Siirt Üniversitesi
İktisadi ve İdari Bilimler Fakültesi
İktisadi Yenilik Dergisi, *Cilt:* 3, *Sayı:* 1, *Temmuz* 2015

Sınıflandırma Ağacı Yaklaşımının R ile Çözümlenmesi: Kayıp Çocuk Profil Örneği

Levent Terlemez®

Anadolu Üniversitesi

ÖZET

Kayıp çocuk sorunu, tüm dünyada olduğu gibi ülkemizde de yaşanan önemli sorunlardan birisidir. Bu sorun, toplumdaki hızlı değişim sürecinden de etkilenmekte ve giderek büyüyen bir sorun haline gelmektedir. Son yıllarda gözlemlenen bu büyüme toplumda da tedirginliğe yol açmaktadır. Bu yüzden kayıp çocuk profilinin ortaya çıkarılması büyük önem taşımaktadır. Bu çalışmada, sınıflandırma ve regresyon ağaçları kullanılarak kayıp çocukların sınıflandırılmasına çalışılmıştır. Kullanılan veri seti, cinsiyet, doğum yılı, kayıp yılı, boy, kilo, ten rengi, göz rengi, saç rengi değişkenlerinden oluşmaktadır. Doğum yılı ve kayıp yılı değişkenleri doğrudan kullanılmamış, bunun yerine, bu değişkenlerden kayıp yaşı değişkeni türetilmiştir. Sınıflandırma ağacı ile gini saflık ölçüsü kullanılarak elde edilen sınıflandırma sonucunda, 7 farklı sınıf elde edilmiştir. Bu sınıflardan, 14 yaşından büyük eşit ve 1.11-1.50,1.51-1.60 ve 1.61-1.70 boy aralıklarındaki kız çocuklarının oluşturduğu sınıf, en baskın sınıf olarak ortaya çıkmıştır ve tüm kayıp çocukların %62'sinin bu sınıfa ait olduğu gözlemlenmiştir.

Anahtar Kelimeler: *CART, Rpart, Sınıflandırma, Kayıp Çocuklar*
JEL Sınıflandırması: C87, Y10, J13

® Yrd.Doç.Dr. Levent TERLEMEZ, Anadolu Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Yunus Emre Kampüsü, Eskişehir/TÜRKİYE, lterlemez@anadolu.edu.tr.

1. GİRİŞ

Kayıp çocuk tanımı, Emniyet Genel Müdürlüğü tarafından, "veli, vasi veya yakınları tarafından nerede olduğu veya akıbeti bilinmeyen, vesayeti veya koruması altında bulunduğu kurumu izinsiz terk eden veya izinli ayrılsa bile kuruma geri dönmeyen ve hakkında polise kayıp müracaatı yapılmış 18 yaşını tamamlamamış kişi" olarak tanımlanmaktadır (EGM, 2014; CİGM, 2014). Başbakanlık İnsan Hakları Başkanlığı'nın 2007 yılında hazırlamış olduğu Kayıp Çocuklar Raporunda ise, "ailesinin bilgisi dışında herhangi bir nedenle evden uzaklaşmış, kaçmış, kaçırılmış ve bu nedenlerle hayatı tehlike altında olan, kendisinden haber alınamayan 0-18 yaş grubu çocuk" olarak ele alınmıştır (BİHB, 2007).

Bir çocuğun kaybolma nedenleri incelendiğinde ise, macera yaşama, zengin veya ünlü olma hayali gibi çeşitli nedenlerden dolayı kendi isteğiyle kaçma; yoksulluk, aile içi şiddet, istismar, kötü ebeveynlik, çocuk işçilik, dilencilik, ideolojik veya suç işleme amaçları için kullanma; cinsel sömürü, zengin ülkelerde evlat edindirme gibi birçok neden sayılabilir (BİHB, 2007; UNICEF, 2011).

Toplum yapısındaki hızlı değişim süreci, sosyal alanda büyük değişimlerin yaşanmasına sebep olmaktadır. Bu durum, toplumu kontrol eden yasal ve sosyal sistemlere, yetersiz kalma ve yaşanan değişime ayak uyduramama olarak yansımaktadır. Yansımanın sonucu ise, sosyal çevrede istenmeyen şartların ortaya çıkması, hukuksal ve sosyal sorunlar meydana gelmesi şeklinde görünmektedir. Bu sorunlardan bir tanesi de kayıp çocuk sorunudur. Bu önemli konuda, Başbakanlık İnsan Hakları Başkanlığı tarafından 2008 yılında "Kayıp Çocuklar Raporu" ve İçişleri Bakanlığı tarafından 2009 yılında "Kayıp Çocuk Rehberi" hazırlanmıştır (Aydın, 2010). TBMM içerisinde kurulan Meclis Araştırma Komisyonu ise 2010 yılında, Kayıp Çocuk Sorunu ve İlgili Risk Faktörlerinin İncelenmesi amacıyla saha araştırması gerçekleştirmiştir.

Bu raporda, kayıp çocuk sorununun dünya genelinde ve ülkemizde çocukların mağdur olduğu sorunların önemli başlıklarından bir tanesi olduğu vurgulanmış, kayıp çocuklar konusunda tam, doğru, sürekli ve düzenli veri toplamayla ilgili bazı zorlukların yaşanmakta olduğu ifade edilmiştir. Bu zorluklar, doğum kayıtlarının tamamına ulaşılamaması, çocuk kayıplarının resmi makamlara zamanında bildirilmemesi, eksik bildirilmesi, kayıp çocuklar hakkındaki toplanan çeşitli bilgileri içeren verinin güvenilir olmaması, elektronik ortamda toplanmaması ve/veya standart yapıda olmaması gibi nedenler olarak belirtilmiştir. Belirtilen bu nedenlerin de, konu hakkında durum tespitinin ve geleceğe yönelik tahminlerin yapılmasını ve uygun mücadele yönteminin seçilmesini zorlaştırdığı, buna bağlı olarak ta, yasal kurumlardan elde edilen istatistiklerin birleştirilmesini, paylaşılmasını ve çeşitli nedenlerden dolayı doğru hesaplanamamasına neden olduğu irdelenmiştir. Çocuk kayıpları nedenleri arasında küresel bir unsur olarak gözlenen insan ticareti konusunda ise, sistematik bilgi toplamayı sağlayacak yerel ya da uluslararası merkezî mekanizmaların bulunmadığı da belirtilmiştir (Aydın, 2010).

Ek olarak kayıp çocuklara ait uluslararası veriler çoğunlukla cinsel sömürü için ülke sınırları dışına kaçırılan kadın ve kız çocuklarıyla sınırlı olduğu, ulusal verilerin ise, esas olarak kolluk kuvvetlerinin kayıp çocuğu aramasında yol göstermeye yönelik olduğu ifade edilmiştir. Arkadaş, aile ve sosyal çevresi ile ilgili özellikler, sağlığıyla ilgili sorunlar vb. çocuğun kaybı ile ilgili olabilecek faktörler hakkında yeterince bilgi vermediği, özellikle yurtlarda yaşayan

çocuklarda daha sık olarak gözlemlenen kayıp olaylarında, çocuğa ait dosyalarda detaylı sosyal incelemeler olsa dahi; bu veriler rutin hizmet sürecinde bir araya getirilmediği, kayıp çocuk profilini tanımlamak, kayıp açısından yüksek riskli çocukları erken dönemde tespit etmek, kayıp çocuk sorununa ait ileriye dönük tahminler yapmak konusunda kullanılamadığı belirtilmiştir (TBMM, 2010).

Bu çalışmada, Emniyet Genel Müdürlüğü Asayiş Dairesi Başkanlığı internet sitesinden 2010 yılında elde edilmiş kayıp çocuklara ait 8 özelliğe göre kayıp statüsünde bulunan 448 çocuğun sınıflandırılması işlemi açık kaynaklı bir yazılım olan R kullanılarak çözümlenmesi gerçekleştirilmiştir. Sınıflandırma için sınıflandırma ve regresyon ağaçları kullanılmıştır.

Sınıflandırma ve regresyon ağacının uygulanması, R yazılımı ve bu yazılımın Rpart paketi kullanılarak yapılmıştır. R, UNIX platformları, Windows ve Mac OS gibi çeşitli işletim sistemleri üzerinde çalışan, içerdiği paket sistemi ile doğrusal ve doğrusal olmayan modelleme, istatistiksel testler, sınıflandırma, kümeleme gibi oldukça geniş istatistiksel ve grafiksel tekniklerin kullanımını sağlayan, ücretsiz istatistiksel hesaplama ve grafikler için yazılım ortamıdır (CRAN, 2015; R Core Team, 2014). Rpart (recursive partition - rpart) ise, ticari marka olan CART (Classification And Regression Trees) programlarının bir uyarlamasıdır. Rpart paketi, iki aşamalı bir izlek kullanarak çok genel bir yapının sınıflandırma veya regresyon modellerini kurar; elde edilen modeller ise ikili ağaçlar şeklinde temsil edilebilir (Therneau vd., 1997).

Bu çalışmanın ikinci bölümünde, araştırmada kullanılan sınıflandırma ağacı metodu hakkında bilgi verilmiştir. Üçüncü bölümde ise, ilgili veri setine ilişkin bilgiler ile elde edilen sonuçlar sunulmuştur. Son bölümde ise araştırma sonuçları vurgulanmıştır.

2. SINIFLANDIRMA VE REGRESYON AĞAÇLARI

Özyinelemeli bölümleme (recursive partition), kategorik (sınıflandırma ağacı) veya sürekli (regresyon ağacı) bir sonucun kestirimi için görselleştirmesi, incelenmesi, anlaşılırlığı ve en önemlisi ifade edilebilirliği kolay olan karar kuralları geliştirirken, bir veri seti yapısının incelenmesinde yardımcı olan çok değişkenli bir metottur (Quick-R, 2014; Graham, 2011).

2.1. Sınıflandırma Ağaçları

Bir sınıflandırma ağacı niteliksel yanıtların kestirimi için kullanılır. Yani, her gözlemin, ait olduğu bölgedeki eğitim gözlemlerinin en sık görülen sınıfına aitliğinin kestirimidir. Ağacın oluşturulması iki aşamadan oluşmaktadır. Birinci aşamada, önce veriyi ikiye en iyi şekilde bölen değişken bulunur, böylelikle kök düğüm elde edilmiş olur. Veri ikiye bölündükten sonra, bu süreç alt gruplara, yani iç düğümlere ayrı ayrı uygulanır. Bölünme işlemi, en küçük alt grup büyüklüğe ulaşılana kadar veya herhangi bir iyileştirme yapılamayana kadar yinelenerek, yani yaprak düğüme ulaşılana kadar devam eder. Elde edilen her bir iç düğüm ile bağlantılı bir test hangi dalın takip edileceğini belirlerken, yaprak düğümler kararları içerir. Elde edilen model çok karmaşık bir model olacağından, ikinci aşamada, çapraz doğrulama kullanılarak tüm ağaç geriye doğru kırpılır (James vd., 2013; Therneau ve Atkinson, 1997; Williams, 2011).

Sınıflandırma ağacında, ait olduğu düğümdeki her gözlemin, düğümdeki eğitim kümesinin en sık gözlemlenen sınıfının hangisine ait olduğu kestirilir. Sınıflandırma sonucunda, elde edilen belirli bir düğüme ilişkin sınıf kestirimi ile birlikte, düğüme düşen eğitim kümesinin içindeki sınıf oranı da önem arz etmektedir.

Sınıflandırma ağacının büyümesi öz yinelemeli ikiye ayırma ile gerçekleşir. İkiye ayırma işlemi için kullanılan ölçüt sınıflandırma hata oranıdır. Amaç, belli bir düğümdeki bir gözlemi, o düğümde yer alan gözlemlerinin en sık görülen sınıfına atamak olduğundan, sınıflandırma hatası o düğümdeki gözlemlerin, en sık gözlemlenen sınıfa ait olmayan gözlemlere oranı olarak ifade edilir. $\hat{p}_{mk}$, k. sınıftan gelen m. düğümdeki eğitim kümesinin oranını ifade ederken, sınıflandırma hatası E;

$$E = 1 - \max_k (\hat{p}_{mk}) \tag{2.1}$$

ile gösterilir. Ancak, yeterli duyarlılığa sahip olmadığı için, uygulamada Gini indeksi, çapraz entropi ve diğer bir alternatif olan twoing indeksinden birisi kullanılmaktadır.

Gini indeksi, K sınıfları arasındaki toplam varyansın bir ölçüsü olarak;

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.2}$$

şeklinde tanımlanır. Küçük değerleri, düğümün baskın olarak tek bir sınıftan gözlemleri içerdiğini belirttiğinden, düğüm saflığının bir ölçüsü olarak ta adlandırılır.

Çapraz entropi;

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{2.3}$$

şeklinde tanımlanır. Çapraz entropi'de, Gini indeksi gibi, küçük değerlerinde m. düğümün saflığını ifade edecektir (James vd., 2013).

Twoing indeksi ise;

$$\varphi = \frac{\hat{p}_L \hat{p}_R}{4} \left[\sum_{k=1}^{K} \left| \hat{p}_{km_L} - \hat{p}_{km_R} \right| \right]^2 \tag{2.4}$$

şeklinde tanımlanır. Twoing indeksi, büyük değerlerinde, etkin bir bölümlemeyi ifade edecektir (Balakrishnan vd., 2009).

Sınıflandırma ağacı oluşturulurken, belirli bir düğümün saflığını değerlendirirken genellikle gini indeksi veya çapraz entropi kullanılır. Bu dört ölçütten herhangi biri, sınıflandırma ağacının budanması esnasında tercih edilebilir. Ancak budama işleminde, budanmış sınıflandırma ağacının kestirim doğruluğu söz konusu ise, sınıflandırma hatası öne çıkmaktadır (James vd., 2013).

2.2. Yazında Yapılmış Çalışmalar

Bu çalışmada kullanılan yöntemle ilgili yapılmış yazında çok sayıda çalışma mevcuttur. Örneğin tıpta multiple sclerosis için Li ve Schwartz (2012), tip I diyabetler için Sampurno (2006); coğrafyada arazi yapısı ve ağaç çeşitleri için Sasaki vd. (2012), ormanlar için Wen vd. (2009), yeraltı suları için Spruill vd. (2002); güvenlikte erken uyarı için Koon ve Petscher (2015); toplum bilimde Quatch vd. (2015); istatistikte Azam vd. (2014); mühendislikte Archer (2010) bu çalışmada kullanılan yöntemin uygulamalarıdır.

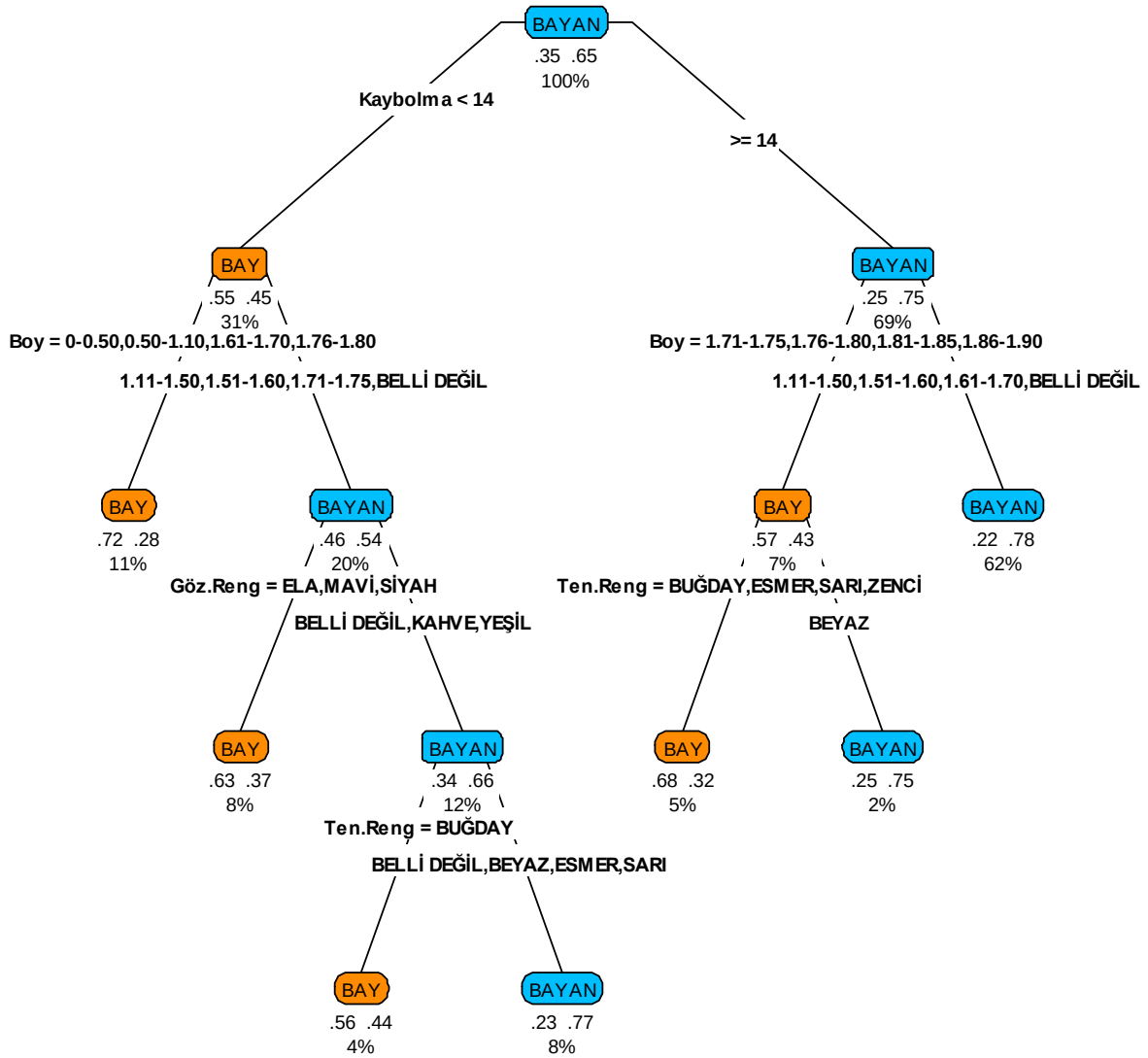
3. UYGULAMA

Bu çalışmada, 8 özelliğe göre kayıp statüsünde bulunan 448 kayıp çocuğun sınıflandırılması işlemi açık kaynak kodlu R yazılımı kullanılarak gerçekleştirilmiştir. Belirlenen 7 özellik sırasıyla kayıp çocuğun cinsiyeti, doğum tarihi ve kaybolma tarihi özelliklerden derlenmiş kayıp yaşı, boyu, kilosu, ten rengi, göz rengi ve saç rengidir. Bu özellikler, İstanbul Emniyet Müdürlüğü Asayiş Dairesi Başkanlığı İnternet Sitesinden 2010 tarihinde elde edilmiştir. Değişkenler Tablo 3.1'deki gibidir.

Değişken	Ölçüm Düzeyi	Rolü
Cinsiyeti	Nominal	Hedef
Kaybolma Yaşı	Oransal	Girdi
Boy	Sıralayıcı	Girdi
Kilo	Sıralayıcı	Girdi
Ten Rengi	Sınıflayıcı	Girdi
Göz Rengi	Sınıflayıcı	Girdi
Saç Rengi	Sınıflayıcı	Girdi

Tablo 3.1 Sınıflandırmada kullanılan değişkenler.

Şekil 3.1 Cinsiyeti = Kaybolma Yaşı + Boy + Kilo + Ten Rengi + Saç Rengi + Göz Rengi modeli için sınıflandırma ağacı



Sınıflandırma ağacı, rpart fonksiyonu aracılığı ile Tablo 3.2'de verilen parametreler kullanılarak oluşturulmuştur.

Parametre	Açıklama
formula	Herhangi bir etkileşim içermeyen, bağımlı değişkeni de barındıran sınıflandırma modelinin formülü
method	Bağımlı değişken, süreç içerikli ise "poisson", kategorik içerikli ise "class", sağkalım içerikli ise "exp", diğer durumlarda "anova" değerini alabilir.
data	Formülde yer alan değişkenlerin yorumlanabilmesi için, analiz edilecek veri setini belirleyen isteğe bağlı parametre.

Tablo 3.2 Rpart fonksiyonunda kullanılan parametreler.

Sınıflandırma için uygun görülen model,

$$Cinsiyeti = Kaybolma\ Yaşı + Boy + Kilo + Ten\ Rengi + Saç\ Rengi + Göz\ Rengi \quad (3.5)$$

şeklindedir ve rpart fonksiyonuna Tablo 3.3'deki gibi verilmiştir.

```
rpart(data=veri1,model=Cinsiyeti~Kaybolma.Yaşı+Boy+Kilo
                                +Ten.Rengi+Saç.Rengi+Göz.Rengi,method="class")
```

Tablo 3.3 Sınıflandırma ağacı için rpart fonksiyonu parametreleri.

Rpart fonksiyonu, bir bölünmenin saflığının ölçülmesi için çapraz entropi ve gini indeksini kullanabilmektedir, varsayılan olarak gini indeksi seçilidir. Bu modele göre sınıflandırma ağacı Şekil 3.1'de gösterildiği gibi elde edilmiştir.

Rpart fonksiyonu, sınıflandırma modelini boy, göz rengi, ten rengi ve kaybolma yaşı değişkenlerini kullanarak oluşturmuştur. İlk ayırıcı değişken kaybolma yaşı olarak ortaya çıkmaktadır ve 14 yaşından küçük kayıp çocukları sol dala, 14 yaş ve daha büyük kayıp çocukları sağ dala ayırmıştır. Sol dalda yer alan ilk düğümdeki 14 yaşından küçük yaşta kaybolan çocuklar, tüm kayıp durumundaki çocukların %31'ini oluşturmaktadır. Erkek çocukların bu düğüm içindeki (sınıf) oranı 0.55, kız çocukların oranı ise 0.45'tir.

Bu düğümden elde edilen ilk yaprakta (sınıf), boy değişkeni ayırmada etkili olmaktadır ve 0.00-0.50, 0.50-1.10, 1.61-1.70 ile 1.76-1.80 boy aralıklarında olan çocuklar bu sınıfa dâhil edilmiştir. Bu sınıftaki kayıp çocuklar tüm kayıp durumundaki çocukların %11'ini oluşturmaktadır. Bu sınıfa dâhil olan kayıp çocukların sınıf olasılıkları erkek çocuk için 0.72, kız çocuk için 0.28 olarak ortaya çıkmaktadır. Bu sonuca göre sınıflandırma kuralı, eğer yaşı 14'ten küçük ve boyu 0.00-0.50, 0.50-1.10, 1.61-1.70, 1.76-1.80 boy aralıklarında ise cinsiyeti erkek şeklinde olacaktır. Bu kuralın destek sayısı 36'dır.

Elde edilen sınıfların içerisinde en büyük öneme sahip olan sınıf ise, yaşı 14 veya daha büyük, 1.11-1.50,1.51-1.60 ve 1.61-1.70 boy aralıklarındaki kız çocuklarının oluşturduğu sınıftır. Bu sınıf, tüm sınıflar içindeki en büyük sınıf olmaktadır ve tüm kayıp çocukların %62'si bu sınıfa dâhil olmuştur. Bu sınıftaki bir çocuğun kız olma olasılığı 0.78, erkek çocuk olma olasılığı ise 0.22 olarak elde edilmiştir. Buna sonuca göre, elde edilen sınıflandırma kural, eğer yaşı 14 veya daha büyük ve 1.11-1.50,1.51-1.60, 1.61-1.70 boy aralıklarında ise cinsiyeti kız şeklinde olacaktır. Bu kuralın destek sayısı 217'dir.

Sınıflandırma ağacından elde edilen tüm sınıflandırma kuralları Tablo 3.4'de ki gibi elde edilmiştir.

Sınıf	Kural	Destek Oranı
1	Eğer Kaybolma Yaşı < 14 ise ve Eğer Boy=0-0.50 veya 0.50-1.10 veya 1.61-1.70 veya 1.76-1.80 ise Cinsiyeti=BAY	0.080
2	Eğer Kaybolma Yaşı < 14 ise ve Eğer Boy=1.11-1.50 veya 1.51-1.60 veya 1.71-1.75 veya BELLİ DEĞİL ise ve Eğer Göz Rengi=ELA veya MAVİ veya SİYAH ise Cinsiyeti=BAY	0.053
3	Eğer Kaybolma Yaşı < 14 ise ve Eğer Boy=1.11-1.50 veya 1.51-1.60 veya 1.71-1.75 veya BELLİ DEĞİL ise ve Eğer Göz Rengi=KAHVE veya YEŞİL veya BELLİ DEĞİL ise ve Eğer Ten Rengi=BUĞDAY ise Cinsiyeti=BAY	0.022
4	Eğer Kaybolma Yaşı < 14 ise ve Eğer Boy=1.11-1.50 veya 1.51-1.60 veya 1.71-1.75 veya BELLİ DEĞİL ise ve Eğer Göz Rengi=KAHVE veya YEŞİL veya BELLİ DEĞİL ise ve Eğer Ten Rengi=BEYAZ veya ESMER veya SARI veya BELLİ DEĞİL ise Cinsiyeti=BAYAN	0.060
5	Eğer Kaybolma Yaşı >= 14 ise ve Eğer Boy=1.11-1.50 veya 1.51-1.60 veya 1.61-1.70 veya BELLİ DEĞİL ise Cinsiyeti=BAYAN	0.484
6	Eğer Kaybolma Yaşı >= 14 ise ve Eğer Boy=1.71-1.75 veya 1.76-1.80 veya 1.81-1.85 veya 1.86-1.90 ise ve Ten Rengi=BEYAZ ise Cinsiyeti=BAYAN	0.013
7	Eğer Kaybolma Yaşı >= 14 ise ve Eğer Boy=1.71-1.75 veya 1.76-1.80 veya 1.81-1.85 veya 1.86-1.90 ise ve Ten Rengi=BUĞDAY veya ESMER veya SARI veya ZENCİ ise Cinsiyeti=BAY	0.033

Tablo 3.4 Sınıflandırma ağacından elde edilen sınıflandırma kuralları.

4. SONUÇ VE DEĞERLENDİRME

Kayıp çocuk problemi tüm dünyada olduğu gibi, ülkemizde de yaşanmaktadır. Bir ülkenin en önemli zenginliklerinden birisi olan çocukların korunması, temel gereksinimlerinin karşılanması, karşılaştıkları sorunların çözümlenmesi, o ülkenin geleceği için yapacağı en iyi yatırım olacaktır. Bu yatırıma karşına çıkan engellerden bir tanesi de kayıp çocuk sorunudur.
Bu sorunun nedenlerinin ve alınacak önlemlerin belirlenmesi, kayıp statüsündeki çocuklarda gözlemlenen özelliklerin tanımlanması ve kaybolma riski yüksek olan çocukların belirlenebilmesi, yüksek risk taşıyan bir çocuk için kaybolma riskini azaltacak önlemlerin alınmasında, kaybolduğunda arama işlemlerinin nasıl yapılacağı, verilecek desteğin nasıl olacağı ve benzeri diğer konularda hazırlıklı olunmasını sağlayacaktır.
Bu çalışmada, İstanbul Emniyet Müdürlüğü internet sitesinde takibi sağlanan kayıp çocuk veri tabanında sunulan 448 kayıp çocuğun kayıt altına alınmış özelliklerinden 8 tanesi kullanılarak bir sınıflandırmalarının yapılmasına çalışılarak, kayıp statüsündeki çocuklarda gözlemlenen özelliklerin tanımlanabilmesi ve kaybolma riski yüksek olan çocukların belirlenebilmesi için farklı bir bakış açısı getirilmeye çalışılmıştır.

Bu amaçla, sınıflandırma tekniklerinden bir tanesi olan sınıflandırma ve regresyon ağaçları metodundan yararlanılmıştır. Sınıflandırma ve regresyon ağaçları görselleştirmesi, incelenmesi, anlaşılırlığı ve en önemlisi ifade edilebilirliği kolay olan çok değişkenli bir metot olarak kendini göstermektedir.

Yapılan uygulama sonucunda, sınıflandırma ağacı kayıp çocukları 7 sınıfa ayırmıştır. Bu sınıflardan en önemlisi, 14 veya daha büyük yaşta, 1.11-1.50,1.51-1.60 ve 1.61-1.70 boy aralıklarındaki kız çocuklarının oluşturduğu sınıftır. Bu sınıf, tüm sınıflar içindeki en büyük sınıf olmaktadır ve tüm kayıp çocukların %62'si bu sınıfa dâhil olmuştur. Bu sınıftaki bir çocuğun kız olma olasılığı 0.78, erkek çocuk olma olasılığı ise 0.22 olarak elde edilmiştir.

Ancak, sağlanan veri kaynağında sunulan bazı sürekli değişkenlerin, nitel ölçekte sunulması, sağlanan özelliklerin sayısının az olması, sınıflandırmanın detaylı ve anlamlı olmasına fazla izin vermemektedir. Kayıp çocukla ilgili kayıt altına alınacak özelliklerin sayısı ve düzeni iyileştikçe, sınıflandırma sonuçlarının daha detaylı ve anlamlı olacağı açıktır.

KAYNAKÇA

Archer, K. J. (2010). rpartOrdinal: An R Package for Deriving a Classification Tree for Predicting an Ordinal Response. *Journal of Statistical Software*, 34 (7), 1–17.

Aydın, H. (2010), Yerel Yönetimlerin Sorumlulukları Çerçevesinde Türkiye'de Kayıp Çocuklar Sorunu. *İdarecilerin Sesi*, 139, 25-30.

Azam, M, M. Aslam ve K. P. Pfeiffer (2014). Three Steps Strategy to Search for Optimum Classification Trees. *Communications in Statistics - Simulation and Computation*, DOI:10.1080/03610918.2013.867991.

Balakrishnan, N., S. Kotz, C. Read, B. Vidakovic ve N. L. Johnson (2006). *T: Twoing Index*. Encyclopedia of Statistical Sciences, 2nd Ed. A John Wiley & Sons, Inc.

BİHB (2007). *Kayıp Çocuklar Raporu*. Başbakanlık İnsan Hakları Başkanlığı. http://www.tihk.gov.tr/www/files/Kayip_cocuklar_raporu_4_8_2008.pdf (10.03.2015).

CİGM (2014). *Kayıp Şahıslar Yönergesi*. Adalet Bakanlığı Ceza İşleri Genel Müdürlüğü. http://www.cigm.adalet.gov.tr/duyurular/2014/kayipsahisyonek.pdf (10.03.2015).

CRAN (2015). *The Comprehensive R Archive Network*. http://cran.r-project.org/ (10.03.2015).

EGM (2014). Emniyet Genel Müdürlüğü Asayiş Dairesi Başkanlığı. http://www.asayis.pol.tr/Sayfalar/kayip_ve_aranan_sahislar.aspx (10.03.2015).

Koon S. ve Y. Petscher (2015). *Comparing Methodologies for Developing an Early Warning System: Classification and Regression Tree Model Versus Logistic Regression*. Applied Research Methods http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=REL2015077 (10.03. 2015).

Li, Y. ve C. E. Schwartz (2012). Erratum to: Data mining for response shift patterns in multiple sclerosis patients using recursive partitioning tree analysis. *Quality of Life Research*, 21 (1), 1543–1553.

James, G., D. Witten, T. Hastie ve R. Tibshirani (2013), *An Introduction to Statistical Learning with Applications in R*, Springer.

Quatch, A., J. Symanzik ve N. Forsgren (2015). Soul of the Community: An Attempt to Assess Attachment to a Community. *Journal of Computational Statistics*, yayın kabul.

Quick – R (2014). *Tree-Based Models*. http://www.statmethods.net/advstats/cart.html (10.03.2015).

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/ (10.03.2015).

Sasaki, T., J. Imanishi, K. Ioki, Y. Morimoto ve K. Kitada (2012). Object-based classification of land cover and tree species by integrating airborne LiDAR and high spatial resolution imagery data. *Landscape and Ecological Engineering*, 8 (2), 157–171.

Sampurno, F. (2006). *Identifying risk factors associated with new onset cardiovascular disease in patients with type I diabetes using Classification Tree*. Melbourne: The University of Melbourne.

Spruill, T. B., W. J. Showers ve S. S. Howe (2002). Application of Classification-Tree Methods to Identify Nitrate Sources in Ground Water. *Journal of Environmental Quality*, 31, 1538–1549 DOI:10.2134/jeq2002.1538.

Stephen Milborrow (2014). rpart.plot: Plot rpart models. *An enhanced version of plot.rpart*, R package version 1.4-5. http://CRAN.R-project.org/package=rpart.plot.

TBMM (2010). *Kayıp Çocuklar Basta Olmak Üzere Çocukların Mağdur Olduğu Sorunların Araştırılarak Alınması Gereken Önlemlerin Belirlenmesi Amacıyla Kurulan Meclis Araştırması Komisyonu Raporu*. http://www.tbmm.gov.tr/sirasayi/donem23/yil01/ss589.pdf (10.03.2015).

Therneau, T. M. ve E. J. Atkinson (2015), *An Introduction to Recursive Partitioning Using the RPART Routines*. Mayo Foundation. http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (10.03.2015).

UNICEF (2011). *Türkiye'de Çocukların Durumu Raporu*. http://unicef.deparyazilim.com/files/bilgimerkezi/doc/sitan-tur-2011.pdf (10.03.2015).

Wen, L., J. Ling, N. Saintilan ve K. Rogers (2009). An investigation of the hydrological requirements of River Red Gum (Eucalyptus camaldulensis) Forest, using Classification and Regression Tree modelling. *Ecohydrology*, 2 (2), 143–155.

Williams, G. (2011). *Data Mining with Rattle and R*. Springer.

Classification Tree Approach via R: A Case Sudy for Missing Child Profiling

Levent Terlemez

Anadolu University

ABSTRACT

Missing children problem is one of the major problems in Turkey as well as all over the world. This problem is affected by the rapid change process in society and is becoming a growing problem. The increase in numbers observed in recent years lead to a concern in the society. Therefore, profiling missing children is of utmost importance. In this study, classification and regression trees are employed to classify missing children. Data set for classification of missing children consists of sex, height, weight, skin color, eye color, hair color and a derived variable, missing age from birth year and missing year. Seven different classes are obtained from classification tree by using Gini purity measure. One of dominant class consists of girls aged greater or equal to 14 and in 1.11-1.50, 1.51-1.60 and 1.61-1.70 height intervals, which covers 62% of the whole missing children subjected to this study.

Keywords: *CART, Rpart, Classification, Missing Children*
JEL Classifications: C87, Y10, J13