T.C. İSTANBUL TİCARET ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

SINIFLAMA VE REGRESYON AĞAÇLARI TEKNİĞİ İLE KALP HASTALIKLARINA ETKİ EDEN BAZI FAKTÖRLERİN BELİRLENMESİ

Onur KÖSE

Danışman

Doç. Dr. Özlem Deniz BAŞAR

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI
İSTANBUL - 2018

KABUL VE ONAY SAYFASI

Onur KÖSE tarafından hazırlanan **"Sınıflama ve Regresyon Ağaçları Tekniği İle Kalp Hastalıklarına Etki Eden Bazı Faktörlerin Belirlenmesi "** adlı tez çalışması 4./7./2018 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **İstatistik Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	**Doç. Dr. Özlem Deniz BAŞAR** **İstanbul Ticaret Üniversitesi**	.....
Jüri Üyesi	**Prof. Dr. Münevver TURANLI** **İstanbul Ticaret Üniversitesi**	.....
Jüri Üyesi	**Dr. Öğr. Üyesi Serpil KILIÇ DEPREN** **Yıldız Teknik Üniversitesi**	.....

Onay Tarihi: 23.07.2018

Prof. Dr. Necip ŞİMŞEK

Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

04.07.2018

Onur KÖSE

İÇİNDEKİLER

Sayfa

ÖZET

Yüksek Lisans Tezi

SINIFLAMA VE REGRESYON AĞAÇLARI TEKNİĞİ İLE KALP HASTALIKLARINA ETKİ EDEN BAZI FAKTÖRLERİN BELİRLENMESİ

Onur KÖSE

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Doç. Dr. Özlem Deniz BAŞAR

2018, 69 sayfa

Karar ağaçlarının algoritmalarından olan CHAID ve CART teknikleri; kullanımının ekonomik olması ve hızlı sonuç vermesi nedeniyle, veri işlemede sıkça kullanılan tekniklerdendir. CART ve CHAID algoritmalarının en önemli özellikleri sürekli ve kategorik verileri aynı anda modele dahil edebilmesi, bağımlı değişkenler üzerinde etkili olan bağımsız değişkenleri bir ağaç diyagramı üzerinde kolayca gösterilip özetlenebilmesidir.

Bu çalışmada; University of California bünyesinde veri setlerini barındıran bir platformdan alınan kalp hastalığına etki eden 38 adet faktör kullanılmıştır. Bu faktörlerin değerlendirilmesi; Sınıflama ve Regresyon Ağaçı (CART) ve Otomatik Ki-Kare Etkileşim Belirleme (CHAID) algoritmaları kullanılarak oluşturulmuştur ve çıkan sonuçlar birbirleriyle karşılaşırılarak yorumlanmıştır.

Anahtar kelimeler: Classification and Regression Tree (CART) ve Chi-Squared Automatic Interaction Detector (CHAID), Kalp hastalığı.

ABSTRACT

M.Sc.Thesis

DETERMINATION OF SOME FACTORS INFLUENCING HEART DISEASES BY USING CLASSIFICATION AND REGRESSION TREES TECHNIQUE

Onur KÖSE

Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Statistics

Supervisor: Doç. Dr. Özlem Deniz BAŞAR

2018, 69 pages

CHAID and CART techniques which are the algorithms of desicion trees and are frequently used techniques in terms of getting quick result and being economic. The most important features of CART and CHAID algorithms that; incorporating the categorized and continuous data at the same time, summarising and showing the independent variables on a diagram tree which are effect on dependent variables.

In this study, 38 factors related to heart diseases are studied and data were taken from a platform which has data sets in University of California. These factors were evaluated by Classification and Regression Tree (CART) and Chi-Squared Automatic Interaction Detector (CHAID) algorithms and the results were interpreted by comparing to one another

Keywords: Classification and Regression Tree (CART) and Chi-Squared Automatic Interaction Detector (CHAID), heart disease.

TEŞEKKÜR

İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü'nde tamamlamış olduğum "Sınıflama ve Regresyon Ağaçları" başlıklı yüksek lisans tezimin hazırlık sürecinde kıymetli zamanını benden esirgemeyen, katkılarıyla beni yönlendiren, yardımını benden esirgemeyen tez danışmanım Sayın Doç. Dr. Özlem Deniz Başar'a en içten teşekkürlerimi sunarım.

Yüksek lisans öğrenimim boyunca akademik bilgi ve birikimleriyle yanımda olup, benden desteğini esirgemeyen hocalarım Sayın Prof. Dr. Münevver TURANLI ve Sayın Prof. Dr. Ünal Halit ÖZDEN 'e sonsuz teşekkürlerimi sunarım.

Yüksek lisans tezimin yazımı esnasında yaşamış olduğum sıkıntılı ve yoğun süreçte beni yalnız bırakmayan aileme sonsuz sevgi ve saygılarımı sunarım.

Onur KÖSE

İSTANBUL, 2018

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER

Sayfa

SİMGE VE KISALTMALAR

AID	Automatic Interaction Detector (Otomatik Etkileşim Belirleyicisi)
ANOVA	Analysis Of Variance (Varyans Analizi)
ASD	Atriyal Septal Defekt
CART	Classification and Regression Trees (Sınıflama ve Regresyon Ağaçları Tekniği)
CHAID	Chi_square Automatic Interaction Detection (Otomatik Ki-Kare Etkileşim Belirleme)
CP	Clark&Pregibon
C4.5	C4.5 Tree
FACT	Fast and Accurate Classification Tree (Hızlı ve Doğru Sınıflandırma Ağacı)
ID3	Iterative Dichotomiser 3
LAD	Least Absolute Deviation (En küçük mutlak sapma)
LS	Least Squares (En küçük kareler)
QUEST	Quick, Unbiased, Efficient Statistical Trees (Hızlı, Tarafsız, Verimli İstatistik Ağaçları)
THAID	Theta Automatic Interaction Detection (Theta Otomatik Etkileşim Algılama)

1. GİRİŞ

Sınıflama ve Regresyon Ağaçları Tekniği (CART) (Classification and Regression Trees) sürekli veya kategorik bağımlı değişkenlerin sayısal karşılıklarını öngörebilmek ve çözümleyebilmek amacıyla oluşturulmuş, dağılımdan bağımsız istatistiksel yöntemlerdendir. CART, kategorik bağımlı değişkenlerde sınıflama ağacı şeklinde, sürekli bağımlı değişkenlerde ise regresyon ağacı şeklinde adlandırılmaktadır (Fu, 2004). CART modelleri, yinelenebilen tahmin ediciler evreninin eş tekrarlı iki alt sınıfa ayrıştırılması temeline dayanan karar ağaçları oluştururlar (Chipman & McCulloch, 2000).

CART yöntemi bütün başlangıç veri setini içinde barındıran kök düğümden başlayarak her düğümü iki küçük düğüme böler ve ikili ağaçlar oluşturur. CART algoritmasının oluşum mekanizması, düğümdeki homejenliği en üst düzeye çıkarabilmek için çalışır. Bir düğümün içinde homojen bir alt kümenin bulunması düğümün safsızlığının bir göstergesi olur. Yani bir uç düğüm her durumda bağımlı değişken için aynı değere sahipse bölünme yapmaz, çünkü artık saf bir düğümdür (Antipov vd., 2010).

CART ve CHAID algoritmaları karar ağacı algoritmalarıdır. Veri seti çok karmaşık olsa bile bağımlı değişkeni etkileyen bağımsız değişkenleri ve bu değişkenleri önem sırasına göre ağaç şeklinde ortaya koyar.

Bu çalışmanın amacı; CART ve CHAID algoritmaları ile yaşları 37 ila 77 arasında değişen 248 hasta üzerinde kalp hastalığına etki eden 38 adet faktörün etkileri incelenmiştir. Çalışmada kalp hastalığına etki eden tansiyon, sigara kullanımı, diyabet ve ilaç kullanımı gibi faktörlerin aralarındaki farklılıklar incelenmiş ve sonuçlar karşılaştırılmalı olarak yorumlanmıştır.

Çalışmanın birinci bölümünde, karar ağaçları yönteminden kısaca bahsedilmiş ve karar ağaçlarının uygulama alanları belirtilmiştir. Daha sonra karar ağaçlarında kullanılan sınıflama yöntemlerinden (algoritmaları) kısaca bilgi verilip bunlardan çalışmada kullanılan CART ve CHAID algoritmaları anlatılmıştır.
Çalışmanın ikinci ve üçüncü bölümünde, sınıflama ve regresyon ağaçları tekniğinin genel olarak anlatımı, tarihi gelişimi, günümüzdeki kullanım alanlarından genel olarak anlatılmış ve optimum ağacın belirlenmesi süreci anlatılmıştır. Burada sınıflama ve

regresyon ağaçlarının teorik çerçevesi içerisinde sınıflama ağaçlarında en önemli ayırma kriterleri olan gini ve twoing bahsedilmiştir. Regresyon ağaçlarında ise hata ölçüm ve tahminleri ve ağaç oluşumundan bahsedilmiştir.
Çalışmanın dördüncü bölümünde, Otomatik Ki-Kare Etkileşim Belirleme (CHAID) (Chi-Squared Automatic Interaction Detector) algoritmasından bahsedilip optimum ağacın belirlenmesi süreci anlatılmıştır.

Ayrıca, kalp hastalığına etki eden faktörler hem sınıflama ve regresyon ağacı (CART) ve CHAID algoritmaları kullanılarak belirlenmeye çalışılmıştır. Bu çalışmada; University of California bünyesinde veri setlerini barındıran bir platformdan alınan ve kalp hastalığına etki eden faktörlerin 248 kişiye uygulanan veri seti kullanılmıştır. İlk olarak kalp rahatsızlıkları üzerine yapılmış diğer çalışmalardan bahsedilmiştir. Daha sonra uygulamaya geçilmiştir. Uygulamanın birinci aşamasında kalp rahatsızlığına etki eden faktörler CART algoritmaları kullanılarak ağaç modellerinin oluşturulması amaçlanmıştır. Bu amaç doğrultusunda sırayla, model verisinin %70, %50 ve %30'luk kısmı analize dahil edilerek üç faklı oranla oluşturulan modellerde, farklı düğüm sayıları ile uygun ağaç modelleri oluşturulmuş ve karşılaştırmalı olarak yorumlanmıştır. İkinci aşamasında kalp rahatsızlığına etki eden faktörler CHAID algoritmaları kullanılarak ağaç modellerinin oluşturulması amaçlanmıştır. Bu amaç doğrultusunda sırayla, model verisinin %70, %50 ve %30'luk kısmı analize dahil edilerek üç faklı oranla oluşturulan modellerde, farklı düğüm sayıları ile uygun ağaç modelleri oluşturulmuş ve karşılaştırmalı olarak yorumlanmıştır.

2. LİTERATÜR ÖZETİ

Yasemin Bahar YÜCEL tarafından 2017 yılında yapılan "yaşam memnuniyetini etkileyen faktörlerin sınıflama ve regresyon ağacı ile belirlenmesi" isimli çalışmada; TÜİK tarafından gerçekleştirilen yaşam memnuniyeti anketi verileri doğrultusunda, kişilerin mutluluk düzeylerini etkileyen faktörler sınıflama ve regresyon ağaçları ile belirlenmeye çalışılmıştır. Sınıflama ve regresyon ağaçlarında CART ve CHAID algoritmaları kullanılmıştır. Ayrıca sınıflama ve regresyon ağaçları, model verisinin %30, %50 ve %70'lik kısmı dahil edilerek oluşturulmuş ve oluşturulan ağaçlar karşılaştırmalı olarak yorumlanmıştır.

Çalışmanın sonucunda oluşturulan sınıflama ağaçlarında; Kişilerin mutluluk düzeylerini etkileyen en önemli faktörün gelecekten umutlu olma olduğu sonucuna ulaşılmıştır. Buna göre; umut seviyesi düştükçe mutluluk oranları da azalmakta olduğu sonucuna ulaşılmıştır. Geleceğinden umutlu olan insanlarda mutlu olanların oranın, geleceğinden umutlu olmayanlarda orta düzey veya mutsuzların oranından daha yüksel olduğu saptanmıştır.

2015 yılında Yeliz SEVİMLİ SAİTOĞLU tarafından yapılan "sınıflama ve regreyon ağaçları" isimli çalışmada ise CART ve MARS yöntemleri, Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde kullanılmış ve her iki yöntemin sonuçları karşılaştırılarak hangi yöntemin daha doğru bir sınıflama yapacağı farklı büyüklükteki başlangıç ve test verisi kullanarak incelenmiştir. İnceleme sonucunda uygun olan başlangıç ve test verisi büyüklüğüne göre farklı büyüklükteki örnek sayıları ile sadece CART kullanılarak modelleme yapılmış ve en başarılı sınıflama modeli oluşturulmaya çalışılmıştır. Yapılan uygulamalar sonucunda, veri setinin yaklaşık %70'inin başlangıç verisi, geri kalan %30'unun da test verisi olarak alınması, en uygun başlangıç ve test verisi büyüklüğü olarak tespit edilmiştir. İlk aşamada bağımlı değişken olan parti tercihinin kategori düzeyi iki olarak ele alınmıştır. Bu aşamada, hem başlangıç hem test verisi için genel doğru sınıflama oranlarında CART yönteminin sonuçlarının, duyarlılık ve özgüllük hesabında ise; MARS yönteminin sonuçlarının nispeten daha yüksek olduğu görülmüştür. İkinci aşamada ise, parti tercihi değişkeninin kategori düzeyi beş düzey olarak ele alınmış ve farklı büyüklükteki örnek sayıları ile bu kez sadece CART ile modelleme yapılmıştır.

Buradan ortaya çıkan sonuç ise, örneklem büyüklüğü arttıkça genel olarak modelin hem başlangıç, hem de test verisinin genel doğru sınıflama oranının arttığı; ayırım gücünün ise azalıp artan bir trend gösterdiği yönünde olmuştur.

Bayram YILDIZ tarafından 2002 yılında "Menemen ilçesinde 35-64 yaş grubu koroner kalp hastalıkları risk faktörleri sıklığının araştırılması" isimli çalışmasında; gelişmekte olan ülkelerde erkek ve kadınlarda mortalite ve mobiditesi giderek artan bir halk sağlığı sorunu olan Koroner Kalp Hastalıkları (KKH) hakkında Menemen bölgesinde 35-64 yaş grubundaki erkek ve kadınlarda KKH risk faktörleri değerlendirilmiştir. Hipertansiyon ve diyabetes mllitus prevalansında yaşla birlikte yükselme, sigara içme prevalansında ise yaşla birlikte düşme olduğu saptanmıştır. Kadınlarda; yaş ve be-boy oranı ile; erkeklerde ise soy geçmişte hipertansiyon öyküsü ile hipertansiyon arasında pozitif ilişki saptanmıştır.

2015 yılında Betül ŞİŞMAN tarafında yapılan "çocuklarda görülen kalp hastalıklarında epidemiyolojik risk faktörlerinin belirlenmesi" isimli çalışmada ise; Dokuz Eylül Üniversitesi Tıp Fakültesi (DEUTF) Çocuk Kardiyoloji Polikliniğine başvuran hastalarda kalp hastalıklarına neden olan epidemiyolojik risk faktörlerinin belirlenmesi için betimsel istatistikler hesaplanmıştır. Elde edilen veriler ki kare analizi ile incelenmiş, doğumsal kalp hastalığı oluşmasını etkileyen maternal ve ailesel nedenler ortaya konulmuştur. Belirlenmişi risk faktörlerine göre hastanın cinsiyeti, akraba evliliği, annenin gebelikte geçirdiği tiroid ve şeker hastalığı ile annenin sigara kullanımı arasında anlamlı bağımlılık ilişkisi bulunmuştur.

3.KARAR AĞAÇLARI VE TEMEL KAVRAMLAR

Karar ağaçları ilk olarak; 1973 yılında Bierman ve Friedman tarafından geliştirilmiş olup değişkenlerin parçalanarak bir ağaç oluşturulması prensibine dayanmaktadır (Ulusoy, 2013).

Veri madenciliğinde karar ağaçları, kurulmasının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile basitçe entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile karar ağacı tekniği sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir (Safavian ve Landgrebe, 1991).
Karar ağaçları, sınıflandırma amacıyla veri madenciliğinde en çok kullanılan tahmin edici bir tekniktir. Genellikle sınıflandırma, kümeleme, tahmin modellerinde ve sorunla ilgili araştırma alanını alt gruplara ayırmak için kullanılmaktadır (Quinlan, 1986).

Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir. Karar ağaçları veri oluşturulduktan sonra ağaç kökten yaprağa doğru inilerek kurallar (if-then) yazılabilir. Karar ağaçlarında kök ve her düğüm bir soruyla etiketlenir. Düğümlerden ayrılan dallar ise ilgili sorunun olası yanıtlarını belirtir. Her dal düğümü de söz konusu sorunun çözümüne yönelik bir tahmini temsil eder. Kök düğüm olarak da adlandırılan ilk eleman en yüksek karar düğümüdür, kullanılan algoritmaya bağlı olarak her düğüm iki veya daha fazla dala sahip olur. İki dala sahip olan karar ağaçları ikili ağaç, daha fazla dala sahip olanlar ise çok yollu ağaç olarak adlandırılır. Her dal bir başka karar düğümüyle, ya da ağacın sonuyla yani yaprak düğümle sonlanır. Karar düğümlerinde gerçekleştirilen her bölünmede oluşturulan gruplar arasındaki mesafenin maksimum olması bir başka değişle elde edilen grupların mümkün olduğu kadar saf olması istenir(Arslan, 2008).

Karar ağacı-karar cetveli, cetvel tekniğinin kompleks prosedürlerin, bilgisayar ve programlamada araç olarak uygun olduğunu tanıdıkları zaman ortaya çıkmıştır. Kavram, karar ağacını veri cetvellerinden ayırt etmek için kullanılmıştır. Uzun zamandır kullanılan bir tekniğin yeni adı, böylece yerini almıştır (Aydın, 1992).
Verilerin sınıflandırılmasında en çok kullanılan yöntemlerden biri karar ağaçları oluşturma yöntemidir. Bu yöntem, sınıfı belli olan verilerin hangi sınıfa dahil

olacağını, bilgi kazancı en iyi olan düğümden başlayarak oluşturmaya çalışır (Sezer, 2008).

Ağaç ile ilgili temel özellikleri;

• Verinin herhangi bir kayba yol açmadan her bir dala ayrılabilmesi,

• Modelin nasıl yapılandırıldığının çok kolay anlaşılması (Sinir ağları veya standart istatistik modellerinin tersine),

• Oluşturulan modelin kolayca kullanılması, ayrıca bazı sezgilerin de modelde yapılandırılmasının mümkün olması,

• Karar ağaçlarının iç içe geçmiş eğer / doğruysa (if / then) kurallarının dizisi olması, görsel olması nedeniyle oldukça kolay anlaşılması ve aynı zamanda kolaylıkla SQL sorgusuna dönüştürülebilir olması,

• Değişken tiplerine göre, farklı yöntemler kullanılabilmesi şeklinde sıralanmaktadır (Koyuncugil A.S., 2007).

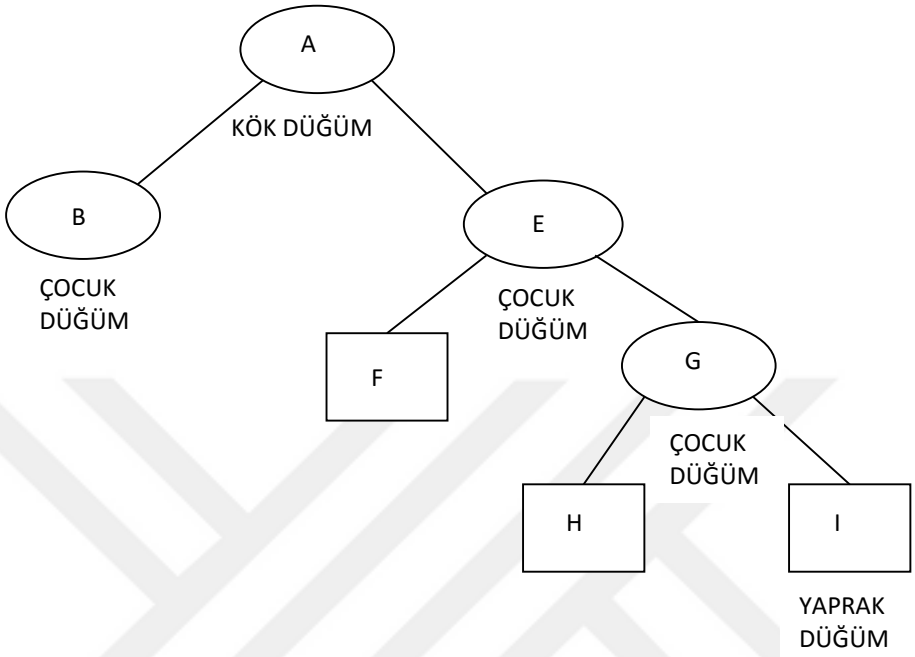
3.1. Karar Ağacı Yapısı

Karar ağacı; kök düğüm, dallar ve yapraklardan oluşur. Bu tipik bir ağacı andıran yapıda; karar düğümleri, yapılacak testi belirtir. Buradaki amaç; ağacın veri kaybetmeden dallara ayrılmasıdır. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir. Bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağaca ait her dal; sınıflama işlemini tamamlamaya yönelik hareket eder. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa o dalın ucunda bir karar düğümü oluşur ki buna yaprak düğüm denir. Bu yaprak düğüm, veri üzerinde belirlenmek istenen sınıflardan birini ifade eder (Özekes, 2002).

Karar ağacının yapısı bir şema yapısındadır. Bu şemada her değişken bir düğüm tarafından ifade edilir. Ağaç yapısı kısaca kök, dallar ve yapraklardan oluşur. En üst yapı kök, en son yapı yaprak ve bunların arasında kalan yapılar ise dal olarak nitelendirilir. Karar ağaçlarının oluşturulmasında en önemli kısım; hangi değişkenin ilk düğüm (kök düğüm) olacağının belirlenmesidir. İlk düğüm; çeşitli kriterler kullanılarak belirlenir (Atılgan ES., 2011).

3.2. Ağaç Modeli

Şekil 3.1 'de ağaç yapılı modele ait bir örnek gösterilmektedir.



Şekil 3.1: Ağaç yapılı model gösterimi

Kök Düğüm: Ağacın başlangıç düğümüdür

Çocuk Düğüm: Bir düğüme doğrudan bağlı olan düğümlerdir.

Yaprak Düğüm: Ağacın en altında bulunan düğümdür.

3.3. Karar Ağaçlarında Kullanılan Sınıflama Yöntemleri

CART haricinde pek çok sınıflandırma yöntemi mevcuttur ve bu yöntemler iki sınıf halinde incelenebilir.

Birinci sınıf içerisinde, Theta Otomatik Etkileşim Algılama (THAID), Hızlı, Tarafsız, Verimli İstatistik Ağaçları (QUEST), Otomatik Ki-Kare Etkileşim Belirleme (CHAID), Hızlı ve Doğru Sınıflandırma Ağacı (FACT), Otomatik Etkileşim Algılama (AID) yöntemleri bulunmaktadır.

İkinci Sınıf içerisinde, Lojistik Regresyon, Yapay Sinir Ağları, Disktiminant Analizi ve Probit Modeller yöntemleri bulunmaktadır.

Birinci sınıfta yer alan yöntemler sınıflandırma ağaçları temeline dayanmaktadır. Sonquist ve Morgan, 1970'li senelerde Michigan Üniversitesinde, iktisadi ve toplumsal olayları, daha doğru biçimde değerlendirebilmek amacıyla istatistik yöntemlerinin haricinde yeni çözümleme yöntemleri oluşturabilmeye yönelik çalışmaları sonucunda, karar ağacı temeline dayanan ilk yazılım ve algoritma olan AID (Automatic Interaction Detector)'i geliştirmişlerdir. Otomatik Etkileşim Algılama yöntemi en doğru ve en sağlam tahmini ortaya koyabilmek için bağımsız ve bağımlı değişkenlerin kendi aralarındaki tüm etkileşimlerin analiz edilmesini esas alır. Karar ağacı yönteminin, rahatlıkla yorumlanabilmesi nedeniyle Automatic Interaction Detector yazılımına, ilk ortaya çıktığında veri analizcileri ve istatistikçiler oldukça ilgi göstermişlerdir (Akpınar, 2000).

G.V Kass'ın 1980 senesinde geliştirdiği CHAID yönteminde, bağımlı değişkenler üzerinde en çok etkisi olan bağımlı değişkenlere ağırlık verilerek populasyon sınıflara ve daha alt sınıflara ayrılmaktadır (Magidson, 1993). Yönteme, değişkenin bağımsız ya da bağımlı olması fark etmeksizin, kategorik ise başvurulabilir (Haughton & Oulabi, 1997).

1988 senesinde Vanichestakul ve Loh, Fast and Accurate Classification Tree yöntemini; 1997 senesinde ise Shin ve Loh, Quick, Unbiased, Efficient Statistical Trees yöntemini geliştirmişlerdir. QUEST yönteminde kategorik değişkenler üzerinde χ^2 testine, sürekli değişkenler üzerinde ise ANOVA F istatistiğine başvurur. FACT ise bütün değişken çeşitleri üzerinde ANOVA F istatistiğine başvurur (Lewis R. j., 2000).

QUEST ve CART yöntemlerine kıyasla FACT yönteminde, kategorik değişkenler söz konusu olduğunda ön yargılar daha fazla bulunur. CART yöntemine kıyasla QUEST ve FACT yöntemlerinde kategorik değişkenler açısından daha hızlı netice elde edilir. Öte yandan ikiden çok sınıf sayısı olan bağımlı değişkenlerde CART yöntemi ile daha hızlı netice elde edilmektedir (lewis R. J., 2000).

4. SINIFLAMA VE REGRESYON AĞAÇLARI

Sınıflama ve Regresyon Ağaçları Tekniği (CART) (Classification and Regression Trees) sürekli veya kategorik bağımlı değişkenlerin sayısal karşılıklarını öngörebilmek ve çözümleyebilmek amacıyla oluşturulmuş, dağılımdan bağımsız istatistiksel yöntemlerdendir. CART, kategorik bağımlı değişkenlerde sınıflama ağacı şeklinde, sürekli bağımlı değişkenlerde ise regresyon ağacı şeklinde adlandırılmaktadır (Fu, 2004). CART modelleri, yinelenebilen tahmin ediciler evreninin eş tekrarlı iki alt sınıfa ayrıştırılması temeline dayanan karar ağaçları oluştururlar (Chipman & McCulloch, 2000). Karar noktalarına ulaşıncaya dek, iki alt sınıfa ayırma işlemi sürdürülür.

CART çözümlemesi bir makine öğrenme (machine learning) yöntemidir. CART çözümlemesi, alışılagelmiş veri çözümleme yöntemlerinden oldukça farklı bir ağaç kurulması yöntemidir (Lewis, 2000). CART, bir aile düğümü oluşturulması ile başlayarak, iki alt gruba ayrışarak diziler halinde ilerleyen, alt gruplara ayrışmanın uç düğümlere dek sürdüğü bir yöntemdir (Bevilacqua, Braglia, & Montanari, 2003).

4.1. CART'ın Tarihi Gelişimi

1963 senesinde Sonquist ve Morgan, CART yöntem bilimini oluşturan ilk adımları atmışlardır (Da Rosa, Veiga, & Medeiros, 2008). Sonquist ve Morgan'ın görüşlerinden günümüze dek oluşturulan tahmin yöntemleri ve bu yöntemlerin gelişim aşamaları aşağıda sunulmuştur.

Fu ve Henrichon (1969), Michalopoulos ve Meisel ikili karar ağaçlarına [öğrenme modelindeki (learning sample) grupların yinelemeli şekilde deney bazlı tekniklerle elde edilmesi] ilişkin tanımı literatüre kazandırmıştır. Stone ve Brieman (1978) geliştirmiş oldukları "budama yöntemi" sayesinde en elverişli ağacı tespit etmek amacıyla asgari karmaşa maliyeti (minimal cost complexity) tekniğini oluşturmuştur. Olshen ve Gordon (1980) Oklit Gözlem Uzayının kendi içinde bölünmesi temeline dayanan karar ilkelerini ortaya atmıştır. Washbrook, Stone ve Mabbet (1980) sınıflama ağaçlarında budama yönteminin uygulanması ve en elverişli ağacın tespiti hususunda çapraz geçerlilik tekniğinin uygulanmasını teklif etmiştir (Sha, 2002).

Stone, Friedman, Breiman ve Olshen 1984 senesinde "Classification and Regression Trees" adlı kitabı yazarak, CART'ın faydalı ve güvenilir bir çözümleme olarak kabul edilmesini sağlamışlardır.

4.2. Sınıflama ve Regresyon Ağaçları Tekniğinin Kullanım Alanları

CART, Sağlık Bilimleri alanında geçtiğimiz yirmi sene içinde önemli bir ilerleme kaydetmiştir. CART, çoğunlukla tıp alanında teşhis ve öngörü amacıyla, karar teorisinde ve botanik alanında uygulanmaktadır. Bununla birlikte iktisadi açıdan risk sınıfındaki işletmelerin kategorize edilebilmesi amacıyla Kao, Altman ve Frydman'ın ekonomi biliminde 1985 senesinde yapmış olduğu çalışmalar ve Patell, Marais ve Wolfson'ın ticari borçların kategorize edilebilmesi amacıyla yine 1985 senesinde yapmış olduğu çalışmalar büyük önem taşımaktadır (Yohannes & Hoddinott, 1999).

Milletlerarası Gıda Politikaları Araştırma Enstitüsü, yerel ve hane halkı seviyesinde yoksunluğun tespiti amacıyla Sınıflama ve Regresyon Ağaçları Tekniğine başvurmuştur (Yohannes & Hoddinott, 1999). Staub 1992 senesinde, Baker 1999 senesinde ve Rejwan 1999 senesinde ekolojik donelerin analizi amacıyla Sınıflama ve Regresyon Ağaçları Tekniğinden yararlanmışlardır (De'ath & Fabricius, 2000).

4.3. Sınıflama ve Regresyon Ağaçları Tekniğinin Avantajları

CART'ın tercih edilmesinde ve faydalı ve güvenilir bir model olarak kabul görmesindeki başlıca sebepler şu şekilde sıralanabilir:

- Sınıflama ve Regresyon Ağaçları Tekniği, istatistiksel dağılımlardan bağımsız bir yöntemdir.
- Modelde; değişkenlerin sıralı, kategorik, sürekli veya karma olması gibi nitelikleri hususunda hiçbir varsayım bulunmaz (Yohannes & Hoddinott, 1999).
- Bir varsayıma ihtiyaç duyulmadığı için değişkenler üzerinde karekök, logaritma gibi işlemlerde bulunulması için de bir gereklilik yoktur.
- Bağımsız ve bağımlı değişkenlerin bağlantısı görsel olarak ortaya koyulduğu için ağaç formundaki model neticeleri, istatistik alanına çok hakim olmayan kişiler tarafından dahi rahatlıkla yorumlanabilir.
- CART, bağımlı değişken tanımlandığında, karşılaşılabilecek tüm bağımsız değişkenleri ve bu değişkenlerin bütün varyasyonlarını modele ekleyerek elde

edilebilecek en yerinde sınıflandırmayı ortaya koyar. Değişkenler ile birlikte bunların varyasyonlarını da incelendiği için verilerin daha geniş bir perspektif ile yorumlanmasını sağlar ve esneklik sunar.

- CART en komplike veri setleri üzerinde dahi yerinde öngörülerde bulunabilir.
- Bağımlı ve bağımsız değişken fark etmeksizin eksik, kayıp ya da ekstrem değerlerin olumsuz etkilerinin gözlemlenmediği bir yöntemdir.
- Alışılagelmiş pek çok istatistik yöntemine (varyans analizi, diskriminant analizi, lojistik regresyon, kümeleme analizi, çoklu regresyon) alternatif olarak kullanılmaktadır.
- Kesinlik arz etmemesine karşın kuvvetli temelleri olan ağaç yöntemlerini de dikkate alır (Lewis, 2000).
- Analizi yapacak olan kişiye yöntem sıralaması üzerinde düzeltme yapma imkanı sağlar (Lewis, 2000).
- Model ile bağımlı değişken üzerinde etkisi olan bağımsız değişkenler ve bu değişkenler arasındaki ilişki (interaction) elde edilir.
- Gereksinim halinde bir bağımsız değişken, aynı ağaç içerisinde başka ayrışma değerleri ile incelenebilir.

Her ne kadar geniş bir perspektif ve büyük bir esneklik sunsa da, CART çözümlemesinde de bazı kısıtlamalar bulunmaktadır. Sınıflama ve Regresyon Ağaçları Tekniğinin en büyük dezavantajı neticelerin bir olasılık modelini baz almıyor oluşudur. Veri seti doğrultusunda hazırlanmış olan bir Sınıflama ve Regresyon Ağacından elde edilecek tahmini sınıflamaya ışık tutabilecek bir güven aralığı veya olasılık derecesi mevcut değildir. CART ile elde edilen neticelerin gerçeği ne derece yansıttığına ilişkin olarak başvurulabilecek tek yol, yalnızca önceki verilerin ne derece yansıttığı ile orantı kurulmasıdır (Yohannes & Hoddinott, 1999).
CART alışılagelmiş çözümleme yöntemleri arasında olmadığından, istatistik uzmanlarını, yöntemin güvenilirliği konusunda ikna etmek çoğunlukla oldukça güçtür (Lewis, 2000).

5. SINIFLAMA AĞAÇLARI VE REGRESYON AĞAÇLARI İLE SINIFLANDIRMA

Karar ağaçlarını sağlama hedefiyle pek çok sınıflandırma modeli hazırlanmıştır. Sınıflandırma algoritmalarının performansları değişkenlik göstermektedir (Gey & Nedelec, 2005). C4.5 ve ID3 "ağaç yapılı sınıflandırma modelleri" haricinde "Sınıflandırma ve Regresyon Ağaçları (Classification And Regression Trees - CART)" olarak adlandırılan modeller de bulunmaktadır.

5.1. Sınıflama Ağaçları

Sınıflandırma ve Regresyon Ağaçları, "ikili ağaç (binary tree)" olarak isimlendirilen yapıda olmaktadırlar. Bir başka değişle, sınıflandırma ve regresyon ağaçlarının bir bölümünün yalnızca iki dalı bulunmaktadır. Bu kısımda, CART bünyesinde Twoing ayırma kriteri, Regresyon ağaçları ve Gini ayırma kriteri anlatılacaktır.

5.1.1. Twoing ayırma kriteri

Twoing ayırma kriteri, bütün basamaklarda eğitim kümesinin iki dala bölünmesi temel alınmaktadır (Duda, Hart, & Stork, 2000). Kümenin bütününün "T eğitim kümesi" tarafından ifade edildiği düşünüldüğünde, kümenin bölünen dallarından soldaki t_L, sağdakiyse t_R tarafından belirtilmektedir. T eğitim kümesinde yer alan kayıtların miktarının N olarak, sol bölüm adına P_L olasılığı şöyle hesap edilecektir (Larose, 2005)

$$P_L = \frac{S_{L1}}{N} \qquad [5.1]$$

Bu formülde pay kısmında yer alan S_{L1} kavramı, t_L olarak ifade edilen kümede bulunan nitelik değerlerinin söz konusu nitelik değerlerindeki tekrar miktarını ifade etmektedir. N ile belirtilen ise eğitim kümesinde bulunan kayıtların miktarıdır. Aşağıdaki koşullu olasılık formülüyse, bir j sınıfı değerinin soldaki dalda bulunma olasılığını belirtmektedir.

$$P(j \setminus t_L) = \frac{S_{L2}}{S_{L3}} \qquad [5.2]$$

Bu ifadede pay kısmında yer alan S_{L2} kavramı, t_L kümesinde bulunan kayıtların j sınıfları miktarını ifade etmektedir. Paydada bulunan S_{L3} kavramıysa, t_L kümesinde yer alan bütün nitelik değerlerinin söz konusu nitelik değerlerindeki tekrar miktarını belirtmektedir.

$$P_R = \frac{S_{R1}}{N} \qquad [5.3]$$

Bu ifadede pay kısmında yer alan S_{R1} kavramı, t_R kümesinde bulunan bütün nitelik değerlerinin söz konusu nitelik değerlerindeki tekrar miktarını ifade etmektedir. $P(j/t_R)$ kavramıysa, bir j sınıf değerinin sağındaki dalda yer alma olasılığını belirtmektedir.

$$P(j \setminus t_R) = \frac{S_{R2}}{S_{R3}} \qquad [5.4]$$

Bu bağıntının pay kısmında yer alan S_{R2} kavramı, t_R kümesinde yer alan kayıtların j sınıflarındaki miktarını belirtirken, S_{R3} kavramı t_R kümesinde bulunan bütün nitelik değerleri içindeki tekrar miktarını göstermektedir.

Seçilecek nitelik hakkında karar alınırken "uygunluk ölçütü değeri" dikkate alınmaktadır. Ø(s/t) kavramı t kısmındaki bölünmelerin "uygunluk ölçütü" olarak ele alınırsa bahsi geçen "uygunluk ölçütü değeri",

$$Ø(s/t) - 2P_L P_R \sum_{j=1}^{n} \left| P(j\setminus t_L)\, P(j\setminus t_R) \right| \qquad [5.5]$$

olarak hesap edilmektedir.

Ø(s/t) değerinin hesap edilmesinin ardından, maksimum olan (maxØ(s/t)) değeri seçilmektedir. Maksimum değeri bulunduran satır, bölünmenin başlayacağı niteliği saptamaktadır.

5.1.2. Gini endeksi

Gini endeksi, toplumda bulunan gelir eşitsizliğini ifade eden bir kavram olarak İtalyan istatistikçi Corrado Gini'nin 1912 senesinde geliştirdiği bir katsayı olarak ifade edilmektedir (Buchan, 2002). Gini endeksi çoğunlukla kullanılan bir eşitsizlik ölçütü olmakla beraber Lorenz eğrisiyle kolaylıkla ifade edilen bir bağı bulunmaktadır. Gini

endeksi 0 ila 1 arasında değişkenlik göstermekte ve 0 olduğunda "tam eşitliliği", 1 olduğundaysa "tam eşitsizliği" ifade etmektedir. Gini endeksine iktisatta haneler arası ve kişiler arası gelir eşitsizliğinin ölçülerek özetlenmesi adına standart bir analiz yöntemi olarak yer verilmektedir (Dumlu & Aydın, 2008). Lorenz eğrisi geometrik bir uygulama olması sebebiyle, takip ettiği seyrin izlenmesinde ve kıyaslamaların gerçekleştirilmesinde yeterli olmamış ve söz konusu eşitsizliğin bir endeksle belirtilmesi daha doğru görülmüştür. Sıklıkla uygulanan metotlardan birisi olan Gini endeksi, Lorenz eğrisi aracılığıyla hazırlanan, gelir eşitsizliğini bir oran ile ifade eden ve gelir eşitsizliğinin dereceni değerlendiren bir endekstir.

Lorenz eğrinin aşağı yönde seyretmesi, kişisel gelir dağılımında yaşanan eşitsizliğinin yükseldiğini belirtmekte ve böylece Gini endeksi de yükselmektedir. Gini katsayısının artması, gelir eşitsizliğinin çoğaldığını belirtmektedir. (Karakayalı , 2005), gini endeksini "mutlak eşitlik doğrusu ile Lorenz eğrisi arasında kalan alanın, mutlak eşitlik doğrusu altında kalan üçgenin alanına oranı" olarak ifade etmiştir. Gelirin kişiler arasında eşit bir şekilde dağılması halinde Gini endeksi 0 olacaktır. Gelir sadece bir şahsa ait olmuş ise, bir başka deyişle eşitsizliğin var olması halinde Gini endeksi 1 olacaktır. Gini endeksi 0'la 1 arasında değişen bir değer almakta ve 1 değerine yakınlaşması eşitsizliğin yükseldiğini, 0 değerine yaklaşması eşitsizliğin düştüğünü belirtmektedir. Gini endeksleri, daha kapsamlı ölçümlere kıyasla daha çok uygulanması nedeniyle, deneysel araştırmalarda sıklıkla yer bulmaktadır (Oğuş, 2004).

Gini endeksi CART bünyesinde yer alan sınıflama yöntemlerinden birisidir. Bir T kümesinde bulunan i sınıflarının "nispi frekans (relative frequency) miktarı" yahut diğer bir ifadeyle, i sınıfındaki değerlerin T kümesinde yer alan eleman miktarına oranı P_i olarak ifade edildiğinde Gini algoritması,

$$\text{gini}(T) = 1 - \sum_{i=1}^{k} P_i^2 \qquad [5.6]$$

şeklinde hesaplanmaktadır (Buchan, 2002).

Gini endeksinden faydalanarak öne sürülen "Gini sınıflandırma algoritması", Twoing modelindeki gibi, nitelik değerlerinin sağ ve sol olarak iki dala ayrılmasını temel almaktadır. İlk olarak, bütün nitelik değerleri "ikili" şekilde gruplar haline getirilir ve

böylece ortaya çıkan sağ ve sol ayrılmaları ifade eden sınıf değerleri grup haline getirilir. Veri kümesinde bulunan bütün nitelikler ile alakalı sağ ve sol ayrılmalar adına gini_R ve gini_L ifadeleri şöyle hesap edilmektedir:

$$\text{gini}_R = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_R|}\right)^2 \quad [5.7]$$

$$\text{gini}_L = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_L|}\right)^2 \quad [5.8]$$

Bu bağıntılarda yer alan k, sınıf miktarını; T, bir bölünmedeki örnekleri, $|T_L|$ kavramı solda bulunan örnek miktarını; L_i kavramı solda bulunan i bölümündeki örnek miktarını; $|T_R|$ kavramı sağda bulunan örnek miktarını ve R_i sağda bulunan i bölümündeki örnek miktarını ifade etmektedir. Algoritmadaki gini_R ve gini_L değerlerinin hesaplanmasının ardından ayrılmanın başlayacağı niteliğe karar verilmesi adına Gini endeksi uygulanır. Eğitim kümesinde bulunan satır miktarı n olarak, bütün j nitelikleri adına Gini endeksi aşağıdaki gibi hesaplanmaktadır:

$$\text{gini}_j = \frac{1}{2}(|T_L|\text{gini}_L + |T_R|\text{gini}_R). \quad [5.9]$$

Hesap edilen bütün ölçütler içinde minimum olan gini değeri, $(\min gini_j)$ seçilmekte, söz konusu değerin yer aldığı nitelik, ayrılmanın başlayacağı bölüm olarak saptanır.

5.2. Regresyon Ağaçları ile Sınıflandırma

İkili bölünme metoduna yer verilen bir başka Sınıflama ve Regresyon Ağaçları metodu "regresyon ağaçları" (regression trees) ismiyle anılmaktadır. Söz konusu yöntem, Breiman ve arkadaşlarınca gerçekleştirilen araştırmalar neticesinde meydana gelmiştir.

Regresyon ağaçlarıyla alakalı araştırmalarda çoğunlukla x "ölçüm vektörü", y ise "yanıt değişkeni" şeklinde adlandırılmaktadır. Bir kestirim kaidesi yahut X reel değerlerini alacak olan kestirimci $d(x)$ fonksiyonu şeklinde belirtilebilir. X'e bağlı olan d(x), "gerçek değerli bir fonksiyon" niteliğindedir. Regresyon analizi ise, L

öğrenme örneği başlangıcı olmak üzere, d(x) tahmin edicisinin saptanması durumuna denk gelen bir genel ifadedir. Bir kestirimci aşağıdaki gibi iki hedefle yaratılabilir:

– Yanıt değişkenlerinin ileride meydana gelecek ölçüm vektörleri ile uyumlu bir şekilde kestirilmesi,

– Yanıt değişkenleriyle ölçüm değişkenleri arasında olan yapısal bağıntısının ifade edilmesidir.

n tane $(x_1,y_1),\ldots,(x_n,y_n)$ gözlemleri ile meydana gelen bir L öğrenme örneği, bir d(x) kestirimcisini yaratmak adına kullanıldığında, kestirimin doğruluğunun ne şekilde saptanacağı problemi meydana gelmektedir. N_2 miktarında oldukça büyük bir $(x'_1, y'_1), \ldots, (x'_{N_2}, y')$ test örneğinin varlığı halinde d(x) kestirimcisinin doğruluğu şu formülden yola çıkılarak hesap edilen "ortalama hata" ile belirlenebilir (Breiman, Freidman, Olshen, & Stone, 1998).

$$\frac{1}{N_2}\sum_{n=1}^{N_2}|y'_n - d(x'_n)|. \qquad [5.10]$$

Bu bağıntıda yer alan d(x'_n) kavramı, n=1,…,N_2 olmakla beraber, y'_n ifadesinin bir tahmincisi niteliğindedir. Söz konusu ölçüm "en küçük mutlak sapma regresyonu"nu temel almaktadır. Fakat daha kolay bir şekilde hesaplayabilmek adına, doğruluk belirlemesinde klasik regresyondaki gibi "hata kareleri ortalaması" ifadesi uygulanır. "Hata kareleri ortalaması" şöyle belirtilebilir:

$$\frac{1}{N_2}\sum_{n=1}^{N_2}\left(y'_n - d(x'_n)\right)^2 \qquad [5.11]$$

d tahmin edicisine ait R^*(d) ise "hata kareleri ortalamasını" ifade etmektedir ve aşağıdaki gibi hesaplanır (Breiman, 1998);

$$R^*(d)=E(Y-d(X))^2. \qquad [5.12]$$

Bir başka deyişle, R^*(d) kavramı, yanıt değişkeninin kestirimcisi şeklinde d(x) kestirimcisini kullanan "beklenen hata kareleridir". R^*(d) kavramı, d sınıflandırıcısının "hatalı sınıflandırma oranı"nı ifade etmektedir. Bu noktada, kestirimcinin doğrulunun

saptanması adına da aynı ifadeye yer verildiği görülmektedir. $R^*(d)$ kavramını en küçük hale getiren d_B kestirimcisi şöyle ifade edilmektedir:

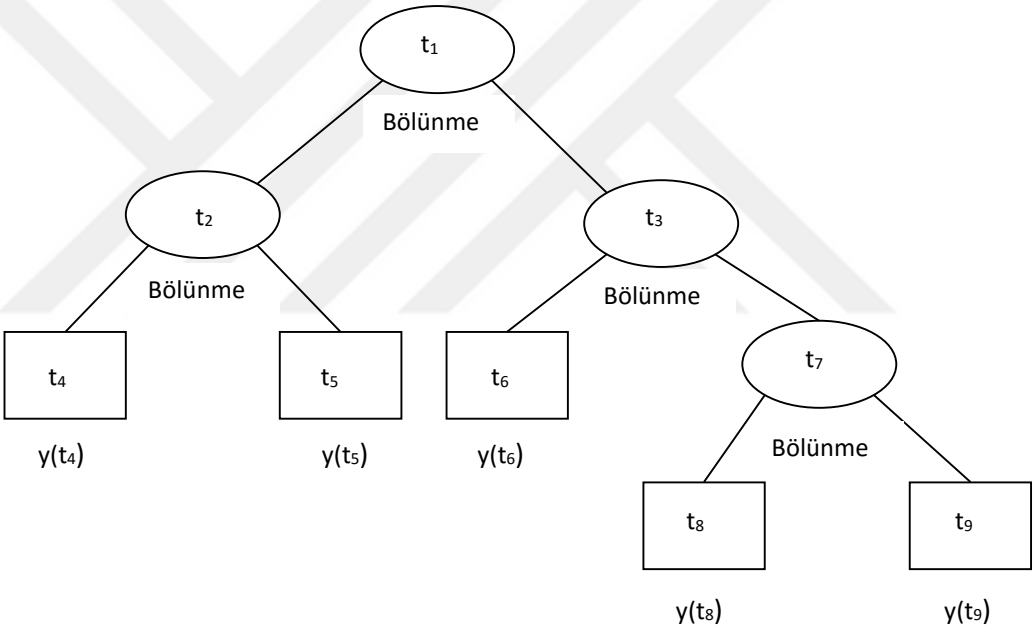
$d_B=E(Y/X=\chi)$. [5.13]

Farklı bir ifadeyle, verilen x ölçüm vektörü adına yanıtın koşullu beklentisi, $d_B(x)$ 'tir.

5.2.1. Regresyon ağacı oluşumu

Ağaç yapılı regresyon, "ağaç yapılı sınıflandırma" esasını temel almaktadır. Terminal düğümlerinde, y(t) cevapları sabittir (Breiman, Freidman, Olshen, & Stone, 1998).

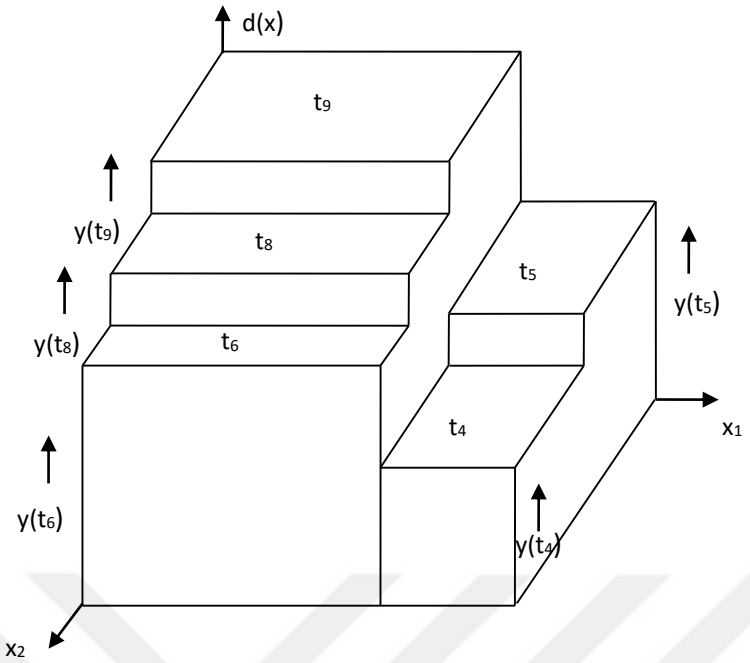
Şekil 5.1'de ağaç yapılı regresyonlara ait bir örnek gösterilmektedir.



Şekil 5.1: Ağaç yapılı regresyon (Breiman vd., 1993)

Bütün terminal düğümlerinde kestirimci d(x)'in sabit olması sebebiyle ağaç, regresyon yüzeyinin bir "histogram tahmini" şeklinde de ifade edilebilir.

Şekil 5.2'de Regresyon yüzeyinin histogram şeklindeki bir örnek gösterilmektedir.



Şekil 5.2: Regresyon yüzeyinin histogram şeklindeki ifadesi (Breiman vd., 1993)

L öğrenme örneğiyle başlayıp, bir ağaç bölünmesini saptamak adına üç faktör bulunmaktadır:

– Ara düğümlerde bölünmenin seçilmesi adına bir yol olması,

– Bir düğümün terminal olduğu zamanı belirlemek için bir kural tanımlanması,

– Her t terminal düğümüne, bir y(t) değerini atamak için bir kural oluşturulmasıdır.

– Bu unsurlar incelendiğinde, sınıflandırmada olduğu gibi, düğüm belirleme kuralının kolayca ortaya konulacağı anlaşılır (Beriman, Fredman, Olshen, & Stone, 1998).

$R^*(d)$ ifadesinin yeniden yerine koyma tahminiyle beraber süreç başlar, tahmin ise şöyledir:

$$R(d) = \frac{1}{N}\sum_n (y_n - d(x_n))^2. \qquad [5.14]$$

Temel olarak regresyon ağaçların oluşumu 4 aşamada sınıflandırılırsa aşağıdaki gibi olacaktır:

1. Veri setindeki soruların oluşumu

Regresyon ağaçlarında örneklem oluşumu sınıflama ağaçlarındaki gibidir. Fakat tek farklılık deney ünitelerinin ait olduğu (J) sınıflar sayısal verilerden oluşur.

2. Ayırma kuralları

Regresyon ağaçlarında amaç en uygun bölme kriterini tespit ederek düğümdeki homojenliği maksimum hale getirerek düğümler arasındaki heterojenliği minimize etmeye çalışmaktır. Böylece çocuk düğümleri arasındaki farklılık maksimum düzeye ulaşmış olur.

Regresyon ağaçlarında üç ayırma kuralı vardır. Bunlar; En küçük kareler (LS), En küçük mutlak sapma (LAD) ve Clark&Pregibon(CP) dir. Bu üç kuralda da amaç düğümlerdeki heterojenliği minimize etmektir. Heterojenlik ölçüsü i(t) sembolüyle gösterilmektedir.

LS ve LAD kurallarının farkı şu şekildedir:

- En küçük kareler kuralına göre heterojenlik ölçüsü i(t) düğümdeki ortalama etrafında bağımlı değişkenlerin karelerinin toplamıdır.

- En küçük mutlak sapma kuralına göre ise heterojenlik ölçüsü i(t) düğümdeki medyan etrafında bağımlı değişkenin kareleri toplamıdır.

Least Squares (LS) Kuralı

Regresyon ağacında dallanma işleminde kullanılan bölme kriteridir

$i(t)=\sum_{i=1}^{N}(Y(i)-\bar{Y}(t))^2$ [5.15]

Burada;

i(t) = t. düğümdeki heterojenlik

Y(i) = t. düğümdeki bağımlı değişkenin değeri

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

Clark&Pregibon (CP) Kuralı

Sapma düğümündeki tüm gözlemlerin sapmalarının toplamıdır. Amaç hata kareler toplamını (RRS) minimize etmektir.

$$(RRS)=\sum_{i\varepsilon L}(y_i-\bar{y}_L)^2+\sum_{i\varepsilon R}(y_i-\bar{y}_R)^2 \quad [5.16]$$

Burada,

y_i = Sol düğümdeki bağımllı değişkenin değeri

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

$\bar{y}_R$=Sağ düğümdeki bağımlı değişkenin ortalama değerini göstermektedir

3. En İyi Ayırma Kriterlerini Tespiti

En iyi ayırmada amaç $\phi(t)$'yi maksimize etmektir. Fonksiyonu aşağıdaki gibidir:

$$\phi(t)=i(t)-i(t_R)-i(t_L) \quad [5.17]$$

$i(t_R)$= Sağ çocuk düğümdeki ortalama etrafındaki kareler toplamı

$i(t_L)$= Sol çocuk düğümdeki ortalama etrafındaki kareler toplamıdır.

4. Regresyon Ağacı Doğruluk Tahmini

Regresyon ağacındaki doğruluk tahmini, doğru olarak sınıflanan deney grubu sayısının toplam deney grubu sayısına bölünmesi ile doğruluk oranı hesaplanır.

5.2.2. Hata ölçüm ve tahminleri

$(x_1,y_1),\ldots,(x_N,y_N)$ şeklinde meydana gelen bir Ω öğrenme örneği, $R^*(d)$ hatasının tahmininde kullanılmasının yanı sıra, d(x) kestirimcinin sağlanmasında da kullanılmaktadır. $R^*(d)$ hatasının tahmininde birkaç yöntem bulunmaktadır. "Yerine koyma (resubstitution) yaygın olarak kullanılan tahmin etme yöntemidir. (Breiman, Freidman, Olshen, & Stone, 1998):

$$R(d)=\frac{1}{N}\sum_n(y_n-d(x_n))^2 \quad [5.18]$$

$R^{ts}(d)$ test örnek tahminlerinin elde edilmesi, L kümesinin L_1 ve L_2 şeklinde ayrılması ile şöyle gerçekleşmektedir:

$$R^{ts}(d) = \frac{1}{N_2}\sum_{(x_n,y_n)\in \mathbb{V}}(y_n - d(x_n))^2. \quad [5.19]$$

$R^{CV}(d)$ ifadesinin çapraz doğrulama tahmini olarak kabul edilirse, L kümesini aynı miktarda olay barındıran v sayısında $L_1,\ldots, L_V$ alt kümelerine bölünmesiyle elde edilir. Bu noktadan hareketle $R^{CV}(d)$ çapraz doğrulama tahmini şöyle belirtilebilir:

$$R^{CV}(d) = \frac{1}{N}\sum_V \sum_{(x_n,y_n)\in \mathbb{V}}\left(y_n - d^{(V)}(x_n)\right)^2 \quad [5.20]$$

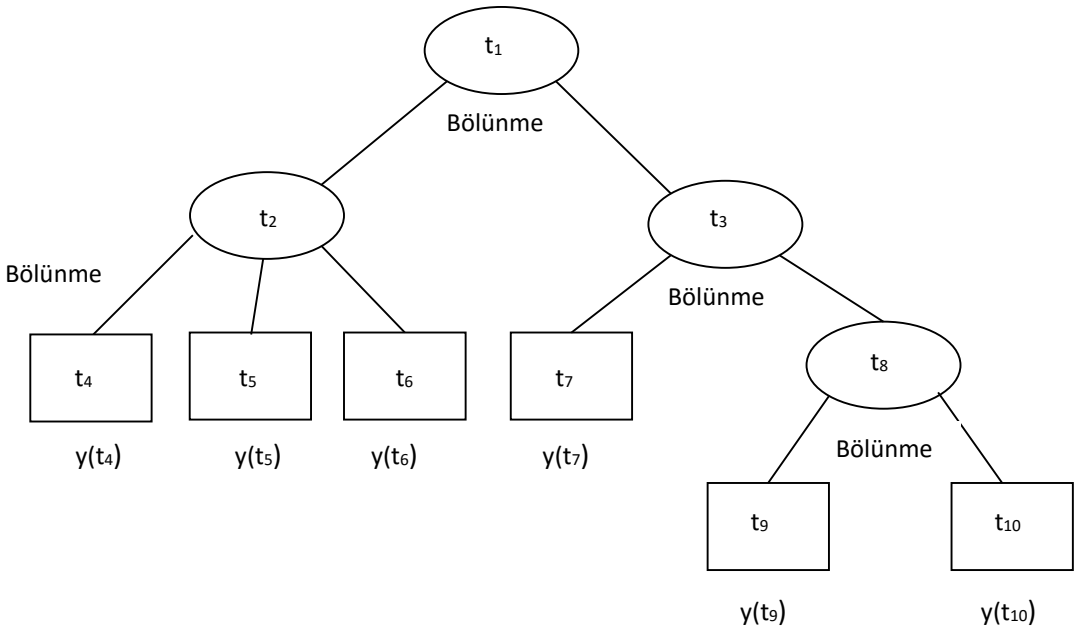
6. CHAID ANALİZİ

1980 yılında Gordon V. Kass tarafından CHAID (Chi-squared Automatic Interaction Detection), en popüler karar ağacı algoritmalarının başında gelmektedir. İlk uygulama alanları tıbbi ve psikiyatrik araştırmalar alanında gerçekleşmiştir. Fakat günümüzde doğrudan pazarlama için tüketici gruplarının seçiminde kullanılmıştır. Her hiyerarşik seviyede ikiden daha fazla dala ayrılabilmesi, bütün ölçü skalaları ile çalışabiliyor olması, nesnelere ağırlık ve frekans değerlerinin atanabilmesi, eksik değerleri ayrı bir kategoride analiz edebilmesi önemli bir avantaj sağlamaktadır. Parametrik olmaması ve verilerin normal dağılıma uyma zorunluluğu olmaması, bu tekniğin daha kullanışlı olmasına sebeptir.

Bağımlı değişkenin sürekli olması durumunda F, kategorik olması durumunda χ^2 testler kullanılmaktadır. Bununla birlikte sürekli bağımsız değişkenler analizinde otomatik olarak kategorik değişkenlere dönüştürülür. CHAID algoritmasının ileri düzey tanımlamalarında Pearson χ^2 veya likelihood-ratio testleri uygulanabilir Akpınar H., 2017).

6.1. CHAID Analizini Genel Yapısı

Şekil 6.1'de CHAID algoritma yapılı ağaç modeline ait bir örnek gösterilmektedir.



Şekil 6.1: CHAID algoritması yapılı ağaç model gösterimi

Otomatik Ki-Kare Etkileşim Belirleme (CHAID) (Chi-Squared Automatic Interaction Detector) Analizi, kategorik bağımlı değişkenler için oluşturulmuş olup, AID analizinin bir uzantısı olarak kabul edilir. Otomatik Ki-Kare Etkileşim Belirleme Analizinin asıl amacı; veriyi daha homojen bir şekilde birden çok alt gruba bölmektir. Devasa bir veri kümesinin homojen bir alt gruba indirgenmesi; bağımlı değişkeni mümkün olduğunca tutumlu olarak açıklayan diğer değişkenleri ve bunlarla ilgili verileri meydana getirmek anlamına gelir. Otomatik Ki-Kare Etkileşim Belirleme Analizi; kategorik değişkenlerle ilgili veri kümesini, bağımlı değişkeni en iyi açıklayan türde detaylı ve homojen alt gruplara böler. Bu alt grupların en önemli özelliği, tahmin edici olmalarıdır. Seçilen tahmin edici gruplar; daha sonra yapılacak ileri analizlerde bağımlı değişkenin tahmininde kullanılır. Otomatik Ki-Kare Etkileşim Belirleme Analizi, regresyon analizlerinde kullanılabileceği gibi karar ağaçlarının oluşturulmasında da kullanılabilir. Değişkenler arasındaki ilişki lineer yapıdan daha karmaşık ise veride gizli olan bu ilişkiyi bulmak için verinin belli kısımlarını eleme yöntemi olan CHAID kullanılır. "Ki-kare" ismini de almasının nedeni algoritmasında birçok çapraz tablonun kullanılması ve istatistiksel önem oranları ile çalışmasıdır (Hoare R., 2004).

CHAID analizinde bağımsız değişkenlerin herbiri için en iyi dallanma hesaplanır. Daha sonra bağımsız değişkenler en iyisi seçilene kadar karşılaştırılır. Seçilen en iyi bağımsız değişkene göre tekrar dallanma işlemi yapılır. Her bir bağımsız değişken kategorilerinin en anlamlı şekilde dallanma işlemi gerçekleştirildikten sonra bağımlı değişkene göre kontenjans tabloları oluşturulurak Bonferroni p değerleri ile $\chi 2$ istatistikleri hesaplanır. Hesaplanan istatistikler doğrultusunda önem derecesine göre kontenjans tabloları şekillenmiş olur. Buradan da anlaşıldığı üzere CHAID analizi ki-kare istatistiklerini, Bonferronni yaklaşımını ve kategori birleştirme algoritmalarını kullanarak araştırmacının ağaç diyagramı ile en iyi açıklayıcı değişkenleri ve bağımlı değişken ile olan etkileşimleri elde etmesine olanak sağlar (Hoare R., 2004).

6.2. CHAID Analizi Algoritması

Bağımlı değişken kategori sayısı $d \geq 2$ olsun. Analiz edilecek olan belirli bir açıklayıcı değişken $c \geq 2$ sayıda kategoriye sahip olsun. Analizdeki amaç, c x d kontenjans tablosunu açıklayıcı değişkenindeki uygun kategorileri birleştirme yolu ile en anlamlı

j x d tablosuna indirgemektir. Kavramsal olarak ilk olarak $Tj^{(i)}$ istatistiğini hesaplanır. $Tj^{(i)}$ j x d tablosunu oluşturmadaki i. yöntem için χ^2 istatistiğidir. (j: 2,3,4.....,c; i'nin değişim aralığı açıklayıcı değişkenin tipine bağlıdır.) $Tj^{*}=maxiTj^{(i)}$ ise en iyi j x d tablo için, χ^2 istatistiği elde edilmiş olunur.

Monotonik ya da dichotomous serbest açıklayıcı değişkeninin varlığında $Tj^{(i)}$ Fisher metoduna göre bulunabilir. Bu dinamik program c^2 hesaplarına dayanır. $d \geq 3$ ve açıklayıcı değişken sıralı kategorilere sahipse Fisher metodundan yararlanılamaz (Kass,G.V., 1980).

Algoritma aşağıdaki gibi üç aşamadan oluşmaktadır.

1. Birleşme

1.Adım: Her bir açıklayıcı değişken için sırasıyla, açıklayıcı değişkenin kategorileri ile bağımlı değişkenin kategorilerini çapraz tablosu bulunur ve adım 2 ve 3 uygulanır.

2.Adım: Sadece açıklayıcı değişkenin tipi tarafından belirlenen uygun çiftler göz önüne alınarak, 2 x d alt tablosunda anlamlılığı düşük olan açıklayıcı değişken kategori çiftleri bulunur. Eğer önem derecesi kritik bir değere ulaşmıyorsa, bu iki kategori birleştirilir. Bu birleşim tek bir kategori olarak ele alınır ve bu adım tekrarlanır. Bu işlem açıklayıcı değişkenin kendi içindeki birleşmeleri anlamsız oluncaya kadar tekrar eder.

3. Adım: Açıklayıcı değişkenin tipi tarafından oluşturulan ve orjinal kategorilerin 3 veya daha fazlasının birleştirilmesi ile meydana gelen; her bir birleşik kategori için, birleşmenin tekrar ayrılabileceği en önemli ikili bölüne bulunur. Eğer önem derecesi kritik değerin üzerindeyse, bölünme gerçekleştirilir ve 2. adıma dönülür.

2. Dağıtma

4. Adım: Optimal bir şekilde birleştirilmiş olan, her bir açıklayıcı değişken için önem derecesi hesaplanarak, en büyük önem derecesine sahip olan, diğerlerinden ayrılır. Eğer önem derecesi, verilen kriter değerlerinden büyük ise, veri kümesi seçilen açıklayıcı değişkenin birleştirilmiş kategorilerine göre alt gruplarına bölünür.

3. Durdurma

5. Adım: Verinin analiz edilememiş her bir grubu için, 1. adıma dönülür. Bu adımda en az sayıda gözleme sahip olan gruplar göz ardı edilir (Kass,G.V., 1980).

7. UYGULAMA

Bu tez çalışmasında; University of California bünyesinde veri setlerini barındıran bir platformdan alınan kalp hastalığına etki eden 38 adet faktör kullanılmıştır (https://archive.ics.uci.edu/ml/datasets.html). Bu faktörlerin değerlendirilmesi; Sınıflama ve Regresyon Ağaçı (CART) ve Otomatik Ki-Kare Etkileşim Belirleme (CHAID) algoritmaları kullanılarak oluşturulmuştur ve çıkan sonuçlar birbirleriyle karşılaştırılarak yorumlanmıştır.

7.1. Kalbin Yapısı

Kalp fibröz perikard ile kaplı, orta mediastinumda bulunan fibromuskuler bir organdır. Kalp kası özelleşmiş liflere sahip bir yapıdadır ve istemsiz çalışır. Kalbin sivri uç kısmına apex, taban kısmına da basis adı verilir. Kalbin 1/3 'ü mid- klaviküler hattın sağına, 2/3 'ü ise soluna bakar. İnsan kalbi dört adet boşluktan meydana gelir ve bu boşluklar bir takım duvar yapılarıyla birbirinden ayrılır. İki adet atrium arasında interatriyal duvar, iki adet ventrikül arasında da interventriküler duvar bulunur.

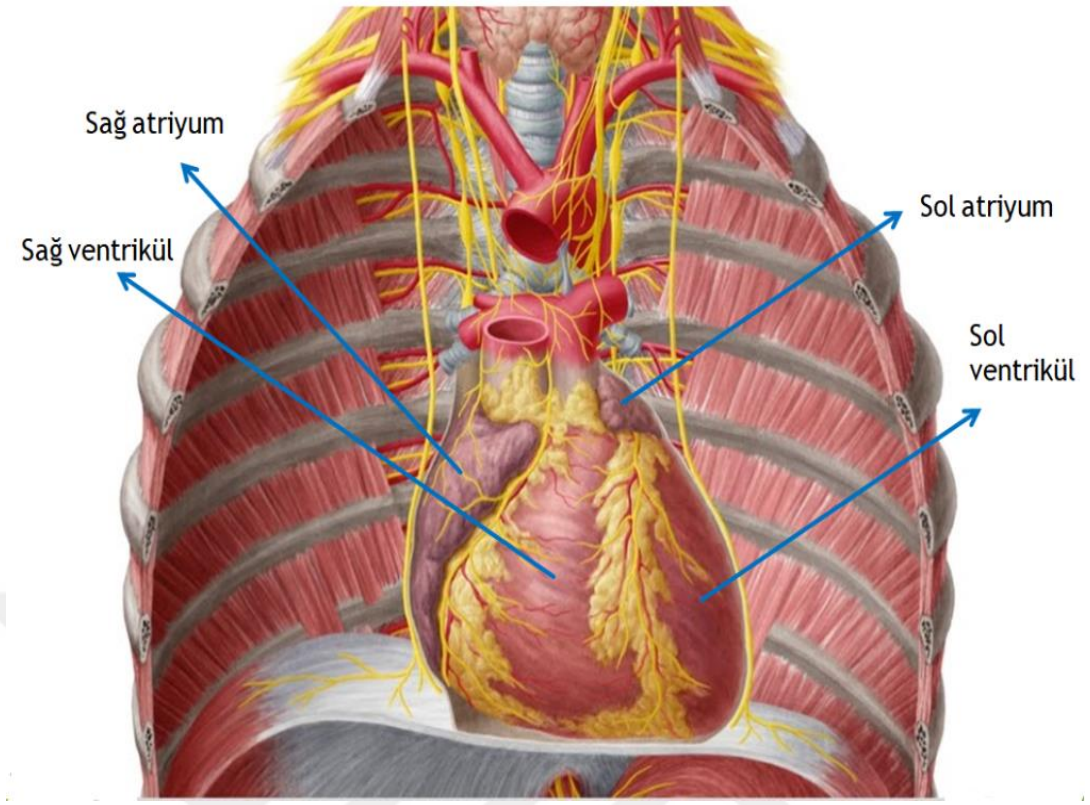
Kalbin boşlukları sırasıyla sağ atriyum, sağ ventrikül, sol atriyum ve sol ventrikülden meydana gelir (Şekil 1). Sağ atriyum, vücudun venöz kanının toplandığı bölümdür. Vena cava superior, vena cava inferior, sinus coronarius gibi venler buraya dökülür. Sağ atriyumun üzeri auricula denen bir kapak ile örtülüdür ve burada bulunan kas yapısına musculus pectinati adı verilir. Anne karnındaki dönemde akciğerler efektif olarak kullanılmadığı için kan, sağ atriyum ile sol atriyum arasında bulunan foramen ovale yardımıyla dolaşımını gerçekleştirir. Ancak doğumdan sonra bu delik kapanarak fossa ovalis adını alır. Doğum sonrasında delik, belli bir süre içerisinde kapanmazsa ASD (atriyal septal defekt) denen hastalık meydana gelir. Sağ atriyumdan sağ ventrüle geçişte triküspid kapak bulunur.

Kan, triküspid kapaktan geçtikten sonra sağ ventriküle gelmiş olur ve sağ ventrikülde bulunan güçlü kas yapısı trabecula carnea sayesinde hızlı bir şekilde kalın bir damar olan truncus pulmonalis 'e yönlendirilir. Ayrıca burada kana yön vermek amacıyla bulunan diğer bir yapı da trabeculoseptomarginalis (moderatör band) dir ve sadece sağ ventrikülde bulunur. Truncus pulmonalis, sağ ve sol akciğere ulaşabilmek için iki dala ayrılır; sağ pulmoner arter ve sol pulmoner arter.

Sağ ventrikülden akciğerlee ulaşan ve oksijene olan kan, sağ ve sol pulmoner venler yardımıyla sol atriyuma gelir. Sol atriyuma gelen kan buradan mitral kapak yardımıyla sol ventriküle geçer ve buradaki kas yapısıyla Aort ‘a yönlendirilir. Aort kalbin dışında, öncelikle kalbi beslemek amacıyla iki önemli dalını verir; sağ koroner arter ve sol koroner arter. Sağ koroner arter aynı zamanda kalbin arka kısmındaki sulcus interventricularis posterior üzerine de dal verir (arteria interventricularis posterior). Sol koroner arter de kalbin ön kısmındaki sulcus interventricularis anterior üzerine doğru dal verir (arteria interventricularis anterior).

Aort; daha sonra vücudun uzantılarına, alt kısma, baş ve boyuna kan iletebilmek adına 3 kısma ayrılır. Aorta ascendens, arcus aorta ve aorta descendens. Daha sonra göğüs bölgesini de diaphragmayı delerek geçtikten sonra, karın bölgesine ulaşır ve adını değiştirerek aorta abdominalis olarak adlandırılır. Arcus aorta dan üç önemli arter ayrılır. Bunların en sağda bulunanı truncus brachiocephalicus ‘tur ve bu da yine başa ve boyuna gitmek üzere ikiye ayrılır. Sol tarafta ise; ilk dal arteria carotis communis sinistra ve diğeri de arteria carotis communis sinistra ‘dır bu da tekrar kafa içine ya da dışına dallar verir.

Kalbi saran perikarda bakıldığında, iki tabakası arasında liquor pericardii denen kaygan bir sıvı bulunur ve bu sıvı kalp kasılıp gevşerken ortamı kayganlaştırmak için bulunur. Kalp zarının iltihabi durumuna perikardit denir. Perikardit aniden gelişebilen bir rahatsızlıktır.



Şekil 7.1: Kalbin önden görünümü ve kalbin boşlukları(https://www.kenhub.com/)

Kalp, otonom sinis sistemi etkisi altında çalışır. Kişi sempatik sistemin etkisi altındayken kalp hızı artar, solunum artar ve bağırsak hareketleri azalır. Ancak kişi parasempatik sistemin etkisine girip rahatladığında tüm bu işlemler tersine döner. Bunun anlamı; kalp ritmi düşer, solunum normale döner ve bağırsakta sindirim tekrar hızlanır(Cumhur, 2014).

7.2. Kalbin İleti Sistemi

Kalp, işlev görebilmesi için gereken enerjiyi kendisi üretebilmektedir. Kalbin sağ üst-dış kısmında bulunan sinoatriyal düğümde (S- A nodülü) uyarılar başlar ve dakikada 60- 90 impuls üretilir. Bu uyartılar daha sonra atrioventriküler düğüme iletilir ve burada dakikada 40- 60 arası impuls üretilir. Yani daha yavaş bir iletim mevcuttur. Uyartılar buradan sağ ve sol atriyoventriküler lif demetlerine iletilerek bütün kalbe dağılması sağlanır ve kalpte kasılıp- gevşeme mekanizması işlev görmüş olur. Eğer S-A nodülünde uyarının gerçekleşmesi mümkün değilse, hastaya yapay kalp pili takılır (Mahadevan V, 2017).

7.3. Kalp Hastalıkları

Günümüzde sık rastlanan kalp hastalıklarının başında koroner kalp hastalıkları, mitral kapak hastalıkları, aort kapak hastalıkları, aort anevrizmaları ve ASD (atriyal septal defekt) gelmektedir. İnsanlarda meydana gelen bu kalp hastalıklarının altında yatan sebepler çok çeşitli olabilir. Bununla birlikte kişinin yaşı, cinsiyeti, egzersiz alışkanlığı, sigara kullanımı, tansiyon, diabet gibi faktörler de kalp hastalıklarının oluşumunu tetikleyebilir. Koroner kalp hastalıklarında kadınlar erkeklere göre 5 kat daha fazla risk altındadır. Bu kişiler genelde mikrovasküler koroner disfonksiyona sahip kişilerdir. (Sedlak, T., 2014) .

Amerika Birleşik Devletleri ‘nde iskemik kalp hastalıklarına ilişkin son yıllarda yapılan istatistiksel çalışmalarda, 1991 yılından bu yana hayatını kaybedenlerin oranlarının %30 un altına düştüğü belirtilmiştir. Ancak halen her dört kadından biri halen kalp rahatsızlıklarından hayatını kaybetmektedir. Özellikle 75 yaş ve üzerindeki kadın hastalarda, erkeklere göre daha riskli bir durum söz konusudur (Murphy, N., 2017).

Kalpte bulunan dört adet kapak oluşumu, kalpteki dolaşımı sağlayan temel yapılardır. Atriyoventriküler kapaklar atriyumları ve ventrikülleri birbirinden ayırır. Bu kapaklar diyastol boyunca açıktır ve ventriküllerin kanla dolmasına yardımcı olur. Aortik ve pulmoner kapaklar ise sistol boyunca açıktır ve kanın kalpten pompalanmasına yardımcı olurlar. Bu kapaklarda meydana gelen birtakım patolojik olaylar, stenoz ya da regürjidasyon olarak tanımlanır. Stenoz durumunda kan akışı normalin altına iner. Regürjidasyonda ise kapalı halde bulunan kapaklardan kan geriye doğru kaçar. Kapaklarda meydana gelen bu anomaliler doğumsal ya da edinilmiş (sonradan ortaya çıkan) olabilir. Kalp kapaklarında meydana gelen bu anomaliler genelde ölümle sonuçlanır. Bazı özel kapak anomalilerinde transtorasik ekokardiyografi sayesinde önemli bulgular saptanabilir. Ancak genelde birçok semptom ortaya çıktıktan sonra tespit edilebilirler. Bunun anlamı; kardiyomiyopati çoktan oluşmuş olabilir. Kapaklarda meydana gelen anomalilerde görülen semptomlardan bazıları, egzersiz anında aşırı yorulma, ağrı, dispne, göğüs ağrısı ve baygınlıktır. Semptomlar genelde 50- 70 yaş arası insanlarda daha sık görülür (Zorana, M., 2017).

Kap hastalıkları arasında sıkça rastlanan bi diğer durum da kalp yetmezliğidir. Kalp yetmezliğinin en temel sebepleri arasında kapaklarda meydana gelen lezyonlar, dilate

olmuş kardiyomiyopati, iskemik kalp hastalıkları, atriyal fibrilasyon ve hipertansiyondur. Kemoterapinin toksik etkileri, aşırı alkol tüketimi ayrıca sistolik sol kapak bozukluğuna sebep olur. Bunların yanısıra; diabetes mellitus, aterosklerozun (damar tıkanıklığı) oluşmasına, hipertansiyona sebep olabilir. Bazı bilimsel veriler tip 2 diabetin diabetik kardiyomiyopatiye sebebiyet verebileceğini göstermiştir. Özellikle diabet ve nefropatisi olan hastalarda, kalp yetmezliği de bir komplikasyon halini almıştır. Diyabeti olan hastalarda, olmayanlara göre kalp yetmezliği sebebiyle ölüm oranı 10 kat daha fazladır. Tip 2 diyabet hastalarında kalp yetmezliğini önlemek amacıyla kilo vermek ve egzersiz yapmak oldukça önemlidir ve risk faktörlerini indirgeyen de bir durumdur (Jorsal A., 2017).

7.4. Analiz ve Bulgular

Bu çalışmada kalp hastalıklarına etki eden faktörlerin kendi aralarındaki ilişkileriyle, bağımlı değişken ve bağımsız değişkenleri arasındaki ilişki CART ve CHAID algoritmaları kullanılarak ayrı ayrı değerlendirilmiştir.
Çalışmada kalp rahatsızlığı durumu bağımlı değişken ve diğer 38 değişken ise bağımsız değişken olarak analize dahil edilmiştir. Araştırmada ki gözlem sayısı 248'dir. İlk önce CART algoritması ve ardından CHAID algoritması ile sınıflandırma ağaçları oluşturulmuş ve sonuçlar yorumlanmıştır.

Sınıflandırma ağacında öncelikle en uygun ağacın hangi ağaç olduğuna karar vermek gerekmektedir. Her iki algoritma için de uygun ağacı belirleyebilmek adına üç farklı deneme yapılmıştır. Bu çalışmada ilk olarak veri setinin % 70'i çalışma örneklemi ve %30'u da test örneklemi olarak alınmıştır. İkinci olarak ise; veri setinin %50'si çalışma örneklemi %50'si de test örneklemi olarak alınmıştır. Son olarak; veri setinin %30'u çalışma örneklemi ve %70'i de test örneklemi olarak alınmıştır. 3 faklı oranla oluşturulan modellerde farklı düğüm sayıları ile uygun ağaç belirlenmeye çalışılmıştır. En yüksek doğru sınıflama oranına sahip olan ağacın en uygun ağaç olduğuna karar verilmiştir.

CHAID algoritması ile oluşturulan en uygun ağacın; veri setinin %70'i çalışma örneklemi %30'u test örneklemi olarak alındığı ağaç olduğu görülmüştür.

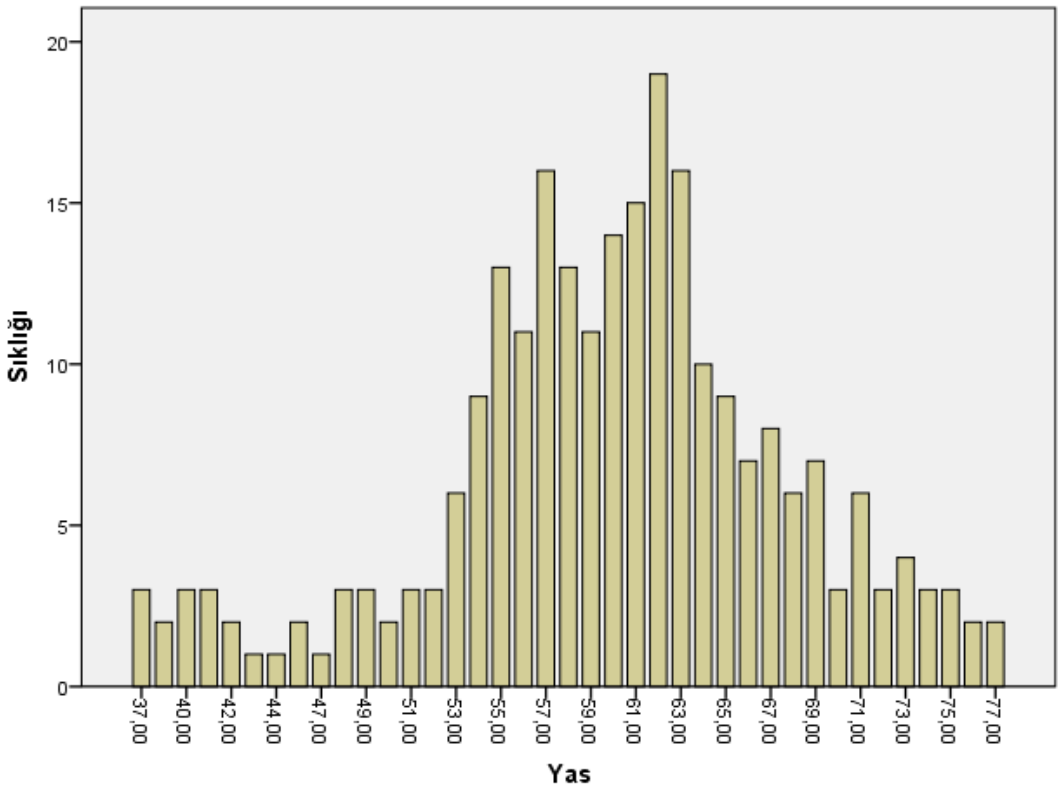
7.4.1. Değişkenlere ait tanımlayıcı istatistikler

Araştırmaya katılan katılımcılara ait bazı bilgiler üzerinde çalışılmıştır.

Çizelge 7.1. Katılımcıların yaş dağılımı

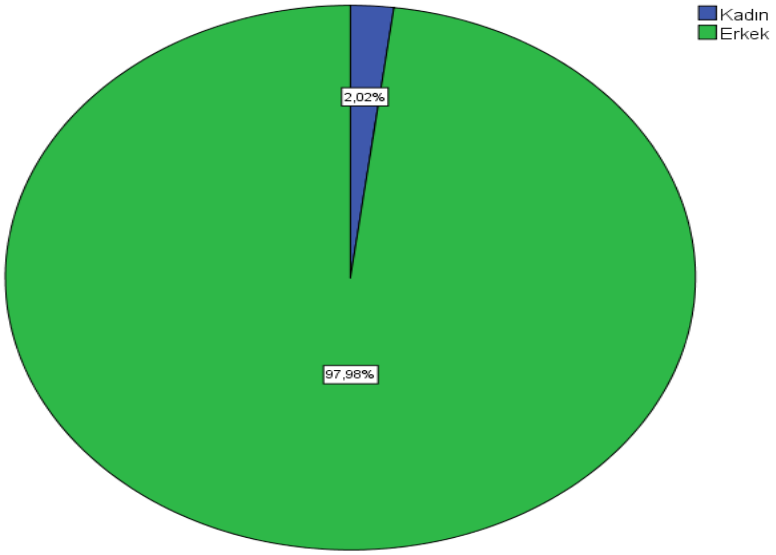
Örneklem	248
Ortalama Değer	59,85
Minimum Değer	37
Maxsimum Değer	77

Katılımcıların yaş dağılımı Çizelge 7.1'de görülmektedir. Burada katılımcılara ait örneklem büyüklüğü, ortalama yaş miktarı ve minimum-maksimum yaş aralığı belirtilmektedir.



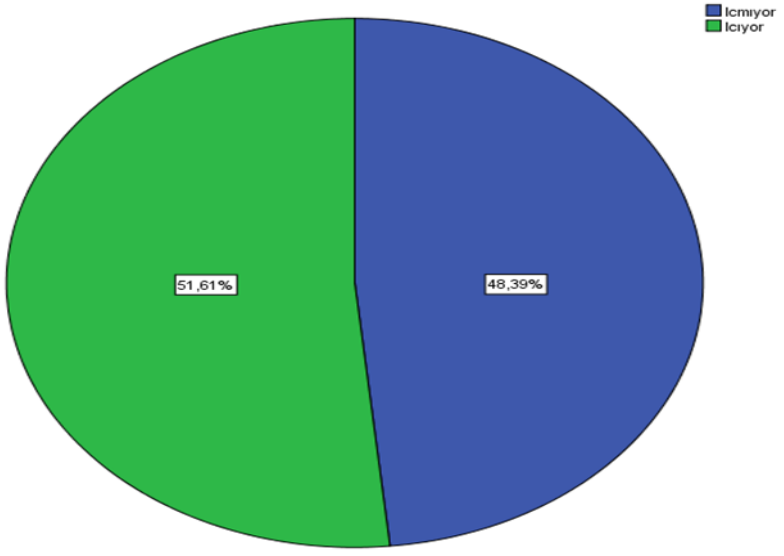
Şekil 7.2: Katılımcıların yaş dağılımı

248 kişilik örneklem ile yapılan çalışmada katılımcıların yaşları Şekil 7.2'de görülmektedir. Katılımcıların yaşları 37 ile 77 arasındadır. Genel itibari ile örneklemdeki yaş ortalamasının 60 olduğu saptanmıştır.



Şekil 7.3: Katılımcıların cinsiyet dağılımı

Katılımcıların cinsiyet durumu Şekil 7.3'te görülmektedir. Araştırmaya katılan katılımcıların büyük çoğunluğunu %97,98 ile erkek hastalar oluşturmaktadır. %2,02'lik kısmını kadın hastalardan oluştuğu görülmektedir.



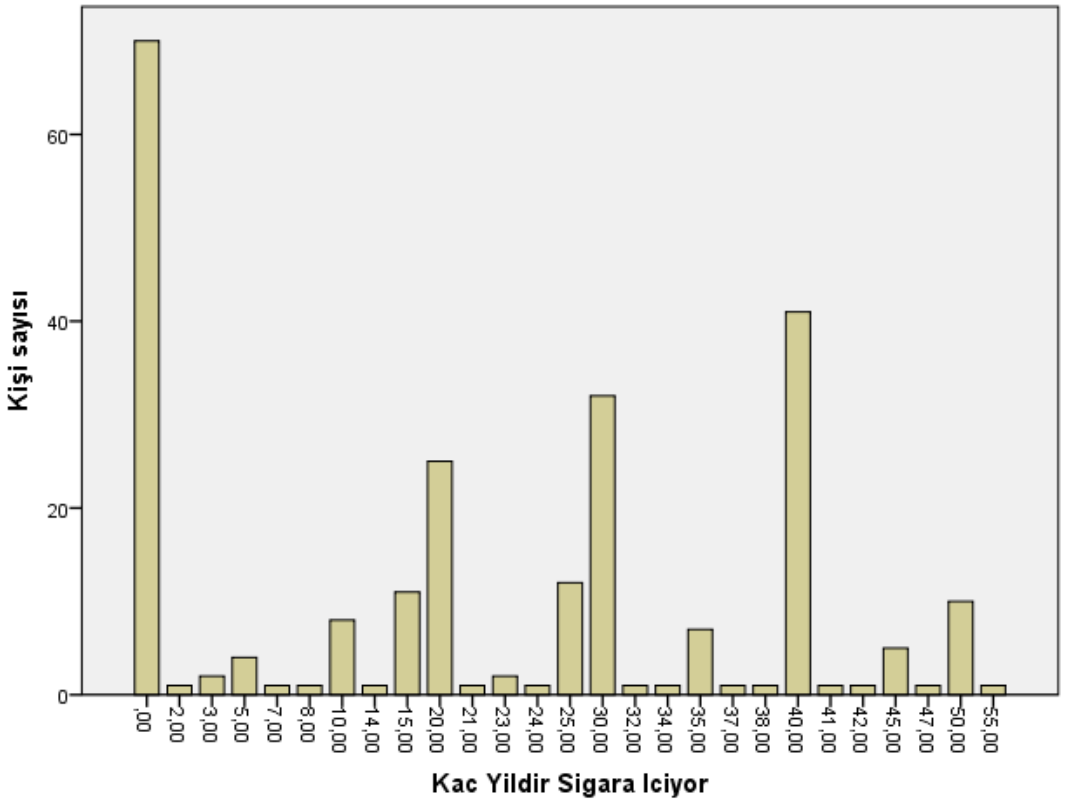
Şekil 7.4: Katılımcıların sigara kullanma durumu

Katılımcıların sigara kulanım durumu Şekil 7.4'te gösterilmektedir. Katılımcıların %51.61'nin sigara kullandığı, %48,39'nun ise kullanmadığı saptanmıştır. Bu sonuca göre çalışmaya katılan bireyler arasında sigara kullanıp kullanmama durumları oranı eşit seviyede olduğu tespit edilmiştir.

Çizelge 7.2'de katılımcıların sigara kullanım sürelerine ait bilgi verilmektedir.

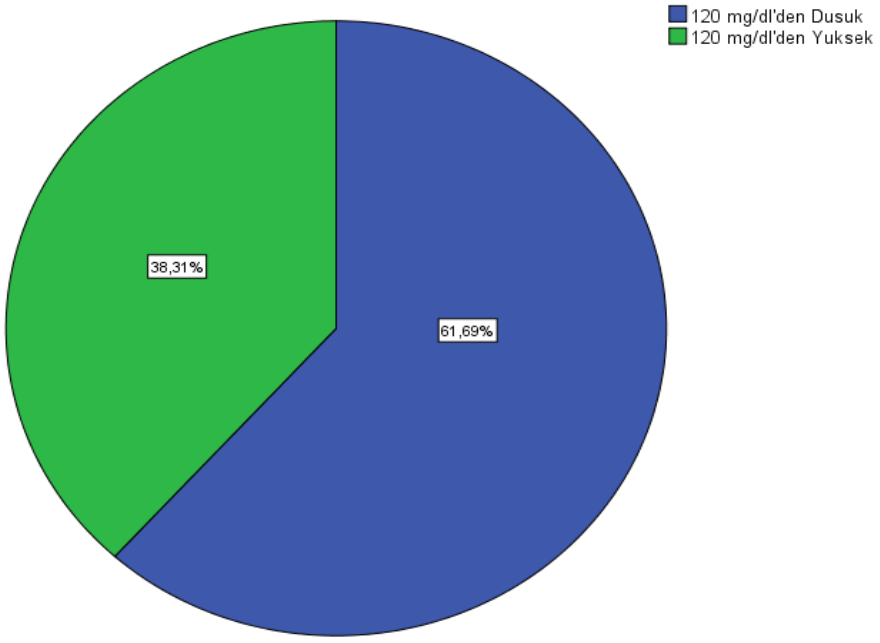
Çizelge 7.2. Katılımcıların sigara kullanım süreleri

Örneklem	243
Ortalama Değer	20,94
Minimum Değer	0
Maxsimum Değer	55



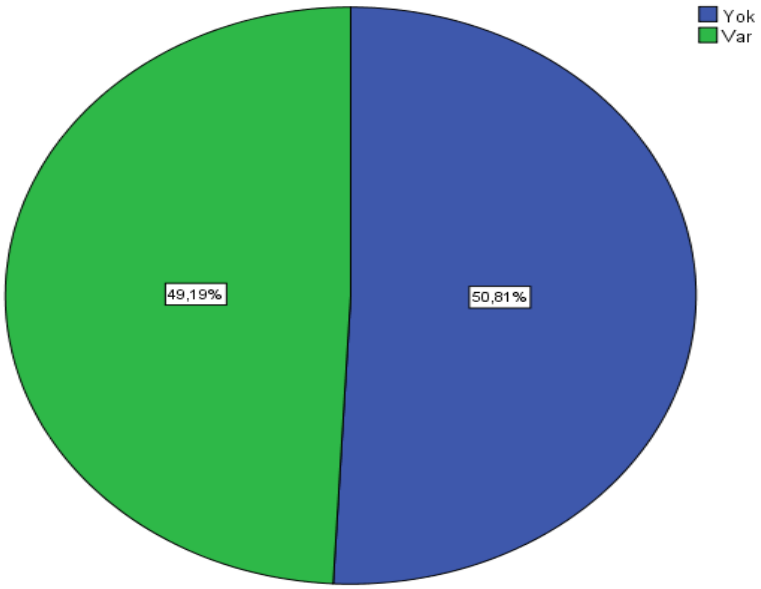
Şekil 7.5: Katılımcıların sigara kullanım süresi

Katılımcıların kaç yıldır sigara içtikleri Şekil 7.5'te görülmektedir. Katılımcıların ortalama sigara kullanım süresi 21 yıl olduğu tespit edilmiştir. Burada en çok sigara kullanan kişinin ise 55 yıldır içtiği saptanmıştır.



Şekil 7.6: Katılımcıların açlık kan şekeri miktarı

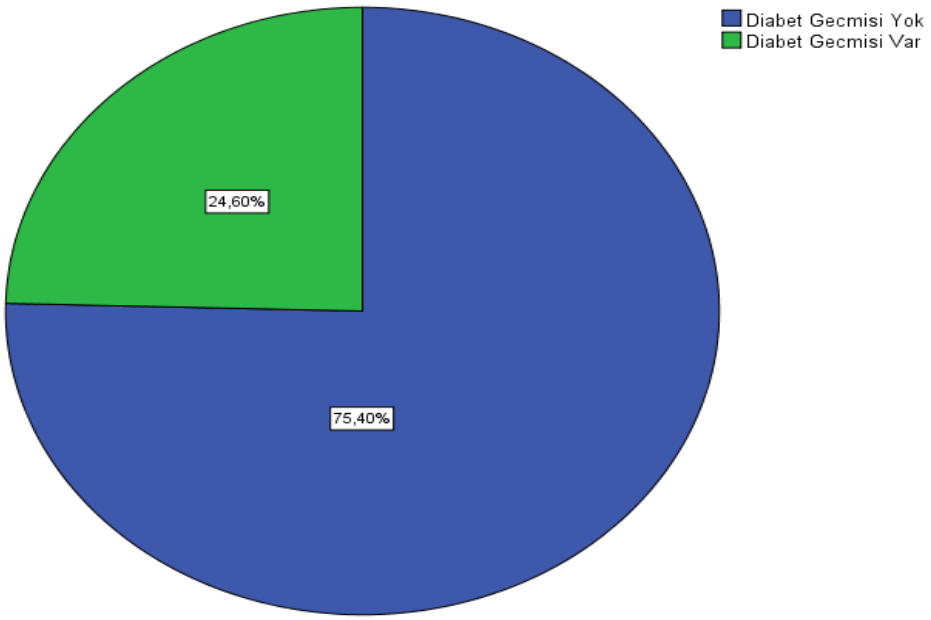
Katılımcıların açlık kan şekeri miktarı Şekil 7.6'da görülmektedir. % 61,69'nun kan şekeri miktarının 120mg/dl'den düşük olduğu, % 38.31'nin ise yüksek olduğu saptanmıştır.



Şekil 7.7: Katılımcıların aile geçmişinde kalp hastalığı durumu

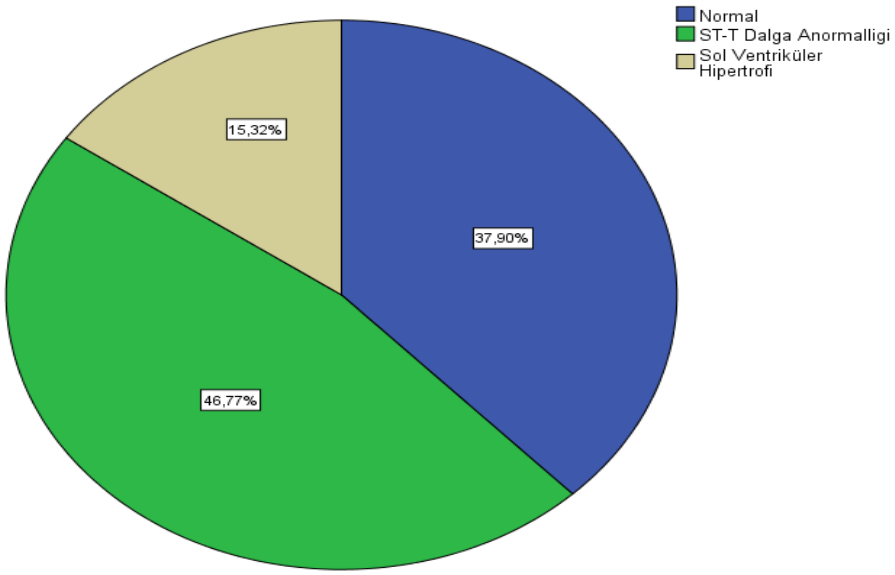
Katılımcıların aile geçmişinde kalp hastalığı teşhisi bulunup bulunmadığı Şekil 7.7 da görülmektedir. Dağılıma göre % 50,81'nin aile geçmişinde kalp hastalığı

bulunmaktadır. Bu sonuca göre çalışmaya katılan bireyler arasında aile geçmişinde kalp hastalığı bulunup bulunmama oranı eşit seviyede olduğu tespit edilmiştir.



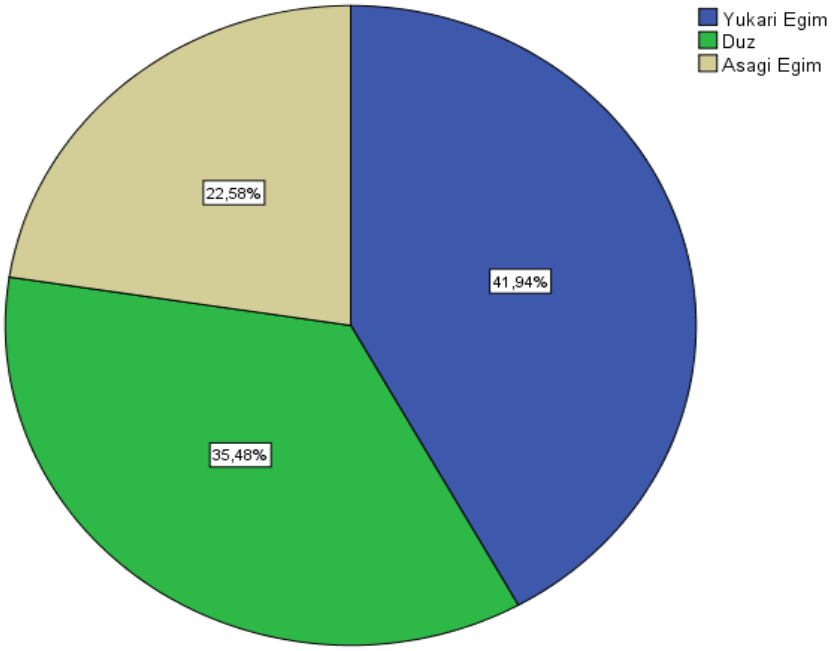
Şekil 7.8: Katılımcıların diabet geçmişi durumu

Katılımcıların diabetik geçmişi durumu Şekil 7.8'de görülmektedir. Dağılıma göre bakıldığında büyük çoğunluğunda yani % 75,40'nın diabet geçmişinin olmadığı görülmektedir. %24.60'nın diabet geçmişinin olduğu gözlenmiştir.



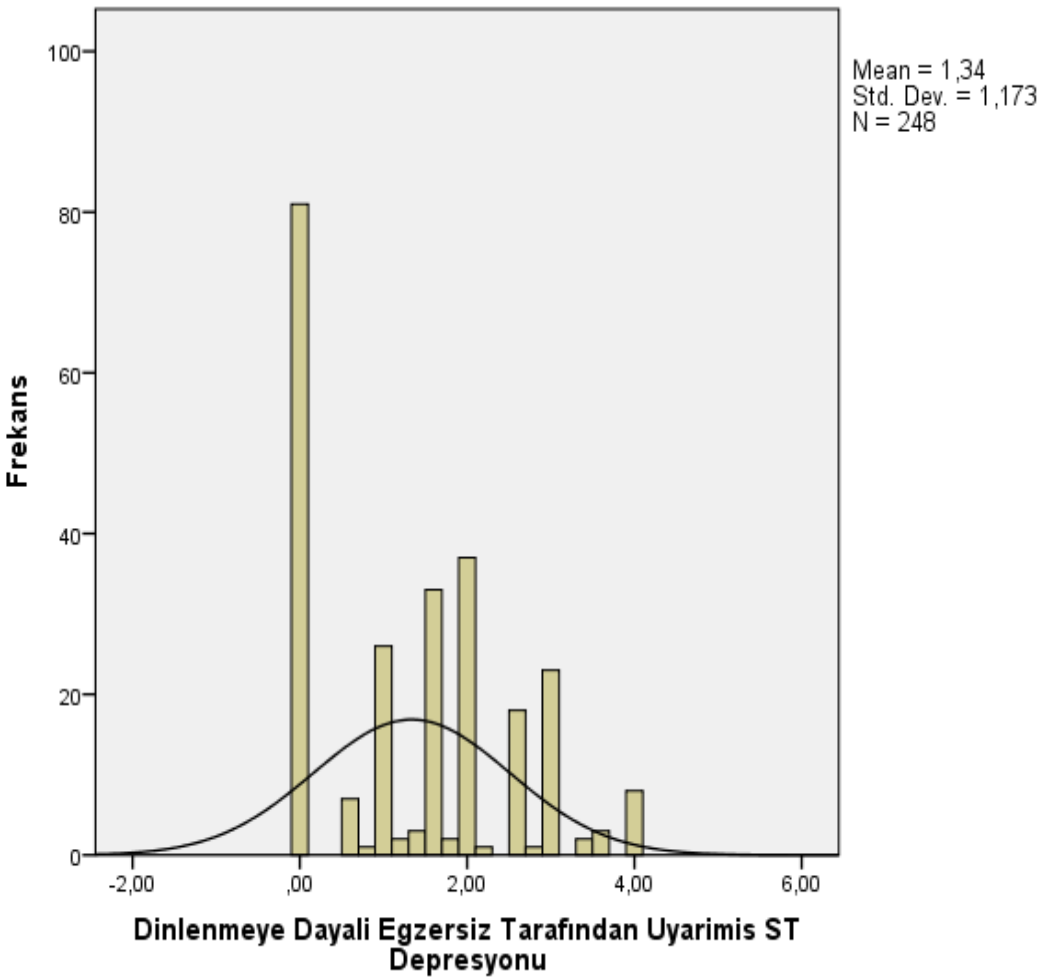
Şekil 7.9: Katılımcıların elektrokardiyografi sonucu

Katılımcıların elektrokardiyografi sonucu Şekil 7.9'da görülmektedir. Dağılıma göre bakıldığında büyük oranda % 46,77'si ST-T dalga anormalliği, % 37,90'nında sol ventriküler hipertrofi ve % 15,32'sinde normal sonuç olduğu gözlenmiştir.



Şekil 7.10: Katılımcıların ST segment egzersizinin zirve eğilimi

Katılımcıların ST Segment Egzersizinin Zirve Eğilimi Şekil 7.10'da görülmektedir. Dağılıma göre bakıldığında % 41,94'ü yukarı eğilim, % 35.38'i düz, % 22,58 oranında aşağı eğilim gözlenmiştir.



Şekil 7.11: Katılımcıların dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu

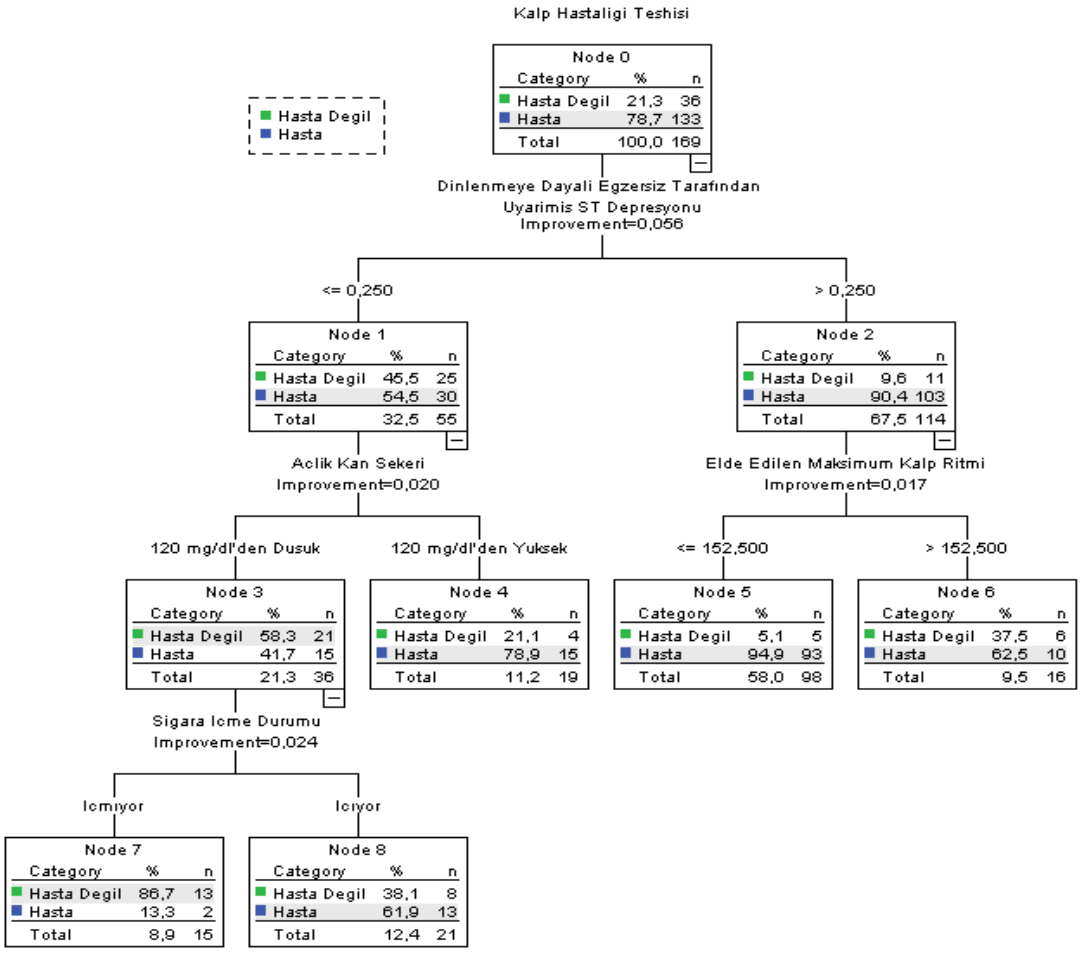
Katılımcıların Dinlenmeye Dayalı Egzersiz Tarafından Uyarılmış ST Depresyonu Şekil 7.11'de görülmektedir. Dağılıma göre bakıldığında ortalama ölçülen değer 1.34'tür. ST depresyon değerinin yükselmesi kalp rahatsızlığı durumunu etkilemektedir.

7.4.2. CART yöntemine ait bulgular

Sınıflandırma ağacında öncelikle en uygun ağacın hangi ağaç olduğuna karar vermek gerekmektedir. CART algoritması için de uygun ağacı belirleyebilmek adına üç farklı deneme yapılmıştır. Bu denemeler sonucu görülmüştür ki; ağacın yapısına en uygun dallanma modeli veri setinin %70'i çalışma örneklemi ve %30 'u test örneklemi alınarak oluşturulan dallanma olmuştur. Bu sonuca varılmadan önce; ilk olarak veri setinin % 70'i çalışma örneklemi ve %30'u da test örneklemi olarak alınmıştır. İkinci

olarak ise; veri setinin %50'si çalışma örneklemi, %50'si de test örneklemi olarak alınmıştır. Son olarak; veri setinin %30'u çalışma örneklemi ve %70'i de test örneklemi olarak alınmıştır. 3 faklı oranla oluşturulan modellerde, farklı düğüm sayıları ile uygun ağaç modeli belirlenmeye çalışılmıştır. CART algoritması ile kurulan modellerde Gini ayırma kriteri kullanılmıştır.

7.4.2.1. CART yöntemi ile oluşturulan sınıflama ağaçlarına ait 1. bulgular



Şekil 7.12: CART birinci ağacı diyagramı

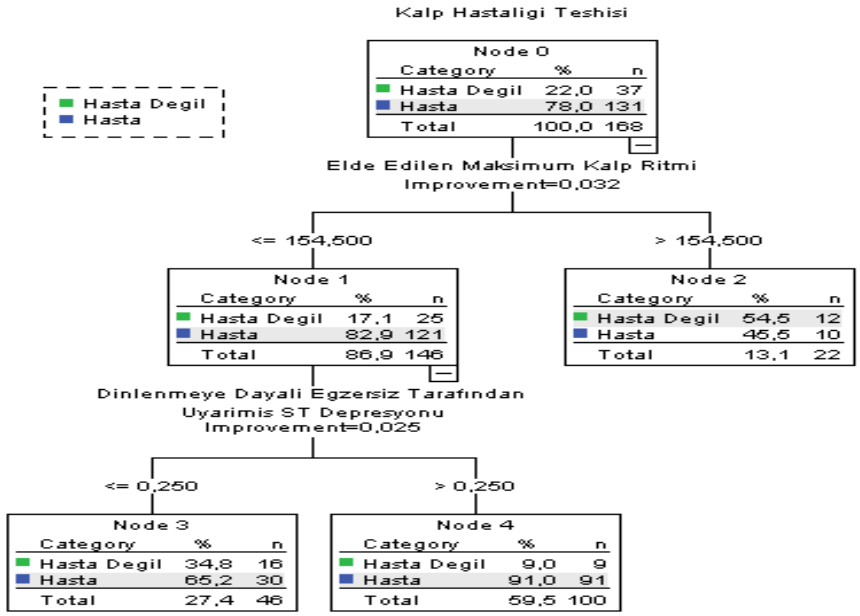
Şekil 7.12 'de CART algoritması ile kalp rahatsızlığını etkileyen değişkenler incelenmiştir.

Şekil 7.12 incelendiğinde kalp hastalığını etkileyen değişkenler; dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu, açlık kan şekeri, kalp ritmi sayısı ve

sigara içme durumudur. Çalışma sonucuna göre bu değişkenler arasında ilişki olduğu saptanmıştır. Kalp hastalıklarını etkileyen en önemli değişkenin ST depresyonu olduğu sonucuna varılmıştır. İkinci olarak önemli olan bağımsız değişken elde edilen maksimum kalp ritmi sayısı, üçüncü olarak önemli bağımsız değişkenin açlık kan şekeri oranı olduğu ve son etkileyen bağımsız değişkenin katılımcıların sigara içip içmeme durumu olduğu belirlenmiştir. Birinci düğüme incelediğimizde ST depresyon grafisi değeri 0.25'ten fazla olanların katılımcıların % 90.4 'ünün hasta, 0,25 ten düşük olanların ise % 54.5 'inin hasta olduğu görülmüştür. ST depresyon değeri düştükçe, kalp rahatsızlığı olanların oranının da düştüğü tespit edilmiştir.

ST depresyon değişkenini açlık kan şekeri ve kalp ritmi değişkenin etkilediği görülmektedir. Kan şekeri oranı 120 mg/dl 'den düşük olanların %41,7 'sinin kalp rahatsızlığı da bulunmaktadır. Kan şekeri oranının 120mg/dl 'den yüksek olan kişilerin % 78,9'u kalp rahatsızlığı teşhisi konulmuşken, maksimum kalp ritmi, 152 'den düşük olanların % 94,9'u kalp rahatsızlığı yaşarken, 152'den yüksek olanlarda ise oranın % 62.5 'e düştüğü tespit edilmiştir. Sigara kullananlarda ise oran % 61,9 'dur.

7.4.2.2. CART yöntemi ile oluşturulan sınıflama ağaçlarına ait 2. bulgular

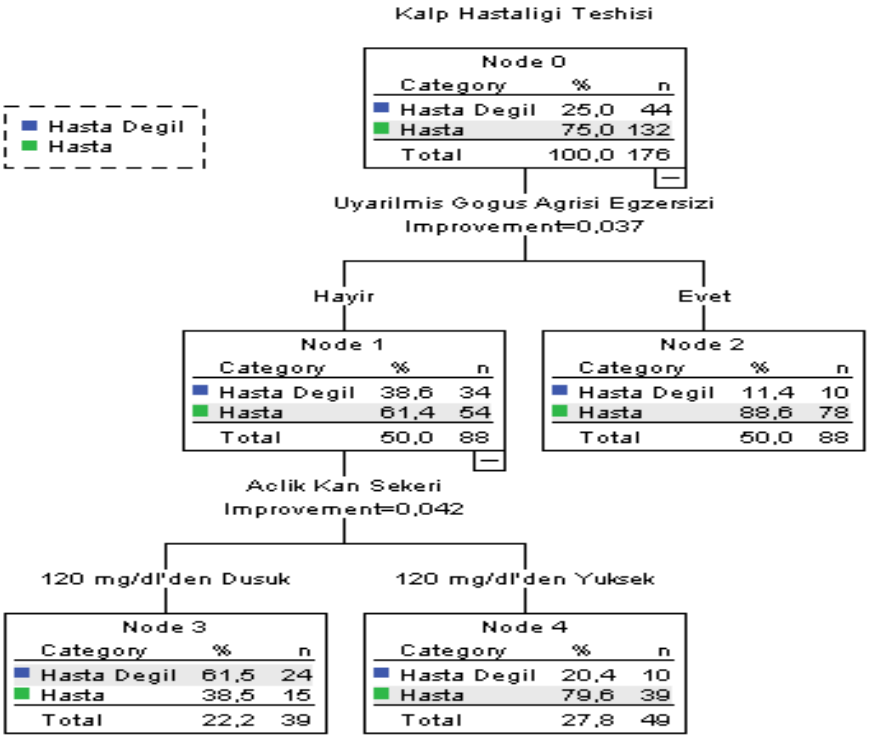


Şekil 7.13: CART ikinci ağacı diyagramı

Şekil 7.13'de CART algoritması ile kalp rahatsızlığını etkileyen değişkenler incelenmiştir.

Şekil 7.13 incelendiğinde kalp hastalığını etkileyen değişkenler; elde edilen kalp ritmi sayısı ve dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu durumudur. Çalışma sonucuna göre bu değişkenler arasında ilişki olduğu saptanmıştır. Kalp hastalıklarını etkileyen en önemli değişkenin kalp ritmi sayısı olduğu sonucuna varılmıştır. Kalp ritmi sayısı 154.5'ten az olanların %82.9'nun hasta, 154.5'ten fazla olanların ise % 45,5'nin hasta olmadı görülmüştür. Kalp ritmi değişkenini dinlenmeye dayalı egzersiz tarafından uyarınmış ST depresyonu değişkenini etkilediği görülmektedir ST depresyon değeri düştükçe, kalp rahatsızlığı olanların oranı da düşmekte olduğu, 0.25'ten büyük olduğu durumda katılımcıların % 91'nin hasta olduğu saptanmıştır.

7.4.2.3. CART yöntemi ile oluşturulan sınıflama ağaçlarına ait 3. bulgular

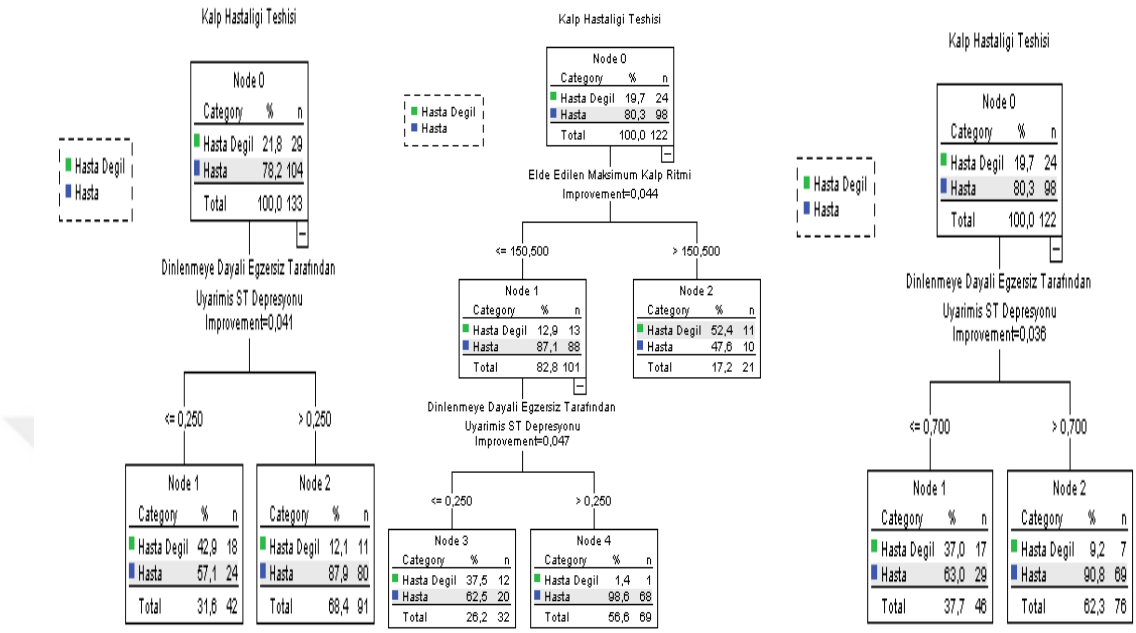


Şekil 7.14: CART üçüncü ağacı diyagramı

Şekil 7.14 ’de CART algoritması ile kalp rahatsızlığını etkileyen değişkenler incelenmiştir.

Şekil 7.14 incelendiğinde kalp hastalığını etkileyen değişkenler; uyarılmış göğüs ağrısı egzersizi ve açlık kan şekeri miktarıdır.. Çalışma sonucuna göre bu değişkenler arasında ilişki olduğu saptanmıştır. Kalp hastalıklarını etkileyen en önemli değişkenin uyarılmış göğüs ağrısı egzersizi olduğu sonucuna varılmıştır. Uyarılmış göğüs ağrısı olanların %88.6'sı hasta, göğüs ağrısı olmayanların %61,4'nün hasta olmadı görülmüştür. Uyarılmış göğüs ağrısı egzersizinin açlık kan şekeri oranını etkilediği görülmüştür. Kan şekeri oranının 120mg/dl 'den yüksek olan kişilerin %79,6 'sı kalp rahatsızlığı teşhisi konulmuşken, 120mg/dl 'den düşük olan kişilerin %38,5'nin hasta olmadığı tespit edilmiştir.

7.4.2.4. CART yöntemi ile oluşturulan sınıflama ağaçlarına ait 4. bulgular



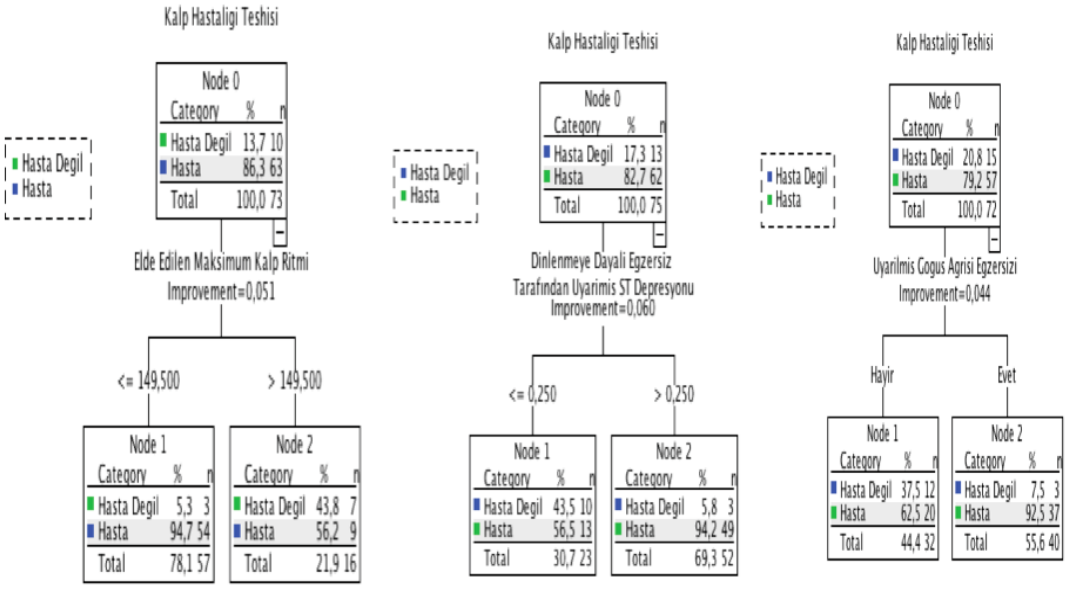
Şekil 7.15: CART dördüncü ağacı diyagramı

Şekil 7.15'de CART algoritmasına ait veri setinin %50'si çalışma örneklemi, %50'si de test örneklemi olarak alınmış olup 3 farklı oranla oluşturulan modellerde farklı düğüm sayıları ile uygun ağaç belirlenmeye çalışılmıştır ve kalp rahatsızlığını etkileyen değişkenler incelenmiştir.

Şekil 7.15 incelendiğinde veri setinin %50'si çalışma örneklemi, %50'si de test örneklemi olarak alındığında 3 farklı oranla da oluşturulsa iyi sonuç vermediğini görülmektedir.

Biri hariç diğer iki şekilde aynı sonuca ulaşıldığı saptanmıştır. Burada iki ağaç modelinde kalp hastalığını etkileyen en önemli değişken dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu çıkarken diğer ağaç modelinde ise elde edilen maksimum kalp ritmi sayısı çıkmıştır. İki ağaç modelinde sadece bir defa dallanırken diğer ağaç modelimi iki defa dallandığı görülmektedir.

7.4.2.5. CART yöntemi ile oluşturulan sınıflama ağaçlarına ait 5. bulgular



Şekil 7.16: CART beşinci ağacı diyagramı

Şekil 7.16'da CART algoritmasına ait veri setinin %30'si çalışma örneklemi, %70'si de test örneklemi olarak alınmış olup 3 farklı oranla oluşturulan modellerde farklı düğüm sayıları ile uygun ağaç belirlenmeye çalışılmıştır ve kalp rahatsızlığını etkileyen değişkenler incelenmiştir.

Şekil 7.16 incelendiğinde veri setinin %30'u çalışma örneklemi, %70'i de test örneklemi olarak alındığında 3 farklı oranla da oluşturulsa iyi sonuç vermediğini görülmektedir. Üç Modelde de ağacın sadece bir defa dallandığı görülmüştür

7.4.2.6. CART modellerini karşılaştırılması

Yapılan çalışmalar sonucunda 9 adet CART modeli oluşturulmuş olup, bu modeller arasında en iyisinin birinci ağaç model olduğu tespit edilmiştir.

İlk grup; veri setinin % 70'i çalışma örneklemi ve % 30'u test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli dört defa dallanmıştır.
İkinci grup; veri setinin % 50'si çalışma örneklemi ve % 50'si test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli sadece bir tanesi 2 defa dallanırken diğer iki ağaç modeli bir defa dallanmıştır.

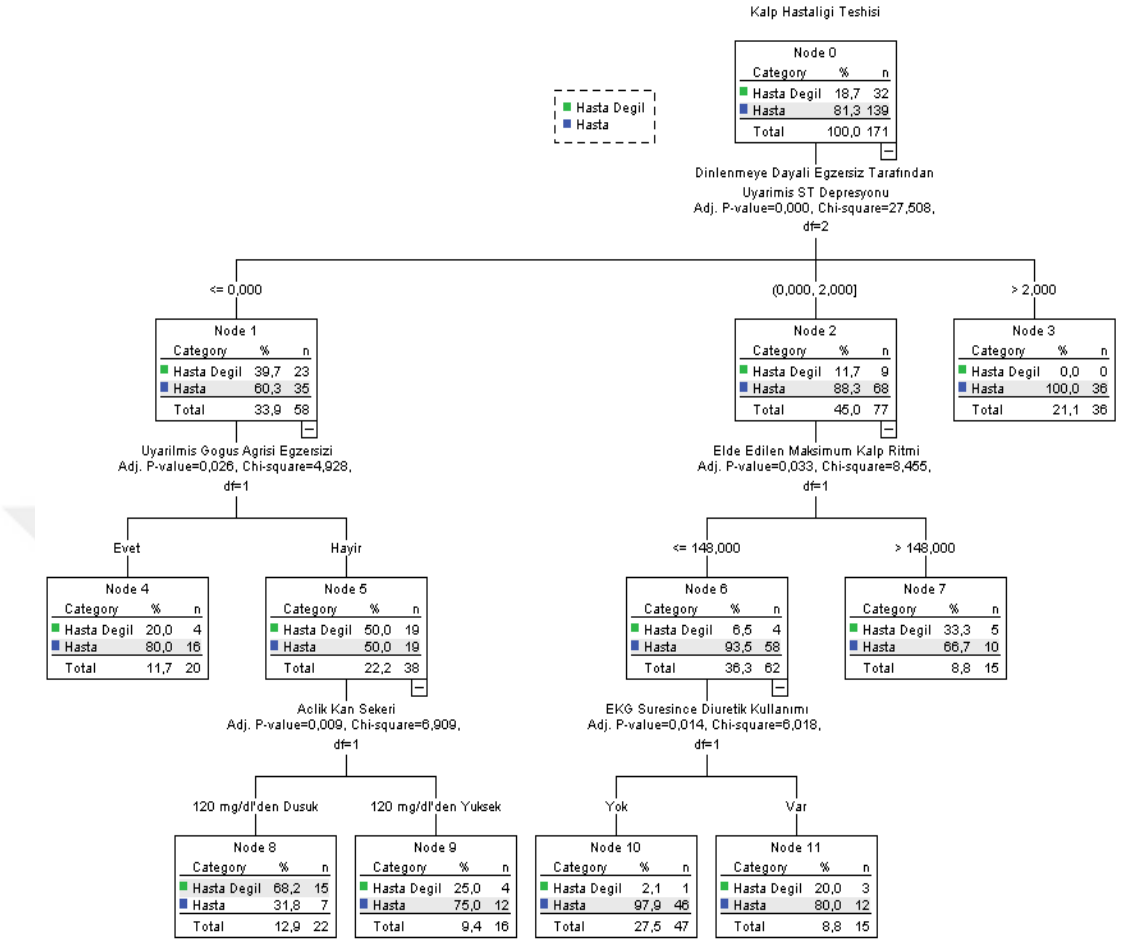
Üçüncü grup; veri setinin % 30'u çalışma örneklemi ve % 70'i test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli bir defa dallanmıştır.

Elde edilen sonuçlar ışığında; birinci gruptaki birinci ağaç diyagramı çalışmanın temelini oluşturan ağaç modelidir.

7.4.3. CHAID yöntemine ait bulgular

Sınıflandırma ağacında öncelikle en uygun ağacın hangi ağaç olduğuna karar vermek gerekmektedir. CHAID algoritması için de uygun ağacı belirleyebilmek adına üç farklı deneme yapılmıştır. Bu denemeler sonucu görülmüştür ki; ağacın yapısına en uygun dallanma modeli veri setinin %70'i çalışma örneklemi ve %30 'u test örneklemi alınarak oluşturulan dallanma olmuştur. Bu sonuca varılmadan önce; ilk olarak veri setinin % 70'i çalışma örneklemi ve %30'u da test örneklemi olarak alınmıştır. İkinci olarak ise; veri setinin %50'si çalışma örneklemi, %50'si de test örneklemi olarak alınmıştır. Son olarak; veri setinin %30'u çalışma örneklemi ve %70'i de test örneklemi olarak alınmıştır. 3 faklı oranla oluşturulan modellerde, farklı düğüm sayıları ile uygun ağaç modeli belirlenmeye çalışılmıştır.

7.4.3.1. CHAID yöntemi ile oluşturulan sınıflama ağaçlarına ait 1. bulgular



Şekil 7.17: CHAID birinci ağacı diyagramı

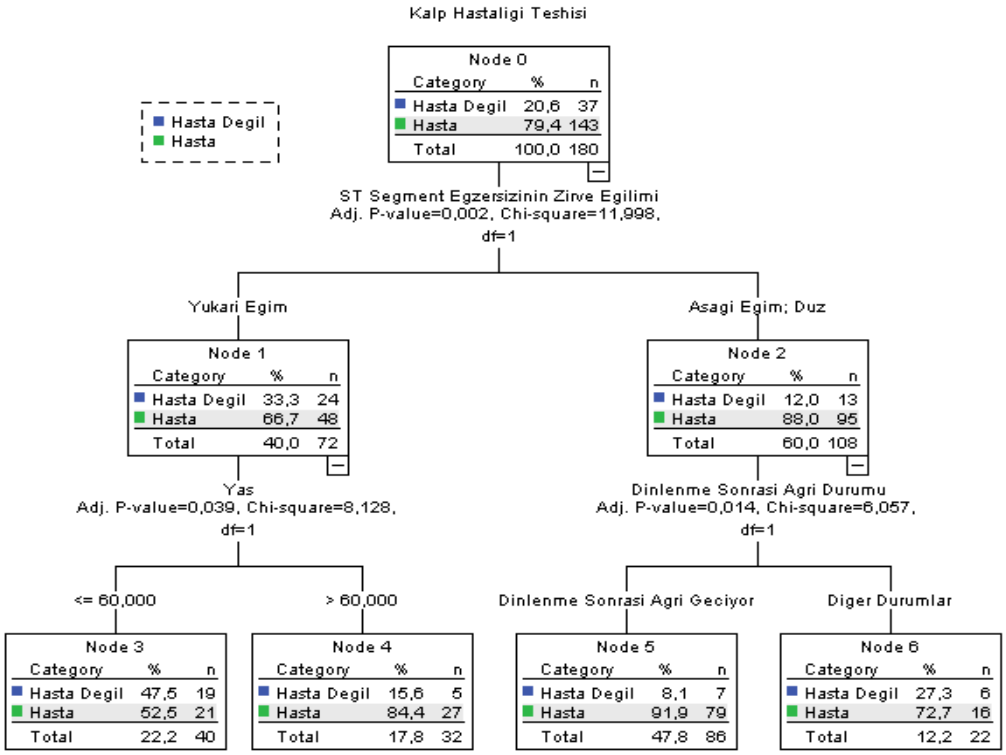
Şekil 7.17 incelendiğinde; kalp hastalığını etkileyen değişkenler; Dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu, elde edilen maksimum kalp ritmi, uyarılmış göğüs ağrısı egzersizi, açlık kan şekeri ve EKG süresince diüretik (idrar söktürücü) kullanımı ve bu değişkenler arasında ilişki olduğu sonucuna varılmıştır.

Kalp hastalığını etkileyen en önemli değişkenin ST depresyonu ($p=0.00$, $\chi 2=27.508$, $sd=2$) olduğu sonucuna varılmıştır. Bu değişken üç düğüm olarak ayrılmıştır. Bu düğümlerden en önemlisi üçüncü düğümdür. ST depresyonu değeri 2 'den yüksek olan kişilerin % 100 'ünün hasta olduğu görülmüştür. Bu sayı azaldıkça hasta sayısının da azaldığı tespit edilmiştir.

ST depresyonu değişkenini uyarılmış göğüs ağrısı egzersizi ve elde edilen maksimum kalp ritmi sayısı değişkenin etkilediği görülmüştür. ST depresyonu değeri (0-2

arasında olan kişi), değişkeni elde edilen maksimum kalp ritmi değişkenini (p=0,033 χ2=8.455, sd=1) etkilemektedir. Maksimum kalp ritmi 148/dk'dan az olan kişilerde kalp rahatsızlığı olan hasta oranı % 93,5 olarak gösterilmiştir. 148 'den fazla olanlarda ise hasta oranı % 66,7'dir. Ayrıca maksimum kalp ritmi değişkeniyle EKG süresince diüretik kullanımı (p=0,014 χ2=6.018, sd=1) değişkeniyle ilişkili olduğu saptanmıştır. EKG süresince diüretik kullanmayanların % 97,9'nun hasta olduğu görülmüştür.

7.4.3.2. CHAID yöntemi ile oluşturulan sınıflama ağaçlarına ait 2. bulgular



Şekil 7.18: CHAID ikinci ağacı diyagramı

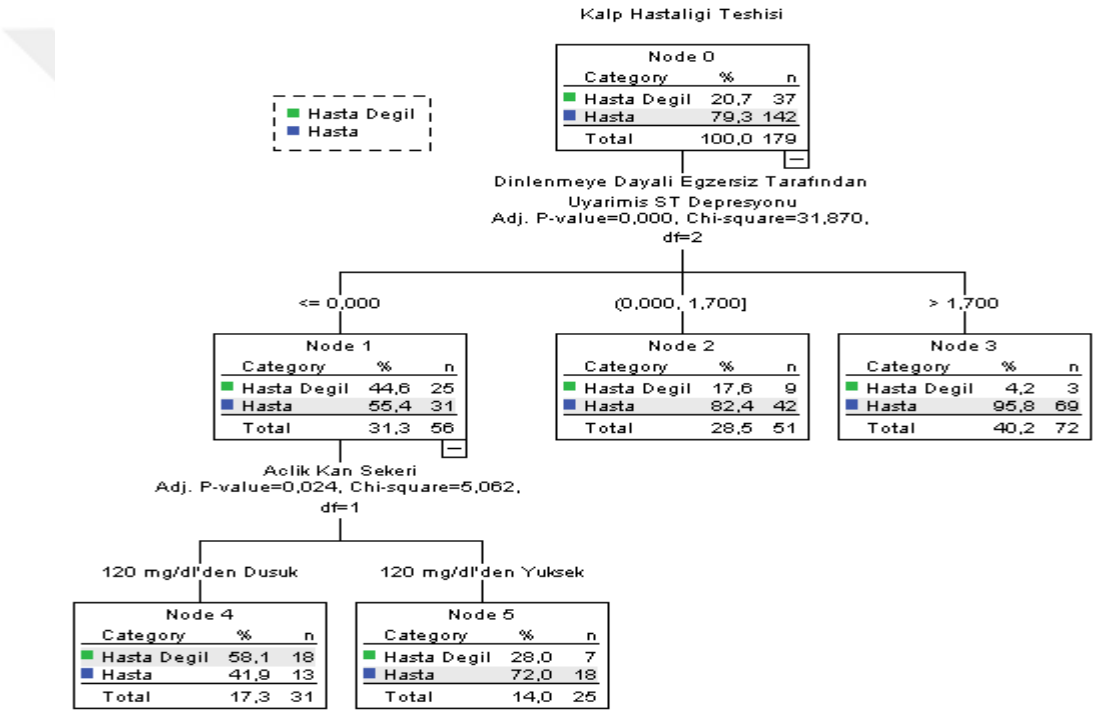
Şekil 7.18 incelendiğinde; kalp hastalığını etkileyen değişkenler; Dinlenmeye ST segment egzersizinin zirve eğilimi, yaş, dinlenme sonrası ağrı durumu ve bu değişkenler arasında ilişki olduğu sonucuna varılmıştır.

Kalp hastalığını etkileyen en önemli değişkenin Dinlenmeye ST segment egzersizinin zirve eğilimi (p=0.002, χ2=11.998, sd=2) olduğu sonucuna varılmıştır. Bu değişken iki düğüm olarak ayrılmıştır. Bu düğümlerden en önemlisi Dinlenmeye ST segment

egzersizinin zirve eğilimi düz olan kişilerin % 88 'nin hasta olduğu görülmüştür. Bu eğilim yukarı yönlü olduğunda hasta sayısının da azaldığı tespit edilmiştir.

Zirve eğilimi düz olan kişilerle dinlenme sonrası ağrı değişkenini (p=0,014 $\chi 2$=6,057, sd=1) etkilemektedir. Dinlenme sonrası ağrı durum geçen kişilerde kalp rahatsızlığı olan hasta oranı %91,9 olarak görülmüştür. Diğer durumlarda ise hasta olma oranı 5 72,7'dir. Ayrıca yaş değişkeni 60'ı geçkin olan kişilerde hasta olma oranı %84,4 olarak hesaplanmıştır.

7.4.3.3. CHAID yöntemi ile oluşturulan sınıflama ağaçlarına ait 3. bulgular



Şekil 7.19: CHAID üçüncü ağacı diyagramı

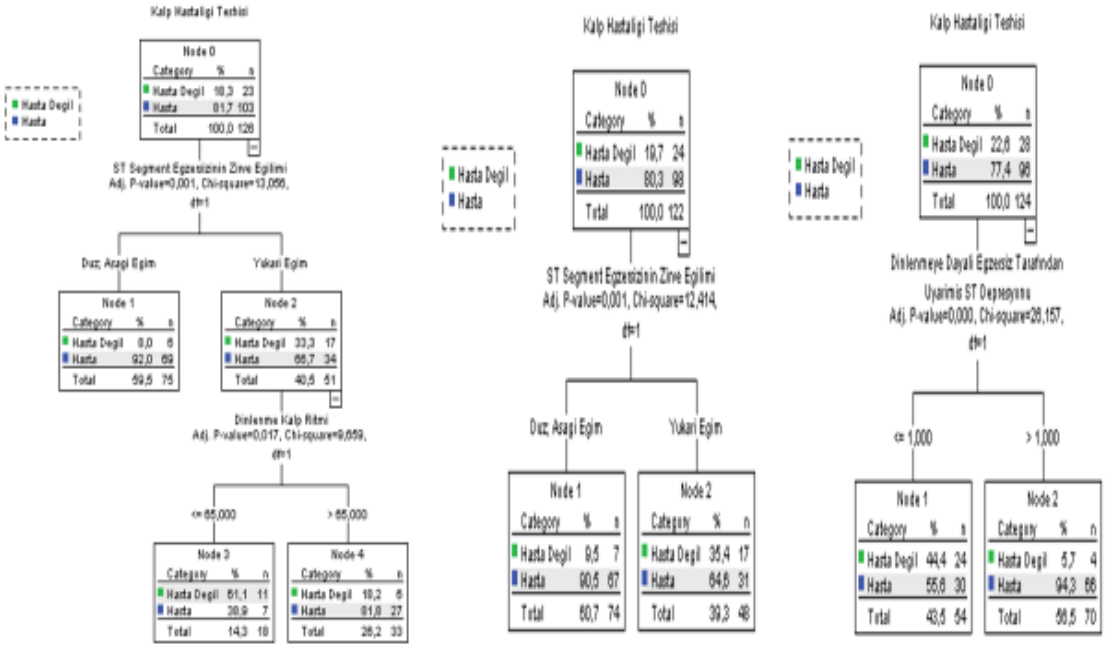
Şekil 7.19 incelendiğinde; kalp hastalığını etkileyen değişkenler; Dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu, açlık kan şekeri oranı ve bu değişkenler arasında ilişki olduğu sonucuna varılmıştır.

Kalp hastalığını etkileyen en önemli değişkenin ST depresyonu (p=0.00, $\chi 2$=31.870, sd=2) olduğu sonucuna varılmıştır. Bu değişken üç düğüm olarak ayrılmıştır. Bu düğümlerden en önemlisi ST depresyonu değeri 1,7'den yüksek olan kişilerin

%95,8'nin hasta olduğu görülmüştür. Bu sayı azaldıkça hasta sayısının da azaldığı tespit edilmiştir.

ST depresyonu değeri (0'dan az olan kişi), değişkeni açlık kan şekeri değişkenini (p=0,024 $\chi 2$=5.062, sd=1) etkilemektedir. Açlık kan şekeri 120mg/dl'den yüksek olan kişilerin hasta olması oranı %72,0 olarak hesaplanmıştır. 120mg/dl'den düşük olan kişilerde ise hasta olma oranı %41,9 olarak hesaplanmıştır.

7.4.3.4. CHAID yöntemi ile oluşturulan sınıflama ağaçlarına ait 4. bulgular



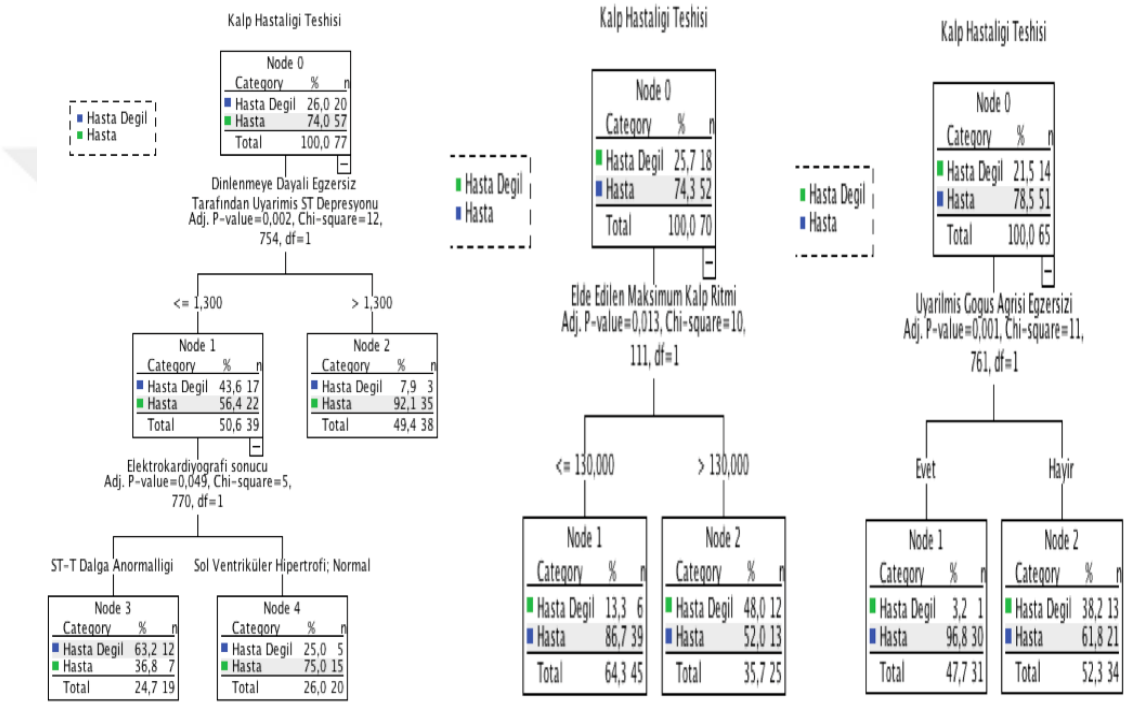
Şekil 7.20: CHAID dördüncü ağacı diyagramı

Şekil 7.20'de CHAID algoritmasına ait veri setinin %50'si çalışma örneklemi, %50'si de test örneklemi olarak alınmış olup 3 farklı oranla oluşturulan modellerde farklı düğüm sayıları ile uygun ağaç belirlenmeye çalışılmıştır ve kalp rahatsızlığını etkileyen değişkenler incelenmiştir.

Şekil 7.20 incelendiğinde veri setinin %50'si çalışma örneklemi, %50'si de test örneklemi olarak alındığında 3 farklı oranla da oluşturulsa iyi sonuç vermediğini görülmektedir.

Biri hariç diğer iki şekilde aynı sonuca ulaşıldığı saptanmıştır. Burada iki ağaç modelinde kalp hastalığını etkileyen en önemli değişken ST segment egzersizinin zirve eğilimi çıkarken diğer ağaç modelinde ise dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu miktarı çıkmıştır. İki ağaç modelinde sadece bir defa dallanırken diğer ağaç modelimi iki defa dallandığı görülmektedir.

7.4.3.5. CHAID yöntemi ile oluşturulan sınıflama ağaçlarına ait 5. bulgular



Şekil 7.21: CHAID beşinci ağacı diyagramı

Şekil 7.21'de CART algoritmasına ait veri setinin % 30'u çalışma örneklemi, % 70'i de test örneklemi olarak alınmış olup 3 farklı oranla oluşturulan modellerde farklı düğüm sayıları ile uygun ağaç belirlenmeye çalışılmıştır ve kalp rahatsızlığını etkileyen değişkenler incelenmiştir.

Şekil 7.21 incelendiğinde veri setinin % 30'u çalışma örneklemi, % 70'i de test örneklemi olarak alındığında 3 farklı oranla da oluşturulsa iyi sonuç vermediğini görülmektedir. Görülmüştür ki sadece bir ağaç modeli 2 defa dallanmıştır. Diğer ağam modelleri bir defa dallanmıştır. Modelleri incelediğimizde Üç farklı sonuç ortaya çıkmıştır. Kalp hastalığını etkileyen değişkenler; dinlenmeye dayalı egzersiz

tarafından uyarılmış ST depresyonu, elde edilen maksimum kalp ritmi ve uyarılmış göğüs ağrısı egzersizi değişkenleri olduğu saptanmıştır.

7.4.3.6. CHAID modellerini karşılaştırılması

Yapılan çalışmalar sonucunda 9 adet CHAID modeli oluşturulmuş olup, bu modeller arasında en iyisinin birinci ağaç model olduğu tespit edilmiştir.

İlk grup; veri setinin %70'i çalışma örneklemi ve %30 'u test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli beş defa dallanmıştır.

İkinci grup; veri setinin %50'i çalışma örneklemi ve %50 'u test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli sadece bir tanesi 2 defa dallanırken diğer iki ağaç modeli bir defa dallanmıştır.

Üçüncü grup; veri setinin %30'i çalışma örneklemi ve %70 'u test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli bir defa dallanmıştır.

Elde edilen sonuçlar ışığında; birinci gruptaki birinci ağaç diyagramı çalışmanın temelini oluşturan ağaç modelidir.

8. SONUÇ

Bu çalışmada veri madenciliğinin bir kolu olan karar ağaçları yönteminin algoritmalarından olan CART ve CHAID yöntemleri karşılaştırılmıştır.

CART yöntemi bağımlı değişkenin kategorik veya sürekli olduğu durumda da kolaylıkla kullanılabilen parametrik olmayan istatistiksel bir yöntemdir. CHAID algoritması veri seti çok karmaşık olsa bile bağımlı değişkeni etkileyen bağımsız değişkenleri ve bu değişkenleri önem sırasına göre ağaç şeklinde ortaya koymaktadır.

Bu çalışmada amaç, kalp hastalığına etki eden 38 faktörün CART ve CHAID yöntemleriyle incelenerek hangi yöntemin diğerinden daha doğru bir sınıflama yapacağını farklı başlangıç verileri kullanarak test etmeye çalışmak. Yani kısaca en başarılı sınıflama modelini oluşturmaya çalışmaktır. Ayrıca model verisinin %70, %50 ve %30'luk kısmı dahil edilerek oluşturulmuş ve oluşturulan ağaçlar karşılaştırmalı olarak yorumlanmıştır.

Bu çalışma kapsamı doğrultusunda, University of California bünyesinde veri setlerini barındıran bir platformdan alınan ve kalp hastalığına etki eden faktörlerin yaşları 37 ila 77 arasında değişen 248 kişiye uygulanan veri seti kullanılmıştır. Bu çalışma kapsamında kişilere kalp hastalığına etki eden yaş, cinsiyet, tansiyon, sigara kullanımı, diyabet ve ilaç kullanımı faktörler incelenmiştir.

CART ve CHAID algoritması ile oluşturulmuş sınıflama ağaçları incelendiğinde Kalp hastalığını etkileyen en önemli değişkenin değişmediğini görülmüştür. Her iki algoritmada da kalp hastalığını tetikleyen bağımsız değişkenler ağaç modellerin genelinde aynı değişkenler çıkmıştır. Bu değişkenler, açlık kan şekeri, elde edilen maksimum kalp ritmi, sigara içme durumu ve uyarılmış göğüs ağrısı egzersizi gibi değişkenler olmuştur.

Çalışma örneklemi arttıkça kalp hastalığını etkileyen değişken sayısı ve düğüm sayısı da artmaktadır. Düğüm ve değişken sayıları arttıkça ağaç modeli daha karmaşık bir yapıya bürünmektedir. Dolayısıyla, kategoriler yani sınıflar hakkında daha ayrıntılı bilgiye ulaşılmaktadır. Sınıflar arasındaki etkileşimlerde daha iyi görülebilmektedir.

Bu durumda da araştırmacı analiz ettiği veriye ait daha detaylı bilgiye sahip olacağından model verisinin yüksek tutulması avantajlıdır.

Çalışmanın sonucunda CART algoritması ile oluşturulan ağaçları modelleri incelendiğinde şu sonuçlara varılmıştır:

En iyi sonucu veren modelin, veri setinin % 70'i çalışma örneklemi ve % 30'u test örneklemi alınarak oluşturulan model olduğu tespit edilmiştir. İkinci olarak ise veri setinin % 50'si çalışma örneklemi, % 50'si de test örneklemi olarak alınmış model. Son olarak; veri setinin % 30'u çalışma örneklemi ve % 70'i de test örneklemi olarak alınan modeller olduğu görülmüştür.

Yapılan çalışmalar sonucunda dokuz adet CART modeli oluşturulmuş olup, bu modeller arasında en iyisinin birinci ağaç model olduğu tespit edilmiştir.

İlk grup veri setinin % 70'i çalışma örneklemi ve % 30'u test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli dört defa dallanmıştır.

İkinci grup veri setinin % 50'si çalışma örneklemi ve % 50'si test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli sadece bir tanesi 2 defa dallanırken diğer iki ağaç modeli bir defa dallanmıştır.

Üçüncü grup veri setinin % 30'u çalışma örneklemi ve % 70'i test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli bir defa dallanmıştır.

CHAID algoritması ile oluşturulan regresyon ağaçları incelendiğinde şunlar elde edilmiştir:
En iyi sonucu veren modelin, veri setinin % 70'i çalışma örneklemi ve %30 'u test örneklemi alınarak oluşturulan model olduğu tespit edilmiştir. İkinci olarak ise veri setinin % 50'si çalışma örneklemi, % 50'si de test örneklemi olarak alınmış model. Son olarak, veri setinin % 30'u çalışma örneklemi ve % 70'i de test örneklemi olarak alınan modeller olduğu görülmüştür.

CHAID algoritması ile oluşturulan ağaca göre kalp hastalığını etkileyen değişkenler; dinlenmeye dayalı egzersiz tarafından uyarılmış ST depresyonu, elde edilen maksimum kalp ritmi sayısı, uyarılmış göğüs ağrısı egzersizi, açlık kan şekeri ve EKG

süresince diüretik kullanımıdır. Bu değişkenler arasında ilişki olduğu sonucuna varılmıştır.

Yapılan çalışmalar sonucunda dokuz adet CHAID modeli oluşturulmuş olup, bu modeller arasında en iyisinin birinci ağaç model olduğu tespit edilmiştir.

İlk grup veri setinin % 70'i çalışma örneklemi ve %30 'u test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli beş defa dallanmıştır.

İkinci grup veri setinin % 50'si çalışma örneklemi ve %50 'si test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli sadece bir tanesi iki defa dallanırken diğer iki ağaç modeli bir defa dallanmıştır.

Üçüncü grup veri setinin % 30'u çalışma örneklemi ve % 70'i test örneklemi alınarak oluşturulmuştur. Oluşturulan ağaç modeli bir defa dallanmıştır.
Elde edilen sonuçlar ışığında, birinci gruptaki birinci ağaç diyagramı çalışmanın temelini oluşturan ağaç modelidir.

Sonuç olarak, CART ve CHAID algoritmalarının ağacı oluşturma ve geliştirme yöntemleri birbirinden farklı olduğundan, bu iki ağaç arasında birtakım farklılıklar görülebilmektedir. CHAID algoritmasında düğümler ikiden fazla sınıflara ayrılabilirken, CART algoritmasında sınıflar sadece ikili olarak ayrılabilmektedir. Dolayısıyla sınıflandırma ağacında CHAID algoritması daha kapsamlı sonuçlara ulaşmayı sağlarken, CART algoritmasında daha genel sonuçlar elde edilebilmektedir.

KAYNAKLAR

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İÜ İşletme Fakültesi Dergisi, 29*(1), 1-22.

Atılgan E. S. (2011), Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi İle İncelenmesi. Yüksek Lisans Tezi. Hacettepe üni. Fen Bilimleri Enstitüsü.

Antipov, Evgeny ve Elena Pokryshevskaya. "Applying CHAID for logistic regression diagnostics and classification accuracy improvement", Journal of Targeting, Measurement and Analysis for Marketing, 2010, Vol.18, No.2, ss.109-117.

Arslan, H., "Sakarya Üniversitesi Web Sitesi Erişim Kayıtlarının Web Madenciliği İle Analizi", Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya, 3-34 (2008).

Aydın D.E. (1992). Karar Ağacı ve Cobol, Doruk Yayınları Ankara, 1. Baskı.

Bahar Y, Y., 2017. Yaşam memnuniyetini Etkileyen Faktörlerin Sınıflma ve Regresyon Ağacı İle Belirlenmesi, İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul

Bayram, Y., 2002. Menemen İlçesinde 35-64 Yaş Grubunda Koroner Kalp Hastalıkları Risk Faktörleri Sıklığının Araştırılması, Ege Üniversitesi, Sağlık Bilimleri Enstitüsü, Doktora Tezi, İzmir

Betül, Ş., 2015. Çocuklarda Görülen Kalp Hastalıklarında Epidemiyolojik Risk Faktörlerinin Belirlenmesi, Dokuz Eylül Üniversitesi, Tıp Fakültesi Çocuk Sağlığı Ve Hastalıkları Anabilim Dalı, Uzmanlık Tezi, İzmir

Bevilacqua, M., Braglia, M., & Montanari, R. (2003). The Classification and Regression Tree Approach to Pump Failure Rate Analysis. Reliability Engineering & System Safety, 79(1), 59-67.

Breiman ve Diğerleri, Classification and Regression Trees, New York:Chapman & Hall, 1993, s.228.

Breiman, L., Freidman, J. H., Olshen, R. A., & Stone, C. J. (1998). Classification And Regression Trees. Londra: Chapman&Hall.

Buchan, I. (2002). *Calculating the Gini coefficient of inequality*. Mayıs 15, 2017 tarihinde Northwest Institute for BioHealth Informatics: https://www. nibhi.org.uk/Training/Forms/AllItems.aspx adresinden alındı

Chipman, H., & McCulloch, R. E. (2000). Hierarchical Priors for Bayesian CART Shrinkage. *Statistics and Computing, 10*(1), 17-24.

Cumhur, M. (2001). Temel Anatomi (1.b.). Ankara: ODTÜ Yayıncılık.

Da Rosa, J. C., Veiga, A., & Medeiros, M. C. (2008). Tree-Structured Smooth Transition Regression Models. *Computational Statistics & Data Analysis, 52*(5), 2469-2488.

De'ath, G., & Fabricius, K. E. (2000). Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology, 81*(11), 3178-3192.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification.* New York: Wiley.

Dumlu, U., & Aydın, Ö. (2008). Ekonometrik Modellerle Türkiye İçin 2006 Yılı Gini Katsayısı Tahmini. *Ege Akademik Bakıs Dergisi, 8*(1), 375-395.

Fu, C. Y. (2004). Combining Loglinear Model with Classification and Regression Tree (Cart): An Application to Birth Data. *Computational Statistics & Data Analysis, 45*(4), 865-874.

Gey, S., & Nedelec, E. (2005). Model Selection for CART Regression Trees. *IEEE Transactions On Information Theory, 51*(2), 658-670.

Haughton, D., & Oulabi, S. (1997). Direct Marketing Modeling with CART and CHAID. *Journal of Interactive Marketing, 11*(4), 42-52.

Jorsal, A., Wiggers, H., McMurray, JJV. (2018). Endocrinology and metabolism clinics of North America. Volume 47, issue 1, 1p: 117- 135.

Karakayalı, H. (2005). *Makro Ekonomi* (5. b.). Manisa: Emek Matbaası.

Koyuncugil, Ali Serhan. "Araştırma Raporu" , Veri Madenciliği ve Sermaye Piyasalarına Uygulanması. Ankara: Sermaye Piyasası Kurulu Araştırma Dairesi, 2007.

Larose, D. T. (2005). *Discovering Knowledge in Data: An introduction to Data* Mining. New York: Wiley.

Lewis, R. J. (2000). An Introduction to Classification and Regression Tree (CART) Analysis. *Annual Meeting of the Society for Academic Emergency Medicine*, (s. 1-14). Kaliforniya.

Magidson, J. (1993). *SPSS.* Chicago.

Mahadevan, V. (2017). Anatomy of the heart surgery.

Murphy, N., Alderman, P., Harvey, KH., Harris, N. (2017). Women and heart disease: An evidence- based update. The journl of nurse practitioners.

Mrsic, Z., Hopkins, S., Antevil, J., Mullenix, P. (2018). Valvular heart disease. Elsevier clinics rewiev articles. Volume 45, issue 1, p: 81- 94.

Oğuş, A. (2004). Türkiye'de Ekonomik Büyüme ve Gelir Dağılımı. *IV. İktisat Kongresi* (s. 1-22). İzmir: İzmir Ekonomi Üniversitesi.

Özekes S. (2002). Veri Madenciliği Uygulaması. Yüksek Lisans Tezi, Marmara Üni. Fen Bilimleri Enstitüsü

Sedlak, T., Izadnegadhar M., Humpries KH. (2014). Sex-spesific factors in microvascular angina. Canadian Journal of cardiology: 747- 755. Elsevier.

Safavian S.R. Landgrebe D., 1991. A Survey of Decision Tree Classifier Methodology,.IEEE Transactions on Systems Man and Cybernetics, 21, 660-674.

Sha, N. (2002). Bolstering Cart and Bayesian Variable Selection Methods for Classification. *Doktora Tezi*. Texas A&M University Department of Statistics.

Sezer Ü. (2008). Karar Ağaçlarının Birliktelik Kuralları İle İyileştirilmesi

Yohannes, Y., & Hoddinott, J. (1999). *Classification and Regression Trees: an* Introduction. International Food Policy Research Institute.

Ulusoy, G., 2013. Karar Ağacı Analizi ile AB GeniĢleme Kriterlerinin Değerlendirilmesi. Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, Ġstatistik Bilim Dalı. Yüksek Lisans Tezi.

Quinlan, J.R., 1986. Induction of Decision Trees. Journal of Machine Learning, Cilt 1, s 81-106.

Yeliz, S. S., 2015. Sınıflama Ve Regresyon Ağaçları, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, İstanbul

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Onur KÖSE
Doğum yılı : 22.07.1989
Doğum yeri : Eminönü
Cinsiyet : Erkek
İletişim : onurrrkose@hotmail.com

Eğitim Durumu

Lise : Hüseyin Kalkavan Lisesi (2002-2006)
Lisans : İstanbul Ticaret Üniversitesi-İstatistik (Burslu)-(2007-2012)
Yüksek Lisans : İstanbul Ticaret Üniversitesi –İstatistik (Tezli) - Yüksek Lisans (2012- Halen)

İş Deneyimleri

İstanbul Ticaret Odası 2014- Halen

Staj Deneyimleri

İstanbul Ticaret Odası
İstatistik Şubesi (15-30 Temmuz 2010)

Türkiye İstatistik Kurumu
Operasyon Bölümü Stajyeri (15-30 Haziran 2011)
- Veri girişi
- Anket düzenleme Müşterilerle telefonda görüşülmesi
- Sahada anket çalışmaları

İstanbul Ticaret Üniversitesi

"Sosyal Güvenlik Kurumuna Bağlı ve Özel Hastanelerden Yararlanan Hastaların Sağlıkta Dönüşüm Projesine Bakış Açıları", 2010 Konulu Araştırma Stajyeri
- Hastanelerde Veri Toplanması
- Toplanan Verilerin Spss Programına Girilmesi

Bilgisayar Programları

- Microsoft Office Programları (iyi derecede)
- Spss Programı (iyi derecede)
- Eviews Programı (orta derecede)
- Sas Programı (orta derecede)
- Ncss Programı (orta derecede)
- R Software (orta derecede)
- Excel Programı (iyi derecede)
- Stata (orta derecede)

Sertifikalar

- Adana Çukurova Üniversitesi İstatistik Kolokyumu 2009 (2 gün)
- Afyon Kocatepe Üniversitesi İstatistik Kolokyumu 2010 (2 gün)
- İzmir Dokuz Eylül Üniversitesi İstatistik Kolokyumu 2011(2 gün)

Yabancı Dil Bilgisi

YÖKDİL: 48