CART ANALİZİ İLE HANEHALKI İŞGÜCÜ ANKETİ SONUÇLARININ ÖZETLENMESİ

Ayşe OĞUZLAR(*)

Özet: Veri madenciliği; istatistik, veri tabanı teknolojisi, makine öğrenimi, örüntü tanıma, yapay zeka ve görselleştirmenin kullanıldığı bir disiplinler arası çalışmadır. Veri madenciliğinin aşamalarından en önemlisi olan modelleme tekniklerinden birisi de karar ağaçları önemli bir yere sahiptir. CART veya aynı anlama gelmek üzere C&RT algoritması, karar ağacı algoritmalarının içinde en fazla kullanılanlardan birisidir. Bu çalışmada, DİE 2002 III. dönem hanehalkı işgücü anketi sonuçları CART algoritması kullanılarak özetlenmeye çalışılmıştır.

Abstract: Data mining is an interdisciplinary exercise and statistics, database technology, machine learning, pattern recognition, artificial intelligence and visualization, all play a role. Modelling is very important phase of data mining. One of the modelling techniques is decision trees. CART or C&RT algorithm is one of the major algorithm for classification. In this study DİE 2002 III. term household manpower survey was used. For this data CART algorithm was used and the rusults were explained and interpreted.

I. Giriş

Veri madenciliğinin uygulanabilmesi için yığın halinde yapılandırılmış verilerin elimizde bulunması ön koşuldur. Veri madenciliği farklı formatlarda çok sayıda kütükte yığın halindeki veriler arasında gizli bir şekilde bulunan mesajları çekip çıkarmamıza yarayan bir araçtır (Muata ve Bryson:2003:1).

Veri madenciliği modelleme aşamasında iki temel kategoriye ayrılmaktadır:

1. Betimsel Modelleme (Descriptive Modeling)
2. Kestirimsel Modelleme (Predictive Modeling)

Veri madenciliğinin ana hedeflerinden bir tanesi de verilerin sınıflandırılması (data classification) konusudur. Verilerin sınıflandırılması konusu hem betimsel ve hem de kestirimsel modelleme içinde yer almaktadır. Kestirimsel sınıflama modellerinden biri de CART yani sınıflandırma ve regresyon ağacıdır. İzleyen bölümlerde karar ağacı teknikleri ile sınıflandırma ve regresyon ağacına ilişkin genel bilgi verilerek, uygulama bölümünde de 2002 hanehalkı işgücü anketi sonuçları CART algoritması ile özetlenmeye çalışılmıştır.

(*) Yrd.Doç.Dr. Uludağ Üniversitesi İİBF Ekonometri Bölümü

II. Karar Ağacı Teniklerinin Gelişimi ve Kullanım Alanları

Sosyal ve ekonomik olaylarını daha güvenilir bir şekilde gösterebilmek için standart istatistik tekniklerin dışında yeni analiz tekniklerinin geliştirilmesi ile 1970'li yıllarda kullanıma alınan AID karar ağacı temelli ilk algoritmadır. Bu teknik en kuvvetli ve en iyi tahmini gerçekleştirebilmek için bağımlı ve bağımsız değişkenler arasındaki tüm ilişkilerin incelenmesine dayanmaktadır. Teknikte, en kuvvetli ilişkiye sahip bağımsız değişken bulunduğunda, veri kümesi bu bağımsız değişken değerlerine göre ikiye ayrılmakta ve süreç mümkün bölünmeler tamamlanıncaya kadar devam etmektedir. İlk temelleri AID yöntemi ile atılan karar ağacı modelleri çeşitli algoritmalar ile sürdürülmüştür (Akpınar, 2000:10).

Karar ağacı temelli analizlerin yaygın olarak kullanıldığı sahalar,

1.Belirli bir sınıfın üyesi olacak elemanların belirlenmesi,

2.Çeşitli vakaların yüksek, orta, düşük risk grupları biçiminde kategorilere ayrılması,

3.Gelecekte gerçekleşebilecek olayların tahmin edilebilmesi için kurallar oluşturulması,

4.Parametrik modellerin kurulmasında kullanılacak çok sayıdaki değişken ve veri kümesinden önemli olanlarının seçilmesi,

5.Yalnızca belirli alt gruplara özgü ilişkilerin tanımlanması,

6.Sürekli değişkenlerin kesikli değişkenlere dönüştürülmesi ve kategorilerin birleştirilmesidir.

Karar ağacının kullanıldığı uygulamalar ise aşağıdaki biçimde sıralanabilir:

1.Demografik grupların hangilerinin mektup aracılığıyla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi,

2.Kredi geçmişlerinin kullanılmasıyla bireylere ilişkin kredi kararlarının verilmesi,

3.İşletmeye en faydalı olan bireylerin özelliklerinin kullanılmasıyla işe alma süreçlerinin belirlenmesi,

4.Tıp ile ilgili gözlem verilerinden hareketle en etkin kararların verilmesi,

5.Satışları hangi değişkenlerin etkilediğinin belirlenmesi,

6.Ürün hatalarına yol açan değişkenlerin belirlenmesidir.

Karar ağaçları sınıflandırma ve kestirim için güçlü ve popüler araçlardır. Karar ağaçlarının çekici olan yönü, bir takım kuralları temsil etmesidir. Kurallar, karar ağacından kolaylıkla okunabilir.

Karar ağaçlarının güçlü yönleri aşağıdaki gibi özetlenebilir:

1.Karar ağaçları anlaşılabilir kurallar üretirler.

2.Karar ağaçları aşırı hesaplamaya gerek kalmadan sınıflandırma yaparlar.

3.Karar ağaçları hem sürekli ve hem de kesikli değişkenler için uygundur.

4.Karar ağaçları sınıflandırma ve kestirim için hangi alanların en önemli olduğunu açık biçimde gösterir (DMS Tutorial, 2001:1).

Karar ağaçları yukarıdan aşağıya, genelden özele doğrultuda eğitilmiş verilerden (training data) türetilmektedir. Bir karar ağacının başlangıç aşaması, kök düğümüdür. Tüm örneklemlerin aynı sınıfa ait olması durumunda herhangi bir karar alınmasına gerek olmadan çözüm bitirilir. Eğer düğümdeki örneklemler bir veya birkaç sınıfa ait ise, düğümde bir test yapılarak bir bölünme oluşur (Roiger ve Geatz, 2003:68).

Karar ağaçları karar kurallarının gösterildiği çizimlerdir. Ağaç örneklemdeki tüm gözlemleri içeren bir kök ile başlar. Ağaç boyunca aşağıya doğru inildiğinde veriler karşılıklı tek alt kümeler içerecek biçimde dallara ayrılır.

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 istatistiksel olarak homojen (benzer) 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 ardıl olarak ağaç tamamıyla büyüyene kadar sürer. Kullanılan istatistiksel test, hedef değişkenin ölçüm düzeyine bağlıdır. Eğer hedef değişken sürekli bir değişken ise, F testi kullanılır. Eğer hedef değişkeni kategorik ise, ki-kare testi kullanılmaktadır. Exhaustive CHAID' ın ise (Ayrıntılı CHAID), hesaplanması uzun zaman alır ve her bir önkestirici için tüm mümkün bölünmeleri araştırır. Ayrıntılı CHAID, CHAID' in modifiye edilmiş şeklidir. CHAID yönteminin zayıf kalan yönlerini gidermek amacıyla geliştirilmiştir. Bazı durumlarda CHAID, bir değişken için optimal bölünmeyi bulamayabilir. Bu durumda tüm kalan kategoriler istatistiksel olarak farklı bulunduğunda, kategorileri birleştirmeyi durdurur. Ayrıntılı CHAID buna çare olarak yalnızca iki süper kategori kalana değin kestirim değişkeninin kategorilerini birleştirmeyi sürdürür. Ardından ön kestirici için birleşim serilerini inceler, hedef değişken ile en güçlü birlikteliği veren kategori kümesini bulur ve bu birliktelik için düzeltilmiş p değerini hesaplar. Bu nedenle Ayrıntılı CHAID, her bir ön kestirici için en iyi bölünmeyi bulur ve bölünme için hangi ön kestiricinin seçileceğine düzeltilmiş p değerlerini kıyaslayarak ulaşır.

Ayrıntılı CHAID, kullandığı istatistiksel testler ve kayıp değerleri değerlendirmesi açısından CHAID' e benzerdir fakat hesaplanması uzun zaman almaktadır. Zamanın sorun olmadığı durumlarda Ayrıntılı CHAID' in kullanılması daha faydalı olacaktır çünkü bazı durumlarda kullanışlı bölünmeler bulabilmektedir. Verilere bağlı olsa da, CHAID ile Ayrıntılı CHAID sonuçları arasında farklılık bulunmamaktadır.

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 (Loh ve Shih, 1997). Göreceli olarak yeni bir iki değerli ağaç büyüme algoritmasıdır. Ayrı ayrı değişken seçimi ve bölünme noktası seçimi ile ilgilidir. QUEST' deki tek değişkenli bölünme yaklaşık olarak yansız değişken seçimini sağlar. Bunun anlamı, tüm kestirim değişkenlerinin hedef değişkene göre eşit düzeyde bilgi sağlayıcı olması durumunda QUEST' in eşit olasılığa sahip kestirim değişkenlerinden herhangi birini seçeceğidir. QUEST hesapsal açıdan etkinlik için yaratılmıştır.

Kullanılan modelin kestirim gücünün test edilmesinde veya uygulanan birden çok modelin hangisinin daha uygun olduğunun belirlenmesinde yanlış sınıflama matrisi (misclassification matrix) kullanılmaktadır. Yanlış sınıflama matrisinde yer alan risk tahmin değeri yanlış bir biçimde sınıflanan örnek yüzdesini göstermektedir.

III.CART' ın Temel Yapısı

CART algoritması, ağaç yapısına dayalı olarak sınıflandırma ve regresyon modellerinin türetilmesi için yaygın olarak kullanılan bir istatistiksel prosedürdür. CART ağaç modeli, tek değişkenli ikili kararların bir hiyerarşisini içerir. CART verileri iki alt kümeye ayırdığı için her bir alt küme içindeki durumlar, bir önceki alt kümeden daha homojen olacaktır. Bu ardışık süreç, homojenlik kriterine ulaşılıncaya veya diğer bazı durma kriterleri sağlanıncaya değin kendini tekrar eder. Aynı kestirim değişkeni ağaçta farklı düzeylerde pek çok kez kullanılabilir. Ağacın yapısı önceden belirlenmemekte, verilerden türetilmektedir (Answer Tree 3.0 User's Guide, 2001:189). CART, kök düğümünde, verilerin iki gruba bölünmesi için en iyi değişkenin seçilmesini sağlar ve farklı bölümlendirme (splitting) kriterleri kullanır. Bu bölümlendirme kriterlerinin tümü, her bir alt kümedeki sınıf etiketlerini mümkün olduğunca homojen olacak biçimde bölümlendirir (Classification and Regression Trees: An Introduction, 2003:12). Bölümlendirme prosedürü çocuk düğümlere (child node) veya alt düğümlerin her birine ardışık olarak uygulanır (Hand, Manila ve Smyth, 2001:147).

CART ağaçları, kesin bir heterojenlik (impurity) ölçüsüne bağlı olarak düğümlere ayrılmış iki değerli (binary) ağaçlardır ve bu nedenle de sonuçta homojen dallar oluşmaktadır (Ahola ve Rinta-Runsala, 2001:17). Ağacın hedefi

benzer veya aynı çıktı değerlerine sahip olma eğiliminde olan alt gruplar yaratmaktır. CART modelleri için bölünmelerin bulunmasında kullanılan dört farklı heterojenlik ölçüsü mevcuttur. Kategorik hedef değişkenler için Gini, Twoing veya (sıralayıcı hedef değişkenleri için) sıralı Twoing, sürekli hedef değişkenler için ise en küçük kareli sapma (LSD) kullanılabilir.

Gini indeksi aşağıdaki şekilde yazılabilir:

$$g(t) = 1 - \sum_j p^2(j/t) \qquad (1)$$

Her hangi bir düğümde durumlar kategoriler arasında eşit biçimde dağıldığında Gini indeksi $1 - \frac{1}{k}$ maksimum değerini alır. Bir düğümdeki durumlar aynı kategoriye ait olduğunda ise Gini indeksi 0' a eşit olacaktır (Apte ve Weiss, 1997:4).

Twoing indeksi, hedef değişken kategorilerinin iki süper sınıfa bölümlendirilmesine dayalıdır ve ardından bu iki süper sınıfa dayalı olarak kestirim değişkenindeki en iyi bölünmeyi bulur. t düğümünde s bölünmesi için Twoing kriter fonksiyonu şu şekilde tanımlanabilir (Answer Tree 3.0 User's Guide, 2001:194):

$$\Phi(s,t) = p_L p_R \left[\sum_j \left| p(j/t_L) - p(j/t_R) \right| \right]^2 \qquad (2)$$

Fonksiyonda yer alan t_L ve t_R, s bölünmesi tarafından yaratılan düğümleri göstermektedir. s bölünmesi, bu kriteri maksimize eden bölünme olarak belirlenir. İki süper sınıf olan C1 ve C2 aşağıdaki biçimde tanımlanabilir:

$$C_1 = \{j : p(j/t_L) \geq p(j/t_R)\}$$
$$C_2 = C - C_1 \qquad (3)$$

Burada C, hedef değişkenin kategori kümesidir.

Sıralı Twoing indeksi, sıralayıcı hedef değişkenleri için Twoing indeksinin değiştirilmiş şeklidir. Sıralı Twoing kriterindeki farklılık yalnızca bitişik kategorilerin süper sınıflar ile birleştirilmesidir. Örneğin bir değişkenin 4 kategorisi olsun. Twoing kriteri 1 ve 4'ü bir süper sınıf ve 2 ve 3' ü de diğer bir süper sınıf olarak belirlemiş olsun. Bununla beraber kategoriler sıralı

olduğundan 1 ve 4 kategorileri birleştirilemez çünkü bunlar bitişik kategoriler değillerdir. Sıralı Twoing indeksi bu durumu göz önüne aldığından 1 ve 4 gibi kategoriler bitişik olmadığından birleştirilemez.

Sürekli hedef değişkenleri için en küçük kareli sapma (LSD) heterojenlik ölçüsü kullanılmaktadır. LSD ölçüsü R(t), t düğümü için basit (ağırlıklandırılmış) düğüm içi varyansıdır ve düğüm için risk tahminine eşittir. R(t)' nin formülü aşağıdaki şekildedir (Answer Tree 3.0 User's Guide, 2001:195):

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_n f_n (y_i - \bar{y}(t))^2 \qquad (4)$$

Nw(t), t düğümündeki ağırlıklandırılmış durum sayısı, w_n i durumu için mevcut ise ağırlıklandırılmış değişken değeri, f_n mevcut ise frekans değişkeninin değerini, y, hedef değişkenin değerini ve $\bar{y}(t)$ ise t düğümü için ağırlıklı ortalamayı göstermektedir.

Sonuçta elde edilen ağacın büyüklüğü, karmaşık budama (pruning) sürecinin bir sonucudur. Çok büyük bir ağaç, uyumun üzerinde (overfitting) ve çok küçük ağaç, yetersiz tahmin gücüne sahip olacaktır. Ağaç yapısının hiyerarşik formu, CART gibi algoritmaları ağaç yapısına dayanmayan diğer sınıflandırma algoritmalarından açık bir şekilde ayırır.

Son yıllarda CART analizinin kullanımında artış gözlemlenmektedir. Ağaca dayalı olan farklı yapısından dolayı kabul görmesi uzun zaman almıştır. Bunun yanında istatistikçilerin tekniğe dayalı tecrübeleri ya çok az veya hiç bulunmamaktadır. CART' ın genel kabul görmesini sınırlandıran diğer faktörler olarak, analizin kompleks oluşu ve yakın geçmişe kadar bu analizi gerçekleştirecek yazılımların kullanma güçlüğü sayılabilir.

CART analizi ağaç yapısına dayalı diğer sınıflama teknikleri ile kıyaslandığında çok sayıda avantaja sahiptir. İlki ve belki de en önemli olan özelliği parametrik olmayışıdır. Diğer bir söyleyişle ön kestirici veya aynı anlama gelmek üzere bağımsız değişken değerlerine ilişkin varsayımlar gerektirmemektedir. Bu nedenle CART analizinde kullanılacak değişkenler çok çarpık sayısal değişkenler olabileceği gibi, sınıflayıcı veya sıralayıcı yapıya sahip kategorik değişkenler de olabilir. Bu önemli bir özelliktir ve analizi yapacak araştırmacıya, normallik araştırma ve dönüşüm yapma gibi işlemler gerektirmediğinden zaman kazandırmaktadır. CART analizi, ele alınan problem yüzlerce mümkün bağımsız değişken içerse bile, bölümlendirilecek tüm mümkün değişkenleri araştırma gücüne sahiptir.

CART, ele alınan veri kümesi eksik değerler içerdiğinde kullanışlı bir analizdir. Eksik değerler çok fazla olduğunda, bu değerler bir vekil değişken olarak ağaç yapısında yer alırlar.

CART analizinin bir başka avantajı da göreceli olarak otomatik bir makine öğrenim tekniği olmasıdır. Diğer bir söyleyişle analizin karmaşıklığı ile kıyaslandığında, araştırmacıya göreceli olarak az miktarda girdi gerekmektedir. Diğer çok değişkenli modelleme yöntemleri araştırmacılara çok fazla girdi gereksinimi yüklemekte, geçici sonuçların analizini gerektirmekte ve ilgili yöntemin modifikasyonu gerekmektedir. Son olarak avantajlarından biri olarak istatistikçi olmayanlar için bile yorumunun çok kolay olduğunu söylemek yanlış olmayacaktır.

CART analizinin avantajlarının yanında dezavantajlarının olduğu da unutulmamalıdır. CART göreceli olarak yeni bir analizdir. Bu nedenle de geleneksel istatistikçiler tarafından kabulünde sorunlar gözlenmektedir. CART temel istatistiksel yazılım paketlerinde standart bir analiz tekniği olmadığından yer almamaktadır. Ayrıca işlem zamanı diğer ağaç yapısına dayalı algoritmalar ile kıyaslandığında daha uzundur. Çoklu değil de iki değerli ağaç tekniği olması da bir dezavantaj olarak sayılabilir. Fakat değişken sayısı çok fazla olduğunda veya değişkenlerin çok fazla kategorisinin olması durumunda iki değerli ağaç yapısı daha yorumlanabilir sonuçlar üretebilir (An Introduction to Classification and Regression Tree (CART) Analysis, 2004:6).

CART analizi uygulama bölümünde 2002 hanehalkı işgücü anketi sonuçlarına uygulanmıştır. CART analizinin bu veri kümesine uygulanışının ana nedeni, veri kümesinin çok fazla eksik değer içermesidir. Yukarıda da belirtildiği gibi CART analizi çok fazla sayıda eksik değer içeren veri kümeleri için en iyi ağaç yapısına dayalı sınıflama tekniği olarak karşımıza çıkmaktadır. CART analizinin diğer bir kullanılma nedeni olarak da veri kümesinin büyüklüğü sayılabilir. Çok büyük veri kümelerinde, iki değerli bir ağaç yapısına sahip olduğundan daha yalın ve anlaşılabilir sonuçlar üretebilmektedir.

IV.Uygulama

Uygulama bölümünde 2002 hanehalkı işgücü anketi veri kümesinden yararlanılmıştır. Bu veri kümesinden alınan 300689 gözlem değeri için kullanılan değişkenler aşağıdaki şekildedir:

1. Cinsiyet
    1. Erkek
    2. Kadın
2. Bitirilen Yaş
3. Hanehalkı Reisine Yakınlık
    1. Hanehalkı reisi
    2. Eşi
    3. Çocuğu
    4. Gelini veya damadı
    5. Torunu
    6. Ebeveyni

7. Diğer Akrabası
8. Akraba Olmayan

4. En Son Bitirilen Okul
1. Bir okul bitirmedi
2. İlkokul
3. İlköğretim
4. Ortaokul
5. Meslek ortaokulu
6. Lise
7. Mesleki Lise
8. 2 yıllık ön lisans
9. 2 yıllık lisans
10. 4 yıllık lisans
11. Mastır, doktora vb.

5. Medeni Durum
1. Hiç evlenmedi
2. Evli
3. Boşandı
4. Eşi öldü

6. İş arama
1. Evet
2. Hayır

değişkenleri ele alınmıştır.

SPSS formatındaki veriler öncelikle AnswerTree 3.0 paket programına aktarılmıştır. Bağlı veya aynı anlama gelmek üzere hedef değişken olarak ISSIZ şeklinde kısaltılmış iş arama değişkeni kullanılmıştır. Bağımsız değişkenler olarak ise; cinsiyet, bitirilen yaş, hanehalkı reisine yakınlık, en son bitirilen okul ve medeni durum değişkenleri ele alınmıştır. İlgili veri kümesinin yaklaşık %10' u test örneklemi olarak ayrılmıştır. İlgili değişkenler için Gini heterojenlik indeksi kullanılarak elde edilen beşinci ve son düzey ağaç grafiğinin yorumları aşağıda belirtilmiş fakat yapısı karmaşık ve büyük olduğundan, şekli makale içerisinde gösterilememiştir.

Elde edilen beşinci düzey ağaç diagramına bakıldığında ortaokul, meslek ortaokulu, lise ve 2 yıllık lisans mezunu olan hiç evlenmeyen, boşanan ve eşi ölen erkeklerin %12,67' si (1472 kişi) iş ararken, %54,52' si ise iş aramamaktadır (6335 kişi). İlkokul , mesleki lise, 4 yıllık lisans, 2 yıllık ön lisans ve mastır, doktora vb. mezunu hiç evlenmemiş, boşanmış ve eşi ölmüş erkeklerin ise %18,89' u iş ararken (2374 kişi), %26,51' i iş aramamaktadır (3331 kişi). İlkokul, meslek ortaokulu ve 2 yıllık lisans mezunu evli erkeklerin ise %9,13' ü (3134 kişi) iş aramakta ve %23,38' i iş aramamaktadır (8024 kişi). Sayılan kategorilerin dışında okul kategorilerinden mezun olan evli

erkeklerin ise %1,08' i iş ararken (126 kişi), %92,16' sı iş aramamaktadır (10785 kişi).

Hanehalkı reisinin ebeveyni olan, evli veya boşanan, bir okul bitirmeyen veya ilköğretim mezunu olan erkeklerin %1,18' i iş ararken (3 kişi), %85,49' u iş aramamaktadır (218 kişi). Hanehalkı reisinin ebeveyni olmanın dışındaki kategorilerin birinde yer alan, evli veya boşanmış, bir okul bitirmeyen veya ilköğretim mezunu erkeklerin ise %6,52' si iş ararken (396 kişi), %46,64' ü (2835 kişi) iş aramamaktadır.

Bir okul bitirmeyen veya ilköğretim mezunları arasından, hiç evlenmeyen veya eşi ölen erkeklerin %1,08' i iş aramakta (126 kişi), %92,16' sı ise iş aramamaktadır (10785 kişi).

2 yıllık ön lisans veya 2 yıllık lisans mezunu, hanehalkı reisi, eşi, ebeveyni veya hanehalkı reisinin akrabası olmayan kadınların %4,93' ü iş aramakta (59 kişi), %48,37' si (579 kişi) iş aramamaktadır. 4 yıllık lisans, mastır, doktora vb. mezunu olan hanehalkı reisi, eşi, ebeveyni veya hanehalkı reisinin akrabası olmayan kadınların %3,25' i iş ararken (75 kişi), %28,26' sı iş aramamaktadır (652 kişi).

Hanehalkı reisi, eşi, ebeveyni veya hanehalkı reisinin akrabası olmayan, ilkokul, lise, mesleki lise veya 2 yıllık ön lisans mezunu olan kadınların %1,72' si (1024 kişi) iş aramakta ve %81,28' si ise iş aramamaktadır (48478 kişi). Hanehalkı reisinin çocuğu, gelini veya damadı, torunu veya diğer akrabası olan ilkokul, lise, mesleki lise veya 2 yıllık ön lisans mezunu olan kadınların %7,30' u iş ararken (1610 kişi), %65,25' i iş aramamaktadır (14383 kişi).

Evli veya boşanan, bir okul bitirmeyen, ortaokul veya ilköğretim mezunu olan kadınların %2,08' iş ararken (153 kişi), %81,48' i ise iş aramamaktadır. Buna karşın hiç evlenmeyen veya eşi ölen, bir okul bitirmeyen, ortaokul veya ilköğretim mezunu olan kadınların %1,86' sı iş aramakta (278 kişi), %91,70' i ise iş aramamaktadır (13738 kişi). Belirtilmesi gereken bir diğer nokta, bağlı değişkenin 1 ve 2 nolu kategorileri dışında yer alan üçüncü bir kategoridir. Bu üçüncü kategori ilgili bölümlendirmede yer alan eksik gözlemleri göstermektedir.

Beşinci düzey ağaç diagramı için bulunan yanlış sınıflama matrisi ise Tablo 1' de gösterildiği şekildedir.

Tablo 1: *Beşinci Düzey Ağaç Diagramı İçin Yanlış Sınıflama Matrisi*

Training Sample					
Misclassification Matrix					
		Actual Category			
		2		1	Total
Predicted Category	2	101355	23818	4786	129959
		19060	113674	7817	140551
	1	0	0	0	0
	Total	120415	137492	12603	270510
		Risk Statistics			
Risk Estimate		0,205098			
SE of Risk Estimate		0,000776329			

Tablo 1' e bakıldığında, 215029 kişinin doğru bir şekilde sınıflandırıldığını söylemek mümkünken, 55481 kişinin ise yanlış bir şekilde sınıflandırıldığını söylemek yanlış olmayacaktır. Risk tahmin değeri olarak bulunan 0,205098 değeri bize ana kütlenin yaklaşık %20' sinin yanlış bir biçimde sınıflandırıldığını göstermektedir. Ağaç diagramında yer alan Improvement değerleri ise, ilgili bölümlendirmenin ağacın kestirim performansını ne derecede arttırdığını göstermektedir.

Beş düzeyden daha fazla ağaç diagramı genişletilemediğinden beşinci düzey ağaç diagramı elde edilmiş ve yorumlanmıştır. Ayrıca test veri kümesi olarak ayrılan ana kütlenin %10' luk dilimi için elde edilen beşinci düzey ağaç daigramına ve yanlış sınıflama matrisine bakıldığında, benzer sonuçların elde edildiğini söylemek yanlış olmayacaktır.

V. Sonuç

Ağaç yapısına dayalı sınıflama tekniklerinin yorumunun, istatistikçi olmayanlar için bile son derece basit olduğundan daha önce söz edilmişti. Çalışmada beş farklı ağaç diagramının tümünün de yorumunun çok basit olduğunu söylemek yanlış olmayacaktır. Ağaç yapısına dayalı sınıflama teknikleri içerisinde yer alan CART analizi; parametrik olmayan bir teknik olduğundan, ele alınan veri kümesi çok fazla eksik değer içerdiğinden ve iki değerli yalın bir ağaç diagramı çıktısı verdiğinden çalışmamızda tercih edilmişti.

Son olarak elde edilen beşinci düzey ağaç diagramına bağlı olarak sonuçların tekrar gözden geçirilmesinde fayda vardır. Bu diagrama göre en fazla

iş arayan grup; ilkokul, meslek lisesi veya yüksek öğrenim görmüş, evlenmemiş, boşanmış veya eşi ölmüş erkekler en fazla iş aramaktadır (%18,89). En fazla iş arayan diğer bir grup da; hanehalkı reisinin çocuğu, gelini, torunu veya diğer akrabası olan ilkokul, lise, mesleki lise veya 2 yıllık ön lisans mezunu olan kadınlardır (%7,30).

En yüksek oranda iş aramayan grup ise; bir okul bitirmeyen veya ilköğretim mezunu, hiç evlenmeyen veya eşi ölen erkeklerin yaklaşık %92' si iş aramamaktadır. Diğer bir yüksek oranda iş aramayan grup ise; hanehalkı reisinin ebeveyni olan, evli veya boşanmış, bir okul bitirmeyen veya ilköğretim mezunu erkeklerin ise yaklaşık %85' i iş aramamaktadır.

Kabaca erkeklerin daha fazla iş aradığını söylemek mümkündür. Ayrıca hanehalkı reisinin çocuğu, gelini, torunu veya diğer akrabası olan kadınlar daha fazla iş aramaktadır. Daha yüksek tahsil görmüş grupta yer alan kişiler de, daha düşük tahsil görenlere nazaran daha çok iş aramaktadır. Ayrıca erkeklerden boşanmış veya eşi ölmüş olanlar daha fazla iş aramaktadır.

Elde edilen beşinci ve son düzey ağaç diagramına göre erkekler ve kadınlar için ilk bölümlenmenin okul değişkeni olarak gösterilen en son bitirilen okul kategorilerine göre yapıldığını görmekteyiz. İş arama veya iş aramama durumu üzerinde en etkili değişkenin en son bitirilen okul değişkeni olduğu söylenebilir. Kadınlar için iş arama değişkeni üzerinde etkili ikinci değişken hanehalkı reisine olan yakınlıkları iken, erkekler için ikinci önemli değişkenin medeni durum olduğu elde edilen beşinci düzey ağaç diagramından rahatlıkla söylenebilir. Aynı veri grubu ve değişkenler için CHAID algoritması da uygulanmıştır. Fakat CHAID iki değerli ağaç yapısı türetmediğinden yorum yapabilmek oldukça güç olmuştur.

2002 hanehalkı işgücü anketinde yer alan başka değişkenler de mevcuttur. Bu değişkenlerin de analize alınması gerektiği söylenebilirse de, değişken sayısı arttıkça, elde edilecek ağacın yapısı bozulmakta ve mevcut verilerden bir kural üretilemediğinden, bir sonuç elde edilemediği denenerek gözlenmiştir.

Kaynaklar

Ahola Jussi ve Rinta-Runsala Esa (2001), "Data Mining Case Studies İn Customer Profiling", VTT Information Technology Research Report, TTE1-2001-29.

Akpınar Haldun (2000), "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İ.Ü. İşletme Fakültesi Dergisi, C:29, S: 1/Nisan 2000, ss. 1-22.

An Introduction To Classification and Regression Tree (CART) Analysis, İnternet Adresi: <http://www.saem.org/download/lewis1.pdf >, Erişim Tarihi: 24.02.2004.

Apte Chidanand ve Weiss Sholom (1997), "Data mining with decision trees and decision rules", Future Generation Computer Systems, 13, ss.197-210.

Answer Tree 3.0 User's Guide (2001), SPSS Inc., USA.

Classification and Regression Trees: An Introduction, İnternet Adresi, <http://www.ifpri.org/themes/mp18/techguid/tg03.pdf >. Erişim Tarihi: 24.02.2004.

DMS Tutorial, İnternet Adresi :<dms.irb.hr/tutorial/tut_dtrees.php>, Erişim Tarihi:05.05.2003.

Hand David, Mannila Heikki ve Smyth Padhraic (2001), Principles of Data Mining, MIT Press, USA.

Muata Kweku ve Bryson Osei (2003), "Evaluation of decision trees:a multi criteria approach", Computers&Operations Research, article in press.

Roiger Richard J. ve Geatz Michael W. (2003), Data Mining A Tutorial-Based Primer, Addison Wesley, USA.