Breiman Algoritması Kullanılarak Homojen Alt Grupların Belirlenmesi: Bir Uygulama

Özge AKŞEHİRLİ, Handan ANKARALI, Şengül CANGÜR, Mehmet Ali SUNGUR

Düzce Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Düzce, Türkiye
ozge_yilmaz85@hotmail.com, hankarali@yahoo.com, sengulcangur@duzce.edu.tr, malisungur@yahoo.com
(Geliş/Received: 28.06.2013; Kabul/Accepted: 29.01.2014)

DOI: 10.12973/bid.2013

Özet – Breiman, birçok verinin birbirine yakın olarak toplandığı "yüksek yoğunluklu" alanları bularak verilerin kümelenebileceğini söylemiştir. Bu çalışmada, Breiman'ın kümeleme algoritmasının işleyiş adımları tanıtılarak bir veri seti üzerinde uygulama adımlarının gösterilmesi ve sonuçlarının yorumlanması amaçlanmıştır. Uygulama bölümünde, hastaneye gece yeme sendromu şikâyetiyle başvuran 433 kişiye ilişkin sosyo-demografik ve klinik özellikler kullanılmıştır. Veri setinde olabilecek kümelerin ortaya konmasında, CART algoritmasından yararlanılmıştır. Elde edilen optimum ağaçta toplam 31 karar noktası bulunmuş ancak bunların 14' ünde yer alan deneklerin kendi içinde kümelenme gösterdiği belirlenmiştir. Çalışmaya alınan kişilerin 350'si oluşturulan 14 küme içine girmiş ve bunların 273 (%78)'ü klinik olarak gece yeme alışkanlığı yoktur tanısı almıştır. Elde edilen 14 kümenin 12'sinde yer alan kişilerin ağırlıklı olarak gece yeme alışkanlığı yok tanısı alanlardan oluştuğu ve bu sonuca göre, bu veri setinden elde edilen kümelerin, genel olarak gece yeme alışkanlığı olmayan bireyleri ayırt edebildiği söylenebilir. Sonuç olarak, hedef veya bağımlı değişkenin bilinmediği durumlarda, veri setinde var olan homojen alt grupların belirlenmesinde, danışmansız öğrenme yöntemlerinden biri olan kümeleme analizinin uygulanması için değişkenlerin dağılım şekli ve tipinden etkilenmeyen Breiman algoritması etkin bir şekilde kullanılabilir.

Anahtar kelimeler – Veri madenciliği, danışmansız öğrenme, kümeleme analizi, Breiman algoritması, CART

Determination of Homogeneous Subgroups Using Breiman's Algorithm: An Application

Abstract – Breiman said that the data can be cluster by finding "high density" areas where lots of data collected in close proximity to each other. In this study, it was aimed to introduce operation steps of Breiman's clustering algorithm, to show application steps of the method using a data set and to interpretation of the results. In the practice section of the study, socio-demographic and clinical characteristics of 433 individuals who admitted to the hospital with complaints of night eating syndrome, were used. CART algorithm was used to produce clusters that may be in the data set. In the obtained optimal tree, 31 decision points were found totally, but it was determined that the subjects located 14 of 31 decision points were clustered within itself. 350 of the individuals included in the study, entered into these created 14 clusters and 273 (78%) of them were diagnosed clinically as there is no habit of eating at night. It can be said that individuals involved in the 12 of 14 obtained clusters have diagnosis of there is no habit of eating at night. And according to this result, we can say that the clusters obtained from this data set, can be distinguish individuals who have not habit of night eating. As a result, when the target or dependent variable is unknown, Breiman's algorithm, which is not affected by the shape of the distribution and type of the variables, can be used effectively.

Keywords – Data mining, unsupervised learning, cluster analysis, Breiman algorithm, CART

1. GİRİŞ (INTRODUCTION)

Veri tabanlarındaki verilerin gruplar veya kümeler altında toplanarak, benzer özelliklere sahip nesnelerin bir araya gelmesini sağlayan kümeleme algoritmaları veri madenciliği alanında oldukça büyük öneme sahiptir. Kümeleme analizi, homojen alt grupların belirlenmesi, daha ileri veri analizleri için bir ön değerlendirme veya örüntü tanıma gibi birçok alanda oldukça sık kullanılmaktadır [1].

Kümeleme analizi, veri matrisinde bulunan ve doğal sınıfları kesin olarak bilinmeyen verileri veya değişkenleri, bunlar arasındaki çeşitli benzerlik ya da farklılıklara dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplara ayırmaya yardımcı olan çok değişkenli yöntemler grubunda da incelenir [2].

Kümeleme analizinde oluşturulan gruplar, kendi içlerinde homojen, kendi aralarında ise heterojen olmalıdır. Bu durumda kümeleme işlemi başarılı olursa, bir geometrik çizim yapıldığında veriler küme içerisinde birbirine çok yakın, elde edilen kümeler ise birbirinden uzak olacaktır [3].

Kümeleme analizi denetimsiz öğrenme metoduna dayalı bir yöntemidir. Denetimsiz öğrenmede ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareketle sınıfların tanımlanması yapılmaktadır. Kısaca denetimsiz öğrenmede bağımlı değişken olmadığından model kurmak için bağımsız değişkenlerden yararlanılır ve önceden belirlenmiş sınıflar olmadığı için, model verilerden üretilir. Yani, nesnelerin kümelere atanması için önceden belirlenmiş kurallar, herhangi bir fonksiyon veya bir model yoktur [4], [5].

Temel olarak, hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleri olmak üzere iki çeşit kümeleme yöntemi bulunmaktadır. Hiyerarşik olmayan kümeleme yöntemleri (k-ortalamalar, vb.), daha çok büyük veri setlerinde kullanılırken; hiyerarşik yöntemler genellikle veri setlerinin çok küçük olduğu durumlarda kullanılır. Veriler arasındaki benzerlik veya farklılıklardan yararlanarak veri setini alt kümelere ayırmakta kullanılan bu yöntemlerde, kümelerin oluşturulması için genellikle bir uzaklık ölçüsünden faydalanılır (Öklid, Manhattan, Minkowski, vb.) [6].

Breiman ise, verideki kümeleri bulmanın yoğunluk tahmini (veri yapısının göreceli sıklığı) ile ilişkili olabileceğini söylemiştir. Yoğunluk tahmininde, değişken değerlerinin tüm mümkün kombinasyonları ile başlanır ve hangi değer yapılarının daha yaygın olduğu (dağılımdaki pikleri) bulunmak istenir. Amaç, birçok verinin birbirine yakın olarak toplandığı "yüksek yoğunluklu" alanları bulmaktır. Eğer veride önemli herhangi bir yapı yoksa bu durumda, anlam taşımayan bir yapı da herhangi başka bir yapı gibi yaygın görülebilir [7], [8].

Breiman'ın danışmansız öğrenme ile ilgili önerisinde algoritmaya, verinin bir yapay halinin (bütün örüntülerin yok olduğu) oluşturulmasıyla başlanır. Verinin yapay hali oluşturulurken, gerçek veri setindeki her bir sütun (değişken) ayrı ayrı alınır ve değerleri bir yerde karıştırılır. Sütun, başladığı tam olarak aynı değer ve frekansla biter. Ancak bu durumda her bir değer, rastgele yerleştirme kuralları uygulanarak "ait olmadığı" bir satıra taşınır. Bu işlem, verinin her bir sütunu için tekrar edilir. Veri setini bağımsızlaştırmak için yapılan bu karıştırma işlemi, herhangi bir diğer sütun referansı olmadan gerçekleştirilir. Yeni verinin tüm tanımlayıcı istatistikleri, gerçek veri ile aynıdır (aynı ortalama, varyans, sıklık dağılımı). Ancak, sütunlar arasındaki ilişki yapısı yok edilmiştir. Breiman'ın fikri; gerçek veriyi ve karıştırılmış yapay veriyi birbirinden ayırmak için tasarlanmış sınıflama modeli ile gerçek verideki yaygın yapıları baskın hale getirmektir [8].

Bu çalışmanın amacı, danışmansız öğrenme tekniklerinden biri olan kümeleme yöntemine ait, Breiman algoritmasının işleyiş adımlarını tanıtmak, gerçek bir araştırmadan elde edilen veri seti üzerinde uygulama adımlarını göstermek ve elde edilen sonuçları yorumlamaktır.

2. GEREÇ VE YÖNTEMLER (MATERIAL AND METHOD)

2.1. Kullanılan veri seti (The data set used)

Çalışmada kullanılan veri seti, Zonguldak Karaelmas Üniversitesi Tıp Fakültesi psikiyatri polikliniğine 1-31 Ocak 2011 tarihleri arasında gece yeme sendromu şikâyetiyle ayakta başvuran ve çalışmaya katılmayı kabul eden 433 hastaya ait çeşitli demografik ve klinik bilgileri içermektedir[*]. Çalışmaya alınan hastalardan, yaş, cinsiyet, eğitim yılı, kardeş sayısı, medeni durum, çocuk sayısı gibi demografik özellikler, bel çevresi, kalça çevresi, beden kitle indeksi, fiziksel hastalık varlığı, psikolojik hastalık varlığı gibi çeşitli değişkenler sorgulanmıştır. Hastalara ayrıca, gece yeme anketi (GYA), beden şekli anketi (BSQ), semptom tarama listesi (SCL-90) ölçekleri de uygulanmıştır.

Bu çalışmada, yukarıda tanımlanan hasta özellikleri kullanılarak veride kümelenmelerin olup olmadığı, bir başka ifadeyle hangi özellikteki bireylerin bir arada toplandığı araştırılmıştır. Başvuru yapan kişilerin 97'sine klinik olarak gece yeme sendromu vardır tanısı konmuştur.

2.2. Breiman'ın danışmansız kümeleme önerisi (Breiman's suggestion of unsupervised clustering)

Danışmanlı öğrenmede, ilk olarak bir hedef veya bağımlı değişken seçimi yapılmadan sınıflama işlemine başlanmaz ve verinin homojen bölümlere ayrılması, temel amaç olan hedef sınıfların ayrıştırılması işlemi ile gerçekleştirilir. Danışmansız öğrenmede ise, bir hedef değişken söz konusu değildir. Burada amaç, verideki benzer özellikteki grupları bulmaktır. Danışmansız öğrenme, bir çeşit veri özetleme yöntemi olarak düşünülebilir. Yani, gerçek veri tabanını özetleyecek veya

[*] Bu veriler amaca uygun olarak çeşitli klasik istatistik yöntemlerle değerlendirilmiş ve yazarları tarafından yayınlanmak amacıyla dergiye gönderilmiştir. Verilerin kullanımı için bu çalışmada ilk isim olarak bulunan yazardan izin alınmıştır.

onun yerine kullanılacak ortalama sayıda değişken bulunmak istenir [9].

Kümeleme analizi için Leo Breiman, danışmansız öğrenmenin yoğunluk tahmini ile birlikte kullanıldığı yeni bir yaklaşım geliştirmiştir. Algoritmada, gerçek veriler ile kopya veriler karşılaştırılarak verideki yapılar kontrol edilir. Herhangi bir çalışma verisindeki kümeleri veya yapıları görsel olarak tanımlamak için, bu veriye ait grafiksel bir çizim yapıldığında da bu yöntem dolaylı olarak kullanılmış olmaktadır [2].

Breiman'ın bu yaklaşımında, ilk olarak gerçek veri setinin bir kopyası oluşturulur ve daha sonra bu kopya set üzerinde bütün örüntülerin yok olduğu bir yapay veri seti elde edilir. Örüntüler, veri setindeki her bir değişken değerinin her defasında diğer değişkenler yok varsayılarak rasgele dağıtımıyla yok edilir. Böylece yapay veri setinde, değişkenler arasındaki bağımlılıklar ortadan kaldırılmış olur. Bu aşamada problem, gerçek veri seti ve yapay veri setinden oluşan iki sınıf problemi olarak ele alınır. Gerçek veri setinde kümelenme olup olmadığını incelemek için, gerçek veri seti ile yapay veri setini ayırt etmede, lojistik regresyon, ayırma analizi, CART, RF, vb. sınıflama yöntemlerinden birisi kullanılabilir. Eğer bu iki veri seti başarılı bir şekilde ayrılabiliyorsa, gerçek veri setinde kümelenme yapılarının olduğundan bahsedilir.

Algoritma aşağıdaki adımlardan oluşmaktadır [2], [9]:

1. İlk olarak gerçek verinin bir kopyası oluşturulur ve kopya veri setinde her bir sütun ayrı ayrı rasgele dağıtımla karıştırılır. Örneğin, bir veri setindeki yaş değişkeninin rasgele dağıtımla karıştırıldığını düşünelim. Bu durumda, her bir hasta kaydı muhtemelen diğer bir hastaya ait olan yaş bilgisini içerecek ve bu şekilde sütundaki veriler rastgele dağıtılmış olacaktır. Bu işlem her sütun için gerçekleştirilir. Breiman, gerçek veriden oluşan bir grup ile bu verilerin her bir sütununun yeniden örneklenmesi (resample) sonucunda oluşturulan kopya verileri kullanır. Daha sonra herhangi bir sınıflama yöntemi yardımıyla gerçek veri grubu ve kopya veri grubunun ayırt edilebilirliği araştırılır.

   Veri, aynı *M* boyutuna sahip, N adet, *X* vektörü içerir. Gerçek veri sınıf I olarak etiketlenir ve sınıf II olarak işaretlenen N büyüklüğünde kopya veri seti oluşturulur. Gerçek veri setinde, n'inci örnekteki m'nci değişkenin değeri $X(m,n)$ olarak gösterilir. Burada her sınıftaki iki örneğin nasıl oluşturulduğu görülmektedir. İlk koordinat N değerden $\{ X(1,n)\}$ rastgele olarak, ikinci koordinat N değerden $\{ X(2,n)\}$ rastgele olarak seçilir ve bütün değişkenlerin rastgele dağıtımı yapılarak kopya veri seti oluşturulur.

   Kopya veri setinin dağılımı, değişkenler arasındaki bağımlılıkları yok eder. Bu sınıf, *M* bağımsız rastgele değişkenin dağılımına sahiptir ve m'nci değişken, gerçek verideki m'nci değişkenle aynı tek değişkenli dağılıma sahiptir.

   Elde edilen bu yeni verinin tüm tanımlayıcı istatistikleri, gerçek veri ile aynıdır (aynı ortalama, varyans, sıklık dağılımı). Ancak, sütunlar arasındaki ilişki yapısı yok edilmiştir.

2. Karıştırılmış veri seti, gerçek veri setine eklenir. Böylece, bu eklenmiş veri setinin sütun sayısı, önceki veri setiyle aynı, satır sayısı ise öncekinin iki katı olur. Verinin üst kısmı gerçek veri, alt kısmı ise karıştırılmış kopya veri olmak üzere iki grup veri olduğunu göstermek için yeni bir sütuna kodlama yapılır.

3. Gerçek ve kopya veri setleri arasındaki ayrımı belirlemek için, lojistik regresyon, CART, RF, vb. sınıflama yöntemlerinden birisi kullanılarak tahmin edici bir model oluşturulur. Bu model ile verilerin hangisinin gerçek veri seti, hangisinin rastgele dağıtımla oluşturulmuş kopya veri seti olduğunu belirlemek mümkün değilse, bu durumda veride önemli bir yapının veya kümelenmenin olmadığı söylenir. Ancak, farklılıklar kolayca belirlenebiliyorsa, bu durumda veride güçlü bir kümelenmenin olduğu kararına varılır. Başka bir ifadeyle; 2. adımdan sonra, veri iki sınıf problemi olarak değerlendirilebilir. Eğer bu iki sınıfı ayırmadaki hata oranı %50'ye yakınsa, herhangi bir sınıflama yöntemi, iki sınıf arasında başarılı bir ayrım yapamaz. Bu durumda, gerçek veri seti *M* tane bağımsız rastgele değişkenden örneklenmiş gibi görünür ve bu dağılım kümelenme olmadığı durumu temsil eden yani ilgilenilmeyen bir dağılımdır. Ancak, eğer bir kümelenme söz konusu ise, gerçek veri setinde sınıflama yöntemleri başarılı bir şekilde ayrım yapabilir.

Danışmansız öğrenmeyle ilgili olan Breiman'ın bu yaklaşımı, kümeleme yönteminde önemli avantajlar sağlar [9]:

- ✓ Değişken seçimi gerekli değildir ve farklı değişken gruplarında farklı kümeler tanımlanabilir.
- ✓ Bu kümeleme yöntemleri verinin nasıl ölçeklendiğinden etkilenmediği için, veride bir ön işleme veya yeniden ölçeklendirme yapmak gerekli değildir.
- ✓ Bu yöntemlerde kayıp veriler otomatik olarak değerlendirildiği için, kayıp değerlerle ilgili herhangi bir zorluk söz konusu değildir.

2.3. Sınıflama ve regresyon ağaçları (CART) algoritması (Classification and regression tree (CART) algorithm)

Parametrik olmayan CART algoritmasında, ana düğüm (iki grubun karıştırıldığı kök yapı) iki yavru düğüme ayrışır ve ikili bölünmeler maksimum ağaç yapısı elde edilinceye kadar devam eder.. Bölünmenin tamamlandığı

düğüme terminal düğüm adı verilir ve içinde hangi grup daha yoğun gözleniyor ise o düğüm söz konusu grubun adını alır. Herhangi bir düğümün heterojenlik değeri safsızlık (impurity) ölçüsü olarak adlandırılır ve bu çalışmada düğüm heterojenliğini ölçmek amacıyla Twoing kriteri kullanılmıştır. Maksimum ağaç elde edildikten sonra, öğrenme kümesindeki gürültülü verilerden oluşan ve test kümesinde hataya neden olan dallar silinerek budama yapılır ve optimum ağaç belirlenir. Optimum ağacın terminal düğümleri içinde hangi gruptaki verinin daha sık gözlendiği, kümelenme olup olmadığının bir göstergesidir. Bu düğümlerde gerçek veri kayıtlarının görülme sıklığı daha yüksek ise, o düğümde potansiyel bir "küme" varlığından bahsedilir. Sınıflama modelinin geçerliliğinin değerlendirilmesinde, 10-katlı çapraz geçerlilik yöntemi kullanılmıştır. Bu yöntemde veri seti 10 eşit parçaya bölünür ve her defasında 9 parçada yer alan veriler kullanılarak (eğitim seti) sınıflama modeli geliştirilir ve geriye kalan 1 parçadaki verilerde (test seti) model performansı test edilir. Bu işlem, her parçanın test verisi olarak kullanılmasıyla sonuçlanır. Hesaplanan performans değerlerinin ortalaması alınarak incelenen ağacın performansı bulunur.

Gece yeme alışkanlığının sorgulandığı araştırmadan elde edilen gerçek ve yapay veri setleri, CART algoritması ile sınıflandırılmıştır [10].

CART algoritması, SPM (demo) paket programı yardımıyla uygulamıştır.

3. BULGULAR (RESULTS)

Başlangıçta gerçek veri seti birinci, kopya veri seti ise ikinci grup olarak belirlenmiş ve iki sınıflı bir sınıflama problemi elde edilmiştir.

Gerçek veri setinden oluşan birinci grup, kümelenme yapılarının olabileceği varsayılan gruptur. Kopya olarak oluşturulan veri setinden elde edilen ikinci grup ise küme yapıları içermeyen ve ratsgele dağılım gösteren özelliklerden oluşmaktadır. Bu problemin çözümü için ileri sürülen sıfır hipotezi, gerçek ve kopya veri setlerinin başarılı bir şekilde ayırt edilemeyeceğini yani gerçek veri setinde bir kümelenme yapısının olmadığını öne sürer. Alternatif hipotez ise, gerçek veri setinde kümelenmeler var şeklinde kurulur. Bu kümeler CART algoritması ile elde edilmiştir. Gerçek ve kopya veri setlerinde aynı sayıda denek bulunduğu için, gruplara sınıflamada başlangıç olasılığı 0.50 olarak alınmıştır.

CART sınıflandırması sonucunda elde edilen 16 farklı ağaç yapısından bir tanesi optimum ağaç olarak seçilmiştir. Optimum ağaç, hatası ve karmaşıklık ölçüsü en düşük ve en dengeli olan ağaç yapısıdır. Optimum ağaçta, çapraz geçerlilik nispi maliyeti (Cross-Validated Relative Cost) 0.693±0.032, tekrar yer değiştirme nispi maliyeti (Resubstitution Relative Cost) 0.372 ve karmaşıklık parametresi (cost complexity parameter) 0.0006 olarak elde edilmiştir. Ayrıca elde edilen optimum ağaçta, eğitim veri setinde ROC eğrisi altında kalan alan 0.8677, test setinde ise 0.6906 olarak bulunmuştur.

Optimum ağaçta toplam 31 terminal düğüm yer almaktadır. Ağaç yapısında yer alan bu düğümler, homojen kümeler olarak adlandırılmıştır. Bu düğümler ağacın karar düğümleri olup 14 tanesi küme varlığını göstermektedir. Bir başka ifadeyle 14 terminal düğüm içinde yer alan denekler kendi içinde kümelenme göstermiştir. Diğer 17 terminal düğüm ise herhangi bir kümelenmenin olmadığı, yani deneklerin rasgele yerleştiği grupları göstermektedir. Bu çalışmada önem taşıyan kümelenme oluşturan 14 terminal düğüm olup bunların yorumları ve özellikleri aşağıda verilmiştir.

Bireylere ait kümelerin elde edilmesinde sorgulanan veya ölçülen değişkenlerin önem sırasına göre sıralaması Tablo 1' de verilmiştir. Tablo 1 incelendiğinde, ilk sırada yer alan kalça çevresi uzunluğunun en etkili ayırt edici değişken olduğu ve kalça çevresinden sonra beden kitle indeksi ve diğer değişkenlerin geldiği görülmektedir. Toplam 15 değişkenin etkisi 0' dan farklı bulunmuş ve bunlar içinde ayırt edicilikte en az etkili değişken ise cinsiyet olarak belirlenmiştir.

Tablo 1. Sınıflamada kullanılan değişkenlerin modele katkılarına göre nispi önem dereceleri
(Degrees of relative importance of the variables used in classification according to their contribution to the model)

Değişken	Nispi Önemlilik
Kalça çevresi	100
BKİ	87.5
Yaş	847
Bel çevresi	82.3
Çocuk sayısı	68.8
Medeni durum	25.1
Kardeş sayısı	3.98
BSQ puanı	3.27
Kilo vermek	2.8
Eğitim yılı	2.3
GYA puanı	1.71
Ort. SCL	1.59
Psikiyatrik tanı	1.42
Fiziksel hast.	1.03
Cinsiyet	0.37
Sigara	0
Antidepresan	0

Çalışmada yer alan toplam 433 kişiden 350'si oluşturulan 14 küme içine girmiştir. Geriye kalan 83 kişinin, ölçülen veya sorgulanan özellikleri bakımından dağılımının rastgele olduğu yani bir küme oluşturmadığı belirlenmiştir. Bunlardan Kümeler içine giren 350 kişiden 273 (%78)'ü klinik olarak gece yeme alışkanlığı yoktur

tanısı alırken, 77 (%22)'si ise gece yeme alışkanlığı vardır tanısı almıştır.

14 kümenin sadece 1 tanesinde (küme no:12) klinik olarak "gece yeme alışkanlığı vardır" tanısı alanların ağırlıklı olarak yer aldığı, bir kümede ise (küme no:4) yine gece yeme alışkanlığı var tanısı alanlarla almayanların yaklaşık benzer oranda yer aldığı gözlenmiştir. Diğer 12 kümede yer alan kişilerin ağırlıklı olarak gece yeme alışkanlığı yok tanısı alanlardan oluştuğu ve bu sonuca göre oluşturulan kümelerin sahip oldukları özelliklerin, genel olarak gece yeme alışkanlığı olmayan bireyleri ayırt edebildiği söylenebilir. Söz konusu 14 homojen kümenin sahip oldukları özellikler birbirlerinden farklı olup Tablo 2'nin ikinci sütununda tanımlanmıştır.

Tablo 2. Ondört terminal düğümün özellikleri
(Properties of fourteen terminal node)

Grup	Kural	N		Toplam N
		GYA yok N (%)	GYA var N (%)	
Küme 1	Yaş ≤ 31.5; medeni durum = bekar; çocuk sayısı ≤ 0.5; Eğitim yılı ≤ 10.5; Fiziksel Hastalık = Yok	19 (70.4)	8 (29.6)	27
Küme 2	Yaş ≤ 31.5; medeni durum = bekar; çocuk sayısı ≤ 0.5; Eğitim yılı > 10.5	43 (79.6)	11 (20.4)	54
Küme 3	Yaş ≤ 31.5; medeni durum = dul, evli; çocuk sayısı = 1; Kalça Çevresi ≤ 99.5	10 (90.9)	1 (9.1)	11
Küme 4	Yaş ≤ 31.5; medeni durum = dul, evli; çocuk sayısı = 1; Kalça Çevresi > 99.5; GYA_Puan > 14.5	8 (57.1)	6 (42.9)	14
Küme 5	Çocuk sayısı > 1.5; 28.5 < Yaş ≤ 31.5; Ort_SCL > 1.33	6 (75)	2 (25)	8
Küme 6	31.5 < Yaş ≤ 58.5; BKİ ≤ 25.25; Bel çevresi ≤ 88.5; Kalça çevresi ≤ 101.5	34 (75.6)	11 (24.4)	45
Küme 7	31.5 < Yaş ≤ 58.5; BKİ ≤ 25.25; Bel çevresi ≤ 88.5; 101.5 < Kalça çevresi ≤ 107	7 (77.8)	2 (22.2)	9
Küme 8	31.5 < Yaş ≤ 58.5; 22.55< BKİ ≤ 25.25; 88.5 < Bel çevresi ≤ 95.5	7 (100)	0 (0)	7
Küme 9	31.5 < Yaş ≤ 58.5; BKİ > 25.25; 98.5 < Kalça çevresi ≤ 101.5; Bel çevresi ≤ 99.5	11 (84.6)	2 (15.4)	13
Küme 10	31.5 < Yaş ≤ 58.5; Kalça çevresi > 101.5; Bel çevresi ≤ 91.5; 25.25 < BKİ ≤ 29.15; Ort_SCL > 0.59	24 (77.4)	7 (22.6)	31
Küme 11	31.5 < Yaş ≤ 58.5; 101.5 < Kalça çevresi ≤ 111.5; 91.5 < Bel çevresi ≤ 103.5; GYA_Puan ≤ 18.5; BKİ > 25.25	27 (100)	0 (0)	27
Küme 12	31.5 < Yaş ≤ 58.5; 91.5 < Bel çevresi ≤ 103.5; GYA_Puan > 18.5; 108.5 < Kalça çevresi ≤ 111.5; BKİ > 25.25	4 (33.3)	8 (66.7)	12
Küme 13	31.5 < Yaş ≤ 58.5; Bel çevresi > 91.5; Kalça çevresi > 111.5; BKİ > 25.25	59 (75.6)	19 (24.4)	78
Küme 14	Yaş > 58.5; Çocuk sayısı > 1.5; GYA_Puan ≤ 14.5	14 (100)	0 (0)	14
Toplam N		**273 (78)**	**77 (22)**	**350**

4. TARTIŞMA ve SONUÇ (DISCUSSION AND CONCLUSION)

Danışmansız öğrenme yöntemlerinden birisi olan kümeleme yöntemleri, sağlık alanında özellikle çeşitli kanser türlerinin genetik yapı ile ilişkisinin araştırıldığı çalışmalarda yaygın kullanılmıştır [11], [12], [13].

Veri madenciliği yöntemleri kullanılarak yapılan kümeleme çalışmalarının sayısında son yıllarda artış görülmektedir [14], [15], [16].

Bu çalışmada, küme oluşturmada avantajlar sağlayan veri madenciliği algoritmalarının, uygulamada kullanımının yaygınlaştırılması hedeflenmiştir. Bu amaçla, gece yeme alışkanlığı olan veya olmayan kişilerde belirli özellikler bakımından bir benzerlik olup olmadığı araştırılmış ve kendi içinde benzer özelliklere sahip, kendi aralarında farklılık gösteren 14 tane küme belirlenmiştir. Bu kümeler, sağlık alanında, koruyucu hekimlikten tedavi yöntemlerinin belirlenmesine kadar çeşitli amaçlarla kullanılabilir.

Sonuç olarak, hedef veya bağımlı değişkenin bilinmediği durumlarda, veri setinde var olan homojen alt grupların belirlenmesinde, değişkenlerin dağılım şekli ve tipinden etkilenmeyen Breiman algoritmasının etkin bir şekilde kullanılabileceği söylenebilir. Ayrıca Breiman ve CART algoritmalarının birlikte kullanımının, sonuçların güvenilirliği ve yorumlama kolaylığı açısından tekrarlanabileceği söylenebilir.

KAYNAKLAR (REFERENCES)

[1] Ş. Koltan Yılmaz ve S. Patır, "Kümeleme Analizi ve Pazarlamada Kullanımı", *Akademik Yaklaşımlar Dergisi*, 2(1), 91-113, 2011.

[2] L. Breiman ve A. Cutler, **RFtools--for Predicting and Understanding Data**, Interface Workshop-April 2004.

[3] Ç. Taşkın ve GG. Emel, "Veri Madenciliğinde Kümeleme Yaklaşımları ve Kohonen Ağları ile Perakendecilik Sektöründe Bir Uygulama", *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 15(3), 395-409, 2010.

[4] Ö. Terzi, EU. Küçüksille, G. Ergin ve A. İlker, "Veri Madenciliği Süreci Kullanılarak Güneş Işınımı Tahmini", *SDU International Technologic Science*, (3)2, 29-37, 2011.

[5] YZ. Ayık, A. Özdemir ve U. Yavuz, "Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği ile Analizi", *Sosyal Bilimler Enstitüsü Dergisi*, 10(2), 441-454, 2007.

[6] Y. Özkan, **Veri Madenciliği Yöntemleri**, Papatya Yayıncılık, 2008.

[7] İnternet: Notes On Setting Up, Using, And Understanding Random Forests, **http://www.stat.berkeley.edu/~breiman/notes_on_random_forests_v2.pdf**, 30.05.2013.

[8] İnternet: Salford Systems Predictive Modeler Unsupervised Learning, **http://1.salford-systems.com/Portals/160602/docs/Unsupervised_Learning_slides.pdf**, 22.05.2013.

[9] İnternet: Unsupervised Learning and Cluster Analysis with CART, **http://www.salford-systems.com/blog/dan-steinberg/item/572-unsupervised-learning-and-cluster-analysis-with-cart**, 02.06.2013.

[10] H. Çamdeviren Ankaralı, AC. Yazıcı, Z. Akkus, R. Bugdayci ve MA.Sungur, "Comparison of logistic regression model and classification tree: An application to postpartum depression data", *Expert Systems with Applications*, 32(4), 987-994, 2007.

[11] LF. Handfield, YT. Chong, J. Simmons, BJ. Andrews ve AM. Moses, "Unsupervised Clustering of Subcellular Protein Expression Patterns in High-Throughput Microscopy Images Reveals Protein Complexes and Functional Relationships Between Proteins", *PLoS Comput Biol*., 9(6), 2013, doi: 10.1371/journal.pcbi.1003085.

[12] MJ. Overman, J. Zhang, S. Kopetz, M. Davies, J. Zhi-Qin, K. Stemke-Hale, P. Rümmele, C. Pilarsky, R. Grützmann, S. Hamilton, R. Hwang, JL. Abbruzzese, G. Varadhachary, B. Broom ve H. Wang, "Gene Expression Profiling of Ampullary Carcinomas Classifies Ampullary Carcinomas in to Biliary-Like and Intestinal-Like Subtypes That are Prognostic of Outcome", *PLoS One*, 8(6), 2013, doi: 10.1371/journal.pone.0065144.

[13] P. Stegmaier, A. Kel, E. Wingender ve J. Borlak, "A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs", *PLoS Comput Biol.*, 2013, doi: 10.1371/journal.pcbi.1002958.

[14] T. Shi, S. Horvath, "Unsupervised Learning With Random Forest Predictors", *Journal of Computational and Graphical Statistics*, 15(1), 118–138, 2006.

[15] T. Shi, D. Seligson, AS. Belldegrun, A. Palotie ve S. Horvath, "Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma", *Mod Pathol*., 18(4), 547-57, 2005.

[16] L. Breiman, "Random forests", *Machine Learning*, 45(1), 5-32, 2001.