TÜRK BİLİM ARAŞTIRMA VAKFI
TÜBAV
1995
TURKISH SCIENCE RESEARCH FOUNDATION

Yıl: 2019, Cilt:12, Sayı:2, Sayfa: 1-17

HANE HALKI İNTERNET HİZMETİ SAHİPLİĞİNİ ETKİLEYEN FAKTÖRLERİN KARAR AĞAÇLARI İLE İNCELENMESİ

Mahmut COŞKUN[1], Halil İbrahim BÜLBÜL[2]

[1]*Türk Telekom, mahmutcoskun84@gmail.com*

[2]*Gazi Üniversitesi, Gazi Eğitim Fakültesi, bhalil@gazi.edu.tr*

Özet

Yapılan işlemler bilgisayarların insan hayatına girmesi ile daha kolay kayıt altına alınmaya başlanmıştır. Kayıt altına alınan bu işlemlerin verileri içinde, keşfedilmeyi bekleyen birbiri ile ilintili bilgiye dönüştürülebilecek bağlantılar bulunmaktadır. Bu gizli bağlantıların keşfedilmesi için veri madenciliği tekniklerinden yararlanılır. Bu makalede, Türkiye İstatistik Kurumu'nun 2016 yılı Hanehalkı Bilişim Teknolojileri Kullanım Anketi verileri kullanılmıştır ve hanehalkının internet hizmetine sahip olma durumu incelenmiştir. Hanehalkı internet kullanımını etkileyen karakteristiklere ilişkin faktörler, karar ağaçları kullanılarak analiz edilmiştir. CHAID, C5.0, C&RT ve QUEST karar ağacı algoritmaları karşılaştırılmıştır ve en başarılı algoritma C5.0 olarak belirlenmiştir.C5.0 algoritması ile analiz yapılmış ve 10 dallanma, 21 düğümden oluşan bir karar ağacı elde edilmiştir. Analiz sonucunda hanelerin internet hizmeti sahipliğini etkileyen en önemli değişkenlerin hane cep telefonu sahipliği, hane bilgisayar kullanımı, hanede 0-25 yaş arasında bireyin olup olmaması, hane tablet sahipliği, hane dizüstü bilgisayar sahipliği, hanehalkı büyüklüğü, hanehalkı reisinin yaşı, hane smarttv sahipliği, hane gelir grubu olduğu görülmüştür.

Anahtar Kelimeler: Veri madenciliği, karar ağacı, C5.0 algoritması, internet hizmeti

INVESTIGATION OF THE FACTORS AFFECTING THE HOUSEHOLD INTERNET SERVICE OWNERSHIP WITH DECISION TREES

Abstract

Processes have been started to be recorded more easily with the introduction of computers into human life. In the records of these transactions recorded, there are links that can be transformed into related information waiting to be explored. Data mining techniques are exploited to discover these hidden links. In this article, Turkey Statistics Institute's 2016 Household Information Technology Usage Survey data is used and the state of households having internet service is examined. Factors affecting the characteristics of household internet use are analyzed by using decision trees. CHAID, C5.0, C&RT and QUEST decision tree algorithms are compared and the most successful algorithm is determined as C5.0. The C5.0 algorithm was used for the analysis and a 10-branch, 21-node decision tree was obtained. As a result of the analyzes, the most important variables affecting the ownership of the internet service of the households are: household cell phone ownership, household computer use, whether the household is between 0-25 years old, household ownership, household laptop ownership, household size, age of household head, household smarttv ownership, household income group.

Key Words: Data mining, decision tree, C5.0 algorithm, internet service

1. GİRİŞ

Veri tabanları günümüzde terabaytlarla ifade edilmektedir. Bu büyük hacimde verinin içinde stratejik önem taşıyan gizli enformasyon yatmaktadır. Ancak bu kadar büyük hacimli veri içerisinde yer alan önemli bilgi ya da bilgilerin nasıl açığa çıkarılacağı en önemli sorudur. Bu önemli soruya en güncel yanıt, hem geliri artıran hem de maliyetleri indirgeyen veri madenciliği alanıdır (Koyuncugil & Özgülbaş, 2009).

Bilişim teknolojileri ile ilgili istatistikler, bilgi toplumunda son yıllarda meydana gelen sosyal, kültürel ve ekonomik gelişmeleri anlamak, bu konuda uygulanan politikaları takip etmek ve piyasaların etkin çalışmasını sağlamak gibi nedenlerle büyük önem kazanmıştır. Bu istatistiklerin üretilmesi amacıyla ülkemizde de Türkiye İstatistik Kurumu (TÜİK) tarafından 2004 yılından bu yana Hanehalkı Bilişim Teknolojileri Kullanım Anketi (HBTKA) gerçekleştirilmektedir. Bu araştırma ile hanelerde bulunan bilgi ve iletişim teknolojileri, bilgisayar, internet, e-ticaret, e-devlet uygulamaları, bilişim güvenliği alanlarında veri derlenmektedir. Bilgi toplumu olma ölçütlerinin oluşturulmasında yürütülen temel araştırmaların başında gelen HBTKA, 2006 yılı hariç olmak üzere 2004 yılından itibaren düzenli olarak TÜİK tarafından gerçekleştirilmektedir (TÜİK, 2016).

Bu makalede TÜİK'in 2016 yılında yaptığı Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması anket verilerine ait veri kümesi ile veri madenciliği sınıflandırma tekniklerinden karar ağaçları kullanılmıştır. Anket soru formu hane geneli ve fertlere ilişkin bölümlerinden oluşmaktadır. Hane geneline ilişkin bölümde; hanede yaşayan tüm fertlerin yaş, cinsiyet, hanelerde bulunan bilişim teknolojileri ürünleri, internet erişim imkânı, internet bağlantısı olan araçlar, kullanılan internet bağlantı türleri ile evden internete bağlanamama nedenleri sorgulanmaktadır. Fertlere ilişkin bölümde; 16-74 yaş arasındaki bireylerin eğitim ve iş gücü durumu ile bilgisayar ve internet kullanımları, kullanım sıklıkları, kullanım amaçları, kamu ile iletişimde internet kullanımı ve e-ticarete ilişkin sorgulama yapılmaktadır (TÜİK, 2016). Bilgi çağının en etkili araçlarından biri olan internet, son yıllardaki gelişimi ile eğitim, sağlık, haberleşme, pazarlama ve ekonomi gibi pek çok alanı etkisi altına alma gücüne sahiptir. İnternet dünyayı küresel bir köy haline getirdikçe ülkemizde yeni hanelerin internet hizmet alma oranı artmaktadır. İnternet çekirdek aileye, ailenin yeni bir üyesi olarak çok hızlı bir şekilde girmiştir (Kuzu, 2011). HBTKA 2016,2017 ve 2018 yılı sonuçlarına göre, 2017 yılının Nisan ayında hanelerin evden internete erişim oranı % 80,7 iken, 2018 yılı Nisan ayında hanelerin % 83,8'i evden internete erişim imkânına sahip olmuştur. Bu oran 2016 yılının aynı ayı için ise % 76,3'tür. Genişbant ile internete erişim sağlayan hanelerin oranı 2018 yılı Nisan ayında % 82,5 olmuştur. Buna göre hanelerin % 44,5'i sabit geniş bant bağlantı (ADSL, kablolu internet, fiber vb.) ile internete erişim sağlarken,% 79,4'ü mobil geniş bant bağlantı ile internete erişim sağlamıştır. Geniş bant internet erişim imkânına sahip hanelerin oranı 2017 yılında % 78,3, 2016 yılında % 73,1'dir (TÜİK, 2016, 2017, 2018)

Bu çalışma HBTKA verileri kullanılarak veri madenciliği ve karar ağaçları ile yapılmış nadir çalışmalardan biridir. HBTKA veri kümesi kullanılarak yapılan çalışmalarda genellikle istatistiksel teknikler kullanılmış ya da anket sonuçları ile ilgili ekonometrik ve istatistiksel analizler yapılmıştır. İnternet hizmeti diğer bilişim teknolojilerinin tamamlayıcısı konumunda olduğu için veri kümesi oluşturulduktan sonra hane internet hizmeti sahipliği hedef değişken olarak seçilmiştir. Ancak hedef değişken değiştirilerek hanenin sahip olduğu diğer bilişim teknolojilerinin sahiplik durumunun ya da bilgisayar kullanımının hane reisi ve hanenin demografik özelliklerine göre nasıl değiştiğinin veri madenciliği teknikleri ile analizi de yapılabilir. Bu çalışma ile internet hizmeti sahipliğinin hanehalkı karakteristiklerine göre ne şekilde değiştiğinin ortaya konması ve sektörün ilgililerinin dikkatine sunulması hedeflenmiştir.

2. MEVCUT LİTERATÜRÜN İNCELENMESİ

Mevcut literatür incelendiğinde HBTKA verileri kullanılarak karar ağaçları ile yapılmış bir çalışma yoktur. Genel olarak TÜİK tarafından yapılan diğer anketlerin sonuçları kullanılarak veri madenciliği çalışmaları yapılmıştır. Aşağıda HBTKA ya da TÜİK'e ait diğer anket sonuçları kullanılarak yapılmış çalışmaların bazıları incelenmiştir.

TÜİK tarafından yapılan 2014 yılına ait HBTKA anketindeki mikro veri kümesi kullanılarak hanelerdeki bilişim ekipmanları sayısı üzerinde etkili olan faktörler poisson regresyon modeliyle araştırılmıştır. TÜİK tarafından her yıl düzenli olarak yapılan bu ankette bilişim ekipmanları masaüstü bilgisayar, taşınabilir bilgisayar (dizüstü, netbook, tablet vb.), cep telefonu (akıllı telefonlar dâhil), sabit hatlı telefon, oyun

konsolu, dijital fotoğraf makinesi/kamera, DVD/VCD/DivX oynatıcı ve internete bağlanabilen TV (Smart TV) olarak belirlenmiştir. Kurulan modelde, hanede bulunan bilişim ekipmanları sayısı, bağımlı değişken olarak alınmıştır. Elde edilen sonuçlara göre, hanede internet erişim imkânının olması bilişim ekipmanı sayısını pozitif yönde etkilemektedir. Aylık gelir arttıkça bilişim ekipmanı sayısı da artmaktadır. Güneydoğu Anadolu'da bulunan illere göre diğer alt bölgelerdeki hanelerin bilişim ekipmanları sayısı fazladır. İki kişilik haneyle kıyaslandığında hanedeki birey sayısının 7 ve üzeri olması bilişim ekipmanı sayısını azaltmaktadır (Alkan, Abar, & Karaaslan, 2015)

Hane otomobil sahiplik durumunun ardışık logit model ile incelendiği bir çalışmada, TÜİK 2013 yılı Bütçe Anketi'ne katılanların tümü 9975, sürekli çalışan 3733 ve geçici veya sabit süreli sözleşmeli ve sözleşmesiz olarak çalışan 618 hanehalkına ait bilgilerden yararlanılarak üç ayrı ardışık logit modeli incelenmiştir. Model tahmini için hanehalkı reisinin cinsiyeti, mesleği, yaşı, çalıştığı süre, yıllık hanehalkının kullanılabilir geliri ve aylık harcamaları ele alınmıştır. Model tahmini sonucunda elde edilen bulgulara göre, hanehalkının otomobile sahip olmamasını en çok etkileyen değişkenin hanehalkının aylık harcamaları olduğu görülmüştür. Hanehalkının aylık harcamaları arttıkça arabaya sahip olma olasılığı azalmaktadır (Tümsel, 2016).

2013 yılı HBTKA mikro veri kümesi kullanılarak yapılan bir çalışmada Türkiye'de hem 6-15 yaş arası çocukların hem de yetişkinlerin sahip olduğu bilişim teknolojileri ürünleri sayısını belirleyen faktörler sayma veri modeli ile incelenmiştir. Çalışmada kullanım amaçları gibi faktörler, hanehalkındaki çocukların ve yetişkinlerin sahip olduğu bilişim teknolojileri ürünleri sayısının analizinde kullanılmıştır. Sonuçta elde edilen modellerde Robust Poisson Regresyon Modelinden faydalanılmıştır. Elde edilen tahminlerin geçerliliğini araştırmak amacıyla bootstrap tekniğine başvurulmuştur. Sonuçta bir hanedeki bilişim teknolojileri ürünleri kullanımını etkileyen en önemli faktörlerin; hanehalkı geliri, yaş, cinsiyet, eğitim seviyesi, meslek ve yerleşim yeri olduğu tespit edilmiştir (Selim & Balyaner, 2017).

2015 yılına ait TÜİK hanehalkı tüketim harcamaları anketleriyle elde edilen verilerin kullanıldığı bir başka çalışmada Türkiye'de hanehalkı telekomünikasyon harcamaları üzerinde etkili olan hanelerin ve hane reislerinin sosyo-demografik-ekonomik özellikleri belirlenmek istenmiştir. Çalışmada ekonometrik modeller olan Çift Sansürlü Model ve Heckman Metodu kullanılmıştır. Modelde yer alan değişkenlerden istatistiki olarak anlamlı bulunan ev sahibi, hane reisi evli, hane reisi emekli, hane reisi zorunlu sağlık sigortasına sahip, devlet ve özel ayni gelire sahip olan, soba ile ısınan, bir kişilik hanelerin diğerlerine göre daha az telekomünikasyon harcamalarında bulunduğu tespit edilmiştir (Börekçi, 2018)

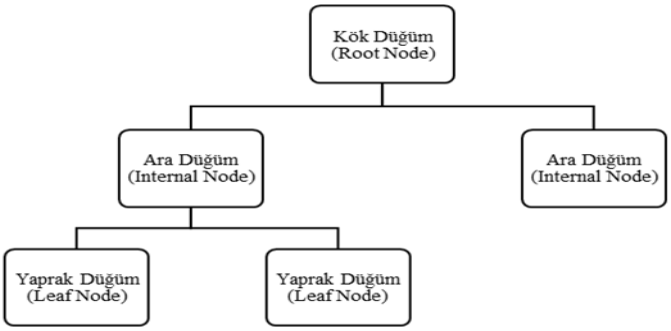
3. YÖNTEM

Çalışmada yöntem olarak veri madenciliği sınıflandırma tekniklerinden karar ağaçları kullanılmıştır.

3.1. Karar Ağaçları

Karar ağaçları ve karar kuralları, birçok gerçek dünya uygulamasında sınıflama problemlerine güçlü bir çözüm olarak uygulanan veri madenciliği metodolojisidir. Verilerden sınıflandırıcılar üretmek için kullanılan en etkili yöntemlerden biri, bir karar ağacı oluşturmaktır (Kantardzic, 2011). Bir sınıflandırma aracı olarak karar ağaçlarının literatürde birçok avantajından bahsedilmektedir. Karar ağaçları kendini açıklayıcı özelliğe sahiptir ve yoğun olduğu zaman bile takip edilmesi kolaydır. Eğer karar ağacı makul sayıda yapraklara sahip ise, kullanıcılar tarafından kolayca anlaşılabilir. Ayrıca, karar ağaçları kurallar kümesine dönüştürülebilir. Böylece, anlaşılması ve yorumlanması daha kolay hale gelir. Karar ağaçları hem nominal (kategorik) hem de sayısal (sürekli) girdi ile işlem yapabilir. Karar ağacının gösterimi herhangi bir ayrık değerli sınıflandırıcıyı ifade etmek için oldukça zengindir. Karar ağaçları parametrik olmayan bir metot olarak kabul edilir. Bundan dolayı, karar ağaçları uzay dağılımı ve sınıflandırıcının yapısı hakkında varsayımlara sahip değildir. Diğer yöntemler verinin normal dağılmamasına veya eksik değerlere karşı hassasken, karar ağacı öğrenimi modellerinde kullanılan veri fazla bir önişleme gerek duyulmadan kullanılabilir. Çeşitli istatistiksel testler yapılarak bir modelin güvenilirliği test edilebilir. Çok büyük boyutlardaki veri kümesine kolayca uygulanabilir. En çok kullanılan veri madenciliği modeli olduğu için bu alan da yayımlanmış birçok doküman ve bilgisayar programı bulunmaktadır (Akpınar, 2014; Diler, 2016; Kuzey, 2012). Bir karar ağacı, doğal bir ağaçta olduğu gibi kök (root), dal (branch) ve yapraklardan (leaf) oluşmaktadır. Bir karar ağacında bu oluşum, kök düğüm, yaprak olmayan veya ara düğüm (non-leaf node / internal node) ve bir karar ağacının sona erdiği noktalar olan yaprak düğüm (leaf node) kavramları ile açıklanır. Şekil 1'de bir karar ağacı çıktısı görülmektedir. Şekilden görüleceği üzere kök düğümden başlayarak her hiyerarşide veri dizisinin belirli kriterlere göre bölünmesi (split)

gerçekleşmektedir. Seçilen karar ağacı algoritmasına göre her aşamada ikili (binary), üçlü (tertiary) ya da çoklu bölünme işlemi gerçekleşebilir (Akpınar, 2014).



Şekil 1. Bir karar ağacı çıktısı

3.2. Karar Ağacı Algoritmaları

Karar ağacı kurulurken eldeki veri tabanının bir kısmı öğrenme bir kısmı da test için kullanılır. Ağaç oluşturulurken oluşturulan modelin çalışıp çalışmadığı belirlenir. Eğer ağaç belirlenen düzeyde çalışıyorsa dallanma durdurulur ve sınıflandırma tamamlanır (Silahtaroğlu, 2016). Karar ağaçlarının oluşturulmasında en önemli nokta hangi değişkenin ilk düğüm, yani kök düğüm olacağının belirlenmesidir. Bunun için çeşitli ölçütler belirlenmiştir ve her farklı ölçüt bir karar ağacı algoritmasına karşılık gelmektedir. Bu algoritmalar şu şekilde sıralanabilir (Atılgan, 2011):

- Entropiye dayalı algoritmalar: ID3, C4.5, C5.0
- Sınıflandırma ve regresyon ağaçları (CART): Twoing, Gini algoritmaları
- Bellek tabanlı sınıflandırma algoritmaları: k-En Yakın Komşu
- İstatistiğe dayalı algoritmalar: Bayesyen sınıflandırma, CHAID

Algoritmaları birbirinden ayıran temel özellikler ise kullanılan ölçü skalası, her düğümden ortaya çıkan yeni düğümlerin sayısı, ağacın büyümesini durdurma kriteri, en iyi bölen özniteliğin seçilmesi ve budama sürecidir (Akpınar, 2014). Karar ağacı algoritmalarının en önemlileri; ID3, C4.5 ve C5.0, C&RT, CHAID ve QUEST algoritmalarıdır.

3.2.1. ID3 Algoritması

J. Ross Quinlan tarafından 1986 yılında geliştirilmiştir. ID3 (Iterative Dichotomiser 3) algoritması çok basit bir karar ağacı algoritması olarak kabul edilir. ID3, bilgi kazancını (information gain) bölme kriteri olarak kullanır. Tüm örnekler, tek bir hedef özellik değerine ait olduğunda veya en iyi bilgi kazanımı sıfırdan büyük olmadığı zaman büyüme duraklar. ID3 herhangi bir budama prosedürü uygulamaz ve sayısal nitelikleri veya eksik değerleri işlemez (Rokach & Maimon, 2005). ID3 algoritmasında sınıflandırma için en ayırıcı özelliğe sahip değişken bulunurken entropi kavramından yararlanılır (Silahtaroğlu, 2016). Bilgiyi ölçmek için kullanılan kavrama entropi denir. Entropi, bir veri kümesindeki belirsizlik, sürpriz veya rastgelelik miktarını ölçmek için kullanılır (Kantardzic, 2011).

3.2.2. C4.5 ve C5.0 Algoritmaları

ID3 algoritmasının geliştirilmiş hali C4.5 algoritmasıdır. C5.0 algoritması ise C4.5'in geliştirilmiş hali olup, özellikle büyük veri setleri için kullanılmaktadır (Çalış, Kayapınar & Çetinyokuş, 2014). C5.0 algoritması, C4.5'in tüm işlevlerini içerir ve modellemede bir dizi yeni teknolojiyi uygular. Bunların arasında en önemli uygulama, örneklerin tanımlanmasının doğruluk oranını arttırmak için kullanılan boosting tekniğidir. Bir diğer önemli uygulama ise maliyet duyarlı karar ağaçlarını oluşturmasıdır (Pang & Gong, 2009). ID3 algoritması değişkenleri birçok alt bölüme ayırır, bu ayırma işlemi aşırı öğrenmeye neden olabileceğinden kazanım yerine kazanım oranı kavramı kullanılmıştır (Silahtaroğlu, 2016). C5.0 algoritması ile C4.5 algoritması kendi içinde karşılaştırıldığında, C5.0 algoritması bazı gelişmiş özelliklere sahiptir. Çok daha hızlı çalışması, daha etkin bellek kullanımı, daha küçük karar ağaçları oluşturması, boosting desteği ve faydasız niteliklerin elimine edilmesini sağlayan winnowing özelliği gelişmiş özellikler olarak sıralanabilir (Akpınar, 2014).

3.2.3. C&RT Algoritması

C&RT algoritması Breiman, Friedman, Olshen ve Stone tarafından 1984 yılında önerilmiştir. Classification and Regression Trees olarak adlandırılan İngilizce ismin baş harflerinden yola çıkarak C&RT algoritması olarak adlandırılmıştır. C&RT algoritması kök düğümden başlayarak her düğüm için olası tüm ayırma şekillerini gözden geçirerek bunlardan en iyisini seçer. C&RT algoritması, her düğümde iki dal üretir. Yani bölünmeler (ya da ayrılmalar) ikilidir (Akküçük, 2011). Ayrılma kriteri için geliştirilmiş yöntemler bulunmaktadır. Bunlar arasında Gini, Twoing, Sıralı Twoing, Simetrik Gini, En Küçük Kareler Sapması (Least Squared Deviation) gibi yöntemler kullanılmaktadır. Bu yöntemlerin kullanımı sınıf hedeflerin sürekli ya da kategorik olmasına göre belirlenmektedir. En yaygın olarak kullanılan ayırma kriterleri Gini ve Twoing kuralıdır (Diler, 2016). C&RT algoritması da ID3 algoritmasında olduğu gibi entropiden yararlanır (Akça, 2014) .

3.2.4. CHAID Algoritması

En popüler karar ağacı algoritmalarından biri olan CHAID (Chi-Squared Automatic Interacton Detection), algoritmasında, eldeki popülasyon bağımlı değişken varyasyonu grup içi minimum, gruplar arası maksimum olacak şekilde ayrıştırarak alt bileşenlere ayrılır (Doğan & Özdamar, 2003). CHAID algoritmasında, diğer istatistiksel yöntemlerde olan normallik, doğrusallık ve homojenlik gibi klasik varsayımlar yoktur. Sürekli ve kategorik bir yapıya sahip tüm değişkenler aynı anda modele dâhil edilir. Bu nedenle bazı araştırmalarda hem parametrik hem de parametrik olmayan analiz yöntemleri içine alınmıştır (Kayri & Boysan, 2007). En iyi açıklayıcı değişkenin ortaya konması için her bir değişken grubu arasında karşılaştırmalar yapılır. Karşılaştırma sürecinin her bir aşamasında veriler, ortaya çıkan değişkene göre yeniden yapılandırılır. Böylece her bir değişken grubunun, bir önceki grubun alt kümesi olduğu ortaya konmuş olur (Kass, 1980). CHAID algoritması, en iyi açıklayıcı değişkenin ortaya konmasında Ki-kare (X^2) test sonuçlarını kullanmaktadır (Doğan & Özdamar, 2003).

3.2.5. QUEST Algoritması

1997 yılında Loh ve Shih tarafından geliştirilen QUEST algoritması tek değişkenli ve doğrusal kombinasyonlu bölünmeleri destekler. Her bölme için, her bir girdi özniteliği ile hedef özniteliği arasındaki ilişki ANOVA F – testi, Levene'nin testi (ordinal ve sürekli öznitelikler için) veya Pearson'ın Ki Kare (nominal öznitelikler) testi kullanılarak hesaplanır. Hedef özniteliği çok terimli ise, iki süper küme oluşturmak için iki yönlü kümeleme kullanılır. Hedef özniteliğiyle en yüksek ilişkilendirmeyi alan öznitelik bölünme için seçilir. Giriş özniteliği ile ilgili en uygun ayırma noktasını bulmak için Kuadratik Diskriminant Analizi (QDA) uygulanır. İkili karar ağaçları oluşturur ve ağaçları budamak için On-kat çapraz doğrulama (Ten cross validation) kullanılır (Rokach & Maimon, 2005). QUEST algoritmasında bölünmüş alan seçimi (split - field selection) ve bölünmüş nokta seçimi (split - point selection) ayrı olarak ele alınır (Kuzey, 2012).

Açıklanan karar ağacı algoritmaları giriş değişkeni, tahmin edici değişken, tahmin türü, bölünme sayısı ve bölme kriterlerine göre Tablo 1'de gösterilmiştir.

Tablo 1. Karar ağacı algoritmalarının karşılaştırılması

Algoritma	Giriş Değişkeni	Tahmin Edici Değişken	Tahmin türü	Bölünme Sayısı	Bölme Kriteri
CHAID	Sürekli ve Kategorik	Sürekli ve Kategorik	Sınıflandırma / Regresyon	≥ 2	Ki-kare/ F testi
QUEST	Kategorik	Sürekli ve Kategorik	Sınıflandırma	=2	Ki-kare / F testi
C&RT	Sürekli ve Kategorik	Sürekli ve Kategorik	Sınıflandırma / Regresyon	=2	Gini / Towing
C4.5/C5.0	Sürekli ve Kategorik	Kategorik	Sınıflandırma	≥ 2	Kazanç Oranı (Entropi)

4. YÖNTEM

Veri kümesinin oluşturulması aşamasında fert ve hane adında iki ayrı veri dosyası için ortak olan bülten numarası değişkeni kullanılarak tek bir veri dosyası oluşturulmuştur. Veri dosyalarını birleştirmek için Microsoft Office Excel 2016 Programından yararlanılmıştır. Fert veri dosyasında hanede yaşı en büyük olan ve cinsiyeti erkek olan kişi hanehalkı reisi olarak kabul edilmiştir. Eğer hanede cinsiyeti ve yaşı en büyük olan kişi erkek değil ise, yaşı en büyük olan ve cinsiyeti kadın olan kişi hanehalkı reisi olarak alınmıştır. Hanehalkı reisi bilgisi alındıktan sonra bu kişiye ait yaş, okuma yazma durumu, eğitim durumu, çalışma durumu ve meslek bilgisi alınmıştır. Hanehalkı reisinin çalışma durumu ve mesleği için referans haftasındaki (28 Mart-03 Nisan 2016) çalışma durumu bilgisi alınarak bu değişkene ait veriler elde edilmiştir. Hanedeki 0-25 Yaş Arası Bireyin Varlığı adında yeni bir değişken oluşturulmuş ve hanede yaşayan fertlerin yaş bilgilerinden bu değişkene ait bilgi doldurulmuştur. 25 yaşında olan fertler değişkene dâhildir. Hanehalkı reisinin mesleği adlı değişkene meslek grupları içinde olmayanlar için 36 numaralı kod ile Diğer adında yeni bir meslek kodu eklenmiştir ve mevcut meslek kodlarına dâhil olmayanlar ya da bu alanın cevabını boş bırakanlar diğer meslek kodu grubuna dâhil edilmiştir. Hanehalkı reisi yaşı adlı ayrık değişken kategorik hale getirilmiş ve 6'ya bölünmüştür. Hanedeki cep telefonu, tablet, masaüstü ve dizüstü bilgisayar, oyun konsolu, smart tv cihazları internete bağlanabilen cihazlardır. Hanehalkına ait aylık gelir bilgisi iki aşamalı kümeleme ile 4 gelir grubuna ayrılmış ve yeni bir değişken olarak veri kümesine dâhil edilmiştir. Gelir değişkenini ayrık hale getirebilmek için iki aşamalı kümeleme analizi ile 4 küme oluşturulmuştur. Ortaya çıkan gelir gruplarına ait hane sayısı ve yüzde bilgileri Tablo 2'deki gibidir.

Tablo 2. Kümeleme sonucunda oluşan gelir grupları

Gelir Grubu	**Hane Sayısı**	**Yüzde (%)**	**Ortalama Aylık Harcanabilir Gelir**
Düşük	5 896	52,5	1 168,11 TL
Orta Alt	4 142	36,9	2 757,34 TL
Orta Üst	1 114	9,9	6 108,17 TL
Yüksek	70	0,6	17 859,57 TL
Toplam	11 222	100,0	

Anket sorularını cevaplayan 11874 haneden kapsam dışı olan haneler çıkarılmış ve 11276 adet hane veri kümesine dâhil edilmiştir. İnternet hizmeti sahiplik durumu için Bilinmiyor şeklinde cevap veren 54 hane veri kümesinden çıkarılmış sonuçta yukarıda belirtilen 19 adet değişkenden ve 11222 haneden oluşan yeni ve tek bir veri kümesi elde edilmiştir. Fert ve hane karakteristiklerinden oluşan Tablo 3'de değişken adları, açıklamaları, aldığı değerler belirtilmiştir.

Tablo 3. Uygulamada kullanılan değişken bilgileri

Sıra No	Değişkenin			
	Adı	Açıklaması	Aldığı Değerler	
1	HHRCINSIYET	Hanehalkı Reisinin Cinsiyeti	1	Erkek
			2	Kadın
2	HHRYAS	Hanehalkı Reisinin Yaşı	1	0-30 Yaş
			2	31-40 Yaş
			3	41-50 Yaş
			4	51-60 Yaş
			5	61-70 Yaş
			6	71 Yaş ve üzeri
3	HHROKUMAYAZMA	Hanehalkı Reisinin Okuma Yazma Durmu	1	Evet
			2	Hayır
4	HHREGITIM		0	Herhangi bir okul bitirmedi

		Hanehalkı Reisinin Eğitim Durumu (Uluslararası Standart Eğitim Sınıflaması)	1	İlkokul Mezunu
			2	İlköğretim/Ortaokul veya Mesleki Ortaokul Mezunu
			3	Lise veya Mesleki Lise Mezunu
			4	İki veya Üç Yıllık Yükseokul Mezunu
			5	Dört Yıllık Yüksekokul veya Fakülte Mezunu
			6	Lisans
			7	Doktora
5	HHRCALISMA	Hanehalkı Reisinin Çalışma Durumu	1	Çalışıyor
			2	Çalışmıyor
6	HHRMESLEK	Hanehalkı Reisinin Mesleği (Uluslararası Standart Meslek Sınıflaması)	0	Silahlı kuvvetlerle ilgili meslekler
			1	Yöneticiler
			4	Büro hizmetlerinde çalışan elemanlar
			5	Hizmet ve satış elemanları
			6	Nitelikli tarım, ormancılık ve su ürünleri çalışanları
			7	Sanatkârlar ve ilgili işlerde çalışanlar
			8	Tesis ve makine operatörleri ve montajcılar
			9	Nitelik gerektirmeyen işlerde çalışanlar

Tablo 3. (devam) Uygulamada kullanılan değişken bilgileri

			21	Bilim ve mühendislik alanlarındaki profesyonel meslek mensupları
			22	Sağlık profesyonelleri
			23	Eğitim ile ilgili profesyonel meslek mensupları
			24	İş ve yönetim ile ilgili profesyonel meslek mensupları
			25	Bilgi ve iletişim teknolojileri ile ilgili profesyonel meslek mensupları
			26	Hukuk, sosyal bilimler ve kültür ile ilgili profesyonel meslek mensupları
			31	Bilim ve mühendislik ile ilgili yardımcı profesyonel meslek mensupları
			32	Yardımcı sağlık profesyonelleri

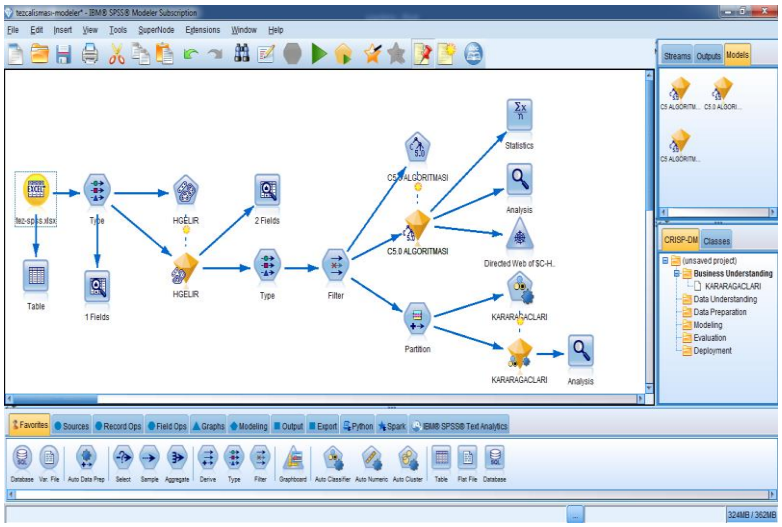
			33	İş ve idare ile ilgili yardımcı profesyonel meslek mensupları
			34	Hukuk, sosyal, kültür ve benzeri alanlar ile ilgili yardımcı profesör
			35	Bilgi ve iletişim teknisyenleri
			36	Diğer
7	HHBUYUKLUK	Hanehalkı Büyüklüğü	1-20	
8	SIFIRYIRMIBESYAS	Hanede 0-25 Yaş Bireyin Varlığı	1	Var
			2	Yok
9	HHBOLGE	Hanenin Bulunduğu Bölge (İstatistiki Bölge Birimleri Sınıflaması Düzey 1)	TR1	İstanbul
			TR2	Batı Marmara
			TR3	Ege
			TR4	Doğu Marmara
			TR5	Batı Anadolu
			TR6	Akdeniz
			TR7	Orta Anadolu
			TR8	Batı Karadeniz
			TR9	Doğu Karadeniz
			TRA	Kuzeydoğu Anadolu
			TRB	Ortadoğu Anadolu
			TRC	Güneydoğu Anadolu
10	HGELIR	Hane Aylık Net Toplam Geliri	HGG-1	Düşük
			HGG-2	Orta Alt
			HGG-3	Orta Üst
			HGG-4	Yüksek
11	HBILGISAYARMASAUSTUSAHIP	Hane Masaüstü Bilgisayar Sahipliği	1	Var
			2	Yok

Tablo 3. (devam) Uygulamada kullanılan değişken bilgileri

12	HBILGISAYARDIZUSTUSAHIP	Hane Dizüstü Bilgisayar Sahipliği	1	Var
			2	Yok
13	HTABLETSAHIP	Hane Tablet Sahipliği	1	Var
			2	Yok
14	HCEPTELSAHIP	Hane Cep Telefonu Sahipliği	1	Var
			2	Yok
15	HOYUNKONSOLSAHIP	Hane Oyun Konsolu Sahipliği	1	Var
			2	Yok
16	HSABITTELSAHIP	Hane Sabit Telefon Sahipliği	1	Var
			2	Yok
17	HSMARTTVSAHIP		1	Var

		Hane Smarttv Sahipliği	2	Yok
18	HBILGISAYARKULLANIM	Hane Bilgisayar Kullanım Bilgisi	1	Evet
			2	Hayır
19	HINTERNETSAHIP	Hane İnternet Himzeti Sahipliği	1	Evet
			2	Hayır

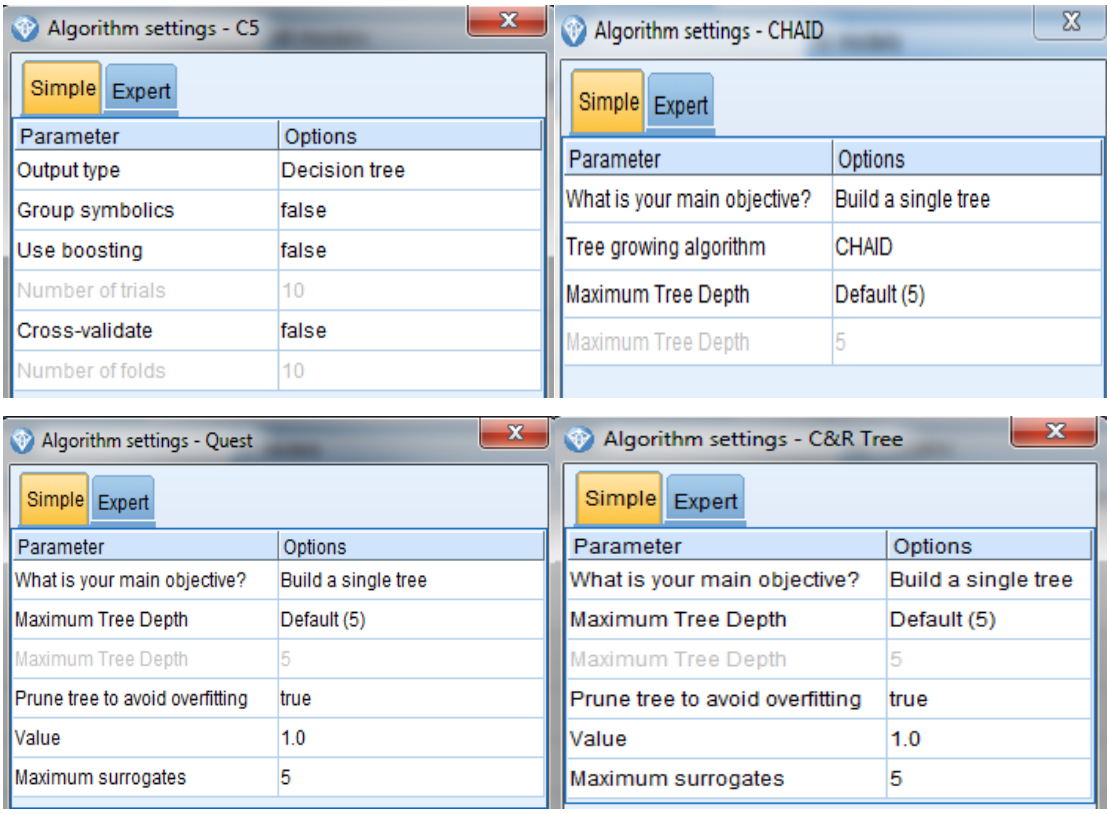
Verilerin analizinde IBM SPSS Modeler 18.1 veri madenciliği programı kullanılmıştır. SPSS modelerde oluşturulan modele ait görüntü Şekil 2'de gösterilmiştir.



Şekil 2. Karar ağacı ile modelleme

Sınıflandırma sürecinde eğitim verileri üzerinde sınıflandırma sürecinde kullanılan algoritmanın doğruluğunun belirlenmesi için bazı yöntemler geliştirilmiştir (Diler, 2016). Hold out yöntemi, analiz edilecek verinin rastgele en az iki alt örnekleme ayrılmasıdır. Algoritmanın parametreleri belirlediği yani eğitildiği bölüme, eğitim kümesi (training set), algoritma sonuçlarının test edildiği kümeye ise, test kümesi (testing set) denir. Algoritmanın eğitimi, eğitim kümesi kullanılarak tamamlandıktan sonra, test kümesi ile modelin doğruluk değeri (accuracy) belirlenmektedir. Eğitim ve test kümesi için genellikle kullanılan oranlar, (0,90 - 0,10), (0,85 - 0,15), (0,80 - 0,20), (0,75 - 0,25) ve (0,70 - 0,30)'dur (Dolgun, 2014). Bir başka yöntem çapraz geçerlilikten türetilen k-katlı çapraz doğrulama olarak adlandırılmaktadır. Veri kümesi k adet alt kümeye ayrılır. Uygulamalarda genel olarak k=10 alınır ve alt kümeler hemen hemen birbirine eşittir (Diler, 2016). Asıl veri kümesinden örneklenerek oluşturulan ve Bootstrap adı verilen diğer bir yöntemde, veri kümeleri modelin oluşturulması için eğitim verisi olarak kullanılır. Bootstrap veri kümesinin dışında kalan veriler test verisi olarak kullanılır. Tesadüfi olarak örneklenen Bootstrap veri kümesinin büyüklüğünün tespitinde modelin hata oranını en iyi temsil edecek oran 0,632 olarak belirlenmiştir (Aydın, 2007). Çalışmada verinin 2/3'ü eğitim kümesine atanırken, kalan kısmı (1/3) test kümesi olarak belirlenmiştir. Hane internet hizmeti sahipliğini etkileyen faktörleri analiz edebilmek ve en iyi performans gösteren algoritmayı seçebilmek için Otomatik Sınıflandırıcı (Auto Classifier) ile dört karar ağacı algoritması seçilmiş ve bu algoritmaların performans sonuçları karşılaştırılmıştır. Algoritmaların parametreleri ile ilgili herhangi bir ayarlama yapılmamış ve tüm algoritmaların varsayılan (default) değerleri alınmıştır. Seçilen algoritmalar için varsayılan değerlere ait ayarlar Şekil 3'de gösterilmiştir. Buna göre algoritmalar seçili olan simple (temel) modda varsayılan ayarlar ile çalışır. C5.0 algoritması için output type parametresi bir karar ağacı, kural kümesi ya da her ikisinin birlikte oluşturulup oluşturulacağının seçimini sağlar. Group symbolics parametresi seçilmezse (false), C5.0, ana düğümü bölmek için kullanılan alanın her değeri için bir alt düğüm oluşturur. Use boosting doğruluk oranını artırmak için C5.0 algoritmasına özel bir parametredir. Cross validate parametresi seçilirse, C5.0, tam veri kümesinde modelin doğruluğunu tahmin etmek için eğitim verilerinin alt kümeleri üzerine kurulmuş bir dizi model kullanacaktır. Tree growing algorithm parametresi CHAID algoritmasına özel

olup CHAID algoritmasının türünün seçilmesini sağlar. Maximum Tree Depth (maksimum ağaç derinliği) parametresi kök düğümün altındaki maksimum seviye sayısının seçilmesini sağlar. Aynı parametre QUEST ve C&RT algoritmaları için de kullanılmaktadır. Varsayılan değeri 5'tir. Prune tree to avoid overfitting (aşırı uyumu önlemek için ağacı budama) parametresi QUEST ve C&RT algoritmalarına özel olup ağacın doğruluğuna önemli katkı sağlamayan alt seviye dallanmaları gidermek için kullanılır. Value parametresi, budanmış ağaç ile risk tahmini açısından en düşük risk taşıyan ağaç arasındaki risk tahminindeki izin verilen farkın boyutunu gösterir. Örneğin, 2 belirlenirse, risk tahmini, tam ağacınkinden büyük olan bir ağaç seçilebilir. Varsayılan durumu True, değeri 1'dir. Maximum surrogates parametresi, eksik değerlerle baş etmek için bir yöntemdir. Ağaçtaki her bölme için algoritma, seçilen bölme alanına en çok benzeyen giriş alanlarını tanımlar. Bu alanlar, bu bölünmenin vekilleridir. Sınıflandırılan bir kayıt bölünmüş bir alan için eksik bir değere sahipse, bölmeyi yapmak için bir vekil alandaki değeri kullanılabilir. Varsayılan değeri 5'tir. Ayrıca Expert (Uzman) ayarları ile algoritmaların diğer parametreleri de ayarlanabilir.



Şekil 3. Seçilen algoritmaların varsayılan değerleri

Aynı veri kümesi üzerinde farklı modeller ile uygulama yapıldıktan sonra kullanılan algoritmaların başarı durumlarının değerlendirmesini yapmak için bazı yöntemler geliştirilmiştir (Akküçük, 2011). Sınıflandırıcıların performansı 3 ölçüt dikkate alınarak karşılaştırılmıştır. Bu ölçütler doğru sınıflandırma yüzdesi (overall accuracy), Alıcı İşletim Karakteristiği (AİK) Eğrisi Altında Kalan Alan (Area Under Receiver Operating Characteristic (ROC) Curve -AUC), kaldıraç değeri (lift ratio)'dir. Doğruluk yüzdesi bir modele ait karışıklık matrisinden hesaplanır. Bu matris Tablo 4'de gösterilmiştir.

Tablo 4. Karışıklık matrisi

		Modelin Tahmin Sınıfı		Toplam Örnek
		Pozitif	Negatif	
Gerçek Sınıf	Pozitif	Doğru Pozitif Sayısı (DP)	Yanlış Negatif Sayısı (YN)	P=DP+YN
	Negatif	Yanlış Pozitif Sayısı (YP)	Doğru Negatif Sayısı (DN)	N=YP+DN

Karışıklık matrisi 4 bölümden oluşmaktadır. DP, pozitif sınıf olarak sınıflandırılan pozitif örnek sayısı; YP, pozitif sınıf olarak sınıflandırılan negatif örnek sayısı; YN, negatif sınıf olarak sınıflandırılan pozitif örnek sayısı; DN, negatif sınıf olarak sınıflandırılan negatif örnek sayısı; P pozitif toplam ve N negatif toplam örnek sayısıdır (Alıç, 2014). Karışıklık matrisi kullanılarak hesaplanan ölçütler; doğruluk (accuracy), hata oranı (error rate), anma (recall), kesinlik (precision) ve F-ölçütü (F-score) ile Alıcı İşletim Karakteristiği (Receiver Operating Characteristic-ROC) eğrisi ve kaldıraç oranı (lift ratio) olarak gösterilebilir (Diler, 2016). Bu değerlerin hesaplanma yöntemleri Tablo 5'de gösterilmiştir (Han, Pei, & Kamber, 2011).

Tablo 5. Performans kriteri hesaplama yöntemleri

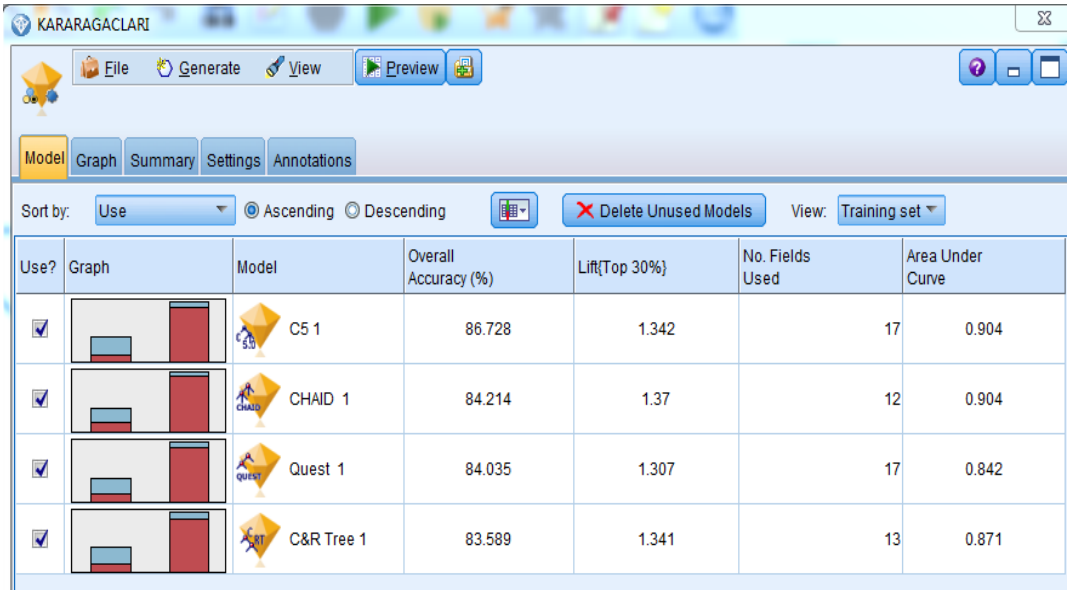
Performans Kriteri	Formül
Doğruluk	Doğruluk = (DP+DN) / (P+N)
Hata Oranı	Hata Oranı = (YP +YN) / (P+N)
Kesinlik	Kesinlik = DP/ (DP+YP)
Anma	Anma = DP / (DP+ YN)
F Ölçütü	F = (2*Kesinlik*Anma) / (Kesinlik+ Anma)

Kaldıraç (lift) değeri, sınıflandırıcı ile tahmin edilen hedef değerin yüzdelik değer içerisindeki oranının hedef değerin ilgilenilen değerinin tüm veri içerisindeki oranına bölümünü ifade eder. Bölme işlemi sonucu elde edilen değerin birden büyük olması ilgili sınıflandırıcı performansının rassal sınıflandırıcı modele (hiç bir sınıflandırıcı modeli kullanılmaması durumunda elde edilecek sonuç) göre ne kadar üstün olduğunu gösterir. Alıcı İşletim Karakteristiği-AİK (Receiving operating characteristics-ROC) grafiği farklı sınıflandırma performanslarını karşılaştırmak için geliştirilen yöntemlerdendir. AİK grafiğinde, doğru pozitif (DP) oranı dikey eksende ve yanlış pozitif (YP) oranı yatay eksende gösterilir (Alıç, 2014). Algoritmaların eğitim (training) verisi için karışıklık matrisleri ile hesaplanan performans kriter değerleri Tablo 6'degösterilmiştir.

Tablo 6. Karışıklık matrisi ile hesaplanan performans kriterleri

	Algoritma			
Performans Kriteri	C5.0	CHAID	C&RT	QUEST
Doğruluk Oranı	**%86,73**	%84,21	%83,59	%84,04
Hata Oranı	%13,27	%15,79	%16,41	%15,96
Kesinlik	%93,00	%94,60	%90,90	%93,00
Anma	%89,00	%85,10	%86,90	%85,90
F-Ölçütü	%90,90	%89,50	%88,80	%89,30

Eğitim veri kümesine göre performans kriterleri değerlendirildiğinde en yüksek doğruluk oranına sahip algoritmanın C5.0 algoritması olduğu görülmektedir. Yine hata oranı olarak en düşük sınıflandırma oranına sahip algoritma C5.0 algoritmasıdır. Kesinlik ve anma değerleri ile hesaplanan F ölçütü değeri, kesinlik ve anma değerlerinin tek başına yorumlanmasına göre daha etkilidir. F ölçütü değerlerine göre en başarılı algoritmanın % 90,90 oranı ile C5.0 algoritması olduğu görülmüştür.



Use?	Graph	Model	Overall Accuracy (%)	Lift(Top 30%)	No. Fields Used	Area Under Curve
☑		C5 1	86.728	1.342	17	0.904
☑		CHAID 1	84.214	1.37	12	0.904
☑		Quest 1	84.035	1.307	17	0.842
☑		C&R Tree 1	83.589	1.341	13	0.871

Şekil 4. Eğitim veri kümesinde karar ağacı algoritma performansları

Şekil 4'e göre eğitim veri kümesinde doğruluk oranı yüzdesi % 86,72 ile en yüksek olan C5.0 algoritması, kaldıraç değeri en yüksek olan 1,37 değeri ile CHAID algoritması, Alıcı İşletim Karakteristiği Eğrisi altında kalan alan (Area Under Curve) değeri 0,904 değerleri ile C5.0 ve CHAID algoritmalarıdır. Sonuç olarak eğitim veri kümesi için en başarılı algoritmalar C5.0 ve CHAID algoritmaları olmuştur. Uygulamada hem test hem veri kümesi için hem de diğer performans analizlerinde en başarılı olan iki algoritmadan biri olan C5.0 algoritması seçilmiş, verilerin modellenmesi, karar ağacının oluşturulması için C5.0 algoritması kullanılmıştır. Çalışmada çapraz doğrulama (cross validate) değeri 10 olarak seçilmiştir. C5.0 parametre değerlerinden budama şiddeti (prunning severity) değeri karar ağacının daha kolay oluşması ve yorumlanabilmesi için 90, her alt dalda minimum 10 kayıt olacak şekilde belirlenmiştir. Son olarak global budama (global prunning) seçeneği de seçilerek model oluşturulmaya hazır hale gelmiştir. 18 adet giriş (input), 1 adet hedef (target) değişkeni ile başlanan analizde 9 adet değişkenin karar ağacında yer aldığı görülmüştür. Bu değişkenler HCEPTELSAHIP, HBILGISAYARKULLANIM, SIFIRYIRMIBESYAS, HTABLETSAHIP, HBILGISAYARDIZUSTUSAHIP, HHBUYUKLUK, HHRYAS, HSMARTTVSAHIP, HGELIRGRUBU değişkenleridir. Hedef değişken olarak seçilen HINTERNETSAHIP değişkenini en iyi açıklayan değişkenin HCEPTELSAHIP değişkeni olduğu görülmüştür. C5.0 algoritması ile HINTERNETSAHIP değişkeni için 8 adet *Evet*, 4 adet *Hayır* değeri olmak üzere 12 adet kural elde edilmiş olup, bu kurallardan 3 adet örnek aşağıda belirtilmiştir.

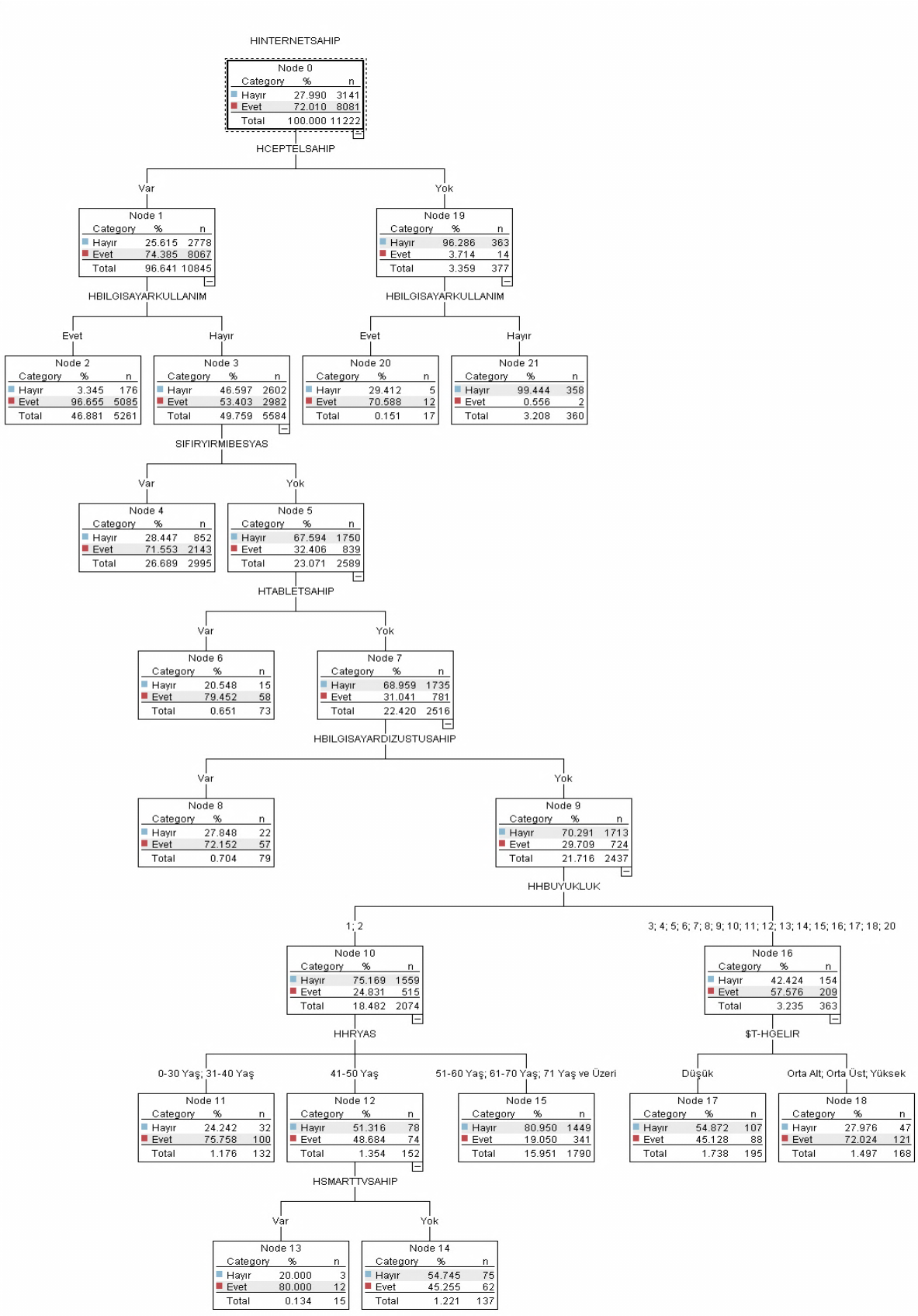
1) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=1,2 ve HHRYAS=41-50 Yaş ve HSMARTTVSAHIP=Var) ise HINTERNETSAHIP=Evet

2) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=1,2 ve HHRYAS=41-50 Yaş VE HSMARTTVSAHIP=Yok) ise HINTERNETSAHIP=Hayır

3) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=3, 4, 5,

6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 ve HGELIR= Orta Alt, Orta Üst, Yüksek) ise HINTERNETSAHIP=Evet

Şekil 5'de C5.0 algoritması ile oluşturulan karar ağacı incelendiğinde hanesinde cep telefonu olanların ve bilgisayar kullananların en yüksek internet sahipliği oranında (% 96,65) olduğu görülmüştür. Cep telefonuna sahip olmayan ve bilgisayar kullanmayan hanelerin ise en düşük internet sahipliği oranında (% 99,44) olduğu görülmüştür. Hanede cep telefonu, tablet, dizüstü bilgisayar, smarttv gibi bilişim ekipmanları sahipliği olması internet hizmetine sahipliğin olmasını da sağlamıştır. Genel olarak hanede 0-25 yaş arası birey olduğu durumda internet hizmeti sahipliğinin de olduğu görülmüştür. Hanehalkı büyüklüğünün 3 ve üzerinde olduğu ve hanenin düşük gelir grubunda olduğu durumda internet hizmeti sahipliğinin düştüğü, orta alt, orta üst ve yüksek gelirli grupta ise arttığı görülmüştür. Hanehalkı büyüklüğünün 3'den az, hanehalkı reisinin yaşının 0-30 ve 31-40 arasında olduğu hanelerde yine internet hizmeti sahipliğinin arttığı görülmüştür.



Şekil 5. C5.0 algoritması ile oluşturulan karar ağacı

Yapılan modelleme sonucunda elde edilen karışıklık matrisi (confusion matrix) Şekil 6'da verilmiştir. Buna göre veri kümesindeki 11222 kayıttan 9577 kayıt doğru sınıflandırılırken, 1645 kayıt yanlış sınıflandırılmıştır. Algoritmanın doğru sınıflandırma yüzdesi % 85,34, yanlış sınıflandırma yüzdesi % 14,66'dır. Bunun anlamı sınıflandırmada internet hizmeti sahipliği durumu *Evet* olan 8081 haneden 7588'i doğru, 493'ü yanlış tahmin edilmiş, internet hizmeti sahipliği *Hayır* olan 3141 haneden 1989'u doğru 1 152'si yanlış tahmin edilmiştir.

Results for output field HINTERNETSAHIP

Individual Models

Comparing $C-HINTERNETSAHIP with HINTERNETSAHIP

Correct	9,577	85.34%
Wrong	1,645	14.66%
Total	11,222	

Coincidence Matrix for $C-HINTERNETSAHIP (rows show actuals)

	1.000000	**2.000000**
1.000000	7,588	493
2.000000	1,152	1,989

Şekil 6. C5.0 algoritması karışıklık matrisi

5. SONUÇ

Bu makale TÜİK'in 2016 yılında yaptığı HBTKA sonucu elde edilen verilerin kullanılması ile hazırlanmıştır. Hedeflenen amaç, hanehalkının internet hizmeti almasını etkileyen faktörlerin, hanehalkının karakteristik özelliklerine göre karar ağaçları ile incelenmesidir. Modelleme için C5.0, CHAID, C&RT, QUEST olmak üzere dört karar ağacı algoritması kullanılmıştır. Veri kümesinin analizi için IBM SPSS Modeler 18.1 programı kullanılmıştır. Programda yer alan Otomatik Sınıflandırıcı (Auto Classifier) kullanılarak belirtilen karar ağacı algoritmalarının performanslarına bakılmıştır. Bu algoritmalar sınıflandırma başarı yüzdeleri, kaldıraç değerleri, alıcı işletim sistemi eğrisi altında kalan alan değerleri bakımından karşılaştırılmıştır . Karşılaştırma sonucunda C5.0 algoritmasının en başarılı algoritma olduğu görülmüştür. Bu nedenle karar ağacı ve kuralları C5.0 algoritması kullanılarak oluşturulmuştur. Yapılan analizde 10 dallanma, 21 düğümden oluşan bir karar ağacı ve 12 adet kural elde edilmiştir. Bu çalışma HBTKA verileri kullanılarak veri madenciliği ve karar ağaçları ile yapılmış nadir çalışmalardan biridir. Sonuç olarak sektörün ilgililerine yol gösterici ve doğru aksiyon almalarını etkileyebilecek bir çalışma yapılmıştır. Hedef değişken değiştirilerek farklı analizler yapılabilir. Çalışma sonucunda elde edilen verilere bakılarak elektronik haberleşme hizmeti veren ve satışını yapan firmaların cep telefonu, tablet, dizüstü bilgisayar satışı gibi kampanyalar yaparak internet sahibi olmayan hanelerin sahip olmasını teşvik edici aksiyonlar almaları beklenebilir.

KAYNAKLAR

Akça, F. (2014). *Veri Madenciliği ile Fen Fakülteleri Öğrenci Profillerinin İncelenmesi: Gazi Üniversitesi Örneği.* Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Akküçük, U. (2011). *Veri Madenciliği: Kümeleme ve Sınıflama Algoritmaları* (Birinci Baskı) İstanbul: Yalın Yayıncılık.

Akpınar, H. (2014). *Data: Veri Madenciliği Veri Analizi* (Birinci Baskı). İstanbul: Papatya Yayıncılık Eğitim.

Alıç, Z. H. G. (2014). *Akut Pankreatit Hastalarının Mortalite Risklerinin Karar Ağacı Yöntemi ile Belirlenmesi.* Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara.

Alkan, Ö., Abar, H., & Karaaslan, A. (2015, 8-10 Ekim). *Hanelerde Bulunan Bilişim Ekipmanları Sayısını Etkileyen Faktörlerin Poisson Regresyon Modeliyle Araştırılması*. Atatürk Üniversitesi 2. Ulusal Yönetim Bilişim Sistemleri Kongresinde sunuldu, Erzurum.

Atılgan, E. (2011). *Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi ile İncelenmesi.* Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Aydın, S. (2007). *Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama.* Yayımlanmamış Doktora Tezi, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.

Börekçi, N. (2018). *Türkiye'de Hanehalkı Telekomünikasyon Harcamalarını Etkileyen Faktörlerin Ekonometrik Analizi.* Yayımlanmamış Yüksek Lisans Tezi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü, Erzurum.

Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014). Veri Madenciliğinde Karar Ağacı algoritmaları ile bilgisayar ve internet güvenliği üzerine bir uygulama. *Journal of Industrial Engineering (Turkish Chamber of Mechanical Engineers), 25*(3-4), 2-19.

Diler, S. (2016). *Veri Madenciliği Süreçleri ve Karar Ağaçları Algoritmaları ile Bir Uygulama.* Yayımlanmamış Yüksek Lisans Tezi, Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü, Van.

Doğan, N., & Özdamar, K. (2003). CHAID Analizi ve Aile Planlaması ile İlgili Bir Uygulama. *Turkiye Klinikleri Journal of Medical Sciences, 23*(5), 392-397.

Dolgun, M. Ö. (2014). *Veri Madenciliği Sınıflama Yöntemlerinin Başarılarının Bağımlı Değişken Prevelansı Örneklem Büyüklüğü ve Bağımsız Değişkenler Arası Ilişki Yapısına Göre Karşılaştırılması.* Yayımlanmamış Doktora Tezi, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and Techniques* (3 ed.). San Francisco: Morgan Kaufman.

Kantardzic, M. (2011). *Data mining: Concepts, Models, Methods, and Algorithms* (2 ed.). New Jersey: John Wiley & Sons.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 119-127.

Kayri, M., & Boysan, M. (2007). Araştırmalarda CHAID analizinin kullanımı ve baş etme stratejileri ile ilgili bir uygulama. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 40*(2), 133-149.

Koyuncugil, A., & Özgülbaş, N. (2009). Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. *Bilişim Teknolojileri Dergisi, 2*(2), 21-32.

Kuzey, C. (2012). *Veri madenciliğinde destek vektör makinaları ve karar ağaçları yöntemlerini kullanarak bilgi çalışanlarının kurum performansı üzerine etkisinin ölçülmesi ve bir uygulama.* Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Kuzu, A. (2011). İnternet ve Aile. *Aile ve Toplum Dergisi, 7*(27), 9-31.

Pang, S.-l., & Gong, J.-z. (2009). C5.0 classification algorithm and application on individual credit evaluation of banks. *Systems Engineering-Theory & Practice, 29*(12), 94-104.

Rokach, L., & Maimon, O. (2005). Decision trees *Data mining and knowledge discovery handbook* (pp. 165-192): Springer.

Selim, S., & Balyaner, İ. (2017). Türkiye'de Hanehalkının Sahip Olduğu Bilişim Teknolojileri Ürünleri Sayısını Belirleyen Faktörlerin Araştırılması: Bir Sayma Veri Modeli. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9*(22), 428-454.

Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları* (Üçüncü Baskı). İstanbul: Papatya Yayınları.

Türkiye İstatistik Kurumu (2016). Hanehalkı Bilişim Teknolojileri Kullanım Araştırması; TÜİK, 21779. Türkiye İstatistik Kurumu Haber Bülteni.

Türkiye İstatistik Kurumu (2017). Hanehalkı Bilişim Teknolojileri Kullanım Araştırması; TÜİK, 24862 Türkiye İstatistik Kurumu Haber Bülteni.

Türkiye İstatistik Kurumu (2018). Hanehalkı Bilişim Teknolojileri Kullanım Araştırması; TÜİK, 27819. Türkiye İstatistik Kurumu Haber Bülteni.

Tümsel, B. (2016). *Hanehalkı otomobil sahip olma durumunun ardışık logit modeli ile tahmini.* Yayımlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.