritmasıdır (Recursive Partitioning Algorithms). Bu

tekniğin ana fikri, başvuru cevaplarını farklı gruplara ayırmak ve her bir grubu grup içindeki çoğunluğa dayanarak iyi veya kötü olarak ayırmaktır. Genel sınıflandırma problemleri için bu teknik Breiman ve Friedman tarafından geliştirilmiştir. Takip eden çalışmalar bu tekniği şu anki durumuna getirmiştir (Thomas vd., 2002).

Başvuru veri seti A, başvuran kişilerin örneklemesini içeren iki alt kümeye bölünür. Bu iki alt kümedeki başvuranların özellikleri, orijinal setin kabul edilmiş riskinin homojenliğinden oldukça uzaktır. Bu iki setin her biri, daha homojen alt kümeler elde etmek için tekrar ikiye ayrılır, ve işlem bu şekilde tekrar ederek devam eder. Tekrarlanan Kısımlara Ayırma Algoritması diye adlandırılmasının nedeni budur. Bu işlem, ağacın en son düğümünde, alt kümeler belirlenmiş gereksinimleri tatmin ettiğinde durur. Her bir uç (terminal) düğüm A_G ve A_B'nin bir üyesi olarak sınıflandırılır. Bütün bu teknik, Şekil 3.1'deki ağaçta gösterilmektedir.



Şekil 3.1 Sınıflandırma Ağacı (Thomas vd., 2002)

Sınıflandırma ağacındaki karar süreçü şöyledir:

- Setleri ikiye ayırma kuralı: Ayırma kuralı
- Bir setin terminal düğüm kararını vermek: Durdurma kuralı
- Terminal düğümleri iyi ve kötü kategorilere nasıl ayrılacak?

İyi-kötü atama kararı en kolay olanıdır. Normalde, düğümdeki örnek olayların çoğunluğu iyi ise bu düğüm iyi olarak adlandırılır. Diğer bir alternatif ise yanlış sınıflama maliyetlerini minimize etmektir. Eğer M kötünün iyi olarak sınıflanma maliyetini, K de iyinin kötü olarak sınıflandırıldığındaki kayıp karı gösterirse, iyi olarak sınıflanan düğümdeki iyi kötü oranının örnekleme genelindeki iyi kötü oranını aştığı maliyeti minimize etmek gerekir.

Ayırma kuralında; her bir özellik için en iyi ayırma bulunur ve bu ayrımın ne derecede iyi olduğunun ölçüsü dikkate alınır. Bu ölçüt altında, en iyi ayırıcı özelliğe karar verilir. Herhangi bir sürekli özellik X_i ve, s'nin bütün değerleri için $\{x_i < s\}$ ve $\{x_i \geq s\}$ ayırıcıları kontrol edilir ve s'nin en iyi değeri bulunur. Eğer X_i kategorik bir değişken ise, kategorileri ikiye ayıran olası ayırıcılara bakılır ve farklı ayırıcılar dikkate alınarak ölçü kontrol edilir. Genellikle artan iyi:kötü oranına göre kategoriler sıralanır, ve grupları ikiye ayıran en iyi ayrıcı tespit edilir. Bunun için kullanılan ölçütler Kolmogorov-Smirnow İstatistiği, Temel Safsızlık İndeksi (Basic Impurity Index), Gini İndeksi (Gini Index), Entropi İndeksi (Entropy Index) ve CHAID Analizi'dir. Bu ölçütlerden uygulamada kullanılacak olan CHAID Analizi'nin teorisine değinilecektir.

3.5.1 CHAID Analizi

CHAID (Chi-squared Automatic Interaction Detection) analizi değişkenlerdeki etkileşim veya kombinasyonları bulan bir yöntem olarak geliştirilmiştir. Yöntem, bir popülasyonu; bağımlı değişkendeki varyasyonu gruplar içi minimum ve gruplar arası maksimum olacak şekilde farklı alt gruplara veya bölümlere tekrarlı olarak ayıran bir tekniktir.

Ayrılan düğümlerdeki iyi oranının aşit olup olmadığını test eden Ki-kare değerinden yararlanılmaktadır. Ki-kare değeri büyükse hipotez reddedilir, yani iki grupta iyi müşteriler oranı eşit değildir. Değer büyüdükçe iki oranın eşit olma olasılığı da azalır.

n(l) ve n(r) sol ve sağ düğümlerdeki (node) toplam sayı ise

$$Chi = n(l)n(r) - \frac{(p(G/l) - p(G/r))^2}{n(l) + n(r)} \tag{3.45}$$

CHAID modeli tüm sürekli ve kategorik değişkenlerle çalışır. Sürekli değişkenler kategorik hale getirilerek model kurulabilir.

Algoritması;

- Her bir X bağımsız değişkeni için en az anlamlı farklılığa sahip olan iki çift bulunur (en

büyük p değerine sahip olan).

- En büyük p değerine sahip olan X değişkeni kategori çiftleri için p değeri belirlenen α değeri ile kıyaslanır. p değeri α değerinden büyük ise bu çift tek kategoride toplanır. Böylelikle X'in yeni kategori seti belirlenmiş olur ve birinci adıma geçilerek işleme devam edilir. p değeri α değerinden küçük ise de 3. adımdan devam edilir.
- X ve Y kategori setleri için ayarlanmış p değerleri hesaplanır.
- En küçük ayarlanmış p değerine sahip X değişkeni (en anlamlı) seçilir. Ayırma için belirlenen α değeriyle kıyaslandığında p değeri eşit ya da küçük ise X değişkeni kategori setlerine bağlı olarak düğüm ayrılır. P değeri ayırma için belirlenen α değerinden büyükse düğüm ayrılmaz. Düğüm bir terminal noktası olarak alınır.
- Durdurma kuralı gerçeklenene kadar da süreç devam eder.

3.5.2 Kolmogorov-Smirnov İstatistiği

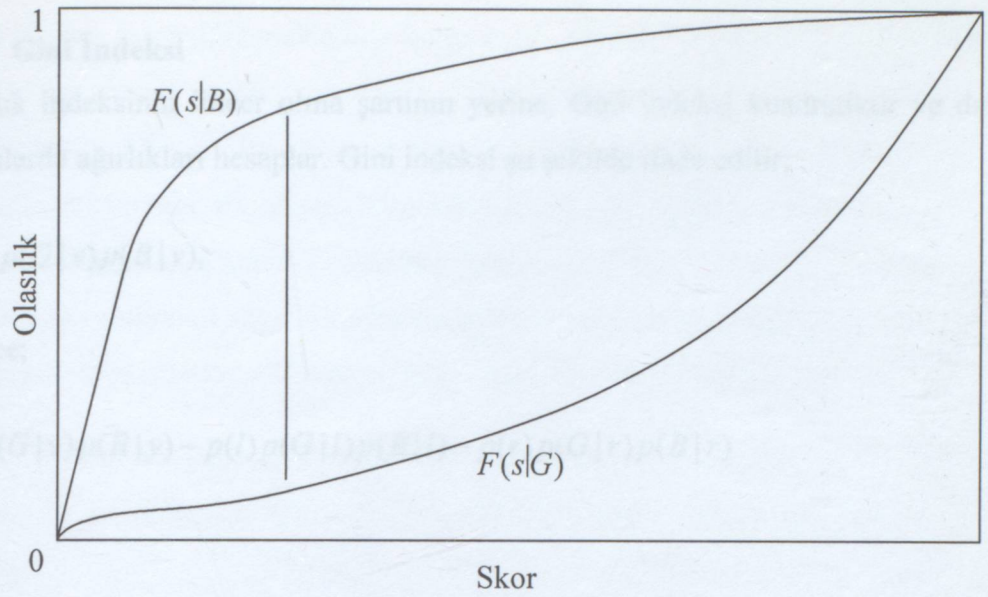
Sürekli bir X_i özelliği için, $F(s|G)$ iyiler için X_i'nin kümülatif dağılım fonksiyonu olarak ve $F(s|\,\mathrm{B})$ kötüler için X_i'nin kümülatif dağılım fonksiyonu olarak alınsın. Kötülerin iyilere göre X_i'nin düşük değerlerine daha büyük bir eğilim gösterdiği kabul edilsin ve daha önce tanımlanan M ve K maliyetleri de dikkate alınsın. Bu durumda s'nin değerini bölmek için kullanılacak olan miyopik kural aşağıdaki formülasyonun minimize edilmesini içerir:

$$KF(s|G)p_G + M(1-F(s|B))p_B \tag{3.46}$$

Eğer $Lp_G = Dp_B$ ise, bu iki dağılım arasındaki Kolmogorov-Smirnov mesafesini seçmek ile aynıdır. Şekil 3.2'de gösterildiği gibi, $F(s|G)-F(s|B)$ yi minimize etmek veya $F(s|\,\mathrm{B})-F(s|G)$ yi maksimize etmek şeklinde olabilir.

Eğer iki alt grup sol (l) ve sağ (r) diye kategorik olarak adlandırılırsa, bu durumda $p(l|B)$ (sol gruptaki kötülerin olması olasılığı–sürekli olaylarda $F(s|\,\mathrm{B})$) ve $p(l|G)$ (sol gruptaki iyilerin olması olasılığı–sürekli olaylarda $F(s|G)$) arasındaki fark maksimize edilebilir. Bayes kuralını da kullanarak sürekli ve kategorik değişkenler için Kolmogorov-Smirnov İstatistiği aşağıdaki ifadenin maksimizasyonu şeklinde yazılabilir:

$$\mathrm{KS}=|p(l|B)-p(l|G)|=\left|\frac{p(B|l)}{p(B)}-\frac{p(G|l)}{p(G)}\right|.p(l) \tag{3.47}$$



Şekil 3.2 Kolmogorov-smirnov mesafesi (Thomas vd., 2002)

3.5.3 Basit Saf Olmama İndeksi (Basic Impurity Index)

Bütün sınıfın safsızlık indeksi, ağaçtaki her bir düğüm *v*'nin ne kadar safsızlık ölçütüne sahip olduğunu belirmeye çalışır. Eğer bir düğüm sol düğüm (*l*) ve sağ düğüm (*r*) ve olasılıkları sırasıyla *p(l)* ve *p(r)* olarak ikiye ayrılırsa, buradaki safsızlık aşağıdaki gibi ölçülür:

$$l = i(v) - p(l)i(l) - p(r)i(r) \tag{3.48}$$

Ne kadar fazla fark varsa, safsızlıktaki değişim de o kadar fazladır. Bu da yeni düğümlerin daha fazla saf olduğu anlamına gelmektedir. Bu ifadenin maksimum olduğu değer bizim ayırma için istediğimiz kriterdir. Bu aynı zamanda $p(l)i(l) + p(r)i(r)$ ifadesinin minimize edilmesine eşdeğerdir. Açıkça görülmektedir ki, eğer pozitif bir fark değeri ortaya çıkmaz ise, düğüm ikiye ayrılmamalıdır.

Bu safsızlık indekslerinin en basiti, düğümdeki $i(v)$ değerinin en küçük gruba oranını dikkate almaktır. Şöyle ki,

$$i(v) = p(G|v) \quad eğer\ p(G|v) \le 0.5, \tag{3.49}$$

$$i(v) = p(B|v) \quad eğer\ p(B|v) < 0.5 \tag{3.50}$$

3.5.4 Gini İndeksi

Safsızlık indeksinin lineer olma şartının yerine, Gini indeksi kuadratiktir ve daha saf olan düğümlerde ağırlıkları hesaplar. Gini indeksi şu şekilde ifade edilir;

$$i(v) = p(G \mid v)p(B \mid v), \tag{3.51}$$

Böylece;

$$G = p(G \mid v)p(B \mid v) - p(l)p(G \mid l)p(B \mid l) - p(r)p(G \mid r)p(B \mid r) \tag{3.52}$$

olur.

3.5.5 Entropi İndeksi

Diğer lineer olmayan indeks entropidir ve şu şekilde hesaplanır;

$$i(v) = -p(G \mid v)\ln(p(G \mid v)) - p(B \mid v)\ln(p(B \mid v)) \tag{3.53}$$

İsminden de anlaşıldığı üzere bu indeks entropi (sistemdeki düzensizlik ve seviye ölçüsü) ile ilgilidir ve düğümdeki iyi ve kötü arasındaki ayrımın bilgi miktarını gösterir. Bu indeks, düğümdeki iyiler ile kötüler arasındaki gerçek ayrımın kaç farklı yol ile yapılabileceğinin ölçüsüdür.

3.6 Yapay Sinir Ağları (Neural Networks)

Sinir ağları aslında, insan beynindeki bilginin işlenmesini ve iletilmesini modelleme denemelerinden geliştirilmiştir. Beyinde, çok sayıda dendrit elektrik sinyallerini nörona (neuron) taşır, nöronlar sinyalleri nabız atışına çevirirler ve aksona (axon) oradan da diğer nöronların dendritlerine (dendrites) bilgiyi ileten çok sayıda uyarım (synapse) gönderirler. İnsan beyni tahmini 10 milyar nörona sahiptir. Beyine benzer bir şekilde, bir sinir ağı çok sayıda girdiden (değişkenlerden) oluşmaktadır. Her bir girdi bir ağırlıklı katsayı ile çarpılmaktadır. Bu beyindeki dendrit yapısına benzemektedir. Ürünler toplanıp bir nöronda dönüştürülür. Elde edilen sonuç diğer nöron için bir girdi değeri olur.

Bağlantısal Mimariler (Connectionist Architectutures), Adaptif Sistemler (Adaptive Systems) veya Paralel Dağıtılmş İşlemciler (Parallel Distributed Processing) olarak da adlandırılan

YSA'lar, oldukça fazla bağlantı içeren ve paralel yapılandırılmış beyin işlevinden esinlenen bir bilgi işlem paradigmasıdır Farklı isimlerle anılmaları, farklılık sağlayan bazı temel özelliklerinden kaynaklanmaktadır. Bağlantısal Mimari (veya Bağlantısal Sistem) olarak anılmalarının temel sebebi, bireysel işlem elemanları (processing nodes) arasındaki bağlantılardır. Ayrıca, bu bağlantıların ağırlıkları değişebildiğinden YSA'lar çalışma sistemlerini daha da etkinleştirebilmektedirler ve bu yüzden Adaptif Sistem olarak da adlandırılmaktadırlar. Paralel Dağıtılmış İşlemciler olarak adlandırılmalarının sebebi ise ağ içinde çok sayıdaki nod veya nöronların hepsinin birbirlerine paralel olarak çalışmalarıdır. Bu yapı, eşanlı bir çözüm üretebilme yeteneği sağlamaktadır (Yurtoğlu, 2005).

YSA'lar, tanımlanmamış girdi veriler hakkında karar verirken genelleme yapabildikleri için iyi birer yapı tanımlayıcısı (pattern recognition engine) ve sağlam sınıflayıcılardırlar (robust classifier).

YSA'ların çok sayıda farklı çeşitleri vardır. Bu farklılıkların kaynağı mimarisi, öğrenme tekniği, bağlantı yapısı vb. olabilmektedir. Genel olarak, YSA'lar üç ana kritere göre sınıflandırılmaktadırlar. Bu kriterlerden biri öğrenme tekniğidir. Temel olarak iki çeşit öğrenme algoritması vardır: yönlendirmeli öğrenme ve yönlendirmesiz öğrenme. Her tekniğin kullandığı öğrenme kuralı değişebilmekteyse de, YSA'lar bu iki algoritmaya göre sınıflandırılırlar.

İkinci bir sınıflandırma, ağın kullandığı veriye göre yapılmaktadır. Temel olarak, kalitatif ve kantitatif olmak üzere iki tür veri vardır. Kalitatif verilerle çalışan ağlar, ister yönlendirmeli ister yönlendirmesiz öğrenme kullansın, sınıflandırma ağları olarak bilinirler. Kantitatif veriler kullanan yönlendirmeli eğitme ise regresyon olarak adlandırılmaktadır.

Son sınıflandırma kriteri ise ağın yapısıdır. Bazı ağlar ileri besleme şeklinde yapılandırılırken, bazı ağlar ise geri besleme yapısı içermektedir. İleri besleme sinir ağlarında, işlem elemanları arasındaki bağlantılar bir döngü oluşturmazlar ve bu ağlar girdi veriye genellikle hızlı bir şekilde karşılık üretirler. Geri beslemeli ağlarda (Recurrent Networks) ise bağlantılar döngü içerirler ve hatta her seferinde yeni veri kullanabilmektedirler. Bu ağlar, döngü sebebiyle girdinin karşılığını yavaş bir şekilde oluştururlar. Bu yüzden, bu tür ağların eğitme süreci daha uzun olmaktadır. Ayrıca, hem ileri besleme hem de geri yayılma olarak tanımlanabilecek ağ yapıları da mevcuttur (Yurtoğlu, 2005).

3.6.1 YSA Avantajları

Teknolojik gelişme olarak da görülmesi gereken yapay sinir ağları tekniği, özellikleri ve yapabildikleri sayesinde önemli avantajlar sunmaktadır.

YSA'ların farklılık ve avantajları;

Doğrusal Olmayan Yapı; YSA'ların en önemli özelliklerinden birisi gerçek hayattaki olası doğrusal olmayan yapıları da dikkate alabilmesidir. Analiz konusunun içerdiği veri setinin doğrusal veya doğrusal olmayan yapı içeriyor olması, analiz sonuçlarını etkileyecek önemli bir faktördür. Bu yüzden, doğrusal olmayan yapıları dikkate alabilmesi YSA'ların önemli bir özelliğidir.

Öğrenme; YSA'ların diğer bir önemli avantajı en önemli özelliğinden kaynaklanmaktadır. Esin kaynağı insan beyninin çalışma sistemi olan bu teknik, eğitme veya başlangıç tecrübesi sayesinde veriyi kullanarak öğrenme yeteneğine sahiptir. Bu özelliği sayesinde ise geleneksel teknikler için çok karmaşık kalan problemlere çözüm sağlayabilmektedirler. Ayrıca, insanların kolayca yapabildiği ama geleneksel metotların uygulanamadığı basit işlemler için de oldukça uygundurlar.

Yerel İşlem ve Esneklik; YSA'lar, geleneksel işlemcilerden farklı şekilde işlem yapmaktadırlar. Geleneksel işlemcilerde, tek bir merkezi işlem elemanı her hareketi sırasıyla gerçekleştirir. YSA modelleri, her biri büyük bir problemin bir parçası ile ilgilenen çok sayıda basit işlem elemanlarından oluşma ve bağlantı ağırlıklarının ayarlanabilmesi gibi özelliklerinden dolayı önemli derecede esnek bir yapıya sahiptirler. Bu esnek yapı sayesinde ağın bir kısmının zarar görmesi modelde sadece performans düşüklüğü yaratır. Modelin işlevini tamamen yitirmesi söz konusu olmaz. Ayrıca, toplam işlem yükünü paylaşan işlem elemanlarının birbirleri arasındaki yoğun bağlantı yapısı sinirsel hesaplamanın temel güç kaynağıdır. Bu yerel işlem yapısı sayesinde, YSA tekniği en karmaşık problemlere bile uygulanabilmekte ve tatminkar çözümler sağlayabilmektedir.

Gerçek Zamanlı İşlem; YSA hesaplamaları paralel olarak yürütülebildiğinden gerçek zamanlı işlem yapabilir.

Genelleme; öğrenme yeteneği sayesinde bilinen örnekleri kullanarak daha önce karşılaşılmamış durumlarda genelleme yapabilmektedir. Yani, hatalı (noisy) veya kayıp veriler için çözüm üretebilmektedir

Hafıza; Bunlara ek olarak, işlem elemanları arasındaki ağırlıklı bağlantılar sayesinde

dağıtılmış hafızada bilgi saklayabildikleri söylenebilir.

Kendi İlişkisini Oluşturma; YSA, bilgilere (verilere) göre kendi ilişkilerini oluştururlar, denklem içermezler.

Sınırsız Sayıda Değişken ve Parametre; YSA modelleri sınırsız sayıda değişken ve parametre ile çalışabilmektedir. Bu sayede mükemmel bir öngörü doğruluğu ile genel çözümler sağlanabilmektedir (Yurtoğlu, 2005).

YSA kullanımının sağladığı avantajlar doğrultusunda sonuçlarının başarısı, bu uygulamayı kullanan şirketlerin rakamsal verileriyle de özetlenmek istenirse; kredi skorlamada YSA kullanan HNC şirketi karlılığını %27 artırmıştır. Bu sistem daha sonra mortgage kredileri için de kullanılmıştır (Stergious ve Siganos, 2005). Finans şirketlerinde YSA kullanan uygulayıcılar birçok faydalarını görmektedirler. American Express ve Security Pasific Bankası hilekarlık tespiti (fraud detection) ve küçük işletmelere verilen krediler için YSA modeli kullanmaktadırlar. Lloyds Bowmaker Motor Finance, otomobil finansmanı için kredi skorlamada YSA kullanmaktadır. Ve YSA uyguladıktan sonra model %10 daha verimli hale gelmiştir (Yegorova, 2001).

3.6.2 Tek Katmanlı Sinir Ağları (Single Layer Neural Networks)

Bir tek katmanlı sinir ağı, yukarıda belirtilen bileşenlerden, yani birden fazla girdiden oluşur. Ancak burada, bir nöronda elde edilen sonucun diğer nöron için girdi değeri olması özelliği yoktur, zaten araştırılan değer dönüştürülen değerdir. Tek katmanlı sinir ağı Şekil 3.3'te gösterilmektedir.

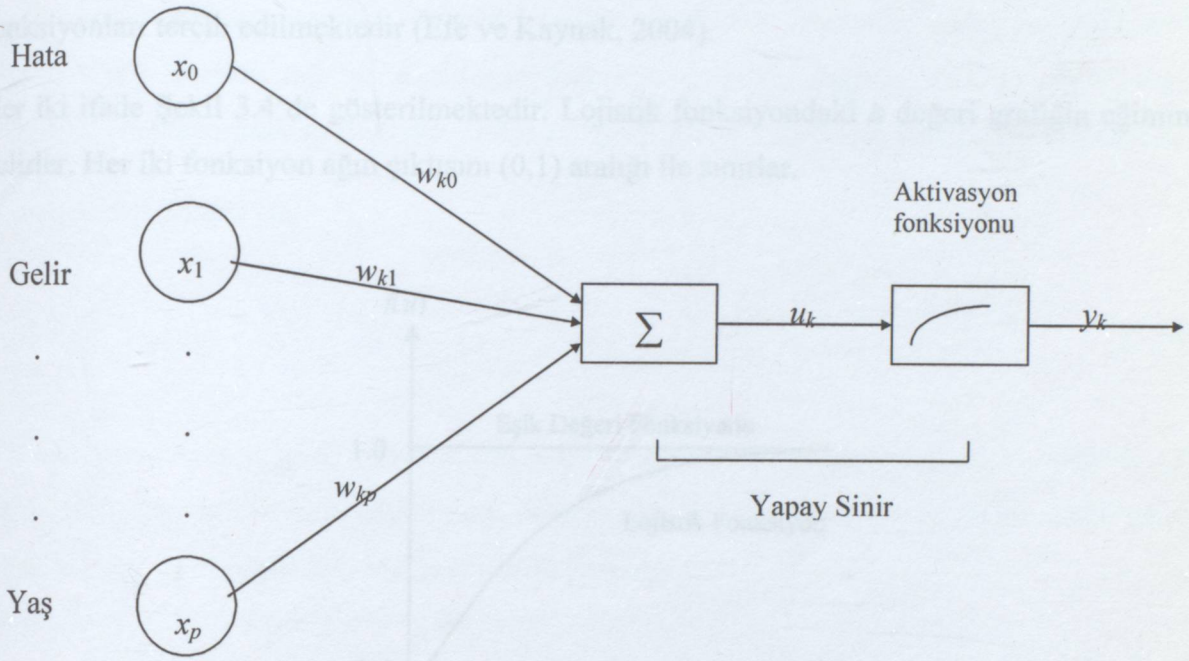
Matematiksel ifadesi ise şöyledir:

$$u_k = w_{k0}x_0 + w_{k1}x_1 + \ldots + w_{kp}x_p = \sum_{q=0}^{p} w_{kq}x_q \tag{3.54}$$

$$y_k = F(u_k) \tag{3.55}$$

Her bir $x_1,\ldots.x_p$ kredi kartı başvurusunun bir karakteristiğini ifade eden bir değişkendir. Her biri sinyal olarak adlandırılan bir değer alır. Sinaptik ağırlıklar, eğer pozitif iseler ilgili değişkeni artırdıkları için uyarıcı olarak bilinirler; eğer negatif iseler pozitif değişkenler için u_k değerini düşürdükleri için yasaklayıcı olarak adlandırılırlar. Her bir ağırlığın gösterimindeki (k, p) harflerinin sırasına dikkat edilirse; k ağırlığın uygulandığı sinir indisini,

p ise değişken indisini gösterir. Tek katmanlı sinir ağında k=1'dir, çünkü yalnızca bir tane sinir vardır. Dikkat edilirse x_0 değerine 1 atanır böylece (3.54) ifadesindeki $w_{k0}x_0$ terimi yalnızca w_{k0} ile gösterilir ve hata terimi olarak bilinir. Bu ifade (3.54) sabit bir değerle artan veya azalan bir u_k fonksiyonuna sahiptir.



Şekil 3.3 Tek katmanlı yapay sinir ağı (Thomas vd., 2002)

u_k değeri bir aktivasyon fonksiyonu kullanılarak dönüşüme tabi tutulur. Eski ağlarda, bu fonksiyon lineerdi ve sadece ilgili ağların sınıflandırma problemleri ile sınırlıydı. Çeşitli alternatif transfer-dönüşüm fonksiyonları kullanılmaktadır.

Nöron davranışını belirleyen önemli etmenlerden biri nöronun aktivasyon fonksiyonudur. Biyolojik nöronlarda, toplam belli bir değeri aştığında nöronun kısa süreli bir darbe gönderdiği bilinmektedir. Bu davranışa benzer bir davranışı yapay nöronlarla elde etmek için kullanılan aktivasyon fonksiyonlarından ikisi;

- Eşik değeri fonksiyonu

$F(u)$ =1 eğer $u \geq 0$,

=0 eğer $u<0$ (3.56)

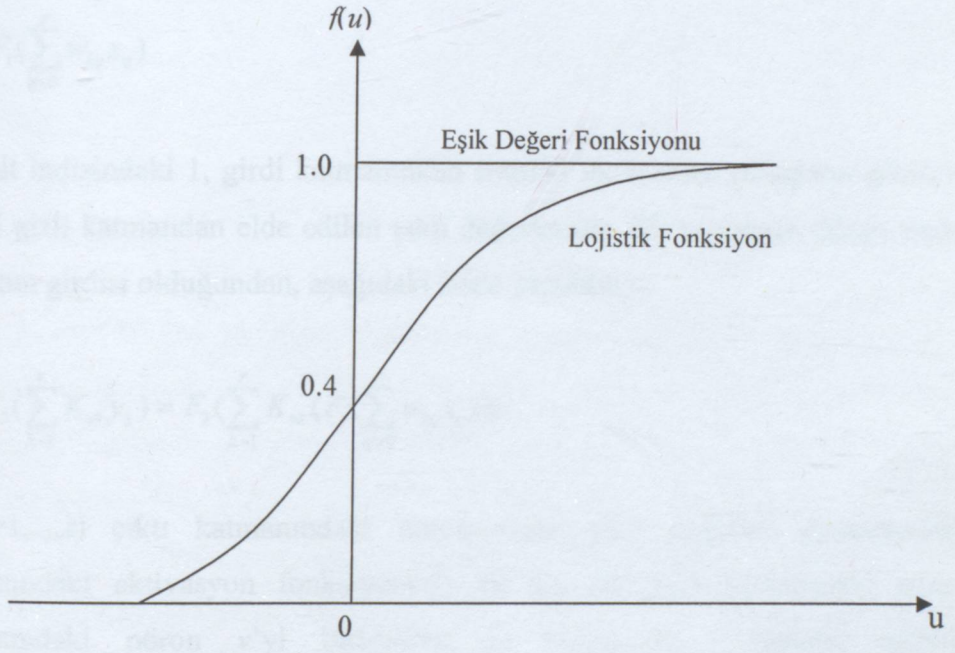
u=0 için en büyük sinir çıktı değeri 1 olur. Aksi taktirde, çıktı değeri 0 olur.

- lojistik fonksiyon

$$F(u)=\frac{1}{1+e^{-au}} \tag{3.57}$$

İkili karar mekanizması gerektiren durumlarda (3.56) ile tanımlan sert geçişli aktivasyon fonksiyonları tercih edilmektedir (Efe ve Kaynak, 2004).

Her iki ifade Şekil 3.4'de gösterilmektedir. Lojistik fonksiyondaki a değeri grafiğin eğimini belirler. Her iki fonksiyon ağın çıktısını (0,1) aralığı ile sınırlar.



Şekil 3.4 Eşik değeri ve lojistik fonksiyonlar (Thomas vd., 2002)

Ağırlıklar ve transfer fonksiyonu için değerler verildiğinde, başvuranın özelliklerini (3.54) ifadesine ekleyerek, (3.55)'den y_k değerini hesaplayıp, bu değeri kesim noktası değeri ile kıyaslayarak bir kredi kartı başvurusu için kabul veya ret tahmini yapılabilir.

1086'da Rumelhart, Hinton ve Williams, lineer olarak ayrılamayan olayları lineer olmayan transfer fonksiyonlarına sahip çok katmanlı ağlar ile sınıflandırma tekniğini geliştirmişlerdir. Hemen hemen aynı zamanlarda, bu tür modellerde ağırlıkların tahminin için geriye doğru yayılma algoritması Rumelhart, Mcclelland, Parker, ve Lecun tarafından ifade edilmiştir (Thomas vd., 2002).

3.6.3 Çok Katmanlı Sinir Ağları (Multilayer Perceptrons)

Çok katmanlı algılayıcı; sinyallerin bir girdi katmanından, çıktı sinyallerine ait (farklı y_v değerleri) bir çıktı katmanından, ve bu katmanlar arasında olan gizli katman olarak adlandırılan çok sayıda sinir tabakasından meydana gelir. Gizli katmandaki her bir nöron, bu katmandaki farklı bir nöron için aynı girdilerden farklı girdilerin meydana geldiği bir ağlık setine sahiptir. Gizli katmandaki her bir nörondaki çıktılar ağırlıklara sahiptir ve bunlar eğer varsa sonraki gizli katmandaki nöronlar için girdi olur. Aksi durumda, çıktı katmanına girdi olurlar. Çıktı katmanı her bir nöron için değerleri verir. Bu değerler, her bir olayı kesme değeri ile karşılaştırarak sınıflamada kullanılır. Şekil 3.5'te 3 katmanlı bir ağ gösterilmektedir.

Matematiksel ifadesi ise;

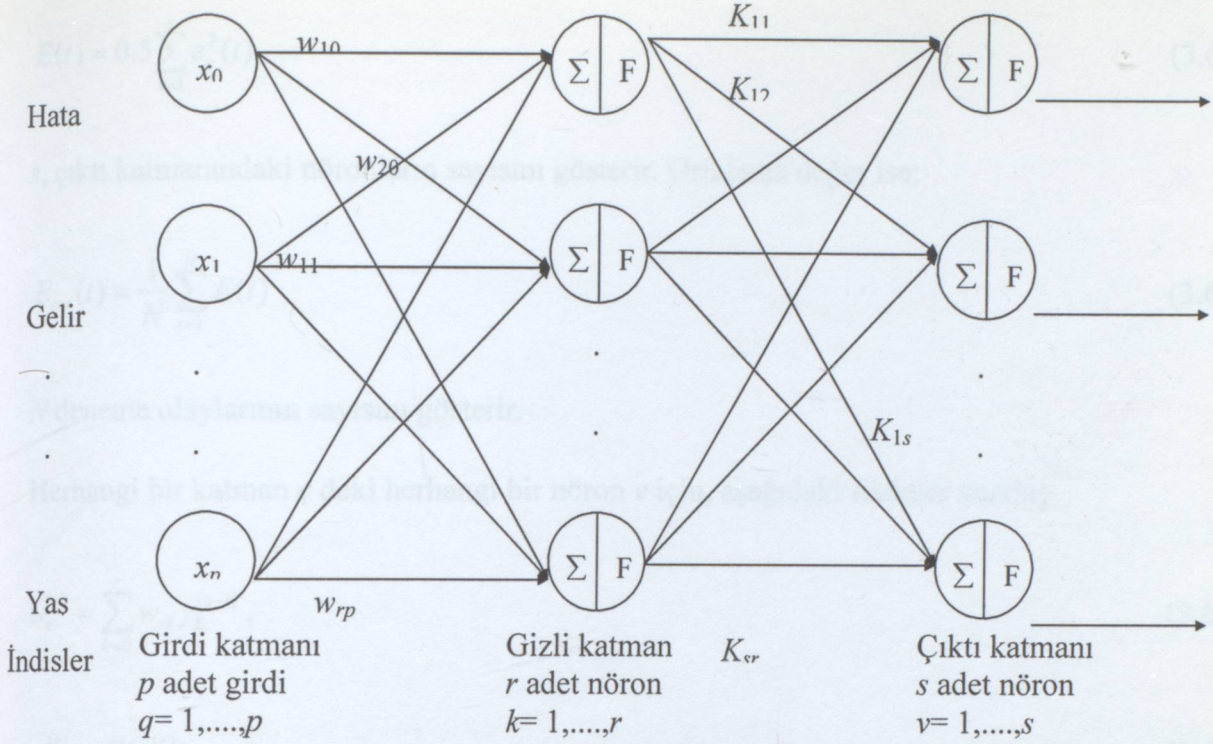
$$y_k = F_1(\sum_{q=0}^{p} w_{kq} x_q) \tag{3.58}$$

F'in alt indisindeki 1, girdi katmanından sonraki ilk katman olduğunu gösterir. y_k (k=1,...,r) birinci gizli katmandan elde edilen çıktı değerleridir. Bir katmanın çıktısı ondan sonra gelen katmanın girdisi olduğundan, aşağıdaki ifade yazılabilir;

$$z_v = F_2(\sum_{k=1}^{r} K_{vk} y_k) = F_2(\sum_{k=1}^{r} K_{vk}(F_1(\sum_{q=0}^{p} w_{kq} x_q))). \tag{3.59}$$

z_v (v=1,...,s) çıktı katmanındaki nöron v'nin çıktı değerini göstermektedir. F_2 çıktı katmanındaki aktivasyon fonksiyonunu ve K_{vk} da gizli katmandaki nöron k ve çıktı katmanındaki nöron v'yi birleştiren y_k katmanına uygulanan ağırlık değerlerini göstermektedir.

Ağırlıkların vektörel hesaplanması deneme olarak adlandırılır. Birçok teknik vardır, fakat en çok bilineni geriye doğru yayılma algoritmasıdır. Deneme çiftleri; durumun her bir girdi değişkeni için değerlerden ve olayın bilinen sınıflamasından oluşur ve hata fonksiyonunu minimize etmek için ağırlıklar tekrar tekrar hesaplanıp ağ üzerinde gösterilir.



Şekil 3.5 Çok katmanlı algılayıcı (Thomas vd., 2002)

3.6.4 Geriye Doğru Yayılma Algoritması (Back Propagation Algorithm)

İlk olarak bütün ağırlıklara rastsal olarak seçilen sayılar atanır. Bir deneme çifti seçilir, ve x_p değerleri kullanılarak z_v hesaplanır. Bilinen o_v değerleri ile hesaplanan z_v değerlerinin farkı alınır. Bu işleme "ileri" adı verilir. "Geri" işlemi ise; hata dağılımını ağ üzerindeki her bir ağırlığa yaptığı katkıya göre oransal olarak geriye doğru dağıtılmasından ve bu hata oranını azaltmak için ağırlıkları ayarlama işleminden meydana gelir. Sonra, ikinci deneme çifti seçilir ve ileri-geri işlemleri aynı şekilde tekrar edilir. Bu işlemler bütün olay seti için tekrar edilir, buna devir (epoch) denir. Bütün bu proses bir durdurma kriterine ulaşılana kadar devam eder.

Yeni bir olay ele alındığında ağırlıklardaki değişim, her bir ağırlığa göre hata teriminin birinci türevine oranı şeklinde ifade edilir. Deneme olayı t ele alındığında, buna ait hata terimi şöyle tanımlanır;

$$e_v(t) = o_v(t) - y_v(t), \tag{3.60}$$

$o_v(t)$, nöron v deki olay t için gözlenen gerçek sonuçları ifade eder, $y_v(t)$ ise tahmin edilen sonuçları gösterir. Amaç, bütün deneme olaylarının ortalama değerini minimize eden bir ağırlıklar vektörü seçmektir. Şöyle ki;

$$E(t) = 0.5\sum_{v=1}^{s} e_v^2(t), \tag{3.61}$$

s, çıktı katmanındaki nöronların sayısını gösterir. Ortalama değer ise;

$$E_{ort}(t) = \frac{1}{N}\sum_{t=1}^{N} E(t) \tag{3.62}$$

N deneme olaylarının sayısını gösterir.

Herhangi bir katman c deki herhangi bir nöron v için, aşağıdaki ifadeler yazılır;

$$u_v^{[c]} = \sum_{k=0}^{r} w_{vk} y_k^{[c-1]}, \tag{3.63}$$

$$y_v^{[c]} = F(u_v^{[c]}), \tag{3.64}$$

Bu ifadeler (3.54) ve (3.55)'ün genelleştirilmiş şeklidir, yani herhangi bir katman ve herhangi bir nöron için yazılmış halidir. ((3.54) ve (3.55) tek bir nöron için yazılmıştır).

Böylece, zincir kuralı uygulanarak, $E(t)$'nin ağırlık w_{vk} ye göre kısmi türevi aşağıdaki şekilde ifade edilir;

$$\frac{\partial E(t)}{\partial w_{vk}(t)} = \frac{\partial E(t)}{\partial e_v(t)} \frac{\partial e_v(t)}{\partial y_v(t)} \frac{\partial y_v(t)}{\partial u_v(t)} \frac{\partial u_v(t)}{\partial w_{vk}(t)} \tag{3.65}$$

3.60'dan,

$$\frac{\partial E(t)}{\partial e_v(t)} = e_v(t) \tag{3.66}$$

3.59'dan,

$$\frac{\partial e_v(t)}{\partial y_v(t)} = -1 \tag{3.67}$$

3.63'den,

$$\frac{\partial y_v(t)}{\partial u_v(t)} = F'(u_v(t)) \tag{3.68}$$

3.62'den

$$\frac{\partial u_v(t)}{\partial w_{vk}(t)} = y_k(t)\,. \tag{3.69}$$

Bu ifadeler (3.64)'de yerlerine konulursa,

$$\frac{\partial E(t)}{\partial w_{vk}(t)} = -e_v(t).F'(u_v(t)).y_k(t) \tag{3.70}$$

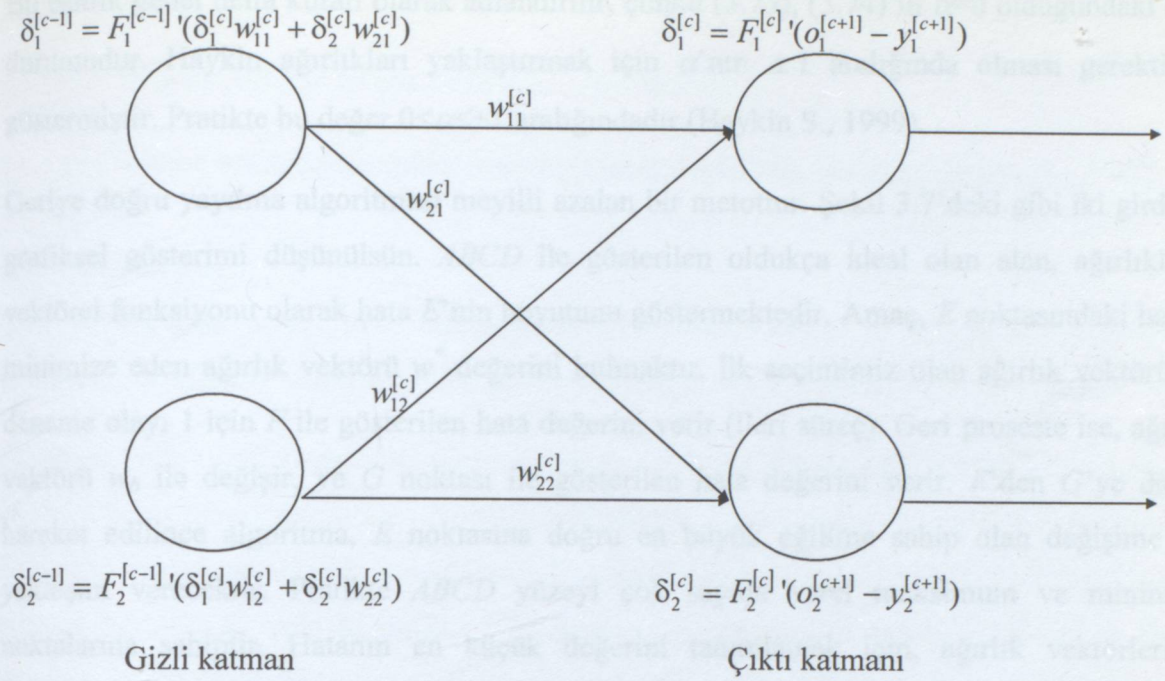
elde edilir.

İleri işlemde hesaplanan ağırlıklar ve geri işlemde hesaplanan ağırlıklar arasındaki değişim ise aşağıdaki gibi olur.

$$\Delta w_{vk}(t) = -\eta \frac{\partial E(t)}{\partial w_{vk}(t)} = \eta \delta_v(t) y_k(t)\,, \tag{3.71}$$

$\delta_v(t) = e_v(t)\, F'(u_v(t))$ olduğunda. η sabiti deneme oranı katsayısı olarak adlandırılır ve bu sabit ayarlanarak w'daki değişimlerin küçük veya büyük olmasını sağlar. Eşitlik (3.71) Delta Kuralı (Delta Rule) yada Widrow-Hole Kuralı olarak bilinir.

Bu kuralın uygulaması, nöron v 'nin çıktı katmanında veya gizli katmanda olmasına göre değişir. Eğer nöron v çıktı katmanındaysa, gözlenen sonuç o_v ve tahmini sonuç y_v bilindiği için e_v değeri direkt olarak elde edilir. Eğer nöron v gizli katmandaysa, e_v terimindeki bir bileşen, o_v, elde edilemez. Bu durumda yine (3.72) kullanılır, ancak $\delta_v(t)$ farklı bir yolla hesaplamalıdır. Çıktı katmanındaki her bir nöron için δ değeri hesaplanır, hesaplama yapılan nöron ile bir önceki katmandaki bir nöronu birleştiren ağırlık değerleri her bir δ ile çarpılır, ve bir önceki katmandaki her bir nöron için bu çarpımların toplamı alınır. Şekil 3.6 bu tekniği çıktı katmanındaki iki nöron için göstermektedir.



Şekil 3.6 Geriye doğru yayılma (Thomas vd., 2002)

Genel olarak;

$$\delta_k^{[c-1]} = F_k^{[c-1]\prime} \sum_{v=1}^{s} \delta_v^{[c]} w_{vk}^{[c]} \tag{3.72}$$

Bir katmandaki her bir nöron için δ değeri, bir önceki katmandaki nöronlara bağlayan ağırlıklara göre dağıtılır, ve sonra önceki katmandaki her bir nöron için, onu bağlayan ağırlıklı δ'ler toplanır. Bu işlem önceki katman için δ değerini verir. Bu prosedür, girdi katmanından orijinal-asıl ağırlıklar seti elde edilene kadar katman katman tekrar edilir.

Herhangi [c-1] katmanını (indisleri k olan), sonraki [c] katmanına (indisleri v) bağlayan ağırlıklardaki değişim (3.71) ve (3.72) eşitliklerine dayanarak şu şekilde hesaplanır;

$$\Delta w_{vk} = \eta \delta_v^{[c]} y_k^{[c-1]} \tag{3.73}$$

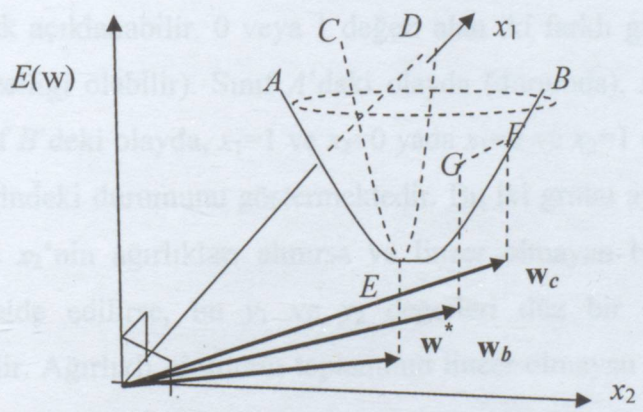
Bu kural genellikle bir moment terimi, $\alpha \Delta w_{vk}(t\text{-}1)$, eklenerek düzeltilir;

$$\Delta w_{vk}(t) = \alpha \Delta w_{vk}(t-1) + \eta \delta_v^{[c]} y_k^{[c-1]} \tag{3.74}$$

Bu eşitlik genel delta kuralı olarak adlandırılır, çünkü (3.73), (3.74)'in α=0 olduğundaki özel durumudur. Haykin ağırlıkları yaklaştırmak için α'nın ± 1 aralığında olması gerektiğini göstermiştir. Pratikte bu değer $0<\alpha<+1$ aralığındadır (Haykin S., 1999).

Geriye doğru yayılma algoritması meyilli azalan bir metottur. Şekil 3.7'deki gibi iki girdinin grafiksel gösterimi düşünülsün. *ABCD* ile gösterilen oldukça ideal olan alan, ağırlıkların vektörel fonksiyonu olarak hata *E*'nin boyutunu göstermektedir. Amaç, *E* noktasındaki hatayı minimize eden ağırlık vektörü w^* değerini bulmaktır. İlk seçimimiz olan ağırlık vektörü w_a deneme olayı 1 için *F* ile gösterilen hata değerini verir (ileri süreç). Geri proseste ise, ağırlık vektörü w_b ile değişir, ve *G* noktası ile gösterilen hata değerini verir. *F*'den *G*'ye doğru hareket edilince algoritma, *E* noktasına doğru en büyük eğilime sahip olan değişime bir yaklaşım verecektir. Pratikte *ABCD* yüzeyi çok sayıda yerel maksimum ve minimum noktalarına sahiptir. Hatanın en küçük değerini tanımlamak için, ağırlık vektörlerinin $\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots$ sıfıra eşit olduğu durumda hata yüzeyinin birinci kısmi türevi alınması gerekir.

Bu elde edildiğinde algoritma durdurulur. Fakat, bu durdurma algoritmasının çözümü çok zaman almaktadır ve bunların türev değerleri için bilgi gerektirir. Durdurma algoritması için alternatif bir teknik ise; $E_{ort}(w)$ değerindeki mutlak değişimin yeteri kadar küçük olması durumudur. Örneğin, her bir epokta (devirde) 0.01'den az olması olabilir.



Şekil 3.7 Geriye doğru yayılma ve hata yüzeyi(Thomas, vd. 2002)

Geri yayılım algoritması;

- En çok kullanılan algoritmalardan biridir.

- Dereceli azalır.
- MLP'lerin eğitiminde en çok kullanılan algoritmadır.
- i ve j katman arasında tanımlanan ağırlıklardaki değişikliği hesaplar (Sağıroğlu, 2002).

3.6.5 Ağ Yapısı

Gizli katmanların sayısı, gizli katmanlardaki nöronların sayısı ve hata fonksiyonu deney tasarımcısı tarafından seçilir. Burada, gizli katmanların sayısı üzerine odaklanılacaktır.

İlk olarak neden gizli katmanlara sahip olunduğu düşünüldüğünde; eğer tek bir katman olursa, yalnızca lineer olarak ayrılabilen gruplar sınıflanabilir. Lineer olmayan aktivasyon fonksiyonu ile birlikte gizli katmanlar da modele eklenirse, en son elde edilen ağ lineer olarak ayrılmayan olayları da doğru bir şekilde sınıflar. Gizli katmanlar, değişkenler arasındaki karmaşık lineer olmayan ilişkileri modeller ve gizli uzaya girdi değişkenlerini dönüştürür. Gizli katmanlar verideki özellikleri ortaya çıkarır. Özellikle, tek lineer olmayan gizli katman sonuçlarını ağa dahil etmek girdi değişkenlerinin konveks alanında bir kesme değerinin altında yada üzerinde bir değer verecektir. İkinci bir gizli katmana sahip olmak konveks alanları birleştirmeye izin verecektir. Ortaya çıkan alan konveks olmayabilir yada tamamen birbirinden ayrı bölgeler olabilir.

Gizli katman ve lineer olmayan aktivasyon fonksiyonunu eklemenin etkisi *X*OR probleminde uygulayarak açıklanabilir. 0 veya 1 değeri alan iki farklı girdi düşünüldüğünde. (kredi kartı başvuru özelliği olabilir). Sınıf *A*'daki olayda (durumda), x_1=0 ve x_2=0 yada x_1=1 ve x_2=1 olsun. Sınıf *B*'deki olayda, x_1=1 ve x_2=0 yada x_1=0 ve x_2=1 olsun. Şekil 3.8 bu dört olasılığın grafik üzerindeki durumunu göstermektedir. Bu iki grubu ayıran düz bir çizgi yoktur. Fakat, eğer x_1 ve x_2'nin ağırlıkları alınırsa ve lineer olmayan bir dönüşüm uygulanıp y_1 ve y_2 değerleri elde edilirse, bu y_1 ve y_2 değerleri düz bir çizgi ile grupları ayırmak için kullanılabilir. Ağırlıklı girdilerin toplamının lineer olmayan dönüşümünü içeren gizli katman modele dahil edilir. Dört durumdaki ağırlık değerlerinin +1 olduğu ve ilgili hataların $-\frac{3}{2}$ ve $-\frac{1}{2}$ olduğu durumda, aşağıdaki eşitlikler yazılabilir:

$$u_1 = x_1 + x_2 - \frac{3}{2}, \qquad u_2 = x_1 + x_2 - \frac{1}{2},$$

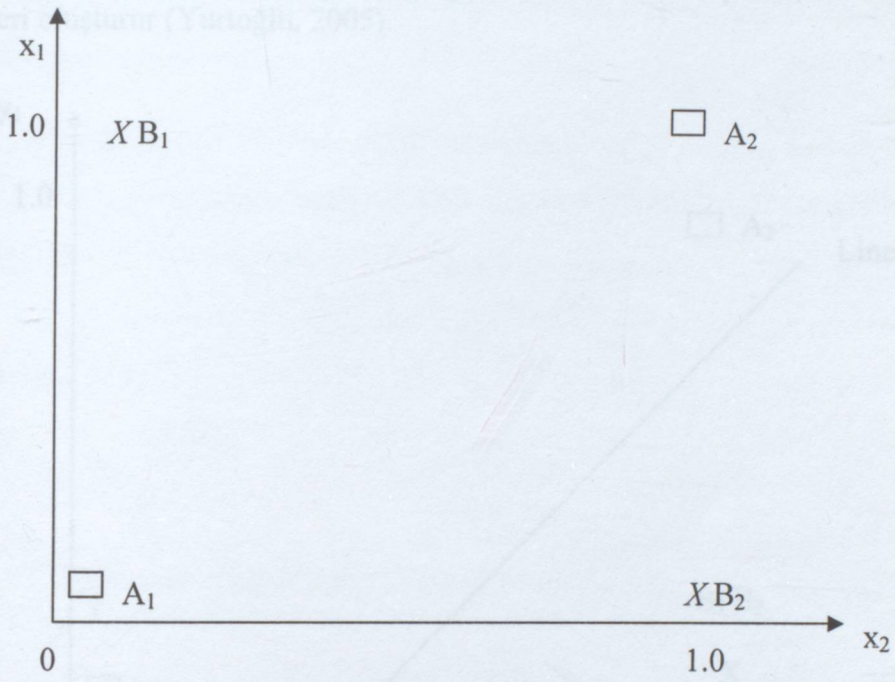
Aktivasyon fonksiyonunu kullanılırsa,

$u_1 < 0 \Rightarrow y_1 = 0,$

$u_1 \geq 0 \Rightarrow y_1 = 1,$

$u_2 < 0 \Rightarrow y_2 = 0,$

$u_2 \geq 0 \Rightarrow y_2 = 1,$



Şekil 3.8 XOR problemi için veriler (Thomas vd., 2002)

Bu dört durum (Şekil 3.9)'da gösterilmiştir. Sınıf *A*, (x_1, x_2) değerleri (0.0) ve (1,1) olan, (y_1,y_2) değerleri (0,0) ve (1,1) olmuştur. *x* uzayındaki değerleri (0,1) ve (1,0) olan sınıf *B* olayı *y* uzayında (0,1) ve (0,1) değerlerini almıştır. Bu durumda düz bir çizgi bu iki grubu ayırabilir. Bu düz çizgi çıktı katmanını gösterir. Fakat, pratikte ikiden fazla katman nadiren kullanılır.

Ek olarak, uygulayıcı her bir gizli katmana kaç tane nöron eklenmesi gerektiğine karar vermelidir. Denemeler başlamadan optimal sayı hakkında fikir sahibi olunamaz. Bazı sezgisel teknikler bunun için geliştirilmiştir.

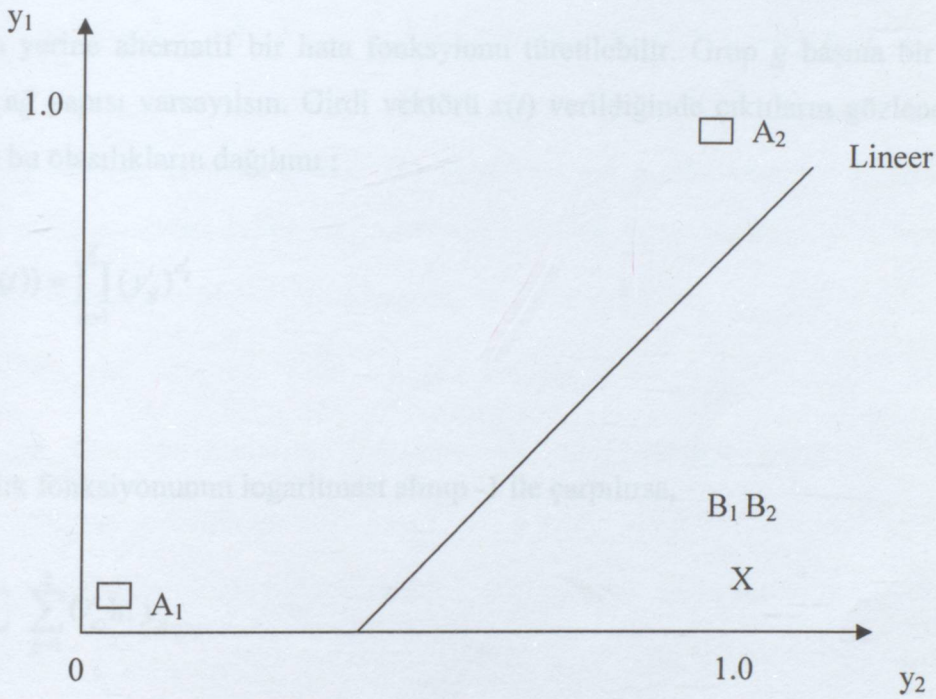
Ağın genel yapısına dönülürse tabaka sayısı ve tabakaların içerdigi işlem elemanı sayısı ağın performansı açısından önemli ve zor kararlardır. Zor karar olmalarının sebebi ise herhangi bir uygulama için net bir seçim kriterinin olmamasıdır. Bunun yerine, uygulamalar sonucunda ortaya çıkmış ve araştırmacılar tarafından benimsenmiş bazı kurallar vardır. Bu kurallar şu

şekilde özetlenebilir:

Kural-1: Girdi ve çıktı veriler arasındaki ilişkinin karmaşıklığı derecesi arttıkça, tabakaların içerdigi işlem elemanı sayısı da artmalıdır.

Kural-2: Modellenen konu değişik safhalara ayrılabiliyorsa, tabaka sayısının artırılması gerekebilir.

Kural-3: Eldeki eğitme verisinin genişliği, gizli tabakalardaki toplam nöron sayısı için bir üst limit kriteri oluşturur (Yurtoğlu, 2005).



Şekil 3.9 Dönüştürülmüş veri (Thomas vd., 2002)

3.6.6 Sınıflandırma ve Hata Fonksiyonları

KS'de, genellikle bireyler farklı gruplara sınıflanmak istenir. Şöyle ki; iyi oyuncular-zayıf oyuncular yada iyi, zayıf ve kötü oyucular. Prensipte, olaylar Z gruba sınıflanmak istendiğinde, çok katmanlı algılayıcıdan Z adet çıktıya ihtiyaç duyulur. White , Richard ve Lipman sonuçlarına dayanarak, sonlu bağımsız olayları ve girdilerin özdeş dağılımını kullanarak (3.62)'in değerini geriye yayılma algoritması ile minimize eden çok katmanlı algılayıcı modeli sınıfların sonraki olasılıklarına asimptotik bir yaklaşımla sonuçlanır (Bishop, 1995). Böylece, deneme örneklemesinin bu koşulları sağlamasıyla, bir olayı bir $C_g, g = 1,....,Z$ grubuna tahsis etmek için bir karar verme kuralı geliştirilir. Eğer üzerinde çalışılan grubun çıktı katmanındaki çıktı değeri , $F_g(x)$, diğer çalışma gruplarının çıktı

katmanındaki değerinden büyük ise, $F_h(x)$,

$$F_g(x) > F_h(x), \qquad g \neq h \tag{3.75}$$

Buna karşın, (3.62)'daki hata fonksiyonunun sınıflandırma problemleri için en iyi hata fonksiyonu olması gerekmez. Bu, bütün ağdaki ağırlıkların ve fonksiyonların ürettiği normal dağılıma sahip belirli çıktı değerlerinin, o_v, olasılıklarının maksimizasyonu ile elde edilir. Fakat, sınıflandırma problemlerinde, çıktı değerleri ikili değişkenlerdir (binary): bir olayın bir grubu ait olması veya olmaması.

(3.62)'in yerine alternatif bir hata fonksyionu türetilebilir. Grup g başına bir çıktının (y_{vg}) olduğu ağ yapısı varsayılsın. Girdi vektörü $x(t)$ verildiğinde çıktıların gözlenen olasılığı y_g olsun ve bu olasılıkların dağılımı ;

$$P(o(t)|x(t)) = \prod_{g=1}^{Z} (y_g^t)^{o_g^t} \tag{3.76}$$

olsun.

Bu olasılık fonksiyonunun logaritması alınıp -1 ile çarpılırsa,

$$E_2 = -\sum_t \sum_{g=1}^{Z} O_g^t \ln y_g^t \tag{3.77}$$

elde edilir.

Bu göreceli entropi kriteridir.

y_v değerleri olasılık olarak yorumlandığı için, bu değerler şu özelliklere sahip olmalıdır. $0 \le y_{vg} \le 1$ ve $\sum_{g=1}^{Z} y_g = 1$. Bunu elde etmek için aşağıdaki gevşek aktivasyon fonksiyonu kullanılır.

$$y_g = \frac{e^{u_g}}{\sum_{g=1}^{Z} e^{u_g}} \tag{3.78}$$

Bu hata ve aktivasyon fonksiyonları aslında Desai tarafından, lojistik regresyon, lineer diskriminant analizi ve sinir ağlarının sınıflandırma performanslarını karşılaştırmak için

kullanılmıştır.

3.7 Doğrusal Programlama (Lineer Programming)

Mangasarian iki ayrı grubun yer aldığı sınıflama problemlerinde doğrusal programlamanın kullanabileceğini belirtmiştir. Freed ve Glover ve Hand, Mutlak Hatalar Toplamını (MSAE) yada Maksimum Hataların Minimizasyonunu (MME) amaç fonksiyonu alan lineer programlama modellerini lineer olması gerekmeyen iki grubu ayırmak için tanımışlardır (Thomas vd, 2002).

Uygulama değişkeni, $X\left(X_1, X_2, ..., X_p\right)$ değerlerinin bütün kombinasyonları A_G ve A_B olarak adlandırılan iki gruba ayrılabilir: A_G, iyi olarak cevaplanan soruları; A_B, kötü olarak cevaplanan soruları göstermektedir. Bir örneklemenin n adet başvuru aldığı kabul edilsin. Örneklemenin n_G adeti iyiler ise, notasyon kolaylığı için, bunların örneklemedeki birinci n_G 'ler olduğu kabul edilsin. Örneklemenin (Örnekleme $i = n_G + 1, ..., n_G + n_B$) geri kalan n_B adedi "kötü" olur. Başvuru i'nin başvuru değişkeni X'e $\left(X_1, X_2, ..., X_p\right)$ olan cevabında $\left(x_{i1}, x_{i2}, ..., x_{ip}\right)$ karakteristiklerine (özelliklerine) sahip olduğu kabul edilsin. Cevapların ağırlıklı ortalama toplamının $w_1X_1 + w_2X_2 + ... + w_pX_p$ iyi başvurular için belirli bir kesme değerinin c (cutoff)üzerinde olduğu ve kötüler için kesme değerinin altında olduğu ağırlıkları veya skorları $\left(w_1, w_2, ..., w_p\right)$ seçilmek istenir. Eğer başvuru değişkenleri ikili (binary) değişkene dönüştürülürse, verilen farklı cevaplar için ağırlıklar (w_i'leri) onların skoru olarak düşünülebilir.

Genellikle iyiler ve kötüler biri birinden kusursuz bir şekilde ayrılamaz. Dolayısıyla, muhtemel hatalar için pozitif veya sıfır değeri alan a_i değişkenleri tanımlanır. Böylece, örneklemede başvuru i iyi değeri alırsa, $w_1x_{i1} + w_2x_{i2} + ... + w_px_{ip} \geq c - a_i$ kısıdı tanımlanır. Buna karşın, eğer başvuru kötü olursa, , $w_1x_{j1} + w_2x_{j2} + ... + w_px_{jp} \leq c + a_j$ kısıdı tanımlanır. Buradaki sapmaların mutlak değerler toplamını (MSD) minimize eden ağırlıklarını $\left(w_1, w_2, ..., w_p\right)$ bulmak için, aşağıdaki lineer programlama modeli çözülmelidir:

Minimize $a_1 + a_2 + ... + a_{n_G + n_B}$ (3.79)

Şu kısıtlara göre: $w_1x_{i1} + w_2x_{i2} + ... + w_px_{ip} \geq c - a_i, \qquad 1 \leq i \leq n_G$

$$w_1 x_{i1} + w_2 x_{i2} + ... + w_p x_{ip} \leq c + a_i, \qquad n_G + 1 \leq i \leq n_G + n_B$$

$$a_i \geq 0, \qquad 1 \leq i \leq n_G + n_B$$

Maksimum Sapmanın Minimizasyonu (MMD) yerine, her bir kısıttaki aynı hatalar teriminin minimizasyonu yapılırsa, aşağıdaki model çözülür:

Minimize a (3.80)

Şu kısıtlara göre: $w_1 x_{i1} + w_2 x_{i2} + ... + w_p x_{ip} \geq c - a$, $\qquad 1 \leq i \leq n_G$

$$w_1 x_{i1} + w_2 x_{i2} + ... + w_p x_{ip} \leq c + a, \qquad n_G + 1 \leq i \leq n_G + n_B$$

$$a \geq 0, \qquad 1 \leq i \leq n_G + n_B$$

KS'de, lineer programlamanın istatistiksel metotlara göre bir avantajı: belirli hata seviyesinde bir skor kart istenirse, lineer programlama skor kart geliştirme prosesine hataları kolaylıkla dahil edebilir. Örneğin, ikili (binary) değişken X_1, 25'in altında olması veya olmaması olarak tanımlansın ve ikili (binary) değişken X_2, 65'in üzerinde olması veya olmaması olarak tanımlansın. 25 yaşın altındakiler için skorun emekliler skorundan yüksek olması istensin. Bir skor kart elde etmek için (3.79) ve (3.80) kısıtlarına $w_1 \geq w_2$ kısıtını eklemek gerekir. Aynı şekilde, başvuru formu değişkenlerinin (X_1'den X_s'e kadar) ağırlıklarının kredi bürosu değişkenleri ağırlıklarını (X_{s+1} den X_s'ye) aştığını her bir başvuru için garanti eden aşağıdaki kısıdı (3.79) ve (3.80)'ye eklemelidir:

$$w_1 x_{i1} + w_2 x_{i2} + ... + w_s x_{is} \geq w_{s+1} x_{is+1} + w_{s+2} x_{is+2} + ... + w_p x_{ip} \quad \text{bütün } i\text{'ler için} \qquad (3.81)$$

Lineer programlama yaklaşımına en büyük eleştiri, istatistiğe dayalı çözümler elde edilmediği için, tahmin edilen parametrelerin istatistiksel olarak anlamlı (significant) olup olmadığının değerlendirilemediği şeklindedir.

Regresyonun lineer programlama üzerindeki diğer avantajı; bilindiği gibi birçok istatistik paket programında en güçlüsünden başlayarak skor karta belirli bir zamanda değişkenler tanımlanabilir. Böylece, istediğiniz karakteristik sayısını (m), regresyonu ve en çok ayrılan m özelliği bulma kararını veren zayıf ve ortalama modeller geliştirebilir. Nath ve Jones jakknife yaklaşımının en güçlü m adet karakteristiğin seçiminde lineer programlamanın nasıl uygulandığını göstermişlerdir (Thomas vd., 2002).

Erenguc ve Koehler, ve Nath, Jackson ve Jones çeşitli data setlerinin sınıflandırılmasında regresyon ve lineer programlama modellerini karşılaştırmışlardır.(Fakat bu data setlerinin hiçbiri kredi ile ilgili değildir). Elde ettiği sonuçlar göstermektedir ki, lineer programlama yaklaşımı istatistiksel yaklaşımlar kadar bir sınıflama yapamamaktadır.

3.8 Tamsayılı Programlama (Integer Programming)

Lineer programlama modelleri yanlış sınıflanan başvuru kredi skorlarındaki standart sapmanın toplamını minimize eder. Buna karşın, daha pratik bir yol yanlış sınıflama sayılarını, yada eğer bir kötüyü iyi olarak yanlış sınıflamanın maliyeti M bir iyiyi kötü olarak yanlış sınıflama maliyetinden K çok farklıysa yanlış sınıflamanın toplam maliyetini, minimize etmektir. Bu kriterleri dikkate alarak skor kartlar oluştururken lineer programlama kullanılabilir, fakat bazı değişkenler tamsayı olmalıdır (0, 1 gibi). Bu teknik tamsayılı programlama olarak adlandırılır. Koehler ve Erenguc aşağıdaki tamsayılı programlama modelini geliştirmişlerdir:

$$\text{minimize } K(d_1 + \ldots + d_{n_G}) + M(d_{n_{G+1}} + d_{n_{G+B}}) \tag{3.82}$$

şu kısıtlara göre: $w_1 x_{i1} + \ldots + w_p x_{ip} \geq c - Md_i, \qquad 1 \leq i \leq n_G$

$$w_1 x_{i1} + \ldots + w_p x_{ip} \leq c + Md_i, \qquad n_G + 1 \leq i \leq n_G + n_B,$$

$0 \leq d_i \leq 1, \qquad d_i$ tamsayı.

Böylece d_i, eğer örneklemede müşteri i yanlış sınıflanırsa 1'e eşit sınıflanmaz ise 0'a eşit olan bir değişken olacak. c=0, w_i =0, i=0,....,p olduğunda (3.82) denklemi tekrardan minimize edilirse, aşağıdaki gibi bir normalizasyon kısıdı eklenmelidir.

$$\sum_{j=1}^{p} \left(s_j^+ + s_j^- \right) = 1, \tag{3.83}$$

$0 \leq s_j^+, \qquad s_j^- \leq 1, \qquad$ ve s_j^+, s_j^- tamsayı, j=1,.....p

$$-1 + 2s_j \leq w_j \leq 1 - 2s_j, \qquad j=1,.....p$$

Bu kısıtlar seti s_i^+ ve s_i^-'den birinin 1'e eşit olmasını ve ilgili w_i değerinin 1'den büyük veya -1'den küçük olmasını gerektirir (c'yi negatif veya pozitif değere eşitlemeye zorlar). Bu durumda, aşağıdaki eşitlik yazılır,

$\sum_{j=1}^{p} w_i = 1$ ve $c = +1$ yada -1

Koehler ve Erenguc (3.82)'te gösterilen tamsayılı programlama modelinin lineer programlama modellerine göre daha iyi bir sınıflama yaptığını göstermişlerdir. Fakat, iki büyük dezavantaj vardır. Birincisi, tamsayılı programlama modellerinin çözümü lineer modellere göre çok daha fazla zaman almakta ve böylece yüzlerce olay setinden yalnızca küçük bir örnekleme ile çözüm yapılabilmektedir. İkincisi, deneme seti üzerinde aynı yanlış sınıflandırma sayısı ile çok sayıda optimal çözüm vardır, fakat incelenen örneklemelerde çok az değişiklik olan performanslar vardır.

Çeşitli araştırmacılar hem incelenen örneklemede güçlü olan bir sınıflamayı garanti etmek hem de deneme örneklemesindeki yanlış sınıflamaların sayısını azaltmak için (3.82) denklemine ekstra kısıtlar ve ikincil amaç fonksiyonları eklemişlerdir. Şu ana kadar tam sayılı programlama minimum yanlış sınıflandırma kriterine sahip sınıflandırma problemlerinin çözümü için kullanılmıştır, fakat hataların mutlak toplamının minimizasyonunda lineer programlama yaklaşımında ortaya çıkan iki güçlüğün üstesinden tamsayılı programlama gelir.

4. UYGULAMALAR

Kredi kartı skorlaması için bir bankanın kredi kartı müşterilerine ait 1806 gözlem ve on üç değişken ile çalışılmıştır. Değişkenlerin kategorileri kredi kartı başvuru formundaki kategoriler baz alınarak kodlanmıştır. AYLIK NET GELİR, MAAŞ gibi parasal değerlerin alt sınırlarının küçük olmasının nedeni kredi kartı müşterilerinin bilgilerinin güncellenmemiş olmasıdır. İyi müşteri-kötü müşteri ayrımı banka tarafından yapılmış veriler üzerinde analizler uygulanmıştır.YSA uygulaması dışındaki uygulamalarda SPSS 13.0 paket programı, YSA uygulamasında ise STATISTICA 7 programı kullanılmıştır. Uygulama sonuçları ilgili tekniklerin başlıkları altında, orijinal program çıktıları ise ekler kısmında yer almaktadır. Uygulamada kullanılan değişkenler Çizelge 4.1'de gösterilmektedir.

Çizelge 4.1 Değişkenler

İYİ-KÖTÜ	Müşterilerin iyi yada kötü oluşu
1	İyi
2	Kötü
MÜŞTERİ TİPİ:	Müşterilerin hangi kampanyalarla banka müşterisi oldukları
1	B (Markette kazanılan müşteriler)
2	C (Mektup1)
3	D (Mektup 2)
4	G (Grup Firmaların Çalışanları)
5	J (Özel Müşteriler)
6	M (Mail Order)
7	N (Normal)
8	O (Okul Kampanyası)
9	R (Alışveriş Merkezi)
10	T (Toplu Pazarlama)
YAŞ:	Müşteri yaşı
1	18-30 yaş arası
2	30-39 yaş arası
3	39+
CİNSİYET:	Müşteri cinsiyeti
1	Erkek
2	Bayan
MEDENİ DURUM:	Müşteri medeni durumu
1	Evli
2	Dul
3	Bekar
ÖĞRENİM DURUMU:	Müşterinin öğrenim durumu
1	Okula Gitmemiş
2	İlkokul Mezunu
3	Ortaokul Mezunu
4	Lise Mezunu
5	Yüksekokul Mezunu
6	Üniversite Mezunu
7	Master/Doktora
ARABA SAHİBİ:	Müşterinin arabası olup-olmaması
1	Var
2	Yok
MAAŞ:	Müşterinin maaşı
1	100-750 YTL
2	750-1000 YTL
3	1000-1900 YTL
4	1900-3000 YTL
5	3000 YTL+
AYLIK NET GELİR:	Müşterinin aylık net geliri
1	700 YTL altı
2	700 – 1000 YTL
3	1000 – 1800 YTL

4	1800 – 3000YTL
5	3000 YTL ve üzeri
BAŞKA BANKA KREDİ KARTI:	Müşterinin başka banka kredi kartı olup olmaması
1	1 tane var
2	2 tane var
3	Yok
BAŞKA BANKA KART LİMİTİ:	Müşterinin başka banka kart limiti
1	Yok
2	<300 YTL
3	300 YTL+
EK KART DURUMU:	Müşterinin ek kartı olup olmaması
1	Var
2	Yok
ÇALIŞMA ŞEKLİ:	Müşterinin işindeki unvanı
1	Diğer
2	Emekli
3	Ev hanımı
4	Kamu sektörü
5	Serbest meslek
6	Özel sektör
7	Öğrenci

4.1 Betimsel İstatistikler

Çalışmada ileri analiz tekniklerine yer verilmesine rağmen ön bilgi sağlaması açısından bazı betimsel istatistikler incelenmiş ve bazı yorumlar yapılmıştır. MÜŞTERİ YAŞI, MAAŞ, AYLIK NET GELİR, BAŞKA BANKA KREDİ KARTI LİMİTİ gibi sürekli değişkenler için betimsel istatistikler Çizelge 4.2'de yer almaktadır.

Çizelge 4.2 Betimsel istatistikler

BETİMSEL İSTATİSTİKLER	İYİ - KÖTÜ			İstatistik	Standart Hata
MÜŞTERİ YAŞI	İYİ	**Ortalama**		**34,99**	0,79
		Ortalama için %95 güven aralığı	Alt Sınır	33,43	
			Üst Sınır	36,56	
		Medyan		33	
		Varyans		100,1	
		Standart Sapma		10	
		Aralık		44	
	KÖTÜ	**Ortalama**		**36,13**	0,74
		Ortalama için %95 güven aralığı	Alt Sınır	34,65	
			Üst Sınır	37,6	
		Medyan		34	
		Varyans		96,97	
		Standart Sapma		9,847	
		Aralık		53	
MAAŞ	İYİ	**Ortalama**		**1.832,00**	188,1
		Ortalama için %95 güven aralığı	Alt Sınır	1461	
			Üst Sınır	2204	
		Medyan		1000	
		Varyans		6,00E+06	
		Standart Sapma		2372	
		Aralık		19774	
	KÖTÜ	**Ortalama**		**3.331,00**	276,6
		Ortalama için %95 güven aralığı	Alt Sınır	2785	
			Üst Sınır	3877	
		Medyan		2000	
		Varyans		1,00E+07	
		Standart Sapma		3638	

BETİMSEL İSTATİSTİKLER	İYİ - KÖTÜ			İstatistik	Standart Hata
AYLIK NET GELİR	İYİ	**Ortalama**		**1.744,00**	189,1
		Ortalama için %95 güven aralığı	Alt Sınır	1.370,00	
			Üst Sınır	2.117,00	
		Medyan		1000	
		Varyans		6,00E+06	
		Standart Sapma		2384	
		Aralık		21.900,00	
	KÖTÜ	**Ortalama**		**3.118,00**	295,5
		Ortalama için %95 güven aralığı	Alt Sınır	2535	
			Üst Sınır	3702	
		Medyan		2000	
		Varyans		2,00E+07	
		Standart Sapma		3887	
		Aralık		26850	
BAŞKA BANKA KREDİ KARTI LİMİTİ	İYİ	**Ortalama**		**2337**	305,6
		Ortalama için %95 güven aralığı	Alt Sınır	1734	
			Üst Sınır	2941	
		Medyan		1500	
		Varyans		1,00E+07	
		Standart Sapma		3854	
		Aralık		44750	
	KÖTÜ	**Ortalama**		**3608**	239,7
		Ortalama için %95 güven aralığı	Alt Sınır	3135	
			Üst Sınır	4082	
		Medyan		2600	
		Varyans		1,00E+07	
		Standart Sapma		3153	
		Aralık		19880	

Çizelge 4.2 sürekli, nümerik değişkenler için özet istatistikleri göstermektedir. Özet istatistik tablosu merkezi eğilim ölçülerini (ortalama, medyan) de içerir. İyi-kötü müşterilerin farklı değişkenler için ortalamalarının farkları incelendiğinde; iyi müşterilerle kötü müşterilerin ikamet süreleri arasında çok fazla bir farklılık olmadığı gözlemlenmektedir. Müşteri yaşları açısından incelediğimizde; kötü müşterilerinin yaş ortalamasının daha yüksek olduğu görülmektedir. Maaş ve buna bağlı olarak aylık net gelir, başka banka kredi kartı limiti açısından da kötü müşterilerin ortalaması ciddi anlamda yüksektir.

Kategorik değişkenlerin frekans tabloları incelendiğinde;

Çizelge 4.3 İyi-kötü müşteriler için müşteri tipi dağılımı

MÜŞTERİ TİPİ	FREKANS					
	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
B	164	17,58	54,49	137	15,09	45,51
C	65	6,97	63,11	38	4,185	36,89
D	16	1,71	57,14	12	1,322	42,86
G	6	0,64	42,86	8	0,881	57,14
J	50	5,36	32,05	106	11,67	67,95
M	85	9,11	37,44	142	15,64	62,56

MÜŞTERİ TİPİ	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
N	401	42,98	57,7	294	32,38	42,3
O	2	0,21	66,67	1	0,11	33,33
R	136	14,58	45,03	166	18,28	54,97
T	8	0,86	66,67	4	0,441	33,33
Toplam	933	100	50,68	908	100	49,32

Çizelge 4.3 incelendiğinde; J kampanyasıyla kazanılan müşteriler %67,95 oranıyla kötü ve %32,05 oranıyla iyi müşteridirler. Bu da J (Özel müşteriler) kampanyasının başarısı hakkında soru işareti oluşturmaktadır. Bu kampanyada kazanılan müşteriler büyük oranda ödeme problemi yaratan müşteriler olmuştur. C kampanyasıyla kazanılan müşteriler ise %63,11 oranı ile iyi müşterilerdir. Bundan dolayı C kampanyasını başarılı bir kampanya olduğu söylenebilir.

Çizelge 4.4 İyi-kötü müşteriler için cinsiyet dağılımı

CİNSİYET	FREKANS					
	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
ERKEK	856	91,75	50,8	829	91,3	49,2
KADIN	76	8,15	49,35	78	8,59	50,65

Çizelge 4.4 incelendiğinde; kredi kartı müşterilerinin büyük oranda erkek müşterilerden oluştuğu görülmektedir. İyi ve kötü müşteriler içerisinde kadın-erkek oranları birbirine yakındır. Bu da cinsiyetin iyi müşteri-kötü müşteri ayırımında belirleyici bir etkisi olmadığı şeklinde yorumlanabilir.

Çizelge 4.5 İyi-kötü müşteriler için medeni hal dağılımı

MEDENİ HAL	FREKANS					
	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
EVLİ	677	72,56	51,6	635	69,93	48,4
DUL	1	0,11	20	4	0,44	80
BEKAR	254	27,22	48,66	268	29,52	51,34
Toplam	932	99,89	50,68	907	99,89	49,32

Çizelge 4.5 incelendiğinde;müşterilerin evli, dul, bekar olmalarının iyi-kötü müşteri ayrımı

yapılmasında etkili bir bilgi olmadığı söylenebilir.

Çizelge 4.6 İyi-kötü müşteriler için öğrenim durumu dağılımı

ÖĞRENİM DURUMU	FREKANS					
	İYİ	İyiler içerisinde %	%	KÖTÜ	Kötüler içerisinde %	%
Eğitim almamış	464	49,73	54,59	386	42,51	45,41
İlkokul	1	0,11	25	3	0,33	75
Ortaokul	348	37,3	50	348	38,33	50
Ön lisans	106	11,36	42,91	141	15,53	57,09
Üniversite	13	1,39	33,33	26	2,86	66,67
Yüksek Lisans/Doktora	1	0,11	50	1	0,11	50
Toplam	933	100	50,68	908	100	49,32

Çizelge 4.6 incelendiğinde; farklı öğrenim durumu seviyeleri için iyi-kötü müşteri oranlarının yakın olduğu gözlemlenmektedir. Öğrenim durumunun ayırıcı bir özellik taşımadığı söylenebilir.

Çizelge 4.7 İyi-kötü müşteriler için araba sahibi dağılımı

ARABA SAHİBİ	FREKANS					
	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
VAR	108	11,58	42,35	147	16,18	57,65
YOK	825	88,42	52,02	761	83,81	47,98
Toplam	933	100	50,68	908	100	49,32

Çizelge 4.7 incelendiğinde; araba sahibi olma kriterinin de ayırıcı bir özellik taşımadığı söylenebilir.

Çizelge 4.8 İyi-kötü müşteriler için başka banka kredi kartı dağılımı

BAŞKA BANKA KREDİ KARTI	FREKANS					
	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
1 TANE VAR	105	11,25	50,48	103	11,34	49,52
2 TANEVAR	80	8,57	50	80	8,81	50
YOK	748	80,17	50,78	725	79,84	49,22
Toplam	933	100	50,68	908	100	49,32

Çizelge 4.8 ,incelendiğinde; Müşterilerin başka bankalara ait kredi kartlarının olup olmaması iyi müşteri-kötü müşteri ayırımı için anlamlı gözükmemektedir.

Çizelge 4.9 İyi-kötü müşteriler için ek kart durumu dağılımı

EK KART	FREKANS					
	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
VAR	22	2,36	50	22	2,42	50
YOK	911	97,64	50,7	886	97,57	49,3
Toplam	933	100	50,68	908	100	49,32

Çizelge 4.9 incelendiğinde; ek kart durumunun da müşterileri iyi müşteri-kötü müşteri olarak ayırmada etkili bir kriter olmadığı söylenebilir.

Çizelge 4.10 İyi-kötü müşteriler için çalışma şekli dağılımı

ÇALIŞMA ŞEKLİ	FREKANS					
	İYİ	İyiler İçerisinde %	%	KÖTÜ	Kötüler İçerisinde %	%
Diğer	264	28,3	53,66	228	25,11	46,34
Emekli	102	10,93	62,96	60	6,61	37,04
Ev hanımı	446	47,8	46,07	522	57,49	53,93
Kamu sektörü	46	4,93	85,19	8	0,88	14,81
Serbest meslek	3	0,32	75	1	0,11	25
Öğrenci	70	7,5	98,59	1	0,11	1,41
Toplam	931	99,79	92,18	79	8,7	7,82

Çaslışma şekli frekans tablosu incelendiğinde;ev hanımı olan müşterilerin kötü müşteri olma eğilimlerinin, öğrenci ve kamu sektörü çalışanlarının ise iyi müşteri olma eğilimlerinin yüksek olduğu görülmektedir. Öğrenciler ise %98,59 oranıyla iyi, %1,41 oranıyla kötü müşteriler içerisinde yer almaktadırlar. Buna göre öğrencilerin büyük oranda iyi müşteri olma eğilimi gösterdikleri söylenebilir.

4.2 Kredi Kartlarında Diskriminant Analizi ile Skorlama

Kredi kartı müşterilerini iyi müşteriler-kötü müşteriler olarak ayırmak için diskriminant analizi uygulanmıştır.

Çizelge 4.11 Kayıp veriler

VERİLER	N	Yüzde
Geçerli Veri	1426	79
Kayıp Veri	379	21
Toplam	1805	100

Kredi kartı müşterileri için BAŞKA BANKAYA AİT KART LİMİTİ, AYLIK NET GELİR, MÜŞTERİ YAŞI değişkenlerine ait 1805 gözlem değeri ile diskriminant analizi yapılmak istenmiştir. Tabloya bakıldığında; gözlemler içinde grup kodu (iyi–kötü) tanımlı olmayan hiçbir gözlem değeri bulunmadığı, 379 gözlem değeri için değişkenlerden en az birinde kayıp veri bulunduğu görülmektedir. Hem grup kodu tanımlı olmayan hem de en az bir değişken değerinin eksik olduğu hiçbir gözlem değeri yoktur. SPSS 13'de iyi–kötü olarak tanımlanan bu gözlem değerlerine uygulanan diskriminant analizi sonuçları incelenmiştir.

Çizelge 4.12 Değişkenler kovaryans-korelasyon matrisi

Kovaryans/Korelasyon Matrisi		MÜŞTERİ YAŞI	AYLIK NET GELİR	BAŞKA BANKA KART LİMİTİ
Kovaryans	MÜŞTERİ YAŞI	98,46	2961,3	3096,28
	AYLIK NET GELİR	2961,3	10596179	4251575,12
	BAŞKA BANKA KART LİMİTİ	3096,28	4251575,12	12293630,3
Korelasyon	MÜŞTERİ YAŞI	1	0,09	0,08
	AYLIK NET GELİR	0,091	1	0,37
	BAŞKA BANKA KART LİMİTİ	0,08	0,37	1

Çizelge 4.12 incelendiğinde; modele giren değişkenler arasında yüksek korelasyon olmadığı gözlemlenmektedir.

Değişken yapıları doğrultusunda, iyi müşteriler ve kötü müşterilere ait kovaryans matrisi;

Çizelge 4.13 İyi-kötü müşteriler için kovaryans matrisi

İYİ-KÖTÜ		MÜŞTERİ YAŞI	AYLIK NET GELİR	BAŞKA BANKA KART LİMİTİ
İYİ	MÜŞTERİ YAŞI	100,08	1.378,75	1.089,44
	AYLIK NET GELİR	1.378,75	5.684.511,24	3.455.191,36

İYİ-KÖTÜ		MÜŞTERİ YAŞI	AYLIK NET GELİR	BAŞKA BANKA KART LİMİTİ
	BAŞKA BANKA KART LİMİTİ	1.089,44	3.455.191,36	14.852.727,08
KÖTÜ	MÜŞTERİ YAŞI	96,97	4.415,06	4.939,78
	AYLIK NET GELİR	4.415,06	15.108.059,89	4.983.136,96
	BAŞKA BANKA KART LİMİTİ	4.939,78	4.983.136,96	9.942.832,18
TOPLAM	MÜŞTERİ YAŞI	98,49	3.342,42	3.447,59
	AYLIK NET GELİR	3.342,42	11.037.280,51	4.676.189,11
	BAŞKA BANKA KART LİMİTİ	3.447,59	4.676.189,11	12.660.979,70

Kovaryans matrislerinin homojenliğine bakarken iki grupta da ilgili kovaryans matrislerinin çok farklılık gösterip göstermediğini incelenir. İlgili değişken çiftlerinin iki grupta da farklı kovaryans değerlerine sahip olması kovaryans matrislerinin homojen olmadığı şeklinde yorumlanabilir.

Çizelge 4.14 Box-M test sonucu

Box's M Testi		51,45
F değeri	Serbestlik Derecesi	6
	Serbestlik Derecesi 2	774036,34
	Anlamlılık Seviyesi	0,00

Box's M testi "H_0: Gruplararası kovaryans matrisleri homojendir" hipotezini test etmek için kullanılır. Test sonucu elde edilen p (Sig)=0,000 belirlenen anlamlılık seviyesinden (0.05) küçük olduğundan dolayı H_0 hipotezi reddedilebilir. Yani gruplar arası kovaryans matrisi homojen değildir ve diskriminant analizi varsayımlarından biri gerçekleşmemiş olur. Bu durumda homojenlik varsayımına gerek duymayan Karesel Diskriminant Analizi uygulanarak sonuçlar aşağıda incelenmiştir.

Çizelge 4.15 Grup ortalamaları eşitlik testi

Grup Ortalamaları Eşitlik Testi	Wilks' Lambda	F	Sd. 1	Sd. 2	Anlamlılık seviyesi (p)
MÜŞTERİ YAŞI	1,00	1,08	1	330	0,03
AYLIK NET GELİR	0,96	14,78	1	330	0,00
BAŞKA BANKA KART LİMİTİ	0,97	10,89	1	330	0,00

Wilk's Lambda çok değişkenli bir anlamlılık testidir ve *U* olarak da adlandırılır. Wilk's Lambda değeri 0 ile 1 arasında değişmektedir. Küçük değerler ilgili değişkenin grupları ayırmada güçlü bir etkisi olduğunu, büyük değerler ise ilgili değişkenin grupları ayırmada daha az etkili olduğunu göstermektedir. Diğer bir istatistik olan F değeri için hesaplanan p belirlenen anlamlılık seviyesinden (0,05) küçük olduğu için AYLIK NET GELİR, ÖĞRENİM DURUMU, MÜŞTERİ TİPİ değişkenlerinin iyi müşterileri kötü müşterilerden ayırmada kullanılan modele anlamlı katkısı olduğu söylenebilir.

Çizelge 4.16 Özdeğer

Özdeğer				
Fonksiyon	Özdeğer	Varyans %	Kümülatif %	Kanonik Korelasyon
1	0,06	100,00	100,00	0,23

Çizelge 4.16 her bir kanonik değişkene ait öz değerin varyansın ne kadarını açıkladığını, kümülatif açıklanan varyansı, ve kanonik korelasyon değerini göstermektedir. Tablo değerlerine bakıldığında varyansın tek bir kanonik değişkenle %100 oranında açıklandığı görülmektedir. Özdeğer, gruplar arası kareler toplamının grup içi kareler toplamına oranıdır. Kanonik korelasyon değeri de gruplarla diskriminant analizi skorları arasındaki başarıyı göstermektedir. 1'e yakın kanonik korelasyon değerleri gruplar ve diskriminant skorları arasında güçlü bir ilişkiyi göstermektedir. Buradaki kanonik korelasyon değeri incelendiğinde modelin tahmin başarısının düşük olduğu söylenebilir.

Çizelge 4.17 Wilks' Lambda değeri

Fonksiyon Testi	Wilks' Lambda	Ki-kare Değeri	Serbestlik Derecesi	Anlamlılık Seviyesi
1	0,95	18,54	3,00	0,00

"H_o: Fonksiyonların gruplar arası ortalamaları eşittir" varsayımını test eder. Wilk's Lambda değeri gruplar arasındaki fark ile açıklanamayan diskriminant fonksiyonlarının toplam varyansının oranını göstermektedir ve 0 ile 1 arasında değişmektedir. 0'a yakın değerler grup ortalamalarının farklı olduğunu, 1'e yakın değerler ise grup ortalamaları arasında fark olmadığını göstermektedir. Wilk's Lambda'nın chi-squre dönüşümü verilen serbestlik derecesinde anlamlılığı test eder. Anlamlılık seviyesi (p) küçük ise grup ortalamaları farklıdır, büyük ise de farklı değildir diye yorumlanabilir. Elde edilen anlamlılık seviyesi belirlenen anlamlılık düzeyinden (0,05) küçük olduğu için grupların ortalamaları farklıdır denilebilir.

Çizelge 4.18 Standardize edilmiş kanonik diskriminant fonksiyonu katsayısı

Standardize Edilmiş Kanonik Diskriminant Fonksiyonu Katsayısı	Fonksiyon
MÜŞTERİ YAŞI	0,13
AYLIK NET GELİR	0,68
BAŞKA BANKA KART LİMİTİ	0,48

Değişkenler farklı ölçüm birimleriyle ölçüldükleri için standardize edilmiş kanonik diskriminant fonksiyonu katsayıları hesaplanır. Diskriminant modelindeki ayırıcı değişkenlerin bağımlı değişken üzerindeki göreceli etkisini gösteren standardize ayırma fonksiyonu katsayılarına göre, gruplar arasındaki en fazla ayırıcı etkide bulunan değişkenlerin önem sırası; AYLIK NET GELİR , BAŞKA BANKA KART LİMİTİ, MÜŞTERİ YAŞI biçimindedir.

Skor=+0,0,68*AYLIK NET GELİR+0,13*MÜŞTERİ YAŞI+0,48*BAŞKA BANKA KART LİMİTİ

şeklinde ifade edilebilir.

Çizelge 4.19 Yapı Matrisi

	Fonksiyon
Yapı Matrisi	1
AYLIK NET GELİR	0,878
BAŞKA BANKA KART LİMİTİ	0,754
MÜŞTERİ YAŞI	0,238

Her bir tahmin edici değişkenin kanonik fonksiyonla grup içi korelasyonlarını içerir. Diskriminant fonksiyonunda her bir değişkenin katkısını ölçmenin bir yoludur. Diskriminant fonksiyonuna katkı sırasına göre AYLIK NET GELİR 0,878 ile en büyük katkıyı yapan değişkendir ve MÜŞTERİ YAŞI 0,238 ile an az katkıyı yapan değişkendir.

Çizelge 4.20 İyi-kötü müşteriler için diskriminant fonksiyonu katsayıları

Sınıflama Fonksiyonu Katsayıları	İYİ-KÖTÜ	
	İyi	Kötü
MÜŞTERİ YAŞI	0,35	0,36
AYLIK NET GELİR	0,00	0,00
BAŞKA BANKA KART LİMİTİ	0,00	0,00
Sabit	-6,98	-7,65

İyi ve kötü müşteri gruplarının sınıflama fonksiyonlarını gösterir.

Çizelge 4.21 Sınıflandırma

Sınıflandırma Sonuçları		İYİ-KÖTÜ	Tahmin		Toplam
			İyi	Kötü	
Orijinal	Sayı	İyi	820	80	900
		Kötü	731	174	905
	%	İyi	91,11	8,89	100
		Kötü	80,77	19,23	100

Bu modelde 900 iyi müşterilerden 820 tanesi (%91,1) iyi olarak sınıflandırılmış, 80 (%8,9) tanesi kötü olarak sınıflandırılmıştır. 905 tane kötü gözlemden 174 tanesi (%19,23) kötü olarak sınıflandırılmış, 731 tanesi (%80,7) iyi olarak sınıflandırılmıştır.

4.2.1 Sonuç

Kredi kartları verileriyle skorlamada diskriminant analizi için modele AYLIK NET GELİR, MÜŞTERİ YAŞI, BAŞKA BANKA KART LİMİTİ değişkenleri girmiştir. Analiz SPSS 13

programında yapılmıştır. Gruplararası kovaryans matrisi homojenliği sağlanamamıştır ve bu nedenle Karesel Diskriminant Analizi uygulanmıştır. Modele giren değişkenlerden modele en yüksek katkıyı AYLIK NET GELİR değişkeninin yaptığı görülmektedir.Varyans tek bir kanonik değişkenler %100 oranında açıklanmıştır ve modelin tahmin başarısı düşüktür. İyi müşteriler %91,11 oranı ile, kötü müşteriler ise %19,23 oranı ile tahmin edilebilmektedirler. Model kötü müşterileri tahmin etmede başarılı değildir.

4.3 Kredi Kartlarında Lojistik Regresyon ile Skorlama

Kredi skorlamada lojistik regresyon uygulamasının amacı, bağımlı değişkendeki değişimi (varyasyonu) en iyi açıklayan yada bağımlı değişkenin çeşitli düzeylerini birbirinden ayırt etmede etkili olabilecek bağımsız değişkenlerin seçimidir (Akkuş vd., 2005). Bu amaçla kredi kartı müşterilerine ait 1806 gözlemden on üç bağımsız değişken MÜŞTERİ TİPİ, ÇALIŞMA ŞEKLİ, ÖĞRENİM DURUMU, AYLIK NET GELİR, MAAŞ, MEDENİ HAL, BAŞKA BANKA KREDİ KARTI, BAŞKA BANKA KART LİMİTİ, CİNSİYET, İKAMET SÜRESİ, ARABA SAHİBİ, YAŞ ve İYİ-KÖTÜ bağımlı değişkeni modele alınarak bu değişkenlere ait odds oranlarının olasılık güven aralıkları, Wald istatistiği olasılığı, değişkenlere ilişkin regresyon katsayılarının serbestlik dereceleri önemlilik düzeyleri incelenmiştir. Söz konusu p değeri 0,25'ten küçük olan değişkenler çok değişkenli lojistik regresyon modeline alınmış, kategorik değişkenlerin frekans değerleri Çizelge 4.22'de incelenmiştir.

Çizelge 4.22 Kategorik değişken kodları

Kategorik Değişken Kodları		Frekans
MÜŞTERİ TİPİ	B	116
	C	94
	D	25
	G	8
	J	106
	M	210
	N	565
	R	295
	T	7
ÇALIŞMA ŞEKLİ	1	460
	2	151
	3	667
	4	47
	5	2
	6	1
	7	98
ÖĞRENİM DURUMU	1	581
	2	3
	3	588
	4	2

		Frekans
	5	217
	6	35
AYLIK NET GELİR	1	300
	2	286
	3	264
	4	301
	5	275
MAAŞ	1	282
	2	287
	3	265
	4	306
	5	286
MEDENİ HALİ	EVLİ	973
	DUL	5
	BEKAR	448
BAŞKA BANKA KREDİ KARTI	1 TANE VAR	194
	2 TANE VAR	141
	YOK	1091
BAŞKA BANKA KART LİMİTİ	2	1083
	4	8
	5	335
YAŞ	1	489
	2	490
	3	447
CİNSİYET	ERKEK	1307
	KADIN	119
İKAMET SÜRESİ	3	1304
	5	122
ARABA SAHİBİ	1	221
	2	1205
EK KART DURUMU	VAR	31
	YOK	1395

Çizelge 4.23 Modele giren değişkenler

Modeldeki Değişkenler		B	Standart Hata	Wald İstatistiği	Serbestlik Derecesi	Anlamlılık Düzeyi	Exp(B)
Adım 1(a)	MÜŞTERİ TİPİ			57,51	8	**0**	
	MÜŞTERİ TİPİ(1)	0,22	1,14	0,04	1	0,85	1,24
	MÜŞTERİ TİPİ(2)	0,33	1,15	0,08	1	0,77	1,39
	MÜŞTERİ TİPİ(3)	1,2	1,2	1	1	0,32	3,33
	MÜŞTERİ TİPİ(4)	1,95	1,37	2,02	1	**0,15**	7
	MÜŞTERİ TİPİ(5)	1,84	1,15	2,56	1	**0,11**	6,32
	MÜŞTERİ TİPİ(6)	1,71	1,13	2,28	1	**0,13**	5,53
	MÜŞTERİ TİPİ(7)	0,74	1,13	0,43	1	0,51	2,1
	MÜŞTERİ TİPİ(8)	1,26	1,14	1,22	1	0,27	3,52
	CİNSİYET(1)	-0,24	0,22	1,19	1	0,28	0,78
	MEDENİ HAL			1,25	2	0,54	
	MEDENİ HAL(1)	-0,14	0,15	0,86	1	0,35	0,87
	MEDENİ HAL(2)	0,65	1,18	0,3	1	0,58	1,91
	ÖĞRENİM DURUMU			18,34	5	**0**	
	ÖĞRENİM DURUMU(1)	-1,49	0,45	11,09	1	**0**	0,23
	ÖĞRENİM DURUMU(2)	20,42	22.306,25	0	1	1	

	Modeldeki Değişkenler	B	Standart Hata	Wald İstatistiği	Serbestlik Derecesi	Anlamlılık Düzeyi	Exp(B)
	ÖĞRENİM DURUMU(3)	-1,18	0,44	7,11	1	**0,01**	0,31
	ÖĞRENİM DURUMU(4)	-0,77	1,57	0,24	1	0,62	0,46
	ÖĞRENİM DURUMU(5)	-0,89	0,46	3,82	1	**0,05**	0,41
	ARABA SAHİBİ(1)	0,49	0,18	7,98	1	**0**	1,64
	BAŞKA BANKA KREDİ KARTI			0,69	2	0,71	
	BAŞKA BANKA KREDİ KARTI(1)	0,49	0,63	0,6	1	0,44	1,63
	BAŞKA BANKA KREDİ KARTI(2)	0,39	0,64	0,36	1	0,55	1,47
	EK KART DURUMU(1)	0,2	0,41	0,25	1	0,62	1,23
	ÇALIŞMA ŞEKLİ			6,06	6	0,42	
	ÇALIŞMA ŞEKLİ(1)	0,27	0,34	0,67	1	0,41	1,32
	ÇALIŞMA ŞEKLİ(2)	0,24	0,38	0,41	1	0,52	1,27
	ÇALIŞMA ŞEKLİ(3)	0,22	0,31	0,52	1	0,47	1,25
	ÇALIŞMA ŞEKLİ(4)	-0,84	0,56	2,24	1	**0,13**	0,43
	ÇALIŞMA ŞEKLİ(5)	0,33	1,94	0,03	1	0,86	1,39
	ÇALIŞMA ŞEKLİ(6)	21,14	40.192,97	0	1	1	
	YAŞ			0,32	2	0,85	
	YAŞ(1)	-0,07	0,17	0,16	1	0,69	0,93
	YAŞ(2)	-0,08	0,15	0,3	1	0,58	0,92
	MAAŞ			18,41	4	**0**	
	MAAŞ(1)	-0,98	0,51	3,7	1	**0,05**	0,37
	MAAŞ(2)	0,25	0,44	0,32	1	0,57	1,28
	MAAŞ(3)	0,5	0,41	1,52	1	**0,22**	1,65
	MAAŞ(4)	0,46	0,35	1,71	1	**0,19**	1,58
	AYLIK NET GELİR			12,14	4	**0,02**	
	AYLIK NET GELİR(1)	-1,37	0,49	7,92	1	**0**	0,25
	AYLIK NET GELİR(2)	-1,29	0,43	8,92	1	**0**	0,27
	AYLIK NET GELİR(3)	-1,34	0,41	10,61	1	**0**	0,26
	AYLIK NET GELİR(4)	-0,72	0,35	4,15	1	**0,04**	0,49
	BAŞKA BANKA KART LİMİTİ			0,43	2	0,81	
	BAŞKA BANKA KART LİMİTİ(1)	0,31	0,62	0,25	1	0,62	1,36
	Sabit	0,77	1,43	0,29	1	0,59	2,16

Çizelge 4.23 incelendiğinde; çok değişkenli lojistik regresyon çözümlemesi sonucunda, değişkenlerin tamamının eleme tekniği kullanılmadan modele alındığı görülmektedir. Bu süreçte Hosmer ve Lemeshow'un önermiş olduğu $p<0,25$ önemlilik düzeyi dikkate alındığında, bağımsız değişkenlerden AYLIK NET GELİR, MAAŞ, ÇALIŞMA ŞEKLİ, ARABA SAHİBİ, ÖĞRENİM DURUMU, MÜŞTERİ TİPİ değişkenlerinin denkleme önemli katkılarda bulunduğu ve bu nedenle denkleme alınması gerektiğine karar verilir. Önemli bulunan 6 değişken dışındaki değişkenlere ait katsayılar dikkate alınan $p<0,25$ ölçütüne uymadıkları için, denkleme alınmazlar (Akkuş vd., 2005). İstatistiksel açıdan önemsiz bulunan bu değişkenlerin denkleme katkılarının olmadığı söylenebilir. Önemli görülen değişkenlerle tekrar model kurulduğunda;

İkili Lojistik Regresyon Hosmer-Lemeshow uyum iyiliği istatistiği;

Çizelge 4.24 Hosmer & Lemeshow istatistiği

Hosmer and Lemeshow Testi			
Adım	Chi-square	Serbestlik Derecesi	Anlamlılık Düzeyi.
1	5,87	8	0,66

Uyum iyiliği istatistiği modelin verilere tatmin edici şekilde uyum sağlayıp sağlamadığına karar vermede kullanılır. Hosmer-Lemeshow istatistiği anlamlılık seviyesi 0,05 altında ise zayıf bir uyumu göstermektedir. Burada anlamlılık seviyesi 0,66 çıktığı için modelin tatmin edici şekilde verilere uyum gösterdiği söylenilebilir.

Daha fazla kontrol için, geriye doğru adımsal lojistik regresyon tekniği kullanılarak model kurulmuştur. Geriye doğru adımsal lojistik regresyon tekniği tüm tahmincileri kullanarak modeli kurmaya başlar. Her adımda modele en az katkıda bulunan tahminci modelden elenir. Modeldeki tüm tahmincilerin anlamlı olduğu noktaya kadar eleme devam eder.

Çizelge 4.25 Sınıflandırma

Sınıflandırma Tablosu	Gözlenen		Tahmin		Tahmin Başarısı
Adım 1			İYİ	KÖTÜ	
	İYİ-KÖTÜ	İYİ	474	239	66,48
		KÖTÜ	197	516	72,37
	%				69,42

Sınıflandırma Tablosu lojistik regresyon modelinin kullanımının pratik sonuçlarını göstermektedir. Her bir değer için, o durumun model tahmin edici olasılığı kesme değerinden daha büyükse tahmin edilen cevap evettir.

- Diagonaldeki hücreler doğru tahminleri gösterir.
- Diagonal dışındaki hücreler yanlış tahminleri gösterir.

Tablo incelendiğinde; gerçekte iyi müşteri olup yine iyi müşteri olarak sınıflanan kişi sayısı 474 (%66,48) ve kötü müşteri olarak sınıflanan kişi sayısı 239'dur. Gerçekte kötü müşteri olup kötü müşteri olarak sınıflanan kişi sayısı 516 (%72,37) ve iyi müşteri olarak sınıflanan kişi sayısı 197'dir. Kötü müşterilerin doğru sınıflandırılma oranı iyi müşterilerin doğru

sınıflandırılma oranından yüksektir. Genel tahmin başarısı ise; %69,42'dir.

Çizelge 4.26 Modele giren değişkenler

Modeldeki Değişkenler		B	Standart Hata	Wald İstatistiği	Serbestlik Derecesi	Anlamlılık Seviyesi	Exp(B)
Adım 1(a)	AYLIK NET GELİR			12,34	4	0,01	-
	AYLIK NET GELİR (1)	-1,37	0,48	8,12	1	0	0,25
	AYLIK NET GELİR (2)	-1,29	0,43	8,99	1	0	0,28
	AYLIK NET GELİR (3)	-1,35	0,41	10,95	1	0	0,26
	AYLIK NET GELİR (4)	-0,74	0,35	4,44	1	0,04	0,48
	MAAŞ			19,01	4	0	-
	MAAŞ (1)	-0,98	0,5	3,78	1	0,05	0,38
	MAAŞ (3)	0,5	0,41	1,51	1	0,22	1,65
	MAAŞ (4)	0,47	0,35	1,86	1	0,17	1,6
	ÇALIŞMA ŞEKLİ			6,03	6	0,42	-
	ÇALIŞMA ŞEKLİ (1)	0,3	0,33	0,82	1	0,36	1,35
	ÇALIŞMA ŞEKLİ (2)	0,21	0,37	0,31	1	0,58	1,23
	ÇALIŞMA ŞEKLİ (3)	0,23	0,31	0,56	1	0,45	1,26
	ÇALIŞMA ŞEKLİ (4)	-0,77	0,55	1,96	1	0,16	0,46
	ÇALIŞMA ŞEKLİ (5)	0,63	1,97	0,1	1	0,75	1,87
	ÇALIŞMA ŞEKLİ (6)	21,07	40.192,97	0	1	1	-
	ARABA SAHİBİ(1)	0,5	0,17	8,25	1	0	1,65
	ÖĞRENİM			20,76	5	0	-
	ÖĞRENİM (1)	-1,53	0,44	12,01	1	0	0,22
	ÖĞRENİM (2)	20,38	22.263,16	0	1	1	-
	ÖĞRENİM (3)	-1,2	0,44	7,53	1	0,01	0,3
	ÖĞRENİM (5)	-0,89	0,45	3,94	1	0,05	0,41
	MÜŞTERİ TİPİ			57,54	8	0	-
	MÜŞTERİ TİPİ (1)	0,21	1,14	0,03	1	0,86	1,23
	MÜŞTERİ TİPİ (2)	0,32	1,15	0,08	1	0,78	1,37
	MÜŞTERİ TİPİ (3)	1,12	1,2	0,87	1	0,35	3,07
	MÜŞTERİ TİPİ (4)	1,9	1,37	1,95	1	0,16	6,72
	MÜŞTERİ TİPİ (5)	1,78	1,15	2,39	1	0,12	5,94
	MÜŞTERİ TİPİ (6)	1,66	1,13	2,15	1	0,14	5,28
	MÜŞTERİ TİPİ (7)	0,71	1,13	0,4	1	0,53	2,04
	MÜŞTERİ TİPİ (8)	1,21	1,13	1,14	1	0,28	3,36
	Sabit	0,86	1,24	0,47	1	0,49	2,35

Maaşı 1000-1900 YTL arası olan müşterilerin maaşı 750-1000 YTL arası olan müşterilere göre kötü müşteri olma riski 1,65 kat daha fazladır. Ve maaşı 1900-3000 YTL arası olan müşterilerin maaşı 1000-1900 YTL arası olan müşterilere göre kötü müşteri olma riski 1,6 kat daha fazladır.

Müşteri tipi G (Grup müşterileri) olan müşterilerin D (Mektup 2 kampanyası) müşterilerine göre kötü müşteri olma riski 6,72 kat daha fazladır. Müşteri tipi J (Özel müşteriler) olan müşterilerin G (Grup müşterileri) müşterilerine göre kötü müşteri olma riski 5,94 kat daha fazladır. Müşteri tipi M (Mail Order) olan müşterilerin J (Özel müşteriler) müşterilerine göre

kötü müşteri olma riski 5,28 kat daha fazladır. MÜŞTERİ değişkeninin mevcut veri yapısı ve lojistik regresyon analizi için etkili bir değişken olduğu söylenebilir.

4.3.1 Sonuç

Kredi kartı müşterilerine ait verilerle lojistik regresyon modeli kurulduğunda modele MÜŞTERİ TİPİ, ÇALIŞMA ŞEKLİ, ÖĞRENİM DURUMU, AYLIK NET GELİR, MAAŞ, MEDENİ HAL, BAŞKA BANKA KREDİ KARTI, BAŞKA BANKA KART LİMİTİ, CİNSİYET, İKAMET SÜRESİ, ARABA SAHİBİ değişkenleri girmiş, $p<0,25$ önemlilik düzeyi dikkate alındığında, bağımsız değişkenlerden AYLIK NET GELİR, MAAŞ, ÇALIŞMA ŞEKLİ, ARABA SAHİBİ, ÖĞRENİM DURUMU, MÜŞTERİ TİPİ değişkenlerinin denkleme önemli katkılarda bulunduğu ve bu nedenle denkleme alınması gerektiğine karar verilmiştir. Hosmer & Lemeshow istatistiği 0,66 ile modelin verilere uyumunun iyi olduğunu göstermektedir. Lojistik regresyon modeline göre iyilerin doğru tahmin edilme oranı %66,48, kötülerin doğru tahmin edilme oranı %72,37'dir. Diskriminant analizine göre lojistik regresyonda kötülerin tahmin oranı nispeten yüksektir. İyilerin doğru tahmin edilme oranı ise %3 düşmüştür. Lojistik modelin genel tahmin başarısı % 69,42 ile diskriminant analizine göre (%63) daha yüksek çıkmıştır. Analizler SPSS 13'te yapılmıştır.

4.4 Kredi Kartlarında En Yakın Komşu Tekniği ile Skorlama

Kredi kartı müşterilerine ait 1806 gözlemden MÜŞTERİ TİPİ, ÇALIŞMA ŞEKLİ, ÖĞRENİM DURUMU, AYLIK NET GELİR, MAAŞ, MEDENİ HAL, BAŞKA BANKA KREDİ KARTI, BAŞKA BANKA KART LİMİTİ, CİNSİYET, İKAMET SÜRESİ, ARABA SAHİBİ, YAŞ değişkenlerine ait son beş veri (1802,1803,1804,1805,1806) X1, X2, X3, X4, X5 olarak tanımlanarak en yakın komşularının iyi yada kötü olmalarına göre sınıflandırılmaya çalışılmıştır. Verilere aşamalı kümeleme tekniklerinden en yakın komşu tekniği uygulanmıştır ve uzaklık ölçü birimi olarak öklityen uzaklık seçilmiştir. Analizler SPSS 13'te yapılmıştır.

X1 için dendogramda bulunan sonuçta en yakın komşular;

Çizelge 4.27 X1 müşterisinin sınıflandırması

1	379
1	930
1	1259
X1	1801

1 725

1259, 930, 379, 725 inci durumlardır. Bu komşular da 1 (iyi) koduyla kodlanmış oldukları için X1'in iyi olduğu yönünde karar alabiliriz ki gerçekte de X1 iyi bir müşteridir.

X2 için dendogramda bulunan sonuçta en yakın komşular;

Çizelge 4.28 X2 müşterisinin sınıflandırması

2 1686
2 1116
2 1506
2 449
2 1612
2 1748
2 959
X2 1802
1 1606
1 348
2 1475
2 1649
2 1160
2 845

1606, 348, 1475, 1649, 959, 1748 inci durumlardır. Bu komşular da hem 2 (kötü) koduyla hem de 1 (iyi) koduyla kodlanmış oldukları için X2'nin durumu için net bir yorum yapılamaz. Ama iki tane iyi kodlanmış komşu dışındakiler kötü kodlanmış olduğu için kötü olmaya eğilimli olduğunu söyleyebiliriz ki gerçekte de X2 müşterisi kötü bir müşteridir.

X3 için dendogramda bulunan sonuçta en yakın komşular;

Çizelge 4.29 X3 müşterisinin sınıflandırması

2 1788
2 740
2 542
X3 1803
2 551

2	1233
2	1542

551, 1233, 1542, 542, 740, 1788 inci durumlarıdır. Bu komşular da 2 (kötü) koduyla kodlanmış oldukları için X3'ün kötü olduğu yönünde karar alabiliriz ki gerçekte de X3 kötü bir müşteridir.

X4 için dendogramda bulunan sonuçta en yakın komşular;

Çizelge 4.30 X4 müşterisinin sınıflandırması

2	182
2	1747
2	1751
2	1162
X4	1804
2	1469
2	1653

1469, 1653, 1162, 1751, 1747, 182 inci durumlardır. Bu komşular da kötü koduyla kodlanmış oldukları için X4'ün kötü olduğu yönünde karar alabiliriz ki gerçekte de X4 kötü bir müşteridir.

X5 için dendogramda bulunan sonuçta en yakın komşular;

Çizelge 4.31 X5 müşterisinin sınıflandırması

2	1051
X5	1805
2	1344
2	1625
2	534
2	1423
2	1258
2	1584

1344, 1625, 534, 1423, 1258, 1584, 1051 inci durumlarıdır. Bu komşular da kötü koduyla kodlanmış oldukları için X5'in kötü olduğu yönünde karar alabiliriz ama gerçekte X5 kötü değil iyi bir müşteridir.

4.4.1 Sonuç

Kredi kartı müşterilerine ait MÜŞTERİ TİPİ, ÇALIŞMA ŞEKLİ, ÖĞRENİM DURUMU, AYLIK NET GELİR, MAAŞ, MEDENİ HAL, BAŞKA BANKA KREDİ KARTI, BAŞKA BANKA KART LİMİTİ, CİNSİYET, İKAMET SÜRESİ, ARABA SAHİBİ, YAŞ değişkenleriyle yapılan kümeleme analizi sonucunda en yakın komşular incelenmiş, ilgili kişilerin yüksek oranda doğru sınıflandıkları gözlemlenmiştir. Burada en yakın komşu sayısı k seçiminde gerekli olan K ve M değerleri bilinmediği için rasgele belirlenen sayıda yakın komşulara bakılarak tahminlerin doğru olup olmadığı hakkında yorumda bulunulmuştur. Uygulama hakkında genel bilgi olması açısından bu teknik izlenmiştir. Fakat en yakın komşu teorisine göre k sayısı belirlenmeden bu tekniğin uygulanması yetersizdir. Sonuçların yorumlanmasında bu eksiklik göz önünde bulundurulmalıdır.

4.5 Kredi Kartlarında Yapay Sinir Ağları ile Skorlama

Çalışmada, bir Geri Yayılma Yapay Sinir Ağı kullanılmaktadır. Geri yayılma ağlar, çok tabakalı perceptron ile aynı yapıya sahiptirler ve öğrenme tekniği olarak geri yayılma algoritması kullanırlar. Dolayısıyla, bu ağlar ileri besleme ağlar sınıfına girmektedirler. Ayrıca, çalışmada kullanılan ağ kantitatif verilerle çalışmaktadır ve yönlendirmeli öğrenme tekniği kullanmaktadır. Bu YSA türünün seçilmesinin temel sebebi öngörü ve sınıflandırma işlemleri için oldukça uygun olmasıdır. Diğer bir önemli neden ise doğrusal olmayan yapılar için oldukça kullanışlı olmasıdır. (Yurtoğlu, 2005)

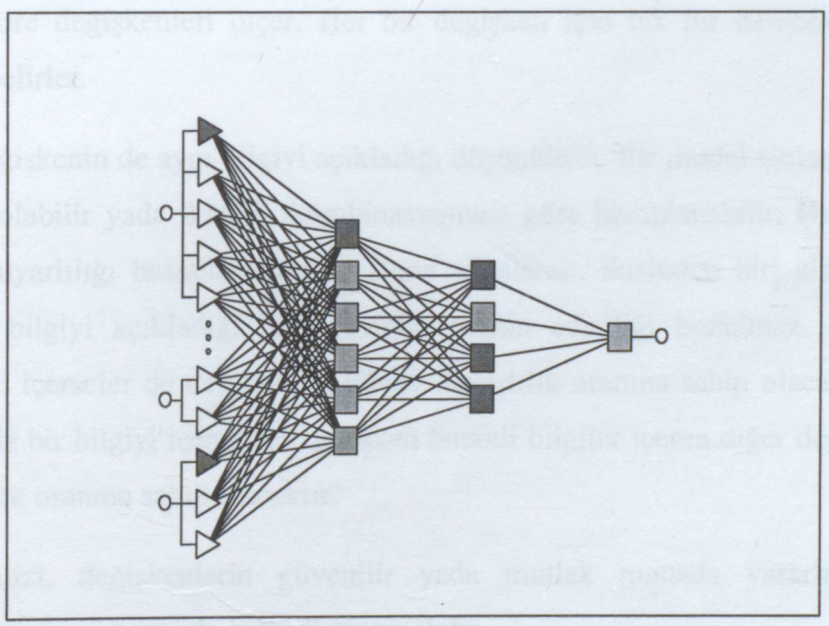
YSA modelinde bağımlı değişken olarak kullanılan iyi-kötü değişkenini sınıflamak için MÜŞTERİ TİPİ, CİNSİYET, MEDENİ HAL, ÖĞRENİM DURUMU, ARABA SAHİBİ, BAŞKA BANKA KREDİ KARTI, EK KART DURUMU, YAŞ, AYLIK NET GELİR, BAŞKA BANKA KREDİ KART LİMİTİ değişkenleri kullanılmıştır. Analizler STATISTICA 7 programında uygulanmıştır.

On girdi değişkeni ve bir çıktı değişkeni içeren bu fonksiyonel yapıyı tahmin etmek için üç tabakalı bir Geri Yayılmalı YSA mimarisi oluşturulmuştur. Şekil 4.1'de sunulan bu mimarinin girdi tabakasında girdi değişkenlerin değerlerinin ağa sunulmasını sağlayan on adet nöron ve çıktı tabakasında ise bağımlı değişkene ait ağ çıktısının alındığı bir adet nöron

bulunmaktadır.

Gizli tabakadaki nöron sayısı için bir kısıt kullanılmamıştır. Bunun yerine, gizli tabakadaki nöron sayısı performans değerlendirmesine göre belirlenmiştir. Değişik sayıda gizli nöronlarla yapılan denemeler sonunda ilk gizli tabakadaki nöron sayısı altı, ikinci gizli tabakadaki nöron sayısı dört olarak belirlenmiştir.

Modelin mimarisi oluşturulduktan sonra eğitme sürecine geçilmiştir. Ağın öğrenme işlemini gerçekleştirdiği bu sürecin başlangıcında, bağlantı ağırlıkları rastsal bir şekilde belirlenmiştir. Eğitme süreciyle birlikte amaç fonksiyonunun (Hata Kareleri Toplamı-SSE) minimize edilebilmesi için bağlantı ağırlıklarının ayarlanması (öğrenme) işlemi gerçekleştirilmiştir. Eğitme süreci için örnekleme seti seçimi kullanılan STATISTICA programına bırakılmıştır. YSA 1000 döngü (epoch) kullanılarak eğitilmiştir. Bağlantı ağırlıklarının ayarlanması, yani eğitme işleminin tamamlanması ile birlikte YSA modelinin tahmin süreci de tamamlanmıştır ve tahmin edilen model simüle edilerek öngörüler alınmıştır.



Şekil 4.1 YSA modeli mimarisi

Çizelge 4.32 Duyarlılık analizi

Duyarlılık Analizi	RASYO	RANK
Maaş	1,006657	1
Aylık Net Gelir	1,002134	2
Yaş	1,001396	3
Başka Banka Kredi Kartı	1,001365	4
Müşteri Tipi	1,001058	5

Duyarlılık Analizi	RASYO	RANK
Başka Banka Kredi Kart Limiti	1,000724	6
Medeni Hal	1,000537	7
Cinsiyet	1,000088	8
Araba Sahibi	1,000083	9
Ek Kart Durumu	1,00006	10
Öğrenim Durumu	0,99949	11

Duyarlılık analizi oluşturulan YSA'da girdi değişkenlerin modele katkılarının önem sırasını göstermektedir. Sadece bilgilendirme amaçlı kullanılmaktadır.

Duyarlılık analizi başlı başına her bir değişkenin gerekliliği konusunda önemli bir bakış açısı kazandırmaktadır. Takip eden analizlerde model dışı tutulacak verileri ve her zaman modelde tutulacak anahtar verileri tanımlar.

Girdi değişkenler genellikle bağımsız değildirler, değişkenler arası bağımlılık mevcuttur. Duyarlılık analizi değişkenler modelde bulunmadığı takdirde model performansındaki kötüleşmeye göre değişkenleri ölçer. Her bir değişken için tek bir derecelendirme değeri (rating value) belirler.

Örneğin, iki değişkenin de aynı bilgiyi açıkladığı düşünülsün. Bir model tamamen birine veya diğerine bağlı olabilir yada ikisinin kombinasyonuna göre hesaplanabilir. Duyarlılık analizi keyfi ilişkili duyarlılığı hesaplamaktadır. Buna ek olarak, ikisinden biri elendiğinde diğer değişken aynı bilgiyi açıkladığı için modelin tatmin ediciliği bozulmaz. Bundan dolayı anahtar bilgileri içerseler de değişkenler düşük duyarlılık oranına sahip olacaklardır. Benzer şekilde, önemsiz bir bilgiyi içeren tek değişken önemli bilgiler içeren diğer değişkenlere göre yüksek duyarlılık oranına sahip olacaktır.

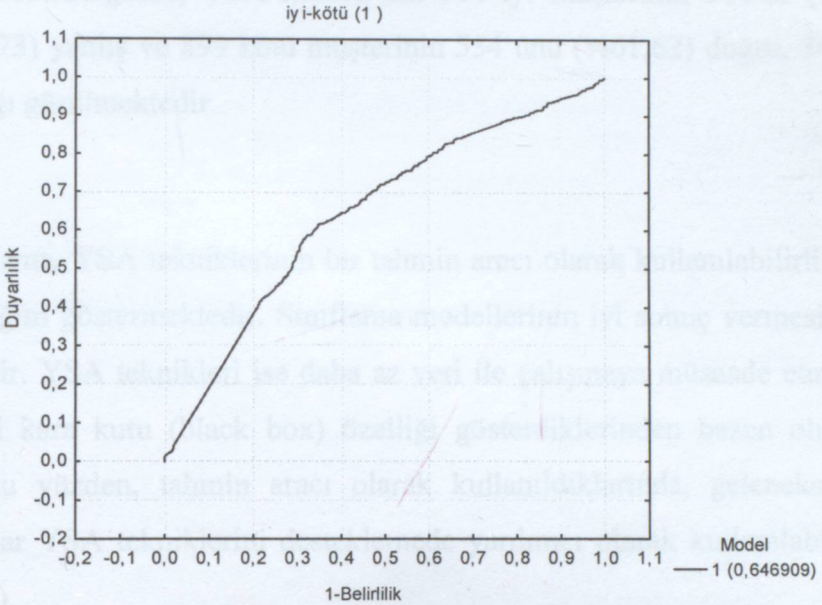
Duyarlılık analizi, değişkenlerin güvenilir yada mutlak manada yararlılığını ölçmez. Değişkenlerin önemi konusunda tedbirli olunmalıdır.

Oran bir yada altındaysa, değişken modelden çıkarıldığında ağın performansına etkisi olmadığı söylenebilir.

Duyarlılık her bir değişken için bir defa hesaplanır ve sıralanır.

Duyarlılık analizi tablosu değişkenlerin önem sırasına göre sıralanmıştır. Tablo incelendiğinde analize en çok katkıyı MAAŞ, en az katkıyı ise ÖĞRENİM DURUMU değişkeninin sağladığı görülmektedir.

Performans değeri çıktı değerlerden yola çıkarak hesaplanır. Regresyon çözümlemesinde performans değeri ölçü birimi standart sapmadır, sınıflama ağ modellerinde ise doğru sınıflanan durumların oranıdır. İlgili modelin öğrenme performansı; 0,665557 olarak hesaplanmıştır. Seçme performansı ise 0,594010 olarak hesaplanmıştır. Test performansı ise 0,628952 olarak hesaplanmıştır.



Şekil 4.2 ROC eğrisi

ROC eğrisi (A Receiver Operating Characteristic Curve) mümkün basamak aralıklarında iki küme sınıflayıcıların performanslarını gösterir. İdeal sınıflayıcılar grafiğin sol kısmında ve üst kısmında, eğrinin 1.0 değerini aldığı yerler altında toplanmışlardır. Rasgele sınıflayıcılar 0.5 (0.5'den küçük alandaki sınıflayıcılar geliştirilebilirler) civarında başarılıdırlar. ROC eğrisi sınıflayıcıları kıyaslamada tavsiye edilir, sadece keyfi olarak seçilen karar basamağındaki performansı değil tüm mümkün karar basamaklarındaki performansı gösterir.

ROC eğrisi optimum basamak kararı seçiminde kullanılır. Bu basamak (her iki sınıfta da yanlış sınıflama oranını eşitleyen) otomatik olarak basamak güven ayarında kullanılabilir.

Çizelge 4.33 Sınıflandırma

Sınıflandırma Tablosu	İYİ	KÖTÜ
Toplam	904	899
Doğru	581	554
Yanlış	323	345
Bilinmeyen	0	0

Sınıflandırma Tablosu	İYİ	KÖTÜ
Doğru %	64,27	61,62
Yanlış%	35,73	38,38
Bilinmeyen %	0	0

Çizelge 4.38 incelendiğinde; YSA modeli'nin 904 iyi müşterinin 581'ni (%64,27) doğru, 323'ünü (%35,73) yanlış ve 899 kötü müşterinin 554'ünü (%61,62) doğru, 345'ini (%38,38) yanlış sınıfladığı görülmektedir.

4.5.1 Sonuç

Yapılan bu çalışma, YSA tekniklerinin bir tahmin aracı olarak kullanılabilirliğini ve oldukça iyi sonuç verdiğini göstermektedir. Sınıflama modellerinin iyi sonuç vermesi, gözlem sayısı ile sıkı ilişkilidir. YSA teknikleri ise daha az veri ile çalışmaya müsaade etmektedir. Ancak YSA teknikleri kara kutu (black box) özelliği gösterdiklerinden bazen olumsuz sonuçlar üretebilirler. Bu yüzden, tahmin aracı olarak kullanıldıklarında, geleneksel metotlar ile bulunan sonuçlar YSA tekniklerini desteklemede yardımcı olarak kullanılabilir (Hamzaçebi ve Kutay, 2004).

Kredi kartı müşterilerine ait MÜŞTERİ TİPİ, CİNSİYET, MEDENİ HAL, ÖĞRENİM DURUMU, ARABA SAHİBİ, BAŞKA BANKA KREDİ KARTI, EK KART DURUMU, YAŞ, AYLIK NET GELİR, BAŞKA BANKA KREDİ KART LİMİTİ değişkenleri ile STATISTICA programında yapılan YSA uygulamasında üç tabakalı bir model kurulmuştur. Girdi tabakasında on adet nöron ve çıktı tabakasında bir adet nöron bulunmaktadır. İlk gizli tabakadaki nöron sayısı 6, ikinci gizli tabakadaki nöron sayısı 4'tür ve YSA 1000 döngü ile eğitilmiştir. Modele en büyük katkıyı MAAŞ, en az katkıyı ise ÖĞRENİM DURUMU değişkenleri yapmıştır. Modelin iyi müşterileri doğru tahmin etme oranı %64,27, kötü müşterileri doğru tahmin etme oranı %61,62'dir. Modelin genel tahmin oranı ise %62,945'tir. İyi müşterilerin tahmin oranı diskriminant analizi ve lojistik regresyona göre düşüktür. Kötü müşterilerin tahmin oranı ise lojistik regresyona göre düşük, diskriminant analizine göre ise yüksektir. Genel tahmin oranı ise hem lojistik regresyon hem de diskriminant analizine göre düşüktür.

4.6 Kredi Kartlarında Sınıflandırma Ağaçları ile Skorlama

Kredi kartı müşterilerine ait MÜŞTERİ TİPİ, CİNSİYET, MEDENİ HALİ, ÖĞRENİM

DURUMU, ARABA SAHİBİ, BAŞKA BANKA KREDİ KARTI, EK KART DURUMU, ÇALIŞMA ŞEKLİ, YAŞ, İKAMET SÜRESİ, MAAŞ, AYLIK NET GELİR, BAŞKA BANKA KART LİMİTİ bağımsız değişkenleri ve İYİ-KÖTÜ bağımlı değişkenine ait 1806 gözlem değerine tüm kategorik yada sürekli değişkenlere uygulanabilen CHAID (Chi-squared Automatic Iteration Detector) analizi uygulanmıştır. Uygulama için SPSS 13 programından yararlanılmıştır.

Çizelge 4.34 Çözümlenen modelin özeti

Model Özeti		
Tanımlar	Büyüme Tekniği	CHAID
	Bağımlı Değişken	İYİ-KÖTÜ
	Bağımsız Değişkenler	MÜŞTERİ TİPİ
		CİNSİYET
		MEDENİ HAL
		ÖĞRENİM DURUMU
		ARABA SAHİBİ
		BAŞKA BANKA KREDİ KARTI
		EK KART DURUMU
		ÇALIŞMA ŞEKLİ
		YAŞ
		İKAMET SÜRESİ
		MAAŞ
		GELİR
		BAŞKA BANKA KART LİMİTİ
	Maksimum Ağaç Uzunluğu	3
	Ana Düğümdeki Minimum Durum Sayısı	90
	Alt Düğümdeki Minimum Düğüm Sayısı	45
Sonuçlar	Modele Giren Bağımsız Değişkenler	AYLIK NET GELİR
		ÖĞRENİM DURUMU
		MEDENİ HAL
		MÜŞTERİ TİPİ
		BAŞKA BANKA KART LİMİTİ
	Düğüm Sayısı	17
	Terminal Düğüm Sayısı	11
	Uzunluk	3

Çizelge 4.34 model hakkında genel bilgileri içermektedir.

Kredi kartı müşterileri incelendiğinde, müşterilerin % 50,1'i (905) kötü, %49,9'u (900) iyidir.

Çizelge 4.35 Kazanç değeri

Kazanç Tablosu						
Düğümler	Düğüm		Kazanç		Cevap	İndeks
	N	%	N	%	%	%
10	139	7,7	111	12,27	79,86	159,27
11	227	12,58	177	19,56	77,97	155,52
8	97	5,37	62	6,85	63,92	127,48
16	48	2,66	30	3,31	62,5	124,65
14	66	3,66	34	3,76	51,52	102,75
5	376	20,83	191	21,1	50,8	101,32
7	221	12,24	108	11,93	48,87	97,47
15	114	6,32	51	5,64	44,74	89,23
12	48	2,66	20	2,21	41,67	83,1
13	168	9,31	57	6,3	33,93	67,67
1	301	16,68	64	7,07	21,26	42,41

Kazanç Tablosu'nda başlıklar;

Düğüm (Node) : Düğümlerin sıra sayısı

Düğüm (n) : Hedef sınıfındaki örnek durumların sayısı

Düğüm(%) : Hedef sınıfına düşen toplam örnek durumlarının yüzdesi

Kazanç(n) : Grup için kazanç değeri (Gruptaki kötü müşteri sayısı)

Kazanç(%) : Grup için kazanç yüzdesi (Gruptaki kötü müşteri oranı)

Cevap(%) : Grup için cevap yüzdesi

İndeks(%): Verilerin hepsi için, grup kazancının kazanca oranı

Kazanç Tablosu (Gain Summary) hedef kategorideki (burada kötü müşteriler olarak seçilmiştir) hangi düğümün en düşük yada en yüksek orana sahip olduğunu göstermektedir. Hangi düğümün kötü müşterileri belirlemede en çok etkili olduğu bilinmek istenir. Bunun için kazanç tablosu incelendiğinde, kötü müşterileri belirlemede onuncu düğümün en belirleyici düğüm olduğu gözlemlenmektedir (%79,9). En az belirleyici olan düğüm ise %21,3 oranıyla birinci düğümdür. Cevap sütunu ise düğümdeki kötü müşteri oranını göstermektedir. İndeks sütunundaki değerler cevap değerinin örneklemede kötü müşteri oranına bölümü şeklinde hesaplanır. Ve indeks değeri yüksek olan düğümlerdeki özellikler kaçınılması gereken müşteri grubunu tanımlar. Burada onuncu düğümdeki müşteri tipi incelendiğinde, aylık net geliri

1000-3000 YTL arası olan ve J (Özel müşteriler), G (Grup personeli), D (Mektup 2 kampanyası), M (Mail order), R (Alışveriş merkezi) kampanyalarıyla kazanılan müşterilere kredi kartı verilirken dikkatli olunması gerektiği yorumu yapılabilir (S.P.S.S., 2001).

CHAID modeliyle iyi-kötü müşterilerin en iyi belirleyicisi olarak AYLIK NET GELİR değişkenini hesaplamıştır. AYLIK NET GELİR değişkeni

700 YTL altı

700–1000 YTL

1000–1800 YTL

1800–3000YTL

3000 YTL ve üzeri

olmak üzere;

1800 YTL ve üzeri maaşı olan müşterilerin %71,6'sı ödeme alışkanlığı kötü müşterilerdir. Beklenenin aksine aylık net gelirin yüksek olduğu müşterilerde ödeme alışkanlığının kötü olması banka politikası açısından geliri yüksek kişilerle ilgili gerekli hassasiyet gösterilmeden kart verildiği şeklinde yorumlanabilir. Buna göre aylık net geliri 700YTL ve altı olan müşteri grupları, kendi içinde %21,26'lık bir kötü oranı ve %7,07'lik bir kazanç değeriyle kredi kartı verilmesi en çok tercih edilen gruptur(1. düğüm). 700YTL–1000 YTL aylık net geliri olan müşteriler %47,3'lük bir kötü oranına sahiptir. 1000 YTL–1800 YTL aylık net geliri olan müşteriler ise %6,8'lik bir kötü oranına sahiptir.

İyi–kötü müşterileri ikinci en iyi tahmin eden değişkenler ise MÜŞTERİ TİPİ ve ÖĞRENİM DURUMU değişkenleridir. ÖĞRENİM DURUMU değişkeni;

Yok

İlkokul

Ortaokul

Lise

M.Y.O

Üniversite

Y.Lisans/Doktora

olmak üzere;

700–1000 YTL arası aylık net geliri olan müşteri grubu için, eğitim durumu ilkokul ve altı olanlar %38,9 kötü oranıyla kredi kartı verilmesi en çok tercih edilen gruptur(6. düğüm). Yine bu müşteri grubu için okula gitmeyen, ilkokul mezunu, ortaokul mezunu müşteriler %48,9 kötü oranı ve %11,93 kazanç oranına sahiptir.

Aylık net geliri 1000–1800 YTL arası olan müşterilerden J (Özel müşteriler), G (Grup şirketlerin personelleri), D (Mektup 2 kampanyası), M (Mail order), R (Alışveriş merkezi) kampanyalarıyla kazanılmış müşteri grubu %79,7 kötü oranı ve %12,27 kazanç oranıyla en riskli gruptur. N (Nomal), B (Market), T (Toplu pazarlama), C (Mektup 1 kampanyası) grubu ise %50 iyi %50 kötü oranı ile risk açısından eşit bir gruptur. Aylık net geliri 1800 YTL ve üzeri olan müşterilerden N, J, M, R kampanyalarıyla kazanılmış müşteriler %78 kötü oranıyla ve %19,56 kazanç oranıyla en riskli gruptur. Aynı müşterilerden B, C kampanyalarıyla kazanılmış müşteriler %41,7 kötü oranı ve %2,21 kazançla en az riskli gruptur.

Aylık net geliri 700–1000 YTL olup, okula gitmemiş yada ilkokul mezunu müşteri grubu için evli olanlar %33,9 kötü oranı ve %6,30 kazanç oranıyla en risksiz gruptur. Aynı müşteri grubu için bekar olanlarda ise kötü oranı %51,5 ve kazan oranı %3,76'dır.

Aylık net geliri 1000-1800 YTL olup, N, B, T, C kampanyalarıyla kazanılmış müşterileri grubu için başka banka kredi kartı limiti 1000 YTL ve altı olan müşteriler %44,7 kötü oranıyla ve %5,64 kazanç oranıyla en risksiz gruptur. Buna karşın başka banka kredi kartı limiti 1000 YTL ve üzeri olan müşteriler %62,5 kötü oranı ve %3,1 kazanç oranıyla en riskli müşterilerdir.

Çizelge 4.36 Sınıflandırma

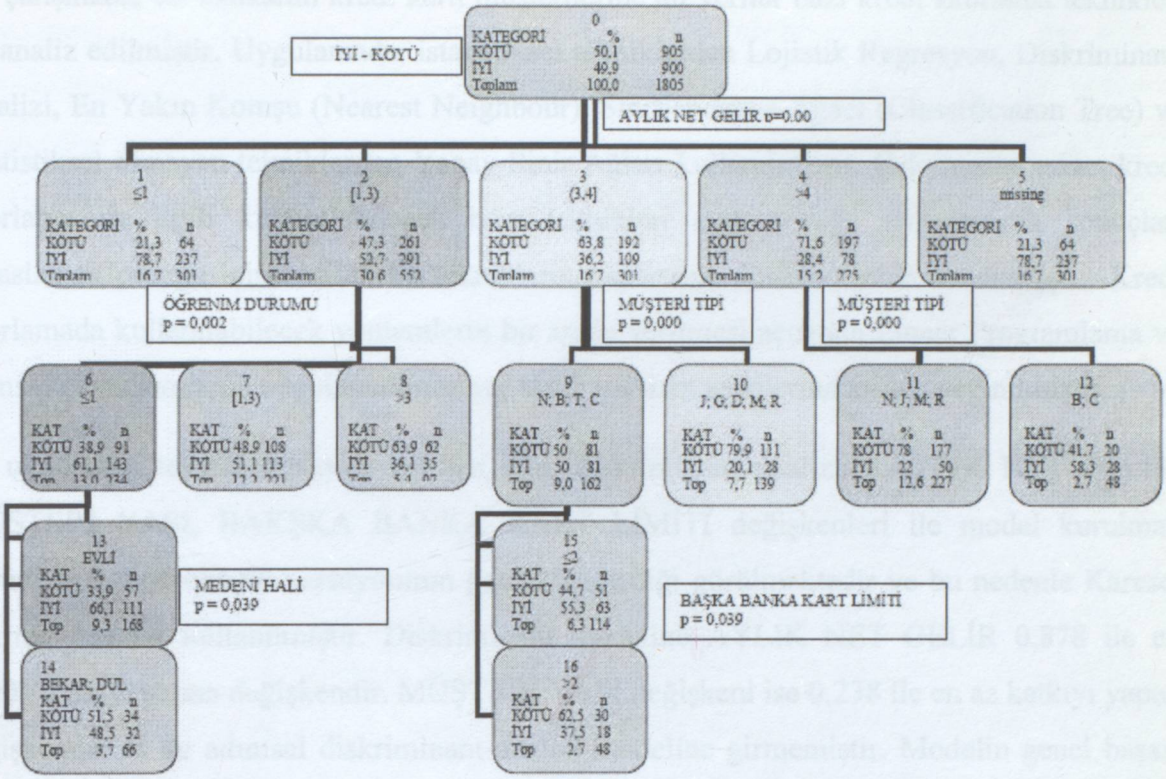
Sınıflandırma			
Gözlem	Tahmin		
	İYİ	KÖTÜ	ORAN
İYİ	552	348	61,33
KÖTÜ	300	605	66,85
Ortalama Oran	47,2	52,8	64,1

Olmak üzere, gerçekte iyi olup yine iyi tahmin edilen kişilerin (552) oranı %61,33 ve gerçekte

iyi olduğu halde kötü tahmin edilen müşteri sayısı 348'dir. Gerçekte kötü olup kötü olarak tahmin edilen müşterilerin (605) oranı %66,85 ve gerçekte kötü olduğu halde iyi olarak tahmin edilen müşteri sayısı ise 300'dür. Sınıflandırma ağacına göre tahmin başarısı iyi ve kötü müşteriler için %64,10'dur.

4.6.1 Sonuç

Aylık net geliri 700–1000 YTL arası, okumayan yada ilkokul mezunu ve evli kişilerin %66,1 iyi oranıyla en risksiz grup olduğu kararı alınabilir. Ayrıca aylık net geliri 1000–1800 YTL arası, J, G, D, M, R kampanyalarıyla kazanılmış müşteriler %79,9 kötü oranıyla en riskli gruptur. Bu sonuçlara bakıldığında; aylık net geliri yüksek, eğitimli, başka banka kredi kart limiti yüksek kişilerin ödeme alışkanlıklarının beklenenin aksine daha kötü çıktığı görülmektedir. Bu da kişilerin eğitim seviyesine bağlı olarak gelir seviyesinin ve dolayısıyla kredi kart limitinin artışı ile harcamalarını kontrol edemeyecek şekilde kart kullanımına yöneldiğini göstermektedir. Eğitim seviyesi düşük, düşük gelire sahip kişiler ise kısıtlı limitler dahilinde kart harcamalarını daha iyi kontrol edebilmektedirler. Bu sonuçtan yola çıkarak bankaların kart limiti belirlemede yada ileriki dönemlerde artırımlarda daha dikkatli davranması gerektiği görülmektedir. Modelin genel tahmin başarısı %64,10'dur. İyi müşteriler için bu oran %61,33, kötü müşteriler için ise %66,85'dir. Model kötü müşterileri daha iyi tahmin etmektedir.



Şekil 4.3 Sınıflandırma ağacı

5. SONUÇLAR VE GENEL DEĞERLENDİRME

Bu çalışmada, bir bankanın kredi kartı müşterilerine ait veriler bazı kredi skorlama teknikleri ile analiz edilmiştir. Uygulamada, istatistiksel tekniklerden Lojistik Regresyon, Diskriminant Analizi, En Yakın Komşu (Nearest Neighbour), Sınıflandırma Ağacı (Classification Tree) ve istatistiksel olmayan tekniklerden Yapay Sinir Ağları kullanılmıştır. Çalışmanın amacı kredi skorlama ile ilgili kullanılabilecek bazı teknikleri aynı örneğe uygulayarak sonuçları kıyaslamak olduğu için seçilen bu tekniklerin teorilerine derinlemesine inilmemiştir. Kredi skorlamada kullanılabilecek yöntemlerin bir arada verilmesi açısında Lineer Programlama ve Tamsayı Programlama uygulamalarına yer verilmeksizin teorilerine kısaca değinilmiştir.

İlk uygulanan teknik varsayım kısıtları olan diskriminant analizidir.AYLIK NET GELİR, MÜŞTERİ YAŞI, BAKŞKA BANKA KART LİMİTİ değişkenleri ile model kurulmak istendiğinde homojenlik varsayımının gerçeklenmediği görülmektedir ve bu nedenle Karesel Ayırma Analizi kullanılmıştır. Diskriminant modeline AYLIK NET GELİR 0,878 ile en büyük katkıyı yapan değişkendir. MÜŞTERİ YAŞI değişkeni ise 0,238 ile en az katkıyı yapan değişkendir ve de adımsal diskriminant analizi modeline girmemiştir. Modelin genel başarı oranı %55,15'tir. Bu başarı oranı iyi müşterileri tahminde %91,1'e kadar çıkarken kötü müşterilerin tahmininde %19,2'dir.

İkinci olarak varsayım kısıtlaması olmayan lojistik regresyon analizi uygulanmıştır. Tüm değişkenler ile analiz yapılmış, analiz sonucuna göre $p<0,25$ kriterini sağlamayan değişkenler modele katkıları yetersiz görüldüğü için modelden elenerek AYLIK NET GELİR, MAAŞ, ÇALIŞMA ŞEKLİ, ARABA SAHİBİ, ÖĞRENİM DURUMU, MÜŞTERİ TİPİ değişkenleriyle tekrar geriye doğru adımsal lojistik regresyon analizi uygulanmıştır. Geriye doğru adımsal lojistik regresyon modeli AYLIK NET GELİR, MAAŞ, ÇALIŞMA ŞEKLİ, ARABA SAHİBİ, ÖĞRENİM DURUMU, MÜŞTERİ TİPİ değişkenleriyle tekrar kurulmuştur. Analiz sonucuna göre müşteri tipinin G (Grup personeli), J (Özel müşteriler), M (Mail order) olması müşterinin iyi-kötü ayrımının yapılmasında büyük rol oynamaktadır ve modelin genel tahmin başarısı %69,42'dir. Tahmin başarısı iyi müşteriler için %66,48, kötü müşteriler için ise %72,37'dir.

En yakın komşu yaklaşımında ise varsayım kısıtlaması olmadığı için tüm değişkenlerle analiz yapılmış, analizde iyi müşteri-kötü müşteri oldukları bilinen fakat bilinmiyor olarak kodlanan beş müşterinin dendogramdaki en yakın komşularına bakılarak tahminde bulunulmuştur. En yakın komşu kümeleme tekniği ve öklit uzaklıklara dayanarak dendogram elde edilmiştir

Burada modelin eksikliği incelenecek en yakın komşu sayısının belirlenememesidir. Bu sayı *M* ve *K* bilgilerine bağlıdır ve bu bilgiler çalışma sırasında elde edilemediği için en yakın komşularla ilgili bir sayı belirlenememiştir. Bu durumda; uygulamanın ne şekilde gelişeceğinin gösterilmesi açısından rasgele en yakın komşular incelenerek bazı yorumlar yapılmıştır ama bunun sadece gösterim amaçlı olduğu unutulmamalıdır. İleride yapılacak çalışmalarda *M, K* değerlerinden yola çıkarak belirlenen sayı ile sağlıklı analizler yapılabilir.

YSA modeli uygulaması da varsayım kısıtlaması olmayan bir tekniktir. Geri yayılma algoritmasıyla kurulan YSA modelinde tüm değişenlerle yapılan analiz sonuçları incelendiğinde modele en büyük katkıyı MAAŞ değişkeninin yaptığı gözlemlenmektedir. ÖĞRENİM DURUMU değişkeni ise modele en az katkıda bulunan değişkendir. Geri yayılma algoritması seçilmesinin nedeni öngörü ve sınıflandırma problemleri için oldukça uygun olması ve doğrusal olmayan yapılar için kullanışlı olmasıdır. YSA mimarisinin girdi tabakasında on, çıktı tabakasında bir adet nöron bulunmaktadır. YSA iki adet gizli tabakaya sahiptir ve birinci gizli tabakada altı, ikinci gizli tabakada dört adet nöron bulunmaktadır. Diskriminant analizi ve lojistik regresyon uygulamalarında modelde etkili olan ÖĞRENİM DURUMU değişkeninin YSA'da az katkı sağlaması bir çelişkidir olarak gözlemlenebilir. Modelin genel tahmin başarısı %62,945'tir. Bu oran iyi müşterilerin tahmininde %64,27, kötü müşterilerin tahmininde %61,62 olarak gözlemlenmektedir.

Son uygulama tekniği olan sınıflandırma ağaçlarında da tüm değişkenlerle CHAID analizi ile çözüm yapıldığında modele AYLIK NET GELİR, ÖĞRENİM DURUMU, MÜŞTERİ TİPİ, MEDENİ HAL, BAŞKA BANKA KART LİMİTİ değişkenleri girmiştir. Sınıflandırma Ağacı çözümlemesinde düğüm sayısı 17, terminal düğüm sayısı 11, uzunluk ise 3 olarak hesaplanmıştır. Kazanç Tablosuna göre kötü müşterileri en iyi tahmin eden düğüm on bir numaralı düğümdür (%19,56). On bir numaralı düğüm incelendiğinde müşteri tipi J (Özel müşteriler), N (Normal müşteriler), M (Mail order), R (Alışveriş merkezi) olan ve aylık geliri 1000-3000 YTL arasında olan müşterilerin büyük oranda kötü müşteriler olacakları yorumu yapılabilir. CHAID modelinde iyi-kötü müşterileri ayırmada en etkili değişken AYLIK NET GELİR'dir. Modelin genel tahmin başarısı %64,1'dir. Bu oran iyi müşteriler için %61,33 iken kötü müşterilerde %66,85'e çıkmaktadır.

Uygulamalardan yola çıkarak bu çalışma için modeller arasında en başarılı genel tahmin yapan model lojistik regresyondur. Lojistik reresyondan sonra sırasıyla sınıflandırma ağacı, diskrimiant analizi ve YSA'dır. İyi müşterilerin doğru sınıflanma başarısı en yüksek olan model sırası; diskriminant analizi, lojistik regresyon, YSA ve sınıflandırma ağacıdır. Kötü

müşterileri doğru sınıflama başarısı en yüksek olan model sıralaması; lojistik regresyon, sınıflandırma ağacı, YSA ve diskriminant analizidir.

Modellere giren değişkenler incelendiğinde; tüm modellere AYLIK NET GELİR değişkeninin büyük katkısı olduğu gözlemlenmektedir. Mevcut veri yapısı ve bu çalışma için bahsedilen değişkenlerin kredi skorlamada etkili değişkenler olduğu söylenebilir.

KAYNAKLAR

Agresti, A., (2001), Categorical Data Analysis, John Willey & Sons, Florida.

Akçapınar, H. ve Gürcan, İ.S., (2002), "Alman Et ve Karacabey Merinosu Koyunlarının Canlı Ağırlık, Vücut Ölçüleri ve Yapağı İnceliği Yönünden Kümeleme Analizi ile İncelenmesi", Turk J. Vet Anim Sci., 26:1255-1261.

Akkuş, Z., Çelik, Y., Satıcı, Ö., Daşdağ, M.M. ve Sanisoğlu, Y., (2005), "Hastane Personelinin Kan Bağışı Hakkındaki Bilgi, Tutum ve Davranışlarının Çok Değişkenli Lojistik Regresyon Tekniğiyle İncelenmesi", İnönü Üniversitesi Tıp Fakültesi Dergisi, 12(1):25-29.

Andersen, E.B., (1994), The Statistical Analysis of Categorical Data, Springer-Verlag, Berlin-Heidelberg.

Arminger, G., Enache D. ve Bone T., (1997), "Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification, Tree Analysis, and Feedforward Networks", Computational Statistics, 12:293-310

Banasik, J., Crook, J.N. ve Thomas, L.C., (1996), "Does Scoring A Subpopulation Make A Difference?"Internat. Rev. Retail Distribution Consumer Res., 6:180-195, Thomas, L.C., Edelman, D.B. ve Crook, J.N., (Derl.), Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, Philadelhia.

Bek, Y., Güler, A., Kaftanoğlu, O. ve Yenihar, H., (1998), "Türkiye'deki Önemli Balarısı Irk ve Ekotiplerinin Morfolojik Karakterler Açısından İlişkilerinin Diskriminant Analiz Tekniğiyle Saptanması", Tr. J. Of Veterinary and Animal Sciences, 23:337-343.

Bishop, C.M., (1995), Neural Networks for Pattern Recognition, Oxford University Press, Oxford, Thomas, L.C., Edelman, D.B. ve Crook, J.N., (Derl.), Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, Philadelhia.

Çelik, H.C., Çelik, M.Y ve Satıcı, Ö., (2005), "Sağlık Personellerinde Kronik Sigara İçme Alışkanlığı Olanların Tutumlarına İlişkin Değişkenlerin Kümeleme Analizi", Dicle Tıp Dergisi, 32(1):20-25.

Doğan, İ., (2002), "Kümeleme Analizi ile Seleksiyon", Turk J. Vet. Anim. Sci., 26:47-53.

Efe, Ö. ve Kaynak, O., (2004), Yapay Sinir Ağları ve Uygulamaları, Boğaziçi Üniversitesi Yayınları, İstanbul.

Gujarati, D.N., (1999), Temel Ekonometri (Çev., Ü., Senesen & G, Şenesen), Literatür Yayıncılık, İstanbul.

Haykin, S., (1999), "Neural Networks: A Comprehensive Foundation", Prentice-Hall International, London, Thomas, L.C., Edelman, D.B. ve Crook, J.N., (Derl.), Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, Philadelhia.

Johnson, R.W., (1992), "Legal, Social and Economic Issues Implementing Scoring in the U.S. in Credit Scoring and Credit Control", Thomas, L.C., Crook, J.N.ve Edelman, D.B., Oxford University Pess, Oxford, Thomas, L.C., Edelman, D.B. ve Crook, J.N., (Derl.), Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, Philadelhia.

Kunter, M.H., Nachtsheim, C.J., Neter, J. ve Wasserman, W., (1996), Applied Linear Statistical Models, The Mc.Graw-Hill Companies, New York.

Lewis, E.M., (1992), An Introduction to Credit Scoring, Athena Press, San Rafael, Thomas, L.C., Edelman, D.B. ve Crook, J.N., (Derl.), Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, Philadelhia.

Malhodra, Naresh, K., (1993), Marketing Research an Applied Orientation, Prentice Hall International.

Menard, S., (1995), Applied Logistic Regression Analysis, Sage Publications, California.

Özdamar, K., (2002a), Paket Programlar ile İstatistikselVeri Analizi 1, Kaan Kitabevi, Eskişehir.

Özdamar, K., (2002b), Paket Programlar ile İstatistikselVeri Analizi 2, Kaan Kitabevi, Eskişehir.

Sharma, S., (1996), Applied Multivariate Techniques, John Willey & Sons, Canada.

S.P.S.S., (2001), Answer Tree 3.0 User's Guide, S.P.S.S. Inc., Chicago.

S.P.S.S., (2001), SPSS Regression Models 11.0, S.P.S.S. Inc., Chicago.

Tatlıdil, H., (1996), Uygulamalı Çok Değişkenli İstatistiksel Analiz, Cem Web Ofset, Ankara.

Thomas, L.C., Edelman, D.B. ve Crook, J.N., (2002), Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, Philadelhia.

Yıldız, D., (1995), Diskriminant Analizine İlişkin Bazı Yöntemler ve Bir Uygulama, İstanbul.

INTERNET KAYNAKLARI

[1] Akpınar, H., (2000), "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", http://www.isletme.istanbul.edu.tr/dergi/nisan2000/1.htm

[2] Aytaç, M. ve Bayram, N., (1999), "Öğretim Elemanlarının Kariyer Tutumlarının Gruplandırılması", http://idari.cu.edu.tr/sempozyum/bil29.htm

[3] Koç, S., (1997), "Türkiye'de İllerin Sosyoekonomik Özelliklere Göre Sınıflandırılması", http://idari.cu.edu.tr/sempozyum/bil20.htm.

[4] Sağıroğlu, Ş., (2002), "Yapay Sinir Ağları ve Mühendislik Uygulamaları", http://www.erciyes.edu.tr/tr/bilgisayar/bilakad/ss/index.htm

[5] Stergiou, C. ve Siganos D., (1996), "Neural Networks", http://www.doc.ic.ac.uk/

[6] Yegorova, I., (2001), "A Successful Neural Network-based Methodology for Predicting Small Business Loan Default", http://www.findarticles.com/p/articles/mi_qa3857/is_200110/ai_n8961377/print

[7] Yurtoğlu, H., (2005), "Yapay Sinir Ağları ile Öngörü Modellemesi:Bazı Makroekonomik Değişkenler İçin Türkiye Örneği", http://www.ekutup.dpt.gov.tr/

[8] http://www.fractalanalytics.com

[9] http://www.statsoft.com/textbook/stcluan.html

EKLER

Ek.1.

Pooled Within-Groups Matrices[a]

		MUSTERI YASI	AYLIK NET GELIR	BASKA BANKA KREDI KARTI KART LIMITI
Covariance	MUSTERI YASI	98,461	2961,309	3096,288
	AYLIK NET GELIR	2961,309	10596179	4251575,129
	BASKA BANKA KREDI KARTI KART LIMITI	3096,288	4251575,1	12293630,344
Correlation	MUSTERI YASI	1,000	,092	,089
	AYLIK NET GELIR	,092	1,000	,373
	BASKA BANKA KREDI KARTI KART LIMITI	,089	,373	1,000

a. The covariance matrix has 330 degrees of freedom.

Ek.2.

Covariance Matrices[a]

IYI - KOTU		MUSTERI YASI	AYLIK NET GELIR	BASKA BANKA KREDI KARTI KART LIMITI
GOOD	MUSTERI YASI	100,082	1378,746	1089,445
	AYLIK NET GELIR	1378,746	5684511,2	3455191,363
	BASKA BANKA KREDI KARTI KART LIMITI	1089,445	3455191,4	14852727,079
BAD	MUSTERI YASI	96,972	4415,060	4939,784
	AYLIK NET GELIR	4415,060	15108060	4983136,959
	BASKA BANKA KREDI KARTI KART LIMITI	4939,784	4983137,0	9942832,180
Total	MUSTERI YASI	98,485	3342,418	3447,593
	AYLIK NET GELIR	3342,418	11037281	4676189,106
	BASKA BANKA KREDI KARTI KART LIMITI	3447,593	4676189,1	12660979,699

a. The total covariance matrix has 331 degrees of freedom.

Ek.3.

Log Determinants

IYI - KOTU	Rank	Log Determinant
GOOD	3	36,517
BAD	3	37,008
Pooled within-groups	3	36,929

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Ek.4.

Test Results

Box's M		51,455
F	Approx.	8,491
	df1	6
	df2	774036,3
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Ek.5.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
MUSTERI YASI	,997	1,081	1	330	,030
AYLIK NET GELIR	,957	14,779	1	330	,000
BASKA BANKA KREDI KARTI KART LIMITI	,968	10,891	1	330	,001

Ek.6.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,058[a]	100,0	100,0	,234

a. First 1 canonical discriminant functions were used in the analysis.

Ek.7.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,945	18,538	3	,000

Ek.8.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
MUSTERI YASI	,131
AYLIK NET GELIR	,685
BASKA BANKA KREDI KARTI KART LIMITI	,487

Ek.9.

Structure Matrix

	Function
	1
AYLIK NET GELIR	,878
BASKA BANKA KREDI KARTI KART LIMITI	,754
MUSTERI YASI	,238

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

Ek.10.

Canonical Discriminant Function Coefficients

	Function
	1
MUSTERI YASI	,013
AYLIK NET GELIR	,000
BASKA BANKA KREDI KARTI KART LIMITI	,000
(Constant)	-1,406

Unstandardized coefficients

Ek.11.

Classification Function Coefficients

	IYI - KOTU	
	GOOD	BAD
MUSTERI YASI	,352	,358
AYLIK NET GELIR	2,96E-005	,000
BASKA BANKA KREDI KARTI KART LIMITI	9,13E-005	,000
(Constant)	-6,978	-7,649

Fisher's linear discriminant functions

Ek.12.

Classification Results[a]

			Predicted Group Membership		
		IYI - KOTU	GOOD	BAD	Total
Original	Count	GOOD	820	80	900
		BAD	731	174	905
	%	GOOD	91,1	8,9	100,0
		BAD	80,8	19,2	100,0

a. 55,1% of original grouped cases correctly classified.

Ek.13.

Categorical Variables Codings

		Frequency	Parameter coding							
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MUSTERI TIPI	B	116	1,000	,000	,000	,000	,000	,000	,000	,000
	C	94	,000	1,000	,000	,000	,000	,000	,000	,000
	D	25	,000	,000	1,000	,000	,000	,000	,000	,000
	G	8	,000	,000	,000	1,000	,000	,000	,000	,000
	J	106	,000	,000	,000	,000	1,000	,000	,000	,000
	M	210	,000	,000	,000	,000	,000	1,000	,000	,000
	N	565	,000	,000	,000	,000	,000	,000	1,000	,000
	R	295	,000	,000	,000	,000	,000	,000	,000	1,000
	T	7	,000	,000	,000	,000	,000	,000	,000	,000
CALISMA SEKLI	1	460	1,000	,000	,000	,000	,000	,000		
	2	151	,000	1,000	,000	,000	,000	,000		
	3	667	,000	,000	1,000	,000	,000	,000		
	4	47	,000	,000	,000	1,000	,000	,000		
	5	2	,000	,000	,000	,000	1,000	,000		
	6	1	,000	,000	,000	,000	,000	1,000		
	7	98	,000	,000	,000	,000	,000	,000		
OGRENIM DURUMU	1	581	1,000	,000	,000	,000	,000			
	2	3	,000	1,000	,000	,000	,000			
	3	588	,000	,000	1,000	,000	,000			
	4	2	,000	,000	,000	1,000	,000			
	5	217	,000	,000	,000	,000	1,000			
	6	35	,000	,000	,000	,000	,000			
NTILES of AYLIK	1	300	1,000	,000	,000	,000				
NET GELIR	2	286	,000	1,000	,000	,000				
	3	264	,000	,000	1,000	,000				
	4	301	,000	,000	,000	1,000				
	5	275	,000	,000	,000	,000				
NTILES of MAAS	1	282	1,000	,000	,000	,000				
	2	287	,000	1,000	,000	,000				
	3	265	,000	,000	1,000	,000				
	4	306	,000	,000	,000	1,000				
	5	286	,000	,000	,000	,000				
MEDENI HALI	EVLÝ	973	1,000	,000						
	DUL	5	,000	1,000						
	BEKAR	448	,000	,000						
BASKA BANKA	1 TANE VAR	194	1,000	,000						
KREDI KARTI	2 TANEVAR	141	,000	1,000						
	YOK	1091	,000	,000						
NTILES of BASKA	2	1083	1,000	,000						
BANKA KART	4	8	,000	1,000						
LIMITI	5	335	,000	,000						
NTILES of YAS	1	489	1,000	,000						
	2	490	,000	1,000						
	3	447	,000	,000						
CINSIYET	ERKEK	1307	1,000							
	KADIN	119	,000							
NTILES of IKAMET	3	1304	1,000							
SURESI	5	122	,000							
ARABA SAHIBI	1	221	1,000							
	2	1205	,000							
EK KART DURUMU	VAR	31	1,000							
	YOK	1395	,000							

Ek.14.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 [a] musteri			57,514	8	**,000**	
musteri(1)	,217	1,143	,036	1	,850	1,242
musteri(2)	,329	1,147	,082	1	,774	1,390
musteri(3)	1,203	1,202	1,002	1	,317	3,331
musteri(4)	1,946	1,368	2,025	1	**,155**	7,002
musteri(5)	1,844	1,153	2,556	1	**,110**	6,320
musteri(6)	1,711	1,133	2,281	1	**,131**	5,533
musteri(7)	,741	1,127	,432	1	**,511**	2,098
musteri(8)	1,258	1,137	1,223	1	,269	3,517
cinsiyet(1)	-,244	,224	1,187	1	,276	,783
medeni_h			1,248	2	,536	
medeni_h(1)	-,138	,148	,860	1	,354	,871
medeni_h(2)	,647	1,181	,300	1	,584	1,910
ogrenim			18,342	5	**,003**	
ogrenim(1)	-1,488	,447	11,087	1	**,001**	,226
ogrenim(2)	20,417	22306,251	,000	1	,999	7E+008
ogrenim(3)	-1,178	,442	7,113	1	**,008**	,308
ogrenim(4)	-,775	1,572	,243	1	,622	,461
ogrenim(5)	-,891	,456	3,824	1	**,051**	,410
araba_sa(1)	,495	,175	7,976	1	**,005**	1,640
b_b_k_k			,692	2	,707	
b_b_k_k(1)	,487	,627	,602	1	,438	1,627
b_b_k_k(2)	,385	,641	,361	1	,548	1,470
ek_kart(1)	,203	,407	,249	1	,618	1,225
calisma			6,059	6	,417	
calisma(1)	,274	,335	,667	1	,414	1,315
calisma(2)	,240	,377	,405	1	,524	1,271
calisma(3)	,221	,307	,517	1	,472	1,247
calisma(4)	-,838	,560	2,238	1	**,135**	,432
calisma(5)	,332	1,944	,029	1	,865	1,393
calisma(6)	21,137	40192,969	,000	1	1,000	2E+009
nyas			,318	2	,853	
nyas(1)	-,069	,171	,161	1	,688	,934
nyas(2)	-,084	,153	,303	1	,582	,919
nikamet_(1)	,068	,254	,072	1	,788	1,071
nmaas			18,410	4	**,001**	
nmaas(1)	-,982	,511	3,697	1	**,055**	,374
nmaas(2)	,250	,438	,325	1	,569	1,284
nmaas(3)	,502	,408	1,516	1	**,218**	1,652
nmaas(4)	,457	,349	1,713	1	**,191**	1,579
naylik_n			12,135	4	**,016**	
naylik_n(1)	-1,370	,487	7,920	1	**,005**	,254
naylik_n(2)	-1,293	,433	8,917	1	**,003**	,275
naylik_n(3)	-1,340	,411	10,607	1	**,001**	,262
naylik_n(4)	-,718	,352	4,155	1	**,042**	,488
nb_bnk_l			,428	2	,808	
nb_bnk_l(1)	,310	,619	,251	1	,617	1,363
nb_bnk_l(2)	-,382	,918	,174	1	,677	,682
Constant	,770	1,429	,290	1	,590	2,160

a. Variable(s) entered on step 1: musteri, cinsiyet, medeni_h, ogrenim, araba_sa, b_b_k_k, ek_kart, calisma, nyas, nikamet_, nmaas, naylik_n, nb_bnk_l.

Ek.15.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5,868	8	,662

Ek.16.

Classification Table[a]

Observed			Predicted		
			IYI - KOTU		Percentage Correct
			GOOD	BAD	
Step 1	IYI - KOTU	GOOD	474	239	66,5
		BAD	197	516	72,4
	Overall Percentage				69,4

a. The cut value is ,500

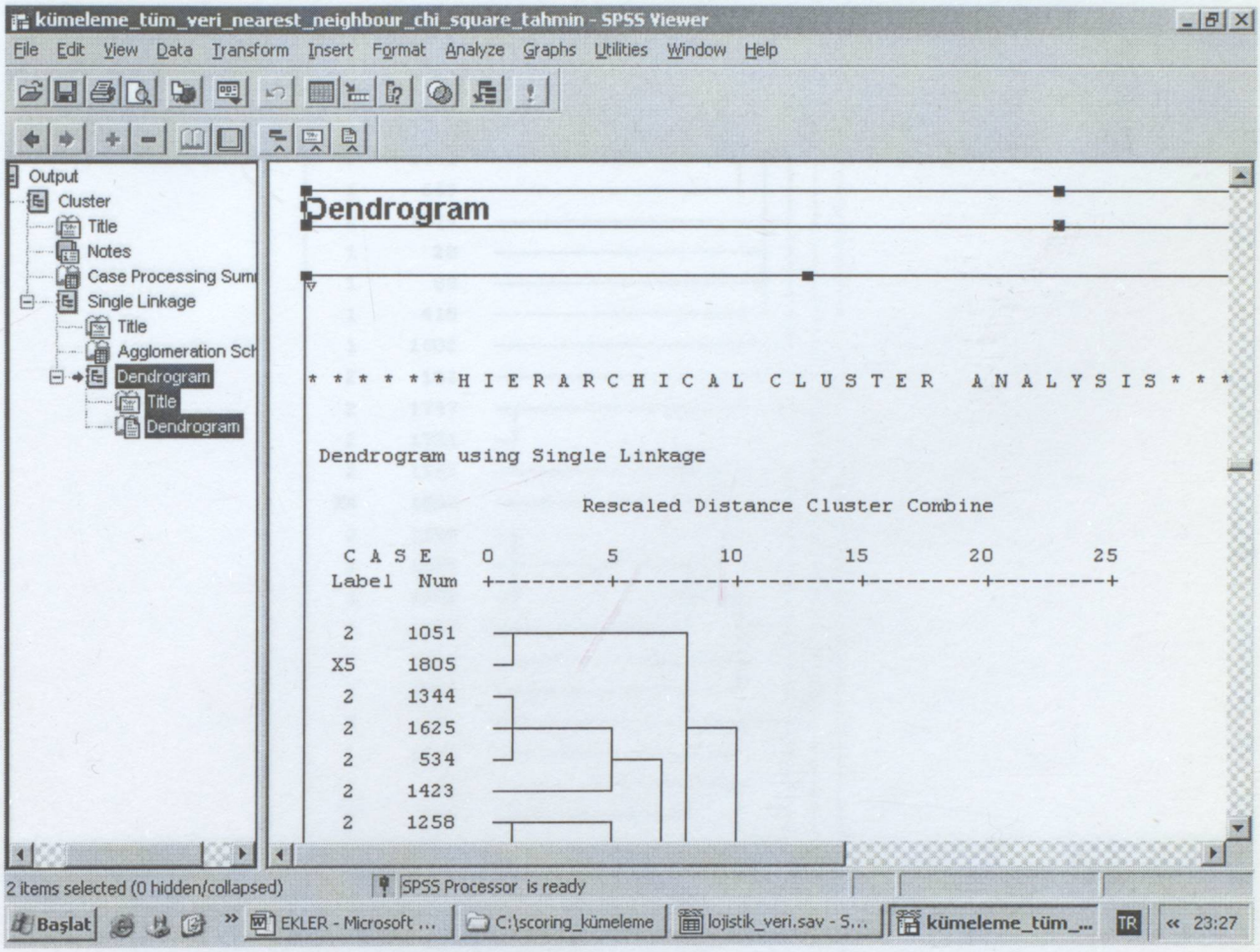
Ek.17.

Variables in the Equation

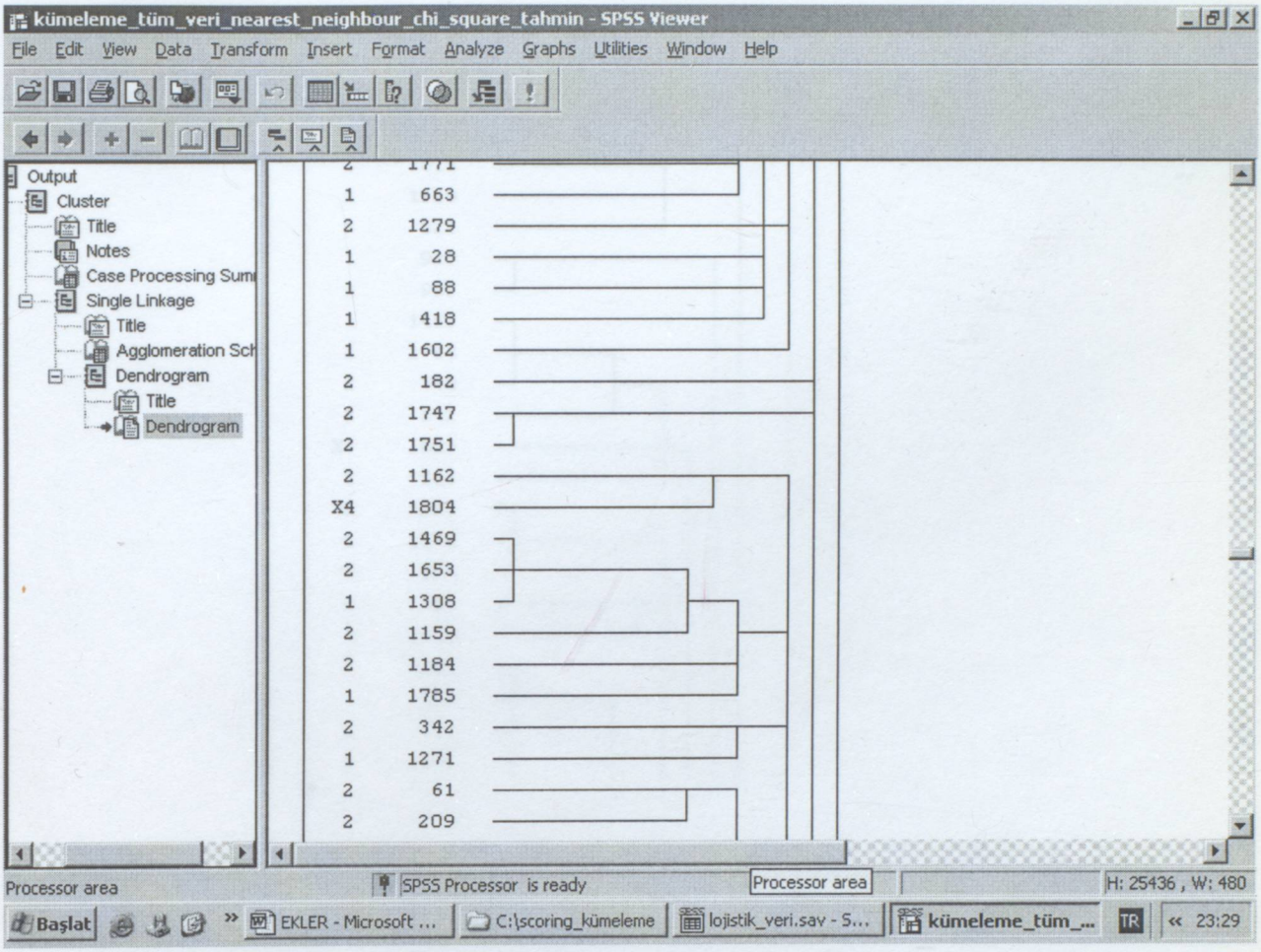
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	naylik_n			12,342	4	,015	
	naylik_n(1)	-1,372	,482	8,121	1	,004	,254
	naylik_n(2)	-1,286	,429	8,986	1	,003	,276
	naylik_n(3)	-1,352	,408	10,953	1	,001	,259
	naylik_n(4)	-,736	,349	4,441	1	,035	,479
	nmaas			19,012	4	,001	
	nmaas(1)	-,979	,504	3,777	1	,052	,376
	nmaas(2)	,259	,435	,353	1	,552	1,295
	nmaas(3)	,498	,406	1,508	1	,220	1,646
	nmaas(4)	,473	,346	1,864	1	,172	1,604
	calisma			6,029	6	,420	
	calisma(1)	,303	,334	,823	1	,364	1,354
	calisma(2)	,209	,375	,312	1	,576	1,233
	calisma(3)	,229	,305	,564	1	,453	1,258
	calisma(4)	-,769	,549	1,958	1	,162	,464
	calisma(5)	,626	1,971	,101	1	,751	1,869
	calisma(6)	21,068	40192,970	,000	1	1,000	1E+009
	araba_sa(1)	,500	,174	8,255	1	,004	1,649
	ogrenim			20,758	5	,001	
	ogrenim(1)	-1,532	,442	12,013	1	,001	,216
	ogrenim(2)	20,383	22263,162	,000	1	,999	7E+008
	ogrenim(3)	-1,198	,437	7,535	1	,006	,302
	ogrenim(4)	-,768	1,582	,236	1	,627	,464
	ogrenim(5)	-,895	,451	3,939	1	,047	,409
	musteri			57,537	8	,000	
	musteri(1)	,205	1,143	,032	1	,857	1,228
	musteri(2)	,318	1,147	,077	1	,782	1,374
	musteri(3)	1,123	1,201	,874	1	,350	3,074
	musteri(4)	1,905	1,365	1,946	1	,163	6,717
	musteri(5)	1,782	1,154	2,387	1	,122	5,943
	musteri(6)	1,664	1,134	2,153	1	,142	5,279
	musteri(7)	,714	1,127	,401	1	,526	2,043
	musteri(8)	1,211	1,132	1,144	1	,285	3,356
	Constant	,856	1,245	,473	1	,491	2,354

a. Variable(s) entered on step 1: naylik_n, nmaas, calisma, araba_sa, ogrenim, musteri.

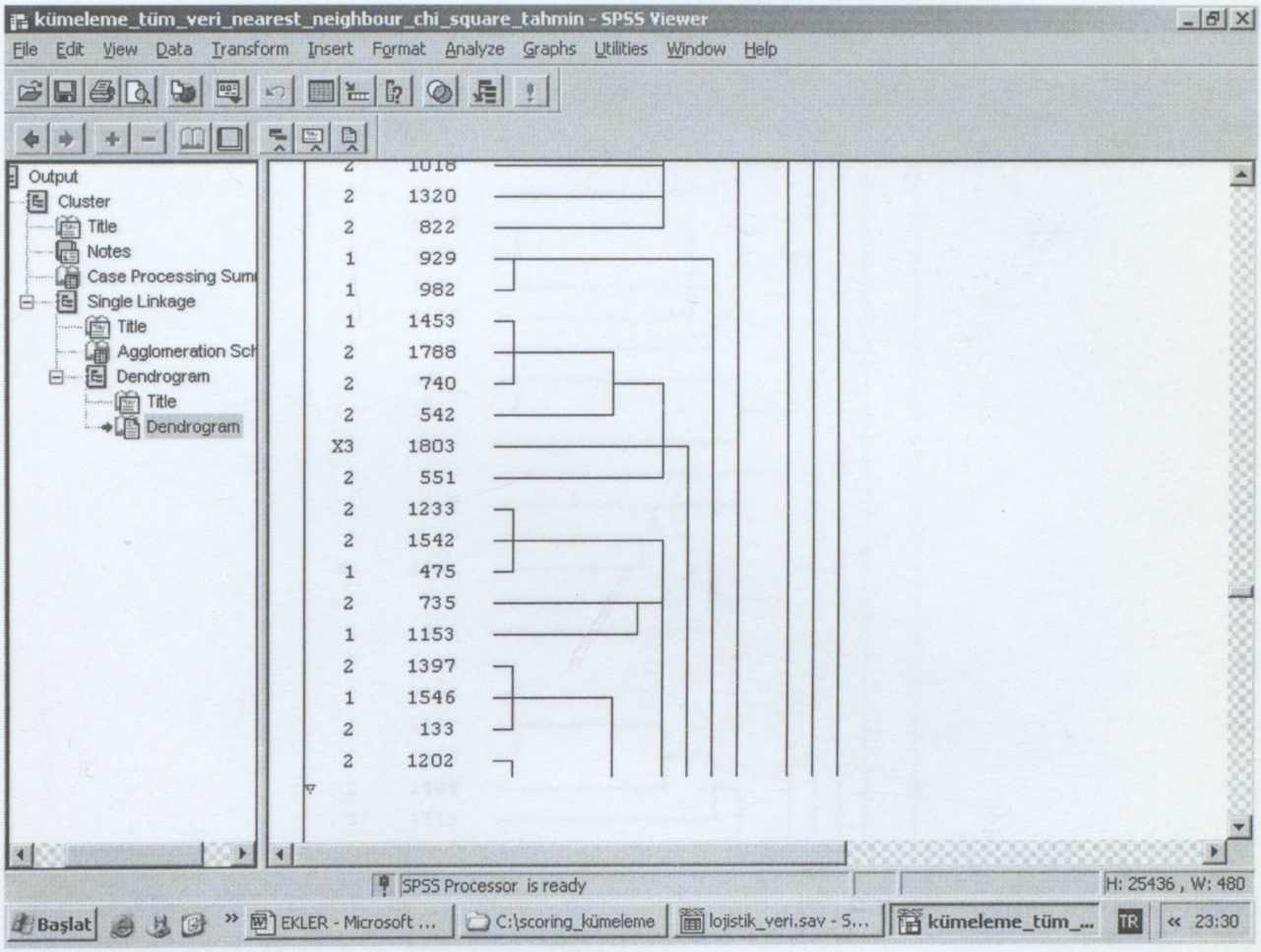
Ek.18.


kümeleme_tüm_veri_nearest_neighbour_chi_square_tahmin - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Cluster
Title
Notes
Case Processing Sum
Single Linkage
Title
Agglomeration Sch
Dendrogram
Title
Dendrogram
Dendrogram
* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * *
Dendrogram using Single Linkage
Rescaled Distance Cluster Combine
C A S E
Label Num
0 5 10 15 20 25
2 1051
X5 1805
2 1344
2 1625
2 534
2 1423
2 1258
2 items selected (0 hidden/collapsed)
SPSS Processor is ready
Başlat
EKLER - Microsoft ...
C:\scoring_kümeleme
lojistik_veri.sav - S...
kümeleme_tüm_...
TR
23:27

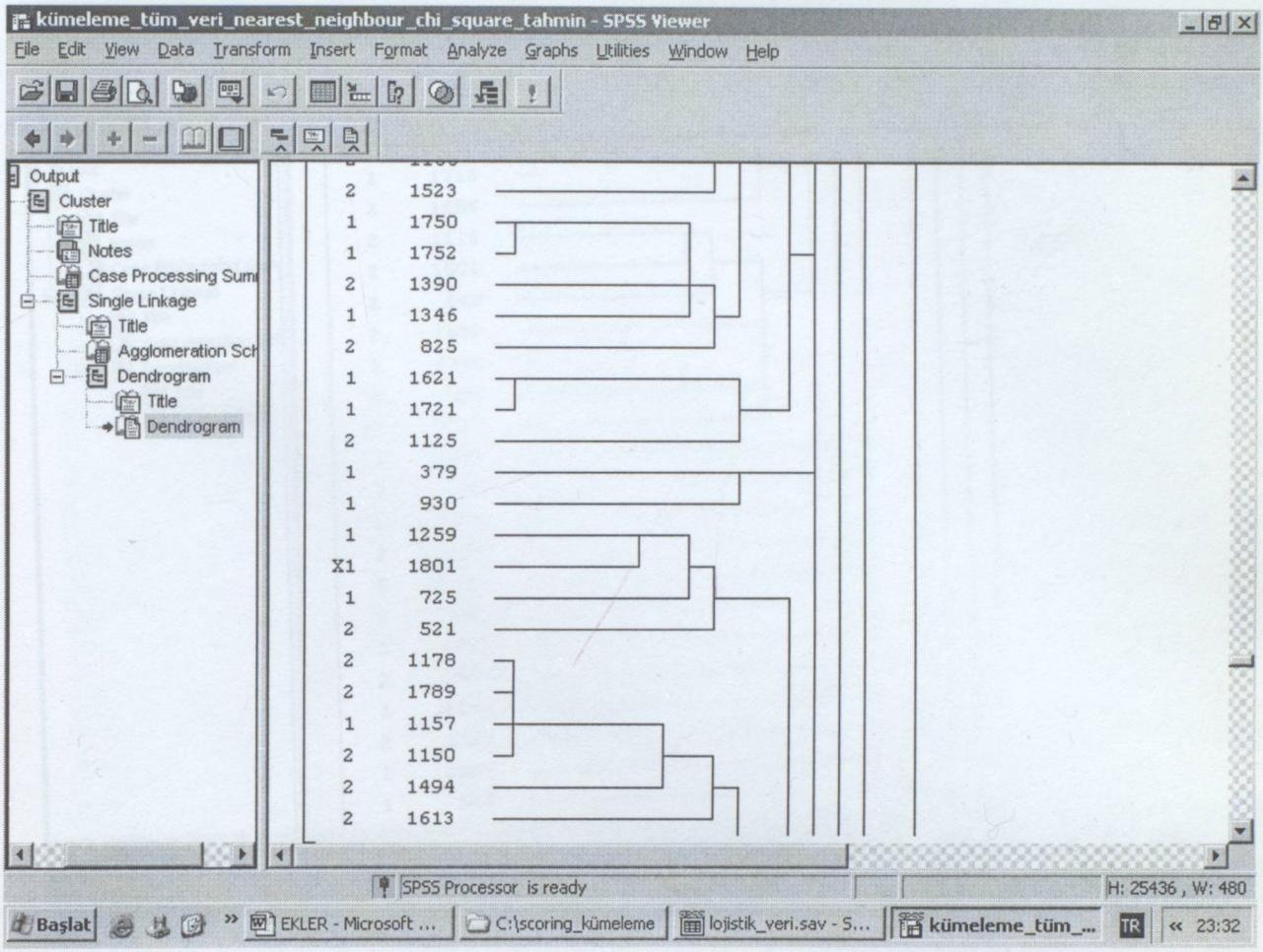
Ek.19.


kümeleme_tüm_veri_nearest_neighbour_chi_square_tahmin - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Cluster
Title
Notes
Case Processing Sum
Single Linkage
Title
Agglomeration Sch
Dendrogram
Title
Dendrogram
1 663
2 1279
1 28
1 88
1 418
1 1602
2 182
2 1747
2 1751
2 1162
X4 1804
2 1469
2 1653
1 1308
2 1159
2 1184
1 1785
2 342
1 1271
2 61
2 209
Processor area
SPSS Processor is ready
Processor area
H: 25436 , W: 480
Başlat
EKLER - Microsoft ...
C:\scoring_kümeleme
lojistik_veri.sav - S...
kümeleme_tüm_...
TR
23:29

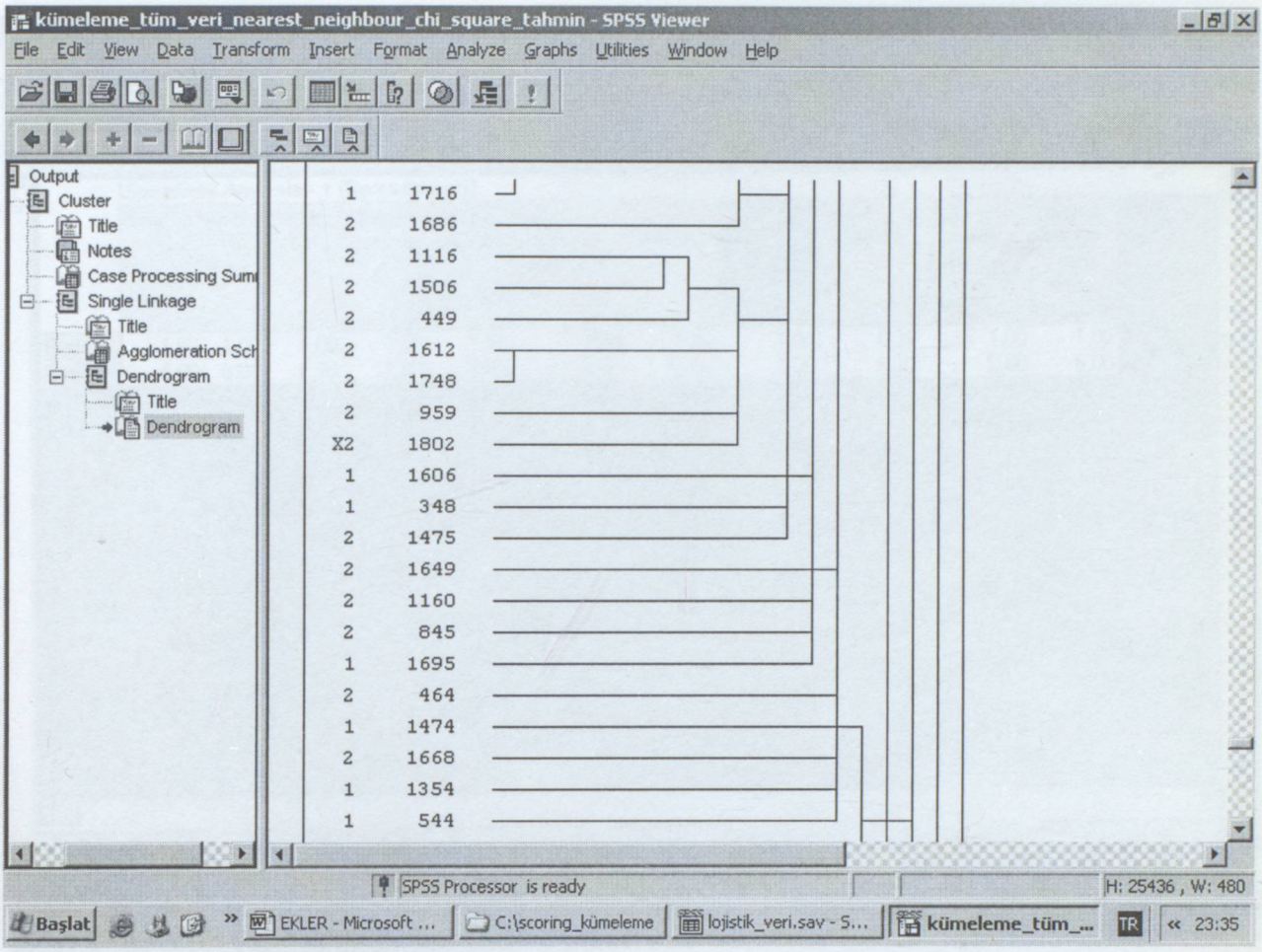
Ek.20.


kümeleme_tüm_veri_nearest_neighbour_chi_square_tahmin - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Cluster
Title
Notes
Case Processing Sum
Single Linkage
Title
Agglomeration Sch
Dendrogram
Title
Dendrogram
2 1320
2 822
1 929
1 982
1 1453
2 1788
2 740
2 542
X3 1803
2 551
2 1233
2 1542
1 475
2 735
1 1153
2 1397
1 1546
2 133
2 1202
SPSS Processor is ready
H: 25436 , W: 480
Başlat
EKLER - Microsoft ...
C:\scoring_kümeleme
lojistik_veri.sav - S...
kümeleme_tüm_...
TR
23:30

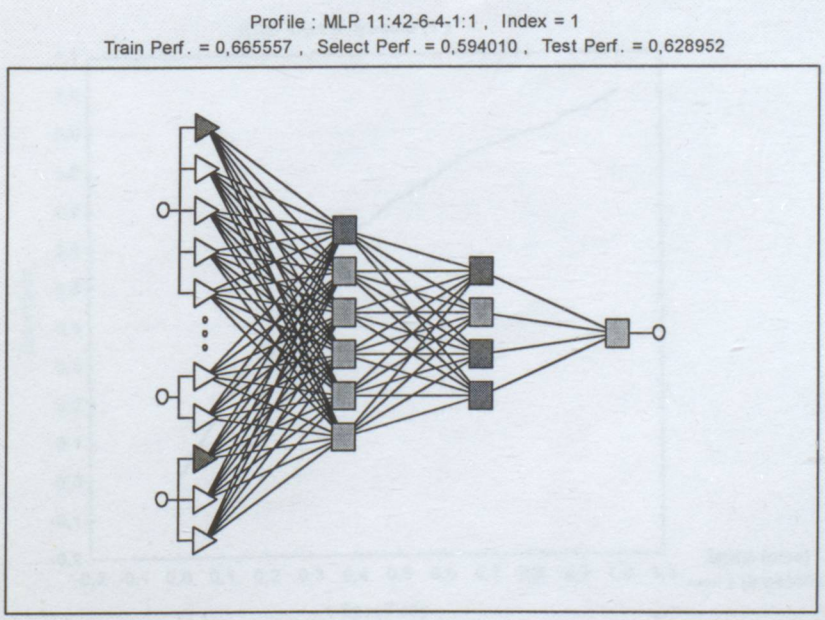
Ek.21.


kümeleme_tüm_veri_nearest_neighbour_chi_square_tahmin - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Cluster
Title
Notes
Case Processing Sum
Single Linkage
Title
Agglomeration Sch
Dendrogram
Title
Dendrogram
2 1523
1 1750
1 1752
2 1390
1 1346
2 825
1 1621
1 1721
2 1125
1 379
1 930
1 1259
X1 1801
1 725
2 521
2 1178
2 1789
1 1157
2 1150
2 1494
2 1613
SPSS Processor is ready
H: 25436 , W: 480
Başlat
EKLER - Microsoft ...
C:\scoring_kümeleme
lojistik_veri.sav - S...
kümeleme_tüm_...
TR
23:32

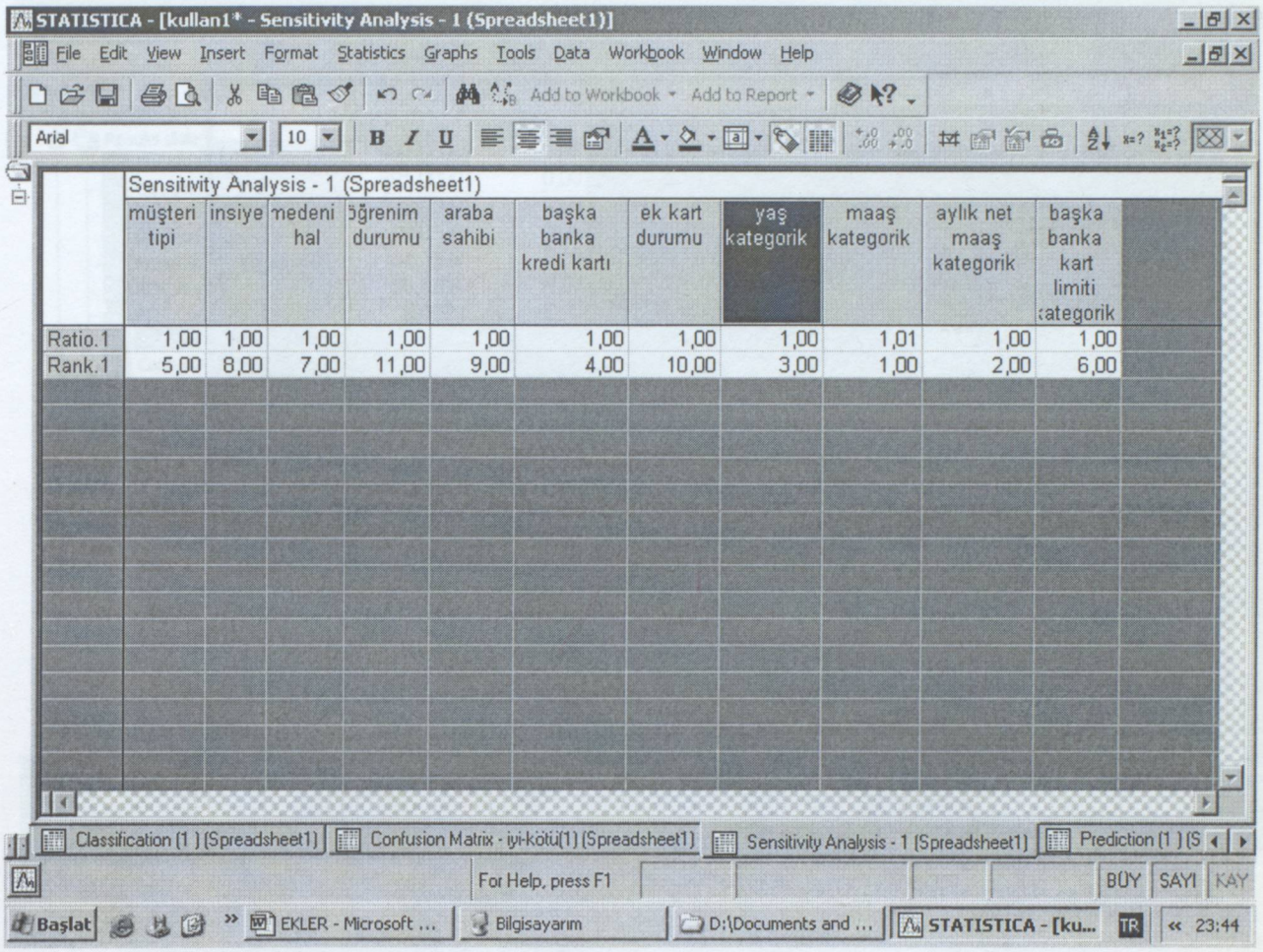
Ek.22.


kümeleme_tüm_veri_nearest_neighbour_chi_square_tahmin - SPSS Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Cluster
Title
Notes
Case Processing Sum
Single Linkage
Title
Agglomeration Sch
Dendrogram
Title
Dendrogram
1 1716
2 1686
2 1116
2 1506
2 449
2 1612
2 1748
2 959
X2 1802
1 1606
1 348
2 1475
2 1649
2 1160
2 845
1 1695
2 464
1 1474
2 1668
1 1354
1 544
SPSS Processor is ready
H: 25436 , W: 480
Başlat
EKLER - Microsoft ...
C:\scoring_kümeleme
lojistik_veri.sav - S...
kümeleme_tüm_...
TR
23:35

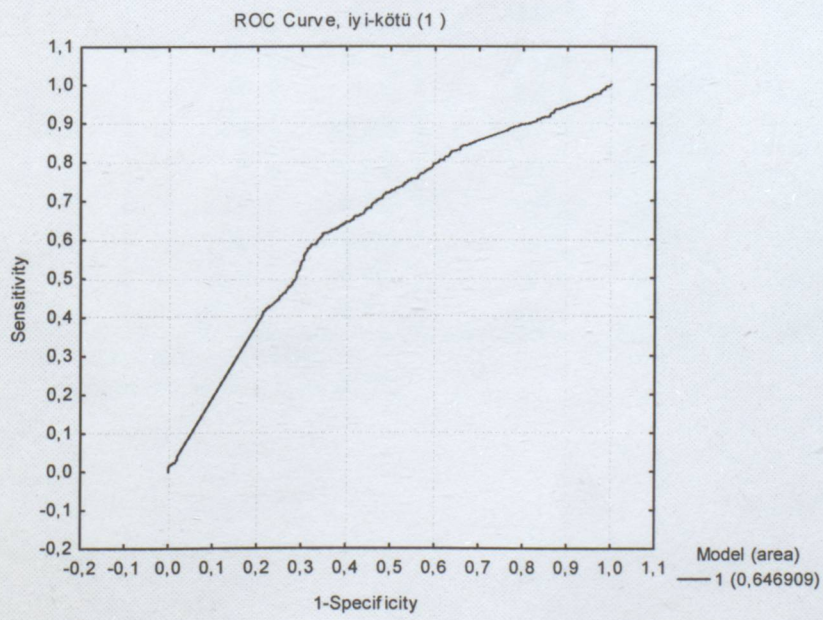
Ek.23.


Profile : MLP 11:42-6-4-1:1 , Index = 1
Train Perf. = 0,665557 , Select Perf. = 0,594010 , Test Perf. = 0,628952

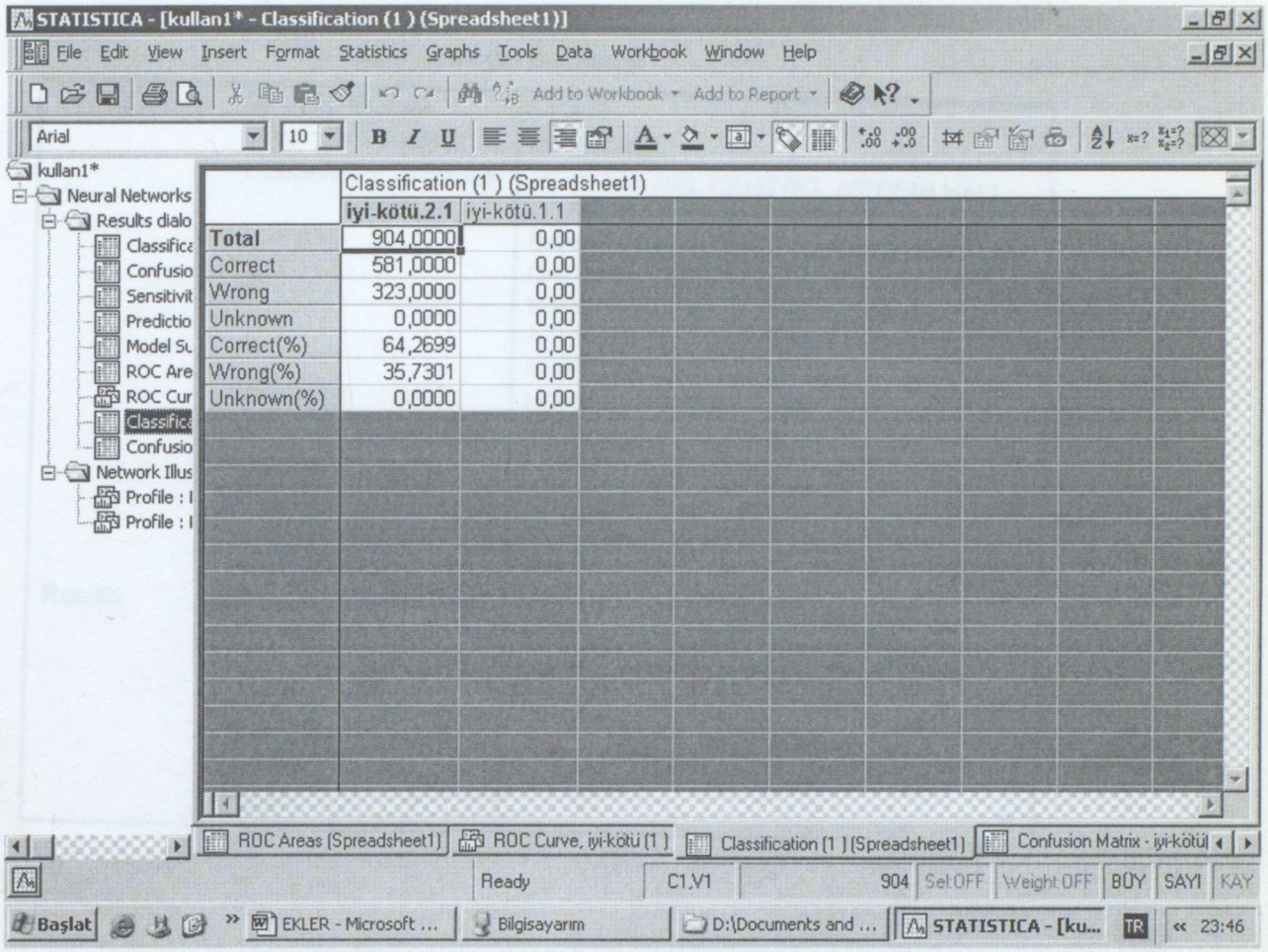
Ek.24.


STATISTICA - [kullan1* - Sensitivity Analysis - 1 (Spreadsheet1)]
File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Add to Workbook
Add to Report
Arial
10
Sensitivity Analysis - 1 (Spreadsheet1)
müşteri tipi
insiye
medeni hal
öğrenim durumu
araba sahibi
başka banka kredi kartı
ek kart durumu
yaş kategorik
maaş kategorik
aylık net maaş kategorik
başka banka kart limiti kategorik
Ratio.1 1,00 1,00 1,00 1,00 1,00 1,00 1,00 1,00 1,01 1,00 1,00
Rank.1 5,00 8,00 7,00 11,00 9,00 4,00 10,00 3,00 1,00 2,00 6,00
Classification (1) (Spreadsheet1)
Confusion Matrix - iyi-kötü(1) (Spreadsheet1)
Sensitivity Analysis - 1 (Spreadsheet1)
Prediction (1) (S
For Help, press F1
BÜY
SAYI
KAY
Başlat
EKLER - Microsoft ...
Bilgisayarım
D:\Documents and ...
STATISTICA - [ku...
TR
23:44

Ek.25.


ROC Curve, iyi-kötü (1)
Sensitivity
1-Specificity
1,1
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-0,1
-0,2
-0,2 -0,1 0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1,0 1,1
Model (area)
1 (0,646909)

Ek.26.


STATISTICA - [kullan1* - Classification (1) (Spreadsheet1)]
File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Add to Workbook
Add to Report
Arial
10
kullan1*
Neural Networks
Results dialo
Classifica
Confusio
Sensitivit
Predictio
Model Su
ROC Are
ROC Cur
Classifica
Confusio
Network Illus
Profile : I
Profile : I
Classification (1) (Spreadsheet1)
iyi-kötü.2.1
iyi-kötü.1.1
Total 904,0000 0,00
Correct 581,0000 0,00
Wrong 323,0000 0,00
Unknown 0,0000 0,00
Correct(%) 64,2699 0,00
Wrong(%) 35,7301 0,00
Unknown(%) 0,0000 0,00
ROC Areas (Spreadsheet1)
ROC Curve, iyi-kötü (1)
Classification (1) (Spreadsheet1)
Confusion Matrix - iyi-kötü
Ready
C1,V1
904
Sel:OFF
Weight:OFF
BÜY
SAYI
KAY
Başlat
EKLER - Microsoft ...
Bilgisayarım
D:\Documents and ...
STATISTICA - [ku...
TR
23:46

Ek.27.

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	IYI - KOTU
	Independent Variables	MUSTERI TIPI, CINSIYET, MEDENI HALI, OGRENIM DURUMU, ARABA SAHIBI, BASKA BANKA KREDI KARTI, EK KART DURUMU, CALISMA SEKLI, NTILES of YAS, NTILES of IKAMET SURESI, NTILES of MAAS, NTILES o AYLIK NET GELIR, NTILES of BASKA BANKA KART LIMITI
	Validation	NONE
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	90
	Minimum Cases in Child Node	45
Results	Independent Variables Included	NTILES of AYLIK NET GELIR, OGRENIM DURUMU, MEDENI HALI, MUSTERI TIPI, NTILES of BASKA BANKA KART LIMITI
	Number of Nodes	17
	Number of Terminal Nodes	11
	Depth	3

Ek.28.

Gains for Nodes

	Node		Gain			
Node	N	Percent	N	Percent	Response	Index
10	139	7,7%	111	12,3%	79,9%	159,3%
11	227	12,6%	177	19,6%	78,0%	155,5%
8	97	5,4%	62	6,9%	63,9%	127,5%
16	48	2,7%	30	3,3%	62,5%	124,7%
14	66	3,7%	34	3,8%	51,5%	102,7%
5	376	20,8%	191	21,1%	50,8%	101,3%
7	221	12,2%	108	11,9%	48,9%	97,5%
15	114	6,3%	51	5,6%	44,7%	89,2%
12	48	2,7%	20	2,2%	41,7%	83,1%
13	168	9,3%	57	6,3%	33,9%	67,7%
1	301	16,7%	64	7,1%	21,3%	42,4%

Growing Method: CHAID
Dependent Variable: IYI - KOTU

Ek.29.

Classification

Observed	Predicted		
	GOOD	BAD	Percent Correct
GOOD	552	348	61,3%
BAD	300	605	66,9%
Overall Percentage	47,2%	52,8%	64,1%

Growing Method: CHAID
Dependent Variable: IYI - KOTU

ÖZGEÇMİŞ

Doğum tarihi	09.03.1980	
Doğum yeri	İstanbul	
Lise	1994-1997	Eyüp Lisesi
Lisans	1997-2001	Yıldız Teknik Üniversitesi Fen Fak. İstatistik Bölümü

Çalıştığı kurum(lar)

2002-2005	Family Finans Kurumu A.Ş.
2005-	Türkiye Finans Katılım Bankası