ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2016-52498

Grup Üyelerini Belirlemede İstatistiksel Sınıflandırma Yöntemleri: Karşılaştırmalı Bir Çalışma

Statistical Classification Methods to Determine Members of the Group: A Comparative Study

Öznur İŞÇİ GÜNERİ,[a]
Dursun AYDIN[a]

[a]İstatistik Bölümü,
Muğla Sıtkı Koçman Üniversitesi,
Fen Fakültesi, Muğla

Geliş Tarihi/*Received:* 24.06.2016
Kabul Tarihi/*Accepted:* 13.12.2016

Yazışma Adresi/*Correspondence:*
Öznur İŞÇİ GÜNERİ
Muğla Sıtkı Koçman Üniversitesi,
Fen Fakültesi,
İstatistik Bölümü, Muğla,
TÜRKİYE/TURKEY
oznur.isci@mu.edu.tr

ÖZET Amaç: Bu çalışmada, önceden tanımlı iki veya daha fazla grubun üyelerini tahmin etmek ve sınıflandırma fonksiyonlarını geliştirmek için diskriminant analizi (DA), çok terimli lojistik regresyon (MLR) ve regresyon ağacı (CART) olarak adlandırılan üç farklı yöntem ele alınmıştır. Bu çalışmanın temel amacı, DA, MLR ve CART analizi kullanılarak sınıflandırma yapmaktır. **Gereç ve Yöntemler:** Bu üç yöntem arasında en temel farklardan biri, CART parametrik olmayan bir teknik olmasına karşın, diğer ikisi bazı temel varsayımlara dayanan parametrik yöntemlerdir. Bir diğer önemli fark, DA ve MLR bireysel türler için grup üyeliğinin olasılıklarını üretirken, CART sadece farklı türler için ortalama olasılıkları üretir. **Bulgular:** Bu yöntemlerin uygulaması, Kuzey-Batı Anadolu, Bozcaada, Gökçeada, Edirne ve Istranca olmak üzere toplam beş farklı bölgeden elde edilen, yedi farklı Apodemus türlerine ait iki ayrı veri seti için sınıflandırma yapılmış ve sonuçlar tablolar halinde sunulmuştur. **Sonuç:** Çalışmada parametrik olmayan bir yöntem olan karar ağacı algoritmalarıyla parametrik yöntemler olarak bilinen diskriminant ve lojistik regresyonun sınıflama özellikleri karşılaştırılmaktadır. Her iki uygulama sonucunda, en yüksek sınıflandırma oranı çok terimli lojistik regresyon analizi ile elde edilmiştir.

Anahtar Kelimeler: Diskriminant analizi; çok terimli lojistik regresyon analizi; CART; Apodemus türleri

ABSTRACT Objective: In this paper, three different methods are considered such as discriminate analysis (DA), multinomial logistic regression (MLR) and regression and classification tree (CART) in order to develop classification functions and predict the membership of objects into two or more predefined groups. The main purpose of the study is to make classification using DA, MLR and CART analysis. **Material and Methods:** . One of the main differences among these three methods is that CART is a nonparametric technique, whereas the other two are the parametric method based on some basic assumptions. Another important difference is that DA and MLR produce the probability of group membership for individual species, whereas CART produces only average probability for different species. In the classification using DA, MLR and CART analysis was conducted and the results are presented in tables. **Results:** The application of the three techniques is illustrated by comparing seven different Apodemus species obtained from five different regions including Northwest Anatolia, Bozcaada, Gökçeada, Edirne and Istranca. **Conclusion:** In this study, a non-parametric methods, which decision tree algorithms and known as parametric methods, discriminant and logistic regression classification features are compared. In a result of both applications, the highest classification rate are obtained using logistic regression analysis.

Keywords: Discriminant analysis; multinomial logistic regression analysis; CART Apodemus species

Sınıflama ve regresyon yöntemleri, bağımlı ve bağımsız değişken arasındaki ilişkiyi tanımlamaya yönelik veri analizlerinin önemli bir parçasıdır. Çok değişkenli istatistik yöntemlerinde sınıflandırmada;

sınıfların önceden bilinmemesi durumuna göre çok boyutlu ölçekleme analizi ve kümeleme analizi kullanılırken, sınıfların önceden bilinmesi durumunda ise diskriminant (ayırma) analizi ve lojistik regresyon analizi kullanılmaktadır. Diğer yandan, bağımsız değişkenler için herhangi bir varsayım olmaksızın kategorik (sınıflı) bağımlı değişkeni tahmin etmek için aynı zamanda karar ağaçları da kullanılmaktadır.[1] Karar ağaçları; kolay anlaşılır olması, görsel sunumunun ön planda olması gibi nedenlerle sıklıkla tercih edilmektedir. Önceden tanımlı grup üyelerinin uygun gruba atanmasında kullanılan Diskriminant ve lojistik regresyon parametrik yöntemler olurken, karar ağaçları parametrik olmayan bir yöntemdir.

Diskriminant analizi, 1936 yılında Fisher tarafından geliştirilen bir sınıflandırma yöntemidir.[2] Diskriminant analizinde olay tamamıyla bir istatistiksel karar vermedir. Yani hatalı sınıflandırma olasılığını en aza indirgeyerek bireyleri ait oldukları gruplara ayırmak, çekilmiş oldukları kitleleri belirlemektir.[3] Diskriminant analizinin temeli incelenen bireyin kitlesinin belirlenmesini sağlayacak bir fonksiyonun bulunmasıdır. Bu fonksiyonun bulunmasında belirlenecek grupların ortalamaları arasındaki farklılığın maksimum olması amaçlanmaktadır. Uygulamada genellikle bağımlı değişkeninin sürekli olduğu doğrusal regresyon modelleri yaygın olsa da, bağımlı değişkenin kategorik olması halinde normallik varsayımının bozulması ve tipik doğrusal modelin uygulanamadığı durumlarda lojistik regresyon modelinin kullanımı standart bir yöntem haline gelmiştir. Bağımlı değişkenin iki ya da çok sınıflı kesikli değişken olması durumunda kullanılabilecek modeller çok çeşitlidir. Bu modellerden doğrusal olasılık modeli, lojistik ve probit modeller arasında en fazla tercih edilen yöntem lojistik regresyondur. Lojistik regresyon analizini, doğrusal regresyon analizinden ayıran en belirgin özellik de lojistik regresyon analizinde bağımlı değişkenin iki ya da çok sınıflı değerler almasıdır. Lojistik regresyon ve doğrusal regresyon analizi arasındaki bu farklılık hem parametrik model seçimine hem de varsayımlara yansımaktadır.[4] Lojistik regresyon, normallik varsayımının bozulması nedeniyle doğrusal regresyon analizine alternatif olmaktadır. Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir. Temel olarak lojistik regresyonda bağımsız değişkenler ile iki ya da çok sınıflı kategorik bağımlı değişken arasındaki ilişkinin tanımlanması için matematiksel modelleme yapmak amaçlanmaktadır.[5]

Bu üç yöntemden diskriminant analizi, sık kullanılan ve çok bilinen yöntemlerden biridir. Diskriminant analizinin sağlaması gereken normallik ve ortak varyans-kovaryans gibi bazı varsayımların bozulması durumunda, lojistik regresyon, alternatif bir yöntem olarak kullanılmaktadır. Lojistik regresyonda bağımlı değişken, 0,1 gibi ikili (binary) ya da ikiden çok düzey içeren (polychotomous) kesikli değişken olduğundan, normallik varsayımı kısıtı olmaması nedeniyle kullanım rahatlığının yanı sıra çözümlemeden elde edilen modelin matematiksel olarak çok esnek olması kolay yorumlanabilir olması yönteme olan ilgiyi arttırmaktadır.[3]

Bu çalışmada, giriş bölümünü takip eden ikinci bölümde sınıflama ve regresyon modellerinden karar ağaçları ve karar ağaçlarında en çok kullanılan analizlerin yapısı ile çalışmanın diğer konusu olan lojistik regresyon modeli ve diskriminant analizine değinilmiştir. Uygulamanın yapıldığı üçüncü bölümde, diskriminant analizi ve lojistik regresyon ile karar ağacı algoritmalarından en çok kullanılan Classification and Regression Trees (CART) algoritmasının, analizi yapılmış ve çalışma yapılan analizlerin karşılaştırılmasının ve açıklanmasının yer verildiği dördüncü bölüm olan sonuç ve öneriler ile sona erdirilmiştir.

GEREÇ VE YÖNTEMLER

DOĞRUSAL SINIFLANDIRMA MODELLERİ GELİŞTİRMEK İÇİN KULLANILAN İSTATİSTİKSEL YÖNTEMLER

Bu bölümde doğrusal sınıflandırma modelini geliştirmek için üç teknik tanımlanmıştır. Diğer bir ifadeyle, türler arsındaki doğrusal sınırlar ile ilişkili modeli tanımlamak için CART, DA ve MLR yöntemleri ele alınmıştır. Bu gibi sınıflandırmalarda sınırın doğası kaç tane açıklayıcı değişkenin kullanıldığına bağlıdır: bir değişken bir sınır noktası (cut-off value) yaratır; iki değişken düz bir çizgi sınırı verir ve üç değişken düzlemsel bir sınır olarak sonuçlanır. Dört veya daha fazla değişken olması durumunda sınır bir hiper düzlem olacaktır.[6]

DİSKRİMİNANT ANALİZİ (DA)

Diskriminant ya da ayırma analizi, önceden sınıflandırılan iki ya da daha fazla grubu birbirinden ayıran çok değişkenli bir analiz türüdür. Bu analiz, grup dışından alınan gözlemin hangi gruba atanabileceğini göstermektedir. Bu analizinin en basit türü, iki grup diskriminant analizidir. Burada, iki grubu ayırmak için grupların merkezinden geçen bir doğrusal diskriminant fonksiyonu (DF) kullanılabilir. Söz konusu diskriminant fonksiyon skoru aşağıdaki gibi tanımlanabilir:

$$DF_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} \quad (1)$$

Burada a_0 sabit terim, $a_1, a_2, ..., a_p$, p tane açıklayıcı değişken $x_1, x_2, ..., x_p$ için regresyon katsayılarını gösterir. Eşitlik (1)'deki *DF* skoru, standartlaştırılmış formda aşağıdaki biçimde ifade edilebilir:

$$SDF_i = b_1 z_{i1} + b_2 z_{i2} + ... + b_p z_{ip} \quad (2)$$

Burada $b_1, b_2, ..., b_p$, p tane standart $z_1, z_2, ..., z_p$ değişkenleri için regresyon katsayılarını gösterir. En büyük değerli standart regresyon katsayılarına sahip değişkenler, grup üyeliğinin kestirimine en çok katkı sağladığı varsayılır. Standartlaştırılan her bir değişkenin ortalaması sıfır olduğundan, tüm bireylere göre SDF 'nin ortalaması sıfır ve her bir SDF_i'nin standart sapması ise 1 olarak bulunur. Bu nedenle, bir birey, SDF skoru sıfırdan büyükse ($SDF_i > 0$) bir gruba sınıflandırılırken, SDF değeri sıfırdan küçükse ($SDF_i \leq 0$) başka bir gruba atanır veya sınıflandırılır.

Diskriminant analizi, üç veya daha fazla grubu ayırmak için kullanıldığında, birden çok *DF*'ye ihtiyaç duyulmaktadır. Söz konusu *DF*'lerin belirlenmesi, gruplar içi varyansın gruplar arası varyansa oranın maksimum olması esasına dayanır. Fisher (1936) tarafından tanımlanan iki varyans oranı,

$$\lambda_i = DF_i(a_1, a_2, ..., a_p) = \frac{a_i' \mathbf{B}\, a_i}{a_i \mathbf{W}\, a_i} \quad (3)$$

biçimindedir. Burada a_i, $(p \times 1)$ boyutlu katsayılar vektörü; $\mathbf{B}$, $(p \times p)$ boyutlu gruplara arası ve $\mathbf{W}$, $(p \times p)$ boyutlu grup içi varyans-kovaryans matrislerini göstermektedir (ayrıntılı bilgi için bkz. Manly(1993)). Eşitlik (3)'de verilen λ_i öz değerleri, $\left|\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}\right| = 0$ denkleminin çözümünden bulunur. Bu özdeğerlere karşılık gelen özvektörlere (katsayılar) göre tanımlanan, DF'lerin ayırma güçleri birincisine oranla azalarak devam etmesi nedeniyle, önemli grup farklılıklarının hemen hemen tümü için ilk birkaç fonksiyonu hesaplamanın yeterli olduğu ümit edilir. [9]

İlke olarak, grupları ayırmak için eşitlik (2)'de verilen *SDF*'ler kullanılır. Ancak, uygulamada bunun yerine özellikle üç veya daha fazla grup olması durumunda grupları ayırmak için aşağıdaki gibi, sınıflandırma fonksiyonları kullanılır:

$$C_i = c_{i0} + c_{i1}x_1 + c_{i2}x_2 + ... + c_{ip}x_p \ , \ i = 1,2,...,m, \ j = 1,2,...,p \qquad (4)$$

Burada c_{i0}, i.grup için sabit terimi; c_{ij}, i.grup için sınıflandırma skorunun hesaplanmasında j. değişkeninin ağırlığını; x_j, j. değişken için ilgili bireyin gözlemlen değerini ve C_i ise ortaya çıkan sınıflandırma skorunu göstermektedir. Eşitlik (4), matris formunda aşağıdaki gibi tanımlanır:

$$\mathbf{C}_i = \left(c_{i1}, c_{i2}, ..., c_{ip}\right) = \mathbf{W}^{-1}\mathbf{M}_i \qquad (5)$$

Burada $\mathbf{W^{-1}}$ gruplar için varyans kovaryans matrisinin tersini; $\mathbf{M}_i = \left(x_{i1}, x_{i2}, ..., x_{ip}\right)$, p değişken üzerinden i.grup için ortalamaların sütun matrisi ve $c_{i1}, c_{i2}, ..., c_{ip}$, i.grup için sınıflandırma katsayılarının sütün matrisini göstermektedir. Eşitlik (4)'de verilen i.grup için c_{i0} sabit terimi ise, aşağıdaki gibi hesaplanır.[10]

$$c_{i0} = \left(-\frac{1}{2}\right)\mathbf{C}_i'\mathbf{M}_i$$

Verilen bir bireyin ait olduğu gruba atanma olasılığını hesaplamak için Mahalanobis uzaklığı kullanılır. Söz konusu uzaklık, açıklayıcı değişkenler ile tanımlanan çok boyutlu uzaydaki grup merkezinden bireyin uzaklığını gösterir. Çok boyutlu uzayı oluşturan değişkenler ilişkili olduğunda, Mahalanobis uzaklığı uygun bir uzaklık ölçüsüdür. Eğer açıklayıcı değişkenler ilişkili değilse, Mahalanobis uzaklığı, Öklid uzaklığı ile aynı olur. Bu "sonsal olasılığın" hesaplanması sadece Mahalanobis uzaklığına değil aynı zamanda "önsel olasılıklara" da bağlıdır. Diskriminant analizi konusunda ayrıntılı bilgi için bkz., Kachigan, (1986), Manly , (1994), Tatlıdil, (2002).[3,7,8]

Çok Terimli Lojistik Regresyon (MLR)

Varsayalım y yanıt değişkeni, $p(x) = p\left(y = 1 | X = x\right)$ ve $1 - p\left(y = 1 | X = x\right)$ olasılıkları ile sırasıyla 1 ve 0 değerlerini alan ve Bernouli dağılan iki terimli kategorik rassal değişken olsun. Burada $X = (x_1, ..., x_k)$, k tane açıklayıcı değişkenler vektörüdür. Ancak burada iki terimli (ya da ikili) yanıt değişkenine dayalı olarak lojistik dağılım kullanılmaktadır.[4,11] Bilinmeyen $\beta = (\beta_0, \beta_1, ..., \beta_k)'$ parametreli *lojistik regresyon modelinin* spesifik şekli,

$$p(x) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\right)} = \frac{\exp\left(X\beta\right)}{1 + \exp\left(X\beta\right)} \qquad (6)$$

biçiminde ifade edilir. Lojistik regresyon modeli, (6) denklemi aracılığı ile açıklayıcı değişkenleri olasılıklara bağlar. Denklem (6) üzerinde yapılan basit cebirsel işlem gösteriyor ki, $p(x)$'in dönüşümü olan *logit model* elde edilir:

$$\text{logit}\left(p(x)\right) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad (7)$$

Eşitlik (7)'de verilen model aynı zamanda iki terimli lojistik regresyon modeli olarak adlandırılır ve $\left(p(x)/1-p(x)\right)$ miktarı, $p(x)$ başarı olasılığını, $1-p(x)$ başarısızlık olasılığı ile ilişkilendirir. (7) nolu modelden, başarının odds'u aşağıdaki gibi ifade edilir:

$$odds = \frac{p(x)}{1-p(x)} = \exp\left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\right) \tag{8}$$

Bu ifadenin logaritması yani, $\log\left(p(x)/1-p(x)\right)$, $p(x)$'nin logit'ini verir ve başarının log odds'unu ifade eder. Lojistik regresyon modeli, başarının log odds'ları için aşağıdaki gibi bir doğrusal modeli belirtir.[12]

$$\text{logit}\left(p(x)\right) = \log\left(\frac{p(x)}{1-p(x)}\right) = \log\left[\exp\left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\right)\right] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{9}$$

Lojistik regresyonda amaç, (6) denklemindeki lojistik regresyon modeli fit etmek için $\beta = (\beta_0, \beta_1, ..., \beta_k)'$ parametre vektörünü, maksimum olabilirlik (ML) yöntemiyle tahmin etmektir. ML denklemi y yanıt değişkeninin olasılık dağılımında elde edilir. Her bir y_i , i.anakütlede binom dağılımı gösterdiği için, y 'nin *bileşik olasılık yoğunluk fonksiyonu*,

$$f\left(\mathbf{y}|\beta\right) = \prod_{i=1}^{N} \frac{n!}{y_i\left(n!-y_i!\right)} p(x)^{y_i}\left(1-p(x)\right)^{n-y_i} \tag{10}$$

olarak ifade edilir. Sabit β değerleri için $f\left(\mathbf{y}|\beta\right)$ olasılık yoğunluk fonksiyonu bilinen bir fonksiyon olarak $\mathbf{y}$ değerlerini ifade eder. Olabilirlik fonksiyonu, fonksiyon parametrelerinin yer değiştirmesi dışında olasılık yoğunluk fonksiyonu ile aynı şekle sahiptir. Böylece *olabilirlik fonksiyonu*,

$$L\left(\beta|\mathbf{y}\right) = \prod_{i=1}^{N} \frac{n!}{y_i\left(n!-y_i!\right)} p(x)^{y_i}\left(1-p(x)\right)^{n_i-y_i} \tag{11}$$

olarak yazılabilir. Maksimum olabilirlik tahminleri (11) denklemini maksimum yapan β değerleridir. Elde edilen β parametresi, S- şeklinde $p(x)$ eğrisinin azalışının veya artışının oranı ile belirtilir. β'nın işareti eğride yükselme $(\beta > 0)$ veya azalma $(\beta < 0)$ olup olmadığı gösterir ve değişim oranı $|\beta|$'deki artışları ifade eder.

Kategorik yanıt değişkeni ikiden çok sınıf ya da terim içermesi durumunda, çokterimli lojistik regresyonu dikkate alınır. Örneğin, $y = (y_1, ..., y_m)$ yanıt vektörü olsun. Ayrıca, $p(y=k) = p_k,\ k = 1,2,..,m$ çok terimli olasılık ile y yanıt vektörünün m olası sonuç (grup) $\{1,2,...,m\}$ içerdiğini varsayalım. Eşitlik (7)'de verilen iki terimli logit modeline benzer olarak, olasılıklar aşağıdaki gibi elde edilir:

$$p_1 = p(y=1) = 1 \Big/ \left[1 + \sum_{j=1}^{m} \exp\left(X\beta_j\right)\right]$$

$$p_k = p(y=k) = \exp\left(X\beta_k\right) \Big/ \left[1 + \sum_{k=1}^{m} \exp\left(X\beta_j\right)\right],\ k = 1,2,.,,,.m \tag{12}$$

Burada $X=(x_1,x_2,...,x_k)'$ açıklayıcı değişkenler vektörü ve β_k, *k*. yanıt kategorisine karşı gelen parametre vektörüdür. (12) denkleminde görüldüğü gibi, standart grup olarak 1. grup dikkate alınmıştır. Ancak, onun yerine diğer herhangi bir grup kullanılabilir.

Lojistik regresyon modelinin log-odds aşağıdaki gibi ifade edilebilir:

$$\log\left(\frac{p_k(x)}{p_1(x)}\right)=X\beta_k,\ k=2,3,...,m \qquad (13)$$

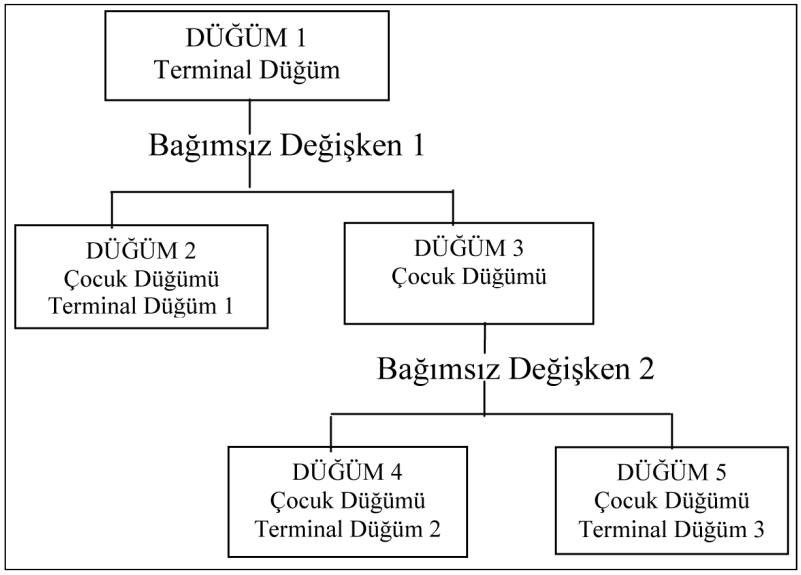
β_k parametre vektörü ve standart hataları (6)'da verilen lojistik regresyon modeli analizi genişletilerek iki terimli durumdaki gibi, maksimum olabilirlik yöntemi ile elde edilirler. Olabilirlik, (12) denkleminde belirtilen *k* tane olasılığı ifade eden çok terimli olasılıkların çarpımıdır. β_k'nın maksimum olabilirlik tahminlerini elde etmek için Newton-Rabson iteratif tahmin yöntemi kullanılır. Burada esas ilgi, parametre tahminlerinin yorumundan ziyade, (12) denkleminde verilen sınıf olasılıklarının tahmini ve verilen bir x değişkenine göre yeni bir gözlemin sınıflandırmasıdır.

SINIFLAMA VE REGRESYON AĞACI (CART)

Diskriminant ve lojistik regresyon analizlerini içeren sınıflama teknikleri ve regresyon modelleri sıklıkla literatürde kullanılmaktadır. Ancak bu tür modellerin gerektirdiği varsayımlar pek çok alanda istatistiksel analiz imkânlarını kısıtlamaktadır. İncelenen veri seti üzerinde hiçbir varsayım gerektirmemesi nedeniyle, CART bu tür parametrik tekniklere karşı güçlü bir alternatif olarak ortaya çıkmaktadır.[13] Söz konusu CART tekniği, hiyerarşik, kural tabanlı, parametrik olmayan bir yöntemdir. Bu yöntem ile hem kategorik hem de sürekli değişkenler modellenebilmektedir. Ele alınan bağımlı değişken kategorik ise yöntem sınıflama ağaçları (Classification Tree), sürekli ise regresyon ağacı (Regression Tree) olarak adlandırılmaktadır.[14] Bu yönüyle CART, hem çoklu regresyon analizini hem de bağımlı değişkenin kategorik olduğu durumlarda kullanılan lojistik regresyon analizini kapsamaktadır. Yapılan çalışmalarda kullanılan CART algoritması, her aşamada ilgili kümeyi kendinden daha homojen olan iki alt kümeye ayırarak ikili karar ağaçları oluşturan bir yapıya sahiptir. En iyi bağımsız değişken safsızlık (impurity) ve değişim ölçülerindeki (Gini, Twoing, en küçük kareler sapması) değişkenliği kullanarak seçilir. Burada, amaç hedef değişkene ilişkin mümkün olabilen en homojen veri alt gruplarını üretmektir.[15]

Karar ağaçlarının kök, dallar ve yapraklardan oluşan ağaca benzeyen bir yapısı olup, örnekteki tüm gözlemleri kapsayan bir kök ile başlayıp aşağıya doğru inildikçe veriyi alt gruplara ayıran dallara ayrılırlar. Bu kökten dallara doğru büyüyen ağaç yapısında her boğum "düğüm" dür, oluşan ağaçlarda homojen olmayan düğümlere "çocuk düğümü (child node)", homojen düğümlere ise "terminal düğüm (parent node)" adı verilir. Düğümler üzerinde niteliklerin test işlemi yapılmakta ve test işleminin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olmaktadır. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşmekte ve sonuç olarak ağaç sınıflar ile son bulmaktadır (ayrıntılı bilgi için bkz., 16,17,18,19). Örnek bir sınıflandırma ağacı Şekil 1'deki gibidir.

Karar ağaçları oluşturulurken kullanılan algoritmanın ne olduğu önemli bir husustur. Kullanılan algoritmaya göre ağacın şekli değişebilir. Bu durumda değişik ağaç yapıları da farklı sınıflandırma sonuçları verecektir. Kök denilen ilk düğümün farklı olması, en uçtaki yaprağa ulaşırken izlenecek yolu ve dolayısıyla sınıflandırmayı da değiştirecektir.[20] Karar ağaçlarına dayalı olarak geliştirilen birçok algoritma vardır. Bu algoritmalar;



ŞEKİL 1: Sınıflandırma ve regresyon ağacı örneği (düğüm sayısı: 5 ve terminal sayısı: 3 için).

- CHAID (Chi-Squared Interaction Detector: Otomatik Ki-Kare Etkileşim Belirleme),
- CART (Classification and Regression Tree: Sınıflama ve Regresyon Ağaçları),
- MARS (Multivariate Adaptive Regression Splines: Çok Değişkenli Regresyon Uzanımları),
- QUEST (Quick, Unbiased, Efficent Statistical Tree: Hızlı, Yansız, Etkili İstatistiksel Ağaç),
- SLIQ (Supervised Learnig in Quest),
- SPRINT(Scable Parallelizable Induction of Decision Tree),
- ID3, C4.5 VE c5.0

Bu algoritmalar kök, düğüm ve dallanma kriteri seçimlerinde izledikleri yol açısından birbirlerinden ayrılırlar. Bu çalışmadaki uygulamada CART algoritması içinde yer alan Gini ve Twoing kuralı kullanılmıştır.

GINI AYIRMA KURALI

Gini ayırma kuralı (ya da Gini indeksi) en yaygın olarak kullanılan kuraldır. İkili bölünmelere dayalı bir sınıflandırma yöntemidir. Bu kural nitelik değerlerinin sol ve sağda olmak üzere iki bölüme ayrılması esasına dayanmaktadır. Gini kuralı şu şekilde uygulanır:

Adım 1: Her nitelik değerleri ikili olarak gruplanır. Bu şekilde elde edilen sol ve sağ bölünmelere karşılık gelen sınıf değerleri gruplandırılır.

Adım 2: Her bir nitelikle ilgili sol ve sağ taraftaki bölünmeler için $Gini_{SOL}$ ve $Gini_{SAĞ}$ değerleri hesaplanır.

$$Gini_{SOL} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{SOL}|} \right)^2 \text{ ve } Gini_{SAĞ} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{SAĞ}|} \right)^2 \quad (14)$$

Bu eşitliklerde *k*, sınıf sayısı; *T*, birim düğümdeki örnekler; $|T_{SOL}|$, sol taraftaki örneklerin sayısı; $|T_{SAĞ}|$ sağ taraftaki örneklerin sayısı; L_i, sol tarafta *i* kategorisindeki örneklerin sayısı ve R_i, sağ tarafta *i* kategorisindeki örneklerin sayısı

Adım 3: Her *j* niteliği için, n eğitim kümesindeki satır sayısı olmak üzere aşağıdaki bağıntının değeri hesaplanır.

$$Gini_j = \frac{1}{n}(|T_{SOL}|Gini_{SOL} + |T_{SAĞ}|Gini_{SAĞ}) \quad (15)$$

Adım 4: Her *j* değeri için hesaplanan $Gini_j$ değerleri arasında en küçük olanı seçilir ve bölünme bu nitelik üzerinden gerçekleştirilir. Karar ağacı en küçük değere göre çizilir.

Adım 5: Adım 1'e dönülerek işlemlere devam edilir.

Gini kuralı en büyük sınıf için örnek öğrenmeyi araştıracak ve geri kalanları izole edecektir. Gini kuralı gürültülü veriler için iyi çalışır.

TWOING AYIRMA KURALI

Twoing ayırmanın anlamı iki süper sınıf içine sınıfları gruplama ve orijinal sınıfların yerine, grupların iki sınıf analizinin yapılması anlamına gelir. Doğal olarak sınıf sayısı büyük olduğunda tüm olasılıklar kontrol edilirse, olası gruplaşmaların sayısı, birleştirici patlamaya neden olabilir. Grabczewski (2014) ve Breiman v.d (1984) tüm olası sınıf gruplarını analiz etmek yerine her bir olası ayırma için optimal süper sınıfı belirlemek için etkili bir yöntem önermiştir.[21,22] Gini kuralının aksine, Twoing verilerin %50'sinden daha fazlasını birlikte yaparak iki sınıfa ayıracaktır. Twoing ayırma kuralının adımları şu şekildedir;

Adım 1: Twoing kuralını uygulamak için, niteliklerin her bir değeri için iki ayrı dizi oluşturulur. İkiye ayrılan bu eğitim kümesine aday bölünme adı verilmektedir. İki diziden sol tarafta bulunanı t_{SOL}, sağ tarafta yer alanı ise $t_{SAĞ}$ dizisi olarak değerlendirilir.

Adım 2: Aday bölünmelerden her biri için P_{SOL} ve $P(j/t_{SOL})$ olasılıkları hesaplanır. Söz konusu olasılıklar aşağıda verilmektedir. Burada $P(j/t_{SOL})$ ifadesi bir *j* sınıf değerinin sol tarafındaki bölünme olma olasılığını vermektedir. Söz konusu *j* değerleri sınıf değerlerinin yer aldığı nitelik olarak göz önüne alınır.

$$P_{SOL} = \frac{t_{SOL}\text{'daki herbir nitelik değerinini ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$P(j/t_{SOL}) = \frac{t_{SOL}\text{'daki kayıtların sayısı}}{t_{SOL}\text{'daki herbir nitelik değerinini ilgili nitelik sütunundaki tekrar sayısı}}$$

Adım 3: Aday bölünmelerden her biri için $P_{SAĞ}$ ve $P(j/t_{SAĞ})$ olasılıkları hesaplanır

$$P_{SAĞ} = \frac{t_{SAĞ}\text{'daki herbir nitelik değerinini ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$P(j/t_{SAĞ}) = \frac{t_{SAĞ}\text{'daki kayıtların sayısı}}{t_{SAĞ}\text{'daki herbir nitelik değerinini ilgili nitelik sütunundaki tekrar sayısı}}$$

Adım 4: $\Phi(s/t)$, t düğümündeki s aday bölünmelerinin uygunluk ölçüsü olsun. Söz konusu uygunluk ölçüsü şu şekilde hesaplanır:

$$\Phi(s/t) = 2P_{SOL}P_{SAĞ}\sum_{j=1}^{n}\left|P(j/t_{SOL}) - P(j/t_{SAĞ})\right| \quad (16)$$

$\Phi(s/t)$ değerleri hesaplandıktan sonra içlerinden en büyük olanı seçilir. Bu değerin ilgili olduğu aday bölünme satırı bize dallanmanın yapılacağı satırı bildirecektir.

Adım 5: Dallanma bu şekilde yapıldıktan sonra bu adıma ilişkin olarak karar ağacı çizilir.

Adım 6: Kuralın birinci adımına dönülerek ağacın alt kümesine aynı işlemler uygulanır.

Twoing kuralı daha dengeli ağaçları inşa etmek için izin verse de, bu kural Gini kuralından daha yavaş çalışır. Gini ve Twoing ayırma kuralı dışında birkaç yöntem daha vardır. Bunlar arasında en çok kullanılanlar Entropi kuralı, χ^2 kuralı, maksimum sapma kuralı sayılabilir.

UYGULAMA

Bu çalışmada temel amaç, diskriminant analizi, çok terimli lojistik regresyon ve CART yöntemleri kullanılarak sınıflandırma yapmaktır. Araştırmaya konu olan veriler, Kuzey-Batı Anadolu (A), Bozcaada (B), Gökçeada (G), Edirne (E) ve Istranca (I) olmak üzere, toplam 5 farklı bölgede yapılan arazi çalışması sonucunda oluşturulan 7 örneklem ya da grubu teşkil eden Apodemus (fare) türlerinden elde edilmiştir. Bu türlere ilişkin bölge tür kodları (BTK) tanımlanmış ve gerekli açıklamalar aşağıda verilmiştir:

BTK	Bölgeler	Apodemus Türleri
1	Anadolu	Apodems hermonensis
2	Bozcaada	Apodems hermonensis
3	Edirne	Apodemus slyvaticus
4	Edirne	Apodemus flavicollis
5	Gökçeada	Apodemus slyvaticus
6	Gökçeada	Apodemus flavicollis
7	Istıranca	Apodemus flavicolis

Farklı bölgelerde bulunan Apodemus türlerini sınıflandırılma ve hangi değişkenler itibariyle farklık gösterdiklerini belirlemek amacıyla iki veri seti dikkate alınmıştır. *Birinci veri seti*, 12 metrik (mm) kafatası değişkenlerinden elde edilirken, *ikinci veri seti* ise aynı Apodemus türlerine ilişkin 6 dış vücut ölçüsü (gr) değişkenlerinden elde edilmiştir. Bu veri setlerini oluşturan değişkenlere ilişkin tanımlar Tablo 1'de verilmiştir.

UYGULAMA 1: BİRİNCİ VERİ SETİ (APODEMUS TÜRLERİNİN KAFATASI ÖLÇÜLERİ)

Diskriminant Analizi Sonuçları

Birinci veri setinden elde edilen diskriminant analizi sonuçları aşağıda verilen Şekil ve Tablolarda özetlenmiştir: (Tablo 2)

Tablo 2 incelendiğinde, büyükten küçüğe doğru sıralanan özdeğerler, sadece diskriminant fonksiyonlarının (*DF*) ayırma güçlerinin önem sıralarını göstermektedir. Buna göre, birinci diskriminant (ayırma) fonksiyonu gruplar arası toplam değişimin (varyasyonun) %71,8'ini ve ikinci diskriminant fonksiyon de-

TABLO 1: Değişken tanımları

Veri Seti	Değişkenler	Kafatası ölçüleri (mm)	N	Minimum	Maksimum
Birinci Veri Seti	m6	Üst Diastema uzunluğu	224	5.40	8.50
	m7	Foramen İnkisiv Uzunluk	224	4.00	7.10
	m8	Nasal Uzunluk	224	7.70	12.00
	m9	Üst Molar Alveolar Uzunluğu	224	3.60	5.10
	m12	İnterorbital Genişlik	224	3.90	5.20
	m13	Oksipital Genişlik	224	9.60	13.00
	m16	Rostrum Genişliği	224	3.00	5.50
	m21	Kafatası Kapsül Genişliği	224	7.00	12.00
	m22	Mandibul Uzunluğu	224	11.00	16.00
	m23	Alt Molar Uzunluğu	224	3.50	4.60
	m24	Üst Kesici Diş Kalınlığı	224	1.00	1.70
	m25	Palatal Köprü Uzunluğu	224	4.00	6.20
İkinci Veri Seti		**Dış vücut ölçüleri (gr)**			
	tl	Toplam Uzunluk	366	137.00	261.00
	bh	Baş ve Vücut Uzunluğu	366	56.00	130.00
	t	Kuyruk Uzunluğu	366	66.00	148.00
	hf	Arka Ayak Uzunluğu	366	19.00	25.00
	e	Kulak Uzunluğu	366	13.00	20.00
	w	Ağırlık	366	11.00	53.00

TABLO 2: Özdeğerler, varyans yüzdeleri ve testler.

Özdeğerlerin İncelemesi					Wilks' Lambda Testi				
DF	Özdeğerler (λ)	Varyans %	Kümülatif %	Kanonik Korelasyon	DF'lerin testi	Wilks' Lambda	Ki-kare (χ^2)	Serbestlik derecesi	p
					0	0.003	1233.799	72	0.000
1	14.970(a)	71.8	71.8	0.968	1	0.049	642.252	55	0.000
2	3.955(a)	19.0	90.8	0.893	2	0.245	300.586	40	0.000
3	1.092(a)	5.2	96.0	0.722	3	0.512	143.021	27	0.000
4	0.636(a)	3.0	99.1	0.623	4	0.837	37.987	16	0.002
5	0.147(a)	0.7	99.8	0.358	5	0.960	8.766	7	0.270
6	0.042(a)	0.2	100.0	0.201	6				

a: Analizde kullanılan ilk 6 diskriminant fonksiyonu

ğişimin %19,0'unu açıklarken, kalan %9,1 toplam değişim ise, son dört diskriminant fonksiyonu tarafından açıklanmaktadır. Açıklanan değişim oranlarının kümülatif sütununa bakıldığında, toplam değişimin %96,0'sı ilk üç diskriminant fonksiyonu tarafından açıklanırken, kalan %4,0'lük kısım diğer üç diskriminant fonksiyonu tarafından açıklanmaktadır. Ayrıca, birinci diskriminant fonksiyonuna karşı gelen korelasyon katsayısı, diğer diskriminant fonksiyonlarına karşı gelen katsayılarından daha yüksek olduğundan, birinci diskriminant fonksiyonu gruplar arası farklılıkların önemli bir kısmını açıklarken, diğer diskriminant fonksiyonları azalan önem sırasında gruplar arası farklılıkları açıklamaktadır.

Diskriminant fonksiyonlarının anlamlılık sınamaları, Bartlett'in ki-kare testiyle yapılmıştır. Birinci diskriminant fonksiyonu istatistiki açıdan anlamlı bulunduktan sonra, en büyük özdeğer (birinci diskriminant fonksiyonu) kaldırılarak, geriye kalan özdeğerler test edilerek, ikinci diskriminant fonksiyonunda anlamlı bulunmuştur. Benzer şekilde, 6. diskriminant fonksiyonunun dışında kalan diğer fonksiyonlarda anlamlı bulunmuştur. Bu fonksiyonlar Tablo 2'de toplu olarak verilmiştir.

Apodemus türlerinin kafatası ölçüsü değerlerinden elde edilen fonksiyonlardan olan birinci diskriminant (ayırma) fonksiyonu,

$$DF_1 = -0.075(m6) + 0.387(m7) - 0.230(m8) - 0.128(m9) - 0.440(m12) - 0.126(m13) + 0.382(m16) + 1.107(m21) - 0.344(m22) - 0.202(m23) - 0.097(m24) + 0.198(m25) \quad (17)$$

olarak elde edilmiştir. Yukarıda ifade edilen DF_1 fonksiyonundaki katsayılar, Tablo 2'de verilen, λ_1=14.970 özdeğerine karşılık gelen özvektör-lerden elde edilmiştir. Buna göre, en büyük özdeğerin karşılığı olan katsayılardan elde edilen DF_1, diğerlerinden bağımsız olarak en büyük ayırma gücüne sahiptir. İkinci diskriminant fonksiyonu ise,

$$DF_2 = 0.152(m6) + 0.578(m7) + 0.120(m8) + 0.573(m9) - 0.290(m12) + 0.099(m13) + 0.268(m16) - 0.313(m21) + 0.016(m22) + 0.230(m23) - 0.513(m24) - 0.496(m25) \quad (18)$$

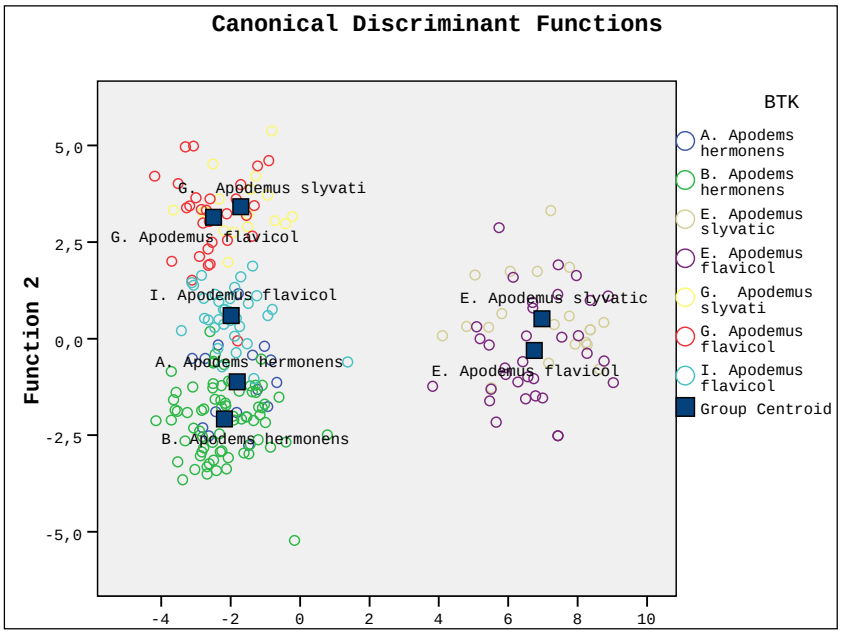
ikinci en büyük özdeğerin (λ_2=3.955) karşılığı olan katsayılardan elde edilen DF_2, diğer diskriminant fonksiyonları ile ilişkisiz ve ikinci derecede ayırma gücüne sahiptir. Benzer biçimde, altıncı diskriminant fonksiyonu;

$$DF_6 = 0.797(m6) - 0.044(m7) + 0.552(m8) + 0.347(m9) + 0.095(m12) - 0.059(m13) - 0.168(m16) + 0.039(m21) - 0.599(m22) - 0.642(m23) - 0.139(m24) + 0.250(m25) \quad (19)$$

en küçük özdeğere (λ_6=0.042) karşı gelen katsayılardan elde edilmiş olup, söz konusu diğer beş diskriminant fonksiyonundan bağımsız olarak, ayırma gücü en az olan diskriminant fonksiyonudur.

Şekil 2'de diskriminant uzayında 7 bölgeye ilişkin bölge tür kodlarına göre grupların dağılımı görülmektedir.

Diskriminant analizine göre sınıflandırma sonuçları Tablo 3'de verilmiştir. Sınıflandırma matrisinin köşegen elemanları doğru sınıflandırma oranlarını göstermektedir. Tablo 3'e göre, Diskriminant analizi türlerin %89,7'sini doğru olarak sınıflandırmıştır.



ŞEKİL 2: Diskriminant uzayında grupların dağılımı

Çok Terimli Lojistik Regresyon Sonuçları

Çok terimli lojistik regresyonda model uyumunu değerlendirmek için birkaç yol vardır. En yaygın kullanılan Tablo 4'te görülen "-***2 log olabilirlik***" testidir. Burada -***2 log olabilirlik*** sadece sabit terimli ve tüm bağımsız değişkenlerin yer aldığı nihai model için hesaplanır. Her biri için -***2 log olabilirlikler*** ki-kareyi üretmek için biri diğerinden çıkartılır (804.760-36.760= 768.000). İki model arasındaki daha büyük bir değişim miktarı model uyumunda daha büyük bir iyileşme önermektedir. Tablo 4'te görüldüğü gibi, son model sabit terimli modelden anlamı bir şekilde farklılaşmaktadır. Böylece bağımsız değişkenler sonuçların kestirimlerine anlamlı bir katkı sağlamaktadırlar. Sabit terimli modelle karşılaştırıldığında, son model için daha düşük bir AIC, aynı zamanda iyi bir uyumu göstermektedir (Tablo 4).

Lojistik regresyona ilişkin parametre tahminleri fazla yer kapladığından burada yer verilmemiştir. Değişkenlerin katkılarının anlamlı olup olmadıklarını incelemek için olabilirlik oran testleri Tablo 5'de verilmiştir. Bu sonuçlara göre 12 değişken içinde m9, m13, m16 ve m23 dışındaki değişkenlerin katkılarının anlamlı olduğu görülmektedir (Tablo 5)

Son modelin yararlı bir göstergesi Tablo 6'de görülen sınıflandırma tablosudur. Doğru olarak sınıflandırılan eleman sayıları köşegen üzerinde yer almaktadırlar. Buradan görülüyor ki, son model elemanların %96,9'unu doğru olarak tahmin etmiştir. Ancak, görülüyor ki, 1, 2, 5, 6 ve 7. gruplar diğer iki grupla karşılaştırıldığında, %100 doğru tahmin edilmişlerdir. Bu çok terimli lojistik regresyonda bazı durumlarda yaygın bir olaydır. Bu sonuçlardan da görüldüğü gibi, genellikle çok terimli lojistik regresyon daha büyük gruplar için daha iyi kestirimler üretir (Tablo 6).[4]

TABLO 3: Diskriminant analizi sınıflandırma sonuçları (%).

	Tahmin edilen grup üyelikleri							
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**Toplam**
1.00	86.7	13.3	0.0	0.0	0.0	0.0	0.0	100
2.00	6.7	92.0	0.0	0.0	0.0	0.0	1.3	100
3.00	0.0	0.0	90.0	10.0	0.0	0.0	0.0	100
4.00	0.0	0.0	18.2	81.8	0.0	0.0	0.0	100
5.00	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100
6.00	0.0	3.3	0.0	0.0	16.7	80.0	0.0	100
7.00	0.0	2.9	0.0	0.0	0.0	0.0	97.1	100

a. Genel olarak doğru olarak sınıflandırılan orijinal gruplardaki bireylerin yüzdesi: % 89.7

TABLO 4: Model uyum bilgisi.

	Model Uyumluluk Kriteri			Olabilirlik Oran Testi		
Model	**AIC**	**BIC**	**-2 Log Olabilirlik**	**Ki-kare (χ^2)**	**Serbestlik derecesi**	**p**
Sadece sabit terimin olduğu model	816.760	837.230	804.760			
Son model	192.760	458.869	36.760	768.000	72	0.000

TABLO 5: Olabilirlik oran testleri.

Etkiler	Model Uyumluluk Kriteri			Olabilirlik Oran Testleri		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Ki-kare (χ^2)	Serbestlik derecesi	p
Sabit	243.461	489.100	99.461(a)	62.701	6	0.000
m6	219.939	465.578	75.939(a)	39.179	6	0.000
m7	242.159	487.797	98.159(a)	61.398	6	0.000
m8	199.568	445.206	55.568(a)	18.807	6	0.005
m9	183.500	429.139	39.500(a)	2.740	6	0.841
m12	226.410	472.049	82.410(a)	45.650	6	0.000
m13	182.066	427.705	38.066(a)	1.306	6	0.971
m16	181.059	426.697	37.059(a)	0.298	6	1.000
m21	194.497	440.136	50.497(a)	13.737	6	0.033
m22	199.780	445.418	55.780(a)	19.019	6	0.004
m23	182.390	428.029	38.390(a)	1.630	6	0.950
m24	201.884	447.522	57.884(a)	21.123	6	0.002
m25	197.417	443.055	53.417(a)	16.656	6	0.011

TABLO 6: Çok terimli lojistik regresyon sınıflandırma sonuçları.

	Öngörülen							
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**Doğru (%)**
1.00	15	0	0	0	0	0	0	100.0
2.00	0	75	0	0	0	0	0	100.0
3.00	0	0	17	3	0	0	0	85.0
4.00	0	0	4	29	0	0	0	87.9
5.00	0	0	0	0	17	0	0	100.0
6.00	0	0	0	0	0	30	0	100.0
7.00	0	0	0	0	0	0	34	100.0
Genel (%)	6.7	33.5	9.4	14.3	7.6	13.4	15.2	**96.9**

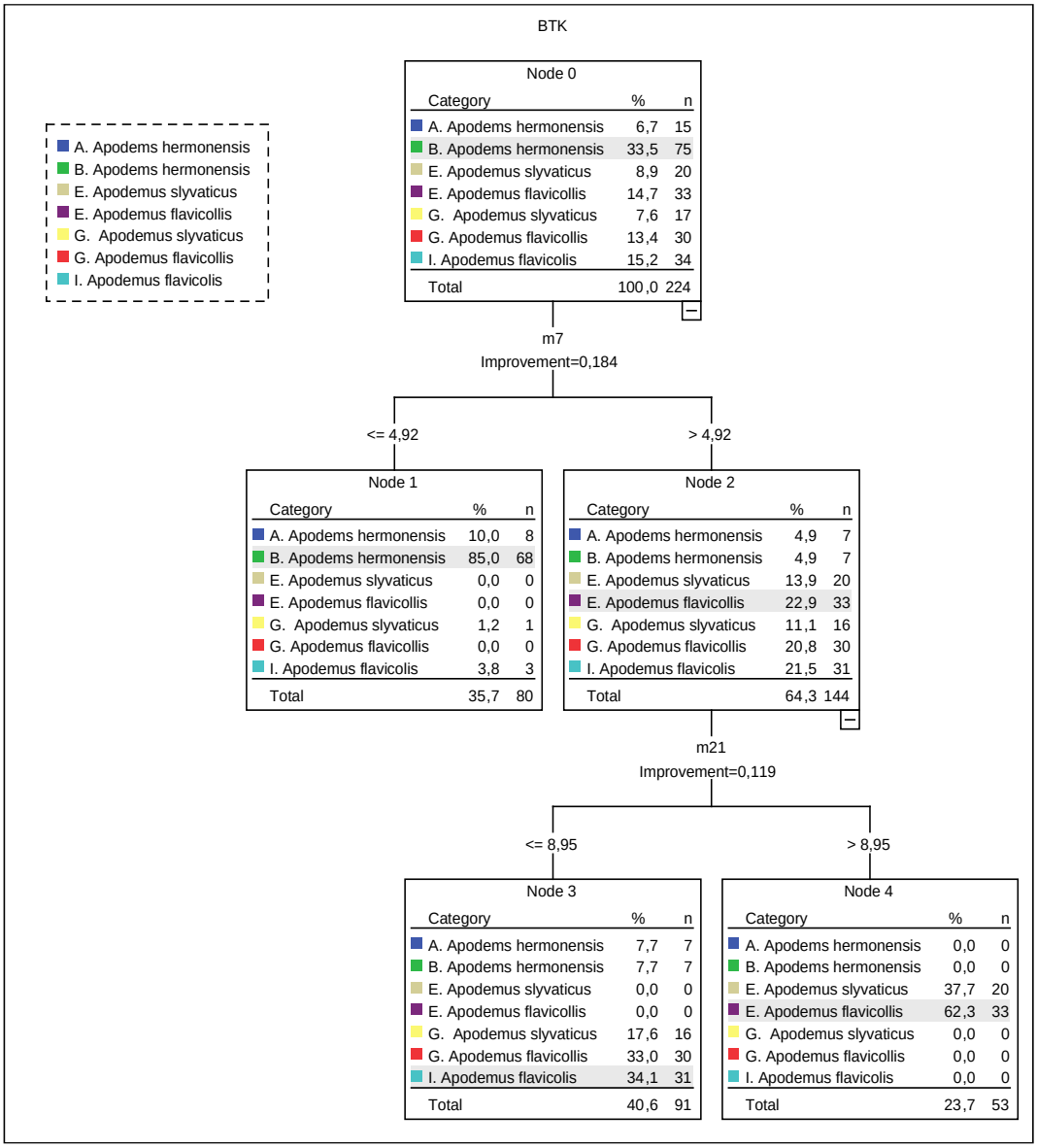
CART Analizi Sonuçları

CART analiz literatürde yaygın olarak kullanılan Gini ve Twoing kurallarına göre ayrı ayrı yapılmıştır. CART ağacı büyütürken bir dalın dağınıklığını yani safsızlığını en çok azaltan tahminleyiciye bölmekte ve ana daldan çocuk dala doğru dağınıklıktaki bu değişime gelişme(improvement) denilmektedir. Analizde bu gelişme değeri için SPSS 15.0'de "default" değeri olan 0,0001 seçilerek model oluşturulmuştur.

Gini algortiması ile oluşturulan modele ilişkin ağaç yapısı Şekil 3'te gösterilmiştir. Analiz sonucunda, parametrik yöntemlerde olduğu gibi, Gini kuralına göre grupları ayırmada en belirleyici değişkenler sırasıyla m7 ve m21 oldukları görülmektedir. Buna göre düğüm sayısı 5, terminal sayısı 3 ve ağacın derinliği yani ağacın büyüyebileceği katman sayısı 2 olarak bulunmuştur (Şekil 3).

Tablo 7'de görüldüğü üzere Gini kuralı kullanılarak yapılan CART sınıflandırma sonucunda doğru sınıflandırma oranı %58,9 olarak bulunmuştur.

Twoing kuralı ile oluşturulan modele ilişkin ağaç yapısı Şekil 4'de verilmektedir. Analiz sonucunda Twoing kuralına göre belirleyici değişkenler sırasıyla m21 ve m7 değişkenleridir. Twoing kuralına göre yine düğüm sayısı 5, terminal sayısı 3 ve ağacın derinliği 2 olarak bulunmuştur (Şekil 4).

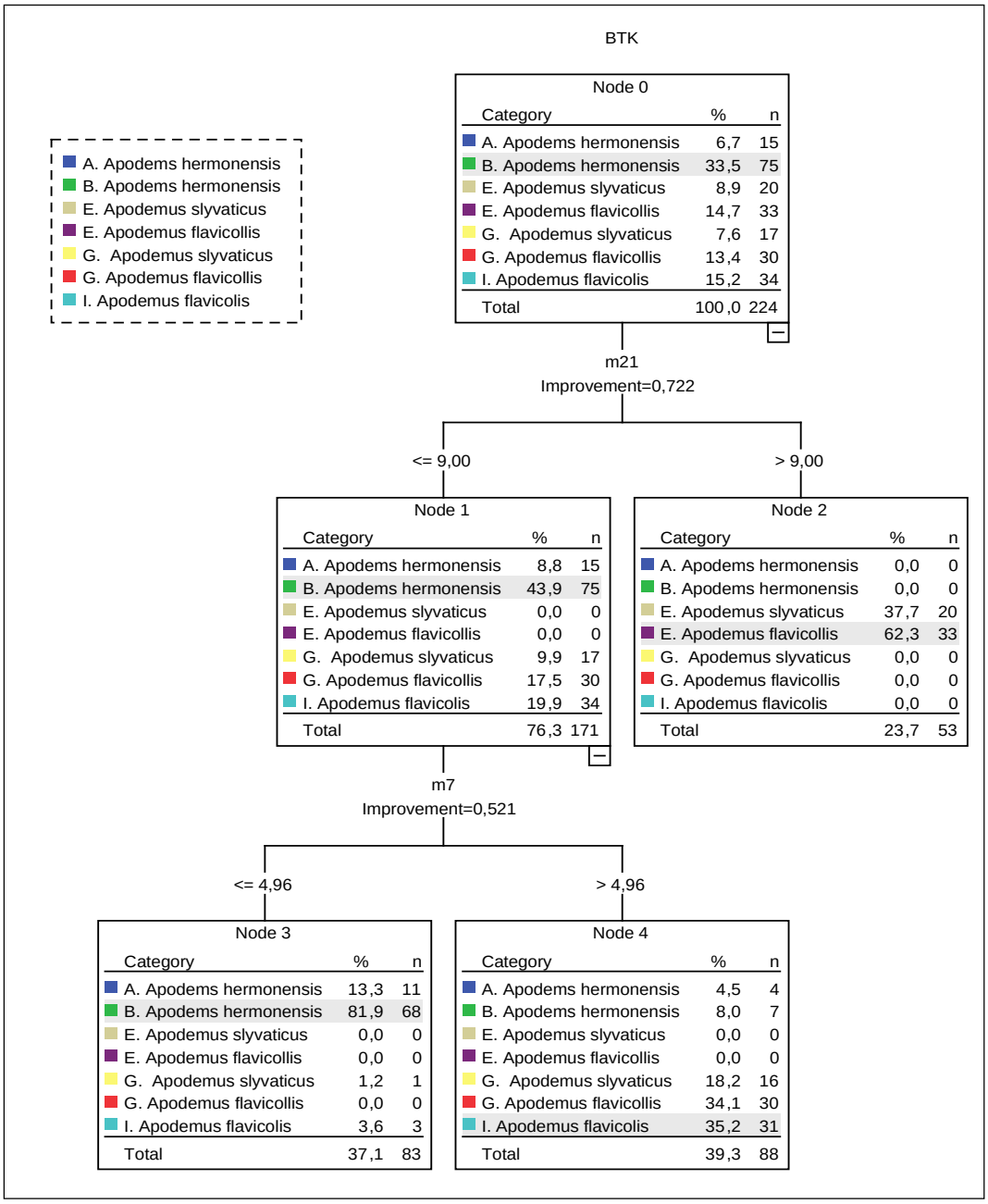


ŞEKİL 3: Gini kuralı ile CART sınıflandırma ağacı.

TABLO 7: CART sınıflandırma sonuçları (Gini kuralı).

	Öngörülen							
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**Doğru (%)**
1.00	0	8	0	0	0	0	7	0.0
2.00	0	68	0	0	0	0	7	90.7
3.00	0	0	0	20	0	0	0	0.0
4.00	0	0	0	33	0	0	0	100.0
5.00	0	1	0	0	0	0	16	0.0
6.00	0	0	0	0	0	0	30	0.0
7.00	0	3	0	0	0	0	31	91.2
Genel(%)	0.0	35.7	0.0	23.7	0.0	0.0	40.6	**58.9**

Tablo 8'de Twoing kuralı ile CART sınıflandırma sonuçları verilmektedir. Twoing kuralı kullanılarak yapılan CART sınıflandırma sonucunda doğru sınıflandırma oranı Gini algortimasında olduğu gibi yine %58,9 olarak bulunmuştur (Tablo 8).



ŞEKİL 4: Twoing kuralı ile CART sınıflandırma ağacı.

UYGULAMA 2: İKİNCİ VERİ SETİ (APODEMUS TÜRLERİNİN DIŞ VÜCUT ÖLÇÜLERİ)

Diskriminant Analizi Sonuçları

Diskriminat analizi sonuçları izleyen tablo grafiklerle özetlenmiştir. Tablo 9 incelendiğinde; büyükten küçüğe doğru sıralanan özdeğerler, sadece diskriminant fonksiyonlarının(DF) ayırma güçlerinin önem sıralarını göstermektedir. Buna göre, birinci diskriminant (ayırma) fonksiyonu gruplar arası toplam değişimin (varyasyonun) %67'sini ve ikinci diskriminant fonksiyon değişimin %15,1'ini açıklarken, kalan %17,9 toplam değişim ise, son dört diskriminant fonksiyonu tarafından açıklanmaktadır. Açıklanan de-

TABLO 8: CART sınıflandırma sonuçları (Twoing kuralı).

	Öngörülen							
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**Doğru (%)**
1.00	0	11	0	0	0	0	4	0.0
2.00	0	68	0	0	0	0	7	90.7
3.00	0	0	0	20	0	0	0	0.0
4.00	0	0	0	33	0	0	0	100.0
5.00	0	1	0	0	0	0	16	0.0
6.00	0	0	0	0	0	0	30	0.0
7.00	0	3	0	0	0	0	31	91.2
Genel(%)	0.0	37.1	0.0	23.7	0.0	0.0	39.3	**58.9**

TABLO 9: Özdeğerler, varyans yüzdeleri ve testler.

Özdeğerlerin İncelemesi					Wilks' Lambda Testi				
DF	**Özdeğerler (λ)**	**Varyans %**	**Kümülatif %**	**Kanonik Korelasyon**	**DF'lerin testi**	**Wilks' Lambda**	**Ki-kare (χ^2)**	**Serbestlik derecesi**	**p**
					0	0.368	358.771	36	0.000
1	0.848(a)	67.0	67.0	0.677	1	0.679	138.633	25	0.000
2	0.191(a)	15.1	82.1	0.400	2	0.809	76.001	16	0.000
3	0.177(a)	14.0	96.1	0.388	3	0.952	17.524	9	0.041
4	0.037(a)	2.9	99.0	0.188	4	0.987	4.600	4	0.331
5	0.013(a)	1.0	100.0	0.112	5	1.000	0.034	1	0.853
6	0.000(a)	0.0	100.0	0.010	6				

a: Analizde kullanılan ilk 6 diskriminant fonksiyonu.

ğişim oranlarının kümülatif sütununa bakıldığında, toplam değişimin %96,1'ini ilk üç diskriminant fonksiyonu tarafından açıklanırken, kalan %3,9'luk kısım diğer üç diskriminant fonksiyonu tarafından açıklanmaktadır. Ayrıca, birinci diskriminant fonksiyonuna karşı gelen korelasyon katsayısı, diğer diskriminant fonksiyonlarına karşı gelen katsayılarından daha yüksek olduğundan, birinci diskriminant fonksiyonu gruplar arası farklılıkların önemli bir kısmını açıklarken, diğer diskriminant fonksiyonları azalan önem sırasında gruplar arası farklılıkları açıklamaktadır (Tablo 9).

Diskriminant fonksiyonlarının anlamlılık sınamaları, Bartlett'in ki-kare testiyle yapılmıştır. Birinci diskriminant fonksiyonu istatistiki açıdan anlamlı bulunduktan sonra, en büyük özdeğer (birinci diskriminant fonksiyonu) kaldırılarak, geriye kalan özdeğerler test edilerek, ikinci diskriminant fonksiyonunda anlamlı bulunmuştur. Benzer şekilde, 6. diskriminant fonksiyonunun dışında kalan diğer fonksiyonlarda anlamlı bulunmuştur. Bu fonksiyonlar Tablo 9'da toplu olarak verilmiştir.

Apodemus türlerinin kafatası ölçüsü değerlerinden elde edilen fonksiyonlardan olan birinci diskriminant (ayırma) fonksiyonu,

$$DF_1 = -0.499(\text{tl}) + 0.218(\text{bh}) + 0.131(\text{t}) + 0.650(\text{hf}) + 0.594(\text{e}) + 0.061(\text{w}) \quad (20)$$

olarak elde edilmiştir. Eşitlik (20)'de verilen katsayılar, Tablo 9'da verilen, λ_1=0,848 özdeğerine karşılık gelen özvektörlerden elde edilmiştir. Buna göre, en büyük özdeğerin karşılığı olan katsayılardan elde edilen DF_1, diğerlerinden bağımsız olarak en büyük ayırma gücüne sahiptir. İkinci diskriminant fonksiyonu ise,

$$DF_2 = 0.573(\text{tl}) + 0.611(\text{bh}) - 1.156(\text{t}) - 0.106(\text{hf}) - 0.250(\text{e}) + 0.457(\text{w}) \quad (21)$$

ikinci en büyük özdeğerin ($\lambda_2 = 0.191$) karşılığı olan katsayılardan elde edilen DF_2, diğer diskriminant fonksiyonları ile ilişkisiz ve ikinci derecede ayırma gücüne sahiptir. Benzer biçimde, altıncı diskriminant fonksiyonu;

$$DF_6 = -5.435(\text{tl}) + 3.608(\text{bh}) + 3.339(\text{t}) + 0.139(\text{hf}) - 0.121(\text{e}) - 0.766(\text{w}) \quad (22)$$

en küçük özdeğere (λ_6=0.000) karşı gelen katsayılardan elde edilmiş olup, söz konusu diğer beş diskriminant fonksiyonundan bağımsız olarak, ayırma gücü en az olan diskriminant fonksiyonudur.

Şekil 5'de diskriminant uzayında 7 bölgeye ilişkin bölge tür kodlarına göre grupların dağılımı görülmektedir.

Sınıflandırma sonuçları Tablo 10'da verilmiştir. Sınıflandırma matrisinin köşegen elemanları doğru sınıflandırma oranlarını göstermektedir. Tablo 10'a göre, diskriminant analizi türlerin %43,7'sini doğru olarak sınıflandırmıştır.

Çok Terimli Lojistik Regresyon Sonuçları

Çok terimli lojistik regresyon analizi sonucunda Tablo 11'de görüldüğü üzere modele ilişkin genel anlamlılığın test edildiği model uygunluk tablosu sonucunda Ki-kare değeri anlamlı bulunmuştur. Olabilirlik oran testleri Tablo 12'de verilmektedir.

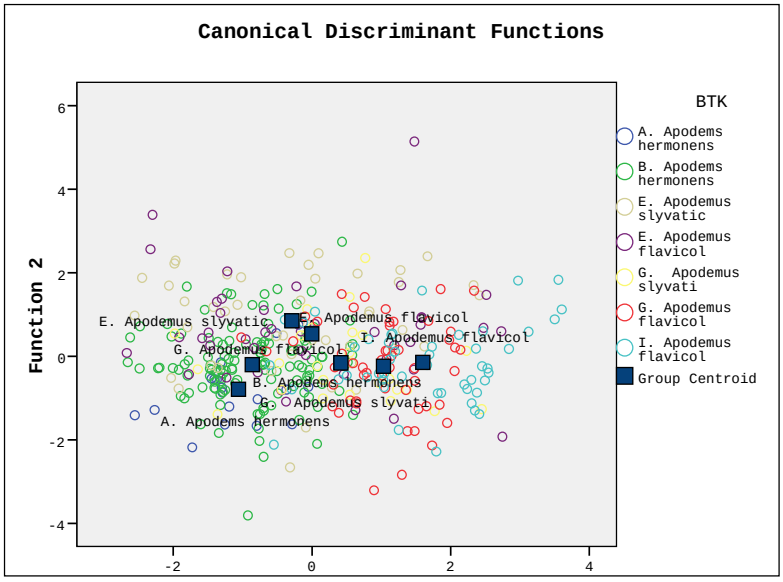
Tablo 12'deki sonuçlara göre 6 değişken içinde tl, bh, ve t dışında değişkenler arasında anlamlı bir ilişki olduğu görülmektedir.

Tablo 13'te görüldüğü üzere çok terimli lojistik regresyon analizi sonucunda doğru sınıflandırma oranı %53,3 olarak bulunmuştur.

CART Analizi Sonuçları

Gini kuralı ile oluşturulan modele ilişkin ağaç yapısı Şekil 6'da verilmektedir. Analiz sonucunda Gini kuralına göre en önemli belirleyiciler sırasıyla hf ve bh değişkenleridir. Buna göre düğüm sayısı 5, terminal düğüm sayısı 3 ve ağacın derinliği yani ağacın büyüyebileceği katman sayısı 2 olarak bulunmuştur (Şekil 6).

Tablo 14'te görüldüğü üzere Gini kuralı kullanılarak yapılan CART sınıflandırma sonucuna göre doğru sınıflandırma oranı %44,5 olarak bulunmuştur.



TABLO 10: Diskriminant analizi sınıflandırma sonuçları (%).

	Tahmin edilen grup üyelikleri							Toplam
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	
1.00	84.2	0.0	10.5	0.0	0.0	5.3	0.0	100
2.00	28.3	34.6	11.0	11.0	11.0	3.1	0.8	100
3.00	8.5	10.6	36.2	14.9	19.1	0.0	10.6	100
4.00	10.9	8.7	17.4	26.1	0.0	19.6	17.4	100
5.00	7.7	3.8	15.4	7.7	30.8	7.7	26.9	100
6.00	0.0	1.9	3.8	5.8	19.2	55.8	13.5	100
7.00	0.0	2.0	6.1	0.0	8.2	14.3	69.4	100

a. Doğru sınıflandırma oranı %43,7.

TABLO 11: Model anlamlılığına ilişkin testin sonucu.

	Model Uyumluluk Kriteri			Olabilirlik Oran Testi		
Model	**AIC**	**BIC**	**-2 Log Olabilirlik**	**Ki-kare (χ^2)**	**Serbestlik derecesi**	**p**
Sadece sabit terimin olduğu model	1313.131	1336.547	1301.131			
Son model	1017.977	1181.888	933.977	367.154	36	0.000

TABLO 12: Olabilirlik oran testleri.

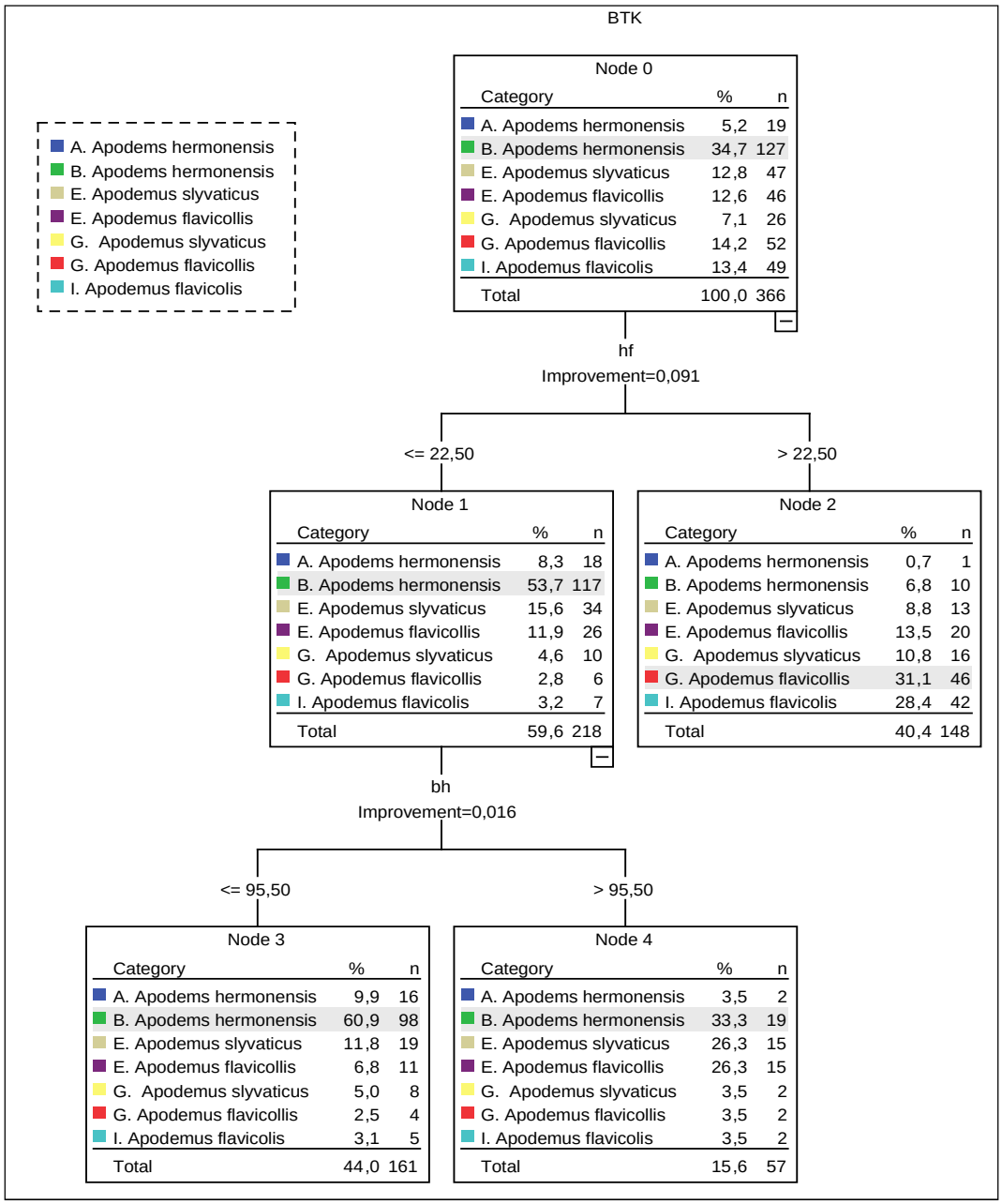
	Model Uyumluluk Kriteri			Olabilirlik Oran Testleri		
Etkiler	**AIC of Reduced Model**	**BIC of Reduced Model**	**-2 Log Likelihood of Reduced Model**	**Ki-kare (χ^2)**	**Serbestlik derecesi**	**p**
Sabit	1160.328	1300.823	1088.328	154.351	6	0.000
tl	1012.224	1152.719	940.224	6.247	6	0.396
bh	1013.643	1154.138	941.643	7.666	6	0.264
t	1013.693	1154.188	941.693	7.716	6	0.260
hf	1079.573	1220.068	1007.573	73.596	6	0.000
e	1070.697	1211.191	998.697	64.719	6	0.000
w	1021.670	1162.165	949.670	15.693	6	0.016

TABLO 13: Çok terimli lojistik regresyon sınıflandırma sonuçları.

	Öngörülen							
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**Doğru (%)**
1.00	1	17	0	0	0	1	0	5.3
2.00	1	110	6	5	0	5	0	86.6
3.00	1	22	12	4	1	1	6	25.5
4.00	0	21	5	6	0	8	6	13.0
5.00	0	11	2	2	1	4	6	3.8
6.00	0	8	2	4	2	30	6	57.7
7.00	0	5	3	0	0	6	35	71.4
Genel(%)	0.8	53.0	8.2	5.7	1.1	15.0	16.1	**53.3**

Twoing kuralı ile oluşturulan modele ilişkin ağaç yapısı Şekil 7'de verilmektedir. Analiz sonucunda Twoing kuralına göre en önemli belirleyici değişkenler hf ve bh değişkenleridir.

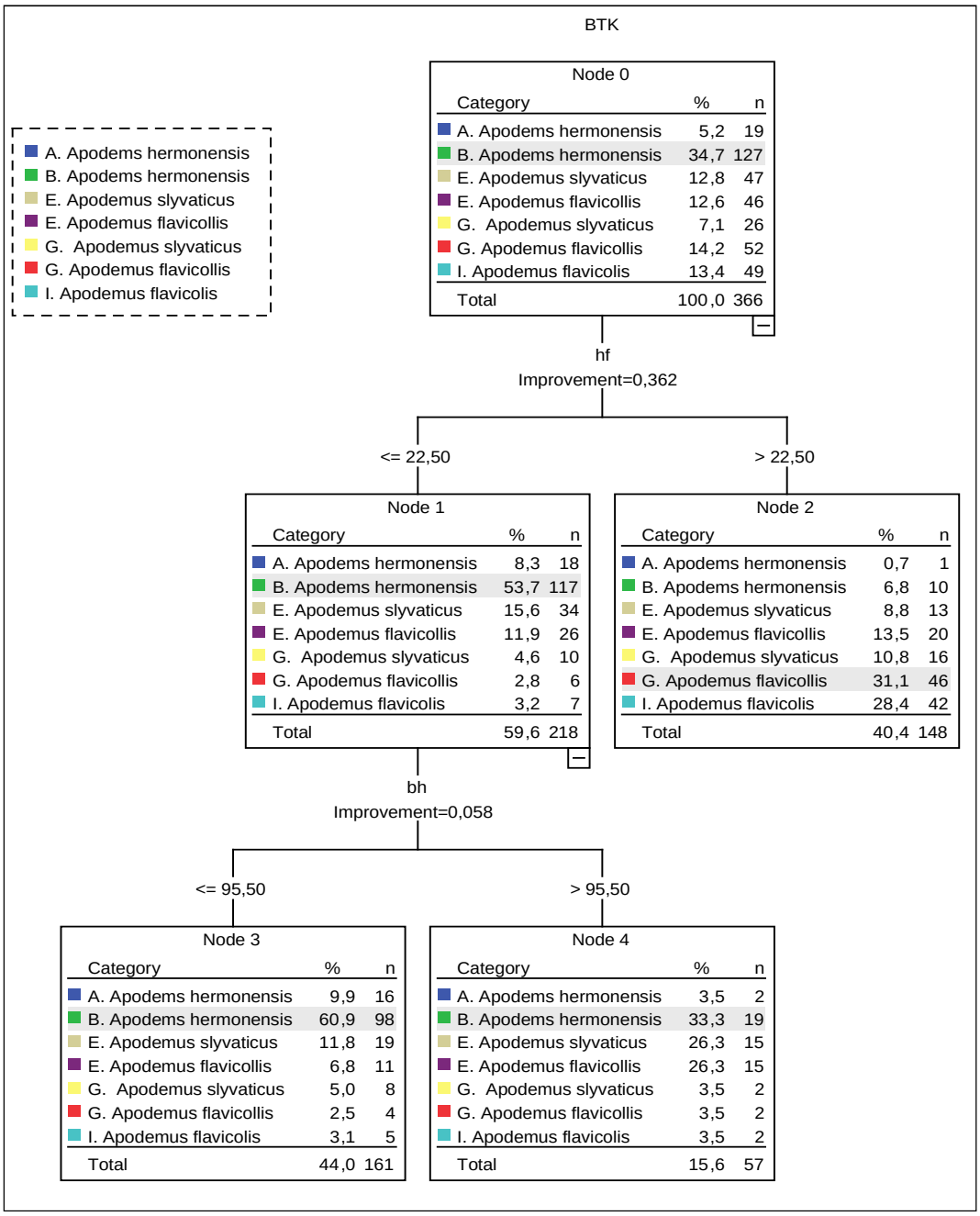
Tablo 15'te Twoing kuralı ile CART sınıflandırma sonuçları verilmektedir. Twoing kuralı kullanılarak yapılan CART sınıflandırma sonucunda doğru sınıflandırma oranı Gini algortimasında olduğu gibi %44,5 olarak bulunmuştur (Tablo 15)



ŞEKİL 6: Gini kuralı ile CART sınıflandırma ağacı.

TABLO 14: CART sınıflandırma sonuçları (Gini kuralı).

	Öngörülen							
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**Doğru (%)**
1.00	0	18	0	0	0	1	0	0.0
2.00	0	117	0	0	0	10	0	92.1
3.00	0	34	0	0	0	13	0	0.0
4.00	0	26	0	0	0	20	0	0.0
5.00	0	10	0	0	0	16	0	0.0
6.00	0	6	0	0	0	46	0	88.5
7.00	0	7	0	0	0	42	0	0.0
Genel(%)	0.0	59.6	0.0	0.0	0.0	40.4	0.0	**44.5**



ŞEKİL 7: Twoing kuralı ile CART sınıflandırma ağacı.

TARTIŞMA VE SONUÇ

Diskriminant analizi ve lojistik regresyon analizi gibi çok değişkenli parametrik yöntemler, verilerin sınıflandırılmasında yaygın olarak kullanılmaktadır. Bu tekniklere göre çok daha yeni olan karar ağacı algoritmaları, parametrik varsayımlarda esneklik sağlayan ve parametrik olmayan bir yöntem olarak, özellikle veri madenciliği alanında çok sık kullanılmaya başlanmıştır. Bu çalışmada, son dönemlerde yaygın olarak kullanılan karar ağaçları algoritmaları, diskriminant analizi ve lojistik regresyon analizi gibi üç farklı yöntem, özellikle sınıflandırma amcacıyla kullanılmış ve iki gerçek veri setinden elde edilen sınıf-

TABLO 15: CART sınıflandırma sonuçları (Twoing kuralı).

	Öngörülen							
BTK	**1.00**	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**Doğru (%)**
1.00	0	18	0	0	0	1	0	0.0
2.00	0	117	0	0	0	10	0	92.1
3.00	0	34	0	0	0	13	0	0.0
4.00	0	26	0	0	0	20	0	0.0
5.00	0	10	0	0	0	16	0	0.0
6.00	0	6	0	0	0	46	0	88.5
7.00	0	7	0	0	0	42	0	0.0
Genel (%)	0.0	59.6	0.0	0.0	0.0	40.4	0.0	**44.5**

TABLO 16: Analizlerin karşılaştırma kriterleri.

İstatistiksel Yöntemler	**Uygulama I**	**Uygulama II**
	Sınıflandırma Başarısı (%)	**Sınıflandırma Başarısı (%)**
Diskriminant analiz i(DA)	89.7	43.7
Çok terimli lojistik regresyon (MLR)	96.9	53.3
Sınıflama ve Regresyon Ağacı (CART)		
■ Gini kuralı	58.9	44.5
■ Twoing kuralı	58.9	44.5

landırma sonuçları karşılaştırılarak hangi yöntemin daha iyi bir performas gösterdiği belirlenmeye çalışılmıştır. Yapılan analizler sonucunda üç yönteme ilişkin iki uygulamadan elde edilen sınıflandırma sonuçları, Tablo 16'da özetlenmiştir.

Tablo 16'da verilen sonuçlar incelendiğinde, en yüksek doğru sınıflandırma oranları, deneklerin kafatası ölçülerine göre %96,9 ve dış vücut ölçülerine göre %53,3 olarak çok terimli lojistik regresyon analizi sınıflandırma sonuçlarından elde edilmiştir. Tablo 16'dan görüldüğü gibi, mm olarak ölçülen kafatası değişkenleri için en yüksek sınıflandırma oranları, parametrik yöntemler olarak ifade edilen diskriminant analizi ve çok terimli lojistik regresyondan elde edilmiştir. Ağırlık olarak ölçülen dış vücut ölçülerine göre, Apodemus türleri dikkate alındığında, en iyi doğru sınıflandırma oranları yine aynı yöntemlerden elde edilmiştir. Buna göre denekler, kafatası değişkenleri açısından çok daha iyi sınıflandırıldığı gözlenmiştir. Ayrıca, 7, 8, 14 ve 15 nolu Tablolarda ayrıntılı olarak ifade edildiği gibi, karar ağaçları algoritması CART yöntemi için yaygın kullanılan Gini ve Twoing kurallarına göre, deneklerin sınıflandırma oranları her iki uygulamada da benzerlik gösterdiği görülmüştür. Ancak parametrik yöntemlerde olduğu gibi, denklerin kafatası değişkenlerine göre çok daha iyi sınıflandırıldığı görülmüştür.

Birinci uygulama verileri dikkate alındığında, gruplara ayırmada en etkili m21 (kafatası kapsül genişliği) olurken, ikinci uygulamada ise, hf (arka ayak uzunluğu) değişkeni olmuştur. Bunlara ilaveten deneklerin birbirlerinden faklı olmalarına katkı sağlayan değişkenleri belirlemek için Eşitlik 17'de belirtilen birinci diskriminant fonksiyonunun katsayıları incelendiğinde, m21 (kafatası kapsül genişliği) ve m7 (foramen inkisiv uzunluk) değişkenleri kafatası ölçülerine göre grupları ayırmada etkili oldukları gözlenmiştir. Diğer yandan ikinci uygulama verileri için elde edilen birinci diskriminant fonksiyona göre, hf (arka ayak uzunluğu) ve e (Kulak Uzunluğu) değişkenleri gruplara ayırmada etkili olmuşlardır (bkz. Eşitlik 20). Buna ilaveten, çok terimli lojistik regresyon yöntemine göre gruplara ayırmada etkili olan değişkenler diskriminant analizine benzerlik gösterdikleri için burada yer verilmemiştir. Ayrıca, Şekil 3 ve 4'ten görüldüğü gibi, hem Gini hem de Twoing kuralları kullanılarak yapılan CART sınıflandırma ağacı yönteminde, birinci uygulama verileri için sınıflara ayırmada en önemli belirleyici m7 ve m21 değişkenleri

olurken, Şekil 6 ve 7'de görüldüğü gibi ikinci uygulamada ise, en iyi belirleyici hf ve bh değişkenleri olmuştur.

Çok terimli lojistik regresyon ve diskriminant analizi ile yapılan analizler sonucunda daha yüksek sınıflandırma oranı gösterse de CART analizi de varsayım gerektirmemesi nedeniyle özelikle uygun veri setleri için sınıflandırma amacıyla kullanılabileceği öngörülmektedir.

Bu makalede üç farklı yöntemin iki gerçek veriye dayalı olarak sınıflandırma performansları karşılaştırılmıştır. Aynı karşılaştırma bir simulasyon çalışması yapılarak da sağlanabilir. Ancak, elde edilen bulgular incelendiğinde, her iki uygulamada da sonuçların birbirleriyle tutarlı oldukları gözlenmiştir. Bu nedenle, ilaveten bir simulasyon çalışması yapılmasına ihtiyaç duyulmamıştır.

Çıkar Çatışması

Yazarlar herhangi bir çıkar çatışması veya finansal destek bildirmemiştir.

Yazar Katkıları

Fikir/Kavram: Araştırma ve/veya makalenin hipotezini veya fikrini oluşturmak: *Öznur İşçi Güneri, Dursun Aydın;* ***Tasarım: Sonuçlara ulaşılmasını sağlayacak yöntemi tasarlamak:*** *Öznur İşçi Güneri, Dursun Aydın;* ***Denetleme/Danışmanlık: Araştırmanın/çalışmanın yürütülmesini organize etmek, ilerlemesini gözetmek ve sorumluluğunu almak:*** *Öznur İşçi Güneri;* ***Veri Toplama ve/veya İşleme: Hastaların takibi, ilgili biyolojik materyallerin toplanması, verilerin düzenlenmesi ve raporlanması, deneylerin yapılması için sorumluluk almak:*** *Dursun Aydın;* ***Analiz ve/veya Yorum: Bulguların mantıklı bir şekilde değerlendirilerek sonuçlandırılmasında sorumluluk almak:*** *Dursun Aydın, Öznur İşçi Güneri;* ***Kaynak Taraması: Çalışma için gerekli kaynak taramasında sorumluluk almak:*** *Dursun Aydın, Öznur İşçi Güneri;* ***Makalenin Yazımı: Çalışmanın tamamının ya da önemli bölümlerinin yazılmasında sorumluluk almak:*** *Öznur İşçi Güneri.*

KAYNAKLAR

1. Kim M. Two-stage logistic regression model. Expert Syst Appl 2009;36(3):6727-34.
2. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics 1936;7(2):179-88.
3. Tatlıdil H. [Applied Multivariate Statistical Analysis]. [Diskriminant (Ayırma) Analizi]. 1. Baskı. Ankara: Cem Web Offset; 2002.p.256-89.
4. Hosmer DW, Lemeshow S. Applied Logistic Regression. . Introduction to the Logistic Regression Model.1st ed. New York: John Wiley and Sons; 1989.p.5-50.
5. Kleinbaum GD, Klein M. Intraduction to logistic regression. A Self-learning Text Logistic Regression. 3rd ed. New York: Springer; 1994. p.1-8.
6. Worth AP, Cronin MTD. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. J Mol Struc (Theochem) 2003;622(1-2):97-111.
7. Kachigan SK. Statistical Analysis: An interdisciplinary intoduction to univariate and multivariate methods. Discriminant Analysis. 1st ed. New York: Radius Press; 1986. p.357-75.
8. Manly FJB. Discriminant function analysis. Multivariate Statistical Methods. 2nd ed. London: Chapman-Hall; 1994. p.107-125.
9. Oğuzhan A, Aydın D. [Discrimination analysis of relationships between different apodemus species living in Trakya and Western Anatolia]. University of Trakya Journal of Social Sciences 2014;1(1):4-7.
10. Tabachnick BG, Fidell S. Using Multivariate statistics. Logistic Regression. 6st ed. Boston: Allyn and Bacon; 2013.p.432-79.
11. Cox DR, Snell EJ. Analysis of binary data. Introduction. 2st ed. London: Chapman and Hall/CRC;1989.p.10-1.
12. Ledolter J. Data mining and business analytics with R. Multinomial Logistic Regression. 1st ed. Newyork: John Wiley and Sons; 2013.p.132-47.
13. Temel GO, Çamdeviren H, Akkuş Z. [Diagnosis for the Restless Legs Syndrome (RLS) Patients by means of Classification Trees]. Journal of Inonu University Medical Faculty 2015;12(2):111-7.
14. Deconinck E, Hancock T, Coomans D, Massart DL, Heyden YV. Classification of drugs in absorption classes using the classification and regression trees (cart) methodology. J Pharm Biomed Anal 2005;39(1-2):91-103.
15. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl 2008;34(1):366-74.

16. Teng JH, Lin KC, Ho BS. Application of classification tree and logistic regression for the management and health intervention plans in a community-based study. J Eval Clin Pract 2007;13(5):741-8.
17. Ayık YZ, Özdemir A, Yavuz U. [Analysis with data mining techniques of the relationship earned high school graduation success and type of high school]. Ataturk University Journal of Social Sciences Institute 2007;10(2):441-54.
18. Berry MJ, Linoff GS. Why and what is data mining? Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 2nded. New York: John Wiley & Sons; 2004. p.1-10.
19. Güner ZB. [Cart and logistic regression analysis in data mining: an application on pharmacy provision system data]. Sosyal Güvenlik Uzmanları Derneği Sosyal Güvence Dergisi 2014;6:59-61.
20. Silahtaroğlu G. [Basic data mining with concepts and algorithms]. Karar ağaçları. 2. Baskı İstanbul: Daisy Publishing Training; 2008.p.45-7.
21. Grabczewski K. Validated Decision Trees versus Collective Decisions. In: Jędrzejowicz P., Nguyen N.T., Hoang K. (eds) Computational Collective Intelligence. Technologies and Applications. ICCCI 2011. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg; 2011. vol 6923. p.342-351.
22. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. The Multiclass Problem: Unit Costs. 1st ed. Belmont, California: Taylor & Francis; 1984. p.103-8.