YELİZ SEVİMLİ SAİTOĞLU

SINIFLAMA VE REGRESYON AĞAÇLARI

EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

İSTANBUL, 2015

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

SINIFLAMA VE REGRESYON AĞAÇLARI

Doktora Tezi

YELİZ SEVİMLİ SAİTOĞLU

İstanbul, 2015

T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

SINIFLAMA VE REGRESYON AĞAÇLARI

Doktora Tezi

YELİZ SEVİMLİ SAİTOĞLU

Danışman: PROF. DR. AHMET METE ÇİLİNGİRTÜRK

İstanbul, 2015

MARMARA ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜ

TEZ ONAY BELGESİ

EKONOMETRİ Anabilim Dalı İSTATİSTİK Bilim Dalı DOKTORA öğrencisi YELİZ SEVİMLİ SAİTOĞLU'nın SINIFLAMA VE REGRESYON AĞAÇLARI adlı tez çalışması, Enstitümüz Yönetim Kurulunun 23.06.2015 tarih ve 2015-22/20 sayılı kararıyla oluşturulan jüri tarafından oy birliği / ~~oy çokluğu~~ ile Doktora Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi ...13../..07../..2015...

Öğretim Üyesi Adı Soyadı			**İmzası**
1.	Tez Danışmanı	Prof. Dr. AHMET METE ÇİLİNGİRTÜRK	
2.	Jüri Üyesi	Prof. Dr. DİLEK ALTAŞ	
3.	Jüri Üyesi	Doç. Dr. ÜNAL ÖZDEN	
4.	Jüri Üyesi	Prof. Dr. ŞAHAMET BÜLBÜL	
5.	Jüri Üyesi	Prof. Dr. DİCLE CENGİZ	

GENEL BİLGİLER

İsim ve Soyadı	: Yeliz Sevimli Saitoğlu
Anabilim Dalı	: Ekonometri
Programı	: İstatistik
Tez Danışmanı	: Prof. Dr. Ahmet Mete Çilingirtürk
Tez Türü ve Tarihi	: Doktora - Temmuz 2015
Anahtar Kelimeler	: CART, MARS, Karar Ağaçları, Türkiye Gençliği ve Siyasi Görüşleri

ÖZET

SINIFLAMA VE REGRESYON AĞAÇLARI

CART yöntemi hem kategorik hem de sürekli değişkenleri kullanarak sınıflama ve regresyon problemlerinin çözümünde karar ağaçlarını kullanan parametrik olmayan istatistiksel bir yöntemdir. Ele alınan bağımlı değişken kategorik ise yöntem sınıflama ağaçları, sürekli ise regresyon ağaçları olarak adlandırılmaktadır. Aynı şekilde MARS yöntemi de nonparametrik ve doğrusal olmayan bir yöntem olup, hem sürekli hem de ikili bağımlı değişkenler için tasarlanmıştır. Bağımlı değişkeninin sürekli olması durumunda kestirim amaçlı olan bu yöntem, kategorik olması durumunda sınıflandırma amacına sahiptir. Her iki yöntemin ortak özelliği bağımlı değişkenin türüne göre hem sınıflama hem de tahmin modeli geliştirebilmesidir.

Bu çalışmada amaç Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde, yani parti tercihlerinde CART ve MARS yöntemlerini karşılaştırıp, uygulamada hangi yöntemin diğerinden daha doğru bir sınıflama yapacağını farklı büyüklükteki başlangıç ve test verisi kullanarak incelemek ve sonrasında en uygun olan başlangıç ve test verisi büyüklüğüne göre farklı büyüklükteki örnek sayıları ile bu kez sadece CART ile modelleme yaparak en başarılı sınıflama modelini oluşturmaya çalışmaktır.

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.

GENERAL KNOWLEDGE

Name and Surname	: Yeliz Sevimli Saitoğlu
Field	: Econometrics
Programme	: Statistics
Supervisor	: Professor Ahmet Mete Çilingirtürk
Degree Awarded and Date	: Doctorate - June 2015
Keywords	: CART, MARS, Decision Trees, Turkish Youth and Their Political Views

ABSTRACT

CLASSIFICATION AND REGRESSION TREES

CART is a nonparametric statistical method that uses decision trees while solving the classification and regression problems using both categorical and continuous variables. If the dependent variable is categorical the method is called as classification tree. If the dependent variable is continuous the method is called as regression tree. Similar to CART, MARS is a nonparametric and nonlinear method that is designed for both continuous and binary dependant variables. If the dependant variable is continuous this method is used for prediction. If the dependant variable is categorical the method is used for classification. The common feature of these two methods is to develope both classification and prediction models according to the type of the dependant variable.

The objective of this study is to compare CART and MARS, practically determining which of them classify better using learning and test data with different sizes and creating the most successful classification model modeling only with CART with different sample sizes due to the best appropriate learning and test data sizes. In this context determining the factors that affect the political views and the political party choice of the Turkish youth is assessed as an implementation in order to define the best method for this case.

According to the information obtained from the implementations the best appropriate learning and test data sizes are determined. By the way approximately 70% of the data set should be taken as the learning data set and the remain part should be

taken as the test set. In the first step the category level of political party choice which is dependant variable is taken as 2. At first according to the general correct classification rate for both learning and test data the results of CART are higher and according to the sensitivity and specificity calculations the results of MARS method are higher. Secondly the category level of political party choice is taken as 5 and modeling with different sample sizes is done only with CART as the sample size increase, the general correct classification rate of both learning and test data increase generally and the values of the AUC ratios fluctuate respectively.

TEŞEKKÜR

Çalışmalarım boyunca bilgi ve deneyimlerini benimle paylaşan, değerli katkılarıyla beni yönlendiren, zor durumlarımda anlayışını ve yardımlarını benden esirgemeyen tez danışmanım ve kıymetli hocam Sayın Prof. Dr. Ahmet Mete ÇİLİNGİRTÜRK'e minnet ve şükranlarımı sunarım.

Tez çalışmalarım esnasında kıymetli zamanını benden esirgemeyerek değerli görüşlerini benimle paylaşan, sevgili hocam Sayın Prof. Dr. Dilek ALTAŞ'a ve tezin son aşamasında göstermiş olduğu ilgi ve destek için kıymetli hocam Sayın Doç. Dr. Ünal ÖZDEN'e en içten teşekkürlerimi sunarım.

Yüksek lisans ve doktora öğrenimim boyunca öğrencisi olmaktan büyük onur ve mutluluk duyduğum; cana yakınlığı ve alçak gönüllüğüyle herkesin sevgi ve saygısını kazanmış kıymetli hocam Sayın Prof. Dr. Şahamet BÜLBÜL'e en içten teşekkürlerimi sunarım.

Tezin savunulması esnasında yapıcı eleştirileriyle katkı sağlayan saygıdeğer hocam Sayın Prof. Dr. Dicle CENGİZ'e teşekkürü bir borç bilirim.

Yüksek lisans öğrenimimde klasik bir eğitimden ziyade bilgiye ulaşmayı, araştırmayı ve sorgulamayı bana öğreten, bana ışık tutan ve kariyerimde akademik olarak ilerlemeyi bana sevdiren kıymetli hocam Sayın Prof. Dr. Nural BEKİROĞLU'na ne kadar teşekkür etsem azdır.

Çalışmalarım esnasında yaşamış olduğum sıkıntılı ve yoğun süreçte desteğini benden esirgemeyen ve bana inanan sevgili eşim Engin SAİTOĞLU'na ve hep yanımda olduklarını hissettiren canım aileme en içten teşekkürlerimi sunarım. İyi ki varsınız…

İstanbul, 2015 Yeliz SEVİMLİ SAİTOĞLU

İÇİNDEKİLER

İKİNCİ BÖLÜM

SINIFLAMA VE TAHMİN MODELLERİ

ÜÇÜNCÜ BÖLÜM

UYGULAMA

TABLO LİSTESİ

Sayfa No.

ŞEKİL LİSTESİ

Sayfa No.

KISALTMALAR

AB	Avrupa Birliği
ADNKS	Adrese Dayalı Nüfus Kayıt Sistemi
AID	Automatic Interaction Detector (Otomatik Etkileşim Belirleyicisi)
BM	Birleşmiş Milletler
AUC	Area Under Curve (Eğri Altındaki Alan)
CART	Classification and Regression Trees (Sınıflama ve Regresyon Ağaçları)
DPO	Doğru Pozitif Oranı
GCV	Generalized Cross Validation (Genelleştirilmiş Çapraz Geçerlilik)
CHAID	Chi-Squared Automatic Interaction Detector (Otomatik Ki-Kare Etkileşim Belirleme)
CT	Classification Trees (Sınıflama Ağaçları)
E-CHAID	Exhaustive Chi-Squared Automatic Interaction Detector (Kapsamlı Otomatik Ki-Kare Etkileşim Belirleme)
EKK	En Küçük Kareler
İBBS	İstatistiki Bölge Birimleri Sınıflandırması
LAD	Least Absolute Deviation (En Küçük Mutlak Sapma)
LSD	Least Squared Deviation (En Küçük Kareli Sapma)
MARS	Multivariate Adaptive Regression Splines (Çok Değişkenli Uyarlanabilir Regresyon Uzanımları)
MDL	Minumum Description Length (En Küçük Uzunluk Tanımlaması)
MSE	Mean Squared Error (Ortalama Karesel Hata)
ROC	Receiver Operating Characteristic (Alıcı İşlem Karakteristiği)
RT	Regression Trees (Regresyon Ağaçları)
SLIQ	Supervised Learning in Quest (Araştırmada Denetimli Öğrenme)

STK	Sivil Toplum Kuruluşu
SPRINT	Scalable Parallelizable Induction of Decision Trees (Ölçeklenebilir, Parelellenebilir Tümevarım Karar Ağacı)
TÜİK	Türkiye İstatistik Kurumu
UNESCO	United Nations Educational, Scientific and Cultural Organization (Birleşmiş Milletler Eğitim, Bilim ve Kültür Kurumu)
YPO	Yanlış Pozitif Oranı
QUEST	Quick, Unbiased, Efficient Statistical Tree (Hızlı, Yansız, Etkin İstatistiksel Ağaç)

GİRİŞ

Birçok disiplin için çok sayıda ve karmaşık bilgi yığını içerisinden açıklayıcı olanları belirleyip en uygun modelin tahmininde bulunmak önemli bir problemdir. Bu problemin çözülebilmesi ihtiyacından yola çıkılarak birçok teknik geliştirilmiştir. Özellikle günümüzde bilgisayar teknolojilerine uygun yazılımlardaki çok hızlı yenilikler ve ilerlemeler bu tür problemlerin çözümüne önemli katkılar sağlamıştır. Veri madenciliği yöntemleri de bu durumlarda başvurulan önemli bir tekniktir. Veri madenciliğinde kullanılan birçok yöntem bulunmaktadır. Karar ağaçları da bu yöntemlerden bir tanesi olarak ortaya çıkmaktadır.

Bu çalışmada veri madenciliği tekniğine kısaca değinilecek ve veri madenciliğinin önemli tekniklerinden biri olan karar ağaçları incelenecektir. Çalışmanın asıl amacı karar ağaçları yöntemlerinden CART (Classification and Regression Trees: Sınıflama ve Regresyon Ağaçları) ve MARS (Multivariate Adaptive Regression Splines: Çok değişkenli Uyarlanabilir Regresyon Uzanımları) yöntemleri hakkında detaylı bilgi vermektir. Her iki yöntemin de ortak özelliği bağımlı değişkenin türüne göre hem sınıflama hem de tahmin modeli geliştirebilmesidir. Bu sebepten dolayı bu iki yöntemin karşılaştırılması amaçlanmaktadır.

CART yöntemi 1984 yılında Breiman, Friedman, Olshen ve Stone tarafından geliştirilmiştir.[1] CART hem kategorik hem de sürekli değişkenleri kullanarak sınıflama ve regresyon problemlerinin çözümünde karar ağaçlarını kullanan parametrik olmayan istatistiksel bir yöntemdir. Ele alınan bağımlı değişken kategorik ise yöntem sınıflama ağaçları (Classification Trees-CT), sürekli ise regresyon ağaçları (Regression Trees-RT) olarak adlandırılmaktadır.[2] Aynı durum MARS yöntemi için de geçerlidir. MARS yöntemi, 1990'ların başında Stanford Üniversitesinden istatistikçi Jerome H. Friedman tarafından geliştirilen nonparametrik ve lineer olmayan bir yöntemdir.[3]

---

[1] Özge Sezgin, "Statistical Method in Credit Banking", (**Yayınlanmamış Yüksek Lisans Tezi,** Orta Doğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, 2006, s.27.

[2] Zeynep Burcu Kıran, "Lojistik Regresyon ve CART Analizi Teknikleriyle Sosyal Güvenlik Kurumu İlaç Provizyon Sistemi Verileri Üzerinde Bir Uygulama", (**Yayınlanmamış Yüksek lisans Tezi,** Gazi Üniversitesi, FBE, 2010), s.19.

[3] Jerome H. Friedman, "Multivariate Adaptive Regression Splines", **The Annals of Statistics,** Vol.19, No.1 (June 1991), s.1.

MARS yöntemi, hem sürekli hem de ikili (binary) bağımlı değişkenler için tasarlanmıştır. Bağımlı değişkenin sürekli olması durumunda kestirim amaçlı olan bu yöntem, bağımlı değişkenin kategorik olması durumunda ise, sınıflandırma amacına sahiptir.[4]

Bu çalışmada amaç, Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde CART ve MARS yöntemlerini karşılaştırarak uygulamada hangi yöntemin diğerinden daha doğru bir sınıflama yapacağını farklı büyüklükteki başlangıç ve test verisi kullanarak inceledikten sonra en uygun olan başlangıç ve test verisi büyüklüğüne göre çeşitli büyüklükteki örnek sayıları ile en başarılı sınıflama modeline karar verilmesidir. Ayrıca uygulamada genellikle insan genetiği, gıda bilimi ve hastalık araştırmaları gibi çeşitli alanlarda kullanılan CART ve MARS yöntemleri bu alanlardan farklı olarak Türkiye gençliği üzerine yapılmış bir anketten toplanan veriler ile değerlendirilmiştir.

Çalışma üç ana bölümden oluşmaktadır. Birinci bölüm "Veri Madenciliğinde Karar Ağaçları", ikinci bölüm "Sınıflama ve Tahmin Modelleri" başlıklarını almakta ve üçüncü bölümde ise CART ve MARS yöntemine ilişkin bir uygulamaya yer verilmektedir.

Çalışmanın birinci bölümünde karar ağaçlarına genel bir giriş yapılmıştır. Genel tanım ve kavramlar ile matematiksel ifadelere yer verilmiştir ve çeşitli karar ağacı algoritmaları çıkış yıllarına göre incelenmiştir. İkinci bölümde ise; CART ve MARS yöntemleri detaylı olarak ele alınmıştır. Her iki yöntemde snıflama ve tahmin modellerinin oluşum yapısı incelenmiştir. Oluşturulmuş olan modeller içerisinde en uygun modellerin seçiminin yapılabilmesi için gerekli olan model seçim kriterlerinden bahsedilmiştir. Sonuç olarak her iki yöntemin avantaj ve dezavantajlarına bakılarak yöntemler karşılaştırılmıştır.

Çalışmanın üçüncü bölümü ise; uygulamadan oluşmaktadır. Bu bölümde KONDA Araştırma ve Danışmanlık şirketinin Nisan 2011 tarihinde yaptığı *"Türkiye*

[4] Mukhopadhyay A ve Iqbal A, "Prediction of Mechanical Property of Steel Strips Using Multivariate Adaptive Regression Splines" *Journal of Applied Statistics*, 2009, Vol.36, No.1, http://www.tandfonline.com/toc/cjas20/current (23 Mart 2009), s.3.

Gençliği Araştırması" adlı bir anketten faydalanılmıştır.[5] Verilerin analizinde ise, karar ağaçlarında kullanılan sınıflama yöntemleri olan CART ve MARS yöntemleri kullanılmıştır. Analiz sonucunda kurulan modellerin karşılaştırılması yapılmış ve her iki yöntemden elde edilen bulgular yorumlanmıştır.

[5] KONDA Araştırma ve Danışmanlık, Türkiye Gençliği Araştırması'2011

BİRİNCİ BÖLÜM

VERİ MADENCİLİĞİNDE KARAR AĞAÇLARI

Bu bölümde ilk olarak veri madenciliği hakkında genel bir bilgi verilmiş, daha sonra veri madenciliğinin önemli bir kolu olan karar ağaçları yöntemlerinden detaylı bir şekilde bahsedilmiştir. Karar ağaçları anlatılmadan önce karar ağaçlarında kullanılacak olan kavramlara yer verilmiştir.

1.1. Genel Tanım ve Kavramlar

Dünyadaki bilgi hacmi her üç yılda bir ikiye katlanarak insanlık tarihinde daha önce hiç görülmediği kadar hızlı artış göstermektedir. Ham maddesi dağ gibi bilgi yığınları olan Big Data (Büyük Veriler) ve bununla ilgili teknolojiler son zamanlarda çok büyük önem kazanmaya başlamıştır.[6] Bu büyük verilerin işlenmesinde veri madenciliği yöntemleri kullanılmaktadır.

Veri madenciliği daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veri tabanlarından elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır.[7] Veri madenciliği ile büyük veri yığınlarından oluşan veri sistemleri içerisinde gizlenmiş bilgilerin çekilmesi sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veritabanı teknolojisi ve çeşitili bilgisayar programları kullanılarak yapılır. Veri madenciliği kendi başına bir çözüm değil çözüme ulaşmak için verilecek karar sürecini destekleyen, problemi çözmek için gerekli bilgileri sağlamaya yarayan bir araçtır.[8]

[6] Bilim ve Teknik, "Rastlantının Bittiği Yer Big Data", Cilt. 46, Sayı. 550 (Eylül 2013), s. 22-23.
[7] Gökhan Silahtaroğlu, **Veri Madenciliği (Kavram ve Algoritmaları),** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.12.
[8] Gamze Pehlivan, "Chaid Analizi ve Bir Uygulama", (**Yayınlanmamış Yüksek lisans Tezi,** Yıldız Teknik Üniversitesi, FBE, 2006), s.13.

Karar verme sürecinde önemli bir yere sahip olan veri madenciliği; pazarlama, bankacılık ve sigortacılık, borsa, perakendecilik, telekominikasyon, endüstri ve mühendislik, sağlık ve ilaç gibi çeşitli sektörlerde kullanılmaktadır.[9]

Veri madenciliğinde kullanılan modeller, "Tanımlayıcı Modeller" ve "Tahmin Edici Modeller" olmak üzere iki başlık altında incelenebilir. Tanımlayıcı modellerde (descriptive models) amaç karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanmasını sağlamaktır. Tanımlayıcı modeller, ilişki analizi ve kümeleme analizi olmak üzere iki grupta incelenebilir. İlişki analizi kapsamında ise birliktelik kuralları (association rules) ile ardışık zamanlı örüntüler yer almaktadır. Tahmin edici modellerin (predictive models) amacı ise, sonuçları bilinen verilerden hareket edilerek bir model geliştilmesi ve kurulan bu modelden yararlanarak, sonuçları bilinmeyen veri kümeleri için sonuç değerlerini tahmin etmektir. Tahmin edici modeler ise sınıflandırma ve istatistiksel tahmin modelleri olmak üzere iki başlık altında incelenmektedir. Karar ağaçları, Yapay sinir ağları ve genetik algoritmalar en yaygın olarak kullanılan sınıflandırma teknikleridir. İstatistiksel tahmin modellerinde ise regresyon analizi, diskriminant analizi ve lojistik regresyon analizi en yaygın kulanılan tekniklerdir.[10] Bu durum aşağıdaki tabloda özet olarak gösterilmiştir.

Tablo 1.1. Veri Madenciliğinde Kullanılan Modeller

Veri Madenciliğinde Kullanılan Modeller			
Tahmin Edici Modeller		**Tanımlayıcı Modeller**	
Sınıflandırma	**İstatistiksel Tahmin Modelleri**	**İlişki Analizi**	**Kümeleme Analizi**
-Karar Ağaçları -Yapay Sinir Ağları -Genetik Algoritmalar -Bellek Tabanlı Sınıflandırma	-Regresyon Analizi -Diskriminant Analizi -Lojistik Regresyon Analizi	-Birliktelik Kuralları -Ardışık Zamanlı Örüntüler	-Hiyerarşik Kümeleme -Hiyerarşik Olmayan Kümeleme

Kaynak: Yazar tarafından çeşitli eserlerden yararlanılarak oluşturulmuştur.

Tahmin edici modellerden olan sınıflandırma modeli, sınıfı tanımlanmış mevcut verilerden yararlanarak, sınıfı belli olmayan verilerin sınıfını tahmin etmek için kullanılan bir veri madenciliği modelidir. Sınıflama modelinin içinde yer alan karar

[9] Pehlivan, s. 14.
[10] Umman Tuğba Şimşek Gürsoy, **Uygulamalı Veri Madenciliği: Sektörel Analizler,** 3. Basım, Ankara: Pegem Akademi, 2012, s.5.

ağacı yöntemi ile verinin içerdiği ortak özelliklere göre veri sınıflanır ve çok sayıda kayıt içeren bir veri kümesi bir dizi karar kuralları uygulanarak daha küçük kümelere bölünür.[11] Örneğin bir banka kredi verdiği müşterilerin risk durumunu karar ağaçları yardımıyla sınıflayarak belirlediğinde, müşterilerin risk profili ortaya çıkmış olacaktır. Buna göre, belirli özelliklere sahip müşterilerinden kredi talebi geldiğinde, karar ağacı bilgilerine dayanarak kredi verip vermeme konusunda karara varmış olacaktır.[12] Tahmin edici modellerde yer alan istatistiksel tahmin modelleri ise, bağımlı değişken ile bağımsız değişkenler arasındaki neden sonuç ilişlikilerinin belirlenmesinde kullanılır.

Tanımlayıcı modellerden olan kümelemenin amacı ise, verilerin kendi arasındaki benzerliklerini göz önüne alarak gruplandırmasıdır. Bu model özellikle Pazar araştırmalarında kullanılmaktadır. Örneğin alışveriş mağazalarında müşterilerin gruplara ayrılması bir kümeleme işlemidir. Bunun dışında desen tanımlama, resim işleme ve uzaysal harita verilerinin analizinde kullanılmaktadır.[13] Kümeleme analizi, hiyerarşik kümeleme ve hiyerarşik olmayan kümeleme olmak üzere ikiye ayrılır. Hiyerarşik olmayan kümelemede kullanılan yöntem k-ortalamalar kümesi yöntemidir ve küme sayısı önceden belirlenir. Hiyerarşik kümelemede ise, küme sayısı analiz sonucunda elde edilen ağaç grafiği ve uzaklık katsayıları bulgularına göre belirlenir. Kümeleme analizininin diskriminant analizinden farkı, diskriminant analizinde gruplar önceden belirlenirken kümeleme analizinde bu belirleme analiz sonucunda elde edilmektedir.[14]

Tanımlayıcı modellerinden olan birliktelik kurallarının amacı ise, veritabanı içinde yer alan kayıtların birbirleriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerini ortaya koymaktır. Bu model özellikle pazarlama alanında kullanılmaktadır. *"Pazar sepet analizleri"* adı verilen bu uygulamalar bu tür veri madenciliği yöntemlerine dayanır. Pazar sepet analizleri yardımıyla müşterilerin alışveriş alışkanlıkları belirlenmeye çalışılır. Bir müşteri herhangi bir ürünü aldığında, sepetine başka hangi ürünleri de koyduğu belirli bir

[11]Ali Sait Albayrak, Şebnem Koltan Yılmaz, "Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri üzerine Bir Uygulama", **Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Dergisi,** Cilt.14, Sayı.1 (2009), s.33.
[12] Yalçın Özkan, **Veri Madenciliği Yöntemleri,** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.45.
[13] Özkan, s. 47.
[14] Aliye Kayış, "Kümeleme Analizi", Şeref Kalaycı (Ed.), **SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri,** içinde (349-376), Ankara: Asil Yayın Dağıtım, 2006, s.349.

olasılığa göre hesaplanır. Birlikte satın alınan ürünler belirlendiğinde, mağazalarda raflar ona göre düzenlenerek müşterilerin bu tür ürünlere daha kolayca erişmeleri sağlanabilir.[15] Tanımlayıcı modellerinden olan ardışık zamanlı örüntüler, birliktelik kurallarıyla aynı amacı sağlamak için kullanılır ve birbirini izleyen olaylardaki örüntüleri belirler.

Sonuç olarak iki başlık altında incelenebilen tahmin edici ve tanımlayıcı modellerden tahmin edici yöntemlerin amacı, verilerden hareket ederek bir model geliştirmek ve kurulan bu model yardımıyla sonuçları bilinmeyen veri kümelerinin sonuç değerlerini tahmin etmek, sınıflamak ve karakterize etmektir. Eğer tahmin edilecek değişken sürekli bir değişken ise tahmin problemi regresyon, kategorik bir değişken ise sınıflama problemi olarak tanımlanmaktadır. Tanımlayıcı yöntemlerin amacı ise, bir bütün olarak veri kümesinin yapısını incelemek ve verilerdeki örüntülerin tanımlanmasını sağlamaktır.[16] Veri madenciliğinde kullanılan modeller burada kısaca anlatıldıktan sonra karar ağaçları daha detaylı olarak aşağıda anlatılmıştır.

Veri madenciliğinde kullanılan önemli bir yöntem olan karar ağaçları yani ağaç temelli yöntemler, parametrik olmayan istatistiksel yöntemlerde kullanılan bir veri analizi yöntemidir. İncelenen öznitelikler uzayını (feature space) bir dizi dörtgene böler ve basit bir modele uydurmaya çalışır.[17] Bilindiği üzere parametrik yöntemler örneklenen kitlenin belirli özellikleri sağladığı varsayımına dayanır. Bu varsayımlar: gözlemlerin bağımsız olması, normal dağılımlı kitleden alınmış olması, en azından aralıklı bir ölçekle ölçülmüş olması ve kitle varyanslarının aynı olmasıdır. Bu varsayımların pratikte her zaman sağlanması zor olduğu için bu varsayımların sağlanamadığı durumlarda parametrik olmayan istatistiksel yöntemler kullanılır.[18] Bu yüzden karar ağaçları, birtakım varsayımları sağlama koşulu taşımaması, kurulmasının ucuz olması, yorumlanmasının kolay olması, veritabanı sistemleri ile kolayca entegre edilebilmeleri, güvenilirliklerinin iyi olması ve ağaç yapısı ile kolay anlaşılabilen

[15] Özkan, s. 48-49
[16] Fürüzan Köktürk, Handan Ankaralı ve Vildan Sümbüloğlu, "Veri Madenciliği Yöntemlerine Genel Bakış", **Türkiye Klinikleri Journal of Biostatistics,** Cilt.1, Sayı.1 (2009), s.22.
[17] Trevor Hastie, Robert Tibshirani ve J erome Friedman, **The Elements of Statistical Learning: Date mining, Inference and Prediction,** New York: Springer, 2001, s.266.
[18] Gülay Kıroğlu, **Uygulamalı Parametrik Olmayan İstatistiksel Yöntemler,** İstanbul: Paymaş, 2001, s.7.

kurallar yaratabilmesi nedeni ile veri madenciliğinde tahmin edici yöntemlerden biri olan sınıflandırmada kullanılan önemli bir yöntemdir.[19]

Karar süreci diyagramlarının sonlu sayıda karar düğümleri, onları birbirine bağlayan dalları ve yaprakları ile ağaca benzeyen bir yapısı vardır. Bu yüzden karar ağacı olarak adlandırılmışlardır. Karar ağacının en tepesinde kök düğümü bulunmaktadır. Kök düğüm verideki ilgili bağımsız değişkene ait bütün gözlemleri kapsamaktadır. Her ara düğüm ya da diğer adlarıyla; yaprak olmayan düğüm, uç olmayan düğüm (internal node-nonleaf node-nonterminal nodes) bir değişkene yapılacak testi belirtir. O yüzden karar düğümü olarak da adlandırılırlar. Bu testin sonucunda karar ağacı herhangi bir veri kaybına uğramadan dallara ayrılmaktadır. Bu durumda her düğümden ayrılan dal, o düğüme uygulanan testin sonucunu ifade eder. Yani dallar düğümler arası bağlantı ilişkilerini ifade eder. Her düğümde test ve dallara ayırma işlemi ardışık olarak devam eder. Eğer bir dalın ucunda sınıflama işlemi gerçekleşmiyorsa bu noktada bir karar düğümü yani ara düğüm meydana gelir. Eğer sınıflama işlemi yapılabiliyorsa bu durum o dalın sonunda yaprak düğümü ya da bir diğer adıyla uç düğüm (leaf node-terminal node) olduğunu ifade eder. İşte bu uç düğüm, veri kümesi üzerinde belirlenmek istenen sınıflardan birisi olmaktadır.[20,21] Buna göre bir karar ağacına ait düğümlerin tanımları aşağıdaki tablo ile özetlenebilir.

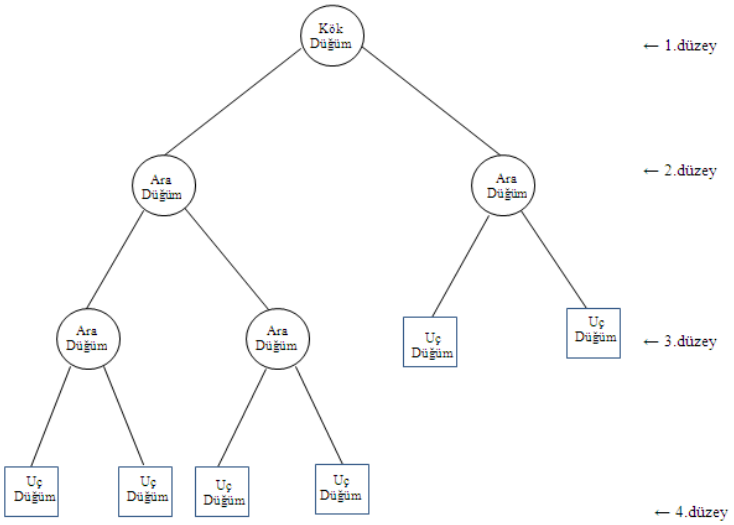
Tablo 1.2. Karar Ağacının Düğümleri

Düğüm Adı	**Tanımı**	**Ağaçta Bulunduğu Yer**
Kök Düğüm	Ana düğümdür. Verideki ilgili bağımsız değişkene ait bütün gözlemleri kapsar.	Başta
Ara Düğüm	Karar düğümüdür. Yaprak değildir. İlgili değişkenin test sonucuna göre yeniden bölünür. Bu yüzden uçta bulunmayan düğümdür.	Ortada
Uç Düğüm	Yaprak düğümdür. Bölünme yapmaz. Sınıfa atama olur.	Sonda

Kaynak: Yazar tarafından çeşitli eserlerden yararlanılarak oluşturulmuştur.

[19] Kıran, s.11.
[20] Jiawei Han ve Micheline Kamber, **Data Mining Concept and Techniques,** 2. Basım, Amerika: Morgan Kaufmann Publications, 2006, s.291.
[21] Aysan Şentürk, **Veri Madenciliği Kavram ve Teknikler,** 1. Basım, Bursa: Ekin Basın Yayın Dağıtım, 2006, s.34-35.

Kısaca karar ağacı süreci kök düğümden başlayarak, yukarıdan aşağıya doğru ardışık düğümler takip edilerek yaprağa ulaşıncaya kadar devam eden bir süreçtir.[22] Yani kök düğüm bölünmeler yaparak kendi alt kümeleri olan ara ve uç düğümleri oluşturmaktadır. Buna göre yukarıdaki tanımlardan yola çıkarak örnek bir karar ağacı modeli aşağıdaki şekil ile gösterilebilir. Şekilde de görüldüğü gibi uç düğümler kare şeklinde gösterilirken, uç olmayan yani bölünmeye devam eden düğümler daire şeklinde gösterilir.[23]



Şekil 1.1: Karar Ağacı Modeli Örneği
Kaynak: Yazar tarafından çeşitli eserlerden yararlanılarak oluşturulmuştur.

Şekildeki ağacın, aynı aile soyağacında olduğu gibi hiyerarşik bir yapısı vardır. Bu yüzden daha önce belirtmiş olduğumuz kök, ara ve uç düğüm ifadeleri tıpkı aile soyağaçlarında kullanılan ata, ana, kardeş, çocuk ve torun gibi birçok kavramı içermektedir.[24] Buna göre bu kavramları sırasıyla açıklayabiliriz. Şekildeki düzey ifadesi, bir düğümüm köke olan uzaklığıdır. Buna derinlik de denilebilir. Kök düğümün düzeyi birdir. Buna göre, bir düğüme doğrudan bağlı olan düğümlere o düğümün çocukları (child) denir. Şekilde de görüldüğü gibi kök düğümün 2. düzeyde iki tane çocuk düğümü vardır. 2. düzeydeki ara düğümlerin de 3. düzeyde görüldüğü üzere,

[22] Şentürk, s.35.
[23] Leo Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s. 21.
[24] Rifat Çölkesen, **Veri Yapıları ve Algoritmalar**, 9. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2014, s.268.

ikişer tane çocuk düğümü vardır. Ama uç düğümlerin çocuk düğümü yoktur. Düğümlerin doğrudan bağlı oldukları düğüm aile (parent) olarak adlandırılır. Örneğin 2. düzeydeki ara düğümlerin ailesi kök düğümdür. Aynı düğüme bağlı olan düğümlere kardeş (sibling) düğüm denir. Örneğin ikinci düzeydeki ara düğümler aynı zamanda birbirinin kardeşidir. Aile düğümünün daha üstünde kalan düğümlere ata (ancestor) denir. Örneğin kök düğümün üstünde başka düğüm olmağından atası yoktur. Ama 3. düzeydeki düğümlerin atası kök düğümdür. 4. düzeydeki düğümlerin atası ise hem 2. düzey hem de 1. düzeyde bulunan düğümlerdir. Bir düğümün çocuğuna bağlı olan düğüme ise torun (descendant) denir. Örneğin kök düğüme 3. düzeyden itibaren bağlı olan tüm düğümler kök düğümün torunudur. Burda bu kavramlar kısaca kök, ara ve uç düğüm olarak kullanılacaktır.

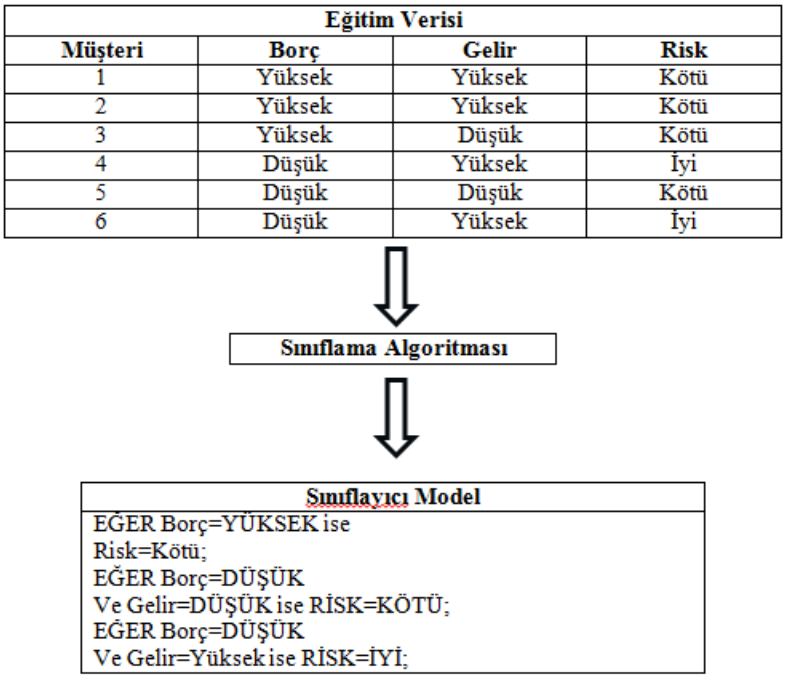
Karar ağacı kurulurken eldeki veri kümesinin bir kısmı öğrenme (learning) işlemi için geri kalanı ise oluşturulan ağacı test etmek için kullanılır.[25] Öğrenme işlemi için kullanılan bu veri, başlangıç veri kümesi (learning sample) ya da eğitim verisi (training data) olarak da adlandırılır.

Başlangıç veri kümesi içinde geçmiş bilgileri barındırır ve bütün gözlemlerin önceden hangi sınıflara atanmış olduğunu bilir. Böylelikle geçmiş veri kümesine (set of historical data) göre karar ağacı oluşturulur ve yeni veri kümesi de bu ağaçtaki bilgilere göre sınıflandırılır.[26] Karar ağaçları bir dizi soru ile temsil edilmektedir. Bu yüzden başlangıç veri kümesi bir dizi sorular aracılığıyla daha küçük parçalara bölünür. Ağaç kurulurken başlangıç veri kümesi yani eğitim verisi ile ağaç oluşturulur. Bu veriler veri kümesi içinden rasgele seçilir. Rasgele seçilen bu veriler ile bir sınıflama kuralı oluşturularak, sınıflama modeli elde edilir. Veri kümesinin geri kalanı ise, test verisi (test data) olarak adlandırılır ve ağaç oluşum sürecinde oluşturulan sınıflama kuralının doğruluğunu tahmin eder. Eğer tahmin edilen doğruluk oranı kabul edilebilecek bir

[25] Silahtaroğlu, s.73.
[26] Roman Timofeev, "Classification and Regression Trees (CART) Theory and Aplications", 2004, Humbolt University, Center of Applied Statistics and Economics, http://tigger.uic.edu/~georgek/HomePage/Nonparametrics/timofeev.pdf (26 Kasım 2012).

düzeyde ise, bu kural yeni veriler üzerinde de uygulanır.[27] Böylelikle karar ağaçları ile bir sınıflandırma modeli oluşturulmuş olur.

Bu durum aşağıdaki örnek bir sınıflandırma modeli ile daha iyi anlaşılabilir. Aşağıdaki şekilde görüldüğü gibi müşterilere ait borç ve gelir değişkenleri bağımsız değişkenler iken, sınıf değişkeni olan risk ise bağımlı değişken olarak değişken niteliklerine yani seviyelerine göre gösterilmiştir. Böylelikle karar ağacınında başlangıç veri kümesi ilk olarak, belirli bir örneğe göre, yani eğitim kümesindeki veriye göre oluşturulmuştur.

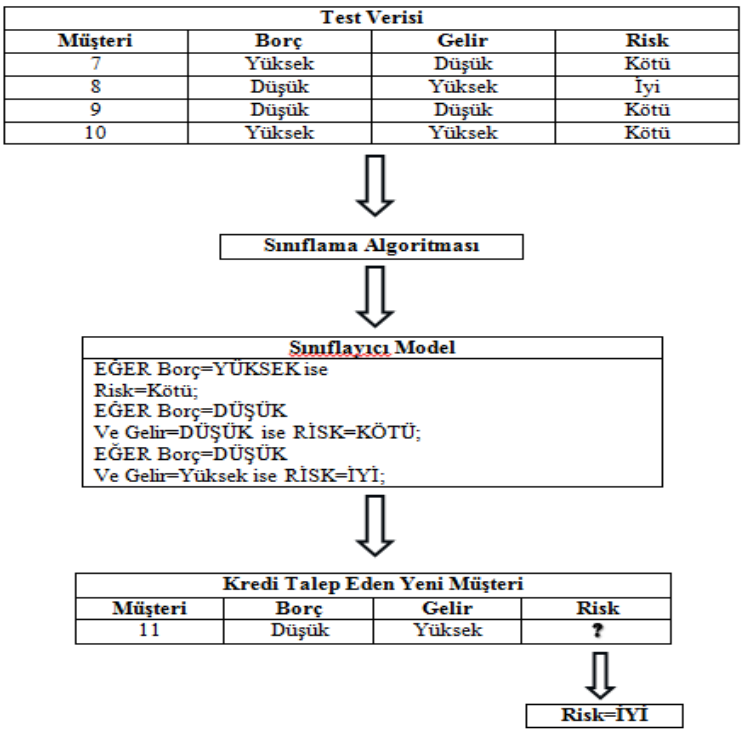


Şekil 1.2: Sınıflama İle İlgili Olarak Model Kurma Süreci
Kaynak: Yalçın Özkan, Veri Madenciliği Yöntemleri, 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.52.

Sınıflandırma sürecinin ikinci aşamasında ise Şekil 1.3'te görüldüğü üzere başlangıç veri kümesinin geri kalan kısmı olan, test verileri üzerinde sınıflama kuralı belirlenir ve söz konusu bu kurallar bu kez test verilerine uygulanarak sınanır. Şekil 1.3'te görüldüğü gibi artık 11 numaralı yeni bir müşteri kredi talebinde bulunduğunda,

[27] Han ve Kamber, s. 287.

bu müşterinin risk durumunu belirlemek için örnek verilerden elde edilen karar kuralı doğrudan uygulanır. Bu müşteri için Borç=Düşük, Gelir=Yüksek olduğu biliniyorsa risk durumunun Risk=İYİ olduğu hemen anlaşılır. Yapılan test sonucunda elde edilen modelin doğru olduğu kabul edilecek olursa, bu model diğer veriler üzerinde de uygulanır. Elde edilen sonuç model mevcut ya da olası müşterilerin gelecekteki kredi talep risklerini belirlemede kullanılır.[28]



Şekil 1.3: Modelin Uygulanma Süreci
Kaynak: Yalçın Özkan, Veri Madenciliği Yöntemleri, 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.52.

Böylelikle uygun sınıflama algoritmaları uygulanarak elde edilen sınıflama modelleri ile olumlu sonuç elde edilirse, bu modeller sayesinde mevcut ya da olası kayıdın gelecekte hangi sınıfta yer alacağı belirlenebilir.

[28] Özkan, s.52.

Buna göre veri madenciliğinde önemli bir yöntem olan karar ağaçları ile igili genel tanım ve kavramlar hakkında bilgi edinildikten sonra, karar ağaçları ile sınıflandırma modellerinin nasıl gerçekleştiğinin anlaşılması için genel matematiksel gösterimlerin bilinmesi gerekir. Bu amaçla bundan sonraki aşamada karar ağaçları ile sınıflamanın matematiksel olarak ifade edilebilmesi için kullanılacak genel gösterimlerin tanımlanması gerekmektedir.

Buna göre **x**; ölçüm vektörü (measurement vektor) olarak adlandırılır ve bir gözlem için ölçülmek istenen değişkenlerinin ölçüm sonuçlarını içinde barındırır. Yani p adet değişkenden oluşan bir gözlemin vektörü olarak düşünülebilir. Buna göre ölçüm vektörü şu şekilde gösterilir:

$$\mathbf{x}^{\mathbf{l}} = (x_1, x_2, ...x_p) \tag{1.1}$$

X; ise ölçüm uzayı (measurement space) olarak adlandırılır ve bütün gözlemlerin ölçüm sonuçlarını yani ölçüm vektörlerini içinde barındıran bir matristir. Ölçüm uzayı içinde hangi ölçümler önemliyse onun ölçüm vektörünü taşır. Bu yüzden birçok ölçüm uzayı matrisi oluşturulabilir. Buna göre (nxp) boyutlu **X** matrisi aşağıdaki gibi gösterilir:

$$\mathbf{X}= \begin{pmatrix} x_{11} & x_{12} & x_{1l}... & x_{1p} \\ . & . & ..... & . \\ x_{g1} & x_{g2} & x_{gl}... & x_{gp} \\ . & . & ..... & . \\ x_{n1} & x_{n2} & x_{nl}... & x_{np} \end{pmatrix} \tag{1.2}$$

Burda g=1, …, n ve *l*=1, …, p olmak üzere; n gözlem sayısını, p ise bağımsız değişken sayısını göstermektedir. Buna göre $\mathbf{X}_{\text{nxp}}$ matrisinde ilk indis gözlem numarasını, ikinci indis ise değişken numarasını gösterir.

Oluşturulan karar ağacı **T** ile ifade edilmek üzere, $\tilde{T}$ ağaçtaki uç düğümlerin oluşturmuş olduğu bir kümeyi ifade etmek için kullanılacaktır. Gözlemlerin J adet sınıfa atandığını varsaydığımızda sınıf sayısı 1,2,…J adet olmak üzere, C sınıf kümesi olarak adlandırılır ve;

$$C = \{1,2,\ldots,J\} \quad (1.3)$$

ile gösterilir. Burda C'ler aslında ***bağımlı değişken*** değeridir. Bağımlı değişkende kaç adet sınıf yani kategori olduğunu ifade eder.

Bir sınıf üyeliğini tespit ederken izlenecek yol $\mathbf{x} \in \mathbf{X}$ olmak üzere her bir **x** vektörünü (yani her bir gözlemi), 1'den J'ye kadar olan bu sınıflardan birine atamaktır. Buna göre bir sınıflayıcı ya da sınıflama kuralı (a classifier or classification rule), d(x) ile gösterilen bir fonksiyon ile tanımlanır. Her problemde sınıflayıcının yani sınıflama kuralının görevi bir gözlemin (vakanın) ya da nesnenin hangi sınıfta yer alacağını sistematik bir şekilde tahmin etmektir.[29] Buna göre d(**x**) fonksiyonu, **X** matrisi ve dolayısıyla **x** vektörü üzerinde tanımlı bir fonksiyondur ve 1'den J'ye kadar olan sınıflardan birine eşittir {d(**x**) = j}. Yani sınıflandırıcının [d(x)] asıl amacı, **X** matrisinin bir alt kümesini olarak tanımlanan A_j'yi oluşturmaktır.

$$A_j = \{\mathbf{x};\ d(\mathbf{x}) = j\} \quad (1.4)$$

(1.4) nolu eşitliğe göre **x** vektörü, d(**x**) olarak tanımlanan bir sınıflandırıcı fonksiyonu aracılığıyla j sınıfına atanmakta ve A_j denilen yani j sınıfını içinde barındıran bir alt küme oluşturmaktadır. Böylelikle her bir gözlemin sınıflara ataması gerçekleştirilir. Bu durum bize başlangıç veri kümesinin nasıl bir yol izleyerek alt parçaya ayrıldığını gösterir.

Çünkü $A_1,\ldots,A_j$'ler **X** matrisinin alt kümeleridir ve $\mathbf{X} = \underset{j}{U} A_j$ olup, her **x** vektörü tahmini bir j sınıfına atanmaya çalışılır. Atandığı sınıf içinde $\mathbf{x} \in A_j$'dir.[30]

Karar ağaçları ile sınıflamanın matematiksel gösterimi ile bir ağacın daha alt kümelerine ayrılırken nasıl bir yol izleyerek sınıflama yaptığı yukarıda detaylı olarak anlatılmıştır.

[29] Breiman ve diğerleri, s.3.
[30] Breiman ve diğerleri, s. 4.

Fakat burdan sonra akıllara gelen bir diğer soru bu ağaçta bölünmeler yani dallanmalar olurken, bu dallanmaların nasıl gerçekleştiği yani hangi kriterlere göre belirlendiğidir. Bu amaçla bundan sonraki aşamada karar ağaçlarının dallanma kritelerin anlatılması gerekmektedir.

1.2. Karar Ağaçlarında Dallanma Kriterleri

Karar ağaçlarının oluşturulması sırasında dallanmaya hangi değişken ile başlanacağı oldukça önemlidir. Çünkü olası tüm ağaç yapılarını ortaya çıkararak içlerinden en uygun olanı ile başlamak mümkün değildir. Bu sebeple karar ağacı algoritmalarının çoğu daha başlangıçta bir takım değerleri hesaplayarak ona göre ağaç oluşturma yoluna gitmektedir.[31]

Buna göre karar ağaçlarında çözülmesi gereken en önemli problemlerden biri, kökten itibaren *bölünmenin* veya bir başka deyişle *dallanmanın* hangi kritere göre yapılacağıdır. Söz konusu bu kriterler ***entropi hesabına dayalı*** kriterler ve onun türevleri olan kazanç ölçütü ve kazanç oranı olabileceği gibi ikinci bölümde detaylı olarak anlatılan Sınıflama ve Regresyon Ağaçlarında (Classification and Regression Tress: CART) kullanılan kriterlerdir. [32] Sınıflama ve regresyon ağaçlarında kullanılan kriterler ise ***safsızlık fonksiyonuna (impurity function)*** "ϕ" dayalı bölünme kriterleridir. Bu kriterler bölme uyum kriteri (goodness of split criteria) olarak da adlandırılır. Sınıflama ağacı problemlerinde kullanılan safsızlık fonksiyonuna dayalı kriterler gini kriteri (ya da gini indeks) ve twoing bölme kriteridir. Regresyon ağacı problemlerinde kullanılan ve safsızlık fonksiyonu dayalı bölme kriteri ise en küçük kareli sapma (least squared deviation: LSD) ve onun alternatif yöntemi olan en küçük mutlak sapma (least absolute deviation: LAD) yöntemidir.[33,34,35]

---

[31] Pınar Tapkan, Lale Özbakır ve Adil Baykasoğlu "Weka ile Veri Madenciliği Süreci ve Örnek Uygulama", Endüstri Mühendisliği Yazılımları ve Uygulamaları Kongresi, İzmir: MMO, 30 Eylül – 01/02 Ekim 2011, s.249.
[32] Özkan, s.54.
[33] Sezgin, s. 31.
[34] IBM SPSS Modeler 15 Algorithms Guide, C&RT Algorithms, 2012, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/AlgorithmsGuide.pdf (29 Mart 2014), s.60.
[35] *Tree-based Regression,* (t.y.) http://www.dcc.fc.up.pt/~ltorgo/PhD/th3.pdf (28.02.2014).

Buna göre herbiri farklı bir hesaba dayalı dallanma kriteri için bir karar ağacı algoritmasının oluşumu söz konusudur. Bu algoritmalar 1.4. kısımda anlatılmıştır. Bu algoritmaların oluşum mantığını anlamak için sırasıyla entropiye ve safsızlık fonksiyonuna dayalı dallanma kriterlerinin anlatılmasında fayda vardır.

1.2.1. Karar Ağaçlarında Entropiye Dayalı Dallanma Kriterleri

Karar ağaçlarında dallanmaların nasıl olacağı belirlenirken kullanılan kriterlerden biri de entropiye dayalı algoritmalardır. Karar ağaçlarında entropiye dayalı bölünme kriterleri olan kazanç ölçütü ve kazanç oranı anlatılmadan önce entropinin ne olduğunun bilinmesi gerekir. Enformasyon teorisinde, entropi rassal bir değişkenin belirsizliğinin ölçüsü olarak tanımlanmaktadır. Entropi 1948 yılında Claude E. Shannon'un "*A Mathematical Theory of Communication*" adlı makalesinde anlatılmıştır.[36] Oysa enformasyon teorisindeki geçmişi Pauli ve Von Neumann'a dayanmaktadır.[37] Entropinin belirlenmesi için enformasyonun bilinmesi gerekir. Enformasyon, rassal bir olayın gerçekleşmesine ilişkin bir bilgi ölçütüdür. Enformasyonun formülü aşağıdaki gibidir.

$$I(x) = \log_a \frac{1}{p(x)} = -\log_a p(x) \qquad (1.5)$$

Shannon' a göre bir x mesajının taşıdığı bilgi, yukarıdaki bağıntıda ifade edildiği gibi, onun gerçekleşmesi olasılığına bağlıdır ve bu olasılığın a tabanına göre eksi işaretli logaritması ile ifade edilir. Bilginin birimi ise, logaritmanın tabanına bağlıdır. Eğer taban 2 ise birim bit'tir, e ise birim nat'tır, 10 ise birim hartley'dir. Bilgisayar dünyasında 0 ve 1'lerle yani bitlerle çalışıldığı için logaritmanın tabanı 2 olarak kabul edilir. Bu bağıntıya göre enformasyon, o bilgiyi ifade etmek için kaç bit kullanılması gerektiğini gösterir. Yukarıdaki eşitlik yüksek olasılığa sahip mesajların

[36] *Entropy (information theory)*,(t.y.) http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Entropy_(information_theory).html (15.11.2013).
[37] *Shannon entropy*, (t.y.) http://www.ueltschi.org/teaching/chapShannon.pdf (15 Kasım 2013).

düşük bilgi içerdiğini, düşük olasılığa sahip mesajların ise yüksek bilgi içerdiğini göstermektedir.[38]

Buna göre, X bir kaynak olmak üzere bu kaynağın $\{ſ_1, ſ_2, \ldots, ſ_w, \ldots, ſ_W\}$ olmak üzere W mesaj üretebildiği varsayılsın. Tüm mesajlar birbirinden bağımsız üretilmekte ve W mesajlarının üretilme olasılıkları p_w olmak üzere, $P = \{p_1, p_2, ..., p_w, ..., p_W\}$ olasılık dağılımına sahip mesajları üreten X kaynağının entropisi H(X) aşağıdaki gibidir:[39]

$$H(X) = E(I(X)) = \sum_{1 \le w \le W} P(x_w).I(x_w)$$
$$H(X) = \sum_{w=1}^{W} p(x_w) \log_2 \frac{1}{P(x_w)} = -\sum_{w=1}^{W} P_w \log_2 P_w \qquad (1.6)$$

Yukarıdaki formülasyondan anlaşılacağı üzere entropi, bir süreç için tüm örnekler tarafından içerilen enformasyonun beklenen değeridir. Yani Shannon Entropisi ile iletilen bir mesajın taşıdığı enformasyonun beklenen değeri bulunmuş olur. Enformasyon ve entropi zıt şeyleri temsil etmelerine rağmen Shannon'a göre maksimum belirsizlik maksimum enformasyon sağladığı için enformasyon ve belirsizlik terimleri benzerdir.[40]

Çünkü düşük olasılığa sahip mesajlar 1.5'teki formülasyonda görüldüğü gibi yüksek bilgi yani maksimum enformasyon içerir. Maksimum enformasyonda 1.6'daki formülde görüldüğü üzere, entropinin yüksek çıkmasını sağlamaktadır. Bu yüzden enformasyon ve entropi terimlerinin benzer olduğu söylenir.

Enropi formülasyonundan anlaşılacağı üzere:

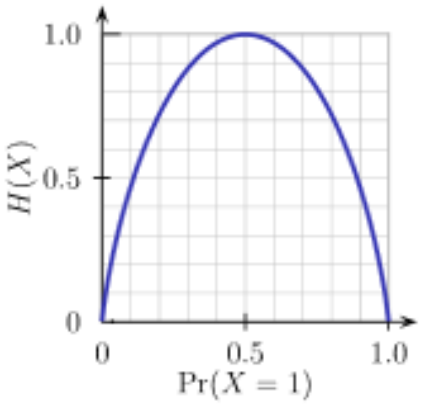
i. Örnekler aynı sınıfa ait ise entropi=0

ii. Örnekler sınıflar arasında eşit dağılmışsa yani eşit olasılıklı durumda entropi=1

[38]Gülser Dondurmacı, "Veri Madenciliği'nde Regresyon Ağaçları ile Sınıflandırma ve Bir Uygulama", (**Yayınlanmamış Doktora Tezi**, Mimar Sinan Güzel Sanatlar Üniversitesi, FBE, 2011), s.68.
[39] Özkan, s.55.
[40] Umut Orhan, "Makine Öğrenmesi", 2012, Çukurova Üniversitesi, Bilgisayar Mühendisliği, http://bmb.cu.edu.tr/uorhan/DersNotu/Ders03.pdf (15 Kasım 2013).

iii. Örnekler sınıflar arasında rastgele dağılmışsa 0<entropi<1 olur. [41]

Buna göre olasılık 1 olduğunda yani kesin durumlarda entropi yani belirsizlik 0 iken, olasılığın eşit olduğu durumlarda ise entropi 1 olup, en yüksek belirsizliğin ortaya çıktığını göstermektedir. Bu durum Şekil 1.4'te gösterilmektedir.



Şekil 1.4: Entropi Değer Aralığı
Kaynak: *Entropy (information theory)*, http://en.wikipedia.org/wiki/Entropy_(information_theory) (15 Kasım 2013).

Buna göre entropinin formülasyonunun anlaşılmasının ardından karar ağaçındaki dallanmanın entropinin alacağı değere göre nasıl belirlendiği şu şekilde ifade edilebilir. Yukarıda da anlatıldığı gibi veri kümesinin sınıf niteliğinin yani bağımlı değişkenin $\{C_1, C_2 ... C_J\}$ olmak üzere J adet sınıfa ayrıldığı varsayılabilir. Bu sınıflarla ilgili olarak ortalama bilgi miktarı belirlenebilir. Buna göre veri kümesindeki gözlem sayısı N ve her j sınıfında bulunan gözlemlerin sayısı da N_j olmak üzere, sınıfların olasılık dağılımı P_j şu şekilde hesaplanır:

$$P_j = \left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_J}{N} \right) \tag{1.7}$$

[41] Dondurmacı, s. 68.

Bu durumda $p_j = \frac{N_j}{N}$ olasılığı ise C_j sınıfında bulunan gözlemlerin küme içindeki oranını verir. O halde C sınıfı için ortalama bilgi miktarı yani entropi şu şekilde gösterilir:[42]

$$H(C) = -\sum_{j=1}^{J} p_j \log_2(p_j) \tag{1.8}$$

1.2.1.1. Kazanç Ölçütü

Karar ağaçlarında hangi değişkene göre dallanmanın yapılacağının belirlenmesinde entoropiye başvurulabileceği anlatılmıştır. Entropi kavramından faydalanarak hesaplanan *"kazanç ölçütü"* dallanma için hangi değişkenin seçilmesi gerektiği belirleyen kriterlerden biridir.

Buna göre; hedef değişkenini yani sınıfı ifade eden C, hedef değişkeni olmayan (yani sınıf değişkeni olmayan) nitel bir X değişkeninin kategori (sınıf) düzeyine (*z*) bağlı olarak $C_1', C_2', ..., C_z'$ *z* adet alt kümeye ayrılır. Buna göre hedef değişkeni C'nin bir sınıfını belirlemek için gerekli bilgi, C_z''nin bir kategori düzeyinin belirlenmesinde gerekli olan bilgilerin ağırlıklı ortalaması olarak kabul edilir. O halde, bu ifadeye göre C'nin bir elemanının sınıfını belirlemek için gerekli bilgi şu şekilde hesaplanır:

$$H(X, C) = \sum_{1}^{z} \frac{N_z}{N} H(C_z') \tag{1.9}$$

Buna göre, C hedef sınıfını nitel bir X değişkenin kategori düzeyine göre ayırma sonucunda elde edilen bilgileri ölçmek için *"kazanç ölçütü"* adı verilen bu ifadeye başvurulur. Bu ölçüt şu şekilde tanımlanır:

$$Kazanç(X, C) = H(C) - H(X, C) \tag{1.10}$$

[42] Özkan, ss. 56-57.

Burada ayırma işlemi yapılırken *Kazanç(X,C)* değerini ***ençoklama*** amaçlanır. Veri kümesi içinde yer alan değişkenlerden en yüksek bilgi kazancını sağlayan, yani kazancı maksimize eden X değişkeni seçilir.[43]

Böylelikle kazanç ölçütü hesabına göre, veri kümesi içerisindeki değişkenlerden hangisinin kazancı en yüksek ise (1.10 nolu formülü en yüksek buluyorsa) dallanma artık o değişkenden devam eder.

1.2.1.2. Kazanç Oranı

Karar ağacının oluşturulması yani dallanmanın hangi değişkene göre gerçekleştileceğinin belirlenmesinde kazanç ölçütü anlatılmıştı. Ancak uygulamada bu yöntemden daha iyi sonuçlar veren entropiye dayalı bir başka yöntem daha kullanılmaktadır. ***Bilgi bölünmesi*** (split information) adı verilen bu kavram Quinlan tarafından ortaya atılmıştır. Burada kazanç ölçütünden faydalanılarak ***kazanç oranı*** hesaplanmaktadır. C hedef sınıfının, nitel bir X değişkeninin değerini belirlemek için gereken bilgi miktarı $H(P_{X,C})$ biçiminde ifade edilir. X değişkeninin kategori düzeyi *z* adet olmak üzere, $P_{X,C}$ ifadesi X değişkeninin her kategori düzeyinde bulunan gözlem sayısına göre hesaplanan olasılık dağılımıdır ve şu şekilde hesaplanır:

$$P_{X,C} = \left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_z}{N} \right) \tag{1.11}$$

Burada $H(P_{X,C})$ miktarı C hedef sınıfına ait veri kümesinde bulunan nitel bir X değişkeni için *bilgi bölünmesi* olarak bilinmektedir. Bu değer şu şekilde hesaplanmaktadır:

$$H(P_{X,C}) = H\left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_z}{N} \right) \tag{1.12}$$

Buna göre entropi formülasyonundan faydalanarak $H(P_{X,C})$ değeri şu şekilde yazılabilir:

[43] Özkan, s.58.

$$H(P_{X,C}) = -\sum_{1}^{z} \frac{N_z}{N} \log_2 \left(\frac{N_z}{N} \right) \quad (1.13)$$

Sonuç olarak, yukarıda verilen $H(P_{X,C})$ değeri ve kazanç ölçütü yardımıyla *kazanç oranı* aşağıdaki gibi hesaplanır:[44]

$$Kazanç\ Oranı(X,C) = \frac{Kazanç(X,C)}{H(P_{X,C})} \quad (1.14)$$

Bu hesaplama sonucu elde edilen kazanç oranlarından en büyük değerli değişken, kök ve sırasıyla bir sonraki düğüm olarak atanır.[45] Yani veri kümesi içerisinde yer alan değişkenlerden hangisinin kazanç oranı en yüksek bulunmuşsa, dallanma için artık o değişken kullanılır. Bu süreç uç düğüme varıncaya kadar devam eder.

1.2.2. Karar Ağaçlarında Safsızlık Fonksiyonuna Dayalı Dallanma Kriterleri

Karar ağaçlarında dallanmaların nasıl olacağı belirlenirken kullanılan kriterlerden bir diğeri ise safsızlık fonksiyonu ya da ölçütüne dayalı kriterlerdir. Teoride birçok safsızlık fonksiyonu olmasına rağmen, iki tanesi yaygın olarak kullanılmaktadır. Bunlar Gini kiteri (ya da gini indeks) ve Twoing Bölme Kriteri olup, ilk olarak Breiman ve ark. tarafından kullanılmıştır.[46] Daha önce de bahsehdildiği gibi bu iki kriter Sınıflama ve Regresyon Ağaçlarında kullanılan ve bölme uyum kriteri olup, bağımlı değişkenin kategorik olması durumunda yani sınıflama ağacı probleminde kullanılır. Bağımlı değişken sürekli olduğunda yani regresyon ağacı problemlerinde ise en küçük kareli sapma yöntemi kullanılır.[47] Bu kriterler anlatılmadan önce safsızlık fonksiyonunun ne olduğunun bilinmesi gerekir.

[44] Özkan, s. 75-76.
[45] Silahtaroğlu, s.82.
[46] Sezgin, s. 31.
[47] IBM SPSS Modeler 15 Algorithms Guide, C&RT Algorithms, 2012, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/AlgorithmsGuide.pdf (29 Mart 2014), s.60.

Safsızlık fonksiyonu düğümün saflığını belirleyen bir fonksiyondur. En iyi bölme uyum kriterlerinin hesaplanabilmesi için ise safsızlık fonksiyonunun belirlenmesi gerekir.

Karar ağaçlarında bir düğümün iki alt kümeye yani iki alt düğüme ayrılabilmesi için bölünmenin yapılacağı en iyi değişkenin bulunması gerekir. Bunun için bir düğüme ait bütün değişkenler ve değişkenlerin olası değerleri araştırılır. Burda her bir düğüm (t) değeri için aday bir bölünme δ değeri vardır ve S bir düğüme ait bölünme değerlerini (δ) barındıran bir aday küme ya da değer kümesi olmak üzere, safsızlık fonksiyonuna dayalı dallanma kriterlerinde en uygun bölünme değerinin bulunması amaçlanır.

Safsızlık fonksiyonu tahmin etmek için bir düğümdeki (t), j sınıfının oranını bilmek gerekir.[48] Bunun için; $\pi(j)$ yani j sınıfının önsel olasılığı hesaplanır.

$$\pi(j) = \frac{N_j}{N} \tag{1.15}$$

Böylelikle her düğüme ait ilgili sınıfa (j) düşen gözlem sayısının, tüm gözlem sayısına bölünmesi ile gözlemin j sınıfında olma olasılığı yani j sınıfının önsel olasılığı bulunmuş olur.

Hesaplanan $\pi(j)$ aşağıdaki formüle yerleştirilirse;

$$p(j,t) = \pi(j)\frac{N_j(t)}{N_j} \tag{1.16}$$

eşitliği bulunmuş olur.

Burda;

$N_j(t)$: j sınıfına ait t düğümünün gözlem sayısıdır.

$p(j,t)$: hem j sınıfı hem de t düğümününde bulunan bir gözlemin olasılığıdır.

[48] Sezgin, s. 30.

Formülasyon (1.15)'te belirtilen $\pi(j)$ değeri, (1.16)'da belirtilen formülasyonda yerine yazılırsa;

$$p(j,t) = \frac{N_j}{N}\frac{N_j(t)}{N_j} \Rightarrow p(j,t) = \frac{N_j(t)}{N} \tag{1.17}$$

değeri bulunmuş olur. Böylelikle $p(j,t)$ formülasyonu ile t düğümününde bulunan ve j sınıfına ait olan bir gözlemin olasılığı hesaplanmış olur.

Ayrıca;

$$p(t) = \sum_j p(j,t) \tag{1.18}$$

Burda;

$p(t)$**:** t düğümüne ait olan herhangi bir gözlemin olasılığıdır.

ve;

$$p(j/t) = \frac{p(j,t)}{p(t)} = \frac{N_j(t)}{N(t)} \tag{1.19}$$

Burda;

$p(j/t)$**:** t düğümündeki j sınıfında olmanın koşullu olasılığıdır. Yani t düğümüne düşmüş olduğu bilinen bir gözlemin j sınıfında olması olasılığı olup bir koşullu olasılıktır. $\sum_j p(j/t) = 1$ olduğu bilinmektedir.[49]

Bir safsızlık fonksiyonu tüm J kümesi ve $(p_1,...,p_J)$ oranları üzerinde tanımlı ϕ 'nin bir fonksiyonu olmak üzere, $p_j \geq 0$, $j = 1,...,J$, $\sum_j p_j = 1$ koşulunu sağlamaktadır. Buna göre safsızlık fonksiyonu aşağıdaki özelliklere sahip $p(j/t)$'nin bir fonksiyonudur.

[49] Breiman ve diğerleri, s.28.

i. $(p_1,...,p_j)$ oranları $(\frac{1}{J},\frac{1}{J}...,\frac{1}{J})$ olduğunda safsızlık fonksiyonu ϕ en büyük depeğerini alır. Yani $p(1/t) = p(2/t) =,...,= p(j/t) = \frac{1}{J}$ iken safsızlık fonksiyonu $\phi(\frac{1}{J},...,\frac{1}{J})$ en büyüktür.

ii. Bir t düğümünde sadece bir sınıfın büyük bir çoğunluğu varsa, o zaman safsızlık fonksiyonu $\phi(p(1/t),..., p(J/t))$ en küçük değerini alır. Bu durum şöyle ifade edilebilir: ϕ, en küçük değerine (1,0,…,0), (0,1,0,…0),…,(0,0,…0,1) sadece bu noktalarda sahip olur.

iii. ϕ, $(p_1,..., p_J)$'nin simetrik bir fonksiyonudur. Yani t düğümündeki j sınıfının olasılığı simetrik bir fonksiyondur.

Buna göre, bir safsızlık fonksiyonunun ϕ, her bir düğümüne (t) ait safsızlık ölçütü i(t) aşağıdaki gibi tanımlanır:[50]

$$i(t) = \phi(p(1/t),..., p(J/t)) \qquad (1.20)$$

(1.20)'de görüldüğü üzere safsızlık fonksiyonu $p(j/t)$'nin bir fonksiyonudur. Bu fonksiyon, bütün sınıflar eşit olasılıklı dağılmış iken maksimum değerini alırken, düğümde sadece bir sınıf olduğunda en küçük değerini alır.

En iyi bölünmenin belirlenebilmesi için, düğümlerin homojenliğinin ölçülmesi gerekir. Safsızlık fonksiyonundaki değişimin maksimum olması ile en iyi bölünme kuralının seçilmesi sağlanır.[51] Buna göre her bir ana düğümün safsızlık derecesi kendi alt düğümlerinin safsızlık derecesi ile karşılaştırılır. Ana ve alt düğümlerin safsızlıkları arasındaki farkın büyük olması istenir.

[50] Breiman ve diğerleri, s.32.
[51] Sezgin, s. 31.

Buna göre, belli bir δ bölünme değeri için bir t düğümünün safsızlık fonksiyonundaki değişim yani safsızlıktaki azalış aşağıdaki gibi tanımlanır:[52]

$$\Delta i(\delta,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{1.21}$$

Burda;

p_L: sol düğümdeki (t_L) gözlemlerin oranı (olasılığı), yani t ana düğümden sol tarafa ayrılan gözlemler.

p_R: sağ düğümdeki (t_R) gözlemlerin oranı (olasılığı), yani t ana düğümden sağ tarafa ayrılan gözlemler.

Böylelikle en iyi bölme kriteri olan $\phi(\delta,t)$ yani bir diğer adıyla safsızlık fonksiyonu, $\Delta i(\delta,t)$'ye yani bir δ bölünme değeri için bir t düğümünün safsızlık fonksiyonundaki değişim ifadesine denk gelmektedir. $\phi(\delta,t)$ değerini maksimize eden δ^* değeri için t düğümünün bölünmesi gerçekleştirilir.

Buna göre karar ağacının oluşması aşamasında ilk düğüm yani kök düğüm olan t_1'den başlıyarak aday bölünme değerleri arasında safsızlıktaki azalışı en büyük yapan δ^* bölünme değeri için t_1 düğümünün bölünmesi gerçekleştirilir.

$$\Delta i(\delta^*,t_1) = \max_{\delta \in S} \Delta i(\delta,t_1) \tag{1.22}$$

Böylelikle kök düğüm olan t_1 en iyi δ^* bölünme değerine göre t_2 ve t_3 düğümlerine bölünür. Aynı süreç, en iyi $\delta \in S$ değeri için t_2 ve t_3 düğümlerine ayrı ayrı uygulanır.[53]

Safsızlık fonksiyonu anlaşıldıktan sonra bundan sonraki aşama en iyi bölünmenin bulunmasını sağlayan safsızlık fonksiyonunu $i(t)$ ölçütleri olan gini ve twoing kriterlerini belirlemektir.

[52] Breiman ve diğerleri, s.25.
[53] Breiman ve diğerleri, s. 26.

1.2.2.1 Gini Kriteri

Sınıflama ve regresyon ağaçlarında her bir düğüm her aşamada belli bir kriter uygulanarak ikiye ayrılmaktadır. Bunun için tüm değişkenlerin sahip olduğu değerler göz önüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir. Bu kriter değişken değerlerinin sol ve sağda olmak üzere iki bölüme ayrılması esasına dayanır.[54] Düğümlerin ikiye ayrılabilmesi için bölünme noktalarının belirlenmesi gerekir. En iyi bölünmeyi seçmek için geliştirilen gini kriteri ya da bir diğer adıyla gini indeksi bir safsızlık ölçütü (fonksiyonu) olarak da anlatılmaktadır. Daha önce belirtildiği gibi $p(j/t)$ ve $j=1,...J$, t düğümündeki j sınıfının olasılığı olmak üzere, t düğümünün safsızlığının ölçütü $i(t)=\phi(p(1/t),...,p(J/t))$ formülasyonu ile tanımlanır. Düğümün ve doğal olarak ağacın safsızlığını düşüren bölünme araştırılır. Böylelikle gini kriterine göre oluşturulan safsızlık fonksiyonu aşağıdaki gibi tanımlanır.[55]

$$i(t)=\sum_{j\neq i} p(j/t)\,p(i/t) \tag{1.23}$$

Gini indeksi aynı zamanda aşağıdaki formülasyon ile de gösterilebilir.

$$i(t)=(\sum_{j} p(j/t))^2-\sum_{j} p^2(j/t)=1-\sum_{j}^{J} p^2(j/t) \tag{1.24}$$

$p(j/t)=\dfrac{N_j(t)}{N(t)}$ olduğu için, (1.24) nolu formülasyonda yerine yazılırsa,

$$i(t)=1-\sum_{j}^{J}\left(\frac{N_j(t)}{N(t)}\right)^2 \tag{1.25}$$

t düğümüne ait gini indeksi bulunur. Bu formül t düğümünün sol ve sağ düğümüne atanan değişkenin kategori değerleri için ayrı ayrı hesaplanır.

İki sınıflı problemlerin çözümü için ise, indeks aşağıdaki formülasyon ile çözümlenebilir.

[54] Özkan, s.89
[55] Breiman ve diğerleri, s. 103.

$$i(t) = 2p(1/t)p(2/t) \tag{1.26}$$

Buna göre, (1.21) nolu formülasyonda verilen safsızlık fonksiyonundaki değişim $\Delta i(\delta,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ olmak üzere, bu formül (1.24) nolu formüle göre yeniden düzenlenirse safsızlık fonksiyonundaki değişim aşağıda gösterildiği gibi olur;

$$\Delta i(t) = 1 - \sum_j^J p^2(j/t) - p_L(1 - \sum_j^J p^2(j/t_L)) - p_R(1 - \sum_j^J p^2(j/t_R)) \tag{1.27}$$

Ayrıca $p_L + p_R = 1$ olduğundan (1.27) nolu formülasyonda yerine yazılırsa;

$$\Delta i(t) = -\sum_j^J p^2(j/t) + p_L \sum_j^J p^2(j/t_L) + p_R \sum_j^J p^2(j/t_R) \tag{1.28}$$

(1.28) nolu eşitlik elde edilir. Sonuçta gini indeksinin $\phi(p_1,...,p_J)$'nin bir fonksiyonu olduğu görülür. $\Delta i(\delta,t)$ yani bir δ bölünme değeri için bir t düğümünün safsızlık fonksiyonundaki değişim en iyi bölme kriteri olan $\phi(\delta,t)$ yani bir diğer adıyla safsızlık fonksiyonu, ifadesine denk gelmektedir. Buna göre, $\phi(\delta,t)$ değerinin maksimum yapan δ^* değeri için bölünme yapılır.

Daha önce de bahsedildiği gibi en iyi bölme kuralı safsızlık fonksiyonundaki değişimi maksimize eden gini değerine göre belirlenir. Maksimum gini değişimini veren değişkenin ilgili kategorisine göre bölünme yapılır.[56]

Eğer gözlemlerin kategorileri bir düğümde aynı oranda dağılmışsa gini indeksi maksimum değerini (1-1/J) alır. Burda J gözlemlerin kategori sayısıdır. Bunun tersi durumunda yani bir düğümdeki gözlemler aynı (eşit) kategori değerine sahipse gini indeksi sıfıra eşit olur.[57]

[56] Sezgin, s.32.
[57] IBM SPSS Modeler 15 Algorithms Guide, C&RT Algorithms, 2012, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/AlgorithmsGuide.pdf (29 Mart 2014), s.61.

1.2.2.2. Twoing Bölme Kriteri

Twoing bölme kuralı her düğüm için sınıfları sol ve sağ olmak üzere iki alt düğüme ayırır ve her seferinde safsızlığı hesaplar. Her düğüm için en iyi bölünme safsızlık fonksiyonundaki değişimi maksimize eden değerdir.

Daha önce bahsedildiği üzere çok sınıflı problemlerde gözlemlerin J adet sınıfa atandığını varsaydığımızda sınıf sayısı 1,2,…J adet olmak üzere, C sınıf seti olarak adlandırılır ve C = {1,2,…,J} ile gösterilir. Her düğümde sınıflar sağ ve sol olmak üzere iki üst sınıfa (superclasses) atanır. Gözlemlerin {g=1,2,…,k,…n} bir kısmının atandığı sınıf C_1, geri kalanının atandığı sınıf C_2 olup; $C_1 = \{1,2,...k\}$, $C_2 = C - C_1$ yani $C_2 = \{n-1, n-2, ..., n-k\}$'dir. Buna göre, $\Delta i(\delta, t)$ yani bir δ bölünme değeri için bir t düğümünün safsızlık fonksiyonundaki değişim C_1'in seçimine bağlı olur ve notasyon $\Delta i(\delta, t; C_1)$ ile ifade edilir. Problem artık $\Delta i(\delta, t; C_1)$ değerini maksimize eden $\delta^*(C_1)$ bölünmesini bulmaya dönüşür ve ilgili C_1^* bulunur. Bu durumda notasyon $\Delta i(\delta^*(C_1), t; C_1)$ ile ifade edilir ve bölünme $\delta^*(C_1^*)$ durumunu sağlayan düğüm için gerçekleştirilir. Bu mantık sınıfları ikiye ayırmak için aynı şekilde uygulanır.[58] Buna göre twoing bölme kuralının teoremini şu şekilde açıklayabiliriz;

İki sınıf kriteri $p(1/t)p(2/t)$ altında bir δ bölünme değeri ve bir üst sınıf $C_1(\delta)$ değeri için $\Delta i(\delta^*(C_1), t; C_1)$ ifadesi maksimize edilir. Burda $C_1(\delta)$:

$$C_1(\delta) = \{j : p(j/t_L) \geq p(j/t_R)\} \tag{1.29}$$

ile ifade edilir. Buna göre twoing bölme kuralına göre hesaplanan safsızlık değişimi aşağıdaki gibi formüle edilir.

$$\max_{C_1} \Delta i(\delta, t, C_1) = \frac{P_L P_R}{4} \left[\sum_{J=1}^{J} p(j/t_L) p(j/t_R) \right]^2 \tag{1.30}$$

[58] Breiman ve diğerleri, s. 104-105.

Neticede herhangi bir t düğümü ve δ bölme değeri için t düğümü t_L ve t_R olarak ikiye ayrılır ve twoing bölme kriteri fonksiyonu $\phi(\delta,t)$:

$$\Delta i(t) = \phi(\delta,t) = \frac{P_L P_R}{4}\left[\sum_{J=1}^{J} p(j/t_L)\,p(j/t_R)\right]^2 \qquad (1.31)$$

formülasyonu ile ifade edilir. Eğer J adet sınıf mevcut ise, algoritma C sınıfını ikiye bölmek için 2^{J-1} adet olası bölme araştırır.[59]

1.2.2.3. En Küçük Kareli Sapma

Sınıflama ve Regresyon Ağaçlarında kullanılan bir diğer bölme uyum kriteri bağımlı değişken sürekli olduğunda yani regresyon ağacı problemlerinde kullanılan ve safsızlık fonksiyonu dayalı bölme kriteri olan en küçük kareli sapma yöntemidir. Regresyon modelleri oluşturmada, model parametrelerini elde etmede bu yöntem ile onun alternatifi olan en küçük mutlak sapma yöntemi kullanılır. En küçük kareli sapma yöntemi ile en küçük kare hata (least squares error) kriteri minimize edilerek parametreler tahmin edilir.[60]

Regresyonda bir gözlem veri setinde (**x**,y) formatındadır. Yani **x** değeri, ilgili gözlemin ölçüm vektörü olup, **X** ölçüm uzayında y gerçek değeri ile bulunmaktadır. Değişken y, yanıt değişkeni ya da bağımlı değişken; **x** değişkenleri ise bağımsız ya da tahmin değişkeni olarak adlandırılmaktadır. Bir kestirim kuralı, **X** ölçüm uzayınında tanımlı bir d(x) fonksiyonudur. Burada notasyonda dikkat edilmesi gereken husus **X** matrisi ve dolayısıyla **x** vektörü üzerinde tanımlanan d(**x**) fonksiyonu; sınıflama ağaçlarında sınıflayıcı (classifier) olarak adlandırılırken, regresyon ağaçlarında kestirimci (predictor) olarak adlandırılmaktadır.

Buna göre amaç başlangıç veri setinden L başlayarak d(x) kestirimcisini oluşturmaktır. d(x) kestirimcisinin oluşturulmasının iki amacı vardır:

[59] Breiman ve diğerleri, s. 107-108.
[60] *Tree-based Regression,* (t.y.) http://www.dcc.fc.up.pt/~ltorgo/PhD/th3.pdf (28.02.2014).

i. Gelecekte ortaya çıkan ölçüm değerlerini mümkün olacak en doğru biçimde tahmin edebilmek için yanıt değişkenini tahmin etmek.

ii. Yanıt değişkeni ile ölçüm değişkeni arasındaki yapısal ilişkiyi ortaya koyabilmek.

Buna göre N adet gözlemden oluşan $\{(\mathbf{x}_1,y_1)_{,\ldots,}(\mathbf{x}_N,y_N)\}$ bir başlangıç veri kümesi ile d(x) kestirimcisinin bulunmaya çalşıldığı düşünüldüğünde, kestirimcinin doğruluğunun nasıl ölçüleceği sorusuna cevap aranır. Bu yüzden eğer N_2 boyutunda çok büyük $(x'_1, y'_1),...,(x'_{N_2}, y')$ bir test örneği mevcut ise, d(x) kestirimcisinin doğruluğu aşağıda formülasyonu verilen ortalama hata (average error) ile ölçülebilir.

$$\frac{1}{N_2}\sum_{n=1}^{N_2}\left|y'_n - d(x'_n)\right| \tag{1.32}$$

Burada $d(x'_n)$, $n=1,...N_2$ olmak üzere y'_n'in bir bir tahmincisidir. Bu ölçüm yukarıda da bahseldiği gibi en küçük kareler yönteminin alternatif yöntemi olan, ***en küçük mutlak sapma*** yöntemidir. Fakat regresyonda hesap kolaylığı sağlamak üzere doğruluk ölçütü için genel olarak hata kareleri ortalaması (averaged squared error) kullanılır. Hata kareler ortalaması aşağıdaki formülasyon ile gösterilir.

$$\frac{1}{N_2}\sum_{n=1}^{N_2}\left(y'_n - d(x'_n)\right)^2 \tag{1.33}$$

Regresyon ağaçları ile ağaç oluşum süreci en küçük kareler yöntemi üzerinde döner. Hata kareler ortalamasının doğruluğunu belirleyebilmek için teorik bir çatıya ihtiyaç vardır. Bunun için başlangıç veri kümesinin (L) ve raslantısal **(X**, Y) vektörünün birbirinden bağımsız olarak aynı dağılımdan alınmış olduğu varsayıldığında, d kestirimcisinin hata kareler ortalaması $R^*(d)$ aşağıdaki gibi ifade edilir.

$$R^*(d) = E\left(Y - d(X)\right)^2 \tag{1.34}$$

Burda *d(**X**),* Y'nin kestirimcisi olup $R^*(d)$ ise, hata karelerin beklenen değeridir. Sınıflama ağaçlarında ise bir sınıf değeri tahmin edildiği için $R^*(d)$ değeri, d kestirimcisinin hatalı sınıflama oranını ifade eder.[61] Oysa regresyon ağaçlarında bir sınıf değerini tahmin etme yerine, kestirimcinin doğruluğunun ölçüsü belirlenir. Bu durumda (1.34)'te verilen formülasyondan hareketle $R^*(d)$ değerini minimize eden optimal d_B kestirimcisi $(\widehat{Y})$;

$$d_B = E(Y / X = x) \tag{1.35}$$

biçiminde ifade edilir ve d_B ilgili **x** ölçüm vektörüne yönelik yanıtın koşullu beklenen değeridir.[62]

Böylelikle regresyon ağaçlarında başlangıç veri kümesi aracılığıyla bir kestirimci *d(**x**)* ve **onun hata tahmini olarak ifade edilen** $R^*(d)$ değeri tahmin edilir. R^* tahmin etmek için birçok yol bulunmaktadır. Yerine koyma (resubstitution) en yaygın olan tahmin yöntemidir.[63] Diğer tahmin yöntemi ise ağırlıklandırılmış varyans yöntemidir. Bunun için ilk olarak yerine koyma tahmin yöntemi anlatılmıştır.

Yerine koyma hatası aşağıdaki gibidir:

$$R(d) = \frac{1}{N}\sum_{n=1}^{N}(y_n - d(x_n))^2 \tag{1.36}$$

Bu eşitlikten anlaşıldığı üzere en iyi bölünme değeri, düğüm içi kareler toplamının ortalamasını veya yerine koyma tahminini minimize eden değerdir.[64] $R(d)$ değerini minimize etmek için y(t) değeri kullanılır. y(t) değeri t düğümüne düşen bütün gözlemler (x_n, y_n) için y_n'in ortalamasıdır. Bu durumda y(t)'nin minimize edilmesi aşağıdaki gibidir:

[61] Breiman ve diğerleri, s. 221-222.
[62] Breiman ve diğerleri, s. 222-223.
[63] Breiman ve diğerleri, s. 223.
[64] Sezgin, s.36.

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{x_n \in t} y_n \tag{1.37}$$

Burada, y_n t düğümündeki bağımlı değişkenlerdir. y_n'in tüm değerlerinin toplamı, $x_n \in t$ ve *N(t)* t düğümünde bulunan gözlemlerin toplam sayısı, olmak üzere $\bar{y}(t)$ değeri, t düğümündeki gözlemlerin ortalaması yani y_n'in ortalamasıdır. (1.37)'de verilen önermenin ispatı $\sum_n (y_n - a)^2$ ifadesini minimize eden a sayısının bulunması esasına dayanır. Buna göre minimum a sayısını elde etmek için $\sum_n (y_n - a)^2$ ifadesinin a'ya göre türevi alınarak sıfıra eşitlenir.

$$-2\sum_{n=1}^{N} y_n + 2aN = 0 \tag{1.38}$$

Buna göre;

$$a = \frac{1}{N} \sum_n y_n \tag{1.39}$$

İfadesi bulunur. Benzer şekilde y_n'in her alt kümesi y'_n için a sayısının, $\sum_{n'} (y_{n'} - a)^2$ ifadesi için minimize edilmesi ile bu kez y'_n'in ortalaması olur. Böylelikle bu aşamadan itibaren her t düğümünün kestirim değeri yerine $a = \bar{y}(t)$'nin kullanılabileceği görülmektedir. Bu durumda $R(d)$ yerine, **T** ile ifade edilen karar ağacının yerine koyma hatası olan *R(T)* notasyonu kullanılır. Sonuç olarak formülasyonlar aşağıdaki gibi düzenlenebilir:

$$R(T) = \frac{1}{N} \sum_{t \in \tilde{T}} \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \tag{1.40}$$

ve,

$$R(t) = \frac{1}{N} \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \tag{1.41}$$

Böylece (1.40)'ta bulunan formülasyon aşağıdaki gibi yazılabilir:

$$R(T) = \sum_{t \in \tilde{T}} R(t) \tag{1.42}$$

(1.41)'deki formülasyonun basit bir yorumu vardır. Her t düğümü için $\sum_{x_n \in t} (y_n - \bar{y}(t))^2$ düğüm içi kareler toplamıdır. Bu ifade, y_n'in yani t düğümündeki bağımlı değişkenlerin, $\bar{y}(t)$ yani t düğümünün tahmin değerinden (yani kendi ortalamasından) sapmasının kareleri toplamıdır. $\tilde{T}$ ağaçtaki uç düğümlerin oluşturmuş olduğu bir kümeyi ($t \in \tilde{T}$) ifade etmek üzere, düğüm içi kareler toplamının N ile bölümü, ortalamayı verir.

Buna göre, S bütün aday bölünme değerlerini içinde barındıran bir küme ve $t \in \tilde{T}$ olmak üzere t düğümüne ait en iyi bölünme δ^* değeri sınıflama ağaçlarında olduğu gibi *R(T)*'deki azalmayı yani değişimi maksimize eden değerdir. Daha detaylı anlatılacak olursa; her bir δ bölünme değeri için bir t düğümünün sol t_L ve sağ t_R düğümlerindeki azalış aşağıdaki gibi tanımlanır:

$$\Delta R(\delta, t) = R(t) - R(t_L) - R(t_R) \tag{1.43}$$

Böylelikle regresyon ağacı *R(T)*'deki azalışı maksimize etmek için tekrarlı yani ardışık bir biçimde bölünerek oluşturulur. Aday bölünme değerleri arasında azalışı en büyük yapan δ^* bölünme değeri için t düğümünün yerine koyma tahmindeki değişim;

$$\Delta R(\delta^*, t) = \max_{\delta \in S} \Delta R(\delta, t) \tag{1.44}$$

biçiminde de ifade edilir.[65] Bu ifadeyi maksimum yapan δ^* değeri için bölünme gerçekleştirilir.

Başlangıçta da belirtildiği üzere yerine koyma tahminin alternatif yolu ağırlıklandırılmış varyansı kullanmaktır. Buna göre, $p(t) = N(t)/N$ ifadesi t düğümüne

[65] Breiman ve diğerleri, s. 230-231.

düşen raslantısal olarak seçilmiş bir gözlemin olasılığı için yerine koyma tahmini olsun. Buna göre t düğümünün varyansı aşağıdaki gibidir.

$$s^2(t) = \frac{1}{N(t)} \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \tag{1.45}$$

Böylece $R(t) = s^2(t)\,p(t)$ ve buradan;

$$R(T) = \sum_{t \in \tilde{T}} s^2(t)\,p(t) \tag{1.46}$$

$s^2(t)$, t düğümündeki y_n değerlerinin örnek varyansıdır. Böylece t düğümüne ait en iyi bölünmenin ağırlıklı varyansı minimize etmesi beklenir.

$$p_L s^2(t_L) + p_R s^2(t_R) \tag{1.47}$$

(1.47)'de görüldüğü gibi ağırlıklar sırasıyla sol (p_L) ve sağ (p_R) düğümde bulunan gözlemlerin oranıdır. Bu ifadeyi minumum yapan bölünmeler seçilerek regresyon ağacı oluşturulur.[66]

Bir başka ifadeyle ağırlıklandırılmış varyanstaki değişim aşağıdaki gibi olmak üzere;

$$\Delta s^2(\delta, t) = s^2(t) - p_L s^2(t_L) - p_R s^2(t_R) \tag{1.48}$$

Ağırlaklandırılmış varyanstaki değişimi maksimum yapan bölünme en iyi bölünme olarak seçilir.[67]

Sonuçta regresyon ağaçlarında bağımlı değişken sürekli olduğu için kategori değerleri içermemektedir. Bu yüzden regresyon ağacı yönteminde sınıflama ağaçlarında kullanılan bölme kiterleri olan, gini ve twoing kriterleri kullanılmaz. Regresyon ağacındaki bölünmeler ikili olarak sonuçlanan düğüm için tahmin edilen toplam

[66] Breiman ve diğerleri, s. 231-232.
[67] Segin, s.37.

varyansın minimize olmasının gerekliliği anlamına gelen "artıkların karelerini azaltma" esasınına göre gerçekleştirilir.[68]

Ağaç oluşum sürecinde, dallanmanın gerçekleştirilebilmesi için kullanılan kriterler genel olarak anlatıldıktan sonra oluşan ağacın budanması devamında anlatılmıştır.

1.3. Karar Ağaçlarının Budanması

Karar ağaçları çoğu kez karmaşık bir görünüme sahip olabilir. Ağaç oluşum sürecinde yapılan önemli adımlardan biri de budama işlemidir. Budama ile ağaçta bulunan sonucu etkilemeyen ve sınıflamaya herhangi bir katkısı olmayan dalların ağaçtan alınması sağlanır. Böylelikle ağacın içinde gereksiz detayların olması engellenir. Çünkü ağaçta birçok düğüm ve dal oluşursa, ağacın alt dalları ve yapraklarına ulaşan veri sayısı da azalacaktır. Bu da ağacın hassasiyetini azaltacaktır.[69] Buna göre, bir karar ağacında bir alt ağacı atarak yerine bir yaprağın yani uç düğümün yerleştirilmesi söz konusu olabilir. Böylelikle ağacın derinliği yani aşağı doğru büyümesi (dallanması) azaltılmış olur. Bu şekilde yapılan işleme *"karar ağacının budanması"* adı verilir. Alt ağacın yerine yaprak yerleştirmek ile, algoritma *"öngörülü hata oranını"* yani hatalı sınıflama oranını azaltmayı ve sınıflandırma modelinin kalitesini arttırmayı amaçlar. Öngörülü hata oranını hesaplamak için artık eğitim verileri kümesi kullanılmaz. Onun yerine test verilerinden olan yeni bir küme ile çalışılır. Test örnekleri ile oluşturulan ağaçta sınıflama doğruluğuna katkısı olmayan ağaçların çıkarılması budama işleminin temelini oluşturur. Böylece daha az karmaşık bir ağaç oluşturulmuş olur. Sınıflama algoritmalarında budama yapmak yani özyinemeli bölme işlemine son vermek için iki yol kullanılır:

i. Bölme işlemine son verme, yani durdurma ölçütü olarak χ^2 gibi istatistik testlerinden faydalanılabilir. Eğer bölünme öncesinde ve sonrasında kayda değer bir fark yoksa, söz konusu düğüm bir yaprak olarak gösterilir. Yani ağacın dallara ayrılmasından önce karar verilir. Bu yüzden bu tekniğe *"ön budama"* denir.

[68] Kürşad Özkan, "Sınıflandırma ve Regresyon Ağacı Tekniği (SRAT) ile Ekolojik Verinin Modellenmesi", Süleyman Demirel Üni. Orman Fak. Dergisi, Cilt.13, Sayı.1-4 (2012), s.2.
[69] Silahtaroğlu, s.74.

ii. Bazı ağaçlarda ise, bir *"doğruluk ölçütü"* seçilerek ağaç budanabilir. Bu yöntemde ise, budama işlemi ağaç oluşturulduktan sonra yapılmış olur.[70]

Budamanın gerçekleştirilmesi için kullanılan algoritmanın işleyiş biçimi budamanın hızı açısından önemlidir. Kullanılan algoritmaların çoğunda varsayılan (default) değer olarak %5-%30 arası değerlerden düşük anlamlılık gösteren değerler budanırken, bu anlamlılığın belirlenmesi kullanıcıya bırakılmaktadır. Bu işlem ile ağacın oluşumundan sonra budama işleminin gerçekleştirilmesi söz konusudur. Yukarıda da bahsedildiği üzere budama, gerek ağacın kurulumu esnasında gerekse de kurulduktan sonra yapılabilir. Eğer, ağaç oluşmadan önce belirli bir saflık düzeyine göre işlem yapılırsa ağacın kurulması esnasında budama işlemi gerçekleştirilmiş olur. Bu işlemde ise ağaç kurulmadan önce belirlenen bir saflık değeri gözönüne alınarak ağacın yaprakları belirli bir yüzde değerdeki, örneğin %70, %95 gibi saflık değerine ulaşınca dek ağaca yaprak ataması yapılır.[71] Buna göre karar ağaçlarının budanmasındaki genel amacın, optimum boyuttaki ağaca ulaşmak olduğu söylenebilir.

Buraya kadar karar ağaçları ile ilgili anlatılan genel bilgilerden yola çıkarak, ağaç oluşum sürecinde farklı yöntemleri kullanan karar ağaçlarının olduğu anlaşılmaktadır. Bundan sonraki aşamada ise, bu karar ağaçlarından bahsedilmiştir.

1.4. Karar Ağacı Yöntemleri

Karar ağaçları yöntemleri, temel anlamda hedef (bağımlı) değişkeni, tahmin edici değişkenlere göre ayırma mantığına dayansa da, bünyesinde değişik amaçlara hizmet eden birbirinden farklı algoritmalara sahiptir.[72] Karar ağaçlarına dayalı olarak geliştirilen birçok algoritma vardır. Bu algoritmalar kök, düğüm ve bölme kriteri seçimlerinde izledikleri yola göre birbirinden farklılık gösterir. Kullanılan algoritmaya göre ağacın şekli değişebilir. Bu durumda değişik ağaç yapıları ortaya çıkar ve farklı sınıflama sonuçları oluşur. Kök denilen ilk düğümün farklı olması, en uçtaki yaprağa ulaşırken izlenecek yolu ve dolayısıyla sınıflamayı da değiştirecektir.[73]

[70] Özkan, s.83.
[71] Silahtaroğlu, s.74.
[72] Dondurmacı, s.26.
[73] Silahtaroğlu, s.74.

Karar ağaçlarında kullanılan bir takım teknik ve algoritmalar sadece bazı türdeki verilerle çalışabilirler. Bazı algoritmalar sadece sayısal değerlerle, bazıları sadece kategorik değerlerle, bazıları ise 0-1 değerlerle işlem yaparlar. Bu durumda mevcut veri kullanılacak algoritmaya uygun hale getirilmelidir.[74] Bu durumda hangi algoritmanın uygulanacağı ise, hem mevcut veri türüne hem de amaç ve uygulamaya bağlıdır. Karar ağaçlarında genel olarak kullanılan algoritmalar aşağıdaki tabloda gösterildiği gibidir.

Tablo 1.3 Karar Ağacı Algoritmaları

Karar Ağacı Algoritması	Gelişim Süreci	Veri tipi	Çalışma Mekanizması	Oluşan Ağaç Modeli	Ağaç Dallanma Kriteri
AID Algoritması	Morgon ve Sonquist-1964	Bağımlı değişken nicel (aralıklı/oransal) , bağımsız değişken kategorik (nominal /ordinal) olmalıdır.	Ağaçta sadece ikili bölünmeler olur. Yani ikili ağaçlar oluşturur.	Bağımlı değişken nicel olduğu için regresyon ağacı oluşur.	Gruplar arası kareler toplamı kriterini kullanır.
CHAID Algoritması	Gordon V. Kass-1980	Bağımlı değişken kategorik (ordinal / nominal) veya sürekli, bağımsız değişken kategorik (nominal /ordinal) olmalıdır.	Ağaç çoklu alt gruplara ayırılır. Yani çoklu ağaçlar oluşturur.	Bağımlı değişken kategorik ise, sınıflama ağacı-sürekli ise regreyon ağacı oluşur.	Bağımlı değişken sürekli ise F testini-Bağımlı değişken nominal ise pearson ki-kare testini, bağımlı değişken ordinal ise en çok olabilirlik testini kullanır.
CART Algoritması	Breiman, Friedman, Olshen ve Stone-1984	Bağımlı değişken kategorik (ordinal / nominal) veya sürekli olabilir. Bağımsız değişken çeşidi ayrımı yoktur.	Ağaçta sadece ikili bölünmeler olur. Yani ikili ağaçlar oluşturur.	Bağımlı değişken kategorik ise, sınıflama ağacı-sürekli ise regreyon ağacı oluşur.	Bağımlı değişken kategorik ise, gini ya da twoing kriteri, bağımlı değişken sürekli ise EKK yöntemi kulanılır.
ID3 Algoritması	Quinlan-1986	Bağımlı ve bağımsız değişkenler kategoriktir(ordinal / nominal).	Ağaç çoklu alt gruplara ayırılır. Yani çoklu ağaçlar oluşturur.	Bağımlı değişken kategorik olduğu için sınıflama ağacı oluşur.	Kazanç ölçütü kriterini kullanır.
MARS Algoritması	Jerome H. Friedman-1991	Bağımlı değişken ikili (binary) veya sürekli olabilir. Bağımsız değişken çeşidi ayrımı yoktur.	Ağaçta sadece ikili bölünmeler olur. Yani ikili ağaçlar oluşturur.	Bağımlı değişken kategorik ise, sınıflama ağacı-sürekli ise regreyon ağacı oluşur.	İleriye ve geriye doğru adım algoritmalarını kullanır.

[74] Tapkan, s.249.

Karar Ağacı Algoritması	Gelişim Süreci	Veri tipi	Çalışma Mekanizması	Oluşan Ağaç Modeli	Ağaç Dallanma Kriteri
Exhaustive CHAID Algoritması	Biggs, de ville ve Suen-1991	Bağımlı değişken kategorik (ordinal / nominal) veya sürekli, bağımsız değişken kategorik (nominal /ordinal) olmalıdır.	Ağaç çoklu alt gruplara ayırılır. Yani çoklu ağaçlar oluşturur.	Bağımlı değişken kategorik ise, sınıflama ağacı-sürekli ise regreyon ağacı oluşur.	Bağımlı değişken sürekli ise F testini-Bağımlı değişken nominal ise pearson ki-kare testini, bağımlı değişken ordinal ise en çok olabilirlik testini kullanır.
C4.5 Algoritması	Quinlan-1993	Bağımlı değişken kategorik (ordinal / nominal), bağımsız değişkenler ise kategorik veya sürekli olabilir.	Ağaç çoklu alt gruplara ayırılır. Yani çoklu ağaçlar oluşturur.	Bağımlı değişken kategorik olduğu için sınıflama ağacı oluşur.	Kazanç oranı kriterini kullanır.
C5.0 Algoritması	Quinlan-1993	Bağımlı değişken kategorik (ordinal / nominal), bağımsız değişkenler ise kategorik veya sürekli olabilir.	Ağaç çoklu alt gruplara ayırılır. Yani çoklu ağaçlar oluşturur.	Bağımlı değişken kategorik olduğu için sınıflama ağacı oluşur.	Kazanç oranı kriterini kullanır.
SLIQ Algoritması	Mehta, Agarwall ve Rissanen-1996	Bağmlı değişken kategorik (ordinal / nominal), bağımsız değişkenler kategorik veya sürekli olabilir.	Ağaç çoklu alt gruplara ayırılır. Yani çoklu ağaçlar oluşturur.	Bağımlı değişken kategorik olduğu için sınıflama ağacı oluşur.	Gini kriterini kullanır.
SPRINT Algoritması	Shafer, Agarwall ve Mehta-1996	Bağmlı değişken kategorik(ordinal / nominal), bağımsız değişkenler kategorik veya sürekli olabilir.	Ağaç çoklu alt gruplara ayırılır. Yani çoklu ağaçlar oluşturur.	Bağımlı değişken kategorik olduğu için sınıflama ağacı oluşur.	Gini kriterini kullanır.
OUEST Algoritması	Loh ve Shih-1997	Bağımlı değişken kategorik (nominal), bağımsız değişkenler kategorik(ordinal / nominal) veya sürekli olabilir.	Bağımlı değişkeni sadece ikiye böler. Yani ikili ağaçlar oluşturur.	Bağımlı değişken kategorik olduğu için sınıflama ağacı oluşur.	Kuadratik diskriminant analizini kullanır.

Kaynak: Yazar tarafından çeşitli eserlerden yararlanılarak oluşturulmuştur.

Buna göre, geçmişten günümüze karar ağaçlarında kullanılan algoritmaları sırasıyla inceleyebiliriz.

1.4.1. AID Algoritması

Karar ağaçlarının ilk temelleri AID (Automatic Interaction Detector: Otomatik Etkileşim Belirleyicisi) yöntemi ile atılmıştır. Bu teknik 1964 yılında Michigan Üniversitesi İnceleme ve Araştırma Merkezi'nde Morgon ve Sonquist tarafından geliştirilmiştir. Bu teknikte amaç veri kümesini, nicel bir bağımlı değişkene göre,

sınıflayıcı (nominal) ve sıralayıcı (ordinal) bağımsız değişkenler kullanarak ikiye bölmektedir.[75]

AID yönteminin kullanım şartı bağımsız değişkenlerin sınıflayıcı ölçekte ve bağımlı değişkenlerin ise aralıklı ya da oransal ölçekte ölçülmüş olmasıdır. Bağımlı değişkenlerin dağılımı ya da bağımsız değişkelerin fonksiyonel formu ile ilgili herhangi bir varsayım AID yönteminde yoktur. AID yönteminin amacı anakütleyi bağımlı değişken ile ilgili en mümkün istatistiksel alt gruplara bölmektir. Belirli segment ya da gruplar için homojenlik gruptaki her bireysel grup ve grup ortalaması arasındaki farkın kareleri toplamının hesaplanması ile ölçülür. Bu da grup içi değişkenliğin ölçüsü olan gruplar içi kareler toplamıdır. Gruplar içi kareler toplamının küçük olması yüksek homojenliği göstermektedir. [76]

AID algoritması bigisayar destekli (computer-intensive) bir regresyon ağacı olup, veri içindeki çok boyutlu örüntüleri ve ilişkileri bulur. Ayrıca hiçbir varsayım içermediğinden regresyon analizinin nonparametrik bir alternatifidir.[77] Regresyon analizinde asıl amaç aşağıdaki formülde gösterilen değişkenler arasındaki ilişkiyi saptamaktır.

$$y = f(x_1,...,x_n) + e \tag{1.49}$$

Regresyon analizinin nonparemetrik bir alternatifi olan AID analizinin amacı ise yukarıdaki formülde gösterilen problemi çözmeye yönelik olmayıp, formülün yapısını tanımlamaktır. Regresyon analizinden farklı olarak AID tekniği bileşenlerin nisbi önemi ya da istatistiksel anlamı hakkında güvenilir bir bilgi vermemektedir. AID teniğinin başlıca fonksiyonu, değişkenlerin arasındaki ilişkilerin doğasını, özelliklerini yani katkılı ya da etkileşimli olup olmadıklarını araştırmaktır. Bu özelliğiyle, AID regresyon analizi ya da benzer teknikler için başlangıç niteliğindedir. Temel prensip bağımlı değişken ve açıklayıcı değişkenler arasındaki mümkün tüm ilişkileri araştırmış önceki çalışmalarla bağımlı değişkenin varyansını açıklamaya çalışmaktır. Araştırmanın

[75] Pehlivan, s. 23.
[76] Pehlivan, s. 26.
[77] Jasna Soldic-Aleksic," Combined Approach Of Kohonen Som And Chaid Decision Tree Model To Clustering Problem: A Market Segmentation Example**", Journal of Economics and Engineering** ,Vol.3, No.1 (April 2012), s.21.

sonuçları ikili ağaçta gösterilir ve dallandırmanın yanı sıra her bölümün genişligi ve modelin açıklayıcı gücü ağaçta yer alır. Algoritmayı özetleyecek olursak, anakütle ile başlar önce anakütleyi iki gruba böler. Bu gruplardan her birini toplam kareler kriterine göre hem grup bölünemeyecek kadar küçük hem de grup homojen oluncaya kadar iki gruba daha böler. Algoritmanın adımları şu şekildedir:

1. Tüm değişkenler dikkate alındıkta sonra, *l.* değişken yani en büyük kareler toplamına (Total Sum of Squares) sahip olan seçilir.

$$TSS_l = \sum_{a-1}^{N_l} Y^2 - \left(\frac{\sum_{a=1}^{N_l} Y_a}{N_l} \right)^2 \qquad (1.50)$$

Burada;

TSS_l = *l.* değişken ya da grup için kareler toplamı

Y = Bağımlı değişkene ait gözlemler

Y_a = *l.* değişken ya da grupla birlikteki gözlem çifti için bağımlı değişken değeri

N_l = *l.* değişken ya da gruba ait gözlem sayısıdır.

2. Bir değişken seçildikten sonra, gözlemler aşağıdaki eşitliği maksimum yapacak iki alt gruba bölünür.

$$BSS_{l,1,2} = (n_1 \overline{Y}_1^{\,2} + n_2 \overline{Y}_2^{\,2}) - n_l \overline{Y}_l^{\,2} \qquad (1.51)$$

n_1, n_2 = bölünme sonrası her gruptaki gözlem sayısıdır.

$\overline{Y}_1, \overline{Y}_2$ = her bölünmüş grup için ortalamadır.

n_l = bölünme öncesi *l.* değişken ya da grup için gözlem sayısıdır.

$\overline{Y}_l$ = bölünme öncesi *l.* değişken ya da grup için ortalamadır.

$BSS_{l,1,2}(B)$ = *1* ve *2* gruplarına bölünen *l.* değişken ya da grup için gruplar arası kareler toplamıdır (Between Sum of Squares).

Bu adımlar sonucu oluşan ağaç aşağıdaki bilgileri içermektedir:

i. Her grup için bir bağımlı değişkenin ortalaması,

ii. Her grup içindeki gözlem sayısı,

iii. Her değişken bölünmesine ait birleştirici güce dayalı sıralayıcı dereceler oluşturan bölünmelerin mertebesi,

iv. Herbir bölünmenin değeri.

v. Analizin toplam açıklayıcı gücünü gösteren R^2 değeri (determinasyon katsayısı).

AID ağacının her bir dalı üç durdurma kuralından bir tanesi ile son verilene kadar dallanır. Dallanma aşağıdaki durmlarda son bulur:

1. Bölünemeyecek kadar küçük sınıflar oluştuğunda,

2. Grubun homojen hale gelmesiyle daha fazla bölünme gereksiz olduğunda,

3. Artık gruplar arası kareler toplamını küçültecek hiçbir olası bölünme kalmadığında durur.[78]

AID yöntemi büyük veri setlerine yönelik uygulamalar için tanımlayıcı metodları geliştirmede popülerliğini tasdik etse de uygulamadaki en büyük eksikliği AID'in çıkarımsal yönleri ile ilgili sağlıklı analizler yapılmamasıdır.[79] Bu yüzden de yanlış kullanılması ve yorumlanması söz konusu olmuştur. Bunu önlemek için veri analizinin başında yanlı tahminlerin ne kadar kabul edilebileceğinin belirlenmesi ve

[78] Pehlivan, s.27-28.
[79] G.V. Kass, "Significance Testing in Automatic Interaction Detection (A.I.D.)", **Applied Statistics,** Vol.24, No.2 (1975), s.178.

doğrulama testleri olmadan kullanımının tatmin edici olmayacağının göz önünde bulundurulması gerekir.[80]

1.4.2. CHAID Algoritması

CHAID (Chi-Squared Automatic Interaction Detector: Otomatik Ki-Kare Etkileşim Belirleme Analizi) algoritması bir çeşit karar ağacı olup, düzeltilmiş anlamlılık testine (Bonferroni testine) dayanır. Bu teknik Güney Afrika'da geliştirilmiş olup, 1980 yılında Gordon V. Kass tarafından doktora tezi olarak basılmıştır.[81] CHAID analizi, AID analizinin bir uzantısı olup, kategorik bağımlı değişkenler için tasarlanmıştır ve AID analizinden farklı olarak bağımlı değişkeni çoklu alt gruplara ayırmaktadır. CHAID analizinde amaç bağımlı değişkeni en iyi şekilde açıklayabilmek için veriyi birbirini kapsamayacak şekilde (homojen) alt gruplara bölmektir. Oluşturulan alt gruplar küçük tahmin edici gruplardan oluşmuştur ve bu tahmin ediciler daha sonraki analizlerde bağımlı değişkenin tahmininde kullanılır.[82] CHAID analizi ile bir bağımlı ve birden fazla bağımsız değişken arasındaki ilişkiler araştırılır. CHAID algoritması bağımlı değişkeni açıklayan bağımsız değişkenler arasındaki çok yönlü etkileşimleri vermekle beraber, bu etkileşimleri ağaç diyagramı yoluyla kolaylıkla yorumlama imkanı da sağlamaktadır.[83] CHAID algoritmasında bağımlı değişken sınıflayıcı, sıralayıcı veya sürekli olabilir. Bağımsız değişkenler ise sadece sınıflayıcı ve sıralayıcı kategorik değişkenkenlerden oluşmalıdır. Eğer bağımsız değişken sürekli ise, algoritmaya yerleştirilmeden önce sıralayıcı kategorik değişkene dönüştürülmesi gerekir.[84]

CHAID analizi ikili olmayan yani çoklu alt gruplar oluşturduğu için (yani bir kök ya da düğüme ikiden fazla dal bağlanabildiği için) büyük veri setleri için nispeten daha uygun bir algoritmaya sahiptir. Ayrıca çok sayıda kategoriden oluşan kategorik bir bağımlı değişkeni, çok sayıda sınıfı bulunan kategorik bağımsız değişkenlerle

---

[80] Pehlivan, s.28.
[81] *CHAID*, (t.y.) http://en.wikipedia.org/wiki/CHAID (14 Ocak 2014).
[82] G.V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data", **Applied Statistics,** Vol.29, No.2, (1980), s.119.
[83] Nezih Dağdeviren, ve Diğerleri, "Akademisyenlerde İş Doyumunu Etkileyen Faktörler", **Balkan Med J,** Cilt.28, Sayı.69-74 (2011), s.71.
[84] CHAID and Exhaustive CHAID Algorithms (t.y.), ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf (18 Ocak 2014).

sınıflayabilmek için, çok yönlü frekans tablolarını (multi-way frequency tables) etkin bir biçimde kullanır.[85] Bu yüzden CHAID analizi istatistikte bilinen Ki-kare test yöntemini kullanarak bir karar ağacı oluşturur. Dolayısıyla hem bir karar ağacı algoritması hem de istatistiğe dayalı bir algoritmadır.[86] "Ki-kare" (χ^2) ismini almasının nedeni algoritmasında birçok çapraz tablonun kullanılması ve istatistiksel önem oranları ile çalışmasıdır.[87] Özellikle bağımsız değişkenlerin birbirleriyle olan ilişki ve etkileşimlerini konu edindiği için Ki-kare istatistiğinden faydalanılır. Çünkü Ki-kare istatistiği değişkenler arasındaki bağımlılığı ele alır.[88]

CHAID analizi ile çok kategorili değişkenlerin yer aldığı büyük bir veri kümesi, benzer kategoriler birleştirilerek ve önemli sayılan değişkenlere göre bölünerek, indirgenir. Her bir bağımsız değişken için kategorilerin anlamlı bir şekilde birleştirilmesinden sonra, bağımlı değişkene göre kontenjans tabloları oluşturularak, Bonferronni p değerleri ile χ^2 istatistikleri hesaplanır. Bağımsız değişkenler birbirleri ile karşılaştırılıp en küçük Bonferronni p değerine sahip değişkenin kategorilerine göre, veriler tekrar alt gruba ayrılır.[89] Bonferronni çarpanı çok katlı testlerde p değerini düzeltmek amacıyla kullanılır. Değişkenlerin bölünmeye uygun olup olmadığına Bonferrroni düzeltilmiş p değeri kullanılarak karar verilir.[90]

Burda dikkat edilmesi gereken husus ise kategoriler birleştirilirken kullanılan anlamlılık değeridir (*p value*). Bunun için bağımlı değişkenin türüne bakılır. Eğer bağımlı değişken sürekli ise F testi kullanılır. Eğer bağımlı değişken nominal kategorik bir değer ise pearson ki-kare istatistiği kullanılır. Eğer ordinal kategorik ise en çok olabilirlik testi (likelihood ratio) kullanılır. Her birleştirilmiş kategori için de bulunan p

---

[85] Evgeny Antipov 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, http://www.palgrave-journals.com/jt/journal/v18/n2/pdf/jt20103a.pdf (31 Ekim 2013), s.111.
[86] Silahtaroğlu, s.104.
[87] http://wiredspace.wits.ac.za/bitstream/handle/10539/7268/thesis.pdf?sequence=2 (18 Ocak 2014), s.4.
[88] Murat Kayri ve Murat Boysan, "Araştırmalarda Chaid Analizinin Kullanımı ve Baş Etme Stratejileri İle İlgili Bir Uygulama", **Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi,** Cilt.40, Sayı.2 (2007), s.140.
[89] Pehlivan, s.32.
[90] CHAID and Exhaustive CHAID Algorithms (t.y.), ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf (18 Ocak 2014).

değerinin anlamlı olup olmadığına bakılır. Bu süreç anlamlı olmayan kategori birleşmesi bulununca son bulur.[91]

Bağımsız değişkenin kategori sayısı e ve birleşme neticesinde oluşan kategori sayısı r olmak üzere, Bonferrroni çarpanı *B*, e kategorisinin, r adet kategori sayısına indirgenmesini sağlayan mümkün olasılıkların sayısı olmak üzere; değişken türüne göre hesaplanan Bonferrroni düzeltilmesi $2 \le r <$e için aşağıdaki gibidir:

$$B = \begin{cases} \binom{e}{r} \\ \sum_{\imath=1}^{r-1} (-1)^{\imath} \frac{(r-\imath)^e}{\imath!(r-\imath)!} \\ \binom{e}{r} + r\binom{e}{r} \end{cases} \qquad (1.52)$$

Yukarıdaki formülasyonda birinci sıradaki eşitilik sıralı, ikinci sıradaki eşitlik nominal ve son olarak üçüncü sıradaki eşitlik eksik kategorili ordinal bağımsız değişkenlerdeki hesaplamalar için kullanılır. Burda r=e için *B*=1 bulunduğu formülasyonda görülmektedir.[92]

CHAID algoritmasında bağımlı değişken için sınıf sayısı $J \ge 2$ olmak üzere. analiz edilecek belirli bir bağımsız değişken ise, $e \ge 2$ sayıda kategori düzeyine sahip olsun. Analizdeki amaç, bağımsız değişkendeki uygun kategorileri birleştirerek *e*xj boyutlu çapraz tablonun en anlamlı rxj tablosuna indirgenebilme problemi olsun. Bu amaçla ilk olarak, ı mümkün olan birleştirmelerin sayısını göstermek üzere, $T_r^{(\imath)}$ istatistiği hesaplanır. Bu istatistik cxe tablosu için (r: 2,3,4,…..,e) bilinen χ^2 istatistiğidir. Eğer $T_r^{(*)} = \max \imath T_r^{(\imath)}$ ise en iyi rxe tablosu için χ^2 istatistiği elde edilmiş olur. Bu durumda $T_r^{(\imath)}$ en anlamlı olarak seçilir. Buna göre algoritmanın tamamı şu şekildedir:

[91] Decision Trees (t.y.), http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf (20 Ocak 2013).
[92] CHAID and Exhaustive CHAID Algorithms (t.y.), ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf (18 Ocak 2014).

Adım 1. Her bir bağımsız değişken için, bağımlı değişkenin kategorileri ile bağımsız değişkenin kategorileri arasında çapraz tablo oluşturulur ve sırasıyla ikinci ve üçüncü adım uygulanır.

Adım 2. Sadece bağımsız değişkenin türüne göre belirlenen uygun çiftler arasından 2xj alt tablosunda anlamlılığı düşük olan bağımsız değişken kategori çiftleri bulunur. Eğer anlamlılık derecesi kritik bir değere ulaşmıyorsa, bu iki kategori birleştirilir. Artık bu birleşim tek bir kategori olarak ele alınır ve bu adım tekrarlanır. Bu işlem bağımsız değişkenin kendi içinde birleşmeleri anlamsız oluncaya kadar devam eder.

Adım 3. Bağımsız değişkenin türü tarafından oluşturulan ve orijinal kategorilerin üç veya daha fazlasının birleştirilmesi ile meydana gelen; her bir birleşik kategori için, birleşmenin tekrar ayrılabileceği en önemli iki bölünme bulunur. Eğer önem derecesi kritik değerin üzerindeyse, bölünme gerçekleştirilir ve ikinci adıma dönülür.

Adım 4. Optimum düzeyde birleştirilen bağımsız değişkenlerin her birinin anlamlılığı hesaplanır, en çok anlamlı olan diğerlerinden ayrılır. Eğer bu anlamlılık verilen kritik bir değerden büyükse veri kümesi seçilen bağımsız değişkenin birleştirilen kategorilerine göre alt gruplarına bölünür.

Adım 5. Henüz analiz edilmemiş gruplar için birinci adıma gidilir. Bu adımda en az gözleme sahip olan gruplar göz ardı edilebilir. Burada bahsedilen ilk üç adım birleştirme, dördüncü adım bölme, beşinci adım ise durdurma olarak açıklanabilir.[93] Bu algoritmanın genel kodu EK 2'de gösterilmiştir.

1.4.3. CART Algoritması

CART (Classification and Regression Trees: Sınıflama ve Regresyon Ağaçları) yöntemi 1984 yılında Breiman, Friedman, Olshen ve Stone tarafından geliştirilmiştir.[94] CART hem kategorik hem de sürekli değişkenleri kullanarak sınıflama ve regresyon

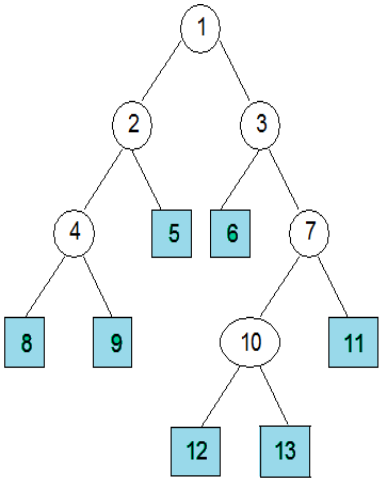
[93] Pehlivan, s.32-33.
[94] Özge Sezgin, "Statistical Method in Credit Banking", (**Yayınlanmamış Yüksek Lisans Tezi,** Orta Doğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, 2006, s.27.

problemlerinin çözümünde karar ağaçlarını kullanan parametrik olmayan istatistiksel bir metottur. Ele alınan bağımlı değişken kategorik ise yöntem sınıflama ağaçları (Classification Trees-CT), sürekli ise regresyon ağaçları (Regression Trees-RT) olarak adlandırılmaktadır.[95]

CART yöntemi, veri setinin çok karmaşık olduğu durumlarda dahi bağımlı değişkeni etkileyen değişkenleri ve bu değişkenlerin modeldeki önemini basit bir ağaç yapısı ile görsel olarak sunabilmektedir.[96] Yöntemin asıl amacı; incelenmekte olan probleme yönelik tahmin yapısını ortaya çıkaran hatasız bir veri seti sınıflayıcısını oluşturmaktır. Sınıflamanın amacı ise, karakterize edilmiş ağaç sayesinde gelecekteki herhangi bir değerin hangi sınıfa düşeceğini belirlemektir. Bu amaçla, hangi değişkenlerin ya da değişkenler arası etkileşimin en iyi sonucu tahmin etmek için gerekli olduğu belirlenir.[97]

CART yöntemi ile ağaç oluşum süreci, üç temel kural içermektedir. Bunlar sırasıyla; "örneklemi bölme kuralı (sample-splitting rule)", "bölme uyum kriteri (goodness of split criteria)" ve "optimum ağacı seçme" kriteridir. CART önceden belirlenmiş bölme kuralına ve düğüm bölme sürecinin her aşamasında uygulanan bölme uyum kriterine göre ağaç oluşturur. Ağaç oluşum sürecinde CART başlangıç veri setini, bağımsız değişkene sorduğu sorulardan alacağı evet/hayır cevaplarına göre ikili alt parçaya (alt örnekleme) böler. Her soru sadece bir bağımsız değişken için sorulur. Alınan cevaba bağlı olarak veri seti sağ ve sol alt parçaya ayrılır.[98] CART algoritmasına örnek olarak aşağıdaki ağaç görülmektedir.

[95] Kıran, s.19.
[96] Gülhan Orekici Temel, Handan Çamdeviren, Zeki Akkuş, "Sınıflama Ağaçları Yardımıyla Legs Syndrome (RLS) Hastalarına Tanı Koyma", **İnönü Üni. Tıp Fak. Dergisi,** Cilt.12, Sayı.2 (2005), s.111.
[97] Yohannes, s. 4.
[98] Yohannes, s. 4-6.



Şekil 1.5: CART Karar Ağacı Modeli Örneği
Kaynak: *CART:Classification and Regression Trees*, http://artax.karlin.mff.cuni.cz/~smetp0am/odkazy/CLASSFINAL.PPT#278,24,Slide 24 (25 Kasım 2012).

Şekilde CART algoritmasına göre oluşturulan bir karar ağacı modeli görülmektedir. Burada, bir numaralı düğüm kök düğümdür. Bölünme sonucu oluşan alt parçalar da birer düğümdür. Yuvarlak olan düğümler (2, 3, 4, 7 ve 10 numaralı düğümler) homojen olmayan düğümler olup, bölünmeye devam ederler. Bunlar daha önce anlatıldığı gibi ara düğüm olarak adlandırılırlar. Kare şeklindeki diğer düğümler ise; (5, 6, 8, 9, 11, 12 ve 13 numaralı düğümler) uç düğüm olarak adlandırılır. Uç düğümler son düğüm olup bölünme yapmazlar. Oysa yuvarlak düğümler, uç düğüm olmadığından bölünmeye devam ederler.[99] Burda her uç düğüm bir sınıfa atanır. Aynı sınıfa atanan iki veya ikiden çok uç düğüm olabilir.[100]

CART algoritması en iyi bölünmeyi bulabilmek için bütün değişkenlerin olası her değerini araştırır. Başlangıç veri kümesine yöneltilen sorulardan alınan cevaplar ile veri kümesi daha küçük parçalara ayırır. Soru sormadaki amaç veriyi maksimum homojenliğe sahip iki parçaya bölebilmektir. Bu süreç bölme sonucu oluşan her bir veri

[99] Roman Timofeew, "Classification and Regression Trees (CART) Theory and Aplications", 2004, Humbolt University, Center of Applied Statistics and Economics, http://tigger.uic.edu/~georgek/HomePage/Nonparametrics/timofeev.pdf (26 Kasım 2012).
[100] Leo Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s.21.

parçası için tekrarlanır ve veri parçası uç düğüme atanınca işlem son bulur.[101] Buna göre veri kümesini en az hata ile sınıflayabilen optimum ağaç CART algoritması ile oluşturulmuş olur.

Sonuç olarak CART yöntemi bütün başlangıç veri setini içinde barındıran kök düğümden başlayarak her düğümü iki çoçuk 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.[102] Burada kategorik bağımlı değişkenler için safsızlığın ölçüsü Gini ve Twoing kriteri iken, sürekli bağımsız değişkenler için ise en küçük kareli sapma yöntemidir. Bu kriterler safsızlık fonksiyonuna bağlı bölünme kriterleri olup, bunlardan daha önce bahsedilmiştir. Buna göre CART ile oluşturululan algoritmanın genel kodu EK 3'te gösterilmiştir. CART algoritmasının formülasyonu ise tezin ikinci bölümünde detaylı olarak anlatılmıştır.

1.4.4. ID3 Algoritması

Karar ağaçları yardımıyla sınıflama işlemlerinin yerine getirilmesi üzerine J. Ross Quinlan (1986) tarafından geliştirilen algoritmalardan biri de ID3 algoritmasıdır. ID3 algoritması entropi tabanlı bir algoritmadır. ID3 algoritmasında hangi değişkene göre dallanmanın yapılacağının belirlenmesinde böl ve yönet (divide-and-conquer) stratejisi ile entropi ölçüsünü kullanır. Böylelikle sınıflamada en ayırıcı özellie sahip değişken belirlenmiş olur.[103,104] ID3 algoritması başlangıçta tavla ve oyunlar için iyi oyun oynama stratejilerini öğrenmek amaçlı kullanılmıştır. Sonrasında akademi ve

[101] Roman Timofeew, "Classification and Regression Trees (CART) Theory and Aplications", 2004, Humbolt University, Center of Applied Statistics and Economics, http://tigger.uic.edu/~georgek/HomePage/Nonparametrics/timofeev.pdf (26 Kasım 2012).
[102] Evgeny Antipov 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, http://www.palgrave-journals.com/jt/journal/v18/n2/pdf/jt20103a.pdf (31 Ekim 2013), s.111.
[103] Özkan, s.54.
[104] Hsinchun Chen, Ganesan Shankaranarayanan ve Linlin She, "A machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms and Simulated Annealing", Journal Of American Society for Information Science, Vol.49, No.8, (Temmuz 1998), s.693.

sanayi gibi çok geniş bir alanda problemlere uygulanmaya başladığından beri oldukça değişmiş ve gelişmiştir.[105]

ID3 algoritması bir veri tabanı bölünmeden önce doğru sınıflandırma yapmak için gelen bilgiyle, veri tabanı bölündükten sonra doğru sınıflandırma için gelen bilgi arasındaki farkı kullanır. Yani verilerin ham (başlangıçtaki) halinin entropisi ile herbir alt bölümün entropilerinin ağırlıklı toplamı arasındaki fark hesaplanır. Bu aradaki fark kazanım olarak adlandırılır. Bu fark hangi alt bölüm için büyükse o alt bölüme doğru dallanma yapılır.[106] ID3 algortitmasında kullanılan dallanma kriteri entropi hesabından faydalanarak hesaplanan *"kazanç ölçütü"* kriteridir. Bu kriterde kazancın maksimum olduğu değişkene göre öncelikli düğüm belirlenir ve dallanmalar o düğümden devam eder.[107] Bununla ilgili bilgiler kısım **1.2.1.1**'de ***kazanç ölçütü*** başlığı altında detaylı olarak anlatılmıştır. ID3 algoritmasının oluşumunda kullanılan genel kod EK 4'te gösterilmiştir.

1.4.5. MARS Algoritması

Çok Değişkenli Uyarlanabilir Regresyon Uzanımları (Multivariate Adaptive Regression Splines: MARS) yöntemi, 1990'ların başında Stanford Üniversitesinden istatistikçi Jerome H. Friedman tarafından geliştirilen nonparametrik ve lineer olmayan bir tekniktir.[108] MARS yöntemi, hem sürekli hem de ikili (binary) bağımlı değişkenler için tasarlanmıştır. Bağımlı değişkeninin sürekli olması durumunda kestirim amaçlı olan bu yöntem, bağımlı değişkeninin kategorik olması durumunda ise, sınıflandırma amacına sahiptir.[109] Bu yöntemin formülasyonu tezin ikinci bölümde detaylı olarak anlatılmıştır.

---

[105] Serhan Koyuncugil, "Veri Madenciliği Yöntemleri ve Uygulamaları", 2010, Başkent Üniversitesi, Ticari Bilimler Fakültesi, http://www.koyuncugil.org/files/ders/bolum9.6.pdf (9 Kasım 2013).
[106] Silahtaroğlu, s.75.
[107] Özkan, s.55
[108] Jerome H. Friedman, "Multivariate Adaptive Regression Splines", **The Annals of Statistics,** Vol.19, No.1 (June 1991), s.1.
[109] Mukhopadhyay A ve Iqbal A, "Prediction of Mechanical Property of Steel Strips Using Multivariate Adaptive Regression Splines" *Journal of Applied Statistics,* 2009, Vol.36, No.1, http://www.tandfonline.com/toc/cjas20/current (23 Mart 2009), s.3.

1.4.6. E-CHAID Algoritması

CHAID algoritmasının kapsamlı incelemesi 1991 yılında Biggs, De ville ve Suen tarafından yapılmıştır.[110] Bu kişiler her düğümdeki bağımsız değişkenlerin kategorilerini (e) birleştirmede daha keşifsel (heuristic) bir yaklaşım kullanarak CAHID algoritmasına önemli bir iyileştirme ile Exhaustive CAHID (Kapsamlı CAHID) algoritmasını geliştirmişlerdir. Bu iyileştirme Bonferroni düzeltme faktöründe nominal değişkenlerin çok sayıda kategoriye ayrılmasını engellemek üzere geliştirilmiş yeni bir Bonferroni çarpanı hesabıdır.[111] Böylelikle CHAID algoritmasından Bonferroni düzeltmesine yönelik getirdiği alternatif bir yaklaşım ile ayrılmaktadır. Buna göre bu algoritma ile en sonda sadece iki kategori kalmak üzere iki kategoriyi ardışık olarak birleştirilir. Böylelikle her iterasyon sonucunda iki kategoriyi birleştirmede mümkün olan Bonferroni çarpanı sayısı değişken türlerine göre aşağıda sıralanmıştır.[112]

$$B = \begin{cases} \dfrac{e(e-1)}{2} \\ \dfrac{e(e^2-1)}{2} \\ \dfrac{e(e-1)}{2} \end{cases} \qquad (1.53)$$

Yukarıdaki formülasyonda birinci sıradaki eşitlik sıralı, ikinci sıradaki eşitlik nominal ve son olarak üçüncü sıradaki eşitlik eksik kategorili sıralı bağımsız değişkenlerdeki mümkün olan birleştirme sayılarını hesaplamada kullanılan formüllerdir. Buna göre E-CHAID algoritmasında CHAID algoritmasında olduğu gibi birleştirme, bölme ve durdurma adımları söz konusudur. CHAID algoritmasından farklı olarak sadece birleştirme adımında, yukarıda verilen birleştirme kombinasyonu açısından farklılık göstermektedir. Bunun dışında E-CHAID algoritmasında CHAID algoritmasında olduğu gibi bağımlı değişken sınıflayıcı, sıralayıcı veya sürekli olabilir.

---

[110] *Answer Tree Algorithm Summary*,(t.y.) http://lyle.smu.edu/~mhd/8331f03/AT.pdf (17 Mart 2014).
[111] Gilbert Ritschard, "CHAID and Earlier Supervised Tree Methods", 2010, University of Geneva, Dept of Econometrics, http://www.unige.ch/ses/metri/cahiers/2010_02.pdf (17 Mart 2014).
[112] CHAID and Exhaustive CHAID Algorithms (t.y.), ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf (18 Ocak 2014).

Bağımsız değişkenler ise sadece sınıflayıcı ve sıralayıcı kategorik değişkenkenlerden oluşmalıdır. Eğer bağımsız değişken sürekli ise, algoritmaya yerleştirilmeden önce sıralayıcı kategorik değişkene dönüştürülmesi gerekir. Yine kategorilerin birleştirilmesi aşamasında kullanılan anlamlılık değeri için bağımlı değişkenin türüne bakılır. Eğer bağımlı değişken sürekli ise F testi kullanılır. Eğer bağımlı değişken nominal kategorik bir değer ise pearson ki-kare istatistiğini, sıralı kategorik ise en çok olabilirlik testini (likelihood ratio) kullanılır.

1.4.7. C4.5 Algoritması

ID3 algortitması yine J. Ross Quinlan (1993) tarafından geliştirilerek C4.5 algoritması elde edilmiştir. Bu algoritma ile sayısal değerlere sahip değişkenler için de karar ağacı oluşturulmasına olanak sağlanmıştır. Sayısal değere sahip bir değişkeni iki aralığa bölmek için en büyük bilgi kazancını sağlayacak biçimde bir "*f" eşik değer* belirlenir. Bunun için herbir değişkene ait ölçüm değerleri $\{x_{11}, x_{21}, ...x_{g1}..., x_{n1}\}$ öncelikle küçükten büyüğe sıralanır. Sıralanmış olan değerlerden $\left[x_{g1}, x_{g1+1}\right]$ aralığının orta noktası alınır ve aşağıda gösterilen eşik değeri belirlenmiş olur.

$$f_i = \frac{x_{g1} + x_{g1+1}}{2} \tag{1.54}$$

Buna göre elde edilen eşik değeri kullanılarak değişken değeri iki parçaya ayrılır. Böylelikle sayısal değere sahip olan bu değişken değerine ikili test uygulanır ve değişkenin aldığı değerin ilgili *f eşik değeri* ile karşılaştırması yapılır. Örneğin bir A değişkeni için $A \leq f$ ve $A > f$ ikili karşılaştırması yapılır.[113] Böylelikle sayısal değere sahip olan değişken, ilgili eşik değerinden küçük eşit veya büyük olmasına göre kategorileştirilir. Bundan sonraki aşama ise, dallanmanın yapılacağı en uygun değişkenlerin belirlenmesidir.

C4.5 algortitmasında kullanılan dallanma kriteri entropi hesabından faydalanarak hesaplanan *"kazanç oranı"* kriteridir. Bu kriterde veri kümesi içinde kazanç oranının maksimum olduğu değişkene göre öncelikli düğüm belirlenir ve

[113] Özkan, s.77.

dallanmalar artık o düğümden devam eder. Bununla ilgili bilgiler kısım **1.2.1.2**'de ***kazanç oranı*** başlığı altında detaylı olarak anlatılmıştır.

C4.5 algoritması ID3 algoritmasına şu konularda üstünlük sağlamaktadır. ID3 algoritmasında karar ağacı oluşturulurken nicel ölçekli değişkenler ve kayıp veriler hesaba katılmazken C4.5 algoritmasında böyle bir kısıtlama yoktur.[114] Böylece daha duyarlı ve daha anlamlı kurallar çıkartan bir ağaç oluşturulabilir. C4.5 algoritmasının oluşumunda kullanılan genel kod EK 5'te gösterilmiştir.

1.4.8. C5.0 Algoritması

C5.0 algoritması, C4.5 algoritması gibi Quinlan tarafından geliştilmiş olup, C4.5 algoritmasının geliştirilmiş halidir ve daha çok ticari amaçlı bir yazılımdır (SEE 5 Yazılımı). C4.5 algoritmasından üstün yanları daha hızlı olması, C4.5'ten daha küçük karar ağaçları oluşturması ve boosting metodunu desteklemesi ile sınıflamanın doğruluğunu arttırır.[115] C5.0 algoritmasında bağımlı değişken kategorik olmak zorundadır. Örneklem maksimum bilgi kazancını verecek sekilde alt örnekleme bölünür. Alt örneklemler ilk bölmenin devamı olarak daha küçük alanlara bölünmeye devam eder. Bu işlem alt örneklemlerin bölünemeyecek duruma gelmelerine kadar devam eder. Son olarak en düşük seviyede olan bölünmeler yeniden incelenir ve modelin değerine önemli bir katkı yapmayanlar budanır.[116,117]

1.4.9. SLIQ Algoritması

SLIQ (Supervised Learning in Quest: Araştırmada Denetimli Öğrenme) algoritması 1996 yılında Uluslararası İş Makinaları Şirketi (International Business Machines Corporation) Almaden Araştırma Merkezinde Mehta, Agarwall ve Rissanen tarafından önerilmiştir.[118] Algoritma hem sayısal hem de kategorik verilerin sınıflandırılmasında kullanılabilmektedir. Sayısal verilerin değerlendirilmesinde

[114] Decision Trees (t.y.), http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf (20 Ocak 2013).
[115] *C4.5 algorithm.* (t.y.) http://en.wikipedia.org/wiki/C4.5_algorithm (30 Ekim 2013).
[116] *C5.0 Node.* (t.y.) http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=%2Fcom.ibm.spss.modeler.help%2Fc50node_general.htm (30 Ekim 2013).
[117] Nilima Patil, Rekha Lathi ve Vidya Chitre, Comparision of C5.0 and CART Classification Algorithms Using Pruning Technique", International Journal of Engineering Research and Technology, Vol.1, No.4, (June 2012), s.2.
[118] Kıran, s.18.

maliyeti azaltmak için ağacın oluşturulması sırasında önceden-sıralama tekniği kullanılır. Sayısal verilerle işlem yapılırken en iyi dallara ayırma kriterini bulmak için verileri sıraya dizme önemli bir faktördür. SLIQ algoritmasında en iyi dallara ayırma kriterinin bulunmasında kullanılan teknik; verilerin sıralama işleminin her düğümde yapılması yerine öğrenme verilerinin sadece bir kere, o da ağacın büyüme aşamasının başlangıcında yapılarak gerçekleştirilmesine dayanır.

ID3 ve C4.5 gibi algoritmalar "önce derinlik" ilkesine (ağacın aşağı doğru büyümesi) göre çalışırken, SLIQ algoritması "önce genişlik" (ağacın sağ ve sola doğru çoklu bölünmeler yapması) düşüncesiyle hareket ederek aynı anda birçok yaprağı oluşturur. Dallara ayrılma işleminde gini kriterini kullanılır. SLIQ, bunun dışında, kategorik verileri alt kümelere ayırmada hızlı bir algoritma kullanır. Ağacın budanması işlemi için ise MDL (Minumum Description Length-En Küçük Uzunluk Tanımlaması) ilkesine dayanan bir strateji izler.

MDL ilkesine göre verileri en iyi temsil edecek model, tanımlanma ve oluşturulma maliyeti en düşük olan modeldir. Bu algoritma hızlı olmasının yanı sıra çok iyi sonuçlar veren karar ağaçları da üretebilmektedir. Ayrıca, SLIQ disk odaklı denilen sadece sabit disk gibi depolama birimlerinde tutulabilen, bellekte tutulması güç olan çok büyük verilere uygulanabilir. Veriler belleğe alınmadan doğrudan bir kerede tek bir ağaç olarak sınıflanırlar.[119]

1.4.10. SPRINT Algoritması

SPRINT (Scalable Parallelizable Induction of Decision Trees: Ölçeklenebilir, Parelellenebilir Tümevarım Karar Ağacı) algoritması 1996 yılında Shafer, Agarwall ve Mehta tarafından önerilmiştir. Daha önce de bahsedildiği üzere ID3, CART ve C4.5 gibi algoritmalar önce derinlik ilkesine göre çalışırlar ve en iyi dallara ayırma kriterine ulaşabilmek için her düğümde verileri sıraya dizerler. SLIQ ise her bir değişken için ayrı bir liste kullanarak bu sıraya dizme işlemini sadece bir kez yapar. SPRINT algoritması bu yönüyle SLIQ algoritmasına benzer ve sözü edilen diğer algoritmalardan ayrılır. Ancak farklı veri yapıları kullanılarak SLIQ algoritmasından ayrılır.

[119] Silahtaroğlu, s.86.

SPRINT ilk olarak her bir değişken için ayrı değişken listesi hazırlar. Her tabloda kullanılacak olan değişken, sınıf ve sıra no bulunur. Böylece veri tabanındaki değişken sayısı kadar tablo oluşur. Sürekli değerleri taşıyan tablolar sürekli değer değişkenine göre sıraya dizilirken, kategorik veri taşıyan diğer tablolar sıra numarasına göre sıralı olarak kalır. Eğitim verilerinden elde edilen ilk listeler sınıflandırma ağacı köküyle ilişkilendirilir. Ağaç büyüyüp düğümler yeni dallara bölündükçe düğümlere ait değişken listeleri de bölünerek yeni dallarla ilişkilendirilir. Liste bölündüğünde içindeki kayıtların sıralaması değiştirilmez, böylece oluşturulan yeni listelerin bir daha kendi içlerinde sıraya dizilmesine gerek kalmaz. Düğümlerden alt dallara ayırma kriteri içinse SLIQ algoritmasında olduğu gibi gini kriterini kullanılır.[120] SPRINT algoritmasının oluşumunda kullanılan genel kod EK 6'da gösterilmiştir.

1.4.11. OUEST Algoritması

OUEST (Quick, Unbiased, Efficient Statistical Tree: Hızlı, Yansız, Etkin İstatistisel Ağaç) algoritması Loh ve Shih tarafından 1997 yılında tanımlanmıştır. OUEST algoritması CART algoritması gibi ikili ağaçlar üretmektedir.[121] Algoritma CHAID ve CART algoritmasından farklı olarak değişken seçimi ve bölünme noktası seçimi işlemlerini ayrı ayrı ele almaktadır. CART algoritmasında ağaç büyüme sürecinde daha fazla bölünmeyi sağlayan kesikli değer ile değişken seçme yoluna gidilmektedir. Bu durum sonuçların genelleştirilmesini azaltarak modelin yanlı olmasına neden olmaktadır. QUEST algoritmasında ise, eğer bağımlı değişken için bağımsız değişkenler aynı bilgi verecekse, bu bağımsız değişkenlerden herhangi biri eşit olasılıkla seçilir. Ayrıca hesaplamalardaki etkinliği açısından da kuvvetli yapıdadır. CART algoritmasının sağladığı bütün avantajlara sahip olmasına rağmen yine de CART gibi hantal bir ağaç yapısına sahiptir. Bağımlı değişken tekdir ve nominaldir. Bir ya da daha fazla bağımsız değişken sınıflayıcı, sıralayıcı ve sürekli türde ölçülmüş olabilir. Kayıp verilerde yerine koyma kuralları uygulanır. Bu durumda QUEST algoritmasına, bağımlı değişkenler kategorik yapıda ise, yansız ağaç önemli ise, büyük ve karmaşık yapıda veri seti varsa ve ağaç tahmininde etkili bir algoritmaya gerek varsa ve ağaç ikili

[120] Silahtaroğlu, s.86.
[121] *QUEST Nod.* (t.y.) http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=%2Fcom.ibm.spss.modeler.help%2Fc50node_general.htm (30 Ekim 2013).

bölünme ile sınırlandırılmışsa başvurulur. CART algoritmasında olduğu gibi ön olasılıklar kullanılmaktadır. QUEST algoritması şu sekildedir; her ayırma için her bağımız değişken ve bağımlı değişken arasındaki ortaklık Anova F testi ya da Levene testi (sıralayıcı ve sürekli bağımsız değişkenler için) ya da Pearson Ki-Kare değeri (sınıflayıcı bağımsız değişkenler için) hesaplanarak bulunur. Değişken seçimi için küçük p degerine sahip olan değişken seçilir. Ayırma işleminde ise eğer bağımlı değişken ikiden fazla kategoriye sahip ise iki süper sınıfı bulabilmek için iki ortalamalı kümeleme algoritması kullanılır. Aksi takdirde bağımsız değişkenin en iyi bölünmesini bulmak için Kuadratik Diskriminant Analizi (QDA) kullanılır. Bu süreç herhangi bir durdurma kuralına rastlayıncaya kadar devam eder. CART algoritmasında olduğu gibi maliyet-karmaşıklık ölçüsü ile optimum ağaç belirlenir.[122]

[122] Pehlivan, s.21.

İKİNCİ BÖLÜM

SINIFLAMA VE TAHMİN MODELLERİ

Birçok disiplin için çok sayıda ve karmaşık bilgi yığını içerisinden açıklayıcı olanları belirleyip en uygun modelin tahmininde bulunmak önemli bir problemdir. Bu problemin çözülebilmesi ihtiyacından yola çıkılarak birçok yöntem geliştirilmiştir. Bu yöntemler ile çok boyutlu verilerin içinde gizlenmiş bilgiler ortaya çıkarılmaktadır. Özellikle günümüzde bilgisayar teknolojisindeki çok hızlı yenilikler ve ilerlemeler bu tür problemlerin çözümüne önemli katkılar sağlamıştır. Bu bölümde bu problemden yola çıkarak geliştirilmiş olan Sınıflama ve Regresyon Ağaçları (CART) ve Çok Değişkenli Uyarlanabilir Regresyon Uzanımları (MARS) yöntemleri üzerinde durulmuştur. Her iki yöntemin ortak özelliği, bağımlı değişkenin türüne göre hem sınıflama hem de tahmin modeli geliştirebilmesidir. Bağımlı değişkenin kategorik olması durumunda sınıflama amacı taşıyan bu yöntemler, bağımlı değişkenin sürekli olması durumunda ise, istatistiksel tahmin amacı taşımaktadır. Bu amaçla sırasıyla bu iki yöntem detaylı olarak açıklanmıştır.

2.1. Sınıflama ve Regresyon Ağaçları'nın Teorik Çerçevesi

Regresyon analizlerinde ağaç kullanım yaklaşımı ilk olarak 1960'lı yıllarda Morgan ve Sonquist tarafından geliştirilen AID yöntemine dayanmaktadır. 1973 yıllarında ise Breiman ve Friedman birbirinden bağımsız olarak sınıflamada ağaç yötemlerinin kullanılmasını yeniden keşfetmek üzere çalışmalar yapmıştır. Daha sonrasında Breiman ve Friedman çalışmalarını biraraya getirmiş ve Stone'un da katılımıyla yöntemin gelişmesine anlamlı katkılar sağlanmıştır. Olshen ise medikal uygulamalarda ağaç yönteminin ilk kullanıcılarından olup, ekibin çalışmalarının teorik gelişmelerine katkılar sağlamıştır. Ağaçlarla ilgili bütün bu çalışmalar Friedman tarafından birleştirilmiş ve 1980'li yıllarda CART (Classification and Regression Trees) adı verilen bu tekniğin doğmasını sağlamıştır.[123]

Bilimsel çalışmalardan elde edilen verilerin analizinde sınıflama (diskriminant, lojistik regresyon, kümeleme analizleri gibi) ve regresyon modelleri sıkça

[123] Leo Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s.ix.

kullanılmaktadır. Ancak bu tür modellerin gerektirdiği varsayımlar pek çok alanda istatistiksel analiz imkânlarını kısıtlamaktadır. İncelenen veri seti üzerinde hiçbir varsayım gerektirmemesi nedeniyle, Sınıflama ve Regresyon Ağaçları (CART) bu tür istatistiksel sınıflama ve regresyon tekniklerine karşı güçlü bir alternatif olarak ortaya çıkmaktadır. Breiman, Friedman, Olshen ve Stone tarafından geliştirilen bu algoritma ile ikili karar ağaçları oluşturulur. Ele alınan bağımlı değişken kategorik yapıda ise, yöntem Sınıflama Ağaçları (Classification Trees: CT) sürekli ise, Regresyon Ağaçları (Regression Trees: RT) olarak adlandırılır.[124]

Buna göre Sınıflama ve Regresyon Ağacı yöntemi bağımlı değişkenin ölçüm türüne göre; Sınıflama Ağaçları ve Regresyon Ağaçları olmak üzere iki başlık altında detaylı bir şekilde anlatılmıştır.

2.1.1. Sınıflama Ağaçları

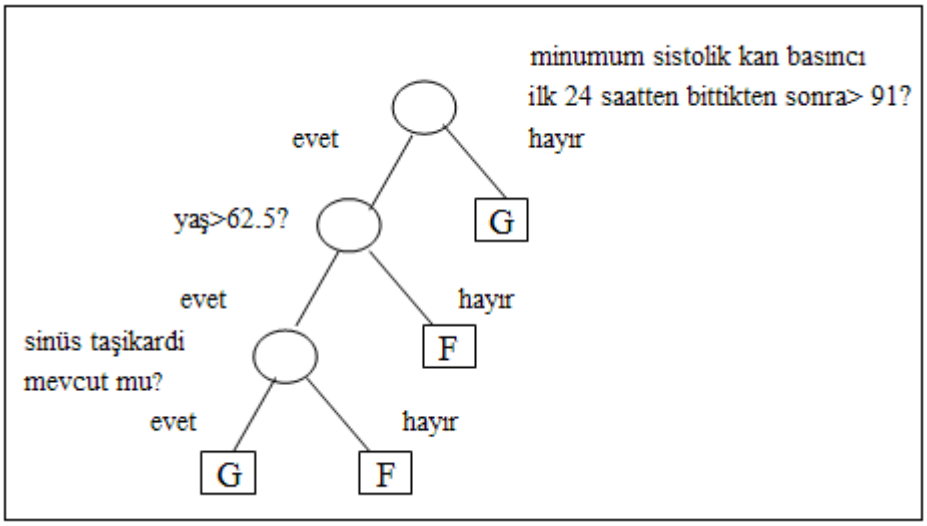
Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi sınıflama olarak adlandırılır. Sınıflama veri madenciliğinin önemli bir konusudur.[125] Sınıflama ağaçları veri madenciliğinin sınıflama ile ilgili konusu arasında yer alır.

Sınıflama çalışmalarının asıl amacı, doğru bir sınıflayıcı üretmek ve problemin altında yatan tahmin edici yapıyı ortaya çıkarmaktır. Bunun için bir gözlemin hangi sınıfta yer alacağının belirlenmesinde; hangi değişkenlerin ya da değişkenler arası etkileşimin olayı çözmede faydalı olacağının anlaşılmaya çalışılması gerekmektedir. Sınıflama sürecindeki en önemli kriter sadece, veri setine yönelik doğru bir sınıflayıcı üretmek değildir. Aynı zamanda tahmin edici yapıda bulunan verilere ışık tutmak ve anlamaktır. Yani ilgili veriye yönelik uzman kişinin geçmiş deneyimi de önemlidir. Bu iki kriter aynı anda önemli olmaktayken, bazen biri diğerinden daha büyük bir etkiye sahiptir.[126]

Buna örnek olarak California Üniversitesi, San Diego Tıp Merkezi'nde yapılan bir sınıflama çalışmasına yer verilebilir. Kalp krizi geçirmiş bir hasta hastaneye kabul edildiğinde İlk 24 saat içinde 19 değişkenin ölçümü yapılır. Kan basıncı, yaş gibi

[124] Temel, "Sınıflama Ağaçları Yardımıyla Legs Syndrome (RLS) Hastalarına Tanı Koyma", s.111.
[125] Özkan, s.51.
[126] Breiman ve diğerleri, s.6.

hastanın durumunu incelemede önemli olduğu düşünülen ikili ve sıralı diğer 17 medikal semptoma bakılır. Burdaki amaç ilk 24 saat içinde elde edilen verilere göre, yüksek risk faktöründeki (en azından 30 günde kurtarılamayan) hastaları belirlemektir. Bu örnekte kalp krizi geçiren yaşlı ve düşük tansiyonlu hastaların yüksek risk grubundaki sınıfa atanması doktorun, geçmiş deneyimine bağlı olarak belirlenmiştir. Aşağıdaki sınıflama ağacı da bu durum için oluşturulmuştur.[127]



Şekil 2.1: San Diego Tıp Merkezi Hastalarının Sınıflama Ağacı
Kaynak: Leo Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s. 2.

Şekilde görüldüğü gibi, hastaların medikal tanısı belirlenmeye çalışılmıştır. G hafi yüksek risk faktöründeki hastaları, F harfi ise düşük risk faktöründeki hastaları nitelemektedir. Burdaki sorulardan alınan evet/hayır yanıtlarına bağlı olarak başlangıç veri seti ikili bölünmeler yapmıştır. Yani hastalar ya G denilen yüksek risk sınıfına ya da F denilen düşük risk sınıfına atanmaya çalışılmıştır. Buna göre, ilk 24 saatin sonunda kan basıncı 91 ve daha küçük olan hastalar yüksek risk sınıfına atanmıştır. Kan basıncı 91'den yüksek olan hastalar için sınıflama işlemi artık yaş değişkenine ve sinüs taşikardinin mevcut olup olmamasına bağlı olarak devam etmiştir. Böylelikle başlangıç veri setinden faydalanılarak sınıflama kuralına bağlı bir ağaç oluşturulmuştur. Oluşan sınıflama kuralına bağlı olarak yeni gelecek hastaların ilk üç soruya vereceği yanıta

[127] Breiman ve diğerleri, s.1.

göre düşük ya da yüksek risk faktöründeki sınıfa atanması sağlanmış olacaktır. Ağaç oluşumundaki bu basitlik, standart istatistiksel sınıflama kurallarının daha doğru sınıflama kuralları vereceğine dair bir şüphe uyandırır. Bunlar denendiğinde ise kuralların, oldukça karmaşık ve daha az doğru sonuçlar ürettiği gözlemlenir.[128]

Yukarıdaki örnekte de görüldüğü gibi sınıflama ağacı oluşum aşamasında, ilgili veri setine yönelik uzman kişinin geçmiş deneyimleri başlangıç veri kümesini oluşturur. Bu veri kümesi N tane gözlemin çeşitli ölçümlerini yani, değişkenlerinin gerçek sınıflamasını içerir.

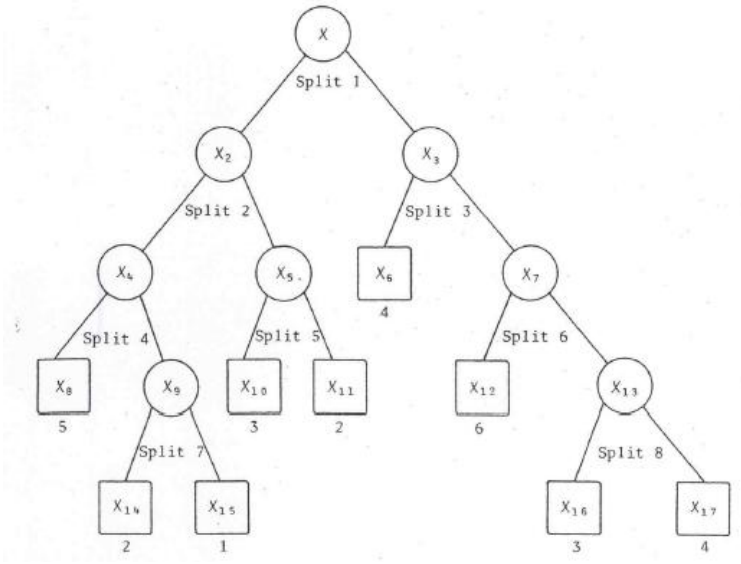
Başlangıç veri kümesi (L) formülasyon olarak şu şekilde açıklanabilir: $\mathbf{x}_n \in \mathbf{X}$ ve $j_n \in \{1,2,\ldots,J\}$, $n = 1,\ldots,N$ olmak üzere $(\mathbf{x}_1, j_1), \ldots, (\mathbf{x}_N, j_N)$ ile gösterilen veri kümesini içerir ve aşağıdaki gibi gösterilir.[129]

$$L = \{ (\mathbf{x}_1, j_1), \ldots,(\mathbf{x}_N, j_N) \} \qquad (2.1)$$

(2.1) nolu formülasyonunda da görüldüğü üzere, başlangıç veri kümesinin içinde **X** ölçüm uzayı (yani bağımsız değişkeler matrisi) ve sınıflar (yani bağımlı değişken vektörünü) bulunur. Sonuç olarak başlangıç veri seti (L) kullanılarak ağaç oluşturabilir.

Buna göre, sınıflama ağaçlarında, ikili ağaç yapısının oluşumu başlangıç veri kümesi olan **X**'in daha homojen iki alt kümeye ayrılması ile başlar. Aşağıdaki örnek şekil ile bu durum gösterilmiştir.

[128] Breiman ve diğerleri, s.1.
[129] Breiman ve diğerleri, s. 5



Şekil 2.2: İkili Sınıflama Ağacı Alt Küme Örneği
Kaynak: Leo Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s. 21.

Şekil 2.1'de görüldüğü üzere, $X = X_2 \cup X_3$ benzer şekilde $X_4 \cup X_5 = X_2$ ve $X_6 \cup X_7 = X_3$ olarak herbir alt küme bir diğey adıyla düğüm, azalan şekilde iki alt kümeye ayrılmaktadır. X_6, X_8, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16} ve X_{17} alt kümelerinin ise bölünme yapmadığı görülmektedir. Bu yüzden bu alt kümeler uç alt kümeler olarak adlandırılır. Uç alt kümeler bitiş noktası olduğu için her uç alt küme, altında numarasının gösterildiği bir sınıfa (C) atanır. Aynı sınıfa atanmış iki veya daha çok alt küme olabilir. Şekilde de görüldüğü gibi X_{10} ve X_{16} aynı sınıf olan 3. sınıfa atanmıştır. Oluşturulan uç alt kümelerden aynı sınıfa denk gelenler aşağıda görüldüğü gibi biraraya getirilir.[130] Bu durum **X** matrisinin bir alt kümesini olarak tanımlanan A_j'lerin nasıl oluştuğunu göstermektedir.

$$A_1 = X_{15} \qquad A_2 = X_{11} \cup X_{14} \qquad A_3 = X_{10} \cup X_{16}$$

$$A_4 = X_6 \cup X_{17} \qquad A_5 = X_8 \qquad A_6 = X_{12}$$

[130] Breiman ve diğerleri, s.21.

Bu bölünmeler $\mathbf{x}=(x_1, x_2, ...)$ ölçüm vektörünün koordinatlarının koşullarına göre oluşturulur. Böylelikle bölünme 1 (Split 1) X'in X_2 ve X_3 alt kümesine ayrılmış hali olduğu için X_2 ve X_3 alt kümlerinin örnek olarak aşağıdaki formda bir yapı ile oluştuğu söylenebilir. Bu durum yukarıda anlatıldığı gibi düğüme uygulanan testin sonucuna göre her alt kümenin nasıl oluştuğunu gösterir.

$$X_2 = \{\mathbf{x};\ x_l \leq c\},\ X_3 = \{\mathbf{x};\ x_l \rangle c\}$$

Aynı durum geri kalan bölünmeler için benzer şekilde gerçekleştirilir. Sonuç itibariyle **x** bir uç kümeye atandığında onun sınıf tahmini, uç kümesinde yazan sınıf etiketi ile gösterilir. Bu noktada ağaç teorisindeki terminoloji yeniden şekillenmektedir ve anlatılmak istenen;

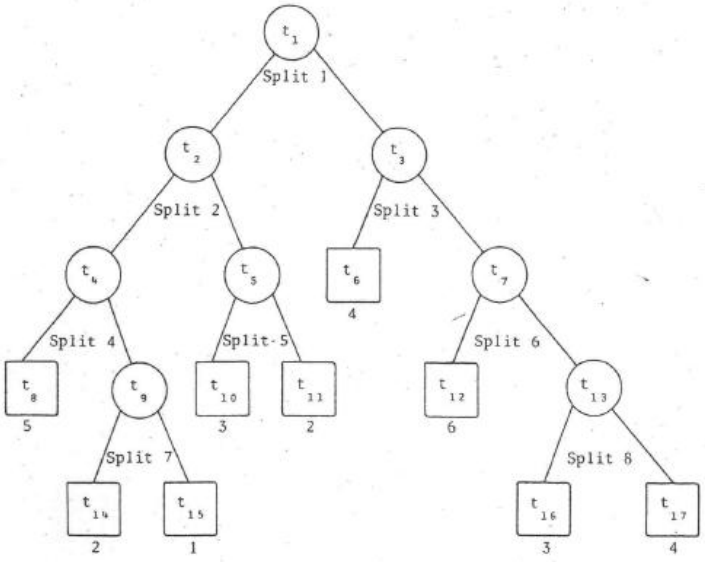
Bir t düğümü = X *matrisinin bir alt kümesi*

ve

Kök düğüm t_1 = X

olur. Bu yüzden uç alt kümeler artık; ***uç düğümler*** olarak adlandırılır. Benzer şekilde uç olmayan alt kümeler; ***uç olmayan düğümler*** olarak adlandırılır.[131] Bu durumu anlatan şekil 2.2 artık şekil 2.3.'te gösterildiği gibi yeniden etiketlendirilir.

[131] Breiman ve diğerleri, s.22.



Şekil 2.3: İkili Sınıflama Ağacı Düğüm Örneği
Kaynak: Leo Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s. 23.

Yukarıda anlatılan örnekten de anlaşılacağı üzere, bir ağacın tamamıyla oluşum süreci üç temel unsur etrafında döner:

1. Bölünmelerin seçiminin yapılması,

2. Bir düğümün uç düğüm olduğuna ya da bölünmeye devam etmesine karar verilmesi,

3. Her bitiş düğümünün bir sınıfa atanmasının yapılmasıdır.

Ağaç oluşumunda çözülmesi gereken problem, bölünmeleri, uç düğümleri ve onların sınıf atamalarını belirlemek için, başlangıç veri kümesinin (L) nasıl kullanılacağıdır. Böylelikle tüm gereken en iyi bölünmeyi bulmak ve bölünmenin ne zaman durudurulacağını bilmektir.[132]

Sınıflama ağacı oluşturmadaki ilk adım maksimum boyuttaki ağacın oluşturulması (T_{max}) ile başlar. Bunun için başlangıç veri kümesi bölünerek son gözleme kadar her bir gözlem bir sınıfa atanır. Maksimum boyuttaki bu ağaç oluşturulduktan

[132] Breiman ve diğerleri, s.22-23.

sonra, ikinci adım olarak ağacın boyutunun optimum olup olmadığına karar verilmesi gerekir. Bu amaçla budama işlemi yapılır. Yüksek boyuttaki ağaç her nekadar düşük bir yanlış sınıflama oranı (misclassification rate) verse de yeni bir veri kümesine uygulandığında yanlış sonuçlar doğurur ve yorumlanması çok karmaşık olur. Ağacın boyutu çok küçük olduğunda ise, başlangıç veri kümesindeki bilgileri içinde barındırmaz ve yüksek bir yanlış sınıflama oranı doğurur. Üçüncü ve son adımda ise optimum boyuttaki ağaç oluşturulduktan sonra, ağaç yeni gözlemleri sınıflayabilmek için kullanılır.[133]

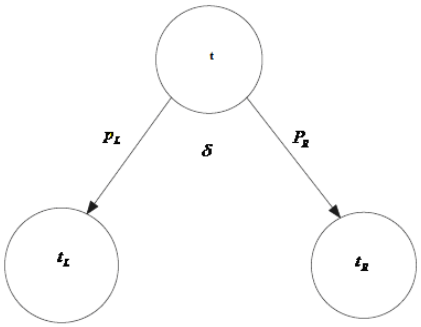
Buna göre, sınıflama ağaçlarının oluşum süreci üç aşamadan oluşmaktadır. Bunlar sırasıyla maksimum boyuttaki ağacın oluştulması, optimum boyuttaki ağacın oluşturulması ve böylelikle optimum boyuttaki ağaç ile yeni bir veri setinin sınıflandırılmasıdır. Maksimum boyuttaki ağaç, ağaç oluşum sürecinde inşa edilen en büyük boyuttaki ilk ağaçtır. Buna göre, bu aşamalar sırasıyla aşağıda anlatılmıştır.

2.1.1.1. Maksimum Boyuttaki Sınıflama Ağacının Oluşturulması

Başlangıç veri kümesinin en son gözleme kadar bölünmesi ile maksimum ağaç oluşturulur. Yani her bir gözlem uç düğümlerde tek bir sınıfta yer alınca maksimum boyuttaki ağaç oluşmuş olur.[134]

Buna göre maksimum ağacın oluşturulması, başlangıç veri kümesinin yani kök düğümün iki alt kümeye bölünmesi ile başlar. İlk olarak bir bağımsız değişkene ait bütün gözlemler kök düğüme yerleştirilir ve bağımsız değişkene ait bütün gözlemler ile başlangıç veri kümesini iki parçaya bölünür. Bu şekilde her bir alt küme bir düğüme atanır. Sonuçta bölünmeler ile başlangıç veri kümesi aşağıda görüldüğü gibi iki düğüme bölünür. Burda t kök düğüm t_L ve t_R ise, sırasıyla sol ve sağ düğümler olup, çocuk düğümü olarak adlandırılabilir.[135]

[133] Sezgin, s.28.
[134] Timofeev, s. 9.
[135] Sezgin, s.28.



Şekil 2.4: CART Bölme Algoritması
Kaynak: Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s.25.

Her bir t düğüm değeri için aday bir bölünme δ vardır. δ, düğümü (t) t_L ve t_R olmak üzere iki parçaya ayıran bir bölünme değeridir. p_L, t düğümündeki gözlemin sol düğüme t_L gitme oranıdır. p_R ise, t düğümündeki gözlemin sağ düğüme t_R gitme oranıdır. S ise, düğüme ait bölünme değerlerini (δ) barındıran bir aday kümedir. Bu aday küme (S) bölünme değerinin belirlenebilmesi için içinde bir dizi soru kümesi (*2*) barındırmaktadır. Soru kümesinde (*2*) bulunan her soru, **x** ϵ A?, $A \subset X$ formuna sahiptir. İligi bölünme δ değerine göre t düğümünde bulunan bütün $\mathbf{x}_n$ değerleri, yanıtlayacağı cevaba göre sol ya da sağ düğüme atanır. Eğer *"evet"* cevabını alırsa sol düğüme t_L , *"hayır"* cebanı alırsa sağ düğüme t_R atanır. [136]

Buna göre, her bölünmede her gözlemin x_i değerinin $x_i \leq \delta$ olup olmadığı araştırılır. Bu eşitsizliğe verilen cevap evet ise gözlem sol düğüme, hayır ise sağ düğüme atanır. Sonra her bölünme için bölme uyum kriteri tahmin edilir ve bu kritere göre en iyi bölünme seçilir. Bu süreç bütün bağımsız değişkenler için ardışık olarak tekrarlanır. Bu tekrarlı bir süreç olduğu için tekrarlı bölünme (recursive partitioning) olarak adlandırılır. Bundan sonra bütün değişkenler safsızlığı azaltacak bir bölme uyum kriterine göre incelenir. Sonlandırma işlemi ise, her düğümün uç düğüm olup olmaması

[136] Breiman ve diğerleri, s.25.

ile neticelenir ve her uç düğüme bir sınıf atanır. Bu süreç, başlangıç veri setindeki her bir gözlem, bir sınıfa atanana kadar devam eder.[137]

Yukarıda bahsedilen sürecin gerçekleşmesi için başlangıç veri kümesinin hangi kritere göre bölünmesi gerektiğinin bilinmesi gerekir. Bu kriterler ise birinci bölümde detaylı olarak anlatılan safsızlık fonksiyonuna dayalı bölünme kriterleridir. Buna göre sınıflama ağaçlarında uygulanan bölme uyum kriterleri olan **gini kriteri** ya da **twoing bölme kriterine** göre ağaçta bölünmeler olur. Bu süreç maksimum boyuttaki ağaç oluşana kadar yani safsızlık fonksiyonundaki değişim maksimum olana kadar devam eder.

Maksimum boyuttaki ağacın oluşum süreci, dört adet temel unsura göre gerçekleşmektedir. Bunlar:[138]

1. Bir soru kümesine (*2*) ait ikili sorular; {Is **x** ϵ A?} , $A \subset \mathbf{X}$ formuna sahip olmalıdır.

2. Her bir t düğüm değeri için aday bir bölünme δ vardır ve en iyi bölme kriteri $\phi(\delta,t)$ her t düğümünün her bölünme δ değeri için değerlendirilir.

3. Bölünmeyi sona erdiren bir durdurma kuralının olması.

4. Her uç düğümü bir sınıfa atayan kuralın belirlenmesi.

Buna göre, birinci temel unsurda belirtildiği üzere, soru kümesinde (*2*) {Is **x** ϵ A?} sorusuna verilen cevap "*evet"* ise, gözlemler sol düğüme t_L atanır ve $t_L = t \cap A$ olur. Verilen cevap *"hayır"* ise gözlemler sağ düğüme t_R atanır ve $t_R = t \cap A^C$ olur. Daha öncede belirtildiği üzere, **X** matrisi bütün gözlemlerin ölçüm sonuçlarını yani ölçüm vektörlerini içinde barındıran bir matristir. $A_1,...,A_j$ 'ler ise, **X** matrisinin alt kümeleridir. Böylelikle sağ düğüme atanan gözlemlerin *A*'nın tümleyeni olduğu görülmektedir.

[137] Sezgin, s. 29.
[138] Breiman ve diğerleri, s.28-29.

Eğer **X** matrisi yani veri seti standart bir yapıya sahipse ve **x** ölçüm vektörü $x = (x_1,...,x_P)$ bütün değişkenler için sabit P boyutunda olup, kategorik ya da ordinal ölçek tipinde ise, soru kümesinde (*2*) aşağıdaki gibi tanımlanır:[139]

1. Her bölünme sadece tek bir değişkenin değerine bağlı olur.

2. Her bir sıralı x_l değişkeni için soru kümesi (*2*), {Is $x_i \leq \delta$?} formundadır ve δ'nin aralığı (-∞, ∞)'dur.

3. Eğer x_l değişkeni $(e_1, e_2,..., e_L)$ gibi kategorik değerler alıyorsa, soru kümesi (*2*), {Is $x_l \in S$? } formundadır ve S $(e_1, e_2,..., e_L)$ değerlerinin alt kümeleridir.

Böylelikle maksimum boyuttaki ağacın oluşumunda değişken türününde dikkate alınarak büyümenin nasıl gerçekleştiği anlatılmıştır. Sınıflama ağaçlarında ağaç oluşum sürecinde bu kez akıllara gelen diğer soru ağacın büyüme sürecinin sonlandırılması ve optimum boyuttaki ağacın seçilmesidir. Bu yüzden öncelikle ağacın büyüme sürecinin nasıl sonlandığının anlatılmasında fayda vardır.

2.1.1.2. Maksimum Boyuttaki Sınıflama Ağacının Büyüme Sürecinin Durdurulması

Başlangıçta da belirtildiği üzere; ağaç oluşturulurken karşılaşılan temel problem başlangıç veri setinin nasıl bölüneceğidir. Bunun için öncelikle veri kümesinin düğümlere nasıl bölündüğü birinci bölümde gini ve twoing bölme kriteri ile anlatılmıştır. Bundan sonraki aşamada ise, bir düğümün nasıl sonlandırılacağı ve bir sınıfın bitiş düğümüne nasıl atanacağı kurallarının belirlenmesidir.

Sonlandırma kuralı basit bir süreci içinde barındırmaktadır. Bir düğümün sonlandırılabilmesi için ağaçtaki bölünmenin bitmiş olması gerekir. Bölme süreci ağacın safsızlığındaki azalış artık imkasızlaşmaya başlayınca son bulmaktadır. Buna göre ağacın safsızlığındaki değişim $\Delta I(T)$ olmak üzere;

$$\Delta I(T) = \Delta i(t)\, p(t) \qquad (2.2)$$

[139] Breiman ve diğerleri, s. 29.

formülasyonu ile belirlenmektedir. Ağacın safsızlığındaki azalış son bulunca ağaç sonlanır ve uç düğümler belirlenmiş olur. Ağaç çiziminde daha önce de belirtildiği gibi uç düğümler dikdörtgen şeklinde, diğer düğümler ise daire şeklinde olur.[140]

Maksimum boyuttaki ağaç oluşumu için gerekli olan son aşama ise, bir sınıfın bitiş düğümüne nasıl atanacağının belirlenmesidir. Bunun için iki farklı yol vardır.

i. Ağacın büyüme sürecini sonlandırmak için ise keşifsel (heuristic) bir kural tasarlanmıştır. Bunlardan ilki çoğunluk (plurality) kuralıdır. Yanlış sınıflama maliyeti aynı olduğunda bu kural uygulanır. Buna göre, bir t düğümünün safsızlığındaki azalış önemsenmeyecek kadar azalınca t düğümünün bölünmesi son bulur ve uç düğüm olur. Yani, gözlemlerin çoğunluğu belirili bir sınıfa ait olduğunda sınıf düğüme atanır. En yüksek olasılığa sahip olan sınıfın (j), düğüme atanması spesifik olarak;

$$p(j_0 / t) = \max_j p(j / t) \qquad (2.3)$$

koşulunu sağlar ve t düğümü artık j_0 uç düğüm sınıfına atanmış olur.[141]

ii. İkinci kurala göre, beklenen yanlış sınıflama maliyetini minimize eden düğüme sınıf ataması yapılır. Eğer sınıfların yanlış sınıflama maliyetleri birbirinden farklı ise çoğunluk kuralı kullanılamaz. Buna göre beklenen yanlış sınıflama maliyeti şu şekilde tanımlanır:

$$r(t) = \sum_{j=1}^{J} c(i / j)\, p(j / t) \qquad (2.4)$$

Burda $c(i / j)$ j sınıfında olduğu bilinen bir gözlemin i sınıfına atanmasının maliyetidir.[142]

Sınıflama ağaçlarında maksimum boyuttaki ağacın oluşum süreci ve bu sürecinin nasıl sonlandırıldığının anlatılmasının ardından optimum boyuttaki ağacın nasıl seçilmesi gerektğine karar verilmesi gerekir. Buna göre optimum boyuttaki ağacın

[140] Sezgin, s. 32.
[141] Breiman ve diğerleri, s. 26.
[142] Sezgin, s. 33.

seçilmesi 2.4. kısım olan model seçim kriterlerinde detaylı olarak anlatılmıştır. Bundan sonraki aşamada ise Sınıflama ve Regresyon Ağaçlarının ikinci başlığı olan regresyon ağaçları anlatılmıştır.

2.1.2. Regresyon Ağaçları

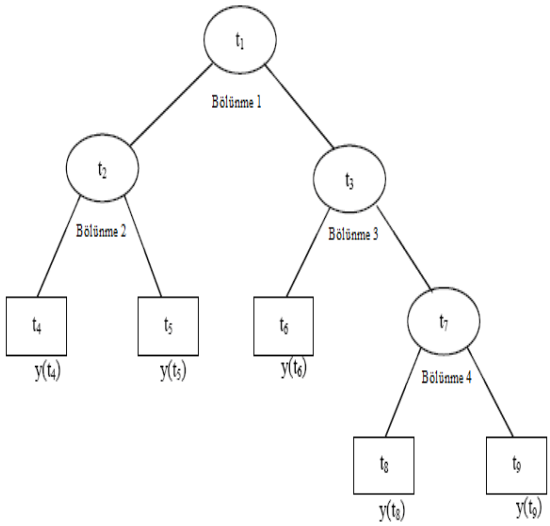
Daha önce bahsedildiği gibi bağımlı değişken sürekli olduğunda karar ağacı regresyon ağacı olarak adlandırılır. Bu süreçteki ağaç oluşum yapısı bazı farklılıklarına rağmen sınıflama ağacı ile benzerlik gösterir. Sınıflama ağaçlarında bağımsız değişkenler, uç düğüm olarak sonlanan bir sınıfa atanmaya çalışır. Oysa regresyon ağaçlarında, bağımlı değişken sürekli olduğu için her uç düğüme atanan değer, bağımlı değişkenin sayısal bir tahminidir. Bu sebepten ötürü regresyon ağaçlarındaki bölme kriteri de farklılık göstermektedir.[143]

Regresyondaki ağaç oluşum süreci, sınıflamadakinden daha basittir ve ağacın büyümesi için kullanılan safsızlık kriteri, aynı zamanda ağacın budanmasında da kullanılmaktadır. Bunun yanısıra regresyon ağacının oluşumu için herhangi bir sınırlama yoktur. Her gözlem eşit ağırlığa sahiptir. Regresyon ağaçlarında kulanılan safsızlık kriteri en küçük kareli sapma hesabına dayanmaktadır. Bu kriter birinci bölümde detaylı olarak anlatılmıştır. En küçük kareler regresyonunda adımsal optimum ağaç yapısının kullanımı AID yöntemine kadar uzanmaktadır. Daha öncede anlatıldığı gibi AID programı 1963 yılında Morgan ve Sonquist tarafından önerilmiştir. AID yönteminin gelişimi 1964 yılında yine Sonquist ve Morgan ile devam edip, bu çalışmalar Sonquist (1970), Sonquist, Baker ve Morgan (1973), Fielding (1977) ve Van Eck (1980) aşamalarını içeren bir yol izlemiştir. AID ve CART arasındaki en temel farklılık, güvenilir bir ağacı oluşturma sürecinde var olan budama ve tahmin sürecine dayanmaktadır. Bir diğer farlılık ise, CART yönteminde bir değişkenin alacağı değer (kategori) sayısı, değişken kombinasyonları, kayıp veriler üzerinde bir sınırlama olmamasıdır.[144]

Regresyon ağaçları ile ağaç yapılı bir tahminin oluşturulması söz konusudur. Ağaç yapılı tahmin, ağaç yapılı sınıflandırma ile benzerdir. Aşağıdaki şekilde de

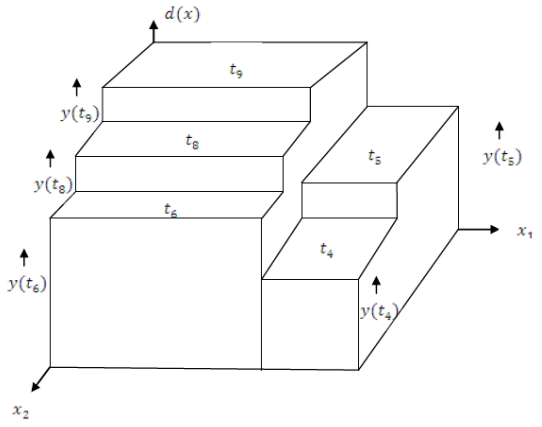
[143] Sezgin, s.36.
[144] Breiman ve diğerleri, s. 217.

görüldüğü gibi X ölçüm uzayı, ardışık ikili bölünmeler ile uç düğümlere ayrılır. Her uç t düğümünde tahmin edilen y(t) yanıt değişkeni sabittir.



Şekil 2.5: Ağaç Yapılı Regresyon
Kaynak: Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s.228.

Regresyon ağaçlarında amaç, başlangıç veri kümesinden L başlayarak d(x) kestirimcisini oluşturmaktır. d(x) kestirimcisi her uç düğümü boyunca sabit olduğundan, ağaç aşağıdaki şekilde de görüldüğü üzere regresyon yüzeyinin bir histogram tahmini olarak düşünülebilir.



Şekil 2.6: Regresyon Yüzeyinin Histogram Tahmini Görünümü
Kaynak: Breiman ve Diğerleri, **Classification and Regression Trees,** New York: Chapman & Hall, 1993, s.228.

Buna göre başlangıç veri kümesinden L başlayarak, bir ağacın tahmin edilmesinde üç unsura gerek duyulmaktadır:

i. Her ara düğümde bir bölünmeyi seçmek için bir yolun belirlenmesi.

ii. Bir düğümün uç düğüm olarak belirlenmesinin kuralını seçilmesi.

iii. Her uç t düğümümüne tahmini bir y(t) değerinin atanması için bir kuralın belirlenmesi.

Regresyon ağaçlarındaki bu süreç sınıflama ağaçlarındaki gibi meydana gelmekle beraber, düğüme atama yapma kuralının çözümü daha rahattır.[145]

Sonuçta, sınıflama ağaçlarında olduğu gibi regresyon ağaçlarında da amaç veri kümesini başlangıç veri kümesinden daha homojen gruplara ayırabilmek için en uygun değişkeni ve o değişkenin bölünme noktasını yani bölünme değerini belirlemektir. Buna göre, üç aşamalı bir süreç söz konusudur. Bunlar sırasıyla maksimum boyuttaki ağacın oluşturulması için gerekli olan temel bölme kriterleri; optimum boyuttaki ağacın oluşturulması için gerekli olan model seçim kriterlerini belirleme ve böylelikle optimum boyuttaki ağaç ile yeni bir veri kümesinin tahminini yapma şeklinde sıralanabilir. Buna göre, bu aşamalar sırasıyla aşağıda anlatılmıştır.

2.1.2.1 Maksimum Boyuttaki Regresyon Ağacının Oluşturulması

Maksimum boyuttaki ağaç, ağaç oluşum sürecinde inşa edilen en büyük boyuttaki ilk ağaçtır. Maksimum boyuttaki regresyon ağacının oluşturulması sınıflama ağacı oluşturulması yaklaşımı ile benzerlik göstermektedir. İlk adımda başlangıç veri kümesini bölmeye yönelik sorular sorulur. En iyi bölünme ise, bölme uyum kriteri ile belirlenmeye çalışılır. Regresyon ağacında sınıflama ağacında olduğu gibi ikili bölünmeler gerçekleştirilir.[146] Regresyon ağaçları gözlem sayısı ve değişken sayısının çok yüksek olduğu verilerin üstesinden etkin ve basit bir şekilde gelebilmektedir.[147]

[145] Breiman ve diğerleri, s. 229.
[146] Sezgin, s.36.
[147] *Tree-based Regression*, (t.y.) http://www.dcc.fc.up.pt/~ltorgo/PhD/th3.pdf (28.02.2014).

Regresyon ağacı oluşum sürecinde bölünmenin yapılacağı değişkenin bölünme değerinin belirlenmesinde en küçük kareli sapma yöntemi kullanılmaktadır. Bu yöntem safsızlık fonsiyonuna dayalı bölme uyum kriteri olup, birinci bölümde detaylı olarak anlatılmıştır. Regresyon ağaçlarında ağaç oluşum sürecinde bu kez akıllara gelen diğer soru ağacın büyüme sürecinin sonlandırılması ve optimum boyuttaki ağacın seçilmesidir. Bu yüzden öncelikle ağacın büyüme sürecinin nasıl sonlandığının anlatılmasında fayda vardır.

2.1.2.2. Maksimum Boyuttaki Regresyon Ağacının Büyüme Sürecinin Durdurulması

Başlangıçta da belirtildiği üzere; ağaç oluşturulurken karşılaşılan temel problem başlangıç veri setinin nasıl bölüneceğidir. Bunun için öncelikle veri setinin düğümlere nasıl bölündüğü en küçük kareleri sapma ve hata tahminleri ile anlatılmıştır. Bundan sonraki aşamada ise, bir düğümün nasıl sonlandırılacağı ve tahmini bir y(t) değerinin, bitiş düğümüne nasıl atanacağı kurallarının belirlenmesidir.

Regresyon ağaçlarında, sınıflama ağaçlarında olduğu gibi bir düğümü uç düğüm kuralına göre sonlandırma gereksinimi vardır. Sınıflama ağaçlarından farklı olarak regresyon ağaçlarında bir düğüm aşağıdaki koşulları sağlayınca sonlandırılır.

$N(t) < N(\min)$

$N(\min)$ genellikle 5 olarak alınır.

Regresyon ağaçlarında düğüme atanmış değer, ilgili düğümdeki yanlış sınıflama kestirimini minimum yapan değerdir. Kareler toplamını minimize eden tahmin değerlerinin kestirimi, düğümdeki bağımlı değişkenin ortalamasıdır.

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{n=1}^{N(t)} y_n \qquad (2.5)$$

Bu yol ile bir düğümdeki gözlemlerin bağımlı değişken değerlerinin ortalaması tahmin değeri olarak atanır.[148]

Regresyon ağaçlarında maksimum boyuttaki ağacın oluşum süreci ve bu sürecinin nasıl sonlandırıldığının anlatılmasının ardından optimum boyuttaki ağacın nasıl seçilmesi gerektiğine karar verilmesi gerekir. Buna göre optimum boyuttaki ağacın seçilmesi için gerekli olan kriterler 2.4. kısım olan model seçim kriterlerinde detaylı olarak anlatılmıştır. Bundan sonraki aşamada ise bir diğer sınıflama ve tahmin modeli olan Çok Değişkenli Uyarlanabilir Regresyon Uzanımları yani MARS yöntemi anlatılmıştır.

2.2. MARS

MARS yöntemi, 1990'ların başında Stanford Üniversitesinden istatistikçi Jerome H. Friedman tarafından geliştirilen nonparametrik ve doğrusal olmayan bir yöntemdir.[149] Bağımlı ve bağımsız değişkenler arasındaki fonksiyonel ilişkiye yönelik bir ön koşul bulundurmayan bu yöntem son zamanlarda insan genetiği, gıda bilimi, hastalık araştırmaları gibi çeşitlli bilim dallarına yönelik yapılan araştırmalar için veri madenciliğinde yaygın olarak kullanılmaya başlanmıştır.[150] Bu yöntem uyarlanabilir (adaptive) bir regresyon yöntemi olup, yüksek boyuttaki problemlere tam anlamıyla uyum sağlamaktadır. MARS yöntemi, regresyon düzenlemelerindeki performansı geliştirmek amacıyla, adımsal regresyonun ve tekrarlamalı ayırma (recursive partitioning) mantığına dayanan CART yönteminin modifikasyonunun geneleştirilmiş bir şekli olarak düşünülebilir.[151] MARS yöntemi, hem sürekli hem de ikili (binary) yanıt değişkenleri için tasarlanmıştır. Yanıt değişkeninin sürekli olması durumunda kestirim

[148] Segin, s.37.
[149] Jerome H. Friedman, "Multivariate Adaptive Regression Splines", **The Annals of Statistics,** Vol.19, No.1 (June 1991), s.1.
[150] Dengju Yao, Jing Yang ve Xiaojuan Zhan, "A Novel Method for Disease Prediction:Hybrid of Random Forest and Multivariate Adaptive Regression Splines**", Journal of Computers**, Vol.8, No. 1, (January 2013), s.1 74.
[151] Travor Hastie, Robert Tibshirani ve Jerome Friedman, **The Elements of Statistical Learning: Date Mining, Inference, and Prediction,** New York: Springer, 2001, s.283.

amaçlı olan bu yöntem, yanıt değişkeninin kategorik olması durumunda ise, sınıflandırma amacına sahiptir.[152]

Nonparametrik ve doğrusal olmayan bu yöntem, doğrusal yöntemlerin aksine değişkenlerin alt kümelerini dikkate almaktadır.[153] Yani tahmin edici (bağımsız) değişkenlerin oluşturmuş olduğu uzay, birbirleriyle örtüşük birçok bölgeye ayrılıp, uzanım fonksiyonlarını (spline functions) oluşturmakta ve bölgesel olan bu regresyon uzanımları da temel fonksiyon (basis function) olarak adlandırılmaktadır.[154] Yani MARS yöntemi ile bağımlı değişken ve bağımsız değişkenler seti arasındaki ilişki belirlenirken düzleştirme uzanımları (smoothing splines) kullanılmaktadır.[155]

Böylelikle, bu yöntem ile her bir değişken bölgelere ayrılır. Bölgeler için gerekli olan dönüşümler belirlenerek, bir dizi temel fonksiyon oluşturulmakta ve her bir temel fonksiyona ait katsayılar hesaplanmaktadır. Oluşan temel fonksiyonlar ise, bağımlı değişken ile parçalı doğrusal (piecewise linear) ilişki içinde bulunmaktadır.[156] Buna göre MARS yöntemi, değişkenlerin en iyi dönüşümlerini ve etkileşimlerini hesaplayarak yüksek boyuttaki verilerin içinde var olan karmaşık ilişkilerin üstesinden gelebilmekte ve yürütmüş olduğu bu süreç ile diğer yöntemler için açıklanması zor ya da imkansız olan veri örüntülerini ve ilişkilerini yorumlayabilmektedir.[157] MARS yöntemi ile temel fonksiyonların oluşumu devam eden bölümde tanımlanmıştır.

2.2.1. MARS ile Temel Fonksiyonların Oluşum Süreci

MARS ile oluşturulan yapısal model, $(x-t)_+$ ve $(t-x)_+$ formunda gösterilen parçalı doğrusal temel fonksiyon (piecewise linear basis functions) açılımını kullanır.

---

[152] Mukhopadhyay A ve Iqbal A, "Prediction of Mechanical Property of Steel Strips Using Multivariate Adaptive Regression Splines" *Journal of Applied Statistics,* 2009, Vol.36, No.1, http://www.tandfonline.com/toc/cjas20/current (23 Mart 2009), s.3.

[153] Q.-S. Xu ve Diğerleri, "Multivariate Adaptive Regression Splines-Studies of HIV Reverse Transcriptase Inhibitors" , **Chemometrics and Intelligent Laboratory Systems,** Vol.72, No.1, (2004), s.28.

[154] R. Put ve Diğerleri, "Multivariate Adaptive Regression Splines (MARS) in Chromatographic Quantitative Structure–Retention Relationship Studies" , **Journal of Chromatography A,** Vol. 1055, No.1-2, (2004), s.12.

[155] K. Batu Tunay, "Türkiye'de Paranın Gelir Dolaşım Hızlarının MARS Yöntemiyle Tahmini", **ODTÜ Gelişme Dergisi,** Vol. 28, No. 3-4, (2001), s.181.

[156] Zhu X ve Davidson I. (Ed.), **Beyond Classification: Challenges of Data Mining for Credit Scoring,** , Hershey PA: Information Science Reference, 2007, s.142.

[157] Shieu-Ming Chou, ve Diğerleri, "Mining the "Breast Cancer Pattern Using Artificial Neural Networks and Multivariate Adaptive Regression Splines", **Expert Systems with Applications,** Vol.27, No.1, (2004), s.136.

"+" alt indisi pozitif kısmı belirtip, istenenen koşulun sağlanamadığı durumda temel fonksiyonunun sıfır sonucunu alacağını belirtir ve aşağıdaki gibi tanımlanır.[158]

$$(x-t)_+ = \begin{cases} x-t, & x > t \\ 0, & diger\ \ durum \end{cases} \tag{2.6}$$

ve

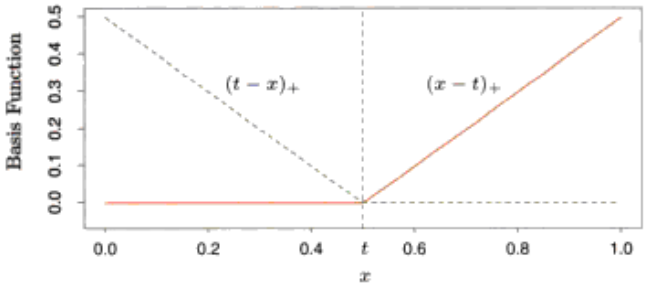
$$(x-t)_- = \begin{cases} t-x, & x < t \\ 0, & diger\ \ durum \end{cases} \tag{2.7}$$

$(x-t)_+$ ve $(x-t)_-$ temel fonksiyonlarının başka bir gösterimi $(x-t)_+ = \max(x-t, 0)$ ve $(x-t)_- = \max(t-x, 0)$ şeklindedir.[159]

$(x-t)_+$ ve $(x-t)_-$ yukarıda belirtildiği gibi parçalı polinom temel fonksiyonların (piecewise polynomial basis functions) doğrusal kombinasyonları olup uzanım fonksiyonları (spline functions) olarak da adlandırılır.[160]

Burada *"**t**"* düğüm değeri olup, her fonksiyon *"**t**"* değerinde parçalı doğrusaldır. $(x-t)_+$ ve $(x-t)_-$ temel fonksiyonlar (lineer uzanımları) aynı zamanda, sırasıyla *"**t**"* düğümünün sağ ve sol bölgelerini ifade edip, yansıtılmış bir çift (a reflected pairs) olarak da adlandırılmaktadır.[161] Şekil 2.7'de $(x-t)_-$ ve $(x-t)_+$ temel fonksiyonları gösterilmektedir.

[158] Hastie, s.283.
[159] Inmaculada Cava Ferreruela, "Explaining Patterns of Broadband Development in OECD Countries", Yogesh K Dwivedi (Ed.), **Handbook of Research on Global Diffusion of Broadband Data Transmission** içinde (756-775), Hershey PA, USA: IGI Global, 2008, s. 760.
[160] Myung Ko, Jan Guynes Clark ve Daijin Ko, "Revisiting The Impact of Information Technology Investments on Productivity: An Empirical Investigation Using Multivariate Adaptive Regression Splines (MARS)", **Information Resources Management Journal,** Vol.21, No.3, (2008), s.20.
[161] Hastie, s.283.



Şekil 2.7: MARS İçin Kullanılan Temel Fonksiyonlar
Kaynak: Travor Hastie, Robert Tibshirani ve Jerome Friedman, **The Elements of Statistical Learning: Date Mining, Inference, and Prediction,** New York: Springer, 2001, s.283.

MARS'ın altında yatan görüş, x_{Kl} düğümlerinin gözlemlendiği her bir X_l girdisi (bağımsız değişken) için yansıtılmış kombinasyonları uygulamaktır. Böylelikle temel fonksiyonların birleşimi aşağıdaki gibi verilir.

$$V = \{(X_l - t)_+, (t - X_l)_+\}, \forall_l = 1,2,..., p \quad t \in \{x_{1l}, x_{2l},..., x_{Kl}\} \tag{2.8}$$

K, her bir X_l değişken uzayını bölen düğüm sayıları olup, p değişken sayısını göstermektedir. Buna göre eğer bütün girdi değerleri belirgin ise, toplamda 2Kp adet temel fonksiyon vardır. Bununla birlikte her temel fonksiyon sadece tek bir X_l değişkenine bağlı olup, $(B(X) = (X_l - t)_+)$ bütün girdi uzayının $(\Re^p)$ bir fonksiyonu olarak ele alınır.[162]

Temel fonksiyonların oluşum süreci EK 7'de gösterilen tekrarlamalı ayırma algoritması ile daha iyi anlaşılabilir. Çünkü bu algoritmanın görevi, sadece verilere en iyi uyum sağlayan katsayı değerlerini belirlemek değil aynı zamanda eldeki verilerden, temel fonksiyonların uygun bir kümesini (alt bölgelerini) türetmektir.[163] Buna göre Ek 7'deki algoritma neticesinde oluşturulan temel fonksiyonlar $[B_m(x)]$ aşağıdaki biçimdedir:

[162] Hastie, s.283.
[163] Friedman, Multivariate Adaptive Regression Splines, s.10-11.

$$B_m(x) = \prod_{k=1}^{K_m} H[s_{km} \cdot (x_{v(k,m)} - t_{km})] \tag{2.9}$$

Burada, $H[\eta]$ aşağıda tanımlandığı üzere pozitif argümanı belirleyen bir adım fonksiyonudur (step function) ve aşağıdaki biçimdedir:

$$H[\eta] = \begin{cases} 1, & \eta > 0 \\ 0, & diger\ \ durum \end{cases} \tag{2.10}$$

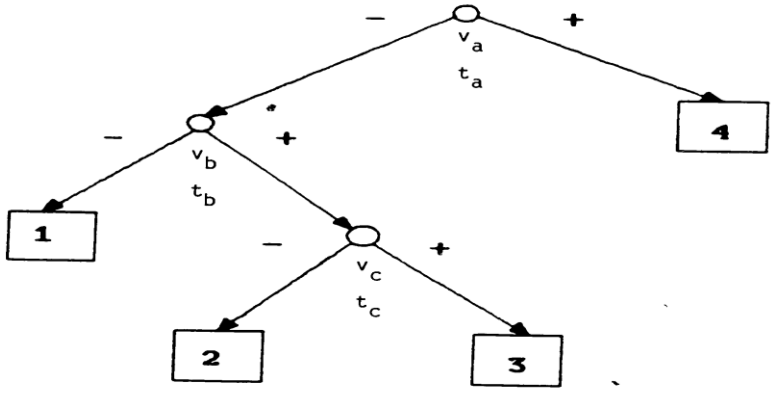
Adım fonksiyonu $H[\eta]$ pozitif argümanı ifade ettiği için aşağıdaki biçimde de gösterilebilir:

$$B_m(x) = \prod_{k=1}^{K_m} [s_{km} \cdot (x_{v(k,m)} - t_{km})]_+ \tag{2.11}$$

Burada her bir temel fonksiyon $[B_m(x)]$ ağacın kökünde başlayan ve karşılık gelen düğümünde sonlanan adım fonksiyonların çarpımıdır. Burada K_m'in adedi bölünmelerin sayısı olup, B_m'in meydana gelmesini sağlar. $s_{km} \in [\pm 1]$ değerlerini alıp, ilgili adım fonksiyonunun sağ/sol kısmını belirler. $v(k,m)$ ise, tahmin edici değişkenleri nitelemekte (etiketlemekte) yani herbir tahmin edici değişkenlerin m. çarpımının k. terimini gösterir. t_{km} ise, ona karşılık gelen değişkenlerdeki değerleri (düğümleri) göstermektedir.[164]

Buna göre Ek 7'de verilen tekrarlamalı ayırma algoritmasının çalıştırılması neticesinde oluşan olası bir, ikili ağaç (binary tree) sonucu ile ona karşılık gelen temel fonksiyonlar aşağıdaki şekilde daha anlaşılır bir biçimde görülmektedir.

[164] Friedman, Multivariate Adaptive Regression Splines, s.11-12.



$$B_1 = H[-(x_{v_a} - t_a)]H[-(x_{v_b} - t_b)]$$
$$B_2 = H[-(x_{v_a} - t_a)]H[+(x_{v_b} - t_b)]H[-(x_{v_c} - t_c)]$$
$$B_3 = H[-(x_{v_a} - t_a)]H[+(x_{v_b} - t_b)]H[+(x_{v_c} - t_c)]$$
$$B_4 = H[+(x_{v_a} - t_a)]$$

Şekil 2.8: Tekrarlamalı Ayırma Regresyonu Modelinin Temel Fonksiyonlarla İlişkisini Gösteren İkili Ağaç
Kaynak: Jerome H. Friedman, "Multivariate Adaptive Regression Splines", **The Annals of Statistics,** Vol.19, No.1 (June 1991), s.12.

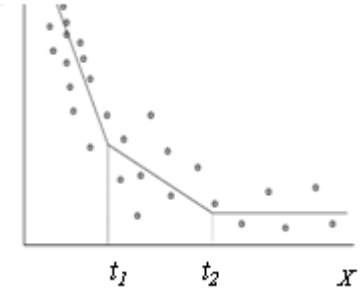
Bu şekilde ikili ağacın iç düğümleri adım fonksiyonlarını, uç düğümleri ise final temel fonksiyonları belirtmektedir. Her bir iç düğümün altında ise, bu düğümle temsil edilen "**v**" değişkeni ve adım fonksiyonu ile ilişkili *"t"* konumu listelenmiştir. Adım fonksiyonundaki **"s"**, düğümün ya sağından ya da solundan olan azalmaların göstergesidir. Şekilde de görüldüğü üzere her bir temel fonksiyon (2.9 nolu formül) ağacın kökünde başlayan ve karşılık gelen düğümünde sonlanan adım fonksiyonların çarpımıdır.[165] MARS yöntemi ile temel fonksiyonların oluşum süreci anlatıldıktan sonra devam eden bölümde düğüm değerlerinin elde edilmesi anlatılmıştır.

2.2.2. MARS ile Düğüm Değerlerinin Elde Edilmesi

MARS yönteminde temel fonksiyonların nasıl oluştuğunun belirlenmesinin ardından, düğüm değerlerinin nasıl elde edileceğinin de bilinmesi gerekmektedir.

[165] Friedman, Multivariate Adaptive Regression Splines, s.12.

MARS yönteminde aynı bağımsız değişken üzerinde ilişkinin şeklinin değiştiği yer düğüm değeri olarak adlandırılır.[166] Şekil 2.9'da görüldüğü üzere t_1 ve t_2 iki düğüm değeridir ve düğüm değerleri arasında farklı doğrusal ilişkilerin tanımlandığı üç aralık bulunmaktadır.

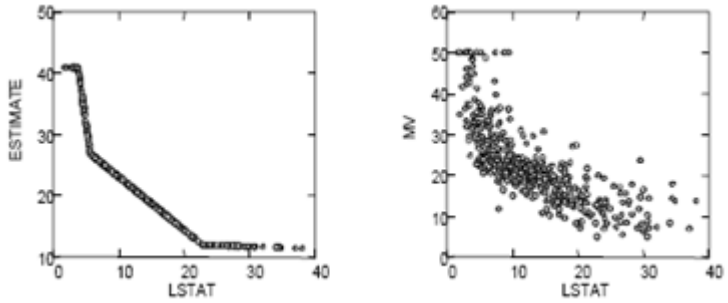


Şekil 2.9: MARS ile Belirlenen Düğüm Değerleri
Kaynak: Lionel C Briand, Bernd Freimut ve Ferdinand Vollei, "Using Multiple Adaptive Regression Splines to Support Decision Making in Code Inspections", **The Journal of Systems and Software**, Vol. 73, No.2, (2004), s. 209.

Buna göre, X değişkenine ait t_1 ve t_2 düğüm değerleri ile bölgeler arasındaki sınırlar yani ilişkinin şeklinin değiştiği yerler belirlenmektedir. Her bir bölge için ise, bağımsız değişkenin bağımlı değişken üzerinde olan ilişkisi farklı olup, eğiminin dolayısıyla ilişki katsayısının bölgeler arasında değiştiği görülmektedir.[167]

Benzer biçimde Şekil 2.10'da gerçek gösterimi sağdaki gibi olan bir veri setinin, üç düğümlü bir MARS uzantısının gösterimi soldaki gibi olmaktadır. Böylelikle veri setinin bölgesel değişimleri düğümler sayesinde belirlenmekte ve düğümler arasında model doğrusal olmaktadır.[168]

[166] Gülhan Orekici Temel, Handan Çamdeviren ve A. Canan Yazıcı, "Regresyon Modellerine Alternatif Bir Yaklaşım: MARS", VIII. Ulusal Biyoistatistik Kongresi, Bursa: Uludağ Üniversitesi, 20-22 Eylül 2005, s.109.
[167] Yeliz Sevimli, "Çok Değişkenli Uyarlanabilir Regresyon Uzanımlarının Bir Split-Mouth Çalışmasında Uygulaması", (**Yayınlanmamış Yüksek Lisans Tezi,** Marmara Üniversitesi, Sağlık Bilimleri Enstitüsü, 2009), s.45.
[168] Salford Systems, **MARS™ User Guide**, San Diego, 2001, s.15.



Şekil 2.10: MARS ile Belirlenen Düğüm Değerleri
Kaynak: Salford Systems, **MARS™ User Guide**, San Diego, 2001, s.15.

MARS yöntemi ile düğüm değerlerinin elde edilmesi anlatıldıktan sonra devam eden bölümde modelin oluşum sürecinde var olan ileriye ve geriye doğru adım algoritmaları anlatılmıştır.

2.2.3 MARS Modeli Oluşum Süreci

Bu süreçte herbir değişken ve değişkenlerin etkileşimi için bir dizi aday temel fonksiyon oluşturulur en uygun temel fonksiyon kümesi seçilir. En uygun temel fonsiyonların seçiminde regresyon modellerindeki ileri ve geriye doğru adım algoritmalarından yararlanılır.[169] Buna göre MARS modelinin oluşumunda iki adımlı bir süreç izlenir. Bunlar, ileriye ve geriye doğru adım algoritmalarıdır.

Bu adımlardan önce modelin karmaşıklığı yani temel fonksiyon sayısı ve "γ" parametresi tanımlanır. Bu parametre, 2.4.kısım olan model seçim kriterinde anlatıldığı üzere, MARS sürecindeki bir düzleştirme parametresidir.

Buna göre Ek 8'de gösterilen İleriye doğru adım algoritması (The forward stepwise algorithm) ile aşağıdaki gibi bir süreç izlenir.

1. En basit bir modelle başlanır. Dolayısıyla, başlangıç olarak modele sabit terim eklenir.

[169] Myung Ko, Jan Guynes Clark ve Daijin Ko, "Revisiting The Impact of Information Technology Investments on Productivity: An Empirical Investigation Using Multivariate Adaptive Regression Splines (MARS)", **Information Resources Management Journal,** Vol.21, No.3, (2008), s.20.

2. Ardından her bir açıklayıcı değişken için, temel fonksiyonların uzayı araştırılır.

3. Temel fonksiyonların uzayı içinde kestirim hatasını minimize edenler yani açıklayıcı değişkenler için artık hata kareler toplamında (sum-of-squares residual error) en çok azalmayı yapan temel fonksiyon çiftleri belirlenir ve modele alınır. (*Kullanılacak temel fonksiyon sayısı kullanıcı tarafından da belirlenebilir.)*

4. Çok büyük bir model bulununcaya kadar (ya da kullanıcının tanımladığı en büyük temel fonksiyon sayısına ulaşılıncaya kadar) yani önceden belirlenmiş maksimum karmaşıklıklı modele ulaşılıncaya kadar 2. adım olan temel fonksiyon araştırma adımına geri dönülür. Böylece, en üst seviyeye ulaşıncaya kadar eklenen temel fonksiyonlarla model büyür.[170]

Yukarıdaki maddelerden de anlaşıldığı üzere, MARS modelinin ileri adımlı sürecinde düğümler otomatik olarak seçilmektedir. Aday düğümler, her bir açıklayıcı değişkenin belirli aralıklarındaki konumlarında, temel fonksiyon çiftlerini tanımlayabilmek için belirlenir. Her adımda model için, artık kareler toplamında (residuals sum of squares) en büyük azalmayı veren düğüm ve ona karşılık gelen temel fonksiyon çiftleri seçilir. Düğüm seçimi maksimum model boyutuna ulaşılıncaya kadar devam etmektedir. Bundan sonraki aşamada ise, geriye doğru adım algoritmasında anlatıldığı gibi modele katkısı en az olan temel fonksiyonların elenmesiyle budama prosedürü uygulanır. Bu sayede kestirim performansı üzerine anlamlı hiçbir katkısı olmayan temel fonksiyonlara ait bağımsız değişkenler tespit edilip, modelden atılabilir.[171] MARS algoritmasında düğüm değerlerinin belirlenmesinde izlenilen yoldan da anlaşılacağı üzere, MARS modeli ile veri kümesindeki her bir açıklayıcı değişken için bir çift uzanım fonksiyonu ve düğüm değeri belirlenip, yanıt değişkeni en iyi şekilde açıklanmaya çalışılmaktır.[172]

---

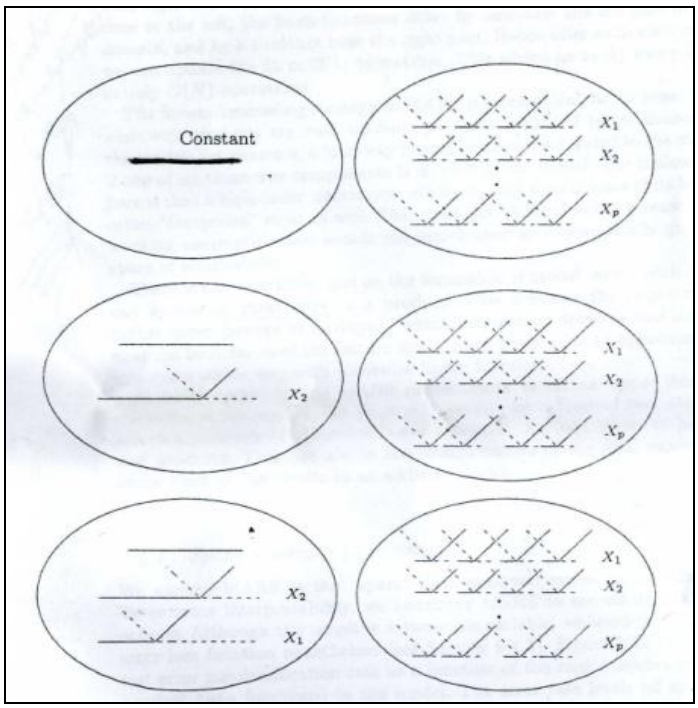
[170] Q.-S. Xu ve Diğerleri, "Multivariate Adaptive Regression Splines-Studies of HIV Reverse Transcriptase Inhibitors" , **Chemometrics and Intelligent Laboratory Systems,** Vol.72, No.1, (2004), s.29.

[171] JR Leathwicka, J Elith ve T Hastie, "Comparative Performance of Generalized Additive Models and Multivariate Adaptive Regression Splines for Statistical Modelling of Species Distributions", **Ecological Modeling,** Vol. 199, No.2, (2006), s.190-191.

[172] E Deconinck ve Diğerleri, "Prediction of Gastro-Intestinal Absorption Using Multivariate Adaptive Regression Splines", **Journal of Pharmaceutical and Biomedical Analysis,** Vol. 39, No.5, (2005), s. 1022.

Sonuç olarak MARS'ın ileri adımlı model oluşturma süreci, maksimum temel fonksiyon sayısına ulaşılıncaya kadar devam eder. Böylelikle oluşan model aşırı uyumlu bir model olur.

Aşağıdaki Şekilde ise MARS'ın ileri adımlı model oluşturma süreci daha anlaşılır bir biçimde görülmektedir.



Şekil 2.11: MARS-İleri Adımlı Model Oluşum Süreci
Kaynak: Travor Hastie, Robert Tibshirani ve Jerome Friedman, **The Elements of Statistical Learning: Date Mining, Inference, and Prediction,** New York: Springer, 2001, s.285.

Bu şekilde solda gösterilen uzanımlar modeldeki temel fonksiyonları ifade edip, sabit fonksiyon olan $h_0(X)=1$ en baştan modele girmiştir. Sağdakiler ise modelin oluşumunda değerlendirilecek olan aday temel fonksiyonları oluşturmaktadır. Görüldüğü üzere bunlar $x_{Kl} \in \{x_{1l}, x_{2l}, \ldots x_{Kl}\}$ düğümlerinin gözlemlendiği her bir X_l bağımsız değişken için parçalı doğrusal temel fonksiyon çiftleridir. Her aşamada

modele aday temel fonksiyon çiftlerinin çarpımları alınır ve artık hatasını en çok azaltan eklenir. Yukarıda ki şekilde bu sürecin ilk üç adımı gösterilmiştir.[173]

Ek 9'da gösterilen geriye doğru adım algoritmasında ise, (The backwise stepwise algorithm) ana amaç ileriye doğru adımda oluşturulmuş olan aşırı uyumlu modeli önlemek için model karmaşıklığını azaltmaya çalışmaktır. Bu amaçla geriye doğru adım algoritmasında, ileriye doğru adım algoritmasının devamı olarak aşağıdaki sıra takip edilir:

5. Modele giren bütün temel fonksiyonlar (sabit temel fonksiyon hariç) kümesi araştırılır. Burda uyum iyiliği kriteri olan ve **2.5.3.** kısımda anlatılan Genelleştirilmiş Çapraz Geçerlilik (Generalized Cross Validation: GCV) kriterine en az katkısı olan temel foksiyonlar elenir.[174] Yani ileriye doğru adım ile oluşturulan maksimum model en uygun varyans-sapma dengesi (optimal balance of bias and variance) bulununcaya kadar budanır.[175] Böylelikle tek tek en az etkili olan yani artıkların hata karelerinde (residual squared error) artışa neden olan temel fonksiyonlar (sabit temel fonksiyon hariç) her adımda elenir.[176]

6. Beşinci adımdaki süreç, en iyi alt model bulununcaya yani genel model için olan GCV kriteri en küçük değerine ulaşıncaya kadar devam eder.[177]

Buna göre MARS modelinin oluşum sürecinde ileriye ve geriye doğru adım algoritmalarının anlatılmasının ardından devam eden bölümde modelin formülasyonu anlatılmıştır.

---

[173] Hastie, s.285.
[174] Q.-S. Xu ve Diğerleri, "Studies of Relationship Between Biological Activities and HIV Reverse Transcriptase Inhibitors by Multivariate Adaptive Regression Splines with Curds and Whey" , Chemometrics and Intelligent Laboratory Systems, Vol.82, No.1-2, (2006), s.25.
[175] Inmaculada Cava Ferreruela, "Explaining Patterns of Broadband Development in OECD Countries", Yogesh K Dwivedi (Ed.), Handbook of Research on Global Diffusion of Broadband Data Transmission içinde (756-775), Hershey PA, USA: IGI Global, 2008, s. 761.
[176] Hastie, s.284.
[177] Q.-S. Xu ve Diğerleri, "Multivariate Adaptive Regression Splines-Studies of HIV Reverse Transcriptase Inhibitors" , **Chemometrics and Intelligent Laboratory Systems,** Vol.72, No.1, (2004), s.29.

2.2.4. MARS Modeli

MARS algoritmasının oluşum aşamasında, en başta anlatıldığı üzere ileri adımlı doğrusal regresyonuna (forward stepwise linear regression) benzer olan bir yöntemden faydalanılır. Fakat burada orjinal girdi değerleri (bağımsız değişkenler) yerine (2.6) ve (2.7)' deki formülasyonda verilen temel fonksiyonların kümesi veya onların çarpımlarından faydalanılır.[178] MARS modeli Ek 8'de gösterilen ileriye doğru adım algoritması ve Ek 9'da gösterilen geriye doğru adım algoritması ile elde edilir. Bu bilgiler neticesinde elde edilen MARS modeli formülasyonu aşağıdaki gibi tanımlanabilir:

$$\hat{y} = \hat{f}_M(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x) \qquad (2.12)$$

$B_m(x)$ (m=1,2,…,M) ise, temel fonksiyon olup (2.9) nolu formülasyonda $B_m(x) = \prod_{k=1}^{K_m} H\left[s_{km} \cdot (x_{v(k,m)} - t_{km})\right]$ ifadesi ile belirtilmişti. Bu ifade (2.12) nolu formülasyonda yerine yazılırsa;

$$\hat{y} = \hat{f}_M(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} H\left[s_{km} \cdot (x_{v(k,m)} - t_{km})\right] \qquad (2.13)$$

Eşitliği elde edilir. Bu eşitlikte β_0, birinci temel fonksiyon (B_0) olan sabit temel fonksiyonunun (constant basis function) katsayısıdır. β_m, ise m. temel fonksiyonun katsayısıdır. Formülasyondaki M temel fonksiyon sayısını, K_m = düğüm sayısını, $s_{km} \in [-1]$ 1 ya da -1 değerlerini alıp adım fonksiyonunun sağ ya da sol taraflarını temsil eder. $v(k,m)$ ise daha önce değinildiği üzere tahmin edici değişkenleri nitelemekte (etiketlemekte) ve t_{km} ona karşılık gelen değişkenlerdeki değerleri (düğümleri) göstermektedir. Buna göre $x_{v(k,m)}$ değeri ise anlaşılacağı üzere bağımsız değişken vektörünü ifade etmektedir.[179]

[178] Hastie, s.283-284.
[179] Friedman, Multivariate Adaptive Regression Splines, s.12.

Sonuç olarak, MARS yöntemiyle, optimal temel fonksiyon sayısı ve düğüm konumları belirlendiğinde seçilen temel fonksiyonlar üzerinden uygun modelin parametre tahminleri, "En Küçük Kareler: EKK" (Least Squares Methods) yöntemi ile olur.[180] Parametre tahminlerine yönelik detaylı bilgiler ise Ek 7 (Tekrarlı ayırma algoritması), Ek 8 (MARS-İleriye doğru adım algoritması), ve Ek 9'da (MARS-Geriye doğru adım algoritması) verilen algoritmalarda görülmektedir.

2.3. CART ile MARS Yönteminin Avantaj ve Dezavantajları

CART ile MARS yöntemleri anlatıldıktan sonra bu iki yöntemin karşılaştırılmasında fayda vardır. Bu iki yöntemin avantaj ve dezavantajları aşağıdaki tabloda açıklanmıştır.

Tablo 2.1. CART ve MARS Yöntemlerinin Karşılaştırılması

MODEL ADI			
CART		**MARS**	
Avantajları	**Dezavantajları**	**Avantajları**	**Dezavantajları**
Bağımlı değişken kategorik (ordinal / nominal) veya sürekli olabilir. Bağımsız değişken çeşidi ayrımı yoktur.	Sınıf sayısı fazla ve başlangıç veri seti sayısı az oldugunda model olusturma çok basarılı değildir.	Bağımlı değişken ikili (binary) veya sürekli olabilir. Bağımsız değişken çeşidi ayrımı yoktur.	Kategorik bağımlı değişkenin kategori sayısı sınırlıdır (sadece ikili).
Hem sınıflama hem de tahmin modeli oluşturabilir.	Uygulanması için genellikle büyük veri setlerine ihtiyaç duyulur.	Hem sınıflama hem de tahmin modeli oluşturabilir.	Uygulanması için genellikle büyük veri setlerine ihtiyaç duyulur.
Bağımlı ve bağımsız değişkenlerin dağılımları ile ilgili herhangi bir varsayım yoktur	En uygun uç düğüm sayısının belirlenmesi kullancının bilgisine bırakılmıştır.	Bağımlı ve bağımsız değişkenlerin dağılımları ile ilgili herhangi bir varsayım yoktur	En uygun temel fonksiyon sayısını belirleme, MARS yaklaşımın eksikliklerinden biridir.
Yüksek boyuttaki veri setindeki karmaşık ilişkilerin üstesiden gelir.	Özellikle eksik verilerde yorumlamalarda dikkat edilmesi grerekir.	Yüksek boyuttaki veri setindeki karmaşık ilişkilerin üstesiden gelebilmektedir.	Özellikle eksik verilerde yorumlamalarda dikkat edilmesi grerekir.

[180] K. Batu Tunay, "Türkiye'de Paranın Gelir Dolaşım Hızlarının MARS Yöntemiyle Tahmini", ODTÜ Gelişme Dergisi, Vol. 28, No. 3-4, (2001), s.182.

MODEL ADI			
CART		**MARS**	
Avantajları	**Dezavantajları**	**Avantajları**	**Dezavantajları**
Modele giren tüm değişkenler arasından en önemliler belirlenerek model oluşturulur. Önemsiz değişkenler modele katılmaz.	-	Modele giren tüm değişkenler arasından en önemliler belirlenerek model oluşturulur. Önemsiz değişkenler modele katılmaz.	Değişkenler arasında ikiden fazla etkileşim uygulamak doğru olmaz.
Ağaç şeklindeki model sonuçları çok fazla istatistik bilgisine gerek duyulmadan kolay bir şekilde yorumlanır.	-	Ağaç görseli yoktur ama modelde bağımsız değişkenin düğüm noktaları grafiklerde görülmektedir.	-
Hem bağımlı hemde bağımsız değişkenler için kayıp ya da eksik değer ile aşırı uç değerlerden model etkilenmez.	-	MARS ile her bir değişken eksik değer için kontrol edilir. Eksik değeri olan her bir değişken için MARS otomatik olarak eksik değer indikatörü oluşturur. Bu indikatör değişken kayıp değer yoksa 0, varsa 1 ile kodlanır. Bu sayede MARS'ın kayıp değerlerden ve aşırı uç değerlerden çok az etkilendiği söylenebilir. Bu yüzden veri seti için hemen hemen hiç ön hazırlık yapmaya gerek yoktur.	-
Diğer modellerin aksine orijinal verilerin alt kümelerini temel aldığından modelin yorumlanması kolaydır	-	Diğer modellerin aksine orijinal verilerin alt kümelerini temel aldığından modelin yorumlanması kolaydır	-

Kaynak: Yazar tarafından çeşitli eserlerden yararlanılarak oluşturulmuştur.

2.4. Model Seçim Kriterleri

Bir modelin oluşturulmasından sonra karşılaşılan en önemli problem modelin doğruluğunun nasıl ölçüleceği ile ilgilidir. Karar ağaçları ile oluşturulmuş olan

modellerde de karşılaşılan temel problem ağacın karmaşıklığının ve doğruluk derecesinin belirlenmesi ile ilgilidir.

Ağacın karmaşıklığı, ağaçtaki uç düğüm sayısı ile belirtilirken, doğruluk derecesi ise, ağacın yanlış sınıflama oranının düşük olması diye ifade edilir. Buna göre ağacın karmaşıklığı ve doğruluğu yani tahmin gücü arasında bir denge kurmak için ağacın budanması yani optimum boyuttaki ağacın oluşturulması gereklidir.[181] Çünkü maksimum boyuttaki ağaç genel olarak aşırı öğrenme (overfiting)* eğilimine sahiptir. Optimum boyuttaki ağacın seçiminin yapılabilmesi için bir takım kriterlerin hesaplanması gerekir. Bu kritelerden elde edilen sonuçlar neticesinde doğru ve karmaşıklığı en az olan optimum ağaç boyutu belirlenmeye çalışılır. Optimum boyuttaki ağacın belirlenmesi için kullanılan model seçim kriterleri aşağıda sırasıyla anlatılmıştır.

2.4.1. En Küçük Maliyet-Karmaşıklık Ölçüsü

En küçük maliyet karmaşıklık ölçüsü (minimal cost complexity measure) optimum boyuttaki ağacı belirlemek için sınıflama ağaçlarında kullanılan etkili bir yöntemdir. Ağacı bir bütün olarak inceleyip, ağacın doğruluğuna önemli derecede katkısı olmayan alt dalların yani zayıf bölünmelerin budanması işleminde kullanılan bir ölçüttür. Hem yanlış sınıflama oranı hem de ağacın karmaşıklığı aynı anda minimize edilmeye çalışılır.[182] Maksimum ağaçtan $(T_{\max})$ bir dizi alt ağaç oluşturulduğu için, her bir alt ağaç $T \le T_{\max}$ koşulunu sağlar. $\tilde{T}$ alt ağacın karmaşıklığı yani alt ağacın toplam düğüm sayısıdır. α her bir ilave uç düğüm için bir ceza karmaşıklık parametresidir. Buna göre her bir alt ağaç için maliyet karmaşıklık ölçüsü $R_\alpha(T)$ aşağıdaki gibi tanımlanır:[183]

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right| \qquad (2.14)$$

[181] Kıran, s.31.
* Ağacın boyutu çok büyük olduğunda eğitim verilerine göre hesaplanan hata oranı düşük çıktığı halde, test verilerine göre hesaplanan hata oranının yüksek olmasına aşırı öğrenme (overfitting) denir.
[182] IBM SPSS Modeler 15 Algorithms Guide, C&RT Algorithms, 2012, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/AlgorithmsGuide.pdf (29 Mart 2014), s.64-65.
[183] Breiman ve diğerleri, s. 66.

$R_\alpha(T)$, ağacın karmaşıklığının ve yanlış sınıflama maliyetinin lineer bir kombinasyonudur. α^* her bir uç düğümün karmaşıklık maliyeti olarak düşünülür ve $\alpha \geq 0$'dır. α'nın her bir değeri için $R_\alpha(T)$ değerini minimize eden bir alt ağaç $T(\alpha) \leq T_{max}$ bulunur.

Eğer oluşturulan maksimum boyuttaki ağaçta T_{max} her uç düğüme bir gözlem düşüyorsa, karmaşıklık maliyeti olmayacağı için $\alpha = 0$ olur. Maksimum ağaç en düşük riske sahiptir. Çünkü her gözlem kesin olarak tahmin ediliyor. α'nın değeri yükseldikçe (α>0) maksimun ağacın alt ağaçları bulunmuş olur. α'nın değeri 0'dan itibaren yükseldikçe sırasıyla her biri daha az sayıda uç düğümden oluşan ve kök düğüme (t) kadar ulaşan alt ağaçlar; T_1, T_2, T_3,...,t oluşur. α'nın dereceli olarak artan her bir değeri içinse $R_\alpha(T)$'yi minimize eden alt ağaç seçilmektedir. Buna göre, alt ağacı $T(\alpha)$ en düşük hesaplayan α aşağıdaki koşulları sağlar:

i. $R_\alpha(T(\alpha)) = \min_{T \leq T_{max}} R_\alpha(T)$

ii. *Eğer(If)* $R_\alpha(T(\alpha)) = R_\alpha(T)$, *sonra(then)* $T(\alpha) \leq T$.

Bu özellikleri sağlayan ağaç en eşsiz ağaç olur.[184] Sonuç olarak bu yöntem ile maksimum ağaçtan (T_{max}) türetilen karmaşıklığı azalmış bir seri daha küçük alt ağaçlar arka arkaya gelen uç dallardan elde edilmekte ve böylece farklı alt ağaçlar en uygun olanla karşılaştırılmaktadır.[185] Bu karşılaştıma ile ağacın karmaşıklığı ve maliyetinin optimum dengesi üzerinde durulmaktadır. Yani optimum yanlış sınıflandırma oranına sahip ve gelecek kestirimler için önemli bir ağaç elde edilmiş olur.

2.4.2. En Küçük Hata-Karmaşıklık Ölçüsü

Regresyon ağaçlarında optimum boyuttaki ağaca ulaşmak için sınıflama ağaçlarında kullanılan maliyet-karmaşıklık ölçüsü yerine, bir hata-karmaşıklık (error-

* α'nın değeri algoritmanın ağaç budama aşamasında hesaplanır.
[184] Breiman ve diğerleri, s. 66-67.
[185] Kıran, s.31.

complexity) ölçüsü kullanılır. Böylelikle regresyon ağacının doğruluğu hata kareler ortalamasının tahmini ile belirlenir. Ölçüm aşağıdaki gibi tanımlanır:

$$R_{\alpha}(T) = R(T) + \alpha\tilde{T} \tag{2.15}$$

ve,

$$R(T) = \sum_{t=1}^{T} R(t) \tag{2.16}$$

Burda sınıflama ağaçlarında olduğu gibi her α için alt diziler, $T_{\max}, T_1, ..., t_0$ tanımlanır ve her alt dizi için hata-karmaşıklık ölçüsü tahmin edilir. Böylelikle bu ölçüyü minimize eden ağaç optimum boyuttaki ağaç olarak kullanılır.[186]

2.4.3. Test Örneği Tekniği

Optimum boyuttaki ağacı oluşturmak için kullanılan model seçim kriterlerinden biride test örneği tekniğidir (test sample technique). Test örneği tekniği ile başlangıç veri kümesi L_1 ve L_2 olmak üzere iki alt örnekleme ayrılmaktadır. L_2 örneklemi genelikle başlangıç veri setindeki gözlemlerin üçte birini içinde barındırmaktadır. Bu teknik hem sınıflama ağaçlarında hem de regresyon ağaçlarında kullanılmaktadır. Her iki yöntemdeki kullanımı aşağıda sırasıyla anlatılmıştır.

2.4.3.1. Sınıflama Ağaçlarında Test Örneği Tekniği

Sınıflama ağaçlarında test örneği tekniği ile L_1 ve L_2 olmak üzere iki alt kümeye ayrılan başlangıç veri kümesinde ilk adım olarak L_1 kümesi maksimum ağacı $(T_{\max})$ ve onun alt ağaçlarını oluşturmak için kullanılır. Her bir alt ağaç için test örneği (test verisi) olan L_2 ise, bağımlı değişkenin değerlerini tahmin etmek için kullanılır. Bağımlı değişkenin gerçek değerleri önceden bilindiği için, her düğüm ve doğal olarak her ağaç için yanlış yapılan sınıflamalar sayılabilir. Aynı zamanda gerçekte j sınıfına ait bir gözlemin her ağaçta i sınıfına ait olarak tahmin edilmesi olasılığı aşağıdaki gibidir:

[186] Segin, s.38.

$$p(i/j,T)=\frac{N_{(2)ij}}{N_{(2)j}} \tag{2.17}$$

Burda;

$N_{(2)ij}$: L_2 'deki i. sınıfa atanan j. sınıf gözlemlerinin sayısıdır.

$N_{(2)j}$: L_2 'deki j. sınıfa düşen gözlemlerin sayısıdır.

Buna göre ağacın yanlış sınıflama maliyeti aşağıdaki gibi tahmin edilir:

$$R^{ts}(T)=\frac{1}{N_{(2)}}\sum_{j=1}^{J}c(i/j)N_{(2)ij} \tag{2.18}$$

Formülasyonda test örneğine ***(ts: test sample)*** ait olarak hesaplanan yanlış sınıflama maliyeti $\{R^{ts}(T)\}$, ağacın doğruluğunun ölçüsünü göstermektedir. Yanlış sınıflama maliyeti minumum olan ağaç optimum boyuttaki ağaç olarak kullanılabilir. Aynı zamanda test verisinin yanlış sınıflama maliyetinin tahmini, minumum maliyet-karmaşıklık ölçüsünde kullanılabilir.[187] Kısacası bu ifade ile, T alt ağacına düşen her L_2 gözlemlerinin yanlış sınıflama maliyeti hesaplanıp, ortalaması alınır şeklinde basit bir yoruma varılabilir.[188]

2.4.3.2. Regresyon Ağaçlarında Test Örneği Tekniği

Regresyon ağaçlarında optimum ağacı oluşturmak için kulllanılan test örneği tekniği, sınıflama ağaçlarında olduğu gibi bir yol izler. Regresyon ağaçlarında da başlangıç veri kümesi L_1 ve L_2 olmak üzere iki alt örnekleme bölünür. L_1 alt diziyi şekillendirmek, L_2 ise doğruluğu tahmin etmek için kullanılır.[189] $T_1 > T_2 > \ldots$,ağaç dizilimi arasından optimum boyutlu ağacın seçilebilmesi için $R(T_k)$'nın doğru bir tahminine gerek duyulur. İlk adımda L_1 artık başlangıç veri kümesi olarak düşünülür ve maksimum ağaç olan (T_k)ile onun alt ağaçlarını oluşturmak için kullanılır. Her bir

[187] Sezgin, s.34-35.
[188] Breiman ve diğerleri, s. 74.
[189] Segin, s.38.

alt ağaç için test örneklemi olan L_2 ise, bağımlı değişkenin değerlerini tahmin etmek için kullanılır. Buna göre $d_k(x)$, T_k ağacına denk gelen kestirimci olup, L_2 örneklemi de N_2 sayıda gözlemden oluşuyorsa, aşağıdaki formülasyon ile ağacın yanlış sınıflama maliyeti tahmin edilir.[190]

$$R^{ts}(T_k) = \frac{1}{N_2} \sum_{(x_n, y_n) \in L_2} (y_n - d_k(x_n))^2 \tag{2.19}$$

formülasyon aşağıdaki gibi düzenlenebilir;

$$R^{ts}(T_k) = \frac{1}{N_2} \sum_{n=1}^{N_2} (y_n - \bar{y})^2 \tag{2.20}$$

Formülasyonda test örneğine ***(test sample:ts)*** ait olarak hesaplanan yanlış sınıflama maliyeti $\{R^{ts}(T_k)\}$, ağacın doğruluğunun ölçüsünü göstermektedir. Yanlış sınıflama maliyeti minumum olan ağaç optimum boyuttaki ağaç olarak kullanılabilir.

2.4.4. Çapraz Geçerlilik Tekniği

Optimum ağaç boyutunu bulmak için veri setinin bir kısmının eğitim diğer kısmının ise test için ayrılığı daha önce anlatılmıştı. Eğitimle elde edilen karar ağacı aracılığıyla, test setinde hem bütün ağaç hem de alt ağaçlar için hata değerleri hesaplanır ve en küçük hata değerine sahip alt ağaç optimal ağaç olarak belirlenir. Ancak bu yolla genellikle ideal ağacın elde edilmesi pek mümkün olmamaktadır. Dahası bu yaklaşımın kullanılması için geniş bir veri setine ihtiyaç vardır. Bunun yerine araştırmacıların daha fazla tercih ettikleri çapraz geçerlilik tekniği (cross-validation technique) uygulanabilir. Bu teknik, veri kümesinin büyük olmadığı durumlarda kullanılır. Buna göre optimum boyuttaki ağacı oluşturmak için kullanılan model seçim kriterlerinden biride çapraz geçerlilik tekniğidir. Bu teknik hem sınıflama ağaçlarında hem regresyon ağaçlarında hem de MARS modeli seçim kriterinde kullanılmaktadır. Her üç yöntemdeki kullanımı aşağıda sırasıyla anlatılmıştır.

[190] Breiman ve diğerleri, s. 234

2.4.4.1. Sınıflama Ağaçlarında Çapraz Geçerlilik Tekniği

V katlı çapraz geçerlilik tekniğinde (V-fold cross validation technique) başlangıç veri seti L raslantısal seçimle boyutları hemen hemen aynı olan *v* adet alt örnekleme $(L_1,...,L_v;\ v=1,...,V)$ bölünür. Her alt örneklem test örneklemi olarak kullanılır. *v.* örneklemi kapsamayan örneklem, başlangıç veri seti olarak kabul edilir. Başlangıç veri seti, $L^{(V)} = L - L_v$ maksimum ağacı (T_{max}) ve onun alt ağaçlarını oluşturmak için kullanılır. Yani *V* katlı çapraz geçerlilik tekniğinde amaç *v* adet yardımcı ağaç oluşturmak ve *v.* yardımcı ağaç ile maksimum ağacı oluşturmaktır. Böylelikle $L^{(v)}$ tüm gözlemlerin *(V-1)/V* kadarını içerir. Genellikle *V,* 10 olarak alınır. Bu durumda başlangıç veri seti $L^{(v)}$ gözlemlerin 9/10'unu içerir. L_v ise, bağımlı değişkeni tahmin etmek için kullanılır.[191] *V,* 10 olarak alındığında ver seti her biri bağımlı değişkenin benzer dağılımını içeren 10 alt örnekleme ayrılır. Bu alt örneklemlerden her biri, diğer 9 alt örneklem ile uyumlu ağacın tahmin hatasını değerlendirmede kullanılır. Sıradaki diğer alt örneklem test örneklemi olarak alınır ve bu işlem doğal olarak 10 defa tekrarlanmış olur.[192] Buna göre her bir ağaç için, i sınıfına yanlış sınıflanan gözlemlerin tamamının sayısı aşağıdaki gibidir:

$$p(i/j,T) = \frac{N_{ij}}{N_j} \tag{2.21}$$

$$N_{ij} = \sum_{v=1}^{V} N_{ij}^{v} \tag{2.22}$$

Burda;

N_{ij}^{v}: L_v'deki i sınıfına atanan j sınıfı gözlemlerinin sayısı

N_{ij}: i sınıfına atanan j sınıfı gözlemlerinin toplam sayısı

N_j: j sınıfı gözlemlerinin sayısı

[191] Breiman ve diğerleri, s. 74.
[192] Kıran, s.32.

Buna göre çapraz geçerliliğe **(cross validation:cv)** göre hesaplanan herhangi bir ağacın yanlış sınıflama maliyeti $\{R^{cv}(T)\}$, aşağıdaki gibidir:

$$R^{CV}(T) = \frac{1}{N}\sum_{j=1}^{J} c(i/J)N_{ij} \qquad (2.23)$$

Bu durumda yanlış sınıflama hatası en küçük olan ağaç seçilebilir. Aynı zamanda test örneği yaklaşımına ait yanlış sınıflama maliyetinin, çapraz geçerlilik tahmini; maliyet-karmaşıklık ölçüsü olarak da kullanılabilir.[193]

Breiman ve diğerlerinin burada üzerinde durdukları konu ise *V'*nin kaç olarak alınacağıdır. Daha önce de belirtildiği gibi ağaç oluşturmada hesaba dayalı temel sıkıntı ilk büyük boyuttaki ağacı oluşturmaktır. *V*'nin alacağı değerle doğrusal olarak çapraz geçerliliğin hesap süresi uzar. Fakat daha büyük *V* değeri daha doğru sonuçlu R^{cv} tahminleri oluşturur. Breiman ve diğerleri yapmış oldukları simülasyon çalışmalarında *V*=10 olarak kullandıklarında yeterli doğruluğa ulaşmışlardır. Başka örneklerde ise daha küçük *V* değeri ile de yeterli doğruluğa ulaşmışlardır. Fakat *V* değerini 10'dan daha büyük alarak da seçilen ağacın doğruluğunda anlamlı bir gelişme ile de karşılaşmamışlardır.[194]

2.4.4.2. Regresyon Ağaçlarında Çapraz Geçerlilik Tekniği

Sınıflama ağaçlarında olduğu gibi regresyon ağaçlarında da optimum ağaç boyutunu bulmak için kullanılan yaklaşımlardan biri çapraz geçerlilik testidir. Çapraz geçerlilik testinde;

i. Veri eşit orana (genellikle on eşit parçaya) ayrılır ve her defasında bir alt grup (verinin %10'u) test için kullanılmak üzere veriden çıkartılır ve diğer kalan veriler ile model inşa edilir.

ii. Bu işlem verinin ayrıldığı parça sayısı kadar (genelde 10 defa) gerçekleştirilir ve böylece bütün veri kullanılmış olur.

[193] Sezgin, s.35-36.
[194] Breiman ve diğerleri, s. 85.

iii. Her defasında inşa edilen ağaçlar ilgili test grupları ile kontrol edilir. Daha sonra bütün alt gruplar birleştirilir, 2. ve 3. adımlar ağacın her boyutu için gerçekleştirilir.

Modellerin değerlendirmesi ile en düşük hata değerine sahip ağaç optimal ağaç olarak kabul edilir.[195]

Sınıflama ağaçlarında daha önce anlatıldığı için burda formülasyonu kısaca anlatılacak olursa başlangıç veri seti v adet alt örnekleme ayrılır. L_v her defasında ölçümün performansını tahmin etmede $L - L_v$ ise ağacı budamada kullanılır. Buna göre çapraz geçerliliğe **(**cross validation: cv**)** göre hesaplanan herhangi bir ağacın yanlış sınıflama maliyeti regresyon ağaçlarında{ $R^{cv}(T_k)$ }, aşağıdaki gibi hesaplanır:

$$R^{cv}(T_k) = \frac{1}{N}\sum_{v=1}^{V}\sum_{(x_n,y_n)\in L_v}(y_n - d_k^{(v)}(x_n))^2 \qquad (2.24)$$

formülasyon aşağıdaki gibi düzenlenebilir;

$$R^{cv}(T_k) = \frac{1}{N}\sum_{v=1}^{V}\sum_{n=1}^{N_v}(y_n - \bar{y})^2 \qquad (2.25)$$

Bu durumda yanlış sınıflama hatası en küçük olan ağaç seçilebilir.

2.4.4.3. MARS Modelinde Çapraz Geçerlilik Tekniği

MARS modeli oluşum sürecinden sonra optimum MARS modelinin belirlenmesi gerekir. Bu amaçla en uygun MARS modelinin seçim kriterinde Craven ve Wahba (1979) tarafında geliştirilen Genelleştirilmiş Çapraz Geçerlilik (Generalized Cross Validation: GCV) ölçümü'nden faydalanılır. Bu ölçüt hata kareler ortalamasının

[195] Özkan, "Sınıflandırma ve Regresyon Ağacı Tekniği (SRAT) ile Ekolojik Verinin Modellenmesi", s.2.

(mean squares errors) doğruluğunu tartan ve modelin karmaşıklığına bağlı bir uyum iyiliği ölçütüdür.[196]

MARS modelinin oluşum sürecinde ikinci adım olan geriye doğru adım algoritması ile ilk adımda oluşturulan maksimum model budanıp, tek tek en az etkili olan değişkenler her adımda elenerek en iyi alt model bulunmaya çalışılır. Bu sürecin işlemesiyle oluşan alt modeller GCV kriteri ile karşılaştırılıp, en iyi kestirime sahip olan alt modelin seçimine bu kriter ile karar verirlir.[197]

Sonuç itibariyle GCV, model karmaşıklığını uyum iyiliği haline dönüştüren bir düzenleme biçimidir. GCV yaklaşımı ile modele, en az katkısı olan temel fonksiyonlar modelden atılır. Böylelikle final modele aşırı sayıda uzanım fonksiyonunun eklenmesi engellenir.[198] Uyum iyiliği kriteri olan GCV kriteri aşağıdaki, gibi tanımlanır:

$$GCV(M) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}_M(x_i) \right]^2 / \left[1 - \frac{\theta(M)}{N} \right]^2 \qquad (2.26)$$

Burada; N gözlem sayısı, $\theta(M)$ sabit temel fonksiyon (2.13'te verilen MARS modelindeki β_0 katsayısı) hariç M tane temel fonksiyon içeren bir modelin maliyet-karmaşıklık ölçütü olup şu şekilde tanımlanır:

$$\theta(M) = M + \gamma M \qquad (2.27)$$

γ ise, her temel fonksiyon optimizasyonuna ait maliyet olup, aynı zamanda MARS sürecindeki düzleştirme (smoothing) parametresidir.[199] Yani bir değişkenin kaç fonksiyondan oluşacağını belirtir.

---

[196] Inmaculada Cava Ferreruela, "Explaining Patterns of Broadband Development in OECD Countries", Yogesh K Dwivedi (Ed.), Handbook of Research on Global Diffusion of Broadband Data Transmission içinde (756-775), Hershey PA, USA: IGI Global, 2008, s. 762.

[197] JR Leathwicka, J Elith ve T Hastie, "Comparative Performance of Generalized Additive Models and Multivariate Adaptive Regression Splines for Statistical Modelling of Species Distributions", Ecological Modeling, Vol. 199, No.2, (2006), s.191.

[198] Q.-S. Xu ve Diğerleri, "Studies of Relationship Between Biological Activities and HIV Reverse Transcriptase Inhibitors by Multivariate Adaptive Regression Splines with Curds and Whey" , Chemometrics and Intelligent Laboratory Systems, Vol.82, No.1-2, (2006), s.25.

[199] Friedman, Multivariate Adaptive Regression Splines, s.20.

Örneğin **"γ"** 2 olarak alındığında ilgili değişken için bir düğüm değeri (kırılma noktası) ve dolayısıyla 2 fonksiyon oluşacak demektir. Yapılan çalışmalar γ için en iyi değerin $2 \le \gamma \le 4$ aralığında olduğunu göstermiştir.[200]

Böylelikle 2.26'daki modelin, pay kısmı M tane temel fonksiyon modelinin $f_M(x_i)$, uyumsuzluğunu (lack of fit) ölçer. Payda kısmı ise, model karmaşıklığını belirtir.[201] En uygun MARS modeli ise en küçük GCV ölçümüne sahip olan değerdir.[202]

2.4.5. ROC Eğrisi Yöntemi

ROC (Receiver Operating Characteristic: Alıcı İşlem Karakteristiği) eğrisi istatistik karar teorisine dayanır. 1950'lilerin başlarında elektronik sinyal tanımlamaları ve radar problemlerinde kullanılmaya başlanmıştır. 1960'lı yıllarda deneysel psikolojide kullanılmıştır. 1967'de Leo Lusted adında bir radyolog tarafından tıpta kullanımı önerilmiş ve 1969 yılında da medikal görüntüleme cihazları ile ilgili karar süreçlerinde kullanılmaya başlanmıştır.[203] ROC analizinde ortaya çıkan bu gelişmeler istatistiksel sonuçların değerlendirilmesi ve kıyaslanmasına duyulan gereksinimin doğal bir sonucudur.[204]

Buna göre oluşturulan sınıflama modelinin sonuçlarının değerlendirilmesinde, testin ayırt etme gücünün belirlenmesinde, testlerin etkinliklerinin karşılaştırılmasında, uygun pozitiflik eşiğinin belirlenmesinde, laboratuar sonuçlarının kalitesinin izlenmesinde ROC eğrisi yaygın olarak kullanılmaktadır. [205]

Böylelikle modelin başarısını ölçmek için öncelikle gerçek sınıf değerleri ile öngörülen sınıf değerleri aşağıdaki gösterilen bir kontenjans tablosu ile düzenlenir.[206]

---

[200] Friedman, Multivariate Adaptive Regression Splines, s.21.
[201] Shieu-Ming Chou, ve Diğerleri, "Mining the "Breast Cancer Pattern Using Artificial Neural Networks and Multivariate Adaptive Regression Splines", Expert Systems with Applications, Vol.27, No.1, (2004), s.136.
[202] Salford Systems, **MARS™ User Guide**, San Diego, 2001, s.37.
[203] Leman Tomak ve Yüksel Bek, "İşlem Karakteristik Eğrisi Analizi ve Eğri Altında Kalan Alanların Karşılaştırılması", 2010, Vol.27, No.2, http://dergi.omu.edu.tr/omujecm/article/view/1009001569 (31 Ekim 2014).
[204] Ayça Deniz Ertorsun ve Diğerleri, "ROC (Receiver Operating Characteristic) Eğrisi Yöntemi ile Tanı Testlerinin Performanslarının Değerlendirilmesi", Başkent Üniveristesi, Tıp Fakültesi, http://tip.baskent.edu.tr/egitim/mezuniyetoncesi/calismagrp/ogrsmpzsnm12/10.2.pdf (25 Ekim 2013), s.3-4.
[205] Yusuf Çelik, "Duyarlılık (Sensitivity) ve Belirleyicilik (Specificity)", 2014, Dicle Üniversitesi, Tıp Fakültesi, http://www.dicle.edu.tr/Contents/3b4d94e3-3582-4384-939b-c5957b348bdd.pdf (31 Ekim 2014).
[206] Dondurmacı, s.36.

Tablo 2.2. Kontenjans Tablosu

Test Sonucu	**Gerçek Durum**	
	Gerçek Pozitifler	**Gerçek Negatifler**
Öngörilen Pozitifler	DP	YP
Öngörülen Negatifler	YN	DN

DP, doğru pozitiflerin sayısını, YP yalış pozitiflerin sayısını, DN doğru negatiflerin sayısını, YN yanlış negatiflerin sayısını ifade etmektedir. Model için kurulan bu kontenjans tablosu sayesinde, model başarısını hesaplamak için aşağıdaki değerler hesaplanır.

$$\textit{Doğruluk (Doğru Sınıflandırma Oranı, DSO)} = \frac{DP + DN}{DP + DN + YP + YN} \quad (2.28)$$

$$\textit{Hata Oranı (Yanlış Sınıflandırma Oranı, YSO)} = \frac{YP + YN}{DP + DN + YP + YN} \quad (2.29)$$

$$\textit{Duyarlılık (Doğru Pozitif Oranı, DPO)} = \frac{DP}{DP + YN} \quad (2.30)$$

$$\textit{Özgüllük (Doğru Negatif Oranı, DNO)} = \frac{DN}{DN + YP} \quad (2.31)$$

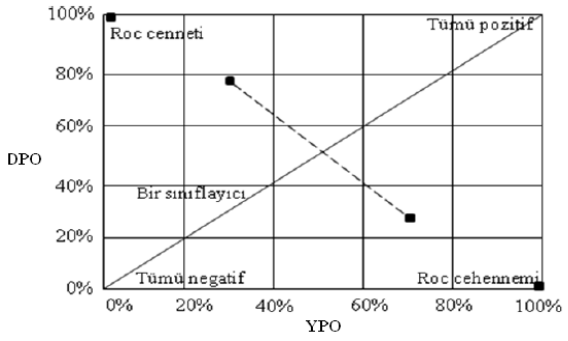
$$\textit{Yanlış Pozitif Oranı (1-Özgüllük; YPO)} = \frac{YP}{DN + YP} \quad (2.32)$$

$$\textit{Yanlış Negatif Oranı (1-Duyarlılık; YNO)} = \frac{YN}{DP + YN} \quad (2.33)$$

Yukarıdaki eşitliklerden de görüldüğü üzere, doğruluk ya da doğru sınıflandırma oranı olarak da adlandırılabilecek olan bu oran gerçekte pozitif ve negatif olan tüm olgular içinden, uygulanan test sonucuna göre pozitif ve negatif olarak kestirilenlerin oranını gösterir. Hata oranı ya da yanlış sınıflandırma oranı olarak

adlandırılan bu oran ise, tüm olgular içinden, uygulanan test sonucuna göre yanlış kestirilenlerin oranını verir. Duyarlılık (sensitivity) gerçekte pozitif olan tüm olgular içinden, uygulanan test sonucunda pozitif bulunan olguların oranını verir. Özgüllük (specificity) ise, gerçekte negatif olan tüm olgular içinden, test sonucunda negatif bulunanların oranını gösterir. Yanlış pozitif oranı (1-özgüllük), gerçekte negatif olan tüm olgular içinden, test sonucunda pozitif bulunanların oranını belirtir. Son olarak yanlış negatif oranı (1-duyarlılık) ise, gerçekte pozitif olan tüm olgular içinden, test sonucunda negatif bulunanların oranını gösterir.[207]

Buna göre ROC eğrisi, sınıflandırma modellerlerinin duyarlılık (sensitivity) ve özgüllüğünü (specificity) ölçmek için kullanılan bir yaklaşımdır. Bunun için araştırılan durumun duyarlılığı ve özgüllüğü arasındaki ilişkiyi tanımlayacak ROC eğrileri kullanılır. Bir ROC grafiğinde yanlış pozitif oranı (YPO) yatay eksende, doğru pozitif oranı (DPO) düşey eksende gösterilir. Her kesim noktasındaki doğru pozitif ve yanlış pozitife karşılık gelen noktalar birleştirilerek ROC eğrisi çizilir. Bu grafiğin (0,1) noktası tüm negatif ve pozitif durumların doğru olarak tahmin edildiği mükemmel bir sınıflamanın gerçekleştiğini (ROC cennetini) ifade eder. (0,0) noktası tüm durumların negatif, (1,1) noktası tüm durumların pozitif ve (1,0) noktası tüm durumların hatalı tahmin edildiği (ROC cehennemini) ifade eder. Bu durum aşağıda verilen şekil ile gösterilmiştir.



Şekil 2.12.: ROC Uzayı
Kaynak: Gülser Dondurmacı, "Veri Madenciliği'nde Regresyon Ağaçları ile Sınıflandırma ve Bir Uygulama", (Yayınlanmamış Doktora Tezi, Mimar Sinan Güzel Sanatlar Üniversitesi, FBE, 2011), s.38.

[207] Burcu Köksal, "Regresyon Analizinde ROC Eğrisi Kestirimi ile Model Seçimi", (Yayınlanmamış Yüksek lisans Tezi, Marmara Üniversitesi, SBE, 2011), s.45-47.

Buna göre, düşey eksene ve üst sınıra yakın olan eğriler başarılı bir testi ifade ederken, 45^o gibi eğime sahip eğriler başarısız bir testi göstermektedir. Böylece ROC eğrileri incelenerek testin başarısı belirlenebilir. Başarılı bir testte eğrinin altında kalan alanın büyük olması beklenir. Test ne kadar iyi ise eğri o kadar yukarıya yani yüksek duyarlılık yani doğru pozitif oranına yaklaşır.[208]

Buna göre ROC eğrileri ile bir teste ilişkin performans değerlendirilmesinde eğri altında kalan alana (Area Under Curve: AUC) gereksinim duyulur. Pozitif ve negatif sınıflarının ayırım etkinliği kestirilen ROC eğrisinin performansının değerlendirilmesinde ve model doğruluklarının karşılaştırılmasında yaygın olarak kullanılan bir göstergedir. AUC' un tanım aralığı $0{,}5 \leq AUC \leq 1$ olup, 0,5 ve 1 sırasıyla alt ve üst sınırları oluşturmaktadır. AUC değerinin alt sınıra eşit bulunması rassal ayrımı ifade eden köşegen çizgisinin altındaki alanı belirtir. AUC değerinin üst sınıra eşit bulunması ise tam ayırımı ifade etmektedir. Modelin ayırım gücü, aşağıdaki tabloda verilen sonuçlara göre belirlenmektedir.[209]

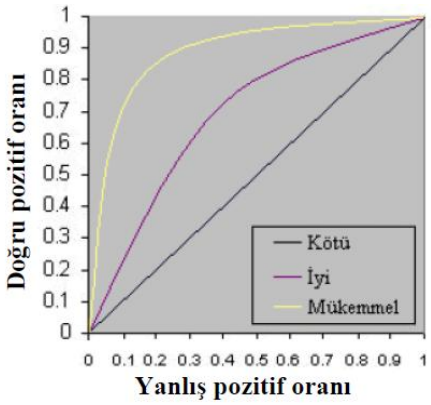
Tablo 2.3. Modelin Ayırım Gücünün AUC Değerlerine Göre Değerlendirilmesi

AUC =0,5	Model ayırım gücüne sahip değildir
0,5 <AUC < 0,7	Model zayıf ayırım gücüne sahiptir
0,7 ≤AUC < 0,8	Model kabul edilebilir bir ayırım gücüne sahiptir
0,8 ≤AUC < 0,9	Model mükemmel bir ayırım gücüne sahiptir
0,9 ≤AUC	Model üstün bir ayırım gücüne sahiptir

Kaynak: Burcu Köksal, "Regresyon Analizinde ROC Eğrisi Kestirimi ile Model Seçimi", (Yayınlanmamış Yüksek lisans Tezi, Marmara Üniversitesi, SBE, 2011), s.50.

Buna göre ideal ve kötü performans gösteren testlere ilişkin ROC eğrileri aşağıdaki şekilde gösterilmiştir.

[208] Dondurmacı, s.37-38.
[209] Burcu Köksal, "Regresyon Analizinde ROC Eğrisi Kestirimi ile Model Seçimi", (Yayınlanmamış Yüksek lisans Tezi, Marmara Üniversitesi, SBE, 2011), s.49-50.



Şekil 2.13: Performanslarına Göre ROC Eğrileri
Kaynak: Ayça Deniz Ertorsun ve Diğerleri, "ROC (Receiver Operating Characteristic) Eğrisi Yöntemi ile Tanı Testlerinin Performanslarının Değerlendirilmesi", Başkent Üniveristesi, Tıp Fakültesi, http://tip.baskent.edu.tr/egitim/mezuniyetoncesi/calismagrp/ogrsmpzsnm12/10.2.pdf (25 Ekim 2013), s.4.

ÜÇÜNCÜ BÖLÜM

UYGULAMA

Bu bölümde Türkiye'deki gençlerin değerleri, beklentileri, hayat tarzları, tüketim tercihleri, aileyle çatışma alanları ve dolayısıyla dünya görüşlerine göre şekillenen siyasi görüşleri incelenmiştir. Bu amaçla Türkiye'deki gençlerin siyasi partilere bakış açılarını modelleyebilmek için, karar ağacı yöntemlerinden olan CART ve MARS modelleri kullanılarak sonuçlar karşılaştırmalı olarak değerlendirilmiştir.

3.1. Türkiye Gençliği ve Siyasi Görüşleri

Türkiye gençliği ve gençliğin siyasi görüşlerini etkileyen faktörler anlatılmadan önce gençlik, siyasal tutum ve siyasal parti kavramlarının tanımlanması gerekmektedir. Gençlik kavramının neyi ifade ettiği konusunda literatürde kabul görmüş bir tanım ve yaş aralığı yoktur. Gençliğin tanımı yapılırken sosyolojik, psikolojik ve biyolojik tanımlara yer verilmektedir. Bu yüzden homojen bir tanım yapılması konusunda sıkıntı yaşanmaktadır. Fakat en homojen gençlik tanımı, güçlü bir devlet tarafından sunulan bir ideal çerçevesinde ideoloji empoze edilen genç kesimler şeklinde ifade edilmektedir. Kimi çalışmalar, 12–24 yaş grubunu, kimi çalışmalar ise 12–26 yaş grubunu, kimileri ise 15–24 ya da 15–30 yaş grubunu genç olarak ele almaktadırlar. Birleşmiş Milletler (BM), Birleşmiş Milletler Eğitim, Bilim ve Kültür Kurumu (UNESCO) ve Dünya Bankası (WB), 15–24 yaş aralığındaki kişileri "genç" olarak tanımlamaktadır. Bununla birlikte, UNESCO, genç insanların değişen ve heterojen bir grup olduğunu ve genç olma tecrübesinin bölgelere ve ülkelere göre büyük değişiklik arz ettiğini vurgulamaktadır. İngiliz Milletler Topluluğu Gençlik Programı gibi örgütler, genç kişileri 15-29 yaş aralığındakiler olarak tanımlamakta; Avrupa Birliği'nin bazı raporlarında da gençler 15-29 yaş aralığındakiler olarak tanımlanmaktadır.[210]

Bireyin belirli bir siyasal konu ile ilgili kanı ve davranışlarının tümü ise, bireyin o konuya ilişkin siyasal tutumunu oluşturur. Sözgelişi, bir birey demokratik bir

[210] Siyaset, Ekonomi ve Toplum Araştırmaları Vakfı (SETA), **Türkiye'nin Gençlik Profili,** 1. Baskı, Ankara: Pelin Ofset, 2012, s.15.

tutum sahibi ise o kişinin siyasi meseleler hakkındaki kanı ve davranışları hakkında çıkarsama yapmak mümkündür. O kişi büyük bir ihtimalle ailesinde herkese söz hakkı tanıyacak, esnek ve hürriyetlerin yaygın olduğu bir siyasal rejimden taraf olacak, insanların eşit ve hür olduklarına inanacaktır. Aynı şekilde otoriter tutum sahibi bir birey ise ata erkil bir aile yapısını savunacak, siyasal rejimin sert ve otoriter olmasından yana olacak, her zaman kendi görüşlerinin doğru, karşıt görüşlerin yanlış olduğunu iddia edecektir.[211] Yani siyasal tutumlar bireyin siyasal bir obje karşısında bireyde var olan duygu, düşünce ve davranışlarından oluşan düzenli bir eğilim olarak ifade edilebilir.[212]

Siyasal partiler ise, belirli bir yapısı, kuruluş biçimi, ortaklaşa karar alan organları, lideri, üyeleri olan örgütlerdir. Birlikte karar alma, parti içerinde kurulmuş çeşitli bağlantılara ve sıralanmalara rağmen partilerin temel etkinliğidir.[213] Siyasal partilere insanlar, parti ilkelerinin ve görüşlerinin paylaşılması, parti liderlerine duyulan bağlılık, partilerin ekonomik ve sosyal konulara yaklaşımları, siyasal erki etkileme isteği, dinsel inançlar, coğrafi bölgenin veya insanların doğduğu ya da yaşadığı yerin özelliği, aile geleneği gibi çeşitli gerekçelerle katılırlar.[214]

"Gençlik çalışmaları", sosyal bilimler alanında pek çok farklı disiplin altında gençlik konusu üzerine yapılmış çalışmaların bir toplamını ifade etmektedir. Türkiye'de yapılan gençlik çalışmalarına türleri açısından sayısal olarak bakıldığında, 5000'in üzerinde olan bu yayınlar arasında, makalelerin (2500'ün üzerinde) ilk sırayı aldığını, ardından tezlerin (1500'ün üzerinde) ve kitapların (1000'e yakın) geldiği görülmektedir. Bu literatür taramasında gençlik çalışmalarının sadece akademik disiplinlerin ilgi

---

[211] Ahmet Taner Kışlalı, **Siyaset Bilimi,** Ankara: İmge Kitapevi, 1996, s.126-127 Aktaran: *Üniversite Gençliğinin Siyasal Tutumları Üzerine Bir İnceleme-Süleymen Demirel Üniversitesi Örneği,* (t.y.) http://www.yerelsiyaset.com/pdf/kasim2007/17.pdf (26 Ocak 2015).

[212] *Üniversite Gençliğinin Siyasal Tutumları Üzerine Bir İnceleme-Süleymen Demirel Üniversitesi Örneği,* (t.y.) http://www.yerelsiyaset.com/pdf/kasim2007/17.pdf (26 Ocak 2015).

[213] Mümtaz Soysal, "Parti ve Hareket" **Milliyet,** 25 Aralık 1994, s.15 Aktaran: Türkay Nuri Tok, "Türkiye'deki Siyasal Partilerin Eğitim Söylemleri ve Siyasaları", **Kuram ve Uygulamada Eğitim Yönetimi,** Cilt.18, Sayı.2 (2012), s.279.

[214] Bülent Daver, **Siyaset Bilimine Giriş,** 5. Basım, Ankara: A.Ü. SBF Yayınları, 1993 Aktaran: Türkay Nuri Tok, "Türkiye'deki Siyasal Partilerin Eğitim Söylemleri ve Siyasaları", **Kuram ve Uygulamada Eğitim Yönetimi,** Cilt.18, Sayı.2 (2012), s.279.

odağında olmadığı, aynı zamanda akademi dışı üretimlerin de gençlik çalışmalarının yarısına yakınını oluşturduğu gözlemlenmiştir.[215]

Sosyolog Ömer Miraç Yaman'ın ifade ettiği gibi:

Akademik çalışmaların tematik olarak daha ziyade gençliğin sorunları, gençlik araştırmaları, kuşak çatışması, gençliğin eğitimi, gençlik kültürü, gençlik ve kimlik konuları etrafında yoğunlaştığı görülmektedir. Akademi dışı üretimlerde ise, gençliğe nasihat verme ve ideolojik olarak bakış açısı kazandırma gayreti ile gençliği ve gençlik dönemini idealleştirme tavrı öne çıkmaktadır. Akademi dışı üretimlerin temelindeki bu yaklaşım, aslında 1923'ten bugüne gençlik konusunu ele alan yazarların pek çoğunda genel bir yaklaşım biçimi olarak belirmektedir. Bu durum Cumhuriyet tarihi boyunca gençlik konusuna olan ilginin, çoğu kez ideolojik bir tutum ve gençliğe dair görev odaklı bir beklenti bağlamında oluştuğunu göstermektedir. Nitekim bu eserlerde, Türkiye'de özellikle siyasal ve toplumsal çalkantıların arttığı askeri darbe dönemleri başta olmak üzere, neredeyse her siyasal iktidarın yapısına bağlı olarak değişen ve yeniden şekillenen bir gençlik imgesinin varlığından söz edilebilir. Dolayısıyla tarihsel olarak gençliğe dair yapılan yayınların dönem dönem azalıp çoğalmasında siyasal ve toplumsal değişimlerin belirleyici olduğu kolaylıkla fark edilebilmektedir.[216]

Buna göre, Türkiye gençliği ve gençliğin siyasi görüşlerini konu alan çalışmaların bazılarına ait literatür özeti çıkış yıllarına göre aşağıdaki tabloda gösterilmiş ve değerlendirilmiştir.

[215] Ömer Miraç Yaman, "Türkiye'de Gençlik Sosyolojisi Çalışmalarına Dair Bibliyografik Bir Değerlendirme", **Alternatif Politika,** Cilt.5, Sayı.2, (2013), s.116.
[216] Yaman, s.117.

Tablo 3.1. Türkiye Gençliği ve Siyasi Görüşleri Araştırması Literatür Özeti

Dönem	Örnek Olarak Seçilen Eserler	Eserlerin Araştırma Konusu	Yöntem	Eserlerin Bulguları
1923-1950	*****Mehmet Emin Erişirgil (1925),"İrtica Karşısında Gençlik" *****İbrahim Alaettin Gövsa (1927), "İlk Gençlik Hakkında Rûhiyat ve Terbiye Tedkikleri" *****Asaf Burhan (1931), "İnkılabımız ve Gençlik" *****Felida Sedat (1932), "Genç Kızlara Muaşeret Usulleri" *****Şevket Süreyya (1932) "Genç Nesil Meselesi" *****Mehmed Çerçi (1936), "Kemalizm ve Türk Gençliği" *****İsmail Hakkı Baltacıoğlu (1939), "Gençler İçin En Büyük Tehlikeler" *Pertev Demirhan (1942), "Türk Çocuğuna Öğütlerim" *Şevket Aziz Kansu (1947), "Gençlikte İrade Eğitimi ve Büyük Adamlar" *Ali Fuat Başgil (1949), "Gençlerle Başbaşa"	Eğitim bilimleri, Gençliği terbiye etmeye dönük adab-ı muaşeret kuralları, Yeni kurulan Cumhuriyetin milli hedeflerini gençlere aktarma gayreti içinde olan, Gençliği tehlikeli fikirlerden uzaklaştıran, ne yapmasını, neye inanmasını, nasıl davranmasını öğütleyen çalışmalar öne çıkmaktadır.	Nitel Araştıma	Cumhuriyetin ilk dönemi sayılan bu yıllarda büyük oranda Cumhuriyet'in temel hedeflerine kararlı bir şekilde bağlı, çoğu zaman Cumhuriyet'in "yılmaz bir bekçisi" olarak motive edilen bir gençlik imgesi öne çıkmaktadır.
1950-1980	*Nermin Abadan (1961), "Üniversiteli Öğrencilerin Boş Zaman Faaliyetleri" *Özer Ozankaya (1966), "Üniversite Öğrencilerinin Siyasal Yönelimleri" *Yavuz Ulusu (1967), "Yüksek Öğrenim Gençliğinin Sorunları ve Aşırı Cereyanlar" *Neriman Öztürkmen (1968), "Kalkınan Türkiye'de Eğitim ve Gençlik-Anket, Röportaj" *Senih Onat (1968), "Üniversite Olayları ve Demirel" *Üniversite Gençlik Anketi: Üniversite Gençliğinin Problem ve Vaziyet Alışları (1970) *Engin Gençtan (1971), "Üniversite Gençlik Sorunları- Gençlik Sorunları Özel İhtisas Komisyonu Raporu, DPT" *İbrahim Öktem (1971), "Gençlik Yeni Bir Düzen İstiyor" *Ahmet Taner Kışlalı (1972), "Öğrenci Ayaklanmalarının Nedenleri ve Rejim Sorunu İle İlişkisi" *Aysel Ekşi (1973), "Gençlerde Kimlik Konusunda Bir Araştırma" *Ahmet Taner Kışlalı (1974), "Öğrenci Ayaklanmaları *Tufan Ata Türkyılmaz (1978),"Gençlik ve Eğitim Üzerine - Demokratik Öğrenci Muhalefetine Bakış" *T. Alkan (1979), "Siyasal Toplumsallaşma, Siyasal Bilincin Gelişmesinde Ailenin, Okulun ve Toplumsal Sınıfların Etkisi"	Bu dönem 1968 öncesi ve sonrası olarak ele alındığında, 1950-1968 aralığında öne çıkan çalışma konuları, üniversite öğrencilerinin sorunları, boş zaman faaliyetleri, siyasal yönelimleri ve tutumları olarak sıralanabilir. 1968 yılı dünya gençliği için bir dönüm noktası olarak belirirken, Türkiye'de de gençlik ve öğrenci hareketlerinin öne çıkmaya başladığı bir dönemin kapısı aralanmaya başlamış bu kapsamda 1968-1980 yılları arasındaki çalışmalarda büyük oranda gençlik-öğrenci hareketlerini analiz etmeye çalışan, özellikle üniversite öğrencilerinin beklenti ve sorunlarına odaklanan, gençlik hareketlerinin siyasal yönlerinden hareketle sağ ya da sol görüşe sahip gençlerin düşünce ve aksiyon dünyalarına dair incelemeler öne çıkmaktadır.	1950-1968 aralığında genelde Nitel Araştımalar yapılmıştır. 1968-1980 aralığında ise genelde gençlik konulu tez çalışmaları üzerinde yoğunlaşılmış. Anket çalışmalarına yer verilmiştir. Nitel ve Nicel Araştırma Yöntemleri kullanılmıştır.	Türkiye'nin tek partili hayattan çok partili siyasal sisteme geçtiği, Cumhuriyet'in kurucu ideolojisinden farklı ideolojik yaklaşımların Türkiye'de etkin bir konuma yükseldiği, üniversite eğitiminin yeni kurulan üniversiteler aracılığıyla yaygınlık kazandığı bu dönemde, gençlik üzerine yapılan çalışmalar nitelik değiştirmeye başlamıştır. Bu değişim sürecinin asıl aktörü olarak gençlerin ve gençlik hareketlerinin belirleyici olduğu çalışmalar öne çıkmaktadır.

Dönem	Örnek Olarak Seçilen Eserler	Eserlerin Araştırma Konusu	Yöntem	Eserlerin Bulguları
1980-1990	*****Atalay Yörükoğlu (1985), "Gençlik Çağı (Ruh Sağlığı ve Eğitimi)" ***** Barlas Tolan (1985), "Gençlik ve Sorunları", Gençliğin Topluma Karşı Daha Faydalı Olmalarında Karşılaştıkları Engeller Paneli *****Milli Eğitim ve Gençlik ve Spor Bakanlığı (1986) , "12-24 Yaş Gençlerin Sosyo-Ekonomik Sorunları, Yayınları" *****Neşet Çağatay (1986) , "Gençliğin Eğitimi", İş Bankası Yayınları *****Cahide Baydilli (1987), "Lise Son Sınıf Öğrencilerinin Meslek Seçimini Etkileyen Bireysel ve Toplumsal Faktörler" ***** Muzaffer Erendil (1987), "Atatürk'ün Güvendiği Gençlik ve Eğitim", Ankara, Türk Tarih Kurumu Basımevi ***** Mahmut Tezcan (1987), "Yurt Dışından Dönen Gençlerin Uyum Sorunları - Eğitim Sistemi ve Topluma Uyum" ***** Milli Eğitim Gençlik Spor Bakanlığı Gençlik Hizmetleri Genel Müdürlüğü (1988), "Çalışmayan Gençliğin Sorunları: 1. Gençlik Şurası Ön Çalışma Raporu", Ankara *****Sinan Erdoğan (1988) "Dış Göç ve F. Almanya'dan Dönüş Yapmış (Sivas Ortaokul ve Liselerinde) Gençlerin Uyum Sorunları Üzerine Sosyolojik Bir İnceleme", Cumhuriyet Üniversitesi, Sosyoloji Bölümü, Y. Lisans Tezi	1980'li yılların ilk yarısında gerek akademide gerekse de akademi dışında yapılan yayınlarda gençlik konusuna çok az değinilmiş, resmi ideoloji tarafından "sorunlu ve tehlikeli" ilan edilen gençlik hareketleri başta olmak üzere neredeyse gençliğe dair pek çok alanda tam bir suskunluk hali yaşanmıştır. 1985'in Birleşmiş Milletler tarafından dünya çapında "Gençlik Yılı" ilan edilmesinin ve bu yıllardan itibaren Türkiye'deki siyasetin ve toplumsal baskının kısmen rahatlaması ile özellikle, gençlik dönemi sorunları, gençlerin ebeveynleri ile ilişkileri, gençliğin gelecek kaygısı-beklentisine dair tutumları, gençlik ve akran ilişkileri, kuşaklararası iletişim problemlerine ait çalışmaların öne çıktığı görülmektedir. Aynı dönemde öne çıkan bir başka temel konu ise, yurt dışında yaşayan Türk vatandaşlarının birinci nesil çocukları ve gençlerinin uyum sorunlarını tespit etmeye yönelik çalışmalardır.	Nitel ve Nicel Araştırma Yöntemleri kullanılmıştır.	1980 darbesi sonrası siyasal atmosferde yaşanan kaotik durum, Türkiye'de siyasetin uzunca bir süre resmen askıya alınması, gençlik hareket ve örgütlerinin darbe sürecini hazırlayan temel etmen olarak nitelendirilerek suçlu ilan edilmesi ve yasal olarak cezalandırılması süreci ile Türkiye'de gençlik meselesinin baştan başa yeniden tanımlandığı bir dönemin kapısını aralamıştır.
1990-2000	*Mahmut Tezcan (1991), "Gençlik Sosyolojisi Yazıları" * Fügen Berkay (1996), "Gençlik Sosyolojisi" *Süreyya Eryaşar (1996), "Atatürkçülük ve Gençlik", Atatürkçü Düşünce, Ankara *Mahmut Tezcan (1997), "Gençlik Sosyolojisi ve Antropolojisi Araştırmaları", Ankara, Ankara Üniversitesi EBF Yayınları *İlker Alp (1997), "Atatürk ve Türk Gençliği", Atatürk Araştırma Merkezi Dergisi, Ankara *Cihan Dura (1997), "Gençler! Atatürk'ü Tanıyın", Atatürkçü Düşünce, Ankara *Esra Burcu (1998), "Gençlik Teorilerinin Sınıflandırılmasına İlişkin Bir Çalışma", Sosyoloji Araştırmaları Dergisi, Ankara *Esra Burcu (2000), "Sosyolojik Bakış Açısından Atatürk'ün Gençliğe Verdiği Önem", Silâhlı Kuvvetler Dergisi, Ankara	Bu dönemde özellikle meslek ve kariyer planlaması ile eğitimde kaliteyi arttırmayı önceleyen yayınlarda bir artış yaşanmış; gençlik ve madde bağımlılığı, alkol ve uyuşturucu madde kullanımı, artan suç oranları, intihar ve şiddet olgusu gençlik çalışmalarında sıkça rastlanan konular arasına girmiştir. Özellikle Türkiye'de yükselen "İslamcı Siyaset"in bir yansıması olarak dindar ve "İslamcı" gençlik üzerine çalışmaların da yoğunlaştığına tanık olunmakta; bu yayınlara paralel bir şekilde artış gösteren "Atatürk ve gençlik" temalı çalışmaların da öne çıktığı gözlenmektedir.	Nitel ve Nicel Araştırma Yöntemleri kullanılmıştır.	1990 sonrası dönem aslında Türkiye gençliği açısından söylemsel planda yeni bir dönemin de habercisi olarak gözükmektedir. Gençlik üzerine çalışan pek çok araştırmacı için 1990'lı yıllarla başlayan yeni dönem; gençliğin liberalleştiği, fikir ve aksiyon anlamında niteliksizleştiği ve "Özal gençliği", "depolitik gençlik" ya da "apolitik gençlik" olarak ifade edilen amaçsız bir gençlik tanımlamasının öne çıktığı bir zaman dilimine karşılık gelmektedir.

Kaynak: Bibliyografik bir değerlendirmeden yararlanılarak oluşturulmuştur. (Yaman, s. 114-138.)

2000'li yıllardan sonra ise gençlik konusunda yürütülen çalışmalar niteliksel olarak çeşitlenmiş ve yeni konular gençlik alanına dahil edilmiştir. Son dönemde özellikle genç işsizliği, gençlik ve değer ilişkisi, gençliğin gelecek kaygısı-beklentisi, teknolojik gelişmeler doğrultusunda gençliğin internet, sosyal medya ve sanal dünya ile ilişkileri, gençlik ve tüketicilik rolü, gençlik altkültürleri öne çıkan çalışma konuları olmuştur. Bu dönemde ayrıca Türkiye gençliği üzerine yapılan saha araştırmalarında da bir artış olmuştur. Gençliğin kimlik arayışını, kültürel alışkanlıklarını, bireysel ve toplumsal taleplerini konu edinen araştırmalar öne çıkmıştır.[217]

Buna göre 2000'li yıllarda Türkiye'de gençlik ve gençliğin ve siyasi görüşlerine yönelik yapılan bir takım saha çalışmaları incelendiğinde ortaya çıkan bulgular incelenebilir. Bunlardan bir tanesi Hasan Tüzer ve Mehmet Meder tarafından 2002 yılında, Pamukkale Üniversitesi Öğrencilerinin Toplumsal, Ekonomik ve Siyasal Eğilimleri üzerine yapmış olduğu bir araştırmadır. Bu araştırma, 2001-2002 öğretim yılında Pamukkale Üniversitesi'nin çeşitli fakültelerinde öğrenim gören 421 öğrenciyi kapsamaktadır. Bu çalışmada öğrencilerin büyük bir çoğunluğu gelir dağılımının eşitsizliğini, çeşitli alanlarda fırsat eşitliğininin bulunmadığını, yasaların tarafsız ve eşit davranılması konusunda eşit olmadığını, Türkiye'nin demokratik bir ülke görünümü veremediği ve mevcut şartlarda politikayı düşünmediğini vurgulamıştır.[218]

Bu konuda yapılmış araştırma raporları ve daha kapsamlı çalışmalara bakacak olursak bunlardan bir tanesi de, Selçuk Şirin tarafından New York Üniversitesi ve Bahçeşehir Üniversitesi ortaklığında gerçekleştirilmiştir. Selçuk Şirin'in "Genç Kimlikler-Siyasal, Kültürel ve Sosyal Kimlikler Bakımından Türkiye Gençliği" başlıklı araştırması, 2009 yılında yaşları 18-25 arasında değişen 1403 gençle birebir görüşme yöntemiyle gerçekleştirilmiştir. Bu çalışma ile gençlerin dünya görüşlerine, mezheplerine ve etnik kimliklerine göre siyasete ve siyasi partilere bakış açısı ortaya konulmaya çalışılmıştır. Yapılan araştırmalar, gençlerin siyasal kimliklerinden dolayı çeşitli zorluklar yaşadığı ortaya koymakla beraber, onların aynı zamanda var olan kamplaşmalar dışında çok yönlü siyasal kimlikler kurma yönünde son derece yaratıcı

[217] Yaman, s.121-122.
[218] Hasan Tüzer ve Mehmet Meder, "Pamukkale Üniversitesi Öğrencilerinin Toplumsal, Ekonomik ve Siyasal Eğilimleri Üzerine Bir Araştırma, **Pamukkale Üniversitesi Eğitim Fakültesi Dergisi,** Cilt.1, Sayı.11, (2002), s.147.

olduklarını göstermektedir. Araştırmada ortaya çıkan önemli sonuçlardan biri de laik kimlik ile laik bilinç arasındaki fark olup, laik kesime ait olduğunu söyleyen gençlerin, laikliğin gereği olan önermelerde tutarsız davranmaktadır. Bu çalışmada ayrıca gençlere klinik psikolojide kullanılan BECK-Umutsuzluk Ölçeği adı verilen bir test de uygulanmıştır. Bu teste çıkan sonuçlar, gençlerin geleceklerinden oldukça kaygılı olduğunu göstermektedir. Özellikle Kemalist ve Kürt kökenli gençler, fırsat bulursanız başka bir ülkeye yerleşir misiniz sorusuna en çok evet diyenleri oluşturmaktadır.[219]

Bu kapsamda yapılan en kapsamlı projeler ise, Avrupa Birliği ve Türkiye tarafından finanse edilen, Şebeke: Gençlerin Katılımı Projesi (ŞEBEKE) çerçevesinde gerçekleştirilen çalışmalardır. Bu Projenin yürütücüsü olan İstanbul Bilgi Üniversitesi Sivil Toplum Çalışmaları Merkezi; STK Eğitim ve Araştırma Birimi, Gençlik Çalışmaları Birimi ve Çocuk Çalışmaları Birimi olmak üzere üç ayrı birim olarak yapılanmıştır. ŞEBEKE'nin amacı, genç yurttaşların ve gençlerle çalışan sivil toplum kuruluşlarının (STK) kamusal tartışmalara ve karar alma mekanizmalarına katılımını güçlendirmektir. Aynı zamanda, gençlerin toplumsal katılımının desteklenmesi hedeflenmektir. ŞEBEKE çerçevesinde gerçekleştirilen çalışmalarda bir yandan gençlerin katılımının siyasal, toplumsal ve ekonomik boyutları ele alınırken, diğer yandan bu katılım konularını birbirinden ayrı araştırılmak ve tartışılmak yerine, birbiri ile bağlantılı; farklı konular üzerinden birbirini etkileyen, hatta kesişen tartışma zeminlerinin oluşturulması gözetilmiştir.[220]

2012 yılında başlayan bu çalışmalar 2 yıllı kapsayan bir uygulama süresinin sonunda tamamlanmıştır. Bu sürede ŞEBEKE kapsamında beş araştırma yapılmıştır. Yapılan çalışmalar sırasıyla incelenecek olursa bunlardan ilki "Türkiye'de Gençlerin Katılımı" adlı çalışmadır. Bu çalışmada KONDA Araştırma ve Danışmanlık şirketi tarafından gerçekleştirilmiştir. Türkiye genelindeki gençleri temsil etmeyi hedefleyen araştırma kapsamında, 18-24 yaş aralığında olan 2508 gençle görüşülmüştür. Genel olarak siyasal, toplumsal ve ekonomik boyutları inceleyen bu araştırmanın gençler ve siyasi partilere katılım değerlendirmesinde şu sonuçlara varılmıştır. Gençlerin yüzde 9'u

[219] *Türküm, Laikim, Müslümanım*, 2009, http://www.hurriyet.com.tr/gundem/12104881.asp (26 Ocak 2015).
[220] Volkan Yılmaz ve Burcu Oy, Türkiye'de Gençler ve Siyasi Katılım: Sosyo-Ekonomik Statü Fark Yaratıyor mu? , **"Şebeke: Gençlerin Katılımı Projesi",** İstanbul, 2014, s.7-10.

siyasi partilerden birine üye olduğunu, üye olmasa da partide veya gençlik kollarında aktif olarak rol aldığını belirtiyor. Olmayı düşünenler ve üyelikten ayrılanlar ile birlikte, dörtte biri siyasi partilerle ilgileniyor. Ancak gençlerin büyük çoğunluğunun siyasi partiye üye olmadığı ve üye olmak da istemediği dikkat çekiyor. Gençlerin çalışanları, aileleri olmadan uzun süre yaşayabileceklerini düşünenleri, geliri yüksek olanları, kısaca ekonomik olarak görece kendine güveni nispeten yüksek olanları siyasi parti üyeliğine daha yatkın bulunmuştur. Diğer bulgularla beraber değerlendirildiğinde, gençlerin bir parti aidiyetiyle var olunan geleneksel siyasete ilgisiz oldukları söylenebilir. Siyasetle ilgilenmiyorum, parti üyeliğini gereksiz buluyorum, sevmiyorum türü cevaplara gençlerin yarıdan fazlasının rağbet ettiği belirtilmiştir.[221]

ŞEBEKE kapsamında ele alınan İkinci araştırma ise, Laden Yurttagüler'in "Meclisin Gençlik Söylemi: 1930-1990" adlı çalışmasıdır. Yurttaşların karar verme süreçlerine doğrudan katıldıkları mekanizma meclis olduğu için, araştırmanın kapsamı mecliste gençlerin hangi bağlamda tartışıldığını görünür kılmayı amaçlamaktadır. Mesliste genç vekil olmadığı için araştırma vekillerin gözünden gençlerin nasıl göründüğüne odaklanmıştır. Buna göre 1930-1990 yılları arasında mecliste, vekiller tarafından yapılan tartışmalar incelendiğinde vekillerin gençlere ilişkin iki temel kaygıyla hareket ettikleri sonucuna varılmıştır. Bunların ilki, gençleri korumaya yönelik çabadır. Gençlerin korunması gerektiğini düşünen ve uygulamalarını bu çerçevede belirleyen bu akıl, gençlerin istenmeyen konu ya da durumlarla temas etmemesini sağlayan görünmez (ve bir o kadar da görünür) bir çerçeve çizilmesine yol açmıştır. Vekillerin gençlerle ilgili olarak varmış oldukları ikinci yargı ise, gençlerin eğitimi alanıdır. Bu yüzden sık sık eğitimin gerekliliğinden söz edilmiş ve aileden zorunlu eğitime, askerlikten üniversiteye kadar pek çok farklı mecrada, birbirinin içine giren ya da birbirleriyle kesişen konularda gençlerin eğitilmeleri için yasalar çıkmıştır. Ancak buradaki temel sorun gençlerin kendi adlarına konuşabilen bireyler olarak bu süreçte yer alamamalarıdır.[222]

[221] KONDA Araştırma ve Danışmanlık, "Türkiye'de Gençlerin Katılımı", İstanbul, 2014, s.89. **"Şebeke: Gençlerin Katılımı Projesi",** İstanbul, 2014, s.89.
[222] Laden Yurttagüler, "Meclisin Gençlik Söylemi: 1930-1990", **"Şebeke: Gençlerin Katılımı Projesi",** İstanbul, 2014, s.113-114.

Bu kapsamda yapılan üçüncü araştırma, Laden Yurttagüler ve diğerlerinin "Özerklik ve Özgürlükler Açısından Türkiye'de Gençlik Politikaları" adlı çalışmasıdır. Bu çalışmanın konusu olan gençler, sosyal haklar konusunda önemli kısıtlar, hatta ihlaller yaşamaktadır. Gençler, gerek eğitimleri süresince gerekse eğitimlerini tamamladıktan sonra iş bulma konusunda zorlu bir süreçten geçmektedirler. Bunun sonucu olarak ise gençlerin, eğitim, sağlık ya da barınma gibi temel ihtiyaç ya da hizmetleri piyasadan almaları sınırlı ya da yetersizdir. Yurttaşlık hakkı olarak, kamu tarafından sağlanması beklenen hizmetler, diğer bir deyişle, destek mekanizmaları ise hem nicelik, hem de nitelik olarak kısıtlıdır. İhtiyaçlarını piyasadan satın alamayan, temel sosyal hakları kamu tarafından yurttaşlık hakkı bağlamında hiç ya da yeterince karşılanmayan gençler için geriye kalan tek güvenlik ağı ailedir. Fakat bu durum gençleri aileye bağımlı hale getirir. Zira ailenin ihtiyaçlarını karşılamadığı durumda gençler, muhtaç durumda kalmaktadırlar. Dolayısıyla, ailenin ilgili kaynakları sağlarken takınacağı tutum, gencin ne denli bağımsız olabileceğinin de belirleyicisidir. Araştırma bağlamında daha yakından bakılırsa, gençlerin haklarına ulaşamamaları sonucunda özerkliklerinin ne denli etkilendiği daha net görülebilir.[223] Bu çerçevede hem genel olarak gençlerin, hem de özel olarak farklı dezavantajları olan gençlerin refahlarını ve katılımlarını geliştirici politikalara ihtiyacın sadece bugünle ilgili olmadığı, nüfus projeksiyonlarınına bakıldığında yarınlar için de geçerli olduğu bu çalışmada belirtilmiştir.[224]

ŞEBEKE kapsamında yapılan dördüncü çalışma ise, Volkan Yılmaz ve Burcu Oy'un "Türkiye'de Gençler ve Siyasi Katılım: Sosyo-Ekonomik Statü Fark Yaratıyor mu?" adlı çalışmasıdır. Bu çalışmada ilk çalışma olan "Türkiye'de Gençlerin Katılımı" adlı çalışmanın devamı niteliğindedir. KONDA Araştırma ve Danışmanlık şirketi'nin 18-24 yaş aralığında olan 2508 gençle yapmış olduğu görüşmenin sonuçlarını sosyo-ekonomik açıdan değerlendirmiştir. Çalışmada gençlerin sosyo-ekonomik statüleri göstergeleri ile çeşitli siyasi katılımları arasındaki ilişki değerlendirilmiştir. ŞEBEKE kapsamında yapılan en son araştırmada ortaya çıkan genel sonuçlar ise; öğrencilerin ortalama siyasi katılım düzeylerinin, öğrenci olmayanlardan yüksek olması, burs ve

[223] Laden Yurttagüler, Burcu Oy ve Yörük Kurtaran, "Özerklik ve Özgürlükler Açısından Türkiye'de Gençlik Politikaları", **Şebeke: Gençlerin Katılımı Projesi,** İstanbul, 2014, s.33.
[224]Yurttagüler ve diğerleri, "Özerklik ve Özgürlükler Açısından Türkiye'de Gençlik Politikaları", s.116.

kredi alan öğrencilerin ortalama siyasi katılım düzeylerinin, burs ve kredi almayan öğrencilerden yüksek olması, eğitim düzeyi lise altında olan gençlerin ortalama siyasi katılım düzeyleri, eğitim düzeyi lise ve üstünde olan gençlerin ortalama siyasi katılım düzeylerinden düşük olması, açık öğretimde okuyan öğrencilerin ortalama siyasi katılım düzeyleri devlet üniversitesi öğrencilerinden düşük, vakıf üniversitesi öğrencilerinden ise neredeyse düşük olması, metropollerde yaşayan öğrencilerin ortalama siyasi katılım düzeyleri kentlerde yaşayan öğrencilerin ortalama siyasi katılım düzeylerinden daha yüksek olması, ailesini maddi açıdan dar gelirli olarak tanımlamış olan gençlerin genel siyasi katılım düzeyleri ailelerini maddi açıdan ortanın üstü ve yüksek gelirli olarak tanımlamış olan gençlerin genel siyasi katılım düzeylerinden daha düşük olması, annesinin üniversite ya da daha yüksek bir okuldan mezun olmuş olduğunu belirten gençlerin genel siyasi katılımları annesinin ortaokul veya lise mezunu olduğunu belirten gençlerin genel siyasi katılımlarından daha yüksek olması ve okuyup çalışmayan gençlerin genel siyasi katılım düzeyleri okumayan çalışmayan gençlerin genel siyasi katılım düzeylerinden yüksek olması gibi sonuçlara ulaşılmıştır.[225]

ŞEBEKE kapsamında Emre Gür ve Devin Bahçeci'nin kaleme aldığı "Karşılaştırmalı Bir Analiz: Avrupa'da Gençlerin Katılımı Ve Gençlik Politikası" adlı bu son araştırmada, daha önceki dört çalışmadan elde edilen bulgu ve öneriler, tercüme edilen kitaplarda yer alan Avrupa coğrafyasındaki gençlik politikaları ve çalışmalarına yönelik sonuçlar ile somut ülke olayları ve deneyimlerine dayandırılarak oluşturulmuştur.[226] Avrupa Konseyi Avrupa'da gençlik politikaları hakkında tartışmaların yoğunlaştığı kurumlardan biridir. Türkiye'nin de üye olduğu Avrupa Konseyi tarafından gençlik alanında yapılan çalışmalar, Avrupa'daki ulusal devletlerin politikalarına ve uygulamalarına yol göstermektedir. Avrupa Konseyi "insan hakları, hukukun üstünlüğü ve çoğulcu demokrasi ilkelerini korumak ve güçlendirmek; azınlıklar, ırkçılık, hoşgörüsüzlük ve yabancı düşmanlığı, sosyal dışlanma, uyuşturucu madde ve çevre konularındaki sorunlara çözüm aramak; Avrupa kültürel benliğinin oluşmasına ve gelişmesine katkıda bulunmak" amacı ile kurulmuş bir işbirliği ve

[225] Yılmaz, s.64.
[226] Emre Gür ve Devin Bahçeci, "Karşılaştırmalı Bir Analiz: Avrupa'da Gençlerin Katılımı Ve Gençlik Politikası", **Şebeke: Gençlerin Katılımı Projesi,** İstanbul, 2014, s.11.

dayanışma örgütüdür.[227] Karşılaştırmalı olarak yapılan bu çalışmada, gençlerin katılımına yönelik örnek olaylara yer verilmiştir. Bunlar arasından siyasi katılım deneyimi, sivil katılım deneyimi ve sosyal hareketler ve katılım deneyimi üç ayrı ülke üzerinden ayrıntılı olarak tartışılmıştır. Bunlar, Birleşik Krallık ile ilgili olarak siyasi katılım deneyimi, Almanya ile ilgili olarak sivil katılım deneyimi ve Finlandiya ile ilgili sosyal hareketler ve katılım deneyimidir. Ayrıca gençlik politikaları açısından Türkiye'ye ışık tutabilecek olan üç ülke ile deneyimi incelenmiştir. Bunlardan Norveç deneyimi, ülkede gençlik teorileri, uygulamaları ve politikalarının nasıl uyumlulaştırıldığına dair örnekler sunarken Çek Cumhuriyeti ve Slokvakya'nın anlatıldığı deneyimde ise AB (Avrupa Birliği) gençlik politikalarının benzer geçmişlere sahip ve her ikisi de 2004 yılında AB üyesi olan iki ülkenin politikalarına nasıl yansıtıldığı irdelenmektedir.[228] Sonuç olarak bir yanda BM, UNICEF gibi kurumların geliştirmiş olduğu politikaların benimsenmesi, diğer yanda Avrupa Birliği ve Avrupa Konseyi gibi kurumların işbirlikleri ve gençlerin katılımı alanında atmış oldukları adımlar, bir diğer yanda da AB'ye üye ülkelerde ve kendi dinamikleri çerçevesinde nasıl yankı bulduğununa yönelik incelemeler, Türkiye gençliği açısından çeşitli modellerin başarısı üzerine düşünüp tartışma imkanı sağlamıştır.[229]

Bu konuda gerek Türkiye gençliği gerekse gençliğin siyasi görüşlerine yönelik literarürde yer alan bazı çalışmalar ve özellikle bu alanda yapılmış projeler incelendikten sonra, Türkiye'de gençlerin siyasi görüşlerinin şekillenmesinde hangi göstergelerin önemli bir yer tuttuğuna yönelik bir araştırma ile Türkiye gençliğinin siyasi profili belirlenmeye çalışılmıştır.

3.2. Genç Nüfus Siyasi Eğilim Araştırması

Gençler, bulundukları gelişimsel aşama itibari ile kimlik oluşumunun gerçekleştiği dönemlerde bulunmaktadır. Kimlik oluşumu gençlerin, bulundukları gelişimsel sürecin en önemli unsurundan biridir. Gençlik ekonomik, sosyal, siyasal

[227] Gür, s.30.
[228] Gür, s.11.
[229] Gür, s.85.

değişim ve dönüşümlerden çok hızlı bir şekilde etkilenebilmektedir.[230] Hem kendi benliği hem de çevresel faktörlerin etkisi altında kalarak siyasi bir eğilim kazanan gençliğe yönelik yapılan, bu araştırmaya ait bilgiler devam eden kısımlarda anlatılmıştır.

3.2.1. Araştırmanın Amacı, Kapsamı ve Önemi

Çalışmanın amacı, Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde CART ve MARS yöntemlerini karşılaştırılıp, uygulamada hangi yöntemin diğerinden daha doğru bir sınıflama yapacağını farklı büyüklükteki başlangıç ve test verisi kullanarak incelemek ve sonrasında en uygun olan başlangıç ve test verisi büyüklüğüne göre, farklı büyüklükteki örnek sayıları ile bu kez sadece CART ile modelleme yaparak ve en başarılı sınıflama modelini oluşturmaya çalışmaktır.

Bu çalışma kapsamında gençlerin önümüzdeki seçimlerde hangi partiye oy vereceklerine yönelik tercihleri bağımlı değişken olarak ele alınmıştır. Her iki yöntem için uygulanan analizlerle sınıflama modelleri oluşturulması hedeflenmiştir. MARS yöntemi ikili bağımlı değişkenler için tasarlanmış bir model olduğu için ve MARS ile elde edilen analiz sonuçlarının CART ile kıyaslanabilmesi için bağımlı değişkenin kategori düzeyi ilk aşamada ikiye indirgenmiştir ve en çok tercih edilen ilk iki parti olan "A Partisi" ve "B Partisi" ele alınmıştır. Böylelikle iki yöntemin sonuçları karşılaştırılmıştır. İkinci aşamada ise en uygun olarak seçilen başlangıç ve test verisi büyüklüğü ile bu kez genel bir değerlendirme yapılmıştır. Bu değerlendirmede bağımlı değişken olan parti tercihi "A Partisi", "B Partisi", "C Partisi" "Diğer" ve "Kararsız" olmak üzere beş kategori düzeyine göre bu kez farklı büyüklükteki örnek sayıları ile ele alınarak, CART ile modellenmiş ve çeşitli örnek büyüklüğüne göre en başarılı sonuç veren sınıflama modeli belirlenmiştir.

Ayrıca uygulamada genellikle insan genetiği, gıda bilimi ve hastalık araştırmaları gibi çeşitli alanlarda kullanılan CART ve MARS yöntemleri bu alanlardan

[230] Siyaset, Ekonomi ve Toplum Araştırmaları Vakfı (SETA), **Türkiye'nin Gençlik Profili,** 1. Baskı, Ankara: Pelin Ofset, 2012, s.13.

farklı olarak Türkiye gençliği üzerine yapılmış bir anketten toplanan veriler ile değerlendirilmiştir.

Çalışmanın kapsamı doğrultusunda, KONDA Araştırma ve Danışmanlık şirketi tarafından 9-10 Nisan 2011 saha tarihinde *"Türkiye Gençliği Araştırması"* adlı bir anket çalışmasından faydalanılmıştır. Bu anket kapsamında Türkiye'de 15 yaş üstü, 30 yaş altı genç nüfusun, saha çalışmasının yapıldığı günlerdeki algıları, beklentileri, değerleri, tercihleri, hayat tarzları, bilgi edinme mecraları ve profillerini yansıtan ve gençlerin kendi ve ailelerinin eğitim durumu gibi demografik özelliklerini içeren sorulara yer verilmiştir. Bu sorular ile gençlerin hayat ve siyasi görüşlerinin nasıl şekillendiği ve dolayısıyla bu temel göstergelerden yola çıkarak bireylerin hayat ve siyasi görüşlerine göre sınıflandırılmasının mümkün olacağı düşünülmektedir. Aslında bu da doğrudan genç nüfusta siyasi seçimin ne şekilde ortaya çıkacağının bir göstergesi olabilmektedir. Bu durum, genç nüfusun toplumun anlamlı bir parçası olarak ele alınmasının ve gençlikle ilgili etkin politikaların oluşturulmasının önemini göstermektedir. Öyle ki nüfus projeksiyonlarına göre, 2025 yılında Türkiye'deki genç nüfus oranının, ABD ve Avrupa ülkelerinden yüksek olacağı ve Türkiye'nin, Avrupa Birliği ülkelerine kıyasla oldukça genç bir nüfusa sahip olduğu göz önüne alındığında; etkili gençlik politikalarının Türkiye açısından ne derece önemli olduğu anlaşılmaktadır.

3.2.2. Anakütle ve Örneklem

Çalışmada kullanılan veriler, KONDA Araştırma ve Danışmanlık şirketi tarafından 9-10 Nisan 2011 saha tarihinde *"Türkiye Gençliği Araştırması"* adlı bir anket çalışması ile toplanmıştır. Buna göre ,yerleşim yerleri önce kır/kent/metropol olarak ayrıştırılmış ve 12 bölge esas alınarak örneklem tespit edilmiştir. Böylelikle, araştırma 35 ilin, 134 ilçesine bağlı 202 mahalle ve köyünde 15-30 yaş arasındaki 2366 kişiyle hanelerinde yüzyüze görüşülerek gerçekleştirilmiştir. Görüşülen kişi sayısına her mahallede 12'şer kişiyle görüşme yapılarak ulaşılmıştır. Bu 12 görüşme için önce yaş filtresi uygulanmış ve görüşmelerde yaş ve cinsiyet kotası uygulanmıştır. Yani yaş grupları 15-20, 21-25 ve 26-30 olacak şekilde her yaş grubundan eşit sayıda kadın ve erkek ile görüşülmüştür.

Araştırmanın kapsamı doğrultusunda Türkiye İstatistik Kurumu (TÜİK) bilgilerine göre 2011 yılında Türkiye nüfusu 73,7 milyon ve 15-30 yaş arasındaki nüfus 17,3 milyon olup, anakütle verisini temsil etmektedir. Ayrıca 2011 Aralık ayı itibariyle, TÜİK Adrese Dayalı Nüfus Kayıt Sistemi (ADNKS) verilerine göre, Türkiye'de toplam nüfusun yarısı 29,7 yaştadır. Yani Türkiye'nin oratanca yaşı 29,7'dir. Ortanca yaş erkeklerde 29,1 iken; kadınlarda 30,3'tür. Bu anakütle verisinden yola çıkarak anketin yapıldığı saha alanı, İstatistiki Bölge Birimleri Sınıflandırması (İBBS), Düzey 1 yani 12 Bölge Birimini kapsamaktadır. Gidilen iller aşağıdaki tabloda gösterilmiştir.

Tablo 3.2. Örneklemin Saha Alanı

Düzey 1 (12 Bölge)	Gidilen İller
İstanbul	İstanbul
Batı Marmara	Tekirdağ, Balıkesir, Çanakkale
Ege	İzmir, Manisa, Denizli
Doğu Marmara	Bursa, Eskişehir, Kocaeli, Düzce
Batı Anadolu	Ankara, Konya
Akdeniz	Antalya, Adana, Mersin, Hatay, Kahramanmaraş
Orta Anadolu	Kayseri, Nevşehir, Sivas
Batı Karadeniz	Samsun, Karabük
Doğu Karadeniz	Trabzon, Rize, Ordu
Kuzeydoğu Anadolu	Erzurum, Kars
Ortadoğu Anadolu	Malatya, Van, Tunceli
Güneydoğu Anadolu	Gaziantep, Şanlıurfa, Diyarbakır, Mardin

Kaynak: KONDA Araştırma ve Danışmanlık, Türkiye Gençliği Araştırması'2011

Çalışmada kullanılan örnekleme tekniği ise, Tesadüfi Örnekleme Tekniklerinden olan Tabakalı Örnekleme tekniğidir. Tabakalı örnekleme, incelenen öznitelik açısından heterojen yapıdaki anakütlenin homojen alt gruplara (tabakalara, zümrelere) ayrıldığı ve bu tabakaların her birinden belirlenen sayılarda ve tesadüfi usülle birimlerin seçildiği örneklemedir.[231] Tabakalı örneklemede, her tabakadan tabaka hacmi ile orantılı (değişken oranlı) veya her tabakadan sabit oranda (n/N) ya da tabakaların değişkenliği ile orantılı birimler tesadüfi olarak seçilerek örneklem

[231] İ.Esen Yıldırım, **Kamuoyu Araştırmaları ve Su Tüketim Bilinci Üzerine Bir Uygulama,** Ankara: Seçkin Yayıncılık, 2010, s.73.

oluşturulur.[232] Aşağıdaki tabloda bulunan oranların incelenmesi neticesinde, hemen hemen her tabaka hacmi ile orantılı örneklemin seçildiği görülmektedir. Bu da kullanılan örnekleme tekniğinin; tabaka hacmi ile orantılı tabakalı örnekleme yöntemi olduğunu göstermektedir.

Tablo 3.3. Örneklem Bilgisi

Araştırma Saha Uygulaması	Araştırma	TÜİK Tüm Nüfus	TÜİK 15-30 Yaş Nüfus
İstanbul	19,0	18,0	20,1
Batı Marmara	5,1	4,3	4,3
Ege	15,5	13,1	13,8
Doğu Marmara	10,2	9,3	7,2
Batı Anadolu	9,3	9,5	10,4
Akdeniz	12,3	12,8	13,9
Orta Anadolu	5,1	5,2	5,7
Batı Karadeniz	6,7	6,1	6,3
Doğu Karadeniz	2,1	3,4	3,5
Kuzeydoğu Anadolu	2,6	3,0	3,6
Ortadoğu Anadolu	3,5	4,9	6,0
Güneydoğu Anadolu	8,7	10,3	5,2
Toplam	100	100	100

Kaynak: KONDA Araştırma ve Danışmanlık, Türkiye Gençliği Araştırması'2011

3.2.3. Araştırma Verisi

Veriler alındıkları kaynağa göre doğrudan/birincil veya dolaylı/ikincil veriler olmak üzere iki ana grupta toplanabilirler. Birincil veriler araştırmayı yapan kişi veya kurum tarafından kaynağından alınan verilerdir. İkincil veriler ise, başka kurum veya kuruluşlar tarafından toplanarak düzenlenen verilerdir.[233] Buna göre araştırma kapsamında ele alınan veriler KONDA Araştırma ve Danışmanlık şirketi tarafından Nisan 2011 tarihinde *"Türkiye Gençliği Araştırması"* adlı bir anket çalışmasından elde edilmiş olup, ikincil veri özelliği taşımaktadır. Anketin uygulandığı katılımcıların sayısı 2366 kişidir. Bu anketlerden veri girişinin yapıldığı anket sayısı ise 2335 gözlemi kapsamaktadır.

[232] Şahamet Bülbül, **Tanımlayıcı İstatistik,** İstanbul: DER Yayınevi, 2013, s.32.
[233] Ahmet Mete Çilingirtürk, **İstatistiksel Karar Almada Veri Analizi,** Ankara: Seçkin Yayıncılık, 2011, s.29.

Bu anket kapsamında Türkiye'de 15 yaş üstü, 30 yaş altı genç nüfusun, saha çalışmasının yapıldığı günlerdeki (9-10 Nisan 2011) algıları, beklentileri, değerleri, tercihleri, hayat tarzları, bilgi edinme mecraları ve profillerini yansıtan ve gençlerin kendi ve ailelerinin eğitim durumu gibi demografik özelliklerini içeren sorulara yer verilmiştir. Bu sorular bağımsız değişken olarak incelenmiştir. Anket kapsamında gençlerin önümüzdeki seçimlerde hangi partiye oy vereceklerine yönelik tercihlerini içeren soru ise, bağımlı değişken olarak ele alınmıştır. Bu soruya yaşı tutmadığı için cevap vermeyen 468 katılımcı çalışmadan çıkarıldığı için öncelikle 1867 anket verisi ele alınmıştır. Daha sonra çalışmada oy oranı en yüksek çıkan iki büyük parti ele alındığında ise 1018 anket verisi ile çalışımıştır. Buna göre ankette yer alan sorular aşağıdaki tabloda genel olarak soruluş amacına göre özetlenmiştir.

Tablo 3.4. Katılımcılara Uygulanan Anket Soruları

Katılımcıların Profiline Ait Sorular	Cinsiyet	Yaş	Eğitim durumu	Baba eğitim durumu
	Doğum yeri (bölge)	Baba doğum yeri (bölge)	Medeni durumu	Çalışma durumu
	Babanız ne iş yapar (dı)?	Aylık geliri	Aylik aile geliri	Gelirin kaynağı
	Kendi dindarlığı	Aile reisinin dindarlığı	Oy verceği parti	Yerleşim Yeri
	Nerede öğrenci?	Oturulan evin tipi	Ne kadar zamandır bu yerde yaşıyorsunuz?	
Katılımcıların Beklentisine Ait Sorular	Benim hayat şartlarım 5 yıl sonra daha iyi olacak		Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak	
Katılımcıların Değerlerine Ait Sorular	Gündelik hayatımda toplumun tüm kurallarına harfiyen uyarım.		Geçmişten gelen geleneklerimiz değişmeden korunmalıdır.	
	Zengin kıza fakir oğlan, fakir kıza zengin oğlan olmaz. Hayatta davul bile dengi dengine çalar.		Hayatımın gidişatını değiştirmek için yapabileceğim pek bir şey yok... Böyle gelmiş böyle gider…	
	Seçim yapmak durumunda olsaydınız, hangisinin en önemli olduğunu söylerdiniz?		En çok hangi kuruma gönülden güvenirsiniz?	
Katılımcıların Fikirlerine Ait Sorular	Sizce ülkemizdeki bu haliyle üniversite eğitimi en çok ne sağlıyor?		Hayata hazırlanırken en çok şeyi nerden öğrencdiniz?	
	Gitme imkânı olsa bile yine Türkiye'de yaşamayı tercih ederdim.		Türkiye Orta Doğu ve Müslüman ülkelerle daha yakın işbirliği içinde olmalıdır.	
	Türkiye Avrupa Birliğine mutlaka üye olmalıdır.			
Katılımcıların Başarı Tanımına Ait Sorular	Çok bilgili olabilmek için en gerekli olan hangisidir?		Hayallerini gerçekleştirebilmek için en gerekli olan hangisidir?	
	İş hayatında çok başarılı olabilmek için en gerekli olan hangisidir?		Güç elde etmek için en gerekli olan hangisidir?	
	Toplumda statü kazanmak için en gerekli olan hangisidir?		Başarı kelimesi sizin için ne ifade ediyor?	
	Hangisi elinizde olursa kendinizi mutlu sayarsınız?		İsteyerek, mutlu olarak çalışacağınız iş konusunda maaşı dışında aşağıdakilerden hangisi sizin için en önemli unsurdur?	

Katılımcıların Ahlaki Değerlerine Ait Sorular	İdolü (mesleğe göre)	Evleneceğiniz insanda aşağıda okuyacaklarımdan hangi özelliğin mutlaka olmasını istersiniz?
	Bir kadın ile erkek aşağıdakilerden hangisi olur ise sizce beraber yaşayabilirler?	Öğretmenin dövdüğü yerde gül biter.
	Kızını dövmeyen dizini döver.	Maceracı sıfatı kime yakışıyor?
	Rekabetçi sıfatı kime yakışıyor?	Uzlaşmacı sıfatı kime yakışıyor?
	Sevecenlik sıfatı kime yakışıyor?	Özgürlük düşkünü sıfatı kime yakışıyor?
Katılımcıların Aile ile Çatışma Durumuna Ait Sorular	Giyim kuşam konusunda aileniz ile hangi sıklıkta çatışma yaşarsınız?	Duygusal arkadaş edinmek konusunda ailenizle hangi sıklıkta çatışma yaşarsınız?
	Gece dışarıya çıkmak konusunda ailenizle hangi sıklıkta çatışma yaşarsınız?	Dini kurallara riayet etme konusunda ailenizle hangi sıklıkta çatışma yaşarsınız?
	Bilgisayar/İnternet kullanma konusunda ailenizle hangi sıklıkta çatışma yaşarsınız?	
Katılımcıların Tüketim Tercihine Ait Sorular	Reklam / promosyon önemli midir? (Ürün satın alırken)	Marka önemli midir? (Ürün satın alırken)
	Moda / trend önemli midir? (Ürün satın alırken)	Ailenin alışkanlıkları önemli midir? (Ürün satın alırken)
	Moda / trend önemli midir? (Ürün satın alırken)	Ailenin alışkanlıkları önemli midir? (Ürün satın alırken)
Katılımcıların Hayat Tarzlarına Ait Sorular	Bilgisayarınız var mı?	İnternete nereden giriyorsunuz?
	İnternete ne sıklıkta giriyorsunuz?	İnternet sosyal paylaşım ağlarına üyelik
	İnterneti en çok hangi iki eylem için kullanıyorsunuz?	En çok izlediğiniz TV kanalı
	En çok izlediğiniz TV programı	TV izleme sıklığı
	Haberleri nereden takip ediyorsunuz?	Okuduğu gazete
	Dernek – vakıf - sosyal kulüp üyeliği	En çok hangisi ilginizi, merakınızı çeker?

Kaynak: KONDA Araştırma ve Danışmanlık, Türkiye Gençliği Araştırması'2011

3.2.4. Araştırma Metodolojisi

Araştırmada verilerin analizi için karar ağaçlarında kullanılan sınıflama yöntemlerinden olan CART ve MARS yöntemlerinden faydalanılmıştır. CART hem kategorik hem de sürekli değişkenleri kullanarak sınıflama ve regresyon problemlerinin çözümünde karar ağaçlarını kullanan parametrik olmayan istatistiksel bir metottur. Ele alınan bağımlı değişken kategorik ise yöntem sınıflama ağaçları, sürekli ise regresyon ağaçları olarak adlandırılmaktadır. Aynı durum MARS yöntemi için de geçerlidir. MARS yöntemi, hem sürekli hem de ikili bağımlı değişkenler için tasarlanmıştır. Bağımlı değişkeninin sürekli olması durumunda kestirim amaçlı olan bu yöntem, bağımlı değişkeninin kategorik olması durumunda ise, sınıflandırma amacına sahiptir.

Çalışma kapsamında gençlerin önümüzdeki seçimlerde hangi partiye oy vereceklerine yönelik tercihleri bağımlı değişken olarak ele alınmıştır. Böylelikle CART ve MARS yöntemleri ile sınıflama modelleri oluşturulmuştur. Her iki yöntem için ilgili analizler SPM 7.0 (Salford Predictive Modeller) veri madenciliği programı ile yapılmıştır. Çalışmada gençlerin demografik özelliklerine ait frekanslar ise, IBM SPSS Paket programı 21 sürümü ile hesaplanmıştır.

3.3. Analiz ve Bulgular

Bu çalışmada Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde CART ve MARS yöntemleri karşılaştırılıp uygulamada hangi yöntemin diğerinden daha doğru bir sınıflama yapacağı farklı büyüklükteki başlangıç ve test verisi kullanılarak incelenmiştir. Sonrasında en uygun olan başlangıç ve test verisi büyüklüğüne göre bu kez sadece CART ile modelleme yapılmış ve farklı büyüklükteki örnek sayıları ile en başarılı sınıflama modeli oluşturulmaya çalışılmıştır.

Çalışma kapsamında gençlerin önümüzdeki seçimlerde hangi partiye oy vereceklerine yönelik tercihleri bağımlı değişken olarak ele alınmıştır. Her iki yöntem için uygulanan analizlerde sınıflama modeli oluştulması hedeflenmiştir. MARS yöntemi ikili bağımlı değişkenler için tasarlanmış bir model olduğu için ve MARS ile elde edilen analiz sonuçlarının CART ile kıyaslanabilmesi için ilk aşamada bağımlı değişkenin

kategori düzeyi ikiye indirgenmiş ve en çok tercih edilen ilk iki parti olan "A Partisi" ve "B Partisi" ele alınmıştır. Böylelikle iki yöntemin sonuçları karşılaştırılmıştır.

Bu amaçla ilk aşamada başlangıçta modeli oluştumak için veri setinin sırasıyla %70'i, %50'si ve %30'u başlangıç veri seti bir diğer adıyla eğitim verisi olarak ele alınmıştır. Buna göre veri setinin sırasıyla geri kalan %30, %50 ve %70'lik kısmı ise modeli test etmek için kullanılmıştır. Bu durumda 3 adet CART ve 3 adet MARS modeli elde edilmiştir.

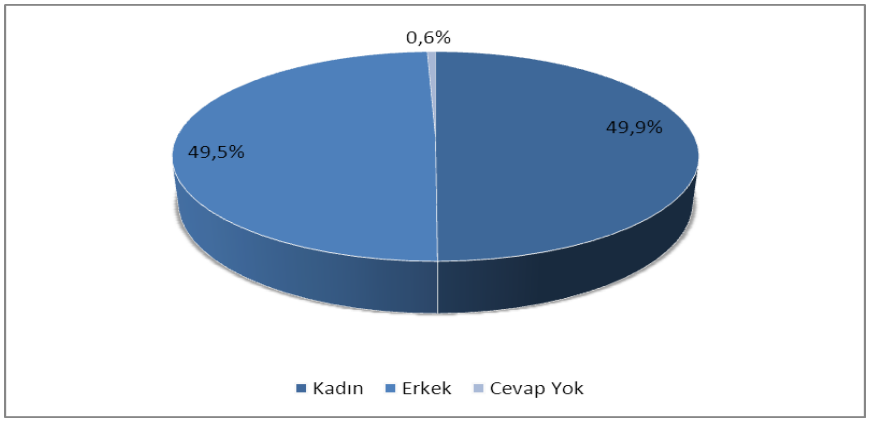
İkinci aşamada ise en uygun olan başlangıç ve test verisi büyüklüğü ile bu kez genel bir değerlendirme yapılmıştır. Bu değerlendirme de bağımlı değişken olan parti tercihi "A Partisi", "B Partisi", "C Partisi" "Diğer" ve "Kararsız" olmak üzere beş kategori düzeyi esas alınarak farklı büyüklükteki örnek sayılarına göre CART yöntemi ile modellenmiş ve çeşitli örnek büyüklükleri içerisinden en başarılı sınıflama modeli belirlenmiştir. Buna göre yapılan analizlere ait bulgular sırasıyla aşağıda verilmiştir.

3.3.1. Katılımcıların Sosyo Demografik Özelliklerine İlişkin Bulgular

Buna göre katılımcılara ait sosyodemografik özelliklerine ilişkin bulgular izleyen tablo ve grafiklerde verilmiştir.

Tablo 3.5. Katılımcıların Cinsiyet Dağılımı

Cinsiyet	Frekans	Yüzde (%)
Kadın	508	49,9
Erkek	504	49,5
Cevap Yok	6	0,6
Toplam	1018	100

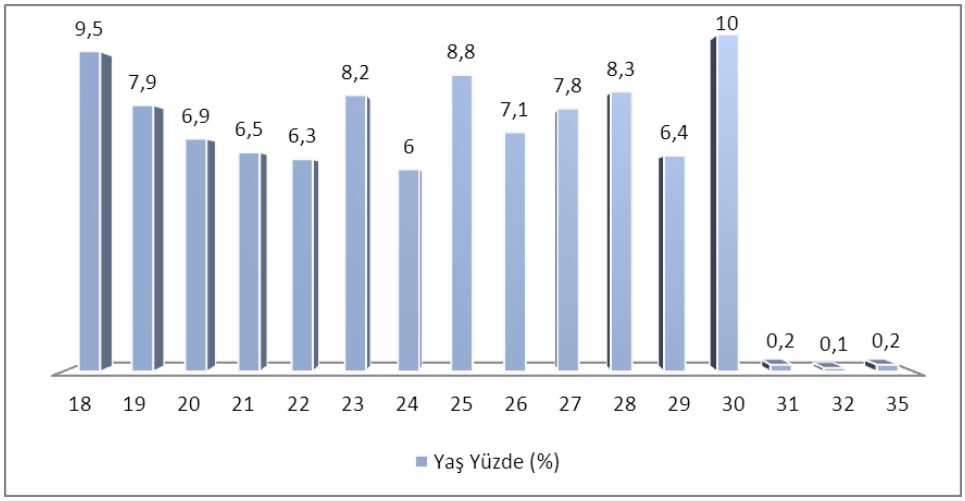


Şekil 3.1: Katılımcıların Cinsiyet Dağılımı

Katılımcıların cinsiyet dağılımlarının hemen hemen eşit çıktığı Tablo 3.5 ve grafik 3.1''de görülmektedir.

Tablo 3.6. Katılımcıların Yaş Dağılımı

Yaş	Frekans	Yüzde (%)
18	97	9,5
19	80	7,9
20	70	6,9
21	66	6,5
22	64	6,3
23	83	8,2
24	61	6
25	90	8,8
26	72	7,1
27	79	7,8
28	84	8,3
29	65	6,4
30	102	10
31	2	0,2
32	1	0,1
35	2	0,2
Toplam	1018	100

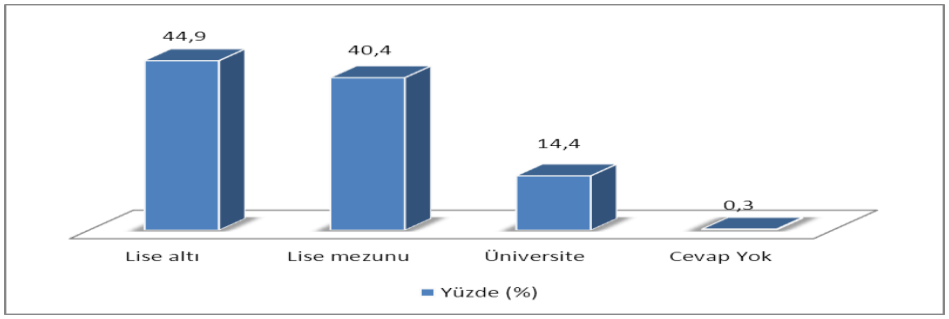


Şekil 3.2: Katılımcıların Yaş Dağılımı

Katılımcıların büyük çoğunlu 18 ve 30 yaşlarında olup, 31-35 yaş arası sadece 5 katılımcı bulunmaktadır.

Tablo 3.7. Katılımcıların Eğitim Durumuna Göre Dağılımı

Eğitim Durumu	Frekans	Yüzde (%)
Lise altı	457	44,9
Lise mezunu	411	40,4
Üniversite	147	14,4
Cevap Yok	3	0,3
Toplam	1018	100

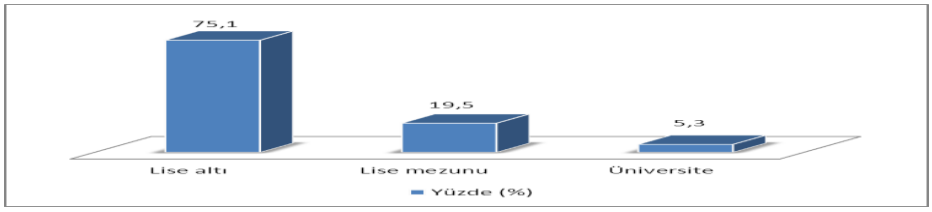


Şekil 3.3: Katılımcıların Eğitim Durumuna Göre Dağılımı

Katılımcıların %44,9'u lise altı, %40,4'ü lise mezunu, %14,4'ü ise üniversite mezunudur.

Tablo 3.8. Katılımcıların Baba Eğitim Durumuna Göre Dağılımı

Baba Eğitim Durumu	Frekans	Yüzde (%)
Lise altı	765	75,1
Lise mezunu	199	19,5
Üniversite	54	5,3
Toplam	1018	100

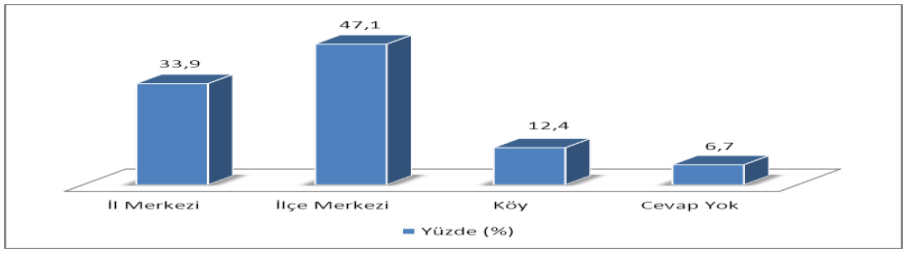


Şekil 3.4: Katılımcıların Baba Eğitim Durumuna Göre Dağılımı

Katılımcıların babalarının %75,1'i lise altı, %19,5'i lise mezunu, %5,3'ü ise üniversite mezunudur.

Tablo 3.9. Katılımcıların Doğum Yerine (İl, İlçe Merkezi, Köy) Göre Dağılımı

Doğum Yeri	Frekans	Yüzde (%)
İl Merkezi	345	33,9
İlçe Merkezi	479	47,1
Köy	126	12,4
Cevap Yok	68	6,7
Toplam	1018	100

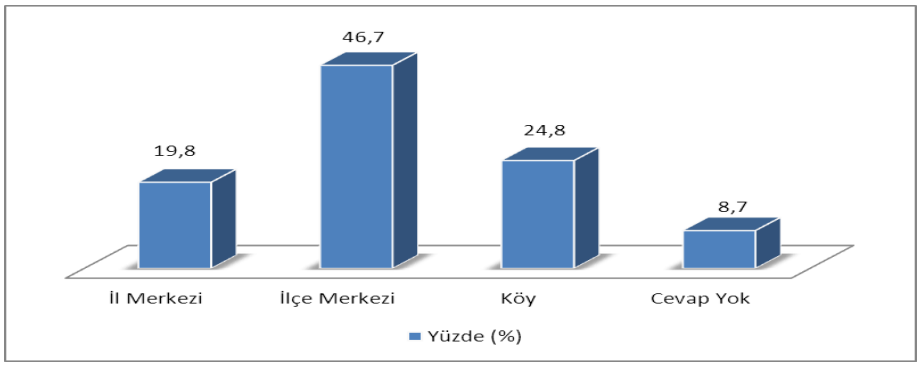


Şekil 3.5: Katılımcıların Doğum Yerine (İl, İlçe Merkezi, Köy) Göre Dağılımı

Katılımcıların %47,1'i ilçe merkezinde, %33,9'u il merkezinde, %12,4'ü ise köyde doğmuş olup, katılımcıların %6,7'si bu soruya cevap vermemiştir.

Tablo 3.10. Katılımcıların Baba Doğum Yerine (İl, İlçe Merkezi, Köy) Göre Dağılımı

Baba Doğum Yeri	Frekans	Yüzde (%)
İl Merkezi	202	19,8
İlçe Merkezi	475	46,7
Köy	252	24,8
Cevap Yok	89	8,7
Toplam	1018	100

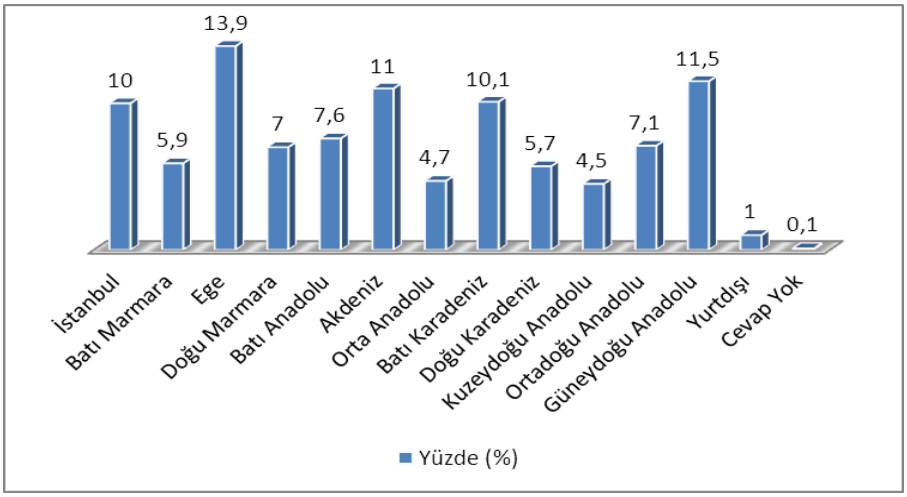


Şekil 3.6: Katılımcıların Baba Doğum Yerine (İl, İlçe Merkezi, Köy) Göre Dağılımı

Katılımcıların babalarının %46,7'si ilçe merkezinde, %24,8'i köyde, %19,8'i ise il merkezinde doğmuş olup, katılımcıların %8,7'si bu soruya cevap vermemiştir.

Tablo 3.11. Katılımcıların Doğum Yerine (Bölge) Göre Dağılımı

Doğum Yeri (Bölge)	Frekans	Yüzde (%)
İstanbul	102	10
Batı Marmara	60	5,9
Ege	141	13,9
Doğu Marmara	71	7
Batı Anadolu	77	7,6
Akdeniz	112	11
Orta Anadolu	48	4,7
Batı Karadeniz	103	10,1
Doğu Karadeniz	58	5,7
Kuzeydoğu Anadolu	46	4,5
Ortadoğu Anadolu	72	7,1
Güneydoğu Anadolu	117	11,5
Yurtdışı	10	1
Cevap Yok	1	0,1
Toplam	1018	100



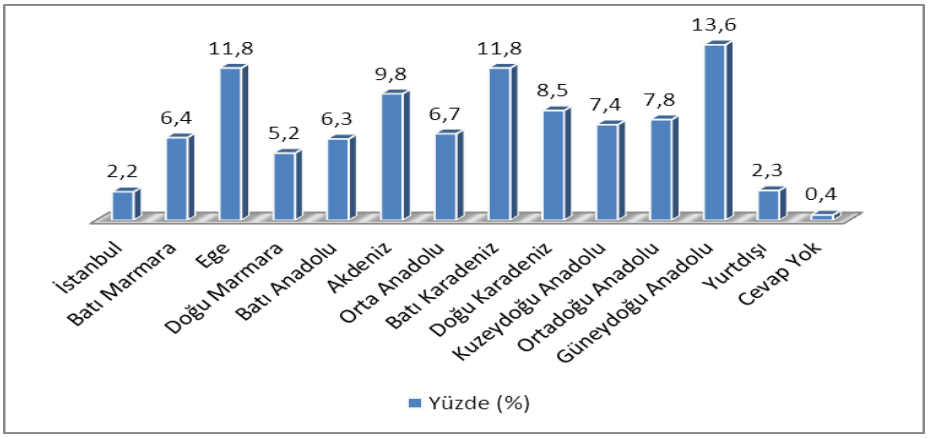
Şekil 3.7: Katılımcıların Doğum Yerine (Bölge) Göre Dağılımı

Katılımcıların %13,9'u Ege, %11,5'i Güneydoğu Anadolu, %11'i Akdeniz, %10,1'i Batı Karadeniz, %10'u İstanbul, %7,6'sı Batı Anadolu, %7,1'i Ortadoğu

Anadolu, %7'si Marmara, %5,9'u Batı Marmara, %5,7'si Doğu Karadeniz, % 4,7'si Orta Anadolu, %4,5'i Kuzeydoğu Anadolu ve %1'i yurtdışı doğumludur.

Tablo 3.12. Katılımcıların Baba Doğum Yerine (Bölge) Göre Dağılımı

Baba Doğum Yeri (Bölge)	Frekans	Yüzde (%)
İstanbul	22	2,2
Batı Marmara	65	6,4
Ege	120	11,8
Doğu Marmara	53	5,2
Batı Anadolu	64	6,3
Akdeniz	100	9,8
Orta Anadolu	68	6,7
Batı Karadeniz	120	11,8
Doğu Karadeniz	87	8,5
Kuzeydoğu Anadolu	75	7,4
Ortadoğu Anadolu	79	7,8
Güneydoğu Anadolu	138	13,6
Yurtdışı	23	2,3
Cevap Yok	4	0,4
Toplam	1018	100



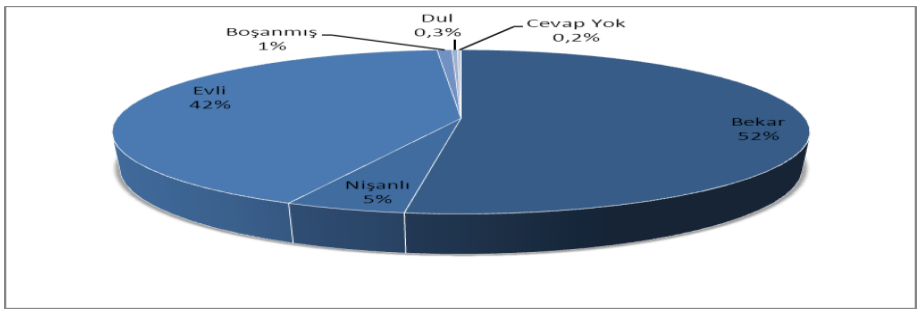
Şekil 3.8: Katılımcıların Baba Doğum Yerine (Bölge) Göre Dağılımı

Katılımcıların babalarının %13,6'sı Güneydoğu Anadolu, %11,8'i Ege ve Batı Karadeniz, %9,8'i Akdeniz, %8,5'i Doğu Karadeniz, %7,8'i Ortadoğu Anadolu, %7,4'ü

Kuzeydoğu Anadolu, %6,7'si Orta Anadolu, %6,4'ü Batı Marmara, %6,3'ü Batı Anadolu, %5,2'si Doğu Marmara, %2,3'ü Yurtdışı ve %2,2'si İstanbul doğumludur.

Tablo 3.13. Katılımcıların Medeni Durumuna Göre Dağılımı

Medeni Durum	Frekans	Yüzde (%)
Bekâr	531	52,2
Nişanlı	49	4,8
Evli	425	41,7
Boşanmış	8	0,8
Dul	3	0,3
Cevap Yok	2	0,2
Toplam	1018	100

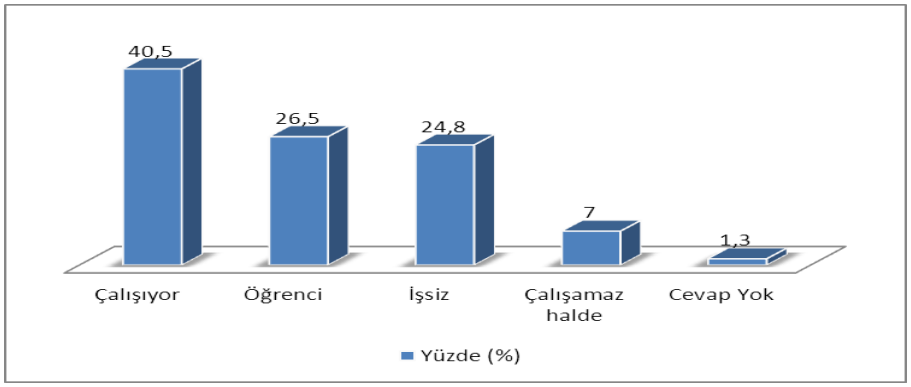


Şekil 3.9: Katılımcıların Medeni Durumuna Göre Dağılımı

Katılımcıların %52,2'si bekâr, %41,7'si evli, %4,8'i nişanlı, %0,8'i boşanmış, 0,3'ü dul olup, katılımcıların %0,2'si ise bu soruya cevap vermemiştir.

Tablo 3.14. Katılımcıların Çalışma Durumuna Göre Dağılımı

Çalışma Durumu	Frekans	Yüzde (%)
Çalışıyor	412	40,5
Öğrenci	270	26,5
İşsiz	252	24,8
Çalışamaz halde	71	7
Cevap Yok	13	1,3
Toplam	1018	100

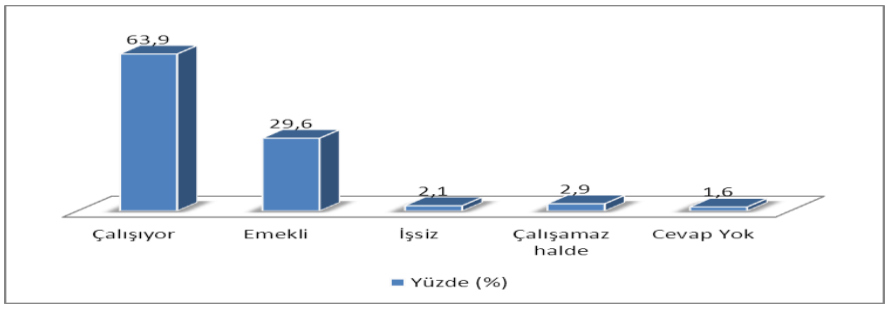


Şekil 3.10: Katılımcıların Çalışma Durumuna Göre Dağılımı

Katılımcıların %40,5'i çalışıyor, %26,5'i öğrenci, %24,8'i işsiz, %7'si çalışamaz halde ve katılımcıların %1,3'ü ise bu soruya cevap vermemiştir.

Tablo 3.15. Katılımcıların Baba Çalışma Durumuna Göre Dağılımı

Baba Çalışma Durumu	Frekans	Yüzde (%)
Çalışıyor	650	63,9
Emekli	301	29,6
İşsiz	21	2,1
Çalışamaz halde	30	2,9
Cevap Yok	16	1,6
Toplam	1018	100

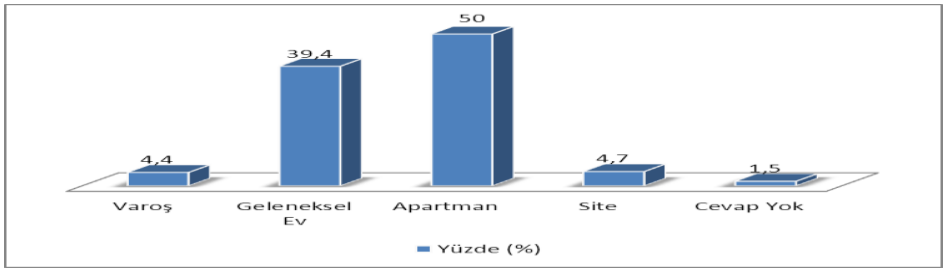


Şekil 3.11: Katılımcıların Baba Çalışma Durumuna Göre Dağılımı

Katılımcıların babalarının %63,9'u çalışıyor, %29,6'sı emekli, %2,9'u çalışamaz halde, %2,1'i işsiz ve katılımcıların %1,6'sı ise bu soruya cevap vermemiştir.

Tablo 3.16. Katılımcıların Oturduğu Evin Türüne Göre Dağılımı

Oturulan Evin Türü	Frekans	Yüzde (%)
Varoş	45	4,4
Geleneksel Ev	401	39,4
Apartman	509	50
Site	48	4,7
Cevap Yok	15	1,5
Toplam	1018	100

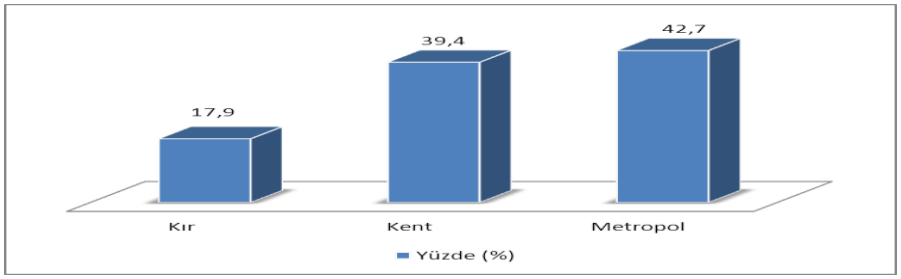


Şekil 3.12: Katılımcıların Oturduğu Evin Türüne Göre Dağılımı

Katılımcıların %50'si apartman dairesinde %39,4'ü geleneksel evde (müstakil ev), %4,7'si sitede, %4,4'ü ise, gecekondu türü varoş evde oturmaktadır. Katılımcıların %1,5'i ise bu soruya cevap vermemiştir.

Tablo 3.17. Katılımcıların Yerleşim Yeri Türüne Göre Dağılımı

Yerleşim Yeri	Frekans	Yüzde (%)
Kır	182	17,9
Kent	401	39,4
Metropol	435	42,7
Toplam	1018	100

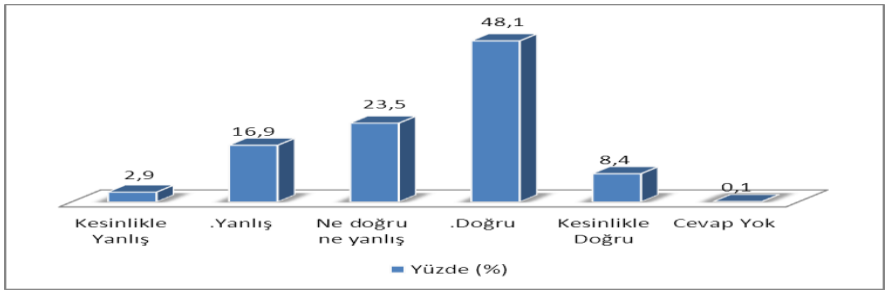


Şekil 3.13: Katılımcıların Yerleşim Yeri Türüne Göre Dağılımı

Katılımcıların %42,7'si metropolde, %39,4'ü kentte ve %17,9'u kır bölgesinde yaşamaktadır.

Tablo 3.18. Katılımcıların "Benim hayat şartlarım 5 yıl sonra daha iyi olacak" Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	30	2,9
Yanlış	172	16,9
Ne doğru ne yanlış	239	23,5
Doğru	490	48,1
Kesinlikle Doğru	86	8,4
Cevap Yok	1	0,1
Toplam	1018	100

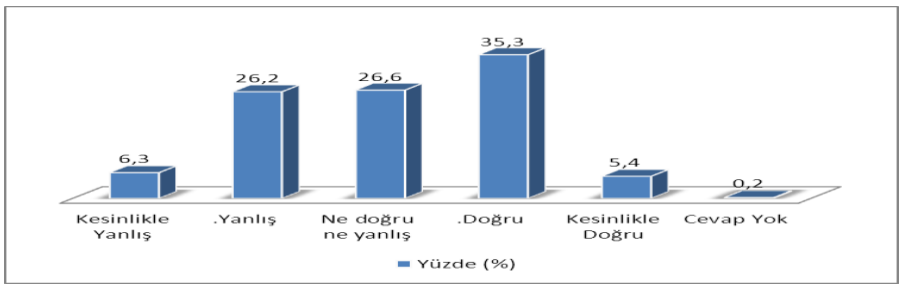


Şekil 3.14: Katılımcıların "Benim hayat şartlarım 5 yıl sonra daha iyi olacak" Görüşü

Gençlerin %48'i "hayat şartlarının 5 yıl sonra daha iyi olacak görüşünü" doğru bulmuşken, %23,5'i kararsız kalmış, %16,9'u bu görüşü yanlış bulmuş, %8,4'ü ise bu görüşü kesinlikle doğru bulurken, %2,9'u kesinlikle yanlış olarak değerlendirmiş. Gençlerin %0,1'i ise bu soruya cevap vermemiştir.

Tablo 3.19. Katılımcıların "Türkiyedeki hayat şartları 5 yıl sonra kesinlikle daha iyi olacak" Görüşü

	Frekans	Yüzde (%)
Kesinlikle Yanlış	64	6,3
Yanlış	267	26,2
Ne doğru ne yanlış	271	26,6
Doğru	359	35,3
Kesinlikle Doğru	55	5,4
Cevap Yok	2	0,2
Toplam	1018	100

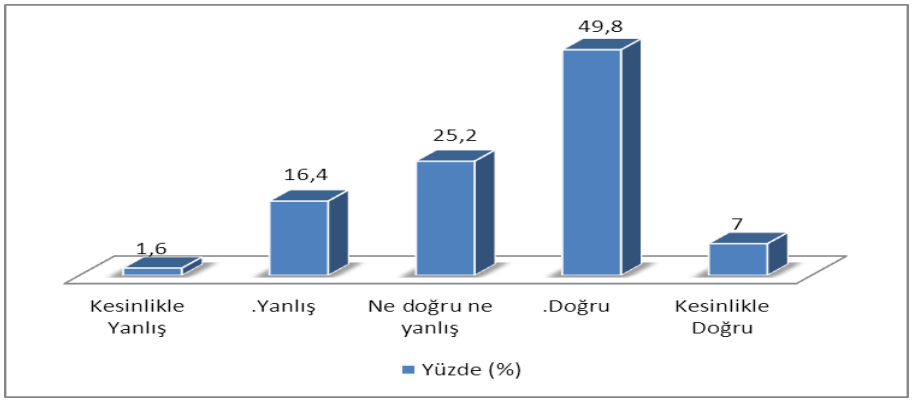


Şekil 3.15: Katılımcıların "Türkiyedeki hayat şartları 5 yıl sonra kesinlikle daha iyi olacak" Görüşü

Gençlerin %35,3'ü "Türkiyedeki hayat şartları 5 yıl sonra daha iyi olacak" görüşünü doğru bulmuşken, %26,6'sı kararsız kalmış, %26,2'si bu görüşü yanlış bulmuş, %6,3'ü ise bu görüşü kesinlikle yanlış olarak bulurken, %5,4'ü kesinlikle doğru olarak değerlendirmiş. Gençlerin %0,2'si ise bu soruya cevap vermemiştir.

Tablo 3.20. Katılımcıların "Gündelik hayatımda toplumun tüm kurallarına harfiyen uyarım" Görüşü

	Frekans	Yüzde (%)
Kesinlikle Yanlış	16	1,6
Yanlış	167	16,4
Ne doğru ne yanlış	257	25,2
Doğru	507	49,8
Kesinlikle Doğru	71	7
Toplam	1018	100

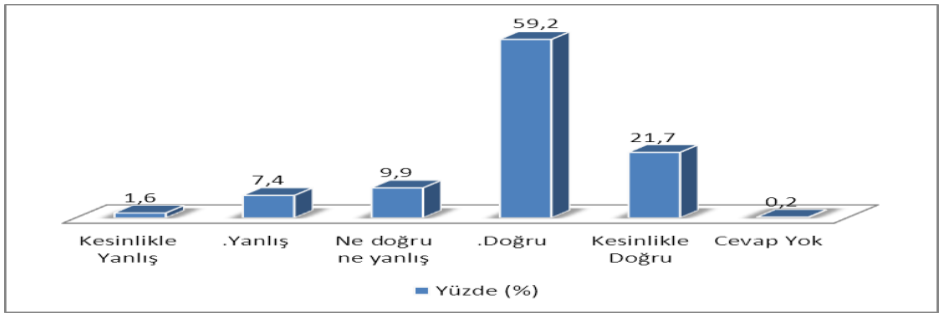


Şekil 3.16: Katılımcıların "Gündelik hayatımda toplumun tüm kurallarına harfiyen uyarım" Görüşü

Gençlerin %49,8'i "gündelik hayatımda toplumun tüm kurallarına harfiyen uyarım" görüşünü doğru bulmuşken, %25,2'si kararsız kalmış, %16,4'ü ise bu görüşü yanlış bulmuş, %7'si bu görüşü kesinlikle doğru olarak bulurken, %1,6'sı ise kesinlikle yanlış olarak değerlendirmiş.

Tablo 3.21. Katılımcıların "Geçmişten gelen geleneklerimiz değişmeden korunmalıdır" Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	16	1,6
Yanlış	75	7,4
Ne doğru ne yanlış	101	9,9
Doğru	603	59,2
Kesinlikle Doğru	221	21,7
Cevap Yok	2	0,2
Toplam	1018	100

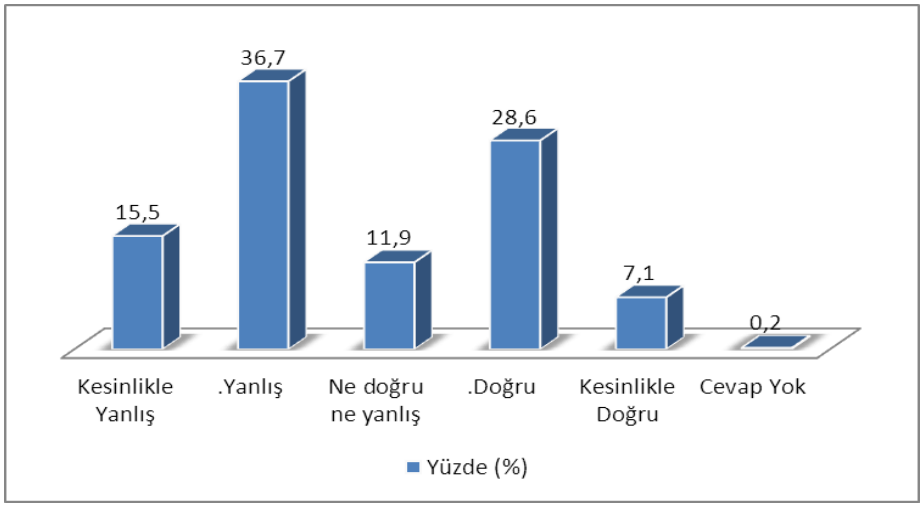


Şekil 3.17: Katılımcıların "Geçmişten gelen geleneklerimiz değişmeden korunmalıdır" Görüşü

Gençlerin %59,2'si "geçmişten gelen geleneklerimiz değişmeden korunmalıdır" görüşünü doğru bulmuşken, %21,7'si kesinlikle doğru bulmuş, %9,9'u ise kararsız kalmış, %7,4'ü bu görüşü yanlış bulurken, %1,6'sı ise kesinlikle yanlış olarak değerlendirmiş. Gençlerin %0,2'si ise bu soruya cevap vermemiştir.

Tablo 3.22. Katılımcıların "Zengin kıza fakir oğlan, fakir kıza zengin oğlan olmaz. Hayatta davul bile dengi dengine çalar." Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	158	15,5
Yanlış	374	36,7
Ne doğru ne yanlış	121	11,9
Doğru	291	28,6
Kesinlikle Doğru	72	7,1
Cevap Yok	2	0,2
Toplam	1018	100

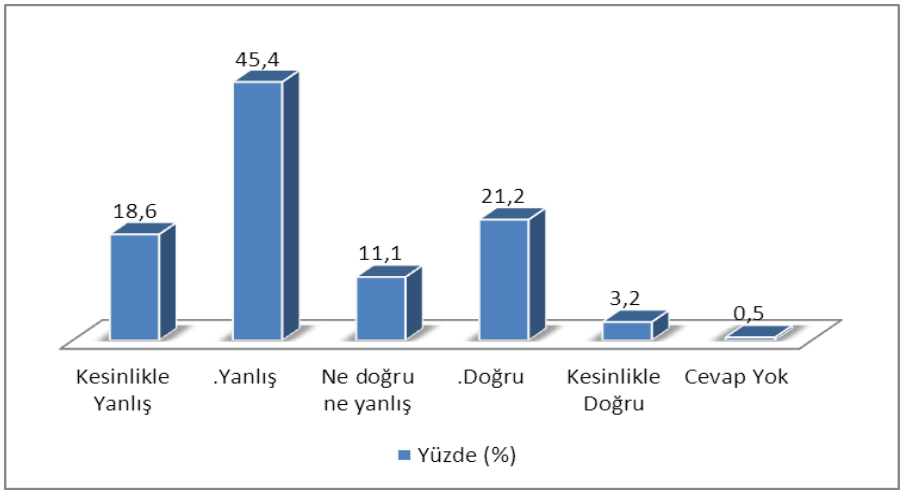


Şekil 3.18: Katılımcıların "Zengin kıza fakir oğlan, fakir kıza zengin oğlan olmaz. Hayatta davul bile dengi dengine çalar." Görüşü

Gençlerin %36,7'si "Zengin kıza fakir oğlan, fakir kıza zengin oğlan olmaz. Hayatta davul bile dengi dengine çalar." görüşünü yanlış bulmuşken, %28,6'sı doğru bulmuş, %11,9'u karasız kalmış, %15,5'i ise bu görüşü kesinlikle yanlış olarak bulmuşken, %7,1'i bu görüşü kesinlikle doğru olarak değerlendirmiş. Gençlerin %0,2'si ise bu soruya cevap vermemiştir.

Tablo 3.23. Katılımcıların "Hayatımın gidişatını değiştirmek için yapabileceğim pek bir şey yok... Böyle gelmiş böyle gider…" Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	189	18,6
Yanlış	462	45,4
Ne doğru ne yanlış	113	11,1
Doğru	216	21,2
Kesinlikle Doğru	33	3,2
Cevap Yok	5	0,5
Toplam	1018	100

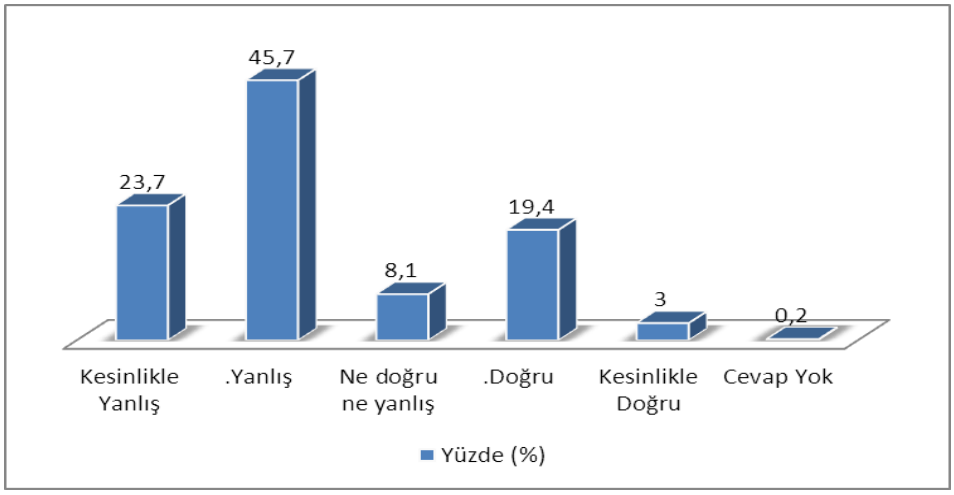


Şekil 3.19: Katılımcıların "Hayatımın gidişatını değiştirmek için yapabileceğim pek bir şey yok... Böyle gelmiş böyle gider…" Görüşü

Gençlerin %45,4'ü "Hayatımın gidişatını değiştirmek için yapabileceğim pek bir şey yok... Böyle gelmiş böyle gider…" görüşünü yanlış bulmuşken, %21,2'si doğru bulmuş, %11,1'i karasız kalmış, %18,6'sı ise bu görüşü kesinlikle yanlış olarak bulmuşken, %3,2'si bu görüşü kesinlikle doğru olarak değerlendirmiş. Gençlerin %0,5'i ise bu soruya cevap vermemiştir.

Tablo 3.24. Katılımcıların "Öğretmenin dövdüğü yerde gül biter" Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	241	23,7
Yanlış	465	45,7
Ne doğru ne yanlış	82	8,1
Doğru	197	19,4
Kesinlikle Doğru	31	3
Cevap Yok	2	0,2
Toplam	1018	100

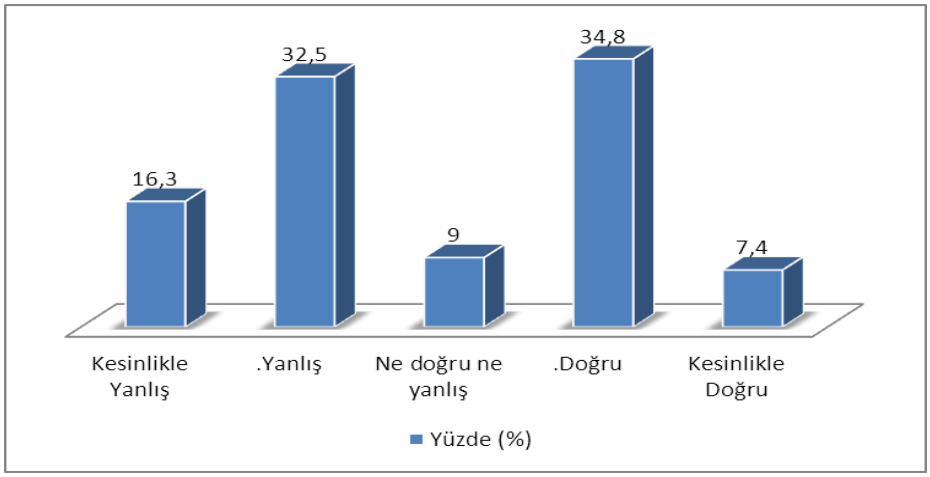


Şekil 3.20: Katılımcıların "Öğretmenin dövdüğü yerde gül biter" Görüşü

Gençlerin %45,7'si "Öğretmenin dövdüğü yerde gül biter" görüşünü yanlış bulmuşken, %19,4'ü doğru bulmuş, %8,1'i karasız kalmış, %23,7'si ise bu görüşü kesinlikle yanlış olarak bulmuşken, %3'ü bu görüşü kesinlikle doğru olarak değerlendirmiş. Gençlerin %0,2'si ise bu soruya cevap vermemiştir.

Tablo 3.25. Katılımcıların "Kızını dövmeyen dizini döver" Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	166	16,3
Yanlış	331	32,5
Ne doğru ne yanlış	92	9
Doğru	354	34,8
Kesinlikle Doğru	75	7,4
Toplam	1018	100

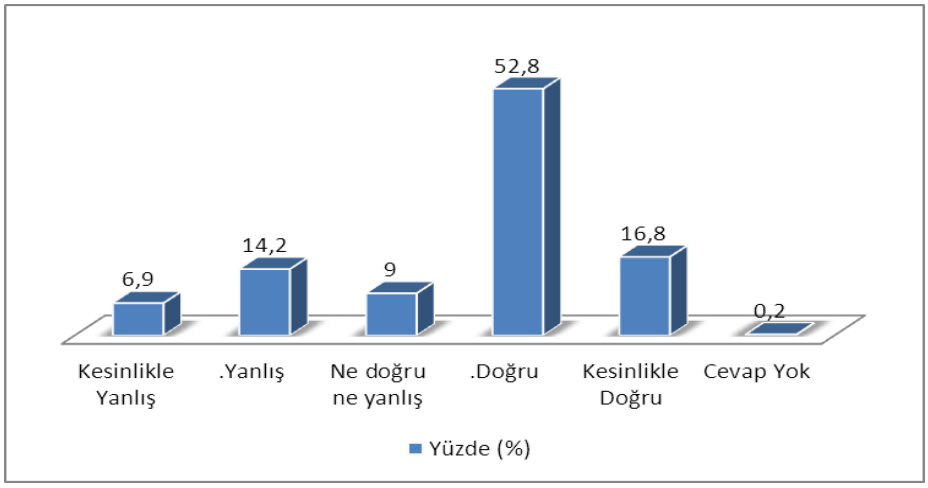


Şekil 3.21: Katılımcıların "Kızını dövmeyen dizini döver" Görüşü

Gençlerin %34,8'i "Kızını dövmeyen dizini döver" görüşünü doğru bulmuşken, %32,5'i yanlış bulmuş, %9'u karasız kalmış, %16,3'ü ise bu görüşü kesinlikle yanlış olarak bulmuşken, %7,4'ü bu görüşü kesinlikle doğru olarak değerlendirmiş.

Tablo 3.26. Katılımcıların "Gitme imkânı olsa bile yine Türkiye'de yaşamayı tercih ederdim." Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	70	6,9
Yanlış	145	14,2
Ne doğru ne yanlış	92	9
Doğru	538	52,8
Kesinlikle Doğru	171	16,8
Cevap Yok	2	0,2
Toplam	1018	100

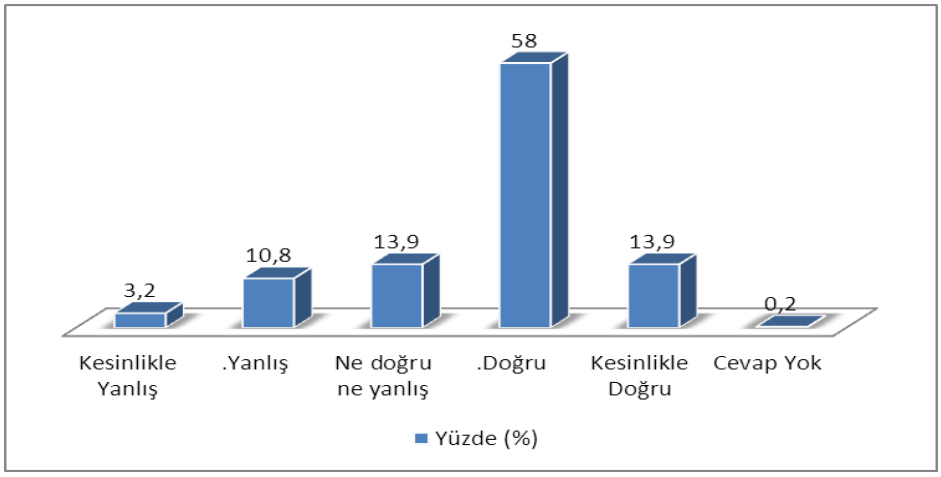


Şekil 3.22: Katılımcıların "Gitme imkânı olsa bile yine Türkiye'de yaşamayı tercih ederdim." Görüşü

Gençlerin %52,8'i "Gitme imkânı olsa bile yine Türkiye'de yaşamayı tercih ederdim." görüşünü doğru bulmuşken, %14,2'si yanlış bulmuş, %9'u karasız kalmış, %16,8'i ise bu görüşü kesinlikle doğru olarak bulmuşken, %6,9'u bu görüşü kesinlikle yanlış olarak değerlendirmiş. Gençlerin %0,2'si ise bu soruya cevap vermemiştir.

Tablo 3.27. Katılımcıların "Türkiye ortadoğu ve müslüman ülkelerle daha yakın bir ilişki içerisinde olmalıdır." Görüşü

Kesinlikle Yanlış	33	3,2
Yanlış	110	10,8
Ne doğru ne yanlış	142	13,9
Doğru	590	58
Kesinlikle Doğru	141	13,9
Cevap Yok	2	0,2
Toplam	1018	100

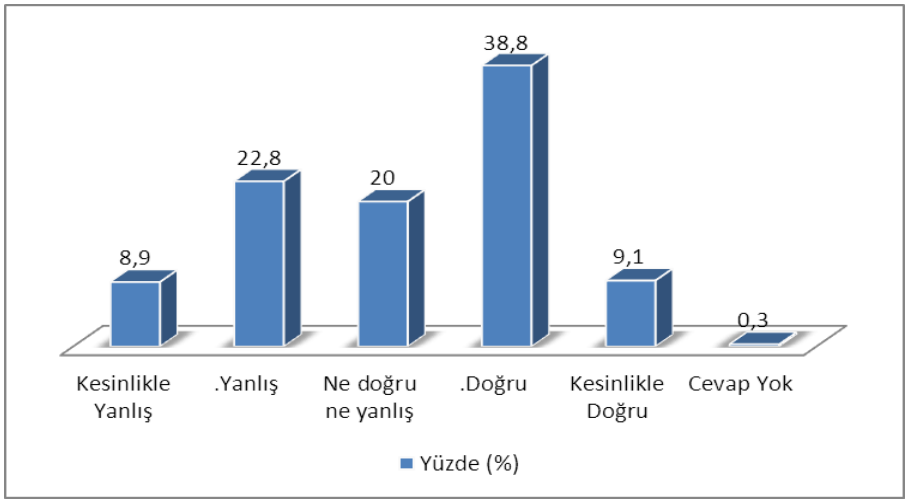


Şekil 3.23: Katılımcıların "Türkiye ortadoğu ve müslüman ülkelerle daha yakın bir ilişki içerisinde olmalıdır." Görüşü

Gençlerin %58'i "Türkiye ortadoğu ve müslüman ülkelerle daha yakın bir ilişki içerisinde olmalıdır." görüşünü doğru bulmuşken, %10,8'i yanlış bulmuş, %13,9'u karasız kalmış, %13,9'u ise bu görüşü kesinlikle doğru olarak bulmuşken, %3,2'si bu görüşü kesinlikle yanlış olarak değerlendirmiş. Gençlerin %0,2'si ise bu soruya cevap vermemiştir.

Tablo 3.28. Katılımcıların "Türkiye Avrupa Birliğine mutlaka üye olmalıdır." Görüşü

	Frekans	**Yüzde (%)**
Kesinlikle Yanlış	91	8,9
.Yanlış	232	22,8
Ne doğru ne yanlış	204	20
.Doğru	395	38,8
Kesinlikle Doğru	93	9,1
Cevap Yok	3	0,3
Toplam	1018	100

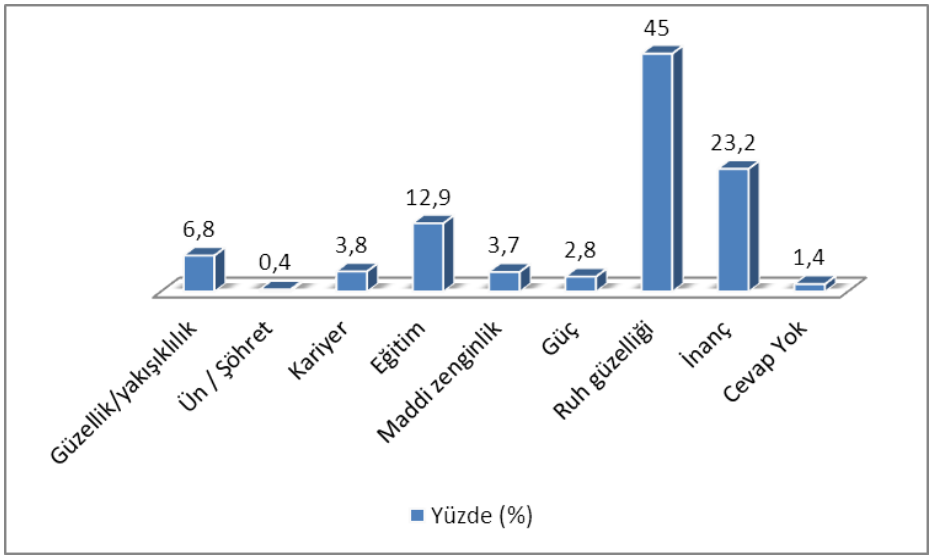


Şekil 3.24: Katılımcıların "Türkiye Avrupa Birliğine mutlaka üye olmalıdır." Görüşü

Gençlerin %38,8'i "Türkiye Avrupa Birliğine mutlaka üye olmalıdır." görüşünü doğru bulmuşken, %22,8'i yanlış bulmuş, %20'si karasız kalmış, %9,1'i ise bu görüşü kesinlikle doğru olarak bulmuşken, %8,9'u bu görüşü kesinlikle yanlış olarak değerlendirmiş. Gençlerin %0,3'ü ise bu soruya cevap vermemiştir.

Tablo 3.29. Katılımcıların "Evleneceğiniz insanda hangi özelliğin mutlaka olmasını istersiniz." Görüşü

	Frekans	**Yüzde (%)**
Güzellik/yakışıklılık	69	6,8
Ün / Şöhret	4	0,4
Kariyer	39	3,8
Eğitim	131	12,9
Maddi zenginlik	38	3,7
Güç	29	2,8
Ruh güzelliği	458	45
İnanç	236	23,2
Cevap Yok	14	1,4
Toplam	1018	100

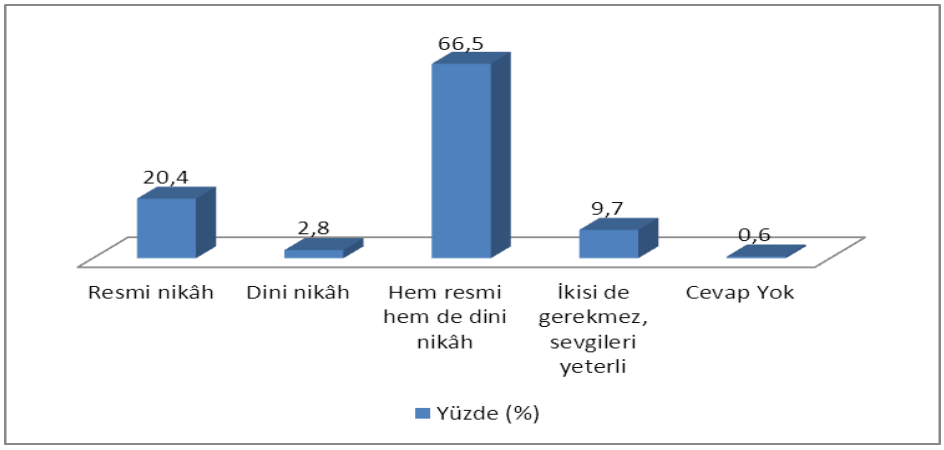


Şekil 3.25: Katılımcıların "Evleneceğiniz insanda hangi özelliğin mutlaka olmasını istersiniz." Görüşü

Gençlerin %45'lik büyük çoğunluğu evleneceği insanda ruh güzelliği aradığını, %23'lik kesim inanç, %12,9'luk oran ise eğitim aradığını belirtmiş, güzellik ve yakışıklılık arayanların oranı ise %6,8'dir. Bunları %3,8'lik oranla kariyer, %2,8'lik oranla güç ve en düşük oran olan %0,4'lük oranla ün ve şöhret izlemiştir. Gençlerin %1,4'ü ise bu soruya cevap vermemiştir.

Tablo 3.30. Katılımcıların "Bir kadınla bir erkek aşağıdakilerden hangisi olur ise sizce beraber yaşayabilirler?" Görüşü

	Frekans	**Yüzde (%)**
Resmi nikâh	208	20,4
Dini nikâh	28	2,8
Hem resmi hem de dini nikâh	677	66,5
İkisi de gerekmez, sevgileri yeterli	99	9,7
Cevap Yok	6	0,6
Toplam	1018	100

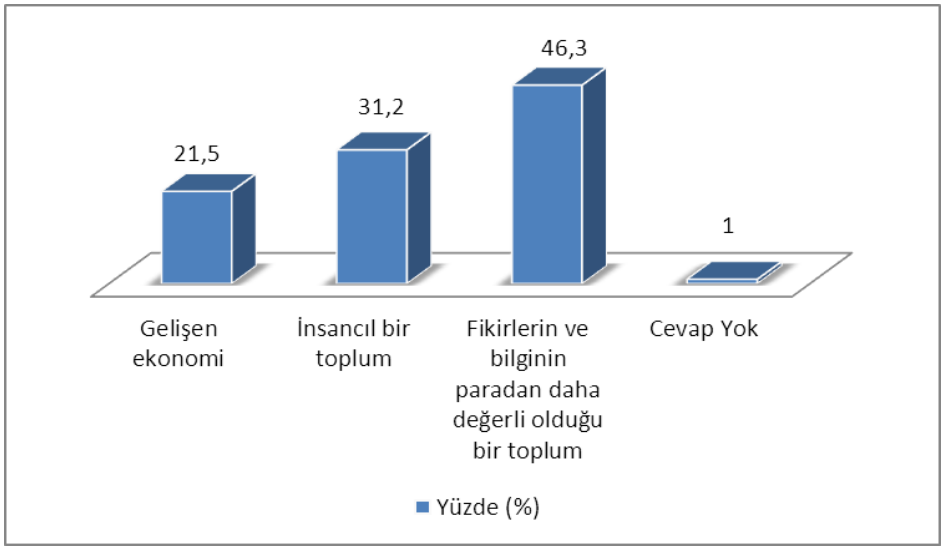


Şekil 3.26: Katılımcıların "Bir kadınla bir erkek aşağıdakilerden hangisi olur ise sizce beraber yaşayabilirler?" Görüşü

Gençlerin %66,5'lik büyük çoğunluğu bir kadınla bir erkek arasında hem resmi hem de dini nikâh olursa beraber yaşayabileceğini, %20,4'lük kesim ise resmih nikâh ile beraber yaşanabileceğini belirtmiştir. %9,7'lik kısmı ise sadece sevginin yeterli olacağını belirtirken, %2,8'lik kısım ise dini nikâhın olmasının yeterli olduğunu belirtmiştir. Gençlerin %0,6'ü ise bu soruya cevap vermemiştir.

Tablo 3.31. Katılımcıların "Seçim yapmak durumunda olsaydınız, hangisinin daha önemli olduğunu söylerdiniz?" Görüşü

	Frekans	**Yüzde (%)**
Gelişen ekonomi	219	21,5
İnsancıl bir toplum	318	31,2
Fikirlerin ve bilginin paradan daha değerli olduğu bir toplum	471	46,3
Cevap Yok	10	1
Toplam	1018	100

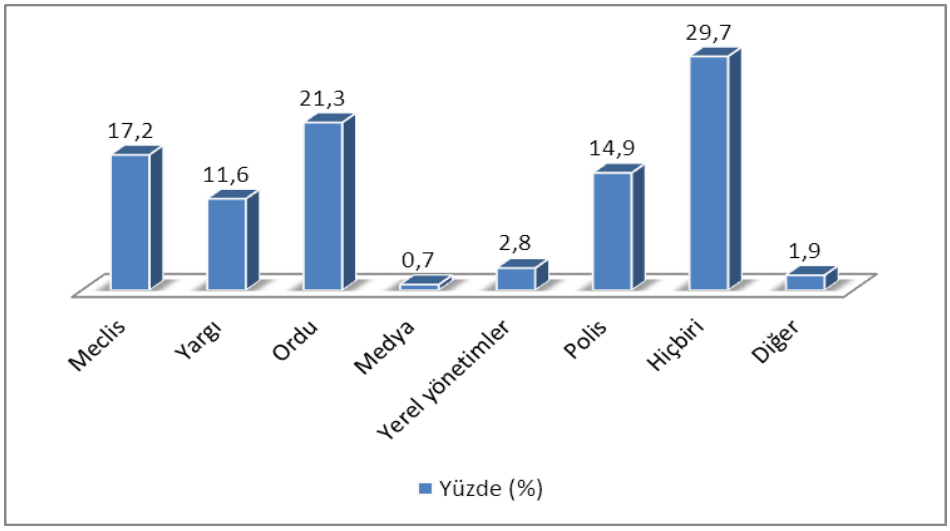


Şekil 3.27: Katılımcıların "Seçim yapmak durumunda olsaydınız, hangisinin daha önemli olduğunu söylerdiniz?" Görüşü

Gençlerin %46,3'lük büyük çoğunluğu fikirlerin ve bilginin paradan daha değerli olduğu bir toplumu önemli olduğunu belirtirken, %31,2'si ise insalcıl bir toplumun önemli olduğunu belirtmiştir. Bunu %21,5 oranla gelişen bir ekonominin önemli olduğu seçeneği takip etmiştir. Gençlerin %1'i ise bu soruya cevap vermemiştir.

Tablo 3.32. Katılımcıların "Ençok hangi kuruma gönülden güvenirsiniz?" Görüşü

	Frekans	**Yüzde (%)**
Meclis	175	17,2
Yargı	118	11,6
Ordu	217	21,3
Medya	7	0,7
Yerel yönetimler	28	2,8
Polis	152	14,9
Hiçbiri	302	29,7
Diğer	19	1,9
Toplam	1018	100

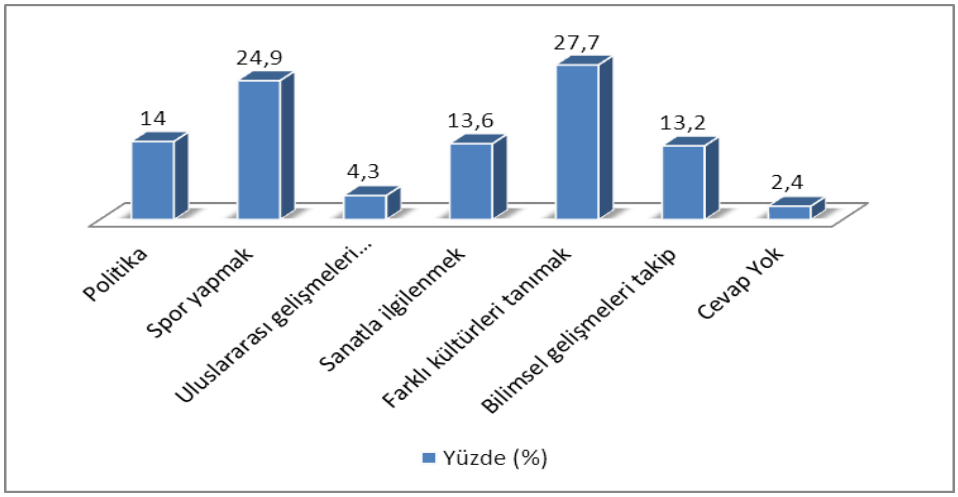


Şekil 3.28: Katılımcıların "Ençok hangi kuruma gönülden güvenirsiniz?" Görüşü

Gençlerin %29,7'lik'lük büyük çoğunluğu bu kurumlardan hiçbirine güvenmezken, %21,3'lük kısımı en çok orduya güvendiğini belirtmiştir. Meclise güvenlerinin oranı %17,2 iken polis güvenenlerin oranı %14,9, yargıya güvenenlerin oranı %11,6 bulunmuştur. Bu oranları %2,8 ile yerel yönetimler, %1,9 ile diğer kurumlar ve %0,7 ile medya izlemiştir.

Tablo 3.33. Katılımcıların "En çok hangisi ilginizi, merakınızı çeker?" Görüşü

	Frekans	**Yüzde (%)**
Politika	143	14
Spor yapmak	253	24,9
Uluslararası gelişmeleri takip	44	4,3
Sanatla ilgilenmek	138	13,6
Farklı kültürleri tanımak	282	27,7
Bilimsel gelişmeleri takip	134	13,2
Cevap Yok	24	2,4
Toplam	1018	100

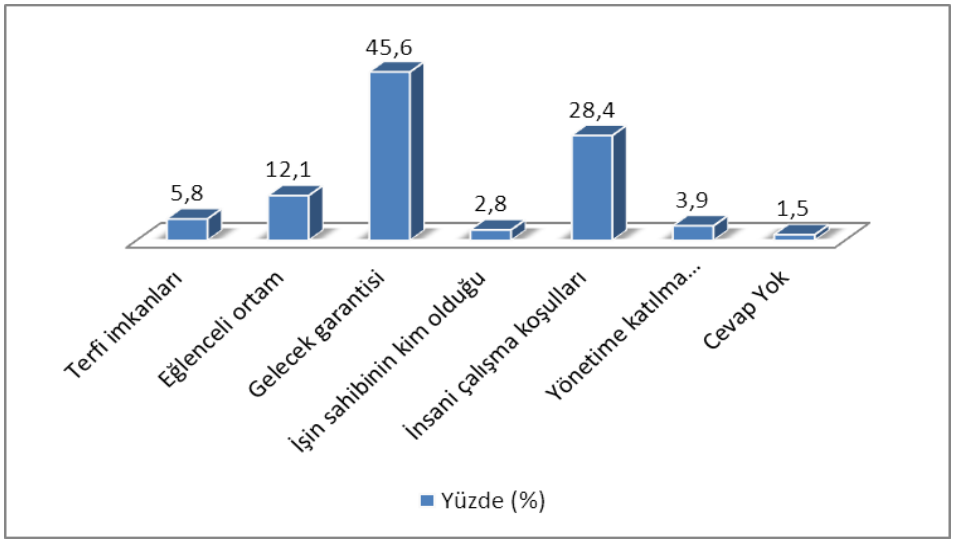


Şekil 3.29: Katılımcıların "En çok hangisi ilginizi, merakınızı çeker" Görüşü

Gençlerin %27,7'lik büyük çoğunluğunun ilgisini farklı kültürleri tanımak çekerken, %24,9'unun spor, %14'ünün politika, %13,6'sının sanat, %13,2'sinin bilimsel gelişmeleri takip etmek ve %4,3'ünün ise uluslararası gelişmeleri takip etmek igisini çekmektedir. Gençlerin %2,4'ü ise bu soruya cevap vermemiştir.

Tablo 3.34. Katılımcıların "İstiyerek mutlu olarak çalışacağınız işten maaşı dışında beklentileriniz" Görüşü

	Frekans	**Yüzde (%)**
Terfi imkânları	59	5,8
Eğlenceli ortam	123	12,1
Gelecek garantisi	464	45,6
İşin sahibinin kim olduğu	28	2,8
İnsani çalışma koşulları	289	28,4
Yönetime katılma olanakları	40	3,9
Cevap Yok	15	1,5
Toplam	1018	100

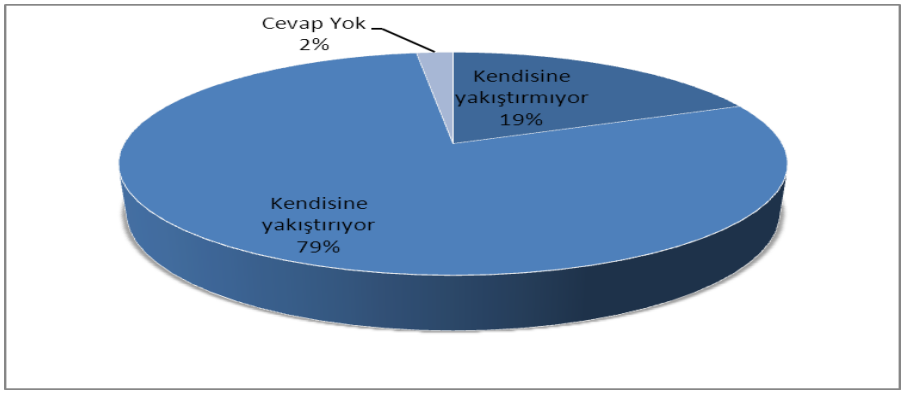


Şekil 3.30: Katılımcıların "İstiyerek mutlu olarak çalışacağınız işten maaşı dışında beklentileriniz" Görüşü

Gençlerin %45,6'lık büyük çoğunluğu işinden maaşı dışında gelecek grantisi beklerken, %28,4'ü insani çalışma koşulları, %12,1'i eğlenceli ortam, %5,8'i terfi imkânı ve %3,9'u yönetime katılma olanakları beklemektedir. Gençlerin %2,8'i işin sahibinin kim olduğuna önem verirken, %2,4'ü bu soruya cevap vermemiştir.

Tablo 3.35. Katılımcıların "Maceracı sıfatını kendinize yakıştırır mısınız?" Görüşü

	Frekans	**Yüzde (%)**
Kendisine yakıştırmıyor	193	19
Kendisine yakıştırıyor	804	79
Cevap Yok	21	2,1
Toplam	1018	100

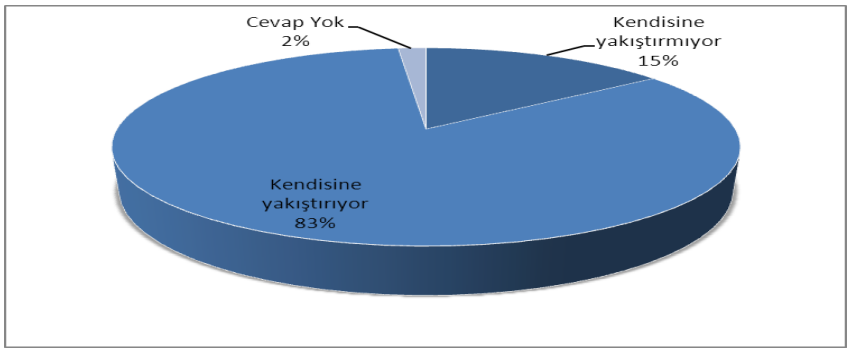


Şekil 3.31: Katılımcıların "Maceracı sıfatını kendinize yakıştırır mısınız?" Görüşü

Gençlerin %79'luk büyük çoğunluğu maceracı sıfatını kendisine yakıştırıyor. %19'u ise, kendisine yakıştırmıyor.

Tablo 3.36. Katılımcıların "Rekabetçi sıfatını kendinize yakıştırır mısınız?" Görüşü

	Frekans	**Yüzde (%)**
Kendisine yakıştırmıyor	154	15,1
Kendisine yakıştırıyor	847	83,2
Cevap Yok	17	1,7
Toplam	1018	100

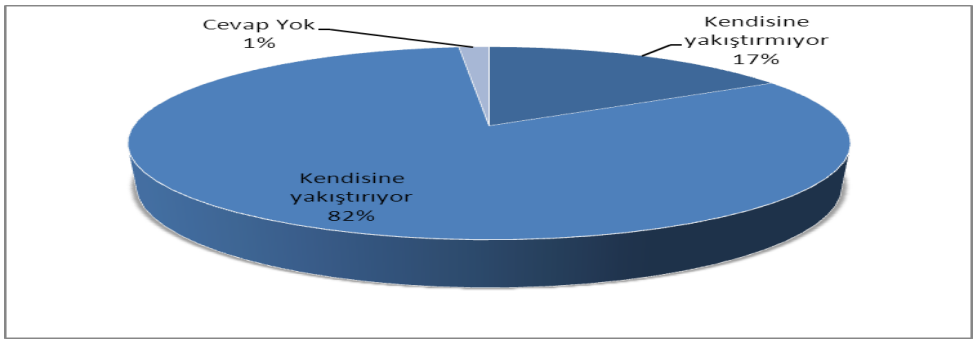


Şekil 3.32: Katılımcıların "Rekabetçi sıfatını kendinize yakıştırır mısınız?" Görüşü

Gençlerin %83'luk büyük çoğunluğu rekabetçi sıfatını kendisine yakıştırıyor. %15'i ise, kendisine yakıştırmıyor.

Tablo 3.37. Katılımcıların "Uzlaşmacı sıfatını kendinize yakıştırır mısınız?" Görüşü

	Frekans	Yüzde (%)
Kendisine yakıştırmıyor	170	16,7
Kendisine yakıştırıyor	832	81,7
Cevap Yok	16	1,6
Toplam	1018	100

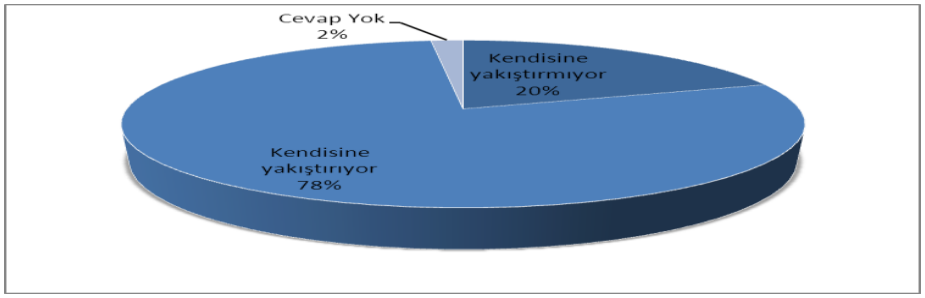


Şekil 3.33: Katılımcıların "Uzlaşmacı sıfatını kendinize yakıştırır mısınız?" Görüşü

Gençlerin %82'lik büyük çoğunluğu uzlaşmacı sıfatını kendisine yakıştırıyor. %17'si ise, kendisine yakıştırmıyor.

Tablo 3.38. Katılımcıların "Sevecenlik sıfatını kendinize yakıştırır mısınız?" Görüşü

	Frekans	Yüzde (%)
Kendisine yakıştırmıyor	203	19,9
Kendisine yakıştırıyor	797	78,3
Cevap Yok	18	1,8
Toplam	1018	100

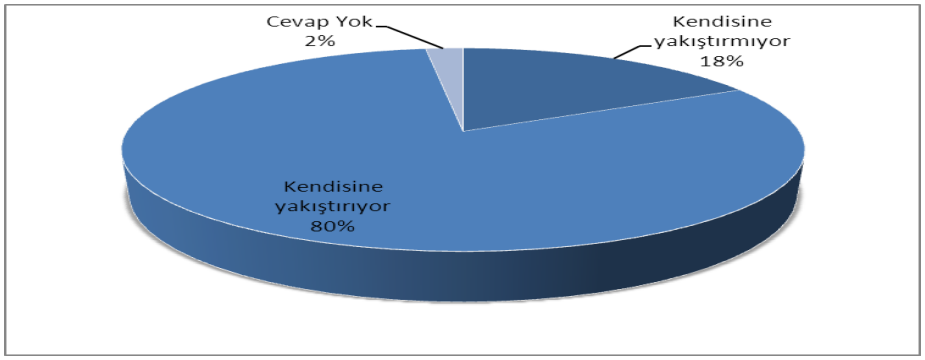


Şekil 3.34: Katılımcıların "Sevecenlik sıfatını kendinize yakıştırır mısınız?" Görüşü

Gençlerin %78'lik büyük çoğunluğu sevencenlik sıfatını kendisine yakıştırıyor. %20'si ise, kendisine yakıştırmıyor.

Tablo 3.39. Katılımcıların "Özgürlük düşkünü sıfatını kendinize yakıştırır mısınız?" Görüşü

	Frekans	**Yüzde (%)**
Kendisine yakıştırmıyor	178	17,5
Kendisine yakıştırıyor	819	80,5
Cevap Yok	21	2,1
Toplam	1018	100

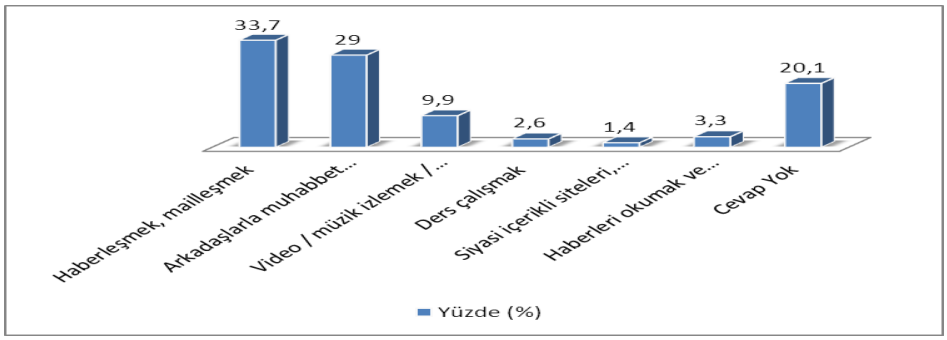


Şekil 3.35: Katılımcıların "Özgürlük düşkünü sıfatını kendinize yakıştırır mısınız?" Görüşü

Gençlerin %80'lik büyük çoğunluğu özgürlük düşkünü sıfatını kendisine yakıştırıyor. %18'i ise, kendisine yakıştırmıyor.

Tablo 3.40. Katılımcıların "İnterneti ençok hangi eylem için kullanıyorsunuz?" Görüşü

	Frekans	Yüzde (%)
Haberleşmek, mailleşmek	343	33,7
Arkadaşlarla muhabbet etmek	295	29
Video, müzik izlemek / oyun oynamak	101	9,9
Ders çalışmak	26	2,6
Siyasi içerikli siteleri, tartışmaları takip	14	1,4
Haberleri okumak ve izlemek	34	3,3
Cevap Yok	205	20,1
Toplam	1018	100

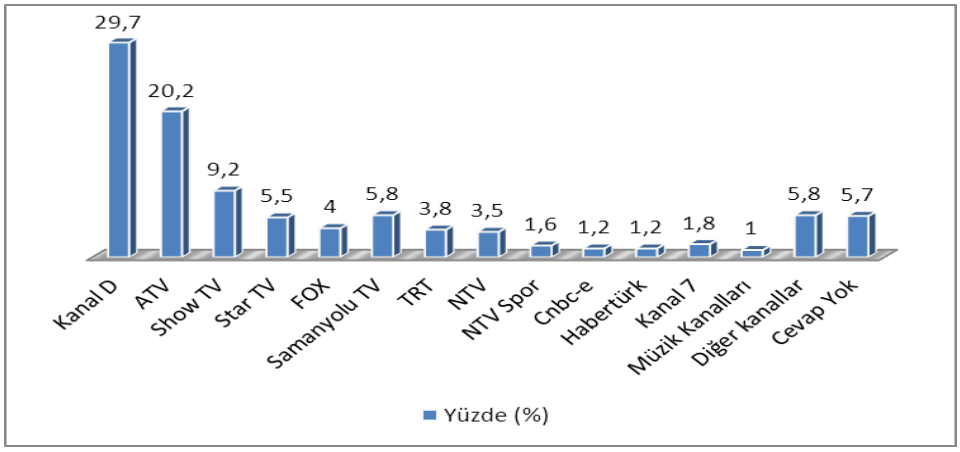


Şekil 3.36: Katılımcıların "İnterneti ençok hangi eylem için kullanıyorsunuz?" Görüşü

Gençlerin %33,7'lik büyük çoğunluğu haberleşmek, mailleşmek için interneti kullanırken, %29'luk kısmı arkadaşlarla muhabbet etmek ve %9,9'luk kısmı ise video, müzik izlemek, oyun oynamak için interneti kullanmaktadır. Bunları oranları %3,3'lük oranla haberleri okumak ve izlemek, %2,6'lık oranla ders çalışmak ve %1,4'lük en az oranla siyasi içerikli siteleri, tartışmaları takip etmek izlemek takip etmektedir. Gençlerin %20,1'lik büyük çoğunluğu ise, bu soruya cevap vermemiştir.

Tablo 3.41. Katılımcıların Ençok izlediği TV kanalı Dağılımı

	Frekans	Yüzde (%)
Kanal D	302	29,7
ATV	206	20,2
Show TV	94	9,2
Star TV	56	5,5
FOX	41	4
Samanyolu TV	59	5,8
TRT	39	3,8
NTV	36	3,5
NTV Spor	16	1,6
Cnbc-e	12	1,2
Habertürk	12	1,2
Kanal 7	18	1,8
Müzik Kanalları	10	1
Diğer kanallar	59	5,8
Cevap Yok	58	5,7
Toplam	1018	100

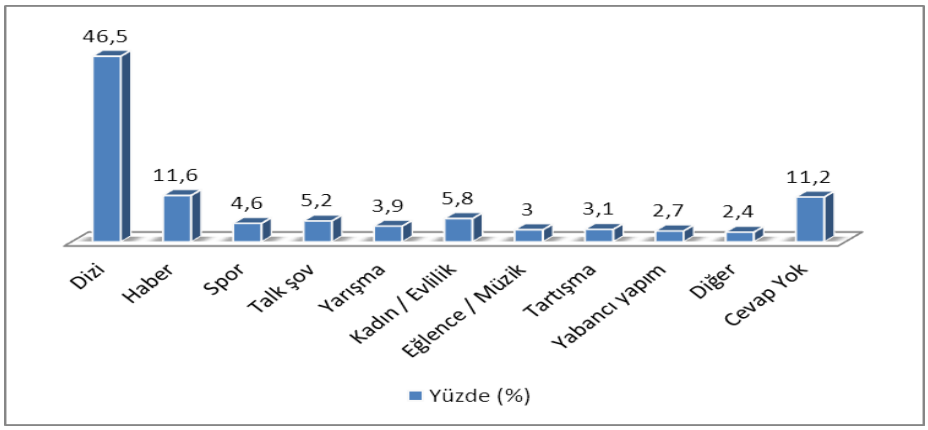


Şekil 3.37: Katılımcıların Ençok izlediği TV kanalı Dağılımı

Gençlerin büyük çoğunluğu sırasıyla %30'luk ve %20'lik oranlarla Kanal D ve ATV'yi izlerken, bu kanalları %9'luk oranla Show TV, %5,8'lik oranla Samanyolu TV %5,5 oranla Star TV ve %4'lük oranla FOX takip etmektedir. Sonrasında sırasıyla TRT ve NTV %3,8 ve %3,5'lik oranlarla gelmektedir. %1 ile %2 arasında değişen az bir kısım ise NTV Spor, Cnbc-e, Habertürk, Kanal 7 ve müzik kanallarını izlemektedir.

Tablo 3.42. Katılımcıların Ençok izlediği TV Programı Dağılımı

	Frekans	Yüzde (%)
Dizi	473	46,5
Haber	118	11,6
Spor	47	4,6
Talk şov	53	5,2
Yarışma	40	3,9
Kadın / Evlilik	59	5,8
Eğlence / Müzik	31	3
Tartışma	32	3,1
Yabancı yapım	27	2,7
Diğer	24	2,4
Cevap Yok	114	11,2
Toplam	1018	100

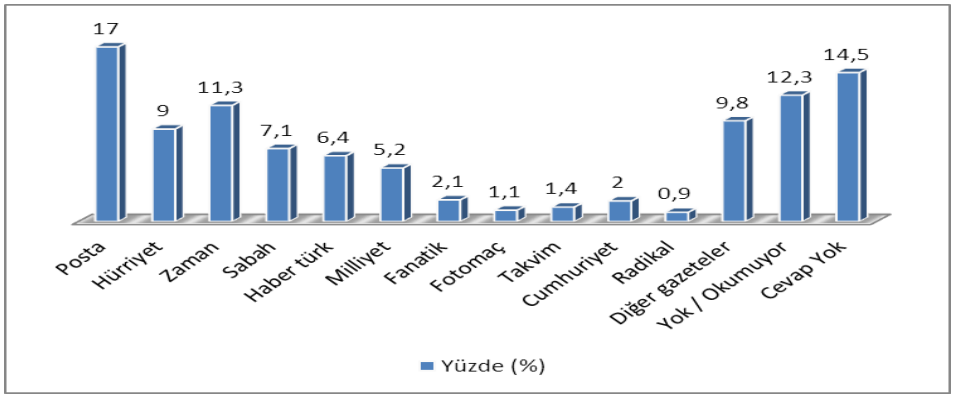


Şekil 3.38: Katılımcıların Ençok izlediği TV Programı Dağılımı

Gençlerin büyük çoğunluğu %46,4'lik oranla dizi izlemektedir. Bunu %11,6'lık oranla haber, %5,8'lik oranla kadın/evlilik programları, %5,2'lik oranla talk şov ve %3,9'luk oranla yarışma programları takip etmektedir. %2 ile %3 arasında değişen kısım ise eğlence/müzik, tartışma ve yabancı yapım programlarını izlemektedir.

Tablo 3.43. Katılımcıların Okuduğu Gazeteye Göre Dağılımı

	Frekans	Yüzde (%)
Posta	173	17
Hürriyet	92	9
Zaman	115	11,3
Sabah	72	7,1
Haber türk	65	6,4
Milliyet	53	5,2
Fanatik	21	2,1
Fotomaç	11	1,1
Takvim	14	1,4
Cumhuriyet	20	2
Radikal	9	0,9
Diğer gazeteler	100	9,8
Yok / Okumuyor	125	12,3
Cevap Yok	148	14,5
Toplam	1018	100

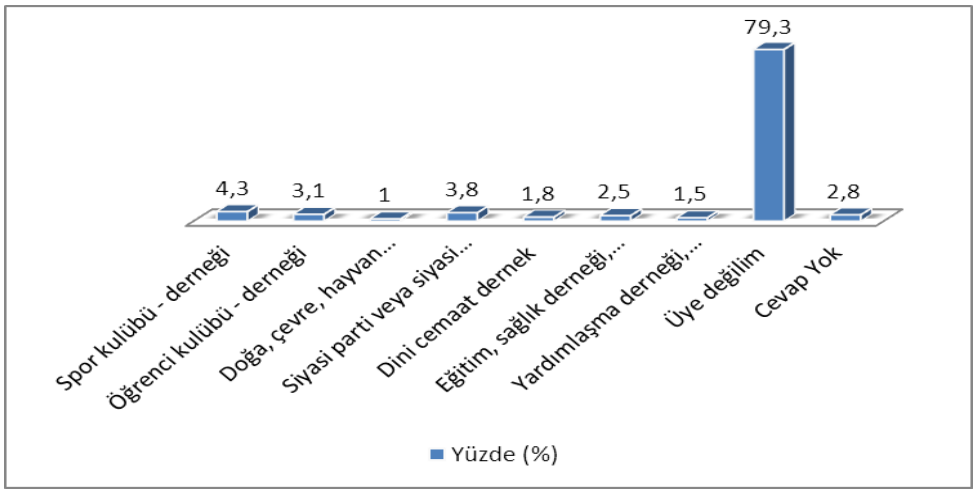


Şekil 3.39: Katılımcıların Okuduğu Gazeteye Göre Dağılımı

Gençlerin büyük çoğunluğu %17'lik oranla Posta okumaktadır. Bunu %11,3'lük oranla Zaman, %7,1'lik oranla Sabah, %6,4'lük oranla Haber türk ve %5,2'lik oranla Milliyet gazetesi takip etmektedir. Cumhuriyet, Radikal, Takvim, Fotomaç ve Fanatik gazetelerinin gençler arasında okunma oranları ise, %1 ile %2 arasında değişmektedir.

Tablo 3.44. Katılımcıların Dernek–Vakıf-Sosyal Kulüp Üyeliği Durumuna Göre Dağılımı

	Frekans	Yüzde (%)
Spor kulübü - derneği	44	4,3
Öğrenci kulübü - derneği	32	3,1
Doğa, çevre, hayvan koruma türü	10	1
Siyasi parti veya siyasi dernek	39	3,8
Dini cemaat dernek	18	1,8
Eğitim, sağlık derneği, vakfı	25	2,5
Yardımlaşma derneği, vakfı	15	1,5
Üye değilim	807	79,3
Cevap Yok	28	2,8
Toplam	1018	100

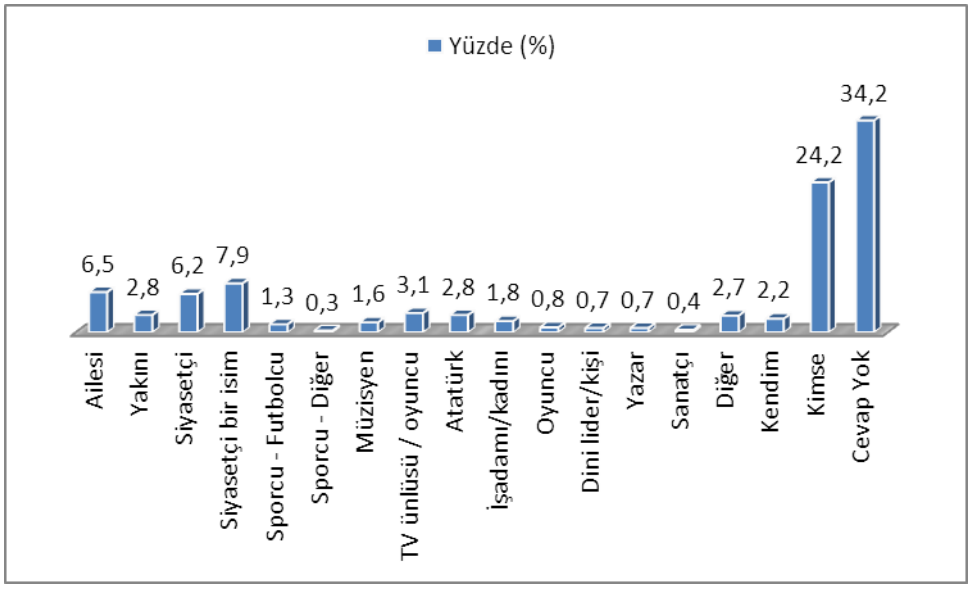


Şekil 3.40: Katılımcıların Dernek–Vakıf-Sosyal Kulüp Üyeliği Durumuna Göre Dağılımı

Gençlerin %79,3'lük büyük çoğunluğunun herhangi bir dernek-vakıf ve sosyal kulüp üyeliği bulunmazken, bu üyelikler içinde en yüksek oranı %4,3 ile spor kulübü ve derneği almaktadır. Bunu %3,8'lik oranla siyasi parti veya siyasi dernek takip etmektedir. Doğa, çevre, hayvan koruma, dini cemaat dernek, eğitim, sağlık derneği ve yardımlaşma derneklerine üyelikler ise %1 ile %2 arasında değişmektedir.

Tablo 3.45. Katılımcıların İdolüne Göre Dağılımı

	Frekans	Yüzde (%)
Ailesi	66	6,5
Yakını	29	2,8
Siyasetçi	63	6,2
Siyasetçi bir isim	80	7,9
Sporcu - Futbolcu	13	1,3
Sporcu - Diğer	3	0,3
Müzisyen	16	1,6
TV ünlüsü / oyuncu	32	3,1
Atatürk	29	2,8
İşadamı/kadını	18	1,8
Oyuncu	8	0,8
Dini lider/kişi	7	0,7
Yazar	7	0,7
Sanatçı	4	0,4
Diğer	27	2,7
Kendim	22	2,2
Kimse	246	24,2
Cevap Yok	348	34,2
Toplam	1018	100



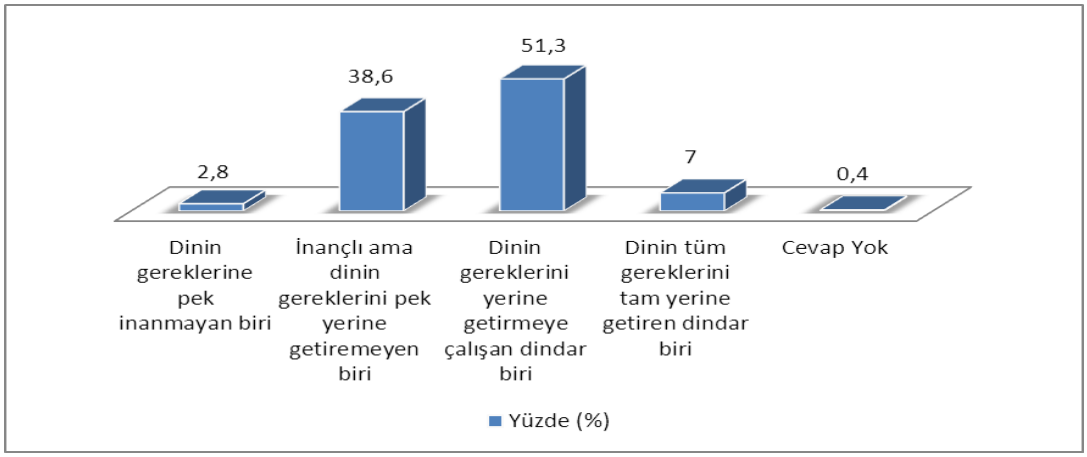
Şekil 3.41: Katılımcıların İdolüne Göre Dağılımı

Gençlerin %24'lük büyük çoğunluğunun herhangi bir idolü bulunmazken, idolü olanlar arasında % 7,9'luk en büyük oranla siyasetçi bir isim yer almaktadır. Bu oranı %6,5 ile aile ve %6,2 ile siyasetçi izlemektedir. İdolü TV ünlüsü / oyuncu olanlar

%3,1 idolü Atatürk ve yakını olanların oranı ise %2,8'dir. Bunları sırasıyla kendisi, İşadamı/kadını, sporcu (futbolcu), oyuncu, dini lider/kişi, yazar, sanatçı ve sporcu (diğer) izlemiştir.

Tablo 3.46. Katılımcıların "Dindarlık açısından kendinizi hangisiyle tarif edersiniz" Görüşü

	Frekans	**Yüzde (%)**
Dinin gereklerine pek inanmayan biri (inançsız)	28	2,8
İnançlı ama dinin gereklerini pek yerine getiremeyen biri (inançlı)	393	38,6
Dinin gereklerini yerine getirmeye çalışan dindar biri (dindar)	522	51,3
Dinin tüm gereklerini tam yerine getiren dindar biri (sofu)	71	7
Cevap Yok	4	0,4
Toplam	1018	100

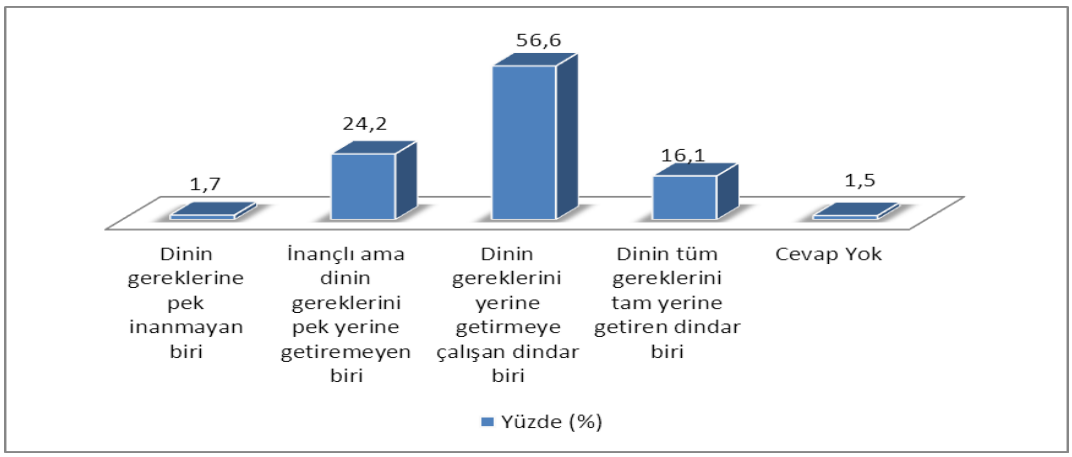


Şekil 3.42: Katılımcıların "Dindarlık açısından kendinizi hangisiyle tarif edersiniz" Görüşü

Gençlerin %51,3'lük büyük çoğunluğu kendini dinin gereklerini yerine getirmeye çalışan dindar biri olarak tanımlarken, %38,6'lık kısmı ise, inançlı ama dinin gereklerini pek yerine getiremeyen biri olarak tanımlamıştır. Bunu %7'lik oranla kendilerini dinin tüm gereklerini tam yerine getiren dindar biri ve %2,8'lük bir oranla dinin gereklerine pek inanmayan biri olarak tanımlayan gençler izlemiştir.

Tablo 3.47. Katılımcıların "Dindarlık açısından ailenizin reisini hangisiyle tarif edersiniz" Görüşü

	Frekans	Yüzde (%)
Dinin gereklerine pek inanmayan biri (inançsız)	17	1,7
İnançlı ama dinin gereklerini pek yerine getiremeyen biri (inançlı)	246	24,2
Dinin gereklerini yerine getirmeye çalışan dindar biri (dindar)	576	56,6
Dinin tüm gereklerini tam yerine getiren dindar biri (sofu)	164	16,1
Cevap Yok	15	1,5
Toplam	1018	100

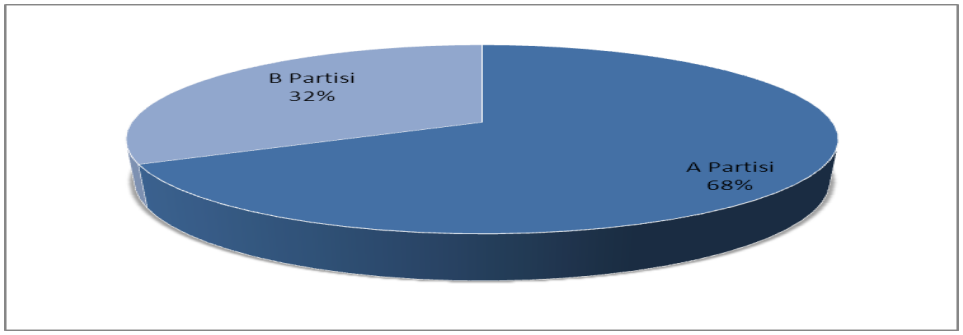


Şekil 3.43: Katılımcıların "Dindarlık açısından ailenizin reisini hangisiyle tarif edersiniz" Görüşü

Gençlerin %56,6'lık büyük çoğunluğu aile reisini dinin gereklerini yerine getirmeye çalışan dindar biri olarak tanımlarken, %24,2'lik kısmı ise, inançlı ama dinin gereklerini pek yerine getiremeyen biri olarak tanımlamıştır. Bunu %16,1'lik oranla aile reislerini, dinin tüm gereklerini tam yerine getiren dindar biri ve %1,7'lik bir oranla dinin gereklerine pek inanmayan biri olarak tanımlayan gençler izlemiştir.

Tablo 3.48. Katılımcıların "Önümüzdeki seçimlerde hangi partiye oy (2 düzey) vereceksiniz?" Sorunun Cevabına Göre Dağılımı

	Frekans	Yüzde (%)
A Partisi	696	68
B Partisi	322	32
Toplam	1018	100



Şekil 3.44: Katılımcıların "Önümüzdeki seçimlerde hangi partiye oy vereceksiniz?" Sorunun Cevabına Göre Dağılımı

Gençlerin %68'lik büyük çoğunlu önümüzdeki seçimlerde A partisine oy vereceğini belirtirken, %32'si ise, B Partisine oy vereceğini belirtmiştir.

3.3.2. CART Yöntemine Ait Bulgular

Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde CART yöntemi farklı büyüklükteki başlangıç ve test verisi kullanılarak incelenmiştir. ilk aşamada bağımlı değişkenin kategori düzeyi ikiye indirgenmiş ve en çok tercih edilen ilk iki parti olan "A Partisi" ve "B Partisi" ele alınmıştır. Bu amaçla ilk aşamada başlangıçta modeli oluştumak için veri setinin sırasıyla %70'i, %50'si ve %30'u başlangıç veri seti bir diğer adıyla eğitim verisi olarak ele alınmıştır. Buna göre veri setinin sırasıyla geri kalan %30, %50 ve %70'lik kısmı ise oluşan modeli test etmek için kullanılmıştır. Bu durumda CART yöntemi ile 3 adet model elde edilmiştir.

İkinci aşamada ise en uygun olan başlangıç ve test verisi büyüklüğü ile bu kez genel bir değerlendirme yapılmıştır. Bu değerlendirme de bağımlı değişken olan parti tercihi "A Partisi", "B Partisi", "C Partisi" "Diğer" ve "Kararsız" olmak üzere beş kategori düzeyi esas alınarak bu kez farklı büyüklükteki örnek sayılarına göre CART yöntemi ile modellenmiş ve çeşitli örnek büyüklükleri içerisinden en başarılı sınıflama modeli belirlenmiştir. Buna göre yapılan analizlere ait bulgular izleyen kısımlarda sırasıyla verilmiştir.

3.3.2.1. CART Birinci Model Bulguları

CART ile kurulan ilk modelde (CART-1) başlangıç veri setinin %70'i modeli oluşturmak %30'u ise modeli test etmek için kullanılmıştır. Veri setine ait sonuçlar aşağıdaki tabloda gösterilmiştir.

Tablo 3.49. Başlangıç ve Test Verisi (CART-1)

Parti	Başlangıç	Yüzde (%)	Test	Yüzde (%)	Toplam	Yüzde (%)
A	490	67.59	206	70.31	696	68
B	235	32.41	87	29.69	322	32
Toplam	725	100.00	293	100.00	1018	100

Tabloda da görüldüğü üzere toplam 1018 gözlemin 696'sı (%68) A partisine oy vereceğini, 322'si (%32) ise B partisine oy vereceğini belirtmiştir. Sonuçta bu gözlemlerin 725'i (%68) başlangıç veri seti olup, modeli oluştumak için kullanılmıştır. 293'ü (%30) ise test verisi olup, kurulan modeli test etmek için kullanılmıştır.

CART-1 modeli sonucunda düğüm sayılarına göre oluşan 7 farklı ağacın maliyet-karmaşıklık tablosu aşağıda verilmiştir.

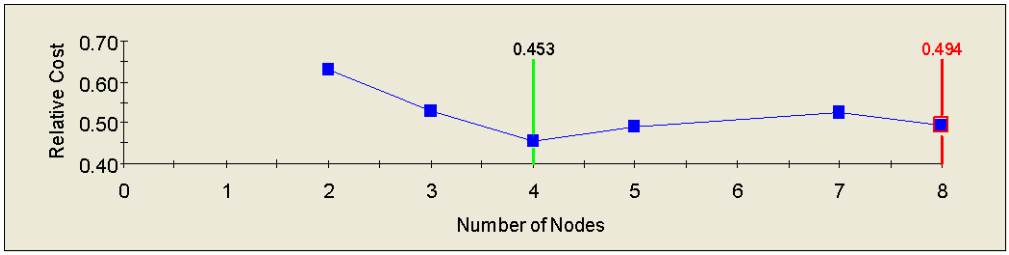
Tablo 3.50. Oluşan Ağaç Dizilerinin Maliyet-Karmaşıklık Tablosu (CART-1)

Ağaç	Uç Düğümler	Test Seti Göreceli Maliyeti	Yeniden Yerine Koyma Maliyeti	Karmaşıklık
1	8	0.49431 ± 0.05590	0.37668	0.00000
2	7	0.52650 ± 0.05634	0.38602	0.00469
3	5	0.49124 ± 0.05605	0.40868	0.00568

Ağaç	Uç Düğümler	Test Seti Göreceli Maliyeti	Yeniden Yerine Koyma Maliyeti	Karmaşıklık
4**	4	0.45319 ± 0.05364	0.43231	0.01182
5	3	0.52673 ± 0.05750	0.52071	0.04421
6	2	0.62951 ± 0.04880	0.61233	0.04582
7	1	1.00000 ± 3.72529E-009	1.00000	0.19384

Tablo 3.50'deki bilgilere göre oluşan ağaçlar arasından 4 numaralı ağaç optimum olarak gösterilmiştir. Fakat 4 adet uç düğümden oluşan, karmaşıklığı fazla ve göreceli maliyeti en az olan bu ağaç yerine daha yüksek maliyet göze alınıp, karmaşıklığı en az olan 1 numaralı ağaç incelemeye alınmıştır. Bu ağaçta 8 adet uç düğüm bulunduğu için sınıflandırma aşamasında daha detaylı kurallarla, uç düğüme yani parti sınıflarına atamalar yapılmaktadır. Bu yüzden çalışmada parti tercihinde önemli olabilecek bağımsız değişkenlerin de modelde kalabilmesi amaçlanmıştır. Bu yüzden 8 adet düğümden oluşan bu ağaç incelemeye alınmıştır.

Bu durum aşağıdaki grafik ile daha iyi anlaşılmaktadır. Bu grafikte 8 düğümden oluşan ağacın göreceli maliyetinin 4 düğümden oluşan ağaçtan biraz daha yüksek olduğu görülmektedir.



Şekil 3.45: Maliyet-Karmaşıklık Karşılaştırması (CART-1)

CART yönteminde çalışmanın literatür bölümünde de belirtildiği gibi modele giren tüm değişkenler arasından en önemliler belirlenerek model oluşturulur. Önemsiz değişkenler modele katılmaz. Buna göre birinci model neticesinde anlamlı bulunan değişkenler, aşağıdaki tabloda gösterilmiştir.

Tablo 3.51. Değişkenlerin Önemlilik Oranları (CART-1)

Değişken	Skor	
X33_KURUM_BAGI (En çok hangi kuruma gönülden güvenirsiniz?)	100,0000	
X5_2DOGUM_BOLGE (Doğum yeriniz nedir? "bölge")	65,0714	
X6_2_BABA_DB (Babanızı doğum yeri- nedir? "bölge")	54,5056	
X63_DINDARLIK (Dindarlık açısından kendinizi hangisiyle tarif edersiniz?)	52,1974	
X64_AILE_DIN (Dindarlık açısından ailenizin resisini hangisiyle tarif edersiniz?)	45,6914	
X8_MEDENI (Medeni durumunuz?)	36,0714	
X13_TUR_HAYATBEK (Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak.)	29,6379	
X9_1_IS (Son bir ay içinde bir işte çalıştınız mı? Çalıştınızsa işiniz?)	28,5500	
X30_EVLILIK (Evleneceğiniz insanda hangi özelliğin mutlaka olmasını istersiniz?	16,4747	
X60_GAZETE (Okuduğunuz gazete hangisidir?)	14,2003	
X3_1_EGITIM (Eğitim durumunuz, yani son bitridiğiniz okul nedir?)	12,5446	
X34_MERAK (En çok hangisi ilginizi, merakınızı çeker?)	9,8804	
X21_TURKIYE_ODOGU (Türkiye Orta Doğu ve Müslüman ülkelerle daha yakın işbirliği içinde olmalıdır.)	6,1364	
X2_YAS (Kaç yaşındasınız?)	4,6329	
X12_HAYAT_BEK (Benim hayat şartlarım 5 yıl sonra daha iyi olacak.)	2,6441	
X_CATISMA (Hangi konularda ailenizle hangi sıklıkta çatışma yaşarsınız?)	1,4215	
X_URUNTERCIH (Bir ürünü satın alırken, hangileri sizin için çok önemlidir?)	0,4727	
X66_1_GELIR (Aylık geliriniz ne kadardır?)	0,2566	

CART-1 modeli sonucunda başlangıç veri seti ile oluşan ağacın doğru sınıflama yüzdesi aşağıdaki tabloda verilmiştir.

Tablo 3.52. Başlangıç Verisi Doğru Sınıflandırma Yüzdesi (CART-1)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	393	97	490	80,2
B	42	193	235	82,13
Genel Sınıflandırma Oranı	435	290	725	80,83

Tablo 3.52'deki 725 gözlemden oluşan başlangıç verisi sonuçlarına göre, A partisini tercih ettiği bilinen 490 katılımcının 393'ü doğru olarak sınıflandırılmış ve 97'si de, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 235 katılımcının 193'ü doğru sınıflandırılmış. 42'si ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre başlangıç verisinde A partisinin doğru sınıflandırma yüzdesi %80,2, B partisinin doğru sınıflandırma yüzdesi %82,1'dir. Toplamda 725 katılımcının

586’sı doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %80,83 olarak bulunmuştur.

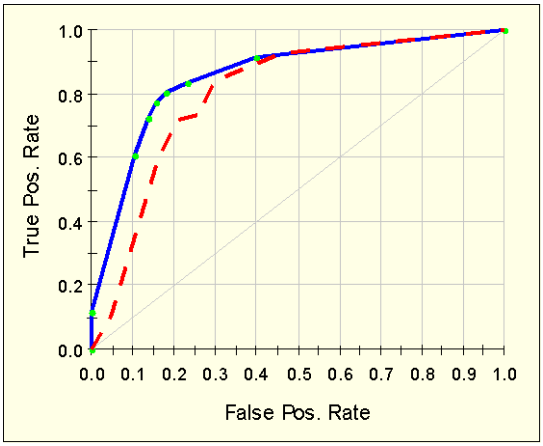
Test verisi ile oluşan ağacın doğru sınıflama yüzdesi de aşağıdaki tabloda verilmiştir.

Tablo 3.53. Test Verisi Doğru Sınıflandırma Tablosu (CART-1)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	161	45	206	78,2
B	24	63	87	72,4
Genel Sınıflandırma Oranı	185	108	293	76,5

Tablo 3.53’teki 293 gözlemden oluşan test verisi sonuçlarına göre, A partisini tercih ettiği bilinen 206 katılımcının 161’i doğru olarak sınıflandırılmış. 45’i ise, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 87 katılımcının 63’ü doğru sınıflandırılmış, 24’ü ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre test verisinde A partisinin doğru sınıflandırma yüzdesi %78,2 ve B partisinin doğru sınıflandırma yüzdesi %72,4’tür. Toplamda 293 katılımcının 224’ü doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %76,5 olarak bulunmuştur.

CART-1 modelinde başlangıç ve test verisine ilişkin ROC eğrileri ise, aşağıdaki şekilde gösterilmiştir.



Şekil 3.46: CART-1 Modeline Ait ROC Eğrisi Sonuçları

CART-1 modelinde başlangıç ve test verisine ilişkin AUC değerleri ise başlangıç veri seti için %85,6 olup mavi çizginin altında kalan alandır. Test verisi için %80,5 olup kırmızı çizginin altında kalan alandır. Bu oran, hem başlangıç hem test verisinde %80 ve üzerinde olduğu için modelin hem başlangıç hem de test verisine ait ayırım gücünün yani duyarlılık (doğru pozitif oranı) ve özgüllüğünün (doğru negatif oranı) oldukça iyi olduğu söylenebilir.

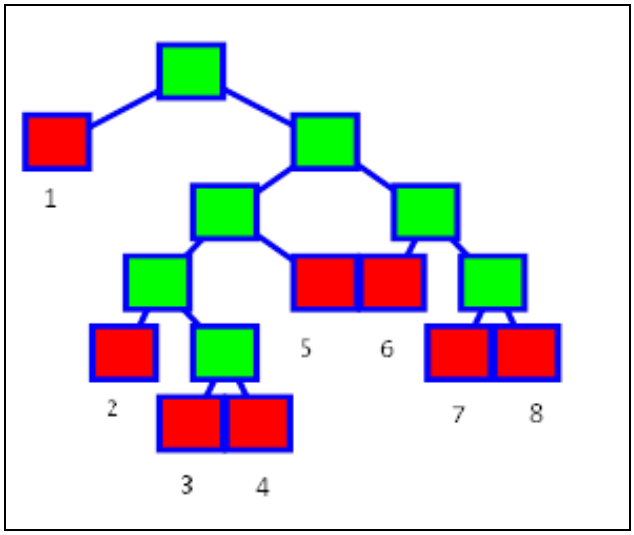
CART-1 modelinde uç düğümlere yapılan atama kuralları aşağıdaki tabloda gösterilmiştir.

Tablo 3.54. Uç Düğüme Atama Kuralları (CART-1)

Uç Düğüm	Kural	Sınıf	Olasılığı (Oranı)
1	**Eğer Kurum Bağı(x33) =** Meclis veya Yerel Yönetim veya Polis veya Diğer	A	0,91
2	**Eğer Kurum Bağı(x33)=** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık (x63)=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak(x13)=** Ne doğru ne yanlış veya Doğru veya Kesinlikle doğru **& Okunulan Gazete(x60)=** Zaman vaya Sabah veya Fanatik veya Fotomaç veya Radikal veya Diğer Gazeteler	A	1
3	**Eğer Kurum Bağı(x33)=** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak(x13)=** Ne doğru ne yanlış veya Doğru veya Kesinlikle doğru **& Okunulan Gazete(x60)=** Posta veya Hürriyet veya Haber Türk veya Milliyet veya Takvim veya Yok-Okumuyor **& İlgi Alanı (x34)=** Politika veya Sanatla İlgilenmek veya Farklı Kültürleri Tanımak vaya Bilimsel Gelişmeleri Takip	A	0,88
4	**Eğer Kurum Bağı(x33)=** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak(x13)=** Ne doğru ne yanlış veya Doğru veya Kesinlikle doğru **& Okunulan Gazete(x60)=** Posta veya Hürriyet veya Haber Türk veya Milliyet veya Takvim veya Yok-Okumuyor **& İlgi Alanı (x34)=** Spor veya Uluslararası gelişmeleri takip	B	0,55
5	**Eğer Kurum Bağı(x33)=** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık (x63)=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak(x13)=** Kesinlikle Yanlış veya Yanlış	B	0,51
6	**Eğer Kurum Bağı(x33)=** Yargı veya Ordu veya Medya vaya Hiçbiri **& Dindarlık (x63)=** İnançsız veya İnançlı **& Doğum Yeri Bölge(x5_2)=** Orta Anadolu veya Kuzeydoğu Anadolu veya Güneydoğu Anadolu	B	0,81
7	**Eğer Kurum Bağı(x33)=** Yargı veya Ordu veya Medya veya Hiçbiri **&Dindarlık (x63)=** İnançsız veya İnançlı **& Doğum Yeri Bölge(x5_2)=** İstanbul veya Batı Marmara veya Ege veya Doğu Marmara veya Batı Anadolu veya Akdeniz veya Batı Karadeniz veya Doğu Karadeniz veya Ortadoğu Anadolu veya Yurtdışı **& Evlilik (x30)=** Ün-Şöhret veya İnanç	A	0,75
8	**Eğer Kurum Bağı(x33)=** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık (x63)=** İnançsız veya İnançlı **& Doğum Yeri Bölge(x5_2)=** İstanbul veya Batı Marmara veya Ege veya Doğu Marmara veya Batı Anadolu veya Akdeniz veya Batı Karadeniz veya Doğu Karadeniz veya Ortadoğu Anadolu veya Yurtdışı **& Evlilik (x30)=** Güzellik-yakışıklılık veya kariyer veya eğitim veya maddiyat veya güç veya ruh güzelliği	B	0,76

Tablo 3.54'te görüldüğü üzere, oluşan ağacın 8 adet uç düğümü bulunmaktadır. Bu atama kurallarına göre her uç düğüme ulaşılırken ilgili kurallar gerçekleşmektedir.

Bu kurallar neticesinde bir akış çizen ağacın görseli aşağıdaki gibidir. Şekilde de görüldüğü gibi kırmızı ile gösterilen düğümler uç düğümler olup, bu düğümlerde sınıflara atama yapılmıştır.



Şekil 3.47: CART-1 Modeline Ait Sınıflama Ağacı

3.3.2.2. CART İkinci Model Bulguları

CART ile kurulan ikinci modelde (CART-2) başlanıç veri setinin %50'si modeli oluşturmak, %50'si ise modeli test etmek için kullanılmıştır. Bu analize ait bulgular aşağıda verilmiştir.

Tablo 3.55. Başlangıç ve Test Verisi (CART-2)

Parti	Başlangıç	Yüzde (%)	Test	Yüzde (%)	Toplam	Yüzde (%)
A	348	67.70	348	69.05	696	68
B	166	32.30	156	30.95	322	32
Toplam	514	100.00	504	100.00	1018	100

Tabloda da görüldüğü üzere toplam 1018 gözlemin 696'sı (%68) A partisine oy vereceğini, 322'si (%32) ise B partisine oy vereceğini belirtmiştir. Sonuçta bu gözlemlerin 514'ü (%50,5) başlangıç veri seti olup, modeli oluştumak için

kullanılmıştır. 504'ü (%59,5) ise test verisi olup, kurulan modeli test etmek için kullanılmıştır.

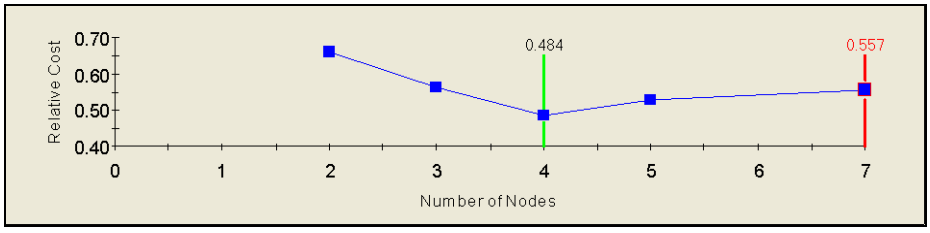
CART-2 modeli sonucunda, düğüm sayılarına göre oluşan 6 farklı ağacın maliyet-karmaşıklık tablosu aşağıda verilmiştir.

Tablo 3.56. Oluşan Ağaç Dizilerinin Maliyet-Karmaşıklık Tablosu-CART 2

Ağaç	Uç Düğümler	Test Seti Göreceli Maliyeti	Yeniden Yerine Koyma Maliyeti	Karmaşıklık
1	7	0.55703 ± 0.04355	0.38734	0.00000
2	5	0.52940 ± 0.04335	0.39011	0.00071
3**	4	0.48431 ± 0.04116	0.39281	0.00136
4	3	0.56477 ± 0.04395	0.48075	0.04398
5	2	0.65915 ± 0.03742	0.57599	0.04763
6	1	1.00000 ± 0.00000	1.00000	0.21201

Tablo 3.56'daki bilgilere göre oluşan ağaçlar arasından 3 numaralı ağaç optimum olarak gösterilmiştir. Fakat 4 adet uç düğümden oluşan, karmaşıklığı biraz fazla ve göreceli maliyeti en az olan bu ağaç yerine daha yüksek maliyet göze alınıp, karmaşıklığı en az olan 1 numaralı ağaç incelemeye alınmıştır. Bu ağaçta 7 adet uç düğüm bulunduğu için sınıflandırma aşamasında daha detaylı kurallarla, uç düğüme yani parti sınıflarına atamalar yapılmaktadır. Bu yüzden çalışmada parti tercihinde önemli olabilecek bağımsız değişkenlerin de modelde kalabilmesi amaçlanmıştır. Bu yüzden 7 adet düğümden oluşan bu ağaç incelemeye alınmıştır.

Bu durum aşağıdaki grafik ile daha iyi anlaşılmaktadır. Bu grafikte 7 düğümden oluşan ağacın göreceli maliyetinin 4 düğümden oluşan ağaçtan biraz daha yüksek olduğu görülmektedir.



Şekil 3.48: CART-2 Maliyet-Karmaşıklık Karşılaştırması

Buna göre ikinci model neticesinde anlamlı bulunan değişkenler, aşağıdaki tabloda gösterilmiştir.

Tablo 3.57. Değişkenlerin Önemlilik Oranları (CART-2)

Değişken	Skor	
X33_KURUM_BAGI (En çok hangi kuruma gönülden güvenirsiniz?)	100,0000	
X5_2DOGUM_BOLGE (Doğum yeriniz nedir? "bölge")	62,0241	
X6_2_BABA_DB (Babanızı doğum yeri nedir? "bölge")	50,5489	
X63_DINDARLIK (Dindarlık açısından kendinizi hangisiyle tarif edersiniz?)	47,5398	
X64_AILE_DIN (Dindarlık açısından ailenizin resisini hangisiyle tarif edersiniz?)	39,9332	
X13_TUR_HAYATBEK (Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak.)	37,4234	
X21_TURKIYE_ODOGU (Türkiye Orta Doğu ve Müslüman ülkelerle daha yakın işbirliği içinde olmalıdır.)	19,0855	
X9_1_IS (Son bir ay içinde bir işte çalıştınız mı? Çalıştınızsa işiniz?)	13,1125	
X8_MEDENI (Medeni durumunuz?)	11,8988	
X62_1_IDOL (İşte benim idolüm,yerinde olmak istediğim kişi dediğiniz kimdir?)	9,2266	
X12_HAYAT_BEK (Benim hayat şartlarım 5 yıl sonra daha iyi olacak.)	8,3102	
X3_1_EGITIM (Eğitim durumunuz, yani son bitridiğiniz okul nedir?)	5,1706	
X57_1_TV_KANAL (En çok izlediğiniz TV kanalı hangisidir?)	2,8692	
X_CATISMA (Hangi konularda ailenizle hangi sıklıkta çatışma yaşarsınız?)	1,7064	
X_URUNTERCIH (Bir ürünü satın alırken, hangileri sizin için çok önemlidir?)	1,3805	
X61_1_UYE (Dernek–vakıf- sosyal kulüp üyeliğiniz var mı? Hangi tür?)	0,0356	
X19_KIZ_EVLAT (Kızını dövmeyen dizini döver?)	0,0267	

CART-2 modeli sonucunda başlangıç veri seti ile oluşan ağacın doğru sınıflama yüzdesi aşağıdaki tabloda verilmiştir.

Tablo 3.58. Başlangıç Verisi Doğru Sınıflandırma Yüzdesi (CART-2)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	250	98	348	72
B	18	148	166	89
Genel Sınıflandırma Oranı	268	246	514	77,4

Tablo 3.58'deki 514 gözlemden oluşan başlangıç verisi sonuçlarına göre, A partisini tercih ettiği bilinen 348 katılımcının 250'si doğru olarak sınıflandırılmış, 98'i de yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 166 katılımcının 148'i doğru sınıflandırılmış, 18'i ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre başlangıç verisinde A partisinin doğru sınıflandırma yüzdesi %72, B partisinin doğru sınıflandırma yüzdesi ise, %89'dur. Toplamda 514 katılımcının 398'i doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %77,4 olarak bulunmuştur. Bu oranlara bakıldığında CART-1 modelinin CART-2 modelinden daha başarılı bir sınıflama yaptığı görülmektedir.

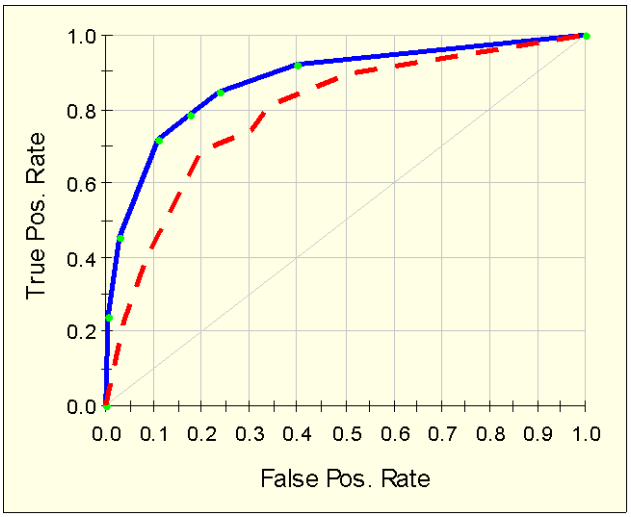
Test verisi ile oluşan ağacın doğru sınıflama yüzdesi de aşağıdaki tabloda verilmiştir.

Tablo 3.59. Test Verisi Doğru Sınıflandırma Tablosu (CART-2)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	239	109	348	68,7
B	31	125	156	80,1
Genel Sınıflandırma Oranı	270	234	504	72,2

Tablo 3.59'daki 504 gözlemden oluşan test verisi sonuçlarına göre, A partisini tercih ettiği bilinen 348 katılımcının 239'u doğru olarak sınıflandırılmış. 109'u ise, yanlış sınıf olan B partisi sınıfına atanmıştır. B partisini tercih ettiği bilinen 156 katılımcının 125'i doğru sınıflandırılmış, 31'i ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre test verisinde A partisinin doğru sınıflandırma yüzdesi %68,7 ve B partisinin doğru sınıflandırma yüzdesi %80,1'dir. Toplamda 504 katılımcının 364'ü doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %72,2 olarak bulunmuştur. Bu oranlara bakıldığında yine CART-1 modelinin CART-2 modelinden daha başarılı bir sınıflama yaptığı görülmektedir.

CART-2 modelinde başlangıç ve test verisine ilişkin ROC eğrileri ise, aşağıdaki şekilde gösterilmiştir.



Şekil 3.49: CART-2 Modeline Ait ROC Eğrisi Sonuçları

CART-2 modelinde başlangıç ve test verisine ilişkin AUC değerleri ise başlangıç veri seti için %87,5 olup mavi çizginin altında kalan alandır. Test verisi için %79,4 olup kırmızı çizginin altında kalan alandır. CART-2 modelinde başlangıç verisi için AUC değeri CART-1 modelinden daha yüksek, test verisinde ise daha düşük çıkmıştır. Buna göre, ikinci model başlangıç verisi için oldukça iyi bir ayırım gücüne sahip çıkmışken, test verisi için birinci modelden daha düşük bir ayırım gücüne sahip çıkmıştır.

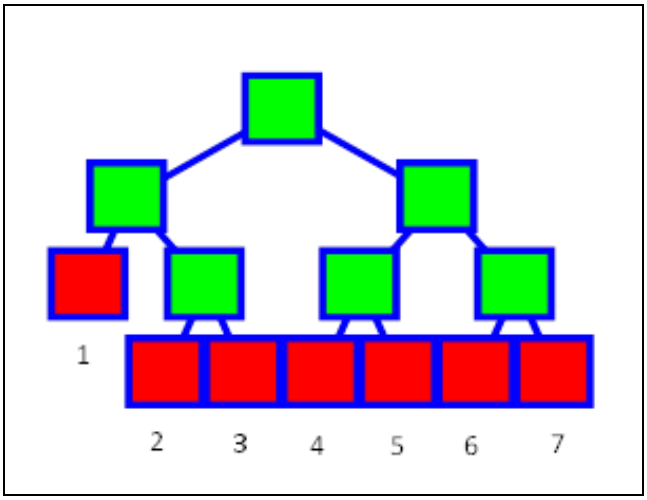
Sınıflama modeli oluşum aşamasında önemli olan kurulan modelin yeni veri seti ile yeniden sınanması neticesinde tekrar iyi bir sınıflama gücüne sahip olmasıdır. İkinci modelin test verisinde hem genel doğru sınıflandırma yüzdesi hem de AUC değeri daha düşük çıktığı için CART-2 modelinin CART-1 modelinden daha iyi bir sınıflama modeli olduğu söylenir.

CART-2 modelinde uç düğümlere yapılan atama kuralları aşağıdaki tabloda gösterilmiştir.

Tablo 3.60. Uç Düğüme Atama Kuralları (CART-2)

Uç Düğüm	Kural	Sınıf	Olasılığı (Oranı)
1	**Eğer Kurum Bağı(x33) =** Meclis veya Yerel Yönetim veya Polis veya Diğer **& İdolü=** Ailesi veya Siyasetçi veya Siyasetçi bir isim veya Sporcu – Futbolcu veya Atatürk veya İşadamı/kadını veya Oyuncu veya Dini lider/kişi veya Yazar veya Sanatçı veya Kendim	A	0,99
2	**Eğer Kurum Bağı(x33) =** Meclis veya Yerel Yönetim veya Polis veya Diğer **& İdolü=** Yakını veya Müzisyen veya TV ünlüsü / oyuncu veya Diğer veya Kimse **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak (x13)=** Ne doğru ne yanlış veya Doğru	A	0,95
3	**Eğer Kurum Bağı(x33) =** Meclis veya Yerel Yönetim veya Polis veya Diğer **& İdolü=** Yakını veya Müzisyen veya TV ünlüsü / oyuncu veya Diğer veya Kimse **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak (x13)=** Kesinlikle yanlış veya Yanlış veya Kesinlikle doğru	B	0,67
4	**Eğer Kurum Bağı(x33) =** Yargı veya Ordu veya Medya veya Hiçbiri**& Dindarlık=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak (x13)=** Ne doğru ne yanlış veya Doğru veya Kesinlikle doğru	A	0,87
5	**Eğer Kurum Bağı(x33) =** Yargı veya Ordu veya Medya veya Hiçbiri**& Dindarlık=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak (x13)=** Kesinlikle yanlış veya Yanlış	B	0,51
6	**Eğer Kurum Bağı(x33) =** Yargı veya Ordu veya Medya veya Hiçbiri**& Dindarlık=** İnançsız veya İnançlı **& Doğum Yeri Bölge(x5_2)=** Orta Anadolu veya Batı Karadeniz veya Kuzeydoğu Anadolu veya Güneydoğu Anadolu	A	0,68
7	**Eğer Kurum Bağı(x33) =** Yargı veya Ordu veya Medya veya Hiçbiri**& Dindarlık=** İnançsız veya İnançlı **& Doğum Yeri Bölge(x5_2)=** İstanbul veya Batı Marmara veya Ege veya Doğu Marmara veya Batı Anadolu veya Akdeniz veya Batı Karadeniz veya Doğu Karadeniz veya Ortadoğu Anadolu veya Yurtdışı	B	0,78

Tablo 3.60'da görüldüğü üzere, oluşan ağacın 7 adet uç düğümü bulunmaktadır. Bu atama kurallarına göre her uç düğüme ulaşılırken ilgili kurallar gerçekleşmektedir. Bu kurallar neticesinde bir akış çizen ağacın görseli aşağıdaki gibidir. Şekilde de görüldüğü gibi kırmızı ile gösterilen düğümler uç düğümler olup, bu düğümlerde sınıflara atama yapılmıştır.



Şekil 3.50: CART-2 Modeline Ait Sınıflama Ağacı

3.3.2.3. CART Üçüncü Model Bulguları

CART ile kurulan üçüncü modelde (CART-3) başlangıç veri setinin %30'u modeli oluşturmak %70'i ise modeli test etmek için kullanılmıştır. Veri setine ait sonuçlar aşağıdaki tabloda gösterilmiştir.

Tablo 3.61. Başlangıç ve Test Verisi (CART-3)

Parti	Başlangıç	Yüzde (%)	Test	Yüzde (%)	Toplam	Yüzde (%)
A	208	66,2	488	69,3	696	68
B	106	33,8	216	30,7	322	32
Toplam	314	100	704	100	1018	100

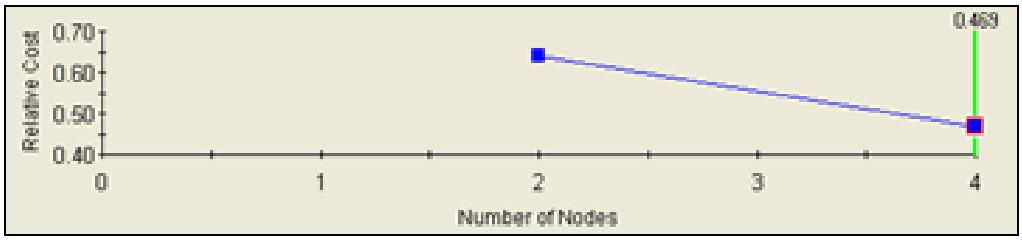
Tabloda da görüldüğü üzere toplam 1018 gözlemin 696'sı (%68) A partisine oy vereceğini, 322'si (%32) ise B partisine oy vereceğini belirtmiştir. Sonuçta bu gözlemlerin 314'ü (%30,7) başlangıç veri seti olup, modeli oluştumak için kullanılmıştır. 704'ü (%69,3) ise test verisi olup, kurulan modeli test etmek için kullanılmıştır.

CART-3 modeli sonucunda sadece bir ağaç oluşturulmuştur. Bu ağacın maliyet-karmaşıklık tablosu aşağıda verilmiştir.

Tablo 3.62. Oluşan Ağaç Dizilerinin Maliyet-Karmaşıklık Tablosu (CART-3)

Ağaç	Uç Düğümler	Test Seti Göreceli Maliyeti	Yeniden Yerine Koyma Maliyeti	Karmaşıklık
1	4	0.46866	-	0.00000

Sadece bir ağaç oluştuğu için yeniden yerine koyma maliyeti söz konusu değildir. Oluşan bu ağaçta 4 adet düğüm bulunmaktadır. Bundan daha büyük başka ağaç oluşmadığı için bu ağacın karmaşıklığı en azdır. Bu durum aşağıdaki grafik ile daha iyi anlaşılmaktadır. Bu grafikte sadece 4 adet düğümden oluşan bir adet optimum ağaç olduğu görülmektedir.



Şekil 3.51: Maliyet-Karmaşıklık Karşılaştırması (CART-3)

CART yönteminde bilindiği üzere modele giren tüm değişkenler arasından en önemliler belirlenerek model oluşturulur. Önemsiz değişkenler modele katılmaz. Buna göre üçüncü model neticesinde anlamlı bulunan değişkenler, aşağıdaki tabloda gösterilmiştir.

Tablo 3.63. Değişkenlerin Önemlilik Oranları (CART-3)

Değişken	Skor	
X33_KURUM_BAGI (En çok hangi kuruma gönülden güvenirsiniz?)	100,0000	
X5_2DOGUM_BOLGE (Doğum yeriniz nedir? "bölge")	65,2009	
X6_2_BABA_DB (Babanızı doğum yeri- nedir? "bölge")	59,8864	
X63_DINDARLIK (Dindarlık açısından kendinizi hangisiyle tarif edersiniz?)	48,4539	
X64_AILE_DIN (Dindarlık açısından ailenizin resisini hangisiyle tarif edersiniz?)	41,0089	
X30_EVLILIK (Evleneceğiniz insanda hangi özelliğin mutlaka olmasını istersiniz?	37,3871	
X13_TUR_HAYATBEK (Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak.)	32,3606	

X19_KIZ_EVLAT (Kızını dövmeyen dizini döver)	20,6923	
X5_1_DOGUM_MERKEZ (Doğduğunuz yer il mi ilçe mi köy müydü?)	16,0350	
X21_TURKIYE_ODOGU (Türkiye Orta Doğu ve Müslüman ülkelerle daha yakın işbirliği içinde olmalıdır.)	14,1238	
X18_OGTMN (Öğretmenin dövdüğü yerde gül biter.)	7,4143	
X2_YAS (Kaç yaşındasınız?)	2,4101	
X3_1_EGITIM (Eğitim durumunuz, yani son bitridiğiniz okul nedir?)	2,3355	

CART-3 modeli sonucunda başlangıç veri seti ile oluşan ağacın doğru sınıflama yüzdesi aşağıdaki tabloda verilmiştir.

Tablo 3.64. Başlangıç Verisi Doğru Sınıflandırma Yüzdesi (CART-3)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	162	46	208	77,8
B	16	90	106	85
Genel Sınıflandırma Oranı	178	136	314	80,3

Tablo 3.64'teki 314 gözlemden oluşan başlangıç verisi sonuçlarına göre, A partisini tercih ettiği bilinen 208 katılımcının 162'si doğru olarak sınıflandırılmış ve 46'sı da, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 106 katılımcının 90'ı doğru sınıflandırılmış. 16'sı ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre başlangıç verisinde A partisinin doğru sınıflandırma yüzdesi %77,8 B partisinin doğru sınıflandırma yüzdesi %85'tir. Toplamda 314 katılımcının 252'si doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %80,3 olarak bulunmuştur. Bu oranlara bakıldığında genel olarak CART-1 daha başarılı bir sınıflama yaptığı görülmektedir.

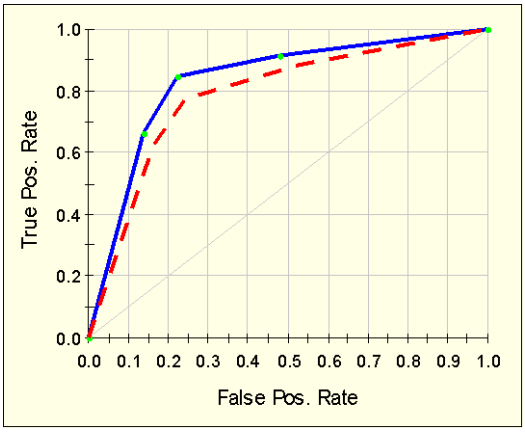
Test verisi ile oluşan ağacın doğru sınıflama yüzdesi de aşağıdaki tabloda verilmiştir.

Tablo 3.65. Test Verisi Doğru Sınıflandırma Tablosu (CART-3)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	370	118	488	75,8
B	49	167	216	77,3
Genel Sınıflandırma Oranı	419	285	704	76,3

Tablo 3.65'teki 704 gözlemden oluşan test verisi sonuçlarına göre, A partisini tercih ettiği bilinen 488 katılımcının 370'i doğru olarak sınıflandırılmış. 118'i ise, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 216 katılımcının 167'si doğru sınıflandırılmış, 49'u ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre test verisinde A partisinin doğru sınıflandırma yüzdesi %75,8 ve B partisinin doğru sınıflandırma yüzdesi %77,3'tür. Toplamda 704 katılımcının 537'si doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %76,3 olarak bulunmuştur. Bu oranlara bakıldığında yine CART-1 modelinin daha başarılı bir sınıflama yaptığı görülmektedir.

CART-3 modelinde başlangıç ve test verisine ilişkin ROC eğrileri ise, aşağıdaki şekilde gösterilmiştir.



Şekil 3.52: CART-3 Modeline Ait ROC Eğrisi Sonuçları

CART-3 modelinde başlangıç ve test verisine ilişkin AUC değerleri ise başlangıç veri seti için %83,3 olup mavi çizginin altında kalan alandır. Test verisi için %78,8 olup kırmızı çizginin altında kalan alandır.

Sınıflama modeli oluşum aşamasında önemli olan kurulan modelin yeni veri seti ile yeniden sınanması neticesinde tekrar iyi bir sınıflama gücüne sahip olmasıdır. CART-3 modelinin test verisinde AUC diğer iki modelden daha düşük çıkmıştır. Test verisine ait genel doğru sınıflandırma yüzdesinde ise CART-2 modeli en az çıkmıştır. Genel olarak bakıldığında ise, CART-1 modelinin en başarılı model olduğu görülmüştür.

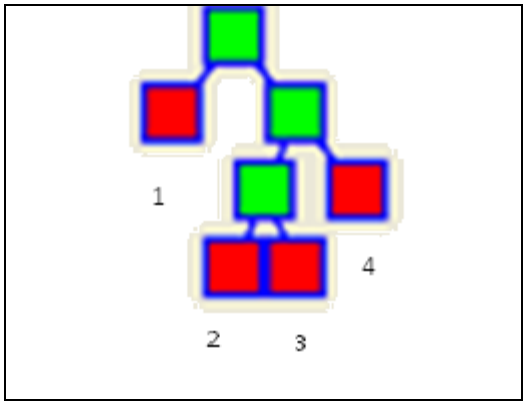
CART-3 modelinde uç düğümlere yapılan atama kuralları aşağıdaki tabloda gösterilmiştir.

Tablo 3.66. Uç Düğüme Atama Kuralları (CART-3)

Uç Düğüm	Kural	Sınıf	Olasılığı (Oranı)
1	**Eğer Kurum Bağı(x33) =** Meclis veya Yerel Yönetim veya Polis veya Diğer	A	0,92
2	**Eğer Kurum Bağı(x33) =** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak (x13)=** Ne doğru ne yanlış veya Doğru veya Kesinlikle doğru	A	0,88
3	**Eğer Kurum Bağı(x33) =** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık=** Dindar veya Sofu **& Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak (x13)=** Kesinlikle yanlış veya Yanlış	B	0,54
4	**Eğer Kurum Bağı(x33) =** Yargı veya Ordu veya Medya veya Hiçbiri **& Dindarlık=** İnançsız veya İnançlı	B	0,71

Tablo 3.66'da görüldüğü üzere, oluşan ağacın 4 adet uç düğümü bulunmaktadır. Bu atama kurallarına göre her uç düğüme ulaşılırken ilgili kurallar gerçekleşmektedir. Bu kurallar neticesinde bir akış çizen ağacın görseli aşağıdaki

gibidir. Şekilde de görüldüğü gibi kırmızı ile gösterilen düğümler uç düğümler olup, bu düğümlerde sınıflara atama yapılmıştır.



Şekil 3.53: CART-3 Modeline Ait Sınıflama Ağacı

3.3.2.4. CART Modellerinin Karşılaştırılması

Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde en çok tercih edilen A ve B partisi için ilk aşamada iki kategori düzeyinde CART modelleri oluşturulmuştur. Bu aşamada farklı büyüklükteki başlangıç ve test verisi kullanılmıştır. CART modellerine ait genel sonuçlar aşağıdaki tabloda gösterilmiştir.

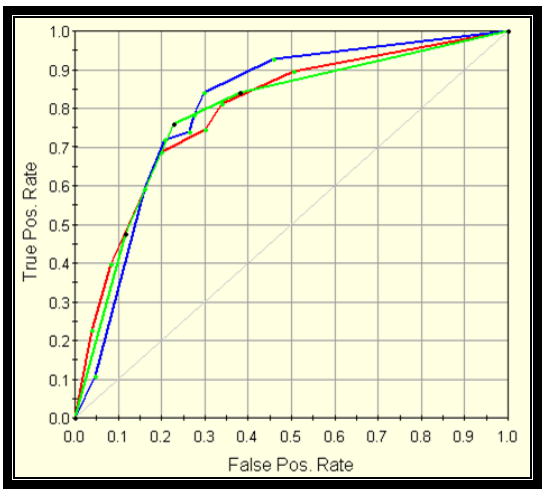
Tablo 3.67. CART (2 düzey) Modellerinin Karşılaştırılması

MODEL	Başlangıç Verisi	Test Verisi	AUC (Başlangıç Verisi)	AUC (Test Verisi)	Genel Doğru Sınıflama Oranı (Başlangıç)	Genel Doğru Sınıflama Oranı (Test)
CART-1	70%	30%	85,6	80,5	80,8	76,5
CART-2	50%	50%	87,5	79,4	77,4	72,2
CART-3	30%	70%	83,3	78,8	80,3	76,3

Sınıflama modeli oluşum aşamasında önemli olan kurulan modelin yeni veri seti ile yeniden sınanması neticesinde tekrar iyi bir sınıflama gücüne sahip olmasıdır. Bu sonuçlar neticesinde CART-1 modelinin test verisine ait AUC değeri ve genel doğru

sınıflandırma oranı en yüksek çıktığı için diğer modellerden daha yüksek bir ayırım gücüne sahip olduğu görülmektedir.

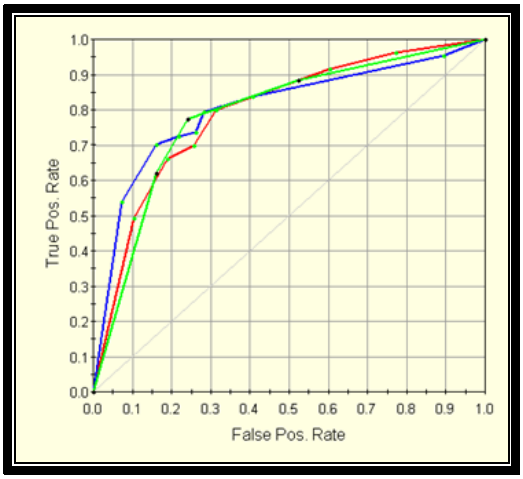
Ayrıca oluşturulan CART modellerinin test verilerine ilişkin ROC eğrileri aşağıdaki şekil ile karşılaştırılabilir. Bu kez modelin genel AUC değerleri yerine sırasıyla A ve B partisine ilişkin karşılaştırmalı ROC eğrileri aşağıda gösterilmiştir.



Şekil 3.54: A Partisine Ait ROC Eğrileri Karşılaştırması-CART

Oluşturulan CART modellerinden mavi ile gösterilen CART-1, kırmızı ile gösterilen CART-2, yeşil ile gösterilen CART-3 olup, A partisi için en iyi ROC değerinin CART-1 yani %70 başlangıç veri seti %30 test verisi ile oluşturulan model olduğu görülmüştür.

Yine aynı şekilde B partisine ilişkin karşılaştırmalı ROC eğrileri de aşağıda gösterilmiştir.



Şekil 3.55: B Partisine Ait ROC Eğrileri Karşılaştırması-CART

Oluşturulan CART modellerinden mavi ile gösterilen CART-1, kırmızı ile gösterilen CART-2, yeşil ile gösterilen CART-3 olup, B partisi için en iyi ROC değerinin CART-2 yani %50 başlangıç veri seti %50 test verisi ile oluşturulan model olduğu görülmüştür.

3.3.2.5. Farklı Örneklem Büyüklüğü ile CART Modellerinin Değerlendirmesi

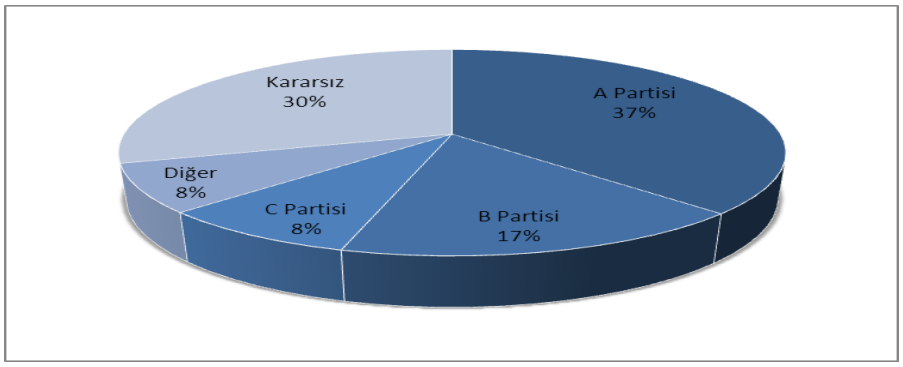
Daha önce belirtildiği gibi çalışmanın ikinci aşamasında en uygun olan başlangıç ve test verisi büyüklüğü ile bu kez genel bir değerlendirme yapılmıştır. İlk aşamada, İlk üç CART modeline ait sonuçlar incelendiğinde en uygun modelin veri setinin %70'inin başlangıç veri seti yani eğitim verisi olarak alınıp, geri kalan %30'unun test verisi olarak alınması gerektiği yönünde olmuştur. Veri madenciliği ile ilgili yapılan çalışmalarda da genel olarak veri setininin 2/3'ünün yani %67'sinin modeli oluşturmak, 1/3'ünün yani %33'ünün ise modeli test etmek için ideal olabileceği yönünde olmuştur.

Bu değerlendirme sonucunda yeni veri setinin %70'i başlangıç, %30'u test verisi olarak alınmıştır. Bu veri setinde bağımlı değişken olan parti tercihi bu kez "A Partisi", "B Partisi", "C Partisi" "Diğer" ve "Kararsız" olmak üzere beş kategori düzeyi esas alınarak belirlenmiştir. Daha önce de belirtildiği gibi bu yeni veri seti 1867

gözlemden oluşmaktadır. Bu yeni veri setinde gözlemlerin parti tercihine göre dağılımı aşağıdaki tabloda gösterilmiştir.

Tablo 3.68. Katılımcıların "Önümüzdeki seçimlerde hangi partiye (5 düzey) oy vereceksiniz?" Sorunun Cevabına Göre Dağılımı

	Frekans	Yüzde (%)
A Partisi	696	37,3
B Partisi	322	17,2
C Partisi	156	8,4
Diğer	142	7,6
Kararsız	551	29,5
Toplam	1867	100



Şekil 3.56: Katılımcıların "Önümüzdeki seçimlerde hangi partiye oy vereceksiniz?" Sorunun Cevabına Göre Dağılımı

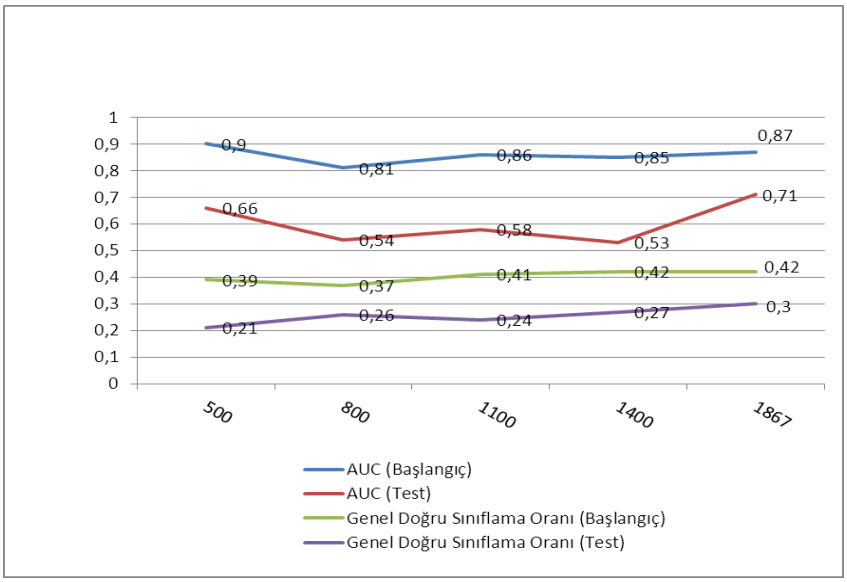
Gençlerin %37,3'lük büyük çoğunlu önümüzdeki seçimlerde A partisine oy vereceğini belirtirken, %17,2'si B partisine ve %8,4'ü ise, C partisine oy vereceğini belirtmiştir. Çalışmaya katılan gençlerin %7,6'sı diğer partilere oy vereceğini belirtirken %29,5'i ise, kararsız olduğunu belirtmişitr.

Böylelikle yeni veri seti, farklı büyüklükteki örnek sayılarına göre yeniden CART yöntemi ile modellenmiştir. Buna göre toplam 1867 gözlemden oluşan örneklemden, excelde rassal sayılar üretilerek, raslantısal olarak sırasıyla 500, 800, 1100 ve 1400 birimlik örneklemler seçilmiştir. Bu farklı örneklemler ile yapılan CART analizlerin karşılaştırılmasına ait bulgular aşağıda verilmiştir.

Tablo 3.69. Farklı Büyüklükteki Örneklemlere Göre CART (5 düzey) Modellerinin Karşılaştırılması

n	AUC (Başlangıç)	AUC (Test)	Genel Doğru Sınıflama Oranı (Başlangıç)	Genel Doğru Sınıflama Oranı (Test)
500	0,90	0,66	0,39	0,21
800	0,81	0,54	0,37	0,26
1100	0,86	0,58	0,41	0,24
1400	0,85	0,53	0,42	0,27
1867	0,87	0,71	0,42	0,30

Farklı büyüklükteki örneklemler ile 5 düzeyi bulunan parti değişkeninin sonuçları incelendiğinde örneklem büyüklüğünün artışına bağlı olarak genel olarak test verisine ait AUC değeri ve genel doğru sınıflandırma oranının da artış gösterdiği görülmektedir. Yalnız 500 birimden oluşan örneklemin test verisine ait AUC değeri, 800, 1100 ve 1400 birimden oluşan örneklemlerden yüksek çıktığı ama genel sınıflama oranının düşük çıktığı görülmektedir. Buradan genel olarak örneklem büyüklüğü arttıkça modelin test verisinin ayırım gücü ve genel doğru sınıflama oranının arttığı sonucuna varılmaktadır. Bu durum aşağıdaki grafik ile görsel olarak da görülmektedir.



Şekil 3.57: Farklı Büyüklükteki Örneklemlere Göre CART(5 düzey) Modellerinin Karşılaştırılması

3.3.3. MARS Yöntemine Ait Bulgular

Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde CART yönteminden sonra bu kez MARS yöntemi farklı büyüklükteki başlangıç ve test verisi kullanılarak incelenmiştir. MARS yöntemi ikili bağımlı değişkenler için tasarlanmış bir model olduğu için ve MARS ile elde edilen analiz sonuçlarının CART ile kıyaslanabilmesi için ilk aşamada bağımlı değişkenin kategori düzeyi ikiye indirgenmiş ve en çok tercih edilen ilk iki parti olan "A Partisi" ve "B Partisi" ele alınmıştır.

Bu amaçla CART yönteminde olduğu gibi MARS yönteminde de ilk aşamada başlangıçta modeli oluştumak için veri setinin sırasıyla %70'i, %50'si ve %30'u başlangıç veri seti bir diğer adıyla eğitim verisi olarak ele alınmıştır. Buna göre veri setinin sırasıyla geri kalan %30, %50 ve %70'lik kısmı ise oluşan modeli test etmek için kullanılmıştır. Bu durumda MARS yöntemi ile 3 adet model elde edilmiştir. Bu modellere ait bulgular izleyen kısımlarda sırasıyla verilmiştir.

3.3.3.1. MARS Birinci Model Bulguları

Çalışmanın ikinci uygulama aşamasında ise, MARS modeli ile çalışılmıştır. Benzer şekilde MARS ile kurulan ilk modelde (MARS-1) başlanıç veri setinin %70'i modeli oluşturmak %30'u ise modeli test etmek için kullanılmıştır. Bu analize ait bulgular aşağıda verilmiştir.

Tablo 3.70. Başlangıç ve Test Verisi (MARS-1)

Parti	Başlangıç	Yüzde (%)	Test	Yüzde (%)	Toplam	Yüzde (%)
A	490	67.59	206	70.31	696	68
B	235	32.41	87	29.69	322	32
Toplam	725	100.00	293	100.00	1018	100

Tabloda da görüldüğü üzere toplam 1018 gözlemin 696'sı (%68) A partisine oy vereceğini, 322'si (%32) ise B partisine oy vereceğini belirtmiştir. Sonuçta bu

gözlemlerin 725'i (%70) başlangıç veri seti olup, modeli oluştumak için kullanılmıştır. 293'ü (%30) ise test verisi olup, kurulan modeli test etmek için kullanılmıştır.

MARS yönteminde CART yönteminde olduğu gibi modele giren tüm değişkenler arasından en önemliler belirlenerek model oluşturulur. Önemsiz değişkenler modele katılmaz. Buna göre birinci model neticesinde anlamlı bulunan değişkenler, aşağıdaki tabloda gösterilmiştir.

Tablo 3.71. Değişkenlerin Önemlilik Oranları (MARS-1)

Değişken	Skor	
X13_TUR_HAYATBEK_mis (Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak.)	100,0	
X33_KURUM_BAGI (En çok hangi kuruma gönülden güvenirsiniz?)	80,57	
X64_AILE_DIN_mis (Dindarlık açısından ailenizin resisini hangisiyle tarif edersiniz?)	79,76	
X64_AILE_DIN (Dindarlık açısından ailenizin resisini hangisiyle tarif edersiniz?)	60,89	
X57_1_TV_KANAL_mis (En çok izlediğiniz TV kanalı)	55,97	
X57_1_TV_KANAL (En çok izlediğiniz TV kanalı)	55,97	
X13_TUR_HAYATBEK (Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak.)	55,63	
X9_1_IS_mis (Son bir ay içinde bir işte çalıştınız mı? Çalıştınızsa işiniz?)	51,21	
X9_1_IS (Son bir ay içinde bir işte çalıştınız mı? Çalıştınızsa işiniz?)	51,21	
X6_2_BABA_DB_mis (Babanızı doğum yeri nedir? "bölge")	50,37	
X6_2_BABA_DB (Babanızı doğum yeri nedir? "bölge")	50,37	
X21_TURKIYE_ODOGU (Türkiye Orta Doğu ve Müslüman ülkelerle daha yakın işbirliği içinde olmalıdır.)	46,67	
X21_TURKIYE_ODOGU_mis (Türkiye Orta Doğu ve Müslüman ülkelerle daha yakın işbirliği içinde olmalıdır.)	46,67	
X31_KADIN_ERKEK_mis (Bir kadın ile erkek aşağıdakilerden hangisi olur ise sizce beraber yaşayabilirler?)	32,82	
X31_KADIN_ERKEK (Bir kadın ile erkek aşağıdakilerden hangisi olur ise sizce beraber yaşayabilirler?)	32,82	

Modelin ilk oluşum aşaması olan ileriye doğru adım yönteminde başlangıçta varsayılan (default) değer olarak alınan 30 adet temel fonksiyonun ele alınması uygun görülmüştür. Fakat geriye doğru yönteminde sadece 8 tane temel fonksiyon anlamlı olarak bulunup, modele alınmıştır. MARS-1 modelinde modele giren değişkenler ile oluşturulan final model ve bu değişkenlere ait düğüm değerleri aşağıdaki tabloda görülmektedir. Tabloda da görüldüğü üzere, sıfırıncı temel fonksiyon sabit terim olup,

modele “3”,”5”, “9”, “13”, ”17”, “21”, “25” ve “29”. temel fonksiyonlar modele alınmıştır.

Tablo 3.72. Final Model (MARS-1)

Modele AlınanTemel Fonksiyonlar	Katsayılar	Değişken	İşareti	Ana Düğüm İşareti	Ana Düğüm	Düğüm Değeri (Kırılma Noktası)
0	0,1465					
3	0,2280	X64_AILE_DIN	+	-	X64_AILE_DIN_mis	"İnançlı", "İnançsız"
5	0,2692	X33_KURUM_BAGI	-			"Ordu", "Medya", "Hiçbiri"
9	0,2226	X21_TURKIYE_ODOGU	+	+	X21_TURKIYE_ODOGU_ mis	"Kesinlikle yanlış", "Yanlış
13	0,1936	X13_TUR_HAYATBEK	-	+	X13_TUR_HAYATBEK_ mis	"Kesinlikle yanlış", "Yanlış"
17	-0,2735	X9_1_IS	+	-	X9_1_IS_mis	"Öğrenci"
21	0,1684	X6_2_BABA_DB	-	-	X6_2_BABA_DB_mis	"Batı Marmara", "Ege", "Batı Anadolu", "Akdeniz", "Yurtdışı"
25	-0,2479	X57_1_TV_KANAL	+	+	X57_1_TV_KANAL_mis	"ATV", "FOX", "Samanyolu TV", "TRT", "NTV Spor", "Kanal 7"
29	-0,1497	X31_KADIN_ERKEK	-	+	X31_KADIN_ERKEK_mis	"Hem resmi hem de dini nikâh"

Buna göre MARS-1 analizi sonucunda Tablo 3.72’de gösterilen temel fonksiyonlardan yararlanılarak oluşturulan final tahmin modeli aşağıda gösterildiği gibidir. Tablo da görüldüğü üzere modelin açıklama yüzdesi (Tanımlayıcılık Katsayısı-R^2) 0,48 olup bağımsız değişkenlerin %48’i parti tercihini açıklyabilmektedir. Bu değerin düşük olması MARS-1 modelinin yeterince iyi bir tahmin modeli olmadığı gösterir.

Tablo 3.73. Final Modele Ait Temel Fonsiyonlar (MARS-1)

BF1 = (X64_AILE_DIN in (“Eksik değer”));
BF3 = (X64_AILE_DIN in ("İnançlı", "İnançsız")) * BF1;
BF5 = (X33_KURUM_BAGI in ("Ordu", "Medya", "Hiçbiri"));
BF7 = (X21_TURKIYE_ODOGU in (“Eksik değer”)) * BF1;
BF9 = (X21_TURKIYE_ODOGU in ("Kesinlikle yanlış","Yanlış)) * BF7;
BF11 = (X13_TUR_HAYATBEK in (“Eksik değer”));
BF13 = (X13_TUR_HAYATBEK in ("Kesinlikle yanlış","Yanlış)) * BF11;
BF15 = (X9_1_IS in (“Eksik değer”)) * BF5;

BF17 = (X9_1_IS in ("Öğrenci")) * BF15; BF19 = (X6_2_BABA_DB in (“Eksik değer”)) * BF11; **BF21** = (X6_2_BABA_DB in ("Batı Marmara", "Ege","Batı Anadolu", "Akdeniz","Yurtdışı"))* BF19; BF23 = (X57_1_TV_KANAL in (“Eksik değer”)) * BF11; **BF25** = (X57_1_TV_KANAL in ("ATV","FOX","Samanyolu TV","TRT", "NTV Spor","Kanal 7") * BF23; BF26 = (X57_1_TV_KANAL in ("Kanal D", "Show TV", "Star TV", "NTV", "Cnbc-e","Habertürk", "Müzik Kanalları", "Diğer kanallar")) * BF23; BF27 = (X31_KADIN_ERKEK in (“Eksik değer”)) * BF26; **BF29** = (X31_KADIN_ERKEK in ("Hem resmi hem de dini nikâh")) * BF27;
MARS-1 Y = 0.146523 + 0.228028 * BF3 + 0.269232 * BF5 + 0.22258 * BF9 + 0.193608 * BF13 - 0.273483 * BF17 + 0.168374 * BF21 - 0.247854 * BF25 - 0.14968 * BF29;
MARS-1 MODEL PARTI = **BF3 BF5 BF9 BF13 BF17 BF21 BF25 BF29;**
R-Kare: 0.48 Düzeltilmiş R-Kare: 0.47 F-İstatistiği: 72,39 MSE: 0,11 Final Modelin GCV Değeri : 0.12

MARS-1 modeli sonucunda başlangıç veri seti ile oluşan ağacın doğru sınıflama yüzdesi aşağıdaki tabloda verilmiştir.

Tablo 3.74. Başlangıç Verisi Doğru Sınıflandırma Yüzdesi (MARS-1)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	346	144	490	70,6
B	20	215	235	91,5
Genel Sınıflandırma Oranı	366	359	725	77,4

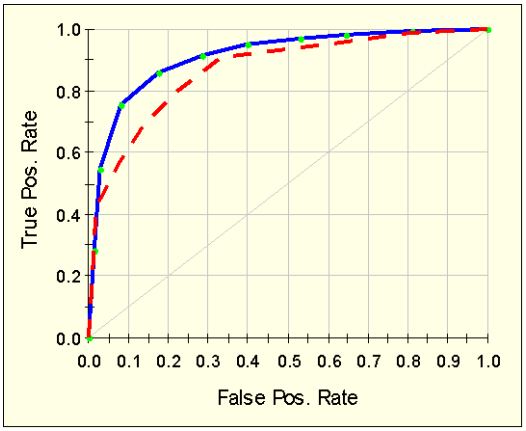
Tablo 3.74’teki 725 gözlemden oluşan başlangıç verisi sonuçlarına göre, A partisini tercih ettiği bilinen 490 katılımcının 346’sı doğru olarak sınıflandırılmış ve 144’ü ise, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 235 katılımcının 215’i doğru sınıflandırılmış, 20’si ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre başlangıç verisinde A partisinin doğru sınıflandırma yüzdesi %70,6 B partisinin doğru sınıflandırma yüzdesi %91,5’tir. Toplamda 725 katılımcının 561’i doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %77,4 olarak bulunmuştur. Test verisi ile oluşan ağacın doğru sınıflama yüzdesi de aşağıdaki tabloda verilmiştir.

Tablo 3.75. Test Verisi Doğru Sınıflandırma Tablosu (MARS-1)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	141	65	206	68,5
B	8	79	87	90,8
Genel Sınıflandırma Oranı	149	144	293	75,1

Tablo 3.75'teki 293 gözlemden oluşan test verisi sonuçlarına göre, A partisini tercih ettiği bilinen 206 katılımcının 141'i doğru olarak sınıflandırılmış. 65'i ise, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 87 katılımcının 79'u doğru sınıflandırılmış, 8'i ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre test verisinde A partisinin doğru sınıflandırma yüzdesi %68,5 ve B partisinin doğru sınıflandırma yüzdesi %90,8'dir. Toplamda 293 katılımcının 220'si doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %75,1 olarak bulunmuştur.

MARS-1 modelinde başlangıç ve test verisine ilişkin ROC eğrileri ise, aşağıdaki şekilde gösterilmiştir.



Şekil 3.58: MARS-1 Modeline Ait ROC Eğrisi Sonuçları

MARS-1 modelinde başlangıç ve test verisine ilişkin AUC değerleri ise başlangıç veri seti için %91 olup mavi çizginin altında kalan alandır. Test verisi için %87,5 olup kırmızı çizginin altında kalan alandır. Bu oran, hem başlangıç hem test verisinde %80 ve üzerinde olduğu için modelin hem başlangıç hem de test verisine ait

ayırım gücünün yani duyarlılık (doğru pozitif oranı) ve özgüllüğünün (doğru negatif oranı) oldukça iyi olduğu söylenebilir.

3.3.3.2. MARS İkinci Model Bulguları

MARS ile kurulan ikinci modelde (MARS-2) başlanıç veri setinin %50'si modeli oluşturmak, %50'si ise modeli test etmek için kullanılmıştır. Bu analize ait bulgular aşağıda verilmiştir.

Tablo 3.76. Başlangıç ve Test Verisi (MARS-2)

Parti	Başlangıç	Yüzde (%)	Test	Yüzde (%)	Toplam	Yüzde (%)
A	348	67.70	348	69.05	696	68
B	166	32.30	156	30.95	322	32
Toplam	514	100.00	504	100.00	1018	100

Tabloda da görüldüğü üzere toplam 1018 gözlemin 696'sı (%68) A partisine oy vereceğini, 322'si (%32) ise B partisine oy vereceğini belirtmiştir. Sonuçta bu gözlemlerin 514'ü (%50,5) başlangıç veri seti olup, modeli oluştumak için kullanılmıştır. 504'ü (%59,5) ise test verisi olup, kurulan modeli test etmek için kullanılmıştır.

MARS yönteminde CART yönteminde olduğu gibi modele giren tüm değişkenler arasından en önemliler belirlenerek model oluşturulur. Önemsiz değişkenler modele katılmaz. Buna göre ikinci model neticesinde anlamlı bulunan değişkenler, aşağıdaki tabloda gösterilmiştir.

Tablo 3.77. Değişkenlerin Önemlilik Oranları (MARS-2)

Değişken	Skor	
X63_DINDARLIK_mis (Dindarlık açısından kendinizi hangisiyle tarif edersiniz?)	100,00	
X33_KURUM_BAGI (En çok hangi kuruma gönülden güvenirsiniz?)	100,00	
X5_2DOGUM_BOLGE_mis (Doğum yeriniz nedir? "bölge")	53,31	
X63_DINDARLIK (Dindarlık açısından kendinizi hangisiyle tarif edersiniz?)	40,67	
X6_2_BABA_DB_mis (Babanızı doğum yeri nedir? "bölge")	39,38	
X6_2_BABA_DB (Babanızı doğum yeri nedir? "bölge")	39,38	
X13_TUR_HAYATBEK_mis (Türkiye'deki hayat şartları 5 yıl sonra daha	38,08	

iyi olacak.)		
X13_TUR_HAYATBEK (Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak.)	38,08	IIIIIIIIIIIIII
X60_GAZETE_mis (Okuduğunuz gazaete hangisidir?)	26,18	IIIIIIIII
X60_GAZETE (Okuduğunuz gazaete hangisidir?)	26,18	IIIIIIIII
X31_KADIN_ERKEK_mis (Bir kadın ile erkek aşağıdakilerden hangisi olur ise sizce beraber yaşayabilirler?)	23,66	IIIIIIII
X31_KADIN_ERKEK (Bir kadın ile erkek aşağıdakilerden hangisi olur ise sizce beraber yaşayabilirler?)	23,66	IIIIIIII

Modelin ilk oluşum aşaması olan ileriye doğru adım yönteminde başlangıçta varsayılan (default) değer olarak alınan 30 adet temel fonksiyonun ele alınması uygun görülmüştür. Fakat geriye doğru yönteminde sadece 6 tane temel fonksiyon anlamlı olarak bulunup, modele alınmıştır. MARS-2 modelinde modele giren değişkenler ile oluşturulan final model ve bu değişkenlere ait düğüm değerleri aşağıdaki tabloda görülmektedir. Tabloda da görüldüğü üzere, sıfırıncı temel fonksiyon sabit terim olup, modele "5","9", "13", "17", "21", ve "27". temel fonksiyonlar modele alınmıştır.

Tablo 3.78. Final Model (MARS-2)

Modele AlınanTemel Fonksiyonlar	Katsayılar	Değişken	İşareti	Ana Düğüm İşareti	Ana Düğüm	Düğüm Değeri (Kırılma noktası)
0	0,0230					
5	-0,2913	X63_DINDARLIK	+	-	X63_DINDARLIK_mis	"Dindar", "Sofu"
9	0,2280	X13_TUR_HAYATBEK	-	-	X13_TUR_HAYATBEK_ mis	"Kesinlikle yanlış", "Yanlış"
13	0,2787	X6_2_BABA_DB	+	+	X6_2_BABA_DB_mis	"Batı Marmara", "Ege", "Batı Anadolu", "Akdeniz", "Yurtdışı"
17	-0,1963	X31_KADIN_ERKEK	-	+	X31_KADIN_ERKEK_mis	"Dini nikâh", "Hem resmi hem de dini nikâh"
21	0,4568	X5_2DOGUM_BOLGE_m is	+	+	X63_DINDARLIK_mis	"Eksik değer"
27	0,2356	X60_GAZETE	-	+	X60_GAZETE_mis	"Hürriyet", "Milliyet", "Cumhuriyet", "Radikal"

Buna göre MARS-2 analizi sonucunda Tablo 3.78'de gösterilen temel fonksiyonlardan yararlanılarak oluşturulan final tahmin modeli aşağıda gösterildiği

gibidir. Tablo 3.79'da görüldüğü üzere modelin açıklama yüzdesi 0,49 olup, MARS-1 modelinden daha yüksek çıkmıştır. Yani bağımsız değişkenlerin %49'u parti tercihini açıklayabilmektedir. Bu değerin düşük olması MARS-2 modelinin de yeterince iyi bir tahmin modeli olmadığını göstermektedir.

Tablo 3.79. Final Modele Ait Temel Fonsiyonlar (MARS-2)

BF1 = (X33_KURUM_BAGI in ("Yargı", "Ordu", "Medya", "Hiçbiri")); BF3 = (X63_DINDARLIK in ("Eksik değer")) * BF1; **BF5** = (X63_DINDARLIK in ("Dindar", "Sofu")) * BF3; BF7 = (X13_TUR_HAYATBEK in ("Eksik değer")); **BF9** = (X13_TUR_HAYATBEK in ("Kesinlikle yanlış", "Yanlış")) * BF7; BF11 = (X6_2_BABA_DB in ("Eksik değer")) * BF3 **BF13** = (X6_2_BABA_DB in ("Batı Marmara", "Ege","Batı Anadolu", "Akdeniz","Yurtdışı")) * BF11; BF15 = (X31_KADIN_ERKEK in ("Eksik değer")) * BF3; **BF17** = (X31_KADIN_ERKEK in ("Hem resmi hem de dini nikâh", "Dini nikâh")) * BF15; **BF21** = (X5_2DOGUM_BOLGE in ("Eksik değer")) * BF3; BF25 = (X60_GAZETE in ("Eksik değer")) * BF3; **BF27** = (X60_GAZETE in ("Hürriyet", "Milliyet", "Cumhuriyet", "Radikal")) * BF25;
MARS-2 Y = 0.0230371 - 0.291316 * BF5 + 0.228038 * BF9 + 0.278698 * BF13 - 0.196266 * BF17 + 0.456767 * BF21 + 0.235632 * BF27;
MARS-2 MODEL PARTI = **BF5 BF9 BF13 BF17 BF21 BF27;**
R-Kare: 0,49 Düzeltilmiş R-Kare:0,49 F-İstatistiği: 78,06 MSE: 0,11 Final Modelin GCV Değeri : 0,12

MARS-2 modeli sonucunda başlangıç veri seti ile oluşan ağacın doğru sınıflama yüzdesi aşağıdaki tabloda verilmiştir.

Tablo 3.80. Başlangıç Verisi Doğru Sınıflandırma Yüzdesi (MARS-2)

Gerçek Durum	**Test Sonucu**			**Yüzde(%)**
	A	**B**	**Toplam**	
A	300	48	348	86,2
B	29	137	166	82,5
Genel Sınıflandırma Oranı	329	185	514	85

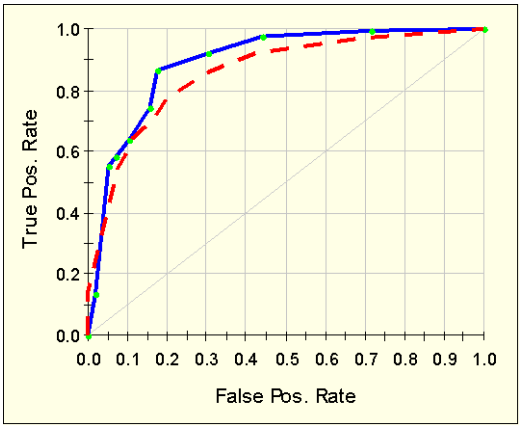
Tablo 3.80'deki 514 gözlemden oluşan başlangıç verisi sonuçlarına göre, A partisini tercih ettiği bilinen 348 katılımcının 300'ü doğru olarak sınıflandırılmış, 48'i de yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 166 katılımcının 137'si doğru sınıflandırılmış, 29'u ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre başlangıç verisinde A partisinin doğru sınıflandırma yüzdesi %86,2, B partisinin doğru sınıflandırma yüzdesi ise, %82,5'tir. Toplamda 514 katılımcının 437'si doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %85 olarak bulunmuştur. Bu oranlara bakıldığında MARS-2 modelinin MARS-1 modelinden daha başarılı bir sınıflama yaptığı görülmektedir. Test verisi ile oluşan ağacın doğru sınıflama yüzdesi de aşağıdaki tabloda verilmiştir.

Tablo 3.81. Test Verisi Doğru Sınıflandırma Tablosu (MARS-2)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	286	62	348	82,2
B	35	121	156	77,6
Genel Sınıflandırma Oranı	321	183	504	80,8

Tablo 3.81'deki 504 gözlemden oluşan test verisi sonuçlarına göre, A partisini tercih ettiği bilinen 348 katılımcının 286'sı doğru olarak sınıflandırılmış. 62'si ise, yanlış sınıf olan B partisi sınıfına atanmıştır. B partisini tercih ettiği bilinen 156 katılımcının 121'i doğru sınıflandırılmış, 35'i ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre test verisinde A partisinin doğru sınıflandırma yüzdesi %82,2 ve B partisinin doğru sınıflandırma yüzdesi %77,6'dir. Toplamda 504 katılımcının 407'si doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %80,8 olarak bulunmuştur. Bu oranlara bakıldığında yine MARS-2 modelinin MARS-1 modelinden daha başarılı bir sınıflama yaptığı görülmektedir.

MARS-2 modelinde başlangıç ve test verisine ilişkin ROC eğrileri ise, aşağıdaki şekilde gösterilmiştir.



Şekil 3.59: MARS-2 Modeline Ait ROC Eğrisi Sonuçları

MARS-2 modelinde başlangıç ve test verisine ilişkin AUC değerleri ise başlangıç veri seti için %89,5 olup mavi çizginin altında kalan alandır. Test verisi için %86,2 olup kırmızı çizginin altında kalan alandır. Bu oran, hem başlangıç hem test verisinde %80 ve üzerinde olduğu için modelin hem başlangıç hem de test verisine ait ayırım gücünün yani duyarlılık (doğru pozitif oranı) ve özgüllüğünün (doğru negatif oranı) oldukça iyi olduğu söylenebilir. MARS-1 ile kıyaslandığında ise daha yüksek çıktığı görülmüştür.

3.3.3.3. MARS Üçüncü Model Bulguları

MARS ile kurulan üçüncü modelde (MARS-3) başlangıç veri setinin %30'u modeli oluşturmak %70'i ise modeli test etmek için kullanılmıştır. Veri setine ait sonuçlar aşağıdaki tabloda gösterilmiştir.

Tablo 3.82. Başlangıç ve Test Verisi (MARS-3)

Parti	Başlangıç	Yüzde (%)	Test	Yüzde (%)	Toplam	Yüzde (%)
A	208	66,2	488	69,3	696	68
B	106	33,8	216	30,7	322	32
Toplam	314	100	704	100	1018	100

Tabloda da görüldüğü üzere toplam 1018 gözlemin 696'sı (%68) A partisine oy vereceğini, 322'si (%32) ise B partisine oy vereceğini belirtmiştir. Sonuçta bu gözlemlerin 314'ü (%30,7) başlangıç veri seti olup, modeli oluşturmak için kullanılmıştır. 704'ü (%69,3) ise test verisi olup, kurulan modeli test etmek için kullanılmıştır.

MARS yönteminde CART yönteminde olduğu gibi modele giren tüm değişkenler arasından en önemliler belirlenerek model oluşturulur. Önemsiz değişkenler modele katılmaz. Buna göre üçüncü model neticesinde anlamlı bulunan değişkenler, aşağıdaki tabloda gösterilmiştir.

Tablo 3.83. Değişkenlerin Önemlilik Oranları (MARS-3)

Değişken	Skor	
X33_KURUM_BAGI (En çok hangi kuruma gönülden güvenirsiniz?)	100,00	
X60_GAZETE_mis (Okuduğunuz gazete hangisidir?)	58,22	
X60_GAZETE (Okuduğunuz gazete hangisidir?)	58,22	
X64_AILE_DIN_mis (Dindarlık açısından ailenizin resisini hangisiyle tarif edersiniz?)	56,18	
X64_AILE_DIN (Dindarlık açısından ailenizin resisini hangisiyle tarif edersiniz?)	56,18	
X6_2_BABA_DB_mis (Babanızı doğum yeri nedir? "bölge")	41,66	
X6_2_BABA_DB (Babanızı doğum yeri nedir? "bölge")	41,66	
X13_TUR_HAYATBEK (Türkiye'deki hayat şartları 5 yıl sonra daha iyi olacak.)	39,56	

Modelin ilk oluşum aşaması olan ileriye doğru adım yönteminde başlangıçta varsayılan (default) değer olarak alınan 30 adet temel fonksiyonun ele alınması uygun görülmüştür. Fakat geriye doğru yönteminde sadece 4 tane temel fonksiyon anlamlı olarak bulunup, modele alınmıştır. MARS-3 modelinde modele giren değişkenler ile oluşturulan final model ve bu değişkenlere ait düğüm değerleri aşağıdaki tabloda görülmektedir. Tabloda da görüldüğü üzere, sıfırıncı temel fonksiyon sabit terim olup, modele "5","9", "13" ve "15." temel fonksiyonlar modele alınmıştır.

Tablo 3.84. Final Model (MARS-3)

Modele AlınanTemel Fonksiyonlar	Katsayılar	Değişken	İşareti	Ana Düğüm İşareti	Ana Düğüm	Düğüm Değeri (Kırılma noktası)
0	0,0115					
5	0,4506	X64_AILE_DIN	+	-	X64_AILE_DIN_mis	"İnançlı", "İnançsız"
9	0,3598	X60_GAZETE	-	-	X60_GAZETE_mis	"Posta", "Hürriyet", "Haber türk", "Milliyet", "Fotomaç", "Cumhuriyet", "Radikal", "Diğer gazeteler"
13	0,2409	X6_2_BABA_DB	+	+	X6_2_BABA_DB_mis	"Batı Marmara", "Ege", "Batı Anadolu", "Akdeniz", "Yurtdışı"
15	0,2510	X13_TUR_HAYATBEK	-			"Yanlış"

Buna göre MARS-3 analizi sonucunda Tablo 3.83'te gösterilen temel fonksiyonlardan yararlanılarak oluşturulan final tahmin modeli aşağıda gösterildiği gibidir. Tablo 3.85'de görüldüğü üzere modelin açıklama yüzdesi 0,51 olup, MARS-1 modelinden daha yüksek çıkmıştır. Yani bağımsız değişkenlerin %51'i parti tercihini açıklayabilmektedir. Bu değerin düşük olması MARS-3 modelinin de yeterince iyi bir tahmin modeli olmadığını göstermektedir.

Tablo 3.85. Final Modele Ait Temel Fonsiyonlar (MARS-3)

BF1 = (X33_KURUM_BAGI in ("Ordu", "Medya", "Hiçbiri")); BF3 = (X64_AILE_DIN in ("Eksik değer")) * BF1; **BF5** = (X64_AILE_DIN in ("İnançlı", "İnançsız")) * BF3; BF7 = (X60_GAZETE ne .) * BF1; **BF9** = (X60_GAZETE in ("Posta", "Hürriyet", "Haber türk", "Milliyet", "Fotomaç","Cumhuriyet", "Radikal", "Diğer gazeteler")) * BF7; BF11 = (X6_2_BABA_DB in ("Batı Marmara", "Ege", "Batı Anadolu", "Akdeniz", "Yurtdışı")); **BF13** = (X6_2_BABA_DB in ("Batı Marmara", "Ege", "Batı Anadolu", "Akdeniz", "Yurtdışı")) * BF11 **BF15** = (X13_TUR_HAYATBEK in ("Yanlış"));
MARS-3 Y = 0.0115098 + 0.450559 * BF5 + 0.35977 * BF9 + 0.240868 * BF13 + 0.251005 * BF15;
MARS-3 MODEL PARTI = **BF5 BF9 BF13 BF15;**
R-Kare:0,51 Düzeltilmiş R-Kare: 0,51 F-İstatistiği: 77,06 MSE: 0,11 Final Modelin GCV Değeri : 0,12

MARS-3 modeli sonucunda başlangıç veri seti ile oluşan ağacın doğru sınıflama yüzdesi aşağıdaki tabloda verilmiştir.

Tablo 3.86. Başlangıç Verisi Doğru Sınıflandırma Yüzdesi (MARS-3)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	172	36	208	82,7
B	17	89	106	84
Genel Sınıflandırma Oranı	189	125	314	83,1

Tablo 3.86'daki 314 gözlemden oluşan başlangıç verisi sonuçlarına göre, A partisini tercih ettiği bilinen 208 katılımcının 172'si doğru olarak sınıflandırılmış ve 36'sı da, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 106 katılımcının 89'u doğru sınıflandırılmış. 17'si ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre başlangıç verisinde A partisinin doğru sınıflandırma yüzdesi %82,7 B partisinin doğru sınıflandırma yüzdesi %84'tür. Toplamda 314 katılımcının 261'i doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %83,1 olarak bulunmuştur. Bu üç modelin oranlarına bakıldığında genel olarak MARS-2'nin en başarılı bir sınıflama yaptığı görülmektedir. Test verisi ile oluşan ağacın doğru sınıflama yüzdesi de aşağıdaki tabloda verilmiştir.

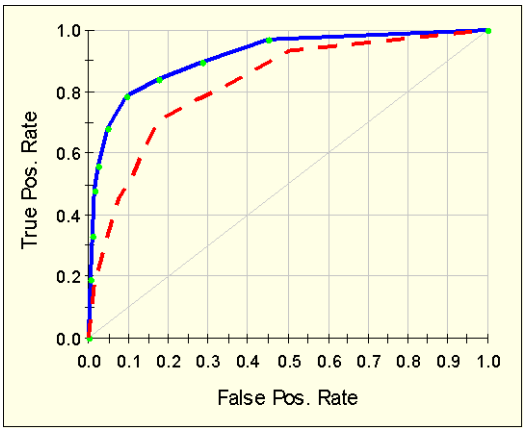
Tablo 3.87. Test Verisi Doğru Sınıflandırma Tablosu (MARS-3)

Gerçek Durum	Test Sonucu			Yüzde(%)
	A	B	Toplam	
A	357	131	488	73,2
B	49	167	216	77,3
Genel Sınıflandırma Oranı	406	298	704	74,4

Tablo 3.87'deki 704 gözlemden oluşan test verisi sonuçlarına göre, A partisini tercih ettiği bilinen 488 katılımcının 357'si doğru olarak sınıflandırılmış. 131'i ise, yanlış sınıf olan B sınıfına atanmıştır. B partisini tercih ettiği bilinen 216 katılımcının 167'si doğru sınıflandırılmış, 49'u ise yanlış sınıf olan A sınıfına atanmıştır. Buna göre test verisinde A partisinin doğru sınıflandırma yüzdesi %73,2 ve B partisinin doğru

sınıflandırma yüzdesi %77,3'tür. Toplamda 704 katılımcının 524'ü doğru sınıfa atanmış olup, genel olarak doğru sınıflandırma oranı %74,4 olarak bulunmuştur. Bu üç modelin oranlarına bakıldığında yine MARS-2 modelinin en başarılı bir sınıflama yaptığı görülmektedir.

MARS-3 modelinde başlangıç ve test verisine ilişkin ROC eğrileri ise, aşağıdaki şekilde gösterilmiştir.



Şekil 3.60: MARS-3 Modeline Ait ROC Eğrisi Sonuçları

MARS-3 modelinde başlangıç ve test verisine ilişkin AUC değerleri ise başlangıç veri seti için %91 olup mavi çizginin altında kalan alandır. Test verisi için %83 olup kırmızı çizginin altında kalan alandır. Bu oran, hem başlangıç hem test verisinde %80 ve üzerinde olduğu için modelin hem başlangıç hem de test verisine ait ayırım gücünün yani duyarlılık (doğru pozitif oranı) ve özgüllüğünün (doğru negatif oranı) oldukça iyi olduğu söylenebilir.

3.3.3.4. MARS Modellerinin Karşılaştırılması

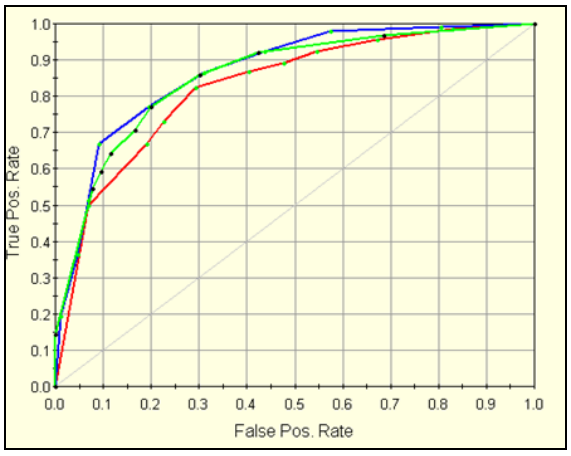
Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde en çok tercih edilen A ve B partisi için bu kez ikili MARS modelleri oluşturulmuştur. Bu aşamada farklı büyüklükteki başlangıç ve test verisi kullanılmıştır. MARS modellerine ait genel sonuçlar aşağıdaki tabloda gösterilmiştir.

Tablo 3.88. MARS Modellerinin Karşılaştırılması

MODEL	Başlangıç Verisi	Test Verisi	AUC (Başlangıç Verisi)	AUC (Test Verisi)	Genel Doğru Sınıflama Oranı (Başlangıç)	Genel Doğru Sınıflama Oranı (Test)
MARS-1	70%	30%	91	87,5	77,4	75,1
MARS-2	50%	50%	89,5	86,2	85	80,8
MARS-3	30%	70%	91	83	83,1	74,4

Sınıflama modeli oluşum aşamasında önemli olan kurulan modelin yeni veri seti ile yeniden sınanması neticesinde tekrar iyi bir sınıflama gücüne sahip olmasıdır. Bu sonuçlar neticesinde MARS-1 modelinin test verisine ait AUC değeri en yüksek, MARS-2 modelinin ise, genel doğru sınıflandırma oranı en yüksek çıktımıştır. Bu yüzden MARS-1 ve MARS-2 modelinin MARS-3 modelinden daha başarılı olduğu söylenebilir.

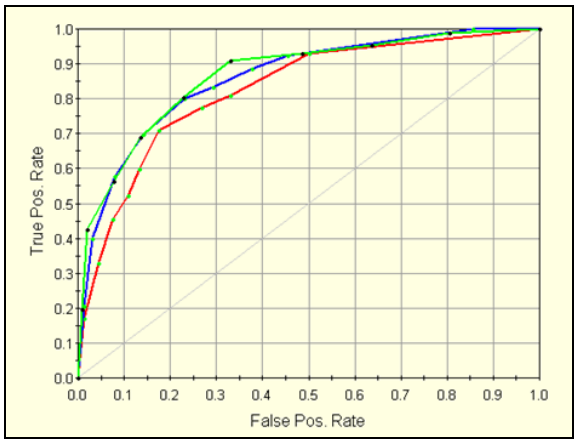
Ayrıca oluşturulan MARS modellerinin test verilerine ilişkin ROC eğrileri aşağıdaki şekil ile karşılaştırılabilir. Bu kez modelin genel AUC değerleri yerine sırasıyla A ve B partisine ilişkin karşılaştırmalı ROC eğrileri aşağıda gösterilmiştir.



Şekil 3.61: A Partisine Ait ROC Eğrileri Karşılaştırması-MARS

Oluştuturulan MARS modellerinden mavi ile gösterilen MARS-1, kırmızı ile gösterilen MARS-3, yeşil ile gösterilen MARS-2 olup, A partisi için en iyi ROC değerinin MARS-1 yani %70 başlangıç veri seti %30 test verisi ile oluşturulan model olduğu görülmüştür.

Yine aynı şekilde B partisine ilişkin karşılaştırmalı ROC eğrileri de aşağıda gösterilmiştir.



Şekil 3.62: B Partisine Ait ROC Eğrileri Karşılaştırması-MARS

B Partisine Ait ROC Eğrilerinde, oluştuturulan MARS modellerinde bu kez mavi ile gösterilen MARS-2, kırmızı ile gösterilen MARS-3, yeşil ile gösterilen MARS-1 olup, B partisi için en iyi ROC değerinin MARS-2 yani %50 başlangıç veri seti %50 test verisi ile oluşturulan model olduğu görülmüştür.

3.3.4. CART ve MARS Modellerinin Karşılaştırılması

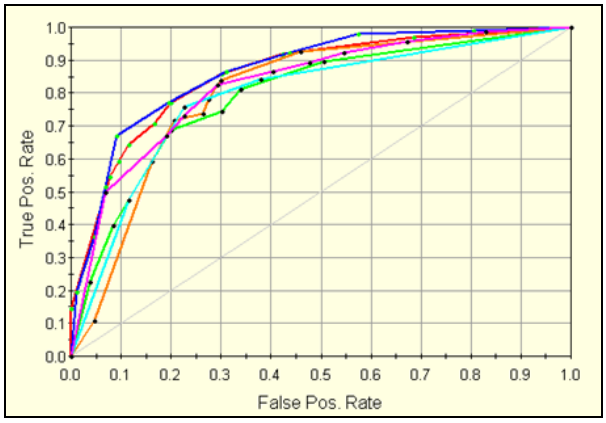
Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde CART ve MARS yöntemleri farklı büyüklükteki başlangıç ve test verisi kullanılarak incelenmiştir. İlk aşamada MARS modeli ikili bağımlı değişkenlerle çalıştığı için bağımlı değişkenin kategori düzeyi en çok tercih edilen ilk iki parti olan "A Partisi" ve "B Partisi" olarak ele alınmıştır. Bu durumda CART ve MARS modellerine genel sonuçlar aşağıdaki tabloda verilmiştir.

Tablo 3.89. CART ve MARS Modellerinin Karşılaştırılması

MODEL	Başlangıç Verisi	Test Verisi	AUC (Başlangıç Verisi)	AUC (Test Verisi)	Genel Doğru Sınıflama Oranı (Başlangıç)	Genel Doğru Sınıflama Oranı (Test)
CART-1	70%	30%	85,6	80,5	80,8	76,5
CART-2	50%	50%	87,5	79,4	77,4	72,2
CART-3	30%	70%	83,3	78,8	80,3	76,3
MARS-1	70%	30%	91	87,5	77,4	75,1
MARS-2	50%	50%	89,5	86,2	85	80,8
MARS-3	30%	70%	91	83	83,1	74,4

CART modellerinin kendi içinde karşılaştırılmasında en başarılı olan model CART-1 olduğu görülmüştür. MARS modellerinde de MARS-1 ve MARS-2 modelleri nin daha başarılı olduğu görülmüştür. İkili sınıflandırma da ise MARS-2 modelinin gerek AUC değerinin gerekse genel doğru sınıflandırma oranının yüksek çıkması ile diğer modellerden daha iyi bir ayırım gücüne sahip olduğu söylenebilir.

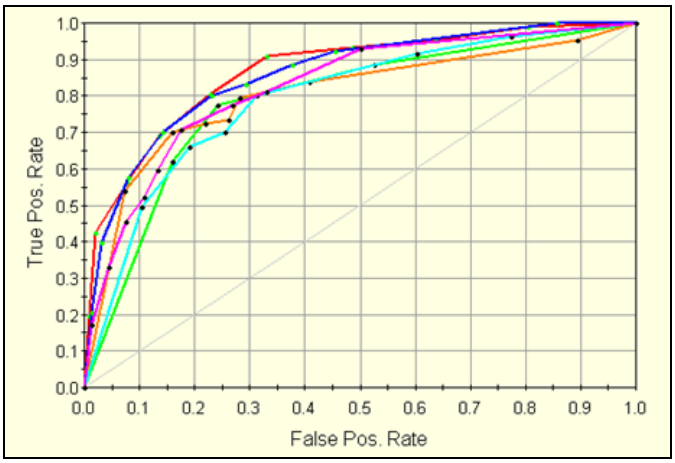
Ayrıca oluşturulan MARS ve CART modellerinin test verilerine ilişkin ROC eğrileri aşağıdaki şekil ile karşılaştırılabilir. Bu kez modelin genel AUC değerleri yerine sırasıyla A ve B partisine ilişkin karşılaştırmalı ROC eğrileri aşağıda gösterilmiştir.



Şekil 3.63: A Partisine Ait ROC Eğrileri Karşılaştırması-CART ve MARS

A partisine ait oluşturulan MARS modellerinden mavi ile gösterilen MARS-1, kırmızı ile gösterilen MARS-2, pembe ile gösterilen MARS-3, turuncu ile gösterilen CART-1, yeşil ile gösterilen CART-2 ve turkuaz rengi ile gösterilen CART-3 olup, A partisi için en iyi ROC değerini mavi ile gösterilen MARS-1 yani %70 başlangıç veri seti %30 test verisi ile oluşturulan model olduğu görülmüştür.

Yine aynı şekilde B partisine ilişkin karşılaştırmalı ROC eğrileri de aşağıda gösterilmiştir.



Şekil 3.64: B Partisine Ait ROC Eğrileri Karşılaştırması-CART ve MARS

B partisine ait oluştuturulan MARS modellerinden kırmızı ile gösterilen MARS-1, mavi ile gösterilen MARS-2, pembe ile gösterilen MARS-3, turuncu ile gösterilen CART-1, turkuaz ile gösterilen CART-2 ve yeşil rengi ile gösterilen CART-3 olup, B partisi için en iyi ROC değerini kırmızı ile gösterilen MARS-1 yani %70 başlangıç veri seti %30 test verisi ile oluşturulan model olduğu görülmüştür.

SONUÇ

Gerçek hayatta karşılaşılan veri setleri çok boyutlu ve karmaşık bir yapıya sahip olması nedeniyle içinde gizli kalmış birçok bilgi yığınını barındırmaktadır. Bu bilgi kirliği içerisinden araştırmacıların faydalı olanları ayıklamasında veri madenciliği oldukça önemli bir yöntem olarak kullanılmaktadır. Bu çalışmada da veri madenciliğinin önemli bir kolu olan karar ağaçları yönteminden faydalanılmıştır. Çalışmanın asıl amacı ise, karar ağaçları yöntemlerinden olan CART ve MARS yöntemlerini karşılaştırmaktır.

CART yöntemi hem kategorik hem de sürekli değişkenleri kullanarak sınıflama ve regresyon problemlerinin çözümünde karar ağaçlarını kullanan parametrik olmayan istatistiksel bir yöntemdir. Ele alınan bağımlı değişken kategorik ise yöntem sınıflama ağaçları, sürekli ise regresyon ağaçları olarak adlandırılmaktadır. Aynı şekilde MARS yöntemi de nonparametrik ve doğrusal olmayan bir yöntem olup, hem sürekli hem de ikili yanıt değişkenleri için tasarlanmıştır. Yanıt değişkeninin sürekli olması durumunda kestirim amaçlı olan bu yöntem, yanıt değişkeninin kategorik olması durumunda ise, sınıflandırma amacına sahiptir. Her iki yönteminde ortak özelliği bağımlı değişkenin türüne göre hem sınıflama hem de tahmin modeli geliştirebilmesidir.

Bu çalışmada amaç, Türkiye'de gençlerin siyasi görüşlerini etkileyen faktörlerin belirlenmesinde yani parti tercihlerinde CART ve MARS yöntemlerini karşılaştırıp, uygulamada hangi yöntemin diğerinden daha doğru bir sınıflama yapacağını farklı büyüklükteki başlangıç ve test verisi kullanarak incelemek ve sonrasında en uygun olan başlangıç ve test verisi büyüklüğüne göre, farklı büyüklükteki örnek sayıları ile bu kez sadece CART ile modelleme yaparak en başarılı sınıflama modelini oluşturmaya çalışmaktır.

Çalışmanın kapsamı doğrultusunda, KONDA Araştırma ve Danışmanlık şirketi tarafından 9-10 Nisan 2011 saha tarihinde *"Türkiye Gençliği Araştırması"* adlı bir anket çalışmasından faydalanılmıştır. Bu anket kapsamında Türkiye'de 15 yaş üstü, 30 yaş altı genç nüfusun, saha çalışmasının yapıldığı günlerdeki algıları, beklentileri, değerleri, tercihleri, hayat tarzları, bilgi edinme mecraları ve profillerini yansıtan ve gençlerin

kendi ve ailelerinin eğitim durumu gibi demografik özelliklerini içeren sorulara yer verilmiştir. Çalışmada gençlerin önümüzdeki seçimlerde hangi partiye oy vereceklerine yönelik soru ise, bağımlı değişken olarak ele alınmıştır.

Her iki yöntem için uygulanan analizlerle sınıflama modelleri oluşturulması hedeflenmiştir. MARS yöntemi ikili bağımlı değişkenler için tasarlanmış bir model olduğu için ve MARS ile elde edilen analiz sonuçlarının CART ile kıyaslanabilmesi için bağımlı değişkenin kategori düzeyi ilk aşamada ikiye indirgenmiştir ve en çok tercih edilen ilk iki parti olan "A Partisi" ve "B Partisi" ele alınmıştır. Böylelikle iki yöntemin sonuçları karşılaştırılmıştır. İkinci aşamada ise, en uygun olarak seçilen başlangıç ve test verisi büyüklüğü ile bu kez genel bir değerlendirme yapılmıştır. Bu değerlendirmede bağımlı değişken olan parti tercihi "A Partisi", "B Partisi", "C Partisi" "Diğer" ve "Kararsız" olmak üzere beş kategori düzeyine göre bu kez farklı büyüklükteki örnek sayıları ile ele alınarak, CART ile modellenmiş ve çeşitli örnek büyüklüğüne göre en başarılı sonuç veren sınıflama modeli belirlenmiştir.

Veri madenciliği ile ilgili yapılan çalışmalarda da genel olarak veri setinin 2/3'ünün yani %67'sinin modeli oluşturmak, 1/3'ünün yani %33'ünün ise modeli test etmek için ideal olabileceği yönünde olmuştur. Yapılan uygulamalardan elde edilen bilgiler doğrultusunda en uygun olan başlangıç yani eğitim verisi ve test verisi büyüklüğünün literatürde de belirtildiği gibi veri setinin yaklaşık %70'inin başlangıç veri seti olarak alınıp, geri kalan %30'unun ise, test verisi olarak alınması gerektiği yönünde olmuştur. İlk aşamada CART ve MARS yöntemlerinde bağımlı değişken olan parti tercihi değişkeninin kategori düzeyi iki olarak alındığında, hem başlangıç hem de 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.

Ayrıca uygulamada genellikle insan genetiği, gıda bilimi ve hastalık araştırmaları gibi çeşitli alanlarda kullanılan CART ve MARS yöntemleri bu alanlardan farklı olarak Türkiye gençliği üzerine yapılmış bir anketten toplanan veriler ile değerlendirilmiştir.

Bu veriler ile gençlerin hayat ve siyasi görüşlerinin nasıl şekillendiği ve dolayısıyla bu temel göstergelerden yola çıkarak bireylerin hayat ve siyasi görüşlerine göre sınıflandırılmasının mümkün olacağı düşünülmektedir. Aslında bu da doğrudan genç nüfusta siyasi seçimin ne şekilde ortaya çıkacağının bir göstergesi olabilmektedir. Bu durum, genç nüfusun toplumun anlamlı bir parçası olarak ele alınmasının ve gençlikle ilgili etkin politikaların oluşturulmasının önemini göstermektedir. Öyle ki nüfus projeksiyonlarına göre, 2025 yılında Türkiye'deki genç nüfus oranının, ABD ve Avrupa ülkelerinden yüksek olacağı ve Türkiye'nin, Avrupa Birliği ülkelerine kıyasla oldukça genç bir nüfusa sahip olduğu göz önüne alındığında; etkili gençlik politikalarının Türkiye açısından ne derece önemli olduğu anlaşılmaktadır.

EKLER

Ek 1: Karar Ağacı Algoritmalarının Genel Kodları

Ek 2: CHAID Algoritmasının Genel Kodu

Ek 3: CART Algoritmasının Genel Kodu

Ek 4: ID3 Algoritmasının Genel Kodu

Ek 5: C4.5 Algoritmasının Genel Kodu

Ek 6: SPRINT Algoritmasının Genel Kodu

Ek 7: Tekrarlamalı Ayırma (Recursive Patitioning) Algoritması

Ek 8: MARS-İleriye doğru adım (MARS-forward stepwise) Algoritması

Ek 9: MARS-Geriye Doğru Adım (MARS-backwards stepwise)Algoritması

Ek 10: Türkiye Gençliği Araştırması'2011 Anketi Verilerinin Kullanım İzni

Ek 1: Karar Ağacı Algoritmalarının Genel Kodları

```
D: Öğrenme veritabanı
T: Kurulacak Ağaç
T = 0            // başlangıçta ağaç boş küme
Dallara ayırma kriterlerini belirle
T = kök düğümü belirle
T = dallara ayırma kuralına göre kök düğümü dallara ayır;
      herbir dal için
do

        Bu düğüme gelecek değişkeni belirle
        if (durma koşuluna ulaşıldı)
                Yaprak ekle ve dur
        else
loop
```

Kaynak: Gökhan Silahtaroğlu, **Veri Madenciliği (Kavram ve Algoritmaları),** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.73.

Ek 2: CHAID Algoritmasının Genel Kodu

Girdiler: Veri Kümesi
Değişken Sayısı
Herbir değişkendeki farklı değer sayısı
Adımlar:
1. DO FOR *herbir tahminleyici (bağımsız) değişken (X)*
2. IF (X> *iki farklı değer*)
DO UNTIL (X= *iki farklı değer*)
Tüm değerleri diğer değerlerle ikili Ki-Kare testine sok ve en büyük p değeri veren ikilileri tek değer olarak etiketle.
3. Tüm iki değerli değişkenler için sınıf değişkenine göre ki-kare değerini hesapla.
4. En küçük *p* değerine sahip değişkenden dal yarat
LOOP
5. DUR

Kaynak: Gökhan Silahtaroğlu, **Veri Madenciliği (Kavram ve Algoritmaları),** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.105.

Ek 3: CART Algoritmasının Genel Kodu

Generate Tree(X)

If Node Entropy(X)<θ

$I^*entopy(X) = \sum_{i=1}^{k} p_m^i \, log_2 p_m^i$

Where m represent node and p is the estimate for probability of class i**l*

Create leaf labeled by majority class in X

Return

i←SplitAttirubute(X)

for each branch of x_i

find X_i falling each branch

Generate Tree(X_i)

SplitAtttribute(X)

MinEnt←MAX

For all attributes i=1,…,d

If x_i is discrete with n values

Split X into $X_1,\ldots,X_n$ by x_i

e←SplitEntropy($X_1,\ldots,X_n$)

$I^*SplitEntopy = -\sum_{j=1}^{n} \frac{N_{mj}}{N_m} \sum_{i=1}^{k} p_{mj}^i \, log_2 p_{mj}^i$ where N_m is number of training instances**I*

If e<MinEntMinEnt←e;bestf←I

Else *I***x_i is numeric $^*I$

For all possible splits

Split X into X_1, X_2 on X_i

e←SplitEntropy(X_1,X_2)

If e<MinEntMinEnt←e; bestf←I

Return bestf

Kaynak: İbrahim Halil Ekşi, "Classification of Firm Failure with Classification and Regression Trees"**, International Research Journal of Finance and Economics,** Vol.76. (2011), s.116.

Ek 4: ID3 Algoritmasının Genel Kodu

- Get dataset:D
- Get Class: C
- Count No.of Fields: f
- **START HER C için genel durum entropisini hesapla** $H(D) = \sum(p_i \log(1/p_i))$

Do for each f

Entropi Hesapla $H(D_i) = \sum(p_i \log(1/p_i)$

Olasılık Hesapla $P(D_i) = \sum_{i=1}^{n} P(D_i)H(D_i)$

Kazanç Ölçütünü Hesapla $H(D) - \sum_{i=1}^{n} P(D_i)H(D_i)$

Loop

En yüksek kazanç değerli f dal olarak atanır.

if durma kriteri= TRUE yaprak ata ve DUR

else

D= Dal ayrımı

Go TO START:

Kaynak: Gökhan Silahtaroğlu, **Veri Madenciliği (Kavram ve Algoritmaları),** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.76.

Ek 5: C4.5 Algoritmasının Genel Kodu

- Get dataset:D
- Get Class: C
- Count No.of Fields: f
- **START:** C sınıfı için genel durum entropisini
- hesapla $H(D) = \sum(p_i \log(1/p_i))$

Do for each f

Bilgi Bölünmesini Hesapla $H\left(\frac{|D_1|}{|D|},\ldots,\frac{|D_s|}{|D|}\right)$

Kazanç Oranını Hesapla Kazanç Oranı(D,S)= Kazanç(D,S)/Bilgi Bölünmesi(D,S)

Loop

En küçük değerli kazanç oranına sahip değişkeni düğüm olarak ata.

if durma kriteri= TRUE yaprak ata ve DUR

else

D= Dal ayrımı

Go TO START:

Kaynak: Gökhan Silahtaroğlu, **Veri Madenciliği (Kavram ve Algoritmaları),** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.81.

Ek 6: SPRINT Algoritmasının Genel Kodu

- Veri Seti: D
- Sınıflar: C
- Değişkenler: f
- Durma Kriteri: *dk*

START
Do for each f

Calculate Splitting Value $gini_{bölünmüş}(K) = \sum_{i=1}^{t} \frac{n_i}{n} gini(K_i)$

Loop
En küçük $GINI_{SPLIT}$ değerli değişkeni düğüm olarak ata

İf durma kriteri = TRUE then yaprak ekle VE DUR
else
D = Next Branch
Go TO START:

Kaynak: Gökhan Silahtaroğlu, **Veri Madenciliği (Kavram ve Algoritmaları),** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013, s.87.

Ek 7: Tekrarlamalı Ayırma (Recursive Patitioning) Algoritması

$B_1(\mathbf{x}) \leftarrow 1$
For $M = 2$ to $M_{\max}$ do: lof* $\leftarrow \infty$
  For $m = 1$ to $M - 1$ do:
    For $v = 1$ to n do:
      For $t \in \{x_{vj} | B_m(\mathbf{x}_j) > 0\}$
        $g \leftarrow \sum_{i \neq m} a_i B_i(\mathbf{x}) + a_m B_m(\mathbf{x}) H[+(x_v - t)] + a_M B_m(\mathbf{x}) H[-(x_v - t)]$
        lof $\leftarrow \min_{a_1, \ldots, a_M} \text{LOF}(g)$
        if lof $<$ lof*, then lof* $\leftarrow$ lof; $m^* \leftarrow m$; $v^* \leftarrow v$; $t^* \leftarrow t$ end if
      end for
    end for
  end for
  $B_M(\mathbf{x}) \leftarrow B_{m^*}(\mathbf{x}) H[-(x_{v^*} - t^*)]$
  $B_{m^*}(\mathbf{x}) \leftarrow B_{m^*}(\mathbf{x}) H[+(x_{v^*} - t^*)]$
end for
end algorithm

Kaynak: Jerome H. Friedman, "Multivariate Adaptive Regression Splines", **The Annals of Statistics,** Vol.19, No.1 (June 1991), s.11.

Bu algorima için, H [illegible] pozitif argümanı belirleyen bir adım fonksiyonudur. LOF(g) ise, g(x) fonksiyonunun dataya uyumsuzluğunu hesaplayan bir işlem olmak üzere belirlenmiştir. Bu durumda Tablo EK 8'de belirtilen ileri adımlı regresyon algoritması, tekrarlamalı ayırma (recursive partitioning) yöntemine eşit olmaktadır.

Ek 8: MARS-İleriye doğru adım (MARS-forward stepwise) Algoritması

$B_1(\mathbf{x}) \leftarrow 1$; $M \leftarrow 2$
Loop until $M > M_{\max}$: lof* $\leftarrow \infty$
  For $m = 1$ to $M - 1$ do:
    For $v \notin \{v(k, m) | 1 \leq k \leq K_m\}$
      For $t \in \{x_{vj} | B_m(\mathbf{x}_j) > 0\}$
        $g \leftarrow \sum_{i=1}^{M-1} a_i B_i(\mathbf{x}) + a_M B_m(\mathbf{x})[+(x_v - t)]_+ + a_{M+1} B_m(\mathbf{x})[-(x_v - t)]_+$
        lof $\leftarrow \min_{a_1, \ldots, a_{M+1}}$ LOF(g)
        if lof < lof*, then lof* $\leftarrow$ lof; $m^* \leftarrow m$; $v^* \leftarrow v$; $t^* \leftarrow t$ end if
      end for
    end for
  end for
  $B_M(\mathbf{x}) \leftarrow B_{m^*}(\mathbf{x})[+(x_{v^*} - t^*)]_+$
  $B_{M+1}(\mathbf{x}) \leftarrow B_{m^*}(\mathbf{x})[-(x_{v^*} - t^*)]_+$
  $M \leftarrow M + 2$
end loop
end algorithm

Kaynak: Jerome H. Friedman, "Multivariate Adaptive Regression Splines", **The Annals of Statistics,** Vol.19, No.1 (June 1991), s.17.
*****Burada LOF(g), g(x) fonksiyonunun dataya uyumsuzluğunu hesaplayan bir işlem olmak üzere belirlenmiştir.

Ek 9: MARS-Geriye Doğru Adım (MARS-backwards stepwise)Algoritması

$J^* = \{1, 2, \ldots, M_{\max}\}$; $K^* \leftarrow J^*$
$\text{lof}^* \leftarrow \min_{\{a_j | j \in J^*\}} \text{LOF}(\sum_{j \in J^*} a_j B_j(\mathbf{x}))$
For $M = M_{\max}$ to 2 do: $b \leftarrow \infty$; $L \leftarrow K^*$
  For $m = 2$ to M do: $K \leftarrow L - \{m\}$
    $\text{lof} \leftarrow \min_{\{a_k | k \in K\}} \text{LOF}(\sum_{k \in K} a_k B_k(\mathbf{x}))$
    if lof $< b$, then $b \leftarrow$ lof; $K^* \leftarrow K$ end if
    if lof $<$ lof*, then lof* $\leftarrow$ lof; $J^* \leftarrow K$ end if
  end for
end for
end algorithm

Kaynak: Jerome H. Friedman, "Multivariate Adaptive Regression Splines", **The Annals of Statistics,** Vol.19, No.1 (June 1991), s.17.

Ek 10: Türkiye Gençliği Araştırması'2011 Anketi Verilerinin Kullanım İzni

T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
GENEL SEKRETERLİK

18.06.2015

Sayı: 1325

Sayın, Öğr.Gör.Yeliz Sevimli SAİTOĞLU

İstanbul Kültür Üniversitesi Nisan 2011 tarihinde KONDA Araştırma ve Danışmanlık şirketine "Türkiye Gençliği Araştırması" anketi yaptırmıştır. Bu ankete ilişkin verileri, Marmara Üniversitesi Sosyal Bilimler Enstitüsü İstatistik Bilim Dalında yapmakta olduğunuz doktora tez çalışmasında kullanmanız uygundur.

Gereğini bilgilerinize rica ederim.

Yrd.Doç.Dr. Ethem TARHAN
Genel Sekreter

İstanbul Kültür Üniversitesi Genel Sekreterlik
Ataköy Yerleşkesi, Bakırköy 34156 / İstanbul
Tel: (0212) 498 47 10 Santral: (0212) 498 41 41 Faks: (0212) 498 43 06

Resmi Yazışmalar: genelsekreterlik@iku.edu.tr
Genel Yazışmalar: kultur@iku.edu.tr
www.iku.edu.tr

KAYNAKÇA

Kitaplar

Breiman, Leo, Friedman, H., Jerome, Olshen, A., Richard, Stone, J., Charles. Classification and Regression Trees. New York: Chapman & Hall, 1993.

Bülbül, Şahamet. **Tanımlayıcı İstatistik.** İstanbul: DER Yayınevi, 2013.

Çilingirtürk, A.Mete. **İstatistiksel Karar Almada Veri Analizi.** Ankara: Seçkin Yayıncılık, 2011.

Çölkesen, Rifat. **Veri Yapıları ve Algoritmalar**. 9. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2014.

Ferreruela, Inmaculada Cava. "Explaining Patterns of Broadband Development in OECD Countries". Yogesh K Dwivedi (Ed.). **Handbook of Research on Global Diffusion of Broadband Data Transmission** içinde. Hershey PA, USA: IGI Global, 2008, ss.756-775.

Gürsoy Şimşek, U., Tuğba. Uygulamalı Veri Madenciliği: Sektörel Analizler. 3. Basım, Ankara: Pegem Akademi, 2012.

Han, Jiawei, Micheline Kamber. **Data Mining Concept and Techniques.** 2. Basım, Amerika: Morgan Kaufmann Publications, 2006.

Hastie, Trevor, Robert Tibshirani ve Jerome Friedman. **The Elements of Statistical Learning: Date mining, Inference and Prediction.** New York: Springer, 2001.

Kayış, Aliye, "Kümeleme Analizi", Şeref Kalaycı (Ed.). **SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri** içinde. Ankara: Asil Yayın Dağıtım, 2006, ss.349-376.

Kıroğlu, Gülay. **Uygulamalı Parametrik Olmayan İstatistiksel Yöntemler.** İstanbul: Paymaş, 2001.

Olecka, Anna. "Beyond Classification: Challenges of Data Mining for Credit Scoring", X. Zhu ve I. Davidson (Eds.). **Knowledge Discovery and Data Mining: Challenges and Realities** içinde. Hershey PA: Information Science Reference, 2007, ss.139-161.

Özkan, Yalçın. **Veri Madenciliği Yöntemleri.** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013.

Ryan, Thomas. **Modern Regression Methods**. New York: John Wiley& Sons, 1997.

Silahtaroğlu, Gökhan. **Veri Madenciliği (Kavram ve Algoritmaları).** 2. Basım, İstanbul: Papatya Yayıncılık Eğitim, 2013.

Şentürk, Aysan. **Veri Madenciliği Kavram ve Teknikler.** 1. Basım, Bursa: Ekin Basın Yayın Dağıtım, 2006.

Tatlıdil, Hüseyin. **Uygulamalı Çok Değişkenli İstatistiksel Analiz.** Ziraat Matbaacılık, 1.Basım, Ankara, 2002.

Yıldırım, İ.Esen. **Kamuoyu Araştırmaları ve Su Tüketim Bilinci Üzerine Bir Uygulama.** Ankara: Seçkin Yayıncılık, 2010.

Süreli Yayınlar

Albayrak, Ali Sait ve Şebnem Koltan Yılmaz. "Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri üzerine Bir Uygulama", **Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Dergisi.** Cilt.14, Sayı.1, 2009, ss.31-52

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. http://www.palgrave-journals.com/jt/journal/v18/n2/pdf/jt20103a.pdf (31 Ekim 2013)

Anirban Mukhopadhyay ve A. Iqbal. "Prediction of Mechanical Property of Steel Strips Using Multivariate Adaptive Regression Splines" *Journal of Applied Statistics*, 2009, Vol.36, No.1, ss.1-9. http://www.tandfonline.com/toc/cjas20/current (23 Mart 2009)

Bilim ve Teknik. "Rastlantının Bittiği Yer Big Data". Cilt. 46, Sayı. 550, 2013, ss.22-26.

Bilim ve Teknik. "Büyük Veri Kahramanı Veri Bilimci". Cilt. 48, Sayı. 569, 2015, ss.76-79.

Chen, Hsinchun, Ganesan Shankaranarayanan ve Linlin She. "A machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms and Simulated Annealing", **Journal Of American Society for Information Science.** Vol.49, No.8, 1998, ss.693-705.

Chou, Shieu-Ming, Tian-Shyug Lee, Yuehjen E. Shao ve I-Fei Chen. "Mining the "Breast Cancer Pattern Using Artificial Neural Networks and Multivariate Adaptive Regression Splines", **Expert Systems with Applications.** Vol.27, No.1, 2004, ss.133-142.

Dağdeviren, Nezih, Zeliha Musaoğlu, İmran Kurt Ömürlü ve Serdar Öztora. "Akademisyenlerde İş Doyumunu Etkileyen Faktörler", **Balkan Med J.** Cilt.28, 2011, ss.69-74.

Deconinck, E, QS. Xu, R. Put, D. Coomans, DL. Massart, Y. Vander Heyden. "Prediction of Gastro-Intestinal Absorption Using Multivariate Adaptive Regression Splines", **Journal of Pharmaceutical and Biomedical Analysis.** Vol. 39, No.5, 2005, ss.1021-1030

Friedman, Jerome H. "Multivariate Adaptive Regression Splines", **The Annals of Statistics.** Vol.19, No.1, 1991, ss.1-67.

Kass, G.V. "Significance Testing in Automatic Interaction Detection (A.I.D.)", **Applied Statistics.** Vol.24, No.2, 1975, ss.178-189.

Kass, G.V. "An Exploratory Technique for Investigating Large Quantities of Categorical Data", **Applied Statistics.** Vol.29, No.2, 1980, ss.119-127.

Kayri, Murat ve Murat Boysan. "Araştırmalarda Chaid Analizinin Kullanımı ve Baş Etme Stratejileri İle İlgili Bir Uygulama", **Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi.** Cilt.40, Sayı.2, 2007, ss.133-149.

Ko, Myung, Jan Guynes Clark ve Daijin Ko. "Revisiting The Impact of Information Technology Investments on Productivity: An Empirical Investigation Using Multivariate Adaptive Regression Splines (MARS)", **Information Resources Management Journal.** Vol.21, No.3, 2008, ss.1-23.

Köktürk, Fürüzan, Handan Ankaralı ve Vildan Sümbüloğlu, "Veri Madenciliği Yöntemlerine Genel Bakış", **Türkiye Klinikleri Journal of Biostatistics,** Cilt.1, Sayı.1, 2009, ss.20-25.

Leathwicka, JR, J. Elith ve T. Hastie. "Comparative Performance of Generalized Additive Models and Multivariate Adaptive Regression Splines for Statistical Modelling of Species Distributions", **Ecological Modeling.** Vol. 199, No.2, 2006, ss.211-232.

Özkan, Kürşad. "Sınıflandırma ve Regresyon Ağacı Tekniği (SRAT) ile Ekolojik Verinin Modellenmesi", **Süleyman Demirel Üni. Orman Fak. Dergisi.** Cilt.13, 2012, ss.1-4.

Patil, Nilima, Rekha Lathi ve Vidya Chitre. "Comparision of C5.0 and CART Classification Algorithms Using Pruning Technique", **International Journal of Engineering Research and Technology.** Vol.1, No.4, 2012, ss.1-5.

R, Put, Xu QS, Massart DL ve Vander Heyden Y. "Multivariate Adaptive Regression Splines (MARS) in Chromatographic Quantitative Structure–Retention Relationship Studies" , **Journal of Chromatography A.** Vol. 1055, No.1-2, 2004, ss.11-19.

Jasna, Soldic-Aleksic. "Combined Approach Of Kohonen Som And Chaid Decision Tree Model To Clustering Problem: A Market Segmentation Example**",** **Journal of Economics and Engineering.** Vol.3, No.1, 2012, 20-27.

Temel, Gülhan, Handan Çamdeviren ve Zeki Akkuş. "Sınıflama Ağaçları Yardımıyla Legs Syndrome (RLS) Hastalarına Tanı Koyma", **İnönü Üni. Tıp Fak. Dergisi.** Cilt.12, Sayı.2, 2005, ss.111-117.

Tomak, Leman, ve Yüksel Bek. "İşlem Karakteristik Eğrisi Analizi ve Eğri Altında Kalan Alanların Karşılaştırılması", 2010, Vol.27, No.2, ss.58-65. http://dergi.omu.edu.tr/omujecm/article/view/1009001569 (31 Ekim 2014).

Tok, Türkay Nuri. "Türkiye'deki Siyasal Partilerin Eğitim Söylemleri ve Siyasaları", **Kuram ve Uygulamada Eğitim Yönetimi.** Cilt.18, Sayı.2, 2012, ss.273-312.

Tunay, K. Batu. "Türkiye'de Paranın Gelir Dolaşım Hızlarının MARS Yöntemiyle Tahmini", **ODTÜ Gelişme Dergisi.** Vol. 28, No. 3-4, 2001, ss.431-454.

Yaman, Ömer Miraç. "Türkiye'de Gençlik Sosyolojisi Çalışmalarına Dair Bibliyografik Bir Değerlendirme", **Alternatif Politika.** Cilt.5, Sayı.2, 2013, ss.114-138.

Xu, M. Daszykowski, B. Walczak, F. Daeyaert, M.R. de Jonge, J. Heeres, L.M.H. Koymans, P.J. Lewi, H.M. Vinkers, P.A. Janssen ve D.L. Massart. "Multivariate Adaptive Regression Splines-Studies of HIV Reverse Transcriptase Inhibitors" , **Chemometrics and Intelligent Laboratory Systems.** Vol.72, No.1, 2004, ss.27-34.

Xu, F. Daeyaert, P. J. Lewi ve D. L. Massart. "Studies of Relationship Between Biological Activities and HIV Reverse Transcriptase Inhibitors by Multivariate Adaptive Regression Splines with Curds and Whey", **Chemometrics and Intelligent Laboratory Systems.** Vol.82, No.1-2, 2006, ss.24-30.

Yao, Dengju, Jing Yang, ve Xiaojuan Zhan. "A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines**",** **Journal of Computers**. Vol.8, No. 1, 2013, ss.170-177.

Diğer Yayınlar

Answer Tree Algorithm Summary. (t.y.) http://lyle.smu.edu/~mhd/8331f03/AT.pdf (17 Mart 2014).

CHAID. (t.y.) http://en.wikipedia.org/wiki/CHAID (14 Ocak 2014).

CHAID and Exhaustive CHAID Algorithms. (t.y.) ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf (18 Ocak 2014).

C4.5 algorithm. (t.y.) http://en.wikipedia.org/wiki/C4.5_algorithm (30 Ekim 2013).

C5.0Node.(t.y.) http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=%2Fcom.ibm.spss.modeler.help%2Fc50node_general.htm (30 Ekim 2013).

Çelik, Yusuf. "Duyarlılık (Sensitivity) ve Belirleyicilik (Specificity)", 2014. Dicle Üniversitesi, Tıp Fakültesi. http://www.dicle.edu.tr/Contents/3b4d94e3-3582-4384-939b-c5957b348bdd.pdf (31 Ekim 2014).

Decision Trees. (t.y.) http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf (20 Ocak 2013).

Dondurmacı, Gülser. "Veri Madenciliği'nde Regresyon Ağaçları ile Sınıflandırma ve Bir Uygulama", **Yayınlanmamış Doktora Tezi.** Mimar Sinan Güzel Sanatlar Üniversitesi, FBE, 2011.

Entropy(informationtheor.,(t.y.) http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Entropy_(information_theory).html (15.11.2013).

Ertorsun, Ayça Deniz, Burak Bağ, Güldeniz Uzar ve Mehmet Ali Turanoğlu. "ROC (Receiver Operating Characteristic) Eğrisi Yöntemi ile Tanı Testlerinin Performanslarının Değerlendirilmesi", Başkent Üniveristesi, Tıp Fakültesi. http://tip.baskent.edu.tr/egitim/mezuniyetoncesi/calismagrp/ogrsmpzsnm12/10.2.pdf (25 Ekim 2013).

IBM SPSS Modeler 15 Algorithms Guide. C&RT Algorithms. 2012. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/AlgorithmsGuide.pdf (29 Mart 2014).

Kıran, Zeynep Burcu. "Lojistik Regresyon ve CART Analizi Teknikleriyle Sosyal Güvenlik Kurumu İlaç Provizyon Sistemi Verileri Üzerinde Bir Uygulama", **Yayınlanmamış Yüksek lisans Tezi,**. Gazi Üniversitesi, FBE, 2010.

KONDA Araştırma ve Danışmanlık, Türkiye Gençliği Araştırması'2011.

Koyuncugil, Serhan. "Veri Madenciliği Yöntemleri ve Uygulamaları", 2010, Başkent Üniversitesi, Ticari Bilimler Fakültesi. http://www.koyuncugil.org/files/ders/bolum9.6.pdf (9 Kasım 2013).

Köksal, Burcu. "Regresyon Analizinde ROC Eğrisi Kestirimi ile Model Seçimi", **Yayınlanmamış Yüksek lisans Tezi.** Marmara Üniversitesi, SBE, 2011.

Orhan, Umut. "Makine Öğrenmesi", 2012, Çukurova Üniversitesi, Bilgisayar Mühendisliği. http://bmb.cu.edu.tr/uorhan/DersNotu/Ders03.pdf (15 Kasım 2013).

Pehlivan, Gamze. "Chaid Analizi ve Bir Uygulama", **Yayınlanmamış Yüksek lisans Tezi.** Yıldız Teknik Üniversitesi, FBE, 2006.

QUESTNod.(t.y.) http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=%2Fcom.ibm.spss.modeler.help%2Fc50node_general.htm (30 Ekim 2013).

Ritschard, Gilbert. "CHAID and Earlier Supervised Tree Methods", 2010, University of Geneva, Dept of Econometrics. http://www.unige.ch/ses/metri/cahiers/2010_02.pdf (17 Mart 2014).

Salford Systems, **MARS™ User Guide**, San Diego, 2001.

Salford Systems, **CART Tree Structured Non-Parametric Data Analysis User Guide**, San Diego, 2001.

Sevimli, Yeliz. "Çok Değişkenli Uyarlanabilir Regresyon Uzanımlarının Bir Split-Mouth Çalışmasında Uygulaması", **Yayınlanmamış Yüksek Lisans Tezi.** Marmara Üniversitesi, Sağlık Bilimleri Enstitüsü, 2009.

Sezgin, Özge. "Statistical Method in Credit Banking", **Yayınlanmamış Yüksek Lisans Tezi,** Orta Doğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, 2006.

Shannon entropy. (t.y.) http://www.ueltschi.org/teaching/chapShannon.pdf (15 Kasım 2013).

Siyaset, Ekonomi ve Toplum Araştırmaları Vakfı (SETA), **Türkiye'nin Gençlik Profili,** 1. Baskı, Ankara: Pelin Ofset, 2012.

Tapkan, Pınar, Lale Özbakır ve Adil Baykasoğlu, "Weka ile Veri Madenciliği Süreci ve Örnek Uygulama", **Endüstri Mühendisliği Yazılımları ve Uygulamaları Kongresi.** İzmir: MMO. 30 Eylül – 01/02 Ekim 2011, ss.247-262.

Temel, G. Orekici, Handan Çamdeviren ve A. Canan Yazıcı. "Regresyon Modellerine Alternatif Bir Yaklaşım: MARS", VIII. **Ulusal Biyoistatistik Kongresi.** Bursa: Uludağ Üniversitesi, 20-22 Eylül 2005, ss.105-123.

Tree-based Regression. (t.y.) http://www.dcc.fc.up.pt/~ltorgo/PhD/th3.pdf (28 Şubat 2014).

Timofeev, Roman. "Classification and Regression Trees (CART) Theory and Aplications", 2004, Humbolt University, Center of Applied Statistics and Economics. http://tigger.uic.edu/~georgek/HomePage/Nonparametrics/timofeev.pdf (26 Kasım 2012).

Türküm, Laikim, Müslümanım. http://www.hurriyet.com.tr/gundem/12104881.asp (26 Ocak 2015).

Üniversite Gençliğinin Siyasal Tutumları Üzerine Bir İnceleme-Süleymen Demirel Üniversitesi Örneğ., (t.y.) http://www.yerelsiyaset.com/pdf/kasim2007/17.pdf (26 Ocak 2015).