Endüstri Mühendisliği Dergisi
Cilt: 25 Sayı: 3-4 Sayfa: (2-19)

VERİ MADENCİLİĞİNDE KARAR AĞACI ALGORİTMALARI İLE BİLGİSAYAR VE İNTERNET GÜVENLİĞİ ÜZERİNE BİR UYGULAMA

Aslı ÇALIŞ, Sema KAYAPINAR, Tahsin ÇETİNYOKUŞ*

Gazi Üniversitesi, Mühendislik Fakültesi, Ankara
aslicalis@gazi.edu.tr, semakyp@gazi.edu.tr, tahsinc@gazi.edu.tr

Geliş Tarihi: 21 Mayıs 2014; Kabul Ediliş Tarihi: 10 Eylül 2014

ÖZET

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 kavramıyla ifade edilmektedir. Karar ağaçları, sınıflandırma ve tahmin için sıkça kullanılan veri madenciliği yaklaşımlarından biridir. Bu çalışmada, bilgisayar ve internet güvenliği üzerine anket düzenlenmiş olup, karar ağaçları kullanılarak farklı demografik özellikteki kişiler için çıkarım yapılması hedeflenmiştir.

Anahtar Kelimeler: Anket, karar ağaçları, veri madenciliği

AN APPLICATION ON COMPUTER AND INTERNET SECURITY WITH DECISION TREE ALGORITHMS IN DATA MINING

ABSTRACT

Developments in computer technologies caused increase in amount of information generated and volume of database systems. By discovering data kept stored in databases as ones having beneficial use potential and creation of meaningful patterns is expressed with the concept of data mining. Decision trees are one of the data mining approaches widely used for classification and forecasting. In this study, a survey was taken on computer and internet security and it was aimed to make an inference for people have different demographic characteristics by using decision trees.

Keywords: Survey, decision trees, data mining.

* İletişim yazarı

1. GİRİŞ

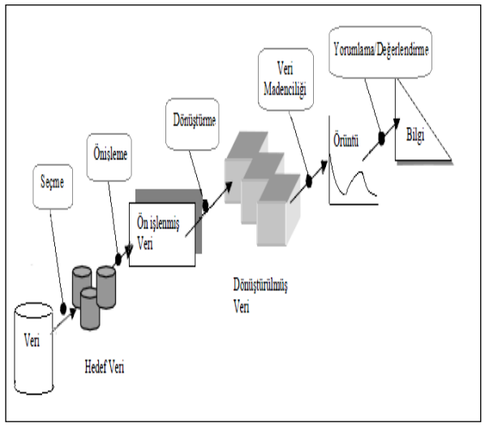
Veri madenciliğinin ortaya çıkışı veri yığınlarının geniş yer kaplamasına ve büyük miktardaki verilerin yararlı bilgilere dönüştürülmesi ihtiyacına dayanmaktadır [1].

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 imkân tanımaktadır [2].

Veri madenciliği, istatistik, 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 [3].

Veri madenciliği için yapılan farklı tanımlardan bazıları şu şekildedir:

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 [4].



Şekil 1. VM'nin Bilgi Keşfi Süreci İçindeki Yeri [7]

Veri madenciliği, büyük miktardaki veriden, anlamlı örüntüler ve kurallar keşfetme sürecidir [5].

En basit tanımıyla veri madenciliği, veri içerisindeki yeni, gizli kalmış veya beklenmeyen örüntüleri bulmak için kullanılan faaliyetler bütünüdür [6].

Veri madenciliği, daha büyük bir süreç olarak adlandırılan bilgi keşfi sürecinin bir bölümüdür [7]. Şekil 1'de VTBK (Veri Transferi Bilgi Keşfi) süreci ve bu sürecin bir parçası olan veri madenciliğine yer verilmiştir.

2. VERİ MADENCİLİĞİ SÜRECİ

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 [8]:

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

3. LİTERATÜR ARAŞTIRMASI

"Türkiye'de Yerli ve Yabancı Ticaret Bankalarının Finansal Etkinliğe Göre Sınıflandırılması" konulu çalışmada [9], 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 [10], 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 hiyerarşik 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, İkinci 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.

"Bankacılık Sektöründe Personel Seçimi ve Performans Değerlendirilmesine İlişkin Veri Madenciliği Uygulaması" [11] adlı çalışmada, bankacılık sektöründe 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.

"Bankaların Gözetiminde Bir Araç Olarak Kümeleme Analizi" konulu çalışmada [12], Türk Bankacılık Sektöründe 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 [13], 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.

"Banka Yatırım Fonu Müşteri Hareketlerinin Belirlenmesine Yönelik Bir Veri Madenciliği Uygulaması" konulu çalışmada [14], 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.

"Veri Madenciliğinde Sınıflandırma Yöntemlerinin Karşılaştırılması" konulu çalışmada [15], 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 karsılaştırılması 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. 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.

4. 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 [16].

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 özel bir 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 [17]. 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 ise tanımlayıcı modellerdir. [18].

4.1 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 [19].

Sınıflandırma, bir veri öğesini, önceden tanımlı sınıflardan birine tasnif ederken, regresyon veri öğesini gerçek değerli bir tahmini değişkene eşler [20].

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)

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 [3].

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 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. Verilen bir Ak alanının Entropi ölçüsünü bulan formüller şu şekildedir [19]:

$$E(C|A_k) = \sum_{j=1}^{M_k} p(a_k, j)x\left[-\sum_{i=1}^{N} p(c_i|a_k, j)\log_2 p(c_i|a_k, j)\right] \quad (1)$$

Bu formülde;

$E(C\backslash A_k) = A_k$ alanının sınıflama özelliğinin Entropi ölçüsü,

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

$p(c_i \backslash a_k, j) = a_k$ 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; $i = 1, 2,\ldots, N$,

$K =$ alanların sayısı; $k = 1,2,\ldots, K$.

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

$$I(S) = -(p_1\log_2(p_1) + p_2\log_2(p_2) + \ldots + p_i\log_2(p_i)) \quad (2)$$

Bu formülde p_i, C_i sınıfına ayrılma olasılığıdır.

Entropi denklemi şu şekilde de ifade edilebilir:

$$E(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} xI(S_i) \quad (3)$$

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

$$Kazanç(A) = I(S) - E(A) \quad (4)$$

Başka bir 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 ID3 algoritmasının geliştirilmiş hali olan C4.5 [25] 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. 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 [21].

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),

Tablo1. Bazı Karar Ağacı Algoritmaları ve Özellikleri [17]

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. Sürekli hedef değişkenleri ile çalışır. 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 sınıflayıcı da 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.

optimal bölünmelerin teşhisi için ki-kare istatistiğini kullanan bir yöntemdir. CHAID, bölümlendirrne amaçlı kullanılan etkili bir istatistiksel tekniktir. İstatistiksel bir 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 [22].

QUEST Algoritması: 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. Tablo 1'de bazı karar ağacı algoritmalarının özellikleri verilmektedir.

5. UYGULAMA

Uygulamada, bütünsel bir görsel modelleme gereci olan SPSS Clementine programı kullanılmıştır. Clementine, veri madenciliği çözümleri ile hem istatistik kökenli algoritmaları hem de yapay zekâ kökenli algoritmaları görsel bir programlama ara yüzü altında sunmaktadır.

Uygulanan ankete ilişkin sorular ve veri tablosu excel de düzenlenmiştir. Bağımsız değişkenler, bilgisayar ve internet güvenliği ile ilgili 10 sorudan oluşurken, bağımlı değişkenler sırasıyla yaş, cinsiyet, eğitim durumu ve internet kullanım süresi olarak alınmıştır. Ankete ilişkin bilgiler EK 1'de sunulmaktadır.

5.1 Anketin Hazırlanması ve Güvenilirliği

Anket sorularının anlaşılabilir ve kolaylıkla cevaplandırılabilir olduğunu tespit etmek ve araştırmanın amacına ne derece hizmet ettiğini görebilmek açısından, XXX üniversitesinde görevli, rassal olarak seçilen 30 akademisyen üzerinde pilot çalışma yapılmıştır. Anket, toplamda rassal olarak seçilen 300 kişiye uygulanmış, Cronbach alfa katsayısı 0,897 olarak bulunmuştur. Cronbach değeri (0,7<0,897<0,90) aralığında bulunduğu için anketin yüksek güvenilirlik düzeyine sahip olduğunu söylemek mümkündür [23].

5.2 Anket Örneklem Büyüklüğünün Belirlenmesi

Örnekleme, bir araştırmanın konusunu oluşturan evrenin bütün özelliklerini yansıtan bir parçasının seçilmesi işlemini belirtir. Örneklem, seçildiği bütünün küçük bir örneğidir. Örneklemin seçildiği grubun tümü ise evreni oluşturur. Örneklem seçilirken, örneklemin temsil yeteneği taşımasına ve yeterli büyüklükte olmasına dikkat etmek gerekir [24]. Örneklem büyüklüğü aşağıda verilen formül yardımıyla bulunur. Ana kütle sayısı bilinmediği durumlarda, p değeri 0,5 olarak belirlenir. %99 güvenilirlik seviyesi için hazırlanan örneklem büyüklüğü aşağıda gösterilmiştir.

n : Örneklem büyüklüğü
p : Ana kütlede X'in gözlenme oranı
q : X'in gözlenmeme oranı
α : Anlam düzeyi
e : Örneklem hatası

$Z_{\alpha/2}$ =2.58, α=0.01

$$N = \frac{p*(1-p)*Z_{\alpha/2}^{2}}{e^{2}} \quad (5)$$

$$N = \frac{0.5*0.5*(2.58)^{2}}{0.1^{2}} = 166.41 \quad (6)$$

Yukarıdaki sonuçtan, yani 166,41 değerinden anket büyüklüğünün yeterli olduğu görülmektedir. Anket, 300 kişiye uygulanmış, 300>166,41 olduğundan, seçilen örneklem büyüklüğü yeterli olduğu görülmüştür.

5.3 Karar Ağacı Algoritmalarının Uygulanması ve Algoritma Sonuçları

Bu aşamada, ilk olarak, yaş, cinsiyet, eğitim durumu ve internet kullanım süresi değişkenleri için SPSS Clementine'de C5.0, C&RT, CHAID ve QUEST algoritmaları uygulanmış ve her bir değişken için algoritmaların doğruluk oranları hesaplanmıştır. Ardından en yüksek doğruluk oranını veren algoritma ile karar ağaçları oluşturularak sonuçlar yorumlanmıştır.

5.3.1 Algoritmaların Doğruluk Oranları

Yaş, cinsiyet, eğitim durumu ve internet kullanımı değişkenleri için algoritmaların doğruluk oranları hesaplanmış olup, her bir değişken için bu oranların

Tablo 2. C5.0 Algoritmasına ait Modelin Doğruluk Oranı

Results for output field cinsiyet
Comparing $C-cinsiyet with cinsiyet

Correct	245	81,67%
Wrong	55	18,33%
Total	300	

Tablo 3. C&RT Algoritmasına ait Modelin Doğruluk Oranı

Results for output field cinsiyet
Comparing $R-cinsiyet with cinsiyet

Correct	230	76,67%
Wrong	70	23,33%
Total	300	

Tablo 4. CHAID Algoritmasına ait Modelin Doğruluk Oranı

Results for output field cinsiyet
Comparing $R-cinsiyet with cinsiyet

Correct	219	73%
Wrong	81	27%
Total	300	

Tablo 5. QUEST Algoritmasına ait Modelin Doğruluk Oranı

Results for output field cinsiyet
Comparing $R-cinsiyet with cinsiyet

Correct	187	62,33%
Wrong	113	37,67%
Total	300	

Coincidence Matrix for $R-cinsiyet (rows show actuals)

	b	e
b	126	24
e	89	61

sıralamasının değişmediği gözlemlenmiştir. Aşağıdaki tablolarda cinsiyet değişkeni için algoritmaların doğruluk oranları verilmiştir.

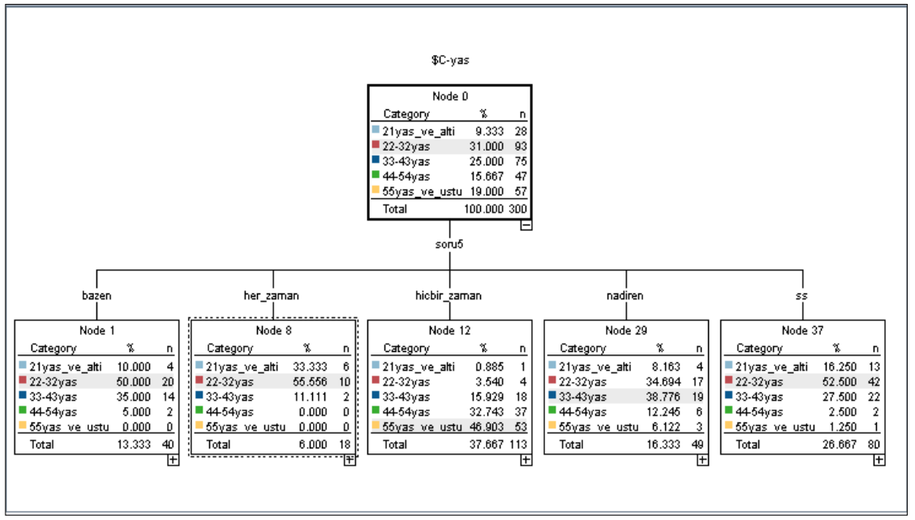
Tablolardan anlaşılacağı gibi, en yüksek doğruluk oranına C5.0 Algoritması ile kurulan modelde ulaşılmıştır. Bu nedenle, karar ağacı ile kural çıkarımı yapılırken C5.0 Algoritmasının kullanılmasına karar verilmiştir. Cinsiyet değişkeni dışındaki diğer üç değişken için de algoritmaların doğruluk oranları test edilmiş olup, sıralamanın değişmediği gözlemlenmiştir.

5.3.2. Yaş Değişkeni İçin C5.0 Algoritması ile Oluşan Karar Ağacı

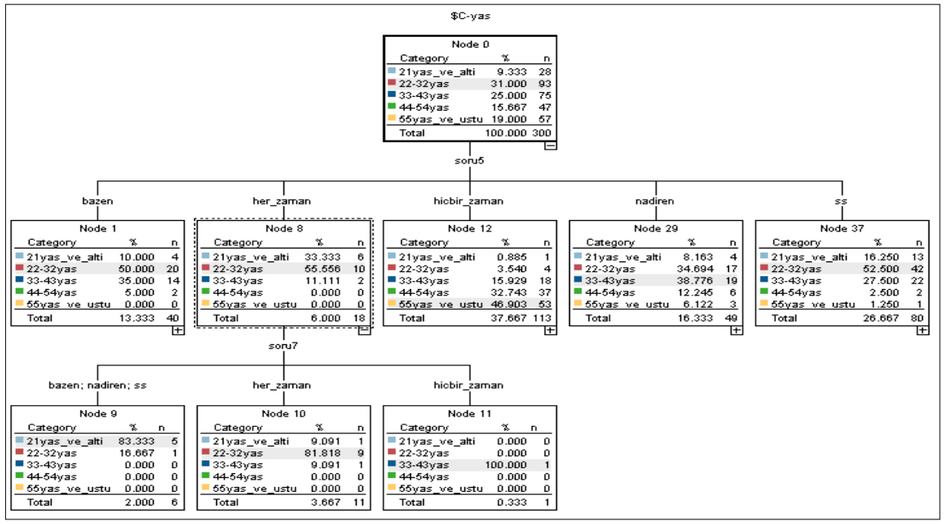
Yaş değişkeni için C5.O Algoritması ile oluşan karar ağacı incelendiğinde, ağacın ilk olarak, 5. sorudan dallandığı görülmektedir. Yani, yaş değişkeni için ağacın dallanmasındaki en etkili sorunun "Sosyal paylaşım sitelerine ne kadar sıklıkta giriyorsunuz?" olduğu görülmektedir. Anketi cevaplayan 300 kişiden 28'i 21 yaş ve altında, 93'ü 22-32 yaş, 75'i 33-43 yaş ve 47'si 44-54 yaş aralığında olup, 57'si ise 55 yaş ve üzerindedir. Buna göre, 5. soruya "bazen" ve "her zaman" cevaplarını verenlerin büyük çoğunluğunun 22-32 yaş arasında olduğu görülmektedir. 55 yaş ve üzeri kişilerin 53'ü 5. soruya "hiçbir zaman" cevabını vermiştir. Yani, 55 yaş ve üstü kişilerin genellikle sosyal paylaşım sitelerini kullanmadığı söylenebilir. "Nadiren" cevabını verenlerin çoğu 33-43 yaş aralığındadır. Bu oran %38,8'dir. 5. soruya "sık sık" cevabını verenlerin %52,5'i 22-32 yaş aralığındadır.

5. soruya "her zaman" cevabını verenler azınlıkta olup, bu kişiler için karar ağacı, dallanmaya 7. soru ile devam etmektedir. Buna göre, sosyal paylaşım

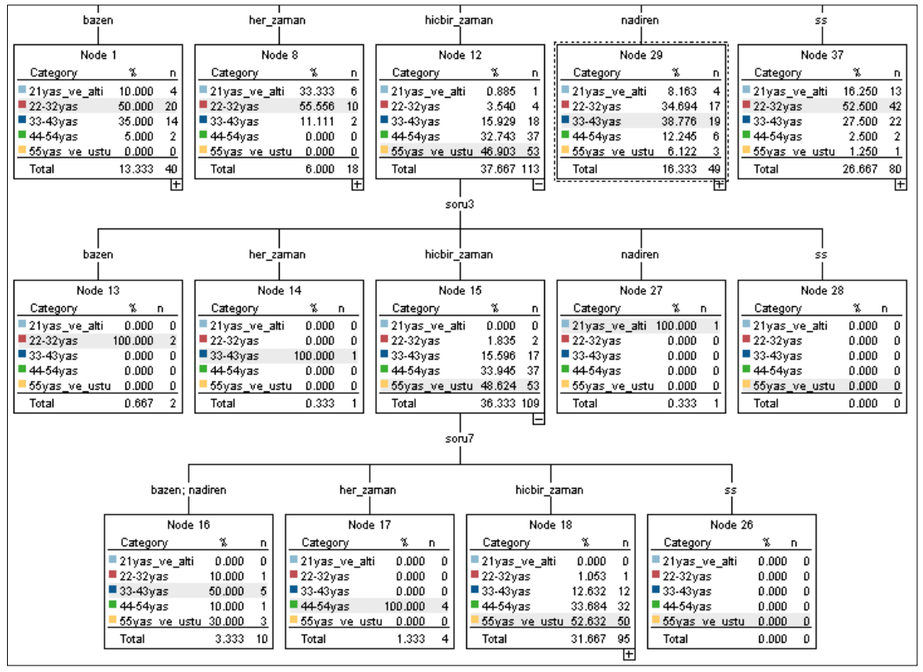
Tablo 6. Yaş Değişkeni İçin Karar Ağacında Oluşan İlk Dal



Tablo 7. 5. Soruya "her zaman" Cevabı Verenler İçin Karar Ağacı Dalları



Tablo 8. 5. Soruya "hiçbir zaman" Cevabı Verenler İçin Karar Ağacı Dalları



sitelerine "her zaman" girenlerin, virüs temizleme, casus yazılım önleme vs. programlarını da her zaman kullandığı görülmektedir. Bu kişilerin büyük çoğunluğu Tablo 7'de görüldüğü gibi 22-32 yaş aralığındadır.

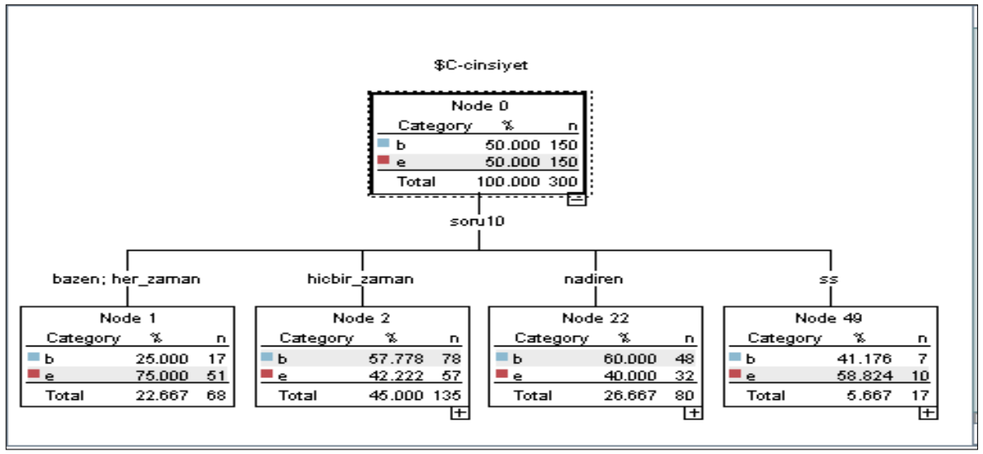
Tablo 8'de görüldüğü gibi 5. soruya "hiçbir zaman" cevabını verenler çoğunlukta olup, bu kişiler için karar ağacı 3. soru ile dallanmaya devam etmektedir. Sosyal paylaşım sitelerine girmeyen kişilerin büyük çoğunluğu internet üzerinden alışveriş de yapmamaktadır. Yine bu kişilerin büyük çoğunluğunun 7. soruya da "hiçbir zaman" cevabı verdiği görülmektedir. Yani bu kişilerin çoğu, virüs temizleme, casus yazılım önleme vs. programlarını da hiçbir zaman kullanmamaktadır. Bu çizelgede dikkat çeken diğer bir unsur da 5 ve 3. soruya "hiçbir zaman" cevabını veren kişiler için de 7. soruya "sık sık" cevabı veren olmamasıdır.

5.3.3 Cinsiyet Değişkeni İçin C5.0 Algoritması ile Oluşan Karar Ağacı

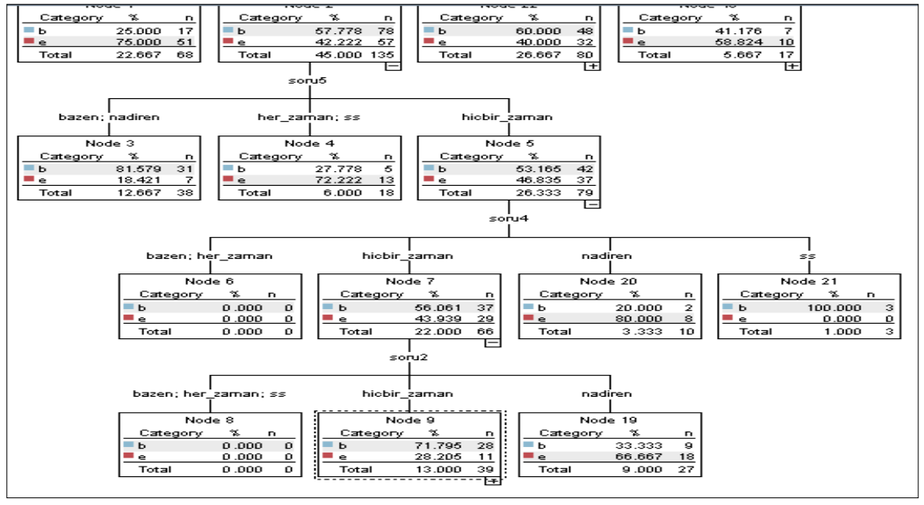
Cinsiyet değişkenine göre karar ağacı algoritması oluşturulduğunda, ağacın 10. soru ile dallanmaya başladığı görülmektedir. Tablo 9'da görüldüğü gibi, ankete katılanların 150'si kadın ve 150'si erkektir. 10. soruya "hiçbir zaman" cevabını verenlerin çoğunlukta olduğu görülmektedir. Bu oran, 135 kişiye denk gelmektedir. Yani, ankete katılanların %45'i bilgisayar ve internet güvenliği ile ilgili hukuki gelişmeleri takip etmemektedir. Bu kişiler içinde kadınların oranı %78 ile çoğunluktadır.

10. soruya "hiçbir zaman" cevabını verenler çoğunlukta olup, bu kişiler için ağaç, dallanmaya 5. soru ile devam etmektedir. Tablo 10'dan anlaşılacağı gibi, 10. soruya "hiçbir zaman" cevabını verenlerin büyük çoğunluğu 5. soruya da "hiçbir zaman" cevabını vermiştir. Yani, bilgisayar ve internet güvenliği ile ilgili hukuki gelişmeleri takip etmeyen 135 kişiden 79'u sosyal paylaşım sitelerine de girmemektedir. Bu kişilerin yaklaşık %53,2'sinin kadınlardan oluştuğu görülmektedir. Bu gruptakiler için karar ağacı oluşumundaki diğer önemli etkenler, sırasıyla 4, 2, ve 6. sorulardır. Büyük çoğunluğun 4 ve 2. soruya da "hiçbir zaman" cevabı verdiği görülmektedir. Yani bu kişiler, kullanıcı şifrelerini değiştirmedikleri gibi internete ulaşım için kişisel bilgisayarlarını da kullanmamaktadırlar. Bu cevabı verenlerin çoğunlukla kadın olduğu görülmektedir. Bu kişiler için ayırt edici diğer soru, 6. sorudur. "İnternet bankacılığını ne sıklıkta kullanırsınız?" sorusuna çoğunlukla "nadiren" cevabı verilmiştir. Bu cevabı verenlerin yaklaşık %95,7'si kadınlardan oluşmaktadır.

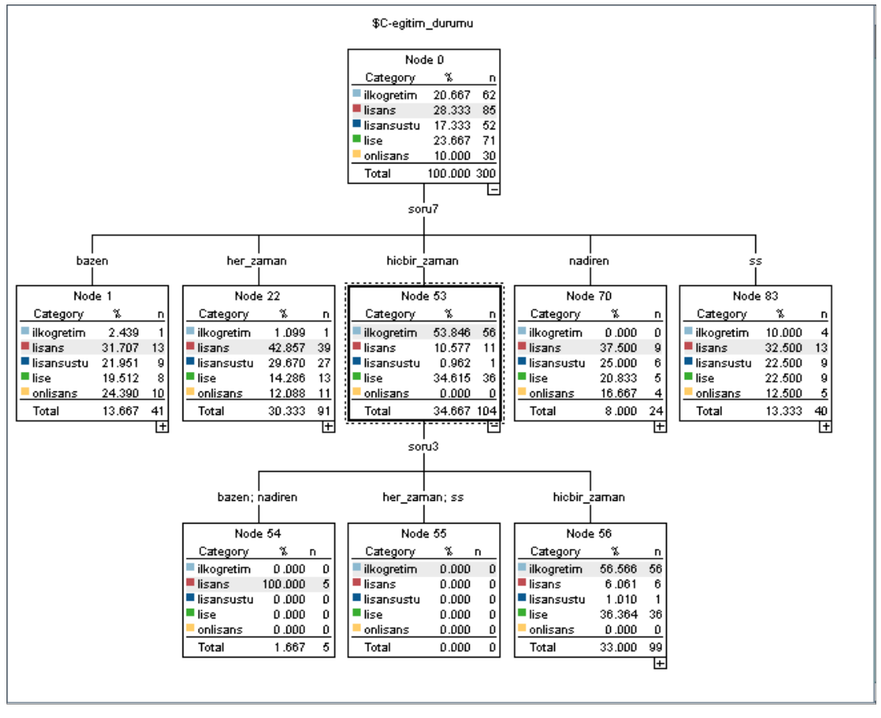
Tablo 9. Cinsiyet Değişkeni İçin Karar Ağacında Oluşan İlk Dal



Tablo 10. 10. Soruya "hiçbir zaman" Cevabı Verenler İçin Karar Ağacı Dalları



Tablo 11. Eğitim Durumu Değişkeni İçin Karar Ağacı



5.3.4 Eğitim Durumu Değişkeni İçin C5.0 Algoritması ile Oluşan Karar Ağacı

Eğitim durumu değişkeni ile karar ağacı oluşturulduğunda, en önemli unsurun 7. soru olduğu görülmektedir. Buna göre, anketin uygulandığı kişilerin %34,6'sı bu soruya "hiçbir zaman" cevabını verirken, %30,3'lük çoğunluğun da "her zaman" cevabını verdiği görülmektedir. Tablodan da anlaşılacağı gibi, virüs temizleme, casus yazılım önleme vs. programını kullanmayanların %53,8'lik oranla ilköğretim mezunu kişilerden oluştuğu görülmektedir. Bun rağmen bu tür programları "her zaman" kullanırım diyenlerin %42.8'lik çoğunlukla lisans mezunu kişilerden oluştuğu görülmektedir.

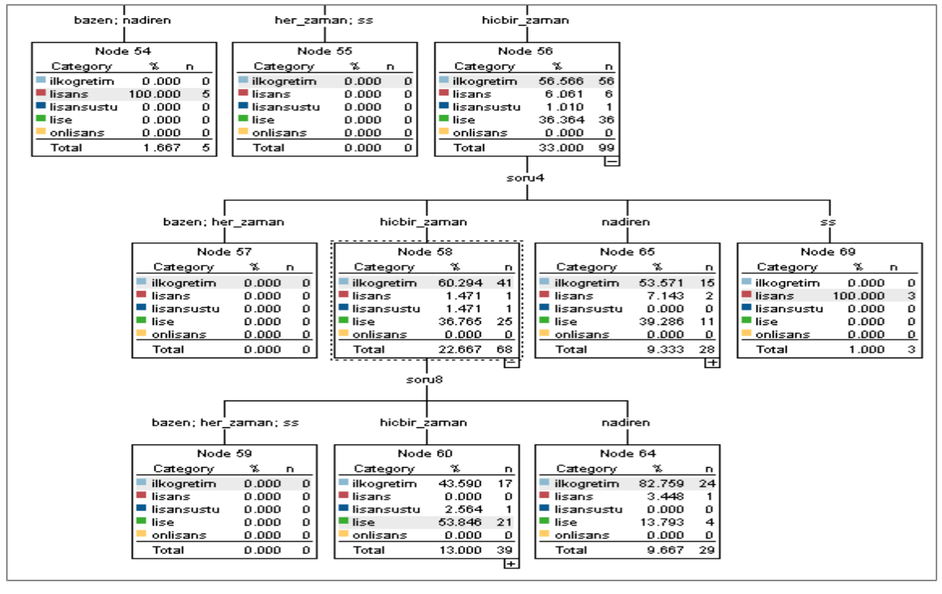
Tablo 12 ve Tablo 13'den anlaşılacağı gibi, 7. soruya "hiçbir zaman" cevabını verenlerin çoğu, sırasıyla 4, 8 ve 1. sorulara da aynı cevabı vermişlerdir. Yani bu kişiler, kullanıcı şifrelerini ve bilgisayar açılış şifrelerini hiçbir zaman değiştirmezken, internete erişim için cep telefonunu da kullanmamaktadırlar. Bu kişiler, çoğunlukla ilköğretim ve lise mezunudurlar.

5.3.5 İnternet Kullanımı Değişkeni İçin C5.0 Algoritması ile Oluşan Karar Ağacı

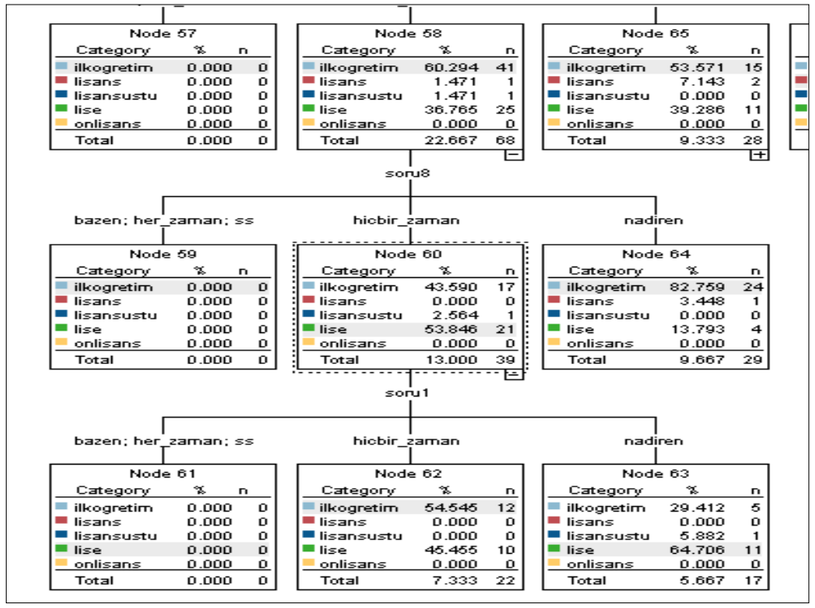
İnternet kullanımı açısından karar ağacı incelendiğinde, ilk dallanmanın 4. soru ile başladığı görülmektedir. Buna göre, "Kullanıcı şifrenizi hangi sıklıkta değiştiriyorsunuz?" sorusuna en fazla, "nadiren" cevabı verildiği görülmektedir. Bu oran toplamda %34'lük kesimi temsil etmekte olup, bu gruptakilerin %41,2'lik çoğunluğunu günde 1-3 saat arası internete girenlerin oluşturduğu görülmektedir. 4. soruya "hiçbir zaman" cevabı verenlerin de %52,6'lık oranla yine günde 1-3 saat arası internete girenlerden oluştuğu görülmektedir. Aynı soruya "her zaman" cevabı veren yalnızca 2 kişi olup, bu kişiler günde 3-5 saat arası internet kullanmaktadırlar.

Soruya nadiren cevabı veren çoğunluk için karar ağacı, dallanmaya 7. soru ile devam etmektedir. Bu soruya "her zaman" ve "hiçbir zaman" cevabı verenlerin oranları eşit olup, büyük çoğunluğu temsil etmektedir. Tablo 15'te görüldüğü gibi, virüs temizleme, casus yazılım önleme vs. programları "hiçbir zaman"

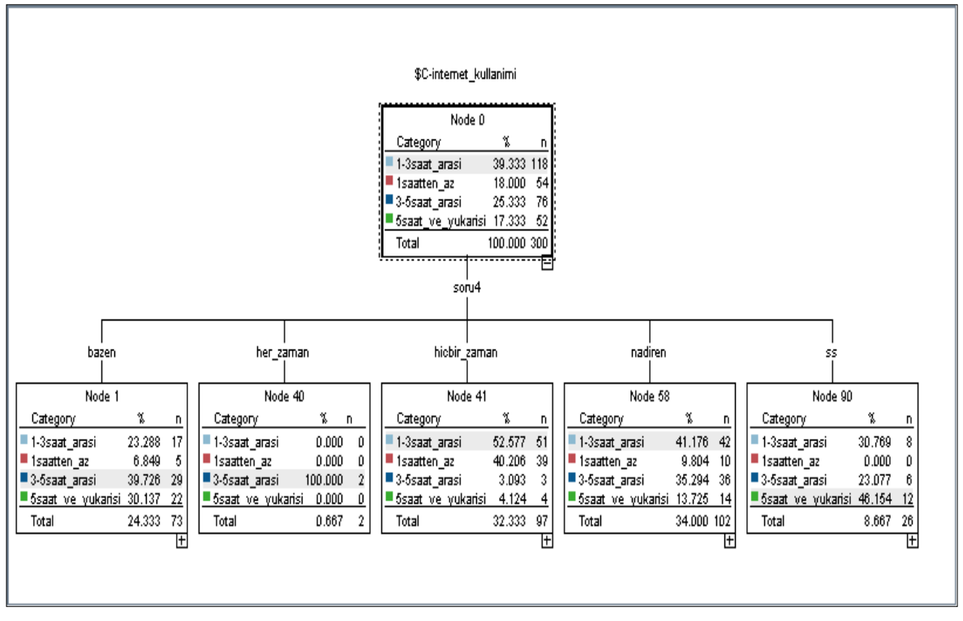
Tablo 12. Eğitim Durumu Değişkeni İçin Karar Ağacı Dalları



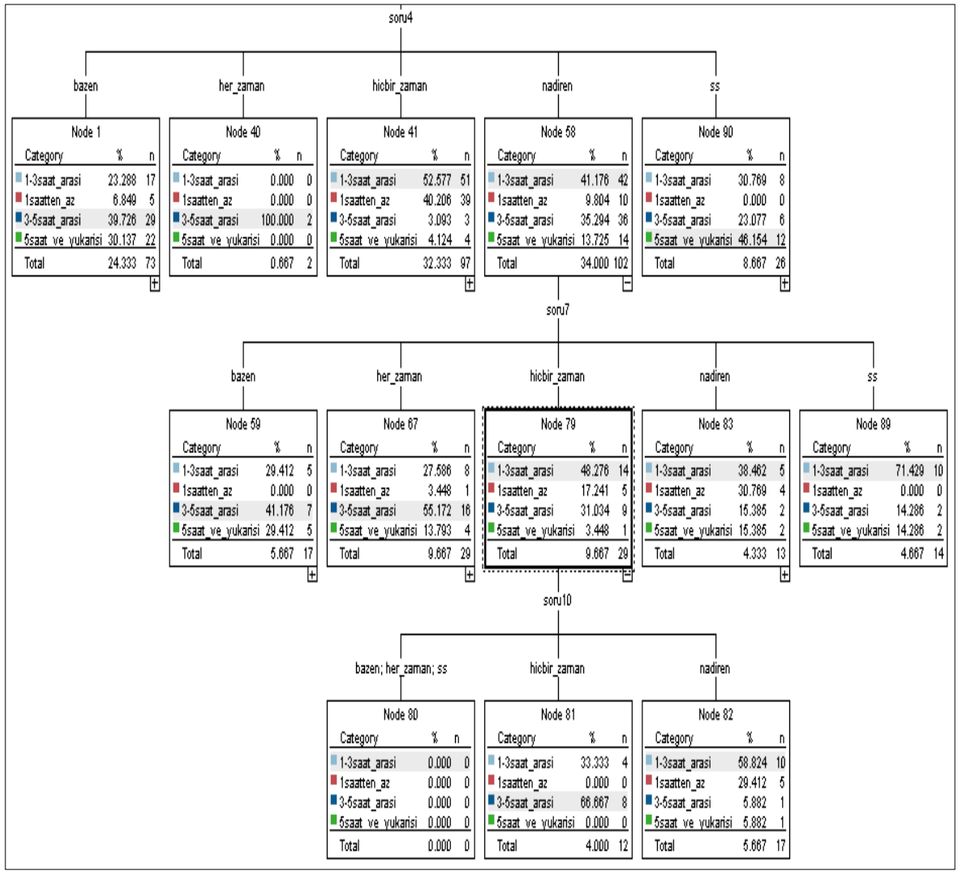
Tablo 13. Eğitim Durumu Değişkeni İçin Karar Ağacındaki Diğer Dallar



Tablo 14. İnternet Kullanımı Değişkeni İçin Karar Ağacında Oluşan İlk Dal



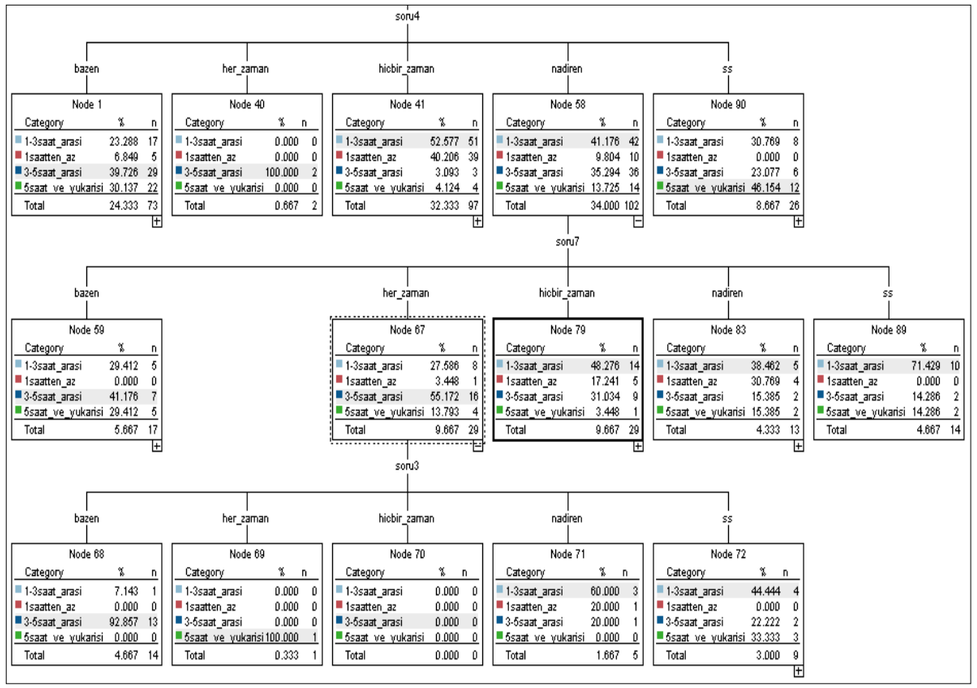
Tablo 15. 7. Soruya "hiçbir zaman" Cevabı Verenler İçin Karar Ağacı Dalları



kullanmayanlar, %48,3'lük çoğunlukla, günde 1-3 saat arası internete girenlerden oluşmakta olup, bu kişileri ayırt edici diğer bir unsur 10. soru olmaktadır. Bu soruya verilen cevaplar, sırasıyla "nadiren" ve "hiçbir zaman"dır.

Tablo 16'da 7. soruya "her zaman" cevabı veren diğer grup için karar ağacının 3. soru ile dallanmaya devam ettiği görülmektedir. Bu grupta 3. soru olan "İnternet üzerinden hangi sıklıkta alışveriş yapıyorsunuz?" sorusuna, en fazla, "bazen" cevabı verilmiş olup, bu cevabı verenlerin %92,9'luk kısmını günde 3-5 saat arası internete girenler oluşturmaktadır. Yani, kullanıcı şifrelerini nadiren değiştirip, virüs temizleme, casus yazılım önleme gibi programları her zaman kullananların çoğu, internet üzerinden bazen alışveriş yapmakta olup, bu kişiler, çoğunlukla, günde 3-5 saat arası internet kullanmaktadırlar.

Tablo 16. 7. Soruya "her zaman" Cevabı Verenler için Karar Ağacı Dalları



6. SONUÇ 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 birçok alanda karar verme sürecine duyulan ihtiyaçtan ötürü veri madenciliği yaygın olarak kullanılmaktadır.

Çalışma kapsamında, ilk olarak, bilgisayar ve internet güvenliği ile ilgili 10 soruluk anket, farklı demografik özelliklere sahip 300 kişiye uygulanmış olup, verilen cevaplar excelde düzenlenerek anket güvenilirliği ölçülmüştür. Ardından SPSS Clementine programında veri madenciliğinde sınıflandırma yöntemlerinden biri olan karar ağaçları kullanılarak kişilerin demografik özelliklerine göre sorulara verilen cevapların doğruluk oranları, 4 farklı karar ağacında test edilmiş ve C5.0 algoritmasının doğruluk oranının her bir bağımlı değişken için diğer algoritmalardan daha yüksek değere sahip olduğu görülmüştür. Bu amaçla, çalışmaya C5.0 algoritması ile devam edilmiştir.

Yaş değişkeni için C5.O Algoritması ile oluşan karar ağacı incelendiğinde, ağacın ilk olarak, 5. sorudan dallandığı görülmüştür. Yani, yaş değişkeni için ağacın dallanmasındaki en etkili soru, "Sosyal paylaşım sitelerine ne kadar sıklıkta giriyorsunuz?" dur. Soruya "bazen" ve "her zaman" cevaplarını verenlerin büyük çoğunluğunun 22-32 yaş arasında olduğu gözlemlenirken, 55 yaş ve üzeri kişilerin 53'ünün 5. soruya "hiçbir zaman" cevabını verdiği görülmüştür. Yani 55 yaş ve üstü kişilerin, genellikle, sosyal paylaşım sitelerini kullanmadığı söylenebilir. Ayrıca, sosyal paylaşım sitelerine her zaman girenlerin, virüs temizleme, casus yazılım önleme vs. programlarını da her zaman kullandığı, sosyal paylaşım sitelerine girmeyen kişilerin çoğunlukla internet üzerinden alışveriş yapmadığı ve bu kişilerin büyük çoğunluğunun virüs temizleme, casus yazılım önleme vs. programlarını da hiçbir zaman kullanmadığı sonucuna ulaşılmıştır. Cinsiyet değişkenine göre karar ağacı algoritması oluşturulduğunda, ağacın 10. soru ile dallanmaya başladığı görülmüştür. Ankete katılanların %45'i bilgisayar ve internet güvenliği ile ilgili hukuki gelişmeleri takip etmemektedir. Bu kişiler içinde kadınların oranı %78 ile çoğunluktadır.

Eğitim durumu değişkeni ile karar ağacı oluşturulduğunda, en önemli unsurun 7. Soru olduğu görülmüştür. Buna göre, anketin uygulandığı kişilerin %34,6'sı bu soruya "hiçbir zaman" cevabını verirken, %30,3'lük çoğunluğun da "her zaman" cevabını verdiği görülmüştür. Virüs temizleme, casus yazılım önleme vs. programını kullanmayanların %53,8'lik oranla ilköğretim mezunu kişilerden oluştuğu görülmektedir. Bun rağmen bu tür programları "her zaman" kullanırım diyenlerin %42,8'lik çoğunlukla, lisans mezunu kişilerden oluştuğu sonucuna ulaşılmıştır.

İnternet kullanımı açısından karar ağacı incelendiğinde ise ilk dallanmanın 4. soru ile başladığı görülmüştür. Buna göre, "Kullanıcı şifrenizi hangi sıklıkta değiştiriyorsunuz?" sorusuna, en fazla, "nadiren" cevabı verilmiştir. Bu oran toplamda %34'lük kesimi temsil etmekte olup, bu gruptakilerin %41,2'lik çoğunluğu, günde 1-3 saat arası internete giren kişilerden oluşmaktadır.

Yaş, cinsiyet, eğitim durumu ve günlük internet kullanımı olarak ifade ettiğimiz bağımlı değişkenler için karar ağaçları, sırasıyla 5, 10, 7 ve 4. sorudan dallanmaya başlamıştır. Yani yaş değişkeni açısından karar ağacı oluşumundaki en önemli soru, "Sosyal paylaşım sitelerine ne kadar sıklıkta giriyorsunuz?" iken; cinsiyet değişkeni için, "Bilgisayar ve internet güvenliğiyle ilgili hukuki gelişmeleri takip eder misiniz?"; eğitim durumu için, "Virüs temizleme, casus yazılım önleme vs. programlarını kullanır mısınız?"; internet kullanımı için ise "Kullanıcı şifrelerinizi hangi sıklıkta değiştiriyorsunuz?" sorusudur. Tüm bunlara rağmen dikkat çeken diğer bir husus da karar ağaçlarının oluşmasında 6 ve 9. soruların çoğunluk açısından ayırt edici bir özellik taşımıyor olmasıdır. Yani, "İnternet bankacılığını ne sıklıkta kullanıyorsunuz?" ve "Tanımadığınız kişilerden gelen e- postaları açar mısınız?" soruları ağaç oluşumundaki kritik sorulardan değildir. Bu sorular yerine farklı sorular kullanılarak anket revize edilebilir.

KAYNAKÇA

1. **Han, J., Kamber, M.** 2006. Data Mining: Concepts and Techniques, Morgan Kaufmann, USA.
2. **Ching, W. K., Michael, K. P.** 2002. Advances in Data Mining and Modeling, World Scientific, Hong Kong.
3. **Chien, C. F., Chen, L. F.** 2008. "Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry," Expert Systems with Applications, vol. 34, p. 280-290.
4. **Larose, D. T.** 2005. Discovering Knowledge in Data: An İntroduction in Data Mining, Wiley, USA.
5. **Linoff, G. S.. Berry, M. J. A.** 2011. Data Mining Techniques for Marketing, Sales and Customer Relationship Management, Wiley, Canada.
6. **Marakas, G. M.** 2003. Decision Support Systems in The 21st Century, Prentice Hall, USA.
7. **Hudairy, H.** 2004. "Data Mining and Decision Making Support in The Governmental Sector," Master Thesis, Faculty of Graduate School of The University of Louisville, Kentucky.
8. **Savaş, S., Topaloğlu N., Yılmaz, M.** 2012. "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri," İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, sayı 21, s. 1-23.
9. **Albayrak, A. S.** 2009. "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, sayı 14, s. 113-139.

10. **Aşan, Z.** 2007. "Kredi Kartı Kullanan Müşterilerin Sosyo-Ekonomik Özelliklerinin Kümeleme Analizi ile İncelenmesi," Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, cilt 17, s. 256-267.
11. **Bilen, H.** 2009. "Bankacılık Sektöründe Personel Seçimi ve Performans Değerlendirilmesine İlişkin Veri Madenciliği Uygulaması," Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
12. **Doğan, B.** 2008. "Bankaların Gözetiminde Bir Araç Olarak Kümeleme Analizi: Türk Bankacılık Sektörü İçin Bir Uygulama," Doktora Tezi, Kadir Has Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
13. **Tosun, T.** 2006. "Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi," Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.
14. **Çil, F.** 2010. "Banka Yatırım Fonu Müşteri Hareketlerinin Belirlenmesine Yönelik Bir Veri Madenciliği Uygulaması," Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.
15. **Çakır, Ö.** 2008. "Veri Madenciliğinde Sınıflandırma Yöntemlerinin Karşılaştırılması: Bankacılık Müşteri Veri Tabanı Üzerinde Bir Uygulama," Doktora Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.
16. **Akpınar, H.** 2000. "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği," İstanbul Üniversitesi İşletme Fakültesi Dergisi, cilt 29, s. 1-22.
17. **Emel, G. G., Taşkın, Ç.** 2005. "Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması," Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, cilt 6, s. 221-239.
18. **Albayrak, A. S., Yılmaz, Ş. K.** 2009. "Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama," Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, cilt 14, s. 31-52.
19. **Özekeş, S., Çamurcu, A. Y.** 2002. "Veri Madenciliğinde Sınıflama Ve Kestirim Uygulaması," Marmara Üniversitesi Fen Bilimleri Dergisi, sayı 18, s. 1-17.
20. **Fayyad, U., Shapiro, G., Smyth, P.** 1996. "From Data Mining to Knowledge Discovery in Databases," American Associatin for Artificial Intelligence, cilt 17, s. 37-54.
21. **Sezer, E. A., Bozkır, A. S., Yağız, S., Gökçeoğlu C.** 2010. "Karar Ağacı Derinliğinin CART Algoritmasında Kestirim Kapasitesine Etkisi: Bir Tünel Açma Makinesinin İlerleme Hızı Üzerinde Uygulama," Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, Kayseri.
22. **Oğuzlar, A.** 2004. "CART Analizi İle Hane Halkı İşgücü Anketi Sonuçlarının Özetlenmesi," Atatürk Üniversitesi İİBF Dergisi, sayı 18, s. 79-90.
23. **Özdamar, K.** 2013. Paket Programları ile İstatistiksel Veri Analizi, Nisan Kitapevi, Ankara, s. 554-555.
24. **Gökçe, B.** 1988. Toplumsal Bilimlerde Araştırma, Savaş Yayınları, Ankara.
25. **Quinlan, J. R.** 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, USA.

Ek 1. Bilgisayar ve İnternet Güvenliği Anketi

1-Yaşınız:

2- Cinsiyetiniz:

3- Eğitim Durumunuz:

☐ İlköğretim ☐ Ön lisans

☐ Ortaokul ☐ Lisans

☐ Lise ☐ Lisansüstü

4- Ortalama günlük internet kullanım süreniz:

☐ 1 saatten az ☐ 1 ile 3 saat arası

☐ 3 ila 5 saat arası ☐ 5 saat ve yukarısı

Aşağıdaki sorulara ilişkin görüşünüzü en iyi yansıtan cevabı (X) içerisine alarak işaretleyiniz. Bu konuda göstermiş olduğunuz ilgiden dolayı şimdiden teşekkür ederiz.

Soru No	Sorular	Dereceler				
		Hiçbir zaman	Nadiren	Bazen	Sık sık	Her zaman
5	İnternete erişim için cep telefonunu ne sıklıkta kullanıyorsunuz?					
6	İnternete erişim için kişisel bilgisayarınızı ne sıklıkta kullanıyorsunuz?					
7	İnternet üzerinden hangi sıklıkta alışveriş yapıyorsunuz					
8	Kullanıcı şifrelerinizi hangi sıklıkta değiştiriyorsunuz?					
9	Sosyal paylaşım sitelerine ne kadar sıklıkta giriyorsunuz?					
10	İnternet bankacılığını ne sıklıkta kullanıyorsunuz					
11	Virüs temizleme, casus yazılım önleme vs. programlarını kullanır mısınız?					
12	Bilgisayara açılış şifresini ne sıklıkta değiştirirsiniz?					
13	Tanımadığınız kişilerden gelen e- postaları açar mısınız?					
14	Bilgisayar ve internet güvenliğiyle ilgili hukuki gelişmeleri takip eder misiniz?					

Varsa görüş ve önerileriniz:

.....

.....