**KOCAELİ ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

**ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI**

**YÜKSEK LİSANS TEZİ**

# VERİ MADENCİLİĞİ YAKLAŞIMI İLE BİREYSEL MÜŞTERİLERİN KREDİ ÖDEME PERFORMANSLARININ DEĞERLENDİRİLMESİ

**ASLI ÇALIŞ**

**KOCAELİ 2013**

**KOCAELİ ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

**ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI**

**YÜKSEK LİSANS TEZİ**

# VERİ MADENCİLİĞİ YAKLAŞIMI İLE BİREYSEL MÜŞTERİLERİN KREDİ ÖDEME PERFORMANSLARININ DEĞERLENDİRİLMESİ

**ASLI ÇALIŞ**

**Yrd.Doç.Dr. Kasım BAYNAL**
**Danışman, Kocaeli Üniv.** .....

**Prof.Dr. Nilgün FIĞLALI**
**Jüri Üyesi, Kocaeli Üniv.** .....

**Doç.Dr. Sermin ELEVLİ**
**Jüri Üyesi, Ondokuz Mayıs Üniversitesi** .....

**Tezin Savunulduğu Tarih: 21.01.2013**

## ÖNSÖZ VE TEŞEKKÜR

Çalışmada, bireysel banka kredisi kullanan müşterilerin geri ödeme performanslarının değerlendirilmesine yönelik veri madenciliği uygulamasına yer verilmiştir. Kümeleme ve sınıflandırma yöntemleri kullanılarak, mevcut müşterilerin analizi yapılmış ve gelecekteki potansiyel müşteriler için çıkarımda bulunulmuştur.

Bu tezin hazırlanması aşamasında yardımlarını esirgemeyen, bana çalışmamın her aşamasında yol gösteren danışmanım Yrd. Doç. Dr. Kasım BAYNAL'a, tezimin son şeklini almasında büyük katkıları olan Hocalarım Doç. Dr. Sermin ELEVLİ ve Öğr. Gör. Dr. Naci MURAT'a, göstermiş oldukları maddi ve manevi desteklerinden ötürü sevgili aileme teşekkürü bir borç bilirim.

Ocak – 2013 Aslı ÇALIŞ

## İÇİNDEKİLER

## ŞEKİLLER DİZİNİ

## TABLOLAR DİZİNİ

## SİMGELER DİZİNİ VE KISALTMALAR

Σ : Birleştirme fonksiyonu
k : Küme sayısı
n : Birim sayısı

### Kısaltmalar

AID : Automatic Interaction Detection (Otomatik Etkileşim Çıkarma)
CHAID : Chi-squared Automatic Interaction Detector (Ki-kare Otomatik Etkileşim Dedektörü)
C&RT : Classification and Regression Tree (Sınıflandırma ve Regresyon Ağacı)
GA : Genetik Algoritma
KGS : Kartlı Geçiş Sistemi
OGS : Otomatik Geçiş Sistemi
OLAP : Online Analytical Processing (Çevrimiçi Analitik İşleme)
QUEST : Quick, Unbiased, Efficient Statistical Tree (Hızlı, Yansız, Etkili . İstatistiksel Ağaç)
SLIQ : Supervised Learning in Quest (Quest Algoritmasında Denetimli Öğrenme)
SPRINT : Scalable Parallelizable Induction of Decision Tree (Karar Ağacının Ölçeklenebilir Paralel İndüksiyonu)
SQL : Structured Query Language (Yapılandırılmış Sorgu Dili)
VM : Veri Madenciliği
VTBK : Veri Tabanında Bilgi Keşfi
YSA : Yapay Sinir Ağları

## VERİ MADENCİLİĞİ YAKLAŞIMI İLE BİREYSEL MÜŞTERİLERİN KREDİ ÖDEME PERFORMANSLARININ DEĞERLENDİRİLMESİ


## ÖZET

Bilgisayar teknolojilerindeki gelişme ile birlikte bilgi miktarında ve veri tabanı sistemlerinin hacminde meydana gelen artış, büyük veri tabanlarında gizli kalmış, anlamlı bilgilerin keşfedilmesi ihtiyacını, dolayısıyla "Veri Madenciliği" kavramını doğurmuştur. Bilginin olağanüstü artışıyla birlikte her alanda strateji geliştirme konusunda ileriye dönük tahmin sistemlerine ihtiyaç duyulmuştur. Bu bağlamda veri madenciliği teknikleri birçok alanda olduğu gibi bankacılık alanında da yaygın bir şekilde kullanılmaktadır. Bankacılık sektöründe yapılan bu çalışmada, veri madenciliği yöntemlerinden kümeleme ve sınıflandırma ile mevcut bireysel kredi müşterilerinin analizi ve gelecekteki potansiyel müşterilerin ödeme durumlarına ilişkin çıkarım yapılması amaçlanmıştır. Çalışmada veri madenciliği yazımlı olarak SPSS Clementine kullanılmış ve bireysel kredi müşterilerinin değerlendirilmesine yönelik bir uygulama gerçekleştirilmiştir.



**Anahtar Kelimeler:** Bireysel Krediler, Kümeleme, Sınıflandırma, SPSS Clementine, Veri madenciliği

**EVALUATION OF INDIVIDUAL CUSTOMERS' CREDIT PAYMENT PERFORMANCES WITH DATA MINING APPROACH**


## ABSTRACT

With developments in computer technologies, amount of information and volume of database systems increased. So it was needed to explore meaningful information which was hidden in large databeses and so "Data Mining" concept arose. Because of the phenomenal rise in information, future forecasting systems about strategy development were needed in each area. Therefore, data mining techniques are used extensively in banking area such as many areas. In this study, conducted in banking sector, it was aimed to analysis of available personal loan customers and estimate potential customers' payment performances with clustering and classification from data mining methods. In the study, SPSS Clementine was used as a software of data mining and an application was done for evaluation of personal loan customers.



**Keywords:** Personal Loans, Clustering, Classification, SPSS Clementine, Data Mining

## GİRİŞ

Ham veri kendi başına değersizdir. Veri, bilgisayar sistemleriyle belirli bir amaç doğrultusunda işlenerek bilgiye dönüşmektedir. Bilgisayar teknolojilerindeki gelişmeler, üretilen bilgi miktarlarında ve veri tabanı sistemlerinin hacminde artış meydana getirmiştir. Veri tabanlarında saklı tutulan, yararlı olma potansiyeline sahip verilerin keşfedilerek anlamlı örüntülerin ortaya çıkarılması, veri madenciliği (VM) kavramıyla ifade edilmektedir.

Günümüzün tüketici odaklı pazarlarında işletmeler süreklilik arz eden yoğun bir rekabetin içindedirler. İşletmelerin bu rekabet şartlarında başarılı olabilmeleri için etkin ve düşük maliyetli pazarlama stratejileri uygulamaları gerekmektedir (Emel ve Taşkın, 2005). Etkin pazarlama stratejilerinin oluşturulabilmesi için doğru bilgilere, doğru bilgilerin elde edilebilmesi için ise verileri çok boyutlu analiz edebilen ileriye dönük tahmin sistemlerine ihtiyaç duyulmaktadır. Bu bağlamda veri madenciliği teknikleri diğer birçok alanda olduğu gibi bankacılık alanında da yaygın bir şekilde kullanılmaktadır.

Bu çalışmada, ülkemizde faaliyet gösteren bir bankanın birinci sınıf şubesinden elde edilen verilerle bir VM uygulaması gerçekleştirilmiştir. Bankaya ait veriler, bireysel kredi müşterilerinin yaş, cinsiyet, medeni hal, öğrenim durumu, aylık gelir, ev, araç, çocuk sahibi olma durumu, eş geliri, ödeme durumu, banka maaş müşterisi olma durumu ve çalışma şekli olmak üzere toplamda on iki farklı değişkene bağlı kişisel özelliklerini içermektedir. Uygulamada öncelikle kümeleme analizi yapılarak mevcut müşterilerin değerlendirilmesi sağlanmıştır. Ardından karar ağacı algoritmaları ile müşterilerin ödeme durumlarına göre sınıflandırılması sağlanarak, gelecekteki potansiyel müşteriler için çıkarım yapılmıştır. Bu süreçte, sınıflandırma ve kümeleme algoritmalarını kolaylıkla uygulayarak, kısa sürede verideki gizli örüntülere ulaşmamızı sağlayan bir veri madenciliği programı olan SPSS Clementine kullanılmıştır.

Tez çalışmasının birinci bölümünde VM' nin farklı kaynaklardan elde edilen tanımları ile tarihsel gelişimi, örnek uygulamaları ve VM süreci gibi veri madenciliğine genel bir bakış sunulabilecek detaylı bilgilere yer verilmiştir. İkinci bölümde VM modellerinden bahsedilerek, sınıflandırma, kümeleme ve birliktelik kuralı algoritmalarına değinilmiştir. Üçüncü bölümde Clementine programı ile bireysel müşterilerin değerlendirilmesine yönelik bir VM uygulaması gerçekleştirilmiş, son bölümde ise çalışma sonuçlarına yer verilmiş ve genel bir değerlendirme yapılmıştır.

# 1. VERİ MADENCİLİĞİNE GENEL BAKIŞ

Veri madenciliğinin ortaya çıkışı veri yığınlarının geniş yer kaplamasına ve büyük miktardaki verilerin yararlı bilgilere dönüştürülmesi ihtiyacına dayanmaktadır (Han ve Kamber, 2006).

Veri madenciliği, karar destek, pazar stratejisi, finansal tahminler gibi birçok alanda uygulanabilir olması nedeniyle son zamanlarda veritabanı kullanıcıları ve araştırmacıların önemli ölçüde dikkatini çekmektedir. Veri madenciliği, makine öğrenme, istatistik ve veri tabanları alanlarındaki teknikleri birleştirerek, büyük veri tabanlarından faydalı ve değerli bilgiyi çıkarmamıza imkan tanımaktadır (Ching ve Pong, 2002).

Veri madenciliği, istatistik, yapay sinir ağları, karar ağaçları, genetik algoritma ve görsel teknikler gibi yıllardır geliştirilen çeşitli teknikleri içermektedir. Veri madenciliği, pazarlama, finans, bankacılık, üretim, sağlık, müşteri ilişkileri yönetimi ve organizasyon öğrenme gibi çoğu alanda uygulanmaktadır (Chien ve Chen, 2008). Veri madenciliği teknikleri büyük veri tabanlarının taranarak, ilginç ve yararlı örüntülerin ortaya çıkarılması için uygulanmaktadır (Tan ve diğ., 2006).

Veri madenciliği için yapılan farklı tanımlardan bazıları şu şekildedir:

Veri madenciliği, veri tabanları veya veri ambarlarında yer alan yığın veri içindeki gizli örüntüleri ve ilişkileri bulmak için istatistiksel algoritmaları ve yapay zeka yöntemlerini kullanan karmaşık bir veri arama yeteneği olarak tanımlanabilir. Veri madenciliği; aynı zamanda bilgisayar bilimini, makine öğrenmesini, veritabanı yönetimini, matematiksel algoritmaları ve istatistiği birleştiren disiplinler arası bir alandır (Emel ve Taşkın, 2005).

Veri madenciliği, büyük veri tabanlarından, yararlı bilgilerin otomatik olarak çıkarılması sürecidir. Veri madenciliği, gelecek trendleri tahmin eder ve davranışları belirler (Hudairy, 2004).

Veri madenciliği, büyük miktardaki veriden, anlamlı örüntüler ve kurallar keşfetme sürecidir (Linoff ve Berry, 2011).

Veri madenciliği, istatistiksel ve matematiksel teknikler ile örüntü tanıma teknolojilerinin kullanılarak, depolama ortamlarında sıkışmış bulunan büyük miktardaki verinin elenmesi ile anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilmesi sürecidir (Larose, 2005).

En basit tanımıyla veri madenciliği, veri içerisindeki yeni, gizli kalmış veya beklenmeyen örüntüleri bulmak için kullanılan faaliyetler bütünüdür (Marakas, 2003).

Veri madenciliği, büyük veri depolarındaki yararlı bilginin otomatik olarak keşfedilmesi sürecidir (Tan ve diğ., 2006).

Veri madenciliği, genellikle büyük ölçüdeki veri setlerindeki, bazı bilinmeyen veya gizli kalmış kuralların keşfine ve analizine yarayan yöntemler ve teknikler kümesidir. Kısaca veri madenciliği, veriden bilgi çıkarma sanatıdır (Tuffery, 2011).

Veri madenciliği, tek başına ham verinin sunamadığı bilgiyi ortaya çıkaran veri analizi sürecidir (Jacobs, 1999)

Veri madenciliği, önceden bilinmeyen, gizli, anlamlı ve yararlı örüntülerin, büyük ölçekli veri tabanlarından otomatik biçimde elde edilmesini sağlayan, veri tabanlarındaki özbilgi keşif ve analiz sürecidir (Karacan ve Yeşilbudak, 2010).

Veri madenciliği, istatistik, veritabanı teknolojisi, örüntü tanıma, makine öğrenme ve diğer alanlarla ilişkili olan bir disiplindir. Önceden tahmin edilemeyen ilişkileri bulmak için büyük veri tabanlarının ikincil analizi ile ilgilidir (Hand, 1998).

Veri madenciliği, büyük miktardaki veriden ilginç bilgi ya da örüntüleri çıkaran süreç veya yöntemi ifade eder (Han ve Kamber, 2006).

Veri madenciliği, büyük miktardaki veri setlerinde saklı durumda bulunan örüntü ve eğilimleri keşfetme işlemidir (Özekes ve Çamurcu, 2002).

Veri madenciliği, yüksek kapasitelere ve yüksek verimlilik ölçümlerine ulaşmak için ihtiyaç duyulan teknolojilerin anahtar bileşenidir (Kittler ve Wang, 1999).

Veri madenciliği, veri içindeki anlamlı örüntüleri otomatik veya yarı otomatik olarak keşfetme sürecidir (Witten ve Frank, 2005).

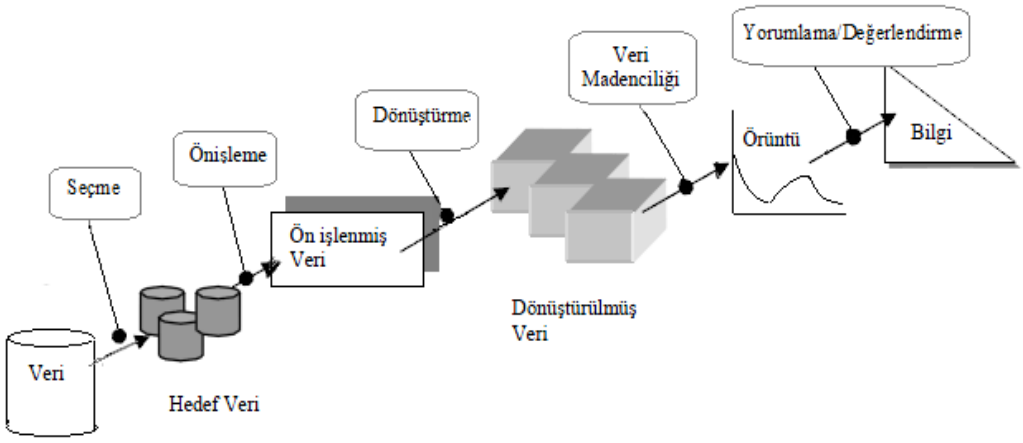
Veri madenciliği, veri ambarlarında yararlı olma potansiyeline sahip, aralarında beklenmedik, bilinmedik ilişkilerin olduğu verilerin keşfedilerek, hem anlaşılır hem de kullanılabilir bir biçime getirilmesine yönelik geliştirilmiş yöntemler topluluğudur (Köktürk ve diğ., 2009).

Veri madenciliği, veriden örüntüleri çekmek için özel algoritmaların kullanımını ifade eder (Fayyad ve diğ., 1996).

Yukarıdaki tanımlardan da anlaşılabileceği gibi veri madenciliği, geleceğe ait tahminlerin yapılabilmesi için büyük veri tabanlarındaki anlamlı, yeni ve gizli kalmış bilgilerin keşfedilerek çeşitli tekniklerle analiz edilmesi sürecidir.

Veri madenciliği, daha büyük bir süreç olarak adlandırılan bilgi keşfi sürecinin bir bölümüdür (Hudairy, 2004).

Şekil 1.1'de veri tabanında bilgi keşfi (VTBK) süreci ve bu sürecin bir parçası olan veri madenciliğine yer verilmiştir.




Şekil 1.1.VM'nin bilgi keşfi süreci içindeki yeri (Hudairy, 2004)

## 1.1. Veri Madenciliğinin Tarihsel Gelişimi

Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek için kullanılmaya başlanmasıyla ortaya çıkmıştır. Bu dönemlerde veri taraması, veri yakalanması gibi isimler verilmiş ve bilgisayar yardımıyla gerekli sorgulama yapıldığında istenilen bilginin elde edilebileceği düşünülmüştür. (Köktürk ve diğ., 2009).

1970'lerde İlişkisel Veri Tabanı Yönetim Sistemleri uygulamaları kullanılmaya başlanmıştır. Bilgisayar uzmanları bununla beraber basit kurallara dayanan uzman sistemler geliştirmişler ve basit anlamda makine öğrenimini sağlamışlardır. 1980'lerde veri tabanı yönetim sistemleri yaygınlaşmış ve bilimsel alanlarda, mühendisliklerde vb. alanlarda uygulanmaya başlanmıştır. Bu yıllarda şirketler, müşterileri, rakipleri ve ürünleri ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Bu veri tabanlarının içerisinde çok büyük miktarlarda veri bulunmaktadır ve bunlara SQL veri tabanı sorgulama dili ya da benzeri diller kullanarak ulaşılabilir. (Savaş ve diğ., 2012).

1990'larda bilgisayar mühendisleri, geleneksel istatistiksel yöntemlerinin yerine algoritmik bilgisayar modülleri ile veri analizinin değerlendirilebileceğini vurgulayarak, veri madenciliği ismini kullanmışladır. Bu yıllarda veri tabanlarındaki veri miktarları katlanarak arttığı için, büyük miktardaki veri içinden yararlı bilgilere nasıl ulaşılması gerektiği üzerinde düşünülmeye başlanmıştır ve VM için ilk yazılım gerçekleştirilmiştir. 2000'li yıllardan itibaren VM sürekli gelişmiş ve geniş bir yelpazede uygulanmaya başlanmıştır.

VM büyük miktardaki verilerin incelenmesini amaçladığı için veri tabanları ile yakından ilişkilidir. Günümüzde yaygın olarak kullanılmaya başlanılan veri ambarları, günlük kullanılan veri tabanlarının birleştirilmiş ve işlenmeye daha uygun durumdaki özetini saklamayı amaçlamaktadır. Günlük veri tabanlarından istenen özet bilgi seçilerek, gerekli ön işlemeden geçtikten sonra veri ambarlarında saklanmaktadır. Hedef doğrultusunda gerekli veriler, veri ambarlarından alınarak

VM için standart bir forma çevrilmektedir. Veri ambarlarının analizi için "Online Analytic Processing (OLAP)" programları kullanılır. OLAP, çok boyutlu veri analizini sağlamaya odaklanmıştır. (Fayyad ve diğ., 1996). Veri madenciliğinin tarihsel gelişim süreci Tablo 1.1'de gösterilmiştir.

Tablo 1.1. Veri madenciliğinin tarihsel gelişim süreci (Yapıcı ve diğ., 2010)

| Gelişim Adımları | Cevaplanan Karar Problemi | Kullanılabilen Teknolojiler | Ürün Sağlayıcıları | Karakteristikler |
|---|---|---|---|---|
| **Veri Toplama (1960'lar)** | "Benim toplam karım geçen 5 yılda ne kadardı?" | Bilgisayarlar, Teypler, Diskler | IBM, CDC | Geriye dönük, statik veri dağıtımı |
| **Veri Erişimi (1980'ler)** | "İngiltere'de geçen Mart ayında birim satışlar ne kadardı?" | İlişkisel Veritabanları, SQL, ODBC | Oracle, Sybase, Informix, IBM, Microsoft | Kayıt düzeyinde, geriye dönük, dinamik veri dağıtımı |
| **Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)** | "İngiltere'de geçen Mart ayında birim satışlar ne kadardı?" | Olap, Çok Boyutlu Veritabanı Sistemleri, Veri Ambarları | Pilot, Comshare, Arbor, Cognos, Microstrategy | Çoklu düzeylerde geriye dönük, dinamik veri dağıtımı |
| **Veri Madenciliği (Bugün)** | "Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir, niçin?" | İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veritabanları | Pilot, Lockheed, IBM, SGI, SPSS, SAS, Microsoft vs. | Geleceğe dönük, proaktif enformasyon dağıtımı |

## 1.2.Veri Madenciliği Kullanım Alanları

Büyük hacimde veri bulunan her yerde VM kullanmak mümkündür. Günümüzde karar verme sürecine ihtiyaç duyulan birçok alanda VM uygulamaları yaygın olarak kullanılmaktadır. (Savaş ve diğ., 2012).

Veri madenciliğinin kullanım alanlarından bazıları şöyledir:

- Bankacılık
- Finans
- Perakendecilik
- Sigortacılık
- Borsa
- Telekomünikasyon

- Bilim ve Mühendislik
- Endüstri
- Sağlık
- Eğitim
- Seyahat/ Konaklama
- Reklamcılık
- Güvenlik
- Web sitesi analizi
- Elektronik Ticaret

Tablo 1.2'de 2010 ve 2011 yılında veri madenciliğinin sektörel bazda kullanım oranlarına ait araştırma sonuçları verilmiştir. Araştırmaya göre 2010 ve 2011 yılında veri madenciliğinin en çok kullanıldığı alan Müşteri İlişkileri Yönetimi olmuştur. 2010 yılında bu alanda veri madenciliğinin kullanım oranı %26,8 iken, 2011 yılında bu oranın %25 olduğu görülmektedir.

Veri madenciliğinin en çok kullanıldığı alanlar sıralamasında ikinci sırada 2010 yılında %19,2'lik ve 2011 yılında %18,9'luk oranla Bankacılık sektörünün yer aldığı görülmektedir.

Sağlık sektörü, veri madenciliğinin kullanım alanları sıralamasında 2010 ve 2011 yılındaki verilere göre üçüncü sırada yer almaktır. Tüketici analitiği ve Bankacılık sektörünün aksine, Sağlık sektöründe 2011 yılında veri madenciliği kullanım oranının bir önceki yıla göre daha yüksek olduğu görülmektedir. 2010 yılında bu oran %13,1 iken, 2011 yılında %16,2'dir.

2010 yılından 2011 yılına kadar en büyük artışlar, sırasıyla; Seyahat (429%), Sosyal Ağlar (% 100), Eğitim (65%), Biyoteknoloji (% 64) ve Kredi Skorlama (% 59) alanlarında görülmüştür.

2010 yılından 2011 yılına kadar VM kullanım oranlarındaki en büyük düşüşler ise; İmalat(-34%), Reklam ( -29%), e-Ticaret (-25%), Yatırım / Stoklar (-22%) ve Web kullanım madenciliği (-21%) alanlarında görülmüştür.

Tablo 1.2. 2010 ve 2011 yıllarında veri madenciliğinin uygulandığı alanlar (URL-1)

| 2011% 2010% | |
|---|---|
| CRM / Tüketici analitiği (57) | 25.0%<br>26.8% |
| Bankacılık (43) | 18,9%<br>19,2% |
| Sağlık / HR (38) | 16.7%<br>13.1% |
| Eğitim (37) | 16.2%<br>9.9% |
| Sahtekarlık Algılama (32) | 14,0%<br>12,7% |
| Bilim (31) | 13.6%<br>10.3% |
| Sosyal Ağlar (30) | 13.2%<br>6.6% |
| Kredi Skorlama (29) | 12.7%<br>8.0% |
| Doğrudan Pazarlama / Kaynak Yaratma (28) | 12,3%<br>11,3% |
| Sigorta (28) | 12,3%<br>10,3% |
| Finans (26) | 11.4%<br>11.3% |
| Telekom / Kablo (25) | 11,0%<br>10,8% |
| Perakende (24) | 10.5%<br>8.0% |
| Tıp / İlaç (22) | 9.6%<br>8.0% |
| Genomik / Biyoteknoloji (21) | 9.2%<br>5.6% |
| Devlet / Askeri (17) | 7.5%<br>6.1% |
| Seyahat / Konaklama (17) | 7.5%<br>1.4% |
| Reklamcılık (16) | 7.0%<br>9.9% |
| Web kullanım madenciliği (16) | 7.0%<br>8.9% |
| Yazılım (16) | 7.0%<br>0.0% |
| E-Ticaret (12) | 5.3%<br>7.0% |
| İmalat (12) | 5.3%<br>8.0% |
| Arama / Web içerik madenciliği (12) | 5.3%<br>6.6% |
| Yatırım / Stoklar (10) | 4.4%<br>5.6% |
| Eğlence / Müzik / TV / Filmler (8) | 3.5%<br>3.3% |
| Güvenlik / Anti-terör (4) | 1.8%<br>1.9% |
| Sosyal Politika / Anket analiz (4) | 1.8%<br>0.9% |
| Önemsiz e-posta / Anti-spam (3) | 1.3%<br>0.9% |
| Diğer (17) | 11.7%<br>7.5% |

Veri madenciliğinin çeşitli alanlardaki kullanım amaçları aşağıdaki gibidir:

• Pazarlama alanında veri madenciliği kullanım amaçları;

Mevcut müşterilerin elde tutulması ve yeni müşterilerin kazanılması, pazar sepeti analizi, satış tahmini, müşteri ilişkileri yönetimi, çapraz satış analizi, tüketicilerin demografik özellikleri arasında bağıntı kurulması, müşteri değer analizi, müşterilerin satın alma örüntülerinin belirlenmesi,

• Borsa alanında veri madenciliğinin kullanım amaçları;

Genel piyasa analizi, hisse senedi fiyatlarının belirlenmesi,

• Bankacılık ve sigortacılık alanında veri madenciliğinin kullanım amaçları;

Sadık müşteri portföyünün oluşturulması, kredi kartı dolandırıcılıklarının tespiti, kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi, kredi taleplerinin değerlendirilmesi, kredi geri ödemelerinin kontrol altında tutulması, çapraz satış ile birim müşteriye yapılan satış miktarının artırılması, müşterilere özgü satış politikalarının oluşturulması, riskli müşteri tiplerinin belirlenmesi, sigorta dolandırıcılıklarının tespiti, yeni poliçe talep edeceklerin belirlenmesi,

• Tıp alanında veri madenciliğinin kullanım amaçları;

Tedavi süreçlerinin belirlenmesi, hasta tepkilerinin tahmin edilip karakterize edilmesi, genetik hastalıkların tespiti, yeni virüs türlerinin keşfi, test sonuçlarının tahmin edilmesi,

• Telekomünikasyon alanında veri madenciliğinin kullanım amaçları;

Hatların yoğunluk tahminleri, servis kalitesinin artırılması, ağ performanslarının yönetimi, kalite ve iyileştirme analizleri.

• Endüstri alanında veri madenciliğinin kullanım amaçları;

Üretim süreçlerinin optimizasyonu, lojistik, kalite kontrol analizleri,

• Eğitim alanında veri madenciliğinin kullanım amaçları;

Öğrencilerin karakteristik özelliklerine göre uygulanacak eğitim modelinin belirlenmesi, eğitimde verimlilik artışını sağlayacak değişikliklerin tespiti,

### 1.3.Veri Madenciliği Örnek Uygulamaları

Veri madenciliği uygulamaları aşağıdaki gibi gruplandırılabilir (URL-2):

• Bağıntı: Amaç mallar arasındaki pozitif veya negatif korelasyonları belirlemektir. Sepet analizinde müşterilerin beraber satın aldığı malların analizi yapılır. Örneğin, "çocuk bezi alan müşterilerin %30'u bira da satın alır." Çocuk bezi alan müşterilerin, mama da satın alacağını veya bira satın alanların cips de alacağını tahmin edebiliriz ancak otomatik bir analiz bütün olasılıkları göz önüne alır ve çocuk bezi ile bira arasındaki gibi kolay düşünülemeyecek bağıntıları da bulmamızı sağlar.

• Sınıflandırma: Amaç bir malın özellikleri ile müşteri özelliklerini eşleştirmektir. Böylece bir müşteri için ideal ürün veya bir ürün için ideal müşteri profili çıkarılabilir. Burada önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. Örneğin bir otomobil satıcısı şirket, geçmiş müşteri hareketlerinin analizi ile, "genç kadınlar küçük araba satın alır, yaşlı, zengin erkekler büyük, lüks araba satın alır" gibi iki kural bulursa genç kadınların okuduğu bir dergiye reklam verirken küçük modelinin reklamını verir.

• Regresyon: "Ev sahibi olan, evli, aynı iş yerinde beş yıldan fazladır çalışan, geçmiş kredilerinde geç ödemesi bir ayı geçmemiş bir erkeğin kredi skoru 825'dir" gibi bağımlı ve bağımsız değişkenler arasındaki ilişkinin çıkarımı söz konusudur. Başvuru skorlamada bir finans kurumuna kredi için başvuran kişi ile ilgili finansal güvenilirliğini notlayan örneğin bir skor hesaplanır. Bu skor kişinin özellikleri ve geçmiş kredi hareketlerine dayanılarak hesaplanır.

• Zaman içinde sıralı örüntüler: "İlk üç taksitinden iki veya daha fazlasını geç ödemiş olan müşteriler %60 olasılıkla kanuni takibe gidiyor" gibi sonuçlar elde edilir. Davranış skoru, başvuru skorundan farklı olarak kredi almış ve taksitleri ödeyen bir kişinin sonraki taksitlerini ödeme veya geciktirme davranışını notlamayı amaçlar.

• Benzer zaman sıraları: "X şirketinin hisse fiyatları ile Y şirketinin hisse fiyatları benzer hareket ediyor" gibi zaman içindeki iki hareket serisi arasında bağıntı kurmayı amaçlar. İki malın zaman içindeki satış miktarlarını örnek verecek olursak, dondurma satışları ile kola satışları arasında pozitif, dondurma satışları ile salep satışları arasında negatif bir bağıntı beklenebilir.

• İstisnalar (Fark saptanması): Amaç önceki uygulamaların aksine kural bulmak değil, kurala uymayan istisnai hareketleri bulmaktır. Örneğin, "normalden farklı davranış gösteren müşterilerim var mı?" sorusuna cevap aranarak, olası sahtekarlıkların saptanması sağlanabilir. Visa kredi kartı için yapılan CRIS sisteminde bir yapay sinir ağı, kredi kartı hareketlerini takip ederek müşterinin normal davranışına uymayan hareketler için müşterinin bankası ile temasa geçerek müşteri onayı istenmesini sağlamaktadır.

• Doküman madenciliği: Veri madenciliği teknikleri ile yazılı belgeler arasındaki ilişkileri bulmayı hedefler. Dokümanlar arasında ayrıca elle bir tasnif gerekmeden benzerlik hesaplayabilmeyi sağlar. Bu amaçla genellikle otomatik olarak çıkarılan anahtar sözcüklerin tekrar sayısı kullanılır. Doküman madenciliği, "arşivimde veya internet üzerinde bu dokümana benzer hangi dokümanlar var?" gibi sorulara cevap bulmamıza yardımcı olur. Günümüzde yaygın olarak kullanılan internet arama motorları, doküman madenciliğini kolaylaştırmıştır.

### 1.4. Veri Madenciliği Uygulamalarında Karşılaşılan Problemler

VM girdi olarak kullanılacak ham veriyi veritabanlarından alır. Bu da veritabanlarının dinamik, eksiksiz, geniş ve net veri içermemesi durumunda sorunlar doğurabilir. Diğer sorunlar da verinin konu ile uyumsuzluğundan doğabilir. Sınıflandırmak gerekirse başlıca sorunlar aşağıdaki gibidir (Akbulut, 2006).

• Sınırlı bilgi: Veritabanları genel olarak veri madenciliği dışındaki amaçlar için tasarlanmışlardır. Bu nedenle, öğrenme görevini kolaylaştıracak bazı özellikler bulunmayabilir.

• Gürültü ve kayıp değerler: Veri özellikleri ya da sınıflarındaki hatalara gürültü adı verilir. Bu hataların sonucu olarak veri tabanında birçok niteliğin değeri yanlış

olabilir. Bu bilgi yanlışlığı, ölçüm hatalarından, ya da öznel yaklaşımdan olabilir. Veri tabanlarındaki eksik bilgi ve bu yanlışlardan dolayı veri madenciliği amacına tam olarak ulaşmayabilir. Bu yüzden, veri madenciliği tekniklerinin gürültülü verilere karşı daha az duyarlı olmalı. Diğer bir ifadeyle, sistem tarafından gürültülü verilerin tanınmaması ve ihmal edilmesi gerekmektedir.

• Artık veri: Verilen veri kümesi, eldeki probleme uygun olmayan veya artık nitelikler içerebilir. Bu durum pek çok işlem sırasında karşımıza çıkabilir. Örneğin, eldeki problem ile ilgili veriyi elde etmek için iki ilişkiyi ortak nitelikler üzerinden birleştirirsek, sonuç ilişkide kullanıcının farkında olmadığı artık nitelikler bulunur. Artık nitelikleri elemek için geliştirilmiş algoritmalar özellik seçimi olarak adlandırılır (Sever ve Oğuz, 2002).

• Boş değerler: Null ifadesiyle de tanımlanabilir. Bu kavram verinin içeriğinin bilinmemesi anlamını taşımaktadır. Boş değerler SQL sorgularında da ele alınması gereken özel değerlerdir. Veri madenciliğinde boş değerler iki yolla ele alınabilir:

1. Boş değerli veriler yok sayılarak, algoritma içinde ihmal edilirler,
2. Boş değerler, olası bir değerle değiştirilebilir.

• Ebat, güncellemeler ve konu dışı sahalar: Veri tabanlarındaki bilgiler, veri eklendikçe ya da silindikçe değişebilir. Veri madenciliği perspektifinden bakıldığında, kuralların hala aynı kalıp kalmadığı ve istikrarlılığı problemi ortaya çıkar. Öğrenme sistemi, kimi verilerin zamanla değişmesine ve verinin zamansızlığına karşın zaman duyarlı olmalıdır.

### 1.5.Veri Madenciliği Süreci

Veri madenciliği, aynı zamanda bir süreçtir. Veri yığınları arasında, soyut kazılar yaparak veriyi ortaya çıkarmanın yanı sıra, bilgi keşfi sürecinde örüntüleri ayrıştırarak bir sonraki adıma hazır hale getirmek de bu sürecin bir parçasıdır. Üzerinde inceleme yapılan işin ve verilerin özelliklerinin bilinmemesi durumunda ne kadar etkin olursa olsun hiç bir veri madenciliği algoritmasının fayda sağlaması mümkün değildir. Bu nedenle, veri madenciliği sürecine girilmeden önce, başarının

ilk şartı, iş ve veri özelliklerinin detaylı analiz edilmesidir. Veri madenciliği sürecinde izlenen adımlar genellikle aşağıdaki şekildedir (Savaş ve diğ., 2012):

1. Problemin tanımlanması,
2. Verilerin hazırlanması,
3. Modelin kurulması ve değerlendirilmesi,
4. Modelin kullanılması,
5. Modelin izlenmesi.

### 1.5.1. Problemin tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın en önemli şartı, uygulamanın hangi amaç için yapılacağının ve elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceğinin tanımlanmasıdır. Bu nedenle, veri madenciliği çalışmalarında öncelikli olarak amaç açık bir şekilde ortaya konulmalı ve durum değerlendirmesi yapılmalıdır.

### 1.5.2. Verilerin hazırlanması

Örneklem kümesi elde edildikten sonra, örneklem kümesinde yer alan hatalı kayıtların çıkarıldığı ve eksik nitelik değerlerinin değiştirildiği aşamadır. Bu aşama seçilen veri madenciliği sorgusunun çalışma zamanını iyileştirir. Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olmaktadır. Veri madenciliğinin en önemli aşamalarından biri olan verinin hazırlanması aşaması, analistin toplam zaman ve enerjisinin %50 - %85 ini harcamasına neden olmaktadır (Çil, 2010).

Verilerin hazırlanması; toplama, birleştirme ve temizleme, dönüştürme ve indirgeme aşamalarından oluşmaktadır.

Veri toplama: Problem için gereken verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi aşamasıdır. Veri toplama aşamasında analist, kendi veri kaynaklarını kullanabileceği gibi, farklı veri tabanlarından da faydalanabilmektedir.

Veri birleştirme ve temizleme: Bu aşamada, toplanan veriler arasında farklılık yaratan, hatalı veya analizin yanlış yönlenmesine sebep olabilecek verilerin

temizlenmesine çalışılır. Bu durum, veri madenciliği sürecinin hızının ve doğruluğunun gelişmesine katkı sağlar. Veri temizleme işlemi ile verideki eksik değerler doldurularak, yanlış değerler giderilir ve tutarsızlıklar düzeltilmeye çalışılır.

Veri Dönüştürme: Kullanılacak model ve algoritma çerçevesinde verilerin tanımlama veya gösterim şeklinin de değiştirilmesi gerekebilir. Veri dönüştürmede, veriler madencilik için uygun olan formlara dönüştürülür veya birleştirilir. Veri dönüştürme aşağıdakileri içerebilir (Bilen, 2009):

- Düzleştirme: Veriden hatalı uç değerlerin silinmesi için çalışır.
- Bütünleştirme: Özetleme veya bütünleştirme işlemlerinin veriye uygulanmasıdır.
- Genelleştirme: Verilerin genelleştirilmesinde alt seviye veri veya ham veri, kavram hiyerarşilerinin kullanılmasıyla daha yüksek seviyelerle değiştirilir.
- Normalizasyon: Bir özelliğe ait veri, normalizasyon ile küçük tanımlanmış bir aralığa düşecek şekilde ölçeklenir.
- Alan Yapılandırma: Veri madenciliği sürecine yardım etmek için verilen alanlar setinden yeni alanlar yapılandırılır ve eklenir.

Veri İndirgeme: Büyük veri tabanları ile yapılan veri madenciliği çalışmalarında çözümleme işlemi çok uzun sürebilir. Orijinal verinin bütünlüğü korunarak, elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir. Bu durum veri indirgeme olarak ifade edilmektedir.

### 1.5.3. Modelin kurulması ve değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle model kurma aşaması, en iyi olduğu düşünülen modele varılıncaya kadar tekrarlanan bir süreçtir.

Bir modelin doğruluğunun test edilmesinde kullanılan en basit yöntem basit geçerlilik testidir. Bu yöntemde tipik olarak verilerin % 5 ile % 33 arasındaki bir kısmı test verileri olarak ayrılır ve kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra, bu veriler üzerinde test işlemi yapılır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile

hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır. (Akpınar, 2000).

Doğruluk Oranı = 1 - Hata Oranı olarak bulunur.

Değerlendirme aşamasında, uygun model ya da modeller kurulduktan sonra, veri madenciliği sonuçlarının araştırma probleminin amaçlarını gerçekleştirip gerçekleştirmediği değerlendirilir. Bu aşama sonuçların değerlendirilmesi, veri madenciliği sürecinin gözden geçirilmesi ve sonraki adımların ne olacağı hususlarını içermektedir. Bu aşamanın sonunda veri madenciliği sonuçlarının kullanımı üzerindeki karara varılmaktadır (Albayrak ve Yılmaz, 2009).

### 1.5.4. Modelin kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi kurumsal uygulamalarda doğrudan kullanılabileceği gibi, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir (Akpınar, 2000).

### 1.5.5. Modelin izlenmesi

Veri madenciliği sürecinin son aşaması, geçerliliği kabul edilen ve kullanılan modelin izlenmesidir. Zaman içerisinde bütün sistemlerin özelliklerinde ve ürettikleri verilerde ortaya çıkan değişiklikler sebebiyle, kurulan modeller sürekli olarak izlenmeli ve gerekirse yeniden düzenlenmelidir.

## 1.6. Bankacılık Alanında Gerçekleştirilen Veri Madenciliği Uygulamalarına Yönelik Literatür Taraması

"Türkiye'de Yerli ve Yabancı Ticaret Bankalarının Finansal Etkinliğe Göre Sınıflandırılması" konulu çalışmada (Albayrak, 2009), yerli ve yabancı olarak önceden grup üyeliği belirlenmiş bankaların sınıflandırmasında yaygın olarak kullanılan veri madenciliği tekniklerinden diskriminant, lojistik regresyon ve karar ağacı modelleri bankalarla ilgili seçilmiş likidite, gelir-gider, karlılık ve faaliyet oranları kullanılarak karşılaştırılmıştır.

Araştırmanın sonuçları, bankaların sınıflandırmasında karar ağacı modelinin geleneksel diskriminant ve lojistik regresyon modellerine üstünlük sağlayarak alternatif etkili bir sınıflandırma teknigi olarak kullanılabileceğini göstermiştir.

"Kredi Kartı Kullanan Müşterilerin Sosyo Ekonomik Özelliklerinin Kümeleme Analiziyle İncelenmesi" adlı çalışmada (Aşan, 2007), kredi kartı kullanan müşterilerin sosyo-ekonomik özelliklerinin gruplanması amaçlanmıştır. Çalışmada öncelikle bireysel bankacılık ve onun bir işlevi olan kredi kartlarının tanımlanmasına, bu kavramların ülkemizdeki yeri ve öneminin belirlenmesine yer verilerek, kredi kartı kullanan banka müşterileri kümeleme analiziyle gruplandırılmıştır. Uygulamada, verilere en uygun teknik olduğu için kümeleme analizinin hiyerarsik olan yöntemlerinden ortalamalar bağlantı tekniği tercih edilmiştir. Bu yöntemle ilgili banka müşterileri sosyo-ekonomik özelliklerine göre üç kümede gruplanmışlardır. İlk kümede en yoğun müşteri topluluğu bulunurken, ikinci kümede daha az müşteri topluluğu yer almış, üçüncü kümede ise azınlıkta olan müşteri grubu yer almıştır. Bu üç kümeye göre müşterilerin on adet sosyo-ekonomik değişkene göre faklılık gösterdiği gözlemlenmiştir. Çalışmanın, sosyo-ekonomik özelliklere göre belli bir müşteri grubunun çeşitli kümelerde gruplanarak, ilgili müşterilere verilecek bireysel bankacılık hizmetlerinde ne tür müşteriyle karşılaşabileceğini bilmek açısından fayda sağladığı, aynı zamanda ileride yapılacak bireysel bankacılıktaki kredi kartı pazarlamasına yönelik planlamalarda ne tür müşterilerin hangi özelliklere ve motivasyonlara sahip olduğunu bilmek açısından da önem arz ettiği vurgulanmıştır.

"Bankacılık Sektöründe Personel Seçimi ve Performans Değerlendirilmesine İlişkin Veri Madenciliği Uygulaması" (Bilen, 2009), adlı çalışmada, bankacılık çalışan satış personellerinin performansları değerlendirilmiş, kümeleme yöntemlerinden k ortalama ile personellerin performans başarı düzeylerine göre sınıflandırılması sağlanmıştır. Elde edilen performans düzeyleri daha sonra sınıflandırma ile karar kuralları oluşturmada çıktı olarak kullanılmıştır. Çalışanların yaş, medeni hal, cinsiyet gibi demografik bilgileri, öğrenim durumu, yabancı dili, SPK belgesi gibi eğitim durumlarına ilişkin bilgileri, çalıştığı şubesine ve iş yaşamındaki pozisyonuna ilişkin bilgileri dikkate alınarak veri madenciliğinde sınıflandırma algoritmaları kullanılmıştır.

WEKA' da gerçekleştirilen madencilik uygulamasında bazı sınıflandırma algoritmaları karşılaştırılmıştır. WEKA çıktılarına göre ID3 algoritması hatalı sınıflandırılan kayıt oranı ve ortalama mutlak hata açısından en iyi sonucu sağlamış ve ID3 algoritmasının sonuçları üzerinde durulmuştur. Karar ağacı algoritmalarıyla elde edilen karar kuralları ile her ildeki personelin performans başarı düzeyleri belirlenmiş, böylece yöneticilerin personel değerlendirme ve personel seçimi sürecinde karar kurallarına sahip olması sağlanarak personel seçimi ve performans değerlendirme sürecinde fayda sağlanmıştır. Veri madenciliği uygulaması neticesinde çalışanların performanslarına göre değerlendirilmesi yapılmış, hangi özelliklerdeki personelin hangi şubede ne oranda başarılı olduğuna yönelik kurallar oluşturulmuştur. Bu kurallar dikkate alınarak, bir personelin özelliklerine göre hangi şubelere atanabileceği ya da ataması düşünülen şubede hangi düzeyde performans gösterebileceğinin öngörülmesi hedeflenmiştir.

"Bankaların Gözetiminde Bir Araç Olarak Kümeleme Analizi" konulu çalışmada (Doğan, 2008), Türk Bankacılık Sektörü" nde (1998–2006) dönemi itibariyle faal olan ticaret bankalarına ait finansal oranlar temel alınarak Kümeleme Analizi uygulamasına yer verilmiştir. Uygulama sonuçlarının bankalar için yapılan finansal analiz sonuçları ile uyumluluğu tartışılarak, elde edilen sonuçlar ışığında Kümeleme Analizi tekniğinin bankaların finansal performanslarını belirlemek ve finansal açıdan benzer bankaları tanımlamak amacıyla, bankaların gözetiminde kullanılan mevcut teknikleri tamamlayıcı bir teknik olarak kullanılabilirliği incelenmiştir.

"Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi" konulu çalışmada (Tosun, 2006), kredi kartı müşterilerinin kaybedilme sebeplerinin bulunabilmesi için veri madenciliği yöntemlerinden faydalanarak sonuçlara ulaşmak amaçlanmıştır. Böylece, müşterinin neden kaybedildiği bilgisinin yanı sıra, hangi tür müşterilerin daha sık kaybedildikleri tahmin edilmeye çalışılmıştır. Karar ağacı uygulamasında denenen kurallardan sonra, karar ağaçlarında eşik değeri kullanıldığında, eşik değeri arttıkça, kullanılan niteliklerin sayısının azalacağı, son hesap hareketi tarihi 12. ayken müşteri son 3 ayda hiç alışveriş yapmamışsa genel olarak kaybedilme olasılığının oldukça yüksek olduğu, son hesap hareket tarihi 10.ay olan bir müşteri, ilk kez hesabını 2005 yılı ve sonrasında açtırdıysa, bu müşterinin kaybedilme olasılığının düşük olduğu gibi sonuçlara ulaşılmıştır.

"Banka Yatırım Fonu Müşteri Hareketlerinin Belirlenmesine Yönelik Bir Veri Madenciliği Uygulaması" konulu çalışmada (Çil, 2010), bir bankanın mevcut fonlarını alıp satan ve belli bir işlem geçmişinden sonra bankadaki hesabını kapatarak banka yatırım fonu müşterisi olmaktan çıkmış müşterilerin, işlem hareket detayının öğrenilmesi, bu işlem hareket detaylarını sergileyerek yatırım hesabını kapatmış müşterilerin sosyo-demografik karakteristiğinin çıkartılması ve bundan sonra hesabını kapatmaya meyilli müşterilerin tespit edilerek kaybedilmesinin önlenmesi üzerinde durulmuştur. Yatırım hesabını kapatarak banka müşterisi olmaktan çıkmış müşterilerin hangi işlem hareket ile hareket ederek yatırım hesabını kapattığı, bu hareketi gösteren müşterilerin sosyo-demografik karakteristiğinin ne olduğu belirlenmiştir. Sonuç olarak, tespit edilen sosyo-demografik ve yatırım fonu işlemi yapma özellikleri ile bankada hesabını kapatmaya yönelen müşterilerin tespit edilebileceği, çeşitli tutundurma faaliyetleri ile proaktif davranılarak müşteri kaybının yaşanmasının engellenebileceği görüşüne ulaşılmıştır.

"Veri Madenciliğinde Sınıflandırma Yöntemlerinin Karşılaştırılması" konulu çalışmada (Çakır, 2008), veri madenciliği standart sürecinin tüm aşamaları bankacılık müşteri veri tabanından rastlantısal olarak seçilmiş veri kümesi üzerinde uygulanmış ve veri madenciliğinin sınıflandırma fonksiyonu üzerinde durulmuştur. Uygulama, birden çok bağımlı değişken üzerinde birden çok sınıflandırma tekniğini kullanarak bu tekniklerin karşılaştırılması üzerine kurgulanmıştır. Bu nedenle, veri madenciliğinin üç önemli bileşeni olan istatistik, yapay öğrenme ve veri tabanı teknolojilerini temsil edecek şekilde lojistik regresyon analizi, yapay sinir ağları ve C5.0 karar kuralı türetme algoritması uygulamada kullanılacak sınıflandırma teknikleri olarak belirlenmiştir. Bu tekniklerin çeşitli bankacılık ürünlerine sahip olma bilgisini içeren üç farklı kategorik değişken üzerinde uygulanması ile toplam dokuz farklı model geliştirilmiştir. Modellerin tarafsız bir şekilde karşılaştırılması için her bağımlı değişkene ilişkin tek bir veri kümesi kullanılmış ve karşılaştırma ölçütleri olarak hız, ölçeklenebilirlik, sınıflandırma kesinliği ve öngörü kesinliği kullanılmıştır. Hız ölçütü açısından yapılan değerlendirmede, C5.0 algoritmasının tartışmasız bir şekilde avantaj sağladığı görülmüştür. Ölçeklenebilirlik açısından yapılan değerlendirmede, yapay sinir ağları ve C5.0 algoritmasının veri sayısına daha az duyarlı olduğu, lojistik regresyon tekniğinin ise veri sayısındaki artıştan

etkilendiği gözlemlenmiştir. Modellerin, geliştirildikleri veri kümesi üzerinde gösterdikleri sınıflandırma başarısının bir ölçüsü olan, sınıflandırma kesinliği açısından anlamlı bir farklılık göstermedikleri görülmüştür. Sonuç olarak, veri madenciliği sürecinin en zorlu kısmının veri hazırlama aşaması olduğu, veri sayısının ve veri kalitesinin uygulamaların başarısında önemli birer faktör olduğu, güncel ve hızlı karar verme ihtiyaçları doğrultusunda en uygun seçimin C5.0 algoritması olacağı görüşü ağırlık kazanmıştır.

# 2. VERİ MADENCİLİĞİ MODELLERİ

Veri madenciliğinde kullanılan modeller, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında incelenmektedir. Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır (Akpınar, 2000).

Veri madenciliğinde tahmin edici modeller ile örüntü tanıma işi sınıflama, regresyon ve zaman serileri yaklaşımlarını içerir. Bu modeller, neyin tahmin edilmesinin istendiğine dayalı olarak farklılaşırlar. Çıktı niteliğinin sürekli değerleri için tahmin istenir ise regresyon analizi, zamanın ayırt edici özellikleri ile ilgileniliyor ise zaman serileri, iyi veya kötü gibi az sayıdaki ayrık kategoriye sahip bir özel veri öğesi için bir tahmin yapılmak isteniyor ise sınıflama gerekir. Eldeki verinin gruplarını bulan kümeleme, birliktelik ve ardışıklık kurallarını elde etmeyi kapsayan birliktelik analizi ve ardışıklık keşfi davranışı ise tanımlama amaçlı kullanılır (Emel ve Taşkın, 2005).

Veri madenciliği modellerini işlevlerine göre üç ana grup altında toplamak mümkündür:

1. Sınıflama (Classification) ve Regresyon (Regression),
2. Kümeleme (Clustering),
3. Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns).

Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir. (Albayrak ve Yılmaz, 2009).

## 2.1. Sınıflama ve Regresyon

Sınıflama ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen analiz yöntemleridir. Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Örneğin, bir sınıflama modeli banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli geliri ve mesleği verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir (Özekes ve Çamurcu, 2002).

Sınıflandırma, bir veri öğesini, önceden tanımlı sınıflardan birine tasnif ederken, regresyon veri öğesini, gerçek değerli bir tahmini değişkene eşler (Fayyad, 1996).

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler:

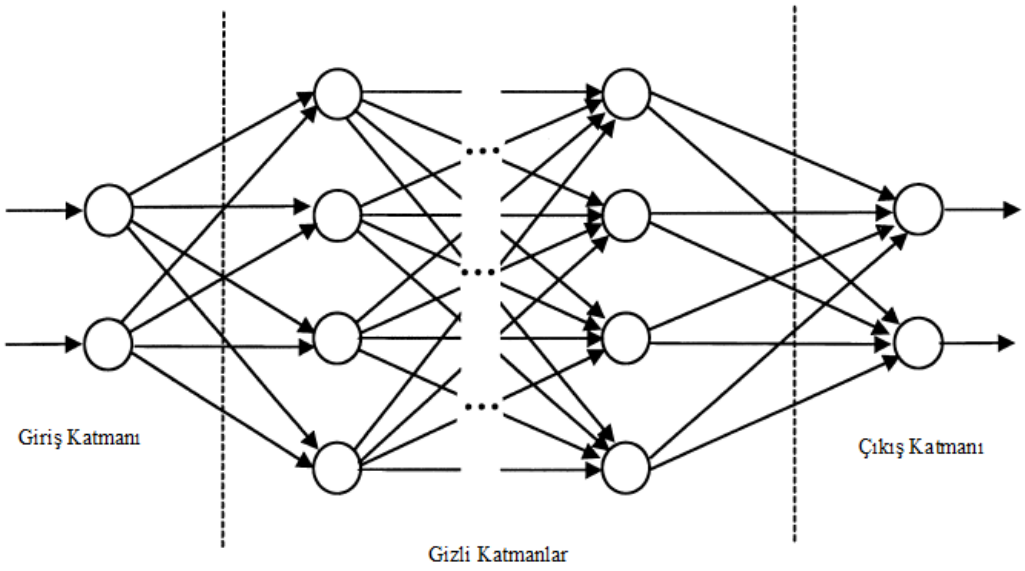
1. Yapay Sinir Ağları (Artificial Neural Networks),
2. Genetik Algoritmalar (Genetic Algorithms),
3. K- En Yakın Komşu (K- Nearest Neighbour),
4. Naive - Bayes sınıflayıcısı,
5. Lojistik Regresyon,
6. Karar Ağaçları (Decision Trees)'dır.

### 2.1.1. Yapay sinir ağları

Yapay sinir ağları öğrenme yeteneğine sahip, gelişmiş matematiksel yapıların hesaplanmasını içeren bir yaklaşımdır. Bu metot sinir sisteminin öğrenmesini model alan akademik araştırmaların bir sonucu olarak ortaya çıkmıştır. Sinir ağları karmaşık ve anlaşılması çok güç olan yapılardan anlam türetme becerisine sahip, dikkate değer yeteneklere sahiptir (Çetinyokuş, 2008).

YSA (Yapay Sinir Ağı), insan beyninin çalışma ilkelerinden ilham alınarak geliştirilmiştir. Ağırlıklı bağlantılar denilen tek yönlü iletişim kanalları vasıtası ile birbirleriyle haberleşirler ve her biri kendi hafızasına sahip birçok işlem elemanından (nöronlardan) oluşurlar. YSA'lar gerçek dünyaya ait ilişkileri tanıyabilir, sınıflandırma, kestirim ve işlev uydurma gibi görevleri yerine getirebilirler (Küçüksille, 2009).

Yapay sinir ağları genellikle bir giriş katmanı, gizli katmanlar ve bir çıkış katmanından oluşmaktadır. Basit şekliyle her bir nöron bir önceki katmanlardaki diğer nöronlara, sinaptik ağırlıkları yoluyla bağlanmaktadır (Kalogirou, 2000). Biyolojik sistemlerde öğrenme, nöronlar arasındaki sinaptik bağlantıların ayarlanması ile gerçekleşmektedir. Yapay sinir ağlarında ise öğrenme, girdi ve çıktı verilerinin işlenmesiyle, yani eğitme algoritmasının bu verileri kullanarak bağlantı ağırlıklarını bir yakınsama sağlanana kadar, tekrar tekrar ayarlamasıyla gerçekleşmektedir. Yapay sinir ağlarının katmanları ve işleyişi Şekil 2.1.'de gösterilmiştir.




Şekil 2.1. Yapay sinir ağlarının katmanları (Kalogirou, 2000)

### 2.1.2. Genetik algoritmalar

Genetik algoritmaların temel ilkeleri ilk kez John Holland tarafından ortaya atılmıştır. Holland evrim süreci kullanılarak, bilgisayara anlayamadığı çözüm yöntemlerinin öğretilebileceğini düşünmüş ve Genetik Algoritma (GA) bu düşüncenin bir sonucu olarak bulunmuştur (Çetinyokuş, 2008).

Genetik algoritmalar bir çözüm uzayındaki her noktayı, kromozom adı verilen ikili bit dizisi ile kodlar. Her noktanın bir uygunluk değeri vardır. Tek bir nokta yerine, genetik algoritmalar bir popülasyon olarak noktalar kümesini muhafaza etmektedir. Her kuşakta, genetik algoritma, çaprazlama ve mutasyon gibi genetik operatörleri

kullanarak yeni bir popülasyon oluşturmaktadır. Birkaç kuşak sonunda, popülasyon daha iyi uygunluk değerine sahip üyeleri içermektedir. Bu, Darwin'in rastsal mutasyona ve doğal seçime dayanan evrim modellerine benzemektedir (Emel ve Taşkın, 2002).

Genetik algoritmalar, çözümlerin kodlanması, uygunlukların hesaplanması, çoğalma, çaprazlama ve mutasyon işlemlerinin uygulanması gibi aşamaları içermektedir. Çözümlerin kodlanması aşamasında, tüm çözümlerin aynı boyutlara sahip bitler dizisi biçiminde gösterilmektedir. Popülasyondaki her üyenin uygunluk değeri hesaplanarak çoğalma aşamasına geçilmekte ve mevcut kuşaktan yeni bir popülasyon yaratılmaktadır. Mevcut gen havuzunun potansiyelini araştırmak için, bir önceki kuşaktan daha iyi nitelikler içeren yeni kromozomlar yaratmak amacıyla çaprazlama operatörü kullanılmakta ve genetik çeşitliliği korumak amacıyla mutasyon işlemi uygulanmaktadır. Tüm bu işlemlerden sonra yeni kuşak oluşturulmakta ve döngü durdurulmaktadır.

Genetik algoritmalar problemlerin çözümü için evrimsel süreci bilgisayar ortamında taklit ederler. Çözüm için tek bir yapının geliştirilmesi yerine, böyle yapılardan meydana gelen bir küme oluştururlar. Problem için olası pek çok çözümü temsil eden bu küme genetik algoritma terminolojisinde nüfus adını almaktadır. Nüfuslar vektör, kromozom veya birey adı verilen sayı dizilerinden oluşmaktadır. Birey içindeki her bir elemana gen denir. Nüfustaki bireyler evrimsel süreç içinde genetik algoritma işlemcileri tarafından belirlenmektedirler. Genetik algoritmalar yapısı gereği, kötü bireyleri yani uygun olmayan çözümleri, operatörleri sayesinde elemektedir. Bu işlemler bir döngü içerisinde durdurma kriteri sağlanana kadar devam etmektedir (Gülçe, 2010).

Genetik algoritmalar, çizelgeleme, tesis yerleşimi, hat dengeleme, atama ve optimizasyon problemlerinin çözümü ile finans, pazarlama ve üretim gibi alanlarda uygulanmaktadır.

### 2.1.3. K- en yakın komşu yöntemi

K En Yakın Komşu yöntemi, sınıflandırma problemini çözen denetimli öğrenme yöntemleri arasında yer almaktadır. Yöntemde; sınıflandırma yapılacak verilerin öğrenme kümesindeki normal davranış verilerine benzerlikleri hesaplanarak; en yakın olduğu düşünülen k verinin ortalamasıyla, belirlenen eşik değere göre sınıflara atamaları yapılır. Önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. Yöntemin performansını k-en yakın komşu sayısı, eşik değer, benzerlik ölçümü ve öğrenme kümesindeki normal davranışların yeterli sayıda olması kriterleri etkilemektedir (Çalışkan ve Soğukpınar, 2008).

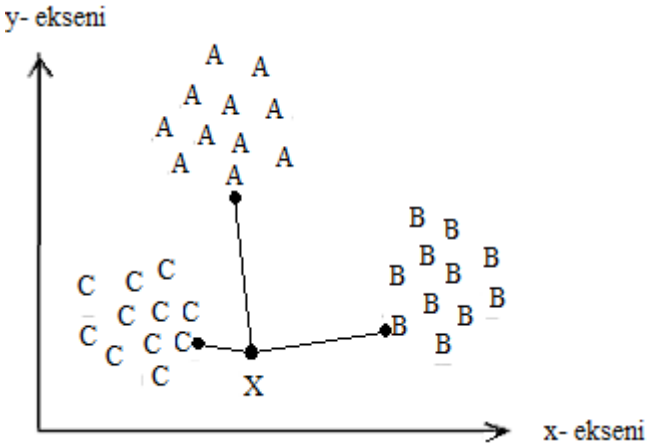
Bu metotta eğitim örnekleri n-boyutlu sayısal niteliklerle tanımlanır. Her örnek n-boyutlu uzayda bir noktayı gösterir. Bu yolla, bütün eğitim örnekleri n-boyutlu örnek uzayda depolanır (Kalıkov, 2006). k-en yakın komşu sınıflayıcısı, bilinmeyen örneğe en yakın k eğitim örneğini bulur. Yakınlık, $X = (x_1,x_2,\ldots,x_n)$ ve $Y = (y_1,y_2,\ldots\ldots,y_n)$ gibi iki nokta arasındaki öklid uzaklığı ile ifade edilmektedir (Bilen, 2009).

İki nokta arasındaki öklid uzaklığı Eşitlik (2.1)' deki gibi hesaplanmaktadır;

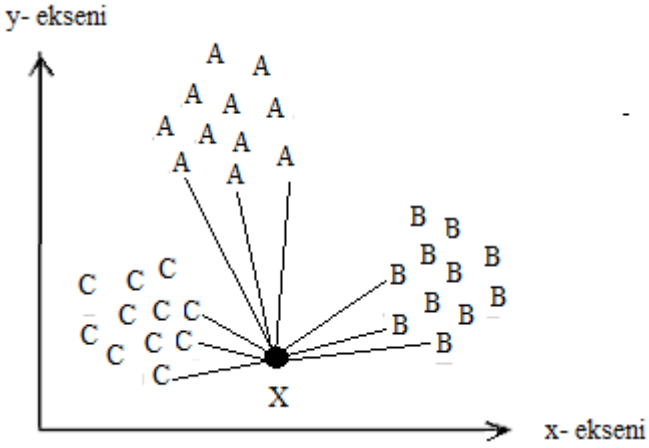
$$-\qquad (2.1)$$

Bilinmeyen örnek, k-komşularının arasında en sık olan sınıfa verilmektedir. k=1 olduğunda, bilinmeyen, örnek uzayında en yakın eğitim örneğinin sınıfına tanımlanmaktadır.

En Yakın Komşu metodu Şekil 2.2'deki örnekle gösterilmiştir. Bu örnekte X, sınıflandırılacak metodu; A, B ve C sınıfları temsil etmektedir. Görüldüğü üzere X dokümanının en yakın olduğu sınıf C sınıfıdır. Şekil 2.3'de ise k=3 olmak üzere K-En Yakın Komşu metodu gösterilmektedir. Burada X dokümanının A, B ve C sınıflarına en yakın olduğu 3 tane en yakın dokümanların uzaklıkları hesaplanır ve her bir sınıfın en yakın uzaklıkları toplamı o sınıfın X'e olan uzaklığını verir. Bu durumda X dokümanı C sınıfı olarak belirlenir (Aşlıyan ve Günel, 2010).




Şekil 2.2. K- en yakın komşu yöntemi




Şekil 2.3. k=3 için K- en yakın komşu yöntemi

### 2.1.4. Navie-Bayes sınıflayıcısı

Naive Bayes, temeli Bayes teorisine dayanan, verileri istatistiksel sınıflandırma tekniklerinden biridir. VM sınıflandırma algoritmalarından olan Bayes, uygulanabilirliği ve hızlı hesaplama performansı ile araştırmacılar tarafından öne çıkan bir algoritmadır. Sınıflandırılacak olayları birbirinden bağımsız olarak ele almaktadır (Olgun ve Özdemir, 2012).

Naïve Bayes Sınıflandırıcısı, örüntü tanıma problemi için kısıtlayıcı görülen bir önerme ile kullanılabilen olasılıkçı bir yaklaşımdır. Bu önerme, örüntü tanımada kullanılacak her bir tanımlayıcı nitelik ya da parametrenin istatistik açıdan bağımsız olmasıdır. Bu durum Naive Bayes sınıflandırıcısının kullanım alanını kısıtlandırsa da, genelde istatistik bağımsızlık koşulu esnetilerek kullanıldığında da daha karmaşık yapay sinir ağları gibi metotlarla karşılaştırabilir sonuçlar vermektedir.

Naive Bayes algoritmasının uygulanmasında bir takım kabuller yapılmaktadır. Bunların en önemlisi niteliklerin birbirinden bağımsız olmasıdır. Nitelikler birbirini etkilediği taktirde, burada olasılık hesaplamak zorlaşacağı için niteliklerin hepsinin aynı derecede önemli olduğu kabul edilmektedir.

Naive Bayes, sürekli veri ile çalışmaz. Bu nedenle sürekli değerleri içeren bağımlı ya da bağımsız değişkenler kategorik hale getirilmelidir. Naive Bayes, modelin öğrenilmesi esnasında, her çıktının öğrenme kümesinde kaç kere meydana geldiğini hesaplamaktadır. Bulunan bu değer, öncelikli olasılık olarak adlandırılır. Naive Bayes aynı zamanda her bağımsız değişken / bağımlı değişken kombinasyonunun meydana gelme sıklığını bulmaktadır. Bu sıklıklar öncelikli olasılıklarla birleştirilmek suretiyle tahminde kullanılır (Akbulut, 2006).

### 2.1.5. Lojistik regresyon

Lojistik regresyon, bağımlı değişkenin tahmini değerlerini olasılık olarak hesaplayarak olasılık kurallarına uygun sınıflama yapma imkanı veren bir istatistiksel yöntemdir. Lojistik regresyon analizinde üç temel yöntem mevcuttur (Özdamar, 2004a):

• İkili Lojistik Regresyon: İkili cevap içeren bağımlı değişkenlerle yapılan lojistik regresyon analizidir. Bir ya da daha fazla değişken ile ikili cevap değişkeni arasındaki bağıntıyı ortaya koyar.

• Sıralı Lojistik Regresyon: Cevap değişkenin sıralı ölçekli olduğu durumlarda uygulanan bir yöntemdir. Sıralı ölçekli cevap değişken, en az üç kategoride gözlenen değerler içermelidir.

• İsimsel Lojistik Regresyon: Cevap değişkenin isimsel ölçekli olduğu durumlarda uygulanan bir yöntemdir. Cevap değişkenin isimsel ölçekli olduğu durumlarda uygulanan bir yöntemdir.

Lojistik Regresyon Analizinin kullanım amacı, istatistikte kullanılan diğer model yapılandırma teknikleri ile aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen ve biyolojik olarak kabul edilebilir bir model kurmaktır (Coşkun ve diğ., 2004).

Lojistik regresyon modelleri, son yıllarda biyoloji, tıp, ekonomi, tarım ve veterinerlik ve taşıma sahalarında yaygın olarak kullanılmaktadır. Lojistik regresyon modellerinin yaygın bir şekilde kullanılır hale gelmesi, katsayı tahmin yöntemlerinin geliştirilmesi ve lojistik regresyon modellerinin daha ayrıntılı incelenmesine sebep olmuştur (Bircan, 2004).

### 2.1.6. Karar ağaçları ve karar ağacı algoritmaları

Karar ağaçları, sınıflandırma ve tahmin için sıkça kullanılan bir veri madenciliği yaklaşımıdır. Sinir ağları gibi diğer metodolojilerin de sınıflandırma için kullanılabilmesine rağmen karar ağaçları, kolay yorumu ve anlaşılabilirliği açısından karar vericiler için avantaj sağlamaktadır (Chien ve Chen, 2008).

Karar ağaçları;

- Düşük maliyetli olması,
- Anlaşılmasının, yorumlanmasının ve veri tabanları ile entegrasyonun kolaylığı,
- Güvenilirliklerinin iyi olması gibi nedenlerden ötürü en yaygın kullanılan sınıflandırma tekniklerinden biridir.

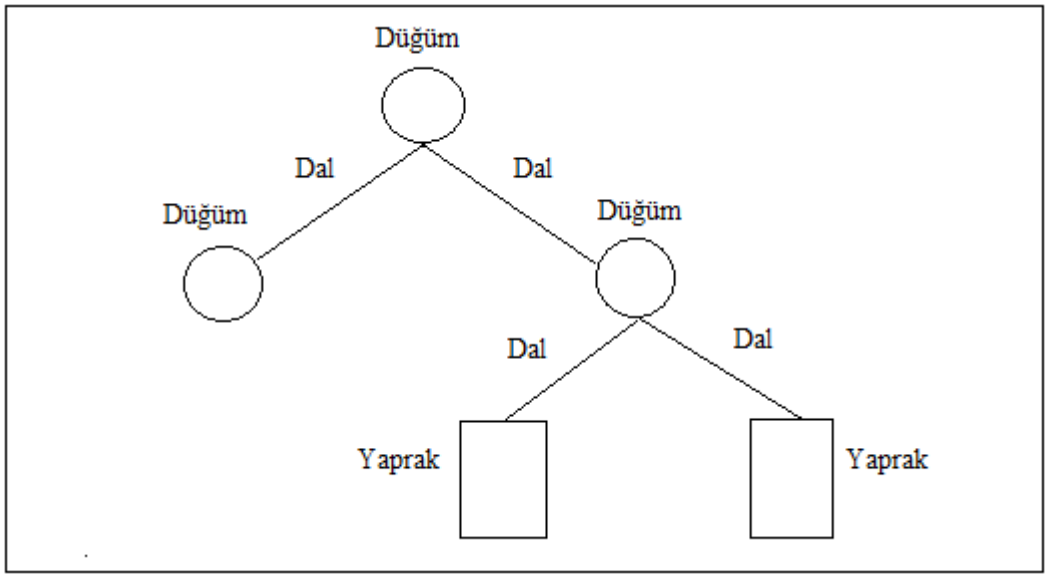
Karar ağaçlarının hedefi bağımlı değişkendeki farklılıkları maksimize edecek şekilde veriyi sıralı bir biçimde farklı gruplara ayırmaktır. Karar ağacı, adında belirtildiği şekilde ağaç görünümünde bir tekniktir. Karar düğümleri, dallar ve yapraklardan oluşmaktadır. Karar ağaçlarının yapısını oluşturan unsurlar (Argüden ve Erşahin, 2008):

- Karar düğümü: Veriye uygulanacak test tanımlanır. Her düğüm bir özellikteki testi gösterir. Test sonucunda ağacın dalları oluşur. Dalları oluştururken veri kaybı yaşanmaması için verilerin tümünü kapsayacak sayıda farklı dal oluşturulmalıdır.

- Dal: Testin sonucunu gösterir. Elde edilen her dal ile tanımlanacak sınıfın belirlenmesi amaçlanır. Ancak dalın sonucunda sınıflandırma tamamlanamıyorsa tekrar bir karar düğümü oluşur. Karar düğümünden elde edilen dalların sonucunda sınıflandırmanın tamamlanıp tamamlanmadığı tekrar kontrol edilerek devam edilir.

• Yaprak: Dalın sonucunda bir sınıflandırma elde edilebiliyorsa yaprak elde edilmiş olur. Yaprak, verileri kullanarak elde edilmek istenen sınıflandırmanın sınıflarından birini tanımlar.

Karar ağacı yapısı Şekil 2.4'te verilmiştir.




Şekil 2.4. Karar ağacının yapısı

Karar ağacı tekniğini kullanarak verinin sınıflanması, öğrenme ve sınıflama olmak üzere iki basamaklı bir işlemdir. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. Sınıflama basamağında ise test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır. Eğitim verisindeki hangi alanların, hangi sırada kullanılarak ağacın oluşturulacağı belirlenmelidir. Bu amaçla en yaygın olarak kullanılan ölçüm Entropi ölçümüdür. Entropi ölçüsü ne kadar fazla ise, o alan kullanılarak ortaya konulan sonuçlar da o oranda belirsiz ve kararsızdır. Bu nedenle karar ağacının kökünde entropi ölçüsü en az olan alanlar kullanılır. (Özekes ve Çamurcu, 2002).

A alanı k farklı değere sahip olsun $\{a_1, a_2, ..., a_k\}$. Verilen bir A alanının entropi ölçüsünü bulan formüller şu şekildedir (Özekes ve Çamurcu, 2002):

- (2.2)

Bu formülde;

E (C\A) = A alanının sınıflama özelliğinin Entropi ölçüsü,

p ($a_k$, j) = $a_k$ alanının j değerinde olma olasılığı,

p ($c_i \setminus a_k$, j) = $a_k$ alanı j. Değerindeyken sınıf değerinin $c_i$ olma olasılığı,

$M_k$ = $a_k$ alanının içerdiği değerlerin sayısı ; j=1,2,…, $M_k$ ,

N = farklı sınıfların sayısı ; i= 1,2,…, N ,

k = alanların sayısı ; k = 1,2,…, k.

Eğer bir S kümesindeki elemanlar kategorik olarak $C_1$, $C_2$, $C_3$, . . . , $C_i$ sınıflarına ayrıştırılırlarsa, S kümesindeki bir elemanın sınıfını belirlemek için gereken bilgi şu formülle hesaplanmaktadır:

- (2.3)

Bu formülde $p_i$, keyfi bir örneğin $C_i$ sınıfına ayrılma olasılığıdır ve $S_i$ / S olarak ifade edilir. $S_i$ ise $C_i$ sınıfında S'nin örneklerinin sayısını temsil etmektedir. Entropi ya da A' ya göre alt kümelerine ayrıştırılmasına dayanan beklenen bilgi denklemi şu şekilde de ifade edilebilir:

E(A) = — x $I(S_i)$ (2.4)

Bu durumda A alanı kullanılarak yapılacak dallanma işleminde, bilgi kazancı şu formülle hesaplanmaktadır:

- (2.5)

Bir başka deyişle Kazanç(A), A alanının değerini bilmekten kaynaklanan entropideki azalmadır.

Karar ağaçlarında kullanılan birçok algoritma mevcuttur. ID3, C4.5, C5.0, CART, CHAID ve QUEST bunlara örnek olarak gösterilebilir.

C4.5 ve C5.0 Algoritmaları: En yaygın kullanılan karar ağacı algoritması Quinlan'ın 1986'da önerdiği ID3 algoritmasının geliştirilmiş hali olan C4.5 algoritmasıdır. C5.0 algoritması ise C4.5'in geliştirilmiş hali olup, özellikle büyük veri setleri için kullanılmaktadır. C5.0 algoritması doğruluğu arttırmak için boosting algoritmasını kullandığından boosting ağaçları olarak da bilinir. C5.0 algoritması C4.5'e göre çok daha hızlı olup, hafızayı daha verimli kullanmaktadır (Sancak, 2008). Her iki algoritmanın sonuçları aynı olsa da C5.0 biçim olarak daha düzgün karar ağaçları elde etmemizi sağlamaktadır.

CART Algoritması: Morgan ve Sonquist'in AID (Automatic Interaction Detection) adlı karar ağacı algoritmasının devamı niteliğine Breiman ve diğerleri tarafından 1984 yılında önerilmiştir. Hem sayısal hem de nominal veri türlerini, girdi ve kestirimsel değişken olarak kabul edebilen CART algoritması, sınıflandırma ve regresyon problemlerinde bir çözüm olarak kullanılabilir. CART karar ağacı, ikili olarak özyinelemeli biçimde bölünen bir yapıya sahiptir. Dallanma kriteri olarak Gini indeksinden yararlanan CART ağacı, kuruluş aşamasında herhangi bir durma kuralı olmaksızın sürekli olarak bölünerek büyümektedir. Artık yeni bir bölünmenin gerçekleşmeyeceği durumda bu sefer uçtan köke doğru budama işlemi başlatılır. Olası en başarılı karar ağacı her budama işlemi sonrası bağımsızca seçilmiş bir test verisi ile değerlendirme yapılarak tespit edilmeye çalışılır (Sezer ve diğ., 2010).

CHAID Algoritması: CART' ın dışında en çok kullanılan karar ağacı algoritmalarından biri de CHAID' dır. CHAID (Chi-squared Automatic Interaction Detector; Ki-kare Otomatik Etkileşim Dedektörü), optimal bölünmelerin teşhisi için ki-kare istatistiğini kullanan bir yöntemdir. CHAID, bölümlendirme amaçlı kullanılan etkili bir istatistiksel tekniktir. Bir istatistiksel testin anlamlılığını kriter olarak kullanarak, bir potansiyel ön kestirici değişkenin tüm değerlerini değerlendirir. Hedef değişkene veya aynı anlama gelmek üzere bağlı değişkene göre homojen olarak değerlendirilen tüm değerleri birleştirir ve diğer tüm değerleri heterojen (benzer olmayan) olarak değerlendirir. Ardından karar ağacındaki ilk dalın formuna göre en iyi ön kestirici değişkenin seçilmesiyle, her bir düğümün seçilen

değişkenin homojen değerlerinin bir grubunu oluşturmasını sağlar. Bu süreç ağaç tamamıyla büyüyene kadar sürer. Kullanılan istatistiksel test, hedef değişkenin ölçüm düzeyine bağlıdır (Oğuzlar, 2004).

QUEST Algoritması: En son geliştirilen karar ağacı olma özelliğini taşıyan QUEST (Quick, Unbiased, Efficient Statistical Tree; Hızlı, Yansız, Etkili İstatistiksel Ağaç), çok sayıda kategoriye sahip ön kestiricileri destekleyen, diğer yöntemlerin yanlılıklarından kaçınılmasını sağlayan ve hızlı hesaplanabilen bir yöntemdir (Oğuzlar, 2004). 1997 yılında Loh and Shih tarafından geliştirilmiştir. İkili karar ağacı yapısı kullanan bir sınıflandırma algoritmasıdır. İkili ağaç kullanılmasının sebebi, ikili ağaçlarda budama ve doğrudan durma kuralı gibi tekniklerin kullanılabilmesidir. QUEST algoritması, ağacın oluşturulması sırasında değişken seçimi ve bölünmeyi eşzamanlı olarak yapan CHAID ve CART'ın aksine hepsi ile ayrı ayrı ilgilenir. QUEST algoritması, ağacın dallanması sırasındaki önyargılı seçimin daha genel hale getirilmesi ve hesaplama maliyetinin düşürülmesi amacıyla geliştirilmiştir. Fakat henüz sınıflandırmadaki doğruluk, ağacın büyüklüğü ve dallanmadaki değişiklik konularında diğerlerine açık bir üstünlük sağlayan sınıflandırma algoritması yoktur (Sancak, 2008).

CHAID, QUEST, C5.0 ve CART algoritmaları dışında geliştirilen diğer algoritmalar arasında Exhaustive CHAID, SLIQ (Supervised Learning in Quest), SPRINT (Scalable Parallelizable Induction of Decision Tree), MARS (Multivariate Adaptive Regression Splines) yer almaktadır (Emel ve Taşkın, 2005). Tablo 2.1'de bazı karar ağacı algoritmalarının özellikleri verilmektedir.

Tablo 2.1. Bazı karar ağacı algoritmaları ve özellikleri (Emel ve Taşkın, 2005)

| KARAR AĞACI ALGORİTMASI | ÖZELLİKLER |
|---|---|
| **C&RT** | Gini'ye dayalı ikili bölme işlemi mevcuttur. Son veya uç olmayan her bir düğümde iki adet dal bulunmaktadır. Budama işlemi ağacın karmaşıklık ölçüsüne dayanır. Sınıflandırma ve regresyonu destekleyici bir yapıdadır. Verinin hazırlanmasına gereksinim duyar. |
| **C4.5 ve C5.0** ( ID3 karar ağacı algoritmasının ileri versiyonları) | Her düğümden çıkan çoklu dallar ile ağaç oluşturur. Dalların sayısı tahmin edicinin kategori sayısına eşittir. Tek bir sınıflayıcıda birden çok karar ağacını birleştirir. Ayırma işlemi için bilgi kazancı kullanır. Budama işlemi her yapraktaki hata oranına dayanır. |
| **CHAID** ( Chi- Squared Automatic Interaction Detector) | Ki-kare testleri kullanarak bölme işlemini gerçekleştirir. Dalların sayısı iki ile tahmin edicinin kategori sayısı arasında değişir. |
| **SLIQ** (Supervised Learning in Quest) | Hızlı ölçeklenebilir bir sınıflayıcıdır. Hızlı ağaç budama algoritması mevcuttur. |
| **SPRINT** (Scalable Parallelizable Induction of Decision Tree) | Büyük veri kümeleri için idealdir. Bölme işlemi tek bir niteliğin değerine dayanır. Tüm bellek sınırlamaları üzerinde nitelik listesi veri yapısı kullanarak işlem yapar. |

## 2.2. Kümeleme

Kümeleme, veri madenciliğinin temel işlemlerinden biridir. Kümeleme, müşteri segmentasyonu ve dolandırıcılık tespiti gibi problemlerin çözümünde yaygın biçimde kullanılır. Kümeleme uygulamalarında üç görevi yerine getirmiş oluruz (Ching ve Pong, 2002):

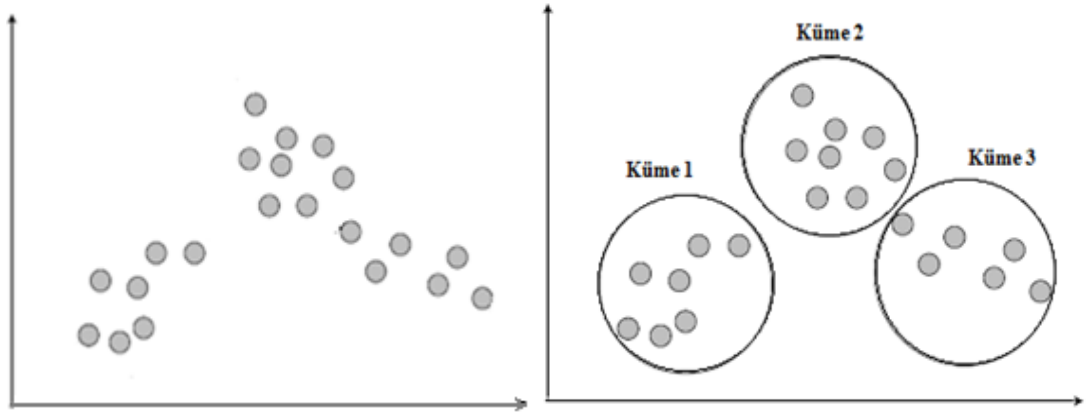
1. Veri setlerini kümeler içinde bölümlere ayırma,
2. Kümeleme sonuçlarını doğrulama,
3. Kümeleri yorumlama.

Kümeleme analizi önceden belirlenen seçme kriterlerine göre birbirine çok benzeyen birey ya da nesneleri aynı küme içinde gruplandırmaktadır. Analizin sonucunda bir kümeyi oluşturan birey veya nesneler birbiriyle benzeşirken, diğer kümelerin birey veya nesneleriyle benzeşmeyeceğinden, kümeler kendi içlerinde homojen iken, kümeler arasında heterojenlik söz konusu olmaktadır. Oluşturulan kümeler çok boyutlu uzayda gösterildiğinde, eğer kümeleme başarılı ise aynı küme içinde yer alan birey veya nesnelerin birbirlerine olukça yakın çıkması, bununla birlikte farklı kümelerin de birbirinden fark edilir düzeyde uzak olması beklenmektedir (Suner ve Çelikoğlu, 2010).

Kullanıcının amacına ve kullanım alanına göre kümeleme analizinin amaçları aşağıdaki gibi sıralanabilir (Çakmak ve diğ., 2005).

- Doğru tiplerin belirlenmesi,
- Model oluşturmak,
- Gruplara dayalı tahmin,
- Hipotez testi,
- Veri araştırma (inceleme),
- Hipotez oluşturma,
- Veri indirgeme.

Kümeleme için basit bir örnek Şekil 2.5'te verilmiştir. Kümeleme öncesinde verilerin dağılımı ve kümeleme sonrası ortaya çıkan üç ayrı küme şekilde gösterilmektedir.




Şekil 2.5. Kümeleme örneği

Kümeleme analizi hemen hemen tüm bilim alanlarında yararlanılan bir yöntemdir. Örneğin Tıp'ta hastalıkların sınıflandırılması, Psikiyatri'de laboratuar bulguları ile klinik bulguların oluşturduğu veri matrislerinden hastalık alt gruplamalarının ya da yeni sendromların tanımlanmasında, Arkeoloji'de canlı fosillerinden elde edilen verilerle tarih öncesi canlıların tiplerinin belirlenmesinde, Eğitim'de kültürel ve eğitsel alt yapılara göre eğitim programları geliştirmek, örnek öğrenme kalıpları oluşturmak, kişilere göre alt kişilik kalıpları belirlemek ve bu kalıplara uyumlu eğitim programları geliştirmek amacıyla kümeleme analizinden yararlanılmaktadır (Özdamar, 2004b).

### 2.2.1. Kümeleme yöntemleri

Birimlerin benzerliklerine göre kümelere dahil edilmesinde kullanılabilecek çeşitli yaklaşımlar mevcuttur. Bu yaklaşımlardan biri, en çok benzer iki birimi aynı gruba atamakla başlayıp tüm birimlerin aynı gruba atanması ile biten hiyerarşik bir yaklaşımdır. Bir başka yaklaşım ise tüm verilerin ortalama değerlerine en yakın değerlere sahip birimlerin aynı kümeye atanmasını esas alan yaklaşımdır. En çok kullanılan bu iki yaklaşım dışında diğer yaklaşımlar da mevcuttur. Tüm yaklaşımlarda en önemli ölçüt, kümeler arası farklar ile kümeler içi benzerliklerin maksimum olmasını sağlamaktır. En çok kullanılan kümeleme algoritmaları

hiyerarşik ve hiyerarşik olmayan kümeleme adı altında iki kategoride toplanmaktadır (Atbaş, 2008).

### 2.2.1.1. Hiyerarşik kümeleme yöntemleri

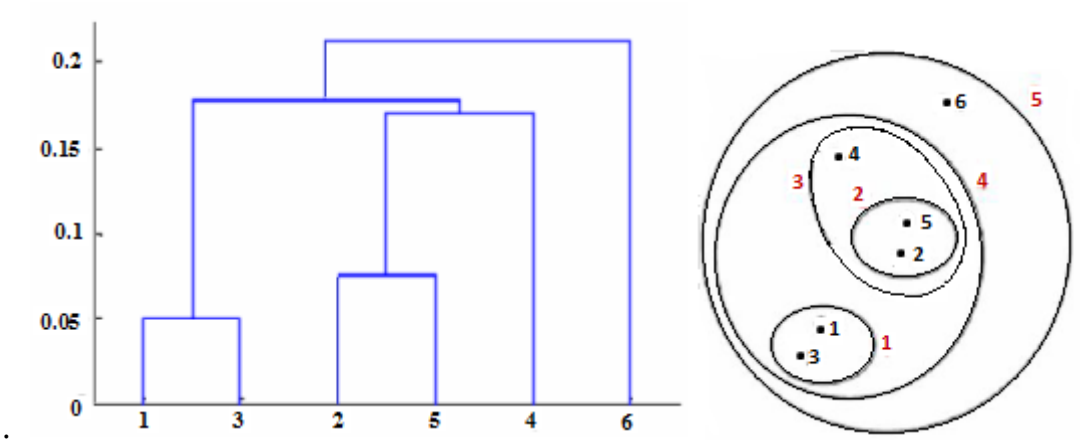
Hiyerarşik kümeleme teknikleri, kümeleri peş peşe birleştirme sürecidir ve bir grup, diğeri ile bir kez birleştirildikten sonra, daha sonraki adımlarda kesinlikle ayrılamaz. Hiyerarşik tekniklerin ağaç diyagramları ile gösterilen sonuçlarına dendogram denir (Çakmak ve diğ., 2005).

Hiyerarşik kümeleme nesnelerin yakınlık ilişkisine göre oluşturulan kümelerden bir ağaç inşa eder. Hiyerarşik kümeleme aşağıdaki özelliklere sahiptir (Çil, 2010):

- Bir veri tabanını bir kaç kümeye ayrıştırır.
- Bu ayrıştırma dendogram adı verilen bir ağaç sayesinde yapılır.
- Bu ağaç, yapraklardan gövdeye doğru veya gövdeden yapraklara doğru kurulabilir.

Dendogram istenen seviyede kesilerek kümeler elde edilir.

Hiyerarşik yöntemle veri kümelemeye ilişkin örnek Şekil 2.6'da verilmiştir.




Şekil 2.6. Hiyerarşik yöntemle veri kümeleme örneği

Bir hiyerarşik kümeleme metodu veri nesnelerini bir küme ağacına gruplayarak çalışır. Hiyerarşik kümeleme yöntemleri, hiyerarşik ayrışmanın yukarıdan-aşağıya veya aşağıdan-yukarıya oluşturulmasına bağlı olarak bütünleştirici ve bölücü hiyerarşik kümeleme olarak sınıflandırılabilir.

Bütünleştirici ve bölücü yaklaşıma göre hiyerarşik kümeleme şu şekildedir (Bilen, 2009):

Aşağıdan yukarıya ya da bir diğer ifadeyle bütünleştirici yaklaşıma göre hiyerarşik kümeleme;

- Her bir nesne için farklı bir grup oluşturarak başla,
- Merkezler arasındaki uzaklık, ortalama gibi kurallara göre grupları birleştir.
- Bir sonlandırma durumuna ulaşılıncaya kadar devam et. Yani, bütün nesneler tek bir küme içinde kalana kadar ya da istenen sayıda küme elde edene kadar birleştirme işlemi devam eder.

Yukarıdan aşağıya ya da bir diğer ifadeyle bölücü yaklaşıma göre hiyerarşik kümeleme;

- Aynı kümedeki bütün nesnelerle başla,
- Bir kümeyi daha küçük kümelere böl,
- Bir sonlandırma durumuna ulaşılıncaya kadar devam et. Yani, her nesne ayrı bir küme oluşturana ya da istenilen küme sayısı elde edilene kadar ayrılma işlemi devam eder.

#### 2.2.1.2. Hiyerarşik olmayan kümeleme yöntemleri

Birimlerin kendi içinde homojen ve kendi aralarında heterojen olan kümelere ayrılmasını hedefleyen ve prototip kümeler aracılığı ile alt populasyonların parametre tahminlerini yapmayı amaçlayan yöntemlerdir. Hiyerarşik olmayan kümeleme yöntemlerinde birimlerin uygun oldukları kümelerde toplanmaları ve n birimin k kümeye parçalanması hedeflenmektedir (Özdamar, 2004b).

Hiyerarşik olmayan kümeleme yöntemleri başlığı altında birçok teknikten söz etmek mümkündür ancak bunlardan en sık kullanılanı k-ortalamalar yöntemidir.

K-ortalamalar yöntemi, verideki kümeleri bulmaya yarayan en açık ve etkili kümeleme algoritmasıdır (Larose, 2005). K- ortalama terimi her bir birimin en yakın

merkezli kümeye atanması süreci anlamında kullanılmıştır. Bu algoritmanın adımlarını aşağıdaki gibi özetlemek mümkündür (Akbulut, 2006):

1. Veri seti rassal olarak k adet başlangıç kümesine ayrılır.
2. Veri setinde yer alan örnekler; merkezi kendisine en yakın olan kümeye atanır
3. Her atamanın sonunda küme merkezi (ortalama) yeniden hesaplanır.
4. Veri setindeki tüm örneklerin ataması yapılana kadar 2. ve 3. adımlar tekrarlanır.

Bu süreç tüm gözlemler gruplara atanıncaya kadar devam eder. Tüm gözlemler gruplara atandıktan sonra atandıkları küme ortalamasından daha yakın küme ortalaması varsa, gözlemlerin yerleri değiştirilmektedir. Amaç diğer kümeleme yöntemlerinde olduğu gibi, gerçekleştirilen kümeleme işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır (Atbaş, 2008).

Bu yöntem çok sayıda birimden elde edilmiş olan sürekli değişkenli veri setlerini küme içi kareler toplamını minimize edecek biçimde k kümeye ayırmayı amaçlamaktadır. Birimlerin az sayıda kümeye yerleştirilmesi iteratif bir biçimde yapılmakta olup, her iterasyonda farklı kümelere atanarak en uygun çözüm permutasyonel bir yaklaşım ile belirlenir. k-ortalamalar yöntemi, birimlerin incelenmesi ile k=2'den başlayarak küme sayılarını her defasında birer arttırarak deneysel olarak en uygun kümelemeyi bulmak şeklinde uygulanabilir. Böyle bir yaklaşımda toplam küme içi varyans matrisi izi minimize edilir (Özdamar, 2004b).

### 2.3. Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

Birliktelik kuralı, geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır (Özçakır ve Çamurcu, 2007).

Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak "pazar sepeti analizi" adı altında veri madenciliğinde yaygın olarak

kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır (Akpınar, 2000).

Birliktelik kuralları, eş zamanlı olarak gerçekleşen ilişkilerin tanımlanmasında kullanılmaktadır. Birliktelik kurallarını aşağıdaki gibi örneklerle ifade etmek mümkündür:

- Müşteriler bira satın aldıklarında %75 olasılıkla kuruyemiş de satın alırlar.
- Düşük yağlı peynir ve yağsız süt alan müşteriler %85 olasılıkla yağsız yoğurt da satın alırlar.

Ardışık zamanlı örüntüler ise birbirleri ile ilişkisi olan ancak birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılmaktadır. Ardışık zamanlı örüntüleri aşağıdaki gibi örneklerle ifade etmek mümkündür:

- Eşofman satın alan bir müşteri, ilk üç ay içerisinde %55 olasılıkla spor ayakkabı alacaktır.
- X ameliyatı yapıldığında, 20 gün içinde % 45 ihtimalle Y enfeksiyonu oluşacaktır.

# 3. UYGULAMA

## 3.1. Uygulamaya Genel Bakış

Bu bölümde özel bir bankanın birinci sınıf bir şubesindeki bireysel kredi müşterilerinin kredi geri ödeme performanslarının değerlendirilmesine yönelik bir veri madenciliği uygulamasına yer verilmiştir. Kredi kullandırma, bankalar için risk arz eden bir durum olduğundan, bu çalışmada karar vermedeki risk oranını en aza indirmek için mevcut verilerden veri madenciliği yoluyla bilgi elde edilmesi hedeflenmiştir. Uygulamada kullanılan veriler, Türkiye'nin finans sektöründe faaliyet gösteren en büyük bankalarından birine ait bir şubedeki bireysel kredi müşterilerinin son altı aylık dönemdeki kanuni takip ve normal ödeme kayıtlarından oluşmaktadır.

Çalışma kapsamında bankanın genel müdürlük bünyesindeki ilgili birimlerinden, uygulamanın yapıldığı şubenin kredi müşterilerine ait müşteri numaraları ve kredi geri ödeme durumlarını içeren veriler temin edilmiştir. Müşterilere ait cinsiyet, medeni hal, yaş, aylık gelir, öğrenim durumu, ev ve araç sahibi olma, çocuk sahibi olma, banka maaş müşterisi olma durumu, çalışma şekli ile ilgili bilgilere ise bankadaki mevcut sistem üzerinden müşteri numaralarına göre inceleme yapılarak ulaşılmıştır. Gizlilik prensiplerine bağlı kalınması nedeniyle müşteri numaraları değiştirilmiştir. Çalışmada kullanılan veriler uzman görüşleri de dikkate alınarak kategorik hale getirilmiştir. Eksik ve hatalı veriler temizlendikten sonra elde edilen veriler Microsoft Excel üzerine aktarılmış ve ön işlemeler yapılmıştır. Uygulama, toplam 200 müşteri kaydından oluşan 200 x 12'lik bir matris ile gerçekleştirilmiştir.

Uygulamada veri madenciliği yöntemlerinden Kümeleme ve Sınıflandırmaya yer verilmiştir. Kümeleme analizi yapılarak mevcut 200 müşteri için oluşan kümelerdeki müşteri profiline göre yapılabilecek çapraz ürün satışları belirlenmiştir. Sınıflandırma

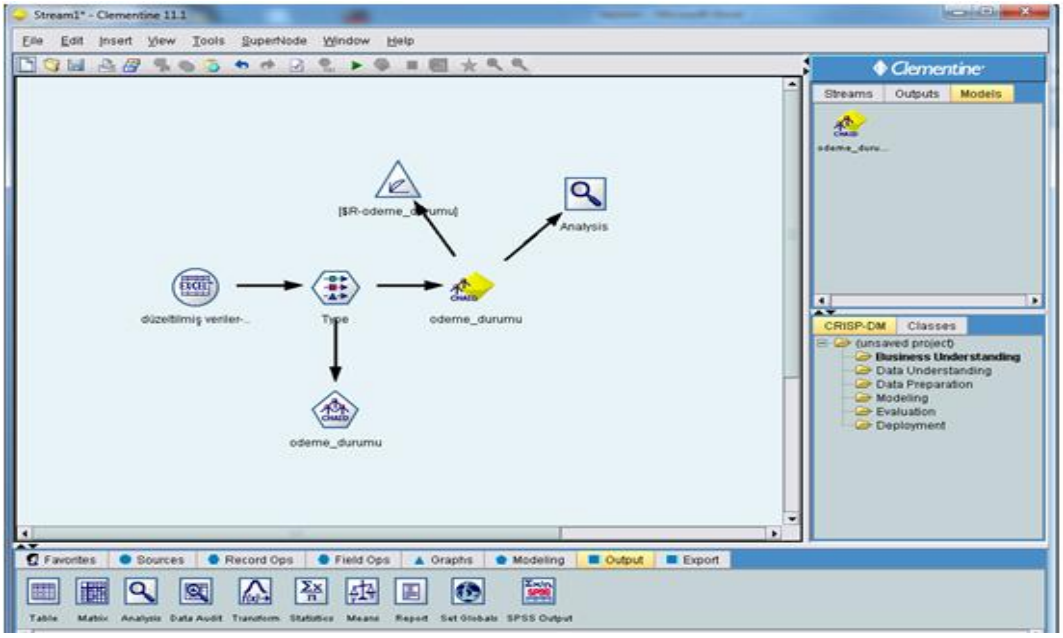
uygulamasında ise karar ağacı algoritmaları uygulanarak, gelecekteki müşteriler için kural çıkarımı yapılması hedeflenmiştir.

### 3.2. Uygulamada Kullanılan Yazılım

Analizde SPSS Clementine yazılımı kullanılmıştır. SPSS Clementine, veri madenciliği uygulamaları için geliştirilmiş bütünsel bir görsel modelleme gerecidir. Veri madenciliği çözümleri ile hem istatistik kökenli algoritmaları hem de yapay zeka kökenli algoritmaları, görsel bir programlama ara yüzü altında sunmaktadır.

SPSS Clementine, analiz yapan kişilere hızlıca tahminsel veri madenciliği modellerini geliştirme yeteneğini kazandırır. SPSS Clementine platformunun son derece kuvvetli bir ara yüzü ve görselliği vardır. Bu sayede kısa zamanda verilerdeki gizli desenler ortaya çıkarılarak verilerdeki etkileşimler görülebilir (URL-3). Veriye kolayca erişme, veriyi modellemeye hazırlama, modelleri hızlı bir şekilde oluşturma, birden fazla modeli ardışık olarak uygulama ve farklı model sonuçlarını kolayca karşılaştırmalarını sağlamaktadır (URL-4).

Clementine'in kullanıcı ara yüzünün kullanımı oldukça kolaydır. Tüm fonksiyonlar ekranın en altında bir palette ikon şeklinde yer almaktadır. Modelin kurulması ve analiz için ikonların ana panele sürüklenmesi, bunların oklarla birleştirilmesi gerekmektedir. Şekil 3.1'de yazılıma ait bir ara yüz görülmektedir.

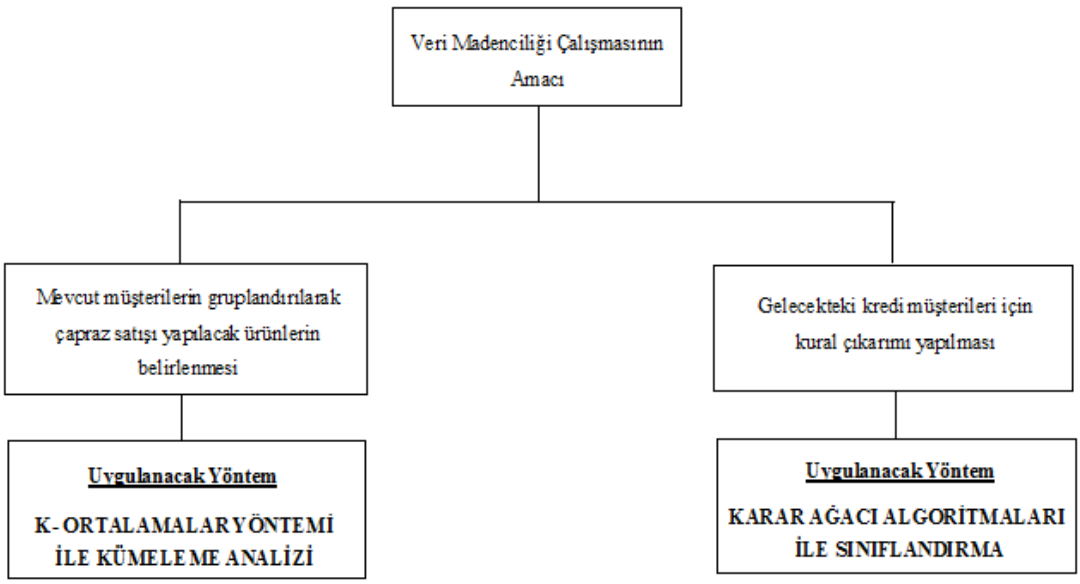



Şekil 3.1. SPSS Clementine yazılımına ait bir arayüz

### 3.3. Veri Madenciliği Probleminin Tanımlanması

Tez kapsamında yapılan veri madenciliği çalışması için özel bir bankanın birinci sınıf bir şubesindeki bireysel kredi müşterilerinin kredi geri ödeme performanslarına ve demografik özelliklerine bağlı veriler ele alınmıştır. Kredilerini üç dönem üst üste aksatan müşteriler, kanuni takibe düşmüştür ve bu sınıftaki müşterilerin ödeme performansı "Kanuni Takip" olarak kategorize edilirken, kredi ödemelerini aksatmayan müşterilerin ödeme durumu ise "Normal Ödeme" olarak adlandırılmıştır.

Uygulamanın amacı, müşterilerin ödeme performanslarına göre gruplandırılarak, kredi geri ödemelerini etkileyen faktörlerin belirlenmesi ve buna bağlı olarak, müşteri gruplarına göre bireysel kredilerin yanı sıra bankanın diğer ürünlerinin de satışa sunulabilmesi, diğer bir ifadeyle "çapraz satış" yapılabilmesidir. Uygulamada, mevcut müşteriler için çapraz satışı yapılacak ürünlerin belirlenmesi ve gelecekte bireysel kredi kullandırılacak müşteriler için kural çıkarımı yapılarak, kriterler oluşturulması hedeflenmiştir. Uygulamanın amacı Şekil 3.2' de görsel olarak ifade edilmiştir.




Şekil 3.2. Uygulamanın amacı

### 3.4. Verilerin Hazırlanması

#### 3.4.1. Veri toplama

Bu aşamada bankanın ilgili departmanlarından müşteri numaralarına göre son altı aylık döneme ait takipli ve normal ödemeli toplam 200 müşterinin bilgileri elde edilmiştir.

#### 3.4.2. Veri birleştirme ve temizleme

Müşterilere ait diğer bilgilere mevcut sistem üzerinden ulaşılarak veriler birleştirilmiştir. Eksik ve hatalı veriler temizlenerek, toplamda 200 müşteriye ait on iki farklı değişkenden oluşan 200x12'lik bir veri kümesi elde edilmiştir. Verilerin, veri dönüştürme öncesindeki hali aşağıdaki tabloda gösterilmektedir.

Tablo 3.1. Dönüştürme öncesinde veri tablosunun bir bölümü

| Müşteri | Cinsiyet | Medeni hal | Yaş | Aylık gelir | Eş geliri | Ev sahibi | Araç sahibi | Çocuk sahibi | Banka maaş müş. | Çalışma Şekli | Öğrenim D. | Kredi Durumu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 87 | B | Bekar | 49 | 1500 | H | H | H | H | H | Ö | L | Kanuni Takip |
| 88 | B | Evli | 72 | 620 | E | E | H | E | H | E | İ | Kanuni Takip |
| 89 | E | Evli | 74 | 820 | H | E | H | E | E | E | İ | Kanuni Takip |
| 90 | E | Evli | 44 | 2100 | H | H | H | E | E | K | İ | Kanuni Takip |
| 91 | E | Evli | 33 | 850 | H | H | H | E | H | Ö | L | Kanuni Takip |
| 92 | E | Evli | 50 | 655 | H | E | H | E | H | E | İ | Kanuni Takip |
| 93 | E | Evli | 53 | 1150 | H | H | H | E | E | E | L | Kanuni Takip |
| 94 | B | Bekar | 29 | 1050 | H | H | H | H | H | Ö | L | Kanuni Takip |
| 95 | B | Bekar | 47 | 2200 | H | H | H | H | H | K | L | Kanuni Takip |
| 96 | E | Evli | 48 | 1750 | H | E | H | E | E | E | İ | Kanuni Takip |
| 97 | E | Evli | 41 | 1400 | H | H | H | E | H | Ö | L | Kanuni Takip |
| 98 | E | Evli | 48 | 1550 | H | H | H | E | H | K | L | Kanuni Takip |
| 99 | E | Bekar | 43 | 1050 | H | H | H | H | H | Ö | İ | Kanuni Takip |
| 100 | E | Evli | 37 | 1100 | H | H | H | E | H | Ö | L | Kanuni Takip |
| 101 | E | Evli | 45 | 2580 | E | E | E | E | E | K | Ü | Normal Ödeme |
| 102 | B | Bekar | 28 | 5400 | H | H | E | H | H | K | Ü | Normal Ödeme |
| 103 | E | Evli | 57 | 1650 | E | E | H | E | E | E | İ | Normal Ödeme |
| 104 | E | Evli | 52 | 1800 | H | E | E | E | E | E | İ | Normal Ödeme |
| 105 | E | Evli | 39 | 1700 | E | H | E | E | E | K | L | Normal Ödeme |
| 106 | E | Evli | 60 | 1250 | H | E | H | E | E | E | Ü | Normal Ödeme |
| 107 | B | Evli | 44 | 1600 | E | H | E | E | E | K | L | Normal Ödeme |
| 108 | E | Evli | 42 | 1690 | E | H | E | E | E | K | L | Normal Ödeme |
| 109 | E | Evli | 41 | 2900 | E | E | E | E | E | K | Ü | Normal Ödeme |
| 110 | E | Bekar | 27 | 1230 | H | H | H | H | E | Ö | L | Normal Ödeme |
| 111 | B | Bekar | 32 | 1850 | H | H | H | H | H | K | Ü | Normal Ödeme |
| 112 | E | Evli | 39 | 2600 | E | H | E | E | H | K | L | Normal Ödeme |
| 113 | B | Evli | 32 | 2300 | E | E | E | E | H | K | Ü | Normal Ödeme |

#### 3.4.3. Veri dönüştürme

Bu aşamada, değişkenler kategorik hale getirilmiştir.

Müşteri Numaraları: Gizlilik politikası çerçevesinde, gerçek müşteri numaraları yerine 1'den 200'e kadar ardışık sayılar kullanılarak yeni müşteri numaraları oluşturulmuştur.

Cinsiyet: Bayan ve erkek müşteriler sırasıyla "B" ve "E" olarak tanımlanmıştır.

Medeni Hal: Medeni hal değişkeni de “EVLI” ve “BEKAR” olmak üzere iki kategoride ele alınmıştır.

Yaş: Müşterilerin doğum tarihleri dikkate alınarak yaşları hesaplanmıştır. En büyük yaş değeri 78, en küçük yaş değeri ise 24’tür. Yaş, sürekli bir değişkendir ve karar ağacının çok fazla dallanmasına neden olmaktadır. Bu amaçla kategorik hale getirilmiştir. Yaş değişkenine ait tanımlamalar Tablo 3.2’de gösterilmektedir.

Tablo 3.2. Yaş değişkenine ait tanımlama

| **YAŞ ARALIĞI** | **TANIMLAMA** |
|---|---|
| 24-30 | 24-30 YAS |
| 31-37 | 31-37 YAS |
| 38-44 | 38-44 YAS |
| 45-51 | 45-51 YAS |
| 52-58 | 52-58 YAS |
| 59-65 | 59-65 YAS |
| 66 | 66 YASVEUSTU |

Aylık gelir: Müşterilerin aylık gelirleri dikkate alınarak, veriler Tablo 3.3’de gösterildiği gibi anlamlı gruplarda kategorize edilmiştir.

Tablo 3.3. Aylık gelire göre tanımlama

| **AYLIK GELİR** | **TANIMLAMA** |
|---|---|
| 750 | 750TLVEALTI |
| 751-1400 | 751-1400TL |
| 1401-2050 | 1401-2050 |
| 2051-2700 | 2051-2700 |
| 2751-3350 | 2751-3350 |
| 3351-4000 | 3351-4000 |
| 4001 | 4001TLVEUSTU |

Eş Geliri: Müşterilerin eş gelirlerinin mevcut olup olmama durumu sırasıyla “VAR” ve “YOK” olarak tanımlanmıştır.

Ev: Müşteriler ev sahibi olma ya da olmama durumuna göre “VAR” ve “YOK” olarak iki kategoride toplanmıştır.

Araç: Araç sahibi olan ve olmayan müşteriler sırasıyla “VAR” ve “YOK” olarak gruplandırılmıştır.

Çocuk: Çocuklu müşterileri, çocuk sahibi olmayan müşterilerden ayırmak için yine “VAR” ve “YOK” şeklinde tanımlamalar yapılmıştır.

Banka maaş müşterisi: Kredi kullanmış olduğu bankadan aynı zamanda maaş alan müşteriler “EVET”, sadece kredi müşterisi olup maaşını farklı bankalardan alanlar ise “HAYIR” tanımlamaları ile kategorize edilmiştir.

Çalışma şekli: Çalışma şekline göre müşteriler üç farklı gruba ayrılmıştır. Kamuda çalışan müşteriler “KAMU”, özel sektör çalışanları “OZEL”, emekli müşteriler ise “EMEKLI” olarak tanımlanmıştır.

Öğrenim durumu: Öğrenim durumuna göre üç farklı tanımlama yapılmıştır. İlköğretim ve altına ait veriler “ILKOĞRETIM”, lise mezunu müşterilere ait veriler “LISE”, lisans ve üstünde öğrenim durumuna sahip müşterilere ait veriler ise “UNIVERSITE” şeklinde kategorize edilmiştir.

Ödeme durumu: Kredi taksitlerini üç dönem boyunca ödemeyen müşteriler kanuni takibe düşmektedir. Kanuni takipteki müşterilere ait veriler “KANUNI_TAKIP”, geri ödemede herhangi bir sorun yaşamayan müşterilere ait veriler ise “NORMAL_ODEME” başlığı altında toplanmıştır.

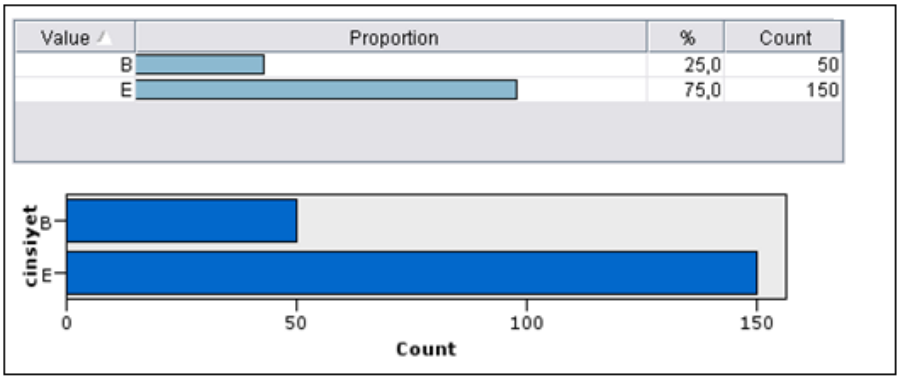
Veri dönüştürme işlemleri de tamamlandıktan sonra oluşan düzenlenmiş veri tablosu Tablo 3.4’ de gösterilmiştir.

Tablo 3.4. Düzenlenmiş veri tablosunun bir bölümü

| | cinsiyet | medeni_hal | yas | aylik_gelir | es_geliri | ev | arac | cocuk | banka_maas_musterisi | calisma_sekli | ogrenim_durumu | odeme_durumu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 82 | E | EVLI | 66YASVEUSTU | 751-1400TL | YOK | VAR | YOK | VAR | HAYIR | EMEKLI | LISE | KANUNI_TAKIP |
| 83 | E | EVLI | 45-51 | 1401-2050TL | YOK | VAR | YOK | VAR | EVET | KAMU | LISE | KANUNI_TAKIP |
| 84 | E | EVLI | 59-65 | 751-1400TL | YOK | VAR | YOK | VAR | HAYIR | OZEL | ILKOGRETIM | KANUNI_TAKIP |
| 85 | E | EVLI | 45-51 | 2051-2700TL | YOK | YOK | YOK | VAR | HAYIR | KAMU | ILKOGRETIM | KANUNI_TAKIP |
| 86 | E | EVLI | 45-51 | 1401-2050TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP |
| 87 | B | BEKAR | 45-51 | 1401-2050TL | YOK | YOK | YOK | YOK | HAYIR | OZEL | LISE | KANUNI_TAKIP |
| 88 | B | EVLI | 66YASVEUSTU | 750TLVEALTI | VAR | VAR | YOK | VAR | HAYIR | EMEKLI | ILKOGRETIM | KANUNI_TAKIP |
| 89 | E | EVLI | 66YASVEUSTU | 751-1400TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | KANUNI_TAKIP |
| 90 | E | EVLI | 38-44 | 2051-2700TL | YOK | YOK | YOK | VAR | EVET | KAMU | ILKOGRETIM | KANUNI_TAKIP |
| 91 | E | EVLI | 31-37 | 751-1400TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP |
| 92 | E | EVLI | 45-51 | 750TLVEALTI | YOK | VAR | YOK | VAR | HAYIR | EMEKLI | ILKOGRETIM | KANUNI_TAKIP |
| 93 | E | EVLI | 52-58 | 751-1400TL | YOK | YOK | YOK | VAR | EVET | EMEKLI | LISE | KANUNI_TAKIP |
| 94 | B | BEKAR | 24-30 | 751-1400TL | YOK | YOK | YOK | YOK | HAYIR | OZEL | LISE | KANUNI_TAKIP |
| 95 | B | BEKAR | 45-51 | 2051-2700TL | YOK | YOK | YOK | YOK | HAYIR | KAMU | LISE | KANUNI_TAKIP |
| 96 | E | EVLI | 45-51 | 1401-2050TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | KANUNI_TAKIP |
| 97 | E | EVLI | 38-44 | 751-1400TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP |
| 98 | E | EVLI | 45-51 | 1401-2050TL | YOK | YOK | YOK | VAR | HAYIR | KAMU | LISE | KANUNI_TAKIP |
| 99 | E | BEKAR | 38-44 | 751-1400TL | YOK | YOK | YOK | YOK | HAYIR | OZEL | ILKOGRETIM | KANUNI_TAKIP |
| 100 | E | EVLI | 31-37 | 751-1400TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP |
| 101 | E | EVLI | 45-51 | 2051-2700TL | VAR | VAR | VAR | VAR | EVET | KAMU | UNIVERSITE | NORMAL_ODEME |
| 102 | B | BEKAR | 24-30 | 4001TLVEUSTU | YOK | YOK | VAR | YOK | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME |
| 103 | E | EVLI | 52-58 | 1401-2050TL | VAR | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | NORMAL_ODEME |
| 104 | E | EVLI | 52-58 | 1401-2050TL | YOK | VAR | VAR | VAR | EVET | EMEKLI | ILKOGRETIM | NORMAL_ODEME |
| 105 | E | EVLI | 38-44 | 1401-2050TL | VAR | YOK | VAR | VAR | EVET | KAMU | LISE | NORMAL_ODEME |
| 106 | E | EVLI | 59-65 | 751-1400TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | UNIVERSITE | NORMAL_ODEME |
| 107 | B | EVLI | 38-44 | 1401-2050TL | VAR | YOK | VAR | VAR | EVET | KAMU | LISE | NORMAL_ODEME |
| 108 | E | EVLI | 38-44 | 1401-2050TL | VAR | YOK | VAR | VAR | EVET | KAMU | LISE | NORMAL_ODEME |
| 109 | E | EVLI | 38-44 | 2751-3350TL | VAR | VAR | VAR | VAR | EVET | KAMU | UNIVERSITE | NORMAL_ODEME |
| 110 | E | BEKAR | 24-30 | 751-1400TL | YOK | YOK | YOK | YOK | EVET | OZEL | LISE | NORMAL_ODEME |
| 111 | B | BEKAR | 31-37 | 1401-2050TL | YOK | YOK | YOK | YOK | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME |
| 112 | E | EVLI | 38-44 | 2051-2700TL | VAR | YOK | VAR | VAR | HAYIR | KAMU | LISE | NORMAL_ODEME |
| 113 | B | EVLI | 31-37 | 2051-2700TL | VAR | VAR | VAR | VAR | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME |
| 114 | E | BEKAR | 24-30 | 2051-2700TL | YOK | YOK | VAR | YOK | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME |
| 115 | E | EVLI | 52-58 | 1401-2050TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | NORMAL_ODEME |
| 116 | E | EVLI | 45-51 | 3351-4000TL | VAR | VAR | VAR | VAR | EVET | KAMU | UNIVERSITE | NORMAL_ODEME |
| 117 | B | BEKAR | 31-37 | 3351-4000TL | YOK | VAR | VAR | YOK | HAYIR | OZEL | UNIVERSITE | NORMAL_ODEME |

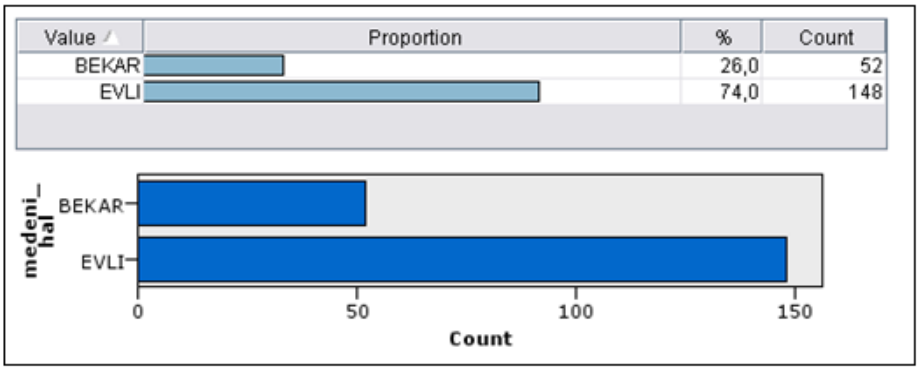
Müşteri bilgilerinin gizliliği açısından gerçek müşteri numaraları yerine, müşterilere 1'den 200' e kadar numara verilmiştir. Müşteri numarası, analiz sonuçlarını etkileyeceği için bir değişken olarak kabul edilmemiştir. Müşterilerin değişkenlere göre dağılımları, her bir değişken için ayrı ayrı incelenerek sonuçlar aşağıda yorumlanmıştır.

Cinsiyet değişkeni incelendiğinde, müşterilerin %25'ini bayanların, geri kalan %75'lik kısmını da erkeklerin oluşturduğu Şekil 3.3'de görülmektedir. Yani toplam 200 müşteriden 50'sinin bayan, kalan 150 kişinin ise erkek olduğu görülmektedir.



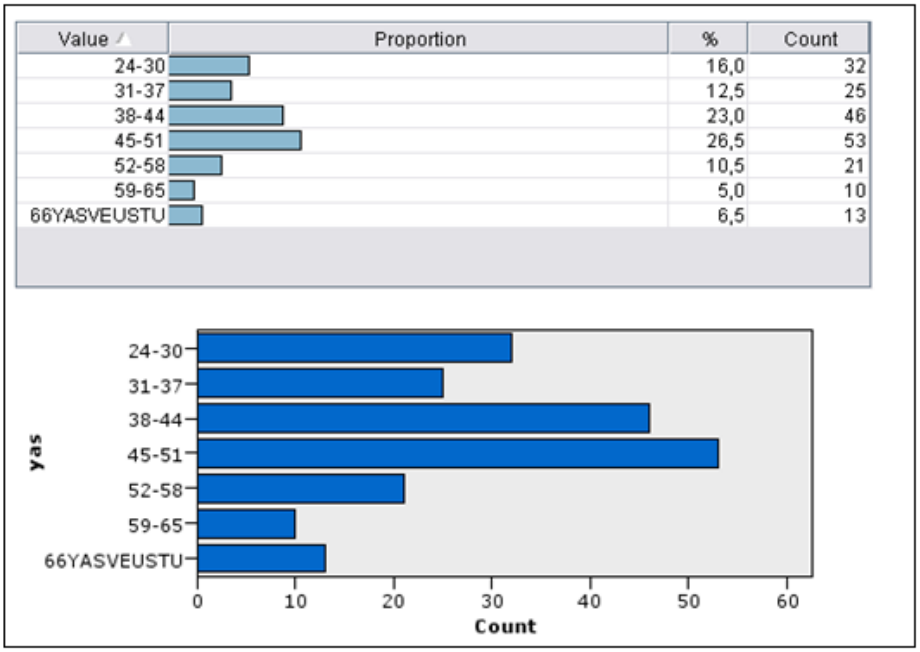

Şekil 3.3. Müşterilerin cinsiyete göre dağılımı

Medeni hale göre müşteriler incelendiğinde, 148 kişinin evli, 52 kişinin ise bekar olduğu görülmektedir. Diğer bir ifadeyle evli müşteriler %74 oranla çoğunlukta iken, bekarların oranı %26'dır.



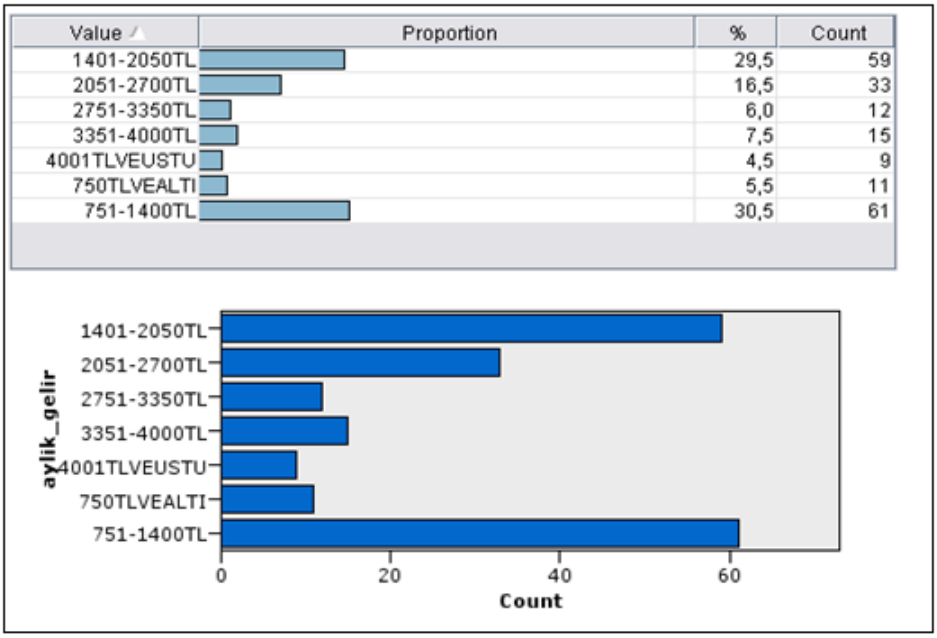

Şekil 3.4. Medeni hal değişkenine göre müşterilerin dağılımı

Yaş değişkeni incelendiğinde, 32 kişinin 24-30 yaş aralığında, 25 kişinin 31-37, 46 kişinin 38-44, 53 kişinin 45-51, 21 kişinin 52-58, 10 kişinin 59-65 yaş aralığında ve 13 kişinin 66 yaş ve üzerinde olduğu görülmektedir. Müşterilerin çoğu %26,5 oranla 45-51 yaş aralığındaki kişilerden oluşturmaktadır.



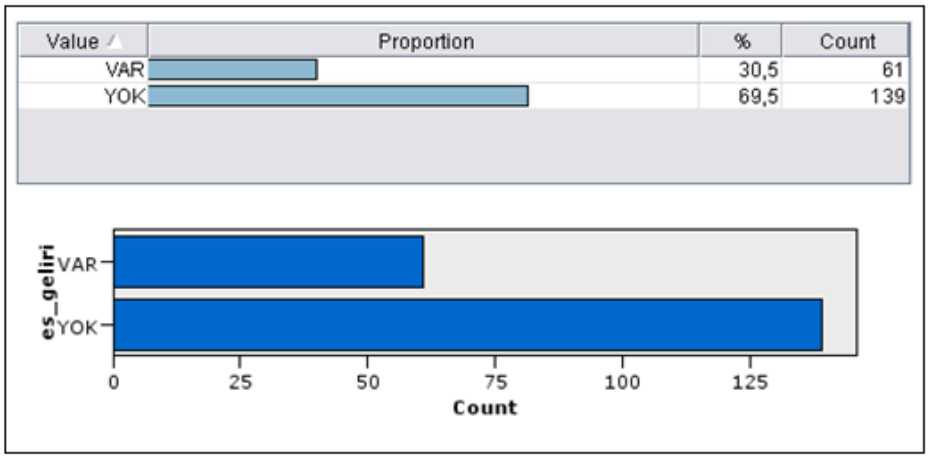

Şekil 3.5. Müşterilerin yaş değişkenine göre dağılımı

Aylık gelirler dikkate alındığında, gelirleri 4001 TL ve üzerindeki müşterilerin %4,5 ile en düşük orana, gelirleri 751 TL ve 1400 TL arasındaki müşterilerin ise %30.5 ile en yüksek orana sahip olduğu görülmektedir.



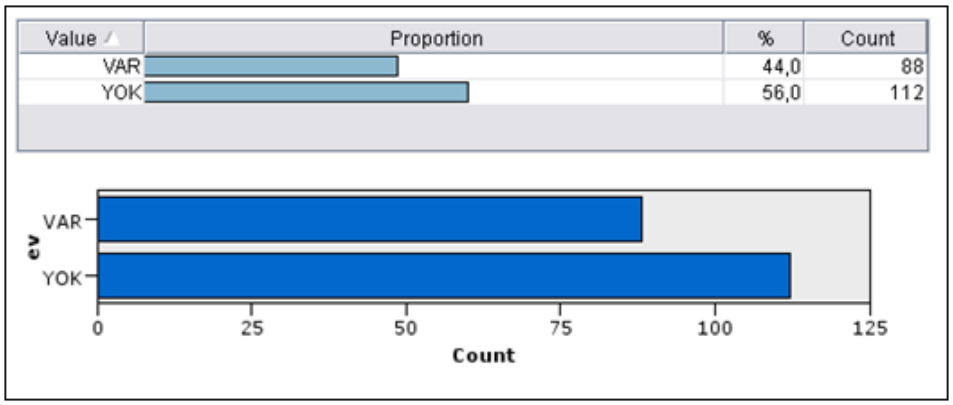

Şekil 3.6. Müşterilerin aylık gelire göre dağılımı

Eş geliri açısından müşteriler incelendiğinde, 139 kişinin eş gelirine sahip olmadığı, geri kalan 61 kişinin ise eş gelirlerinin mevcut olduğu Şekil 3.7'de görülmektedir.



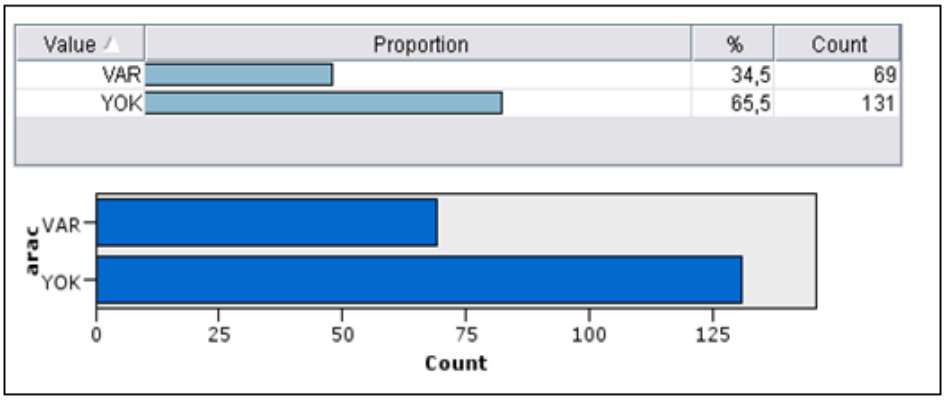

Şekil 3.7. Eş geliri değişkenine göre müşterilerin dağılımı

Müşterilerin %56'lık kısmının ev sahibi olmadığı, geri kalan %44'lük kısmının ise kendilerine ait evlerinin olduğu Şekil 3.8'de görülmektedir.



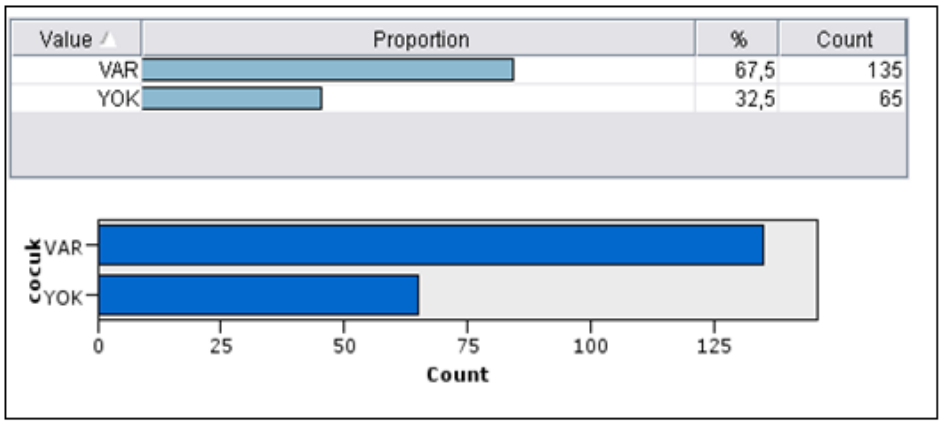

Şekil 3.8. Ev sahibi olma durumuna göre müşterilerin dağılımı

Araç sahibi olup olmama durumuna göre müşteriler incelendiğinde, 69 kişide araç mevcut iken, kalan 131 kişinin ise araç sahibi olmadığı görülmektedir.



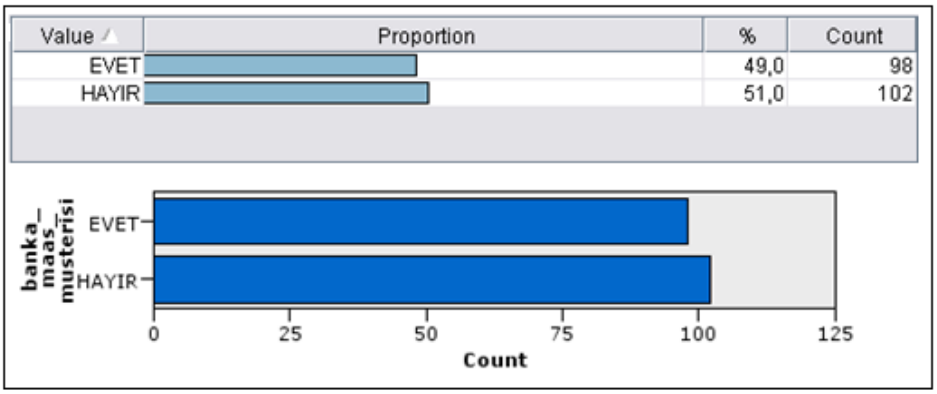

Şekil 3.9. Araç sahibi olma durumuna göre müşterilerin dağılımı

Şekil 3.10'de görüldüğü gibi çocuk sahibi olan kişiler %67,5 oranla çoğunluktadır. Başka bir ifade ile 135 kişi çocuk sahibi iken kalan 65 kişi çocuk sahibi değildir.



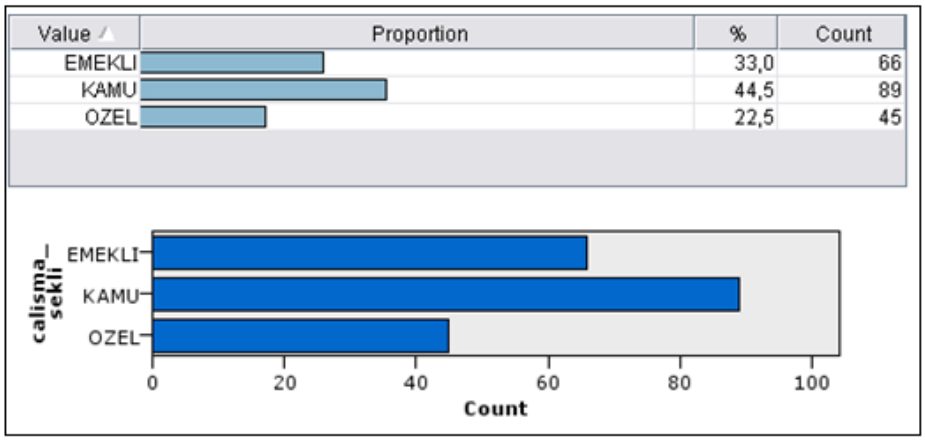

Şekil 3.10. Çocuk sahibi olma durumuna göre müşterilerin dağılımı

Müşteriler banka maaş müşterisi olup olmaması açısından kategorize edildiğinde, %49'luk kısmın banka maaş müşterilerinden, kalan %51'lik kısmın ise maaşını farklı bankalardan alan kredi müşterilerinden oluştuğu görülmektedir. Maaş alan kredi müşterileri ile maaş almayan müşterilerin oranının birbirine çok yakın olduğu Şekil 3.11'de gösterilmiştir.



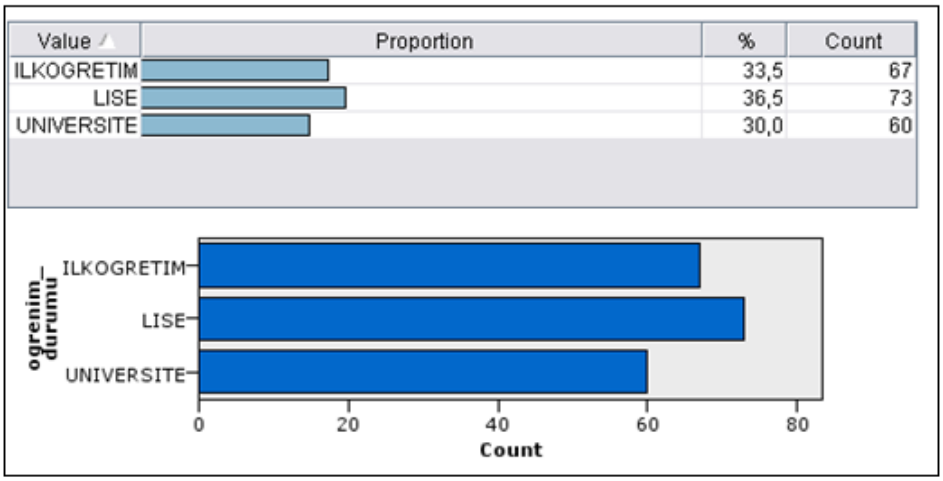

Şekil 3.11. Banka maaş müşterisi olma değişkenine ait dağılımlar

Çalışma şekli açısından müşteriler, "Emekli, Kamu ve Özel sektör çalışanları" olmak üzere üç gruba ayrılmıştır. Çalışma şekline göre müşterilerin dağılımı Şekil 3.12'de verilmektedir. %44,5 oranla kamu çalışanı olan kredi müşterilerinin çoğunlukta olduğu görülmektedir. Kredi müşterilerinin %33'lük kısmının emekli müşterilerden, kalan %22,5'lik kısmının ise özel sektör çalışanlarından oluştuğu Şekil'de görülmektedir.



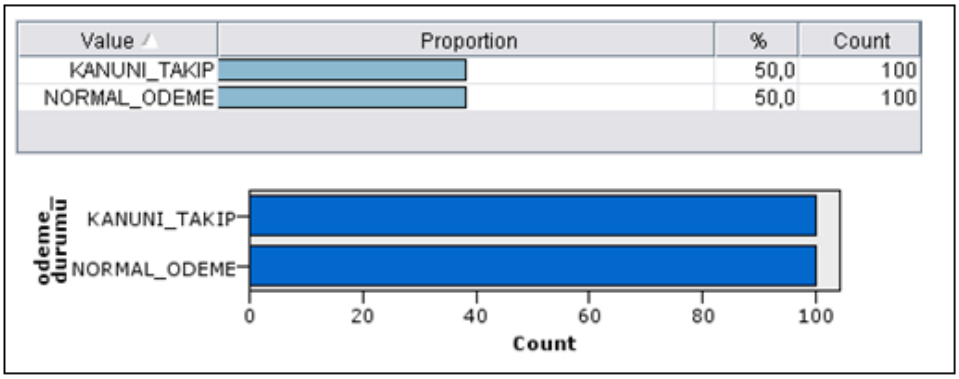

Şekil 3.12. Çalışma şekline göre müşterilerin dağılımı

Öğrenim durumu değişkeni incelendiğinde, lise mezunu müşterilerin %36,5 oranla çoğunlukta olduğu görülmektedir. Üniversite mezunu müşteriler ise %30 ile en düşük orana sahiptir. Toplamda bakıldığında 200 müşteri içinden, 67 kişinin ilköğretim mezunu, 73 kişinin lise mezunu ve 60 kişinin üniversite mezunu olduğu görülmektedir.




Şekil 3.13. Öğrenim durumuna göre müşterilerin dağılımı

Ödeme durumlarına göre müşterilerin eşit oranda dağılım gösterdiği görülmektedir. Şekil 3.14'e göre, toplamda 100 müşteri kredilerini geri ödemede sorun yaşamazken, kalan 100 kişiden kredi tahsilatı kanuni takip yoluyla sağlanmaktadır.




Şekil 3.14. Ödeme durumuna göre müşterilerin dağılımı

### 3.5. Modelin Kurulması ve Değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir (Terzi ve diğ., 2011).

Tez çalışmasında kümeleme ve sınıflandırma modellerine yer verilmiştir. Mevcut durumu değerlendirmek için k-ortalamalar yöntemi kullanılarak kümeleme analizi yapılmış, geleceğe yönelik tahminde bulunabilmek için ise sınıflandırma modellerinden karar ağacı algoritmaları ile kural çıkarımı yapılmıştır.

#### 3.5.1. Kümeleme analizi

Kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir (Akpınar, 2000).

Tezin üçüncü bölümünde kümeleme analizi ile ilgili ayrıntılı bilgilere yer verilmiştir. Bu bölümde kümeleme analizi tekniklerinden k- means yöntemi ile değişkenlerin kümeler üzerindeki etkileri ayrı ayrı incelenerek, mevcut müşterilere yönelik değerlendirmeler yapılmıştır.

K-Ortalamalar Algoritmasının Uygulanması:

K-ortalamalar algoritmasında K değeri probleme göre belirlenebilir veya belirlenmez. Hata kareler ölçütü gibi bir kümeleme ölçütünün olması gerekir. K-ortalamalar algoritması k kümelerini, her bir kümeyi temsil edecek bir nesnenin keyfi seçimiyle başlatır. Kalan her nesne bir kümeye atanır ve kümeleme kriteri küme ortalamasını hesaplayabilmek için kullanılır. Bu ortalamalar yeni küme noktaları olarak kullanılır ve her bir nesne kendisine en benzer olan kümeye yeniden atanır. Bu kümeler yeniden hesaplanır ve kümelerde hiç bir değişim gözlenilmediği duruma ve değişim istenen hata düzeyinin altına düşürülünceye kadar bu döngü devam ettirilir (Bilen, 2009).

Kümeleme Analizinin en kritik konusu küme sayısına karar vermektir. Araştırmacının küme sayısına karar vermede öznelliği minimize etmesi gerekmektedir. Ancak günümüzde yayınlanan birçok makalede bu konuda kesin bulunmuş sonuçlar yoktur. İlk önerilen yaklaşımlardan en çok bilinen eşitlik:

$$k= (n/2)^{1/2} \quad (3.1)$$

biçiminde hesaplanmaktadır. Burada "k" küme sayısını, "n" birim sayısını göstermektedir. Küçük örneklemli araştırmalarda kullanılması tavsiye edilir. Büyük örneklemli araştırmalarda kullanılması durumunda sağlıklı sonuçlara ulaşılması zorlaşır (Atbaş, 2008).

Uygulamada küme sayısını belirlemek için iki farklı yöntem kullanılmıştır. İlk olarak küme sayısı:

$k= (200/2)^{1/2}$ 'den 10 olarak bulunmuştur. Burada 200 değeri müşteri sayısını göstermektedir.

İkinci aşamada, k= 2'den k=10'a kadar küme sayısı birer arttırılmış ve her değer için hata kareleri toplamı bulunmuştur. Yukarıdaki formüle göre bulunmuş olan k= 10 değeri ile diğer küme sayılarına ilişkin hata kareleri karşılaştırılmış olup, hata kareleri toplamı en küçük olan değer küme sayısı olarak kabul edilmiştir.

SPSS Clementine'de k- ortalama yöntemi ile kümeleme analizi "Sources" kısmında yer alan "Excel" düğümü kullanılarak, excel dosyasındaki verilerin programa yerleştirilmesi ile başlamıştır. Değişkenlerin tipini belirlemek için ise "Type" düğümü kullanılmıştır. Type düğümüne, "Modeling" bölümünde yer alan "K-Means" düğümü eklenerek "Execute" seçeneği ile analiz sonuçları elde edilmiştir.

Tablo 3.5' de her küme sayısı için hata kareleri toplamına ilişkin değerler görülmektedir. Hata kareleri toplamı en az olan küme sayısı 3 olarak bulunmuştur ve bir sonraki aşamada k=3 için değişkenlerin kümeler üzerindeki önem dereceleri belirlenmiştir.

Tablo 3.5. K-means için küme sayısı ve hata kareleri toplamı

| Küme Sayısı | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| Hata Kareleri Toplamı | 2,22372 | 1,7355 | 1,78784 | 1,90777 | 2,11905 | 1,80539 | 1,98876 | 2,07131 | 1,92548 |

Tablo 3.6'da küme sayısının 3 ve 10 olması durumunda iterasyonlar sonucunda oluşan hatalar görülmektedir.

Tablo 3.6. Küme sayısının 3 ve 10 olması durumunda oluşan hatalar

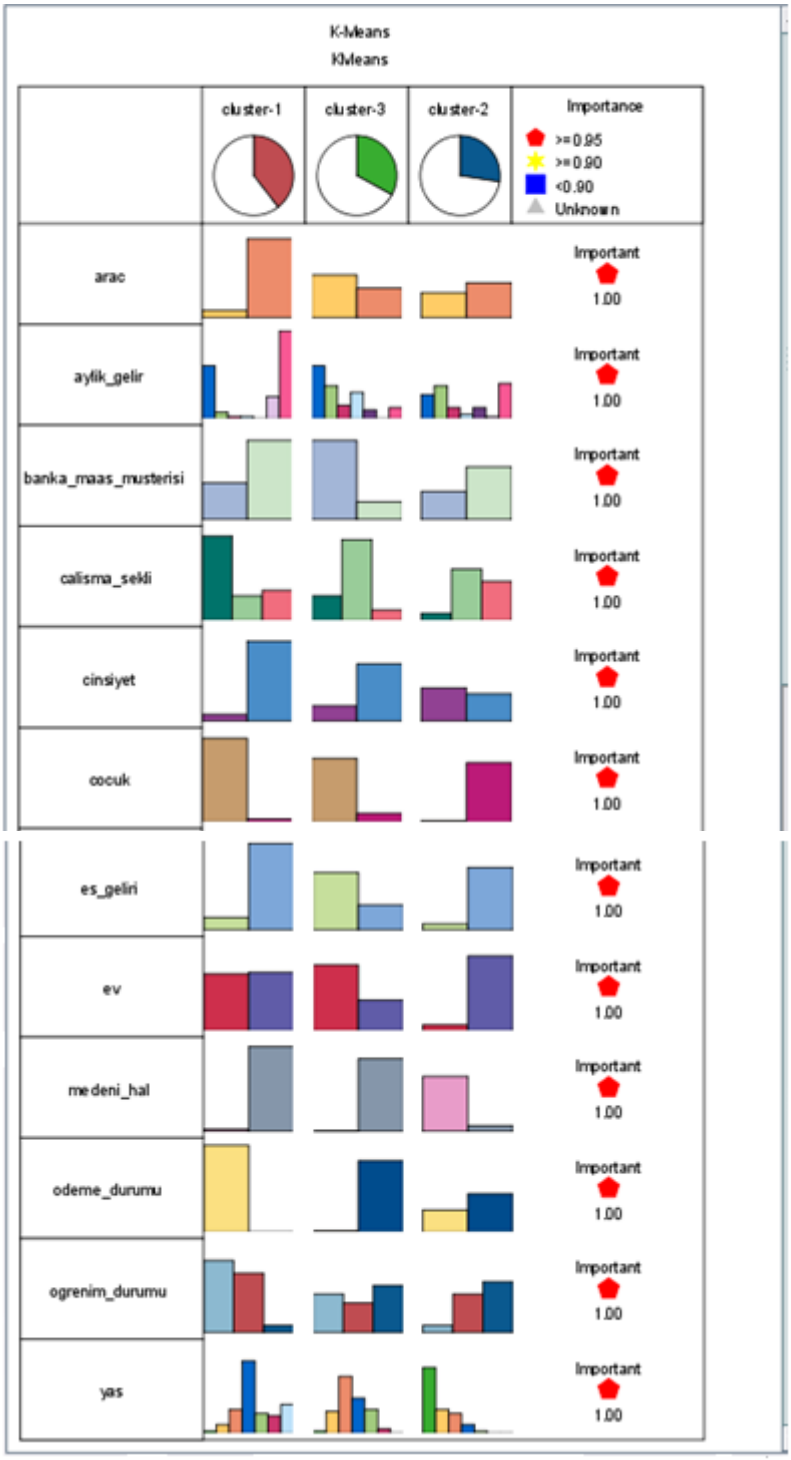
Number of clusters: 3

| Iteration | Error |
|---|---|
| 1 | 1,274 |
| 2 | 0,183 |
| 3 | 0,093 |
| 4 | 0,14 |
| 5 | 0,123 |
| 6 | 0,186 |
| 7 | 0,031 |
| 8 | 0,0 |

Number of clusters: 10

| Iteration | Error |
|---|---|
| 1 | 1,27 |
| 2 | 0,395 |
| 3 | 0,215 |
| 4 | 0,212 |
| 5 | 0,213 |
| 6 | 0,103 |
| 7 | 0,097 |
| 8 | 0,0 |

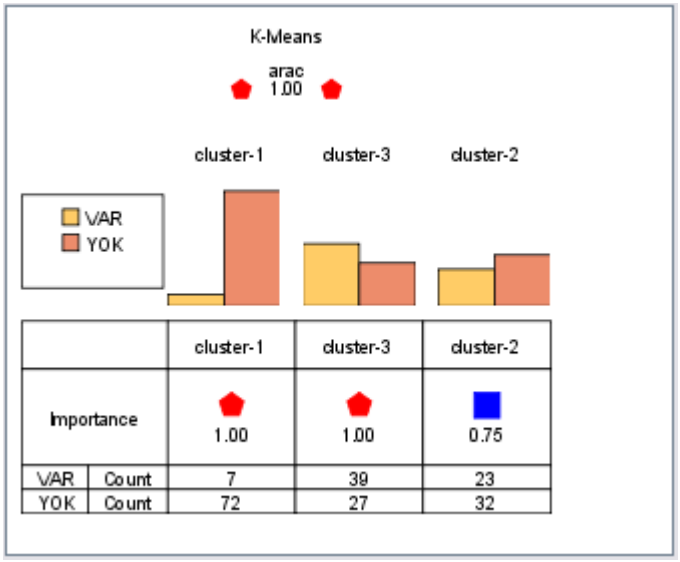
K-Means yöntemi ile elde edilen kümeler ve değişkenlerin kümeler üzerindeki etkileri Şekil 3.15'de gösterilmektedir. SPSS Clementine'de sürekli değişkenler için t-testi, kategorik değişkenler için ise ki-kare testi kullanılarak, değişkenlerin kümelerin belirlenmesinde etkili olup olmadığı test edilir. Eğer kategorik ya da sürekli bir değişkenin kümeler üzerinde etkisi varsa $H_0$ hipotezi reddedilir. Bu durumda anlamlılık düzeyi 0'a, önem endeksi ise 1'e yakın olur.




Şekil 3.15. K-Ortalamalar yöntemi ile elde elden kümeler

Program tarafından önem dereceleri 0.90'ın altında olan değişkenlerin kümeler üzerindeki etkisinin önemli düzeyde olmadığı tespit edilmiştir. Yukarıdaki şekilde görüldüğü gibi üç küme üzerinde de herhangi bir etkisi olmayan değişken yoktur.

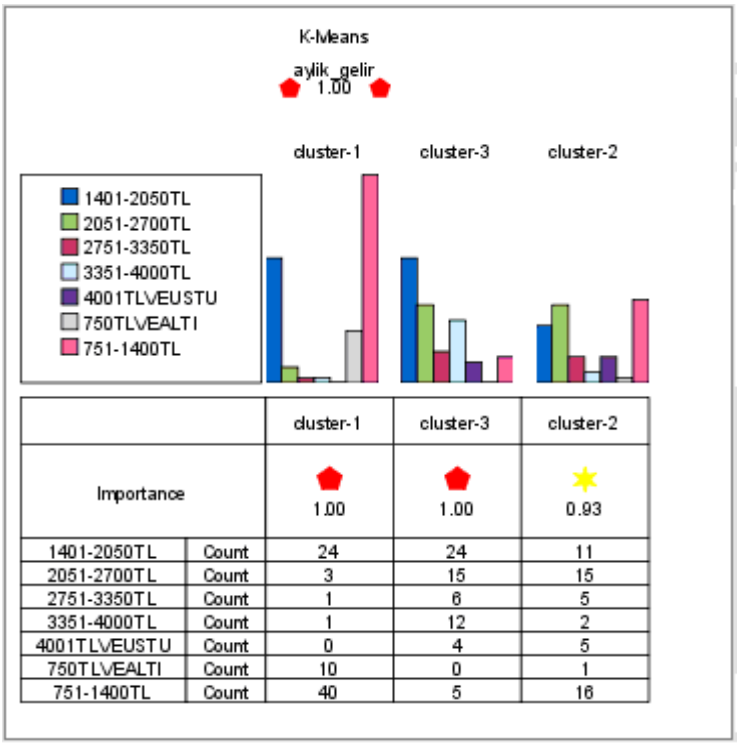
Bundan sonraki aşamada, değişkenlerin her birinin kümeler üzerindeki etkileri ayrı ayrı değerlendirilmiştir. Öncelikle ilk sıradaki araç değişkeni dikkate alınmıştır. Buna göre, araç değişkeni 1. ve 3. kümeler için önemli iken, 2. küme için ayırt edici bir niteliğe sahip değildir. Birinci kümedeki müşterilerin %91,14'ü araç sahibi değilken, üçüncü kümedeki müşterilerin %59,09'unda araç mevcuttur. Şekil 3.16' da görüldüğü gibi birinci kümede toplam 79, ikinci kümede 55 ve üçüncü kümede 66 kişi bulunmaktadır.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 0.75 |
| VAR | Count | 7 | 39 | 23 |
| YOK | Count | 72 | 27 | 32 |

Şekil 3.16. Araç değişkeninin kümelere etkisi

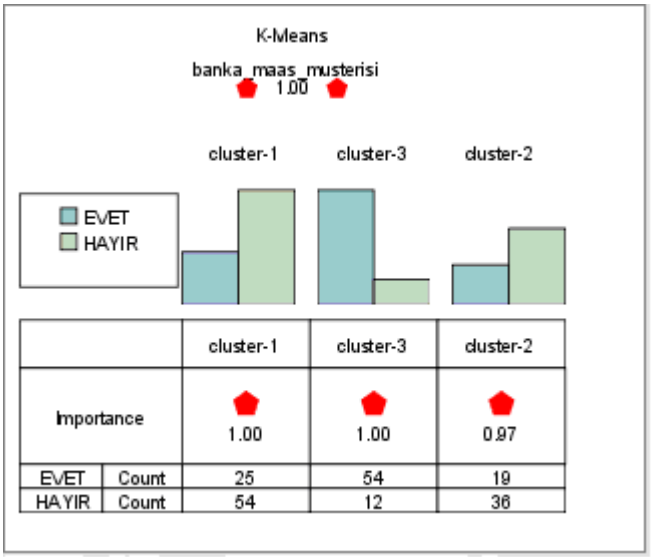
Aylık gelir değişkeni açısından kümeler incelendiğinde, bu değişkenin her üç küme için de önemli olduğu görülmektedir. Birinci kümedeki müşterilerin çoğunun aylık geliri 751- 1400 TL arasındadır. Bu gelir aralığına sahip kişilerin birinci kümedeki oranı %50,63'tür. İkinci kümenin aylık gelir oranları dikkate alındığında, müşterilerin %29,09'unun aylık gelirinin 751- 1400 TL aralığında, %27,27'sinin ise 2051- 2700 TL arasında olduğu görülmektedir. Üçüncü kümedeki müşterilerin aylık gelirleri %36,36'lık oranla 1401- 2050 TL arasındadır. Birinci kümede aylık geliri 4001 TL ve üstünde olan, üçüncü kümede ise aylık geliri 750 TL ve altında olan müşteri bulunmamaktadır.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 0.93 |
| 1401-2050TL | Count | 24 | 24 | 11 |
| 2051-2700TL | Count | 3 | 15 | 15 |
| 2751-3350TL | Count | 1 | 6 | 5 |
| 3351-4000TL | Count | 1 | 12 | 2 |
| 4001TLVEUSTU | Count | 0 | 4 | 5 |
| 750TLVEALTI | Count | 10 | 0 | 1 |
| 751-1400TL | Count | 40 | 5 | 16 |

Şekil 3.17. Aylık gelir değişkeninin kümelere etkisi

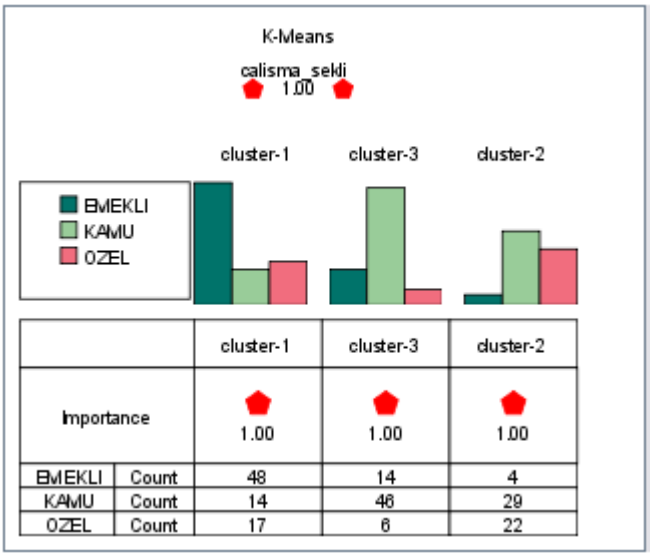
Banka maaş müşterisi olup olmaması açısından kümeler incelendiğinde, bu değişkenin üç küme üzerinde de önemli bir etkisinin olduğu görülmektedir. Birinci ve ikinci kümeler sırasıyla, %68,35 ve %65,45'lik oranla farklı bankalardan maaş alan kredi müşterilerinden, üçüncü küme ise %81,82 'lik oranla kredi kullandıkları bankadan maaş alan müşterilerden oluşmaktadır.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 0.97 |
| EVET | Count | 25 | 54 | 19 |
| HAYIR | Count | 54 | 12 | 36 |

Şekil 3.18. Banka maaş müşterisi olma durumunun kümeler üzerindeki etkisi

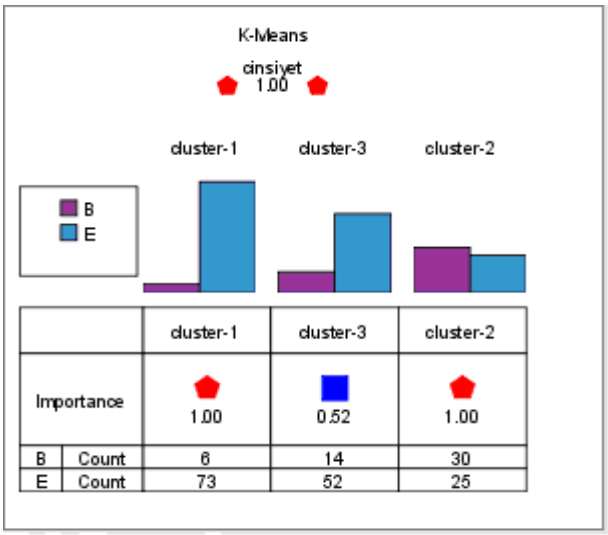
Çalışma şekli değişkeninin kümeler üzerindeki etkisi incelendiğinde, bu değişkenin üç küme açısından da önem arz ettiği görülmektedir. Birinci kümenin %60,76 oranla emekli müşteri ağırlıklı olduğu, ikinci ve üçüncü kümelerin ise sırasıyla, %52,73 ve %69,7 oranında kamu çalışanlarından oluştuğu Şekil 3.19'da gösterilmiştir.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 1.00 |
| EMEKLI | Count | 48 | 14 | 4 |
| KAMU | Count | 14 | 46 | 29 |
| OZEL | Count | 17 | 6 | 22 |

Şekil 3.19. Çalışma şekli değişkeninin kümeler üzerindeki etkisi

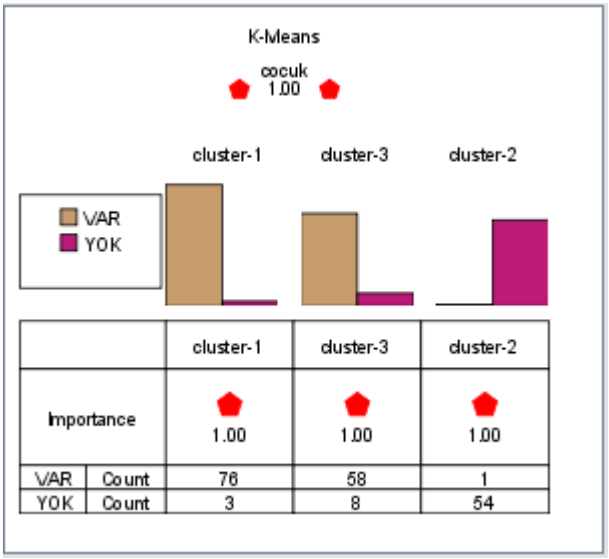
Şekil 3.20' de görüldüğü gibi cinsiyet değişkeni birinci ve ikinci kümeler için ayırt edici nitelikteyken, üçüncü kümeyi diğer kümelerden ayıracak bir özelliğe sahip olmadığından bu küme için önemli bir unsur değildir. Birinci küme %92,41 oranla erkek, ikinci küme ise %54,55 oranla bayan müşterilerden oluşmaktadır.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 0.52 | 1.00 |
| B | Count | 6 | 14 | 30 |
| E | Count | 73 | 52 | 25 |

Şekil 3.20. Cinsiyet değişkeninin kümeler üzerindeki etkisi

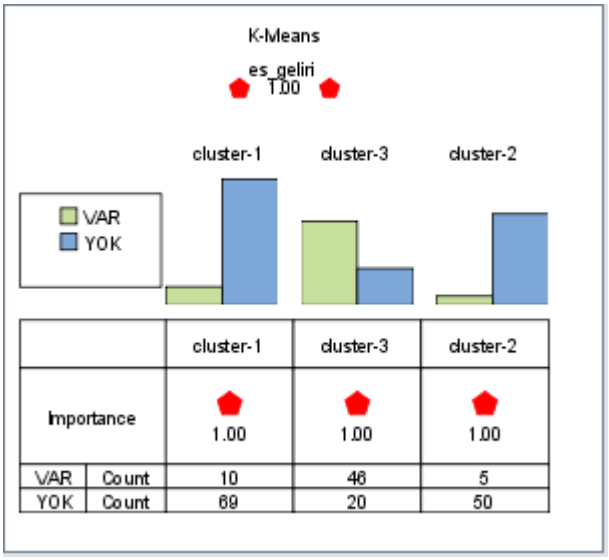
Çocuk sahibi olma durumunun üç küme için de önemli olduğu Şekil 3.21'de gösterilmektedir. Birinci ve üçüncü kümelerdeki müşterilerin sırasıyla %96,2 ve %87,88'i çocuk sahibi iken, ikinci kümedeki müşterilerin %98,18'i çocuk sahibi değildir.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 1.00 |
| VAR | Count | 76 | 58 | 1 |
| YOK | Count | 3 | 8 | 54 |

Şekil 3.21. Çocuk sahibi olma durumunun kümeler üzerindeki etkisi

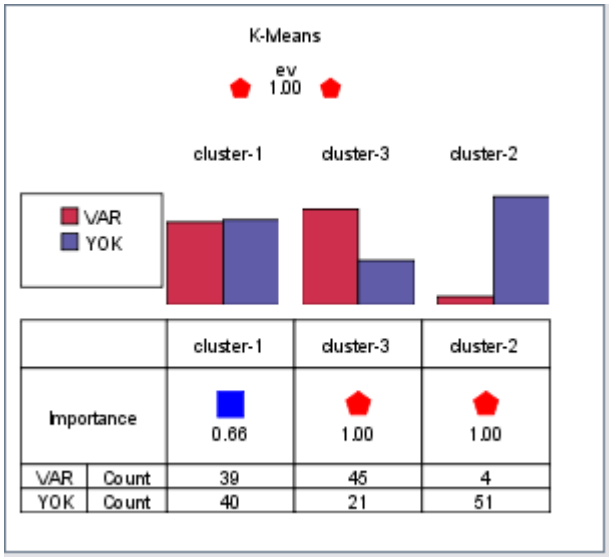
Eş geliri üç kümeyi de etkileyen önemli bir değişkendir. Birinci ve ikinci kümelerdeki müşterilerin sırasıyla %87,34 ve %90,91'lik kesiminde eş geliri mevcut değilken, üçüncü kümedeki müşterilerin %69,7'si eş gelirine sahiptir.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 1.00 |
| VAR | Count | 10 | 46 | 5 |
| YOK | Count | 69 | 20 | 50 |

Şekil 3.22. Eş geliri değişkeninin kümeler üzerindeki etkisi

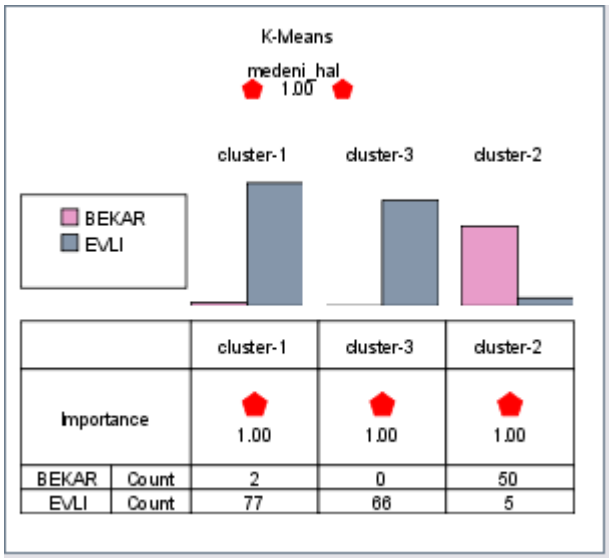
Ev sahibi olma değişkeni ikinci ve üçüncü kümeler için önemliyken, birinci küme için önem taşımamaktadır. Birinci, ikinci ve üçüncü kümelerdeki müşterilerin %49,37, %7,27 ve %68,18' inin kendilerine ait evleri mevcuttur.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 0.66 | 1.00 | 1.00 |
| VAR | Count | 39 | 45 | 4 |
| YOK | Count | 40 | 21 | 51 |

Şekil 3.23. Ev sahibi olma durumunun kümeler üzerindeki etkisi

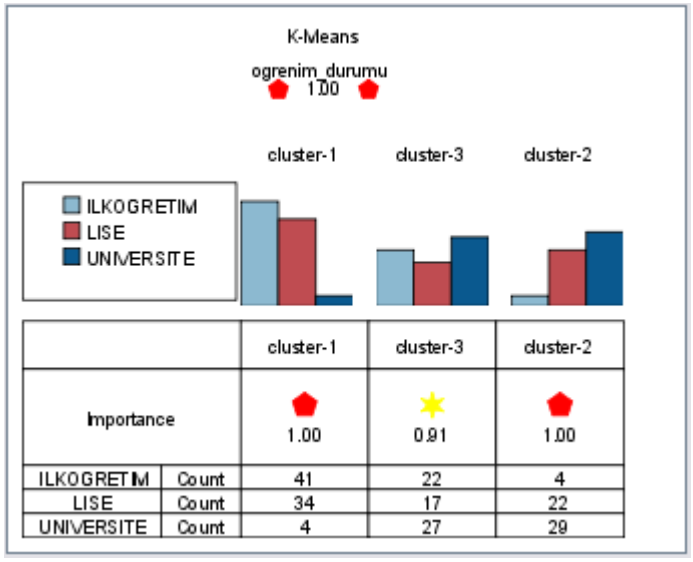
Medeni hal değişkeni her üç küme için de önem taşımaktadır. Birinci kümede %97,47 oranla evli, ikinci kümede ise %90,91 oranla bekar müşteriler çoğunluktadır. Üçüncü kümenin ise tamamı evli müşterilerden oluşmaktadır.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 1.00 |
| BEKAR | Count | 2 | 0 | 50 |
| EVLI | Count | 77 | 66 | 5 |

Şekil 3.24. Medeni hal değişkeninin kümeler üzerindeki etkisi

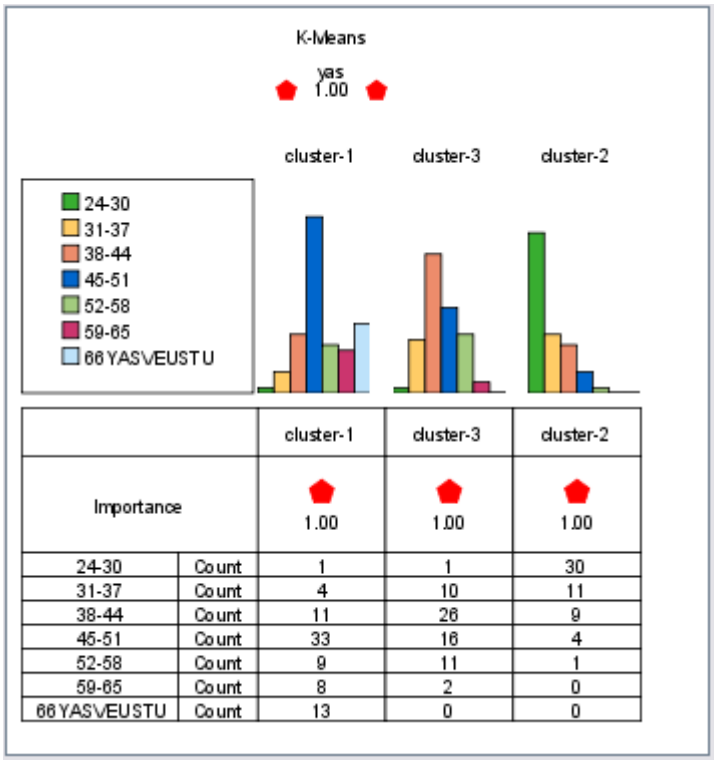
Öğrenim durumu değişkeninin üç küme için de önemli olduğu görülmektedir. Birinci kümede %51,9'luk oranla ilköğretim mezunu müşteriler, ikinci ve üçüncü kümelerde ise %52,73 ve %40,91'lik oranla üniversite mezunu müşteriler çoğunluktadır.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 0.91 | 1.00 |
| ILKOGRETIM | Count | 41 | 22 | 4 |
| LISE | Count | 34 | 17 | 22 |
| UNIVERSITE | Count | 4 | 27 | 29 |

Şekil 3.25. Öğrenim durumu değişkeninin kümeler üzerindeki etkisi

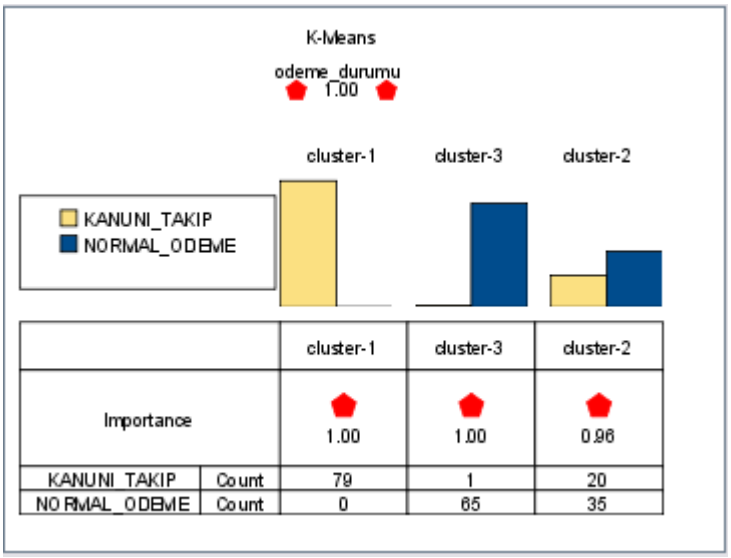
Yaş değişkeninin kümeler üzerindeki etkisi incelendiğinde, bu değişkenin üç küme için de önemli bir unsur olduğu görülmektedir. Birinci kümeyi oluşturan müşterilerin %41,77'lik çoğunluğu 45-51 yaş aralığındadır. İkinci kümede %54,55 oranla 24-30 yaş aralığındaki kişiler çoğunluktadır. Bu kümede 59-65 yaş aralığında ve 66 yaş üstünde müşteri bulunmamaktadır. Üçüncü kümede ise %39,39'luk oranla 38-44 yaş aralığındaki kişilerin çoğunlukta olduğu görülmektedir ve bu kümede de 66 yaş ve üstünde müşteri mevcut değildir.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 1.00 |
| 24-30 | Count | 1 | 1 | 30 |
| 31-37 | Count | 4 | 10 | 11 |
| 38-44 | Count | 11 | 26 | 9 |
| 45-51 | Count | 33 | 16 | 4 |
| 52-58 | Count | 9 | 11 | 1 |
| 59-65 | Count | 8 | 2 | 0 |
| 66YAS\EUSTU | Count | 13 | 0 | 0 |

Şekil 3.26. Yaş değişkeninin kümeler üzerindeki etkisi

Son olarak ödeme durumu değişkeni açısından kümeler incelendiğinde, bu değişkenin üç küme için de önem arz ettiğini görülmektedir. Şekil 3.27'de birinci kümedeki müşterilerin tamamının kredi ödemelerinde sorun yaşamış ve kanuni takibe düşmüş, ikinci ve üçüncü kümelerdeki müşterilerin sırasıyla %63,64 ve %98,48'inin kredi geri ödemelerinde sorun yaşamamış kişilerden oluştuğu görülmektedir.




| | | cluster-1 | cluster-3 | cluster-2 |
|---|---|---|---|---|
| Importance | | 1.00 | 1.00 | 0.96 |
| KANUNI_TAKIP | Count | 79 | 1 | 20 |
| NORMAL_ODEME | Count | 0 | 65 | 35 |

Şekil 3.27. Ödeme durumu değişkeninin kümelere etkisi

Bu veriler ışığında kümelerdeki müşteri profilleri değerlendirildiğinde, birinci kümenin çoğunlukla 45-51 yaş aralığında, kendilerine ait evleri, araçları olmayan, aylık gelirleri 751-1400 TL aralığında olan maaşını farklı bankalardan alan ilköğretim mezunu emekli erkek müşterilerden oluştuğu görülmektedir. Bu kümedeki müşterilerin tamamı kredi ödemelerini aksatarak kanuni takibe düşmüştür. Bu kişilerin tekrar kredi kullanabilmeleri için merkez bankası kayıtlarındaki olumsuzlukların giderilmesi gerekmektedir. Bunun için de ortalama beş yıllık bir süreye ihtiyaç vardır. Birinci kümedeki müşterilerin, bu süre sonunda tekrar kredi talep etmeleri durumunda başvuruları olumsuz değerlendirilebilir ya da risk oranını azaltmak için ipotek veya kefil talep edilebilir.

İkinci küme genellikle 24-30 yaş aralığındaki kamu ve özel sektör çalışanı bekar müşterilerden oluşmaktadır. Diğerlerinin aksine bu kümede bayanların sayısı erkeklere göre daha fazladır. Dikkat edilecek diğer bir husus da bu kümedeki müşterilerin %92,73'lük kısmının kendilerine ait evlerinin bulunmamasıdır. Bu durumda taksitlerini geciktirmeyen müşterilerin konut kredisi kullanmaları teşvik edilebilir.

Üçüncü küme ele alındığında, bu kümenin çoğunlukla aylık gelirleri 1401-2050 TL aralığında olan, maaşlarını kredi kullandıkları bankadan alan, 38-44 yaş aralığında ev ve araç sahibi, kamu çalışanı, emekli erkek müşterilerden oluştuğunu görmekteyiz. Bu kümedeki müşterilerin büyük bir kısmının eş geliri mevcuttur ve %98.48'i ödemelerini düzenli bir şekilde gerçekleştirmektedir. Bu kümedeki müşteriler özel müşteri olarak değerlendirilebilir ve bu müşterilere internet bankacılığı, hesap işletim ücreti olmayan yatırım hesabı, döviz hesabı, herhangi bir kart aidatı olmayan kredi kartları, KGS ve OGS cihazları ile kaza, deprem, yangın gibi durumları içeren sigorta ürünlerinin çapraz satışları yapılabilir. Ayrıca, yeniden kredi talep etmeleri durumunda özel faiz indirimi uygulanabilir. Bu şekilde müşterilerin ilgili bankaya olan bağlılıkları korunabilir.

Kümeleme analizi sonucunda oluşan veri seti Tablo 3.7'de verilmiştir.

Tablo 3.7. Kümeleme analizi sonucu oluşan veri tablosu

| | cinsiyet | medeni_hal | yas | aylik_gelir | es_geliri | ev | arac | cocuk | banka_maas_musterisi | calisma_sekli | ogrenim_durumu | odeme_durumu | $KM-K-Means |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 82 | E | EVLI | 66YASVEUSTU | 751-1400TL | YOK | VAR | YOK | VAR | HAYIR | EMEKLI | LISE | KANUNI_TAKIP | cluster-1 |
| 83 | E | EVLI | 45-51 | 1401-2050TL | YOK | VAR | YOK | VAR | EVET | KAMU | LISE | KANUNI_TAKIP | cluster-1 |
| 84 | E | EVLI | 59-65 | 751-1400TL | YOK | VAR | YOK | VAR | HAYIR | OZEL | ILKOGRETIM | KANUNI_TAKIP | cluster-1 |
| 85 | E | EVLI | 45-51 | 2051-2700TL | YOK | YOK | YOK | VAR | HAYIR | KAMU | ILKOGRETIM | KANUNI_TAKIP | cluster-1 |
| 86 | E | EVLI | 45-51 | 1401-2050TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP | cluster-1 |
| 87 | B | BEKAR | 45-51 | 1401-2050TL | YOK | YOK | YOK | YOK | HAYIR | OZEL | LISE | KANUNI_TAKIP | cluster-2 |
| 88 | B | EVLI | 66YASVEUSTU | 750TLVEALTI | VAR | VAR | YOK | VAR | HAYIR | EMEKLI | ILKOGRETIM | KANUNI_TAKIP | cluster-1 |
| 89 | E | EVLI | 66YASVEUSTU | 751-1400TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | KANUNI_TAKIP | cluster-1 |
| 90 | E | EVLI | 38-44 | 2051-2700TL | YOK | YOK | YOK | VAR | EVET | KAMU | ILKOGRETIM | KANUNI_TAKIP | cluster-1 |
| 91 | E | EVLI | 31-37 | 751-1400TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP | cluster-1 |
| 92 | E | EVLI | 45-51 | 750TLVEALTI | YOK | VAR | YOK | VAR | HAYIR | EMEKLI | ILKOGRETIM | KANUNI_TAKIP | cluster-1 |
| 93 | E | EVLI | 52-58 | 751-1400TL | YOK | YOK | YOK | VAR | EVET | EMEKLI | LISE | KANUNI_TAKIP | cluster-1 |
| 94 | B | BEKAR | 24-30 | 751-1400TL | YOK | YOK | YOK | YOK | HAYIR | OZEL | LISE | KANUNI_TAKIP | cluster-2 |
| 95 | B | BEKAR | 45-51 | 2051-2700TL | YOK | YOK | YOK | YOK | HAYIR | KAMU | LISE | KANUNI_TAKIP | cluster-2 |
| 96 | E | EVLI | 45-51 | 1401-2050TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | KANUNI_TAKIP | cluster-1 |
| 97 | E | EVLI | 38-44 | 751-1400TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP | cluster-1 |
| 98 | E | EVLI | 45-51 | 1401-2050TL | YOK | YOK | YOK | VAR | HAYIR | KAMU | LISE | KANUNI_TAKIP | cluster-1 |
| 99 | E | BEKAR | 38-44 | 751-1400TL | YOK | YOK | YOK | YOK | HAYIR | OZEL | ILKOGRETIM | KANUNI_TAKIP | cluster-2 |
| 100 | E | EVLI | 31-37 | 751-1400TL | YOK | YOK | YOK | VAR | HAYIR | OZEL | LISE | KANUNI_TAKIP | cluster-1 |
| 101 | E | EVLI | 45-51 | 2051-2700TL | VAR | VAR | VAR | VAR | EVET | KAMU | UNIVERSITE | NORMAL_ODEME | cluster-3 |
| 102 | B | BEKAR | 24-30 | 4001TLVEUSTU | YOK | YOK | VAR | YOK | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME | cluster-2 |
| 103 | E | EVLI | 52-58 | 1401-2050TL | VAR | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | NORMAL_ODEME | cluster-3 |
| 104 | E | EVLI | 52-58 | 1401-2050TL | YOK | VAR | VAR | VAR | EVET | EMEKLI | ILKOGRETIM | NORMAL_ODEME | cluster-3 |
| 105 | E | EVLI | 38-44 | 1401-2050TL | VAR | YOK | VAR | VAR | EVET | KAMU | LISE | NORMAL_ODEME | cluster-3 |
| 106 | E | EVLI | 59-65 | 751-1400TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | UNIVERSITE | NORMAL_ODEME | cluster-3 |
| 107 | B | EVLI | 38-44 | 1401-2050TL | VAR | YOK | VAR | VAR | EVET | KAMU | LISE | NORMAL_ODEME | cluster-3 |
| 108 | E | EVLI | 38-44 | 1401-2050TL | VAR | YOK | VAR | VAR | EVET | KAMU | LISE | NORMAL_ODEME | cluster-3 |
| 109 | E | EVLI | 38-44 | 2751-3350TL | VAR | VAR | VAR | VAR | EVET | KAMU | UNIVERSITE | NORMAL_ODEME | cluster-3 |
| 110 | E | BEKAR | 24-30 | 751-1400TL | YOK | YOK | YOK | YOK | EVET | OZEL | LISE | NORMAL_ODEME | cluster-2 |
| 111 | B | BEKAR | 31-37 | 1401-2050TL | YOK | YOK | YOK | YOK | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME | cluster-2 |
| 112 | E | EVLI | 38-44 | 2051-2700TL | VAR | YOK | VAR | VAR | HAYIR | KAMU | LISE | NORMAL_ODEME | cluster-3 |
| 113 | B | EVLI | 31-37 | 2051-2700TL | VAR | VAR | VAR | VAR | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME | cluster-3 |
| 114 | E | BEKAR | 24-30 | 2051-2700TL | YOK | YOK | VAR | YOK | HAYIR | KAMU | UNIVERSITE | NORMAL_ODEME | cluster-2 |
| 115 | E | EVLI | 52-58 | 1401-2050TL | YOK | VAR | YOK | VAR | EVET | EMEKLI | ILKOGRETIM | NORMAL_ODEME | cluster-3 |
| 116 | E | EVLI | 45-51 | 3351-4000TL | VAR | VAR | VAR | VAR | EVET | KAMU | UNIVERSITE | NORMAL_ODEME | cluster-3 |
| 117 | B | BEKAR | 31-37 | 3351-4000TL | YOK | VAR | VAR | YOK | HAYIR | OZEL | UNIVERSITE | NORMAL_ODEME | cluster-2 |

### 3.5.2. Karar ağacı algoritmalarının uygulanması ve algoritma Sonuçları

Kümeleme analizi ile mevcut müşteriler değerlendirildikten sonra, gelecekteki müşterilere yönelik çıkarım yapmak için karar ağacı algoritmalarından yararlanılmıştır. Bu amaçla SPSS Clementine'de C&RT, C5.0, CHAID ve QUEST algoritmaları uygulanmış ve sonuçlar değerlendirilmiştir.

#### 3.5.2.1. C&RT algoritmasına ilişkin sonuç özeti

C&RT algoritması kesikli ve sürekli veriler üzerinde çalışabilen, her dallanmada iki yeni düğüm oluşturan ikili bir karar ağacıdır. Veriyi iki alt kümeye ayırmaktadır ve bir sonraki adımda oluşacak olan alt küme, bir öncekinden daha homojen olmaktadır. C&RT algoritmasından tezin üçüncü bölümünde bahsedilmiştir.

C&RT karar ağacı tekniği uygulaması sonucunda ödeme durumlarına göre müşteriler sınıflara ayrılmıştır. Algoritma sonuçları Ek-1'de detaylı olarak verilmiştir.Buna göre müşterilerin ödeme durumları açısından aylık gelir değişkeni, karar ağacında anlamlı fark yaratan bir değişkendir.

Aylık gelirleri 750 TL ve altı ile 751-1400TL arasında olan müşterilerin %88,89'luk kısmı kanuni takibe düşmüştür ve bu oran 64 kişiye denk gelmektedir. Aylık gelirler baz alındığında bu sınıfta yer alan müşteriler için karar ağacının bir sonraki dalında öğrenim durumu önemli bir ayırıcı kriterdir. Bu gruptaki üniversite mezunu kişilerin tamamı kredilerini aksatmadan geri ödemektedir. İlköğretim ve lise mezunu kişiler için ise ayırt edici diğer kriter banka maaş müşterisi olma durumudur. Bu grupta olup aynı zamanda maaşını farklı bankalardan alan kişilerin tamamı kanuni takiptedir.

Aylık geliri diğer aralıklarda olan müşteriler için yaş değişkeninin, karar ağacının bir sonraki dalında etkili olduğu görülmüştür. Bu grupta yer alan 38-44 yaş aralığındaki, özel sektör çalışanı ve maaşını farklı bankalardan alan kişilerin tamamı kanuni takiptedir. C&RT algoritması sonucunda doğru sınıflandırılan kayıt oranının %95 olduğu görülmektedir.

### 3.5.2.2. C5.0 algoritmasına ilişkin sonuç özeti

C5.0 algoritması C4.5'in geliştirilmiş hali olup, özellikle büyük veri setleri için kullanılan, hızlı ve hafızayı etkili bir şekilde kullanabilen bir karar ağacı algoritmasıdır.

C5.0 algoritması uygulaması ile elde edilen model Ek-2'de verilmiştir. Uygulamada yine ödeme durumu değişkeni bağımlı, geri kalan değişkenlerin tümü bağımsız değişken olarak seçilmiştir. C5.0 algoritması ile ödeme durumuna göre müşterilerin sınıflandırılması yapılırken, karar ağacındaki ilk dallanmanın araç değişkeni ile başladığı görülmüştür. Araç sahibi olmayan müşterilerin %67,94'ü takipte iken bu oran araç sahibi olan müşterilerde %15,95'tir.

Araç değişkeninden sonra karar ağacının dallanmasında etkili olan diğer kriterin banka maaş müşterisi olma durumu olduğu görülmüştür. Buna göre araç sahibi olup, maaşını ilgili bankadan alan müşterilerin hiçbiri kanuni düşmemiştir. C5.0 algoritmasında banka maaş müşterisi olma durumu "Hayır" seçeneği ile ifade edildiğinde oluşan son daldaki kriterin öğrenim durumu olduğu görülmüştür. Banka maaş müşterisi olma durumu "Evet" seçeneğini gösterdiğinde ise dallanmanın sırasıyla eş geliri ve çalışma şekli kriterlerine göre oluştuğu görülmüştür. C5.0 algoritması sonucunda doğru sınıflandırılan kayıt oranının %91 olduğu görülmektedir.

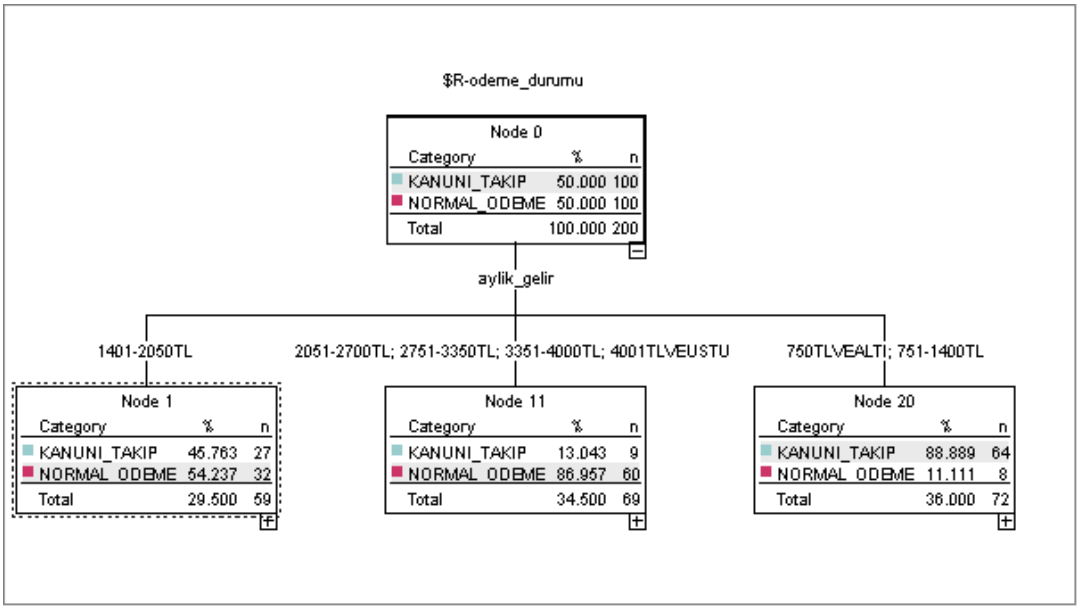
### 3.5.2.3. QUEST algoritmasına ilişkin sonuç özeti

QUEST, ikili karar ağacı yapısı kullanan bir sınıflandırma algoritmasıdır. QUEST algoritması sonuçları Ek-3'de verilmiştir. Ödeme durumu değişkenine göre müşterilerin sınıflandırıldığı uygulamada, karar ağacında oluşan dallanmaların C&RT algoritmasında olduğu gibi aylık gelir değişkeni ile başladığı görülmüştür. Aylık gelirleri 750 TL ve altı ile 751-1400 TL aralığında olan müşteriler için ayırt edici değişken öğrenim durumu olmuştur. Aylık gelir değerleri bu aralıkta olan üniversite mezunu müşterilerin tamamı kredi taksitlerini düzenli olarak ödemektedir. Aylık gelirleri diğer aralıklarda olan müşteriler için ayırt edici kriterler yaş, banka maaş mşterisi olma durumu ve çalışma şekli olmuştur. QUEST algoritması ile oluşturulan modelin doğruluk payının %92,5 olduğu görülmektedir.

### 3.5.2.4. CHAID algoritmasına ilişkin sonuç özeti

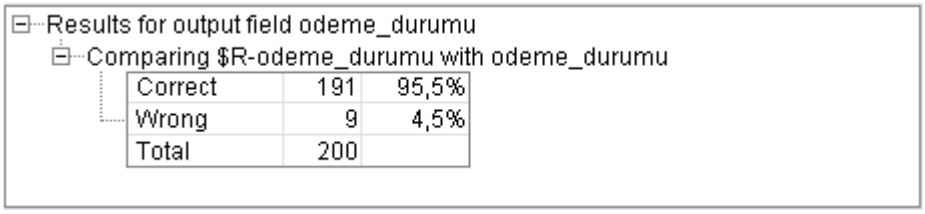
CHAID algoritması, optimal bölünmelerin teşhisi için ki-kare istatistiğini kullanan bir yöntemdir. CHAID, sürekli ve kategorik tüm değişken tipleriyle çalışabilmekte ve ki-kare metriği vasıtasıyla, ilişki düzeyine göre farlılık rastlanan grupları ayrı ayrı sınıflamaktadır. Ağacın yaprakları ikili değil, verideki farklı yapı sayısı kadar dallanmaktadır (Albayrak ve Yılmaz, 2009). CHAID algoritmasından tezin üçüncü bölümünde bahsedilmiştir.

Ödeme durumunun hedef değişken olarak seçildiği uygulamada karar ağacındaki ilk dalın formuna göre en iyi ön kestirici değişkenin aylık gelir olduğu görülmüştür. Aylık gelir değişkenine göre müşteriler Şekil 3.28'deki gibi üç farklı şekilde sınıflandırılmıştır. Buna göre aylık gelirleri 750TL ve altı ile 751-1400 TL aralığında olan müşterilerin ise %88,89'u, 1401-2050 TL aralığında olan müşterilerin %45,76'sı kanuni takiptedir. Diğer gelir aralıklarındaki müşteriler %86,96'luk oranla kredi taksitlerini düzenli bir şekilde ödemektedirler. CHAID algoritması sonucuna göre doğru sınıflandırılan kayıt oranı %95,5 olmuştur.




Şekil 3.28. CHAID algoritması ile karar ağacında oluşan ilk dal

CHAID algoritmasına ait doğruluk oranı Şekil 3.29'da verilmiştir. Algoritma sonucunda doğru sınıflandırılan kayıt oranı %95,5 olarak tanımlanmıştır. Bu durum toplam 200 müşteriden 191'inin sınıflandırılmasının doğru bir şekilde yapıldığını ifade etmektedir.




Şekil 3.29. CHAID algoritması ile elde edilen modelin doğruluk oranı

### 3.5.2.5. Algoritma sonuçlarının karşılaştırılması

Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır (Akpınar, 2000).

Uygulamada kullanılan karar ağacı algoritmalarına ilişkin sonuçlar Tablo 3.9'da verilmiştir. Buna göre doğruluk payının en yüksek olduğu model %95,5'lik oranla CHAID algoritması ile elde edilmiştir.

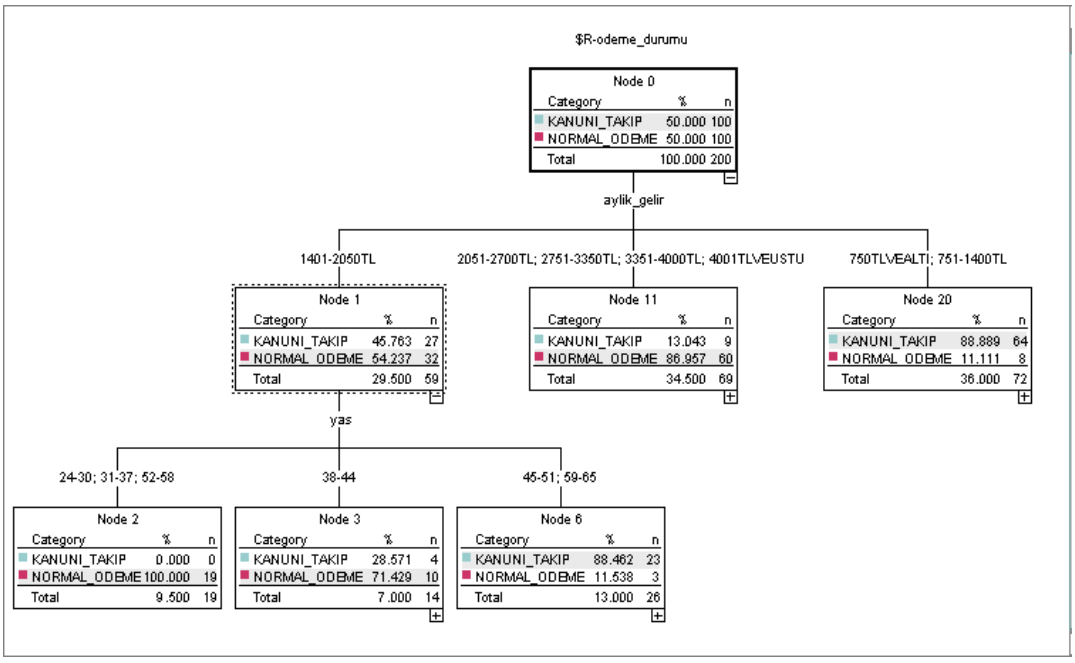
Tablo 3.8. Algoritma sonuçlarına ilişkin değerler

| | C&RT | C5.0 | QUEST | CHAID |
|---|---|---|---|---|
| Doğru sınıflandırılan müşteri sayısı | 190 | 182 | 185 | 191 |
| Hatalı sınıflandırılan müşteri sayısı | 10 | 18 | 15 | 9 |
| Modelin Doğruluk Oranı(%) | 95 | 91 | 92,5 | 95,5 |

### 3.5.2.6. CHAID algoritmasına ait sonuçlarının yorumlanması

CHAID algoritması ile oluşan karar ağacındaki ilk dallanma aylık gelir değişkeni ile başlamıştır. Aylık gelir değişkenine göre müşterilerin üç sınıfta toplandığı Şekil3.30'da gösterilmektedir.

Aylık gelirleri 1401-2050 TL aralığında olan müşteriler için ayırt edici ikinci kriter yaş değişkeni olmuştur. Buna göre "24-30, 31-37 ve 52-58" yaş aralığındaki müşterilerin tamamı kredi taksitlerini aksatmadan ödemektedir. Yaş aralığı "45-51 ve 59-65" olan müşterilerin ise %88,46'sı kanuni takiptedir. 38-44 yaş aralığındaki kişilerin %71,43'ü ödemelerini aksatmadan yapmaktadır. Bu gruptaki müşterileri ödeme durumlarına göre ayıran diğer kriter banka maaş müşterisi olma durumudur. Yaş aralığı 38-44 olan ve maaşlarını farklı bankalardan alan müşterilerin tamamı kanuni takipte iken bu yaş aralığında olan ve maaşını ilgili bankadan alan müşterilerin kredi taksitlerini geri ödemede sorun yaşamadıkları görülmektedir.

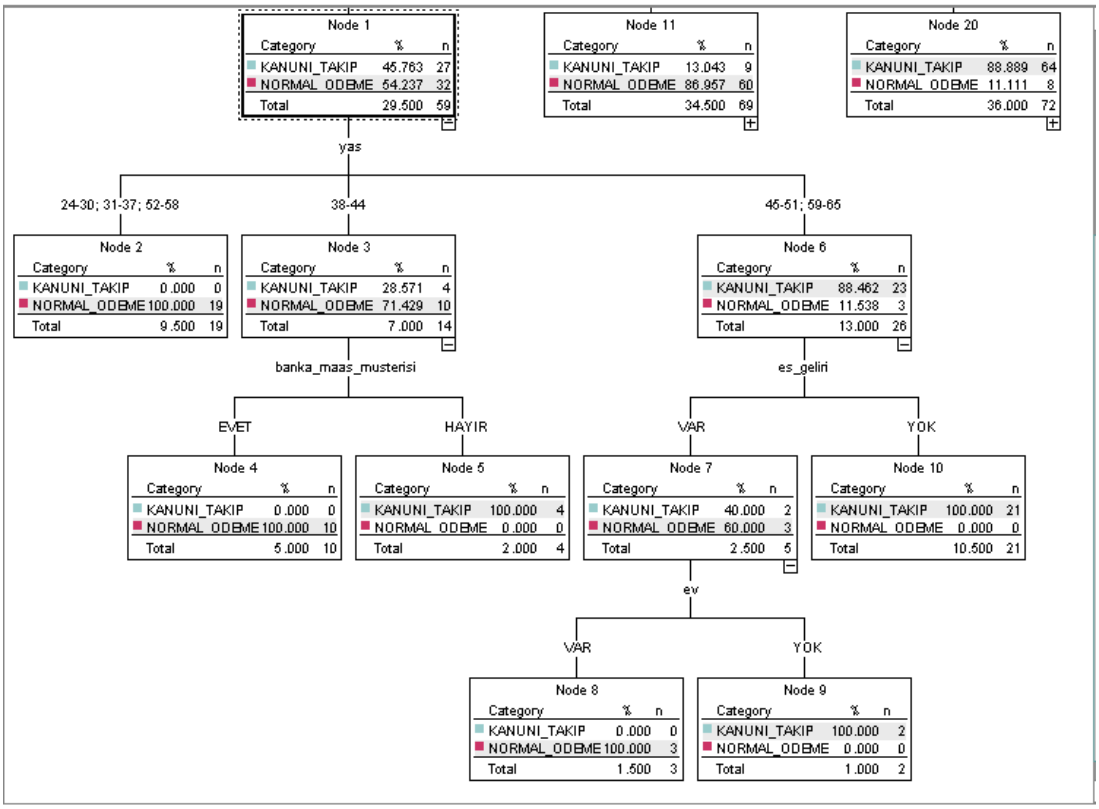



Şekil 3.30. 1401-2050 TL aralığındaki aylık gelir durumuna ilişkin karar ağacı

1401-2050 TL aylık gelire ve 45-51 ile 59-65 yaş aralığına sahip müşterilerin sınıflandırılmasında etkili olan diğer bir değişken eş geliridir. Yaş aralıkları bu şekilde olup, eş geliri mevcut olmayan müşterilerin tamamı kanuni takiptedir. Şekil3.31'den bu kategorideki müşteri sayısının 21 olduğu görülmektedir. Eş geliri mevcut olan müşteriler son olarak ev sahibi olma durumuna göre kendi içinde sınıflara ayrılmaktadır. Kendilerine ait evi bulunan müşterilerin tamamının "Normal Ödeme", ev sahibi olmayan müşterilerin tamamının ise "Kanuni Takip" kategorileri altında toplanmıştır.

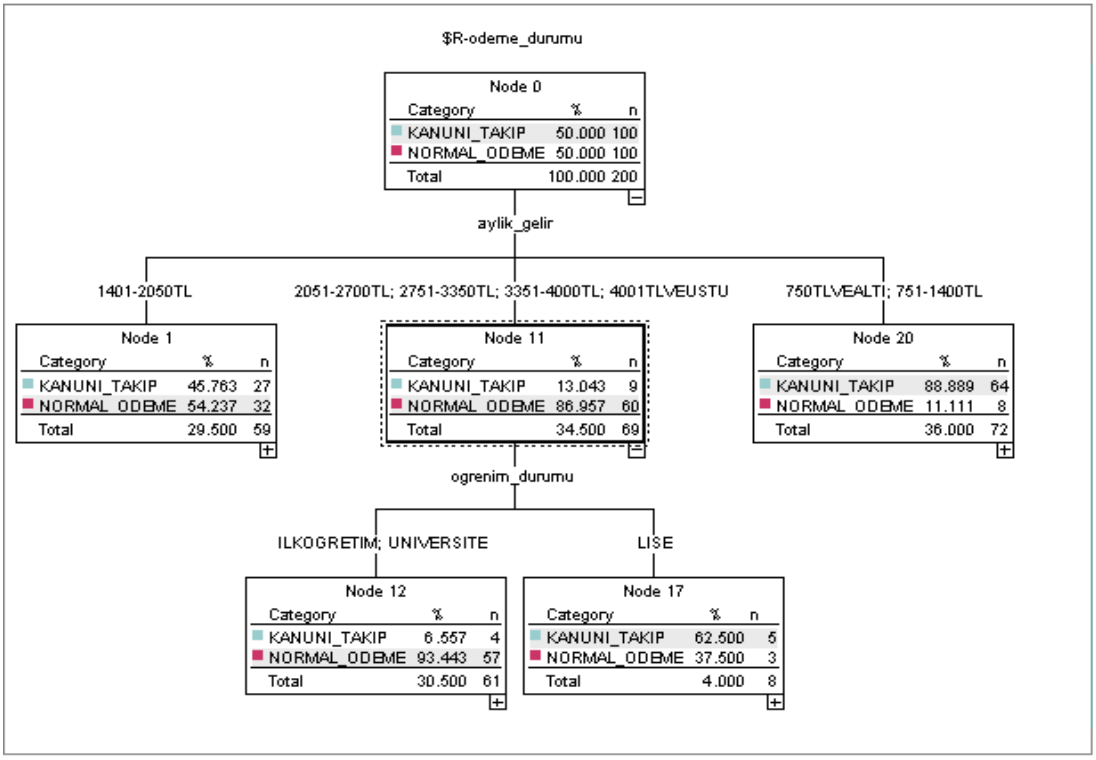
Karar ağacı yapısına göre aşağıdaki gibi kural çıkarımı yapmak mümkündür:

Eğer kredi kullanan müşterinin aylık geliri 1401-2050 TL aralığında ise; müşteri 45-51 yada 59-65 yaş aralığına sahipse ve eş geliri mevcut değilse kanuni takiptedir. Eğer müşterinin eş geliri ve evi mevcut ise kanuni takipte değildir.



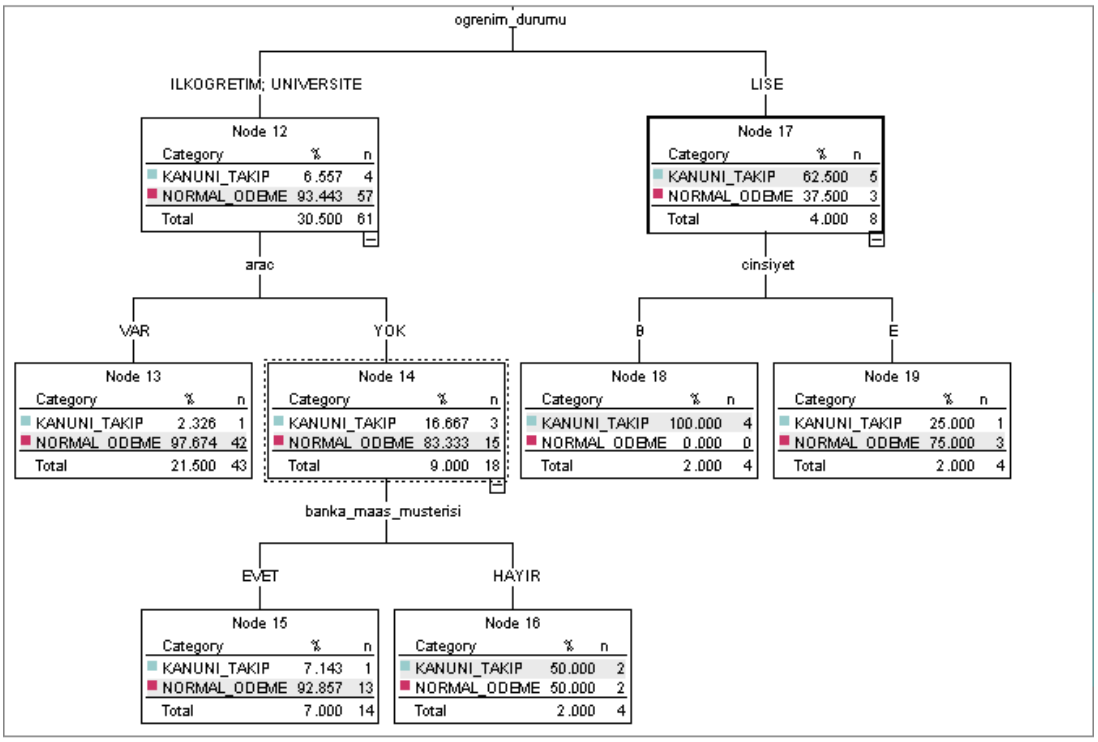

Şekil 3.31. Yaş değişkenine ilişkin karar ağacı

Aylık geliri 2051-4000 TL ile 4001 TL ve üstünde olan müşterilerin ödeme durumu hedef değişkenine göre sınıflandırılmasını sağlayan diğer bir bağımsız değişken öğrenim durumudur. Şekil 3.32 incelendiğinde bu gruptaki ilköğretim ve üniversite mezunu müşterilerden %93,44'ü ödemelerini aksatmazken, lise mezunu olanların %62,5'i kanuni takipte olduğu görülmektedir. Lise mezunu müşterilerin kategorize edilmesindeki diğer önemli etken cinsiyet değişkeni olmuştur. Buna göre, lise mezunu bayan müşterilerin tamamı, erkek müşterilerin ise %25'lik kısmı kanuni takip grubunda yer almaktadır.




Şekil 3.32. 2051-4001 TL ve üzerindeki aylık gelir durumuna ilişkin karar ağacı

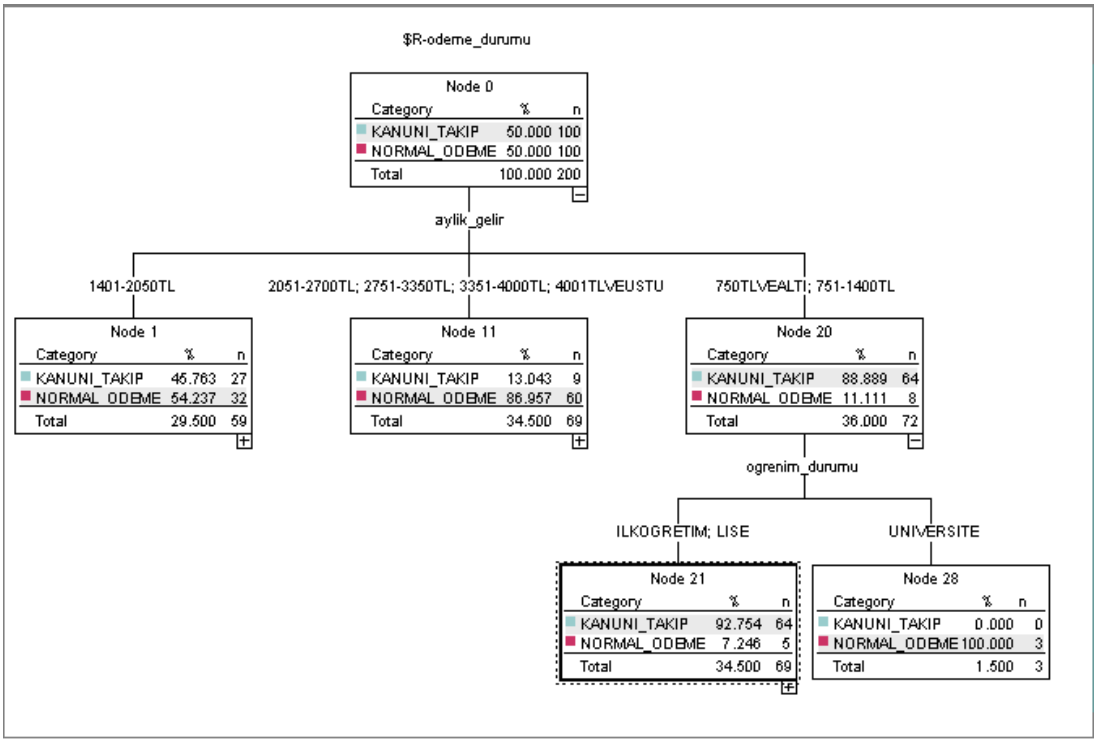
2051-4000 TL ile 4001 TL ve üstünde aylık gelire sahip olan ilköğretim ve üniversite mezunu müşterilerin sınıflandırılmasında etkili olan kriterler ise sırasıyla araç sahipliği ve ilgili bankanın maaş müşterisi olma durumudur. Bu değişkenlerin hedef alan üzerindeki etkileri Şekil 3.33'deki gibidir. Buna göre araç sahibi olan 43 müşteriden sadece 1'i kanuni takiptedir ve bu oran %2,33'e denk gelmektedir. Araç sahibi olmayan müşterilerin %83,33'lük kısmı kredi eri ödemelerini düzenli bir şekilde yapmaktadır. Bu grupta bulunup maaşını farklı bankalardan alan müşterilerin yarısı kanuni takiptedir. Maaşlarını ilgili bankadan alan müşterilerden sadece biri kanuni takiptedir ve bu oran %7,14'e karşılık gelmektedir.




Şekil 3.33. 2051-4001 TL ve üzerinde gelire sahip müşterilerin öğrenim durumu değişkenine göre sınıflandırılmasına ilişkin ağaç yapısı

Şekil 3.34 incelendiğinde aylık geliri 750 TL ve altı ile 751-1400 TL aralığında olan müşterilerin %88,89'luk çoğunluğu kanuni takipte olduğu görülmektedir. Bu gruptaki müşterilerin hedef değişkene göre sınıflandırılmasını sağlayan diğer bir bağımsız değişken öğrenim durumudur. Gelir durumu bahsedilen aralıklarda olan üniversite mezunu müşterilerin hiçbiri kanuni takibe düşmemiştir. İlköğretim ve lise mezunu müşterilerde ise kanuni takip oranı %92,75'tir. Bu oran toplamda 69 kişiye denk gelmektedir.

Üniversite mezunu kişilerin sınıflandırılması öğrenim durumu kriteri ile tamamlanırken, ilköğretim ve lise mezunu kişiler için karar ağacı ilgili bankanın maaş müşterisi olma durumuna at değişken ile dallanmaya devam etmektedir.

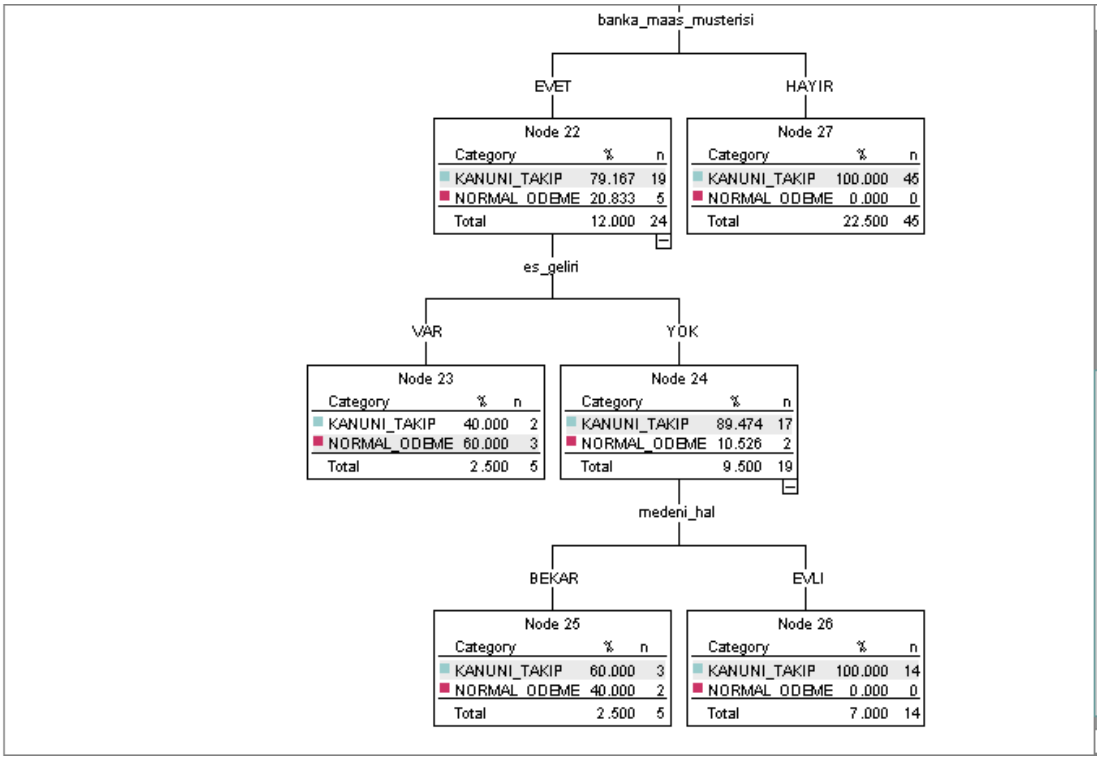



Şekil 3.34. 750 TL ve altı ile 751-1400 TL gelir aralığına ilişkin karar ağacı

İlköğretim ve lise mezunu olup, maaşını ilgili bankadan almayan müşterilerin tamamının kanuni takibe düştüğü Şekil 3.35'te görülmektedir. Bu grupta toplam 45 kişi mevcuttur. Kredi kullandıkları bankanın aynı zamanda maaş müşterisi olan kişilerin ise %79,17'lik çoğunluğu takiptedir.

İlgili bankanın maaş müşterilerini sınıflandırmada etkili olan diğer bir kriter ise eş geliridir. Eş geliri olan müşterilerin % 40'ı kanuni takipte iken bu oranın eş geliri olmayan müşterilerde %89,47'ye yükseldiği görülmektedir.

Eş geliri mevcut olmayan müşteriler son olarak medeni hal değişkeni ile sınıflara ayrılmıştır. Bu gruptaki bekar müşterilerin %40'ı ödemelerini aksatmadan gerçekleştirirken, evli müşterilerin ise tamamı kredi taksitlerini ödemede zorlanarak kanuni takibe düşmüştür.




Şekil 3.35. 750 TL ve altı ile 751-1400 TL gelir aralığındaki ilköğretim ve lise mezunu müşterilere ilişkin karar ağacı

CHAID algoritmasında bağımsız değişkenlerden "çocuk sahibi olma durumu" ve "çalışma şekli" değişkenlerine rastlanmamıştır. Bu durum, iki değişkenin de hedef değişken olan ödeme durumu üzerinde önemli bir etkisinin olmamasından kaynaklanmaktadır.

### 3.6. Modelin Kullanılması

Kümeleme analizi sonucu k-ortalamalar yöntemi ile elde edilen modelde, mevcut müşteriler kümelere ayrılmıştır. Bu model, kümelerdeki müşteri profillerine göre çapraz satışı yapılacak ürünlerin tespiti için kullanılabilir. Böylece ilgili banka şubesinin çapraz satış oranlarında artış yaratılması sağlanarak, şubenin performans sıralamasındaki yeri üst seviyelere taşınabilir.

Karar ağacı algoritmaları ile yapılan analizler sonucunda doğruluk oranları dikkate alındığında, diğer algoritmalara kıyasla CHAID algoritması ile elde edilen modelin geçerliliğinin daha yüksek olduğu görülmektedir. Bu model kullanılarak, kredi başvurusunda bulunacak olan müşterilerin belirli özellikleri modele yerleştirilip, ödeme durumunun ne olacağı konusunda bir çıkarım yapılabilir ve duruma göre

müşteriden kefil talep edilebilir. Başka bir alternatif ise, müşterilerin ödeme durumları için yapılacak değerlendirmede risk oranı çok düşük seviyede ise müşteri için özel faiz indirimi uygulanabilir ve bankaya bağlılığı sağlanabilir.

### 3.7. Modelin İzlenmesi

Veri madenciliği sürecinin son aşaması, geçerliliği kabul edilen ve kullanılan modellerin izlenmesidir. Zaman içerisinde kredi başvurularında dikkat edilecek kriterlerin değişmesine bağlı olarak, kurulan modeller sürekli izlenmeli ve gerekirse düzenlemeler yapılmalıdır.

# 4. SONUÇLAR VE ÖNERİLER

Veri madenciliği içerdiği tekniklerle, veri yığınları içerisinde gizli kalmış olan anlamlı bilgilere ulaşmayı sağlayan bir süreçtir. Pazarlama, finans, üretim, sağlık, müşteri ilişkileri yönetimi gibi alanlarda olduğu gibi bankacılık alanında da karar verme sürecine duyulan ihtiyaçtan ötürü veri madenciliği yaygın olarak kullanılmaktadır.

Sürdürülebilir bir büyüme sağlamak amacıyla işletmelerin hem yeni müşteri kazanmak hem de mevcut müşterileri elde tutmak için faaliyetlerde bulunmaları gerekmektedir. Araştırmalara göre yeni bir müşteri kazanmanın maliyeti mevcut bir müşteriyi elde tutma maliyetinin çok üstündedir. Bu nedenle öncelikli olarak mevcut müşteriler tanınmalı ve bu müşterilere yönelik çalışmalar düzenlenerek işletmeler için pazarda rekabet avantajı sağlayacak olan sadık müşteri portföyü oluşturulmalıdır.

Bu çalışmada bankacılık sektöründe bir uygulamaya yer verilerek, bireysel kredi müşterilerinin ödeme performansları değerlendirilmiştir. Çalışma kapsamında öncelikle, mevcut kredi müşterileri için k-ortalamalar yöntemi ile kümeleme analizi yapılmıştır. SPSS Clementine ile gerçekleştirilen analiz sonucunda verideki gizli bilgiler açığa çıkarılarak, müşteriler davranışlarına göre gruplandırılmıştır. Müşterilerin kümelere ayrılmasında yaş, cinsiyet, aylık gelir, medeni hal, öğrenim durumu, ödeme durumu gibi on iki farklı değişkenden yararlanılmıştır. Tüm değişkenlere göre kümelerdeki müşterilerin dağılımları değerlendirilmiştir.

Ödeme durumlarına göre kümeler incelendiğinde ilk kümeyi oluşturan müşterilerin tamamının kanuni takipte olduğu, ikinci ve üçüncü kümedeki müşterilerin sırasıyla %63,64 ve %98,48'lik oranlarla ödemelerini aksatmadıkları sonucuna ulaşılmıştır. Bu veriler ışığında mevcut müşterilerin tekrar kredi talep etmeleri durumunda, bulundukları kümelere bağlı olarak başvuruları değerlendirmeye alınabilir. İlk kümedeki müşteriler için bu sonuç olumsuz olarak değerlendirilebileceği gibi, risk

oranını azaltmak için kefil de talep edilebilir. Üçüncü kümedeki müşterilerin başvurularına kısa süre içerisinde olumlu dönüş yapılabilir ve bu kümedeki müşterilerin sadakatini kazanmak adına faiz indirimi gibi özel kampanyalar düzenlenebilir. Analize konu olan diğer değişkenlerin müşteriler üzerindeki etkisi dikkate alınarak, ödemelerini düzenli yapan müşteriler için tüketici kredilerinin yanında kredi kartları OGS cihazları, internet bankacılığı, mevduat hesapları gibi bankanın diğer ürünlerinin sunumu da yapılabilir. Bu şekilde müşteriler için faydalı olabilecek ürünlerin satışı yapılarak şube ve dolayısıyla banka performansında artış da sağlanabilir.

Uygulamanın ikinci kısmında C&RT, C5.0, QUEST ve CHAID algoritmaları ile müşteriler sınıflandırılmış ve geleceğe yönelik tahminlerde bulunabilmek için karar ağaçları ile kural çıkarımı sağlanmıştır. Clementine çıktılarına göre algoritmalar karşılaştırılmış ve CHAID algoritmasının %95,5 'lik doğru sınıflandırılan kayıt oranı ile en iyi sonucu sağladığı tespit edilmiştir. Aykırı değerlerin belirlenip çıkarılması ve karar ağacı seviyesinin arttırılması ile daha da yüksek tahmin başarısı elde edilebilir.

CHAID algoritmasının çıktıları incelendiğinde ödeme durumunun hedef değişken olarak seçildiği uygulamada, karar ağacında ilk dallanmanın aylık gelir ile başladığı gözlemlenmiştir. Yani karar kuralları oluşturmada ilk dikkat edilecek nokta olarak "aylık gelir" ön plana çıkmıştır. Daha sonraki dallanmalar ise aylık gelir aralıklarına göre değişiklik göstermiştir. Diğer algoritmaların aksine CHAID algoritmasında ağacın yapraklarının ikili değil verideki farklı yapı sayısı kadar dallanabileceği görülmüştür . Bu özelliği nedeniyle, CHAID algoritması ile daha fazla alt gruplarla değerlendirme yapmak ve daha homojen gruplardan sonuç çıkarmak mümkün olmaktadır.

CHAID algoritmasında müşterilerin ödeme durumlarına ilişkin kural çıkarımı yapılırken, bağımsız değişkenlerden "çocuk sahibi olma durumu" ve "çalışma şekli" değişkenlerinin kullanılmadığı gözlemlenmiştir. Bu durum, iki değişkenin de hedef değişken olan ödeme durumu üzerinde önemli bir etkisi olmadığını göstermektedir. Kredi müşterilerinin çalışma şekillerinden ziyade ilgili bankanın maaş müşterisi olma durumları kredi geri ödemelerinde daha önemli bir etkendir. Çünkü müşterilerin

takibe düşme olasılığının azaltılması adına kredi ödemeleri, taksi müşterilerin maaş hesaplarından otomatik ödeme yoluyla gerçekleşmektedir. Düzenli ödeme tarihi ile müşterilerin maaş tarihleri aynı güne denk getirilerek kredi taksitinin hesaptan otomatik olarak çekilmesi sağlanmaktadır. Ayrıca bazı nedenlerden dolayı maaş alan müşterilerin kredi taksitlerinde aksama meydana gelmesi durumunda sistem tarafından hesaba bloke koyulmakta ve müşterinin bir sonraki maaşından yada hesaba ilk yatan tutar üzerinden tahsilat gerçekleştirilerek aksamanın ortadan kaldırılması sağlanmaktadır.

Sonuç olarak, bankaların sektörde rekabet avantajı sağlayarak uzun süreler ayakta kalabilmeleri için müşterilerini doğru bir şekilde tanımaları ve riskli müşterileri diğerlerinden ayırabilmeleri gerekmektedir. Tez kapsamında gerçekleştirilen çalışma ile veri madenciliği yöntemlerinden kümeleme ve sınıflandırma algoritmaları kullanılarak bireysel kredi müşterilerinin ödeme performanslarına göre profilleri çıkarılmış ve gelecekteki kredi müşterileri için kurallar oluşturulmuştur. Veriler ışığında bireysel kredilerde kanuni takibe düşme oranının azaltılması hedeflenmiştir. Verimliliği artırılabilmek amacıyla bankanın diğer departmanlarında da benzer uygulamalar gerçekleştirilebilir.

## KAYNAKLAR


Akbulut S., Veri madenciliği teknikleri ile bir kozmetik markanın ayrılan müşteri analizi ve müşteri segmentasyonu, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2006, 180239.

Akpınar H., Veritabanlarında bilgi keşfi ve veri madenciliği, *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 2000, **29**, 1-22.

Albayrak A. S., Türkiye'de yerli ve yabancı ticaret bankalarının finansal etkinliğe göre sınıflandırılması: karar ağacı, lojistik regresyon ve diskriminant analizi modellerinin bir karşılaştırılması, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 2009, **14**, 113-139.

Albayrak A. S., Yılmaz Ş. K., Veri madenciliği: Karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 2009, **14**, 31-52.

Argüden Y., Erşahin B., *Veri madenciliği*, 1. Baskı., ARGE Danışmanlık Yayınları, İstanbul, 2008.

Aşan Z., Kredi kartı kullanan müşterilerin sosyo ekonomik özelliklerinin kümeleme analizi ile incelenmesi, *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 2007, **17,** 256-267.

Aşlıyan R., Günel K., Metin içerikli Türkçe dökümanların sınıflandırılması, XII. *Akademik Bilişim Konferansı*, Muğla, Türkiye, 10-12 Şubat 2010.

Atbaş A. C., Kümeleme analizinde küme sayısının belirlenmesi üzerine bir çalışma, Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, 2008, 233366.

Bilen H., Bankacılık sektöründe personel seçimi ve performans değerlendirilmesine ilişkin veri madenciliği uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 2009, 233733.

Bircan H., Lojistik regresyon analizi: tıp verileri üzerine bir uygulama, *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 2004, **8**, 185-208.

Chien C. F., Chen L. F., Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry, *Expert Systems with Applications*, 2008 , **34**, 280-290.

Ching W. K., Pong M. K.., *Advances in data mining and modeling*, 1st ed., World Scientific, Hong Kong, China, 2002.

Coşkun S., Coşkun A., Kartal M., Bircan H., Lojistik regresyon analizinin incelenmesi ve diş hekimliğinde bir uygulaması, *Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi*, 2004, **7**, 41-50.

Çakır Ö., Veri madenciliğinde sınıflandırma yöntemlerinin karşılaştırılması: bankacılık müşteri veri tabanı üzerinde bir uygulama, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, 2008, 226629.

Çakmak Z., Uzgören, N., Keçek, G., Kümeleme analizi teknikleri ile illerin kültürel yapılarına göre sınıflandırılması ve değişimlerinin incelenmesi, *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 2005, **12**, 15-36.

Çalışkan S. K. ve Soğukpınar, İ., KxKNN: K-means ve K en yakın komşu yöntemleri ile ağlarda nüfuz tespiti, *2. Ağ ve Bilgi Güvenliği Sempozyumu*, Girne, KKTC, 16-18 Mayıs 2008.

Çetinyokuş T., Veri küplerinin bütünleşik kullanımına yönelik yeni bir OLAP mimarisi, Doktora Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 2008, 233888.

Çil F., Banka yatırım fonu müşteri hareketlerinin belirlenmesine yönelik bir veri madenciliği uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 2010, 303522.

Doğan B., Bankaların gözetiminde bir araç olarak kümeleme analizi: Türk bankacılık sektörü için bir uygulama, Doktora Tezi, Kadir Has Üniversitesi, Sosyal Bilimler Enstitüsü, 2008, 214722.

Emel G. G., Taşkın Ç., Veri madenciliğinde karar ağaçları ve bir satış analizi uygulaması, *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi*, 2005, **6**, 221-239.

Emel G. G., Taşkın Ç., Genetik algoritmalar ve uygulama alanları, *Uludağ Üniversitesi İ.İ.B.F. Dergisi*, 2002, **21**, 129–152.

Fayyad U., Shapiro G., Smyth P., From data mining to knowledge discovery in databases, *American Associatin for Artificial Intelligence*, 1996, **17**, 37-54.

Gülçe G., Veri ambarı ve veri madenciliği teknikleri kullanılarak öğrenci karar destek sistemi oluşturma, Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, 2010, 275300.

Han J., Kamber M., *Data mining: concepts and techniques*, 2nd ed., Morgan Kaufmann, USA, 2006.

Hand D. J., Data mining: statistics and more? , *The American Statistician*, 1998, **52**, 112-118.

Hudairy H., Data mining and decision making support in the governmental sector, Master Thesis, Faculty of Graduate School of The University of Louisville, Kentucky, 2004.

Jacobs P., Data mining: what general managers need to know, *Harvard Management Update*, 1999, **4** , 8-9.

Kalıkov A., Veri madenciliği ve bir e-ticaret uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 2006, 180400.

Kalogirou S. A., Applications of artifcial neural-networks for energy systems, *Applied Energy* , 2000, **67**, 17-35.

Karacan H., Yeşilbudak M., Kullanıcı merkezli interaktif veri madenciliği: bir literatür taraması, *Bilişim Teknolojileri Dergisi*, 2010, **3**, 17-22.

Kittler R., Wang W., The emerging role in data mining, *Solid State Technology*, 1999, **42**, 45-58.

Köktürk F., Ankaralı H., Sümbüloğlu V., Veri madenciliği yöntemlerine genel bakış, *Türkiye Klinikleri*, 2009, **1**, 20-25.

Küçüksille E., Veri madenciliği süreci kullanılarak portföy performansının değerlendirilmesi ve İMKB hisse senetleri piyasasında bir uygulama, Doktora Tezi Süleyman Demirel Üniversitesi, Sosyal Bilimler Enstitüsü, 2009, 231614.

Larose D. T., *Discovering knowledge in data: an introduction in data mining*, 1st ed., Wiley, USA, 2005

Linoff G. S., Berry M. J. A., *Data mining techniques for marketing, sales and customer relationship management*, 3rd.ed.,Wiley, Canada, 2011

Marakas G. M ., *Decision Support Systems in the 21st century*, 2nd ed. ,Prentice Hall, USA, 2003.

Oğuzlar A., CART analizi ile hanehalkı işgücü anketi sonuçlarının özetlenmesi", *Atatürk Üniversitesi İİBF Dergisi*, 2004, **18,** 79-90.

Olgun M. O., Özdemir G., İstatistiksel özellik temelli Bayes sınıflandırıcı kullanarak kontrol grafiklerinde örüntü tanıma, *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 2012, **27**, 303-311.

Özçakır F. C., Çamurcu A.Y., Birliktelik kuralı yöntemi için veri madenciliği yazılımı tasarımı ve uygulaması, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 2007, **12**, 21-37.

Özdamar K., *Paket programlar ile istatistiksel veri analizi I*, 5. Baskı, Kaan Kitabevi, Eskişehir, 2004.

Özdamar K., *Paket programlar ile istatistiksel veri analizi II*, 5. Baskı, Kaan Kitabevi, Eskişehir, 2004.

Özekes S., Çamurcu A.Y., Veri madenciliğinde sınıflama ve kestirim uygulaması, *Marmara Üniversitesi Fen Bilimleri Dergisi*, 2002, **18**, 1-17.

Sancak S., Saldırı tespit sistemleri tekniklerinin karşılaştırılması, Yüksek Lisans Tezi, Gebze İleri teknoloji Enstitüsü, Sosyal Bilimler Enstitüsü, 2008, 220150.

Savaş S., Topaloğlu N., Yılmaz M., Veri madenciliği ve Türkiye'deki uygulama örnekleri, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 2012, **21**, 1-23.

Sever H., Oğuz B., Veri tabanlarında bilgi keşfine formel bir yaklaşım, *Informatıon World*, 2002, **3**, 173-204.

Sezer E. A., Bozkır A. S., Yağız S., Gökçeoğlu C., Karar ağacı derinliğinin CART algoritmasında kestirim kapasitesine etkisi: bir tünel açma makinesinin ilerleme hızı üzerinde uygulama, *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*, Kayseri, Türkiye, 21-24 Haziran 2010.

Suner A., Çelikoğlu C. C., Toplum tabanlı bir çalışmada çoklu uygunluk analizi ve kümeleme analizi ile sağlık kurumu seçimi, *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 2010, **25**, 43-55.

Tan P. N., Steinbach M., Kumar V., *Introduction to Data Mining*, 1st ed., Pearson International Edition, USA, 2006.

Terzi Ö., Küçüksille E. U., Ergin G., İlker A., Veri Madenciliği Süreci Kullanılarak Güneş Işınımı Tahmini, *SDU International Technologic Science*, 2011, **3**, 29-37.

Tosun T., Veri madenciliği teknikleriyle kredi kartlarında müşteri kaybetme analizi, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, 2006, 223366.

Tuffery S., *Data mining and statistics for decision making*, 1st ed., Wiley, USA, 2011.

Yapıcı A. P., Özel A., Ayça C., Oracle data miner ile kredi ödemeleri üzerine bir veri madenciliği uygulaması, Bitirme Ödevi, 2010.

URL-1: http://www.kdnuggets.com/polls/2011/industries-applied-anaytics-data-mining.html (Ziyaret tarihi: 5 Aralık 2012).

URL-2:Alpaydın E., Zeki veri madenciliği: Ham veriden atın bilgiye ulaşma yöntemleri, http://www.cmpe.boun.edu.tr/~ethem/files/.../veri-maden_2k-notlar.doc (Ziyarte tarihi: 10 Nisan 2012).

URL-3: http://www.spss.com.tr/Veri.html (Ziyaret tarihi: 14 Kasım 2012).

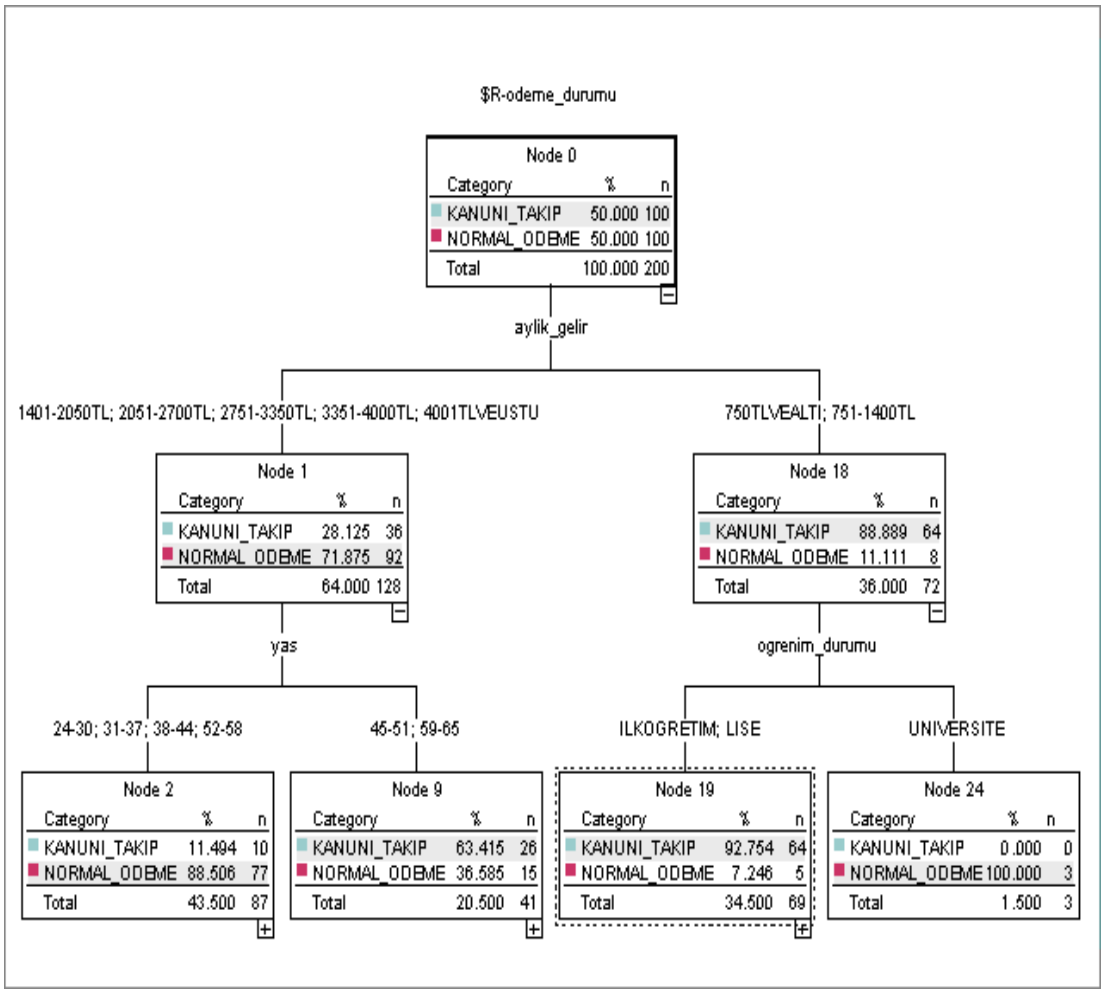
URL-4:http://sbu.saglik.gov.tr/Ekutuphane/kitaplar/biyoistatistik%20(6).pdf, (Ziyaret tarihi: 8 Kasım 2012).

Witten I. H., Frank E., Data Mining: practical machine learning tools and techniques, 2nd ed., Morgan Kaufmann, USA, 2005.

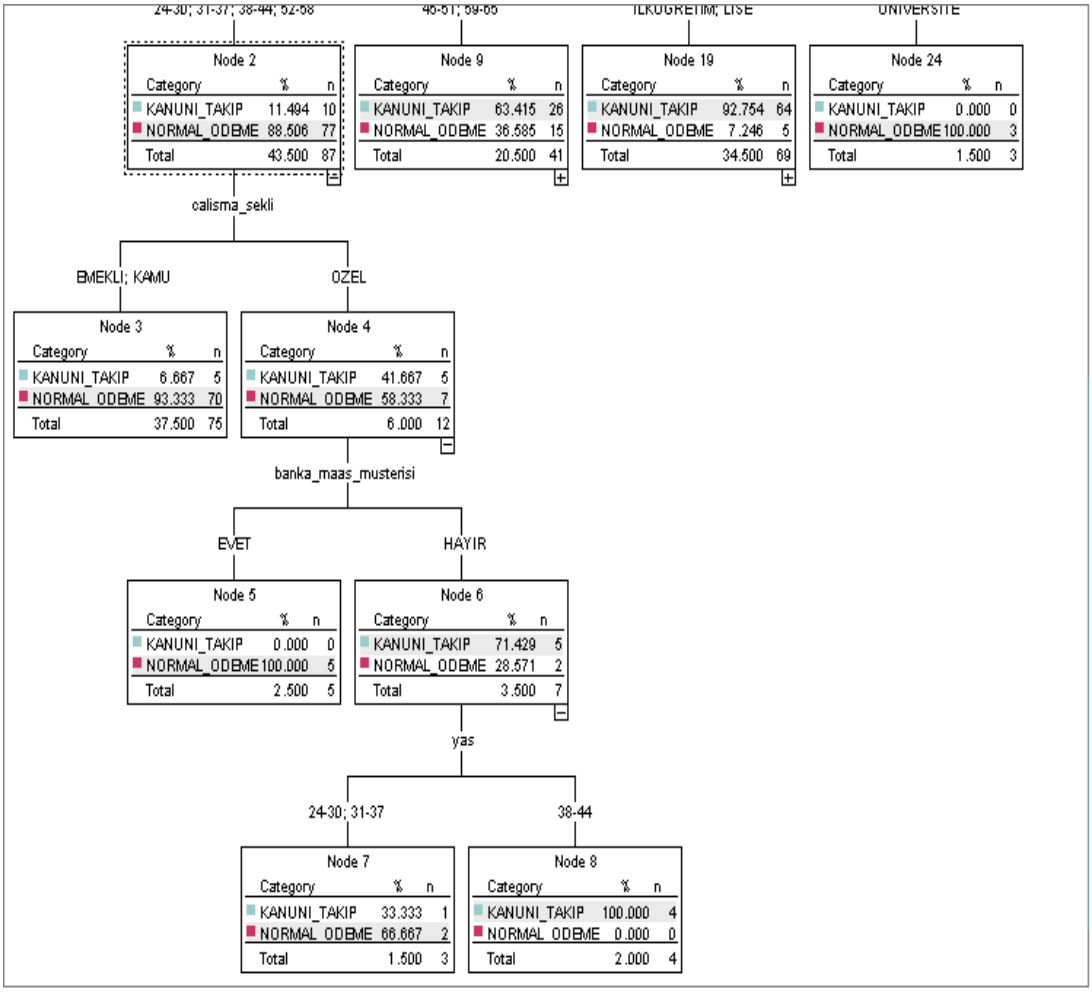
# EKLER

**Ek-A**

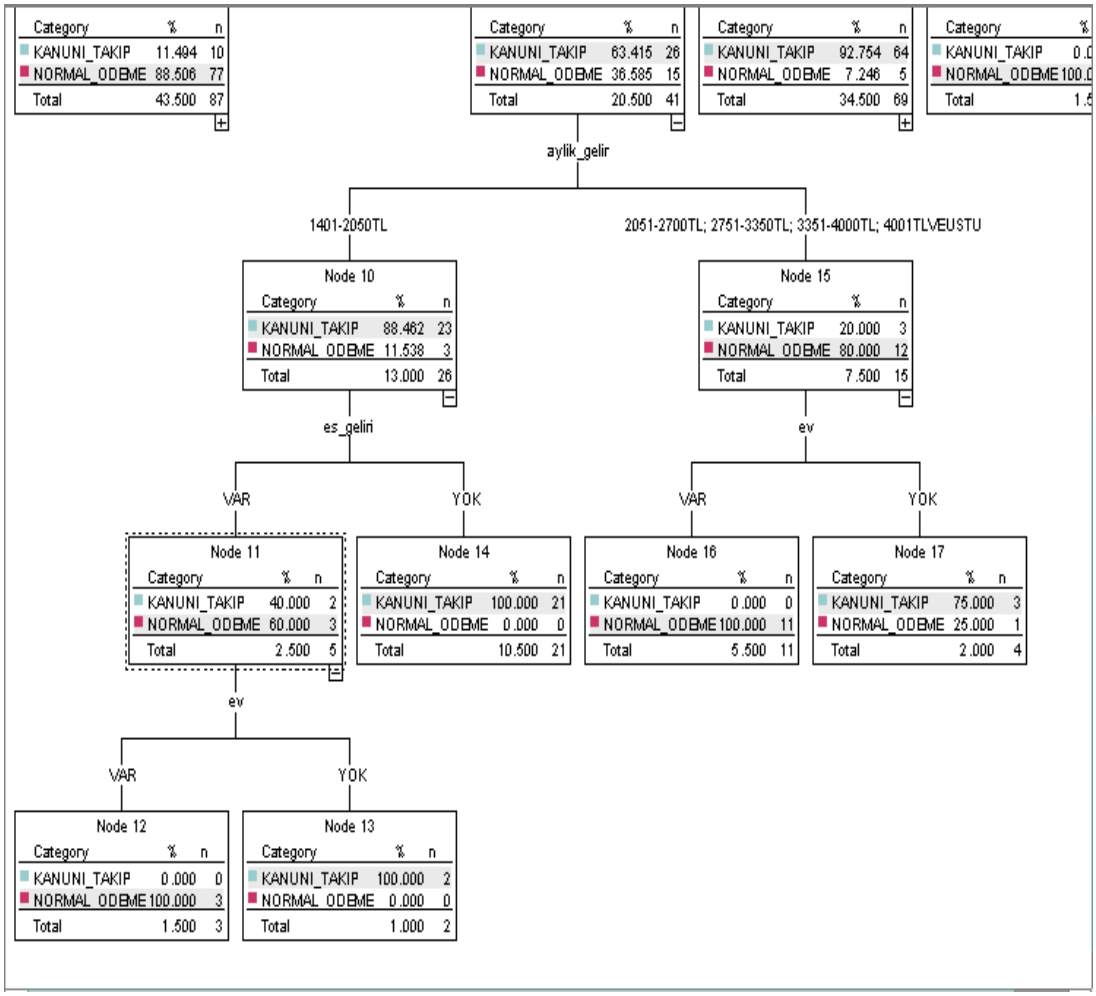
C&RT algoritmasına ilişkin karar ağacı



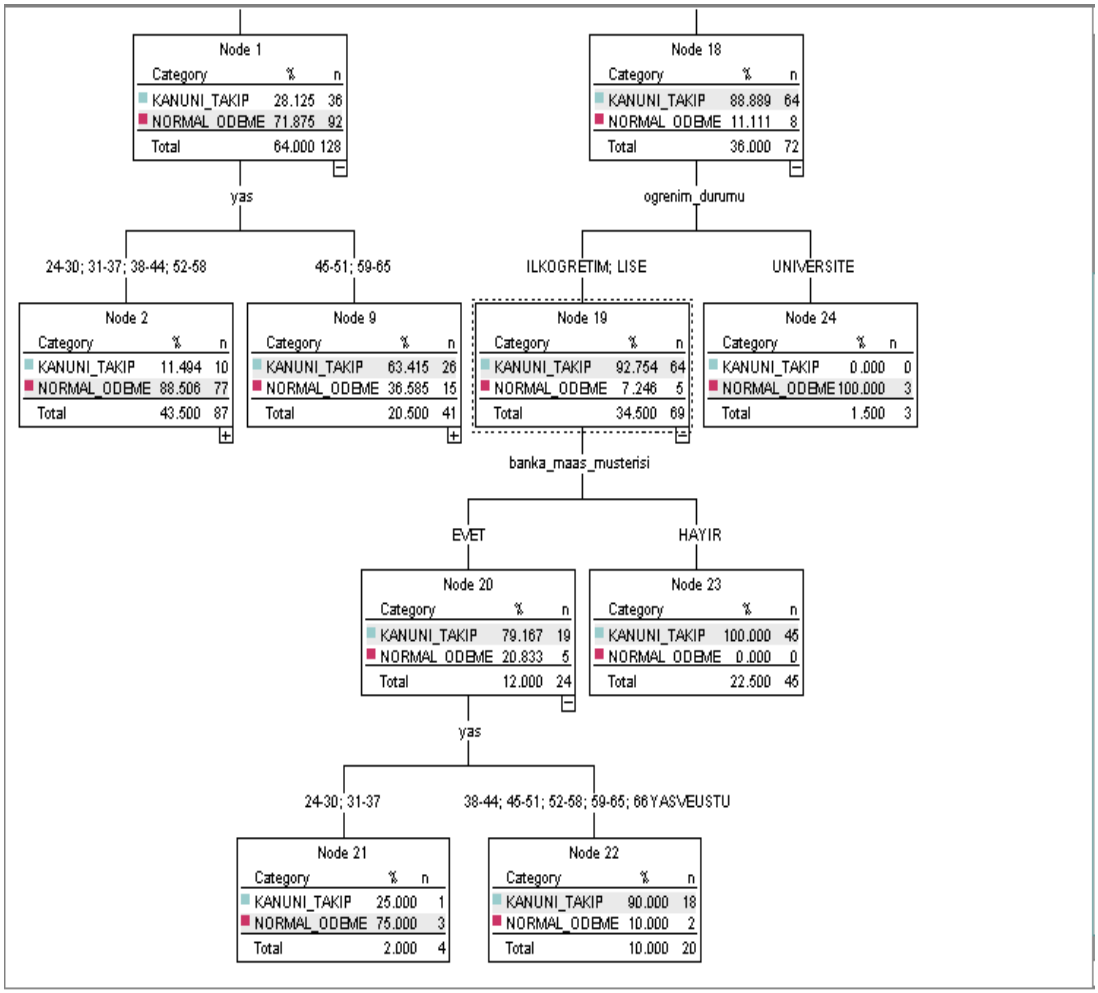

Şekil A.1. C&RT algoritması sonuçlarına ilişkin karar ağacı yapısı




Şekil A.1. (Devam) C&RT algoritması sonuçlarına ilişkin karar ağacı yapısı



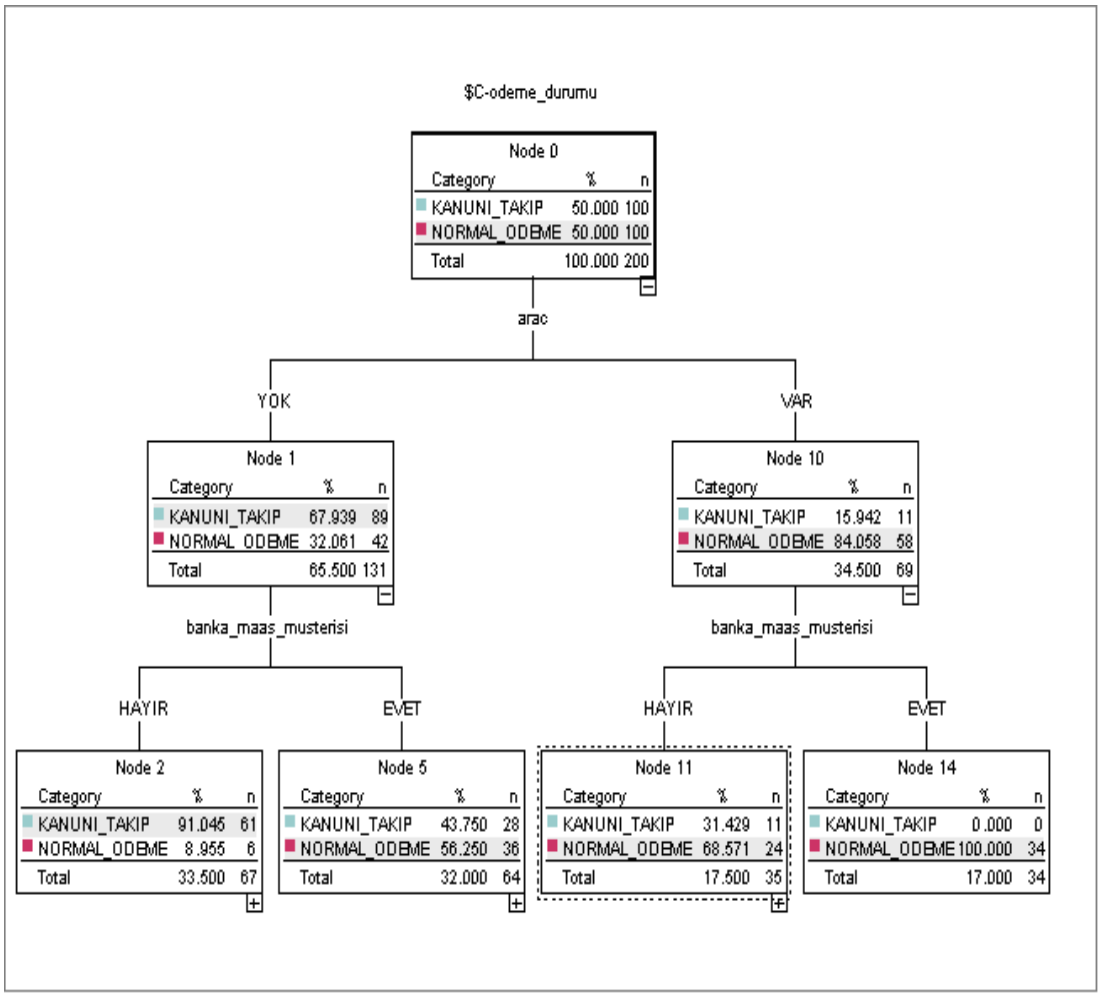

Şekil A.1. (Devam) C&RT algoritması sonuçlarına ilişkin karar ağacı yapısı



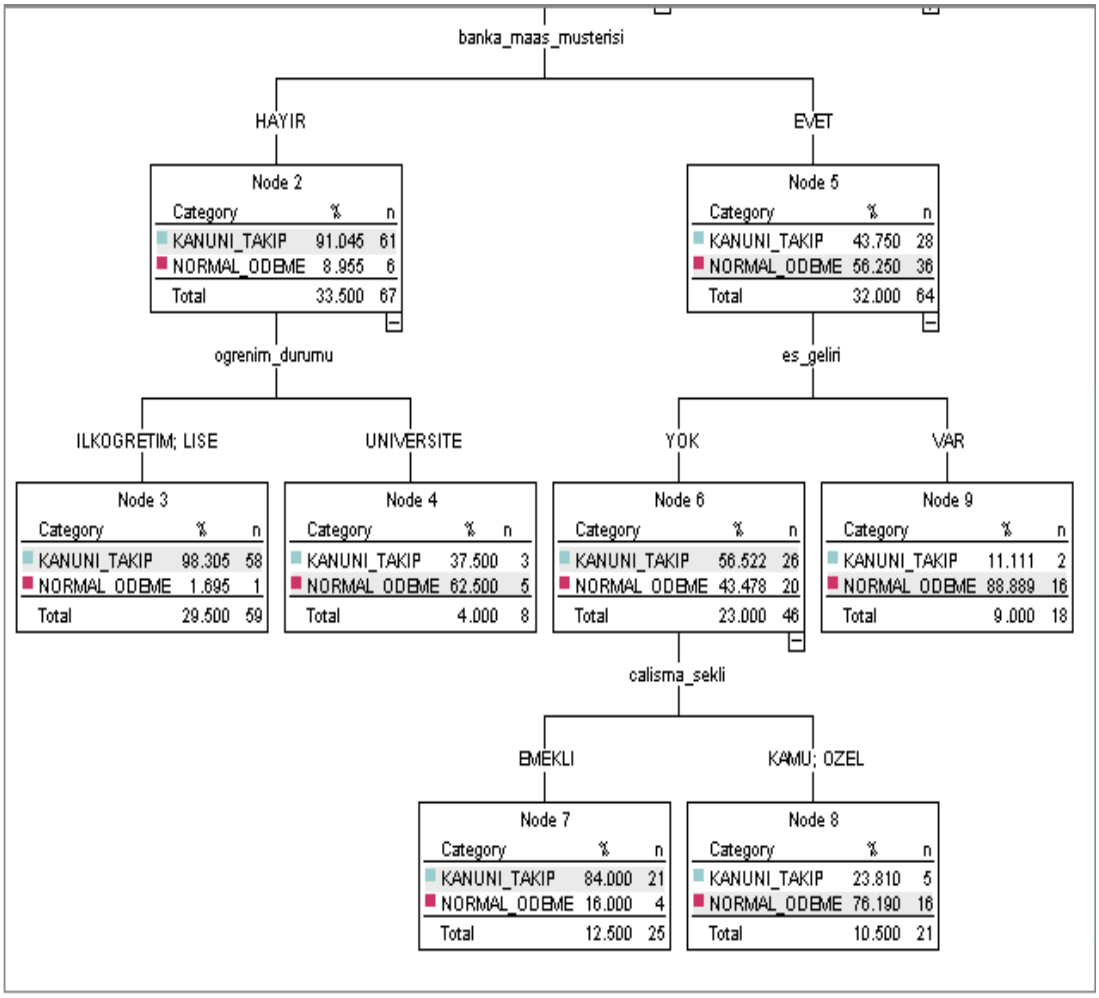

Şekil A.1. (Devam) C&RT algoritması sonuçlarına ilişkin karar ağacı yapısı

**Ek-B**

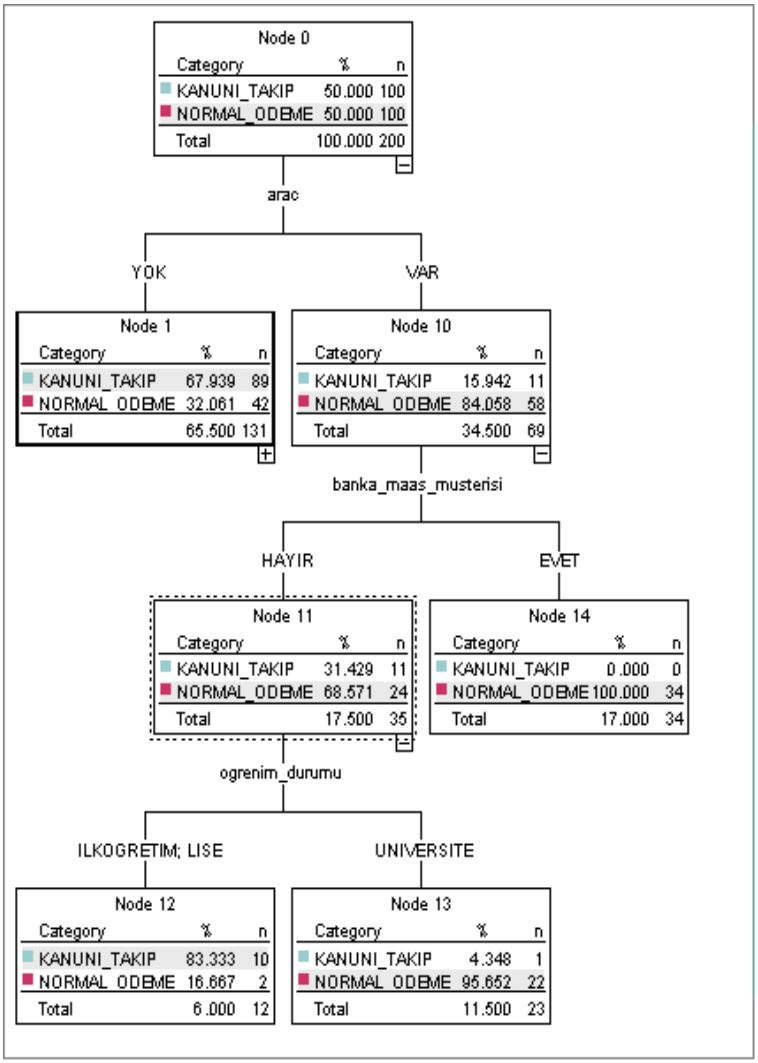
C5.0 algoritmasına ilişkin karar ağacı




Şekil B.1. C5.0 algoritması sonuçlarına ilişkin karar ağacı yapısı



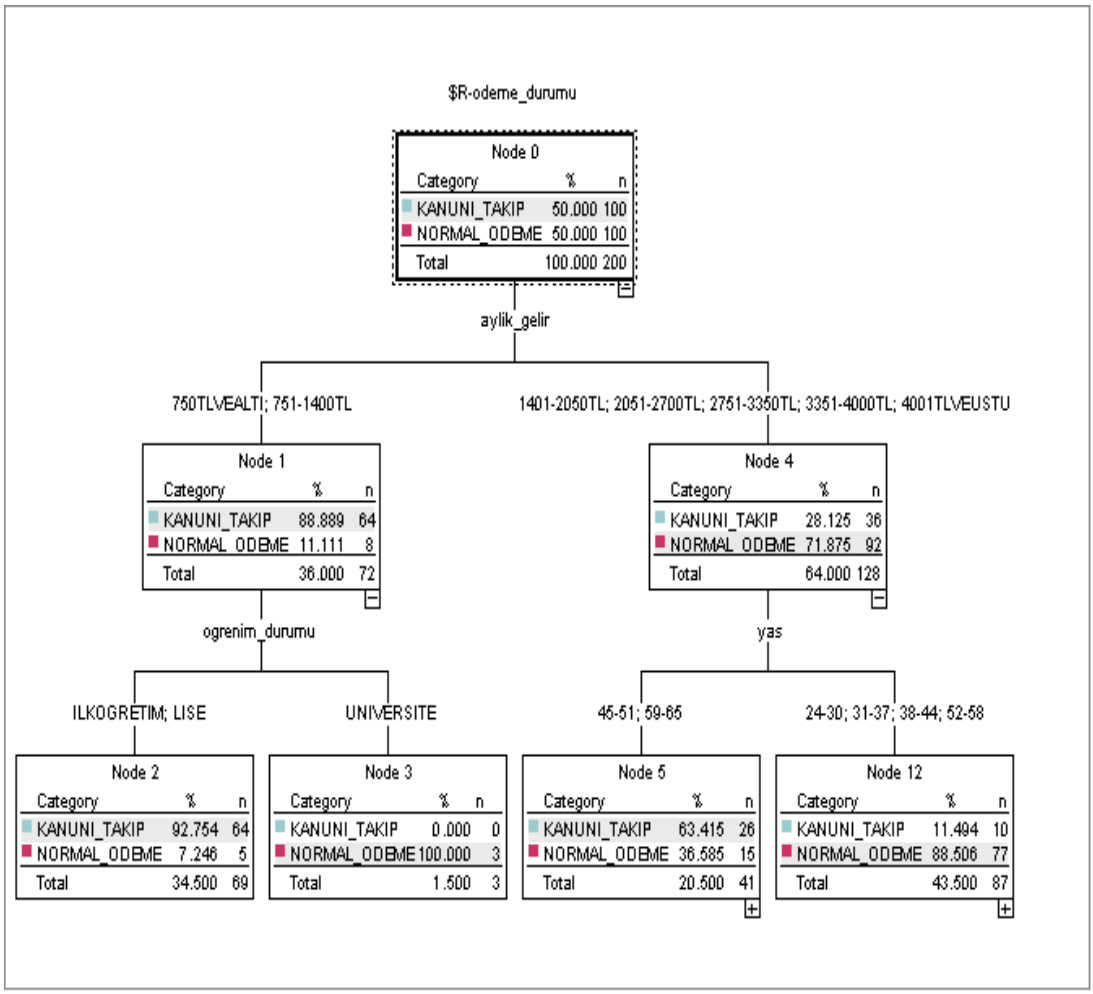

Şekil B.1. (Devam) C5.0 algoritması sonuçlarına ilişkin karar ağacı yapısı



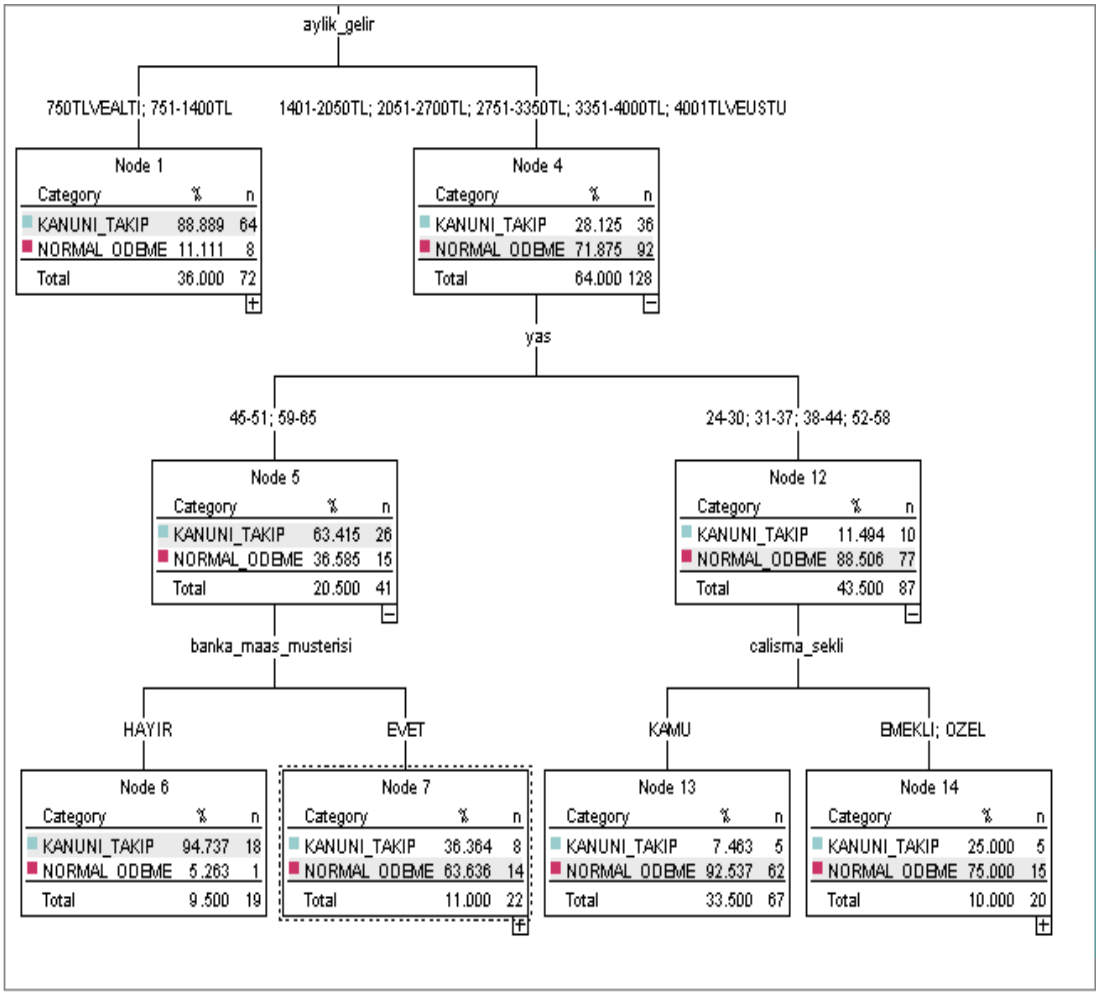

Şekil B.1. (Devam) C5.0 algoritması sonuçlarına ilişkin karar ağacı yapısı

**Ek-C**

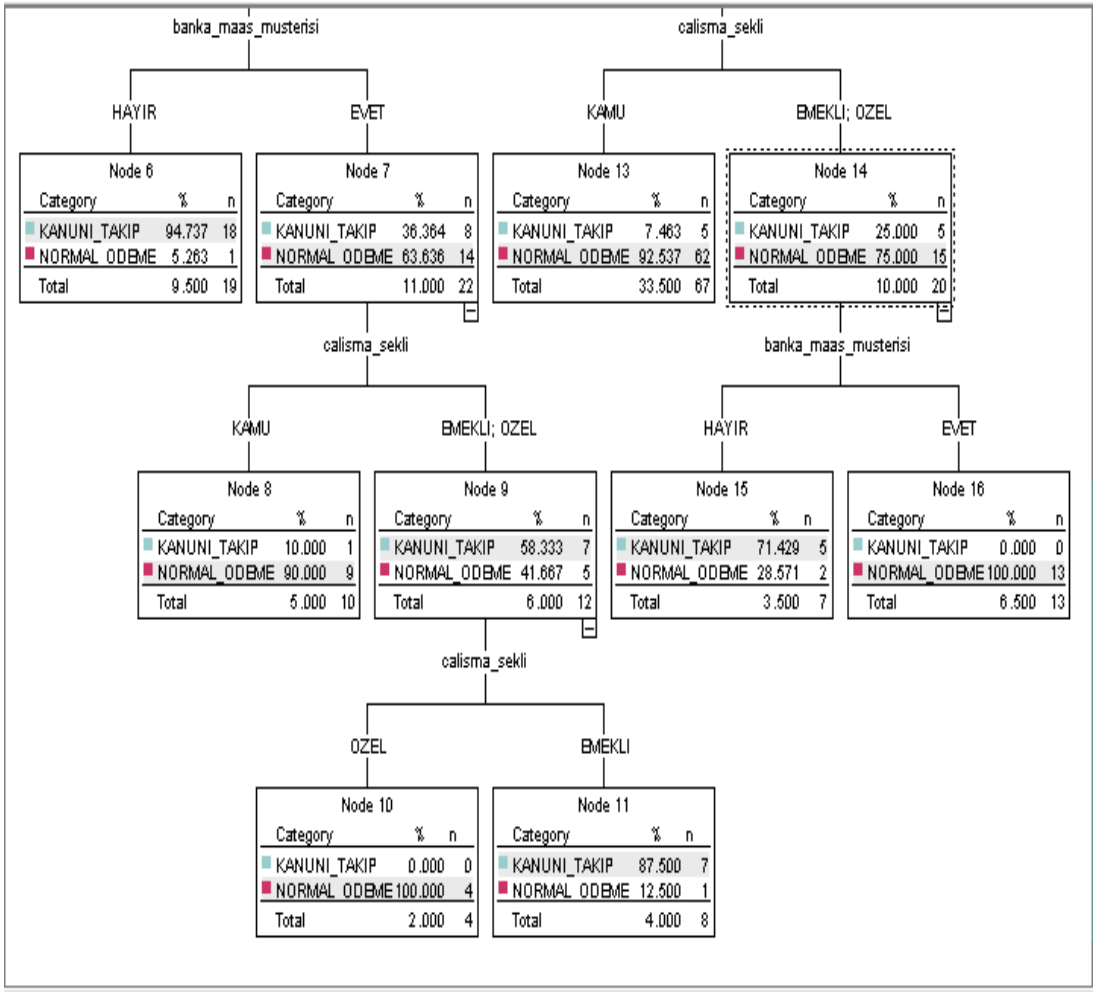
QUEST algoritmasına ilişkin karar ağacı




Şekil C.1. QUEST algoritması sonuçlarına ilişkin karar ağacı yapısı




Şekil C.1. (Devam) QUEST algoritması sonuçlarına ilişkin karar ağacı yapısı




Şekil C.1. (Devam) QUEST algoritması sonuçlarına ilişkin karar ağacı yapısı

## ÖZGEÇMİŞ

1986 yılında Ordu'da doğdu. İlk, orta ve lise öğrenimini Ordu'da tamamladı. 2004 yılında girdiği Kocaeli Üniversitesi Mühendislik Fakültesi Endüstri Mühendisliği Bölümü'nden 2008 yılında Endüstri Mühendisi olarak mezun oldu. 2010 yılında Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı'nda Yüksek Lisans eğitimine başladı. 2009-2012 yılları arasında özel bir bankanın bireysel krediler departmanında çalıştı. 2012 yılının Eylül ayından itibaren Ondokuz Mayıs Üniversitesi Mühendislik Fakültesi'nde araştırma görevlisi olarak görev yapmaktadır.