T.C.
MERSİN ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

SINIFLAMA ve REGRESYON AĞAÇLARI

Gülhan OREKİCİ TEMEL

YÜKSEK LİSANS TEZİ

DANIŞMAN
Yrd. Doç. Dr. Handan ÇAMDEVİREN

MERSİN - 2004

Mersin Üniversitesi Sağlık Bilimleri Enstitüsü

Biyoistatistik Yüksek Lisans Programı Çerçevesinde yürütülmüş olan Sınıflama ve Regresyon Ağaçları adlı çalışma, aşağıdaki jüri tarafından Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Savunma Tarihi..09/02/2004

Yrd.Doç.Dr.Arzu KANIK
Mersin Üniversitesi
Jüri Başkanı

Yrd.Doç.Dr.Handan ÇAMDEVİREN
Mersin Üniversitesi
Jüri Üyesi

Yrd.Doç.Dr.Resul BUĞDAYCI
Mersin Üniversitesi
Jüri Üyesi

Yukarıdaki tez, Enstitü Yönetim Kurulunun 17.02.2004. tarih ve 2004/27 sayılı kararı ile kabul edilmiştir.

Enstitü Müdürü

TEŞEKKÜR

Mersin Üniversitesi Sağlık Bilimleri Enstitüsünde tamamlamış bulunduğum "Sınıflama ve Regresyon Ağaçları" başlıklı yüksek lisans tezimin hazırlık sürecinde ve çalışma aşamalarında yönlendirme ve bilimsel desteği ile katkıda bulunan danışman hocam Sayın Yrd.Doç.Dr. Handan Çamdeviren'e sonsuz teşekkürlerimi sunarım. Ayrıca Anabilim Dalı Başkanı olarak desteğini esirgemeyen Sayın Yrd.Doç.Dr. Arzu Kanık'a gerek önerileri gerekse akademik çalışma ortamını hazırlaması dolayısıyla teşekkür ederim. Tezin uygulama bölümünde kullanılan veri setinin temininde Sayın Yrd.Doç.Dr. Serhan Sevim' e, ayrıca çalışmamın başından sonuna kadar konuyla ilgili bilgi birikimini benimle paylaşan çalışma arkadaşım Arş.Gör.Dr. Tevfik Aytemiz'e ve göstermiş olduğu anlayış ve sabırdan dolayı sevgili eşim Devrim Temel'e, bütün öğrenim hayatımı borçlu olduğum sevgili aileme sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

ÖZET

Sınıflama ve Regresyon Ağaçları

Sınıflama ve Regresyon Ağaçları (Classification and Regression Tree, CART) parametrik olmayan bir metoddur. Herhangi bir objenin sınıf üyeliğini bir veya daha fazla bağımsız değişken kullanarak tahmin etmeye yarayan ağaç algoritmasıdır. Sınıflama ve Regresyon Aağaçlarında bağımlı değişken sürekli ise regresyon ağacı, bağımlı değişken kategorik ise sınıflama ağacı ismini alır. CART analizi, bir sınıflama ağacı oluşturarak obje veya bireylerin gelecekte hangi sınıfa gireceklerini belirler. Model kurulduktan ve ağaç oluşturulduktan sonra modelin yorumlanması ve kullanımı oldukça basittir.

CART analizi ülkemizde çok sık kullanılan bir analiz tekniği değildir. Bu noktadan hareketle bu tezin amacı; CART hakkında önemli teorik bilgileri özetlemek ve söz konusu metodun tıpta uygulanabilirliğini, tanı koyma problemine uygun bir veri seti kullanarak göstermektir. Bu amaçla, Nöroloji bölümünün 206 denek üzerinde yaptığı anket çalışmasının sonuçları kullanılmış ve deneklerin RLS (Restless Legs Symptoms) hastası olup olmama durumuna etki eden değişkenler sınıflama ağaçları analizi ile tespit edilmiştir. Analiz sonuçlarına göre, RLS hastalığını belirleyen değişkenler literatürde yer alan faktörlerle paralellik göstermektedir. CART hesaplamalarında Statistica®6.0 paket programı kullanılarak analiz edilmiş ve sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Sınıflama ve Regresyon Ağaçları (CART), Sınıflama, Tahmin, Karar ağaçları, Hatalı sınıflama oranı.

ABSTRACT

Classification and Regression Trees

Classification and Regression Trees (CART) belong to the class of non-parametric methods. They are tree algorithms that forecast the class membership of an object with one or more independent variables. In the case of continuous dependent variable CART produce a regression tree; otherwise (i.e., categorical dependent variable) they produce a classification tree. CART analysis constructs a classification or regression tree that enables one to forecast the unknown class membership of an object in the future. Once the model is built and the classification tree is constructed, it is then very easy to use and examine the model.

CART is not a very common technique in our country. From this point, the objectives of this thesis is to provide a summary of theoretical background on CART, and to demonstrate the applicability of this method on medical sciences, using an appropriate data set for diagnostics problem. For this purpose, data is collected from a questionnaire research conducted by the Neurology Department on 206 samples, and significant factors affecting the RLS (Restless Legs Symptoms) were determined using the classification trees. Statistica®6.0 computer software was used for the analysis. Significant factors that resulted from this analysis on RLS syndrome agree with the literature research.

Keywords: Classification and Regression Trees (CART), Classification, Forecasting, Classification Trees, Misclassification Rate.

1. GİRİŞ

Kategorik ya da sürekli, bir ya da birden fazla bağımsız değişkenin kombinasyonları kullanılarak, tekrarlamalı ikili homojen bölünmelerle, bağımlı değişkendeki değişimi ortaya çıkarmaya ve bağımlı değişkenin değerlerini tahmin etmeye yarayan ve görsel olarak ters ağaç şeklindeki modellere ağaç modelleri denir (1).

İstatistiksel verilerin görsel olarak sunulması, aralarındaki etkileşimin belirlenebilmesi ve bu etkileşimden yararlanılarak tahminler yapılabilmesi için karar ağaç modelleri sıkça kullanılmaktadır.

Ağaç modellerinin işleyiş yapısı, bağımsız değişkene ait temel basit sorulardan alınan cevapların yarattığı yolları (ağaç dalları) takip etmektir. Ağaç dalları, bağımlı değişkeni hangi bağımsız değişken ya da değişkenlerin etkilediğini gösterir.

Sınıflama ve Regresyon Ağaçları (Classification and Regression Tree, CART) bağımsız değişkene ilişkin hiçbir ön koşul öne sürmeden, kategorik ya da sürekli bağımlı değişkenin sınıf üyeliğini tahmin eden ters ağaç şeklindeki modellerdir (2).

Bilimsel çalışmalardan elde edilen verilerin analizinde sınıflama (diskriminant, lojistik regresyon, kümeleme analizleri gibi) ve regresyon modelleri sıkça kullanılmaktadır. CART bu tür istatistiksel sınıflama ve regresyon tekniklerine alternatif bir metoddur. Veri seti çok karmaşık olsa bile bağımlı değişkeni etkileyen değişkenler ve bu değişkenlerin modeldeki önemi, karmaşık bir model kurmadan görsel sunumla yapılabilir.

CART analizi son yıllarda yurt dışında yaygın olarak kullanılmasına karşın ülkemizde sık kullanılan bir analiz tekniği değildir. Bu noktadan hareketle bu tezin amacı; CART hakkında önemli teorik bilgileri özetlemek ve söz konusu metodun uygulanabilirliğini tıpta tanı koyma problemine uygun bir veri seti kullanarak göstermektir. Bu amaçla bu tezde RLS (Restless Legs Symptoms) hastası olup olmama durumuna etki eden değişkenler sınıflama ağaçları kullanılarak tespit edilmiştir. İncelenen bağımlı değişkenin iki seviyeli kategorik değişken olması nedeniyle sınıflama ağaçları daha detaylı bir şekilde ele alınarak regresyon ağaçları sadece teorik çerçevede incelenmiştir.

Bu tezin çalışma planı şu şekildedir. İkinci bölümde; örnek bir karar ağaç modeli üzerinde karar ağaç modellerinin genel yapısı tanıtılmış, ağaç modellerinin amaçları üzerinde durulmuştur. Üçüncü bölümde; CART tanıtılmış, tarihsel gelişimi, kullanım alanları, avantaj ve dezavantajları anlatılmış ve alternatif sınıflama modellerine değinilmiştir. Dördünü bölümde; sınıflama modeli kurulurken, kullanılabilecek ayırma kriterleri, atama kuralları ve sınıflama ağaçlarının doğruluk tahminleri üzerinde durulmuştur. Dördüncü bölümde; regresyon ağaçlarının oluşumu, ayırma kuralları ve regresyon ağaçlarının doğruluk tahminleri üzerinde durulmuştur. Son bölümde CART analizinin uygulamasını göstermek amacıyla Mersin Üniversitesi Tıp Fakültesi Hastanesinden uygun bir veri seti seçilerek Statistica®6.0 paket programı kullanılarak analiz edilmiş ve sonuçlar yorumlanmıştır.

2. TANIM VE TEORİK TEMELLER

2.1. Örnek Ağaç Modeli

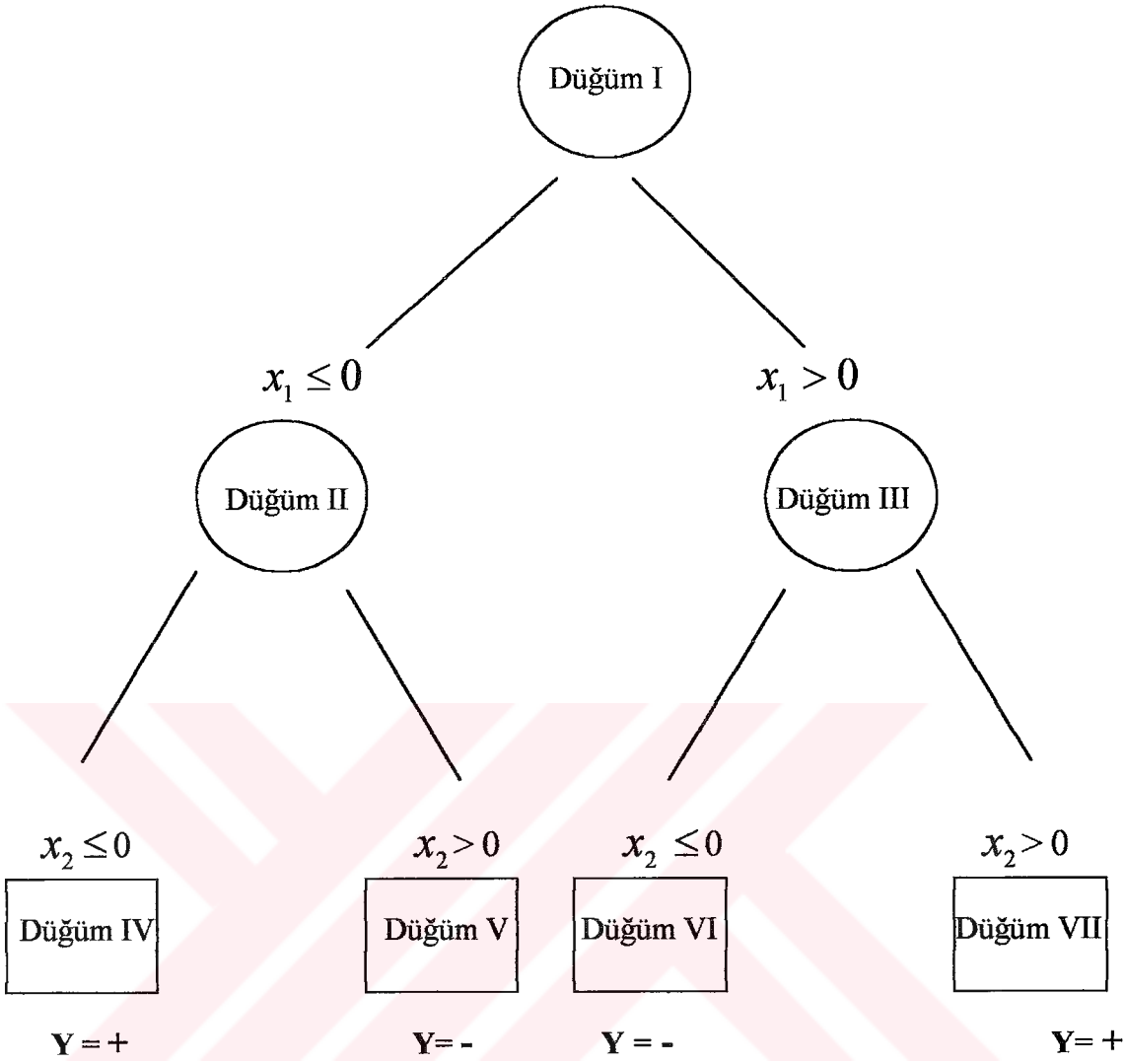
Giriş bölümünde tanımı verilen ağaç modeline uygun bir örnek aşağıdaki gibi gösterilebilir.

Basit bir ağaç modelinde, x_1 ve x_2 ; [-1;+1] aralığında değişen düzgün (uniform) olasılık dağılımdan tesadüfi olarak seçilen n_1 ve n_2 büyüklükteki örneklerin içerdiği bağımsız değişkenler olduğunda, bağımlı değişken Y ise bağımsız değişken değerlerine çarpma kuralı uygulanarak elde edilen sonuç değişkeni ise; çarpma kuralına göre;

$x_1 \cdot x_2 \geq o$ ise Y 'nin değeri pozitiftir.

$x_1 \cdot x_2 \leq o$ ise Y 'nin değeri negatiftir.

Böylece bağımlı değişkenin pozitif ve negatif olmak üzere iki seviyesi vardır. Ölçme düzeylerine bakılmaksızın, Şekil 2.1 yukarıdaki örneğin ağaç yapısını göstermektedir.



Şekil 2.1.: Örnek bir ağaç modeli

Ağaç modellerinde karar verme noktalarına düğüm denir. Şekil 2.1.'deki ağaç modelinde başlangıç düğümü Düğüm I'dir (3). Bu düğüm gözlem değerlerinin tümünü içine aldığından en karmaşık düğümdür. Bu düğüme aile düğümü yada kök düğümü de denir. Kök düğümü iki alt çocuk düğümüne bölünür. Örnek ağaç modelindeki çocuk düğümleri Düğüm II ve Düğüm III'tür. Tersten ifade edilecek olursa, her aile düğümü çocuk düğümlerinin yükselmesinden meydana gelir. İkili olarak bölünen çocuk düğümlerinin birleşmesinden aile düğümü meydana gelir. Yani her çocuk düğümü bir sonraki adımda aile düğümü olur. Bu ifade simgesel olarak aşağıdaki gibi ifade edilir.

Düğüm I = Düğüm II $\cup$ Düğüm III

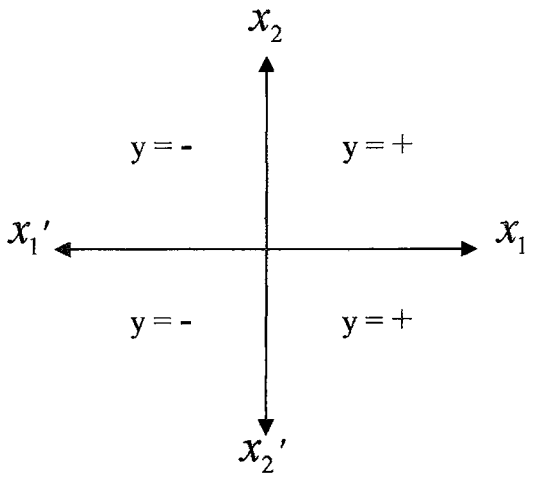
Çocuk düğümlerinde henüz karar verme gerçekleşmemiştir ve dolayısıyla çocuk düğümleri henüz saf değildir. Aile düğümünden her çocuk düğümüne bölünme gerçekleştiğinde çocuk düğümü aile düğümüne göre daha homojendir. Amaç çocuk düğümlerini ayırma kriterleri doğrultusunda tekrarlamalı ikili parçalara ayırarak karar noktalarına yani terminal düğümlere ulaşmaktır. Terminal düğümlerde obje ya da nesnelerin sınıf üyelikleri tanımlanır ve düğümde var olan obje ya da nesnelerin sapmasının sıfır olduğu kabul edilir (4). Dolayısıyla terminal düğümlerinin ağaçtaki en homojen düğümler olduğu söylenebilir. Terminal düğümlerden sonra bölünme gerçekleşmez.

Ağaç modellerinde genellikle çocuk düğümleri dairelerle, terminal düğümler ise kare ile gösterilir. Örnek ağaç modelinde Düğüm IV, Düğüm V , Düğüm VI ve Düğüm VII terminal düğümlerdir.

Ağaç modellerinde temel amaç; başlangıç düğümünden başlayarak ikili tekrarlı ayırmalarla daha homojen alt gruplara ulaşıp karar noktalarında bağımlı değişkenin durumunu tanımlamaktır. Bir başka ifade ile her bir düğüm kendi içinde homojen ikili bölünmelere uğrar ve bu süreç ağaç inşasının sonuna kadar sürdürülür. Bu süreçte, sınıflama (karar) ve regresyon ağaçlarındaki düğüm noktalarında yer alan gözlemler sahip oldukları bağımsız değişkenin değerlerine göre iki çocuk düğümünden birine atanırlar. Düğüm I'deki gözlemler eğer sıfırdan küçük ya da sıfıra eşit iseler Düğüm II'ye aksi takdirde Düğüm III'e atanırlar. Aynı şekilde Düğüm II ve Düğüm III'deki gözlemler ait oldukları bağımsız değişken değerine göre Düğüm IV, Düğüm V, Düğüm VI veya Düğüm VII'ye atanırlar.

Şekil 2.1'den görüleceği üzere bağımlı değişkene ait grupları sınıflandırırken genellikle birden fazla bağımsız değişken kullanılmaktadır. Böylece, bağımsız değişkenler arasında varolan interaksiyon (etkileşim) ağaç modelleriyle görülebilmektedir (5).

Bağımsız değişken sayısı iki ya da üç olduğunda görsel sınıflara ayırma kolaylıkla yapılabilir. Bağımsız değişken sayısı 5'in üstünde olduğu durumlarda bağımlı değişkenin değerlerinin grafikle tahmin edilmesi hemen hemen imkansızdır.



Şekil 2.2: İki bağımsız değişken arasındaki etkileşimin grafik gösterimi.

2.2. Ağaç Modellerinin Amaçları

Verilerdeki varyasyonun açıklanması ve modelin elde edilmesinden sonra tahminlerin yapılabilmesi için istatistiksel analizlere başvurulur. Ağaç modelleri, verilerin karmaşık ilişkilerini görsel olarak sunar ve istatistiksel özet bilgilerini verir. Ağaç modelleri, karar vericiye problemin ayrı ayrı her bir aşamasını ağaç üzerinde inceleme olanağı verir ve karmaşık problemlerin olası alt kümelerinin aşamalı olarak değerlendirilmesine olanak sağlar. Aynı zamanda bu modeller, ağaç yapısından yararlanarak obje veya bireylerin (deney ünitesi) gelecekte hangi risk sınıfına gireceklerini belirleme ve bağımsız değişkenlerin etkilerini aşamalı olarak açıklama şansı verir.

Ağaç modellerinin *sınıflama* ve *tahmin* olmak üzere iki temel amacı vardır. Sınıflama, basit veri yapısını sistematik bir şekilde modelde sunar. Tahmin ise gözlenmeyen verileri bu modelden güvenli olarak tahmin eder. Sınıflama ve tahmin özelliklerinden dolayı ağaç modelleri, hem verilerin tanımlanmasında hem de gelecekteki verilerin tahmin edilmesinde kullanılabilir.

3. SINIFLAMA VE REGRESYON AĞAÇLARI

3.1.Tanım

Sınıflama ve Regresyon Ağaçları (CART) kategorik yada sürekli bağımlı değişkenlerin alacağı değerleri analiz ve tahmin etmek üzere geliştirilen parametrik olmayan istatistiksel bir metodolojidir. Bağımlı değişken kategorik ise CART sınıflama ağacı ismini, bağımlı değişken sürekli ise CART regresyon ağacı ismini alır. (6). CART modelleri tekrarlanabilen tahmin ediciler uzayının homojen tekrarlı ikili alt gruplara ayrılması üzerine kurulmuş olan karar ağaçları inşa ederler (7). İkili alt gruplara ayrıştırma karar noktalarına kadar devam eder.

CART analizi bir mekanik öğrenme metodudur. CART analizi geleneksel veri analiz tekniklerine benzemeyen ağaç inşası tekniğidir(5). CART bir aile düğümünden başlayan, ikili ayırmalar dizisinden oluşan, ayırmaların terminal düğümlere kadar devam ettiği bir metot olarak tanımlanabilir (8).

3.2. CART'ın Tarihsel Gelişimi

CART metodolojisi ilk defa 1963 yılında Morgan ve Sonquist tarafından ortaya konulmuştur (9). Yaklaşımdan bu zamana geliştirilen tahmin metodlarının gelişim süreci aşağıdaki gibidir.

Henrichon ve Fu (1969), Meisel ve Michalopoulos ikili karar ağaçlarının (learning sample'daki sınıfların tekrarlamalı olarak deneye dayalı stratejilerle ortaya çıkarılması) tanımını ortaya çıkarmıştır. Brieman ve Stone (1978) budama fikri ile en uygun ağacı seçmek için minimal cost complexity (en az karmaşa maliyeti) yöntemini geliştirmiştir. Gorden ve Olshen (1980) bir Oklit Gözlem Uzayının parçalara ayrılması üzerine kurulmuş karar kurallarını önermiştir. Mabbet, Stone ve Washbrook (1980) sınıflama ağaçlarının budanması ve en uygun ağacın seçimi konusunda cross-validation metodunun kullanılmasını önermiştir (10).

1984 Yılında Breiman, Friedman, Olshen ve Stone yazmış olduğu "Classification and Regression Trees" isimli kitap ile sınıflama ve regresyon ağaçları güvenilir ve yararlı bir analiz halini almıştır.

3.3. Kullanım Alanları

Sağlık Bilimlerinde son 20 yıl içerisinde CART tekniği büyük bir gelişme göstermiştir. Sınıflama ve Regresyon Ağaçları yaygın olarak tıp biliminde (tanı koyma ve tahmin), botanikte (sınıflama) ve karar teorisinde kullanılmıştır. Ayrıca ekonomik olarak risk altındaki firmaların sınıflandırılması için Frydman, Altman ve Kao tarafından 1985 yılında finans alanında, ticari borçların sınıflandırılması için Marais, Patell ve Wolfson tarafından 1985 yılında kullanılmıştır (11).

Uluslararası Gıda Politikası Araştırma Enstitüsü (IFPRI) CART'ı bölgesel ve hane halkı düzeyindeki kıtlık belirlemesi için 1985 yılında kullanmıştır(11). Ekolojik verilerin değerlendirilmesi için 1992 yılında Staub, 1993 yılında Baker ve 1999 yılında Rejwan tarafından kullanılmıştır (12).

3.4. CART' ın Avantajları

Sınıflama ve regresyon ağaçlarını (CART) cazip bir model haline getiren temel nedenler şunlardır.

- CART parametrik olmayan bir modeldir.
- Modelde değişkenlerin türü (sürekli, kategorik, sıralı ya da bunların karışımı) konusunda herhangi bir varsayım yoktur (2).
- Modelde bağımlı ve bağımsız değişkenlerin dağılımı ile ilgili bir varsayım gerektirmediğinden değişkenlere logaritma, karekök gibi dönüştürmelerin uygulanmasına gerek kalmaz.
- Bağımlı ve bağımsız değişkenler arasındaki ilişki görsel sunuma sahip olduğundan ağaç şeklindeki model sonuçları çok fazla istatistik bilgisine gerek duyulmadan kolay bir şekilde yorumlanabilir.
- CART, tanımlanan bağımlı değişken için olabilecek bütün bağımsız değişkenleri ve onların tüm kombinasyonlarını modele katar ve mümkün olan en doğru sınıflandırmayı yapar. Değişkenlerin kombinasyonlarına da bakıldığı için model, esnek ve daha geniş bir bakış açısı ile elde edilebilir. Böylece bağımlı değişkeni etkileyen önemli bağımsız değişkenlerde belirlenmiş olur.
- Çok karmaşık veri setlerinde bile doğru tahmin yapabilir.
- Hem bağımlı hem de bağımsız değişkenler için kayıp veya eksik değerler ile aşırı uç değerlerden etkilenmeyen bir metottur.

- Geleneksel birçok istatistik tekniğine (çoklu regresyon, varyans analizi, logistik regresyon, diskriminant analizi, kümeleme analizi) alternatiftir.
- Kesin olmayan ancak sağlam temellere dayanan ağaç metotlarını da hesaba katar (5).
- Analizciye metot sıralamasını düzeltme olanağı tanır (5).
- Model bağımlı değişkeni etkileyen bağımsız değişkenleri ve bağımlı değişkenlerin bağımsız değişkenlerle arasındaki interaksiyonu (etkileşim) ortaya çıkarır.
- Eğer ihtiyaç duyulursa aynı bağımsız değişken aynı ağaçta farklı ayırma değerleriyle (cut off) kullanılabilir.

Sınıflama ve regresyon ağaçları analizinin de sınırlamaları vardır. CART tekniğinin en önemli zayıflığı sonuçların bir olasılık modeline dayanmıyor olmasıdır. Veri setine uygun bir CART ağacından alınmış tahmini sınıflandırmaya yardım edebilecek bir olasılık derecesi ya da güven aralığı yoktur. CART tarafından üretilen sonuçların doğruluğuna duyulabilecek güven tamamen geçmiş verilere dayalı doğruluğuyla orantılıdır (2).

CART klasik bir analiz tekniği değildir. İstatistikçilere CART'ın yeterli güvenirliliğe sahip olduğunu kabul ettirmek genellikle zordur (5).

3.5. Diğer Sınıflama Metotları

CART'ın dışında birçok sınıflandırma metodu vardır. Bunlar iki gruba ayrılır:

Grup I: AID (Automatic Interaction Detection), THAID (Theta Automatic Interaction Detection), CHAID (Chi_square Automatic Interaction Detection), QUEST (Quick, Unbiased, Effecient Statistical Trees), FACT (Fast and Accurate Classification Tree).

Grup II: Diskirminant Analizi, Logistik Regresyon, Probit Modeller, Yapay Sinir Ağları.

Grup I'deki metotların temeli sınıflandırma ağaçlarıdır. Sosyal ve ekonomik olayları, daha güvenilir bir şekilde gösterebilmek için standart istatistik tekniklerin dışında yeni analiz tekniklerinin geliştirilmesi ile ilgilenen Morgan ve Sonquist

tarafından University of Michigan'da 1970'li yılların başlarında kullanıma alınan *Automatic Interaction Detector* – AID karar ağacı temelli ilk algoritma ve yazılımdır. AID tekniği en kuvvetli ve en iyi tahmini gerçekleştirebilmek için bağımlı ve bağımsız değişkenler arasındaki mümkün bütün ilişkilerin incelenmesine dayanmaktadır. Karar ağacı tekniğinin sağladığı kuruluş ve yorumlama kolaylıkları, AID yazılımının başlangıçta istatistikçiler ve veri analistleri tarafından büyük coşku ile karşılanmasına neden olmuştur (13).

1980 yılında G.V. Kass tarafından geliştirilen CHAID algoritmasında, bağımlı değişkeni en fazla etkileyen bağımsız değişkenler üzerinde durarak populasyonu gruplara ve daha alt gruplara ayırır (14). Bu algoritma hem bağımlı hem de bağımsız değişken kategorik olduğunda uygulanabilir (15).

QUEST algoritması 1997 yılında Loh ve Shin, FACT algoritması ise 1988 yılında Loh ve Vanichestakul tarafından ortaya atılmışlardır. QUEST sürekli değişkenler için ANOVA F istatistiğini, kategorik değişkenler için χ^2 testini kullanır. FACT ise tüm değişken türleri için ANOVA F istatistiğini kullanır (16).

CART ve QUEST 'e göre FACT kategorik değişkenlerde daha çok ön yargılara (bias) yer verir. FACT ve QUEST kategorik değişkenler için CART'a göre daha hızlı sonuç verir. Ancak, bağımlı değişkenin sınıf sayısı ikiden fazla olduğunda CART daha hızlıdır (16).

4.SINIFLAMA AĞACININ OLUŞTURULMASI

Sınıflama ağaçlarının (classification trees) temel amacının, bağımsız değişkenlerin belli değerler alması durumunda bağımlı değişkenin alacağı değeri tahmin etmek olduğunu belirtmiştik. Sınıflama ağaçları, bu amacı gerçekleştirmek için, N adet deney ünitesine ait ölçüm bilgilerini içeren bir başlangıç veri setini kullanır. Bu veri setine *Learning Sample* denir ve terminolojide L ile gösterilir.

$X = \{x_1, x_2, ..., x_n, ..., x_N\}$

$J = \{j_1, j_2, ..., j_n, ..., j_N\}$

$L = \{(x_1, j_1), (x_2, j_2), ..., (x_n, j_n), ..., (x_N, j_N)\}$

$x_n \in X$; $j_n \in \{1,2,...,C\}$ ve $n = 1,...,N$'dir.

Burada ;

L : Başlangıç veri seti (bağımsız değişkenlere ait bilgilerinin tümünü içine alan ölçüm vektörüdür) .

X : Bağımsız değişkenlere ait ölçüm vektörü.

J : Deney ünitelerinin ait olduğu sınıflar (kategoriler) vektörü.

x_n : n. birey yada objeye ait bağımsız değişkenin değeri.

j_n : n. birey yada objenin ait olduğu sınıf (kategori).

C : Deney ünitelerinin ait olduğu sınıf (kategori) sayısı.

N : Toplam deney ünite sayısıdır.

Ölçüm vektöründe yer alan değişkenler gerçel sayılardan oluşuyorsa o vektör sürekli değişkenler vektörü, sayımla belirtilen veya yapay verilerden oluşuyorsa kategorik değişken vektörü olarak adlandırılır.

Sürekli değişkene tıbbi verilerden örnek olarak boy uzunluğu, kandaki hemoglobin miktarını kategorik değişkene örnek olarak cinsiyet, diyabet varlığı, tümör evreleri verilebilir.

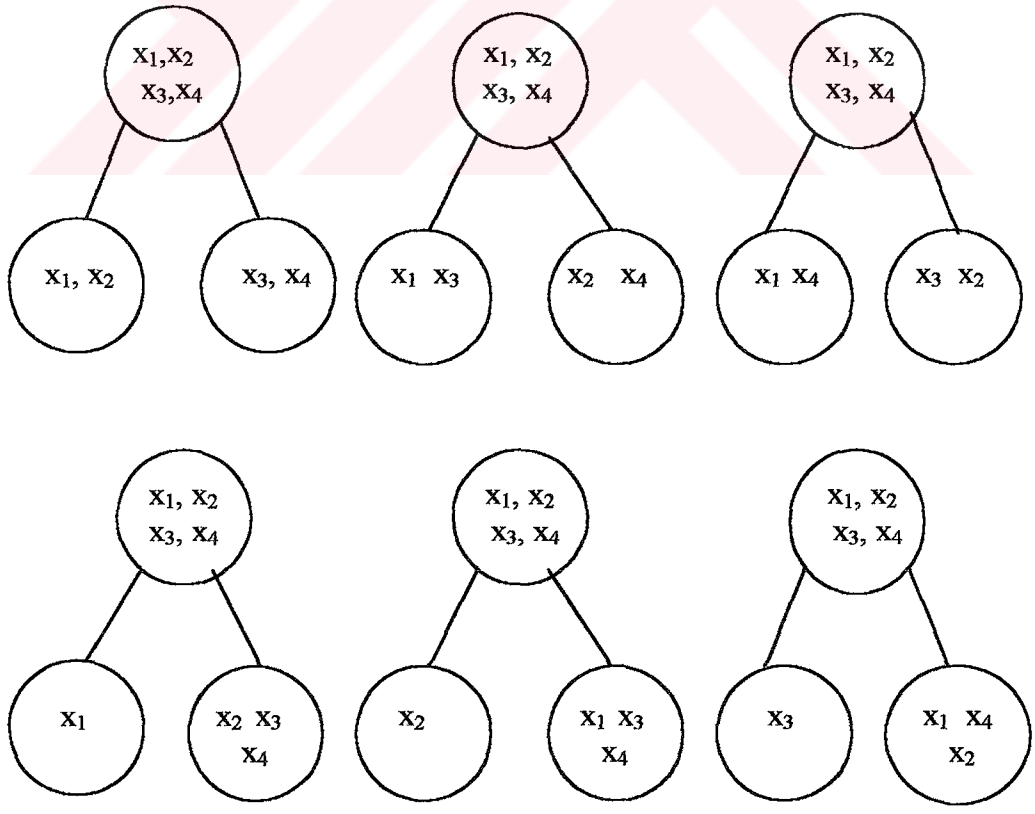
Sınıflama ağaçları, kök düğümden başlayarak devam eden ve her düğümde o düğüme ait deney ünitelerine uygulanan basit sorulardan alınan evet/hayır cevaplarının yarattığı yollardan oluşur. Her düğümde uygulanan bu sorulara *ayıraç* denir. Her ayıraç t düğümüne ait deney ünitelerini evet/hayır cevaplarına göre iki alt kümeye ayırır. Bu işlem *ayırma* olarak adlandırılır ve terminolojide $\delta(t)$ ile gösterilir. Sınıflama

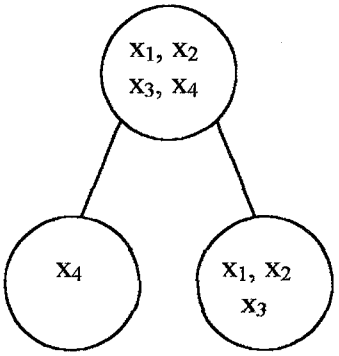
ağaçlarında ayırmalar X bağımsız değişkeninin sürekli veya kategorik oluşuna göre iki şekilde tanımlanabilir.

Eğer X sürekli bir değişken ise ayıraç "s (cut off değeri) reel bir sayı olmak üzere, $x_n \leq s$ midir?" sorusudur. Deney üniteleri eğer soruya alınan yanıt evet ise sol düğüme, hayır ise sağ düğüme atanır. Söz konusu sürekli bağımsız değişkenin k adet farklı değeri var ise bu durumda en fazla k-1 adet ayıraç ve dolayısı ile k-1 adet farklı ayırma mümkündür.

Eğer X kategorik bir değişken ise ayıraç "$A \subset X$ olmak üzere, $x_n \in A$ mıdır?" sorusudur. Deney üniteleri eğer soruya alınan yanıt evet ise sol düğüme, hayır ise sağ düğüme atanır. Söz konusu kategorik bağımsız değişken k adet farklı kategori içeriyorsa bu durumda X'in en fazla 2^{k-1} adet (boş küme hariç) alt kümesi vardır ve dolayısı ile 2^{k-1}-1 adet farklı ayırma mümkündür (12).

Örnek olarak, X tesadüfi kategorik bir bağımsız değişken ve bu değişkenin aldığı değerler $\{x_1, x_2, x_3, x_4\}$ (k=4) olsun. Bu durumda, mümkün olan ayırmalar aşağıdaki gibidir.





Şekil 4.1: Bir Sınıflama Ağacında Mümkün Olan Ayırmalar

Tek bir bağımsız değişkene sahip sınıflama ağaçları için yukarıda yapılan tanımlamalar birden fazla bağımsız değişkene sahip sınıflama ağaçları için de aynen kullanılabilir. Deney ünitelerinin birden fazla bağımsız değişken içermesi durumunda değişen tek şey ayıraçların tüm değişken ve değişken kombinasyonlarını tek tek ele almasıdır. Bu durumda, deney ünitelerinin içerdiği bağımsız değişkenler ve bu değişkenlerin birbirleri ile kombinasyonlarının tanımlı bulunduğu aralıklardaki tüm olası değerler birer ayıraç olarak düşünülüp, mümkün olan tüm olası ayırmalar belirlenir.

Sınıflama ağaçlarında, herhangi bir t düğümünde mümkün olan tüm olası ayırmalar belirlendikten sonra, her bir olası ayırma için *ayırmanın uygunluk derecesi* hesaplanır. Herhangi bir t düğümünde olası bir ayırmanın, $\delta_i(t)$, uygunluk derecesi hesaplanırken *ayırma fonksiyonu* kullanılır. Ayırma fonksiyonu matematiksel olarak aşağıdaki şekilde gösterilir;

$$\Delta(\delta_i(t)) = i(t) - P_L\, i(t_L) - P_R\, i(t_R)$$

Burada;

P_L : t. düğümden sol çocuk düğümüne atanan deney ünitelerinin (gözlemlerin) oranı.

P_R : t. düğümden sağ çocuk düğümüne atanan deney ünitelerinin (gözlemlerin) oranı.

$i(t)$: t düğümünün safsızlık ölçüsü.

$i(t_L)$: Sol çocuk düğümünün safsızlık ölçüsü.

$i(t_R)$: Sağ çocuk düğümünün safsızlık ölçüsüdür (17).

Her bir olası ayırmanın uygunluk derecesi, ayırma fonksiyonu yardımıyla hesaplandıktan sonra, maksimum uygunluk derecesine sahip ayırma, $\delta^*(t)$, *en iyi ayırma* olarak seçilir ve t düğümü bu şekilde ayrılır.

Herhangi bir düğümün heterojenlik değeri *safsızlık (impurity) ölçüsü* olarak adlandırılır ve bu değer *safsızlık fonksiyonu* kullanılarak hesaplanır. Safsızlık ölçüsü sıfır değerini alıyorsa düğüm tamamen homojendir.

Kategorik bağımlı bir değişken için sınıf numaraları j = 1,2...,k olsun. Herhangi bir t düğümü için safsızlık ölçüsü matematiksel olarak aşağıdaki gibi tanımlanabilir;

$i(t) = \Phi\{p(1|t), p(2|t), ...,p(k|t)\}$.

Burada;

i(t) : t düğümünün safsızlık ölçüsü.

Φ : Safsızlık (impurity) fonksiyonu.

p(j|t) : t. düğümde, bağımlı değişkenin j. sınıfına atanan deney ünitelerinin oranıdır ($\sum_{j=1}^{k} p(j|t) = 1$).

Eğer t. düğümdeki her sınıf eşit orana sahip olursa safsızlık fonksiyonu maksimum değerine ulaşır ($\Phi\{p(1|t), p(2|t),...,p(k|t)\} = \Phi\{1/k, 1/k,..., 1/k\}$ = maksimum).
Aynı şekilde, eğer t. düğüm tek bir sınıfa ait deney ünitelerini kapsarsa safsızlık fonksiyonu minimum değerine ulaşır ($\Psi = \{x; (1, 0,..., 0),(0, 1,..., 0),..., (0, 0,..., 1)\}$ ve $\Phi\{p(1|t), p(2|t),...,p(k|t)\} = 0, \forall x \in \Psi$).

Sınıflama ağaçlarında kullanılabilecek birçok alternatif safsızlık ölçüsü (Gini, Twoing, Chi-square, G-square) vardır. Ayırma fonksiyonundan anlaşılacağı gibi, kullanılan safsızlık ölçüsü herhangi bir t düğümü için en iyi ayırmanın seçimini önemli bir şekilde etkilemektedir. Bu nedenle safsızlık ölçüleri literatürde *en iyi ayırma kriterleri (ya da ayırma kuralları)* olarak da bilinirler. En yaygın olarak kullanılan ayırma kriterleri Gini Diversity Index (Gini) ve Twoing Kuralı'dır.

4.1. En Yaygın Kullanılan Ayırma Kriterleri

4.1.1. Gini Diversity Index (Gini)

Sınıflama ağaçlarında herhangi bir düğümde mümkün olan en iyi ayırmanın bulunmasında yaygın olarak kullanılan Gini, veri tabanındaki en geniş sınıfı diğer bütün sınıflardan ayırmaya çalışır. Matematiksel olarak Gini safsızlık ölçüsü aşağıdaki şekilde tanımlanabilir;

$$i(t) = \sum_{i \neq j} p(i \mid t) p(j \mid t)$$

$$= \sum_{j} p(j \mid t)(1 - p(j \mid t))$$

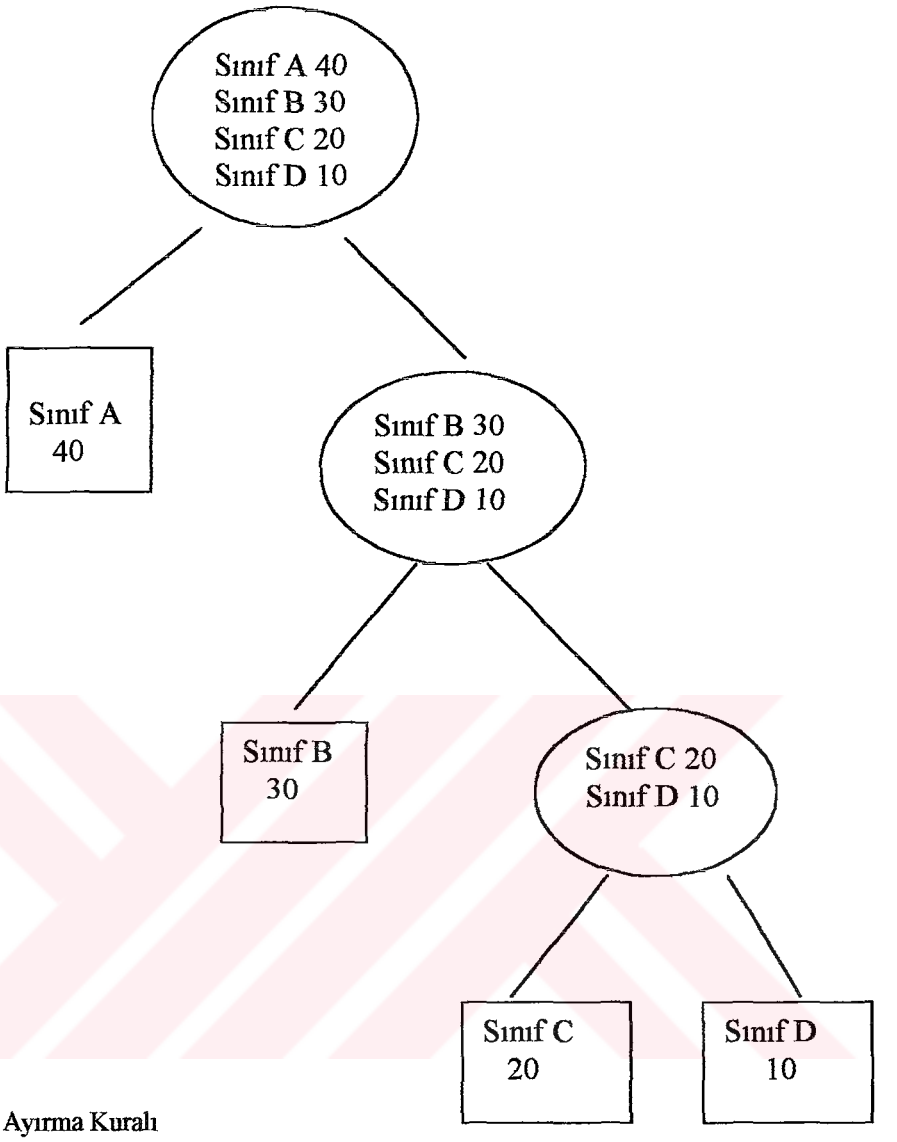
$$= 1 - \sum_{j} (p(j \mid t))^2 \text{ (eğer hatalı sınıflama maliyeti açıkça belirtilmediyse) veya}$$

$$i(t) = \sum_{j \neq i} C(i \mid j) p(j \mid t) p(i \mid t) \text{ (eğer hatalı sınıflama maliyeti açıkça belirtildiyse)}$$

Burada C(i|j), gerçekte j sınıfına ait bir deney ünitesini i sınıfı gibi (hatalı) sınıflamanın maliyetidir (17).

Örnek olarak, Şekil 4.2' deki sınıflama ağacı üzerinde Gini ayırma kuralını açıklayalım. A, B, C ve D sınıflarının populasyon içerisindeki oranları sırasıyla %40, %30, %20 ve %10 olsun. Gini ayırma kuralının amacı ayırma fonksiyonunu maksimum yapacak şekilde bir sınıfı diğerlerinden ayırmaktır. Gini safsızlık ölçüsünün tanımına göre sadece bir sınıfa ait deney ünitelerini içeren düğümün safsızlık ölçüsü (heterojenlik) minimum (sıfır) dır. Bu durumda, hangi sınıf diğerlerinden ayrılırsa ayrılsın sol çocuk düğümünün safsızlık ölçüsü sıfır olur. Ancak, ayırma fonksiyonunun maksimum yapılabilmesi için sağ çocuk düğümünün safsızlık ölçüsü de düşük olmalıdır. Bu amaçla, populasyon içerisinde en yüksek orana sahip olan sınıf A diğer sınıflardan ayrılacaktır. Bu durumda, sınıf A terminal düğüm olarak adlandırılır. Sınıf A diğerlerinden ayrıldıktan sonra sınıf B sınıf C ve D'den ve daha sonra sınıf C sınıf D'den ayrılır.

Fakat aşağıdaki ağaç modelinde olduğu gibi deney ünitesindeki tek bir sınıfı tamamen homojen bir şekilde diğer sınıflardan ayırabilecek bir ayıraç bulmak normal şartlarda zor yada imkansızdır (18). Ancak Gini bu ideale mümkün olduğunca yaklaşır.

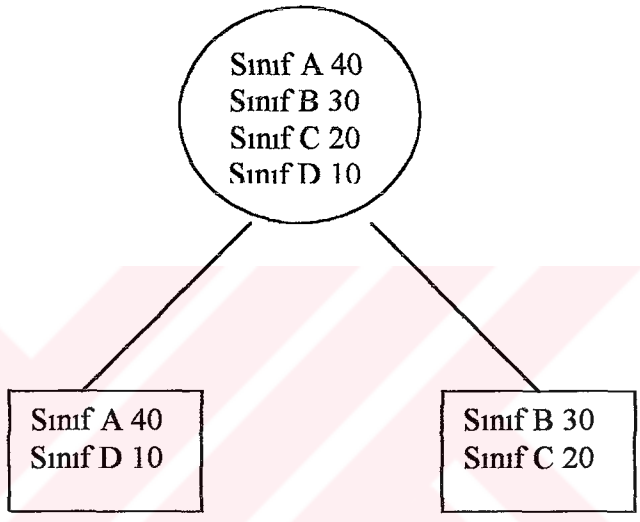


Şekil 4.2: Gini Ayırma Kuralı

4.1.2. Twoing Kuralı

Twoing kuralı Gini'den oldukça farklıdır. Twoing ilk olarak, tek bir sınıfı diğerlerinden ayırmak yerine, düğüme ait verinin %50'sini kapsayan ve birbirine benzemeyen sınıfları ayırmaya çalışır. Örnek olarak, Şekil 4.3'de kullanılan sınıflama ağacı üzerinde Twoing ayırma kuralını açıklayalım. A, B, C ve D sınıflarının populasyon içerisindeki oranları sırasıyla %40, %30, %20 ve %10 olsun. Twoing ayırma kuralının ilk amacı her bir düğümdeki verileri, verilerin %50'si dağılacak şekilde iki çocuk düğümüne ayırmaktır. Bu ayırma gerçekleştirilirken çocuk düğümlerinde mümkün olduğunca farklı sınıfların bulunmasına dikkat eder. Eğer A ile B sınıfı, C ile D sınıfından ayrılacak olursa verilerin %70'i sağ düğüme, %30'u ise sol

düğüme atanır. Twoing ayırma kriteri her bir çocuk düğümüne verilerin %50'sini atamayı amaçladığı için böyle bir atama uygun değildir. Ancak, A ile D sınıfı, B ile C sınıfından ayrılacak olursa Twoing ayırma kriterine uygun olarak verilerin %50'si sağ, %50'si ise sol çocuk düğümüne atanır. Aynı şekilde, sağ çocuk düğümünde sadece A ile D sınıfına ait veriler, sol çocuk düğümünde ise B ile C sınıfına ait veriler yer alır. Ağaç oluşumunun son safhalarına doğru ise, Twoing ayırma kriteri her bir düğüme bir sınıf gelecek şekilde verileri ayırır (18).



Şekil 4.3: Twoing Ayırma Kuralı

Herhangi bir t düğümünde olası tüm ayırmalar bulunduktan ve en iyi ayırma, ayırma fonksiyonu ve Gini yada Twoing ayırma kuralları yardımıyla, seçildikten sonra seçilen bu ayırma düğüme uygulanır. Bu işlem sonrasında ortaya çıkan çocuk düğümlere mevcut sınıflardan en uygun olanı tahmini olarak atanır. Bir düğüm için en uygun sınıfın tahmin edilmesi esnasında düğümde *mevcut deney ünitelerinin sınıflara dağılımı, ön olasılıklar* (prior probabilities) ve *kayıp yada zarar (risk) matrisi* (decision loss or cost matrix) gibi faktörler göz önünde bulundurulur.

4.1.3. Gini ve Twoing Ayırma Kurallarının Karşılaştırılması

Düğümlerdeki mümkün olan en iyi ayırmayı gerçekleştirmek için ayırma kuralı seçilirken aşağıdaki faktörler göz önünde bulundurulmalıdır.

- Kategorik bağımlı değişkenin seviye sayısı iki ise ve analizin 0.50'den daha az hata oranına sahip olacağı tahmin ediliyorsa ayırma kuralı olarak Gini kuralı tercih edilmelidir.

- Kategorik bağımlı değişkenin seviye sayısı iki ise ve analizin 0.80'den daha az hata oranına sahip olacağı tahmin ediliyorsa ayırma kuralı olarak Twoing kuralı tercih edilmelidir.
- Bağımlı değişkenin seviye sayısı 4' ten daha büyük olduğu koşullarda Twoing kuralı Gini'den daha doğru bir seçimdir (18).

4.2. Mevcut Deney Ünitelerinin Sınıflara Dağılımı

Ayırma sonunda ortaya çıkan bir çocuk düğümüne en uygun sınıfın atanması esnasında Learning Sample'da yer alan deney ünitelerinin sınıflara dağılımı önemli bir rol oynar. Kategorik bağımlı bir değişken için sınıf numaraları j = 1,2...,k olsun. Deney ünitelerinin sınıflara dağılımı, her bir j sınıfına ait olan deney ünitelerinin sayısı, N_j, olarak ifade edilir.

4.3. Ön Olasılıklar

CART metodunda bir sınıflama ağacı oluşturulurken ön olasılıklar (prior probabilities) kullanır. Ön olasılıklar deney ünitelerinin ait olacağı sınıfın belirlenmesini etkiler (19). j sınıfı için ön olasılık değeri (π_j) ile gösterilir ve bu değerler ya veri setinden hesaplanır yada araştırmacı tarafından bildirilir. Ön olasılık değerleri 3 alternatif şekilde hesaplanır.

a) Örneklemden hesaplanan ön olasılık : Populasyondaki bağımlı değişkenlerin dağılımının Learning Sample'daki sınıfların dağılımı ile aynı olduğunu ($\pi_j = \frac{N_j}{N}$) varsayar.

b) Eşit Ön olasılık: Bağımlı değişkenin her bir sınıfının eşit gerçekleşme olasılığının var olduğunu varsayar. Örneğin, bağımlı değişken 2 sınıfa sahip ise;

$\pi_1 = \pi_2 = \frac{1}{2}$ olduğu varsayılır.

c) Priors mixed: Herhangi bir sınıf için örneklemden hesaplanan ön olasılık ve eşit ön olasılık değerlerinin ortalamasıdır.

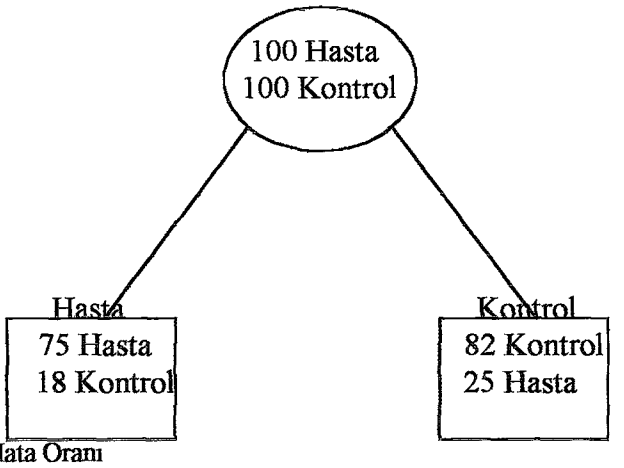
4.4. Kayıp yada Zarar (Risk) Matrisi

Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır (Doğruluk Oranı = 1 - Hata Oranı). Verilerin sınıflandırılması için oluşturulan modellerin hata oranlarına karar vermek için risk matrisi kullanılmaktadır. Bu matris, Diskriminant Analizi, Lojistik Regresyon v.b. sınıflama modellerinde kullanıldığı gibi, sınıflama ağaçlarında da aynen kullanılır.

Aşağıda Çizelge 4.1'de verilen matris örnek bir risk matrisidir. Satırlarda gerçekte olması gereken sınıf değerlerini sütunlarda ise model sonucunda elde edilen tahmini sınıflama değerleri yer almaktadır. Örnek olarak, gerçekte hasta grubunda 100 ve kontrol grubunda 100 birey olması gerekirken kurulan tahmin modeli (yada sınıflama ağacı) sonucunda hasta grubunda 93 birey, kontrol grubunda ise 107 birey yer almıştır. Kurulan modelin hata oranı %21.5 ((18+25)/200), doğruluk oranı ise %78.5'dur ((75+82)/200).

Çizelge 4.1: Örnek Risk Matrisi

	TAHMİN		
GERÇEK	***HASTA***	***KONTROL***	***TOPLAM***
HASTA	75	25	100
KONTROL	18	82	100
TOPLAM	93	107	200



Şekil 4.4: Sınıflama Ağacı Hata Oranı

Ayırma sonucunda ortaya çıkan herhangi bir düğüme atanacak olan en uygun sınıf aşağıdaki gibi tahmin edilir;

C(j/i) : i sınıfını j sınıfı gibi sınıflamanın maliyeti (risk matrisi katsayıları),

π_i : i sınıfının önceki olasılığı,

N_i : Learning Sample'da i sınıfında bulunan deney ünitelerinin sayısı,

$N_i^{(t)}$: t düğümünde i sınıfında bulunan deney ünitelerinin sayısı olmak üzere;

$\frac{C(j/i)\pi_i N_i^{(t)}}{C(i/j)\pi_j N_j^{(t)}} > \frac{N_i}{N_j}$ eşitsizliği j'nin bütün değerleri (j = 1,2,...,k ve $j \neq i$) için sağlanıyorsa t düğümüne en uygun olarak i sınıfı atanır (5).

Düğümün yapısına göre bazı durumlarda birden fazla sınıf yukarıda belirtilen eşitsizliği sağlayarak en uygun sınıf konumuna girer yada hiçbir sınıf bu eşitsizliği sağlayamaz. Böyle bir durumda en uygun sınıfın belirlenmesi için iki alternatif kural mevcuttur.

4.5. Çoğulluk Kuralı

Çoğulluk kuralı hatalı sınıflama maliyetini göz önüne almaksızın (eşit varsayarak) düğüm içerisinde en büyük orana sahip olan sınıfı en uygun sınıf olarak atar (19).

4.6. Minimum Risk Kuralı

Minimum risk kuralı düğüm içerisinde deney ünitelerinin sınıflara dağılımını göz önüne almaksızın (eşit varsayarak) düğüm içerisinde hatalı sınıflama maliyetini minimum yapan sınıfın en uygun sınıf olarak seçilmesidir.

Örnek olarak, bir problemde iki sınıf (Sınıf 1 ve Sınıf 2) var olsun ve;

$r_1^{(t)}$: Sınıf 1'in t düğümüne atanma maliyeti,

r_2^t : Sınıf 2'nin t düğümüne atanma maliyeti,

π_1 : Sınıf 1'in ön olasılığı,

π_2 : Sınıf 2'nin ön olasılığı,

$r_1^{(t)}$ $= \pi_1 . C(2/1)$

$r_2^{(t)}$ $= \pi_2 . C(1/2)$ olarak tanımlanmış olsun.

Eğer;

$r_1^{(t)} < r_2^{(t)}$ ise düğüm t, Sınıf 1'e, aksi halde Sınıf 2'ye atanır.

C(2/1) = C(1/2) ise çoğulluk kuralına başvurulur. Ön olasılıkları en yüksek olan Sınıf 1 bu düğüme atanır.

Bu tanımlamalardan sonra, bir sınıflama ağacının oluşturulması için gerekli adımlar aşağıdaki şekilde özetlenebilir.

Ağaç üzerinde herhangi bir t düğümü için;

1. Düğümde yer alan deney ünitelerinin içerdiği bağımsız değişkenler ve bu değişkenlerin bir birleri ile kombinasyonlarının tanımlı bulunduğu aralıklardaki tüm olası değerleri birer ayıraç olarak varsayıp, mümkün olan tüm olası ayırmaları belirlenmesi.
2. Mümkün olan her bir ayırma için o ayırmanın uygunluk derecesini, ayırma fonksiyonu yardımıyla hesaplayarak maksimum uygunluk derecesine sahip ayırmanın belirlenmesi.
3. En iyi ayırmayı yapan ayıracın, t düğümüne uygulanması ve ortaya çıkacak sol ve sağ çocuk düğümlerinin her birine en uygun sınıfın tahmin edilmesi.

Yukarıda sıralanan adımlar kök düğümden başlayarak, daha sonra ortaya çıkacak her düğüm için tekrarlanır. Sınıflama ağacı, her bir düğüm noktası bu şekilde ikiye ayrılarak büyür. Bu büyüme;

1. Her çocuk düğümündeki gözlem sayısı
   - Sadece bir gözlem ise veya on gözlem ise (5, 17).
2. Her düğümde grup içi homojenlik söz konusu ise,
3. Ağacın düzey sayısında analizi yürüten kişi tarafından bir sınırlama yapıldıysa,
4. Yeni oluşacak düğümlerde fazla bir değişiklik yaratmıyorsa durur.

Sınıflama ağacının büyümesini durduran bu şartlardan (*stopping criteria*) herhangi birinin gerçekleşmesi sonucunda ağacın oluşması (büyümesi) safhası sona erer. Ağaç inşası sonunda elde edilen ağaç *büyük (maximal) ağaç* olarak adlandırılır ve Learning Sample'daki deney ünitelerine en uygun ağaçtır. Ancak maximal ağaç pratikte iki probleme neden olur;

1. Maximal ağaç Learning Sample'ı kusursuz biçimde tanımlar çünkü eklenen her bağımsız değişken hatalı sınıflama oranını düşürür. Bu durumda, maximal ağaç Learning Sample için olması gerekenden daha iyi bir tahmin modeli (overfitting) sunar. Ancak, Learning Sample'a aşırı uyumlu maximal ağaçlar farklı bir veri seti (örneğin Test Sample) söz konusu olduğunda iyi bir tahmin sağlayamazlar.
2. Bir sınıflama ağacının karmaşıklık ölçüsü o ağacın terminal düğüm sayısına eşittir. Terminal düğüm sayıları ve dolayısıyla karmaşıklığı yüksek olan maximal ağacın anlaşılması ve yorumlanması güçtür.

Maximal ağacın pratikte ortaya çıkardığı bu sorunların çözümü için maximal ağacın budanması yani maximal ağaçtan oluşturulan daha küçük bir ağacın seçilmesi gereklidir (20).

Maximal ağacın budanması daha küçük ağaçlar dizisi oluşturur ve oluşturulan bu dizi içerisinden optimal ağaç seçilir. Optimal ağaç maximal ağaçtan daha az karmaşıklığa sahiptir ancak, optimal ağaç Learning Sample'a maximal ağaçtan daha az uyumludur ve hatalı sınıflama oranı daha yüksektir. Optimal ağacın seçimi için kullanılan *maliyet-karmaşıklık budama metodu (cost-complexity pruning method)* hatalı sınıflama oranı ile ağacın karmaşıklığı arasındaki dengeyi sağlar ve matematiksel olarak;

$R_\alpha^{(T)} = R(T) + \alpha.T$ şeklinde ifade edilir.

Burada;

$R_\alpha^{(T)}$: Maliyet-karmaşıklık ölçüsünü,

$R(T)$: T ağacı için hesaplanan hata oranını,

T : T ağacındaki terminal düğüm sayısını,

α : Ağaçtaki her terminal düğüm için belirlenen ceza katsayısını ($\alpha \geq 0$) gösterir.

Maliyet-karmaşıklık budama metoduna göre maximal ağaç, maliyet-karmaşıklık ölçüsü minimum değerine ulaşıncaya kadar budanır ve optimum ağaç elde edilir. Maliyet-karmaşıklık ölçüsünde α değerinin artması optimal ağaçta daha az terminal düğümünün yer almasına yol açar. Bir başka ifade ile α değeri arttıkça budama artar.

Maliyet-karmaşıklık ölçüsünde yer alan $R(T)$ değerinin (hata oranının) alternatif hesaplanma yöntemleri aşağıda verilen başlık altında daha detaylı olarak incelenecektir.

4.7. Sınıflama Ağaçlarında Doğruluk Tahmini

Daha önce belirtildiği gibi, bir sınıflama ağacında yanlış olarak sınıflanan deney ünitesi sayısının toplam deney ünitesi sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan deney ünitesi sayısının toplam deney ünitesi sayısına bölünmesi ile ise doğruluk oranı hesaplanır. Sınıflama ağaçlarında bağımlı değişken kategorik olduğunda üç alternatif doğruluk tahmin yöntemi vardır.

4.7.1. Revizyon veya Yeniden Yerine Koyma Tahmini (Resubstitution Estimate)

Bu yöntemde Learning Sample'ın tümü alınarak ağaca uygulanır ve tekrar sınıflandırılır. Bu sınıflandırma sonucunda hatalı sınıflandırılmış deney ünitelerinin oranı (hata oranı, $R(T)$) hesaplanır.

$$R(T) = \frac{1}{N} \sum_{i=1}^{N} X(d(x_n) \neq J_n)$$

Burada:

$X(.)$ = İndikatör (gösterge) fonksiyondur. Bağımlı değişkenin tahmin edilen sınıf üyeliği gerçekte ait olduğu sınıf üyeliğine eşit ise "1" değerini değilse "0" değerini alır.

$d(x)$ = Deney ünitesinin tahmin edilen bağımlı değişken sınıfıdır.

J_n = Deney ünitesinin gerçekte ait olduğu bağımlı değişken sınıfıdır.

N = Toplam deney ünite sayısıdır.

4.7.2. Test Sample Tahmini (Test Sample Estimation)

Bir sınıflama ağacının doğruluğunun test edilmesinde kullanılan diğer bir yöntem de test sample tahminidir. Bu yöntemde tipik olarak, deney ünitelerinin yaklaşık olarak % 33'ü Test Sample, %67'i Learning Sample olarak ayrılır. Sınıflama ağacı Learning Sample kullanılarak inşa edilir ancak, ağacın hata oranı ($R(T_{ts})$ değeri) Test Sample kullanılarak hesaplanır. Bu yöntem deney ünitelerinin (yani mevcut veri setinin)

bir kısmının Test Sample olarak ayrılmasını ve dolayısı ile büyük bir veri setini gerektirir.

$$R(T_{ts}) = \frac{1}{N_2} \sum_{i=1}^{N} X(d(x_n) \neq J_n)$$

N_2 = Test Sample'daki toplam deney ünite sayısıdır.

4.7.3. Çapraz Geçerlilik Testi (Cross Validation Test)

Sınırlı miktarda veri olduğu durumda, kullanılabilecek diğer bir yöntem ise çapraz geçerlilik testidir. Bu yöntemde, mevcut deney üniteleri tesadüfi olarak a ve b olmak üzere iki eşit parçaya ayrılır. İlk aşamada a parçası Learning Sample ve b parçası Test Sample olarak; ikinci aşamada ise b parçası Learning Sample ve a parçası Test Sample olarak düşünülür ve bu şekilde elde edilen iki hata oranın ortalaması ağacın hata oranı ($R(T_{ts}^{c2})$ değeri) olarak kullanılır.

Eğer Learning Sample $L_1, L_2, ..., L_V$ olmak üzere v eşit parçaya ayrılırsa ve çapraz geçerlilikteki adımlar gerçekleştirilirse v katlı çapraz geçerlilik uygulanmış olur ve bu şekilde elde edilen v tane hata oranın ortalaması ağacın hata oranı ($R(T_{ts}^{cv})$ değeri) olarak hesaplanır.

$$R(T_{ts}^{cv}) = \frac{1}{N_v} \sum_{i=1}^{N} X(d^v(x_n) \neq J_n)$$

$$N_v = \frac{N}{v}$$

5. REGRESYON AĞAÇLARI

Bağımlı değişken sayısal ölçümler (kesikli veya sürekli değişken) aldığı zaman CART regresyon ağacı üretir. Regresyon ağaçlarının iki amacı vardır.

1. Bağımsız değişkenlere ait ölçüm vektöründen doğru ve güvenilir bir şekilde bağımlı değişkenin değerini tahmin etmek,
2. Bağımlı ve bağımsız değişken arasındaki yapısal ilişkiyi ortaya çıkarmaktır.

Regresyon ağaçlarının oluşumu sınıflama ağaçlarına benzer. Fakat regresyon ağaçlarında sınıf atama kurallarına, ön olasılıklara, hatalı sınıflama maliyetlerine ihtiyaç yoktur. Ayırma kuralları, en uygun kriterin seçimi ve ağacın doğruluğunun tahmini sınıflama ağaçlarından farklıdır. Regresyon ağaçlarında ek olarak terminal düğümler özet istatistik (ortalama ve standart sapma) değerlerine sahiptir.

5.1.Regresyon Ağaçları Oluşumu

Ağaçların oluşumu dört adımda gerçekleşir.

5.1.1. Başlangıç veri setindeki soruların oluşumu

Regresyon ağaçlarında Learning Sample oluşumu sınıflama ağaçlarındaki gibidir. Tek farklılık deney ünitelerinin ait olduğu (J) sınıflar sayısal verilerden oluşur.

5.1.2. Ayırma kuralları

Regresyon ağaçlarında da amaç en uygun bölme kriteri tespit edilerek düğümdeki heterojenlik maksimum şekilde giderilmeye çalışılarak iki çocuk düğümündeki homojenlik maksimum hale getirilir. Bu şekilde çocuk düğümleri arasındaki farklılık maksimum seviyeye ulaşır.

Regresyon ağaçlarında Least Squares (LS) , Least Absolute Deviation (LAD) ve Clark&Pregibon (CP) olmak üzere üç ayırma kuralı vardır. Bu üç kuralda da amaç düğümlerdeki (çocuk ve terminal) heterojenlik minimize etmektir. Heterojenlik ölçüsü i(t) ile tanımlanır.

LS ve LAD kurallarının farkı;

LS kuralına göre heterojenlik ölçüsü i(t) düğümdeki ortalama etrafında bağımlı değişkenin karelerinin toplamıdır.

LAD kuralına göre heterojenlik ölçüsü i(t) düğümdeki medyan etrafında bağımlı değişkenin karelerinin toplamıdır.

5.1.2.1. Least Squares (LS) Kuralı:

$$i(t) = \sum_{i=1}^{N} (Y(i) - \bar{Y}(t))^2$$

Burada;

$i(t) =$ t. düğümdeki heterojenlik.

$Y(i) =$ t. düğümdeki bağımlı değişkenin değeri

$\bar{Y}(t) =$ t. düğümdeki bağımlı değişkenin ortalama değerini göstermektedir.

5.1.2.2. Clark&Pregibon (CP) Kuralı:

Bu kurala göre sapma düğümdeki bütün gözlemlerin sapmalarının toplamıdır. Amaç hata kareler toplamını (RRS) minimize etmektir.

$$RRS = \sum_{i \in L} (y_i - \bar{y}_L)^2 + \sum_{i \in R} (y_i - \bar{y}_R)^2$$

Burada;

$y_i =$ Sol düğümdeki bağımlı değişkenin değerini

$\bar{y}_L =$ Sol düğümdeki bağımlı değişkenin ortalama değerini

$\bar{y}_R =$ Sağ düğümdeki bağımlı değişkenin ortalama değerini göstermektedir(4).

5.1.3. En İyi Ayırma Kriterlerinin Tespiti:

Aşağıdaki fonksiyonla en iyi ayırma ölçülebilir.

$$\phi(t) = i(t) - i(t_R) - i(t_L)$$

$i(t_R) =$ Sağ çocuk düğümdeki katışıklık (sağ çocuk düğümündeki ortama etrafındaki kareler toplamı)

$i(t_L)$ = Sol çocuk düğümdeki katışıklık(sol çocuk düğümündeki ortama etrafındaki kareler toplamı)

En iyi ayırmada amaç $\phi(t)$'yi maksimize etmektir. Yani $i(t_R)+i(t_L)$'yi minimize etmektir.

5.1.4. Regresyon Ağaçlarında Doğruluk Tahmini

Regresyon ağaçlarında doğruluk tahminlerinin işleyişi sınıflama ağaçlarında olduğu gibidir. Sadece formüller aşağıdaki gibi değişir.

5.1.4.1. Resubstitution Estimate (Revizyon Tahmini veya Yeniden Yerine Koyma Tahmini)

$$R(d)=\frac{1}{N}\sum_{i=1}^{N}(Y_i-d(x_i))$$

Burada:

$R(d)$ = Hata oranıdır.

Y_i = Sürekli bağımlı değişkenin gerçek değeridir.

$d(x_i)$ = Bağımlı değişkenin tahmin edilen değeridir.

5.1.4.2.Test Sample Estimate

$$R_{(d)}^{ts}=\frac{1}{N_2}\sum_{(x_i,y_i)\in L_2}^{N}(Y_i-d(x_i))^2$$

5.1.4.3. V-Fold Cross Validation

$$R_{(d)}^{cv}=\frac{1}{N_v}\sum_{v}\sum_{x_n,y_n}(Y_i-d_{(x_n)}^{v})^2$$

6.UYGULAMA

Bu çalışmanın uygulama bölümünde, Mersin Üniversitesi Tıp Fakültesi Hastanesi Nöroloji Bölümünün 206 denek üzerinde yaptığı anket çalışmasının sonuçları veri olarak kullanılmıştır. Anket çalışmasında kullanılan anket formları Ek 1'de sunulmuştur. Bu anket çalışmasının uygulandığı 206 deneğin 103'ü Huzursuz Bacak Sendromu (RLS) hastasıdır. Geriye kalan 103 denekte ise RLS hastalığı yoktur. Bu şekilde denekler, Hasta ve kontrol grubu olarak ikiye ayrılmış ve deneğin ait olduğu grup sınıflama ağacı uygulamasında iki seviyeli (Hasta, Kontrol) kategorik bağımlı değişken olarak kullanılmıştır. Ankette her iki gruba da sorulan ortak sorular kullanılarak Hasta ve Kontrol grubunun ayrımını önemli ölçüde etkileyen bağımsız değişkenler Statistica® 6.0 paket programı yardımıyla tespit edilmiştir (21).

Sadece kadınları ilgilendiren gebelik ve menopoz (22) RLS Hastalığı için önemli birer risk faktörü taşıdığından, genel ve kadınlara özel olmak üzere iki ayrı analiz gerçekleştirilmiştir.

6.1.Birinci Analiz

Birinci analizde, cinsiyet ayrımı yapılmaksızın Hasta ve Kontrol grubunun ayrımını önemli ölçüde etkileyen bağımsız değişkenler tespit edilmiştir. Fakat bu analiz uygulanırken anket formunda yer alan kadınlara ait sorular (gebelik ve menopozla ilgili sorular) erkekleri ilgilendirmediğinden bu sorular analiz dışı bırakılmış ve hem kadın hem erkek deneklere uygun olan ortak sorular analizde kullanılmıştır. Analizde kullanılan; sürekli bağımsız değişkenler Çizelge 6.1.1'de, kategorik bağımsız değişkenler ise Çizelge 6.1.2'de tanımlayıcı istatistikleri ile verilmiştir.

Çizelge 6.1.1: Analiz I'de kullanılan sürekli bağımsız değişkenlere ait tanımlayıcı istatistikler.

	RLS			**KONTROL**		
Sürekli Değişken	$\overline{X} \mp SD$	Min	Max	$\overline{X} \mp SD$	Min	Max
Yaş	43.25±15.31	18	79	43.10±15.21	19	75
Kilo	68.87±12.53	45	105	68.41±14.37	45	125
Boy(cm.)	1.63±0.08	1.50	1.87	1.62±0.09	1.47	1.96
Öğrenim Süresi(yıl)	4.57±3.98	0	18	4.73±3.57	0.0	15

Çizelge 6.1.2. Analiz I'de kullanılan kategorik bağımsız değişkenlere ait tanımlayıcı istatistikler

Kategorik Bağımsız Değişkenler	Kategorik Bağımsız Değişkenlerin Seviyeleri	RLS		KONTROL	
		n	%	n	%
Cinsiyetiniz:	Kadın	64	62.13	64	62.13
	Erkek	39	37.86	39	37.86
Mesleğiniz:	Ev hanımı	60	58.25	61	59.22
	Öğrenci	1	0.97	2	1.94
	Çiftçi	9	8.73	9	8.73
	Devlet memuru	5	4.85	5	4.85
	Emekli	7	6.79	7	6.79
	Esnaf	8	7.71	10	9.71
	Diğer	11	10.67	8	7.76
Yaşadığınız yer:	İl	54	52.42	54	52.42
	İlçe	28	27.18	28	27.18
	Köy	21	20.38	21	20.39
Yaşadığınız yerin deniz kıyısına olan uzaklığı:	0-100m.	77	74.75	78	75.27
	101-500m.	15	14.56	14	13.59
	501-1000m.	9	8.73	9	8.73
	1001-2000m.	2	1.94	2	1.94
Medeni haliniz:	Evli	91	88.34	89	86.41
	Bekar	12	11.65	14	13.59
Sigara içiyor musunuz :	İçmiyorum	57	55.33	72	69.91
	İçiyorum	46	44.66	31	30.09
Günde ne kadar sigara içiyorsunuz:	Hiç içmedim	49	47.57	65	63.10
	Şimdi içmiyor	8	7.76	7	6.79
	Günde 10 adetten az	12	11.65	13	12.62
	Günde 10-19 adet	12	11.65	6	5.83
	Günde 1-2 paket	20	19.41	11	10.67
	Günde2 paketten fazla	2	1.94	1	0.97
Geçmişte sigara alışkanlığınız var mı:	Hiç içmedim	54	52.43	65	63.11
	Var	49	47.47	38	36.89
Alkol kullanıyor musunuz:	Kullanmıyorum	7	6.79	11	10.68
	Ayda 10 duble rakıdan az	90	87.37	90	87.37
	Ayda 10 duble rakıdan fazla	6	5.82	2	1.95
Antidepressan ilaç kullanıyor musunuz :	Hayır	98	95.14	94	91.26
	Evet	5	4.85	9	8.74
Antiparkinson ilaç kullanıyor musunuz :	Hayır	100	97.09	103	100
	Evet	3	2.91	0	0
Beyin ya da omurilik ya da bu bölgelerle ilgili başka hastalık geçirdiniz mi:	Hayır	91	88.34	96	93.20
	Evet	12	11.65	7	6.80
Kansızlık hastalığınız var mı:	Hayır	95	92.23	99	96.11
	Evet	8	7.77	4	4.89

Çizelge 6.1.2 (Devam). Analiz I'de kullanılan kategorik bağımsız değişkenlere ait tanımlayıcı istatistikler

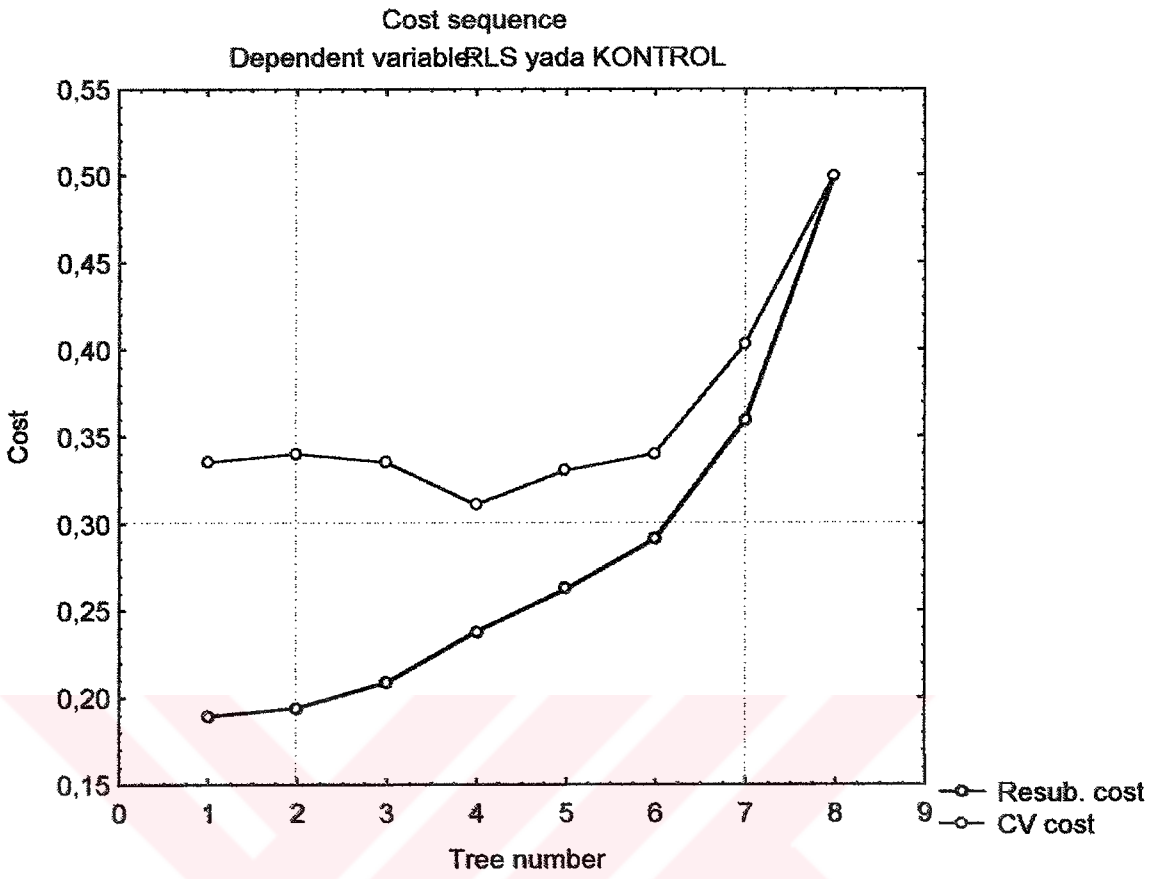
Böbrek yetmezliğiniz var mı:	Hayır	94	91.26	102	99.03
	Evet	9	8.74	1	0.07
Hipertansiyon hastalığınız varmı:	Hayır	92	89.32	89	86.40
	Evet	11	10.68	14	13.60
Diyabet hastalığınız var mı:	Hayır	100	97.08	100	97.08
	Evet	3	2.92	3	2.92
Migren hastalığınız var mı:	Hayır	102	99.03	103	100
	Evet	1	0.07	0	0
Depresyon hastalığınız var mı:	Hayır	98	95.14	103	100
	Evet	5	4.86	0	0
Ayda ortalama kaç gün gündüz saatlerinde uyuklarsınız:	0 gün	49	47.57	60	58.25
	1-5 gün	24	23.30	31	30.09
	6-15 gün	13	12.62	5	4.86
	15 günden fazla	17	16.51	7	6.80
Ayda ortalama kaç gece uyurken uykudan uyanırsınız:	Hiç	12	11.65	26	25.24
	1-5 gece	30	29.12	47	45.63
	6-15 gece	24	23.30	16	15.53
	15 geceden fazla	37	35.93	14	13.60
Günde ortalama kaç saat uyursunuz:	2 saatten az	0	0	1	0.97
	2-4 saat	4	3.88	2	1.94
	5-7 saat	53	51.45	44	42.72
	8-10 saat	42	40.77	50	48.55
	10 saatten fazla	4	3.88	6	5.82
Ayda ortalama kaç gece rüya görürsünüz:	Hiç	10	9.71	8	7.77
	1-5 gece	42	40.78	50	48.54
	6-15 gece	26	25.24	25	24.27
	15 geceden fazla	25	24.27	20	19.42
Sağlığınız genel olarak nasıldır:	Mükemmel	0	0	8	7.76
	Çok iyi	21	20.38	32	31.06
	İyi	29	28.15	35	33.98
	Orta	45	43.70	25	24.27
	Kötü	8	7.76	3	2.91
Son 1 ay içinde kaç gün moraliniz bozuktu:	0-10 gün	30	29.12	38	36.89
	11-20 gün	26	25.24	40	38.83
	21-30 gün	47	45.63	25	24.27
1.Derece akrabalarınızda bu tür şikayetler var mı:	Evet	63	61.16	11	10.67
	Hayır yada bilmiyorum.	40	38.84	92	89.33

Sınıflama Ağacı analizinde ayırma kriteri olarak Gini ayırma kriteri, budama yöntemi olarak 10 katlı çapraz geçerlilik yöntemi tercih edilmiştir. Hasta ve Kontrol gruplarının sayıları eşit olduğu için önsel olasılıkları eşit (0.5) olarak alınmıştır. Statistica® 6.0 başlangıçta 8 sınıflama ağacı üretmiştir. Bu ağaçlara ait maliyet-karmaşıklık bilgileri çizelge 6.1.3'de sunulmuştur.

Çizelge 6.1.3. Analiz I için oluşturulan 8 sınıflama ağacına ait maliyet-karmaşıklık bilgileri

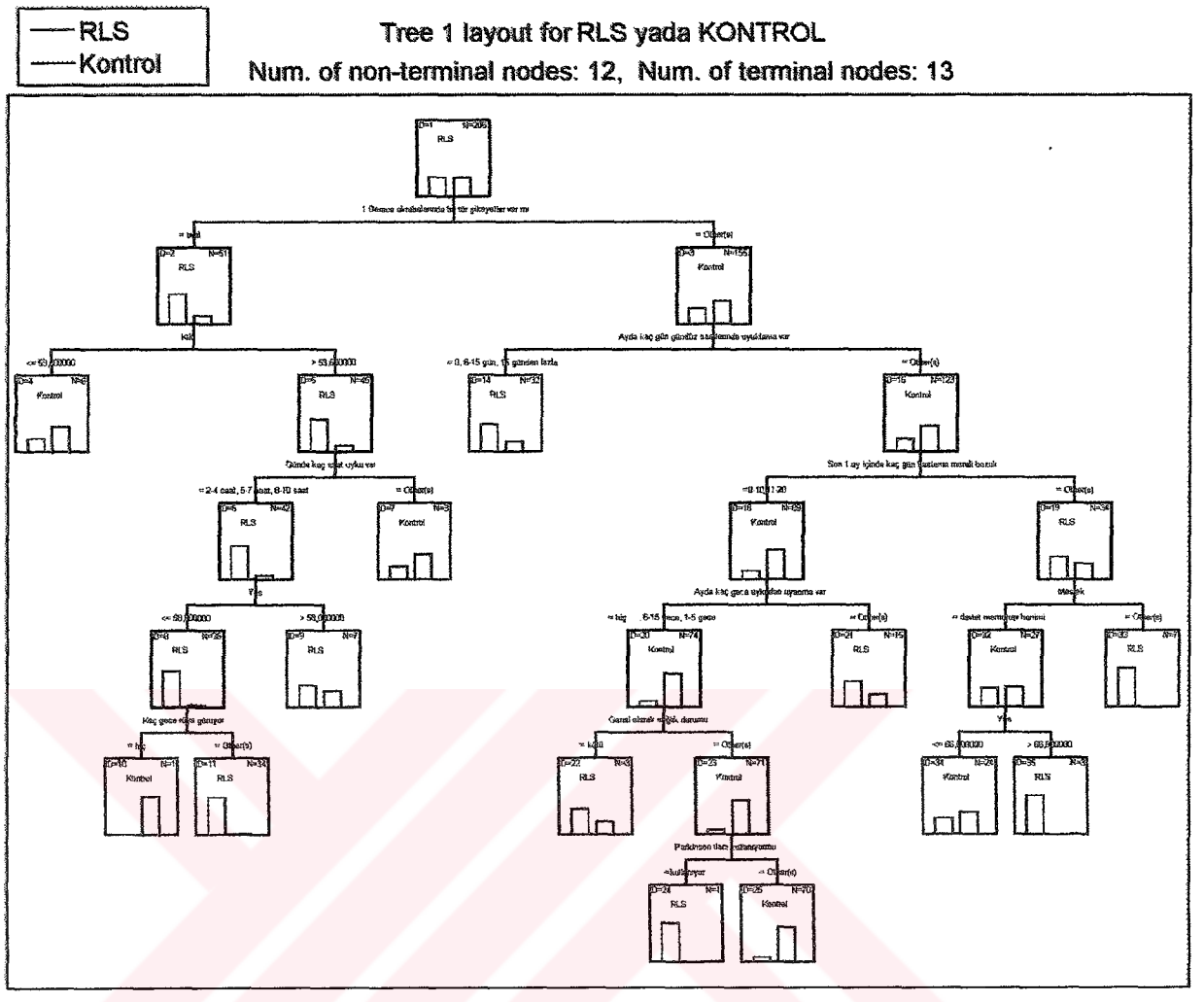
	Terminal nodes	**CV Cost (%)**	**CV std. Error (%)**	**Resubstitution Cost (%)**	**Node complexity**
Tree 1	13	0,334951	0,032884	**0,189320**	0,000000
Tree 2	11	0,334951	0,032884	0,194175	0,002427
Tree 3	8	0,320388	0,032379	0,208738	0,004854
Tree 4	5	0,315534	0,032511	0,237864	0,009709
Tree 5	4	0,330097	0,032764	0,262136	0,024272
***Tree 6**	3	0,339806	0,033000	**0,291262**	0,029126
Tree 7	2	0,383495	0,033878	0,359223	0,067961
Tree 8	1	0,500000	0,034837	0,500000	0,140777

Ağaç 1 (Tree 1) maximal ağaçtır ve 13 adet terminal düğüme sahiptir. Amaç maliyet-karmaşıklık ölçüsünü minimize etmek olduğundan, hatalı sınıflama maliyetleri (CV cost ve Resubstitution cost), ceza katsayısı (Node complexity, α) ve terminal düğüm sayısını (T) dengeleyen Çizelge 6.1.3'de * ile işaretli olan 6 nolu ağaç (*Tree 6) optimal ağaç olarak seçilmiştir. Budama artıkça terminal düğüm sayısı azalmıştır fakat hatalı sınıflama maliyetleri artmıştır. Analize giren bağımsız değişkenlerin daha fazla sayıda olması nedeniyle en iyi sınıflamanın yapıldığı maximal ağaçta hatalı sınıflama maliyetleri en düşüktür.



Şekil 6.1.1: Analiz I için oluşturulan ağaçların hatalı sınıflama maliyetleri

Şekil 6.1.1' de, oluşturulan 8 sınıflama ağacına ait hatalı sınıflama maliyetleri verilmektedir. Başlangıçta hatalı sınıflama maliyeti düşüktür. Budama arttıkça terminal düğüm sayısı ve dolayısı ile modele giren bağımsız değişken sayısı azaldığı için hatalı sınıflama maliyetleri yükselmiştir. En son oluşturulan ağacın (Tree 8) hatalı sınıflama maliyetleri maksimumdur. Bu ağaçta 206 deneğin tamamı RLS hastası olarak sınıflandırılmış ve böylece hatalı sınıflama oranı %50 olmuştur.



Şekil 6.1.2. Analiz I için oluşturulan maximal sınıflama ağaç diyagramı.

Şekil 6.1.2'de sunulan sınıflama ağacı, değişkenler arasındaki ilişkileri en ayrıntılı biçimde gösteren maximal ağaçtır. Şekil 6.1.2'de düğümler kırmızı (terminal düğümleri) ve mavi (çocuk düğümleri) renkte kareler olarak şekillendirilmiş ve her bir düğüm içerisinde o düğümün hangi sınıfa ait olduğu belirtilmiştir. Maximal ağaçta 13 terminal düğüm, 12 çocuk düğüm vardır. Düğüm 4, Düğüm 7, Düğüm 9, Düğüm10, Düğüm 11, Düğüm 14, Düğüm 21, Düğüm 22, Düğüm 24, Düğüm 25, Düğüm 33, Düğüm 34, Düğüm 35 terminal düğümleri, diğerleri ise çocuk düğümleridir. Düğümlerin içerisindeki sağ kutucuklarda N harfi ile o düğümde kaç deneğin bulunduğu, sol kutucuklarda ise D harfleri ile düğüm numarası gösterilmektedir. Düğümler içerisinde ayrıca, o düğümde yer alan deneklerin ait oldukları sınıflar bar grafiği ile sunulmuş ve çoğulluk kuralına göre o düğüme atanan sınıf belirtilmiştir.

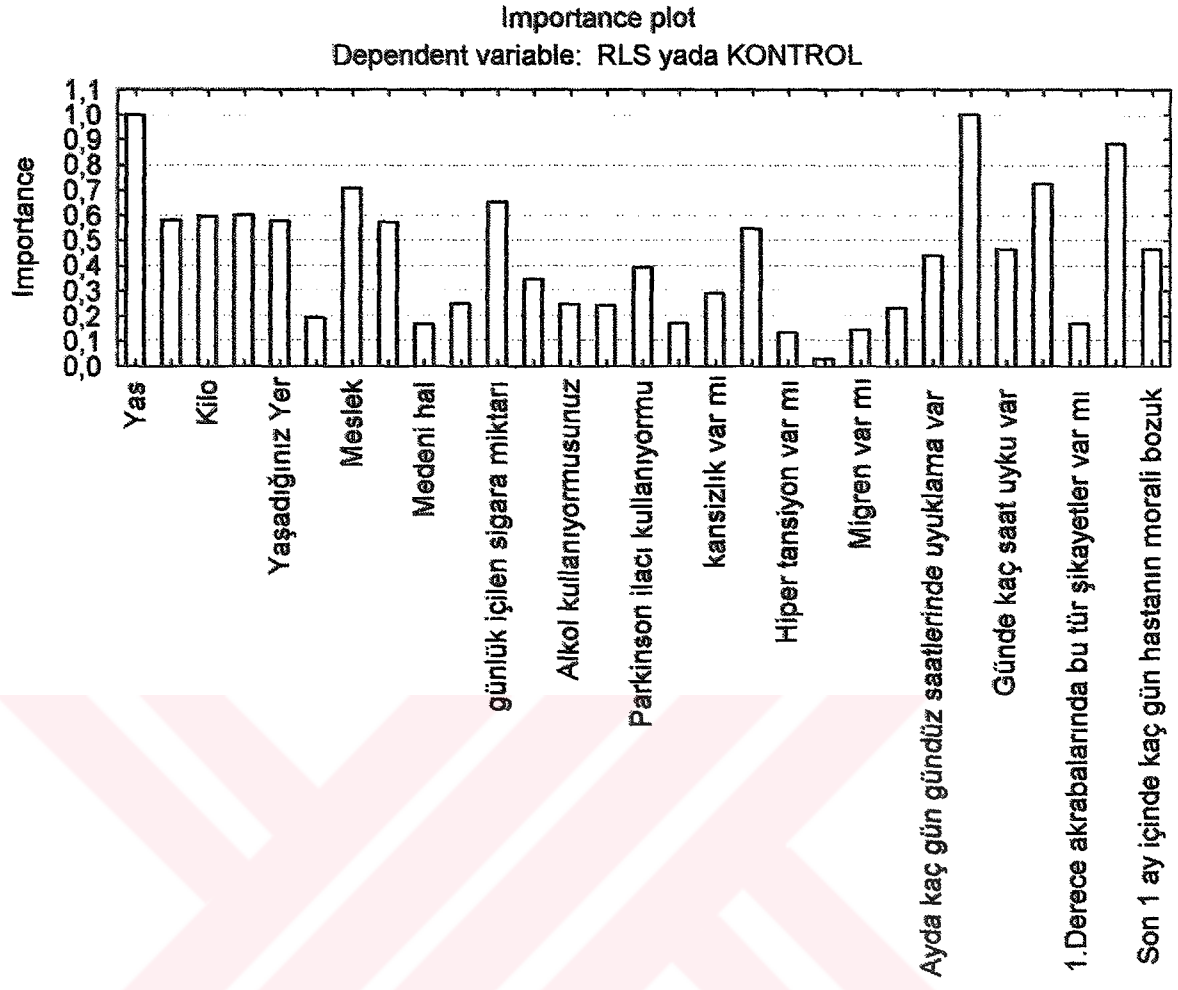
Şekil 6.1.2'de sunulan sınıflama ağacı başlangıçta 206 deneğin tümünü aynı grupta kabul ederek analize başlamıştır.

Aile düğümünü iki çocuk düğümüne ayıran ilk ayıraç *1. derece akrabalarınızda bacaklarda uyuşma, karıncalanma ve hareket ettirdikçe geçen bu tür şikayetlerin olup olmadığı* sorusudur. Bu soruya verilen *evet, hayır ya da bilmiyorum* cevapları ile aile düğümü iki çocuk düğümüne ayrılmıştır. Toplam 206 denek içerisinden, bu soruya cevabı *evet* olan 51 denek sol çocuk düğümüne (Düğüm 2), cevabı *hayır ya da bilmiyorum* olan 155 denek ise sağ çocuk düğümüne (Düğüm 3) ayrılır. Düğüm 2 ve Düğüm3 henüz saf düğüm olmamakla birlikte, bu düğümler içerisinde yer alan deneklerin ait oldukları sınıflar çoğulluk kuralına göre atanır.

Henüz saf olmayan 2 nolu düğümü saflaştırmak için kullanılan ayıraç *kilo* sorusudur. Kilosu *53,5 ve daha küçük* olan 6 denek 4 nolu sol terminal düğüme Kontrol grubu olarak atanmışlardır. Bu düğümde karar verme gerçekleşmiştir ve tekrar bölünme olmaz. *Kilosu 53,5 altında* olanlar 45 denek RLS grubu olarak 5 nolu sağ çocuk düğümüne atanmışlardır.Bu düğümde henüz karar verme gerçekleşmemiştir. 5 nolu çocuk düğümünü homojenleştirmek için ayıraç olarak *günde kaç saat uyku uyursunuz* sorusu kullanılmıştır. *Günde 2-4 saat,5-7 saat ve 8-10saat* uyuyan 42 denek RLS grubu olarak 6 nolu sol çocuk düğümüne atanmıştır. *Günde 2 saat'ten az ve 10 saatten fazla* uyuyan 3 denek 7 nolu terminal düğüme Kontrol grubu olarak atanmıştır. Bu düğümde de karar verme gerçekleşmiştir. 6 nolu çocuk düğümü henüz saf olmadığından *Yaş* sorusu ayıraç olarak kullanılır. *Yaşı 58 ve daha küçük* olan 35 denek sol çocuk düğümüne RLS grubu olarak atanmıştır. *Yaşı 58'den büyük* olan 7 denek 9 nolu terminal düğüme RLS grubu olarak atanmıştır. Bu düğümde de karar verme gerçekleşmiştir. 8 nolu çocuk düğümünü homojen hale getirmek için *1 ay içinde deneğin kaç gece rüya görürsünüz* sorusu kullanılır. *Hiç rüya görmeyen* 1 denek 10 nolu terminal düğüme Kontrol grubu olarak, *1-5gün,6-15gün,15günden fazla rüya gören* 34denek ise 11 nolu terminal düğüme RLS grubu olarak atanırlar. 10 ve 11 nolu düğümlerde karar verme gerçekleşmiştir. Artık ilk ayıraçla RLS grubu olarak ayrılan 51 deneğin 41 (34+7) tanesi RLS, 10 (6+1+3) tanesi Kontrol grubu olarak sınıflanmıştır.

Çocuk düğümü olan 3 nolu çocuk düğümünü saflaştırmak için deneklere *ayda kaç gün gündüz saatlerinde uyursunuz* sorusu sorulmuştur. *0, 6-15 gün,15 günden fazla* olarak cevap veren 32 denek 14 nolu sol terminal düğüme RLS grubu olarak, *1-5gün* olarak

cevap veren 123 denek ise 15 nolu düğüme Kontrol grubu olarak atanmıştır. Düğüm 14'de karar verme gerçekleşmiştir. 15 nolu Kontrol düğümü içerisindeki 123 Kontrol grubunu ayırmak için deneğe *son 1 ay içinde kaç gün moraliniz bozuktu* sorusu ayıraç olarak sorulmuştur. Cevabı *0-10,11-20* olan 89 birey 18 nolu sol çocuk düğümüne Kontrol grubu olarak atanır. Cevabı 21-30 olan 34 birey 19 nolu sağ çocuk düğümüne RLS grubu olarak atanır. 18 ve 19 nolu düğümlerde henüz karar verme gerçekleşmemiştir. Bu düğümleri homojenleştirmek için 18 nolu sol çocuk düğümüne *ayda kaç gece uykudan uyanma var* sorusu ayıraç olarak sorulmuştur. 18 nolu düğümde cevabı *hiç,1-5,6-15gece* olan 74 denek 20 nolu sol çocuk düğümüne Kontrol grubu olarak, cevabı *15 geceden fazla* olan 15 denek 21 nolu terminal düğüme RLS grubu olarak atanmıştır. Bu düğümde de karar verme gerçekleşmiştir. 20 nolu çocuk düğümünü safsızlaştırmak için, *genel olarak sağlık durumunuz nasıldır* sorusu ayıraç olarak sorulmuştur. *Sağlık durumunun kötü* olduğunu söyleyen3 denek 22 nolu sol terminal düğümüne RLS grubu olarak atanmıştır. Bu düğümde de karar verme gerçekleşmiştir. Sağlık durumunun *iyi,orta.çok iyi, mükemmel* olduğunu söyleyen 71 denek 23 nolu sağ çocuk düğümüne Kontrol grubu olarak atanmıştır. Bu düğümü safsızlaştırmak için ayıraç olarak sorulan soru deneklerin *Parkinson ilacı kullanıp kullanmadıkla*rıdır. *Kullanan* 1 denek 24 nolu sol terminal düğümüne RLS grubu olarak, *kullanmayan* 70 denek ise 25 nolu sağ terminal düğümüne Kontrol grubu olarak atanmıştır. 19 nolu düğüme RLS grubu olarak atanan 34 deneğe *meslek* sorusu ayıraç olarak sorulmuştur. Meslekleri *devlet memuru ve ev hanımı* olan 27 denek 32 nolu sol çocuk düğümüne Kontrol grubu olarak, *diğer mesleklerden* olan 7 denek 33 nolu sol terminal düğüme RLS grubu olarak atanmıştır. 33 nolu düğümde karar verme gerçekleşmiştir. 32 Nolu çocuk düğümü saflaştırmak için *yaş* sorusu tekrar sorulur. Yaşı *66,5'den büyük* olan 3 denek RLS grubu olarak 36 nolu sağ terminal düğümüne , yaşı *66,5'den küçük ve eşit* olan 24 denek Kontrol grubu olarak sol terminal düğümüne atanmıştır. İlk ayıraçla Kontrol grubu olarak ayrılan 155 deneğin 61 (32+3+1+15+3+7) tanesi RLS grubu olarak, 94 (70+24) tanesi Kontrol grubu olarak sınıflanmıştır.



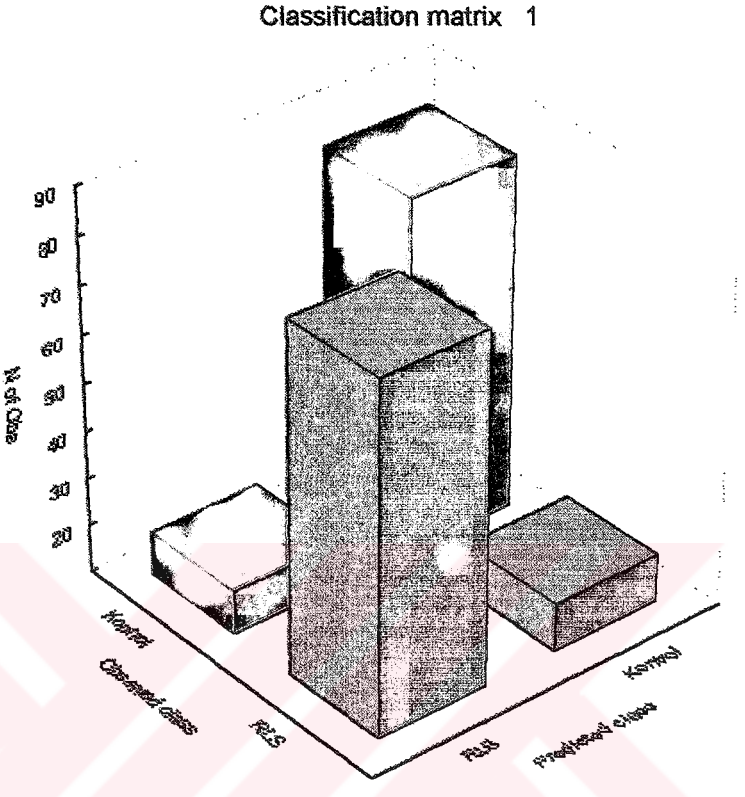
Şekil 6.1.3: Analiz I için oluşturulan maximal sınıflama ağacı oluşumunda kullanılan bağımsız değişkenlerin sınıflamada önemlilik grafiği.

Kategorik Bağımsız Değişkenler	Önem Derecesi
Yaşınız:	0,999066
Öğrenim Durumunuz:	0,579969
Kilonuz:	0,595107
Boyunuz:	0,600061
Yaşadığınız Yer:	0,576457
Cinsiyetiniz:	0,193112
Mesleginiz:	0,705905
Evinizin deniz kıyısından uzaklığı:	0,571130
Medeni haliniz:	0,166485
Sigara içiyor musunuz:	0,247546
Günlük içilen sigara miktarı:	0,651880
Geçmişte sigara alışkanlığınız var mı:	0,346515
Alkol kullanıyor musunuz:	0,245528
Antidepresan ilac kullanıyor musunuz:	0,240816
Parkinson ilacı kullanıyor musunuz:	0,392948
Beyin yada omiriliğe ait bir rahatsızlık geçirdiniz mi:	0,170288
Kansizlık var mı:	0,289887
Böbrek yetmezliği var mı:	0,546239
Hipertansiyon var mı:	0,133960
Diyabet var mı:	0,030154
Migren var mı:	0,145035
Depresyon var mı:	0,230257
Ayda kaç gündüz saatlerinde uyuklama var:	0,439431
Ayda kaç gece uykudan uyanma var:	1,000000
Günde kaç saat uyursunuz:	0,465177
Kaç gece rüya görüyorsunuz:	0,725590
1.Derece akrabalarınızda bu tür şikayetler var mı:	0,188598
Genel olarak sağlık durumunuz nasıl:	0,883795
Son 1 ay içinde kaç gün moraliniz bozuktu:	0,464573

Çizelge 6.1.4: Analiz I için oluşturulan maximal sınıflama ağacı oluşumunda kullanılan bağımsız değişkenlerin sınıflamada önemlilik dereceleri

Şekil 6.1.3 ve çizelge 6.1.4'de maximal ağaçta kullanılan bağımsız değişkenleri sınıflamadaki önem derecelerine ait değerler sunulmaktadır. Bu grafikte yer alan bağımsız değişkenlerin önem dereceleri 0 ile 1 arasında değişen olasılık değerleridir ve söz konusu değişkenin tanı koymadaki başarısının 100 puan üzerinden değerlendirme

sonucudur. Önem dereceleri "1" değerine yakın değişkenler ayırmada önemli, "0" değerine yakın değişkenler ise ayırmada önemsiz değişkenler olarak değerlendirilir.



Şekil 6.1.4: Analiz I için oluşturulan maximal sınıflama ağacına ait sınıflama bar grafiği

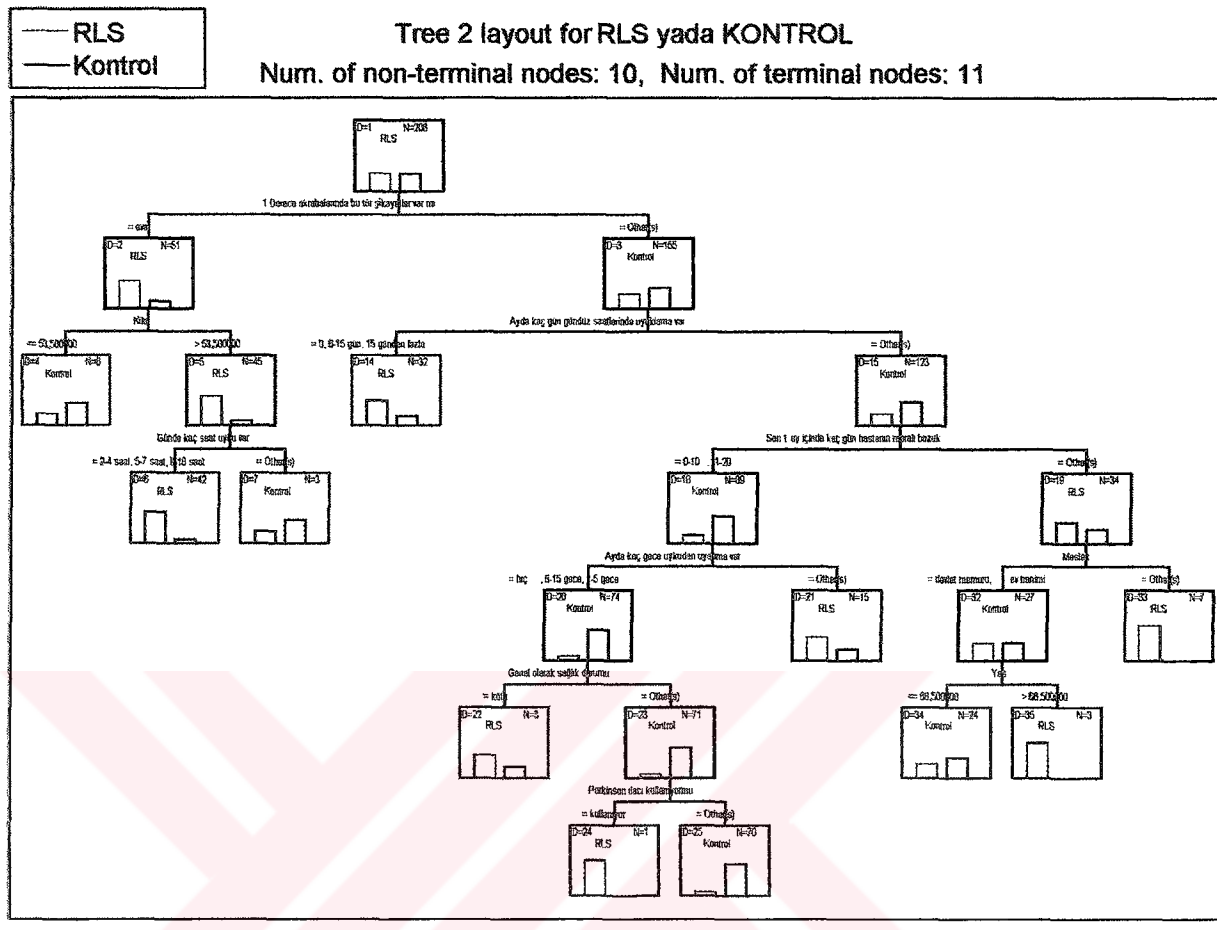
Çizelge 6.1.5: Analiz I için oluşturulan maximal sınıflama ağacına ait sınıflama matrisi

	Geçek Sınıf		
Tahmin sınıfı	RLS	Kontrol	Toplam
RLS	83	19	102
Kontrol	20	84	104
Toplam	103	103	206

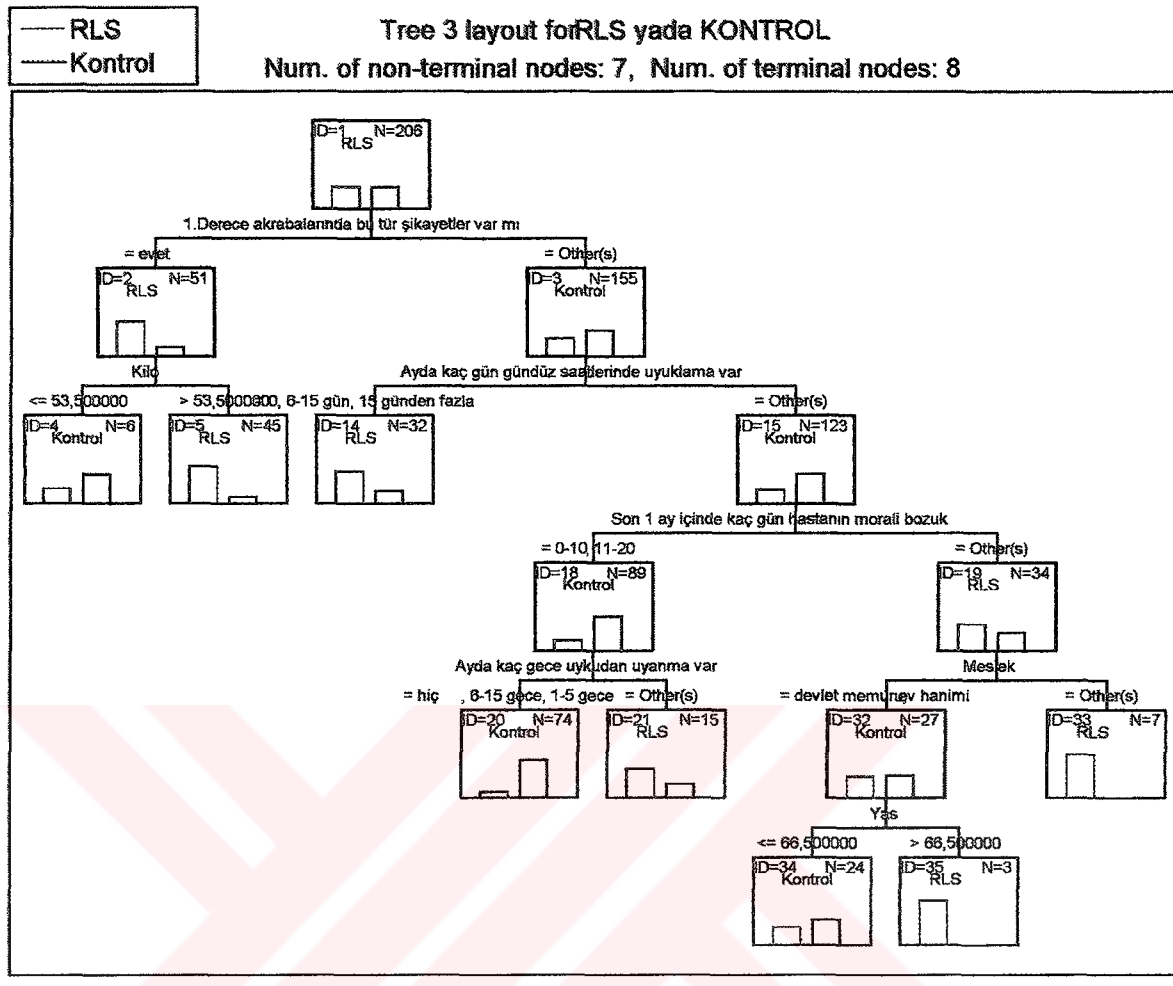
Şekil 6.1.4 ve Çizelge 6.1.5'den yararlanarak maximal ağacın hatalı sınıflama ve doğru sınıflama oranını aşağıdaki gibi hesaplayabiliriz.

Hatalı Sınıflama oranı=(19+20)/206=0,189

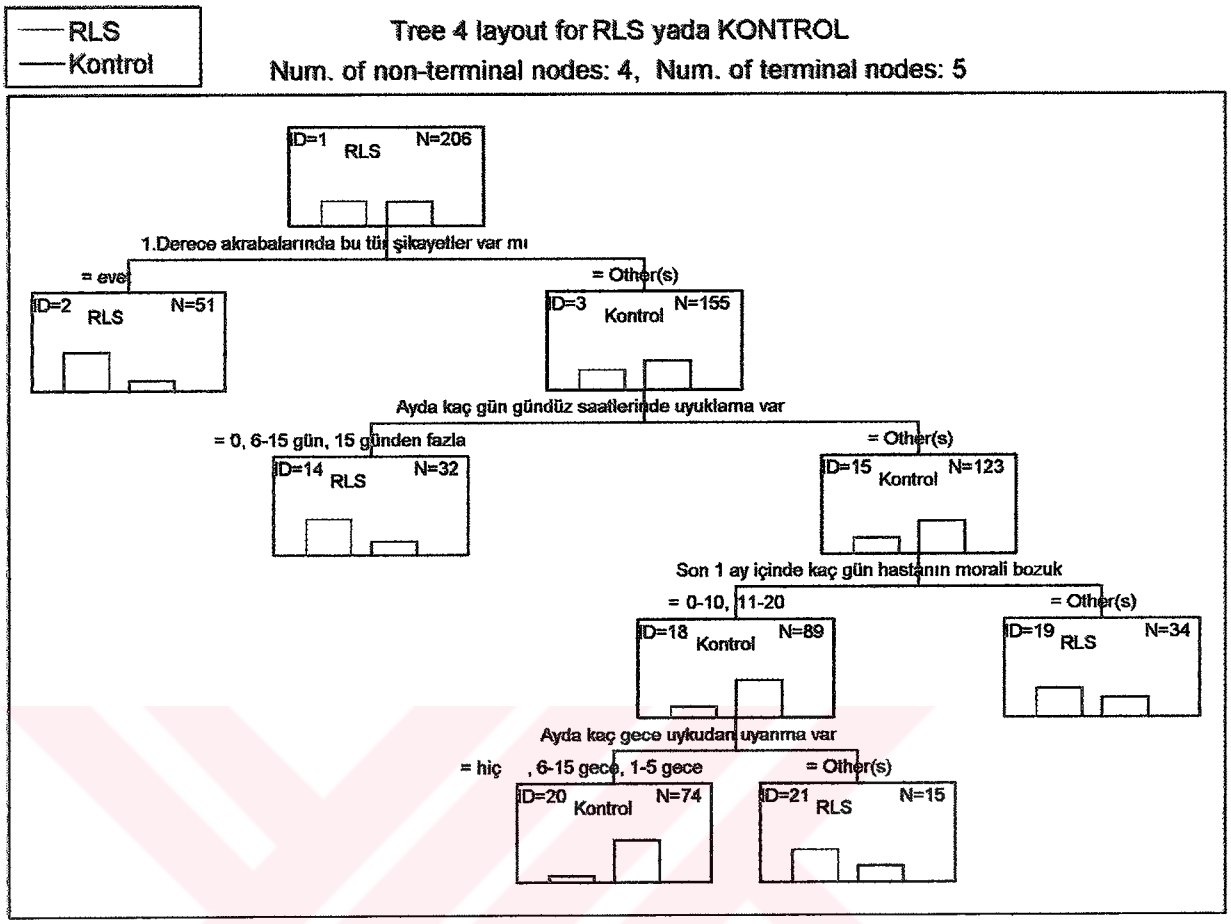
Doğru Sınıflama Oranı=1-0,189=0,81



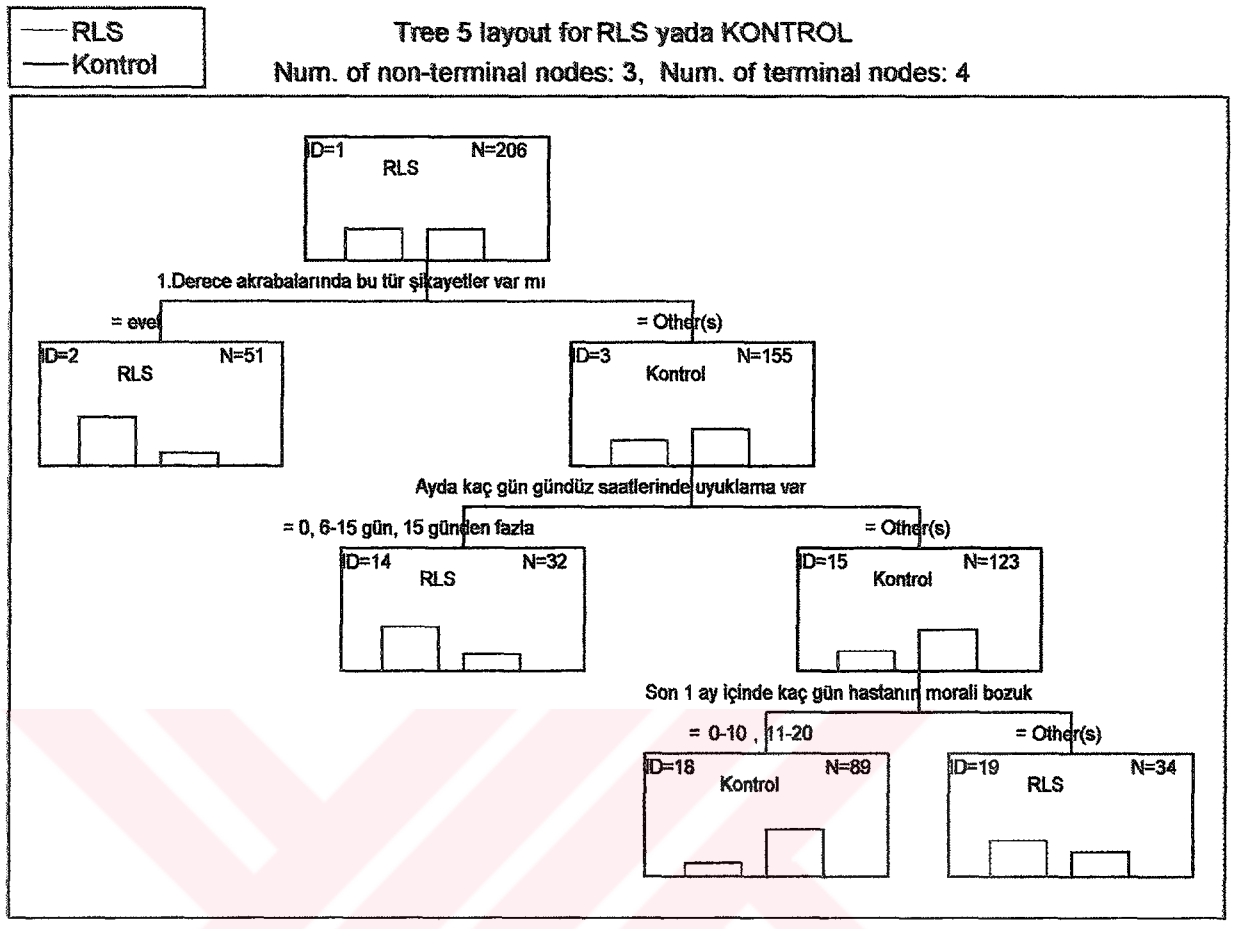
Şekil 6.1.5 : Analiz I için oluşturulan 2 nolu budanmış sınıflama ağaç diyagramı.



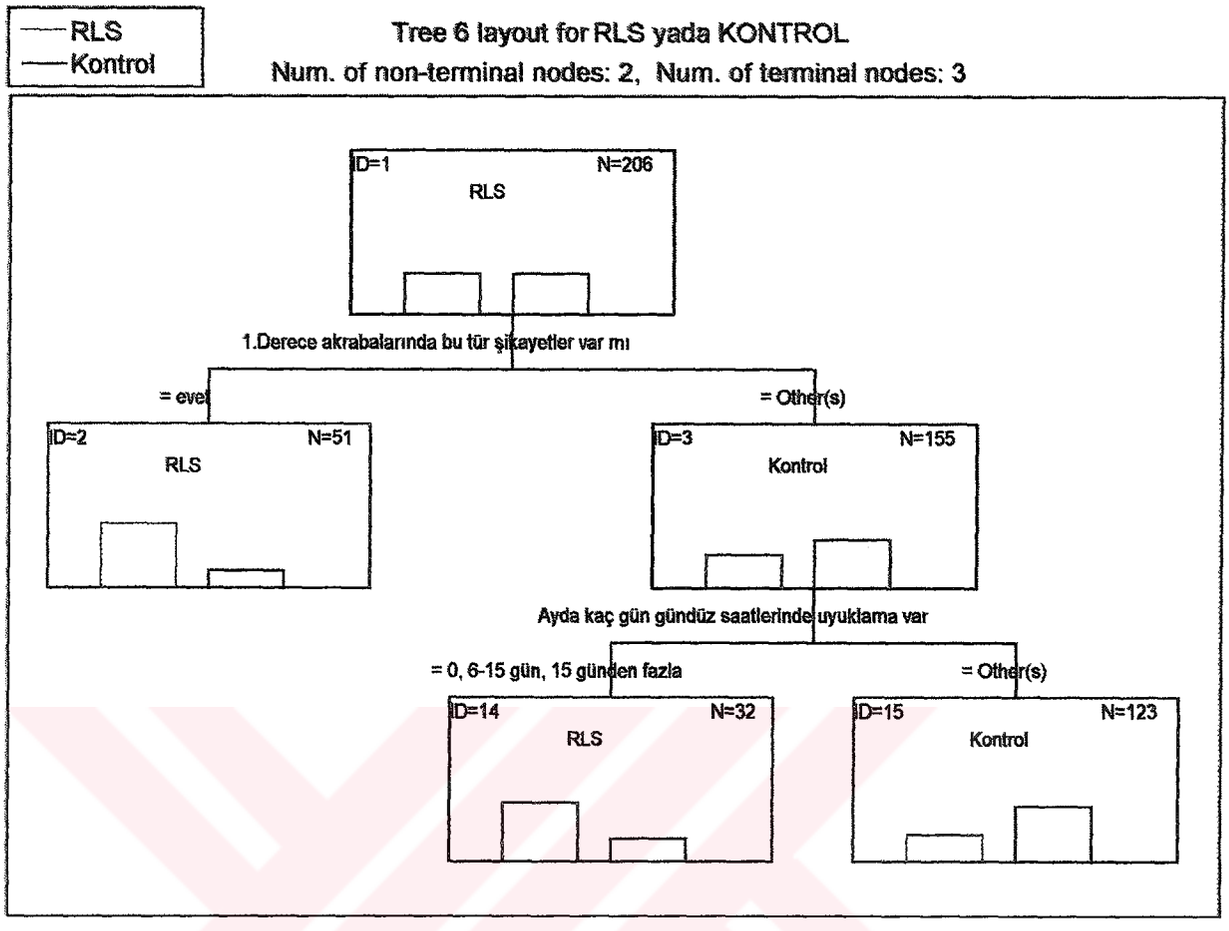
Şekil 6.1.6: Analiz I için oluşturulan 3 nolu budanmış ağaç diyagramı.



Şekil 6.1.7: Analiz I için oluşturulan 4 nolu budanmış ağaç diyagramı.

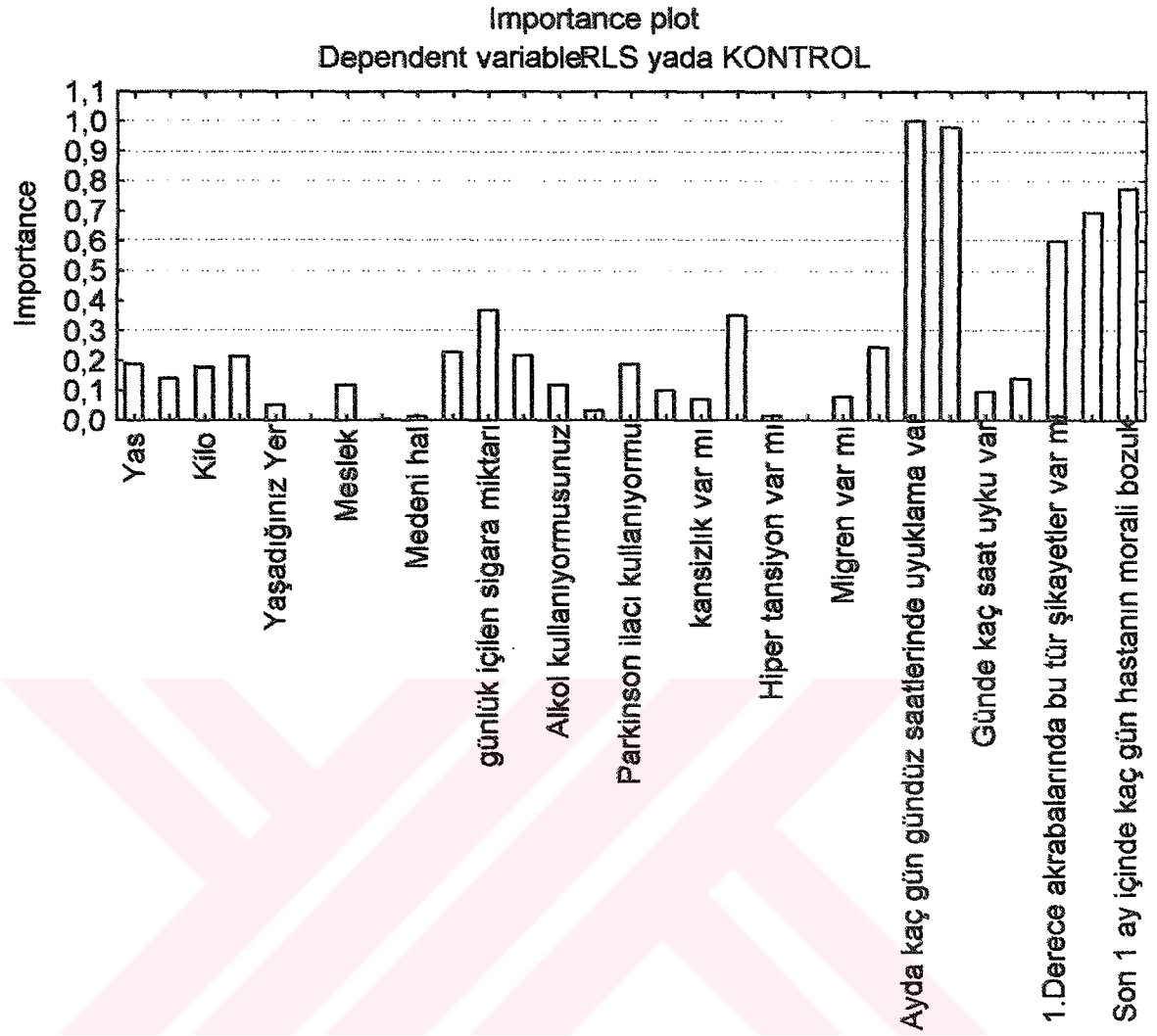


Şekil 6.1.8: Analiz I için oluşturulan 5 nolu budanmış ağaç diyagramı



Şekil 6.1.9: Analiz I için oluşturulan Optimal sınıflama ağacına ait diyagram.

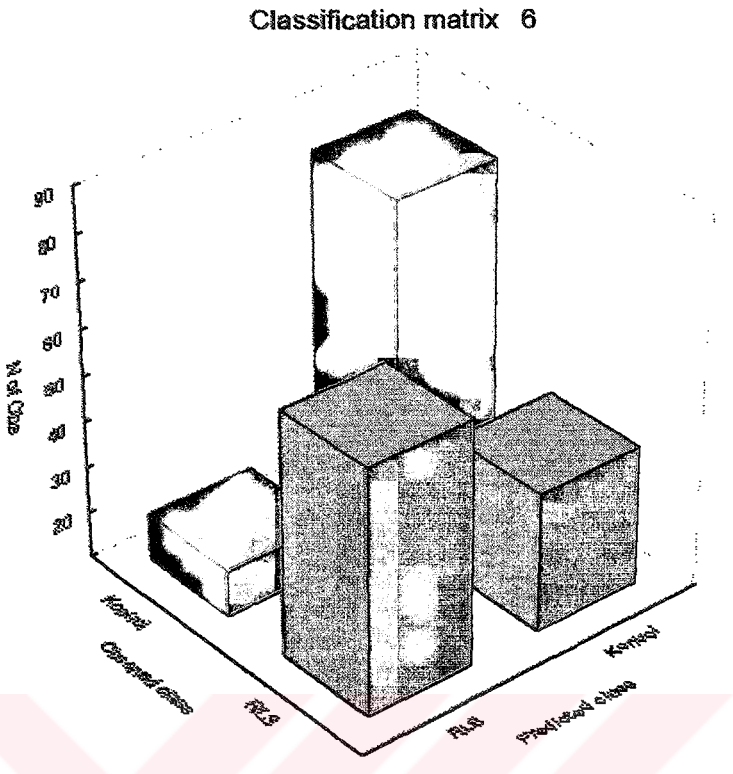
Şekil 6.1.9 sunulan maximal sınıflama ağacı;
10 katlı çapraz geçerlilikle budanarak Şekil 6.1.9'da sunulan optimal sınıflama ağacı inşa edilmiştir.



Şekil 6.1.10: Analiz I için oluşturulan optimal sınıflama ağacı oluşumunda kullanılan bağımsız değişkenlerin sınıflamada önemlilik grafiği

Çizelge 6.1.6: Analiz I için oluşturulan optimal sınıflama ağacı oluşumunda kullanılan bağımsız değişkenlerin sınıflamada önemlilik oranları

	Predictor importance 6 Responses: RLS yada I Options: Categorical re	
	Variable rank	Importance
Yas	19	0,187097
Öğrenim durumu	14	0,138504
Kilo	18	0,176268
Boy	21	0,212924
Yaşadığınız Yer	5	0,052019
Cinsiyet	0	0,001321
Meslek	12	0,119455
Evinizin deniz kıyısından uzaklığı	0	0,003477
Medeni hal	1	0,012768
Sigara içiyorsunuz	23	0,227926
günlük içilen sigara miktarı	37	0,368522
Geçmişte sigara alışkanlığı var mı	22	0,215936
Alkol kullanıyormusunuz	12	0,118390
Antidepresan ilacı kullanıyormusunuz	3	0,033610
Parkinson ilacı kullanıyormu	19	0,187097
Beyin yada omurilik ameliyatı geçirmiş mi	10	0,098838
kansizlik var mı	7	0,069626
Bobrek yetmezliği var mı	35	0,349435
Hiper tansiyon var mı	2	0,015764
Diyabet var mı	0	0,000031
Migren var mı	8	0,079060
Depresyon var mı	24	0,244001
Ayda kaç gün gündüz saatlerinde uyuklama var	100	1,000000
Ayda kaç gece uykudan uyanma var	98	0,979276
Günde kaç saat uyku var	10	0,095641
Kaç gece rüya görüyor	14	0,138935
1.Derece akrabalarında bu tür şikayetler var mı	60	0,599559
Genel olarak sağlık durumu	69	0,694929
Son 1 ay içinde kaç gün hastanın moralı bozuk	77	0,772234



Şekil 6.1.11: Analiz I için oluşturulan optimal sınıflama ağacına ait sınıflama bar grafiği

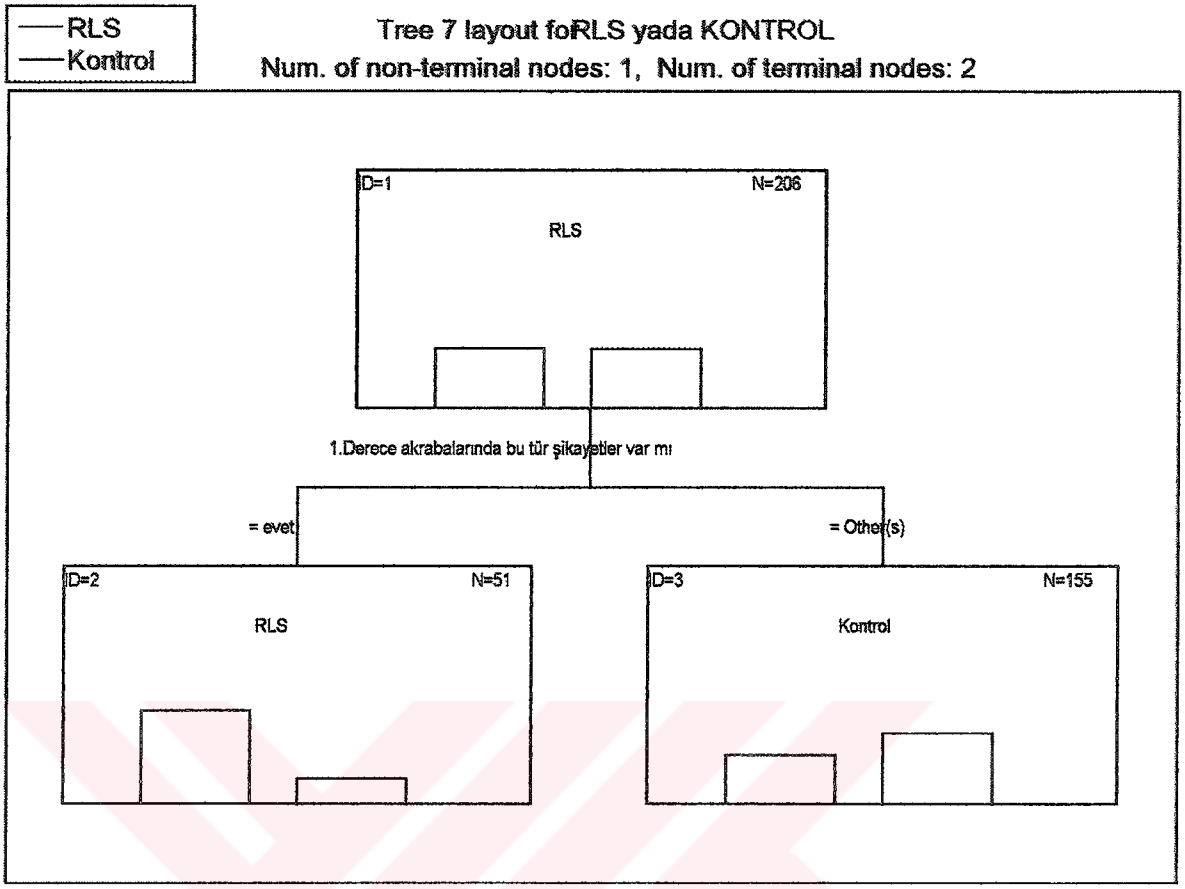
Çizelge 6.1.7: Analiz I için oluşturulan optimal sınıflama ağacına ait sınıflama matrisi.

	Geçek Sınıf		
Tahmin sınıfı	RLS	Kontrol	Toplam
RLS	63	20	83
Kontrol	40	83	123
Toplam	103	103	206

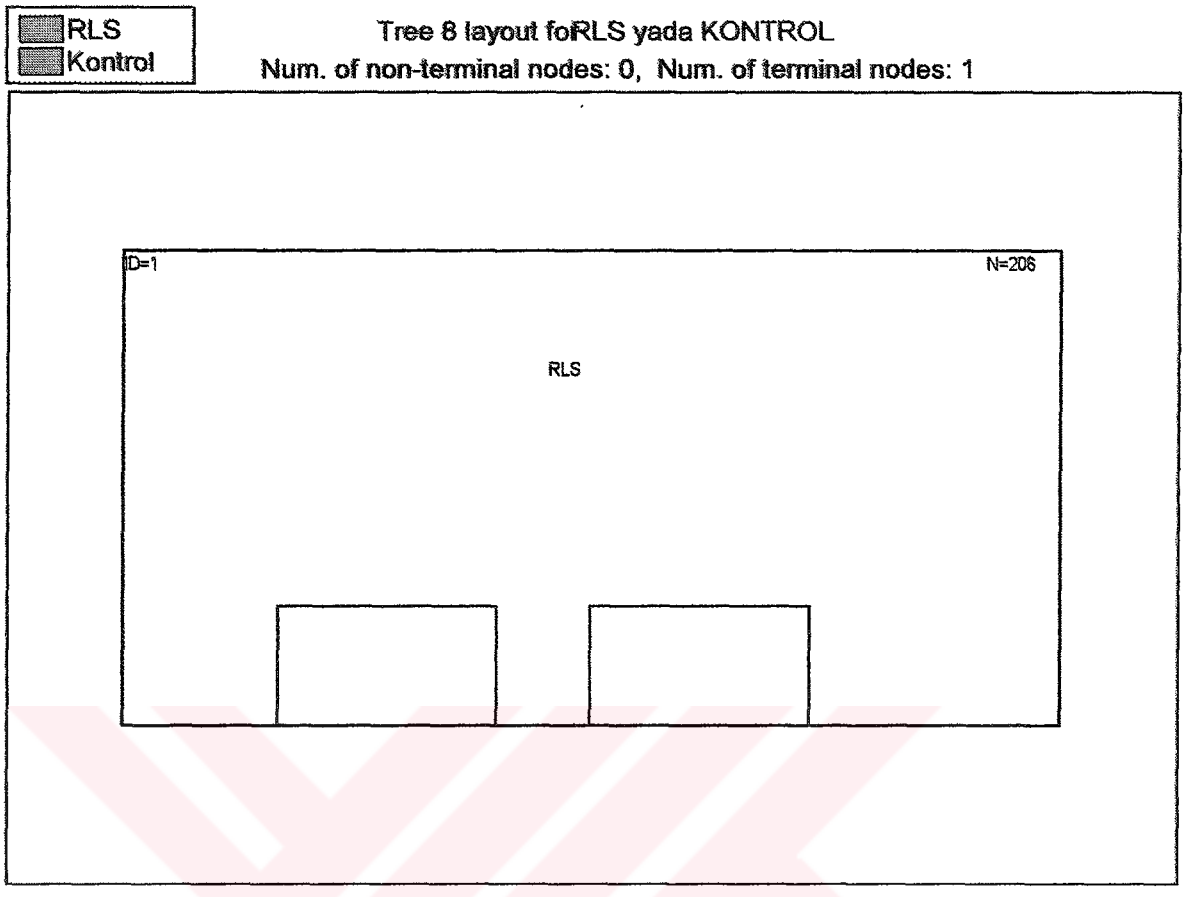
Şekil 6.1.11ve Çizelge 6.1.7''den yararlanarak optimal ağacın hatalı sınıflama ve doğru sınıflama oranını aşağıdaki gibi hesaplayabiliriz.

Hatalı sınıflama oranı:40+20/206=0,291

Doğru sınıflama oranı=1-0,291=0,709



Şekil 6.1.12: Analiz I için oluşturulan 7 nolu budanmış ağaç diyagramı



Şekil6.1.13: Analiz I için oluşturulan 8 nolu budanmış ağaç diyagramı

6.2. İKİNCİ ANALİZ:

Bu analizin amacı, anket formunda yer alan ve RLS hastalığı için önemli risk unsuru olan, kadınlara ait soruları (gebelik ve menopozla ilgili sorular) analize dahil etmektir.

Veriler ilk olarak cinsiyet faktörüne göre (kadın-erkek) ayrılmıştır. Analize toplam 128 kadın denek katılmış ve bu deneklerden 64 tanesinin RLS grubu, 64 tanesinin de Kontrol grubu olduğu tespit edilmiştir. Analize RLS ve Kontrol grubu 2 seviyeli kategorik bağımlı değişken olarak alınmış ve analizde kullanılan; sürekli bağımsız değişkenler Çizelge 6.2.1, kategorik bağımsız değişkenler ise Çizelge 6.2.2'de tanımlayıcı istatistikleri ile verilmiştir.

Çizelge 6.2.1: Analiz II'de kullanılan sürekli bağımsız değişkenlere ait tanımlayıcı istatistikler.

	RLS			KONTROL		
Sürekli Değişken	$\bar{X} \mp SD$	Min	Max	$\bar{X} \mp SD$	Min	Max
Yaş	42.5±15.05	18	79	42.20±15.36	19	75
Kilo	66.93±12.34	45	100	65.76±14.13	45	100
Boy(cm.)	1.58±0.05	1.50	1.75	1.57±0.04	1.47	1.70
Öğrenim Süresi(yıl)	3.64±3.37	0	13	4.07±3.38	0	14

Çizelge6.2.2:Analiz II'de kullanılan kategorik bağımsız değişkenlere ait tanımlayıcı istatistikler.

Kategorik Bağımsız Değişkenler	Kategorik Bağımsız Değişkenlerin Seviyeleri	RLS		KONTROL	
		n	%	n	%
Mesleğiniz:	Ev hanımı	60	93.75	61	95.31
	Öğrenci	0	0	2	3.123
	Çiftçi	1	1.56	0	0
	Devlet memuru	1	1.56	0	0
	Emekli	1	1.56	1	1.56
	Esnaf	0	0	0	0
	Diğer	1	1.56	0	0
Yaşadığınız yer:	İl	36	56.25	35	54.68
	İlçe	17	26.57	17	26.57
	Köy	11	17.18	12	18.75
Yaşadığınız yerin deniz kıyısına olan uzaklığı:	0-100m.	52	81.25	52	81.25
	101-500m.	7	10.93	6	93.38
	501-1000m.	5	7.81	5	7.81
	1001-2000m.	0	0	1	1.56
Medeni haliniz:	Evli	58	90.63	57	89.06
	Bekar	6	9.37	7	10.94
Sigara içiyor musunuz :	İçmiyorum	44	65.62	53	82.81
	İçiyorum	22	34.38	11	17.19
Günde ne kadar sigara içiyorsunuz:	Hiç içmedim	41	64.06	52	81.25
	Şimdi içmiyor	1	1.57	1	1.57
	Günde 10 adetten az	10	15.62	7	10.93
	Günde 10-19 adet	5	7.82	2	3.12
	Günde 1-2 paket	7	10.93	1	1.57
	Günde2 paketten fazla	0	0	1	1.57
Geçmişte sigara alışkanlığınız var mı	Hiç içmedim	41	64.06	52	81.25
	Var	23	35.94	12	18.75
Alkol kullanıyor musunuz:	Kullanmıyorum	4	6.25	3	4.68
	Ayda 10 duble rakıdan az	59	92.18	61	95.32
	Ayda 10 duble rakıdan fazla	1	1.57	0	0
Antidepressan ilaç kullanıyor musunuz :	Evet	5	7.81	6	9.38
	Hayır	59	92.19	58	90.62
Antiparkinson ilaç kullanıyor musunuz :	Evet	3	95.31	0	0
	Hayır	61	4.69	64	100
Beyin yada omurilik yada bu bölgelerle ilgili başka hastalık geçirdiniz mi:	Evet	7	10.93	5	7.82
	Hayır	57	89.07	59	92.18
Kansızlık hastalığınız var mı:	Evet	8	12.5	4	6.25
	Hayır	56	87.5	60	93.75
Böbrek yetmezliğiniz var mı:	Evet	7	10.93	1	1.57
	Hayır	57	89.07	63	98.43
Hipertansiyon hastalığınız var mı:	Evet	8	12.5	10	84.37
	Hayır	56	87.5	54	15.63
Diyabet hastalığınız var mı:	Evet	1	1.57	2	3.13
	Hayır	63	98.43	62	97.87

Çizelge6.2.2(Devam):Analiz II'de kullanılan kategorik bağımsız değişkenlere ait tanımlayıcı istatistikler

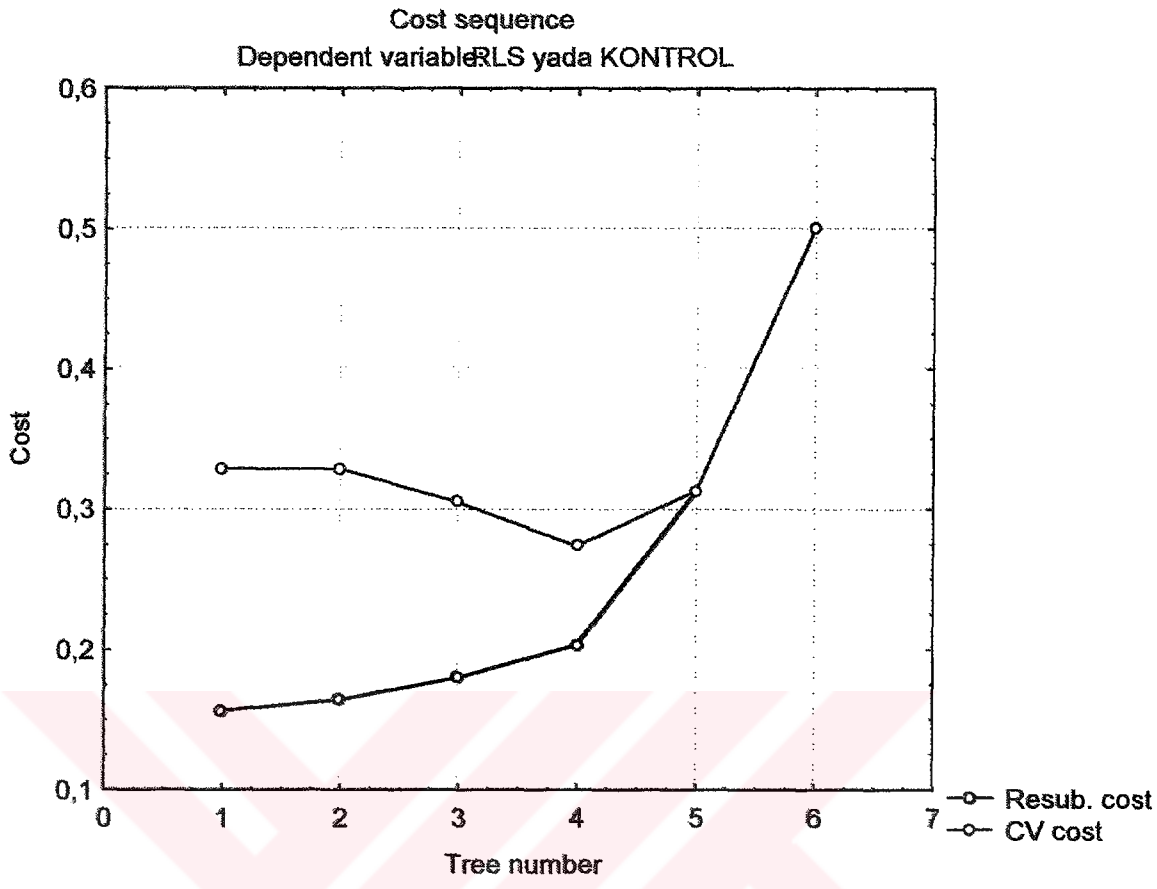
Migren hastalığınız var mı:	Evet	1	1.57	0	0
	Hayır	63	98.43	64	100
Depresyon hastalığınız var mı:	Evet	5	7.81	0	0
	Hayır	59	92.19	64	100
Ayda ortalama kaç gün gündüz saatlerinde uyuklarsınız:	Hiç	34	53.12	38	59.37
	1-5 gün	13	20.31	18	28.13
	6-15 gün	8	12.51	3	4.68
	15 günden fazla	9	14.06	5	7.82
Ayda ortalama kaç gece uyurken uykudan uyanırsınız:	Hiç	8	12.5	15	23.43
	1-5 gün	23	35.94	28	43.75
	6-15 gün,	12	18.75	10	15.64
	15 günden fazla	21	32.81	11	17.18
Günde ortalama kaç saat uyursunuz:	2 saatten az	0	0	1	1.56
	2-4 saat	2	3.13	1	1.56
	5-7 saat	33	51.56	28	43.75
	8-10 saat	27	42.18	30	46.88
	10 saatten fazla	2	3.13	4	6.25
Gebelik geçirdiniz mi:	Gebelik yok	9	15	6	9.68
	Gebelik geçirmiş	51	85	56	90.32
Menopoza girdiniz mi:	Girmiş	18	28.12	25	39,06
	Girmemiş	42	65.63	37	57.81
	Geçiş:	4	6.25	2	3.13
Ayda ortalama kaç gece rüya görürsünüz:	Hiç	5	7.81	2	3.13
	1-5 gece	24	37.5	34	53.13
	6-15 gece	16	25	14	21.87
	15 geceden fazla	19	29.69	14	21.87
Sağlığınız genel olarak nasıldır:	Mükemmel	0	0	4	6.26
	Çok iyi	12	18.75	19	29.68
	İyi	18	28.13	19	29.68
	Orta	29	45.31	19	29.68
	Kötü	5	7.81	3	4.68
Son 1 ay içinde kaç gün moraliniz bozuktu:	0-10 gün	16	25	19	29.69
	11-20 gün	16	25	26	40.62
	21-30 gün	32	50	19	29.69
1.Derece akrabalarınızda bu tür şikayetler var mı:	Evet	24	37.5	7	10.94
	Hayır yada bilmiyorum.	40	62.5	57	89.06

Sınıflama Ağacı analizinde ayırma kriteri olarak Gini ayırma kriteri, budama yöntemi olarak 10 katlı çapraz geçerlilik yöntemi tercih edilmiştir.Hasta ve Kontrol gruplarının sayıları eşit olduğu için önsel olasılıkları eşit (0.5) olarak alınmıştır. Statistica® 6.0 başlangıçta 6 sınıflama ağacı üretmiştir. Bu ağaçlara ait maliyet-karmaşıklık bilgileri Çizelge 6.2.3 'de sunulmuştur.

Çizelge 6.2.3: Analiz II için oluşturulan 6 sınıflama ağacına ait maliyet-karmaşıklık bilgileri

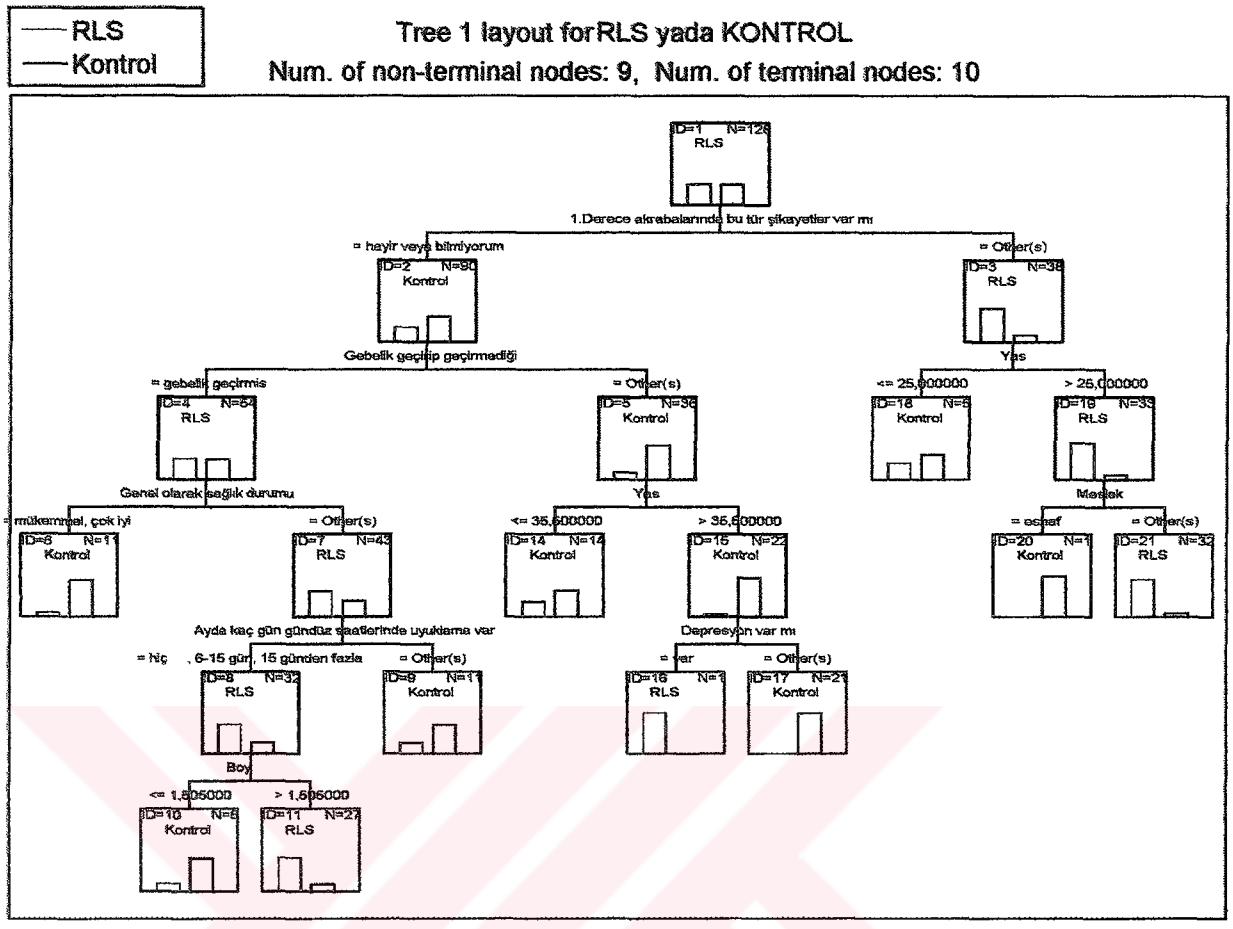
	Tree sequence (KADIN VERİSİ) Responses: RLS yada KONTROL Optimal tree denoted by *				
	Terminal nodes	CV cost	CV std. error	Resubstitution cost	Node complexity
Tree 1	10	0,328125	0,041501	**0,156250**	0,000000
Tree 2	8	0,328125	0,041501	0,164063	0,003906
Tree 3	6	0,304688	0,040683	0,179688	0,007813
Tree 4	5	0,273438	0,039397	0,203125	0,023438
*Tree 5	2	0,312500	0,040969	**0,312500**	0,036458
Tree 6	1	0,500000	0,044194	0,500000	0,187500

Ağaç 1(Tree 1) maximal ağaçtır ve 10 adet terminal düğüme sahiptir. Amaç maliyet-karmaşıklık ölçüsünü minimize etmek olduğundan, hatalı sınıflama maliyetleri (CV cost ve Resubstitution cost) ceza katsayısı (Node complexity, α) ve terminal düğüm sayısını (Terminal nodes,T) dengeleyen Çizelge 6.2.3'de, * ile işaretli olan 5 nolu ağacı (*Tree 5) optimal ağaç olarak seçmiştir. Budama artıkça terminal düğüm sayısı azalmıştır fakat hatalı sınıflama maliyetleri artmıştır. Analize giren bağımsız değişkenlerin daha fazla sayıda olması nedeniyle en iyi sınıflamanın yapıldığı maximal ağaçta hatalı sınıflama maliyetleri en düşüktür.



Şekil 6.2.1: Analiz II için oluşturulan ağaçların hatalı sınıflama maliyetleri

Şekil 6.2.1' de, oluşturulan 6 sınıflama ağacına ait hatalı sınıflama maliyetleri verilmektedir. Başlangıçta hatalı sınıflama maliyeti düşüktür. Budama arttıkça terminal düğüm sayısı ve dolayısı ile modele giren bağımsız değişken sayısı azaldığı için hatalı sınıflama maliyetleri yükselmiştir. En son oluşturulan ağacın (Tree 6) hatalı sınıflama maliyetleri maksimumdur. Bu ağaçta 128 deneğin tamamı RLS hastası olarak sınıflandırılmış ve böylece hatalı sınıflama oranı %50 olmuştur.



Şekil 6.2.2: Analiz II için oluşturulan maximal sınıflama ağaç diyagramı.

Şekil 6.2.2'de sunulan sınıflama ağacı değişkenler arasındaki ilişkileri en ayrıntılı biçimde gösteren maximal ağaçtır. Şekil 6.2.2'de sunulan sınıflama ağacı başlangıçta 128 deneğin tümünü RLS hastası kabul ederek analize başlamıştır.

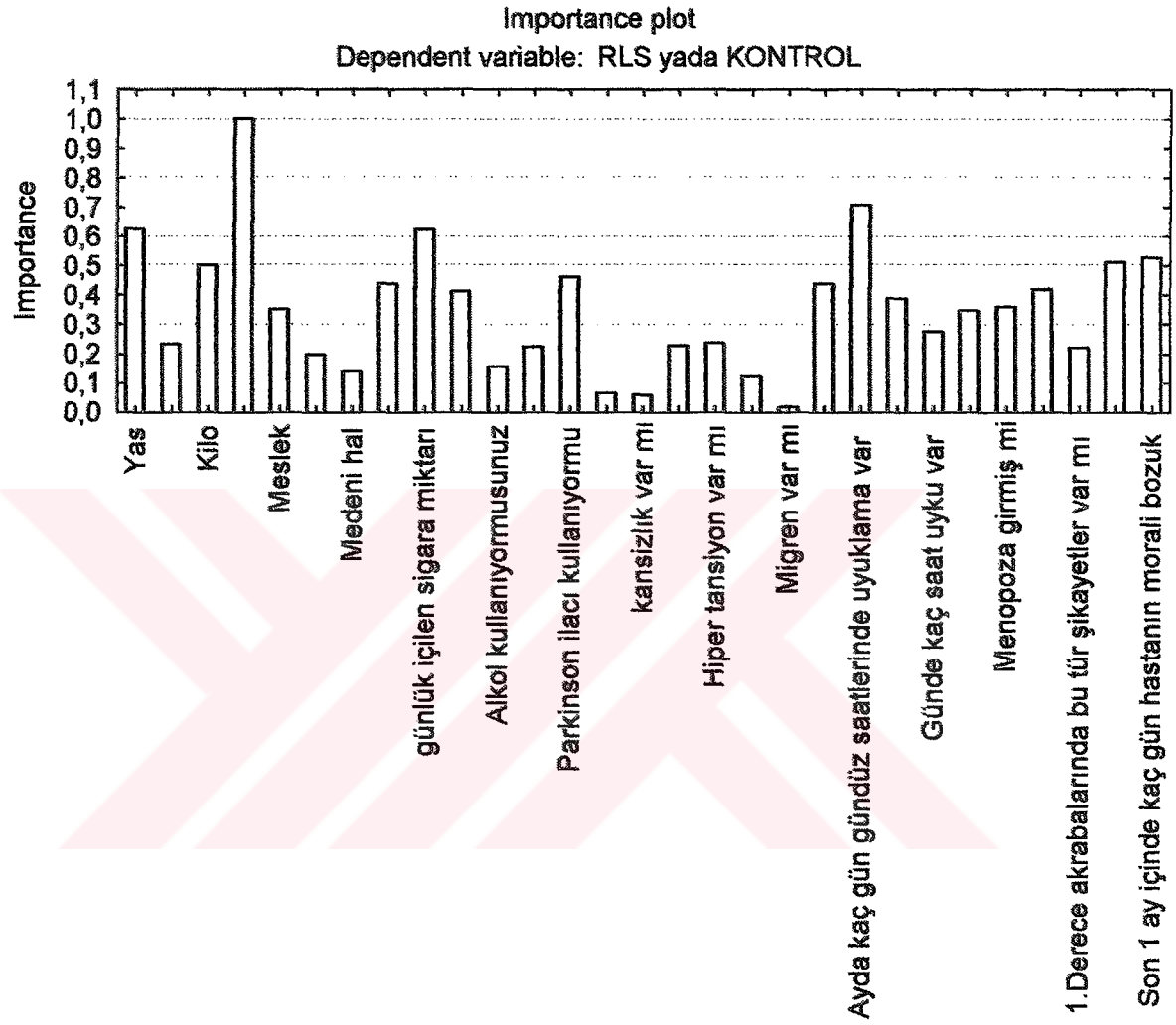
Aile düğümünü iki çocuk düğümüne ayıran ilk ayıraç *1. derece akrabalarınızda bacaklarda uyuşma, karıncalanma ve hareket ettirdikçe geçen bu tür şikayetlerin olup olmadığı* sorusudur. Bu soruya *Evet* cevabı veren 38 denek RLS grubu olarak 3 nolu sağ çocuk düğümüne, *hayır yada bilmiyorum* cevabı veren 90 denek 2 nolu sol çocuk düğümüne Kontrol grubu olarak atanmıştır. 2 ve 3 nolu düğümler henüz saf değildirler. 2 nolu düğümü saflaştırmak için kullanılan ayıraç deneklerin *gebelik geçirip geçirmediği* sorusudur. Bu soruya *gebelik geçirmiş* yanıtını veren 54 denek 4 nolu çocuk düğümüne RLS grubu olarak atanmış, *gebelik geçirmemiş* cevabını veren 36 denek 5 nolu sağ çocuk düğümüne Kontrol düğümü olarak atanmıştır. 4 nolu çocuk düğümünü saflaştırmak için kullanılan ayıraç, *genel olarak sağlık durumunuz nasıldır* sorusudur. Sağlık durumuna *mükemmel, çok iyi* yanıtını veren 11 denek 6 nolu sol

terminal düğümüne Kontrol grubu olarak atanmıştır. Sağlık durumuna iyi, orta, kötü yanıtını veren 43 denek 7 nolu sağ çocuk düğümüne RLS grubu olarak atanmıştır. 6 nolu düğümde karar verme tamamlanmıştır ve bu düğümün saf olduğuna karar verilir. 7 nolu çocuk düğümünü safsızlaştırmak için ayıraç olarak deneklere *1 ayda kaç gün gündüz saatlerine uyursunuz* sorusu sorulmuştur. Cevabı 1-5gün olan 11 denek Kontrol grubu olarak 9 nolu terminal düğüme atanmıştır. Bu düğümde de karar verme tamamlanmıştır.7 nolu çocuk düğümünde cevabı *hiç,6-15gün ve 15'günden fazla* olan 32 denek, 8 nolu çocuk düğümüne RLS grubu olarak atanmıştır. 8 nolu düğümü safsızlaştırmak için ayıraç olarak *boyunuz* sorusu sorulmuş ve cevabı *1,50cm.'den kısa ve eşit* olanlar 10 nolu sol terminal düğümüne Kontrol grubu olarak, cevabı *1,50cm.'den uzun* olanlar 11 nolu sağ terminal düğümüne RLS grubu olarak atanmıştır. Gebelik geçirmeyen ve 5 nolu çocuk düğümüne Kontrol olarak atanan 38 deneği saflaştırmak için *yaş* sorusu ayıraç olarak sorulur.Yaşı *35,5 eşit ve küçük* olan 14 denek 14 nolu sol terminal düğümüne Kontrol grubu olarak atanmıştır. Yaşı *35,5'dan büyük* olan 22 denek 15 nolu sağ çocuk düğümüne Kontrol grubu olarak atanmıştır. Bu düğümü safsızlaştırmak için deneklere *doktor tarafından tanısı konmuş depresyon rahatsızlıklarının olup olmadığı* sorusu sorulmuştur. Cevabı *evet* olan 1denek 16 nolu sol terminal düğüme RLS grubu olarak, cevabı *hayır* olan 21 denek 17 nolu sağ terminal düğüme Kontrol grubu olarak atanmıştır.

Başlangıçta ilk ayıraçla 2 nolu çocuk düğümüne Kontrol grubu olarak atanan 90 denek, maximal sınıflama ağaç inşasının sonunda , 28 (27+1) denek RLS grubu ve 62 (11+5+11+14+21) denek Kontrol grubu olarak sınıflanmıştır.

1.Derece akrabalarında şikayet olan ve ilk ayıraçla 3 nolu sağ çocuk düğümüne RLS grubu olarak atanan 38 deneği saflaştırmak için yaş sorusu ayıraç olarak sorulmuştur. Yaşı 25'e eşit ve daha küçük olan 5 denek 18 nolu sol terminal düğümüne Kontrol grubu olarak, yaşı 25'den küçük olan 33 denek 19 nolu sağ çocuk düğümüne RLS grubu olarak atanmıştır. 18 nolu düğümde karar verme tamamlanmıştır. 19 nolu düğümü safsızlaştırmak için mesleğiniz sorusu ayıraç olarak kullanılmış ve mesleği esnaf olan 1 denek 20 nolu sol terminal düğüme Kontrol grubu olarak, mesleği esnaflık dışında diğer mesleklerden olan 32 denek de 21 nolu sağ terminal düğüme RLS grubu olarak atanmıştır.

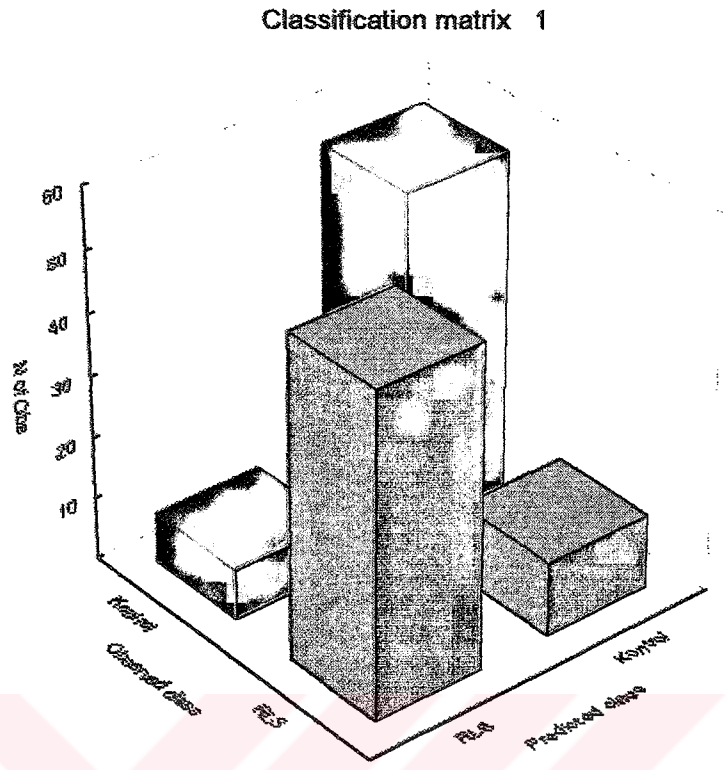
Başlangıçta ilk ayıraçla 3 nolu çocuk düğümüne RLS grubu olarak atanan 38 denek, maximal sınıflama ağaç inşasının sonunda , 32 denek RLS grubu ve 6 (5+1) denek Kontrol grubu olarak sınıflanmıştır.



Şekil 6.2.3: Analiz II için oluşturulan maximal sınıflama ağacı oluşumunda kullanılan bağımsız değişkenlerin sınıflamada önemlilik grafiği.

Çizelge6.2.4:AnalizII'dekullanılanbağımsızdeğişkenlereaitönemlilikoranları

	Predictor importance 1 Responses: RLS yada I Options: Categorical re	
	Variable rank	Importance
Yas	62	0,624434
Ogrenim durumu	23	0,231340
Kilo	50	0,500454
Boy	100	1,000000
Meslek	35	0,351934
Evinizin deniz kiyisindan uzakligi	20	0,196236
Medeni hal	14	0,138745
Sigara iciyorsunuz	44	0,437942
gunluk icilen sigara miktari	62	0,621773
Gecmiste sigara alişkanligi var mi	41	0,412537
Alkol kullaniyormusunuz	16	0,155650
Antidepresan ilaci kullaniyormusunuz	22	0,222261
Parkinson ilaci kullaniyormu	46	0,460914
Beyin yada omurilik ameliyati gecirmiş mi	6	0,064697
kansizlik var mi	6	0,059177
Bobrek yetmezligi var mi	23	0,228297
Hiper tansiyon var mi	24	0,236037
Diyabet var mi	12	0,121433
Migren var mi	2	0,017487
Depresyon var mi	44	0,438733
Ayda kaç gün gündüz saatlerinde uyuklama var	71	0,708027
Ayda kaç gece uykudan uyanma var	39	0,387048
Günde kaç saat uyku var	27	0,274890
Kaç gece rüya görüyor	35	0,347453
Menopoza girmiş mi	36	0,358771
Gebelik geçirip geçirmediği	42	0,417508
1 Derece akrabalarında bu tür şikayetler var mı	22	0,220131
Genel olarak sağlık durumu	51	0,511873
Son 1 ay içinde kaç gün hastanin morali bozuk	53	0,527583



Şekil 6.2.4: Analiz II için oluşturulan maximal sınıflama ağacına ait sınıflama bar grafiği

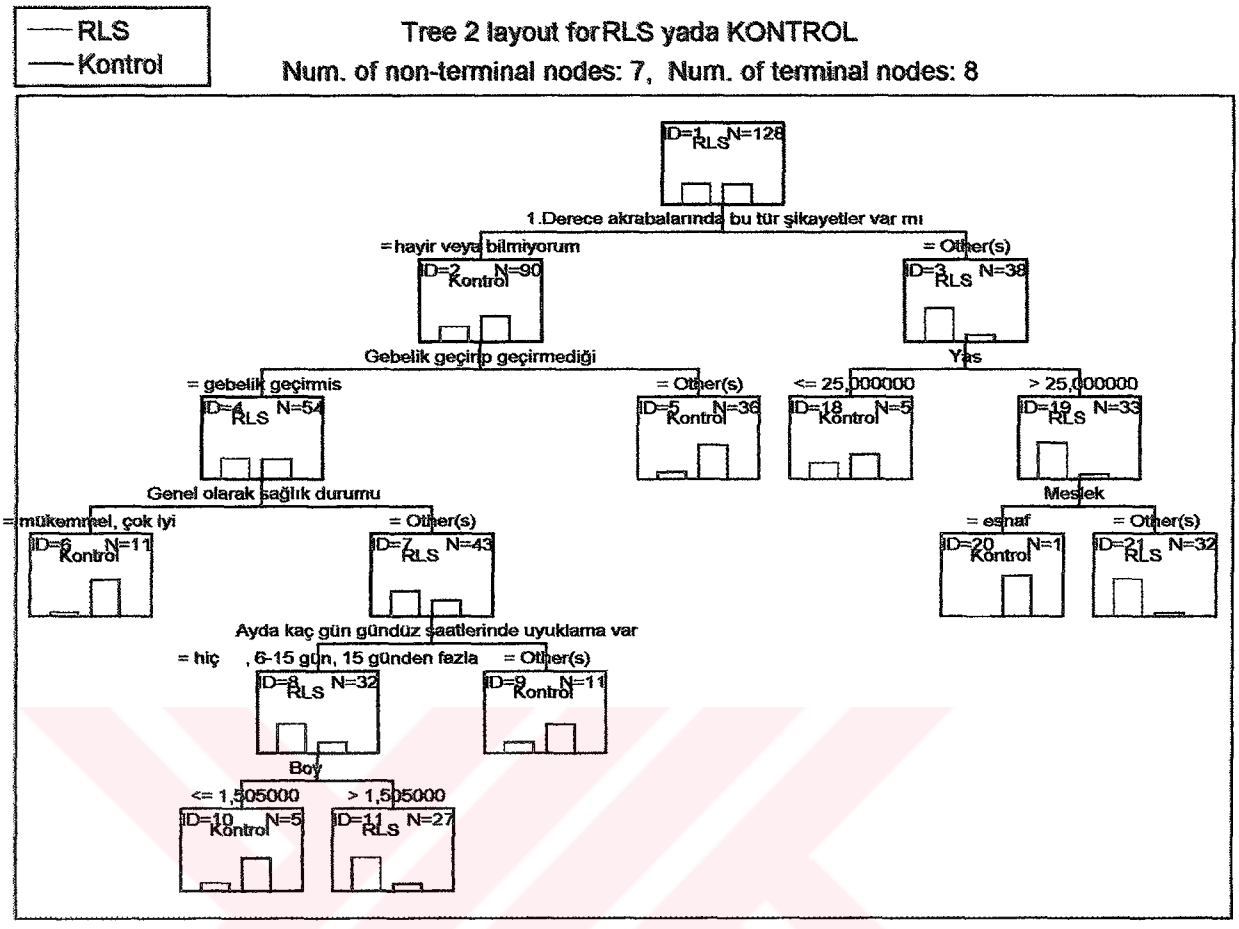
Çizelge 6.2.5: Analiz I için oluşturulan maximal sınıflama ağacına ait sınıflama matrisi

	Geçek Sınıf		
Tahmin sınıfı	RLS	Kontrol	Toplam
RLS	52	8	60
Kontrol	12	56	68
Toplam	64	64	128

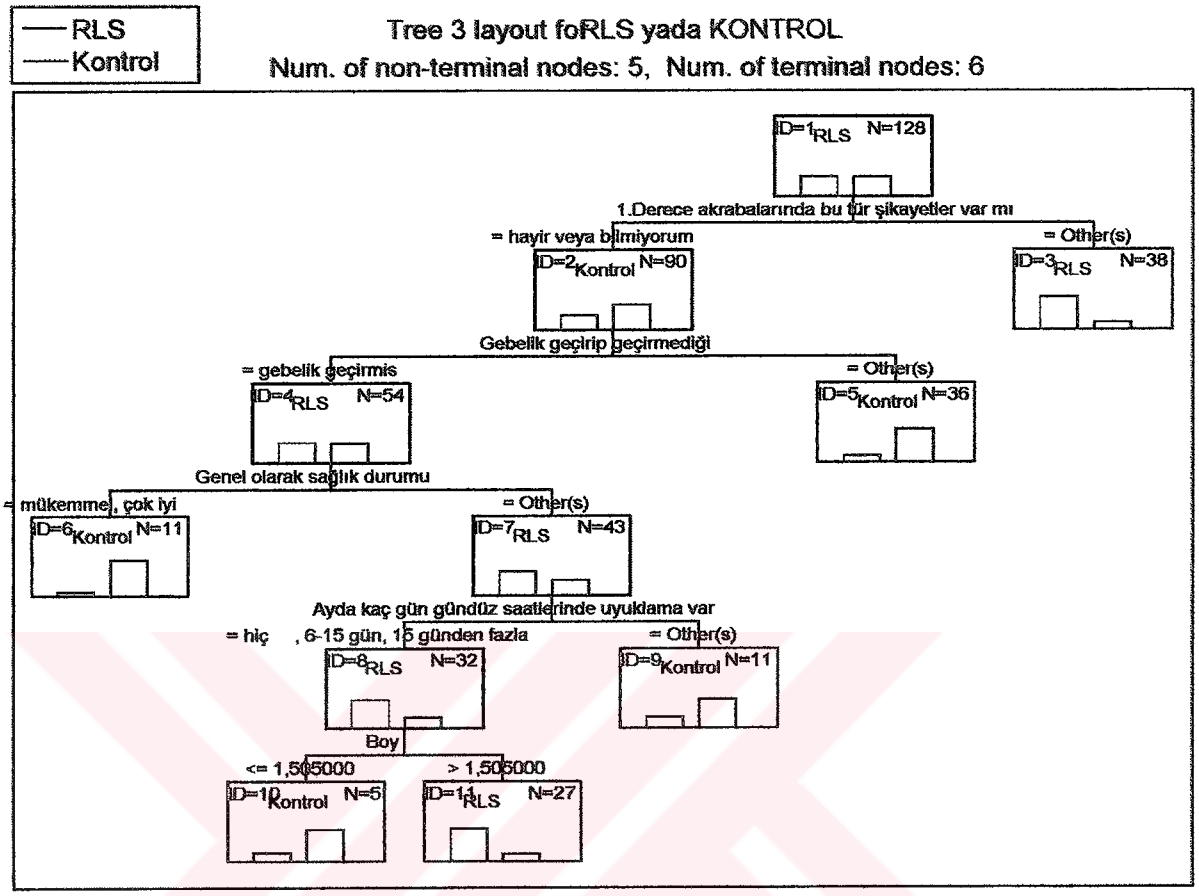
Şekil 6.2.4 ve Çizelge 6.2.5'den yararlanarak maximal ağacın hatalı sınıflama ve doğru sınıflama oranını aşağıdaki gibi hesaplayabiliriz.

Hatalı Sınıflama oranı=(8+12)/128=0,156

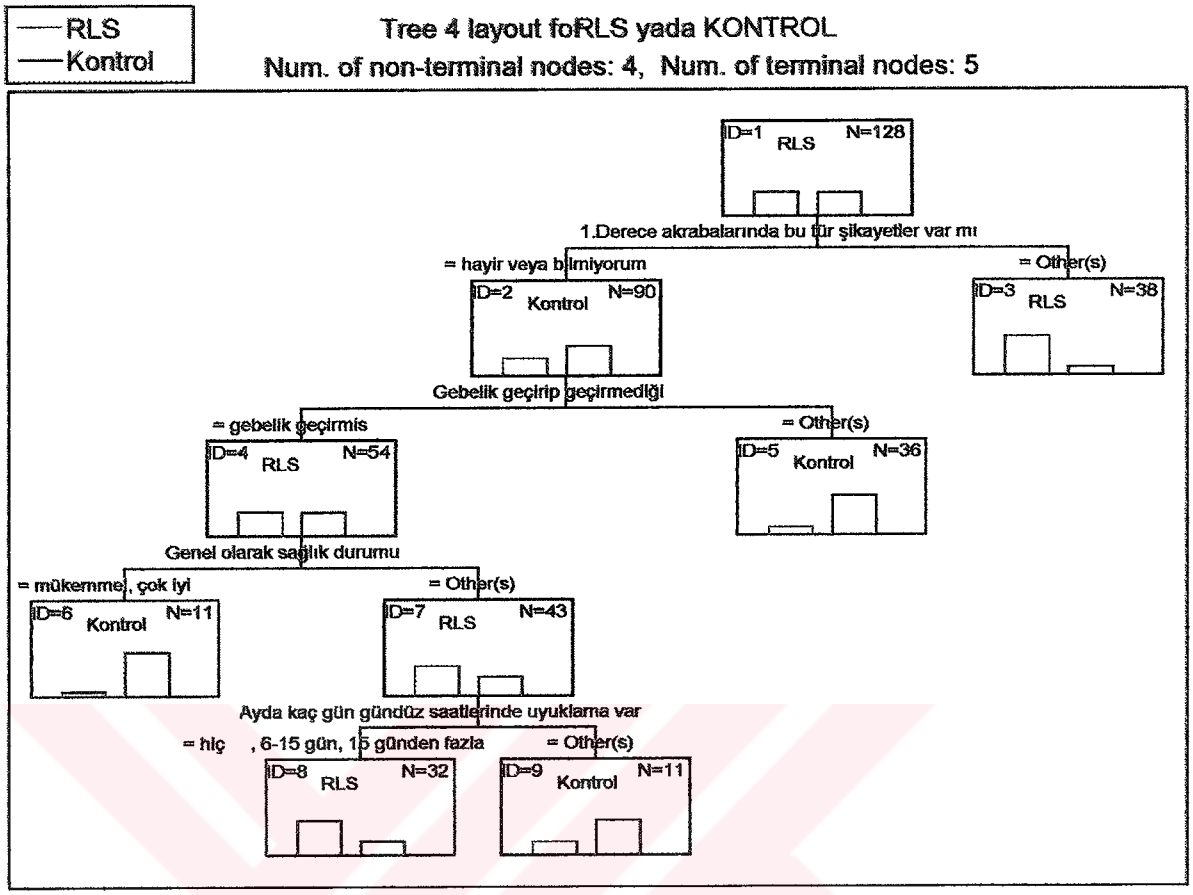
Doğru Sınıflama Oranı=1-0,156=0,844



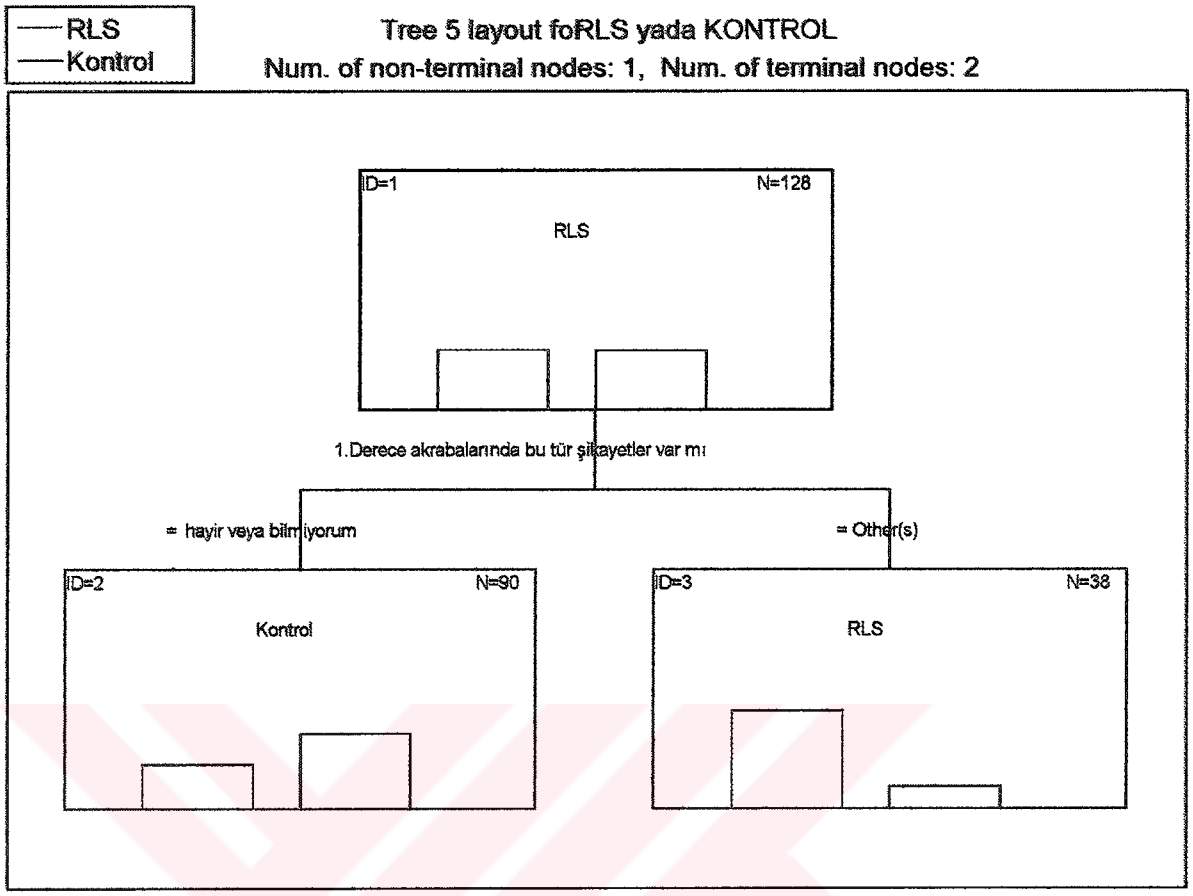
Şekil 6.2.5 : Analiz II için oluşturulan 2 nolu budanmış sınıflama ağaç diyagramı.



Şekil 6.2.6: Analiz II için oluşturulan 3 nolu budanmış sınıflama ağacına ait diyagram

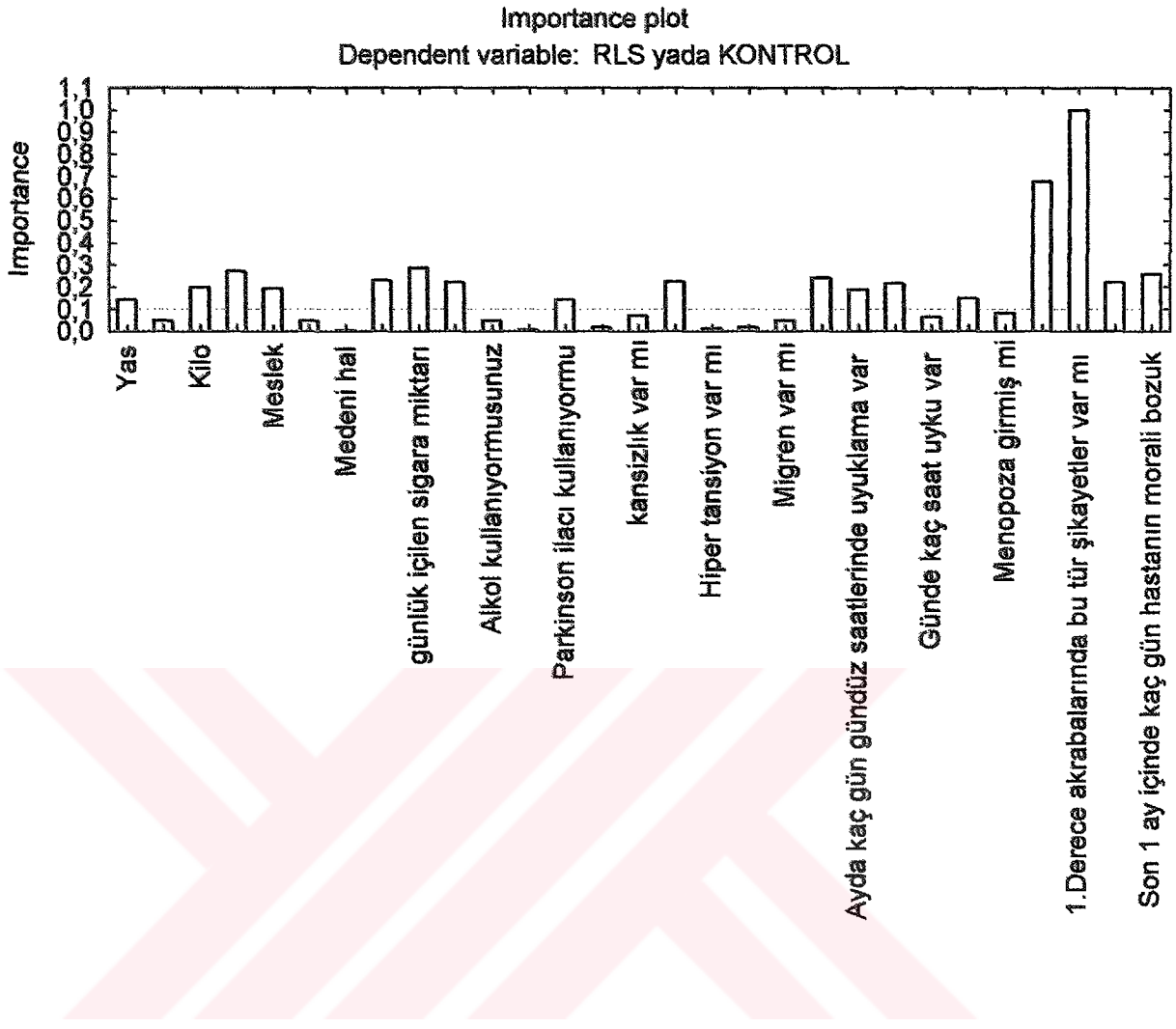


Şekil 6.2.7: Analiz II için oluşturulan 4 nolu budanmış sınıflama ağacına ait diyagram



Şekil 6.2.8: Analiz II için oluşturulan optimal sınıflama ağaç diyagramı

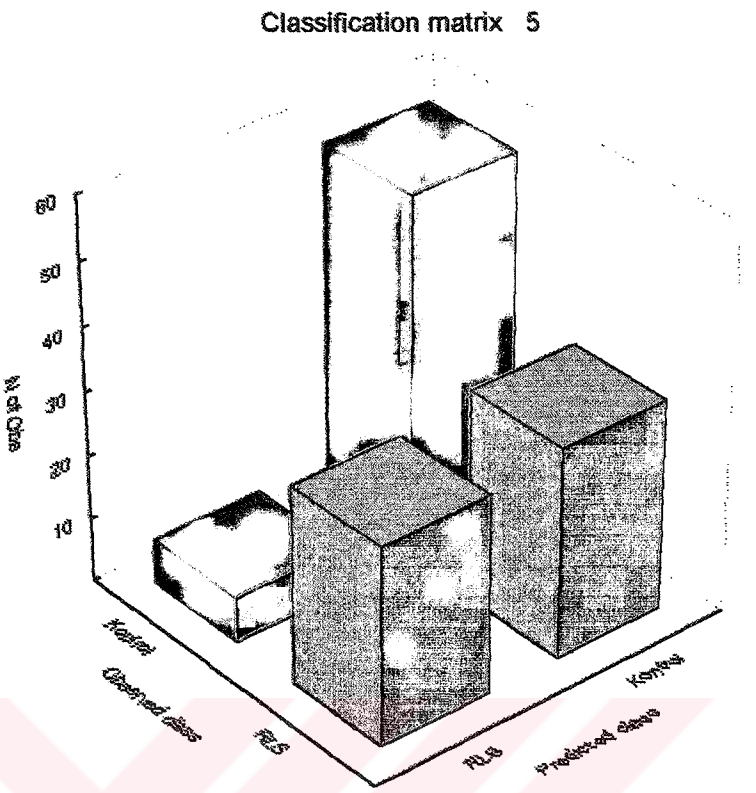
Şekil 6.2.4'de sunulan maximal sınıflama ağacı, 10 katlı çapraz geçerlilikle budanarak Şekil 6.2.8'de sunulan optimal sınıflama ağacı inşa edilmiştir.



Şekil 6.2.9: Analiz II için oluşturulan optimal sınıflama ağacı oluşumunda kullanılan bağımsız değişkenlerin sınıflamada önemlilik grafiği

Çizelge 6.2.6: Analiz II için oluşturulan optimal sınıflama ağacı oluşumunda kullanılan bağımsız değişkenlerin sınıflamada önemlilik oranları.

	Predictor importance 5 Responses: RLS yada I Options: Categorical re	
	Variable rank	Importance
Yas	14	0,142500
Öğrenim durumu	5	0,046752
Kilo	20	0,199146
Boy	27	0,271650
Meslek	19	0,191532
Evinizin deniz kiyisindan uzakligi	5	0,046752
Mederi hal	0	0,003972
Sigara iciyorsunuz	23	0,229167
günlük icilen sigara miktari	28	0,283739
Gecmisle sigara aliskanligi var mi	22	0,220718
Alkol kullaniyormusunuz	5	0,046752
Antidepresan ilaci kullaniyormusunuz	0	0,004613
Parkinson ilaci kullaniyormu	14	0,142500
Beyin yada omurilik ameliyati gecirmis mi	2	0,017062
kansizlik var mi	7	0,068247
Bobrek yetmezligi var mi	22	0,222656
Hiper tansiyon var mi	1	0,011995
Diyabet var mi	2	0,015833
Migren var mi	5	0,046752
Depresyon var mi	24	0,241362
Ayda kaç gün gündüz saatlerinde uyuklama var	19	0,186772
Ayda kaç gece uykudan uyanma var	21	0,213964
Günde kaç saat uyku var	6	0,063090
Kac gece rüya görüyor	15	0,146244
Menopoze girmis mi	8	0,079600
Gebelik geçirip geçirmediği	68	0,676424
1 Derece akrabalarında bu tür sikayetler var mı	100	1,000000
Genel olarak saglik durumu	22	0,220718
Son 1 ay içinde kaç gün hastanın moral bozuk	26	0,255523



Şekil 6.2.10: Analiz II için oluşturulan optimal sınıflama ağacına ait sınıflama bar grafiği

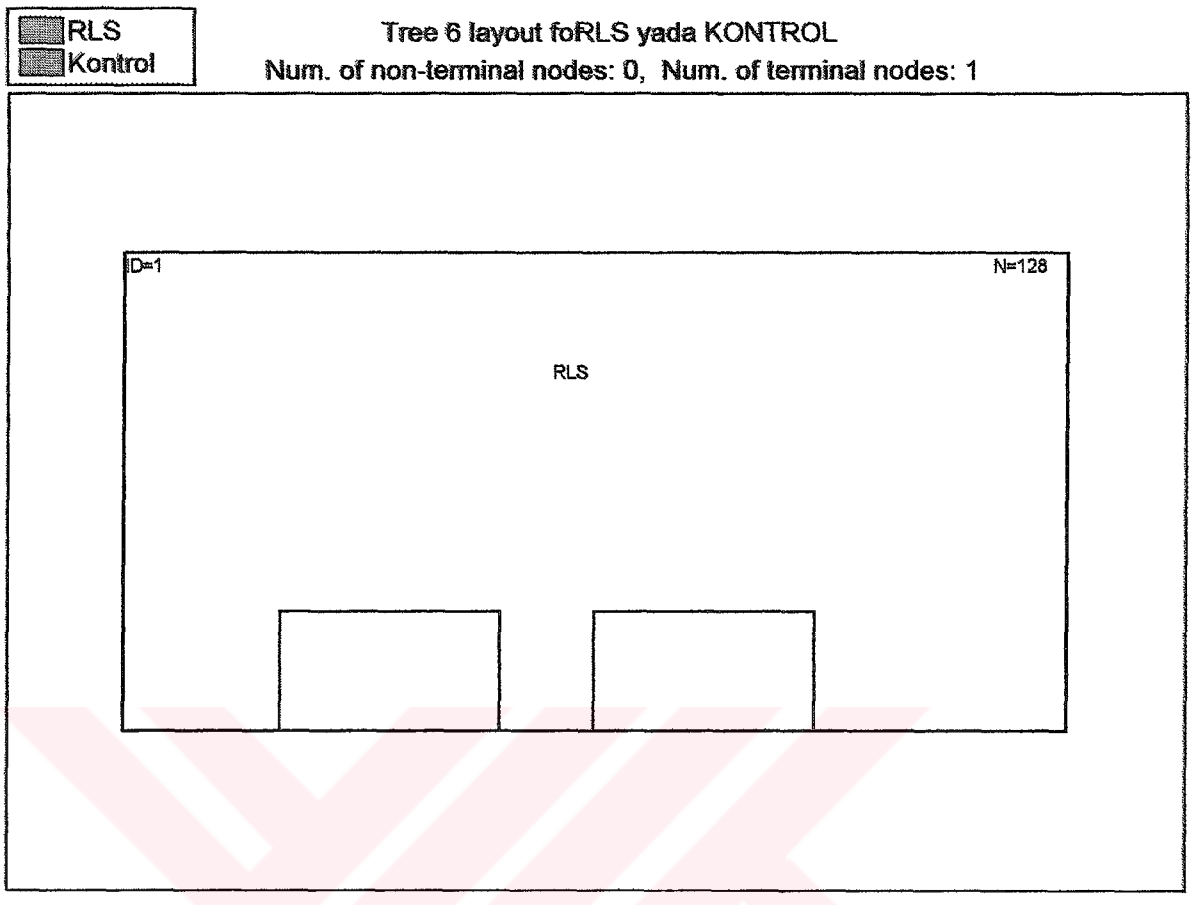
Çizelge 6.2.7: Analiz II için oluşturulan optimal sınıflama ağacına ait sınıflama matrisi.

	Geçek Sınıf		
Tahmin sınıfı	RLS	Kontrol	Toplam
RLS	31	7	38
Kontrol	33	57	90
Toplam	64	64	128

Şekil 6.2.10 ve Çizelge 6.2.7'den yararlanarak optimal ağacın hatalı sınıflama ve doğru sınıflama oranını aşağıdaki gibi hesaplayabiliriz.

Hatalı Sınıflama oranı=(33+7)/128=0,312

Doğru Sınıflama Oranı=1-0,312=0,688



Şekil 6.2.11: Analiz II için oluşturulan 6 nolu budanmış sınıflama ağaç diyagramı.

7. BULGULAR

Huzursuz Bacak Sendromu (Restless Legs Sendrome, RLS) hekimler tarafından da yaygın olarak atlanabilen ancak oldukça sık rastlanan ve en önemli uykusuzluk nedenlerinden biri olup, özelikle de bacaklarda istirahat sırasında veya yatarken ortaya çıkan nahoş duygular sonucunda bacakları sürekli hareket ettirme ihtiyacı duyma ve bu nedenle uykuya dalamama ile karakterize bir rahatsızlıktır (23).

RLS populasyonda % 5-10 oranında görülen ve ilerleyen yaşla birlikte görülme sıklığı artan bir nörolojik sendromdur (24).

Kadınlarda RLS hastalığının görülme sıklığı daha yüksektir. Hastalık her yaşta başlayabilmekle birlikte orta ve ileri yaşlarda belirgin olarak daha fazla görülmektedir. Hastalığın iki tipi vardır.

1. Tipte: RLS hastalığına neden olabilecek herhangi bir hastalık veya durum yoktur. Bu tipin büyük bir bölümünde ailesel özellik söz konusudur.

2.Tipte: Bu tipteki RLS hastalığının başlangıç yaşı 1.tipe göre daha erkendir ve bazı durumlar veya hastalıklarla ikincil olarak ortaya çıkar. Böbrek hastalıklarında (%15-20), gebelikte (%11), sinir hastalıklarında, bazı antidepressan ilaçları kullanımı sırasında, Parkinson hastalığı gibi nörolojik rahatsızlıklarda sıkça rastlanmaktadır (23).

Analiz I'de deneğin, 1.derece akrabalarında RLS hastalığının olup olmadığı(25), kilosu, günde kaç saat uyuduğu(26), ayda kaç gece uykudan uyandığı, ayda kaç gün gündüz saatlerinde uyukladığı, ayda kaç gün moralinin bozuk olduğu, genel olarak sağlık durumu, mesleği ve Parkinson ilacı kullanıp kullanmaması soruları RLS grubunu(hastalığını) ve Kontrol grubunu birbirinden ayıran önemli bağımsız değişkenler olarak tespit edilmiştir. Analiz II sonuçlarına bakıldığında kadın deneklerin, 1.derece akrabalarında RLS hastalığının olup olmadığı, gebelik geçirip geçirmediği(27), günde kaç saat uyuduğu, ayda kaç gece uykudan uyandığı, ayda kaç gün gündüz saatlerinde uyukladığı, ayda kaç gün moralinin bozuk olduğu, genel olarak sağlık durumu, yaşı(28), boyu, mesleği ve depresyona girip girmediği(29) soruları RLS ve Kontrol grubunu ayırmada önemli değişkenler olmuştur. Her iki sınıflama ağacı analizi de, bu uygulamada literatürle paralellik göstererek oldukça iyi sonuçlar vermiştir.

Eğer eldeki denek sayıları daha yüksek tutulursa analiz daha gerçekçi sonuçlar verebilir.

Analiz I'de oluşturulan sınıflama ağacı modeli ile gelecek bir veri seti için hatalı sınıflama oranı %19'dur. Kadınlar için uygulanan Analiz II'deki sınıflama ağacı modeli ile gelecek bir veri seti için hatalı sınıflama oranı %16'dır.

Hatalı sınıflama oranlarını görmek dışında, ağaç modeline bakarak RLS hastalığını başka hangi faktörlerin etkilediğini ve bu faktörlerin ayıraç değerlerini görmek (hangi sınırından sonra riskin artığını görmek) bir hekim için son derece önemlidir.

Ayrıca diğer sınıflama ve tahmin problemlerinde sorun yaratan bağımsız değişkenler arasındaki etkileşim, CART için bir sorun teşkil etmemekte aksine bu sorun modeli daha güvenilir kılmaktadır.

8. SONUÇ ve ÖNERİLER

1. Bu metot klasik sınıflama ve regresyon modellerine alternatif olarak kullanılabilir.
2. Özellikle tıbbi araştırmalarda, eldeki veri sayısı fazla olduğu durumlarda, bireylere tanı koymada kolay uygulanabilir ve yorumlanabilir bir tekniktir.
3. Modelle giren değişken çeşidi ne olursa olsun yöntemin kullanılmasında sakınca yoktur.
4. CART analiziyle oluşturulan sınıflama ağacına (modele) bakıldığında bağımlı değişkenleri hangi bağımsız değişkenlerin etkilediğini, bağımsız değişkenlerin cutt of değerlerini ve bağımsız değişkenler arasındaki etkileşimi görmek mümkündür. Bir hekime, daha önce etkisi yok yada olmadığını düşündüğü bir bağımsız değişkeni modelde görerek, tekrar düşünme yada üzerinde çalışma şansını verir.
5. Ülkemizde henüz bir tez (30) ve bir uygulama (9) yapılmış olan bu metot klasik istatistik tekniklerine yeni bir yaklaşım getirmiştir. Bilgisayar teknolojisindeki ilerleme ile bu metodun kullanılmasına paralellik göstermiştir.

KAYNAKLAR

1. **Marian M, Nestler I, Bernd H.** GIS-based regionalization of soil profiles with classification and regression trees. *J Plant Nutr. Soil.Sci,* **2002**;39-43.

2. **Yohannes Y, Hoddinott J.** Classification and regression trees:an ıntroduction. Erişim: http://www.ifpri.org/themes/mp18/techquid/tg03.pdf. Erişim tarihi:10.06.2003

3. **Speybroeck N, Berkvens D, Mfoukou-Ntsakala A, Aerts M, Hens N, Huylenbroeck G, Thys E.** Classification trees versus multinomial models in the analysis of urban farming systems in central Africa. *Agricultural Systems,* **2003;1-17**

4. **Bremner A P, Taplin R.** Modified classification and regression tree splitting criteria for data with interactions. *Australian Statistical Publishing Association*, **2002**;44(2):169-176.

5. **Lewis R.** An introduction to classification and regression tree(cart) analysis. Academic Emergency Medicine. California, **2000** :1-14.

6. **Fu C.** Combing loglinear model with classification and regression tree (CART): an application to birth data. *Computational Statistics&Data Analysis*, **2003**; 1-11.

7. **Chipman H A, George E I, McCulloch R E.** Hierarchical priors for bayesian CART Shrinkage. *Statistics and Computing,* **2000**; 10:17-24

8. **Bevilacqua M, Braglia M, Montanari R.** The classification and regression tree approach to pump failure rate analysis. *Reliability Engineering and System Safety*, **2003**;79:59-67.

9. **Rosa J, Veiga A, Medeiros M.** Tree-structured smooth transition regression models based on cart algorithm. Erişim:http://www.econ.puc-rio.br/PDF/td469.pdf. Erişim tarihi:20.06.2003 .

10. **Sha N.** Bolstering cart and bayesian variable selection methods for classification. Doctor, America, **2002.**

11. **Yohannes Y.** Classification and RegressionTrees, Cart™ . Erişim: http://www.ifpri.org/pubs/microcom/micro3.pdf Erişim tarihi:20.06.2003

12. **De'ath G, Fabricius K.** Classification and regression trees: a powerful yet simple technique for eqological data analysis. *Ecology*, **2000**; 81(11):3178-3192

13. **Akpınar H.** Veri tabanlarında bilgi keşfi süreci. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, **2000**;29(1):1-22.

14. **Magidson J.** SPSS, America,**1999**

15. **Haughton D, Oulabi S.** Direct marketing modeling with cart and chaid. *Journal of Direct Marketing,* **1997**; 4(11): 42-52

16. **Bai L.**Comparison of cart, fact and quest. Erişim:http://www.isye-gatech.edu. Erişim tarihi:24.09.2003

17. **Breiman L, Friedman J H, Olshen R A, Stone C J,** Classification and regression trees.Chapman&Hall,1993:32-104.

18. Salford Systems. Salford systems white paper series. Erişim:http://www.salford_systems.com/whitepaper.html Erişim tarihi: 27.09.2003

19. StatSoft. Classification and regression trees. Erişim:http://www.statsoft.com/textbook/scart.html. Erişim tarihi:18.03.2003

20. **Put R, Questier F, Coomans D, Massart D L, Heyden Y**. Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure-retention relationship studies. *Journal of Chromatography A,* **2003**;988: 261-276.

21. Statistica Inc. Statistica for Windows. Version 6.0,US: Statistica Inc., 2001.

22. Genç O. Erişim:www.milliyet.com.tr/content/saglık/sag011/sag17.html. Erişim tarihi:01.12.2003

23. Erişim: www.internationalhospital-com.tr Erişim tarihi:18.09.2003

24. Huzursuz bacak sendromu. Erişim:www.sinaps.org/klinik/hbs.asg. Erişim tarihi:03.10.2003

25. Erdal S. Huzursuz bacak sendromu.Erişim: www.genetikbilimi.com/tip/huzursuz.html. Erişim tarihi:01.12.2003

26. Kaynak H. Uykusuzluk. Erişim: www.ntvmsnbc.com/news. Erişim tarihi:27.10.2003

27. Genç B.Erişim:www.milliyet.com.tr/content/saglık/sag011/sag17/html Erişim tarihi:01.12.2003

28. Ardıç S. Erişim:www.saglık.tr.net/ruh_saglıgı_uyku_bozuklukları1.shtmı. Erişim tarihi:10.11.2003

29. **Özbel H, Ağargün Y M.** Yeni nesil antidepressan ilaçlar ve uyku üzerine etkileri.*Klinik Psikofarmakoloji Bülteni*, **2001**;11(4):1-20

30. **Gölbaşı G.** Sınıflama ve regresyon ağaçları ve bir uygulama. Yüksek lisans tezi, Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, **2000.**

EKLER

EK1: ANKET FORMU

Soru1: Adınız ve soyadınız?

Soru 2: Telefon numaranız?

Soru 3: Yaşınız?

Soru 4: Cinsiyetiniz?

a.) Kadın b.) Erkek

Soru 5: Mesleğiniz?

a.) Ev Hanımı b.) Öğrenci c.) Çiftçi d.) Devlet Memuru

e.) Emekli f.) Esnaf g.) İşsiz h.) Diğer

Soru 6: Yaşadığınız yer?

a.) İl b.) İlçe c.) Köy

Soru 7: Yaşadığınız yerin deniz kıyısına olan uzaklığı?

a.) 0-100m. b.) 101-500m c.) 501-1000m. d.) 1001-2000m.

Soru 8: Medeni haliniz?

a.) Evli b.) Bekar

Soru 9: Eğitim düzeyiniz?

Soru 10: Kilonuz?

Soru 11: Boyunuz(cm.) ?

Soru 12: Sigara içiyor musunuz ?

a.) Hiç İçmedim b.) Günde 10 Adetten Az

c.) Günde 10-19 Adet d.)Günde 1-2 Paket

e.) Günde 2 Paketten Fazla f.) Şimdi İçmiyor

Soru13: Geçmişte içmiş iseniz ?

13.a.) Günde kaç adet içtiniz?

13.b.) Kaç yıl boyunca içtiniz?

13.c.)Kaç yıldır içmiyorsunuz?

Soru14: Alkol kullanıyor musunuz?

a.) Kullanmıyorum

b.) Ayda 10 duble rakı veya 4 şişe şarap veya 10 şişe bira veya eş değerinden az

c.) Ayda 10 duble rakı veya 4 şişe şarap veya 10 şişe bira veya eş değerinden fazla

Soru15: Kaç yıldır kullanıyorsunuz?

Soru16: Kullandığınız ilaç var mı ?

a.)Hayır yok b.) Antidepressan c.) Antipsikotik

d.) Antihipertansif e.) Antidiyabetik f.) Doğum kontrol hapı

g.) Kullanıyor ama adını bilmiyor h.) Diğer:

Soru 17: Beyin yada omurilik yada bu bölgelerle ilgili başka hasatlık geçirdiniz mi?

a.) Hayır b.) Evet

Soru 18: Doktor tarafından saptanmış önemli bir rahatsızlığınız var mı?

A.) Hayır yok b.) Hıpertansiyon c.) Diyabet

D.) Kansızlık e.) Migren

F.) Böbrek hastalığı g.) Depresyon h.)Diğer:

Soru 19: Ayda ortalama kaç gün gündüz saatlerinde uyuklarsınız?

a.) 0 gün b.) 1-5 gün c.) 6-15 gün d.) 15 günden fazla

Soru20: Ayda ortalama kaç gece uyanırsınız?

A.) Hiç b.) 1-5 gün c.) 6-15 gün d.) 15 günden fazla

Soru 21: Günde ortalama kaç saat uyursunuz?

A.) 2 saatten az b.) 2-4 saat c.) 5-7 saat

D.) 8-10 saat e.) 10 saatten fazla

Soru 22: Ayda ortalama kaç gece rüya görürsünüz?

A.) Hiç b.) 1-5 gece c.) 6-15 gece d.) 15 geceden fazla

Soru 23: Menopoza girdiniz mi (kadınlar için) ?

A.) Evet b.) Hayır

Soru 24: Kaç yıldır menopozdasınız?

Soru 25: Kaç gebelik geçirdiniz?

Soru 26: Birinci derece akrabanızda daha çok geceleri ve hareketsizlikten ortaya çıkan ve bacaklarını hareket ettirme isteği uyandıran bacaklarda iğnelenme, ağrı, uyuşma gibi yakınmalar var mı ?

a.) Evet b.) Hayır yada bilmiyorum

Soru 27: Sağlığınız genel olarak nasıldır?

a.) Mükemmel b.) Çok iyi c.) İyi d.) Orta e.) Kötü

Soru 28: Son 1 ay içinde kaç gün moraliniz bozuktu?

a.) 0-10 gün b.) 11-20 gün c.) 21-30 gün

Soru 29: Kollarınızda da hareket etme isteği uyandıran iğnelenme, ağrı, uyuşma gibi geceleri artan şikayetleriniz olur ve hareket ettirmekle hafifler mi?

a.) Evet b.) Hayır c.) Bilmiyorum

Soru 30: bacaklarınızdaki bu şikayetleriniz kaç yaşında başladı?

Soru 31: Bacaklarınızdaki bu şikayetleriniz kaç yıldır var?

Soru 32: Bacaklarınızdaki bu şikayetleriniz son 1 ay içinde ortalama kaç dakika sürüyor?

Soru 33: Bacaklarınızdaki bu şikayetleriniz ilk ortaya çıktığında kaç dakika sürüyordu?

Soru 34: Bu şikayetleriniz giderek arttı mı, aynı mı kaldı, yoksa azaldı mı?

a.) Arttı b.) Aynı kaldı c.) Azaldı yada kayboldu

Soru 35: Bacaklarınızdaki bu şikayetiniz ayda kaç gece olur?

A.) 5 geceden az b.) 5-15 gece c.) 15 geceden az

Soru 36: Gebelikleriniz sırasında, daha çok geceleri ve hareketsizlikten ortaya çıkan ve bacaklarınızı hareket ettirme isteği uyandıran, bacaklarda iğnelenme, ağrı veya uyuşma gibi yakınmalar ortaya çıktı mı?

a.) Evet b.) Hayır c.) Varolan yakınmalar artmış

d.) Varolan yakınmalar azalmış

Soru 37: Böyle şikayetleriniz varsa gebeliğin hangi döneminde daha çok oldu?

a.) Olmamış b.) İlk 3 ay c.) İkinci 3 ay d.) Son 3 ay

Soru 38: Ailenizin toplam geliri:

Soru 39: Aile içindeki kişi sayısı:

Soru 40: İkiziniz var mı?

a.) Evet b.) Hayır

Soru 41: Varsa telofon numarası:

Soru 42: 18 yaşından büyük ve sağ olan birinci derece aile fertlerinin adları ve tlf numaraları:

Soru 43: Bu hastalıkla ilgili tedavi görmek ister misiniz?

a.) Evet b.) Hayır

ÖZGEÇMİŞ

1976 yılında Tarsus'ta doğdu. İlk, orta,lise öğrenimi Tarsus'ta tamamladı. 1998 yılında 19 Mayıs Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümünden mezun oldu. Aynı yıl Mersin Üniversitesi İktisadi ve İdari Bilimler Fakültesinde Öğretim Görevlisi olarak göreve başladı. 2000 yılında Mersin Üniversitesi Sağlık Bilimleri Enstitüsünde İngilizce hazırlığa başladı. 2001 yılında Mersin Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalında Yüksek lisansa başladı.

Halen Mersin Üniversitesi İktisadi ve İdari Bilimler Fakültesinde Öğretim Görevlisi olarak çalışmaktadır.