
GAZİ ÜNİVERSİTESİ
Gazi
1926

HANEHALKI BİLİŞİM TEKNOLOJİLERİ KULLANIMININ VERİ MADENCİLİĞİ TEKNİKLERİ İLE ANALİZİ

Mahmut COŞKUN

YÜKSEK LİSANS TEZİ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

HAZİRAN 2019

Mahmut COŞKUN tarafından hazırlanan "HANEHALKI BİLİŞİM TEKNOLOJİLERİ KULLANIMININ VERİ MADENCİLİĞİ TEKNİKLERİ İLE ANALİZİ" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Yönetim Bilişim Sistemleri Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. Halil İbrahim BÜLBÜL

Bilgisayar ve Öğretim Teknolojisi Eğitimi, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ...……

Başkan: Prof. Dr. Cevriye GENCER

Endüstri Mühendisliği, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Üye: Doç. Dr. Nergiz ÇAĞILTAY

Bilgisayar Mühendisliği, Atılım Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 17/06/2019

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……..…….

Doç. Dr. Aslıhan TÜFEKCİ

Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Mahmut COŞKUN

17/06/2019

HANEHALKI BİLİŞİM TEKNOLOJİLERİ KULLANIMININ VERİ MADENCİLİĞİ TEKNİKLERİ İLE ANALİZİ

(Yüksek Lisans Tezi)

Mahmut COŞKUN

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

Haziran 2019

ÖZET

Bilgisayarların insan hayatına girmesi ile yapılan çoğu işlem kayıt altına alınmaya başlanmıştır. Kayıt altına alınan bu işlemlerin verileri arasında keşfedilmeyi bekleyen birbiri ile ilintili bilgiye dönüştürülebilecek bağlantılar bulunmaktadır. Bu gizli bağlantıların keşfedilmesi için veri madenciliği tekniklerinden yararlanılır. Bu tezde, Türkiye İstatistik Kurumu'nun 2016 yılı Hanehalkı Bilişim Teknolojileri Kullanım Araştırması Anketi verileri kullanılmıştır ve hanehalkının internet hizmetine sahip olma durumu incelenmiştir. Hanehalkı internet hizmeti sahipliğini etkileyen karakteristiklere ilişkin faktörler, karar ağaçları kullanılarak analiz edilmiştir. Verilerin analizinde IBM SPSS Modeler 18.1 yazılımı kullanılmıştır. CHAID, C5.0, C&RT ve QUEST karar ağacı algoritmaları karşılaştırılmış ve C5.0 en başarılı algoritma olarak belirlenmiştir. C5.0 algoritması ile analiz yapılmış ve 10 dallanma, 21 düğümden oluşan bir karar ağacı elde edilmiştir. Analiz sonucunda hanelerin internet hizmeti sahipliğini etkileyen en önemli değişkenlerin hane cep telefonu sahipliği, hane bilgisayar kullanımı, hanede 0-25 yaş arasında bireyin olup olmaması, hane tablet sahipliği, hane dizüstü bilgisayar sahipliği, hanehalkı büyüklüğü, hanehalkı reisinin yaşı, hane smart tv sahipliği, hane gelir grubu olduğu görülmüştür. Karar ağacı ile internet hizmeti sahipliğinin "Evet" değeri için 8 adet, "Hayır" değeri için 4 adet olmak üzere toplam 12 adet kural elde edilmiştir. Modelleme sonucunda 11 222 haneden 9 577'si doğru sınıflandırılmış ve hanelerin doğru sınıflandırılma oranı % 85,34'tür. Bu çalışma ile internet hizmeti sahipliğinin hanehalkı karakteristiklerine göre ne şekilde değiştiğinin incelenmesi amaçlanmış ve sektörün ilgililerinin dikkatine sunulması hedeflenmiştir.

Bilim Kodu	:	114611
Anahtar Kelimeler	:	Veri madenciliği, Türkiye İstatistik Kurumu, hanehalkı bilişim teknolojileri kullanımı araştırması, karar ağacı, C5.0 algoritması
Sayfa Adedi	:	110
Danışman	:	Prof. Dr. Halil İbrahim BÜLBÜL

ANALYSIS OF THE USE OF HOUSEHOLD INFORMATION TECHNOLOGIES WITH DATA MINING TECHNIQUES

(M. Sc. Thesis)

Mahmut COŞKUN

GAZİ UNIVERSITY

INSTITUTE OF INFORMATICS

June 2019

ABSTRACT

Processes have been started to be recorded more easily with the introduction of computers into human life. In the records of these transactions recorded, there are links that can be transformed into related information waiting to be explored. Data mining techniques are exploited to discover these hidden links. In this thesis, Turkey Statistics Institute's 2016 Household Information Technology Usage Research Survey data is used and the state of households having internet service is examined. Factors affecting the characteristics of household internet service ownership were analyzed using decision trees. IBM SPSS Modeler 18.1 data mining software program was used for data analysis. CHAID, C5.0, C&RT and QUEST decision tree algorithms are compared and the most successful algorithm is determined as C5.0. The C5.0 algorithm was used for the analysis and a 10 branch, 21 node decision tree was obtained. As a result of the analyzes, the most important variables affecting the ownership of the internet service of the households are: household cell phone ownership, household computer use, whether the household is between 0-25 years old, household tablet ownership, household laptop ownership, household size, age of household head, household smart tv ownership, household income group. A total of 12 rules have been obtained with the decision tree, 8 for the "Yes" value and 4 for the "No" value of internet service ownership. As a result of the modeling, 9 577 of 11 222 households were classified correctly and the correct classification rate of households was % 85,34. With this study, it is aimed to examine how the ownership of internet services has changed according to household characteristics and to consider to the attention of the sector concerned.

Science Code	:	114611
Key Words	:	Data mining, Turkey Statistical Institute, household information technology usage survey, decision tree, C5.0 algorithm
Page Number	:	110
Supervisor	:	Prof. Dr. Halil İbrahim BÜLBÜL

TEŞEKKÜR

Bu çalışmanın gerçekleşmesinde tecrübesini ve değerli bilgilerini benimle paylaşan saygıdeğer danışman hocam Prof. Dr. Halil İbrahim BÜLBÜL'e, desteğini esirgemeyen saygıdeğer hocam Prof. Dr. Şeref SAĞIROĞLU'na, tez aşamasında iken kaybettiğim değerli babam, benden desteğini hiçbir zaman esirgemeyen annem ve kardeşime, çalışma süresince tüm zorluklara rağmen bana destek ve yardımcı olan kızım Irmak ve değerli eşim Özlem COŞKUN'a teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

ÖZET iv
ABSTRACT v
TEŞEKKÜR vi
İÇİNDEKİLER vii
ÇİZELGELERİN LİSTESİ x
ŞEKİLLERİN LİSTESİ xii
RESİMLERİN LİSTESİ xiii
SİMGELER VE KISALTMALAR xv
1. GİRİŞ 1
2. MEVCUT LİTERATÜRÜN İNCELENMESİ 5
3. KAVRAMSAL ÇERÇEVE 11
3.1. Veri Madenciliği Nedir? 11
3.2. Veri Madenciliği Süreci 12
3.2.1. CRISP-DM süreç modeli 13
3.3. Veri Madenciliği Yöntemleri 16
3.3.1. Sınıflandırma 18
3.3.2. Kümeleme 20
3.3.3. Birliktelik Kuralları 22
3.4. Karar Ağaçları 24
3.4.1. Karar ağaçlarının avantaj ve dezavantajları 26
3.5. Karar Ağacı Algoritmaları 27
3.5.1. ID3 algoritması 30
3.5.2. C4.5 ve C5.0 algoritmaları 35
3.5.3. C&RT algoritması 36

Sayfa

3.5.4. Chaid algoritması 38
3.5.5. Quest algoritması 39
3.6. Sınıflandırma Doğruluk Belirleme Yöntemleri 41
3.6.1. Hold out 41
3.6.2. K-fold cross validation 42
3.6.3. Bootstrap 42
3.7. Sınıflandırma Performans Karşılaştırma Yöntemleri 43
3.7.1. Karışıklık matrisi 43
3.7.2. Kaldıraç değeri ve grafiği 44
4. UYGULAMA VE BULGULAR 47
4.1. Uygulama Kullanılan Veriler 47
4.2. Uygulamada Kullanılan Değişkenlere Ait İstatistiksel Bilgiler 52
4.3. Uygulamada Kullanılan Yöntem ve Veri Madenciliği Aracı 59
4.4. Karar Ağacı Algoritmaları ile Modelleme 61
4.5. Sınıflandırma Modellerinin Performanslarının Karşılaştırılması 65
4.5.1. Doğru sınıflandırma yüzdesi 65
4.5.2. Alıcı işletim karakteristiği ve grafiği 67
4.5.3. Kaldıraç değeri ve grafiği 69
4.6. C5.0 Algoritması ile Karar Ağacı Modelleme 71
4.6.1. C5.0 algoritması başarı durumu 87
4.6.2. C5.0 algoritmasındaki önemli değişkenler için web grafiği 88
4.6.3. Farklı hedef değişken seçimi ile C5.0 algoritmasının kullanılması 89
5. SONUÇ VE ÖNERİLER 93
KAYNAKLAR 96
EKLER 103
EK-1. IBM SPSS Modeler 18.1 programı karar ağaçları analizi 103

Sayfa

EK-2. IBM SPSS Modeler 18.1 C5.0 algoritması karar ağacı analizi 108

EK-3. TÜİK B Grubu Mikro Veri Kullanımı İzin Formu 109

ÖZGEÇMİŞ 110

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 1.1. Hanelerin internete erişim imkânı ve internet bağlantı türleri 2
Çizelge 3.1. Diğer karar ağacı algoritmaları 29
Çizelge 3.2. Hava durumu verileri 31
Çizelge 3.3. Hava niteliği için kazanç ölçütü 32
Çizelge 3.4. İlk dallanma için kazanç değerleri 33
Çizelge 3.5. Güneşli değeri için nitelik tablosu 33
Çizelge 3.6. Isı niteliği için kazanç ölçütü 33
Çizelge 3.7. İkinci dallanma için kazanç değerleri 34
Çizelge 3.8. Karar ağacı algoritmalarının karşılaştırılması 41
Çizelge 3.9. Karışıklık matrisi 43
Çizelge 3.10. Performans kriteri hesaplama yöntemleri 44
Çizelge 4.1. Uygulamada kullanılan değişken bilgileri 50
Çizelge 4.2. Hanehalkı reisinin cinsiyeti 52
Çizelge 4.3. Hanehalkı reisinin yaşı 52
Çizelge 4.4. Hanehalkı reisinin okuma yazma durumu 52
Çizelge 4.5. Hanehalkı reisinin eğitim durumu 53
Çizelge 4.6. Hanehalkı reisinin çalışma durumu 53
Çizelge 4.7. Hanehalkı reisinin mesleği 53
Çizelge 4.8. Hanehalkı büyüklüğü 54
Çizelge 4.9. Hanede 0-25 yaş bireyin varlığı 55
Çizelge 4.10. Hanenin bulunduğu bölge 55
Çizelge 4.11. Hane masaüstü bilgisayar sahipliği 56
Çizelge 4.12. Hane dizüstü bilgisayar sahipliği 56
Çizelge 4.13. Hane tablet sahipliği 56

Çizelge **Sayfa**

Çizelge 4.14. Hane cep telefonu sahipliği 56
Çizelge 4.15. Hane oyun konsolu sahipliği 57
Çizelge 4.16. Hane sabit telefon sahipliği 57
Çizelge 4.17. Hane smart tv sahipliği 57
Çizelge 4.18. Hane bilgisayar kullanımı 57
Çizelge 4.19. Hane internet hizmeti sahipliği 58
Çizelge 4.20. Kümeleme sonucunda ortaya çıkan gelir grupları 59
Çizelge 4.21.C5.0 algoritması tahmin sayı ve yüzdesi 65
Çizelge 4.22. CHAID algoritması tahmin sayı ve yüzdesi 66
Çizelge 4.23. C&RT algoritması tahmin sayı ve yüzdesi 66
Çizelge 4.24. QUEST algoritması tahmin sayı ve yüzdesi 66
Çizelge 4.25. C5.0 algoritması karışıklık matrisi 66
Çizelge 4.26. CHAID algoritması karışıklık matrisi 66
Çizelge 4.27. C&RT algoritması karışıklık matrisi 66
Çizelge 4.28. QUEST algoritması karışıklık matrisi 67
Çizelge 4.29. Karışıklık matrisi ile hesaplanan performans kriterleri 67
Çizelge 4.30. Eğitim veri kümesi için en başarılı ve başarısız algoritmalar 71
Çizelge 4.31. Test veri kümesi için en başarılı ve başarısız algoritmalar 71

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 3.1. Veri madenciliği süreci....12
Şekil 3.2. CRISP-DM süreç diyagramı....14
Şekil 3.3. Veri madenciliğine etki eden disiplinler....16
Şekil 3.4. Veri madenciliğinde kullanılan tanımlayıcı ve tahmin edici yöntemler....17
Şekil 3.5. Kümeleme teknikleri....21
Şekil 4.1. Uygulamada kullanılan algoritmaların kaldıraç grafiği....69
Şekil 4.2. C5.0 algoritması karışıklık matrisi....87
Şekil 4.3. Önemli değişkenler için web grafiği....88
Şekil 4.4. Farklı bir hedef değişken için C5.0 algoritması karışıklık matrisi....91

RESİMLERİN LİSTESİ

Resim **Sayfa**

Resim 4.1. Gelir değişkeni için iki aşamalı kümeleme işlemi sonucu oluşan kümeler 58
Resim 4.2. Model özeti ve iki aşamalı kümeleme kalitesi 59
Resim 4.3. IBM SPSS Modeler ana ekranı 60
Resim 4.4. IBM SPSS Modeler tpye modülü 61
Resim 4.5. Karar ağacı ve iki aşamalı kümeleme ile modelleme 61
Resim 4.6. İki aşamalı kümeleme parametre seçim ekranı 62
Resim 4.7. Veri kümesinin eğitim ve test verisi olarak ayrılması 62
Resim 4.8. Karar ağacı algoritma seçimi 63
Resim 4.9. Seçilen algoritmaların varsayılan değerleri 64
Resim 4.10. Eğitim veri kümesinde karar ağacı algoritma performansları 70
Resim 4.11. Test veri kümesinde karar ağacı algoritma performansları 70
Resim 4.12. C5.0 algoritması için parametre seçim ekranı 72
Resim 4.13. Karar ağacı haritası 75
Resim 4.14. Değişken önem durumları 76
Resim 4.15. C5.0 algoritması ile oluşturulan karar ağacı 77
Resim 4.16. Birinci dallanmaya ait C5.0 analizi sonuçları 78
Resim 4.17. İkinci dallanmaya ait C5.0 analizi sonuçları 79
Resim 4.18. Üçüncü dallanmaya ait C5.0 analizi sonuçları 80
Resim 4.19. Dördüncü dallanmaya ait C5.0 analizi sonuçları 80
Resim 4.20. Beşinci dallanmaya ait C5.0 analizi sonuçları 81
Resim 4.21. Altıncı dallanmaya ait C5.0 analizi sonuçları 82
Resim 4.22. Yedinci dallanmaya ait C5.0 analizi sonuçları 82
Resim 4.23. Sekizinci dallanmaya ait C5.0 analizi sonuçları 83
Resim 4.24. Dokuzuncu dallanmaya ait C5.0 analizi sonuçları 84
Resim 4.25. Onuncu dallanmaya ait C5.0 analizi sonuçları 85

Resim **Sayfa**

Resim 4.26. Farklı bir hedef değişken için C5.0 algoritması karar ağacı.....90

Resim 4.27. Farklı hedef değişken için değişken önem durumları.....91

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklamalar**
dp	Doğru Pozitif
yp	Yanlış Pozitif
dn	Doğru Negatif
yn	Yanlış Negatif
%	Yüzde

Kısaltmalar	**Açıklamalar**
AIC	Akaike's Information Criterion (Akaike'nin Bilgi Kriteri)
AİK	Alıcı İşletim Karakteristiği
AUC	Area Under Curve (Eğri Altında Kalan Alan)
BIC	Bayesian Information Criterion (Bayes Bilgi Kriteri)
CRISP-DM	Cross Industry Standart Process for Data Mining (Çapraz Endüstri Veri Madenciliği Süreci)
C&RT	Classification and Regression Trees (Sınıflandırma ve Regresyon Ağaçları)
CHAID	Chi-Squared Automatic Interacton Detection (Ki-kare Otomatik Etkileşim Algılama)
HBTKA	Hanehalkı Bilişim Teknolojileri Kullanım Araştırması
ID3	Iterative Dichotomiser 3 (Tekrarlı İkilikçi Ağaç)
QUEST	Quick, Unbiased, Efficient, Statistical Tree (Hızlı, Yansız, Etkili İstatistiksel Ağaç)

Kısaltmalar	Açıklamalar
ROC	Receiver Operating Characteristic (Alıcı İşletim Karakteristiği)
TÜİK	Türkiye İstatistik Kurumu
YSA	Yapay Sinir Ağları

1. GİRİŞ

Bilgi çağının en etkili araçlarından biri olan internet, son yıllardaki gelişimi ile eğitim, sağlık, haberleşme, pazarlama ve ekonomi gibi pek çok alanı etkisi altına alma gücüne sahip olmuştur. İnternet dünyayı küresel bir köy haline getirdikçe ülkemizde yeni hanelerin internet hizmeti alma oranı artmaktadır. İnternet çekirdek aileye âdeta ailenin yeni bir üyesi olarak çok hızlı bir şekilde girmiştir [1]. Küresel dijital medya ajansı *We Are Social* tarafından her yıl yayınlanan *Digital in 2018* raporuna göre 2017 yılı verileri Türkiye'de aşağıdaki gibidir [2, 3].

- Türkiye'de 82 milyona yaklaşan ülke nüfusunun 54,3 milyonu internet kullanmaktadır.
- Mobil internet kullanıcısı sayısı 51,5 milyondur.
- Mobil telefon kullanıcısı sayısı toplam 59 milyon ile nüfusun % 73'üne ulaşmıştır.
- Türkiye'de mobil telefon kullanımı % 98, akıllı telefon kullanımı % 77 ve bilgisayar kullanımı %48'dir.
- Türkiye'de herhangi bir cihazla internette geçirilen toplam süre 7 saattir.
- İnternet kullanıcılarının %84'ü her gün internete girmektedir.
- İnternet trafiğinin % 62'si mobil cihazlardan, % 36'sı masaüstü ve dizüstü bilgisayarlardan gelmektedir.
- İnternet kullanıcılarının % 56'sı bir ürünü internette araştırmakta, % 43'ü ise satın alma işlemini gerçekleştirmektedir.

Bilişim teknolojileri ile ilgili istatistikler, bilgi toplumunda son yıllarda meydana gelen sosyal, kültürel ve ekonomik gelişmeleri anlamak, bu konuda uygulanan politikaları takip etmek ve piyasaların etkin çalışmasını sağlamak gibi nedenlerle büyük önem kazanmıştır. Bu istatistiklerin üretilmesi amacıyla ülkemizde de Türkiye İstatistik Kurumu (TÜİK) tarafından 2004 yılından bu yana Hanehalkı Bilişim Teknolojileri Kullanım Araştırması (HBTKA) gerçekleştirilmektedir. Bu araştırma ile hanelerde bulunan bilgi ve iletişim teknolojileri, bilgisayar, internet, e-ticaret, e-devlet uygulamaları, bilişim güvenliği alanlarında veri derlenmektedir [4-6]. 2016, 2017, 2018 yıllarında yapılan HBTKA sonuçlarına göre [4-6];

- 2018 yılında 16-74 yaş grubundaki bireylerde bilgisayar kullanım oranı % 59,6, internet kullanım oranı ise % 72,9'dur. Bu oran 2017 yılında bilgisayar kullanımı için % 56,6, internet kullanımı için % 66,8'dir. 2016 yılında ise bilgisayar kullanım oranı % 54,9 ve internet kullanım oranı % 61,2'dir
- Bilgisayar kullanım oranı 16-74 yaş grubundaki erkeklerde % 68,6, kadınlarda % 50,6'dır. İnternet kullanım oranı ise erkeklerde % 80,4, kadınlarda % 65,5'dir. 2017 yılı için ise bilgisayar ve internet kullanım oranları 16-74 yaş grubundaki erkeklerde % 65,7 ve % 75,1 iken, kadınlarda % 47,7 ve % 58,7'dir.
- 2017 yılının Nisan ayında hanelerin evden internete erişim oranı % 80,7 iken, 2018 yılı Nisan ayında hanelerin % 83,8'i evden internete erişim imkânına sahip olmuştur. Bu oran 2016 yılının aynı ayı için ise % 76,3'tür (Çizelge 1.1).
- Genişbant ile internete erişim sağlayan hanelerin oranı 2018 yılı Nisan ayında % 82,5 olmuştur. Buna göre hanelerin % 44,5'i sabit geniş bant bağlantı ile internete erişim sağlarken,% 79,4'ü mobil geniş bant bağlantı ile internete erişim sağlamıştır. Geniş bant internet erişim imkânına sahip hanelerin oranı 2017 yılında % 78,3 iken 2016 yılında bu oran % 73,1'dir (Çizelge 1.1).

Çizelge 1.1. Hanelerin internete erişim imkânı ve internet bağlantı türleri [4-6]

İnternet Erişim ve Bağlantı Türleri	2016 (%)	2017 (%)	2018 (%)
İnternet Erişimi	76,3	80,7	83,8
Genişbant Bağlantı	73,1	78,3	82,5
Darbant Bağlantı	3,2	2,4	7,6
Genişbant Bağlantı Türleri			
Sabit Genişbant (3G, 4,5G Cep Telefonu/Akıllı Telefon ya da Modem üzerinden)	39,5	40,0	44,5
Mobil Genişbant (3G, 4,5G Cep Telefonu/Akıllı Telefon ya da Modem üzerinden)	65,2	72,4	79,4

TÜİK'in 2016 yılında yaptığı HBTKA sonucu elde edilen verilerin kullanılması ile hazırlanan bu tezde, veri madenciliği sınıflandırma tekniklerinden karar ağaçları ve kümeleme tekniklerinden iki aşamalı küme kullanılmıştır. Tezin ilk bölümünde HBTKA araştırması ile ilgili genel bilgiler, dünyada ve Türkiye'de internet kullanımı ve hanelerin internet erişimi ile ilgili bilgiler verilmiştir. İkinci bölümde mevcut literatürün taraması yapılmış, HBKTA kullanılarak yapılan çalışmalar, TÜİK'in yaptığı diğer anket çalışmaları

ile karar ağaçları kullanılarak yapılan çalışmalar ile karar ağaçlarının karşılaştırıldığı çalışmalar incelenmiştir. Ayrıca iki aşamalı kümeleme tekniği kullanılan bazı çalışmalarda incelenmiştir. Üçüncü bölüm olan kavramsal çerçeve bölümünde, veri madenciliği kavramı, süreci, yöntemleri, karar ağaçlarının avantaj ve dezavantajları, karar ağaçları ve algoritmaları, sınıflandırma doğruluk belirleme ve performans karşılaştırma yöntemleri anlatılmıştır. Dördüncü bölüm uygulama ve bulgular bölümü olup, uygulamada kullanılan veriler, değişkenlere ait istatistiksel bilgiler, kullanılan yöntem ve veri madenciliği aracı, karar ağacı algoritmaları ile modelleme, sınıflandırma modellerinin performansları, modelleme sonucunda en başarılı olan algoritmanın seçimi, seçilen algoritma ile yapılan uygulama ve elde edilen bulgular açıklanmıştır. Ayrıca hedef değişken değiştirilerek farklı bir hedef değişken ile yeniden bir karar ağacı oluşturulması ve incelenmesi bu bölümde açıklanmıştır. Son bölüm ise sonuç ve öneriler olup, çalışma sonucunda elde edilen sonuçlar ve uygulama özetlenmiş, çalışma sonuçlarına göre önerilerde bulunulmuştur.

HBTKA anketi sonuçlarına ait veri kümesi kullanılarak yapılan çalışmalarda genellikle ekonometrik modeller kullanılmış ya da anket sonuçları ile ilgili istatistiksel analizler yapılmıştır. Bu çalışma HBTKA anketi veri kümesi ile veri madenciliği ve karar ağaçları teknikleri kullanılarak yapılmış nadir çalışmalardan biridir. Bu çalışma ile internet hizmeti sahipliğinin hanehalkı karakteristiklerine göre ne şekilde değiştiğinin karar ağaçları kullanılarak incelenmesi amaçlanmış ve sektörün ilgililerinin dikkatine sunulması hedeflenmiştir.

2. MEVCUT LİTERATÜRÜN İNCELENMESİ

Literatür taraması sırasında incelenen çalışmalar aşağıda özetlenmiştir.

Albayrak ve Yılmaz (2009), sanayi ve hizmet sektöründe faaliyet gösteren 173 işletmenin 2004-2006 yıllarına ait yıllık finansal göstergelerini kullanmış, firmaları ayıran en önemli değişkenlerin araştırılması için karar ağacı tekniklerinden CHAID analizini kullanmıştır. Çalışmada, firmaları ayırmada etkili olan değişkenlerin, işletme sermayesinin net satışlara oranı, stok devir hızı ve ekonomik rantabilite oranı olduğu görülmüştür [7].

Bir başka çalışmada Atılgan (2011), 2006 ve 2007 yıllarında Ankara ilinde meydana gelen ölümlü, yaralanmalı ve maddi hasarlı trafik kazalarını incelemiş, kaza sonucunda sürücünün ve yayanın ölü, yaralı ya da sağlam olma durumuna etki eden faktörleri veri madenciliği yöntemlerinden karar ağaçları ve birliktelik kuralları kullanarak belirleme çalışması yapmıştır. Çalışma sonucunda yolun bölünmüşlüğü ve yönü, sürücü ve yaya kusurları, yaş ve sürücünün eğitim durumlarının kazalar üzerinde etkili olduğu görülmüştür [8].

Kuzey (2012) doktora tezi çalışmasında, çalışan memnuniyetini etkileyen faktörleri Destek Vektör Makinaları ve karar ağaçları yöntemleri ile incelemiş, kullanılan algoritma performanslarını karşılaştırmış ve sonuçta başarı oranı en yüksek olan Destek Vektör Makinaları tekniği olmuştur. Farklı modellerle yapılan incelemeler sonucunda, çalışan memnuniyeti için, çalışma şartları ve ücret/ödüllendirme faktörlerinin en yüksek göreceli önem oranlarına sahip faktörler olduğu görülmüştür [9].

Heshmati, Al-Hammadany ve Bany-Mohammed (2013) makalelerinde, Irak Haberleşme Bakanlığı tarafından 15 834 kişiye Irak'ın farklı beş bölgesinde yapılan internet durum anketinden elde edilen verileri kullanmıştır. Katılımcılara ait yaş, cinsiyet, eğitim, istihdam, internet kullanmama nedenleri, internet kullanım şekilleri, internet yoğunluğu, yer ve internet edinme maliyeti değişkenlerini kullanarak değişkenler arasındaki bağlantıları Lojistik Regresyon Modeli ile analiz etmişlerdir. Sonuçlar internet maliyeti ve gelir seviyesi gibi ekonomik nedenlerin düşük seviye kullanım yoğunluğu oranları için temel neden olduğunu göstermiştir [10].

Coşkun ve Baykal (2013) yayınladıkları makalede, 1973 yılı itibarı ile başlanmış olan, kanser verileri farklı yılları kapsayan, farklı tümör tiplerine göre gruplar altında metin formatında, 118 nitelikten oluşan SEER olarak adlandırılan veri kaynağının 2008 yılına ait versiyonunu kullanarak göğüs kanseri hastalarının kayıtlarını incelemiş, hastaların hayatta olup olmadıkları, hayatta değil iseler ne kadar süre hayatta kaldıkları ve ölüm sebepleri göz önünde tutularak, herhangi bir hastanın hastalığı yenip yenemeyeceği ile ilgili C4.5, Naive Bayes, Lojistik Regresyon ve Kstar kullanarak modeller oluşturmuş ve oluşturulan modellerin başarım derecelerini karşılaştırmıştır. Çalışma sonuçlarına göre J48 algoritması %86,36 doğruluk derecesiyle en iyi sonucu vermiştir. Bu ölçüte göre J48 algoritmasını sırasıyla KStar, Lojistik Regresyon ve NaiveBayes algoritmaları izlemiştir. Bir diğer kriter olan F ölçütüne bakıldığında sıralamanın J48, KStar, Lojistik Regresyon ve NaiveBayes olduğu görülmüştür [11].

Karadağ (2014) yaptığı tez çalışmasında C&RT, CHAID, C4.5 ve Lineer Regresyon analizlerinin performanslarını benzetim verileri kullanarak karşılaştırmıştır. Bağımsız değişkenin tümü kategorik olduğunda duyarlılık, özgüllük, pozitif kestirim, negatif kestirim, doğruluk ve AUC kriterlerine göre doğru sınıflamada en yüksek oranın C4.5 algoritmasında olduğunu gözlemlemiştir. Sonuç olarak bağımsız değişkenin yapısı ve simülasyon deneme sayısı değişse de J48 algoritması kullanılan yöntem diğer yöntemlerden daha yüksek bir performans göstermiştir [12].

Alıç (2014) tez çalışmasında, Ankara ilindeki bir kamu hastanesindeki 206 adet Akut Pankreatit hastasına ait verileri kullanmıştır. Hastalığın riskinin ortaya konulmasına yönelik sınıflandırma yapılabilmesi için C5.0, C&R, CHAID, QUEST, Diskriminant Analizi, Lojistik Regresyon, Bayes Ağı, Destek Vektör Makinası, Yapay Sinir Ağı ve k-En Yakın Komşu sınıflandırıcılarını kullanmıştır. Tüm algoritmalar içinde en iyi performansı veren C5.0 algoritması olmuştur. Çalışma sonucunda elde edilen kuralların, hastalık üzerinde çalışan hekimlere önemli farkındalıklar ve açılımlar sağlayacağı düşünülmektedir [13].

Alkan, Abar ve Karaaslan (2015) yaptıkları çalışmada, TÜİK tarafından yapılan 2014 yılına ait HBTKA anketindeki mikro veri kümesini kullanarak hanelerdeki bilişim ekipmanları sayısı üzerinde etkili olan faktörleri Poisson regresyon modeliyle araştırmışlardır. TÜİK tarafından her yıl düzenli olarak yapılan bu ankette bilişim ekipmanları masaüstü bilgisayar, taşınabilir bilgisayar (dizüstü, netbook, tablet vb.), cep telefonu (akıllı telefonlar dâhil), sabit

hatlı telefon, oyun konsolu, dijital fotoğraf makinesi/kamera, DVD/VCD/DivX oynatıcı ve internete bağlanabilen TV (Smart TV) olarak belirlenmiştir. Kurulan modelde, hanede bulunan bilişim ekipmanları sayısı, bağımlı değişken olarak alınmıştır. Elde edilen sonuçlara göre, hanede internet erişim imkânının olması bilişim ekipmanı sayısını pozitif yönde etkilemektedir. Aylık gelir arttıkça bilişim ekipmanı sayısı da artmaktadır. Güneydoğu Anadolu'da bulunan illere göre diğer alt bölgelerdeki hanelerin bilişim ekipmanları sayısı fazladır. İki kişilik haneyle kıyaslandığında hanedeki birey sayısının 7 ve üzeri olması bilişim ekipmanı sayısını azaltmaktadır [14].

Demircan (2015) tez çalışmasında, TÜİK' in 2012 yılında yaptığı Yaşam Memnuniyeti Anketi sonuçlarına ait verileri kullanarak insanların genel memnuniyet ve umut düzeyini etkileyen faktörlerin belirlenmesi için sınıflandırma algoritmalarını kullanmış ve bu algoritmaların performanslarını değerlendirmiştir. Çalışma sonucunda bireylerin umut düzeyi, sağlık hizmetlerinden ve asayiş hizmetlerinden memnuniyet durumları sınıflandırma kuralları ile ortalama % 70 doğruluklarla elde edilmiştir [15].

Ercan (2016) doktora tezi çalışmasında Türkiye İstatistik Kurumu tarafından 2009, 2010, 2011 ve 2012 yıllarında yapılan Hanehalkı Bütçe Anketi verilerini kullanarak hanehalkının katı ve sıvı yağ tüketiminin etkileyen karakteristikleri ile ilgili faktörleri karar ağacı yöntemlerinden CHAID kullanarak analiz etmiştir. Sonuç olarak zeytinyağı ve tereyağı gibi zengin besin öğelerine sahip, sağlıklı ve pahalı yağ türlerinin tüketimini etkileyen faktörler genel olarak birbirine benzerdir. İki yağ türünde de en önemli faktörler gelir, hanehalkı reisinin yaşı ve eğitim düzeyi, hanehalkı büyüklüğü ve otomobil sahipliği olduğu görülmüştür. Margarin ve diğer bitkisel yağlar ile 216 yenilebilir yağ (mısır, ayçiçeği, kolza, vb.) gibi yağ türlerinin tüketimini etkileyen faktörlerin de birbirine benzer olduğu görülmüştür. İki yağ türünde de en önemli faktörlerin hanehalkı büyüklüğü, kır/kent durumu, hanehalkı reisinin yaşı ve cinsiyeti, gelir, sıfır-beş yaş birey sahipliği ve müstakil konut sahipliği olduğu görülmüştür [16].

Bir başka tez çalışmasında Diler (2016), Türkiye İstatistik Kurumu tarafından yayımlanan Yaşam Memnuniyeti Araştırmalarından elde edilen veri kümesi ile karar ağacı tekniklerinden CHAID ve C&RT algoritmalarını kullanarak mutluluk düzeyinin durumu ile ilgili çalışma yapmıştır. CHAID ve C&RT algoritmaları ile oluşturulan karar ağaçlarının sınıflama doğruluk oranlarının birbirine yakın olduğu görülmüştür [17].

Hane otomobil sahiplik durumunun ardışık logit model ile incelendiği bir çalışmada Tümsel (2016), TÜİK 2013 yılı Bütçe Anketi'ne katılanların tümü 9975, sürekli çalışan 3733 ve geçici veya sabit süreli sözleşmeli ve sözleşmesiz olarak çalışan 618 hanehalkına ait bilgilerden yararlanarak üç ayrı ardışık logit modeli incelemiştir. Model tahmini için hanehalkı reisinin cinsiyeti, mesleği, yaşı, çalıştığı süre, yıllık hanehalkının kullanılabilir geliri ve aylık harcamaları ele alınmıştır. Model tahmini sonucunda elde edilen bulgulara göre, hanehalkının otomobile sahip olmamasını en çok etkileyen değişkenin hanehalkının aylık harcamaları olduğu görülmüştür. Hanehalkının aylık harcamaları arttıkça arabaya sahip olma olasılığı azalmaktadır [18].

Bir diğer çalışmada Lakshmi, Indumathi ve Ravi (2016) aylık kontrol için hastanelere başvuran 600 hamile kadının yaş, boy, hamilelik süresi gibi kişisel bilgileri, sağlık bilgileri ve aile geçmişi bilgilerini kullanarak hamilelik dönemindeki risklerin tahmin edilmesi için karar ağacı yöntemlerinden C4.5 algoritmasını kullanmıştır. Çalışma sonuçları C4.5 karar ağacı sınıflayıcısının, standardize edilmemiş verilerden ziyade, standartlaştırılmış verilere uygulandığında hamilelikteki risk seviyelerini tahmin etmede daha yüksek bir potansiyele sahip olduğunu göstermiştir [19].

Ceylan, Gürsey ve Bulkan (2017) çalışmalarında, Türkiye'de faaliyet gösteren bir emeklilik şirketinin 2014 yılı içinde bireysel emeklilik sistemine dâhil olmuş, 2017 yılına kadar ödemelerine devam eden aktif 171 625 müşterisinin demografik, sosyokültürel bilgileri, ödeme ve tasarruf bilgilerini kullanmıştır. İki aşamalı kümeleme analizi tekniği ile müşterilere ait üç küme oluşturulmuştur. Çalışmanın sonuçları şirketlerin satış stratejileri sadece sözleşme adedi odaklı olarak değil müşterilerinin cinsiyet, medeni durum, yaş, meslek, öğrenim durumu, yaşadığı şehir bilgilerinin yansıra ödeme ve tasarruf bilgilerini de göz önünde bulundurarak oluşturması gerektiğini göstermiştir [20].

2013 yılı HBTKA mikro veri kümesi kullanılarak yapılan bir çalışmada Selim ve Balyaner (2017), Türkiye'de hem 6-15 yaş arası çocukların hem de yetişkinlerin sahip olduğu bilişim teknolojileri ürünleri sayısını belirleyen faktörleri sayma veri modeli ile incelemiştir. Çalışmada kullanım amaçları gibi faktörler, hanehalkındaki çocukların ve yetişkinlerin sahip olduğu bilişim teknolojileri ürünleri sayısının analizinde kullanılmıştır. Sonuçta elde edilen modellerde Robust Poisson Regresyon Modelinden faydalanılmıştır. Elde edilen tahminlerin geçerliliğini araştırmak amacıyla bootstrap tekniğine kullanılmıştır. Çalışma ile, bir

hanedeki bilişim teknolojileri ürünleri kullanımını etkileyen en önemli faktörlerin; hanehalkı geliri, yaş, cinsiyet, eğitim seviyesi, meslek ve yerleşim yeri olduğu tespit edilmiştir [21].

Bir diğer çalışmada Ahmed, Asif, Hina ve Muzammil (2017), 2010-2015 yılları arasında Pakistan ülkesinde yerel hisse piyasasını doğrudan etkileyen bazı ticari, politik ve siyasal olayların verilerini Google trendleri aracılığıyla toplamıştır. Bu veriler aracılığıyla internet aramaları ile sonraki borsa hareketlerinin nasıl etkilendiğini karar ağaçları ve Yapay Sinir Ağları yöntemleri ile incelemiş ve sonuçta karar ağaçları ile % 94 oranında başarı yakalanmıştır [22].

Topirceanu ve Grosseck (2017), öğrencilerin çevrimiçi kurslara katılım profilinin belirlenmesi amacıyla Romanya ülkesinde 632 öğrenciye sorulan anket sorularına verilen cevaplardan yararlanarak karar ağaçı tekniklerini kullanmıştır. Çalışma ile sınıflandırma ağaçlarının herhangi biri kullanılarak, eğitimcilerin her öğrenciyi neyin harekete geçirdiğini, neyin daha iyi etkileşime sokabileceğini tespit etmesi, böylece öğretmen ve öğrenci arasındaki mevcut mekânsal ve kültürel boşluğun kapatılması hedeflenmiştir [23].

Güzel (2018) tez çalışmasında, 155 hepatit hastasına ait hastalık bilgileri, kandaki değerler ve ultrasonik tetkikle ilgili bilgilere ait verileri kullanmış ve hastalığın ölümcül olup olmadığı ile ilgili teşhis analizi yapmıştır. Analizde karar ağacı, k-En Yakın Komşu, Naive Bayes, Çok Katmanlı Yapay Sinir Ağları, Yerine Koyarak Örnekleme (Bagging) ve Rastgele Altuzaylar (Random Subspaces) teknikleri kullanılmış, en başarılı algoritma % 92,90'lık başarı oranına sahip olan Yerine Koyma Yöntemi olmuştur. Yapılan çalışmada elde edilen sonuçlar önerilen analiz sisteminin hepatit hastalığı için tanı koymada tıbbi karar destek sistemi olarak kullanılabileceğini göstermiştir [24].

Börekçi (2018), 2015 yılına ait TÜİK hanehalkı tüketim harcamaları anketleriyle elde edilen verileri kullanarak, Türkiye'de hanehalkı telekomünikasyon harcamaları üzerinde etkili olan hanelerin ve hane reislerinin sosyo-demografik ve ekonomik özelliklerini belirlemek istemiştir. Çalışmada ekonometrik modeller olan Çift Sansürlü Model ve Heckman Metodu kullanılmıştır. Modelde yer alan değişkenlerden istatistiki olarak anlamlı bulunan ev sahibi, hane reisi evli, hane reisi emekli, hane reisi zorunlu sağlık sigortasına sahip, devlet ve özel ayni gelire sahip olan, soba ile ısınan, bir kişilik hanelerin diğerlerine göre daha az telekomünikasyon harcamalarında bulunduğu tespit edilmiştir [25].

Jimenez, Anupol, Cajal ve Gervilla (2018), 9 300 öğrencinin katıldığı alkol, sigara, esrar ve kokain kullanımı ile ilgili bir ankete verilen cevapları kullanarak veri madenciliği çalışması yapmıştır. Çalışmada karar ağaçları, Yapay Sinir Ağları, k-En Yakın Komşu, Naive Bayes ve Lojistik Regresyon tekniklerini kullanmışlardır. Sonuçlar, analiz edilen tüm tekniklerde alkol ve sigaranın, daha düşük doğru sınıflandırma yüzdesine sahip olduğunu göstermiştir ancak bu kullanım motifleri, esrar ve kokain kullanımının öngörülmesinde daha iyi sınıflandırma performansına sahiptir. Kullanılan veri madenciliği teknikleri ile madde kullanım riskini ve koruyucu faktörleri analiz etmek için yol gösterici bir çalışma yapılmıştır [26].

Mevcut literatür incelendiğinde HBTKA anket verileri kullanılarak yapılmış çalışmalarda genellikle istatistiksel teknikler ve ekonometrik modeller kullanılmıştır. Bazı çalışmalarda regresyon tekniği kullanılmış ve genelde hanede yer alan bilişim cihazlarının sayısını belirleyen faktörler araştırılmıştır. HBTKA anketi verileri ile karar ağaçları teknikleri kullanılarak yapılmış çalışma yoktur. TÜİK tarafından yapılan yaşam memnuniyeti araştırmaları, hanehalkı tüketim harcamaları anketi, hanehalkı bütçe anketi gibi diğer veri kümeleri kullanılarak yapılmış veri madenciliği çalışmaları mevcuttur. Bunun haricinde farklı karar ağaçları kullanılarak en başarılı karar ağacının seçildiği ve buna göre modelleme yapıldığı çalışmalar incelenmiştir.

3. KAVRAMSAL ÇERÇEVE

Yapılan işlemler bilgisayarların insan hayatına girmesi ile daha kolay kayıt altına alınmaya başlanmıştır. Marketlerden alınan ve iade edilen her ürün, o ürün ile birlikte alınan ürünler ile ilgili kayıtlar bilgisayarlarda, veri tabanlarında tutulmaya başlanmıştır. Bütün bu bilgiler bir yığın halinde depolanırken içlerinde birbiri ile ilişkili gizli bilgileri de tutmaktadır. Tüm bu bilgiler veri tabanlarından çıkarılıp işlenmeyi bekleyen tıpkı bir maden gibidir ve bilgiye dönüşmeyi beklemektedir [27]. Bilgisayar dosyalarında ve veri tabanlarında tutulan veri miktarı olağanüstü bir oranda büyümeye devam ederken bu verilerin kullanıcıları da onlardan daha karmaşık bilgi beklemektedir. Basit yapılandırılmış sorgu dili sorguları, bu artan bilgi taleplerini desteklemek için yeterli gelmemektedir [28]. Veri tabanları günümüzde terabaytlarla ifade edilmektedir. Bu büyük hacimde verinin içinde stratejik önem taşıyan gizli enformasyonlar yatmaktadır. Ancak bu kadar büyük hacimli veri içerisinde yer alan önemli bilgi ya da bilgilerin nasıl açığa çıkarılacağı en önemli sorudur. Bu önemli soruya en güncel yanıt, hem geliri artıran hem de maliyetleri indirgeyen veri madenciliği teknikleridir [29].

3.1. Veri Madenciliği Nedir?

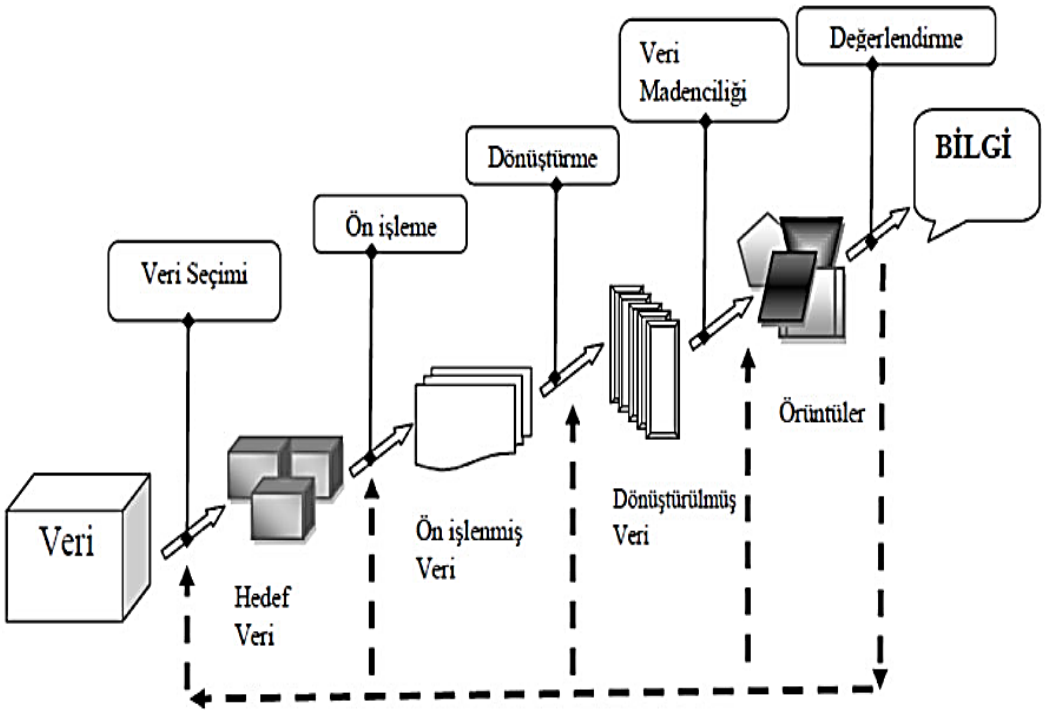
Veri madenciliği ile ilgili birçok akademisyen, kuruluş, araştırmacı ve yazar tarafından farklı tanımlamalar yapılmaktadır. Bu tanımlamalardan bazıları aşağıda belirtilmiştir.

Veri madenciliği genellikle bir veri tabanında gizli bilgi bulmak olarak tanımlanır. Alternatif olarak, araştırıcı veri analizi, veri odaklı keşif ve tümden gelim öğrenmesi olarak adlandırılmıştır [28]. Büyük miktarda veriden ilginç desenleri ve bilgileri keşfetme işlemidir [30]. Veri madenciliği daha önceden bilinmeyen, keşfedilmemiş, geçerli ve uygulanabilir bilgilerin veri tabanları kullanılıp elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır [27]. Veri madenciliği, çeşitli kaynaklardan çeşitli şekillerde toplanan verilerin üzerinde işlem yapılarak anlamlı bilgilerin çıkarılmasıdır [31]. Veri madenciliği; önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından kümeleme, veri özetleme sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve aykırılık tespiti gibi farklı birçok tekniğin kullanıldığı dinamik bir süreç ile elde edilmesidir [32].

Veri madenciliği, belirli bir veri koleksiyonundan çeşitli modelleri, özetleri ve türetilmiş değerleri keşfetme sürecidir [33]. Anlamlı desen ve kuralları keşfetmek için büyük miktarlardaki verilerin araştırılması ve analizidir [34]. İstatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerini kullanarak, depolama ortamlarında saklanmış bulunan veri yığınlarının elenmesi ile anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilmesi sürecidir [35]. Veri madenciliği, operasyonel olarak, çok miktarda veride otomatik veya yarı otomatik olarak kalıpları keşfetme sürecidir [36].

3.2. Veri Madenciliği Süreci

Veri madenciliği ve veri tabanlarında bilgi keşfi konusunda, veriden bilgiye ulaşmak için gidilen yolda geliştirilmiş olan çeşitli süreçler bulunmaktadır. Özellikle SPSS'in liderliğini yaptığı CRISP-DM bu süreçlerden biridir. Fayyad, Piatetsky-Shapiro ve Smith tarafından 1996 yılında yayımlanan From Data Mining to Knowledge Discovery in Databases isimli makalede belirtilen veri madenciliği süreci Şekil 3.1'de gösterilmiştir [37, 38].



Şekil 3.1. Veri madenciliği süreci [38]

Belirtilen bu süreç 5 adımdan oluşmaktadır.

Veri Seçimi (Data Selection)

Veri tabanından belirlenen hedef verilerin üretilmesi için yapılan işlemdir. Buradaki amaç hedef veri kümesini oluşturacak verileri, veri tabanından tipik olarak ayıklama işlemidir [16].

Ön İşleme (Preprocessing)

Bir süreçte kullanılacak veriler yanlış veya eksik olabilir. Farklı veri türleri ve metrikleri içeren birden çok kaynaktan alınmış anormal veriler olabilir. Hatalı veriler (genellikle veri madenciliği araçları kullanılarak) önişleme sürecinde düzeltilebilir veya kaldırılabilir, ancak eksik veriler sağlanmalı veya tahmin edilmelidir [28].

Dönüştürme (Tranformation)

Veri birleştirme olarak da bilinir, seçilen verilerin madencilik prosedürüne uygun formlara dönüştürüldüğü aşamadır [39].

Veri Madenciliği (Data Mining)

Potansiyel olarak yararlı kalıpları çıkarmak için akıllı tekniklerin uygulandığı en önemli adımdır [39]. Çeşitli yöntemler kullanılarak, bu aşamaya kadar gelen veriden, anlamlı sonuçları olan, yorumlamaya uygun ve çıkarılmak istenen bilgiye hizmet edecek işlemler yapılır [31].

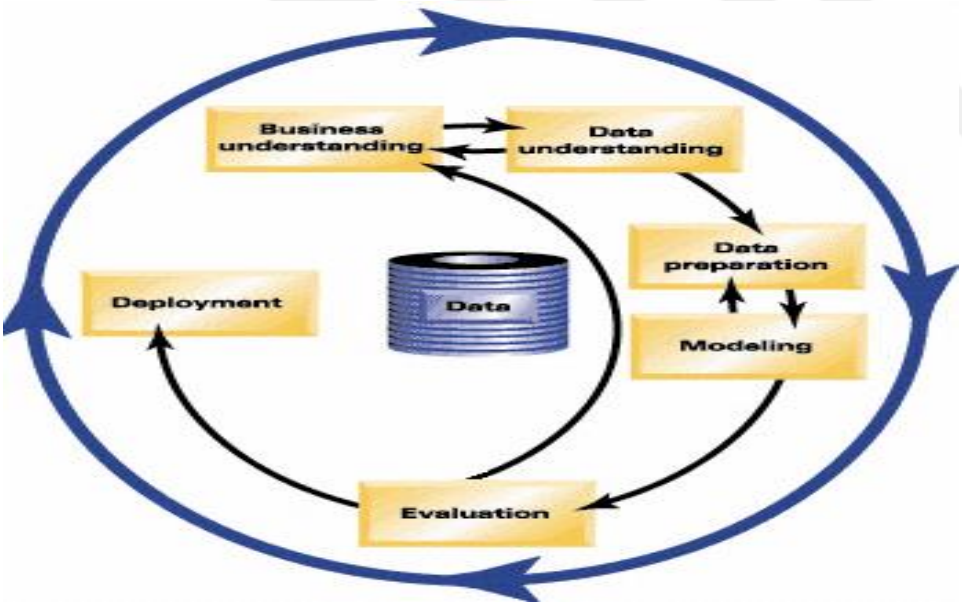
Değerlendirme (Evaluation)

Bu aşamada oluşan örüntüler değerlendirilerek birinci aşamada belirlenen amaç ile alaka kurulmalıdır ve elde edilen bilgi amaca yönelik kullanılmalıdır [40].

3.2.1. CRISP-DM süreç modeli

Veri madenciliği, bir dizi farklı beceri ve bilgi gerektiren yaratıcı bir süreçtir. Veri madenciliği projeleri yürütmek için standart bir çerçeve olmaması bir veri madenciliği

projesinin başarısının ya da başarısızlığının, onu yürüten kişi ya da kuruluşa büyük ölçüde bağlı olduğu ve başarılı uygulamaların işletme genelinde tekrarlanmasının gerekmediği anlamına gelmektedir [41]. Bazı danışmanlık şirketleri temelde benzer adımları içeren kendine özgü yöntemler geliştirmiştir. Bu yöntemlerden biri de 1996 yılında Daimler Chrysler, SPSS, AG, NCR ve OHRA gibi lider veri madenciliği kullanıcıları ve tedarikçilerinden oluşan bir konsorsiyum tarafından geliştirilmiş olan CRISP-DM (Cross Industry Standart Process for Data Mining)'dir. Türkçe karşılığı *Endüstri Bağımsız Veri Madenciliği Süreci Standardı* olan CRISP-DM otomotiv, uzay endüstrisi, telekomünikasyon, eğitim ve sağlık gibi birçok sektörde kullanım alanı bulmuştur [17]. CRISP-DM' ye göre, belirli bir veri madenciliği projesi Şekil 3.2'de gösterildiği gibi altı aşamadan oluşan bir yaşam döngüsüne sahiptir. Döngüde sıradaki sonraki aşama genellikle bir önceki aşama ile ilişkili sonuçlara bağlıdır. Fazlar arasındaki en önemli bağımlılıklar oklarla gösterilir [42]. Döngüdeki aşamalar aşağıda anlatılmıştır.



Şekil 3.2. CRISP-DM süreç diyagramı [41]

İşin anlaşılması (Business understanding)

Başlangıç aşaması olan bu aşama, proje hedeflerinin ve gereksinimlerinin iş perspektifinden anlaşılmasına ve ardından bu bilginin bir veri madenciliği problemi tanımına dönüştürülmesine odaklanır. Ayrıca hedeflere ulaşmak için tasarlanmış bir ön proje planı tasarlanmasını sağlar [41].

Verinin anlaşılması (Data understanding)

Verinin anlaşılması aşaması öncelikle verilerin toplanması ile başlar. Daha sonra verilerin tanınması, keşfedilmesi ve kalitesinin değerlendirilmesi aşamasına geçilir. Son olarak eğer istenirse farklı alt desenler seçilerek çeşitli amaçları gerçekleştirebilecek faaliyetler yapılır [42].

Veriyi hazırlama (Data preparation)

Mevcut veri kaynakları belirlendikten sonra veri seçilmeli, temizlenmeli, istenilen şekle getirilmeli ve biçimlendirilmelidir. Veri modelleme hazırlığında veri temizleme ve veri dönüşümü bu aşamada gerçekleşmektedir. Bu aşamada ek modeller uygulanması sebebiyle, iş anlayışına dayalı kalıpları görme fırsatı ortaya çıkmaktadır. Verilere ön işleme yapılmasının amacı, seçilen veriyi temiz ve kaliteli olması için hazırlamaktır [43].

Modelleme (Modelling)

Modelleme aşaması, veri madenciliği yazılımları ile değişik durumlar için sonuçlar üretme aşamasıdır. İlk olarak kümeleme analizi ve verilerin görsel olarak keşfi yapılır. Verinin türüne bağlı olarak değişik modeller daha sonra uygulanabilir [16]. Verilerin formunu belirli bir veri madenciliği tekniğinin özel gereksinimleri ile uyumlu hale getirmek için veri hazırlama aşamasına geri dönüş yapılması gerekebilir [42].

Değerlendirme (Evaluation)

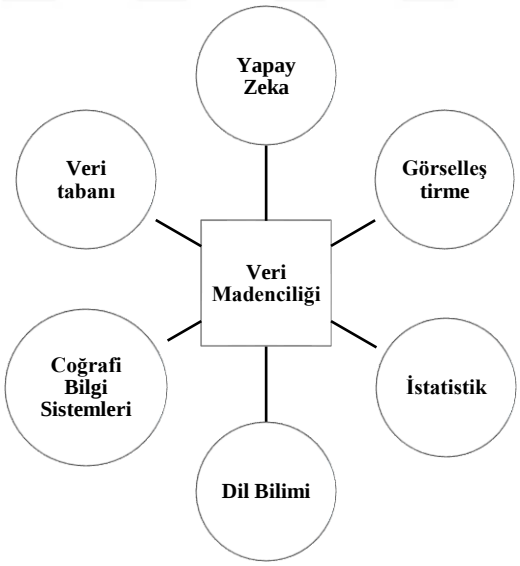
Veri madenciliği yazılım araçlarından görselleştirme (veri ve ilişki kurma) ve kümeleme analizi (hangi değişkenler birlikte daha ilişkili ise) ilk analiz için yararlıdır. Veri kavrandıktan sonra veri tipine uygun daha ayrıntılı model ve kurallar geliştirilmektedir. Verilerin türüne bağlı olarak, çeşitli modeller uygulanabilir [43]. Modelleme aşamasında kullanılan model ya da modeller sahada kullanılmak üzere dağıtılmadan önce kalite ve etkinlik açısından değerlendirilmelidir. Son olarak veri madenciliği sonuçlarının nasıl kullanılması gerektiği ile ilgili bir karara varılmalıdır [42].

Kullanıcılarla paylaşım (Deployment)

Modelin oluşturulması genellikle projenin sonu değildir. Kazanılan bilginin, müşterinin kullanabileceği şekilde organize edilmesi ve sunulması gerekecektir. Gereksinimlere bağlı olarak, kullanıcılarla paylaşım aşaması bir rapor oluşturmak kadar basit veya tekrarlanabilir bir veri madenciliği sürecinin uygulanması kadar karmaşık olabilir. Çoğu zaman paylaşım adımlarını gerçekleştirecek olan veri analisti değil, kullanıcı olacaktır. Her durumda yaratılan modellerden faydalanabilmek için hangi eylemlerin gerçekleştirilmesi gerektiğini açıklamak önemlidir [41].

3.3. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri genellikle istatistik, yapay zekâ ve yapay zekânın uzantısı olan makine öğrenmesi olmak üzere iki ana kökten beslenmektedir. Bu iki disiplinin yanında Şekil 3.3'de görüldüğü gibi veri tabanı yönetim sistemleri, dil bilim, görselleştirme ve coğrafi bilgi sistemleri de veri madenciliği yöntemlerinde önemli role sahiptir [37].

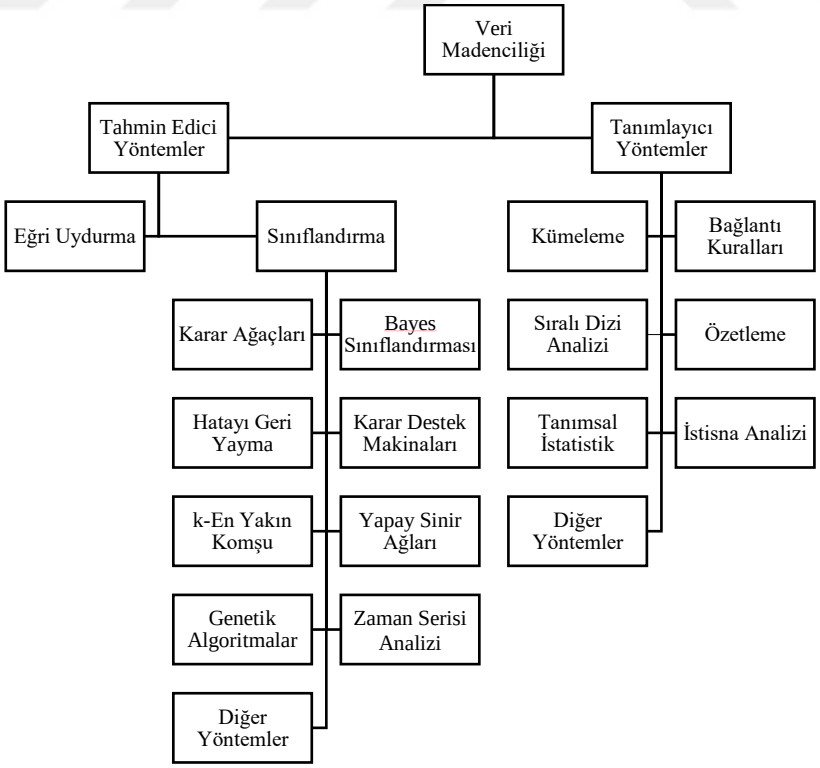


Şekil 3.3. Veri madenciliğine etki eden disiplinler [37]

Veri madenciliği ile ilgili kullanılan pek çok yöntemin yanına hemen her geçen gün yeni yöntem ve algoritmalar eklenmektedir. Bunlardan bir kısmı onlarca yıldır kullanılan klasik teknikler denilen ağırlıklı olarak istatistiksel yöntemlerdir [44]. Veri madenciliğinde

kullanılan yöntemler, Tahmin edici (Predictive) ve Tanımlayıcı (Descriptive) olmak üzere iki ana başlığa ayrılabilir.

Tahmin edici yöntemler, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanmaktadır. Örneğin, bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişkenler kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun olarak kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceği tahmininde kullanılmaktadır. Tanımlayıcı yöntemlerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki bağlantıların tanımlanması sağlanmaktadır. X/Y aralığında geliri ve iki veya daha fazla arabası olan çocuklu aileler ile çocuğu olmayan ve geliri X/Y aralığından düşük olan ailelerin satın alma örüntülerinin birbirlerine benzerlik gösterdiğinin belirlenmesi tanımlayıcı modellere bir örnektir [45]. Tanımlayıcı ve tahmin edici yöntemler Şekil 3.4'de gösterilmiştir [15].



Şekil 3.4. Veri madenciliğinde kullanılan tanımlayıcı ve tahmin edici yöntemler [15]

Veri madenciliği yöntemleri ana işlevlerine göre üç ana başlık altında toplanmaktadır.

1) Sınıflandırma (Classification)
2) Kümeleme (Clustering)
3) Birliktelik Yöntemleri (Association Rules)

Sınıflandırma modeli tahmin edici, kümeleme ve birliktelik kuralları tanımlayıcı modellerdir [45].

3.3.1. Sınıflandırma

Sınıflandırma (classification), günlük hayatta çok sık meydana gelen bir görevdir. Esas olarak, nesneleri, her biri, sınıflar olarak bilinen birbirini tamamlayan ve özel kategorilerden birine atanacak şekilde bölmeyi içerir. Her bir nesnenin tam olarak bir sınıfa verilmesi, yani asla birden fazla ve hiç bir zaman hiç bir sınıfa verilmemesi gerektiği anlamına gelir [46]. Genel olarak sınıflandırma için gerek istatistik gerekse makine öğrenimi disiplinlerinde birbirinden bağımsız olarak fakat hemen hemen aynı amaca hizmet eden çeşitli yöntemler geliştirilmiştir. Matematikte y = f(x), y değerinin x' in veya x değerlerinin bir fonksiyonu olduğunu gösterir. Sınıflandırma amaç, bu bağlantıyı sağlayacak fonksiyonun tespit edilmesidir. Çok değişkenli yöntemlerde bu fonksiyon Eş. 3.1'deki gibi yazılabilir.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + \varepsilon \quad (3.1)$$

Bu eşitlikte yer alan x_i değerleri öznitelikleri ifade eder. Ayrıca makine öğreniminde feature, istatistikte bağımsız değişken (independent variable) kavramları disipline özel olarak kullanılmaktadır. Eşitlikte yer alan b değerleri ise kat sayıları ifade etmektedir. Bu durumda tüm yöntemlerin ortak amacı, y değeri ile x değerleri arasındaki bağıntıyı sağlayacak b kat sayılarının hesaplanmasıdır. Bu eşitliğin tam olarak sağlanamaması durumunda ortaya çıkan fark ise ε gösterilmiştir. x ve y değerlerinin hangi ölçü skalasında oldukları analiz yönteminin belirlenmesinde büyük önem taşır. Bu değişkenler kategorik ve sürekli ölçü skalasında olabilirler. Ayrıca kategorik ölçü skalasında yer alan değişkenler nominal ve ordinal, sürekli ölçü skalasında yer alan değişkenler aralıklı ve oransal olabilirler [37].

Sınıflandırma süreci, bir sınıflandırma modelinin oluşturulduğu bir öğrenme aşamasından ve modelin verilen veriler için sınıf etiketlerini tahmin etmek için kullanıldığı bir sınıflandırma adımından oluşan iki aşamalı bir süreçtir [30]. Sınıflandırma işleminde, sınıf veya istatistik alanındaki deyimiyle bağımlı değişken hem sınıfsal hem de sürekli değer taşıyabilir. Bu anlamda regresyon ve çok terimli regresyona yaklaşmaktadır. Ancak veri madenciliği çerçevesinde bu istatistiksel yöntemlerin dışında sınıflandırma işleminde Bayesyen sınıflandırma algoritması, karar ağaçlarına dayalı algoritmalar, Yapay Sinir Ağları temelli algoritmalar ve k-En Yakın Komşu algoritması gibi birçok teknik ve algoritma geliştirilmiştir [27]. Örneğin, bir sınıflama modeli banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli geliri ve mesleği verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir [45]. Sınıflandırma modellerinde kullanılan başlıca teknikler [47];

- Karar Ağaçları (Decision Tree)
- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic Algoritms)
- k-En Yakın Komşu (k-Nearest Neighbor)
- Bellek Temelli Nedenleme (Memory Based Reasoning)
- Naive Bayes
- Lojistik Regresyon (Logistic Regression) olarak gösterilebilir.

Birçok pratik karar verme görevi, sınıflandırma problemleri olarak formüle edilebilir; diğer bir deyişle, çeşitli kategorilerden birine insan veya nesne atamak bir sınıflandırmadır [46]. Örneğin [46];

- Bir süpermarkette belirli bir ürünü satın alması veya almaması muhtemel müşteriler
- Belli bir hastalığa yakalanma riski yüksek, orta veya düşük olan kişiler
- Öğrenci projeleri arasında başarılı veya başarısız notu için bir ayrım yapmak
- Araçlara, insanlara, binalara veya ağaçlara karşılık gelen bir radar ekranındaki nesneler
- Suça karışmış bir insana yakından benzeyen, az benzeyen veya hiç benzemeyen insanlar
- On iki ay içinde değer olarak yükselmesi muhtemel olan, değer düşüşüne uğrayan ya da değişmeyen bir değere sahip olan evler
- Gelecek on iki ay içinde yüksek, orta veya düşük trafik kazası riski olan kişiler

- Her bir siyasi parti için oy kullanma ihtimali olan ya da olmayan kişiler
- Bir hava durumu tahmini için ertesi gün yağmurun büyük olasılıkla, muhtemelen, olası değil, çok olası durumlarına göre yağma ihtimali birer sınıflandırmadır.

3.3.2. Kümeleme

Kümeleme (clustering) analizi X veri matrisinde yer alan ve doğal grupları kesin olarak bilinmeyen birimleri, değişkenleri ya da birim ve değişkenleri birbirleri ile benzer olan alt kümelere (grup, sınıf) ayırmaya yardımcı olan yöntemler topluluğudur. Kümeleme analizi, birimleri değişkenler arası benzerlik ya da uzaklıklara dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplar oluşturmaya çalışır [48]. Kümeleme modelinde, sınıflama modelinde olan veri sınıfları yoktur yani verilerin herhangi bir sınıfa aitliği bulunmamaktadır. Sınıflama modelinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiğinde bu verinin hangi sınıftan olabileceği tahmin edilmektedir. Kümeleme modelinde ise, sınıfları bulunmayan veriler gruplar halinde kümelere ayrılırlar. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir ön işlemi gibi görev alabilmektedir [45]. Kümeleme çeşitli işletme ve devlet yönetimi uygulamalarının yanında antropoloji, telekomünikasyon, uzay araştırmaları, uzaktan algılama, biyoenformatik, tıp, kimya, sosyoloji ve jeoloji gibi alanlarda önemini artırmaktadır [37]. Marketlerde farklı müşteri gruplarının keşfedilmesi ve bu grupların alışveriş örüntülerinin ortaya konması, biyolojide bitki ve hayvan sınıflandırmaları ve işlevlerine göre benzer genlerin sınıflandırılması, şehir planlanmasında evlerin tiplerine, değerlerine ve coğrafik konumlarına göre gruplara ayrılması, web üzerinde bilgi keşfi için dokümanların sınıflanması gibi uygulamalar tipik kümeleme uygulamalarıdır. Veri kümeleme güçlü bir gelişme göstermektedir. Veri tabanlarında toplanan veri miktarının artmasıyla orantılı olarak, kümeleme analizi son zamanlarda veri madenciliği araştırmalarında aktif bir konu haline gelmiştir [45]. Kullanıcının amacına ve kullanım alanına göre kümeleme analizinin amaçları aşağıdaki gibi sıralanabilir [15]:

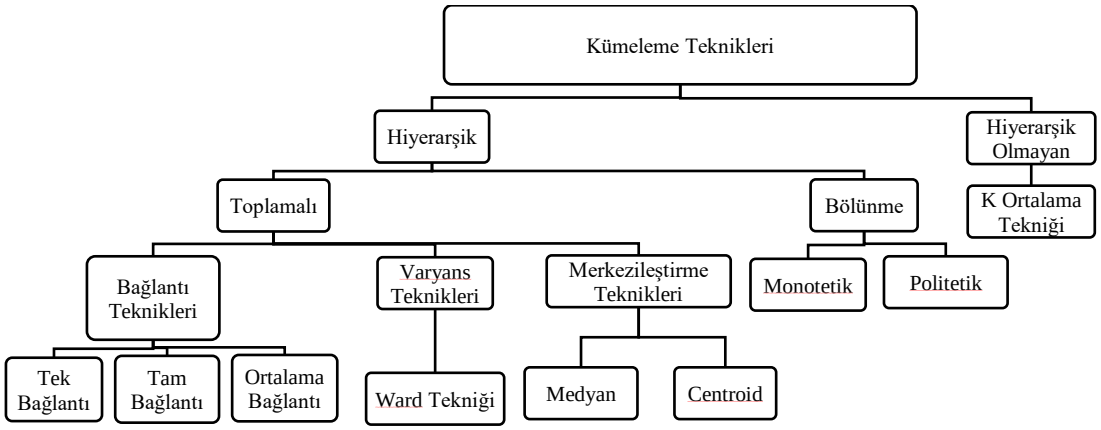
- Doğru tiplerin belirlenmesi
- Model oluşturmak
- Gruplara dayalı tahmin
- Hipotez testi

- Veri araştırma (inceleme)
- Hipotez oluşturma
- Veri indirgeme

Kümelemenin matematiksel tanımı şu şekilde yapılabilir. D= { X_1, X_2, X_3, … , X_n }, n=1, 2…m değerlerini alırken D kümesi veri tabanı olmak üzere her bir X_n bir kaydı temsil etmektedir. X= { x_1, x_2, x_3, … , x_i }, i=1, 2…m değerlerini alırken x_i, ad, soyad, yaş ve gelir gibi özellikleri temsil etmektedir. Kümelemedeki amaç D veri tabanını, j adet K kümesine bölmek ve $K_j \subseteq D$ koşulunun sağlanmasıdır. Kümeleme analizinde sınıflandırmadan farklı olarak belirlenecek kümelerin özellikleri önceden bilinmemektedir ve ortaya çıkacak küme sayısı da belli değildir. Ancak algoritmaların zaman karmaşıklığını, alınacak sonuçların kullanılabilirliğini artırmak için, literatürdeki algoritmaların bir kısmı ya küme sayısını ya da her bir kümede bulunacak eleman veya bu elemanlar arasındaki minimum-maksimum benzerlik uzaklık ölçütünü kullanıcıdan ister [27]. Kümeleme analizi için birçok algoritma öne sürülmüştür. Ancak literatürde bu algoritmalar iki başlık altında toplanmıştır:

1) Hiyerarşik kümeleme teknikleri
2) Hiyerarşik olmayan kümeleme tekniği

Her iki teknikte de ortak amaç kümeler arasındaki farklılıkları ve kümeler içi benzerlikleri en yüksek düzeye çıkarmaktır. Yani, küme içi homojenlik arttırılırken kümeler arası homojenlik ise azaltılmaktadır. Hangi tekniğin kullanılacağı küme sayısına bağlı olmakla birlikte her iki tekniğin beraber kullanılması çok daha yararlıdır. Kümeleme tekniklerinin genel yapısı Şekil 3.5'de gösterilmiştir [49].



Şekil 3.5. Kümeleme teknikleri [49]

İki aşamalı kümeleme

Punj ve Steward (1983) tarafından önerilen ve hiyerarşik olmayan bir kümeleme analiz yöntemi olan iki aşamalı kümeleme analizi, veri yapısında hem kategorik hem de sürekli değişkenlerin olması durumunda kullanılabilen ve Ward'ın Minimum Varyans Yöntemi ile *k ortalamalar* yönteminden oluşan hibrid bir kümeleme yaklaşımıdır. Bu yöntemin en önemli özelliği, büyük veri setlerinde uygulanabilmesi, elde edilen kümelerle uyum sağlamayan gözlemlerin istendiğinde veriden ayıklanabilmesi ve AIC (Akaike's Information Criterion) ya da BIC (Schwarz's Bayesian Information Criterion**)** kriterlerine göre en uygun küme sayısını otomatik olarak belirleyebilmesidir [50]. İki aşamalı kümeleme analizinde, süreç ön kümeleme ve kümeleme olmak üzere iki ana aşamadan oluşmaktadır [32]. Her iki aşamada da kullanılabilecek iki uzaklık ölçüsü Log-Olabilirlik uzaklığı ve Öklit uzaklığıdır. Log-Olabilirlik uzaklığı veride sürekli ve kategorik değişkenlerin olması durumunda kullanılabilmektedir. Olasılık temelli bir yaklaşıma dayanmaktadır [50].

3.3.3. Birliktelik Kuralları

Veri tabanlarındaki bilgi arttıkça, birçok kuruluş sahip oldukları bilgiler arasındaki ilişkileri ortaya çıkarma çabası içine girmiştir. Yığınlar halindeki bilgilerin arasındaki ilişkilerin keşfedilmesi, kuruluşlar için önemli sonuçlar oluşturacak kararlar alınmasını sağlayabilir. İlişki analizi satış, pazarlama, ürün katalog tasarımları gibi birçok alanda kullanılmaktadır. Örneğin herhangi bir ürünü satın alırken, bu ürünün yanında başka ürün ya da ürünlerin satın alınması bu ürünler arasındaki bağlantıyı ortaya koyar. Bu bağlantıların ortaya çıkarılması ve bunun bir kural olarak belirlenmesi birliktelik kuralları (association rules) ya da ilişki analizinin konusudur. Literatürde bu çalışmalara *Pazar ya Market Sepeti Analizi* denilmektedir [27]. Market sepet analizi denmesinin nedeni; birlikteliklerin modellenerek müşterilerin hangi ürünleri bir arada aldıkları bilgisinin ortaya çıkması ve market yöneticilerinin de bu bilgi ışığında daha etkili satış stratejileri geliştirebilmesidir. Bağlantı Analizi aynı zamanda kişi davranışlarından hareketle sahtekârlık tespitinde, hangi davranışın hangi eylem ile beraber yapıldığında olağan dışı değerlendirilmesi gerektiğine öngörü sağlamaya da yardımcı olur [51]. Birliktelik kuralları, bir veri kümesinde bir veya birden fazla değişkenin diğer değişkenlerle olan birlikteliğini gösterir. Böylece eğer-sonra (if-then) durumları üretilerek değişkenler arasında gizli kalmış önemli ilişkiler ortaya çıkarılır [52]. Birliktelik kurallarının kullanım alanlarına örnek olarak [42];

- Ebeveynleri kendileri için iyi okuyucular olan çocukların okuma oranlarını incelemek;
- Telekomünikasyon ağlarında bozulmayı tahmin etmek;
- Bir süpermarkette hangi öğelerin birlikte satın alındığını ve hangi öğelerin birlikte satın alınmadığını öğrenmek;
- Yeni bir ilacın tehlikeli yan etkiler göstereceği vakaların oranını belirlemek gösterilebilir.

Birliktelik kuralları için güven, destek, eşik değer kavramlarından bahsedilmesi gereklidir. D kümesi, T işlemlerinin oluşturduğu bir küme olsun. Bu D kümesi içindeki her T işlemi, I'nın içerdiği bir nesneler kümesini temsil etmektedir. Örneğin A nesnelerinden oluşan bir küme ve B nesnelerinden oluşan başka bir küme olduğu düşünülürse; *Eğer A ise, B* biçiminde A'nın öncül, B'nin ise sonuç olarak adlandırıldığı bir birliktelik kuralı oluşabilmektedir. A→ B biçiminde bir birliktelik kuralının destek değeri (D kümesinde) hem A'yı hem B'yi içeren işlemlerin oranıdır (Eş 3.2). Güven değeri ise kuralın doğruluğunun bir ölçüsüdür (Eş 3.3).

$$\text{Destek} = P(A \cup B) = \frac{\text{A ve B'yi içeren işlemlerin sayısı}}{\text{Toplam işlemlerin sayısı}} \tag{3.2}$$

$$\text{Güven} = P(B / A) = \frac{P(A \cap B)}{P(A)} = \frac{\text{A ve B'yi içeren işlemlerin sayısı}}{\text{A'yı içeren işlemlerin sayısı}} \tag{3.3}$$

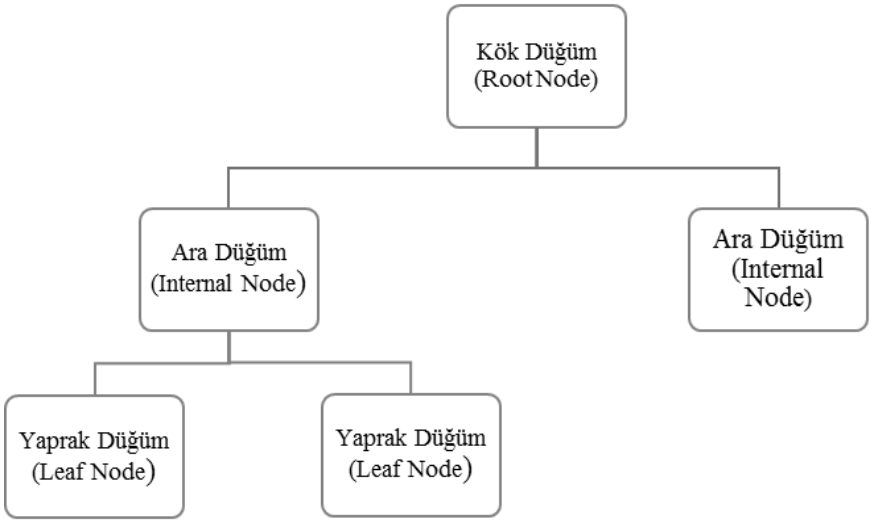
Birliktelik kuralları belirlenirken yukarıda sözü edilen destek ve güven ölçütlerinin yanı sıra bu değerleri karşılaştırmak için bir eşik değere ihtiyaç vardır. Hesaplanan destek veya güven ölçütlerinin bu eşik destek ve eşik güven değerlerinden büyük olması beklenir. En küçük destek değerine sahip nesne kümeleri geniş, diğerleri ise küçük nesne kümesidir. Hesaplanan destek veya güven ölçütleri ne kadar büyük ise birliktelik kuralları da o derece güçlüdür [8,53]. Veriler arasında ilişkilerin ortaya çıkarılması ve birliktelik kuralları oluşturulabilmesi için çeşitli algoritmalar geliştirilmiştir. Bu algoritmalar aşağıda belirtilmiştir [27, 37].

- AIS Algoritması
- SETM Algoritması
- Apriori Algoritması
- AprioriTid Algoritması

- Eclat (Equivalence Class Transformation)
- Fp Growth (Frequent Pattern)

3.4. Karar Ağaçları

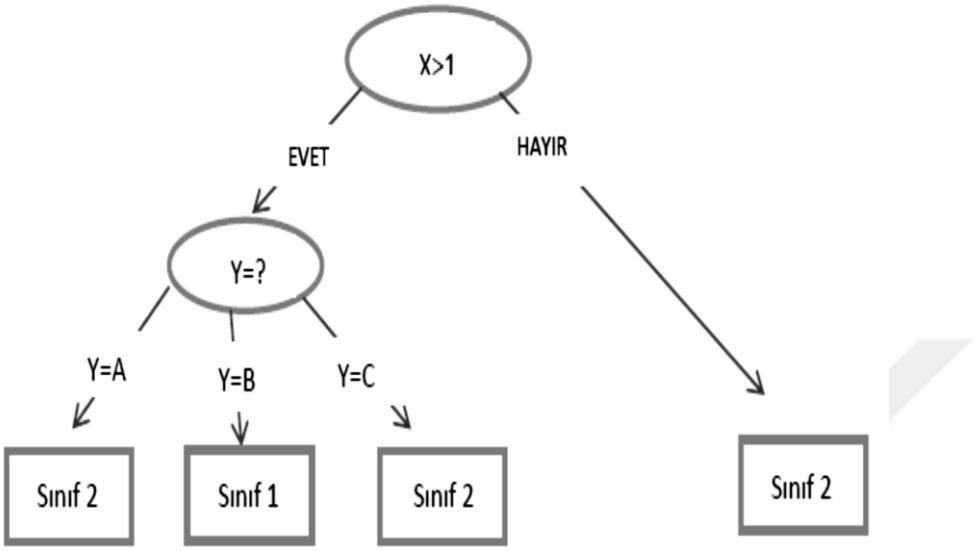
Karar ağaçları, 1973 yılında Bierman ve Friedman tarafından önerilmiş olan değişkenlerin parçalanarak bir ağaç oluşturulması mantığına dayanan, literatürde yaygın kullanıma sahip bir istatistiksel sınıflandırma algoritmasıdır [24]. Karar ağaçları ve karar kuralları, birçok gerçek dünya uygulamasında sınıflama problemlerine güçlü bir çözüm olarak uygulanan veri madenciliği metodolojisidir. Verilerden sınıflandırıcılar üretmek için kullanılan en etkili yöntemlerden biri, bir karar ağacı oluşturmaktır [33]. Bir karar ağacı, doğal bir ağaçta olduğu gibi kök (root), dal (branch) ve yapraklardan (leaf) oluşmaktadır. Bir karar ağacında bu oluşum, kök düğüm, yaprak olmayan veya ara düğüm (non-leaf node / internal node) ve bir karar ağacının sona erdiği noktalar olan yaprak düğüm (leaf node) kavramları ile açıklanır. Şekil 3.6' da bir karar ağacı çıktısı görülmektedir. Şekilden görüleceği üzere kök düğümden başlayarak her hiyerarşide veri dizisinin belirli kriterlere göre bölünmesi (split) gerçekleşmektedir. Seçilen karar ağacı algoritmasına göre her aşamada ikili (binary), üçlü (tertiary) ya da çoklu bölünme işlemi gerçekleşebilir [37].



Şekil 3.6. Bir karar ağacı çıktısı [37]

Aşağıda tipik bir karar ağacı örneği yer almaktadır (Şekil 3.7). Burada X ve Y'den oluşan iki giriş niteliğine sahip bir örnek sınıfının basit karar ağacı görülmektedir. X> 1 ve Y= B

değerini taşıyan örnekler Sınıf 1'de; yine X> 1, Y= A ve Y= C koşullarına uygun olanlar Sınıf 2'de yer almaktadır. Ancak Y'nin değerini göz önüne almadan X≤1 koşuluna uygun örnekler Sınıf 2'de yer alır [12].



Şekil 3.7. X ve Y nitelikleri üzerine uygulanan testleri içeren basit bir karar ağacı [12]

Karar ağaçları kullanılarak verinin sınıflandırılması iki basamaklı bir işlemdir. İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması ile analiz edilir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci basamak olan sınıflama basamağında ise test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir bir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır. Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır. Eğer modelin doğruluğu kabul edilebilir bir değerde ise model, sınıfı bilinmeyen yeni verilerin sınıflanması amacıyla kullanılabilir. Örneğin, bir eğitim verisi incelenerek kredi durumu sınıfını tahmin edecek bir model oluşturulacak olsun. Bu modeli oluşturan bir sınıflama kuralı;

IF yas = "41...50" AND gelir = yüksek THEN kredidurumu = mükemmel

şeklinde ise bu kurala göre yaşı "41...50" kategorisinde olan (yaşı 41 ile 50 arasında olan) ve gelir düzeyi yüksek bir kişinin kredi durumunun mükemmel olduğu söylenebilir.

Oluşturulan bu modelin doğruluğu, bir test verisi ile onaylandıktan sonra model, sınıfı belli olmayan yeni bir veriye uygulanabilir ve sınıflama kuralına göre yeni verinin sınıfı *mükemmel* olarak belirlenebilir [45]. Karar ağacı temelli analizlerin genel olarak kullanıldığı alanlar şunlardır [7]:

- Belirli bir sınıfın olası üyesi olacak ya da olmayacak elemanların belirlenmesi,
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi farklı kategorilere ayrılması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok sayıdaki değişkenden en önemli olanlarının seçilmesi,
- Gelecekteki olayların gerçekleşme durumlarının tahmin edilebilmesi için çeşitli kuralların oluşturulması,
- Sadece belirli bir alt gruba özgü olan ilişkilerin tanımlanması,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikli değişkenlere dönüştürülmesidir.

3.4.1. Karar ağaçlarının avantaj ve dezavantajları

Bir sınıflandırma aracı olarak karar ağaçlarının literatürde birçok avantajından bahsedilmektedir [9,37].

- Karar ağaçları kendini açıklayıcı özelliğe sahiptir ve yoğun olduğu zaman bile takip edilmesi kolaydır. Eğer karar ağacı makul sayıda yapraklara sahip ise, kullanıcılar tarafından kolayca anlaşılabilir. Ayrıca, karar ağaçları kurallar kümesine dönüştürülebilir. Böylece, anlaşılması ve yorumlanması daha kolay hale gelir [9].
- Karar ağaçları hem nominal (kategorik) hem de sayısal (sürekli) girdi ile işlem yapabilir [9].
- Karar ağacının gösterimi herhangi bir ayrık değerli sınıflandırıcıyı ifade etmek için oldukça zengindir [9].
- Karar ağaçları parametrik olmayan bir metot olarak kabul edilir. Bundan dolayı, karar ağaçları uzay dağılımı ve sınıflandırıcının yapısı hakkında varsayımlara sahip değildir [9].

- Diğer yöntemler verinin normal dağılmamasına veya eksik değerlere karşı hassasken, karar ağacı öğrenimi modellerinde kullanılan veri fazla bir önişleme gerek duyulmadan kullanılabilir [37].
- Çeşitli istatistiksel testler yapılarak bir modelin güvenilirliği test edilebilir [37].
- Çok büyük boyutlardaki veri kümesine kolayca uygulanabilir [17].
- En çok kullanılan veri madenciliği modeli olduğu için bu alan da yayımlanmış birçok doküman ve bilgisayar programı bulunmaktadır [17].

Bununla birlikte, karar ağaçları bazı dezavantajlara da sahiptir.

- Birçok karar ağacı algoritması, hedef nitelik için sadece kesikli ayrık değerler (discrete values) gerektirmektedir [9].
- Karar ağaçlarının eğitme kümesine, alakasız niteliklere karşı aşırı duyarlı olması bir dezavantajdır [9].
- Karar ağaçlarında "böl ve yönet" metodu kullandığı gibi, eğer çok sayıda çok alakalı nitelik mevcut ise çok iyi bir performans göstermektedirler, fakat eğer birçok karmaşık etkileşim mevcut ise daha az performans gösterebilirler [9].
- Karar ağaçlarında veri dizisinin bölünmesi nitelik seçiminde kullanılan kriterlerin yetkinliğine bağlıdır. Sezgisel özelliği sahip bu kriterlerin en uygun bölme gerçekleştireceğinin bir garantisi olmadığı için elde edilen sonuçlar genel optimum sonuçlar olmayabilir [37].
- Karar ağaçları algoritmaları çoğu zaman eğitim verisini genelleştiremediğinden, ezbere öğrenme olabilir. Bu durum budama yöntemlerinin etkin kullanımını gerektirebilir [37].

3.5. Karar Ağacı Algoritmaları

Karar ağacını temel alarak geliştirilen algoritmalar genel olarak aşağıdaki kaba kodlardan oluşur (Şekil 3.8).

```
D: Öğrenme veri tabanı
T: Kurulacak Ağaç
T = 0              // başlangıçta ağaç boş küme
Dallara ayırma kriterlerini belirle
T = kök düğümü belirle
T = dallara ayrılma kurallarına göre kök düğümü dallara ayır;
     her bir dal için;
do
   Bu düğüme gelecek değişkeni belirle
   if ( durma koşuluna ulaşıldı )
         Yaprak ekle ve dur
   else
loop
```

Şekil 3.8. Karar ağacı kaba kodları [27]

Karar ağacı kurulurken eldeki veri tabanının bir kısmı öğrenme bir kısmı da test için kullanılır. Ağaç oluşturulurken oluşturulan modelin çalışıp çalışmadığı belirlenir. Eğer ağaç belirlenen düzeyde çalışıyorsa dallanma durdurulur ve sınıflandırma tamamlanır. Programdaki durdurma kriterine göre ağacın hassasiyeti de belirlenmiş olur. Geç durdurulan bir ağaç fazla dallanacak, ağaç daha geniş olacak ve çalışma süresi uzayacaktır. Bu durumda ağaç daha duyarlı sonuçlar üretecektir. Erken durdurulan bir ağaç ise hızlı çalışmış olsa da tam öğrenmenin gerçekleşmeme durumu ortaya çıkabilir [27]. Aşırı uyum (overfitting) karar ağacı modelleri için önemli bir sorundur. Öğrenme algoritmasını etkileyecek şekilde eğitim seti hatalarını azaltmaya devam edildiğinde aşırı uyum olur. Bir karar ağaç inşasında aşırı uyumdan kaçınmak için genel olarak iki yaklaşım kullanılır.

- Ön budama (Pre-pruning): Ağacın büyümesini sınıflandırma işleminden önce durdurmak
- Sonraki budama (Post-pruning): Ağaçtaki gereksiz kısımları öncelikle tüm ağacı oluşturup daha sonra çıkarmak

Bir modelde ne zaman pruning (budama) işlemi yapılacağının belirlenmesinin zor olması sebebiyle ilk yaklaşım pek kullanılmamaktadır. İkinci yaklaşım çok daha başarılıdır [54, 55].

Bu yaklaşımda aşağıdaki adımlara dikkat edilmelidir [54, 55]:

- Budama işlemine karar vermek için eğitim verisinden farklı bir veri kümesi kullanılmalıdır. Bu veri kümesine doğrulama veri kümesi (validation dataset) denir. Doğrulama veri kümesi gereksiz düğümlere karar vermek için kullanılır.
- Bir karar ağacı elde ettikten sonra, hata tahmini (error estimation) ve önem testi (significance testing – chi square testing) gibi istatiksel metotlar kullanarak eğitim verisi üzerinde budama ve genişleme olup olmayacağına karar verilir.

Karar ağaçlarının oluşturulmasında en önemli nokta hangi değişkenin ilk düğüm, yani kök düğüm olacağının belirlenmesidir. Bunun için çeşitli ölçütler belirlenmiştir ve her farklı ölçüt bir karar ağacı algoritmasına karşılık gelmektedir. Bu algoritmalar şu şekilde sıralanabilir [8]:

- Entropiye dayalı algoritmalar: ID3, C4.5, C5.0
- Sınıflandırma ve regresyon ağaçları (C&RT): Twoing, Gini algoritmaları
- Bellek tabanlı sınıflandırma algoritmaları: k-En Yakın Komşu
- İstatistiğe dayalı algoritmalar: Bayesyen sınıflandırma, CHAID

Algoritmaları birbirinden ayıran temel özellikler ise kullanılan ölçü skalası, her düğümden ortaya çıkan yeni düğümlerin sayısı, ağacın büyümesini durdurma kriteri, en iyi bölen özniteliğin seçilmesi ve budama sürecidir [37]. Karar ağacı algoritmalarının en önemlileri ID3, C4.5 ve C5.0, C&RT, CHAID ve QUEST algoritmalarıdır. Karar ağaçları için geliştirilen diğer algoritmalar Çizelge 3.1'de özetlenmiştir.

Çizelge 3.1. Diğer karar ağacı algoritmaları [27, 37, 56]

Algoritma	**Açıklama**	**Referans**
CAL5	Özellikle sayısal değerli nitelikler için tasarlanmıştır.	Muller ve Wysotzki (1994)
FACT	QUEST algoritmasının daha eski sürümüdür. Her düğümü ayırmak adına bir öznitelik seçmek için istatistiksel testler kullanır ve ardından ayrılma noktasını bulmak için diskriminant analizi kullanır.	Loh ve Vanichsetakul (1988)
LMDT	Çok değişkenli testlere dayalı bir karar ağacı niteliklerin doğrusal kombinasyonlarını oluşturur.	Brodley ve Utgoff (1995)

Çizelge 3.1. (devam) Diğer karar ağacı algoritmaları [27, 37, 56]

T1	Yalnızca bir özelliği kullanarak örnekleri sınıflandıran tek düzey bir karar ağacı algoritmasıdır. Eksik değerler *özel değer* olarak kabul edilir. Hem sürekli hem de nominal nitelikleri destekler.	Holte (1993)
PUBLIC	Sayısal karmaşıklığı azaltmak için MDL maliyetini kullanarak büyümeyi ve budamayı bütünleştirir.	Rastogi ve Shim (2000)
MARS	Melez lineer regresyonların uyumlandırılması ile esnek modellerin kurulmasını sağlar.	Friedman (1991)
OCI	Sürekli değere sahip öznitelikler için tasarlanmıştır.	Murty ve Salzberg (1994)
SLIQ	Önce genişlik ilkesi ile aynı anda birçok yaprak oluşturur. Kategorik verileri alt kümelere ayırmak için hızlı bir algoritma kullanır.	Rissanen, Agrawal ve Mehta (1989)
SPRINT	SLIQ algoritmasına benzer. Her bir değişken için ayrı bir değişken listesi hazırlar.	Khoshgoftaar ve Seliya (2003)

3.5.1. ID3 algoritması

Bu algoritma J. Ross Quinlan tarafından 1986 yılında geliştirilmiştir. ID3 (Iterative Dichotomiser 3) algoritması çok basit bir karar ağacı algoritması olarak kabul edilir. ID3, bilgi kazancını (information gain) bölme kriteri olarak kullanır. Tüm örnekler, tek bir hedef özellik değerine ait olduğunda veya en iyi bilgi kazanımı sıfırdan büyük olmadığı zaman büyüme duraklar. ID3 herhangi bir budama prosedürü uygulamaz ve sayısal nitelikleri veya eksik değerleri işlemez [56]. ID3 algoritmasında sınıflandırma için en ayırıcı özelliğe sahip değişken bulunurken entropi kavramından yararlanılır [27]. Bilgiyi ölçmek için kullanılan kavrama entropi denir. Entropi, bir veri kümesindeki belirsizlik, sürpriz veya rastgelelik miktarını ölçmek için kullanılır. Bir kümedeki tüm veriler tek bir sınıfa ait olduğunda, hiçbir belirsizlik yoktur. Bu durumda entropi sıfırdır [28]. Entropi 0-1 arasında bir değer alır. Bütün olasılıklar $< p_1, p_2 \ldots p_n >$ olasılıkları ifade etmek üzere tüm olasılıkların toplamı 1 olmalıdır (Eş 3.4).

$$\sum_{i=1}^{n} p_i = 1 \tag{3.4}$$

Bu durumda entropi aşağıdaki gibi hesaplanır (Eş 3.5).

$$\text{Entropi} = H\left(p_1, p_2 \ldots\ p_n\right) = \sum \left(p_i \log\left(1/p_i\right)\right) \quad (3.5)$$

ID3 algoritması veri tabanı bölünmeden önce gelen bilgi ile bölündükten sonra gelen bilgi arasındaki farkı kullanarak öncelikli düğüm ve dallanmalara karar verir [27]. Hesaplanan entropi değeri ile alt bölüme ait entropi değerleri arasındaki fark kazanım (gain) olarak adlandırılır. Bu değer aşağıdaki gibi hesaplanır (Eş 3.6).

$$\text{Kazanım} = G(D) = H(D) - \sum_{i=1}^{n} P(D_i)\, H(D_i) \quad (3.6)$$

Eş 3.6'ya göre D bir değişkeni temsil ederken D_i değişkenin i. düzeyini temsil etmektedir. Kazanım değeri hangi değişken için daha büyükse o değişken üzerinden dallandırma yapılır [57]. ID3 algoritması Çizelge 3.2'deki 15 günlük hava durumu verilerinin alındığı bir örnekle aşağıda açıklanmıştır [13].

Çizelge 3.2. Hava durumu verileri

Özellikler				**Hedef**
Hava	**Isı**	**Nem**	**Rüzgâr**	**Oyun**
Güneşli	Sıcak	Yüksek	Hafif	Hayır
Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
Bulutlu	Sıcak	Yüksek	Hafif	Evet
Yağmurlu	Ilık	Yüksek	Hafif	Evet
Yağmurlu	Soğuk	Normal	Hafif	Evet
Yağmurlu	Soğuk	Normal	Kuvvetli	Hayır
Bulutlu	Soğuk	Normal	Kuvvetli	Evet
Güneşli	Ilık	Yüksek	Hafif	Hayır
Güneşli	Soğuk	Normal	Hafif	Evet
Yağmurlu	Ilık	Normal	Hafif	Evet
Güneşli	Ilık	Normal	Kuvvet	Evet
Bulutlu	Ilık	Yüksek	Kuvvet	Evet
Bulutlu	Sıcak	Normal	Hafif	Evet
Yağmurlu	Ilık	Yüksek	Kuvvetli	Hayır

Çizelge 3.2'deki hava, ısı, nem ve rüzgâr değişkenlerinin değerlerine göre, hedef değer olan Golf oyununun oynanıp oynanamayacağının belirlenmesinde bir karar ağacı oluşturulacağı için hedef değerin entropisi hesaplanmıştır.

Oyun nitelik değerlerinden oluşan küme T kümesi olsun. T kümesinde 9 adet Evet, 5 adet Hayır değeri vardır. Buna göre T kümesinin entropisi;

$$\text{Entropi (T)} = -\left(\frac{5}{14}\log_2\left(\frac{5}{14}\right) + \frac{9}{14}\log_2\left(\frac{9}{14}\right)\right) = 0{,}940$$

Hava niteliği için kazanç ölçütü hesaplanmasında, her bir değer için nitelik içinde tekrar sayıları aşağıdadır. Her bir nitelik için kazanç değerleri hesaplanır.

Çizelge 3.3. Hava niteliği için kazanç ölçütü

Hava	**Oyun**
Güneşli	Hayır
Güneşli	Hayır
Güneşli	Hayır
Güneşli	Evet
Güneşli	Evet
Yağmurlu	Hayır
Yağmurlu	Hayır
Yağmurlu	Evet
Yağmurlu	Evet
Yağmurlu	Evet
Bulutlu	Evet
Bulutlu	Evet
Bulutlu	Evet
Bulutlu	Evet

$T_{\text{güneşli}}$ değişkeninde 5 kayıt için 2 Evet, 3 Hayır değeri vardır. $T_{\text{yağmurlu}}$ değişkeninde5 kayıt için 3 Evet, 2 Hayır değeri vardır. T_{bulutlu} değişkeninde 4 kayıt için 4 Evet, 0 Hayır değeri vardır.

$$\text{Entropi}\left(T_{\text{güneşli}}\right) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) = 0{,}971$$

$$\text{Entropi}\left(T_{\text{yağmurlu}}\right) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) = 0{,}971$$

$$\text{Entropi}\left(T_{\text{bulutlu}}\right) = -\left(\frac{4}{4}\log_2\left(\frac{4}{4}\right) + \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right) = 0{,}000$$

Bu durumda Entropi (Hava, T):

$$\text{Entropi (Hava, T)} = \frac{5}{14}\text{Entropi}\left(T_{\text{güneşli}}\right) + \frac{5}{14}\text{Entropi}\left(T_{\text{yağmur}}\right) + \frac{4}{14}\text{Entropi}\left(T_{\text{bulutlu}}\right) = 0{,}694$$

Kazanç (Hava, T) = Entropi (T) - Entropi (Hava, T) = 0,246

Benzer şekilde diğer nitelikler için de kazanç hesabı yapılır. Son durumda nitelikler için kazanç değerleri aşağıdaki tabloda gösterilmiştir (Çizelge 3.4).

Çizelge 3.4. İlk dallanma için kazanç değerleri

Nitelik	Kazanç
Hava	0,246
Isı	0,029
Nem	0,151
Rüzgâr	0,048

Bu durumda en büyük kazanç 0,246 değeri ile Hava niteliğine aittir. Dolayısıyla hava niteliği kök düğümdür ve güneşli, yağmurlu, bulutlu olmak üzere 3 dala ayrılır. İkinci adımda *güneşli, bulutlu ve yağmurlu* değerleri için dallanmaların hangi nitelikler seçilerek devam edeceğini belirlemek için hesaplamalar yapılır.

Çizelge 3.5. Güneşli değeri için nitelik tablosu

Hava	**Isı**	**Nem**	**Rüzgâr**	**Oyun**
Güneşli	Sıcak	Yüksek	Hafif	Hayır
Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
Güneşli	Ilık	Yüksek	Hafif	Hayır
Güneşli	Soğuk	Normal	Hafif	Evet
Güneşli	Ilık	Normal	Kuvvetli	Evet

T kümesinde 5 değer vardır. Bunların 3 adeti Hayır, 2 adeti Evet değeridir.

$$\text{Entropi (T)} = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) = 0{,}971$$

Karar ağacının *Güneşli* değeri için ikinci dallanmasında hangi niteliğe göre dallanacağı belirlenir. Bu nedenle her bir niteliğin kazanç değeri hesaplanır.

Çizelge 3.6. Isı niteliği için kazanç ölçütü

Isı	**Oyun**
Sıcak	Hayır
Sıcak	Hayır
Soğuk	Evet

Çizelge 3.6. (devam) Isı niteliği için kazanç ölçütü

Ilık	Hayır
Ilık	Evet

$T_{sıcak}$ değişkeninde 2 kayıt için 2 Hayır değeri, $T_{soğuk}$ değişkeninde 1 kayıt için 1 Evet değeri vardır. $T_{ılık}$ değişkeninde 2 kayıt için 1 Evet 1 Hayır değeri vardır.

$$\text{Entropi}\ (T_{sıcak}) = -\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0{,}000$$

$$\text{Entropi}\ \left(T_{soğuk}\right) = -\left(\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) = 0{,}000$$

$$\text{Entropi}\ (T_{ılık}) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1{,}000$$

$$\text{Entropi}\ (\text{Isı, T}) = \frac{2}{5}\text{Entropi}\ (T_{sıcak}) + \frac{1}{5}\text{Entropi}\ \left(T_{soğuk}\right) + \frac{2}{5}\text{Entropi}\ (T_{ılık}) = 0{,}400$$

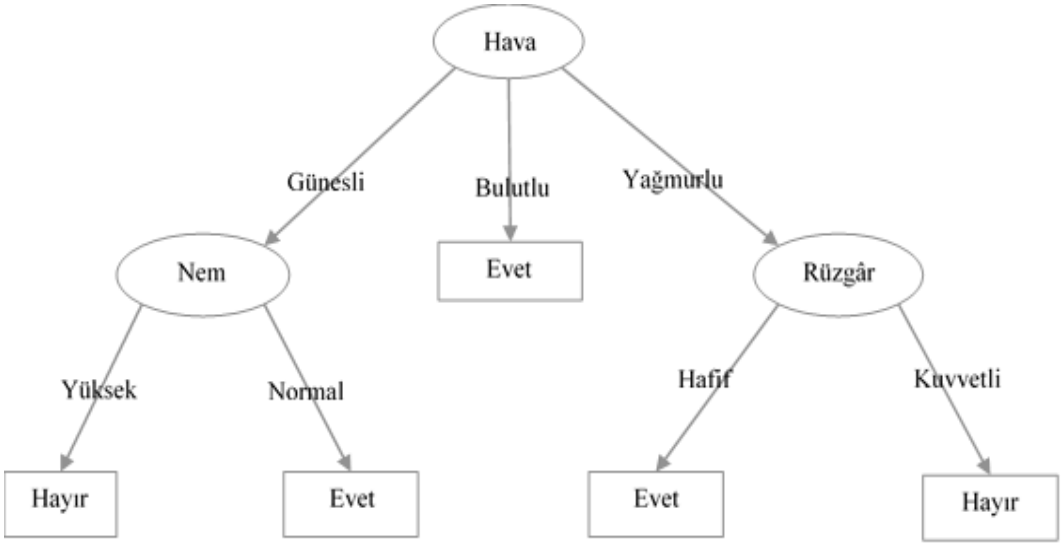
$$\text{Kazanç}\ (\text{Isı, T}) = \text{Entropi}\ (\text{T}) - \ \text{Entropi}\ (\text{Isı, T}) = 0{,}571$$

Benzer şekilde diğer nitelikler için de kazanç hesabı yapılır. Son durumda nitelikler için kazanç değerleri aşağıdaki çizelgede gösterilmiştir (Çizelge 3.7).

Çizelge 3.7. İkinci dallanma için kazanç değerleri

Nitelik	**Kazanç**
Isı	0,571
Nem	0,971
Rüzgâr	0,020

Tabloda en büyük kazanç 0,971 değeri ile Nem niteliğine aittir. Dolayısıyla Nem niteliğine göre karar ağacı dallandırılır. Yukarıda belirtilen süreç, tüm veriler sınıflandırılana kadar devam edecektir. Son durumdaki karar ağacı Şekil 3.9'da gösterilmiştir.



Şekil 3.9. ID3 algoritması karar ağacı

Oluşan karar ağacına göre aşağıdaki kurallar elde edilir.

Eğer (Hava=”Güneşli” ve Nem=”Yüksek”) ise Oyun=”Hayır”
Eğer (Hava=”Güneşli” ve Nem=”Normal”) ise Oyun=”Evet”
Eğer (Hava=”Bulutlu”) ise Oyun=”Evet”
Eğer (Hava=”Yağmurlu” ve Rüzgâr=”Hafif”) ise Oyun=”Evet”
Eğer (Hava=”Yağmurlu” ve Rüzgâr=”Kuvvetli”) ise Oyun=”Hayır”

3.5.2. C4.5 ve C5.0 algoritmaları

ID3 algoritmasının geliştirilmiş hali C4.5 algoritmasıdır. C5.0 algoritması ise C4.5’in geliştirilmiş hali olup, özellikle büyük veri setleri için kullanılmaktadır [58]. C5.0 algoritması, C4.5'in tüm işlevlerini içerir ve modellemede bir dizi yeni teknolojiyi uygular. Bunların arasında en önemli uygulama, örneklerin tanımlanmasının doğruluk oranını arttırmak için kullanılan *boosting* tekniğidir. Bir diğer önemli uygulama ise maliyet duyarlı karar ağaçlarını oluşturmasıdır [59]. C4.5 algoritmasında ID3’e göre aşağıdaki iyileştirmeler yapılmıştır [16].

- C4.5 budama işlemi yapar.
- C4.5 eksik veriler ile çalışma yeteneğine sahiptir.
- C4.5 sürekli öznitelikleri, özniteliğin değer aralığını iki alt kümeye bölerek (ikili bölme) kullanır.

- C4.5 kazanım oranı kriterlerini maksimuma yükselten en iyi eşiği arar. Eşiğin üzerindeki her değer ilk alt kümeyi, diğer tüm değerler ise ikinci alt kümeyi oluşturur.

ID3 algoritması değişkenleri birçok alt bölüme ayırır, bu ayırma işlemi aşırı öğrenmeye neden olabileceğinden kazanım yerine kazanım oranı kavramı kullanılmıştır [27]. C4.5 ve C5.0 algoritmalarında hedef değişken üzerinde etkili olabilecek her bir değişken için ayırma değeri (split value) ve kazanım oranı (gain ratio) Eş 3.7 ve Eş 3.8'deki gibi hesaplanmaktadır [49].

$$\text{Ayırma Değeri} = H\left(\frac{|D_1|}{|D|}, \frac{|D_2|}{|D|}, \ldots \frac{|D_s|}{|D|}\right) \tag{3.7}$$

$$\text{Kazanım Oranı} = \frac{\text{Kazanım}}{\text{Ayırma Değeri}} \tag{3.8}$$

C5.0 algoritması ile C4.5 algoritması kendi içinde karşılaştırıldığında, C5.0 algoritması bazı gelişmiş özelliklere sahiptir. Çok daha hızlı çalışması, daha etkin bellek kullanımı, daha küçük karar ağaçları oluşturması, boosting desteği ve faydasız niteliklerin elimine edilmesini sağlayan winnowing özelliği gelişmiş özellikler olarak sıralanabilir [37].

3.5.3. C&RT algoritması

C&RT algoritması Breiman, Friedman, Olshen ve Stone tarafından 1984 yılında önerilmiştir. *Classification and Regression Trees* olarak adlandırılan İngilizce ismin baş harflerinden yola çıkarak C&RT algoritması olarak adlandırılmıştır. C&RT algoritması kök düğümden başlayarak her düğüm için olası tüm ayırma şekillerini gözden geçirerek bunlardan en iyisini seçer. C&RT algoritması, her düğümde iki dal üretir. Yani bölünmeler ya da ayrılmalar ikilidir [60]. Ayrılma kriteri için geliştirilmiş yöntemler bulunmaktadır. Bunlar arasında Gini, Twoing, Sıralı Twoing, Simetrik Gini, En Küçük Kareler Sapması (Least Squared Deviation) gibi yöntemler kullanılmaktadır. Bu yöntemlerin kullanımı sınıf hedeflerin sürekli ya da kategorik olmasına göre belirlenmektedir. En yaygın olarak kullanılan ayırma kriterleri Gini ve Twoing kuralıdır [17]. Towing ve Gini kriterleri aşağıda açıklanmıştır. C&RT algoritması da ID3 algoritmasında olduğu gibi entropiden yararlanır. En iyi ayırma kriterinin belirlenmesinde ise farklı bir fonksiyon kullanır. Bu fonksiyona Twoing Kriter Fonksiyonu denir [57]. Bu amaçla kullanılan formül Eş. 3.9'da verilmiştir.

$$\psi\left({}^{s}/_{t}\right) = 2\, P_L P_R \sum_{j=1}^{M}\left[P\left(C_j \middle| t_L\right) - P\left(C_j \middle| t_R\right)\right] \quad (3.9)$$

Eş. 3.9'a göre;

$\psi\left({}^{s}/_{t}\right)$: Dallandırma kriteri
t: Dallanmanın yapılacağı düğüm
c: Kriter
L: Ağacın sol tarafı
R: Ağacın sağ tarafı
P_L ve P_R: Bir kaydın sağda veya solda olma olasılığı
$P\left(C_j \middle| t_L\right)$ ve $P\left(C_j \middle| t_R\right)$: C_j sınıfındaki bir kayıtın ağacın sağında veya solunda olma olasılığını ifade etmektedir.

Bir niteliğin Gini değeri hesaplanmadan önce niteliğin $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanmalıdır (Eş 3.10 ve Eş 3.11).

$$Gini_{sol} = 1 - \sum_{i=1}^{k}\left[\frac{L_i}{|T_{sol}|}\right]^2 \quad (3.10)$$

$$Gini_{sağ} = 1 - \sum_{i=1}^{k}\left[\frac{R_i}{|T_{sağ}|}\right]^2 \quad (3.11)$$

Eş. 3.10 ve Eş 3.11'deki sembollerin açıklamaları;

k: Sınıfların sayısı
T: Bir düğümdeki örnek sayısı
T_{sol}: Sol koldaki örneklerin sayısı
$T_{sağ}$: Sağ koldaki örneklerin sayısı
L_i: Sol kolda i kategorisindeki örneklerin sayısı
R_i: Sağ kolda i kategorisindeki örneklerin sayısıdır.

Bir niteliğin Gini değeri ise Eş. 3.12 kullanılarak bulunmaktadır.

$$Gini_j = \frac{1}{n}\left(|T_{sol}|\, Gini_{sol} + |T_{sağ}|\, Gini_{sağ}\right) \quad (3.12)$$

Her bir nitelik için Gini değerleri hesaplanır, en küçük olanı seçilir ve bölünme bu nitelik üzerinden gerçekleşir. Kalan veri kümesi üzerinde yukarıdaki bahsedilen adımlar tekrar uygulanır ve diğer bölünme hesaplanır [61].

3.5.4. Chaid algoritması

1980 yılında Gordon V. Kass tarafından geliştirilen CHAID (Chi-Squared Automatic Interacton Detection), en popüler karar ağacı algoritmalarından biridir [54]. CHAID algoritmasında, eldeki popülasyon bağımlı değişken varyasyonu grup içi minimum, gruplar arası maksimum olacak şekilde ayrıştırarak alt bileşenlere ayrılır [62]. CHAID analizinde, diğer istatistiksel yöntemlerde olan normallik, doğrusallık ve homojenlik gibi klasik varsayımlar yoktur. Etkili bir öteleme algoritmasına sahip olan CHAID analizinin veri kümesini kararlı alt bileşenlere ayırabilmesi bunun temel nedenidir. Sürekli ve kategorik bir yapıya sahip tüm değişkenler aynı anda modele dâhil edilir. Bu nedenle bazı araştırmalarda hem parametrik hem de parametrik olmayan analiz yöntemleri içine dâhil edilmiştir [63]. En iyi açıklayıcı değişkenin ortaya konması için her bir değişken grubu arasında karşılaştırmalar yapılır. Karşılaştırma sürecinin her bir aşamasında veriler, ortaya çıkan değişkene göre yeniden yapılandırılır. Böylece her bir değişken grubunun, bir önceki grubun alt kümesi olduğu ortaya konmuş olur [64]. CHAID analizinde, en iyi açıklayıcı değişkenin ortaya konmasında ise Ki-kare (X^2) test sonuçlarını kullanmaktadır. Teste ilişkin sonuçların çözümlenmesinde ise aşağıdaki algoritma kullanılmaktadır [62].

Her bir tahmin edici değişken için X' in, Y hedef değişkenini dikkate alan p değeri en büyük kategori çifti bulunur. Bu adımda, bağımlı değişken sürekli yapıda ise F testi kullanılır. Yine bu adımda bağımlı değişken nominal yapıda ise bağımlı değişkenin (X) kategorileri satırlarda, bağımsız değişkenin kategorileri ise sütunlarda yer alacak şekilde sınıflandırması yapılır. Bu işlem için de Pearson Ki-Kare testi veya Pearson Olabilirlik testi kullanılır. Son olarak bu adımda, Y ordinal yapıda ise, Y birliktelik modeli için Olabilirlik testi kullanılır. En yüksek p değerli X' in kategori çifti için, p değeri önceden belirlenmiş $\alpha_{birleştirme}$ ölçütüyle kıyaslanır. Bu adımda, p değerinin $\alpha_{birleştirme}$ değerinden büyük olması durumunda, söz konusu çift tek kategori altında birleştirilir. Bu işlem sonrası, X' in yeni kategori sınıfı birinci adımdan tekrar başlatılır. Yine bu adımda, p değerinin $\alpha_{birleştirme}$ değerinden küçük olması durumunda, üçüncü adıma geçilir. X ve Y'nin kategori sınıfı için *Bonferroni* düzeltmesi

kullanılarak düzeltilmiş p değeri hesaplanır. En küçük düzeltilmiş p değerine sahip X tahmin edici değişkeni seçilerek, p değeri önceden tanımlanmış alfa düzeyi $\alpha_{böl}$ ile kıyaslanır. Bu adımda, en küçük düzeltilmiş p değeri $\alpha_{böl}$ değerine eşit veya küçük ise düğüm X' in kategori kümesi temel alınarak bölümlenir. Yine bu adımda, en küçük düzeltilmiş p değeri $\alpha_{böl}$ değerinden büyük ise, bu düğüm uç düğüm olduğundan ötürü bölünme gerçekleşmeyecektir. Her bir adımdaki ağaç büyütme işlemi, durma kuralları sağlanana kadar tekrarlanır. Ki-kare testi aşağıdaki gibi hesaplanır (Eş 3.13).

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(G_{ij} - B_{ij})^2}{B_{ij}} \tag{3.13}$$

Eş 3.13'de, G_{ij}, i. satır j. sütun için gözlenen frekansı, B_{ij}, i. satır j. sütun için beklenen frekansı, c sütun sayısını, r ise satır sayısını göstermektedir [57].

3.5.5. Quest algoritması

1997 yılında Loh ve Shih tarafından geliştirilen QUEST algoritması tek değişkenli ve doğrusal kombinasyonlu bölünmeleri destekler. Her bölme için, her bir girdi özniteliği ile hedef özniteliği arasındaki ilişki ANOVA F – testi, Levene'nin testi (ordinal ve sürekli öznitelikler için) veya Pearson'ın Ki Kare (nominal öznitelikler) testi kullanılarak hesaplanır. Hedef özniteliği çok terimli ise, iki süper küme oluşturmak için iki yönlü kümeleme kullanılır. Hedef özniteliğiyle en yüksek ilişkilendirmeyi alan öznitelik bölünme için seçilir. Giriş özniteliği ile ilgili en uygun ayırma noktasını bulmak için Kuadratik Diskriminant Analizi (QDA) uygulanır. İkili karar ağaçları oluşturur ve ağaçları budamak için On-kat çapraz doğrulama (Ten cross validation) kullanılır [56]. QUEST algoritmasında bölünmüş alan seçimi (split - field selection) ve bölünmüş nokta seçimi (split - point selection) ayrı olarak ele alınır. QUEST algoritmasının çalışma prensibi aşağıda gösterilmiştir [9]. Bölünmüş alan seçimi;

1) Her bir belirleyici alan X için, eğer X kategorik bir alan ise, nominal ya da sıralı, bir X ve bağımlı alan arasındaki bağımsızlığın bir Pearson Ki-kare testinin p-değeri hesaplanır. Eğer X değeri sürekli ise (ölçek - seviyeli) F testi kullanılarak p değeri hesaplanır.
2) En küçük p değerinin önceden belirlenen (Benferroni-adjusted alpha level α_B) bir değer ile kıyaslanması

a) Eğer en küçük α_B, p değerinden küçük ise, karşılık gelen belirleyici alan düğümü bölmek için seçilir. Sonraki adıma gidilir.
b) Eğer en küçük α_B, p değerinden küçük değil ise, o zaman sürekli olan her bir X için bir p değeri hesaplanır ve eşit olmayan varyanslar için Levene's test'i kullanılır. Bu durum X' in hedef alanın farklı seviyelerinde eşit olmayan varyanslara sahip olup olmadığının test edilmesi anlamına gelir.
c) Leven's testinden olan küçük p değeri yeni bir Bonferroni – düzeltilmiş alfa düzeyi (Bonferroni –adjusted alpha level) olan α_L ile kıyaslanır.
d) Eğer p değeri α_L 'den küçük ise düğümü bölmek için Leven testinden en küçük p değerli olan ilgili açıklayıcı alan seçilir.
e) Eğer p değeri α_L'den büyük ise düğüm bölünmez.

Bölünmüş nokta seçimi;

1) Eğer Y sadece iki kategoriye sahip ise, bir sonraki adıma geçilir. Yoksa Y'nin kategorileri iki üst sınıflara aşağıdaki gibi gruplara ayrılır:
a) Y'nin her bir kategorisi için X' in ortalaması hesaplanır
b) Eğer tüm ortalamalar aynı ise, bir üst sınıf olarak en büyük ağırlıklı frekans kategori seçilir ve diğer üst sınıfları oluşturmak için diğer tüm kategoriler birleştirilir.
c) Eğer ortalamalar aynı değil ise, başlangıç küme merkezlerinin iki en uç sınıf ortalaması olarak ayarlanması ile k-ortalamalar kümesi algoritması Y'nin iki üst sınıfını elde etmek için o ortalamalara uygulanır (burada k-ortalamalar kümesinin bir özel durumunda, k=2 olarak alınır)
2) Kuadratik Diskriminant Analizi (QDA) bölme noktasının (split point) belirlemek için uygulanır. QDA genellikle iki kesme noktası (cut-off point) üretir, ilk üst sınıfın örnek ortalamasına en yakın olanı seçilir.

Açıklanan karar ağacı algoritmaları giriş değişkeni, tahmin edici değişken, tahmin türü, bölünme sayısı ve bölme kriterlerine göre Çizelge 3.8'de gösterilmiştir.

Çizelge 3.8. Karar ağacı algoritmalarının karşılaştırılması [53]

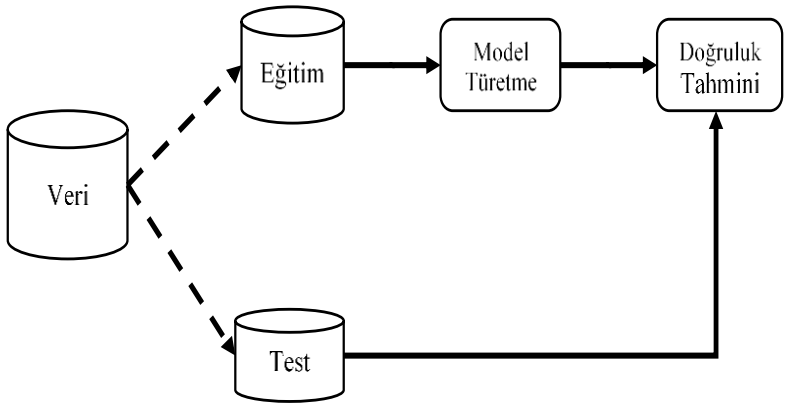
Algoritma	Giriş Değişkeni	Tahmin Edici Değişken	Tahmin türü	Bölünme Sayısı	Bölme Kriteri
C4.5/C5.0	Sürekli ve Kategorik	Kategorik	Sınıflandırma	≥ 2	Kazanç Oranı (Entropi)
C&RT	Sürekli ve Kategorik	Sürekli ve Kategorik	Sınıflandırma / Regresyon	=2	Gini / Towing
CHAID	Sürekli ve Kategorik	Sürekli ve Kategorik	Sınıflandırma / Regresyon	≥ 2	Ki-kare/ F testi
QUEST	Kategorik	Sürekli ve Kategorik	Sınıflandırma	=2	Ki-kare / F testi

3.6. Sınıflandırma Doğruluk Belirleme Yöntemleri

Karar ağacının eğitilmesinde kullanılacak olan verilerin sınıflandırma sürecinde seçilmesi gerekmektedir. Veri hazırlığı aşaması sonucunda elde edilen verinin analizi için temizlenmiş verilere bu aşamada ihtiyaç duyulmaktadır. Ancak bütün veriler ağacın eğitilmesinde kullanılmaz. Veriler ikiye bölünür; biri öğrenme ya da eğitim (training) diğeri ise deneme ya da test (testing)'dir. Bu ikiye bölme işlemi oluşturulan ağacın doğruluğunu ölçmek için gerçekleştirilir. Bazı durumlarda veri kümesi bölünebilecek kadar büyük olmayabilir böyle durumlarda doğruluğu test etmek için geliştirilmiş yöntemler vardır. Sınıflandırma sürecinde eğitim verileri üzerinde sınıflandırma sürecinde kullanılan algoritmanın doğruluğunun belirlenmesi için kullanılan bu yöntemlerden en yaygın olanları aşağıda açıklanmıştır [17].

3.6.1. Hold out

Hold out yöntemi, analiz edilecek verinin rastgele en az iki alt örnekleme ayrılmasıdır. Algoritmanın parametreleri belirlediği yani eğitildiği bölüme, eğitim kümesi (training set), algoritma sonuçlarının test edildiği kümeye ise, test kümesi (testing set) denir. Algoritmanın eğitimi, eğitim kümesi kullanılarak tamamlandıktan sonra, test kümesi ile modelin doğruluk değeri (accuracy) belirlenmektedir. Eğitim ve test kümesi için genellikle kullanılan oranlar, (0,90 - 0,10), (0,85 - 0,15), (0,80 - 0,20), (0,75 - 0,25) ve (0,70 - 0,30)'dur. Bu oranlar, analiz edilecek verinin büyüklüğüne ve yapısına göre değişim gösterebilir [65]. Doğruluk oranının belirlenmesi için Hold out yönteminin kullanılması Şekil 3.10'da gösterilmiştir.



Şekil 3.10. Hold out yöntemi ile doğruluk oranının belirlenmesi [53]

3.6.2. K-fold cross validation

Çapraz geçerlilikten türetilen k-katlı çapraz doğrulama olarak adlandırılmaktadır. Veri kümesi k adet alt kümeye ayrılır. Uygulamalarda genel olarak k değeri 10 alınır ve alt kümeler hemen hemen birbirine eşittir. Alt kümelerin belirlenmesinin ardından test kümesinin yeri değiştirilerek sırayla hata oranları elde edilir. Verilerin 10 gruba bölündüğü varsayıldığında ilk grup test için kalan dokuz grup ise model oluşturulmada kullanılır. Model verinin %90'ı tarafından oluşturulur, bu işlem her seferinde test grubunun değiştirilmesiyle 10 kez tekrar edilir. Elde edilen 10 adet bağımsız hata değerinin ortalaması, kurulan modelin hata oranını vermektedir [17]. Live one out yöntemi ise, k-fold cross validation yönteminin özel bir durumudur [30].

3.6.3. Bootstrap

Asıl veri kümesinden örneklenerek oluşturulan ve Bootstrap adı verilen veri kümeleri modelin oluşturulması için eğitim verisi olarak kullanılır. Bootstrap veri kümesinin dışında kalan veriler test verisi olarak kullanılır. Tesadüfi olarak örneklenen Bootstrap veri kümesinin büyüklüğünün tespitinde modelin hata oranını en iyi temsil edecek oran 0,632 olarak belirlenmiştir. Asıl verinin % 63,2'lik bir oranını temsil eden Bootstrap veri kümesi her yinelemede test kümesindeki verilerle yer değiştirir. Böylece her veri en az bir kez eğitim verisinde yer alır. Bazen bu işlem binin üzerinde tekrarlanır. Oluşturulan modelin hata oranı her Bootstrap örneğinin hata oranının ortalaması ile hesaplanmaktadır [67].

3.7. Sınıflandırma Performans Karşılaştırma Yöntemleri

CRISP-DM ya da diğer benzer modellerde, modelleme aşamasından sonra model değerlendirme aşaması gelir. Aynı veri kümesi üzerinde farklı modeller ile uygulama yapıldıktan sonra kullanılan algoritmaların başarı durumlarının değerlendirmesini yapmak için bazı yöntemler geliştirilmiştir [60].

3.7.1. Karışıklık matrisi

Veri madenciliği yöntemlerinde kullanılan pek çok sınıflandırma algoritması var olduğundan, algoritma seçimi yapmak için geliştirilmiş birkaç kıstas bulunmaktadır. Bunlardan biri karışıklık matrisi (confusion matrix)'dir. Bir modele ait karışıklık matrisi Çizelge 3.9'da gösterilmiştir.

Çizelge 3.9. Karışıklık matrisi [30]

		Modelin Tahmin Sınıfı		Toplam Örnek
		Pozitif	Negatif	
Gerçek Sınıf	Pozitif	Doğru Pozitif Sayısı (DP)	Yanlış Negatif Sayısı (YN)	P=DP+YN
	Negatif	Yanlış Pozitif Sayısı (YP)	Doğru Negatif Sayısı (DN)	N=YP+DN

Karışıklık matrisi 4 bölümden oluşmaktadır. DP, pozitif sınıf olarak sınıflandırılan pozitif örnek sayısı; YP, pozitif sınıf olarak sınıflandırılan negatif örnek sayısı; YN, negatif sınıf olarak sınıflandırılan pozitif örnek sayısı; DN, negatif sınıf olarak sınıflandırılan negatif örnek sayısı; P pozitif toplam ve N negatif toplam örnek sayısıdır [13]. Burada DP ve DN değerlerinin toplamı gerçek sınıfa ait ne kadar doğru tahmin yapıldığını göstermektedir. Karışıklık matrisi kullanılarak hesaplanan ölçütler; doğruluk (accuracy), hata oranı (error rate), anma (recall), kesinlik (precision) ve F-ölçütü (F-score) ile Alıcı İşletim Karakteristiği (Receiver Operating Characteristic-ROC) eğrisi ve kaldıraç oranı (lift ratio) olarak gösterilebilir [17].

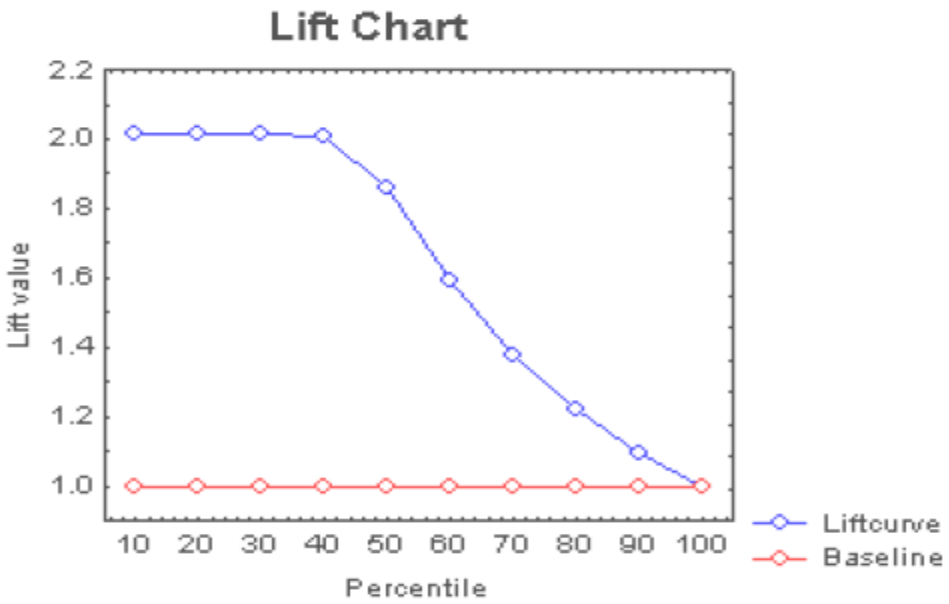
Bu değerlerin hesaplanma yöntemleri aşağıda gösterilmiştir (Çizelge 3.10).

Çizelge 3.10. Performans kriteri hesaplama yöntemleri [30]

Performans Kriteri	Formül
Doğruluk	$\text{Doğruluk} = \frac{DP + DN}{P + N}$
Hata Oranı	$\text{Hata Oranı} = \frac{YP + YN}{P + N}$
Kesinlik	$\text{Kesinlik} = \frac{DP}{DP + YP}$
Anma	$\text{Anma} = \frac{DP}{DP + YN}$
F Ölçütü	$F = \frac{(2 * \text{Kesinlik} * \text{Anma})\ )}{(\text{Kesinlik} + \text{Anma}}$

3.7.2. Kaldıraç değeri ve grafiği

Kaldıraç (Lift) değeri, sınıflandırıcı ile tahmin edilen hedef değerin yüzdelik değer içerisindeki oranının hedef değerin ilgilenilen değerinin tüm veri içerisindeki oranına bölümünü ifade eder. Bölme işlemi sonucu elde edilen değerin birden büyük olması ilgili sınıflandırıcı performansının rassal sınıflandırıcı modele (hiç bir sınıflandırıcı modeli kullanılmaması durumunda elde edilecek sonuç) göre ne kadar üstün olduğunu gösterir [13]. Kaldıraç grafiği her bir yüzdelik dilimdeki isabet edilen (hits) kayıtların yüzdeleri ile eğitme veri kümesindeki tüm isabet edilen kayıtların (hits) oranları ile kıyaslama yapmayı sağlar. İyi bir model için, *lift* grafiği sol tarafta %100 (1,0) çok üzerinde başlar, sağa doğru hareket ederken yüksek bir düz plato şeklinde devam eder ve sağ tarafa doğru yani %100 (1,0) 'a kadar hızlı bir şekilde düşer [9]. Kümülatif olarak çizilen Şekil 3.13'deki örnek kaldıraç grafiğinde x ekseni yüzde, y ekseni kaldıraç oranı olmak üzere üstte seyreden sınıflandırıcıya ait eğri, o sınıflandırıcının performansının yüksek olduğunu gösterir [13].



Şekil 3.13. Kaldıraç grafiği [13]

4. UYGULAMA VE BULGULAR

Gerçekleştirilen uygulamanın amacı; hanehalkının internet hizmeti sahipliği durumunu etkileyen faktörlerin, hanehalkı karakteristik özelliklerine göre karar ağaçları ile incelenmesidir. Günlük hayatın bir parçası haline gelen, her türlü bilgiye erişimi sağlayan ve sürekli büyüyen bir iletişim ağı olan internet hizmetinin sahiplik durumu hedef değişken olarak seçilmiş ve buna göre modelleme yapılmıştır. İnternet hizmeti sahipliğini etkileyen faktörler ile ilgili karar ağaçları performansları karşılaştırılmış ve buna göre çalışma yapılmıştır ancak hedef değişken değiştirilerek diğer bilişim cihazlarının sahiplik durumu içinde farklı bir çalışma yapılabilir. Tezin uygulama bölümünde verilerin veri madenciliği uygulamasına uygun hale getirilmesi, kullanılan değişkenler, uygulamada kullanılan veri madenciliği aracı, karar ağacı algoritmaları kullanılıp model performansları karşılaştırılarak en uygun modelin nasıl seçildiği açıklanmıştır. Seçilen modele göre bir karar ağacı elde edilmiş ve elde edilen karar ağacının dallanmaları anlatılmıştır. Ayrıca hedef değişken olan hane internet hizmeti sahipliği değiştirilerek hane bilgisayar kullanımı hedef değişken olarak seçilmiş ve buna göre bir model oluşturulup değerlendirilmiştir.

4.1. Uygulama Kullanılan Veriler

Bilgi toplumu olma ölçütlerinin oluşturulmasında yürütülen temel araştırmaların başında gelen HBTKA 2006 yılı hariç olmak üzere 2004 yılından itibaren düzenli olarak TÜİK tarafından gerçekleştirilmektedir. Araştırmanın örneklem tasarımı 2 aşamalı tabakalı sistematik küme örneklemesi yöntemidir. Birinci aşamada ortalama 100 hane ihtiva eden kümeler (bloklar) PPS yöntemi (büyüklüğü ile orantılı ve olasılıklı) ile örneğe seçilmektedir. İkinci aşamada ise örneğe seçilen bu kümelerden sistematik seçim yöntemi kullanılarak, örnek adresler belirlenmektedir. İstatistiki Bölge Birimleri Sınıflaması 1. Düzey tabakalama kriteri olarak kullanılmaktadır. Hazırlanan bu tezde Türkiye İstatistik Kurumunun (TÜİK) 2016 yılında yaptığı HBTKA anket verilerine ait veri kümesi kullanılmıştır. Anket çalışmasında kullanılan soru formuna Avrupa Birliği İstatistik Ofisi'nin (EUROSTAT) AB üye ve aday ülke istatistik ofisleri ve OECD teşkilatı ile iş birliği içerisinde gerçekleştirdiği model soru formu altlık teşkil etmektedir. 2004 ve 2005 yıllarında Hanehalkı İşgücü Anketi'ne modül olarak eklenen çalışma, 2006 yılında uygulanmamış olup, 2007 yılından itibaren ise ayrı bir araştırma olarak yıllık periyotlarda gerçekleştirilmektedir. Soru formu

hane geneli ve fertlere ilişkin bölümlerinden oluşmaktadır. Hane geneline ilişkin bölümde; hanede yaşayan tüm fertlerin yaş, cinsiyet, hanelerde bulunan bilişim teknolojileri ürünleri, internet erişim imkânı, internet bağlantısı olan araçlar, kullanılan internet bağlantı türleri ile evden internete bağlanamama nedenleri sorgulanmaktadır. Fertlere ilişkin bölümde; 16-74 yaş arasındaki bireylerin eğitim ve iş gücü durumu ile bilgisayar ve internet kullanımları, kullanım sıklıkları, kullanım amaçları, kamu ile iletişimde internet kullanımı ve e-ticarete ilişkin sorgulama yapılmaktadır. Veri kümesinde Türkiye sınırları dâhilinde bulunan tüm yerleşim yerlerindeki toplam 13 700 hane ankete dâhil edilmiş olup 11 874 hane soruları cevaplamıştır. Araştırmadaki toplam fert sayısı 39 754 olup 11 820 fert araştırma kapsam dışındadır. Araştırma kapsamına dâhil olan 16-74 yaş arası 27 934 fert ankete katılmış 25 058 fert anketi cevaplamıştır. Kurumsal nüfus olarak tanımlanan okul, yurt, otel, çocuk yuvası, huzurevi, hastane ve hapishanede bulunanlar ile kışla ve ordu evlerinde ikamet edenler kapsanmamıştır. Ayrıca nüfusu toplam nüfusun % 1'ini geçmeyecek şekilde yeterli (küçük köyler, oba, mezra vb.) örnek hane sayısına ulaşılamayacağı düşünülen yerleşim yerleri kapsam dışı bırakılmıştır [4]. 2016 yılına ait HBTKA B grubu mikro verileri TÜİK'ten talep edilmiş, TÜİK değerlendirme kurulunca talep uygun bulunmuş ve verilere ulaşılmıştır. İlk aşamada *Hane* ve *Fert* adında iki ayrı veri dosyası için ortak olan bülten numarası değişkeni kullanılarak tek bir veri dosyası oluşturulmuştur. Veri dosyalarını birleştirmek için bir hesap işlemci programından yararlanılmıştır.

Hane veri dosyasından;

1) Hanehalkı Büyüklüğü,
2) Hanehalkı Bölgesi,
3) Hanede Masaüstü Bilgisayar Sahiplik Durumu,
4) Hanede Dizüstü Bilgisayar Sahiplik Durumu,
5) Hanede Tablet Sahiplik Durumu,
6) Hanede Cep Telefonu Sahiplik Durumu,
7) Hanede Oyun Konsolu Sahiplik Durumu,
8) Hanede Sabit Telefon Sahiplik Durumu,
9) Hanede Smart TV Sahiplik Durumu,
10) Hanede Bilgisayar Kullanımı,
11) Hanehalkının Gelir Düzeyi,
12) Hanedeki İnternet Hizmeti Sahiplik Durumu değişkenleri alınmıştır.

Fert veri dosyasından ise; *Hane* dosyası ile ortak olan *Bülten No* adlı değişken kullanılmış ve aşağıdaki değişkenler elde edilmiştir.

1) Hanehalkı Reisinin Cinsiyeti,
2) Hanehalkı Reisinin Yaşı,
3) Hanehalkı Reisinin Okuma Yazma Durumu,
4) Hanehalkı Reisinin Eğitim Durumu,
5) Hanehalkı Reisinin Çalışma Durumu,
6) Hanehalkı Reisinin Mesleği
7) Hanede 0-25 Yaş Arası Bireyin Varlığı

Değişkenler oluşturulurken hanenin demografik yapısı, hanehalkı reisinin özellikleri ve hanenin bazı bilişim teknolojileri cihazlarına sahip olup olmadığı ile ilgili veriler kullanılmıştır. Ayrıca hanelerin satın alma karar davranışlarının ortaya çıkmasında, kişisel birer faktör olarak hanehalkı reislerinin yaşları, cinsiyetleri, gelir durumları, eğitim seviyeleri, medeni durumları, çalışma durumları gibi demografik pek çok değişkenin etkisinde kalacakları düşünülmüştür. Veri dosyasında hanede yaşı en büyük olan ve cinsiyeti erkek olan kişi hanehalkı reisi olarak kabul edilmiştir. Eğer hanede cinsiyeti ve yaşı en büyük olan kişi erkek değil ise, yaşı en büyük olan ve cinsiyeti kadın olan kişi hanehalkı reisi olarak kabul edilmiştir. Hanehalkı reisi bilgisi alındıktan sonra bu kişiye ait yaş, okuma yazma durumu, eğitim durumu, çalışma durumu ve meslek bilgisi alınmıştır. Hanehalkı reisinin çalışma durumu ve mesleği için referans haftasındaki (28 Mart-03 Nisan 2016) çalışma durumu bilgisi alınarak bu değişkene ait veriler elde edilmiştir. Genç nüfusun daha fazla internet kullandığı düşüncesi ile sahiplik durumunu etkileyeceği düşünüldüğünden *Hanede 0-25 Yaş Arası Bireyin Varlığı* adında yeni bir değişken oluşturulmuş ve hanede yaşayan fertlerin yaş bilgilerinden bu değişkene ait bilgi doldurulmuştur. 25 yaşında olan fertler değişkene dâhildir. Son olarak *Hanehalkı Reisinin Mesleği* adlı değişkene belirtilen meslek grupları içinde olmayanlar için 36 numaralı kod ile *Diğer* adında yeni bir meslek kodu eklenmiştir. Mevcut meslek kodlarına dâhil olmayanlar ya da bu alanın cevabını boş bırakanlar Diğer meslek kodu grubuna dâhil edilmiştir. *Hanehalkı Reisi Yaşı* adlı ayrık değişken kategorik hale getirilmiş ve 6'ya bölünmüştür. Hanedeki cep telefonu, tablet, masaüstü ve dizüstü bilgisayar, oyun konsolu, smart tv cihazları internete bağlanabilen cihazlardır. Hanehalkına ait aylık gelir bilgisi *iki aşamalı kümeleme analizi* ile 4 gelir

grubuna ayrılmış ve yeni bir değişken olarak veri kümesine dâhil edilmiştir. Değişkeni kategorik bir değişken yapmak için ve daha homojen kümeleme yapmayı sağladığından iki aşamalı kümeleme analizi kullanılmıştır. Hanehalkı gelirine ait istatistiki bilgiler kısmında kümeler ile ilgili bilgi verilmiştir. Anket sorularını cevaplayan 11 874 haneden kapsam dışı olan haneler çıkarılmış ve kalan 11 276 adet hane veri kümesine dâhil edilmiştir. İnternet hizmeti sahiplik durumu için *Bilinmiyor* şeklinde cevap veren 54 hane veri kümesinden çıkarılmış sonuçta 19 adet değişken, 11 222 haneden oluşan yeni ve tek bir veri kümesi elde edilmiştir. Fert ve hane karakteristiklerinden oluşan değişken adları, açıklamaları, aldığı değerler ve tipleri Çizelge 4.1'de belirtilmiştir.

Çizelge 4.1. Uygulamada kullanılan değişken bilgileri

Sıra No	Değişkenin				Tipi
	Adı	Açıklaması	Aldığı Değerler		
1	HHRCINSIYET	Hanehalkı Reisinin Cinsiyeti	1	Erkek	Nominal
			2	Kadın	
2	HHRYAS	Hanehalkı Reisinin Yaşı	1	0-30 Yaş	Nominal
			2	31-40 Yaş	
			3	41-50 Yaş	
			4	51-60 Yaş	
			5	61-70 Yaş	
			6	71 Yaş ve üzeri	
3	HHROKUMAYAZMA	Hanehalkı Reisinin Okuma Yazma Durmu	1	Evet	Flag
			2	Hayır	
4	HHREGITIM	Hanehalkı Reisinin Eğitim Durumu (Uluslararası Standart Eğitim Sınıflaması)	0	Herhangi bir okul bitirmedi	Nominal
			1	İlkokul Mezunu	
			2	İlköğretim/Ortaokul veya Mesleki Ortaokul Mezunu	
			3	Lise veya Mesleki Lise Mezunu	
			4	İki veya Üç Yıllık Yükseokul Mezunu	
			5	Dört Yıllık Yüksekokul veya Fakülte Mezunu	
			6	Lisans	
			7	Doktora	
5	HHRCALISMA	Hanehalkı Reisinin Çalışma Durumu	1	Çalışıyor	Flag
			2	Çalışmıyor	
6	HHRMESLEK	Hanehalkı Reisinin Mesleği (Uluslararası Standart Meslek Sınıflaması)	0	Silahlı kuvvetlerle ilgili meslekler	Nominal
			1	Yöneticiler	
			4	Büro hizmetlerinde çalışan elemanlar	
			5	Hizmet ve satış elemanları	
			6	Nitelikli tarım, ormancılık ve su ürünleri çalışanları	
			7	Sanatkârlar ve ilgili işlerde çalışanlar	
			8	Tesis ve makine operatörleri ve montajcılar	
			9	Nitelik gerektirmeyen işlerde çalışanlar	

Çizelge 4.1. (devam) Uygulamada kullanılan değişken bilgileri

			21	Bilim ve mühendislik alanlarındaki profesyonel meslek mensupları	
			22	Sağlık profesyonelleri	
			23	Eğitim ile ilgili profesyonel meslek mensupları	
			24	İş ve yönetim ile ilgili profesyonel meslek mensupları	
			25	Bilgi ve iletişim teknolojileri ile ilgili profesyonel meslek mensupları	
			26	Hukuk, sosyal bilimler ve kültür ile ilgili profesyonel meslek mensupları	
			31	Bilim ve mühendislik ile ilgili yardımcı profesyonel meslek mensupları	
			32	Yardımcı sağlık profesyonelleri	
			33	İş ve idare ile ilgili yardımcı profesyonel meslek mensupları	
			34	Hukuk, sosyal, kültür ve benzeri alanlar ile ilgili yardımcı profesör	
			35	Bilgi ve iletişim teknisyenleri	
			36	Diğer	
7	HHBUYUKLUK	Hanehalkı Büyüklüğü	1-20		Ordinal
8	SIFIRYIRMIBESYAS	Hanede 0-25 Yaş Bireyin Varlığı	1	Var	Flag
			2	Yok	
9	HHBOLGE	Hanenin Bulunduğu Bölge (İstatistiki Bölge Birimleri Sınıflaması Düzey 1)	TR1	İstanbul	Ordinal
			TR2	Batı Marmara	
			TR3	Ege	
			TR4	Doğu Marmara	
			TR5	Batı Anadolu	
			TR6	Akdeniz	
			TR7	Orta Anadolu	
			TR8	Batı Karadeniz	
			TR9	Doğu Karadeniz	
			TRA	Kuzeydoğu Anadolu	
			TRB	Ortadoğu Anadolu	
			TRC	Güneydoğu Anadolu	
10	HGELIR	Hane Aylık Net Toplam Geliri	HGG-1	Düşük	Nominal
			HGG-2	Orta Alt	
			HGG-3	Orta Üst	
			HGG-4	Yüksek	
11	HBILGISAYARMASAUSTUSAHIP	Hane Masaüstü Bilgisayar Sahipliği	1	Var	Flag
			2	Yok	
12	HBILGISAYARDIZUSTUSAHIP	Hane Dizüstü Bilgisayar Sahipliği	1	Var	Flag
			2	Yok	
13	HTABLETSAHIP	Hane Tablet Sahipliği	1	Var	Flag
			2	Yok	
14	HCEPTELSAHIP	Hane Cep Telefonu Sahipliği	1	Var	Flag
			2	Yok	
15	HOYUNKONSOLSAHIP	Hane Oyun Konsolu Sahipliği	1	Var	Flag
			2	Yok	
16	HSABITTELSAHIP	Hane Sabit Telefon Sahipliği	1	Var	Flag
			2	Yok	

Çizelge 4.1. (devam) Uygulamada kullanılan değişken bilgileri

17	HSMARTTVSAHIP	Hane Smarttv Sahipliği	1	Var	Flag
			2	Yok	
18	HBILGISAYARKULLANIM	Hane Bilgisayar Kullanım Bilgisi	1	Evet	Flag
			2	Hayır	
19	HINTERNETSAHIP	Hane İnternet Hizmeti Sahipliği	1	Evet	Flag
			2	Hayır	

4.2. Uygulamada Kullanılan Değişkenlere Ait İstatistiksel Bilgiler

Değişkenlere ait istatistiksel bilgiler aşağıdaki çizelgelerde gösterilmiştir.

Çizelge 4.2. Hanehalkı reisinin cinsiyeti

	Kişi Sayısı	Yüzde (%)
Erkek	8 376	74,6
Kadın	2 846	25,4
Toplam	11 222	100,0

Çizelge 4.2'ye göre, hanehalkı reisinin % 74,6'sı erkek; % 25,4'ü ise kadındır.

Çizelge 4.3. Hanehalkı reisinin yaşı

	Kişi Sayısı	Yüzde (%)
0-30 Yaş	837	7,5
31-40 Yaş	2 241	20,0
41-50 Yaş	2 464	22,0
51-60 Yaş	2 390	21,3
61-70 Yaş	1 986	17,7
71 ve üzeri Yaş	1 304	11,6
Toplam	11 222	100,0

Çizelge 4.3'e göre, hanehalkı reisi yaş dağılımına göre % 7,5'inin 0-30; % 20'sinin 31-40; % 22'sinin 41-50; % 21,3'ünün 51-60; % 17,7'sinin 61-70 ve % 11,6'sının ise 71 ve üzeri yaşta olduğu ortaya çıkmıştır.

Çizelge 4.4. Hanehalkı reisinin okuma yazma durumu

	Kişi Sayısı	Yüzde (%)
Evet	9 900	88,2
Hayır	1 322	11,8
Toplam	11 222	100,0

Çizelge 4.4'den de anlaşılacağı üzere, hanehalkı reisinin % 88,2'sinin okuma yazmasının olduğu; buna karşılık % 11,8'inin okuma yazmasının olmadığı görülmüştür.

Çizelge 4.5. Hanehalkı reisinin eğitim durumu

	Kişi Sayısı	Yüzde (%)
Herhangi Bir Okul Bitirmemiş	2 990	26,6
İlkokul Mezunu	3 933	35,0
İlköğretim/Ortaokul veya Mesleki Ortaokul Mezunu	1038	9,2
Lise veya Mesleki Lise Mezunu	1 617	14,4
2 veya 3 Yıllık Yüksekokul Mezunu	471	4,2
4 Yıllık Yüksekokul veya Fakülte Mezunu	961	8,6
Yüksekokul Lisans (5 veya 6 Yıllık Fakülteler Dâhil) Mezunu	175	1,6
Doktora Mezunu	37	0,3
Toplam	11 222	100,0

Çizelge 4.5'deki, hanehalkı reisinin eğitim durumuna göre, % 26,6'sının herhangi bir okul bitirmediği; % 35'inin ilkokul mezunu; % 9,2'sinin ilköğretim/ortaokul veya mesleki ortaokul mezunu; % 14,4'ünün lise veya mesleki lise mezunu; % 4,2'sinin 2 veya 3 yıllık yüksekokul mezunu; % 8,6'sının 4 yıllık yüksekokul veya fakülte mezunu; % 1,6'sının yüksekokul lisans (5 veya 6 yıllık fakülteler dâhil) mezunu ve % 0,3'ünün doktora mezunu olduğu ortaya çıkmıştır.

Çizelge 4.6. Hanehalkı reisinin çalışma durumu

	Kişi Sayısı	Yüzde (%)
Çalışıyor	5 344	47,6
Çalışmıyor	5 878	52,4
Toplam	11 222	100,0

Çizelge 4.6'ya göre, hanehalkı reisinin % 47,6'sının çalıştığı, buna karşılık % 52,4'ünün ise çalışmadığı ortaya çıkmıştır.

Çizelge 4.7. Hanehalkı reisinin mesleği

	Hane Sayısı	Yüzde (%)
Silahlı Kuvvetler ile İlgili Meslekler	179	1,6
Yöneticiler	245	2,2
Büro Hizmetlerinde Çalışan Elemanlar	458	4,1
Hizmet ve Satış Elemanları	933	8,3
Nitelikli Tarım. Ormancılık ve Su Ürünleri Çalışanları	473	4,2

Çizelge 4.7. (devam) Hanehalkı reisinin mesleği

Sanatkârlar ve İlgili İşlerde Çalışanlar	293	2,6
Tesis ve Makine Operatörleri ve Montajcılar	268	2,4
Nitelik Gerektirmeyen İşlerde Çalışanlar	1 752	15,6
Bilim ve Mühendislik Alanlarındaki Profesyonel Meslek Mensupları	92	0,8
Sağlık Profesyonelleri	77	0,7
Eğitim ile İlgili Profesyonel Meslek Mensupları	248	2,2
İş ve Yönetim ile İlgili Profesyonel Meslek Mensupları	105	0,9
Bilgi ve İletişim Teknolojileri ile İlgili Profesyonel Meslek Mensupları	27	0,2
Hukuk. Sosyal Bilimler ve Kültür ile İlgili Profesyonel Meslek Mensupları	26	0,2
Bilim ve Mühendislik ile İlgili Yardımcı Profesyonel Meslek Mensupları	38	0,3
Yardımcı Sağlık Profesyonelleri	32	0,3
İş ve İdare ile İlgili Yardımcı Profesyonel Meslek Mensupları	70	0,6
Hukuk, Sosyal, Kültür ve Benzeri Alanlar ile İlgili Yardımcı Profesör	15	0,1
Bilgi ve İletişim Teknisyenleri	13	0,1
Diğer	5 878	52,4
Toplam	11 222	100,0

Çizelge 4.7 incelendiğinde, hanehalkı reisi mesleklerinin daha çok diğer (işsiz, emekli, mevsimlik işçi, eğitim öğretime devam eden, yaşlı ya da bu alanı ankette boş bırakanlar vs.) grupta olanlar olduğu ortaya çıkmıştır. Meslek grubu için diğer dağılımda hanehalkı reislerinin, % 1,6'sının silahlı kuvvetler ile ilgili meslekler; % 2,2'sinin yöneticiler; % 4,1'inin büro hizmetlerinde çalışan elemanlar; % 8,3'ünün hizmet ve satış elemanları; % 4,2'sinin nitelikli tarım ormancılık ve su ürünleri çalışanları; % 2,6'sının sanatkârlar ve ilgili işlerde çalışanlar; % 2,4'ünün tesis ve makine operatörleri ve montajcılar; % 15,6'sının nitelik gerektirmeyen işlerde çalışanlar; % 0,8'inin bilim ve mühendislik alanlarındaki profesyonel meslek mensupları; % 0,7'sinin sağlık profesyonelleri olduğu ortaya çıkmıştır.

Çizelge 4.8. Hanehalkı büyüklüğü

	Hane Sayısı	Yüzde (%)
1	959	8,5
2	2 841	25,3
3	2 472	22,0
4	2 456	21,9
5	1 329	11,8
6	587	5,2
7	263	2,3
8	147	1,3

Çizelge 4.8. (devam) Hanehalkı büyüklüğü

9	70	0,6
10	49	0,4
11	20	0,2
12	3	0,0
13	10	0,1
14	8	0,1
15	2	0,0
16	1	0,0
17	1	0,0
18	3	0,0
20	1	0,0
Toplam	11 222	100,0

Çizelge 4.8 incelendiğinde, hanehalklarının daha çok 2 kişilik (%25,3) aileler olduğu ortaya çıkmıştır. Bunu sırasıyla, 3 kişilik (% 22) ve 4 kişilik (% 21,9) ailelerin takip ettiği görülmüştür.

Çizelge 4.9. Hanede 0-25 yaş bireyin varlığı

	Hane Sayısı	Yüzde (%)
Var	7 043	62,8
Yok	4 179	37,2
Toplam	11 222	100,0

Çizelge 4.9'a göre bir hanede 0-25 yaş arası bireyin varlığına ilişkin verilerde, 7 043 hanede 0-25 yaş bireyin olduğu; 4 179 hanede ise 0-25 yaş bireyin olmadığı ortaya çıkmıştır.

Çizelge 4.10. Hanenin bulunduğu bölge

	Hane Sayısı	Yüzde (%)
İstanbul	1 488	13,3
Batı Marmara	739	6,6
Ege	1 273	11,3
Doğu Marmara	1 101	9,8
Batı Anadolu	1 084	9,7
Akdeniz	1 193	10,6
Orta Anadolu	777	6,9
Batı Karadeniz	822	7,3
Doğu Karadeniz	643	5,7
Kuzeydoğu Anadolu	520	4,6
Ortadoğu Anadolu	628	5,6
Güneydoğu Anadolu	954	8,5
Toplam	11 222	100,0

Çizelge 4.10 incelendiğinde, hanelerin daha çok *İstanbul* (% 13,3) bölgesinde yaşadığı ortaya çıkmıştır. Bunu sırasıyla, *Ege* (% 11,3) kişilik ve *Akdeniz* (% 10,6) bölgelerinin takip ettiği görülmüştür.

Çizelge 4.11. Hane masaüstü bilgisayar sahipliği

	Hane Sayısı	Yüzde (%)
Var	2 435	21,7
Yok	8 787	78,3
Toplam	11 222	100,0

Çizelge 4.11'deki hane masaüstü bilgisayar sahipliği durumuna ilişkin verilerde 2 435 hanede masaüstü bilgisayar olduğu; 8 787 hanede masaüstü bilgisayar olmadığı ortaya çıkmıştır.

Çizelge 4.12. Hane dizüstü bilgisayar sahipliği

	Hane Sayısı	Yüzde (%)
Var	3 775	33,6
Yok	7 447	66,4
Toplam	11 222	100,0

Çizelge 4.12'deki hane dizüstü bilgisayar sahipliği durumuna ilişkin verilerde 3 775 hanede dizüstü bilgisayar olduğu; 7 447 hanede dizüstü bilgisayar olmadığı ortaya çıkmıştır.

Çizelge 4.13. Hane tablet sahipliği

	Hane Sayısı	Yüzde (%)
Var	3 184	28.4
Yok	8 038	71.6
Toplam	11 222	100.0

Çizelge 4.13'deki hane tablet sahipliği durumuna göre 3 184 hanede tablet olduğu; 8 038 kişinin ise hanesinde tablet olmadığı ortaya çıkmıştır.

Çizelge 4.14. Hane cep telefonu sahipliği

	Hane Sayısı	Yüzde (%)
Var	10 845	96,6
Yok	377	3,4
Toplam	11 222	100,0

Çizelge 4.14'deki hane cep telefonu sahipliği durumuna ilişkin verilerde 10 845 hanede cep telefonu olduğu; 377 hanede cep telefonu olmadığı ortaya çıkmıştır.

Çizelge 4.15. Hane oyun konsolu sahipliği

	Hane Sayısı	Yüzde (%)
Var	534	4,8
Yok	10 688	95 2
Toplam	11 222	100,0

Çizelge 4.15'deki hane oyun konsolu sahipliği durumuna ilişkin verilerde 534 kişinin ailesinde oyun konsolu olduğu; 10 688 kişinin ise hanesinde oyun konsolu olmadığı ortaya çıkmıştır.

Çizelge 4.16. Hane sabit telefon sahipliği

	Hane Sayısı	Yüzde (%)
Var	2 918	26,0
Yok	8 304	74,0
Toplam	11 222	100,0

Çizelge 4.16'daki hane sabit telefon sahipliği durumuna ilişkin verilerde 2 918 kişinin ailesinde sabit telefon olduğu; 8 304 kişinin ise hanesinde sabit telefon olmadığı ortaya çıkmıştır.

Çizelge 4.17. Hane smart tv sahipliği

	Hane Sayısı	Yüzde (%)
Var	2 581	23,0
Yok	8 641	77,0
Toplam	11 222	100,0

Çizelge 4.17'deki hane smart tv sahipliği durumuna ilişkin verilerde 2 581 hanede smart tv olduğu; 8 641 hanede smart tv olmadığı ortaya çıkmıştır.

Çizelge 4.18. Hane bilgisayar kullanımı

	Hane Sayısı	Yüzde (%)
Evet	5 278	47.0
Hayır	5 944	53.0
Toplam	11 222	100.0

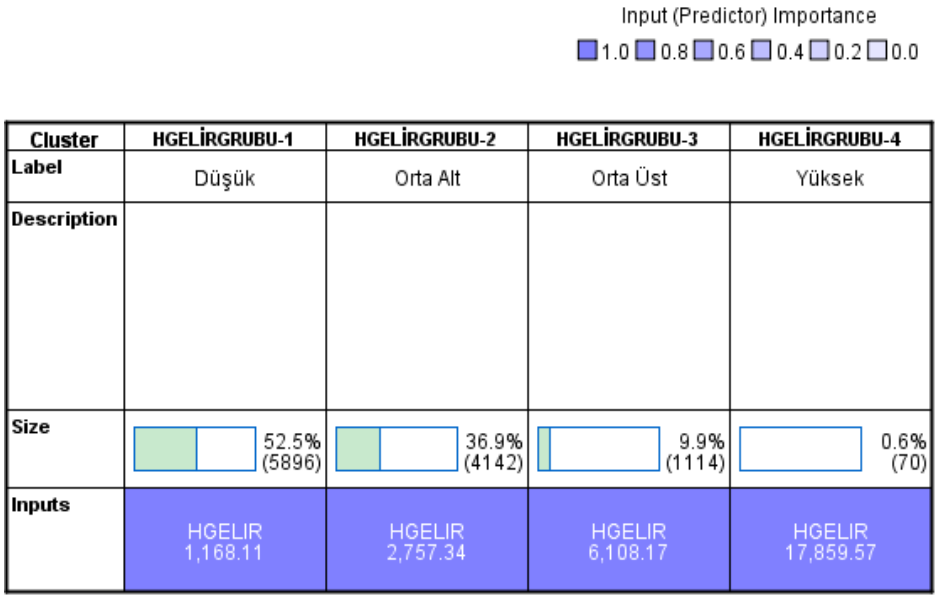
Çizelge 4.18'deki hane bilgisayar kullanımına ilişkin verilerde 5 278 hanede bilgisayar kullanıldığı, 5 944 hanede bilgisayar kullanılmadığı ortaya çıkmıştır.

Çizelge 4.19. Hane internet hizmeti sahipliği

	Hane Sayısı	Yüzde (%)
Evet	8 081	72,0
Hayır	3 141	28,0
Toplam	11 222	100,0

Çizelge 4.19'daki hane internet hizmet sahipliği durumuna ilişkin verilerde 8 081 hanede internet olduğu; 3 141 hanede ise internet olmadığı ortaya çıkmıştır.

Veri kümesinde bulunan Hane Aylık Net Toplam Geliri değişkeni *iki aşamalı kümeleme* ile kümeleme işlemine tabi tutulmuş ve kümeleme sonucunda 4 adet gelir grubu oluşmuştur. Kümeleme işlemi sonucunda ortaya çıkan kümelere ait bilgiler ve model özeti Resim 4.1 ve Resim 4.2'de gösterildiği gibidir.



Resim 4.1. Gelir değişkeni için iki aşamalı kümeleme işlemi sonucu oluşan kümeler



Resim 4.2. Model özeti ve iki aşamalı kümeleme kalitesi

İki aşamalı kümeleme analizi sonucunda ortaya çıkan Gelir gruplarına ait hane sayısı ve yüzde bilgileri Çizelge 4.20'deki gibidir.

Çizelge 4.20. Kümeleme sonucunda ortaya çıkan gelir grupları

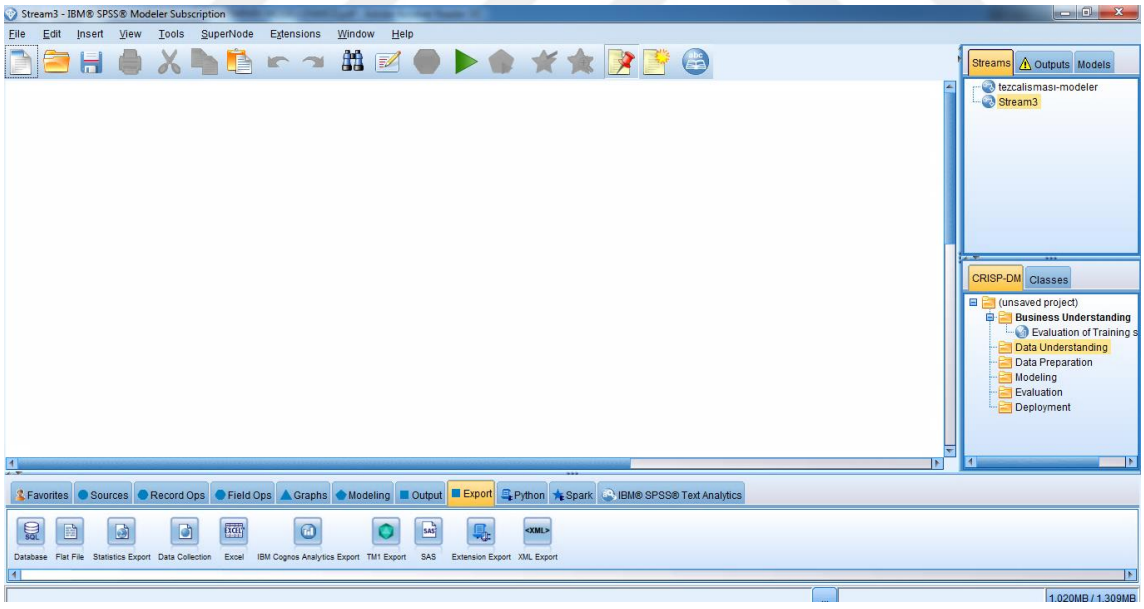
Gelir Grubu	**Hane Sayısı**	**Yüzde (%)**	**Ortalama Aylık Harcanabilir Gelir**
Düşük	5 896	52,5	1 168,11 TL
Orta Alt	4 142	36,9	2 757,34 TL
Orta Üst	1 114	9,9	6 108,17 TL
Yüksek	70	0,6	17 859,57 TL
Toplam	11 222	100,0	

Buna göre *Düşük* gelir grubunda 5 896 hane (% 52,5), *Orta Alt* gelir grubunda 4 142 hane (% 36,9), *Orta Üst* gelir grubunda 1 114 hane (%9.9), *Yüksek* gelir grubunda 70 (% 0.6) hane bulunmaktadır. Düşük gelir grubunda bulunan hanelerin Ortalama Aylık Harcanabilir Geliri 1 168,11 TL, Orta Alt gelir grubunda bulunan hanelerin Ortalama Aylık Harcanabilir Geliri 2 757,34 TL, Orta Üst gelir grubunda bulunan hanelerin Ortalama Aylık Harcanabilir Geliri 6 108,17 TL, Yüksek gelir grubunda bulunan hanelerin Ortalama Aylık Harcanabilir Geliri 17 859,57 TL'dir.

4.3. Uygulamada Kullanılan Yöntem ve Veri Madenciliği Aracı

Çalışmada yöntem olarak veri madenciliği sınıflandırma tekniklerinden karar ağaçları kullanılmıştır. Kullanıcılar tarafından kolayca yorumlanabilmesi, kurallar kümesine dönüştürülebilmesi, diğer yöntemler verinin normal dağılmamasına veya eksik değerlere

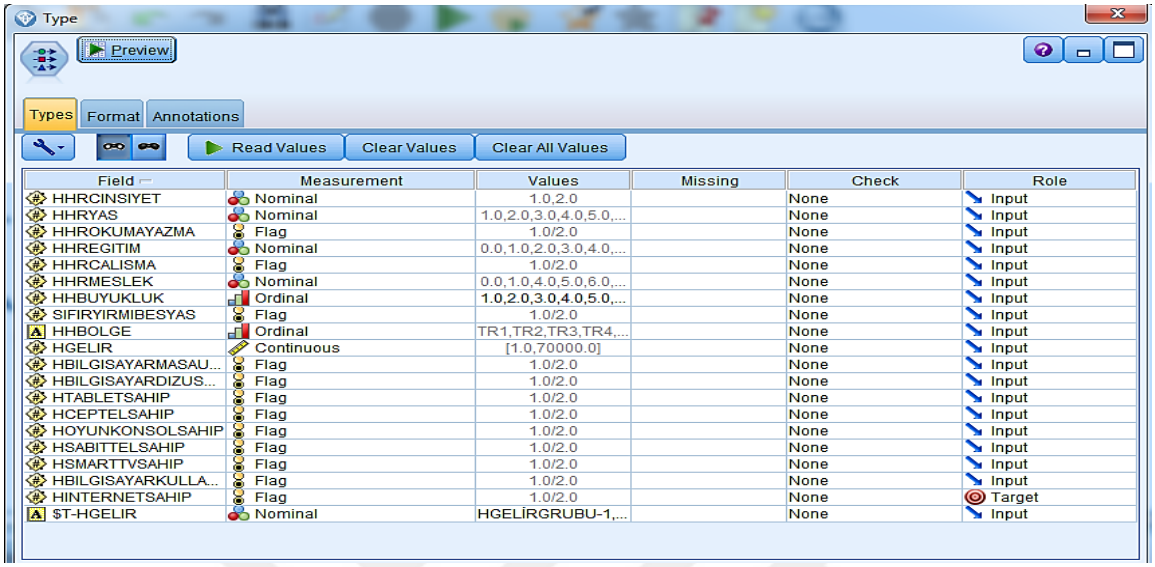
karşı hassasken, karar ağaçlarında kullanılan verinin fazla bir önişleme gerek duyulmadan kullanılabilmesi, çok büyük veri setlerine kolayca uygulanabilmesi, en çok kullanılan yöntemlerden biri olduğu için çok sayıda kaynak ve program olması gibi nedenlerden yöntem olarak karar ağaçları tercih edilmiştir. Verilerin analizinde IBM SPSS Modeler 18.1 veri madenciliği yazılımı kullanılmıştır. Modeler, hızlı bir şekilde öngörü modelleri geliştirilebilen, bireylerin karar alma yeteneğini yükselten ve bunları iş dünyasının içerisine entegre edebilen veri madenciliği araçları bütünüdür. CRISP-DM modeli etrafında tasarlanmış, daha iyi iş sonuçlarına götüren IBM SPSS Modeler tüm veri madenciliği sürecini desteklemektedir. YSA, birliktelik kuralları, sınıflandırma modelleme seçeneklerini içermektedir. Her bir model, ayrı güce sahip ve problemlerin ayrı türleri için uygundur. Resim 4.3, IBM SPSS Modeler programının ana ekranını göstermektedir. Resimde alt kısımda görülen Sources, Record Options, Modeling ve diğer sekmeler veri madenciliği uygulamasının gerçekleştirildiği ve yürütüldüğü sekmelerdir. Şeklin ortasında yer alan büyük alan ise veri madenciliği sürecinin ve veri akışının tanımlandığı alandır. Bütün eklenen nesneler bu alanda görülecektir [16].



Resim 4.3. IBM SPSS Modeler ana ekranı

Resim 4.4, kullanılan değişkenlerin türlerinin belirlendiği *Type* modülünü göstermektedir. Burada her değişkenin türleri ve modelleme aşamasında modele ne şekilde dâhil olacağı belirlenebilmektedir. *Field* kısmında kullanılan değişkenlerin adları, *Measurement* kısmında değişkenlerin tipleri (continuous, flag, nominal, ordinal, categorical, typless), *Values*

kısmında ilgili değişkenlerin değerleri ya da hangi aralıkta değerler aldığı, *Role* kısmında ise hangi değişkenin giriş hangi değişkenin hedef değişken olacağı seçilmektedir.

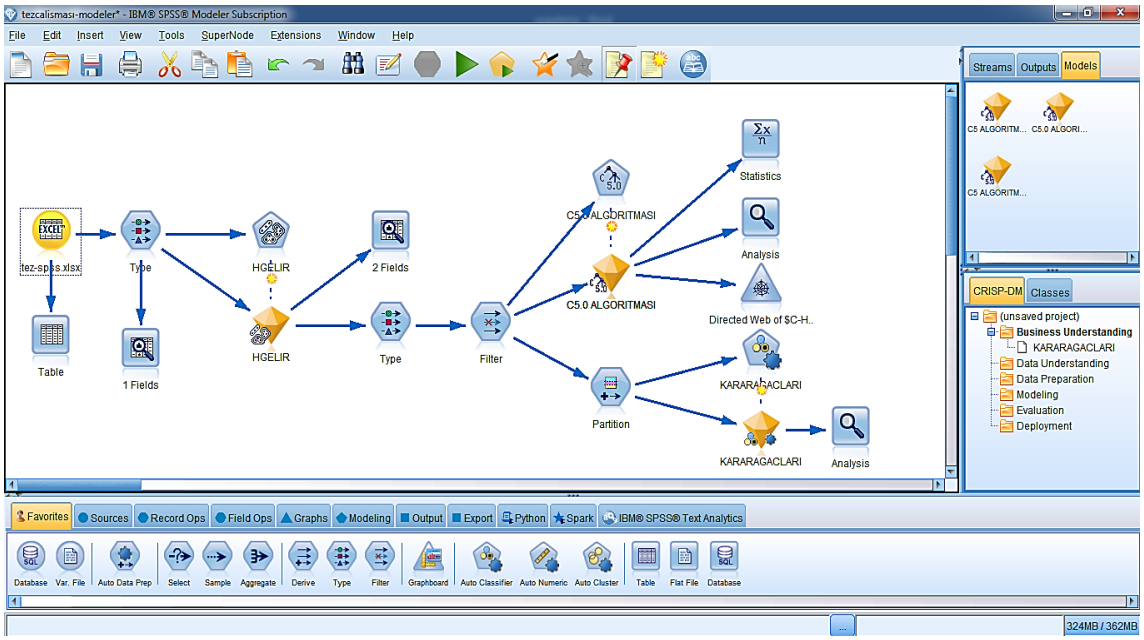


Field	Measurement	Values	Missing	Check	Role
HHRCINSIYET	Nominal	1.0,2.0		None	Input
HHRYAS	Nominal	1.0,2.0,3.0,4.0,5.0,...		None	Input
HHROKUMAYAZMA	Flag	1.0/2.0		None	Input
HHREGITIM	Nominal	0.0,1.0,2.0,3.0,4.0,...		None	Input
HHRCALISMA	Flag	1.0/2.0		None	Input
HHRMESLEK	Nominal	0.0,1.0,4.0,5.0,6.0,...		None	Input
HHBUYUKLUK	Ordinal	1.0,2.0,3.0,4.0,5.0,...		None	Input
SIFIRYIRMIBESYAS	Flag	1.0/2.0		None	Input
HHBOLGE	Ordinal	TR1,TR2,TR3,TR4,...		None	Input
HGELIR	Continuous	[1.0,70000.0]		None	Input
HBILGISAYARMASAU...	Flag	1.0/2.0		None	Input
HBILGISAYARDIZUS...	Flag	1.0/2.0		None	Input
HTABLETSAHIP	Flag	1.0/2.0		None	Input
HCEPTELSAHIP	Flag	1.0/2.0		None	Input
HOYUNKONSOLSAHIP	Flag	1.0/2.0		None	Input
HSABITTELSAHIP	Flag	1.0/2.0		None	Input
HSMARTTVSAHIP	Flag	1.0/2.0		None	Input
HBILGISAYARKULLA...	Flag	1.0/2.0		None	Input
HINTERNETSAHIP	Flag	1.0/2.0		None	Target
$T-HGELIR	Nominal	HGELİRGRUBU-1,...		None	Input

Resim 4.4. IBM SPSS Modeler tpye modülü

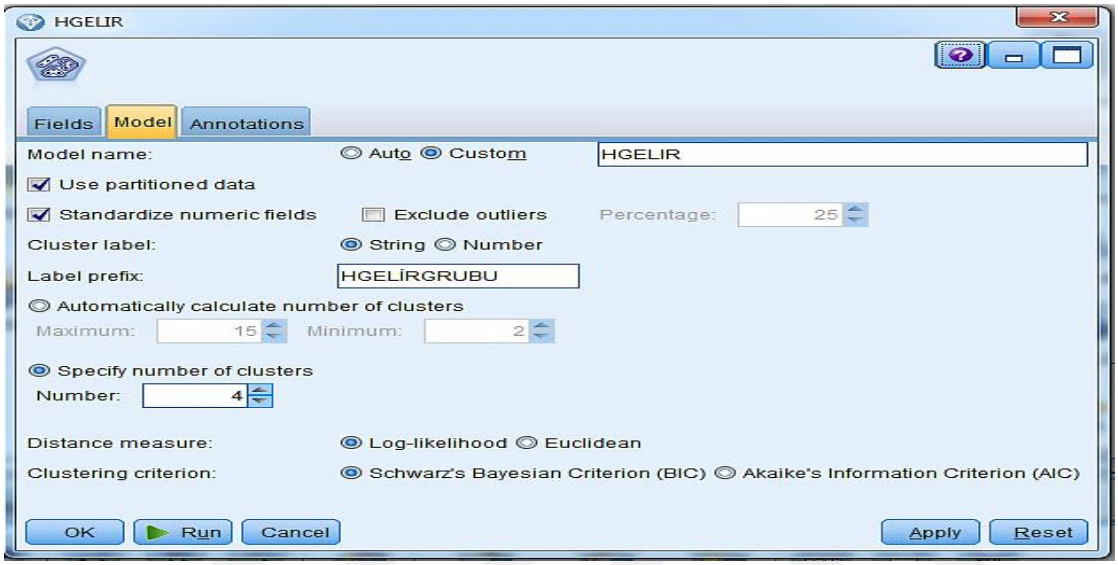
4.4. Karar Ağacı Algoritmaları ile Modelleme

Çalışmada hedef değişken olarak *hane internet sahipliği durumu* seçilmiştir. Çalışma için uygulanan modelleme ekran görüntüsü Resim 4.5'de gösterilmiştir.



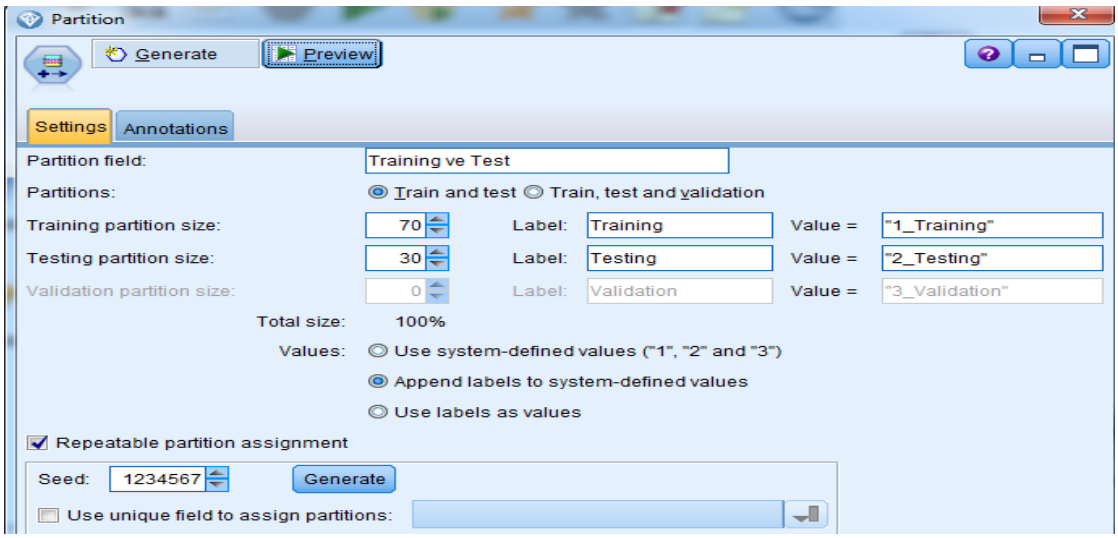
Resim 4.5. Karar ağacı ve iki aşamalı kümeleme ile modelleme

Gelir değişkeni için iki aşamalı kümeleme analizi yapılmıştır. Bu analiz için aşağıdaki şekilde belirtilen parametreler seçilmiştir. Küme sayısı 4 olarak belirlenmiş, uzaklık ölçümü tekniği olarak Log - Olabilirlik (Log - Likelihood), kümeleme kriteri olarak Schwarz'ın Bayesyen Kriteri (Schwarz's Bayesian Criterion - BIC) seçilmiştir. Seçim yapılan ekranın görüntüsü Resim 4.6'da gösterilmiştir.



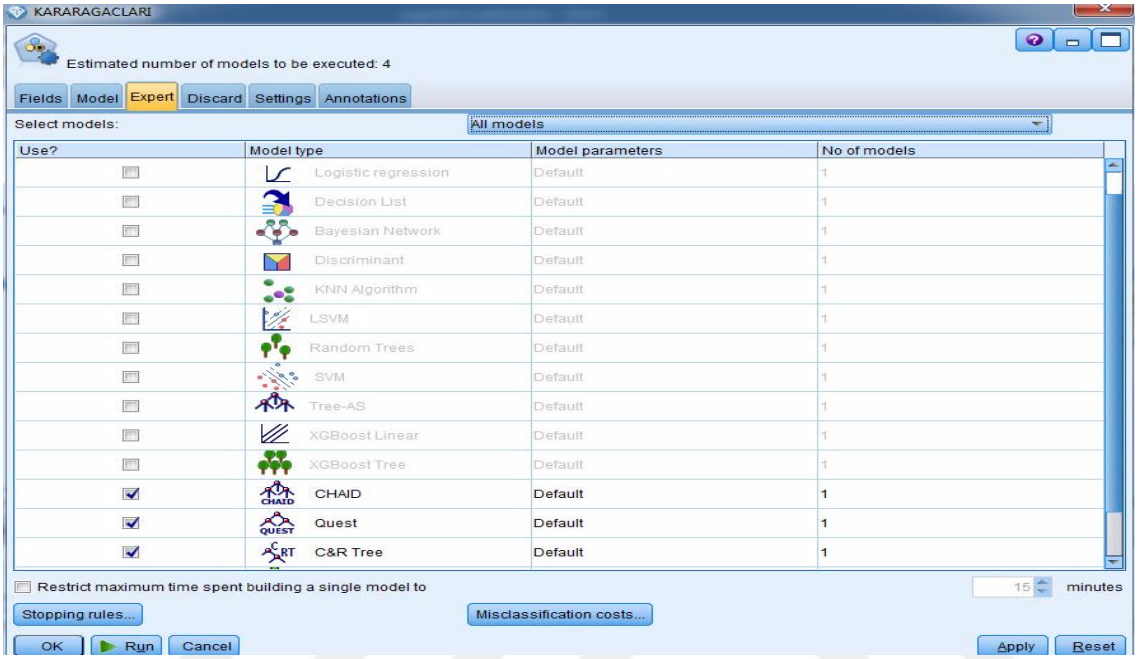
Resim 4.6. İki aşamalı kümeleme parametre seçim ekranı

Buna göre iki aşamalı kümeleme analizi ile yeniden oluşturulan gelir değişkeni ile diğer değişkenler alınarak oluşturulan veri kümesinin son hali Resim 4.7'de görüldüğü gibi *Partition* ile % 70 eğitim (training), % 30 test (test) olacak şekilde ikiye bölünmüştür.



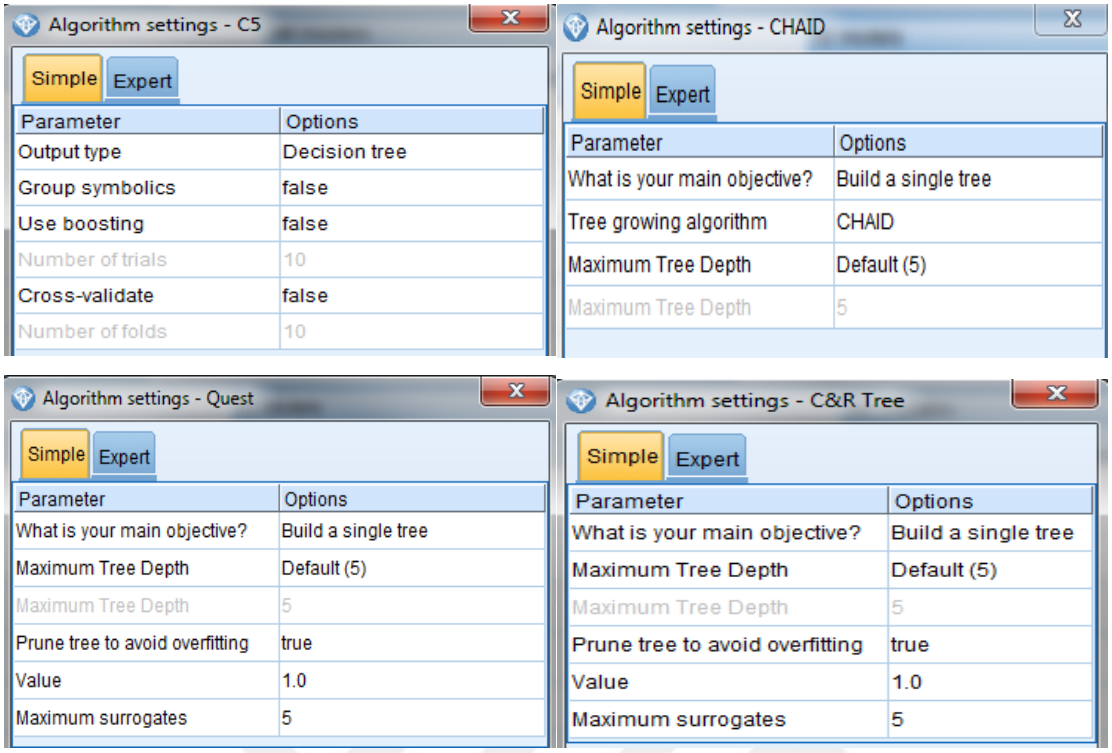
Resim 4.7. Veri kümesinin eğitim ve test verisi olarak ayrılması

Çalışmada hane internet hizmeti sahipliğini etkileyen faktörleri analiz edebilmek ve en iyi performans gösteren algoritmayı seçebilmek için *Otomatik Sınıflandırıcı (Auto Classifier)* ile dört karar ağacı algoritması seçilmiş ve bu algoritmaların performans sonuçları karşılaştırılmıştır. Karşılaştırma için seçilen karar ağacı ekranı Resim 4.8'de gösterilmiştir.



Resim 4.8. Karar ağacı algoritma seçimi

Seçilen algoritmaların parametreleri ile ilgili herhangi bir ayarlama yapılmamış ve tüm algoritmaların *Varsayılan (Default)* değerleri alınmıştır. Seçilen algoritmalar için varsayılan değerlere ait ayarlar Resim 4.9'da gösterilmiştir.



Resim 4.9. Seçilen algoritmaların varsayılan değerleri

Buna göre algoritmalar seçili olan simple (temel) modda varsayılan ayarlar ile çalışır. C5.0 algoritması için output type parametresi bir karar ağacı, kural kümesi ya da her ikisinin birlikte oluşturulup oluşturulacağının seçimini sağlar. Group symbolics parametresi seçilmezse (false), C5.0, ana düğümü bölmek için kullanılan alanın her değeri için bir alt düğüm oluşturur. Use boosting doğruluk oranını artırmak için C5.0 algoritmasına özel bir parametredir. Cross validate parametresi seçilirse, C5.0, tam veri kümesinde modelin doğruluğunu tahmin etmek için eğitim verilerinin alt kümeleri üzerine kurulmuş bir dizi model kullanacaktır. Tree growing algorithm parametresi CHAID algoritmasına özel olup CHAID algoritmasının türünün seçilmesini sağlar. Maximum Tree Depth (maksimum ağaç derinliği) parametresi kök düğümün altındaki maksimum seviye sayısının seçilmesini sağlar. Aynı parametre QUEST ve C&RT algoritmaları için de kullanılmaktadır. Varsayılan değeri 5'tir. Prune tree to avoid overfitting (aşırı uyumu önlemek için ağacı budama) parametresi QUEST ve C&RT algoritmalarına özel olup ağacın doğruluğuna önemli katkı sağlamayan alt seviye dallanmaları gidermek için kullanılır. Value parametresi, budanmış ağaç ile risk tahmini açısından en düşük risk taşıyan ağaç arasındaki risk tahminindeki izin verilen farkın boyutunu gösterir. Örneğin, 2 belirlenirse, risk tahmini, tam ağacınkinden büyük olan bir ağaç seçilebilir. Varsayılan durumu True, değeri 1'dir. Maximum surrogates parametresi, eksik değerlerle baş etmek için kullanılan bir yöntemdir. Ağaçtaki her bölme için algoritma,

seçilen bölme alanına en çok benzeyen giriş alanlarını tanımlar. Bu alanlar, bu bölünmenin vekilleridir. Sınıflandırılan bir kayıt bölünmüş bir alan için eksik bir değere sahipse, bölmeyi yapmak için bir vekil alandaki değeri kullanılabilir. Varsayılan değeri 5'tir. Ayrıca Expert (Uzman) ayarları ile algoritmaların diğer parametreleri de ayarlanabilir.

4.5. Sınıflandırma Modellerinin Performanslarının Karşılaştırılması

Sınıflandırıcıların performansı 3 ölçüt dikkate alınarak karşılaştırılmıştır. Bu ölçütler aşağıda sıralanmıştır.

1) Doğru Sınıflandırma Yüzdesi (Overall Accuracy)
2) Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan (Area Under Receiver Operating Characteristic (ROC) Curve -AUC)
3) Kaldıraç Değeri (Lift Ratio)

4.5.1. Doğru sınıflandırma yüzdesi

Karışıklık matrisinde yer alan değerleri (doğru pozitif, doğru negatif, yanlış pozitif, yanlış negatif) ve bu matrise göre hesaplanan ölçütlere değinilmişti (Bkz. Çizelge 3.9 ve Çizelge 3.10). Buna göre tezde kullanılan algoritmaların eğitim (training) ve test (testing) verilerine ait sınıflandırma tahmin sayıları ile yüzdeleri Çizelge 4.21, Çizelge 4.22, Çizelge 4.23, Çizelge 4.24'de gösterilmiştir. Eğitim (training) veri kümesi için en yüksek doğru tahmin sayısı 6 796 ve doğru (correct) sınıflandırma yüzdesi % 86,73 ile C5.0 algoritması model çıktısındadır. Test veri kümesi için ise en yüksek doğru tahmin sayısı 2 897 ve doğru sınıflandırma yüzdesi % 85,56 ile yine C5.0 algoritması sonucunda alınmıştır.

Çizelge 4.21.C5.0 algoritması tahmin sayı ve yüzdesi

'Training ve Test '	1_Training		2_Testing	
Correct	6,796	86.73%	2,897	85.56%
Wrong	1,040	13.27%	489	14.44%
Total	7,836		3,386	

Çizelge 4.22. CHAID algoritması tahmin sayı ve yüzdesi

'Training ve Test '	1_Training		2_Testing	
Correct	6,599	84.21%	2,868	84.7%
Wrong	1,237	15.79%	518	15.3%
Total	7,836		3,386	

Çizelge 4.23. C&RT algoritması tahmin sayı ve yüzdesi

'Training ve Test '	1_Training		2_Testing	
Correct	6,550	83.59%	2,850	84.17%
Wrong	1,286	16.41%	536	15.83%
Total	7,836		3,386	

Çizelge 4.24. QUEST algoritması tahmin sayı ve yüzdesi

'Training ve Test '	1_Training		2_Testing	
Correct	6,585	84.04%	2,860	84.47%
Wrong	1,251	15.96%	526	15.53%
Total	7,836		3,386	

Kullanılan algoritmaların eğitim (training) verileri için karışıklık matris tabloları Çizelge 4.25, Çizelge 4.26, Çizelge 4.27, Çizelge 4.28'deki gibidir.

Çizelge 4.25. C5.0 algoritması karışıklık matrisi

	1,000000	2,000000
1,000000	5 248	395
2,000000	645	1 543

Çizelge 4.26. CHAID algoritması karışıklık matrisi

	1,000000	2,000000
1,000000	5 340	303
2,000000	934	1 259

Çizelge 4.27. C&RT algoritması karışıklık matrisi

	1,000000	2,000000
1,000000	5 130	513
2,000000	773	1 420

Çizelge 4.28. QUEST algoritması karışıklık matrisi

	1,000000	2,000000
1,000000	5 251	392
2,000000	859	1 334

Algoritmaların eğitim (training) verisi için karışıklık matrisleri ile hesaplanan performans kriter değerleri Çizelge 4.29'da gösterilmiştir.

Çizelge 4.29. Karışıklık matrisi ile hesaplanan performans kriterleri

Performans Kriteri	Algoritma			
	C5.0	CHAID	C&RT	QUEST
Doğruluk Oranı	**%86,73**	%84,21	%83,59	%84,04
Hata Oranı	**%13,27**	%15,79	%16,41	%15,96
Kesinlik	%93,00	**%94,60**	%90,90	%93,00
Anma	**%89,00**	%85,10	%86,90	%85,90
F-Ölçütü	**%90,90**	%89,50	%88,80	%89,30

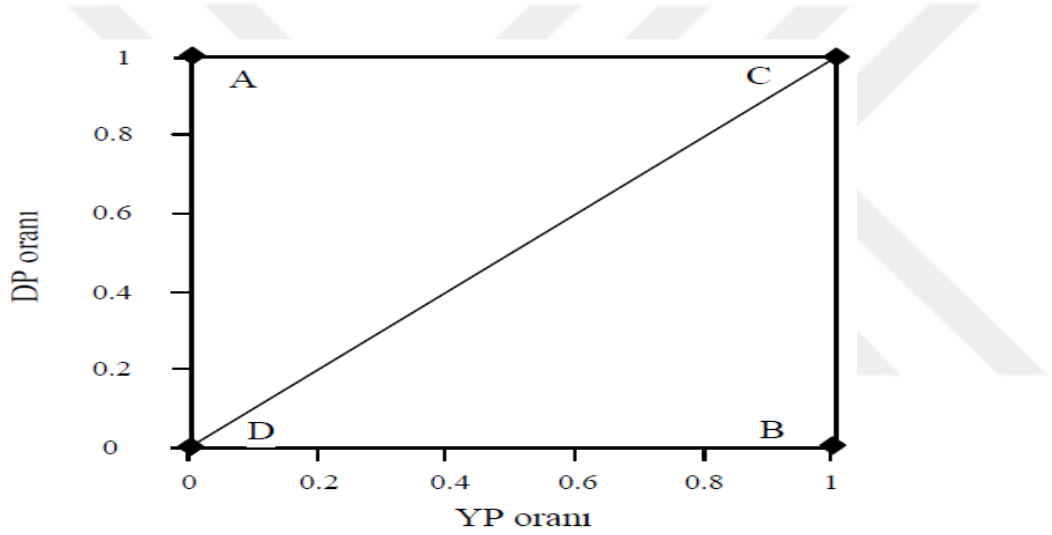
Eğitim veri kümesine göre performans kriterleri değerlendirildiğinde en yüksek doğruluk oranına sahip algoritmanın C5.0 algoritması olduğu görülmektedir. Yine hata oranı olarak en düşük sınıflandırma oranına sahip algoritma C5.0 algoritmasıdır. Kesinlik ve anma değerleri ile hesaplanan F ölçütü değeri, kesinlik ve anma değerlerinin tek başına yorumlanmasına göre daha etkilidir. F ölçütü değerlerine göre en başarılı algoritmanın % 90,90 oranı ile C5.0 algoritması olduğu görülmüştür.

4.5.2. Alıcı işletim karakteristiği ve grafiği

Alıcı İşletim Karakteristiği (AİK) (Receiving Operating Characteristics-ROC) grafiği farklı sınıflandırma performanslarını karşılaştırmak için geliştirilen yöntemlerdendir. AİK grafiğinde, DP oranı dikey eksende ve YP oranı yatay eksende gösterilir. Yanlış pozitif oranı Eş 3.14'deki gibi hesaplanır.

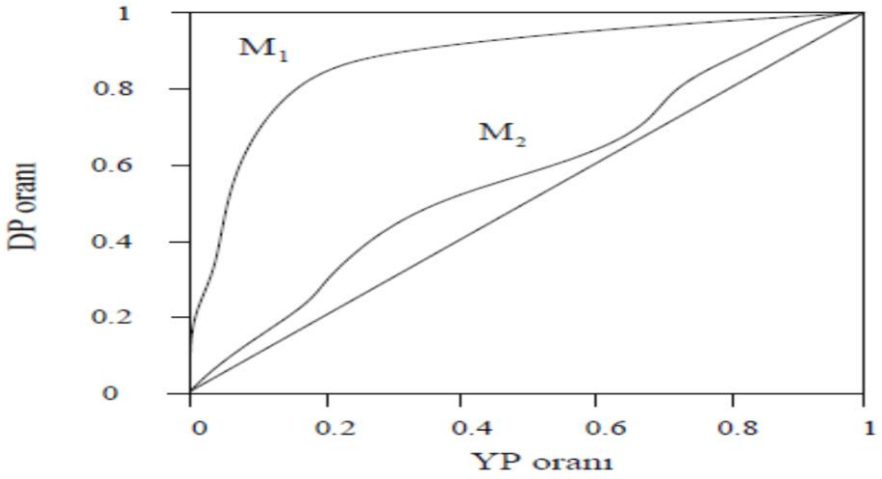
$$\text{Yanlış Pozitif Oranı (YP)} = \frac{\text{Yanlış Pozitif Sayısı (YP)}}{\left(\text{Doğru Pozitif Sayısı (DP)}+\text{Yanlış Negatif Sayısı (YN)}\right)} \quad (3.14)$$

AİK grafiğinde gösterilen her bir (x, y) noktası sırasıyla YP oranını ve DP oranını göstermektedir. AİK grafiği üzerindeki noktalar (0, 1), (1, 0), (1, 1) ve (0, 0) Şekil 3.11'de gösterilmiştir. Bu noktalar dört özel duruma (A, B, C ve D) karşılık gelmektedir. A, mükemmel sınıflandırıcı olarak tanımlanmaktadır ve grafik üzerinde, sol üst köşe olan en iyi noktaya karşılık gelmektedir. B, en kötü olası sınıflandırıcı olarak tanımlanmaktadır ve grafik üzerinde sağ alt köşede yer almaktadır. C, ultra-liberal sınıflandırıcı olarak tanımlanmakta ve anlamı sınıflandırıcının nesneleri her zaman pozitif sınıfa ayırması durumudur; D, ultra-muhafazakâr sınıflandırıcı olarak tanımlanmakta ve anlamı ise sınıflandırıcının nesneleri her zaman negatif sınıfa ayırması durumudur. Sonuçta, sol üst köşeye yakın olan grafik daha iyi sınıflandırıcı performansına sahiptir [13].



Şekil 3.11. ROC grafiğinde önemli noktalar [67]

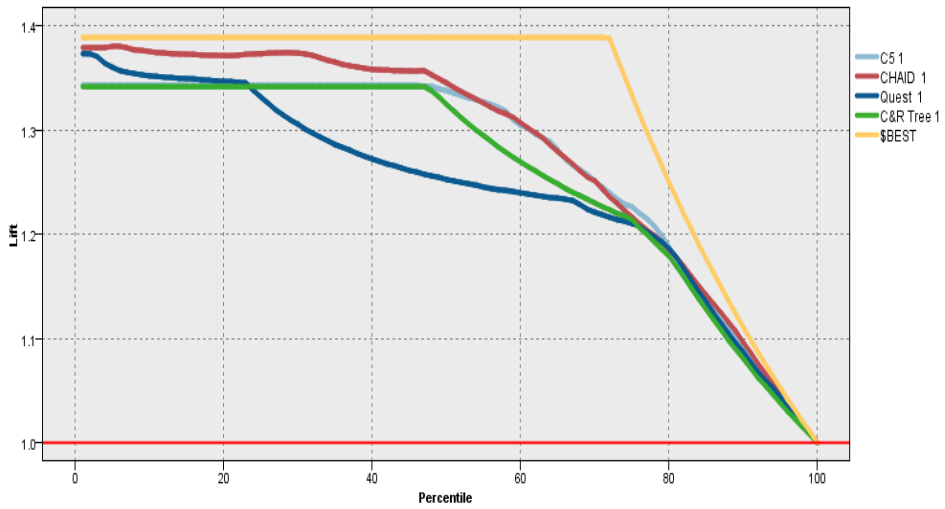
Şekil 3.12, iki sınıflandırma modelinin ROC eğrilerini göstermektedir. Şekilde sınıflandırma modellerinin ROC eğrileri M_1 ve M_2'dir. Rastgele tahminleri temsil eden çapraz çizgi de ayrıca gösterilmiştir. Bir modelin ROC eğrisi diyagonal olan çapraz çizgiye yakınsa model doğruluğu daha az olacaktır. Burada M_1 modeli daha yüksek doğruluğa sahiptir. Model gerçekten iyi ise, doğru pozitiflerle karşılaşma olasılığı daha yüksektir. Böyle bir durumda, eğri dik olarak yukarı doğru hareket eder. Daha sonra, daha az doğru pozitif ve daha fazla yanlış pozitifle karşılaşmaya başlanıldığı zaman, eğri gevşer ve daha fazla yatay olur. Bir modelin doğruluğunu değerlendirmek için eğrinin altındaki alan ölçülebilir. Birçok yazılım paketi böyle bir hesaplamayı yapabilir. Alan 0,5'e yaklaştıkça, model daha az doğrudur. Mükemmel doğrulukta bir modelin alanı 1.0 olacaktır [30].



Şekil 3.12. İki sınıflandırma modelinin ROC eğrileri [30]

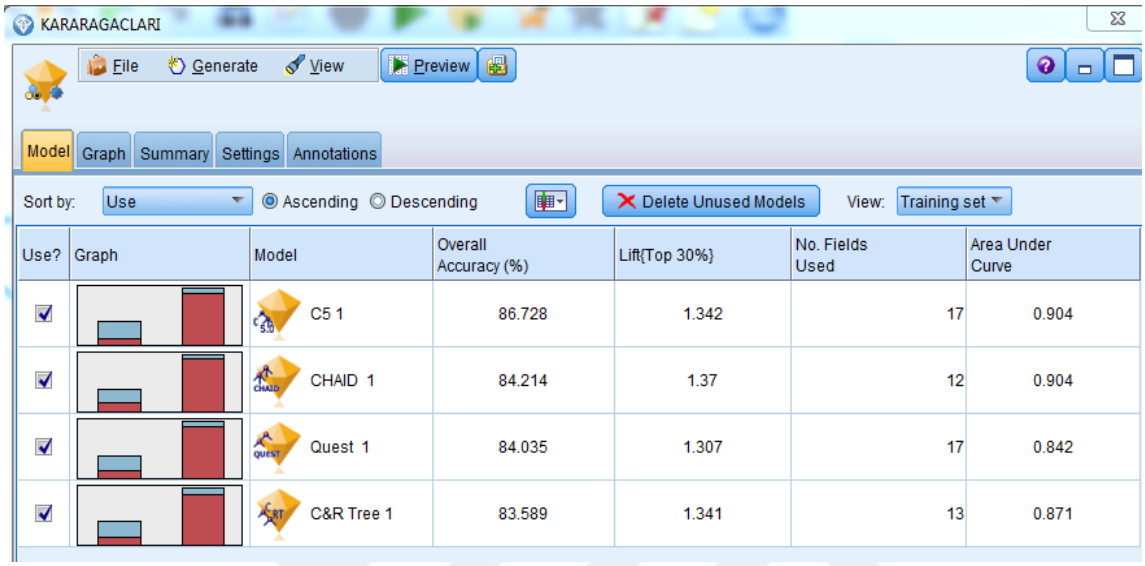
4.5.3. Kaldıraç değeri ve grafiği

Uygulamada kullanılan algoritmalara ait %30'luk değere karşılık gelen kaldıraç (lift) oranı ile alınan kaldıraç grafiği Şekil 4.1'de gösterilmiştir. Buna göre kullanılan algoritmalar için, sarı çizgi en iyi çizgi olarak tanımlanan en üstteki modeli, kırmızı olan çizgi temel düzeyi, aradaki çizgiler ise modellerin yaptığı tahmin doğruluğunu göstermektedir. Grafik yorumlandığında en iyi modele en yakın olan algoritmanın CHAID algoritması olduğu görünmektedir. Daha sonra sırasıyla C5.0, C&RT ve QUEST algoritmaları en iyi modele yakınlık olarak sıralanabilir.



Şekil 4.1. Uygulamada kullanılan algoritmaların kaldıraç grafiği

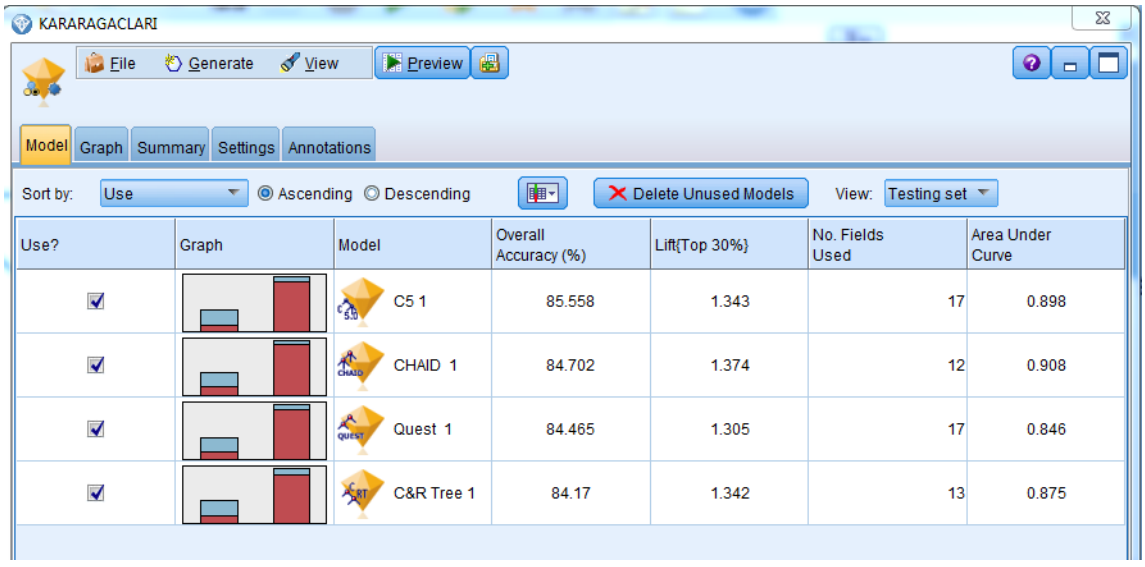
Seçilen sınıflandırma algoritmalarına göre eğitim (%70) ve test verileri (%30) üzerinde bu algoritmalara ait performans kriterleri ve sonuçları Resim 4.10 ve Resim 4.11'de gösterilmiştir.



Use?	Graph	Model	Overall Accuracy (%)	Lift{Top 30%}	No. Fields Used	Area Under Curve
☑		C5 1	86.728	1.342	17	0.904
☑		CHAID 1	84.214	1.37	12	0.904
☑		Quest 1	84.035	1.307	17	0.842
☑		C&R Tree 1	83.589	1.341	13	0.871

Resim 4.10. Eğitim veri kümesinde karar ağacı algoritma performansları

Eğitim veri kümesinde doğruluk oranı yüzdesi % 86,728 ile en yüksek olan C5.0 algoritmasıdır. Kaldıraç değeri en yüksek olan 1,37 değeri ile CHAID algoritması, Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan (Area Under Curve) değeri, 0,904 değerleri ile C5.0 ve CHAID algoritmalarıdır. Sonuç olarak eğitim veri kümesi için en başarılı algoritmalar C5.0 ve CHAID algoritmaları olmuştur.



Use?	Graph	Model	Overall Accuracy (%)	Lift{Top 30%}	No. Fields Used	Area Under Curve
☑		C5 1	85.558	1.343	17	0.898
☑		CHAID 1	84.702	1.374	12	0.908
☑		Quest 1	84.465	1.305	17	0.846
☑		C&R Tree 1	84.17	1.342	13	0.875

Resim 4.11. Test veri kümesinde karar ağacı algoritma performansları

Test veri kümesinde doğruluk oranı yüzdesi % 85,558 ile en yüksek olan C5.0 algoritması, Kaldıraç değeri en yüksek olan 1,37 değeri ile CHAID algoritması, Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan (Area Under Curve-AUC) değeri 0,908 değeri ile CHAID algoritması olmuştur. Sonuç olarak test veri kümesi için de en başarılı algoritmalar yine C5.0 ve CHAID algoritmaları olmuştur. Eğitim ve test veri setleri için sınıflandırma ölçüt değerlerine göre en başarılı ve başarısız algoritmalar Çizelge 4.30 ve Çizelge 4.31'de gösterilmiştir.

Çizelge 4.30. Eğitim veri kümesi için en başarılı ve başarısız algoritmalar

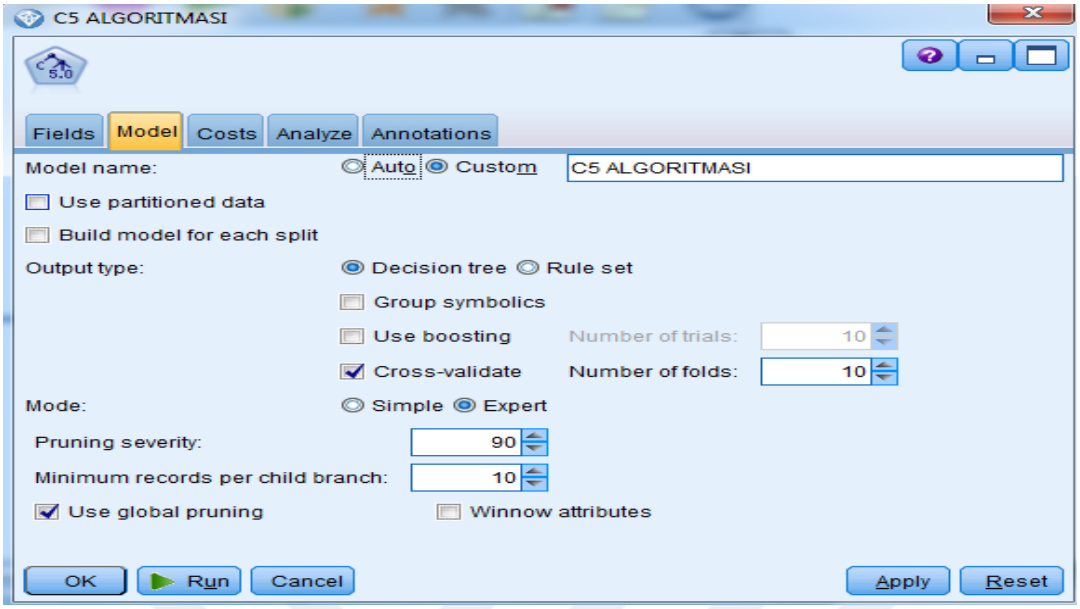
Sınıflandırma Ölçütü	En Başarılı Algoritma	Ölçüt Değeri	En Başarısız Algoritma	Ölçüt Değeri
Overall Accuracy (%)	C5.0	86,728	C&RT	83,598
Lift Ratio (%30)	CHAID	1,370	QUEST	1,307
Area Under Curve	C5.0 ve CHAID	0,904	QUEST	0,842

Çizelge 4.31. Test veri kümesi için en başarılı ve başarısız algoritmalar

Sınıflandırma Ölçütü	En Başarılı Algoritma	Ölçüt Değeri	En Başarısız Algoritma	Ölçüt Değeri
Overall Accuracy (%)	C5.0	85,558	C&RT	84,17
Lift Ratio (%30)	CHAID	1,374	QUEST	1,305
Area Under Curve	CHAID	0,908	QUEST	0,846

4.6. C5.0 Algoritması ile Karar Ağacı Modelleme

Doğruluk oranı ve F- ölçütü için en başarılı algoritma olduğundan, lift ratio, AUC değerleri içinde yine en başarılı iki algoritmadan biri olduğundan verilerin modellenmesi, karar ağacının ve kuralların oluşturulması için C5.0 algoritması kullanılmıştır. C5.0 algoritması ile modelleme için Resim 4.12'de gösterilen parametreler seçilmiştir. Bu parametrelerin ne olduğu ve hangi değerlerin seçildiği aşağıda açıklanmıştır [68].



Resim 4.12. C5.0 algoritması için parametre seçim ekranı

Model adı (Model name): Üretilecek modelin adını belirtmek için kullanılır.

- Otomatik (Auto): Bu seçenek seçildiğinde, hedef alan adlarına göre model adı otomatik olarak oluşturulacaktır. Bu seçenek varsayılan olarak seçilmiştir.
- Özel (Custom): Bu seçenek oluşturulacak model için kendi isminizi belirlemek için kullanılır.

Bölümlenmiş verileri kullan (Used partitioned data): Bir bölüm alanı tanımlanmışsa, bu seçenek modeli oluşturmak için yalnızca eğitim bölümünden alınan verilerin kullanılmasını sağlar.

Bölünmüş modeller oluştur (Build model for each split): Bölme alanları olarak belirtilen giriş alanlarının olası her değeri için ayrı bir model oluşturur.

Çıkış tipi (Output type): Ortaya çıkan modelin bir karar ağacı ya da kural kümesi olması seçiminin yapılmasını sağlar.

- Karar ağacı (Decision tree): Bu seçenek seçilirse modelin çıktısı bir karar ağacı olur.
- Kural kümesi (Rule set): Bu seçenek seçilirse modelin çıktısı bir kural kümesi olur.

Grup sembolik (Group symbolics): Bu seçenek seçilirse, C5.0 algoritması, çıktı alanına göre benzer desenlere sahip sembolik değerleri birleştirmeyi deneyecektir. Bu seçenek seçilmezse, C5.0 ana düğümü ayırmak için kullanılan alanın her değeri için bir alt düğüm oluşturacaktır.

Artırmayı kullan (Use boosting): C5.0 algoritmasının doğruluk oranını artırabilmek için kullanılan boosting olarak adlandırılan özel bir metot vardır. Birden fazla model oluşturarak çalışır. İlk model her zamanki şekilde inşa edilmiştir. İkinci model, ilk model tarafından yanlış sınıflandırılan kayıtlara odaklanacak şekilde inşa edilmiştir. Daha sonra, ikinci modelin hatalarına odaklanmak için üçüncü bir model oluşturulur. Son olarak, vakalar, ayrı tahminleri tek bir genel tahminle birleştirmek için ağırlıklı bir oylama prosedürü kullanarak, tüm model setini onlara uygulayarak sınıflandırır. Arttırma, bir C5.0 modelinin doğruluğunu önemli ölçüde artırabilir, ancak daha uzun bir eğitim gerektirir. Denemeler sayısı (Number of trials) seçeneği model doğruluğunu artırmak için kaç kez deneme yapılacağı sayısının girileceği alandır.

Çapraz doğrulama (Cross validate): Bu seçenek seçilirse, C5.0, tam veri kümesinde modelin doğruluğunu tahmin etmek için eğitim verilerinin alt kümeleri üzerine kurulmuş bir dizi model kullanacaktır. Bu, veri kümesi geleneksel eğitim ve test kümelerine bölünemeyecek kadar küçükse yararlıdır. Doğruluk tahmini hesaplandıktan sonra çapraz doğrulama modelleri atılır. Kat sayısı (Number of folds) katlama sayısının veya çapraz doğrulama için kullanılan model sayısının belirtildiği alandır.

Mod (Mode): C5.0 algoritmasının temel ya da uzman modunda ayarlarının yapılmasını sağlar.

- Temel (Simple): Algoritma için temel modun seçilmesini sağlar ve varsayılan ayarlar ile çalışır.
- Uzman (Expert): Algoritma için uzman modunun seçilmesini sağlar. Eğitim parametreleri üzerinde daha doğrudan kontrol sağlar.

Temel (Simple) mod seçenekleri;

- İyileştirme (Favor): C5.0 algoritması mümkün olan en doğru ağacı üretmeye çalışacaktır. Bu durum bazı durumlarda, model yeni verilere uygulandığında zayıf performansa neden olabilen aşırı çalışmaya neden olabilir. Bu soruna daha az duyarlı olan algoritma ayarlarını kullanmak için Genellik (Generality) seçeneği seçilmelidir.
- Beklenen gürültü (Expected noise): Eğitim setindeki gürültülü veya hatalı verilerin beklenen oranını belirtmekte kullanılır.

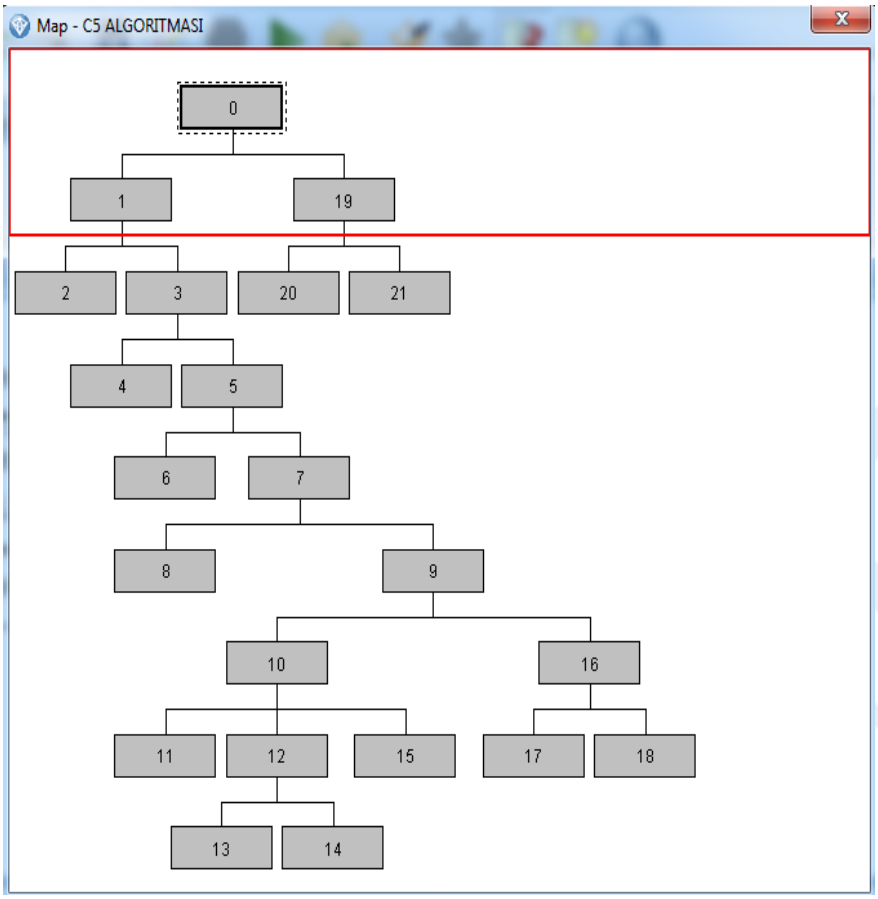
Uzman (Expert) modu seçenekleri;

- Budama şiddeti (Prunning severity): Karar ağacının veya kural kümesinin budanma derecesini belirler. Daha küçük, daha özlü bir ağaç elde etmek için bu değerin artırılması daha doğru bir ağaç elde etmek için ise azaltılması gereklidir. Bu ayar sadece yerel budamayı etkiler.
- Alt dallar için minimum kayıt (Minimum records per childs brunch): Alt grupların büyüklüğü, ağacın herhangi bir dalındaki bölme sayısını sınırlamak için kullanılabilir. Ağacın bir dalı, ancak, sonuçta ortaya çıkan alt dalların iki veya daha fazlası, eğitim kümesinden en az bu kadar çok kayıt içerdiğinde bölünecektir. Varsayılan değeri 2'dir. Gürültülü verilerle aşırı yüklenmeyi önlemek için bu değerin artırılması gerekmektedir.

Global budama bullan (Use global prunning): Ağaçlar iki aşamada budanır: Birincisi, alt tipleri inceleyen ve modelin doğruluğunu arttırmak için dalları daraltan yerel bir budama aşamasıdır. İkincisi, genel bir budama aşaması olup ağacı bir bütün olarak ele alır ve zayıf alt altlıklar da çökebilir. Global budama varsayılan olarak gerçekleştirilir.

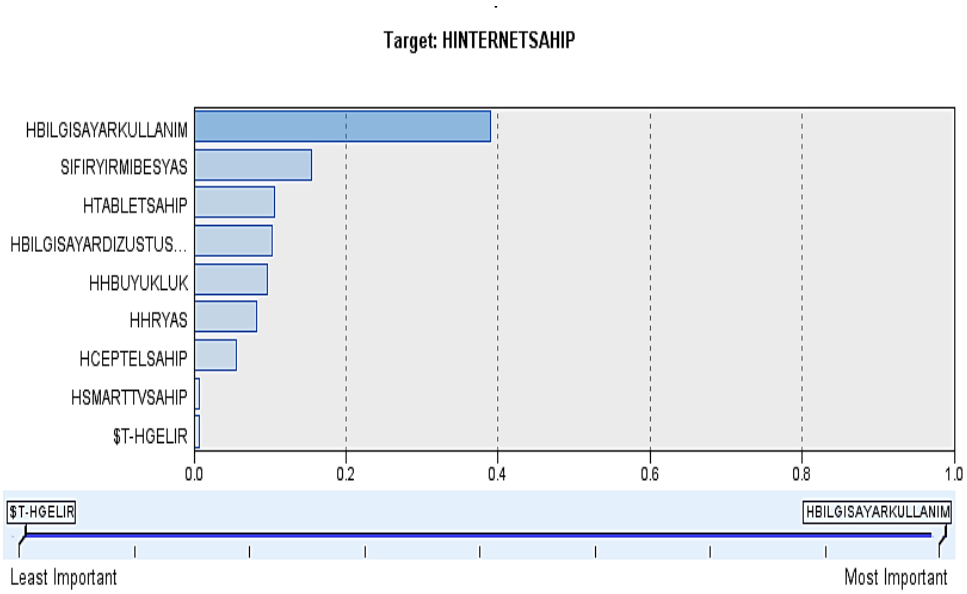
Winnow öznitelikleri (Winnow attiributes): Bu seçenek seçilirse, C5.0 modeli oluşturmaya başlamadan önce değişkenlerin kullanışlılığını inceleyecektir. Konuyla ilgisi olmayan değişkenler daha sonra model oluşturma sürecinden çıkarılır. Bu seçenek birçok değişken alanı olan modeller için yararlı olabilir ve çok sayıda değişkene sahip olma özelliğini azaltır. Çalışmada çapraz doğrulama (cross validate) değeri 10 olarak seçilmiştir. Çapraz doğrulama yöntemi, tüm veriyi kullanarak sınıflandırma yapmayı sağlayan bir metottur. Diğer parametre değerlerinden Budama şiddeti (Prunning severity) değeri karar ağacının daha kolay oluşması ve yorumlanabilmesi için 90, her alt dalda minimum 10 kayıt olacak şekilde

belirlenmiştir. Son olarak Global budama (Global prunning) seçeneği de seçilerek model oluşturulmaya hazır hale gelmiştir. Seçilen C5.0 algoritması parametre değerleri ile 10 dallanma 21 düğümden oluşan karar ağacı haritası Resim 4.13'de gösterilmiştir.



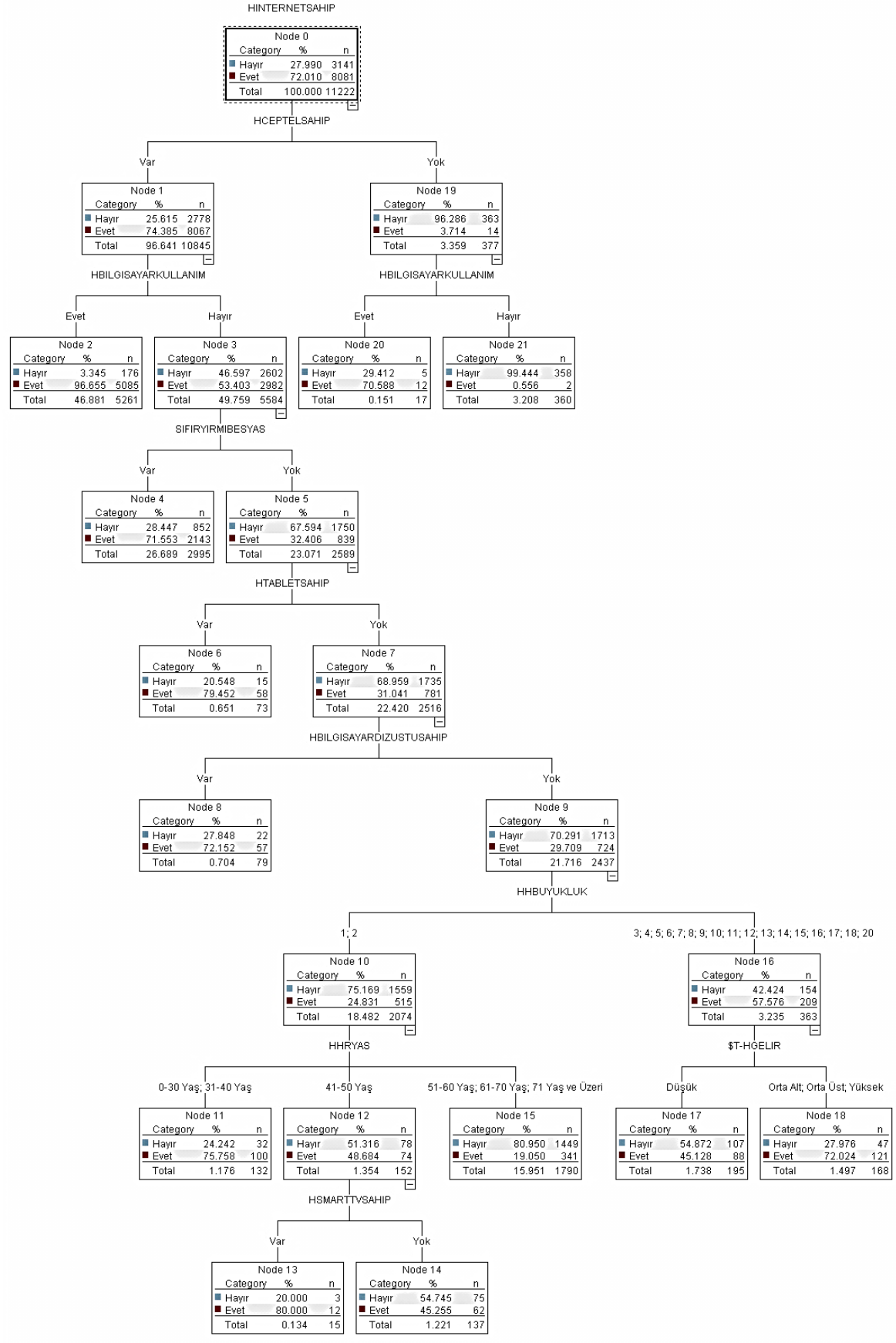
Resim 4.13. Karar ağacı haritası

18 adet giriş (input) 1 adet hedef (target) değişkeni ile başlanan analizde 9 adet değişkenin karar ağacında yer aldığı görülmüştür. Bu değişkenler HCEPTELSAHIP, HBILGISAYARKULLANIM, SIFIRYIRMIBESYAS, HTABLETSAHIP, HBILGISAYARDIZUSTUSAHIP, HHBUYUKLUK, HHRYAS, HSMARTTVSAHIP, HGELIRGRUBU olup önem derecelerine göre Resim 4.14'de görünmektedir.



Resim 4.14. Değişken önem durumları

C5.0 algoritması ile modelleme yapıldığında elde edilen 10 dallanma, 21 düğümden oluşan karar ağacı Resim 4.15'de gösterilmiştir.

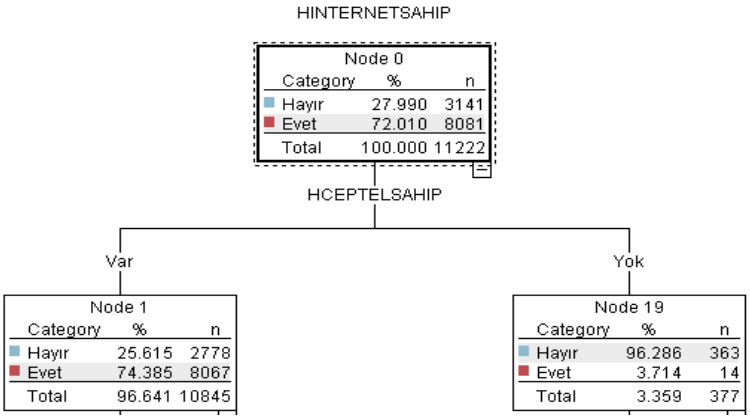


Resim 4.15. C5.0 algoritması ile oluşturulan karar ağacı

Resim 4.15'de yer alan her bir dal ve bu dallara ait düğümler detaylı bir şekilde ele alınarak aşağıda anlatılmıştır. Hedef değişken olarak seçilen HINTERNETSAHIP değişkenini en iyi açıklayan değişkenin HCEPTELSAHIP değişkeni olduğu görülmüştür.

Birinci Dallanma: HCEPTELSAHIP (Hane Cep Telefonu Sahipliği)

Resim 4.16'ya göre HINTERNETSAHIP değişkeni kök düğümde (Node 0) yer almaktadır. Bu düğüme göre 8 801 hanede internet olduğu, 3 141 hanede internet olmadığı görülmüştür. Hanesinde internet olanların oranı % 72,01 iken hanesinde internet olmayanların oranı % 27,99'dur.

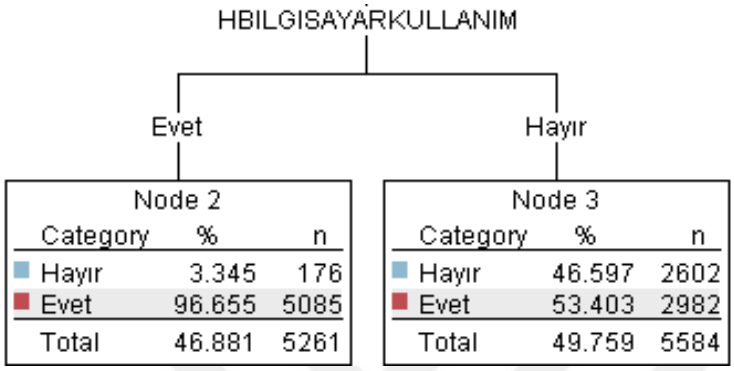


Resim 4.16. Birinci dallanmaya ait C5.0 analizi sonuçları

Oluşturulan ağaca göre dallanma HCEPTELSAHIP değişkenine göre olmuştur. Hanede cep telefonu olup olmama durumuna göre dallanma gerçekleşmiştir. Düğüm 1'e (Node 1) göre cep telefonu sahibi olan hane sayısı 8 067, oranı % 72,385'dir. Cep telefonu olmayan hane sayısı 2 778, oranı % 25,615'dir. Bu düğüm hane internet hizmeti sahipliği oranının % 96,641'ni oluşturmaktadır. Düğüm 19'a (Node 19) göre cep telefonu sahibi olmayan hanelerden internet sahibi olanların sayısı 14, oranı % 3,714, internet sahibi olan hane sayısı 363, oranı % 96,286'dır. Bu düğüm hane internet hizmeti sahipliği oranının % 3,359'nu oluşturmaktadır.

İkinci Dallanma: HBILGISAYARKULLANIM (Hane Bilgisayar Kullanımı)

C5.0 algoritması analizi sonucunda HCEPTEL değişkenini en iyi açıklayan değişkenin HBILGISAYARKULLANIM değişkeni olduğu görülmüştür. Hanede bilgisayar kullanımı olup olmama durumuna göre Resim 4.17'deki gibi dallanma gerçekleşmiştir.

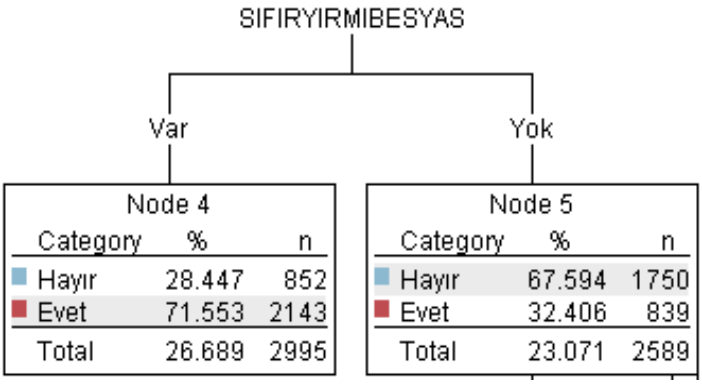


Resim 4.17. İkinci dallanmaya ait C5.0 analizi sonuçları

Düğüm 2'ye (Node 2) göre cep telefonu sahibi olan ve bilgisayar kullanımı olan hanelerden internet sahibi olan hane sayısı 5 085, oranı % 96,655'dir. İnternet sahibi olmayan hane sayısı ise 176, oranı % 3,345'dir. Bu düğüm hane bilgisayar kullanımı oranının % 46,881'ni oluşturmaktadır. Düğüm 3'e (Node 3) göre cep telefonu sahibi olan ve bilgisayar kullanımı olmayan hanelerden internet sahibi olan hane sayısı 2 982, oranı % 53,403, internet sahibi olmayan hane sayısı ise 2 602, oranı % 46,597'dir. Bu düğüm hane bilgisayar kullanım oranının % 49,759'nu oluşturmaktadır.

Üçüncü Dallanma: SIFIRYIRMIBESYAS (Hanede 0-25 Yaş Arası Bireyin Varlığı)

C5.0 algoritması analizi sonucunda HBILGISAYARKULLANIMI değişkenini en iyi açıklayan değişkenin SIFIRYIRMIBESYAS değişkeni olduğu görülmüştür. Hanede 0-25 yaş arasında birey olup olmama durumuna göre Resim 4.18'deki gibi dallanma gerçekleşmiştir.

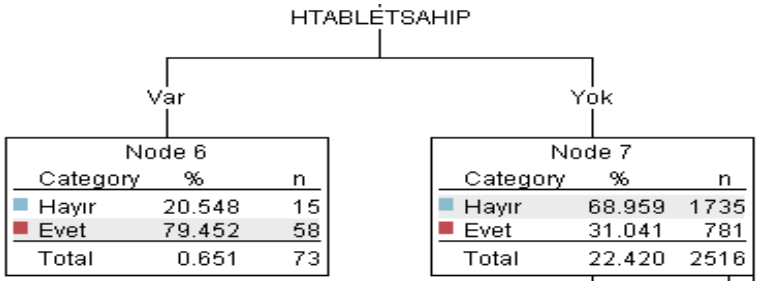


Resim 4.18. Üçüncü dallanmaya ait C5.0 analizi sonuçları

Düğüm 4'e (Node 4) göre bilgisayar kullanımı olmayan ve 0-25 yaş arası bireye sahip olan hanelerden internet sahibi olan hane sayısı 2 143, oranı % 71,553'tür. İnternet sahibi olmayan hane sayısı ise 852, oranı % 28,447'dir. Bu düğüm 0-25 yaş arası bireye sahip olan hanelerin % 26,689'nu oluşturmaktadır. Düğüm 5'e (Node 5) göre bilgisayar kullanımı olmayan ve 0-25 yaş arası bireye sahip olmayan hanelerden internet sahibi olan hane sayısı 839, oranı % 32,406, internet sahibi olmayan hane sayısı ise 1 750, oranı % 67,594'tür. Bu düğüm 0-25 yaş arası bireye sahip olmayan hanelerin % 23,071'ni oluşturmaktadır.

Dördüncü Dallanma: HTABLETSAHIP (Hane Tablet Sahipliği)

C5.0 algoritması analizi sonucunda SIFIRYIRMIBESYAS değişkenini en iyi açıklayan değişkenin HTABLETSAHIP değişkeni olduğu görülmüştür. Hanenin tablet sahibi olup olmama durumuna göre Resim 4.19'daki gibi dallanma gerçekleşmiştir.



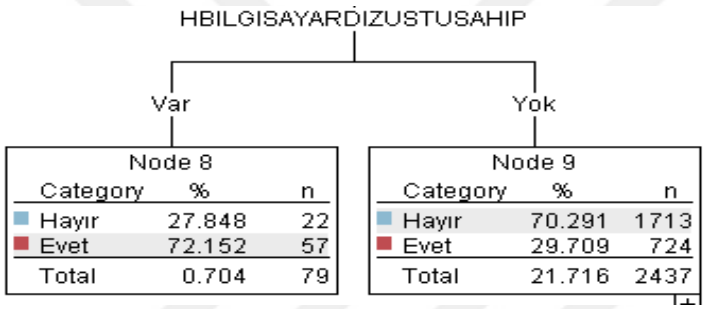
Resim 4.19. Dördüncü dallanmaya ait C5.0 analizi sonuçları

Düğüm 6'ya (Node 6) göre 0-25 yaş arası bireye sahip olmayan ve tablet sahibi olan hanelerden internet sahibi olan hane sayısı 58, oranı % 79,452'dir. İnternet sahibi olmayan hane sayısı ise 15, oranı % 20,548'dir. Bu düğüm tablet sahibi olan hanelerin % 0,651'ni

oluşturmaktadır. Düğüm 7'ye (Node 7) göre 0-25 yaş arası bireye sahip olmayan ve tablet sahipliği olmayan hanelerden internet sahibi olan hane sayısı 781, oranı % 31,041'dir. İnternet sahibi olmayan hane sayısı ise 1 735, oranı % 68,959'dur. Bu düğüm tablet sahibi olmayan hanelerin % 22,420'sini oluşturmaktadır.

Beşinci Dallanma: HBILGISAYARDIZUSTUSAHIP (Hane Dizüstü Bilgisayar Sahipliği)

C5.0 algoritması analizi sonucunda HTABLETSAHIP değişkenini en iyi açıklayan değişkenin HBILGISAYARDIZUSTUSAHIP değişkeni olduğu görülmüştür. Hanede dizüstü bilgisayar olup olmama durumuna göre Resim 4.20'deki dallanma gerçekleşmiştir.

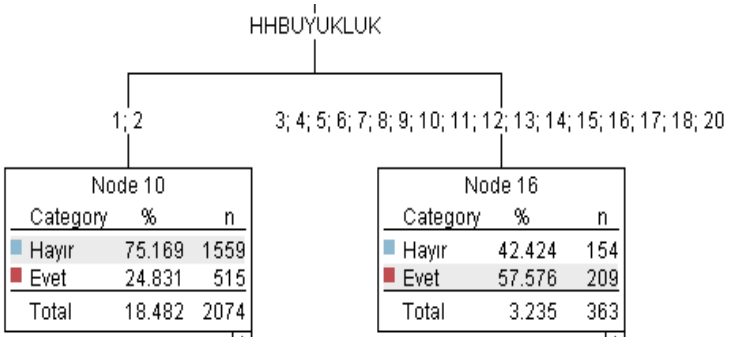


Resim 4.20. Beşinci dallanmaya ait C5.0 analizi sonuçları

Düğüm 8'e (Node 8) göre tableti sahibi olmayan ve dizüstü bilgisayarı olan hanelerden internet sahibi olan hane sayısı 57, oranı % 72,152'dir. İnternet sahibi olmayan hane sayısı ise 22, oranı % 27,848'dir. Bu düğüm tablet sahibi olan hanelerin % 0,704'nü oluşturmaktadır. Düğüm 9'a (Node 9) göre tablet sahibi olmayan ve dizüstü bilgisayarı olmayan hanelerden internet sahibi olan hane sayısı 724, oranı % 29,704'tür. İnternet sahibi olmayan hane sayısı ise 1 713, oranı % 70,291'dir. Bu düğüm dizüstü bilgisayarı olmayan hanelerin % 21,716'sını oluşturmaktadır.

Altıncı Dallanma: HHBUYUKLUK (Hanehalkı Büyüklüğü)

C5.0 algoritması analizi sonucunda HDIZUSTUBILGISAYARSAHIP değişkenini en iyi açıklayan değişkenin HHBUYUKLUK değişkeni olduğu görülmüştür. Hanehalkı büyüklüğüne göre Resim 4.21'deki dallanma gerçekleşmiştir.

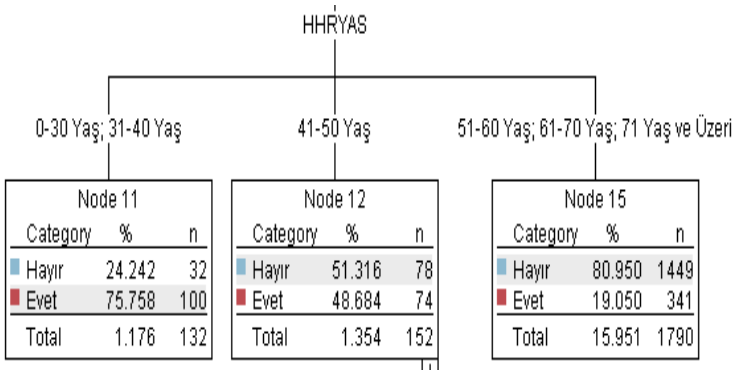


Resim 4.21. Altıncı dallanmaya ait C5.0 analizi sonuçları

Düğüm 10'a (Node 10) göre dizüstü bilgisayarı olmayan ve hanehalkı büyüklüğü 1 ve 2 olan hanelerden internet sahibi olan hane sayısı 515, oranı % 24,831'dir. İnternet sahibi olmayan hane sayısı ise 1 559, oranı % 75,169'dur. Bu düğüm hanehalkı büyüklüğünün % 18,482'ni oluşturmaktadır. Düğüm 16'e (Node 11) göre dizüstü bilgisayarı olmayan ve hanehalkı büyüklüğü 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 olan hanelerden internet sahibi olan hane sayısı 209, oranı % 57,576'dır. İnternet sahibi olmayan hane sayısı ise 154, oranı % 42,424'dür. Bu düğüm hanehalkı büyüklüğünün % 3,235'ni oluşturmaktadır.

Yedinci Dallanma: HHRYAS (Hanehalkı Reisinin Yaşı)

C5.0 algoritması analizi sonucunda HHBUYUKLUK değişkenini en iyi açıklayan değişkenin HHRYAS değişkeni olduğu görülmüştür. Hanehalkı reisinin yaşına göre Resim 4.22'deki dallanma gerçekleşmiştir.



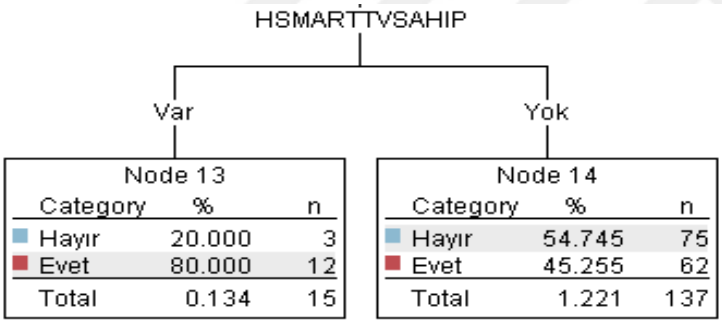
Resim 4.22. Yedinci dallanmaya ait C5.0 analizi sonuçları

Düğüm 11'e (Node 11) göre hanehalkı büyüklüğü 1 ve 2 olan ve hanehalkı reisinin yaşı 0-30, 31-40 yaş aralığında olan hanelerden internet sahibi olan hane sayısı 100, oranı % 75,758'dir. İnternet sahibi olmayan hane sayısı ise 32, oranı % 24,242'dur. Bu düğüm

hanehalkı reisi yaşlarının % 18,482'ni oluşturmaktadır. Düğüm 12'e (Node 12) göre hanehalkı büyüklüğü 1 ve 2 olan ve hanehalkı reisini yaş aralığı 41-50 olan hanelerden internet sahibi olan hane sayısı 74, oranı % 48,684'tür. İnternet sahibi olmayan hane sayısı ise 78 oranı % 51,316'dır. Bu düğüm hanehalkı reisi yaşlarının % 1,354'nü oluşturmaktadır. Düğüm 15'e (Node 15) göre hanehalkı büyüklüğü 1 ve 2 olan ve hanehalkı reisinin yaş aralığı 51-60, 61-70, 71 yaş ve üzeri olan hanelerden internet sahibi olan hane sayısı 341, oranı % 19,050'dir. İnternet sahibi olmayan hane sayısı ise 1 449, oranı % 80,950'dur. Bu düğüm hanehalkı reisi yaşlarının % 15,951'ni oluşturmaktadır.

Sekizinci Dallanma: HSMARTTVSAHIP (Hane Smart Tv Sahipliği)

C5.0 algoritması analizi sonucunda HHRYAS değişkenini en iyi açıklayan değişkenin HSMARTTVSAHIP değişkeni olduğu görülmüştür. Hanede smart tv olup olmama durumuna göre Resim 4.23'deki dallanma gerçekleşmiştir.

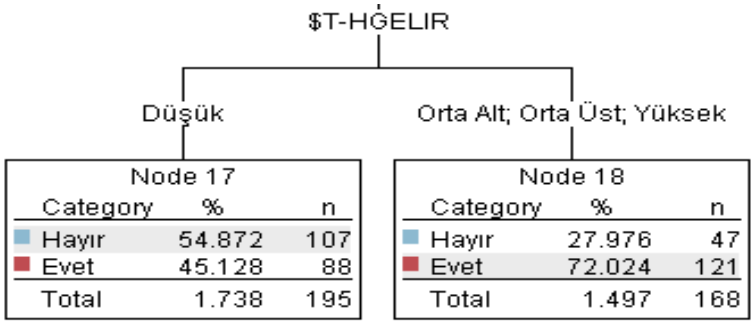


Resim 4.23. Sekizinci dallanmaya ait C5.0 analizi sonuçları

Düğüm 13'e (Node 13) göre hanehalkı reisinin yaş aralığı 41-50 olan ve smart tv sahibi olan hanelerden internet sahibi olan hane sayısı 12, oranı % 80'dir. İnternet sahibi olmayan hane sayısı ise 3, oranı % 20'dir. Bu düğüm smart tv sahipliğinin % 0,134'nü oluşturmaktadır. Düğüm 14'e (Node 14) göre hanehalkı reisinin yaş aralığı 41-50 olan ve smart tv sahibi olmayan hanelerden internet sahibi olan hane sayısı 62, oranı % 45,255'dir. İnternet sahibi olmayan hane sayısı ise 75 oranı % 54,745'dir. Bu düğüm smart tv sahipliğinin % 1,221'ni oluşturmaktadır.

Dokuzuncu Dallanma: HGELIR (Hanehalkı Aylık Net Toplam Geliri)

C5.0 algoritması analizi sonucunda HGELIR değişkenini en iyi açıklayan değişkenin HHBUYUKLUK değişkeni olduğu görülmüştür. Hanehalkı aylık net toplam geliri ile oluşturulan gelir gruplarına göre Resim 4.24'deki dallanma gerçekleşmiştir.

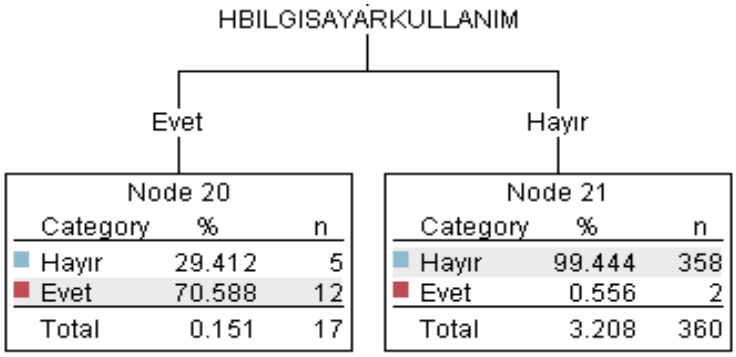


Resim 4.24. Dokuzuncu dallanmaya ait C5.0 analizi sonuçları

Düğüm 17'e (Node 17) göre hanehalkı büyüklüğü 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 olan ve düşük gelir grubundaki hanelerden internet sahibi olan hane sayısı 88, oranı % 45,128'dir. İnternet sahibi olmayan hane sayısı ise 107, oranı % 54,872'dir. Bu düğüm gelir gruplarının % 1,738'ni oluşturmaktadır. Düğüm 18'e (Node 18) göre hanehalkı büyüklüğü 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 olan ve orta alt,orta üst, yüksek gelir grubundaki hanelerden internet sahibi olan hane sayısı 121, oranı % 72,024'dir. İnternet sahibi olmayan hane sayısı ise 47 oranı % 27,976'dır. Bu düğüm gelir gruplarını % 1,497'sini oluşturmaktadır.

Onuncu Dallanma: HBILGISAYARKULLANIM (Hane Bilgisayar Kullanımı)

C5.0 algoritması analizi sonucunda HCEPTEL değişkenini en iyi açıklayan değişkenin HBILGISAYARKULLANIM değişkeni olduğu görülmüştür. Hanede bilgisayar kullanımı olup olmama durumuna göre Resim 4.25'deki gibi dallanma gerçekleşmiştir.



Resim 4.25. Onuncu dallanmaya ait C5.0 analizi sonuçları

Düğüm 20'ye (Node 20) göre cep telefonu olmayan ve bilgisayar kullanımı olan hanelerden internet sahibi olan hane sayısı 12, oranı % 70,588'dir. İnternet sahibi olmayan hane sayısı ise 5, oranı % 29,412'dir. Bu düğüm hane bilgisayar kullanımı oranının % 0.151'ni oluşturmaktadır. Düğüm 21'e (Node 21) göre cep telefonu olmayan ve bilgisayar kullanımı olmayan hanelerden internet sahibi olan hane sayısı 2, oranı % 0,556'dır, internet sahibi olmayan hane sayısı ise 358, oranı % 99,444'tür. Bu düğüm hane bilgisayar kullanım oranının % 3,208'ni oluşturmaktadır.

C5.0 algoritması ile HINTERNETSAHIP değişkeni için 8 adet *Evet*, 4 adet *Hayır* değeri olmak üzere 12 adet kural elde edilmiş olup, kurallar aşağıda belirtilmiştir.

1) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Evet) ise HINTERNETSAHIP=Evet
2) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Var) ise HINTERNETSAHIP=Evet
3) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Var) ise HINTERNETSAHIP=Evet
4) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Var) İSE HINTERNETSAHIP=Evet
5) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=1,2 ve HHRYAS=0-30 Yaş, 31-40 Yaş) ise HINTERNETSAHIP=Evet

6) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=1,2 ve HHRYAS=41-50 Yaş ve HSMARTTVSAHIP=Var) ise HINTERNETSAHIP=Evet
7) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=1,2 ve HHRYAS=41-50 Yaş VE HSMARTTVSAHIP=Yok) ise HINTERNETSAHIP=Hayır
8) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=1,2 ve HHRYAS=51-60 Yaş, 61-70 Yaş, 71 Yaş ve Üzeri) ise HINTERNETSAHIP=Hayır
9) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 VE HGELIR=Düşük) ise HINTERNETSAHIP=Hayır
10) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20 ve HGELIR= Orta Alt, Orta Üst, Yüksek) ise HINTERNETSAHIP=Evet
11) Eğer (CEPTELSAHİP=Yok ve HBILGISAYARKULLANIM=Evet) ise HINTERNETSAHIP=Evet
12) Eğer (CEPTELSAHİP=Yok ve HBILGISAYARKULLANIM=Hayır) ise HINTERNETSAHIP=Hayır

C5.0 algoritması ile oluşturulan karar ağacı incelendiğinde hanesinde cep telefonu olanların ve bilgisayar kullananların en yüksek internet sahipliği oranında (% 96,65) olduğu görülmüştür. Cep telefonuna sahip olmayan ve bilgisayar kullanmayan hanelerin ise en düşük internet sahipliği oranında (% 99,44) olduğu görülmüştür. Burada hanede cep telefonu sahipliğinin olması ve bilgisayar kullanımının olması hanenin internet hizmeti satın alma durumunu olumlu etkilemektedir. Hanede cep telefonu, tablet, dizüstü bilgisayar, smarttv gibi bilişim ekipmanları sahipliğinin olması internet hizmetinin alınmasına olumlu katkıda bulunmuştur. Aynı şekilde hanede cep telefonu, dizüstü bilgisayar, tablet sahipliği

olmaması, bilgisayar kullanımının olmaması, genç nüfus olarak tabir edilebilecek 0-25 yaş arasında bireyin olmaması internet hizmetine ihtiyaç duyulmadığını da göstermektedir. Genel olarak hanede 0-25 yaş arası birey olduğu durumda internet hizmeti sahipliğinin de olduğu görülmüştür. Hanehalkı büyüklüğünün 3 ve üzerinde olduğu ve hanenin düşük gelir grubunda olduğu durumda internet hizmeti sahipliğinin düştüğü görülmüştür. İnternet bağlantı ücretinin yüksek olmasının düşük gelir grubunda hizmete sahip olma oranını azalttığı düşünülmektedir. Orta alt, orta üst ve yüksek gelirli grupta ise internet hizmetine sahip olma oranını arttırdığı görülmüştür. Hanehalkı büyüklüğünün 3'den az, hanehalkı reisinin yaşının 0-30 ve 31-40 arasında olduğu hanelerde yine internet hizmeti sahipliğinin arttığı görülmüştür. Hanehalkı büyüklüğünün 1,2 olduğu, hanehalkı reisinin 41-50 yaş aralığında olduğu hanelerde smarttv sahipliğinin olması internet hizmeti satın alma durumunu artırmıştır. Aynı veri kümesi ile istatistiksel ve ekonometrik modeller kullnılarak yapılan çalışmalarda da bilişim ekipmanlarının sayısının internet hizmeti kullanımı oldukça arttığı görülmektedir. Yapılan bu çalışma bu sonucu destekler niteliktedir. *Digital in 2018* raporuna göre 2017 yılı verilerine göre internet trafiğinin % 62'sinin mobil cihazlardan geldiği görülmüştür. Bu sonuç karar ağacında hane cep telefonu sahipliğinin değişkeninin kök düğüm olması ve hanesinde cep telefonu olanların en yüksek internet hizmeti sahipliğinde olması ile birbirini desteklemektedir.

4.6.1. C5.0 algoritması başarı durumu

Çapraz Doğrulama (Cross Validate) değeri 10 seçilerek yapılan modelleme sonucunda elde edilen karışıklık matrisi (confusion matrix) Şekil 4.2'de verilmiştir.

Results for output field HINTERNETSAHIP

Individual Models

Comparing $C-HINTERNETSAHIP with HINTERNETSAHIP

Correct	9,577	85.34%
Wrong	1,645	14.66%
Total	11,222	

Coincidence Matrix for $C-HINTERNETSAHIP (rows show actuals)

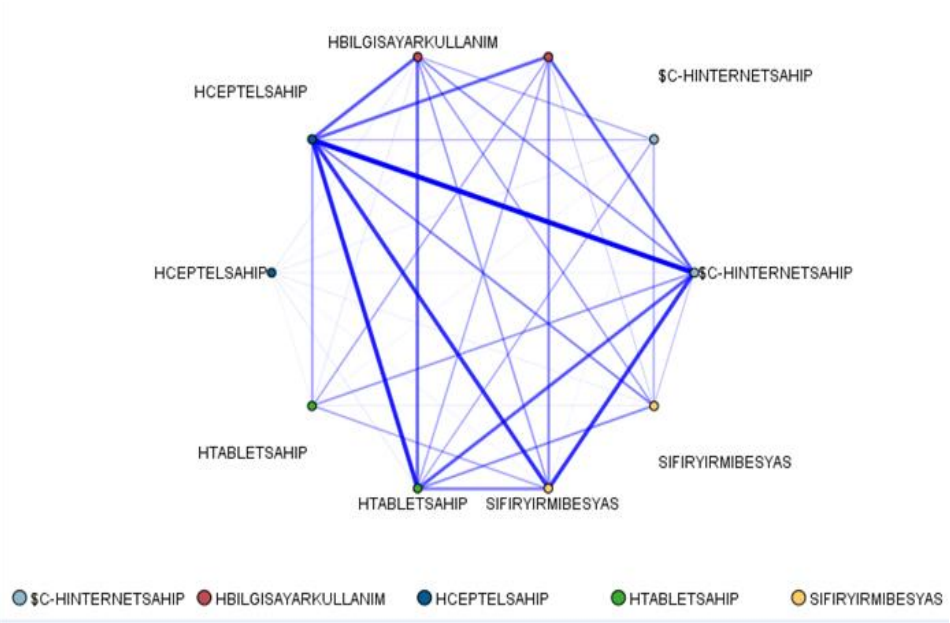
	1.000000	**2.000000**
1.000000	7,588	493
2.000000	1,152	1,989

Şekil 4.2. C5.0 algoritması karışıklık matrisi

Buna göre veri kümesindeki toplam 11 222 kayıttan 9 577 kayıt doğru sınıflandırılırken, 1 645 kayıt yanlış sınıflandırılmıştır. Bunun anlamı sınıflandırma sonucunda, internet hizmeti sahipliği durumu *Evet* olan 8 081 haneden 7 588'i doğru 493'ü yanlış tahmin edilmiş, internet hizmeti sahipliği *Hayır* olan 3 141 haneden 1 989'u doğru 1 152'si yanlış tahmin edilmiştir. Algoritmanın doğru sınıflandırma yüzdesi % 85,34, yanlış sınıflandırma yüzdesi % 14,66'dır.

4.6.2. C5.0 algoritmasındaki önemli değişkenler için web grafiği

İki ya da daha fazla kategorik veri arasındaki ilişkinin gücünü görmek için kullanılan web grafiği ile HINTERNETSAHIP değişkeni ile bu değişkeni etkileyen önemli değişkenlerden HCEPTELSAHIP, HBILGISAYARKULLANIM, HTABLETSAHIP, SIFIRYIRMIBESYAS değişkenleri web grafiği ile incelenmiştir (Şekil 4.3). Nokta ile gösterilen değişkenler arasındaki düz çizgilerle birbirlerine bağlanırlar ve çizgilerin kalın olması iki değişken arasında güçlü bir ilişki olduğunu gösterir. İnce olması zayıf bir ilişki olduğunu gösterir. Web grafiği incelendiğinde en güçlü ilişkilerin hane internet hizmeti sahipliği ile cep telefonu sahipliği arasında olduğu görülmektedir. Daha sonra en güçlü ilişkilerin sırasıyla hane cep telefonu sahipliği ile tablet sahipliği, hane cep telefonu sahipliği ile hanede 0-25 yaş arası bireyin olup olmaması, hane internet hizmeti sahipliği ile hanede 0-25 yaş bireyin olup olmaması şeklindedir.

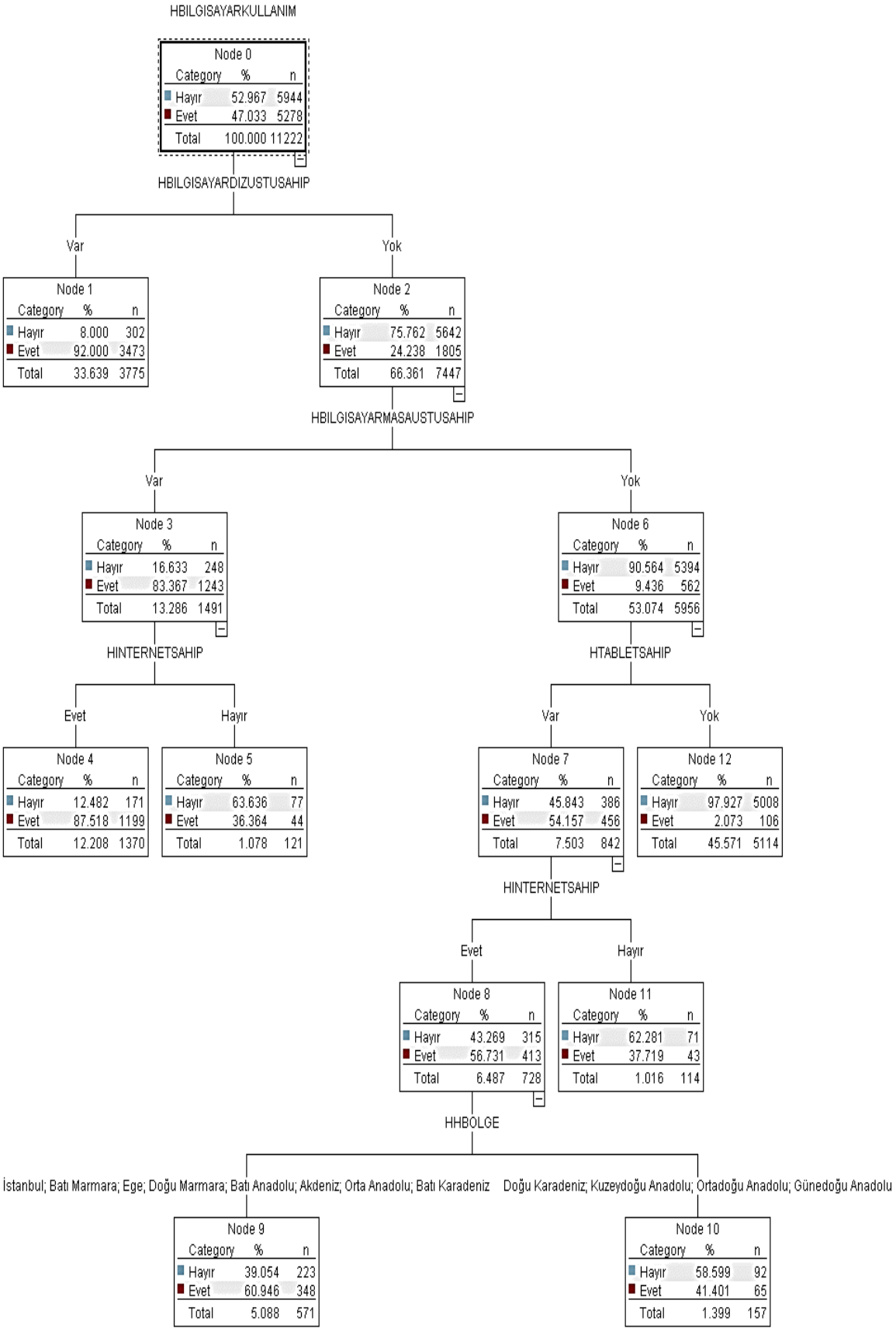


Şekil 4.3. Önemli değişkenler için web grafiği

4.6.3. Farklı hedef değişken seçimi ile C5.0 algoritmasının kullanılması

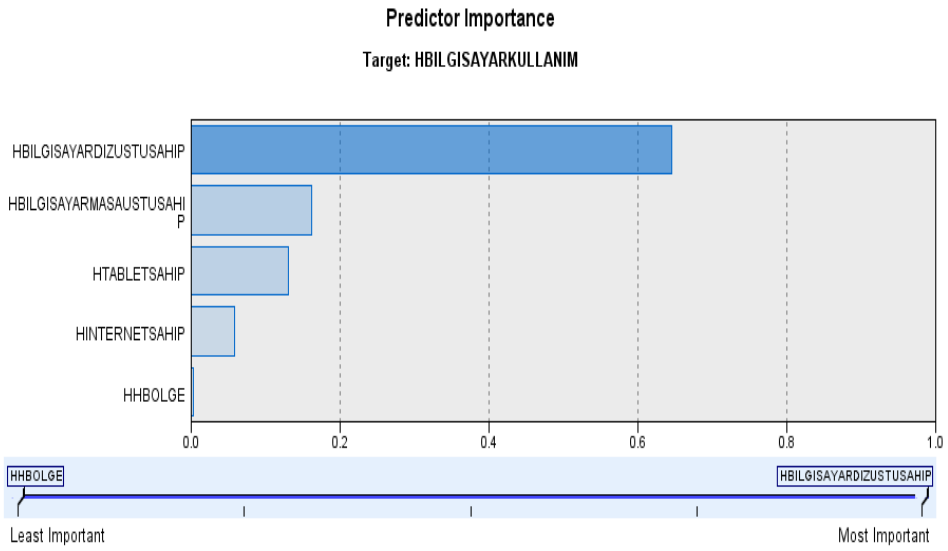
Hedef değişken olarak seçilen HINTERNETSAHIP yerine farklı değişkenler seçilerek farklı karar ağaçları ve kurallar oluşturulabilir. Örneğin aynı parametre değerleri ile HBILGISAYARKULLANIM değişkeni hedef değişken olarak seçilirse Resim 4.26'daki karar ağacı elde edilmektedir.

6 dallanma 12 düğümden oluşan bu karar ağacında HBILGISAYARKULLANIM hedef değişkenini en çok etkileyen faktörler HBILGISAYARDIZUSTUSAHIP, HBILGISAYARMASAUSTUSAHIP, HTABLETAHIP, HINTERNETSAHIP, HHBOLGE değişkenleridir.



Resim 4.26. Farklı bir hedef değişken için C5.0 algoritması karar ağacı

Değişkenlerin önem dereceleri Resim 4.27'de gösterilmiştir.



Resim 4.27. Farklı hedef değişken için değişken önem durumları

Oluşturulan bu modelin sınıflandırma doğruluk sayıları ve yüzdeleri aşağıda belirtilmiştir (Şekil 4.4). Buna göre 11 222 kayıttan doğru sınıflandırılan kayıt sayısı 10 268, sınıflandırma başarı oranı ise % 91,5'dir. Yanlış sınıflandırılan kayıt sayısı 954 olup başarısız sınıflandırma oranı ise % 8,5'dir.

Results for output field HBILGISAYARKULLANIM

Individual Models

Comparing $C-HBILGISAYARKULLANIM with HBILGISAYARKULLANIM

Correct	10,268	91.5%
Wrong	954	8.5%
Total	11,222	

Coincidence Matrix for $C-HBILGISAYARKULLANIM (rows show actuals)

	1.000000	2.000000
1.000000	5,020	258
2.000000	696	5,248

Şekil 4.4. Farklı bir hedef değişken için C5.0 algoritması karışıklık matrisi

5. SONUÇ VE ÖNERİLER

TÜİK ’in 2016 yılında yaptığı HBTKA anketi sonucu elde edilen verilerin kullanılması ile hazırlanan bu tezde hedeflenen amaç, hanehalkının internet hizmeti almasını etkileyen faktörlerin, hanehalkının karakteristik özelliklerine göre nasıl değiştiğinin karar ağaçları ile incelenmesidir. Günlük hayatın bir parçası haline gelen, her türlü bilgiye erişimi sağlayan ve sürekli büyüyen bir iletişim ağı olan internet hizmeti diğer bilişim teknolojilerinin tamamlayıcısı konumunda olduğu için sahiplik durumu hedef değişken olarak seçilmiş ve buna göre modelleme yapılmıştır. Kullanıcılar tarafından kolayca yorumlanabilmesi, kurallar kümesine dönüştürülebilmesi, diğer yöntemler verinin normal dağılmamasına veya eksik değerlere karşı hassasken, karar ağaçlarında kullanılan verinin fazla bir önişleme gerek duyulmadan kullanılabilmesi, çok büyük veri setlerine kolayca uygulanabilmesi, en çok kullanılan yöntemlerden biri olduğu için çok sayıda kaynak ve program olması gibi avantajlarından yöntem olarak karar ağaçları tercih edilmiştir. TÜİK ’ten alınan HBTKA B grubu mikro veriler içinde yer alan *Hane* ve *Fert* adındaki iki veri dosyası ortak alan olan *Bülten No* alanı adı bir hesap işlemci programı kullanılarak birleştirilmiş ve 11 222 hanenin bilgilerinden oluşan, 19 değişkeni olan yeni bir veri dosyası elde edilmiştir. Değişkenler oluşturulurken hanenin demografik yapısı, hanehalkı reisinin özellikleri ve hanenin bazı bilişim teknolojileri cihazlarına sahip olup olmadığı ile ilgili veriler kullanılmıştır. Ancak HBTKA anketindeki değişkenlerin tamamı seçilip SPSS yazılımı yardımı ile en önemli değişkenlerin seçimi yapılarak bir modelleme yapılabilir. Elde edilen yeni veri dosyası IBM SPSS Modeler 18.1 veri madenciliği yazılımı ile analiz edilmiştir. Analizde C5.0, CHAID, C&RT, QUEST olmak üzere dört karar ağacı algoritması kullanılmıştır. Öncelikle IBM SPSS Modeler 18.1 veri madenciliği programında yer alan Otomatik Sınıflandırıcı (Auto Classifier) kullanılarak belirtilen karar ağacı algoritmalarının performansları karşılaştırılmıştır. Bu algoritmalar; sınıflandırma başarı yüzdeleri, kaldıraç değerleri, Alıcı İşletim Karakteristiği eğrisi altında kalan alan değerleri bakımından karşılaştırılarak değerlendirme yapılmıştır. Karşılaştırma sonucunda en başarılı algoritmanın C5.0 ve CHAID algoritmaları olduğu görülmüştür. Karar ağacı ve kuralları C5.0 algoritması kullanılarak oluşturulmuştur. Analiz sonucunda 11 222 haneden 9 577’si doğru sınıflandırılmış ve % 85,34’lük doğru sınıflandırma başarısı elde edilmiştir. Bunun anlamı internet hizmeti sahipliği durumu *Evet* olan 8 081 haneden 7 588’i doğru, 493’ü yanlış sınıflandırılmış, internet hizmeti sahipliği *Hayır* olan 3 141 haneden 1 989’u doğru 1 152’si

yanlış sınıflandırılmıştır. Toplam doğru sınıflandırılan 9 577 hanenin, toplam hane sayısına yani 11 222'ye oranı doğru sınıflandırma oranı olan % 85,34 değerini vermektedir. Yapılan analizde 10 dallanma, 21 düğümden oluşan bir karar ağacı elde edilmiştir. Bu karar ağacının incelenmesi sonucunda hanenin internet hizmeti sahipliğini etkileyen en önemli değişkenlerin hane cep telefonu sahipliği, hane bilgisayar kullanımı, hanede 0-25 yaş arasında bireyin olup olmaması, hane tablet sahipliği, hane dizüstü bilgisayar sahipliği, hanehalkı büyüklüğü, hanehalkı reisinin yaşı, hane smart tv sahipliği, hane gelir grubu olduğu görülmüştür. Oluşturulan karar ağacı sonunda internet hizmeti sahipliğinin Evet değeri için 8 adet, Hayır değeri için 4 adet olmak üzere 12 adet kural elde edilmiştir. Aşağıda örnek kurallardan bazıları gösterilmiştir.

1) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Evet) ise HINTERNETSAHIP=Evet
2) Eğer (CEPTELSAHİP=Var ve HBILGISAYARKULLANIM=Hayır ve SIFIRYIRMIBESYAS=Yok ve HTABLETSAHIP=Yok ve HBILGISAYARDIZUSTUSAHIP=Yok ve HHBUYUKLUK=1,2 ve HHRYAS=51-60 Yaş, 61-70 Yaş, 71 Yaş ve Üzeri) ise HINTERNETSAHIP=Hayır

Hedef değişken değiştirilerek hanenin sahip olduğu diğer bilişim teknolojisi cihazlarının sahipliğini ya da kullanımını etkileyen faktörler ile ilgili de farklı çalışmalar yapılıp karar ağaçları oluşturulabilir ya da diğer veri madenciliği teknikleri kullanılabilir. Bunun için yapılan çalışmada hedef değişken olarak HBILGISAYARKULLANIM değişkeni seçilmiş, 6 dallanma, 12 düğümden oluşan bir karar ağacı elde edilmiştir. Bu sınıflandırma modeli sonucunda 11 222 haneden 10 268'si doğru sınıflandırılmış ve % 91,50'lik doğru sınıflandırma başarısı elde edilmiştir.

HBTKA anketi sonuçlarına ait veri kümesi kullanılarak yapılan çalışmalarda genellikle ekonometrik modeller kullanılmış ya da anket sonuçları ile ilgili istatistiksel analizler yapılmıştır. Bu çalışma HBTKA anketi veri kümesi ile veri madenciliği ve karar ağaçları teknikleri kullanılarak yapılmış nadir çalışmalardan biridir. Sonuç olarak sektörün ilgililerine yol gösterici ve doğru aksiyon almalarını etkileyebilecek bir çalışma yapılmıştır. Çalışma sonucunda elde edilen verilere bakılarak elektronik haberleşme hizmeti veren ve satışını yapan firmaların cep telefonu, tablet, dizüstü bilgisayar satışı gibi kampanyalar yaparak internet sahibi olmayan hanelerin sahip olmasını teşvik edici aksiyonlar almaları

beklenebilir. Özellikle karar ağacındaki en önemli değişken olan cep telefonuna sahipliğin internete hizmetine sahip olunmasını en önemli oranda etkilemesi nedeniyle sabit internetin satışlarında cep telefonu ile birlikte satış yapılması sektörün ilgilileri için bir kampanyalı satış taktiği olabilir. HBTKA veri kümesi içinde yer alan internet kullanımı ile ilgili diğer bilgiler kullanılarak internet kullanım alışkanlıkları ile ilgili bir çalışma yapılabilir. Çalışmada sadece 2016 yılına ait HBTKA veri kümesi kullanılmış olup diğer yıllara ait veri kümeleri de alınarak daha kapsamlı bir çalışma yapılabilir. Çalışmada seçilen bağımlı ve bağımsız değişkenlerin kişinin kendi kararı ve tecrübesi ile değil SPSS yazılımı aracılığı ile seçilmesi sağlanabilir. Hanede bulunan tüm bilişim ekipmanlarının sahiplik durumunu etkileyen karakteristikler bağımlı değişken değiştirilerek araştırılabilir ve modellenebilir. Daha önceki çalışmalara paralel olarak hanede bulunan bilişim ekipmanları sayısını etkileyen faktörler veri madenciliği teknikleri ile modellenebilir.

KAYNAKLAR

1. Kuzu, A. (2011). İnternet ve Aile. *Aile ve Toplum Dergisi,* 7(27), 9-31.

2. İnternet: Kemp, Simon. We are Social. URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwearesocial.com%2Fblog%2F2018%2F01%2Fglobal-digital-report-2018&date=2019-01-14, Son Erişim Tarihi: 14.01.2019.

3. İnternet: Türkay, İmdat. "Digital in 2018" Raporunda Dünyada ve Türkiye'de Durum. URL:http://www.webcitation.org/query?url=https%3A%2F%2Fvergialgi.net%2Fekonomimaliye%2Fdigital-in-2018-raporunda-dunyada-ve-turkiye dedurum%2F&date=2019-01-14, Son Erişim Tarihi: 14.01.2019.

4. Türkiye İstatistik Kurumu (2016). Hanehalkı Bilişim Teknolojileri Kullanım Araştırması; TÜİK, 21779. Türkiye İstatistik Kurumu Haber Bülteni.

5. Türkiye İstatistik Kurumu (2017). Hanehalkı Bilişim Teknolojileri Kullanım Araştırması; TÜİK, 24862 Türkiye İstatistik Kurumu Haber Bülteni.

6. Türkiye İstatistik Kurumu (2018). Hanehalkı Bilişim Teknolojileri Kullanım Araştırması; TÜİK, 27819. Türkiye İstatistik Kurumu Haber Bülteni.

7. Albayrak, A. S., Yılmaz, Ş. K. (2009). Veri madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* 14(1), 31-52.

8. Atılgan, E. (2011). *Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi ile İncelenmesi.* Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

9. Kuzey, C. (2012). *Veri Madenciliğinde Destek Vektör Makinaları ve Karar Ağaçları Yöntemlerini Kullanarak Bilgi Çalışanlarının Kurum Performansı Üzerine Etkisinin Ölçülmesi ve Bir Uygulama.* Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

10. Heshmati, A., Al-Hammadany, F. H., Bany-Mohammed, A. (2013). Analysis of Internet Usage Intensity in Iraq: An Ordered Logit Model. *Journal of Knowledge Management, Economics and Information Technology*, 3(3), 1-21.

11. Coşkun, C., Baykal, A. (2011, 2-4 Şubat). *Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması*. XIII. Akademik Bilişim konferansında sunuldu, Malatya.

12. Karadağ, M. (2014). *Karar Ağaçları ile Lojistik Regresyon Analizinin Performanslarının Simülasyon Çalışması ile Karşılaştırılması*. Yayımlanmamış Yüksek Lisans Tezi, Trakya Üniversitesi Sağlık Bilimleri Enstitüsü, Edirne.

13. Alıç, Z. H. G. (2014). *Akut Pankreatit Hastalarının Mortalite Risklerinin Karar Ağacı Yöntemi ile Belirlenmesi.* Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara.

14. Alkan, Ö., Abar, H., Karaaslan, A. (2015, 8-10 Ekim). *Hanelerde Bulunan Bilişim Ekipmanları Sayısını Etkileyen Faktörlerin Poisson Regresyon Modeliyle Araştırılması*. Atatürk Üniversitesi 2. Ulusal Yönetim Bilişim Sistemleri Kongresinde sunuldu, Erzurum.

15. Demircan, S. (2015). *Tüik Yaşam Memnuniyeti Anketleri Üzerine Veri Madenciliği Uygulaması.* Yayımlanmamış Yüksek Lisans Tezi, Erciyes Üniversitesi Fen Bilimleri Enstitisü, Kayseri.

16. Ercan, U. (2016). *Veri Madenciliği ile Hanehalkı Yağ Tüketiminin Modellenmesi.* Yayımlanmamış Doktora Tezi, Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü, Antalya.

17. Diler, S. (2016). *Veri Madenciliği Süreçleri ve Karar Ağaçları Algoritmaları ile Bir Uygulama.* Yayımlanmamış Yüksek Lisans Tezi, Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü, Van.

18. Tümsel, B. (2016). *Hanehalkı otomobil sahip olma durumunun ardışık logit modeli ile tahmini.* Yayımlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

19. Lakshmi, B., Indumathi, T., Ravi, N. (2016). A study on C. 5 decision tree classification algorithm for risk predictions during pregnancy. *Procedia Technology*, 24, 1542-1549.

20. Ceylan, Z., Gürsev, S., Bulkan, S. (2017). İki Aşamalı Kümeleme Analizi ile Bireysel Emeklilik Sektöründe Müşteri Profilinin Değerlendirilmesi. *Bilişim Teknolojileri Dergisi,* 10(4), 475-485.

21. Selim, S., Balyaner, İ. (2017). Türkiye'de Hanehalkının Sahip Olduğu Bilişim Teknolojileri Ürünleri Sayısını Belirleyen Faktörlerin Araştırılması: Bir Sayma Veri Modeli. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9*(22), 428-454.

22. Ahmed, F., Asif, R., Hina, S., Muzammil, M. (2017). Financial Market Prediction Using Google Trends. *International Journal of Advanced Computer Science and Applications,* 8(7), 388-391.

23. Topîrceanu, A., Grosseck, G. (2017). Decision Tree Learning Used for The Classification of Student Archetypes in Online Courses. *Procedia Computer Science,* 112, 51-60.

24. Güzel, S. (2018). *Veri Madenciliğinde Sınıflandırma Algoritmaları Kullanılarak Hepatit Hastalığının Tespiti.* Yayımlanmamış Yüksek Lisans Tezi, Sütçü İmam Üniversitesi Fen Bilimleri Enstitüsü, Kahramanmaraş.

25. Börekçi, N. (2018). *Türkiye'de Hanehalkı Telekomünikasyon Harcamalarını Etkileyen Faktörlerin Ekonometrik Analizi*. Yayımlanmamış Yüksek Lisans Tezi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü, Erzurum.

26. Jiménez, R., Anupol, J., Cajal, B., Gervilla, E. (2018). Data Mining Techniques for Drug Use Research. *Addictive behaviors reports,* 8, 128-135.

27. Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları* (3. Baskı). İstanbul: Papatya Yayınları, 11-12, 67,73-74, 81, 86, 139-140, 155-156.

28. Dunham, M. H. (2006). *Data mining: Introductory and Advanced Topics*. India: Pearson Education, 3, 10, 97.

29. Koyuncugil, A., Özgülbaş, N. (2009). Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. *Bilişim Teknolojileri Dergisi,* 2(2), 21-32.

30. Han, J., Pei, J., Kamber, M. (2011). *Data Mining: Concepts and Techniques* (Third Edition). San Francisco: Morgan Kaufman, 8, 328, 365-366, 370-371, 377.

31. Şeker, Ş. E. (2013). *İş Zekası ve Veri Madenciliği* (1. Baskı). İstanbul: Cinius Yayınları, 25-28.

32. Baykal, A. (2006). Veri Madenciliği ve Uygulama Alanları. *DÜ Ziya Gökalp Eğitim Fakültesi Dergisi,* 7, 95-107.

33. Kantardzic, M. (2011). *Data mining: Concepts, Models, Methods, and Algorithms* (Second Edition). New Jersey: John Wiley and Sons, 6,139.

34. Berry, M., Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (Second Edition). Indiana: Wiley Publishing, 7.

35. İnternet: Gartner IT Glossary Data Mining. Gartner Group. URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwww.gartner.com%2Fitglossary%2Fdata-mining.&date=2019-01-14, Son Erişim Tarihi: 14.01.2019.

36. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2005). *Data Mining: Practical machine learning tools and techniques* (Second Edition). USA: Morgan Kaufmann, 8.

37. Akpınar, H. (2014). *Data: Veri Madenciliği Veri Analizi* (Birinci Baskı). İstanbul: Papatya Yayıncılık Eğitim, 50, 70, 72, 163, 186, 204-206, 218-223, 286, 333.

38. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data mining to Knowledge Discovery in Databases. *AI magazine*, 17(3), 37-54.

39. Gupta, S., Kumar, D., Sharma, A. (2011). Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. *Indian Journal of Computer Science and Engineering (IJCSE),* 2(2), 188-195.

40. Özdemir, A., Aslay, F. Y., Çam, H. (2009). Veri Tabanında Bilgi Keşfi Süreci: Gümüşhane Devlet Hastanesi Uygulaması *Sosyal Ekonomik Araştırmalar Dergisi,* 10(20), 347-366.

41. Wirth, R., Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining.* Paper presented at the Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 29-39.

42. Larose, D. T., Larose, C. D. (2015). *Data Mining and Predictive Analytics* (Second Edition). New Jersey: John Wiley and Sons, 71-73.

43. Kılıçalan, M. B., (2018). *Hanehalkı İşgücü Araştırma Verileri ile Veri Madenciliği Yöntemlerinin Uygulanması ve Modellerin Karşılaştırılması.* Yayımlanmamış Yüksek Lisans Tezi, Hacettepe Fen Bilimleri Enstitüsü, Ankara.

44. Savaş, S., Topaloğlu, N., Yılmaz, M. (2012). Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 1-23.

45. Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Dergisi,* 3, 65-82.

46. Bramer, M. (2007). *Principles of Data Mining* (Vol. 180). UK: Springer, 3-4.

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

48. Yılmaz, Ş. K., Patır, S. (2011). Kümeleme Analizi ve Pazarlamada Kullanımı. *Akademik Yaklaşımlar Dergisi,* 2(1), 91-112.

49. Akın, Y. K. (2008). *Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi.* Yayımlanmamış Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

50. Özdemir, A., Orçanlı, K. (2012). İki Aşamalı Kümeleme Algoritması ile Pazar Bölümlemesi, Müşteri Profillerinin Belirlenmesi ve Niş Pazarların Tespiti. *Uşak Üniversitesi Sosyal Bilimler Dergisi,* 5(3), 1-26.

51. Özcan, C. (2014). *Veri Madenciliğinin Güvenlik Uygulama Alanları ve Veri Madenciliği ile Sahtekarlık Analizi.* Yayımlanmamış Yüksek Lisans Tezi, Bilgi Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

52. Gemici, B. (2012). *Veri Madenciliği ve Bir Uygulaması.* Yayımlanmamış Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal Bilimleri Enstitüsü, İzmir.

53. Agrawal, R., Srikant, R. (1994). *Fast Algorithms for Mining Association Rules.* Paper presented at the Proc. 20th int. conf. very large data bases, VLDB, San Jose, CA.

54. İnternet: Sayad, Saed. Decision Tree - Overfitting. URL: http://www.webcitation.org/query?url=http%3A%2F%2Fwww.saedsayad.com%2Fdecision_tree_overfitting.htm&date=2019-01-14, Son Erişim Tarihi: 14.01.2019.

55. İnternet: Uzun, Erdinç. Decision Tree (Karar Ağacı): ID3 Algoritması – Classification (Sınıflama). URL:http://www.webcitation.org/query?url=https%3A%2F%2Fwww.e-adys.com%2Fmakine_ogrenmesi%2Fdecision-tree-karar-agaci-id3-algoritmasi classification-siniflama%2F&date=2019-01-14, Son Erişim Tarihi: 14.01.2019.

56. Rokach, L., Maimon, O. (2005). Decision trees *Data Mining and Knowledge Discovery Handbook* (pp. 165-192): Springer, Boston, MA.

57. Akça, F. (2014). *Veri Madenciliği ile Fen Fakülteleri Öğrenci Profillerinin İncelenmesi: Gazi Üniversitesi Örneği.* Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

58. Çalış, A., Kayapınar, S., Çetinyokuş, T. (2014). Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama. *Journal of Industrial Engineering (Turkish Chamber of Mechanical Engineers),* 25(3-4), 2-19.

59. Pang, S., Gong, J. (2009). C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Systems Engineering-Theory and Practice,* 29(12), 94-104.

60. Akküçük, U. (2011). *Veri Madenciliği: Kümeleme ve Sınıflama Algoritmaları* (1. Baskı). İstanbul: Yalın Yayıncılık, 74, 83.

61. Adak, M. F., Yurtay, N. (2013). Gini Algoritmasını Kullanarak Karar Ağacı Oluşturmayı Sağlayan Bir Yazılımın Geliştirilmesi. *Bilişim Teknolojileri Dergisi,* 6(3), 1-6.

62. Doğan, N., Özdamar, K. (2003). CHAID Analizi ve Aile Planlaması ile İlgili Bir Uygulama. *Turkiye Klinikleri Journal of Medical Sciences,* 23(5), 392-397.

63. Kayri, M., Boysan, M. (2007). Araştırmalarda CHAID Analizinin Kullanımı ve Baş Etme Stratejileri ile İlgili Bir Uygulama. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi,* 40(2), 133-149.

64. Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied statistics*, 29(2), 119-127.

65. Dolgun, M. Ö. (2014). *Veri Madenciliği Sınıflama Yöntemlerinin Başarılarının Bağımlı Değişken Prevelansı Örneklem Büyüklüğü ve Bağımsız Değişkenler Arası İlişki Yapısına Göre Karşılaştırılması.* Yayımlanmamış Doktora Tezi, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

66. Kartal, E. (2015). *Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama.* Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

67. Aydın, S. (2007). *Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama.* Yayımlanmamış Doktora Tezi, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.

68. İnternet: C5.0 Node Model Options. IBM Knowledge Center. 2019-01-18. URL: http://www.webcitation.org/query?url=https%3A%2F%2Fwww.ibm.com%2Fsupport%2Fknowledgecenter%2Fen%2FSS3RA7_17.0.0%2Fclementine%2Fc50_modeltab.html&date=2019-01-18, Son Erişim Tarihi: 18.01.2018.

EKLER

EK-1. IBM SPSS Modeler 18.1 programı karar ağaçları analizi

Auto Classifier
Fields
Target
HINTERNETSAHIP
Inputs
HCEPTELSAHIP
HBILGISAYARKULLANIM
SIFIRYIRMIBESYAS
$T-HGELIR
HBILGISAYARDIZUSTUSAHIP
HSMARTTVSAHIP
HHRMESLEK
HHROKUMAYAZMA
HHBUYUKLUK
HHREGITIM
HHRCINSIYET
HHRYAS
HBILGISAYARMASAUSTUSAHIP
HHBOLGE
HTABLETSAHIP
HOYUNKONSOLSAHIP
HHRCALISMA
HSABITTELSAHIP
Build Settings

Use partitioned data: true
Partition: Training ve Test

Training Summary

Application: IBM® SPSS® Modeler 18.1
Planned models : 4
Completed models : 4
Models discarded from final results : 0
Models which failed to build or score : 0
Models not completed due to interruption : 0
Elapsed time for model build: 0 hours, 0 mins, 17 secs

Model Details
C5.0 1
Analysis
Tree depth: 12
Fields
Target
HINTERNETSAHIP

EK-1. (devam) IBM SPSS Modeler 18.1 programı karar ağaçları analizi

Inputs
HCEPTELSAHIP
HBILGISAYARKULLANIM
SIFIRYIRMIBESYAS
$T-HGELIR
HBILGISAYARDIZUSTUSAHIP
HSMARTTVSAHIP
HHRMESLEK
HHROKUMAYAZMA
HHBUYUKLUK
HHREGITIM
HHRCINSIYET
HHRYAS
HBILGISAYARMASAUSTUSAHIP
HHBOLGE
HTABLETSAHIP
HOYUNKONSOLSAHIP
HHRCALISMA

Build Settings

Use partitioned data: true
Partition: Training ve Test
Calculate predictor importance: false
Calculate raw propensity scores: false
Calculate adjusted propensity scores: false
Use weight: false
Output type: Decision tree
Group symbolics: false
Use boosting: false
Cross-validate: false
Mode: Simple
Favor: Accuracy
Expected noise (%): 0
Use misclassification costs: false

Training Summary

Algorithm: C5.0
Model type: Classification
Application: IBM® SPSS® Modeler Subscription 18.1
Elapsed time for model build: 0 hours, 0 mins, 5 secs

CHAID 1

Analysis

Tree depth: 5

Fields

Target

EK-1. (devam) IBM SPSS Modeler 18.1 programı karar ağaçları analizi

HINTERNETSAHIP
Inputs
HHRYAS
HHREGITIM
HHBUYUKLUK
SIFIRYIRMIBESYAS
HHBOLGE
HBILGISAYARMASAUSTUSAHIP
HBILGISAYARDIZUSTUSAHIP
HTABLETSAHIP
HSABITTELSAHIP
HSMARTTVSAHIP
HBILGISAYARKULLANIM
$T-HGELIR
Build Settings
Use partitioned data: true
Partition: Training ve Test
Calculate predictor importance: true
Calculate raw propensity scores: false
Calculate adjusted propensity scores: false
Continue training existing model: false
Use frequency: false
Use weight: false
Method: CHAID
Levels below root: 5
Alpha for Splitting: 0.05
Alpha for Merging: 0.05
Epsilon For Convergence: 0.001
Maximum iterations for convergence: 100
Use Bonferroni adjustment: true
Allow splitting of merged categories: false
Chi-Square method: Pearson
Stopping criteria: Use percentage
Minimum records in parent branch (%): 2
Minimum records in child branch (%): 1
Use misclassification costs: false

Training Summary

Algorithm: CHAID
Model type: Classification
Application: IBM® SPSS® Modeler Subscription 18.1
Elapsed time for model build: 0 hours, 0 mins, 5 secs
Quest 1
Analysis
Tree depth: 5
Fields
Target

EK-1. (devam) IBM SPSS Modeler 18.1 programı karar ağaçları analizi

HINTERNETSAHIP
Inputs
HHRYAS
HHROKUMAYAZMA
HHREGITIM
HHRCALISMA
HHRMESLEK
HHBUYUKLUK
SIFIRYIRMIBESYAS
HHBOLGE
HBILGISAYARMASAUSTUSAHIP
HBILGISAYARDIZUSTUSAHIP
HTABLETSAHIP
HCEPTELSAHIP
HOYUNKONSOLSAHIP
HSABITTELSAHIP
HSMARTTVSAHIP
HBILGISAYARKULLANIM
$T-HGELIR
Build Settings
Use partitioned data: true
Partition: Training ve Test
Calculate predictor importance: true
Calculate raw propensity scores: false
Calculate adjusted propensity scores: false
Use frequency: false
Levels below root: 5
Mode: Expert
Maximum surrogates: 5
Alpha for Splitting: 0.05
Stopping criteria: Use percentage
Minimum records in parent branch (%): 2
Minimum records in child branch (%):
Prune tree: true
Use standard error rule: false
Prior probabilities: Based on training data
Adjust priors using misclassification costs: false
Use misclassification costs: false
Training Summary
Algorithm: QUEST
Model type: Classification
Application: IBM® SPSS® Modeler Subscription 18.1
Elapsed time for model build: 0 hours, 0 mins, 5 secs
C&R Tree 1
Analysis
Tree depth: 4
Fields

EK-1. (devam) IBM SPSS Modeler 18.1 programı karar ağaçları analizi

- Target
  - HINTERNETSAHIP
- Inputs
  - HHRYAS
  - HHREGITIM
  - HHRCALISMA
  - HHRMESLEK
  - HHBUYUKLUK
  - SIFIRYIRMIBESYAS
  - HHBOLGE
  - HBILGISAYARMASAUSTUSAHIP
  - HBILGISAYARDIZUSTUSAHIP
  - HTABLETSAHIP
  - HOYUNKONSOLSAHIP
  - HBILGISAYARKULLANIM
  - $T-HGELIR

Build Settings
- Use partitioned data: true
- Partition: Training ve Test
- Calculate predictor importance: true
- Calculate raw propensity scores: false
- Calculate adjusted propensity scores: false
- Use frequency: false
- Use weight: false
- Levels below root: 5
- Mode: Expert
- Maximum surrogates: 5
- Minimum change in impurity: 0.0
- Impurity measure for categorical targets: Gini
- Stopping criteria: Use percentage
- Minimum records in parent branch (%): 2
- Minimum records in child branch (%): 1
- Prune tree: true
- Use standard error rule: false
- Prior probabilities: Based on training data
- Adjust priors using misclassification costs: false
- Use misclassification costs: false

Training Summary
- Algorithm: C&R Tree
- Model type: Classification
- Application: IBM® SPSS® Modeler Subscription 18.1
- Elapsed time for model build: 0 hours, 0 mins, 5 secs

EK-2. IBM SPSS Modeler 18.1 C5.0 algoritması karar ağacı analizi

Analysis

- Tree depth: 8
- Cross Validation
- Mean: 85.0
- Standard Error: 0.3

Fields

- Target
  - HINTERNETSAHIP
- Inputs
  - HCEPTELSAHIP
  - HBILGISAYARKULLANIM
  - SIFIRYIRMIBESYAS
  - HTABLETSAHIP
  - HBILGISAYARDIZUSTUSAHIP
  - HHBUYUKLUK
  - HHRYAS
  - HSMARTTVSAHIP
  - $T-HGELIR

Build Settings

- Use partitioned data: false
- Calculate predictor importance: true
- Calculate raw propensity scores: true
- Calculate adjusted propensity scores: false
- Use weight: false
- Output type: Decision tree
- Group symbolics: false
- Use boosting: false
- Cross-validate: true
- Number of folds: 10
- Mode: Expert
- Pruning severity: 90
- Minimum records per child branch: 10
- Winnow attributes: false
- Use global pruning: true
- Use misclassification costs: true

Training Summary

- Algorithm: C5.0
- Model type: Classification
- Application: IBM® SPSS® Modeler Subscription 18.1
- Elapsed time for model build: 0 hours, 0 mins, 1 secs

EK-3. TÜİK B Grubu Mikro Veri Kullanımı İzin Formu

T.C.
TÜRKİYE İSTATİSTİK KURUMU BAŞKANLIĞI
Bilgi Dağıtım Ve İletişim Daire Başkanlığı

Sayı : 27964695-622.03-E.111 02.01.2018
Konu : Hanehalkı Bilişim Teknolojileri Kullanım Araştırması

Sayın Mahmut COŞKUN

İlgi : 20.12.2017 tarihli yazınız.

Gazi Üniversitesi, Bilişim Enstitüsü adına, "Hanehalkı Bilişim Teknolojileri Kullanımı Veri Madenciliği Teknikleri İle Analizi" konulu araştırma projesi için talep ettiğiniz B Grubu Hanehalkı Bilişim Teknolojileri Kullanımı Araştırması 2016 yılı mikro verileri Değerlendirme Kurulunca uygun bulunmuştur. Ekte verilen taahhütnamenin imzalanması durumunda talep ettiğiniz mikro veriler tarafınıza elektronik ortamda iletilecektir.

Mikro Veri kullanılarak yapılan tüm araştırmalara, Kurum web sitesi ana sayfası kütüphane linki aracılığıyla veya http://kutuphane.tuik.gov.tr/yordambt/yordam.php linkinde arama bölümünde "TÜİK Mikro Verileri ile Hazırlanan Çalışmalar" seçilerek erişim sağlayabilirsiniz.

Bilgilerinizi rica ederim.

e-imzalıdır
Abdullah ÖZDEMİR
Başkan a.
Daire Başkan V.

Ek : Taahhütname (1 Sayfa)

Not: 5070 sayılı elektronik imza kanununun 5.maddesi gereği bu belge elektronik imza ile imzalanmıştır.

Devlet Mahallesi Necatibey Caddesi No:114 Çankaya/Ankara
Tel: (0312) 410 02 22 Faks: (0312) 417 04 32
KEP: tuikbilgidagitimveiletisimdairebaskanligi@hs01.kep.tr

Bilgi için:Ümit ÖĞÜT
Uzman
Telefon No:(312) 410 02 47



ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, Adı : COŞKUN, Mahmut

Uyruğu : T.C.

Doğum tarihi ve yeri : 25.02.1984, Kayseri

Medeni hali : Evli

Telefon : 0 (506) 692 86 57

E-posta : mahmutcoskun84@gmail.com

Eğitim Derecesi	Okul/Program	Mezuniyet Yılı
Yüksek Lisans	Gazi Üniversitesi / Bilişim Enstitüsü Yönetim Bilişim Sistemleri Anabilim Dalı	Devam ediyor
Lisans	Erciyes Üniversitesi / Elektrik-Elektronik Mühendisliği	2008
Lise	Kayseri Sema Yazar Anadolu Lisesi	2002

İş Deneyimi

Yıl	Yer	Görev
2008-Halen	Türk Telekom Ankara Bölge Müdürlüğü	Ekip Lideri

Yabancı Dil

İngilizce

Hobiler

Robotik, spor, kitap okuma, gezmek



GAZİLİ OLMAK AYRICALIKTIR.