T.C. İSTANBUL TİCARET
ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

SINIFLANDIRMA AĞACI ANALİZİ İLE BULUT DEPO KULLANIMI YAPAN BİREYLERİN PROFİLLERİNİN İNCELENMESİ

GÜNER GÖZDE TEKSİN

Danışman
PROF. DR. MÜNEVVER TURANLI

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI
İSTANBUL - 2018

KABUL VE ONAY SAYFASI

GÜNER GÖZDE TEKSİN tarafından hazırlanan **"Sınıflandırma Ağacı Analizi İle Bulut Depo Kullanımı Yapan Bireylerin Profillerinin İncelenmesi"** adlı tez çalışması 11/07/2018 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **İstatistik Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	**Prof. Dr. Münevver TURANLI** İstanbul Ticaret Üniversitesi	
Jüri Üyesi	**Prof. Dr. Ünal Halit ÖZDEN** İstanbul Ticaret Üniversitesi	
Jüri Üyesi	**Prof. Dr. Şahamet BÜLBÜL** Marmara Üniversitesi	

Onay Tarihi : 23./7./2018

Prof. Dr. Necip ŞİMŞEK
Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

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

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

beyan ederim.

23.07.2018

GÜNER GÖZDE TEKSİN

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER i
ÖZET iii
ABSTRACT iv
TEŞEKKÜR v
ŞEKİLLER DİZİNİ vi
ÇİZELGELER DİZİNİ vii
SİMGELER VE KISALTMALAR DİZİNİ viii

1. GİRİŞ 1

2. LİTERATÜR ÖZETİ 4

3. BÜYÜK VERİNİN TANIMI VE ÖZELLİKLERİ 7

4. VERİ MADENCİLİĞİ 11
4.1. Veri Madenciliği Kavramı 11
4.2. Veri Madenciliğinin Uygulama Alanları 12
4.3. Veri Madenciliği Süreci 15
4.4. Veri Madenciliğinde Makine Öğrenimi Kavramı 16
4.5. Makine Öğreniminde Kullanılan Modeller 19
4.5.1. Tahmin Edici Modeller 19
4.5.2. Tanımlayıcı Modeller 20

5. SINIFLANDIRMA VE REGRESYON MODELLERİ 22
5.1. Makine Öğreniminde Karar Ağaçları 22
5.1.1. Karar Ağacı Algoritmaları 24
5.1.1.1. ID3 Algoritması 24
5.1.1.2. C4.5 Algoritması 25
5.1.1.3. C5.0 Algoritması 26
5.1.1.4. CART Algoritması 28
5.1.1.5. CHAID Algoritması 31
5.1.1.6. SLIQ Algoritması 33
5.1.1.7. SPRINT Algoritması 33
5.1.1.8. MARS Algoritması 34
5.1.1.9. QUEST Algoritması 35
5.2. Karar Ağacı Algoritmalarının Avantaj ve Dezavantajları 36

6. SINIFLANDIRMA VE REGRESYON AĞAÇLARI 37
6.1. Sınıflandırma Ağaçları 37
6.1.1. Gini Ayırma Kriteri 38
6.1.2. Twoing Ayırma Kriteri 39
6.1.3. Torbalama Ayırma Kriteri 40
6.1.4. Gini ve Twoing Ayırma Kriterlerinin Karşılaştırılması 41
6.2. Regresyon Ağaçları 41
6.3. Sınıflandırma Ağacı ve Regresyon Ağacının Karşılaştırılması 42

7. UYGULAMA 45
7.1. Bulut Bilişim ve Bulut Depo Kullanımı 45
7.2. 2015-2016-2017 Hanehalkı Bilişim Teknolojileri Anket Sonuçları 46
7.3. CART Algoritması ile Oluşturulan Sınıflandırma Ağaçları ve Bulgular 69
7.4. Sınıflandırma Ağaçlarından Elde Edilen Sınıflandırma Kuralları 85
8. SONUÇ 86
KAYNAKLAR 88
ÖZGEÇMİŞ 94

ÖZET

Yüksek Lisans Tezi

SINIFLANDIRMA AĞACI ANALİZİ İLE BULUT DEPO KULLANIMI YAPAN BİREYLERİN PROFİLLERİNİN İNCELENMESİ

Güner Gözde TEKSİN

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Prof. Dr. Münevver TURANLI

2018, 94 sayfa

Büyük veri, gelişen ve değişen teknoloji ile birlikte günümüzde sıkça kullanılan bir kavram haline gelmiştir. Büyük verilerin işlenebilmesi ve bilgiye dönüştürülebilmesi için, veri madenciliği ve makine öğrenimi gibi alanlar verinin analiz edilebilmesi için yöntemler sunmakta ve algoritmalar geliştirmektedir. Veri madenciliği ve makine öğrenimi içerisinde, sıkça kullanılan yöntem karar ağaçlarıdır. Karar ağaçları, parametrik olmayan yöntemdir ve bu nedenle istatistiksel anlamda kısa sürede analiz olanağı sağlamaktadır. Karar ağacı algoritmaları içerisinde en yaygın kullanıma sahip algoritma, sınıflandırma ve regresyon ağacı (CART) algoritmasıdır. Bağımlı değişkenin kategorik yapıda olması durumunda sınıflandırma ağacı, bağımlı değişkenin sürekli olması durumunda ise regresyon ağacı oluşmaktadır.

Bu çalışmada, uygulama olarak bulut depo kullanımı yapan bireylerin kişisel olarak internette yaptıkları faaliyetleri, demografik özellikleri ve yazılım faaliyetleri incelenmiştir. Çalışmada kullanılan bağımlı değişken kategorik yapıda olduğu için, farklı eğitim verileriyle sınıflandırma ağaçları oluşturulmuştur. Dolayısıyla, ağaçlar arasındaki farklılıklar ile hatalı sınıflandırma oranından yararlanarak optimum ağaca karar verilmiştir.

Anahtar Kelimeler: Büyük veri, veri madenciliği, makine öğrenimi, karar ağacı algoritmaları, sınıflandırma ve regresyon ağacı

ABSTRACT

M.Sc. Thesis

ANALYSIS OF INDIVIDUALS' PROFILES USING CLOUD STORAGE WITH CLASSIFICATION TREE ANALYSIS

Guner Gozde TEKSIN

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

Supervisor: Prof. Dr. Munevver TURANLI
2018, 94 pages

Big data, together with developing and changing technology nowadays has become a commonly used concept. Machine learning and Data mining provides methods and algorithms for analyzing data, so that big data can be processed and information can be transformed. In data mining and machine learning, decision trees are frequently used. Decision trees are non-parametric methods and therefore provide statistical analysis in a short time. Among the decision tree algorithms, classification and regression tree (CART) algorithm is the most widely used. If the dependent variable is a categorical structure, it is a classification tree. If the dependent variable is continuous, a regression tree is formed.

This study focuses on cloud computing and cloud storage, which has become a cost-reducing concept in the analysis of large data sets. Individuals who use cloud storage over the Internet; their personal activities on the Internet, their demographic characteristics and software activities has been examined. Since the dependent variable used in the study is categorical, classification trees have been created with different training data. Therefore, optimum tree size was determined by taking advantage of the differences between the trees and the faulty classification ratio.

Keywords: Big data, data mining, machine learning, decision tree algorithms, classification and regression tree

TEŞEKKÜR

Lisans ve yüksek lisans öğrenimim boyunca kendisine saygı duyduğum, her çalışmamda bana yol gösteren, her zaman olumlu enerjisiyle mesleğimi daha çok sevmemi sağlayan, anne sevgisi ve şefkatini bana okulda da yaşatan, güler yüzüyle içimi aydınlatan, her sorunumda bana fazlasıyla destek olan, tez dönemim boyunca gece gündüz demeden sorularıma yanıt veren tez danışmanım ve en büyük destekçim, kendime örnek aldığım, geleceğime yön veren, çok sevdiğim kıymetli hocam Sayın Prof. Dr. Münevver TURANLI'ya en içten minnetimi ve teşekkürlerimi sunarım.

Üniversite hayatım boyunca bilgi ve birikimlerini paylaşmaktan çekinmeyen, babacan tavrı nedeniyle güven duyduğum, anlayışlı ve destekçi yapısıyla öğrencinin yanında olduğunu hissettiren, güler yüzünü asla esirgemeyen ve öğrenci dostu, çok değerli hocam Sayın Prof. Dr. Ünal Halit ÖZDEN'e en içten teşekkürü borç bilirim.

Lisans öğrenimimin ilk dersinde bende unutulmaz olumlu izler bırakan, değerli görüşleri ile yol gösterici davranış sergileyen, disiplinli yapısını benimsediğim, araştırmacı kimliği sayesinde kendime farklılıklar kattığım, değerli vaktini hiçbir zaman benden esirgemeyen çok değerli hocam Sayın Doç. Dr. Özlem Deniz BAŞAR'a en içten teşekkürlerimi sunarım.

Üniversite hayatım boyunca gerek bilgi ve tecrübelerini, gerekse değerli zamanını esirgemeyen, çalışmalarımda faydalı olabilmek için elinden geleni yapan, samimi ve içten tavırlarıyla her anlamda destek olan, anlayışı ve güler yüzünü benden esirgemeyen saygıdeğer hocam Sayın Yrd. Doç. Dr. Seda Bağdatlı KALKAN'a içten teşekkürlerimi sunarım.

Üniversite yıllarımda güler yüzünü hiç esirgemeyen, dertleşebildiğim, iyi niyetine inandığım, fikir ve düşüncelerine güvendiğim, değerli vaktini hep faydalı olmak için harcayan çok değerli hocam Prof. Dr. Necip ŞİMŞEK'e en içten teşekkürlerimi sunarım.

Tez savunması sırasında yapıcı eleştirileri ve fikirleriyle değerli katkılar sunan, kısa zamanda çok şey öğrendiğim saygıdeğer hocam Prof. Dr. Şahamet BÜLBÜL'e teşekkürü borç bilirim.

Okula başladığım ilk günden beri benden desteklerini asla esirgemeyen, yaşadığım tüm sıkıntılarda her zaman yanımda olan, yoğun sınav dönemlerimde hep sabırlı olan, sevgisini ve ilgisini göstermekten asla çekinmeyen, güler yüzüyle en zorlu dönemleri bile aşmamı sağlayan canım annem başta olmak üzere, tecrübeleriyle geleceğime ışık tutan sevgili babama, küçük yaşımdan beri bana yoldaş olan kardeşime, beni her zaman olumlu yönde motive eden, desteğini ve ablalığını benden hiç esirgemeyen Güner Güleç ELİK'e ve desteklerini her zaman yanımda hissettiğim dostlarıma teşekkürlerimi sunarım.

Güner Gözde TEKSİN

İSTANBUL, 2018

ŞEKİLLER DİZİNİ

Şekil 1. Karar Ağacı Yapısı 22
Şekil 2. Ağacın Kökü, Dalları ve Yaprakları 37
Şekil 3. Cinsiyet Dağılımları 46
Şekil 4. Yaş Dağılımları 47
Şekil 5. Okuma Yazma Dağılımları 47
Şekil 6. Eğitim Durumu Dağılımları 48
Şekil 7. Meslek Dağılımları 49
Şekil 8. Haneden Herhangi Birinin Bilgisayar Kullanma Durumu 50
Şekil 9. Hanede İnternet Erişim Durumu 50
Şekil 10. Hanede İnternet Kullanım Durumu 51
Şekil 11. Hanede Bulunan Bilişim Ekipmanları 51
Şekil 12. Evde Kullanılan İnternet Bağlantı Türleri 52
Şekil 13. Evden İnternete Bağlanmama Nedenleri 53
Şekil 14. Hanelerin Aylık Net Ortalama Geliri 54
Şekil 15. Bilgisayar Kullanım Sıklığı 55
Şekil 16. İnternet Kullanım Sıklığı 55
Şekil 17. Kişisel Amaçlarla İnternette Yapılan Faaliyetler (Son Üç Ay) 56
Şekil 18. İnternet Üzerinden Bulut Depo Kullanımı 57
Şekil 19. Kamu Kurum/Kuruluşları İle İletişimde Bulunma 58
Şekil 20. İnternet Üzerinden Yapılan Öğrenim Faaliyetleri 58
Şekil 21. Kamu Kurum/Kuruluşlarının Web Siteleri Üzerinden Form Göndermeme Nedenleri 59
Şekil 22. İnternet Üzerinden Alınan Mal veya Hizmet Türleri 60
Şekil 23. İnternet Üzerinden İndirilen ya da Satın Alınan Mal veya Hizmet Türleri 61
Şekil 24. İnternet Üzerinden Mal/Hizmet Alım/Sipariş Sırasında Karşılaşılan Sorunlar 62
Şekil 25. İnternet Üzerinden Gerçekleştirilen Finansal İşlemler 63
Şekil 26. İnternet Üzerinden Alışveriş Yapmama Nedenleri 64
Şekil 27. Bilgisayar ya da Mobil Cihazla Yapılan İşlemler 65
Şekil 28. Yazılım İle İlgili Yapılan Faaliyetler 66
Şekil 29. İnternette Paylaşılan Kişisel Bilgiler 67
Şekil 30. Çerezler Hakkında Bilgi Sahibi Olma Durumu 68
Şekil 31. İnternet Tarayıcı Ayarlarında Çerezlerin Devre Dışı Bırakılması Durumu 68
Şekil 32. 1. Sınıflandırma Ağacı Budama Kararı 70
Şekil 33. CART Algoritması İle Oluşturulan 1. Sınıflandırma Ağacı 71
Şekil 34. 2. Sınıflandırma Ağacı Budama Kararı 73
Şekil 35. CART Algoritması İle Oluşturulan 2. Sınıflandırma Ağacı 76
Şekil 36. 3. Sınıflandırma Ağacına Ait Budama Kararı 78
Şekil 37. CART Algoritması İle Oluşturulan 3. Sınıflandırma Ağacı 79
Şekil 38. 4. Sınıflandırma Ağacına Ait Budama Kararı 82
Şekil 39. CART Algoritması İle Oluşturulan 4. Sınıflandırma Ağacı 83

TABLOLAR DİZİNİ

Tablo 1. Karar Ağacı Algoritmaları Avantajları ve Dezavantajları 36
Tablo 2. Uygulamada Kullanılan Değişkenler ve Açıklamaları 69
Tablo 3. 1. Sınıflandırma Ağacı Modeli 71
Tablo 4. 2. Sınıflandırma Ağacı Modeli 74
Tablo 5. 3. Sınıflandırma Ağacı Modeli 78
Tablo 6. 4. Sınıflandırma Ağacı Modeli 82
Tablo 7. Sınıflandırma Ağacı Karar Kurallarından Bazıları 85
Tablo 8. Hatalı Sınıflandırma Oranları 87

SİMGELER VE KISALTMALAR

CART	Classification and Regression Trees
CT	Classification Tree
RT	Regression Tree
CRM	Customer Relations Management
CHAID	Chi-Squared Automatic Interaction Detection
SLIQ	Supervised Learning in QUEST Algorithm
SPRINT	Scalable Paralellizable Induction of Decision Trees
MARS	Multivariate Adaptive Regression Splines
QUEST	Quick, Unbiased, Efficient, Statistical Tree
QDA	Quadratic Discriminant Analysis
LSD	Least Squared Deviation

1. GİRİŞ

Günümüzde bilişim teknolojilerinin hayatın her alanında yer almasıyla veri boyutlarında artışlar meydana gelmiştir. Veri boyutlarındaki artış, anlamlı veriye ulaşmayı zorlaştırmış ve büyük veri kavramının oluşmasına neden olmuştur. Büyük veriler, ilişki bakımından birçok farklı verinin bir arada bulunması anlamına da gelmektedir. Büyük verileri anlamlı veriler ve anlamsız veriler şeklinde inceleyecek olursak, anlamlı verilere ulaşma çabası daha değerli hale gelecektir. Verilerin artmasıyla, analiz edilmesi ve yorumlanması insan gücünün ötesine geçmiş ve bu durum nitelikli iş gücüne ihtiyaç duyulmasına neden olmuştur. Bu ihtiyaç, "veri madenciliği" kavramının doğmasına ve gelişmesine katkı sağlamıştır. Veri madenciliği, anlamlı ve anlamsız verilerin bir arada bulunduğu yapı olan veri madeninden anlamlı verilerin keşfedilmesidir. Anlamlı verilerin analiz edilmesi ve yorumlanması için bilgisayar algoritmaları geliştirilmiştir. Geliştirilen bilgisayar algoritmalarına da makine öğrenimi denilmektedir. Bu kavramlar, bilginin analiz edilmesi ve anlamlı verilerden tahminleme yapma işlemi olarak açıklanabilir.

Veri madenciliğinde, analiz edilecek veri için herhangi bir sınırlama bulunmamaktadır. Sağlık ile ilgili veriler, araştırma verileri, internetten elde edilen veriler, banka verileri, ekonomi ile ilgili veriler analiz edilebilmektedir. Videolar, fotoğraflar, ses kayıtları, yazılar dahil olmak üzere birçok şey "veri" olarak adlandırılabilmektedir. Bu nedenle, artık veri olarak adlandırdığımız yapı karmaşık olduğu için büyük veri projeleri genellikle yapay zeka ve makine öğrenimi içeren analitik yapıyı kullanmaktadır. Bilgisayarlar verilerin temsil ettiği yapıyı çeşitli algoritmalarla daha kolay analiz edebilmektedir. Bu durum, iş dünyasında devrim yaratmaktadır. Şirketler, müşterilerin belirli segmentlerini, ürünü ne zaman satın almak isteyebileceklerini ya da tercih kriterlerini hata payı en az olacak şekilde tahmin etmektedirler. Bu nedenle, şirketler operasyonlarını daha verimli şekilde yürütmektedirler.

Veri madenciliği konusu olan Sınıflandırma ve Regresyon Ağacında (CART), bağımlı değişkenin kategorik olduğu durumlarda Sınıflandırma Ağacı (Classification Tree, CT), sürekli olması durumunda ise Regresyon Ağacı (Regression Tree, RT) kullanılmaktadır. Bu yöntemlerde verilerin sınıflandırılması işlemi iki aşamadan

oluşmaktadır. Öğrenme basamağı adı verilen birinci aşamada, eğitim verisi sınıflama algoritması tarafından çözümlenmekte ve model oluşmaktadır. İkinci aşamada ise eğitim verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla test edilmektedir. Doğruluk kabul edilebilir bir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılmaktadır (Silahtaroğlu, 2008). Büyük ve karmaşık veri setlerinde karar ağacı algoritmalarının kullanılmasıyla görsel ve anlaşılır bilgilerin elde edilmesi kolaylaşmaktadır (Aitkenhead, 2008). Bu nedenle, veri madenciliği yöntemlerinden olan CART; endüstri, mühendislik, sağlık, finans olmak üzere birçok bilimsel alanda yaygın olarak kullanılmaktadır.

Bu çalışmanın amacı, sınıflandırma ve regresyon ağacı algoritması olan CART ile 2017 yılı Türkiye geneli internet üzerinden bulut depo kullanımı yapan bireyleri demografik, sosyal ve kültürel kriterler bakımından incelemektir.

Çalışmanın birinci bölümünde veri madenciliği, sınıflandırma ve regresyon ağaçlarına ilişkin çalışmalar yıllara göre kısaca anlatılmaya çalışılmıştır.

Çalışmanın ikinci bölümünde, büyük verinin tanımı ve yapısından bahsedilerek, günümüzde büyük veri ile meydana gelen değişimler anlatılmıştır.

Çalışmanın üçüncü bölümünde, veri madenciliği kavramı, uygulama alanları, makine öğrenimi kavramı ve kullanılan modeller anlatılmaya çalışılmıştır.

Çalışmanın dördüncü bölümünde, makine öğreniminde karar ağacı algoritmaları teorik olarak anlatılarak algoritmaların avantaj ve dezavantajları ele alınmıştır.

Çalışmanın beşinci bölümünde, uygulamada kullanılan sınıflandırma ağacı ve regresyon ağacı teorik olarak açıklanmış ve ayırma kriterleri hakkında bilgiler verilmiştir.

Çalışmanın altıncı bölümünde yapılacak uygulama, bulut depo kullanımı yapan bireylere ilişkin olduğu için önce bulut bilişim ve bulut depo kullanımı ile ilgili anlatıma yer verilmiştir. Türkiye'de internet üzerinden bulut depo kullanımı yapan bireylere ait özellikler CART algoritması ile incelenmiştir. Uygulamada Türkiye İstatistik Kurumu (TÜİK)'in 2017 yılında Türkiye genelinde 16-74 yaş arası bireylere uyguladığı "Hanehalkı Bilişim Teknolojileri Kullanımı Anketi" verileri kullanılmıştır. Uygulamanın ilk aşamasında, 2015-2016 ve 2017 yılları "Hanehalkı

Bilişim Teknolojileri Kullanımı Anketi" sorularına ait analizler karşılaştırmalı olarak yapılmış ve yıllara göre değişim yüzdeleri gösterilmiştir. Uygulamanın ikinci aşamasında, %60, %70, %80 ve %90' eğitim verileriyle analiz sonucu sınıflandırma ağaçları oluşturulmuş ve ağaçlar karşılaştırmalı olarak yorumlanmıştır.

2. LİTERATÜR ÖZETİ

Sınıflandırma ve regresyon ağacı ile yapılmış birçok çalışma bulunmaktadır. Araştırma kapsamında incelenen çalışmalar literatür özeti olarak açıklanmaya çalışılmıştır.

Oğuzlar (2004), Türkiye İstatistik Kurumu'nun 2002 Hanehalkı işgücü anketi sonuçlarını sınıflandırma ve regresyon ağacı ile analiz etmiştir. Çalışma sonucunda en yüksek oranda iş arayan ve aramayan gruplar ve cinsiyetlerin dağılımları hakkında bilgi vermiştir.

Dener ve diğerleri (2009), lisansüstü öğrencilere ait verileri kullanarak çeşitli algoritmalarla sınıflandırma yapmıştır ve açık kaynak kodlu yazılımlara ait başarıların derecelerini karşılaştırmalı incelemişlerdir.

Emel ve Taşkın (2005), bir perakendeci işletmenin gerçekleşen toplam satış hasılatının ürünlere göre dağılımının ayrıntılarını ve ürünlerin toplam satış değerlerinin toplam satış hasılatı üzerindeki göreceli önemi müşteriler ile ilişkilendirilerek sınıflandırmıştır.

Temel ve diğerleri (2005), huzursuz ayak sendromu olan ve olmayan 206 kişiye 10 soru sorularak sendromu olan ve olmayan bireyleri ayırmada etkili olan risk faktörleri sınıflandırma ağaçları yardımıyla sınıflandırmıştır.

Ayık ve diğerleri (2007), Atatürk Üniversitesi öğrencilerine ait veritabanında bulunan tüm verileri kullanarak sınıflandırma analizi yapmışlardır.

Kayri ve Boysan (2008), Yüzüncü Yıl Üniversitesi öğrencileri arasından seçilen 437 lisans öğrencisine birkaç test yapılması sonrası elde edilen verilerle bilişsel yatkınlık ile depresyon düzeyleri ilişkisini sınıflandırma ve regresyon ağacı ile incelemiştir.

Albayrak ve Yılmaz (2009), İMKB verilerinden yararlanarak karar ağaçları ile sınıflandırma analizi metoduyla veri madenciliği yapmışlardır.

Dolgun ve diğerleri (2009), Bir telekomünikasyon şirketine ait verileri karar ağacı algoritmalarından yararlanarak terk eden müşteriye ait profil modeli oluşturmuşlardır.

Sezer ve diğerleri (2010), Karar ağacı derinliklerinin sınıflandırma ve regresyon ağacı algoritmasının kestirim kapasitesine etkisinin incelemişlerdir.

Aktürk ve diğerleri (2012), Nisan 2012 tarihinde Çanakkale şehir merkezinde tüketicilerle yüz yüze anket yöntemi ile verileri toplamıştır. Tüketicilerin sosyo-ekonomik özellikleri ve zeytinyağı tüketim davranışlarını etkileyen özellikler sınıflandırma ve regresyon ağacı yöntemiyle analiz etmişlerdir.

Kavzaoğlu ve diğerleri (2012), Trabzon iline ait heyelan duyarlılık haritasının üretilmesinde sınıflandırma ve regresyon ağacı yönteminden yararlanmıştır. Bu çalışmada, sınıflandırma ve regresyon ağacından elde edilen sonuçlarla lojistik regresyon yöntemiyle yapılan analiz sonuçları karşılaştırılmıştır.

Kaya ve diğerleri (2012), Epileptik EEG işaretlerini karar ağaçları ve karar kurallarını kullanarak sınıflandırarak tanı performanslarının oldukça yüksek olduğunu tespit etmiştir.

Özkan (2012), Yukarıgökdere yöresinde dikdörtgen şeklinde kesilen belirli bir alanda yazım aşamasında olan iki çalışmanın ve tür çeşitliliğinin ağaç ve coğrafi dağılım modelleri sınıflandırma ve regresyon ağacı yöntemi kullanarak analiz etmiştir.

Alan (2014), Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Fakültesi'nde bulunan 8 bölümde öğrenim gören 4106 öğrenciye ait verilerle sınıflandırma yapmıştır. Annenin ve babanın çalışma durumu, cinsiyet, yaş, gelir gibi değişkenlerin öğrenim kredisi alıp almamaya dair sınıflandırma analizi yapmıştır.

Avcı ve Altay (2014), 1990-2010 dönemi Türkiye, Arjantin, Meksika, Malezya ve Tayland'da yaşanan finansal krizlerin öngörüsünde regresyon ağaçları yöntemini kullanmıştır.

Akşahan ve Keskin (2015), Türkiye'nin Afyon ili Bolvadin ilçesinde yetiştirilen 103 baş tosunun besi sonucu ağırlığını etkileyen bazı vücut ölçülerini regresyon ağacı yöntemi ile belirlemiştir. Bağımlı değişken olarak da besi sonu canlı ağırlığı kullanılmıştır.

Ersöz ve Özseven (2015), kobilerin finansal sorunlarını etkileyen faktörler sınıflandırma ve regresyon ağacı ile modellemişlerdir.

Şatır ve diğerleri (2016), Hacıefendioğlu'nun yüksek lisans tezinden alınan hasta bilgileri veri kümesini kullanmış ve nitelik sayısı kaba kümeler yöntemi ile 8'den 5'e düşürülmüştür. İndirgenen veriler karar ağaçları ve yapay sinir ağları kullanılarak sınıflandırılmış ve ardından çapraz doğrulama ve performans değerlendirmesi gerçekleştirilmiştir.

Güneri ve Aydın (2017), Toplam beş farklı bölgeden yedi farklı Apodemus türlerine ait iki farklı veri setiyle grup üyelerini tahmin etmek ve sınıflandırma yapabilmek için diskriminant analizi, çok terimli lojistik regresyon ve regresyon ağacı kullanmışlardır.

Takma ve diğerleri (2017), Türkiye'de yetiştirilen bir yumurtacı sürüsünden toplamda 1980 adet yumurtacının verimleri ve özelliklere ilişkin katsayıları çalışmada kullanmıştır. Bu araştırmada yumurta adedi üzerine kafes, sıra, eşeysel olgunluk yaşı ve ağırlığı ile kuluçka dönemi özelliklerinin etkisi sınıflandırma ve regresyon ağacı ile analiz edilmiştir.

Uysal ve diğerleri (2014), kalp hastalığı verilerini twoing algoritması ile sınıflandırmıştır. Sınıflandırma sonucunda karar kuralları oluşturularak kalp hastalığının belirtileri açıklanmıştır. Çalışmanın sonunda cinsiyet, maksimum kalp hızı, açlık kan şekeri, göğüs ağrısı tipi, talasemi, yaş, büyük damarlar ve anjine bağlı depresyon değişkenlerinin kalp hastalığını ne derece etkilediği belirlenmiştir.

Alan ve Dündar (2017), yatırım teşvik verilerinden yararlanarak verileri en başarılı sınıflandıran algoritma ve bu algoritmanın ürettiği sınıfları belirlemeye çalışmışlardır.

Sınıflandırma ve regresyon ağaçları ile yapılan çalışmalardan da anlaşılacağı gibi, karar ağaçları büyük veri setlerine kolayca uygulanabilmektedir. Elde edilen sonuçların yorumu, diğer yöntemlere göre daha kolay ve açıklayıcıdır. İlişkilerin yönünü ve önem sırasını da görsel olarak ortaya koymaktadır (Albayrak vd., 2009).

3. BÜYÜK VERİNİN TANIMI VE ÖZELLİKLERİ

Büyük veri terimi, 1990'ların başında kullanılmaya başlanmış ve önemi yıllar geçtikçe artmıştır. Büyük veriler, genellikle veri stratejisinin ayrılmaz bir parçası olarak da görülmektedir.

Büyük veri, içerisinde birçok bilgiyi barındırmaktadır. Bu yüzden verinin doğru değerlendirilmesi büyük önem taşımaktadır.

Günümüzde dijital ortamlarda gerçekleşen eylemler, hakkımızda birçok verinin toplanmasına neden olmaktadır. Ancak, bu veriler anlamlı veya anlamsız olabilmektedir. Bu veriler içinden anlamlı verilerin çıkarılmaya çalışılması, büyük veri kavramını ortaya çıkarmaktadır. Özellikle ticari açıdan düşünüldüğünde işletmeler, veriyi daha anlamlı hale getirebilmek için yeni yollar arayışı içindedirler. Yenilikler yapmak, üretimi daha akılcıl hale getirebilmek ve işletmelerin sürdürülebilirliğini sağlayabilmek için verilerin toplanacağı ve analizlerinin yapılacağı departmanlar gün geçtikçe artış göstermektedir.

Büyük verinin asıl ilkesi, herhangi bir durum hakkında edinilmiş bilgiye bağlı olarak, daha güvenli ve anlamlı bir şekilde yeni kavramlarla gelecekte ne olacağı konusunda tahminlerde bulunmaktır. Büyük veri, sayıca fazla verinin karşılaştırılmasına olanak sağlar ve daha önceden fark edilmemiş birçok kavramın ve gerçeğin fark edilmesini sağlayarak daha doğru kararlar almamıza olanak sağlamaktadır. Günümüzde gelişen teknoloji ile birlikte birden fazla algoritma yardımıyla büyük veriler analiz edilerek modeller oluşturulmakta ve tahminler üretilmektedir. Büyük verileri analiz eden ve bu analiz sonuçlarını gelecek stratejilerini belirlemek için kullanan şirketler büyüyecek, veriyi göz ardı eden şirketler ise ya küçülecek ya da gerileyecektir. Büyük veri, veri girişimlerinin hem avantajlarını hem de dezavantajlarını anlamamıza yardımcı olmaktadır.

Büyük veri, yaşamımızda terabaytları, perabaytları ve verileri tanımlamak için kullanılır. Veriler, hayatımızı büyük bir oranda değiştirmektedir. Veri yapısı arttıkça, verileri analiz edebileceğimiz bilgisayar algoritmaları da artacaktır. Yalnızca işletmeler için değil, hayatın her alanında büyük veri yıllar geçtikçe çok kritik bir hale gelecektir.

Büyük veride en büyük sorun, mevcut verilerin hacminin büyük olması sebebiyle anlamlı verilerin oluşturulması istenirken maliyetinin artış göstermesidir. Bu yüzden işletmeler ve kişiler kısa süreli ya da bazı zamanlarda kullanabilecekleri veri depolama araçlarına yatırım yapmak istememektedir.

Günümüzde, atılan her adım dijital iz bırakmaktadır. Sosyal medyada çevrimiçi sohbetler, lokasyon paylaşımları, sohbet uygulamaları ve online yapılan alışverişler kişiler hakkında verilerin toplanabildiği araçtır. Yalnızca insanların değil insan gücüyle yapılan makineler de veri miktarında artışa neden olmaktadır. Endüstriyel makineler, verilerin toplanması ve aktarılması için daha donanımlı üretilmektedir. Büyük veri, verinin geniş yelpazeden toplanmasını sağlar ve bu durum bazı yönleriyle gelecek açısından bir avantajdır. Büyük veriler, ticari satış kayıtları, bilimsel deneylerin sonuçları ve internette kullanılan bilgiler gibi birbirinden farklı kaynaklardan elde edilebilmektedir. Ancak, veriler daha öncesinde farklı yazılım araçlarıyla verimli hale getirilmiş ya da getirilmemiş olabilmektedir.

Diğer taraftan büyük veri, verilerin yapısından kaynaklanan birçok özelliklere sahiptir. Bu özellikler aşağıda yazıldığı gibi açıklanmıştır.

Variety (Çeşitlilik): Üretilen verilerin tek bir yapısı yoktur ve birçok farklı ortamdan elde edilen formatlardan oluştukları için verilerin birbirlerine dönüştürülebiliyor olması gerekmektedir. Ses, resim, video dosyaları, çerezlerin oluşturduğu dosyalar, tıklama verileri de çeşitliliğe örnek gösterilmektedir.

Velocity (Hız): Büyük veri üretimi gün geçtikçe hız kazanmaktadır. Bu veriler kısa zamanda inanılamayacak boyutlara ulaşmaktadır. Verinin hızla artmasıyla birlikte işlem sayısı ve hızı da artış göstermektedir. Hem yazılımsal hem de donanımsal hız, verinin artışıyla doğru orantılı olmak durumundadır. Google, günde ortalama 3,5 milyardan fazla arama yapıldığını, bu durumun saniyede 40.000 aramaya karşılık geldiğini açıklamıştır.

Volume (Veri Büyüklüğü): Verilerin hızla büyümesi gelecekte depolanması gereken veri sayısının artacağına işaret etmektedir. Bu nedenle verilerin artmasına göre oluşan veri yığınının nasıl depolanacağı konusu da önemli bir strateji gerektirmektedir. Her dakika 300 saat video youtube'a yüklenmektedir. Bu durum mevcut veri büyüklüğünü göstermektedir.

Verification (Doğrulama): En önemli veri karakteridir. Verinin hızla artışına bağlı olarak çok farklı stratejiler geliştirilebilir, ancak verinin güvenliği tüm stratejilerden daha önemlidir. Verilerin kimler tarafından üretildiği, saklandığı, görüldüğü ve aynı zamanda doğru veriler olup olmadığının incelenmesi gerekmektedir.

Value (Değer): Verinin bir diğer karakter bileşeni ise değerdir. Verinin üretilmesi ve işlenmesi aşamalarından sonra çalışmalar için anlamlı ve verimli olması, verinin değer bakımından karakterini yansıtmaktadır.

Variability (Değişkenlik): Büyük veride değişkenlik birkaç farklı anlama gelmektedir. İlk olarak verilerde oluşan tutarsızlıklar incelenebilmektedir. Var olan verilerle anlamlı bir analiz yapılabilmesi için, verilerde normalliğin ve aykırı değerlerin incelenmesi gerekir. Büyük veri, birden fazla farklı veri türünü bünyesinde barındırması ve bu veri türlerinin boyutları nedeniyle de değişkenlik göstermektedir.

Veracity (Doğruluk): Büyük verinin en önemli özelliklerinden biridir. Büyük verinin çeşitliliği, hızı, veri büyüklüğü, değeri ve değişkenliği arttıkça doğruluk düşmektedir. Doğruluk, veri kaynağı ile analizin anlamlılığının ve güvenilirliğinin ne kadar önemli olduğunu ifade etmektedir. Doğruluk, büyük veriye ilişkin riskleri daha iyi anlamamıza yardımcı olmaktadır.

Validity (Geçerlilik): Geçerlilik, verilerin kullanım amaçlarının ne kadar doğru olabileceğini ifade etmektedir. Forbes yaptığı bir açıklamada, bir veri bilimcinin analiz yapmadan önce zamanının %60'ını verileri temizlemek için harcadığını belirtmiştir.

Vulnerabilitiy (Güvenlik Açığı): Büyük veri, güvenlik endişelerini de beraberinde getirmektedir. Büyük veri içeren bir konuda yapılan en ufak bir ihlal, geri dönülmez sonuçlar doğurabilmektedir.

Volatility (Oynaklık): Büyük veri kavramı ortaya çıkmadan önce, birçok kuruluş verilerin süresiz depolanması taraftarıydı. Ancak büyük verinin hacmi ve hızından ötürü oynaklık kavramına çok dikkat etmek gerekmektedir. Büyük veride, her verinin bir ömrü bulunmaktadır. Bu yüzden, elde edilen verinin ne kadar süre sonra geçersiz olacağının bilinmesi gereklidir.

Teknolojinin gelişmekte olduğu ülkelerde büyük veri analizi yapılmaktadır. Bu analizler, savunma sanayiden seçim sonuçlarının tahminine kadar önemli alanlarda kullanılmaktadır. Bu nedenle, anlamlı verilere ulaşabilmek ve gerekli analizleri yapabilmek için veri madenciliğine ihtiyaç duyulmaktadır.

Visulization (Görselleştirme): Büyük veride görselleştirme oldukça zor bir kavramdır. Çok sayıda veriye ait grafikler oluşturmaya çalışırken geleneksel yöntemler yeterli olmayacaktır. Bu nedenle, veri kümeleme ve ağaç grafikleri gibi verileri temsil edecek farklı görselleştirme tekniklerine ihtiyaç duyulmaktadır.

Sonuç olarak, yukarıda özellikleri açıklanan büyük veri, bilgilerin toplanması ve ilişkilendirilmesi, analiz edilmesi ve ardından da her türlü soruya cevap alınabilecek şekilde anlamlı verinin oluşturulmasıdır.

4. VERİ MADENCİLİĞİ

Bilgisayarların üretilmesiyle verilerin saklanması ve depolanması süreçlerinde hızlı gelişim sağlanmış, günümüzde büyük veriler kolaylıkla depolanabilir hale gelmiştir. Depolanan verilerin zaman geçtikçe büyümesi, anlaşılmasını da zorlaştırmıştır. Diğer taraftan, veri içerisinde saklı kalmış yararlı bilgilerin elde edilmesi oldukça zor hale gelmiştir. Geleneksel sorgu ve raporlama araçlarının veriler karşısında yetersiz kalması, veri madenciliği ve veri madenciliği altında yapılan sınıflamalar gibi yeni araştırmaların yapılmasına neden olmuştur (Kuyucu, 2012). Bununla birlikte, yeni yöntem ve teknolojilerin geliştirilmesi ihtiyacı da ortaya çıkmıştır. Veri madenciliği; istatistik, makine öğrenimi, enformasyon teknikleri, veritabanı teknolojileri ve ilgili diğer disiplinlerdeki tekniği bir araya getirmektedir. Veri madenciliği yöntemi, anlamlı örüntülerin otomatik olarak keşfedilmesi amacıyla da kullanılmaktadır (Pang-Ning Tan vd., 2006).

4.1. Veri Madenciliği Kavramı

Veri madenciliği, büyük verilerden daha önce fark edilmemiş potansiyel olarak kullanışlı bilginin ya da anlamlı verinin çıkarılması demektir. Bu durum, kümeleme ve verilerdeki sapmaların tespiti gibi birçok istatistiksel araştırmayı içermektedir.

Veri madenciliği, şirketler tarafından işlenmemiş verilerin faydalı bilgilere dönüştürülmesi amacıyla kullanılmaktadır. Büyük verilerden istenilen bilgileri alabilmek amacıyla çeşitli yazılımlar kullanılarak, şirketler müşterileri hakkında daha fazla gerekli bilgiye ulaşabilirler ve buna uygun bir stratejik planlama ve pazarlama haritası çıkarabilirler. Veri madenciliği, şirketlerin istenilen bilgiye ulaşmalarını kısa sürede ve kolay yollarla sağlamaktadır. Bu yüzden veri madenciliği, az maliyetle kısa zamanda amaca ulaşılmasını desteklemektedir.

Veri madenciliğini istatistiksel yöntemler dizisi olarak görmek mümkündür. Ancak, istatistik ve veri madenciliği arasında farklılıklar bulunmaktadır. Veri madenciliğinde mantık kuralları çerçevesinde görsel olarak destekleyici nitel modeller oluşturulmaktadır. Veri madenciliği, insanı temel olarak gören bir alan olmakla birlikte insan ve bilgisayar birlikteliğini de temeline yerleştirmektedir.

Veri madenciliğinin tarihsel gelişimi, 1950'li yıllarda bilgisayarların sayımlarda kullanımıyla başlamıştır. 1960'lı yıllarda hiyerarşik ve ağ modeller ile veritabanları

oluşmaya başlamış, 1970'li yıllarda ilişkisel veri modelleri ortaya konulmuş ve 1980'li yıllarda bu modeller uygulanmaya başlanmıştır. 1990'lı yıllarda büyük verilerin değerlendirilmesiyle ilgili çalışmalar yapılmış ve 1992 yılında veri madenciliğiyle ilgili ilk yazılım geliştirilmiştir. Bu tarihten itibaren, 2000'li yıllarda veri madenciliği gelişerek uygulama alanları da gittikçe yaygınlaşmıştır.

Veri madenciliğinin amacı, toplanmış verilerin bir takım istatistiksel yöntemlerle incelenmesidir. Veri madenciliği sürecinde büyük verinin verimli ve etkin hale getirilmesi amaçlanmaktadır.

Büyük verilerde ortaya çıkmamış örüntüler bulunmakta ve bu örüntüleri ortaya çıkarma süreci de veri madenciliği kapsamında yer almaktadır. Bu süreçte, ortaya çıkmamış örüntüleri ortaya çıkararak veri etkin hale gelir ve değer kazanır.

Veride iki ya da daha fazla değişken varsa ve bu değişkenler arasındaki ilişki açıklanmak isteniyorsa veri madenciliği kullanılmaktadır. İçinde bulunduğumuz çağda bilgi kirliliği ve teknolojinin gelişmesiyle üretilen bilgiye bakıldığında doğru bilgiye erişmenin gerekli ve zorunlu olduğu gerçeği ortaya çıkmıştır. Veri madenciliği, anlaşılması zor olan kapalı bilgilerin kolaylıkla ve doğru olarak ortaya çıkmasını sağlar.

İnternet üzerinde bilgilerin teknolojik gelişmelerle birlikte artması, genetik araştırmalarda gelişmelerin artması, kullanılan eski tekniklerin gerçek bilgiye ulaşma konusunda yetersiz kalması, bilimsel hesaplamalar ve modellerin giderek gelişmesi, verilerin sınıflandırılmasına duyulan ihtiyaç, sosyal medyanın kullanılmasıyla bilgilerin sürekli artışı, bankacılık işlemlerinde kişilerin bilgilerinin depolanması ve bilgilerde oluşan artış, müşteri memnuniyetlerinin sağlanması amacıyla yapılan çalışmalar veri madenciliğinin kullanım nedenleri arasında sayılabilmektedir.

4.2. Veri Madenciliğinin Uygulama Alanları

Veri madenciliği içinde bulunduğumuz çağda bilgi kirliliği ve teknolojinin gelişmesiyle üretilen bilgiye bakıldığında doğru bilgiye erişmenin gerekli ve zorunlu olduğu gerçeği ortaya çıkmıştır. Veri madenciliği, anlaşılması zor olan kapalı bilgilerin kolaylıkla ve doğru olarak ortaya çıkmasını sağlar.

Gerek kamu gerekse özel kurum ve kuruluşlarda alınan kararlarda riski azaltmak ve zarar etmemek amaçlı çok fazla bilgiye ulaşma ihtiyacı doğmaktadır. Verinin hacmi kadar verinin niteliği de çok önemlidir ve kullanılan veri kümesinin yeterliliği ne kadar iyi olursa, alınan kararlarda ya da yapılan çalışmalarda riskin azalması kolaylaşacaktır.

Günümüzde veri madenciliğinden; bankacılık, sigortacılık, perakendecilik, borsa, telekomünikasyon, sağlık ve spor gibi birçok alanda yararlanılmaktadır. Veri madenciliğinden yararlanan bu alanlar aşağıda görüldüğü şekilde açıklanmıştır.

- Bankacılık

Bankacılık alanında veri madenciliği, kredi kartı dolandırıcılığının tespitinde, kredi kartı sahiplerinin yaşam olasılıklarının belirlenmesinde, finansal tablolarda hilelerin tespit edilmesinde, müşteri sadakatini sağlamak amaçlı CRM stratejilerinin değerlendirilmesinde, risk analizlerinin yapılmasında, kredi taleplerinin değerlendirilmesinde, bankaların gelir-gider, likidite, karlılık ve faaliyet oranlarına göre sıralanmasında kullanılmaktadır.

- Sigortacılık

Sigortacılık alanında, müşterilerin gelecekteki davranışlarının ve müşterilerin tercihlerinin tespit edilmesinde, risk değerlendirmesinde, dolandırıcılık tespitinde, hak taleplerinin toplanması ve puanlanmasında veri madenciliğinden yararlanılmaktadır.

- Perakendecilik

Perakendecilik alanında veri madenciliği, özellikle kozmetik, giyim, market alışverişlerinin yapıldığı mağazalar tarafından kullanılmaktadır. Mağazalar tarafından ücret karşılığında müşterilere verilen alışveriş kartları bulunmakta ve bu kartlar sayesinde müşteriler indirim olanaklarından yararlanmaktadır. Bu kartlar sayesinde müşterilerin satın aldıkları ürünlerin fiyatları, alışveriş yapılan tarihler, hangi ürünlerin hangi müşteri tarafından sürekli olarak satın alındığını takip etmek kolaylaşmaktadır. Mağazalar tarafından bu veriler analiz edilerek müşterilerin satın alma alışkanlıklarına göre ürünün satış zamanı, satışa sunulması gereken fiyatı ve

müşterinin doğum günü, özel günlerine indirimler gibi çeşitli yollar veri madenciliğinden yararlanılarak yapılmaktadır.

Ayrıca; satış ve stok tahmininde, markalar arası rekabet artışını etkileyen faktörlerin incelenmesinde, kar marjlarının düşme sebeplerinin değerlendirilmesinde, alışveriş sepeti analizlerinde, mağazayı terk edecek müşterilerin belirlenmesinde de veri madenciliği kullanılmaktadır.

- Borsa

Borsada veri madenciliği, hisse senetlerinin analizinde, genel piyasa analizlerinde, borsa şirketlerinin sektörel risk profillerinin belirlenmesinde, fiyat değişimlerinde paralellik gösteren hisse senetlerinin bulunmasında kullanılmaktadır.

- Telekomünikasyon

Telekomünikasyon alanında veri madenciliği, uzun vadeli müşterilerin tespitinde, hisse tespitlerinde, müşteri davranışlarının tespitinde, müşteri odaklı anahtar performans göstergelerinin oluşturulmasında, GSM şebekelerinin performans analizlerinin değerlendirilmesinde kullanılmaktadır.

- Sağlık ve İlaç

Sağlık alanında, solunum fonksiyon testlerinin analizinde, kanserli hücrelerin tespitinde, test sonuçlarının tahmininde, tedavi sürecinin belirlenmesinde veri madenciliğinden yararlanılmaktadır (Pehlivan, 2006).

- Spor

Spor alanında, futbol karşılaşmalarının analizinde, tenis maçında oyuncuların hangi alana isabetli şutlar attığının belirlenmesinde veri madenciliğinden yararlanılmaktadır (Altunkaynak, 2017).

4.3. Veri Madenciliği Süreci

Veri madenciliği bir süreç olmakla birlikte bilgiyi keşfetme olarak da adlandırılır. Öncelikle şirketler, verileri toplayarak veri ambarına yüklemektedirler. Veriler, şirket içi sunucularda veya bulutta depolanmakta ve yönetilmektedir. Şirketlerde bu verilerin erişimi ve düzenlenmesi için iş analistleri, bilgi teknolojisi uzmanları ve yönetim ekipleri görevlendirilmektedir. Yazılımlar, istenilen verileri sıralayarak grafik ya da tablo gibi kolay anlaşılacak biçimde kullanıcıya sunmaktadır.

Veri madenciliği süreci, aşağıda görülen adımlardan oluşmaktadır.

- Veri Toplama

Veri madenciliğinde araştırma yapabilmek için öncelikle büyük veriye ihtiyaç vardır. Büyük veriler için de veri toplamak gereklidir. Bir araştırma için analiz yapmak istenildiğinde veri tabanlarından veya veri ambarlarından veri toplama işlemi gerçekleştirilmektedir. Veri toplandıktan sonra, veriler eğitim verisi ve test verisi olmak üzere ikiye ayrılarak gerekli analizler ve çalışmalar yapılmaktadır.

- Veri Temizleme

Veri toplama aşamasında verilerin tutarsız veriler olup olmadığına dikkat edilmemektedir. Veri tabanında yer alan hatalı ve tutarsız verilere gürültü adı verilmektedir. Hatalı ve tutarsız verileri temizlemek için birçok yöntem bulunmaktadır. Hatalı değer içeren gözlemlerin veri setinden çıkarılması, verilere uygun regresyon ya da karar ağacı uygulaması ile hatalı gözlemlerin tamamlanması yapılabilir, diğer verilerin ortalaması ile hatalı veriler yerine bu değerler yazılabilir.

- Veri Bütünleştirme

Birbirinden farklı veri kaynaklarından ya da veritabanlarından alınan verilerin birlikte değerlendirmeye alınabilmesi için verilerin aynı türde gösterilmesi gerekmektedir. Veri bütünleştirmeye, veritabanlarında medeni durumun evli/bekar ya da (evli:0, bekar:1) şeklinde gösterilmesi örnek olarak verilebilir.

- Veri İndirgeme

Veri madenciliğinde genellikle büyük veri setleriyle çalışmalar yapılmaktadır. Bu sebeple, uygulamalarda büyük veri setlerinde sonuç değişmeyecek şekilde gözlem ya

da değişken sayısı azaltılabilmektedir. Örnekleme, boyut indirgeme yöntemlerden bazılarıdır.

- Veri Dönüştürme

Veri dönüştürme, modele uygun olarak veri içeriğinin korunarak şeklini dönüştürme işlemidir. Her verinin uygun model ve algoritma yapısı farklılık göstermektedir. Bu nedenle, yapılacak dönüştürme işleminde veri yapısına ve modele dikkat etmek gerekir.

- Veri Madenciliği Algoritmasının Uygulanması

Veriye gerekli işlemlerin yapılmasının ardından konuya uygun veri madenciliği algoritma ya da algoritmaları uygulanır.

- Sonuçları Sunma ve Değerlendirme

Algoritmaların uygulanmasının sonucunda elde edilen değerler düzenlenerek ilgili yerlere sunulmaktadır. Regresyon ağacında, verilerin ağaç grafiğinde sınıflandırılması örnek verilebilir (Çalış vd., 2014).

4.4. Veri Madenciliğinde Makine Öğrenimi Kavramı

Makine öğrenimi, sistemlere açık bir şekilde programlanmayan deneyimlerden faydalanarak otomatik olarak öğrenme ve gelişme yeteneği sağlayan yapay zeka ürünüdür.

Makine öğrenimi, verilere erişebilen bilgisayar algoritmalarının geliştirilmesine olanak sağlamaktadır.

Büyük verilerin kullanımının artmasıyla birlikte oluşturulması gereken modellemeler bilgisayar algoritmalarıyla yapılmaktadır. Verilerin modellenmesini sağlayan bilgisayar algoritmalarına makine öğrenimi denilmektedir. İstatistiksel ve matematiksel analizler yapılırken, analiz edilecek verilere uygun bilgisayar algoritmaları kullanılmaktadır.

Makine öğreniminde amaç, mevcut veri ve kullanılan bilgisayar algoritmalarıyla oluşturulan algoritmanın en yüksek performansı sergilemesidir. Makine öğreniminde yüksek performans, tahmine dayanmaktadır.

Makine öğreniminde, verinin yapısını belirlemek ve verilen örneklere dayanarak gelecekte daha iyi kararların alınabilmesi için deneyim, gözlem ve verilere ihtiyaç olduğu söylenebilir.

Makine öğreniminde öncelikli amaç bilgisayarlara, insan zekası ve müdahalesi olmaksızın, otomatik olarak öğrenimlerine izin vermek ve oluşan eylemleri ve karar alma mekanizmalarını buna göre ayarlamaktır.

Genel olarak makine öğrenimi, büyük verilerin analizinde kolaylık sağlamakta ve fırsatlar ile risklerin daha kolay tanımlanmasını ve fark edilmesini hızlandırmaktadır. Diğer taraftan makine öğrenimi, karar alma mekanizmalarında da doğru sonuçlar verebilmektedir. Makine öğrenimini yapay zeka ve yazılımsal teknolojik algoritmalar ile birleştirmek, büyük verilerin incelenmesinde ve analizinde daha da kolaylık sağlayacak ve etkili olacaktır. Ancak, doğru algoritmayı seçmek oldukça zor bir yöntemdir. Deneme yanılma yolu ile doğru algoritma seçimi sağlanabilmektedir. Kullanılacak algoritma seçimi analiz yapılacak verinin türüne, hacmine, veriden almak istenilen sonuçlara ve analiz tekniklerine göre de değişmektedir.

Makine öğrenimi algoritmaları, veri analizinde gerekli kararların alınması anlamını taşımaktadır. Kararların alınması aşamasında ise karar kuralları, karar ağaçları ve sinir ağlarından faydalanılmaktadır. Bu yöntemlerde ise çeşitli öğrenim stratejileri bulunmaktadır ve bu stratejiler; denetimli makine öğrenimi, denetimsiz makine öğrenimi, yarı denetimli makine öğrenimi, takviyeli makine öğrenimi ve yoğun makine öğrenimi şeklinde açıklanabilir.

Denetimli makine öğrenimi algoritmaları, geçmişte öğrenilenlerden gelecekteki olayları tahmin etmek için etiketli örnekler kullanımıdır. Denetimli makine öğreniminde veriler birbirleriyle etkileşim halindedir ve oluşan modelde analiz sonucu elde edilen sonuçlar ile olması hedeflenen sonuçların birbirine yakın olacak şekilde üretimi amaçlanmaktadır (Atalay ve Çelik, 2017).

Denetimli makine öğrenimi algoritmalarında, daha önce uygulanmış eğitim verisinin analizinden yola çıkılmaktadır. Oluşan algoritma, değerler hakkında bir tahmin üretmektedir. Sistem, yeterli eğitim sonrasında herhangi bir yeni verinin analizinde hedefler sağlayabilmektedir. Denetimli öğrenimde, modeli uygun hale getirmek için

oluşan değerlerin, amaçlanan değerlerle örtüşüp örtüşmediği kontrol edilir ve oluşan hatalar bu sayede bulunabilir.

Denetimsiz makine öğrenimi algoritmalarında modeli oluşturan veriler arasındaki ilişki ortaya çıkarılarak birbirine yakın değerler gruplandırılmaya çalışılmaktadır (Atalay ve Çelik, 2017). Ancak, eğitim verisi sınıflandırılmadığında denetimsiz makine öğrenimi algoritmaları kullanılamamaktadır.

Denetimsiz makine öğreniminde sistemler, gizli bir yapıyı etiketlenmemiş verilerden tanımlamak için nasıl bir işlevin ortaya çıktığını incelemektedir. Bu sayede gizli ve etiketlenmemiş verilerden çıkarım yapılabilmektedir. En yaygın denetimsiz makine öğrenim yöntemi, kümelemedir. Kümeleme teknikleri, genetik yapı araştırmalarında ve pazar araştırmalarında yaygın olarak kullanılmaktadır.

Yarı denetimli makine öğrenimi, denetimli ve denetimsiz makine öğrenimini birlikte kapsamaktadır. Çünkü gelecekteki olayları tahmin etmek için eğitim veri setinde etiketli ve etiketsiz veriler birlikte kullanılmaktadır. Yarı denetimli makine öğrenimi, büyük miktarda denetimsiz öğrenme koşulunu sağlayan veri ile küçük miktarda denetimli öğrenme koşulunu sağlayan verinin oluşturduğu modeldir. Bu yöntem kullanıldığı takdirde tahmin sonuçları büyük ölçüde doğru çıkacaktır.

Takviyeli makine öğrenimi, modelin oluşumundan ortaya çıkan sonuca göre verilerin iyi ya da kötü olmak üzere yorumlanması şeklinde gerçekleşmektedir (Atalay ve Çelik, 2017). Takviyeli makine öğrenimi, deneme ve hata bulma yolunu seçmektedir. Bu yöntem, makinelerin ve yazılım firmalarının performansını en üst düzeye çıkarmak için kullandıkları bir yöntemdir. Çünkü, takviyeli makine öğreniminde temel amaç var olan yapıyı güçlendirmektir.

Yoğun makine öğreniminde ise matematiksel ve istatistiksel olarak doğrusal ve doğrusal olmayan dönüşümler yer almaktadır. İstenilen modelin oluşumu için veriler birden fazla algoritma ile analiz edilmelidir.

Tahmin yapmak için model oluşturmak gerekiyorsa makine öğrenimi tekniklerinden en uygun ve en doğru olanı denetimli makine öğrenimi algoritmasını seçmektir. Satış ve stok tahminleri gibi süreçler bu algoritmalara örnek olarak gösterilebilmektedir.

Diğer taraftan eğer verilerden daha önce yararlanılmadıysa ve verileri kümelere ayırmak gerekiyorsa denetimsiz öğrenim algoritmalarını kullanmak daha doğru olacaktır.

4.5. Makine Öğreniminde Kullanılan Modeller

Makine öğreniminde kullanılan modeller, tahmin edici modeller ve tanımlayıcı modeller olmak üzere iki şekilde incelenmektedir. Çalışmada önce tahmin edici modeller, daha sonra da tanımlayıcı modeller anlatılmıştır.

4.5.1. Tahmin Edici Modeller

Tahmin edici modeller, regresyon ve sınıflandırma teknikleri ile tahminler içermekle birlikte tahmin edilecek yapıya göre farklılık gösterebilmektedir (Dunham, 2003).

Sınıflandırma, veri madenciliğinde en çok kullanılan yöntemlerden biridir. Sınıflandırmada veri setinin %70'i eğitim verisinden, %30'u test verisinden oluşmaktadır. Ancak, eğitim verisini %80, test verisini %20 ya da eğitim verisini %90, test verisini de %10 olarak almak mümkündür.

Sınıflandırmada süreç iki aşamada gerçekleşmektedir. İlk aşamada veri setinden rastgele eğitim verisi seçilir ve veri kümesine uygun model kurulur. Modelde veri tabanındaki isimler kullanılmaktadır. İkinci aşamada test verisi üzerinden sınıflandırma yapılır. Kurallar analiz verisi üzerinden sınanır ve karar kuralları ortaya konur.

Regresyon, tahmine dayalı bir sınıflandırma yöntemidir. Sınıflandırma, bir veri öğesini, önceden tanımlı sınıflardan birine tasnif ederken, regresyon veri öğesini gerçek değerli bir tahmini değişkene eşler (Fayyad vd., 1996).

Tahmin edici modellerde kullanılan başlıca yöntemler:

- Karar Ağaçları (Decision Trees)
- Bayes Sınıflandırması (Bayesian Classification)
- Hatayı Geri Yayma (Backpropagation)
- Karar Destek Makineleri (Support Vector Machines)
- K- En Yakın Komşu (K- Nearest Neighbour)
- Yapay Sinir Ağları (Neural Networks)

- Genetik Algoritmaları (Genetic Algorithms)
- Zaman Serisi Analizi (Time Series Analysis)
- Diğer Metotlar (Diskriminant Analizi, Faktör Analizi, Lojistik Regresyon Analizi)

4.5.2. Tanımlayıcı Modeller

Tanımlayıcı modeller, veriler arasındaki ilişki, benzerlik ya da sapmaların ortaya konmasını sağlamaktadır. Tanımlayıcı modeller tahmin için değil, mevcut durum analizi için kullanılmaktadır (Çelikten ve Karacan, 2013).

Tanımlayıcı modellerde kullanılan başlıca yöntemler:

- Kümeleme Analizi
- Birliktelik Kuralları Analizi

Kümeleme analizi, gruplanmış verileri benzerliklerine göre sınıflandırmada sıklıkla kullanılan çok değişkenli istatistiksel bir yöntemdir (Kalaycı, 2016). Kümeleme analizinde önemli olan nokta veriler arasındaki uzaklıklardır. Hiyerarşik ve hiyerarşik olmayan şeklinde iki farklı kümeleme yöntemi bulunmaktadır. Bu yöntemler, değişkenlerin aralarındaki farklılıklardan yararlanarak kümelere ait alt kümeler oluşturmayı sağlar.

Küme sayısına bağlı olarak kümeleme yöntemi de değişiklik göstermektedir. Hiyerarşik kümeleme yönteminde küme oluşumlarının daha iyi açıklanması için dendrogram grafiğinden yararlanılmaktadır. Hiyerarşik olmayan kümeleme yöntemi ise küme sayısı hakkında ön bilgiye sahip olma ve küme sayısına karar verme durumlarında tercih edilmektedir. Bu nedenle hiyerarşik ve hiyerarşik olmayan kümeleme yöntemlerinin birlikte kullanılması daha yararlıdır. Kümeleme analizinde verilerin normal dağılım varsayımı bulunmaktadır. Ancak, uygulamada verilerin normal dağılım varsayımı göz ardı edilmekte ve uzaklık değerlerinin normal dağılımı yeterli bulunmaktadır.

Tanımlayıcı modellerde kullanılan kümeleme analizi, homojen veri grupları oluşumu için tahmin edici modellerde de verilerin ön işleme aşaması olarak da kullanmaktadır.

Diğer taraftan, denetimsiz makine öğreniminde birbirine yakın değerlerin gruplandırılması için de kümeleme analizi yapılmaktadır.

Birliktelik kuralları analizi ise, veri kümesindeki yaygın örüntüler ve nesneleri oluşturan nitelikler arasında ilişki bulunma durumudur. Perakende satışlarında müşterilerin satın alma eğilimlerinin belirlenmesinde birliktelik analizleri kullanılmaktadır. Birliktelik analizi, istatistiksel olarak değişkenler arasındaki korelasyonun araştırılması anlamına da gelmektedir. Birliktelik kurallarının üretilmesinde kullanılan en yaygın uygulama apriori algoritmasıdır (Witten ve Frank, 2005). Apriori algoritması, Agrawal ve diğerleri tarafından 1994 yılında geliştirilmiştir. Ayrıca, birliktelik kurallarında en çok kullanılan algoritma olmuştur (Han, Jiawei ve Kamber, 2006). Veri madenciliği çalışmalarında büyük miktardaki veri tabanlarında apriori algoritması kullanılmaktadır. Birliktelik kurallarında müşterilerin satın aldığı tüm ürünlerden yola çıkarak satın alma eğilimlerini ortaya koyan uygulamaya "pazar sepet çözümlemesi" adı verilmektedir. Sepetlerin tümü incelendiğinde alışveriş yapan kişilerin aldıkları ürünlerden yola çıkılarak, daha sonra alışveriş yapan kişilerin sepetindeki ürünlerin neler olabileceği konusunda tahmin yürütülebilir, ürün stok durumu tahminlemesi ile oluşabilecek zarar en aza indirgenebilir ve buna uygun satış politikaları geliştirilebilir.

Yukarıda makine öğreniminde kullanılan modellere ilişkin gerekli açıklamalar verilmiştir. Makine öğreniminin tarihsel geçmişine bakılacak olursa, yapay zeka arayışı ile ortaya çıkan bir kavram olduğu görülmektedir. Yapay zeka, insan zekasının bilgisayar algoritmalarına aktarımıyla makinelerin karmaşık problemlere çözüm üretmesini sağlamaktadır. Yapay zeka, insan beyni gibi çalışmakta ve özellikle makine öğreniminde kullanılan modeller için kullanılmaktadır.

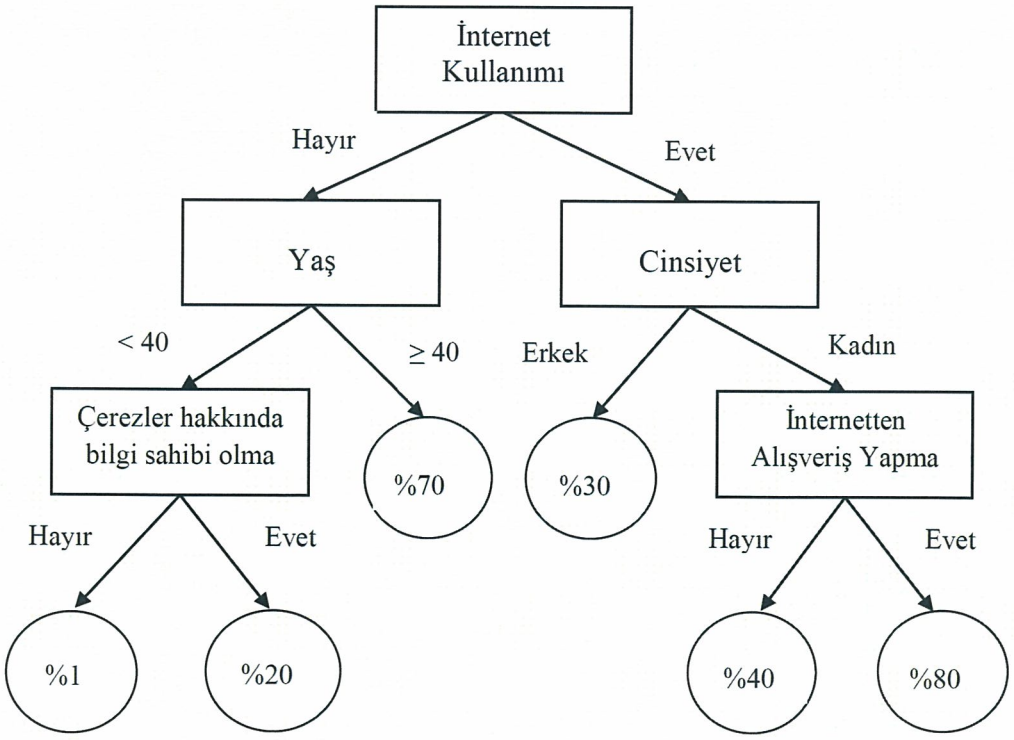
5. SINIFLANDIRMA VE REGRESYON MODELLERİ

5.1. Makine Öğreniminde Karar Ağaçları

Makine öğreniminde karar ağaçları, hem sınıflandırma hem de tahminleme için kullanılan bir yöntemdir. Karar ağaçlarının kolay yorumlanması ve anlaşılabilirlilik açısından ve karar vericiler için avantaj sağlaması bakımından yapay sinir ağları gibi birçok yönteme rağmen daha çok tercih edilmektedir (Chien ve Chen, 2008).

Diğer taraftan, karar ağaçlarının basit yapılı olması, oluşan sınıflandırma modelinin anlaşılabilir ve kolay olması, bilgi keşfine uygun bir yapı sunması, diğer karar ağaçlarına nazaran daha kısa sürede oluşturulması ve karar ağaçlarının parametrik olmayan yöntem olması sebebiyle karar ağaçlarında sınıflandırma teknikleri daha çok tercih edilmektedir (Gehrke, 2003).

Karar ağacı analizlerinde karar vermeyi görsel olarak ve açık bir şekilde temsil etmek için Şekil 1'de görüldü gibi bir karar ağacı kullanılabilmektedir.



Şekil 1. Karar Ağacı Yapısı

Şekil 1'deki ağaçta görüldüğü gibi, değişkenler ağacın dallara bölündüğü koşulları yani iç düğümleri temsil etmektedir. Ayrılmayan dallar ise, karara dahil edilmeyen ölü dalları göstermektedir.

Bir karar ağacının oluşumunda hangi özelliklerin seçileceğine ve hangi bölüme ayrılacağına karar vermek ve ağacın nerede bitirileceğinin bilinmesi de gerekmektedir. Eğer bir ağaç kontrolsüz bir şekilde oluşturulursa, ağacın anlaşılmasını kolaylaştırmak için ağaca bölme işleminin yapılması gerekmektedir. Bu işlemler yapılırken, ilk olarak tüm özellikler dikkate alınır ve tüm düğüm noktalarının maliyeti test edilir, en düşük maliyetle en iyi bölünmenin gerçekleşmesi gerekmektedir. Maliyeti düşüren kök düğümü en iyi sınıflandırıcı özelliğini taşır. Karar ağaçları hem kategorik hem de sürekli verilerin sınıflandırmasında rol oynamaktadır.

Veri setlerinden kolaylıkla karar ağacı oluşturmak amaçlı birçok karar ağacı algoritması geliştirilmiştir. Bu algoritmalar genellikle hatayı en aza indirgemekle birlikte optimum karar ağacı yapısının oluşumunu amaçlamaktadır. Karar ağacı algoritmalarıyla birlikte oluşturulan büyük ağaçlar, optimum ağaç özelliğini taşımamaktadır ve genelleştirme bakımından düşük başarıya sahiptir. Karar ağacı oluştururken kullanılan yaklaşımlardan biri düğüm ayırmada kullanılan ölçütlerdir. Bilgi kazancı, ki-kare istatistiği ve gini indeksi başlıca kullanılan düğüm ayırma ölçütlerindendir (Kothari ve Dong, 2001). Diğer taraftan, karar ağacı algoritmalarında önemli olan diğer bir konu da budama yöntemidir. Budama yönteminde amaç, düşük istatistiksel geçerliliğe sahip ağaç dallarını ortadan kaldırmak ve optimum ağaç yapısının oluşumu sağlamaktır. Budama yöntemleri, maliyet-karmaşıklık budama ve azaltılmış hata budama şeklinde iki şekilde gerçekleşmektedir. Maliyet- karmaşıklık budama yöntemi de kendi içerisinde ikiye ayrılmaktadır. Maliyet-karmaşıklık budama yönteminde ilk olarak eğitim verisi üzerinde kök, ağacı temsil edecek şekilde ağaçlar oluşturulmaktadır. İkinci aşamada ise bu ağaçlar içerisinden genelleme hata tahminine dayalı bir ağaç da budanmış ağaç olarak seçilmektedir. Azaltılmış hata budama yönteminde ağacın düğümlerinden aşağıdan yukarı doğru gezinmesiyle, her bir iç düğümün çoğunlukla görülen sınıf ile yer değişimi sonrası ağacın doğruluğunu azaltıp azaltmadığını kontrol edilir ve buna uygun olarak düğümleri budanır. Bu yöntemde karamsar budama, en düşük hata budaması ve en uygun budama gibi yöntemler de bulunmaktadır. Değerlendirme

yapılırken maliyet-karmaşıklık budama yöntemleri ve azaltılmış hata budama yönteminden bahsedilme nedeni, gereğinden fazla budama yapılması ve doğruluk oranının düşük olmasıdır. Bunun yanı sıra karamsar budama, en düşük hata budaması ve en uygun budama yöntemlerinin az budama yaptıkları gözlemlenmektedir (Kotsiantis, 2013). Bununla birlikte tüm bu koşulların incelenmesi sonucu kabul görmüş optimum ağacın oluşumunu sağlayan budama yönteminin varlığı gözlemlenmemiştir.

5.1.1. Karar Ağacı Algoritmaları

Karar ağacı algoritmaları diğer sınıflandırma algoritmalarına kıyasla çok daha kolaydır. Bu nedenle karar ağacı algoritmaları; ID3, C4.5, C5.0, CART, CHAID, SLIQ, SPRINT, MARS ve QUEST algoritmaları olmak üzere 9 şekilde incelenmektedir.

5.1.1.1. ID3 Algoritması

ID3 algoritması, entropi ve bilgi kazanımı üzerine kurulmuştur. 1986 yılında J. R. Quinlan tarafından kategorik değişkenler için geliştirilen ID3 (Iterative Dichotomiser 3) algoritması yaygın kullanılan ve en temel sınıflandırma algoritmalarından biridir (Altunkaynak, 2017). ID3 algoritması, C4.5 algoritmasının da temelini oluşturmaktadır. Örneklerin tamamı homojen olduğunda entropi ölçümü sıfır, örnek değerleri birbirine eşit olması durumunda entropi ölçümü 1 olmaktadır.

Entropi ölçüm formülü;

$$E(C \mid A_k) = \sum_{j=1}^{Mk} p(a_k, j) x \left[- \sum_{i=1}^{N} p(c_i \mid a_k, j) log_2 p(c_i \mid a_k, j) \right] \quad (4.1)$$

Bu formülde;

$E(C \mid A_k)$ = A_k alanı Sınıflandırma özelliğinin entropi ölçüsü,

$p(a_k,j)$ = a_k alanının j değerinde olma olasılığı,

M_k = a_k alanının değerleri sayısı; j=1, 2,…, M_k,

N = farklı sınıf sayıları; i=1, 2, …, N,

K = alanların sayısı; k=1, 2, …, K.

ID3 algoritmasında kazanım değeri yüksek olan bağımsız değişken bağımlı değişken üzerinde en yüksek belirleyiciliğe sahip olarak adlandırılır ve dallandırma için bu değişken seçilir. Karar ağacı düğümleri içerisinden en baskın olanı karar ağacı düğümü olarak ağaca konulmaktadır. Bunun üzerine bilgi kazanımı diğer özellikler bakımından tekrar hesaplanmaktadır.

Bilgi kazanımı;
Kazanç(A)=I(S)-E(A) görüldüğü gibi hesaplanmaktadır. (4.2)

Bilindiği gibi, makine öğrenim algoritmalarının ortak noktası nominal değerlerin kullanılamamasıdır. Ancak, ID3 algoritmasında sayısal değişkenler nominal değere dönüştürülmektedir. Bu nedenle, eğitim verilerinin yapıya uyum göstermesi kısa bir zaman almaktadır.

5.1.1.2. C4.5 Algoritması

C4.5 algoritması, ID3 algoritmasından türeyen bir algoritma olmakla birlikte en çok bilinen karar ağacı algoritmalarından biridir. C4.5 algoritmasını ID3 algoritmasından ayıran özellik normalleştirmenin kullanılıyor olmasıdır. ID3 algoritması ile oluşan ağaçta entropi ölçüm hesabı yapılarak belirlenen karar kurallarına göre düğüm noktaları belirlenirken C4.5 algoritmasında entropi ölçümleri oran olarak verilmektedir. Bununla birlikte ID3 algoritmasında dallandırma yapıldıktan sonra optimum ağaç için budama yapılmazken, C4.5 algoritması ile oluşan ağaçta budama işlemi yapılmaktadır. C4.5 algoritmasında kayıp veriler hesaba katılmazken, ID3 algoritmasında böyle bir durum söz konusu değildir. Diğer taraftan, C4.5 algoritması daha anlamlı ve duyarlı kuralların oluşumunu sağlayarak ağaç ürettiği için kullanımı tercih edilen algoritmaların başında gelmektedir.

C4.5 algoritması, karar ağacı şeklinde bir sınıflandırıcı oluşturmaktadır. Bunun için de C4.5'te önceden değişkenlerin sınıflandırıldığı bir veri seti kullanılmaktadır.

C4.5 algoritmasında, eğitim veri seti kullanılmaktadır. Eğitim veri seti, sınıflarla etkilendiği için denetimli makine öğrenimi algoritmasıdır.

C4.5 algoritması da diğer karar ağacı algoritmaları gibi yorumlama ve açıklama bakımından kolaylık sağlamaktadır.

5.1.1.3. C5.0 Algoritması

1970'li yılların sonlarında J. Ross Quinlan, ağaç temelli modeller geliştirmekle uğraşırken 1980'lerde bu yöntemler sınıflandırma ağacı modellerine dönüşmüştür. 1993 yılında Quinlan tarafından C4.5 algoritması geliştirilmiştir. 1994 yılında, Sydney Üniversitesi'nde J. Ross Quinlan tarafından geliştirilen C5.0 algoritması, C4.5 ve ID3 algoritmasına dayanmaktadır (Quinlan, 1986). C4.5 ve C5.0, ID3 karar ağacı algoritmasının ileri versiyonları niteliğindedir (Li vd., 2009). Bu sebeple, C5.0 algoritması daha fazla tercih edilmektedir.

C5.0 pek çok alanda kullanılmaya başlayan sınıflandırma tekniklerindendir. Bu teknik, karar ağaçlarının anlaşılma ve yorumlanma bakımından kolaylık sağlamaktadır. C5.0 algoritmasında, ikili ağaçlar oluşmakta ve budama sonrası optimum ağaç oluşmaktadır.

C5.0 algoritmalarında bağımlı değişken kategorik (nominal/ordinal) yapıdadır. Bağımsız değişkenler ise kategorik veya sürekli olabilmektedir. Bağımsız değişkenler ile ilgili herhangi bir sınırlama bulunmamaktadır.

C5.0 algoritması, parametrik olmayan bir yöntemdir. Bu nedenle; normallik, doğrusallık gibi parametrik yöntemlere özgü varsayımlara ihtiyaç duyulmamaktadır.

C5.0 algoritmasında her düğümden çoklu dallar oluşmaktadır. Dalların sayısı, bağımlı tahmin edicinin kategori sayısına bağlı olarak değişmektedir. Ayırma kriteri olarak bilgi kazancı (information gain) kullanılmaktadır. Budama işlemi her yapraktaki hata oranına bağlıdır (Bounsaythip vd., 2001).

C5.0 algoritması genellikle büyük veri setlerinde kullanılmaktadır. C5.0 algoritması, doğruluğu artırmak için boosting algoritmasını da kullanmaktadır. Bu nedenle diğer bir adına boosting ağacı da denilmektedir. C5.0 algoritması, C4.5 algoritmasından daha gelişmiş olduğu için biçim bakımından daha düzgün karar ağaçları elde etmemizi sağlamaktadır (Çalış vd., 2014).

C5.0 algoritması, bellek tabanlı bir algoritmadır ve çalışmaların verimli olması açısından C4.5 algoritmasından daha iyidir. Aynı zamanda, karar ağaçlarını en aza indirerek anlaşılır olmakta ve karar kuralları üretimi açısından da daha iyi sonuçlar vermektedir (Shahnaz, 2006).

C5.0 algoritması düğümlerde ayırım kriteri olarak entropi ölçümü ve bilgi kazancını kullanmaktadırlar. Ayırma kriteri, eğitim verisiyle ağaç düğümlerinin bölünmesini sağlamaktadır. Entropi ölçüm sonuçlarının düşük olması beklenmektedir. Entropi ölçüsü düşük olan alanlar, doğru sınıflandırma oranı en yüksek olan alanlardır. Sınıfların olasılıklarının dengeli olması, entropinin yüksek olmasını sağlayacaktır. Sınıf olasılıkları arasındaki farklar, bölünmeden önceki değerlere göre önemli olduğu durumlarda yüksek bilgi kazancı meydana gelecektir.

X değişkeni için k adet olasılıklar sırasıyla p_1, p_2, p_3, ... , p_k olarak adlandırılmaktadır. Entropi ölçümü ise aşağıdaki şekilde yapılmaktadır (Quinlan, 1993).

$$Entropi = H(T) = -\sum_{j=1}^{k} p_j \log_2 (p_j) \quad (4.3)$$

Eğitim veri setinde yer alan X değişkenine bağlı olarak T alt kümeleri de T_1, T_2, T_3, ..., T_k olarak ayrılmaktadır. Her bir T için sınıfın belirlenmesi gerekmektedir. Sınıfların belirlenmesi için gereken bilgilerin ağırlıklı ortalaması entropilerin ağırlıklandırılmış toplamı şeklinde hesaplanmaktadır. Gerekli bilgilerin ağırlıklı ortalaması ve bilgi kazancı aşağıdaki gibi hesaplanmaktadır.

Ağırlıklı ortalama;

$$H_S(T) = \sum_{i=1}^{k} p_i H_S (T_i) \quad (4.4)$$

Bilgi kazancı;

$$Bilgi\ Kazancı\ (S) = H(T) - H_S(T) \quad (4.5)$$

olarak hesaplanmaktadır. Kısaca, bilgi kazancı bölünme öncesi ve bölünme sonrasındaki entropi ölçüm farkları anlamına da gelmektedir.

C5.0 algoritmasında budama oldukça etkili ve verimlidir. Bu algoritmalar, budama bakımından teorik olarak temeli sağlam olmamasına karşın iyi sonuçlar verebilmektedir. Bu algoritmalarda çapraz doğrulama tekniği kullanılmaktadır. Çapraz doğrulama, eğitim verisiyle kurulan modelin test verisiyle doğruluğunun ve güvenilirliğinin incelenmesidir.

C5.0 algoritması, budama tekniği olarak binom güven sınırı yöntemini kullanmaktadır. Çünkü, binom güven sınırı yöntemi eksik değerlerin ele alınması durumunda bu değerlerin tahmin edilip tahmin edilemeyeceğine izin vermektedir.

C5.0 algoritmasında birçok zayıf sınıflandırıcı bulunabilmektedir. Bu zayıf sınıflandırıcıların güçlü bir sınıflandırıcıda birleştirilmesine boosting (artırma) denilmektedir. Artırma, ön yargının ve varyansın azalması için kullanılmaktadır. Boosting algoritması, 1990'ların başında geliştirilen AdaBoost algoritması ile benzerlik göstermektedir (Kuhn ve Johnson, 2013).

C5.0 algoritmasında, ağaç ve kural tabanlı modellerde kategorik değişken verilerini işlemenin iki seçeneği vardır. Bu iki seçenek de C5.0 algoritmasında mevcuttur. Değişkenler, ya gruplandırılmış şekilde ya da bağımsız olarak işlenmektedir. Gruplandırılmış kategorik değişkenlerde, her bir kategorik değişken tek olarak girer ve bu şekilde modelin nasıl bölüneceği kararlaştırılır. Bağımsız kategorik değişkenler de değişkeni iki kukla değişkene ayrıştırır ve her biri bağımsız olarak kabul edilir. Değişkenlerin yalnızca bir kısmı yüksek tahmin niteliği taşıdığında, genellikle gruplandırılmış seçenek yaklaşımı kullanılmaktadır. Bu iki yaklaşımın da avantajları bulunmaktadır. Öncelikle veriler ile iki şekilde de model kurulur ve daha sonra sonuçlar karşılaştırılır. Ancak bu şekilde, dezavantajlı olan yaklaşım veriye uygunluk açısından değerlendirilebilmektedir.

5.1.1.4. Sınıflandırma ve Regresyon Ağacı (CART) Algoritması

Sınıflandırma ve regresyon ağaçları, veri madenciliğinin en önemli konuları arasında yer almaktadır.

İlk olarak 1984 yılında Breiman ve arkadaşları tarafından sınıflandırma ve regresyon ağacı ile ilgili çalışmalara başlanmıştır. Sınıflandırma ve regresyon ağacı algoritmasında esas olan, bağımsız değişkenlerin birbirleriyle ve bağımlı değişken ya da değişkenlerle olan ilişkilerini, ağaç şeklinde bir modelde incelemektir. Sınıflandırma ve regresyon ağacı, verileri alt gruplara ayırmakla birlikte bütün bağımsız değişkenleri kullanmaktadır.

Sınıflandırma ve regresyon ağacı, parametrik olmayan bir yöntemdir. Bağımlı değişkenin sürekli olduğu durumda regresyon ağacı, bağımlı değişken kategorik olduğu durumda sınıflandırma ağacı kullanılmaktadır (Chang ve Wang, 2006).

Bağımlı değişkenin kategorik (ikili; 0-1) olduğu durumlarda lojistik regresyon analizi de sınıflandırma ve regresyon ağacı analizine dahil olmaktadır. Sınıflandırma ve regresyon ağacında hem kategorik değişkenleri hem de sürekli değişkenleri modellemek mümkün olmaktadır.

Sınıflandırma ve regresyon ağacında, bağımsız değişken ya da değişkenler arasındaki ilişkiyi inceleyen basit ve çoklu regresyon analizindeki gibi homojenlik, doğrusallık, normallik gibi varsayımlara gerek duyulmamaktadır.

Sınıflandırma ve regresyon analizinde oldukça güçlü bir algoritma kullanılmaktadır. Bu algoritma, bağımsız değişkenlerin bağımlı değişkenle olan etkisini incelemekle kalmayıp aynı zamanda model içerisindeki genel etkileşimi de incelemektedir. Bu durumda sınıflandırma ve regresyon ağacında doğrusal olmayan ilişkilerin açıklanması da mümkün olmaktadır.

Sınıflandırma ve regresyon ağacı, literatür incelemesinde birçok şekilde ifade edilmektedir. CART, SRAT ya da SRT şeklinde ifadeler de sınıflandırma ve regresyon ağacı anlamına gelmektedir.

Sınıflandırma ve regresyon ağaçlarının en tepesinde, bağımlı değişken yer almaktadır. Ağacın yapısında bağımlı değişkene kök düğümü adı verilmektedir. Öncelikle kök düğüm, ağaç yapısının oluşabilmesi için iki dala ayrılmaktadır. Bu dallara, ebeveyn dalı denilmektedir. Kök düğümü etkileyen ebeveyn dalları alt kümelere ayrıldıklarında yavru düğüm olarak adlandırılmaktadır. Ancak ebeveyn dallarını etkileyen alt düğümler alt kümeler oluşturmuyorsa bu durumda alt düğümlere terminal düğüm adı verilmektedir.

Her bir parçalanmada oluşan düğümler, alt kümedir (Keskin vd., 2007). Ağacın oluşturulmasında temel amaç, yavru düğümde homojenliğin mümkün olduğunca sağlanmasıdır. Bu durumda amaç, modele alınan tüm bağımsız değişkenleri test ederek, yeni oluşacak her düğümde en yüksek homojenliği sağlamak ve açıklayıcı değişkenin kesim değerini belirlemektir (Keskin vd., 2007).

CART algoritmasında ağaç yapısının oluşması üç temel unsurdan meydana gelmektedir. Bunlar, "ağacın oluşturulması", "budama", "en uygun ağaç yapısının seçimi" şeklindedir.

Sınıflandırma ve regresyon ağacı algoritması, maksimum düzeyde alt sınıfların oluşması esasına dayanmaktadır. Ağacın alt sınıfları ile bağımlı değişken arasında önemli ilişkilerin olması beklenmektedir.

Sınıflandırma ve regresyon ağaçlarının oluşabilmesi için öncelikle ağacın büyümesi, daha sonra da optimal ağacın oluşabilmesi için de ağacın budanması gerekmektedir (Küçükoğlu, 2010). Ağaçta oluşan fakat sonucu etkilemeyen ve sınıflandırmada katkısı olmayan dalların ağaçtan alınması işlemine budama işlemi denilmektedir. Sınıflandırma ve regresyon ağacında budama işleminin uygulanmasının amacı, ağacın oluşumunda en başından itibaren modele dahil edilen değişkenlerin ağacın büyümesiyle tekrar modele dahil olmasını engellemektir. Budama sürecine, en az katkı sağlayan düğümden başlanmaktadır. Budama işleminin amacı, ağaca önemli katkı sağlayan düğümlerin kalmasını sağlamaktır.

Sınıflandırma ve regresyon ağacı gibi ağaç yapılarında ikili dallanmalar mevcuttur. Kısaca, her düğümden sadece iki dallanma oluşmaktadır. Her dallanmada yeni bölünmelerin oluşabilmesi için her düğüme belli bir kriter uygulanmaktadır. Bunun için öncelikle değişkenlerin sahip olduğu nitelikler göz önüne alınmaktadır (Özkan, 2013).

Sınıflandırma ve regresyon ağacı analizinde ağaç yapısı, bağımlı değişkeni en fazla etkileyen bağımsız değişkenlerin yukarıdan aşağıya doğru ikili dallanmalar oluşturulması şeklinde meydana gelmektedir. Bağımlı değişkeni etkileyen bağımsız değişkenlerin etkileri ilerleme katsayısı ile belirlenmektedir. Katsayı, ağaç yapısına uygun olarak yukarıdan aşağı inildikçe küçülmektedir. İlerleme katsayısının alt ve üst sınırı bulunmamaktadır.

Sınıflandırma ve regresyon ağacı analizi uygulanması sırasında farklı iki yol izlenmektedir. Bunlar, CART ve CHAID analizleridir.

5.1.1.5. CHAID Algoritması

CHAID algoritması, cevap ağacı yöntemine bağlı olarak kullanılan yöntemlerden biridir. 1980 yılında G.V. Kass tarafından bağımlı değişkeni en iyi açıklayacak şekilde geliştirilen bir algoritmadır.

CHAID algoritması, daha büyük veri setlerinin analizi için oldukça uygun olan basit bir algoritmaya dayanmaktadır. Bu algoritma, aynı zamanda birden çok kategorik değişkeni sınıflandırmaktadır. Bu nedenle pazar araştırmasında oldukça kullanılan bir yöntemdir.

Hem CHAID hem de CART algoritmalarında, her bir düğümün en iyi tahmini (sürekli bağımlı ve bağımsız değişkenlerinin) veya sınıflandırılmasını (kategorik bağımlı ve bağımsız değişkenleri için) elde etmek için bölünmüş bir durumu tanımlayan ağaçlar oluşacaktır.

Bağımlı değişkenin nominal, ordinal kategorik veya sürekli ya da kategorik; bağımsız değişkenlerin de kategorik veya nominal kategorik ve sürekli olduğu durumlarda CHAID (Chi-Squared Automatic Interaction Detection) algoritması kullanılmaktadır (SPSS, 1998).

CHAID algoritmasında, model kurma aşamasında bağımlı ve bağımsız değişkenlerin veri yapıları konusunda herhangi bir sınırlama getirilmemektedir. Kısaca, CHAID algoritmasında değişkenlerin yapısı bakımından bir sınama yoktur. Bağımsız değişkenlerin aynı tip ölçeğe uygun olması gibi bir zorunluluk da bulunmamaktadır (Satıcı vd., 2009). Bu yönüyle CHAID analizi önemli bir avantaja sahiptir (Ratner, 2000).

Ağacın oluşumunda ilk adım, sürekli değişkenleri yaklaşık olarak eşit sayıda gözlem ile birkaç kategoriye ayırarak, herhangi bir sürekli değişkenden kategorik değişken oluşturmaktır.

CHAID analizinde, regresyon analizinin normallik, doğrusallık, homojenlik ve toplanabilirlik varsayımları sınanmamaktadır. Çünkü CHAID analizinde güçlü bir öteleme algoritması vardır ve bu şekilde kararlı alt düğümlere bölünme kolaylıkla sağlanmaktadır.

CHAID analizinin kullanımını artıran sebepler arasında sürekli ve kategorik değişkenlerin aynı anda modele alınabiliyor olmasıdır. CHAID algoritması, yapısı gereği parametrik olmayan bir yöntem gibi gözükse de algoritma veriyi kararlı alt düğümlere böldüğü için, aynı zamanda homojenlik ve normallik varsayımlarını da sağlamaktadır. Bu sebeple CHAID analizine yarı parametrik yöntem demek daha doğru olacaktır. CHAID analizini diğer karar ağacı algoritmalarından ayıran farklılık ağacın türemesidir. Diğer karar ağacı algoritmaları ikili ağaç türetirken, CHAID analizi çoklu ağaçlar türetmektedir (Üngüren ve Doğan, 2010.)

CHAID analizi, bağımsız değişkenlerin birbirleriyle olan etkileşim ve ilişkilerini ki-kare test istatistiği ile incelemektedir. Bu bağlamda, CHAID analizinin ki-kare istatistiği tabanına bağlı modeller kurması analizin tutarlı olduğunu göstermektedir. Sonuç olarak da, ilişki tabanlı modeller kurulması test istatistiklerinin doğru ve yansız olmasını sağlamaktadır. Ki-kare test istatistiğinin kullanım nedeni, en iyi tahmin sonucunun elde edilebilmesi için başlangıç değişkenlerin bağımsız olarak kategorileştirilmesidir. Bu kategorileştirme işlemi istatistiksel olarak anlamlı değişkenler bulunamayana kadar devam eder (Kuşakcı, 2010). Dikkat edilmesi gereken nokta, bağımlı değişkenin kategorik olduğu durumlarda en iyi bölünmeyi belirlemek için ki-kare istatistiğinin kullanılmasıdır. Kategoriler anlamlı bir şekilde birleştikten sonra bağımlı değişkene uygun kontenjans tabloları oluşur ve Benferroni düzeltilmiş p değerleri ile ki-kare istatistikleri hesaplanır. Bağımlı değişkenin sürekli olduğu durumda F testi kullanılmaktadır. CHAID analizinde bağımlı değişkenin türü bakımından en iyi bölünme için kullanılan test istatistiğinin değişmesi, sınıflandırma ve regresyon ağacında yöntemin bağımlı değişkenin türüne göre değişmesine benzemektedir.

Bağımlı değişkenin kategorik (0-1; ikili) olduğu durumlarda lojistik regresyon yöntemi kullanılmaktadır ve modele ilişkin risk faktörlerinin tahminlemesi yapılabilmektedir (Hair vd., 1998).

CHAID analizi risk faktörlerine ilişkin tahminlemelerin yapılması bakımından daha güçlü bir yöntemdir. Bunun nedeni yapısal olarak homojenlik varsayımını algoritmasında taşıması ve yarı parametrik bir algoritma olması sebebiyle ağacın en üst düğümünde oluşan etkileşimi alt düğümlere de homojen olarak taşımasıdır.

CHAID analizinde tahminlerin belli bir karar kuralı bağlamında oluşması, oluşan regresyon denklemine ait parametrelerin yansız ve güvenilir sonuçlar vermesi olasıdır. Bu analizde, bağımlı ve bağımsız değişkenler arasındaki ilişkiler daha ayrıntılı değerlendirilebilmekte ve değişkenlerin birbirleri üzerindeki etkiler kolaylıkla yorumlanabilmektedir (Üngüren ve Doğan, 2010).

5.1.1.6. SLIQ Algoritması

Hem sayısal hem de kategorik değişken türlerinin sınıflandırması ile kullanılan bir karar ağacı yöntemidir.

SLIQ algoritmasında, sayısal değişken türlerinin değerlendirilmesinde maliyeti azaltmak için ağaç oluşumu aşamasında, önceden sınıflandırma tekniği kullanılmaktadır. Bu nedenle en iyi dallanma kriteri, verileri sıraya dizmektir. Sürekli değer taşıyan tablolar, sürekli değişkene göre kategorik veri taşıyan tablolar da sıra numarasına göre dizilir. Bu nedenle ağaç, budamayı da hızlı şekilde gerçekleştiren bir algoritmaya sahiptir.

SLIQ algoritmasında eğitim verileri kullanılmaktadır.

Karar ağacı algoritmaları genellikle derinlik ilkesine göre hareket etmektedir. Ancak, SLIQ algoritmasında bu durum farklıdır ve öncelikli olarak genişlik ilkesine göre hareket edilmektedir. Ağacın dallara ayrılmasında gini indeksi kullanılmaktadır.

$$gini(K) = 1 - \sum p_j^2 \qquad (4.6)$$

p_j = K kümesi içinde j sınıfının sıklığıdır.

Bu algoritmada verileri en iyi temsil eden model, ağacın oluşumu sağlayan en az maliyetli modeldir. Özellikle büyük verilere karar ağacı uygulanacağı durumlarda kullanılan bir yöntemdir.

5.1.1.7. SPRINT Algoritması

SPRINT algoritması, büyük veriler için oldukça ideal bir algoritmadır. Bu algoritma da SLIQ algoritması gibi, her bir değişken için ayrı liste oluşturarak sıraya dizme işlemini yalnızca bir kez yapar. Dallara ayırma kriteri olarak da SLIQ algoritmasında olduğu gibi gini indeksi kullanılmaktadır.

5.1.1.8. MARS Algoritması (Değişkenli Uyumlu Regresyon Uzanımları)

MARS algoritması, 1991 yılında Standford'da istatistikçi ve fizikçi Jerome Friedman tarafından geliştirilmiştir. MARS algoritmasının temel amacı, diğer regresyon modellerinde olduğu gibi, bağımlı değişkenin değerini bağımsız değişkenler yardımıyla kestirmeye çalışmaktır.

MARS algoritması, sürekli ve kategorik yapıdaki bağımlı değişkenlere uygulanabilmektedir. Bağımsız değişkenlerin sürekli ya da kategorik olması bakımından sınırlama yoktur. Fakat, MARS algoritmasının temeline bakılacak olursa bağımsız değişkenlerin sürekli yapıda olduğu görülecektir. MARS algoritması, bir ya da birden fazla bağımlı değişkene uygulanabilmektedir.

MARS algoritması, varsayımlara dayalı bir karar ağacı algoritması olmaması kaynaklı parametrik olmayan bir yöntemdir. Model, regresyon modeli ile açıklanan hata varyansının oranı hakkında bilgi vermektedir. Regresyon modellerinde amaç, kurulan modelin gerçeği yansıtmasını sağlamaktır. Ancak, gelişen teknolojiyle birlikte karmaşık yapıda kullanılan algoritmalar oluşmaktadır. MARS da karmaşık yapıda bir algoritmadır ve parametrik olmayan bir yöntem olması ve matematiksel bir ilişki aranmaması sebebiyle tercih edilen bir yöntem haline gelmiştir. Fakat, modelin karmaşık yapıda olması, aşırı kestirim değerlerinin oluşumuna neden olabilmektedir. Bu durumda karmaşıklığı gidermek amacıyla budama yöntemi kullanılmaktadır.

MARS algoritmasının stratejisi, parçala ve çözümle şeklindedir. Bu yöntem, veri uzayını önce bölgelere ayırmakta ve her birine regresyon modeli oluşturmaktadır.

MARS, diğer yöntemler için çok boyutlu sorun oluşturabilecek çok değişkenli regresyon problemleri için bir çözümdür (Temel vd., 2005).

Doğru kestirimin yanında modelin yapı özelliklerinin ve değişkenler arası ilişkilerin bütünüyle açıklanması isteniyorsa MARS algoritması yeterli tüm olanağı sağlayacaktır.

5.1.1.9. QUEST Algoritması

En son geliştirilen karar ağacı algoritması olma özelliğini taşıyan QUEST (Quick, unbiased, efficient statistical tree; hızlı, yansız, etkili istatistiksel ağaç), çok sayıda kategoriye sahip ön kestiricileri destekleyen, diğer yöntemlerin yanlılıklarından kaçınılmasını sağlayan ve hızlı hesaplanabilen bir yöntemdir (Loh ve Shih, 1997).

QUEST algoritmasında bağımlı değişkenin nominal, bağımsız değişkenlerin ise ordinal, sürekli veya nominal olduğu durumlarda kullanılan bir karar ağacı yöntemidir (SPSS, 1998). Bağımlı değişken sürekli olduğu zaman QUEST algoritması kullanılamamaktadır. Kısaca, bağımlı ve bağımsız değişkenler bakımından sınırlandırma getirmektedir. İkili karar ağacı yapısı ile sınıflandırma yapan bir algoritma türüdür.

QUEST algoritması, ağacın oluşumu sırasında değişken seçimi ve bölünme noktası ile tek tek ilgilenir. QUEST algoritması, hızlı bir ağaç oluşturma yeteneğine sahip olmasının yanında hesaplama maliyetini de oldukça düşürmektedir.

QUEST algoritmasında sınıflandırma ve regresyon ağacında olduğu gibi ön olasılıklar kullanılmaktadır. Sıralayıcı ve sürekli değişkenler için F testi ya da Levene testi kullanılırken, sınıflayıcı açıklayıcı değişkenler için Pearson ki-kare değeri kullanılmaktadır. Değişken seçiminde küçük p değerine sahip değişkenler seçilmektedir.

Diğer taraftan, ayırma işleminde bağımlı değişkenin ikiden fazla kategoriye sahip olması durumunda iki sınıf bulmak için iki ortalamalı kümeleme algoritması ve açıklayıcı değişkenin en iyi bölünmesini bulmak için de Kuadratik Diskriminant Analizi (QDA) kullanılmaktadır (Pehlivan, 2006).

5.2. Karar Ağacı Algoritmalarının Avantajları ve Dezavantajları

Karar ağacının avantajları ve dezavantajları aşağıda görüldüğü gibi Tablo 1'de özetlenmiştir.

Tablo 1. Karar Ağacı Algoritmaları Avantajları ve Dezavantajları

Avantajlar	Dezavantajlar
1. Karar ağacı oluşturmak fazla zaman almayan bir yöntemdir. Küçük ağaçların oluşturulması ve yorumlanması oldukça kolaydır.	1. Ağaç oluştururken ve ağacın budanmasında algoritmaların yapısından kaynaklı karmaşıklıklar oluşmaktadır.
2. Karar ağaçları sürekli ve kategorik yapıdaki değişkenlere uygulanabilir.	2. Sürekli değişkenlerin tahmini konusunda başarısızdır.
3. Karmaşık modellerin sınıflandırılmasıyla anlaşılır hale gelmesini sağlamaktadır.	3. Parametreler arasındaki doğrusal olmayan ilişkiler ağaç performansını etkilemektedir.
4. Uygun modelin kurulabilmesi için veri yapısının ve sınıf sayısının yeterli düzeyde olması gerekmektedir.	4. Karar ağacı, verileri genelleştirmeyen aşırı karmaşık ağaçlar oluşturabilir.
5. Karar ağaçları, veri hazırlama sürecinde az çaba gerektirmektedir.	5. Karar ağaçları kararsız bir yapı sergileyebilir, çünkü verilerdeki en ufak değişiklikler farklı bir ağacın türetilmesine sebep olabilmektedir.

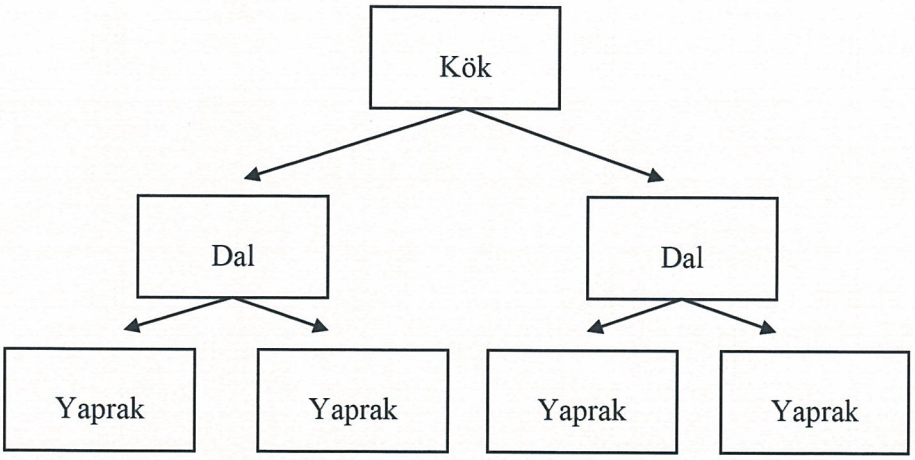
6. SINIFLANDIRMA VE REGRESYON AĞAÇLARI

Çalışmamızın bu bölümünde, sınıflandırma ve regresyon ağaçları teorik olarak anlatılmaya çalışılmıştır.

6.1. Sınıflandırma Ağaçları

Verilerin ortak özelliklerinin kullanımı yardımıyla sınıflandırma yapılması mümkündür. Sınıflandırma bir öğrenme algoritmasına dayanmaktadır. Öğrenmenin amacı, bir sınıflandırma modelinin yaratılmasıdır (Özkan, 2013).

Karar ağaçlarında sınıflandırma, akış şemalarına benzeyen yapılardır. Dallar ve yapraklar ağaç yapısının elemanlarıdır. Ağaçta, en son yapı yaprak adını, en üst yapı da kök adını almaktadır. Kök ile yapraklar arasında kalan alan ise dal olarak adlandırılmaktadır. Şekil 2'de kök, dallar ve yapraklar görülmektedir.



Şekil 2. Ağacın Kökü, Dalları ve Yaprakları

Bağımlı değişkenin kategorik olması (evet-hayır, var-yok) durumunda sınıflandırma ağacı yöntemi kullanılmaktadır (Chang ve Wang, 2006). Bu nedenle, sınıflandırma ağacı dalların ikili ayrılması esasına dayanmaktadır. Sınıflandırma ağacı yöntemlerinde yapılacak olan parametre tahminleri ve değişkenler arasındaki ilişkilerin ortaya çıkarılması, sağlam bir zemine dayanmaktadır (Kayri ve Gökdaş, 2006).

Sınıflandırma ve regresyon ağacının kullanmakta olduğu algoritma, bağımsız değişkenlerin bağımlı değişkenle olan ilişkisini incelemenin yanı sıra model içerisindeki etkileşimi de çözümlemektedir. Sınıflandırma ve regresyon ağacına sahip algoritma, 3 unsurdan meydana gelmektedir. Bunlar sırasıyla, "ağacın oluşturulması", "budama (prunning)" ve "en uygun ağaç yapısının seçimi" şeklindedir. Ağacın oluşumu aşamasında en fazla sayıda alt ağaçların oluşumu gözlemlenmektedir. Ancak, alt ağaçlarda bağımlı değişkenle önemli derecede ilişkiler içeren ağaçların seçilmesi gerekmektedir. Bu nedenle algoritmada "budama" modülü devreye girmektedir. "En uygun ağaç yapısının seçimi" modülü ile de sınıflandırma ağacı elde edilebilmektedir (Kayri ve Boysan, 2008).

Sınıflandırma ağacında kullanılan ayırma kriterleri yardımıyla her bir düğüm noktası ikiye ayrılarak büyümektedir ve her düğümde grup içi homojenlik söz konusu olduğunda, ağacın düzey sayısı ile ilgili analizi yürüten kişi tarafından sınırlama yapıldıysa ve yeni oluşan düğümlerde farklılıkların oluşmadığı gözlemleniyorsa, ağacın büyümesi durdurulmaktadır (Temel, 2004).

6.1.1. Gini Ayırma Kriteri

Gini algoritması, sınıflandırma ağaçlarında genellikle ikili bölünme şeklinde gerçekleşmektedir. Bu nedenle en çok kullanılan sınıflandırma ağacı algoritmalarındandır. Sınıflandırma ağacının kök ve yapraklarının sağ ve sol olmak üzere ikili bölünmeler şeklinde gruplandırılmasıyla hesaplamalar yapılmaktadır.

Sağ ve sol bölünmeler için $Gini_{sol}$ ve $Gini_{sağ}$ formülleri aşağıda görüldüğü gibidir.

L_i : Sol daldaki i grubunda örnek ya da örneklerin sayısı,
R_i : Sağ daldaki i grubunda örnek ya da örneklerin sayısı,

k : Sınıfların sayısı,
T : Düğümdeki örnekler,

$|T_{sol}|$: Sol daldaki örnek ya da örneklerin sayısı,
$|T_{sağ}|$: Sağ daldaki örnek ya da örneklerin sayısı.

$$\text{Gini}_{sol} = 1 - \sum_{i=1}^{k} \left(\frac{Li}{|T_{sol}|}\right)^2 \quad (5.1)$$

$$\text{Gini}_{sağ} = 1 - \sum_{i=1}^{k} \left(\frac{Ri}{|T_{sağ}|}\right)^2 \quad (5.2)$$

Her bir j sınıfı için gini indeksi hesaplanması olarak yapılmaktadır.

$$\text{Gini}_j = \frac{1}{n}\left\langle |T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ} \right\rangle \quad (5.3)$$

Gini algoritması, kayıp verinin çok olduğu büyük verilerde olumlu sonuçlar vermektedir.

6.1.2. Twoing Ayırma Kriteri

Twoing algoritması Gini'den oldukça farklıdır ve iki adımdan oluşmaktadır.

1. Adım;

- Niteliklerin içerdiği değerler göz önüne alınarak eğitim kümesi iki ayrı dala ayrılmaktadır. Bu duruma aday bölünme denilmektedir. Bir t düğümünde "sağ" ve "sol" olmak üzere iki farklı dal bulunmaktadır. Bu bölünen kümeler t_{sol} ve $t_{sağ}$ biçimindedir.
- Aday bölünmelerin her biri için P_{Sol} ve $P(j/t_{Sol})$ olasılıkları hesaplanmaktadır. Burada $P(j/t_{sol})$ ifadesi her bir j sınıf değerinin sol taraftaki bölünmede olma olasılığını vermektedir. j değerleri, sınıf değerlerinin yer aldığı nitelik olarak göz önüne alınmaktadır.

$$P_{Sol} = \frac{t_{sol}\text{daki her bir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}} \quad (5.4)$$

$$P_{(j/tsol)} = \frac{t_{sol}\text{ daki kayıtların j sınıfları sayısı}}{t_{sol}\text{ daki her bir nitelik değerinin nitelik sütunundaki tekrar sayısı}} \quad (5.5)$$

- Aday bölünmelerin her biri için $P_{sağ}$ ve $P(j/t_{Sağ})$ olasılıkları hesaplanmaktadır. Burada $P(j/t_{Sağ})$ ifadesi bir j sınıf değerinin sağ taraftaki bölünme olma olasılığını verir.

$$P_{Sağ} = \frac{t_{sağ}\text{daki her bir nitelik değerininilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}} \quad (5.6)$$

$$P_{(j/tsağ)} = \frac{t_{sağ}\text{ daki kayıtların j sınıfları sayısı}}{t_{sağ}\text{ daki her bir nitelik değerinin nitelik sütunundaki tekrar sayısı}} \quad (5.7)$$

- $\phi(s/t)$, t düğümündeki s aday bölünmelerinin uygunluk (goodness) ölçüsü olsun. Uygunluk ölçüsünün hesaplanması aşağıdaki gibidir.

$$\phi(s/t) = 2P_{sol}\,P_{sağ}\sum_{j=1}^{n}\left|(P(j/t_{sol}) - P(j/t_{sağ})\right| \quad (5.8)$$

- $\phi(s/t)$ değerleri hesaplandıktan sonra içlerinde en büyük olanı seçilmektedir. Bu değerin ilgili olduğu aday bölünme satırı dallanmanın yapılacağı satırı bildirmektedir.
- Dallanma bu şekilde yapıldıktan sonra, bu adıma ilişkin olarak karar ağacı çizilir.

2. Adım;

Algoritmanın en başındaki adıma dönülerek ağacın alt kümesine aynı işlemler uygulanmaktadır (Uysal, Bilen ve Ulukuş, 2014)

6.1.3. Torbalama (Bagging) Algoritması

Torbalama algoritması, 1994 yılında Breiman tarafından ortaya atılmıştır. Bu algoritma, istatistiksel sınıflandırma yöntemlerinin doğruluğunu artıran makine öğrenimi temelli bir algoritmadır. (Breiman, 1996). Torbalama yöntemi, varyansı düşürmekte ve yüksek dereceli öğrenmenin engellenmesini sağlamaktadır. Genelde eğitim verisinin farklı kombinasyonları ile eğitim örneklerinin oluşturulması amaçlanmaktadır. Bazı eğitim verisinde bazı örnekler yer almazken, bazıları birden fazla yer almaktadır. Bu şekilde üretilmiş birbirinden farklı örnekler içeren eğitim verileri ile kullanılan bu yönteme torbalama (bagging) algoritması denilmektedir (Atasever ve Özkan, 2012).

6.1.4. Gini ve Twoing Ayırma Kriterlerinin Karşılaştırılması

Kategorik bağımlı değişkenler için gini ve twoing algoritması kullanılmaktadır. Sınıflandırma ağaçlarında düğümlerde mümkün olan en iyi ayırmayı gerçekleştirmek için ayırma kuralı seçilirken aşağıdaki faktörler göz önünde bulundurulmalıdır.

- Kategorik bağımlı değişkenin seviye sayısı iki ise ve analizin %50'den daha az hata oranına sahip olacağı tahmin ediliyorsa, ayırma kuralı olarak Gini algoritması tercih edilmelidir.
- Kategorik bağımlı değişkenin seviye sayısı iki ise ve analizin %80'den daha az hata oranına sahip olacağı tahmin ediliyorsa, ayırma kuralı olarak Twoing algoritması tercih edilmelidir.
- Bağımlı değişkenin seviye sayısı 4'ten daha büyük olduğu koşullarda Twoing kuralı, Gini'den daha doğru bir seçimdir (Temel, 2004).

6.2. Regresyon Ağaçları

Regresyon ağaçlarının oluşumu sınıflandırma ağaçlarına benzemektedir. Sınıflandırma ağaçlarında, bağımlı değişkenin kategorik olmasından kaynaklı ortalama ve standart sapma değerleri hesaplanamamaktadır. Ancak, regresyon ağaçlarında bağımlı değişken sürekli olduğu için ortalama ve standart sapma değerleri hesaplanabilmektedir.

Ağaç modelinde karar verme noktalarına düğüm denilmektedir. Ağaç başlangıç düğümü, kök veya aile düğümü ile başlamaktadır. Bağımsız değişkenler aralarındaki ilişkiye göre her defasında ikili dallanma ile birbirleri arasında heterojen kendi içinde homojen alt düğümlere (çocuk düğümlere) ayrılmaktadır. Çocuk düğümlerden sonra bölünmenin gerçekleşmediği, ağacın en homojen yapısı olan terminal düğümlere ulaşılmaktadır. Bu yapı, bağımlı değişkenin sürekli olduğu durumlarda regresyon ağacı adını almaktadır (Jarosik, 2011)

Regresyon ağacı yönteminde değişkenlerin alt düğümlere ayrılmasında uygulanan azaltma (minimizasyon) problemi aşağıdaki gibi çözülmektedir.

$$argmin_{\,xj \le xj^R, j=1,\ldots,M} [P_l Var(Y_l) + P_r\, Var(Y_r)] \qquad (5.9)$$

Burada P_l ve P_r sırası ile sol ve sağ düğümlerin olasılıklarıdır. M ise eğitim verisinde yer alan değişken sayısıdır. Değişken j "x_j" olarak gösterilmektedir. x_j^R ise değişken x_j'nin en iyi ayırım değerini, Var(Y_l) ve Var(Y_r) karşılıklı sağ ve sol alt düğümlerin sorumlu olduğu vektörleri göstermektedir (Takma vd., 2017)

Regresyon ağaçlarında ayırma kriteri olarak en küçük kareli sapma (LSD) heterojenlik ölçüsü kullanılmaktadır.

$$R(t)=\frac{1}{N_w(t)}\sum w_n f_n(y_1-\bar{y}_{(t)})^2 \qquad (5.10)$$

olarak hesaplanmaktadır. Burada,

R(t) : LSD ölçüsü,
$N_w(t)$: t düğümündeki ağırlıklandırılmış durum sayısı,
w_n : i durumu için mevcut frekans değişkeni değeri,
y_l : hedef değişken değeri,
$\bar{y}_{(t)}$: t düğümü için ağırlıklı ortalama
göstermektedir.

En küçük kareli sapma sonucunda elde edilen ağacın büyüklüğü, budama (prunning) sonucudur (Oğuzlar, 2004).

6.3. Sınıflandırma Ağacı ve Regresyon Ağacının Karşılaştırılması

Sınıflandırma ağacı, sınıflandırma amaçlı kategorik yapıdaki bağımlı değişken değerlerinin tahmin edilmesini amaçlayan parametrik olmayan bir yöntemdir. Sınıflandırma ağaçları, parametrik olmayan bir yöntem olduğu için çoklu ya da doğrusal regresyondaki varsayımlara gerek duyulmamaktadır. Sınıflandırma ağacı, analiz ve yorumlama açısından oldukça görsel içeriklidir. Özellikle, bağımlı değişkenin alacağı değerler kolaylıkla tahmin edilebilmekte ve kolaylıkla yorumlanabilmektedir. Sınıflandırma ağacında bağımlı değişkenin kategorik yapıda olduğu bilindiği halde, bağımsız değişkenler hakkında belirli bir yargı

bulunmamaktadır. Bu sebeple; bağımsız değişkenler; kategorik, sürekli ya da sıralı yapıda olabilmektedir.

Sınıflandırma ağacı, karar ağacı algoritmasıdır. İstatistikte çok değişkenli istatistiksel yöntemler de karar alma aşamalarında bu ağacın, çok önemli bir yeri vardır. Bu sebeple, sınıflandırma ağacı birçok istatistiksel teknik yerine kullanılabilmektedir.

Regresyon modelleri, modelde yer alan değişkenlerdeki kayıp gözlemlerden ve uç değerlerden etkilenmektedir. Ancak, sınıflandırma ağacında bağımlı ya da bağımsız değişkenler bu konuda regresyon modelleri kadar hassas değildir ve modelde yer alan kayıp gözlemlerden ve uç değerlerden etkilenmezler. Özellikle büyük veriyle birlikte veri madenciliğinin oluşması ve gerekli bilgiye en az maliyetle ulaşabilme isteği sınıflandırma ağacının Türkiye ve dünya çapında kullanımını olabildiğince artırmaktadır.

Sınıflandırma ağacı, lojistik regresyon yöntemine alternatif iken; regresyon ağacı, çoklu ya da çok değişkenli regresyona alternatiftir (Kayri, 2014). Sınıflandırma ağacında, tahminleme yapabilmek için belirli bir yapı bulunmamaktadır. Birçok yöntem olmasına karşın, tahminlemenin doğruluğu ancak geçmişteki veriler yardımıyla belirlenebilmektedir.

Sınıflandırma ağaçlarında, birçok safsızlık ölçüsü bulunmaktadır. Bu safsızlık ölçüleri; gini, twoing ve ki-kare şeklindedir (Kurt, Türe ve Kurum, 2008). Safsızlık ölçüleri sıfır değerini aldığında bağımlı değişken homojen yapıda olmaktadır. Ancak, genelde en çok kullanılan safsızlık ölçüleri ise gini ve twoing algoritmalarıdır.

Bağımlı değişkenin, sayısal ya da sürekli olduğu durumlarda regresyon ağacı kullanılmaktadır. Regresyon ağaçlarında, sınıflar yoktur. Bu sebeple regresyon ağacında ayırma kriteri olarak gini indeksi kullanılmamaktadır. Regresyon ağacında, sınıflandırma ağacında olduğu gibi bağımsız değişkenler hem kategorik hem de sayısal olabilmektedir.

Regresyon ağacında bağımsız değişkenlerin her biri için bir model kurulabilmektedir. Bu modelde, veriler her bir bağımsız değişkene göre bölünmektedir. Bu durumda, her bölünme noktasında tahmin edilen değerler ile gerçek değerler arasında hatalar oluşmaktadır. Burada amaç, en düşük hataya uygun olarak ağaç bölünmesini gerçekleştirmektir.

Regresyon ağaçları, sınıflandırmaya dayalı kullanılan modellerin başında gelmektedir. Regresyon ağaçlarında, sınıflandırma ağacında olduğu gibi ilk olarak ağacın kurulması gerekmektedir. Ağaç yapısı oluştuktan sonra da eldeki verinin ağaca yerleştirilmektedir.

7. UYGULAMA

Bu tez çalışmasında, Türkiye İstatistik Kurumu tarafından "Hanehalkı Bilişim Teknolojileri Kullanımı Anketi" kapsamında 2017 yılında Türkiye'yi temsil edebilecek nitelikte örneklem üzerinde toplanan veriler kullanılmıştır. Çalışmada, bulut depo kullanımı yapan bireylerin profili, sınıflandırma ağacı tekniği ile incelenmiştir. Çalışmada kullanılan veriler, bulut depo kullanımı yapan bireylere ait olduğu için bulut bilişim ve bulut depo kullanımı hakkında da kısaca bilgi verilmiştir.

7.1. Bulut Bilişim ve Bulut Depo Kullanımı

Bilişim teknolojileri; her türlü bilginin ve verinin toplandığı, verilerin işlendiği, depolandığı, internet ve ağ sistemleri ile birbirinden farklı yerlere iletildikten sonra kullanıcıların hizmetine sunulmasında kullanılan iletişim teknikleri ve bilgisayarlar dahil tüm teknolojileri kapsayan bir sistemdir. Amerikan Bilişim Teknolojisi Topluluğu, bilişim teknolojilerini "çalışma, tasarım, geliştirme, uygulama, destek, kısmi yazılım uygulamaları ve bilgisayar donanımları" olarak açıklamıştır.

Bilgi ve iletişim teknolojilerinde yaşanan gelişmeler, bilginin verilere dönüşmesini sağlamaktadır. Bu da büyük verilerin internet üzerinden depolanabilirliğini ve erişilebilirliğini olanaklı hale getiren bulut bilişim teknolojisinin oluşumuna zemin hazırlamıştır.

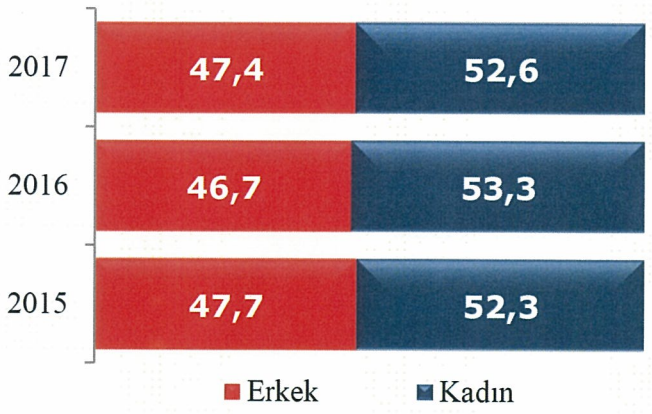
Verilerin bir bulutta depolandığı ve internet bağlantısı olan herhangi bir ortamda cihazlar aracılığı ile verilere ve bilgilere kolayca ulaşabildiğimiz hizmetlerin tümüne Bulut Bilişim ya da Bulut Teknolojisi adı verilmektedir. Bulut genellikle interneti temsil eden yapı olarak canlandırıldığından bulut bilişim adını almıştır (Sultan, 2010).

Bulut bilişim teknolojisi; çeşitli şirketler, üniversiteler ve büyük kuruluşlar tarafından kullanılmaktadır. Bulut bilişim, kişisel bilgisayarların depolama alanlarında yükü azaltmakta ve herhangi bir yerde ve zamanda bilgiye kolaylıkla ulaşmayı sağlamaktadır.

Bulut bilişim, içerisinde birçok servis ve altyapıyı barındırmaktadır. Bu sebeple, günümüz işletmeleri yüksek bir oranda bilişim teknolojilerini kullanarak verimlilik artışı sağlamaktadır (Baschab ve Piot, 2007).

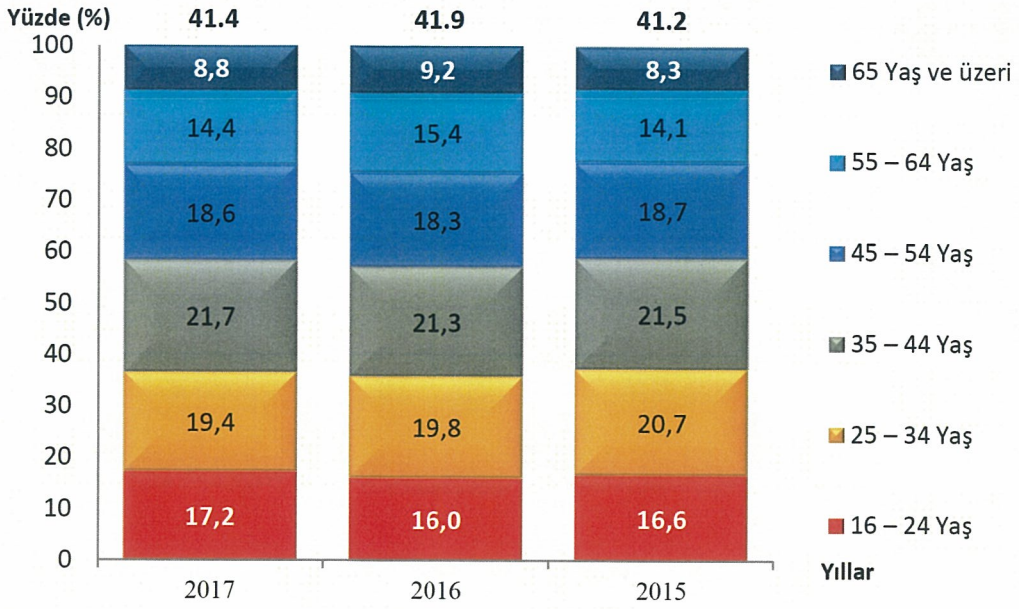
Bulut bilişim gelişmekte olan bir teknolojidir ve bu bilişim modelinin hayata geçebilmesi web hizmetleri, sanallaştırma (virtulization) ve ızgara (grid) bilişim adı verilen teknolojiler sayesinde olmuştur. Web hizmetleri, internet üzerinden erişilebilen yapılardır ve yer alan hizmetler açık standartlara göre yazıldığından programlama dilleri ve işletim sistemleri birbirinden bağımsız olarak ilerlemektedir. Web hizmetleri ile yer alan hizmetler yazılımcılar sayesinde geliştirilebilmekte, hatta yazılımcıların ve yazılımla uğraşan kişilerin kullandıkları kodların da web üzerinden yayınlanmasıyla var olan sistemin daha da gelişmesi sağlanmaktadır (Papazoglou, 2008).

7.2. 2015-2016-2017 Hanehalkı Bilişim Teknolojileri Kullanımı Anketi Sonuçlarının Karşılaştırılması



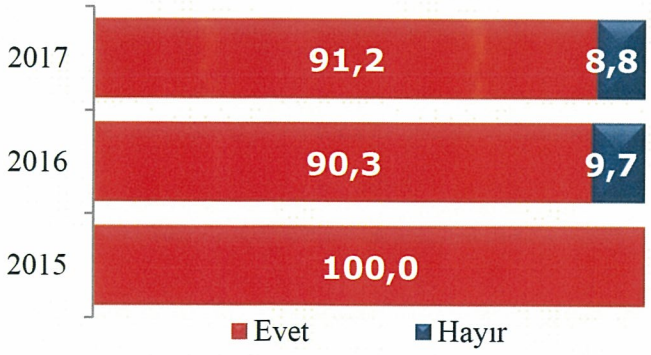
Şekil 3. Cinsiyet Dağılımları

2015 yılında ankete katılan 29359 kişinin %47,7'si erkek, %52,3'ü kadındır. 2016 yılında ankete katılan 25058 kişinin %46,7'si erkek, %53,3'ü kadındır. 2017 yılında ankete katılan 22977 kişinin %47,4'ü erkek, %52,6'sı kadındır. Ankete katılım oranı en yüksek 2015 yılında, en düşük 2017 yılında gerçekleşmiştir.



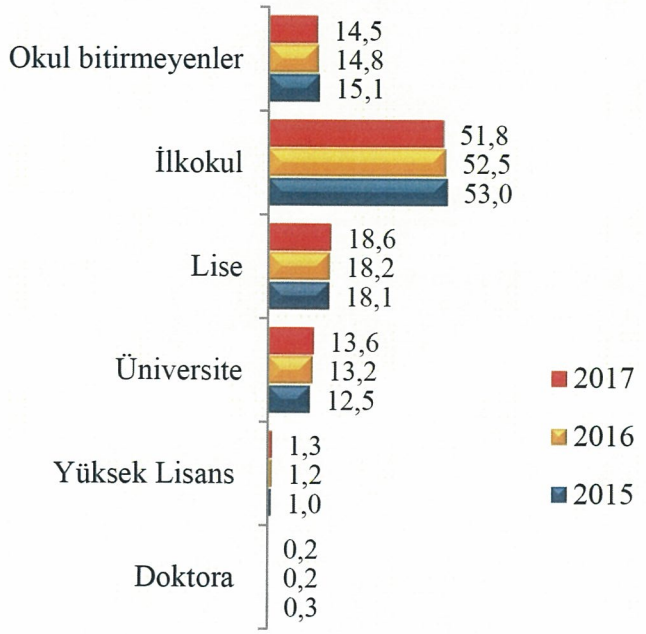
Şekil 4. Yaş Dağılımları

Ankete katılan birey sayısı 2015 yılında 29359, 2016 yılında 25058 ve 2017 yılında 22977 kişidir. 2015 yılında bireylerin %16,6'sı 16-24 yaş aralığında, %20,7'si 25-34 yaş aralığında, %21,5'i 35-44 yaş aralığında, %18,7'si 45-54 yaş aralığında, %14,1'i 55-64 yaş aralığında, %8,3'ü 65 yaş ve üzeridir. 2016 yılında bireylerin %16,0'ı 16-24 yaş aralığında, %19,8'i 25-34 yaş aralığında, %21,3'ü 35-44 yaş aralığında, %18,3'ü 45-54 yaş aralığında, %15,4'ü 55-64 yaş aralığında, %9,2'si 65 yaş ve üzeridir. 2017 yılında bireylerin %17,2'si 16-24 yaş aralığında, %19,4'ü 25-34 yaş aralığında, %21,7'si 35-44 yaş aralığında, %18,6'sı 45-54 yaş aralığında, %14,4'ü 55-64 yaş aralığında, %8,8'i 65 yaş ve üzeridir. Yıllara göre yaş dağılımlarına genel olarak bakıldığında, yüzdelerde yüksek bir değişim gözlenmemiştir.



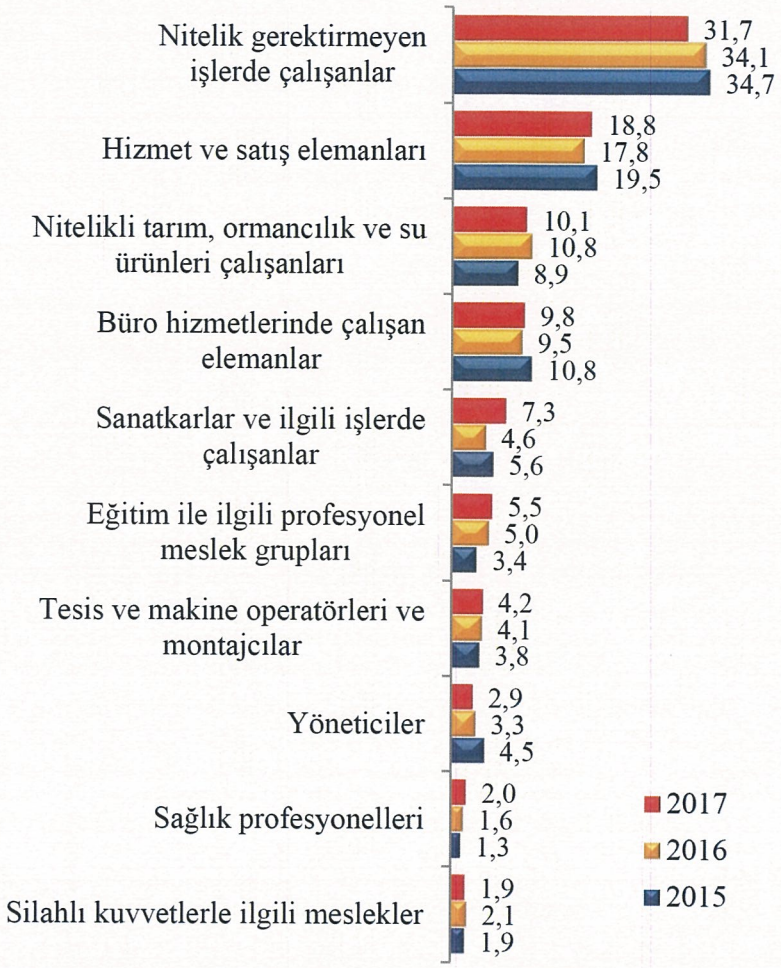
Şekil 5. Okuma Yazma Dağılımları

2015 yılında bireylerin tamamı okuma ve yazma biliyorken, 2016 yılında ankete katılan bireylerin yalnızca %90,3'ü okuma yazma bilmekte ve %9,7'si okuma yazma bilmemektedir. 2017 yılında bireylerin %91,2'si okuma ve yazma bilmektedir.



Şekil 6. Eğitim Durumu Dağılımları

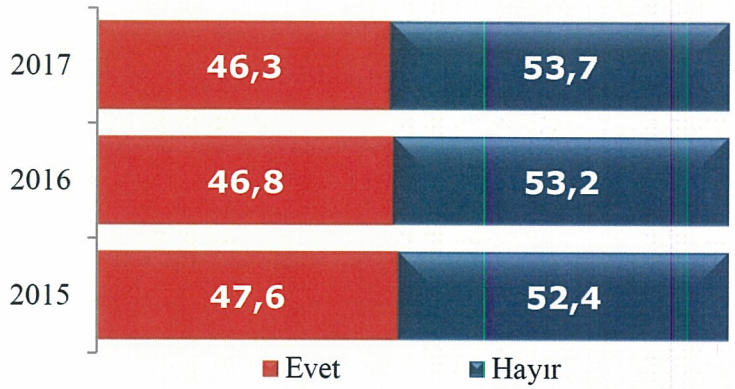
Katılımcıların eğitim durumlarına bakıldığında, 2015 yılında bireylerin %15,1'i okul bitirmemişken, %53'ü ilkokul, %18,1'i lise, %12,5'i üniversite, %1'i yüksek lisans ve %0,3'ü doktora mezunudur. 2016 yılında bireylerin %14,8'i okul bitirmemişken, %52,5'i ilkokul, %18,2'si lise, %13,2'si üniversite, %1,2'si yüksek lisans ve %0,2'si doktora mezunudur. 2017 yılında bireylerin %14,5'i okul bitirmemişken, %51,8'i ilkokul, %18,6'sı lise, %13,6'sı üniversite, %1,3'ü yüksek lisans ve %0,2'si doktora mezunudur.



Şekil 7. Meslek Dağılımları

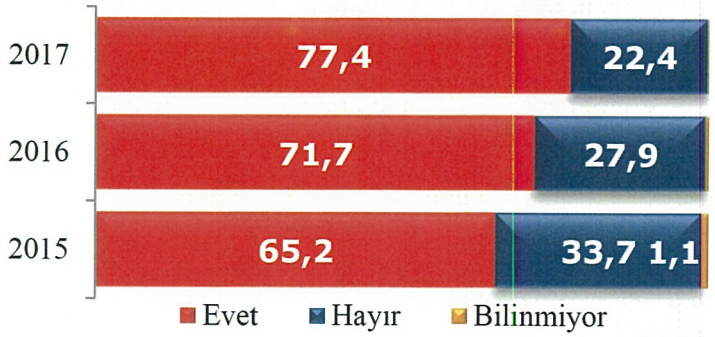
Uluslararası meslek sınıflandırmasına göre, 2015 yılında katılımcıların %34,7'si nitelik gerektirmeyen işlerde çalışanlardan, %19,5'i hizmet ve satış elemanlarından; %8,9'u nitelikli tarım, ormancılık ve su ürünleri çalışanlarından, %10,8'i büro hizmetlerinde çalışan elemanlardan, %5,6'sı sanatkarlar ve ilgili işlerde çalışanlardan, %3,4'ü eğitim ile ilgili profesyonel meslek gruplarından, %3,8'i tesis-makine öperatörleri ve montajcılarından, %4,5'i yöneticilerden, %1,3'ü sağlık profesyonellerinden, %1,9'u silahlı kuvvetlerle ilgili meslek çalışanlarından oluşmaktadır. 2016 yılında katılımcıların %34,1'i nitelik gerektirmeyen işlerde çalışanlardan, %17,8'i hizmet ve satış elemanlarından; %10,8'i nitelikli tarım, ormancılık ve su ürünleri çalışanlarından, %9,5'i büro hizmetlerinde çalışan elemanlardan, %4,6'sı sanatkarlar ve ilgili işlerde çalışanlardan, %5,0'ı eğitim ile ilgili profesyonel meslek gruplarından, %4,1'i tesis-makine operatörleri ve

montajcılarından, %3,3'ü yöneticilerden, %1,6'sı sağlık profesyonellerinden, %2,1'i silahlı kuvvetlerle ilgili meslek çalışanlarından oluşmaktadır. 2017 yılında katılımcıların %31,7'si nitelik gerektirmeyen işlerde çalışanlardan, %18,8'i hizmet ve satış elemanlarından; %10,1'i nitelikli tarım, ormancılık ve su ürünleri çalışanlarından, %9,8'i büro hizmetlerinde çalışan elemanlardan, %7,3'ü sanatkarlar ve ilgili işlerde çalışanlardan, %5,5'i eğitim ile ilgili profesyonel meslek gruplarından, %4,2'si tesis-makine operatörleri ve montajcılarından, %2,9'u yöneticilerden, %2,0'ı sağlık profesyonellerinden, %1,9'u silahlı kuvvetlerle ilgili meslek çalışanlarından oluşmaktadır.



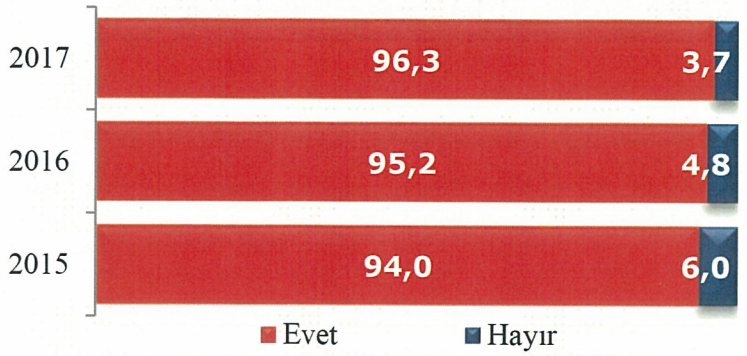
Şekil 8. Haneden Herhangi Birinin Bilgisayar Kullanma Durumu

Haneden herhangi birinin bilgisayar kullanım durumu, cevaplı tüm hanelere sorulan bir sorudur. Ankete katılan hane sayısı 2015 yılında 9847, 2016 yılında 11276 ve 2017 yılında 12781'dir. 2015 yılında hanelerin %47,6'sında herhangi biri bilgisayar kullanırken, 2016 yılında hanelerin %46,8'inde herhangi biri bilgisayar kullanmakta ve 2017 yılında hanelerin %46,3'ünde herhangi biri bilgisayar kullanmaktadır.



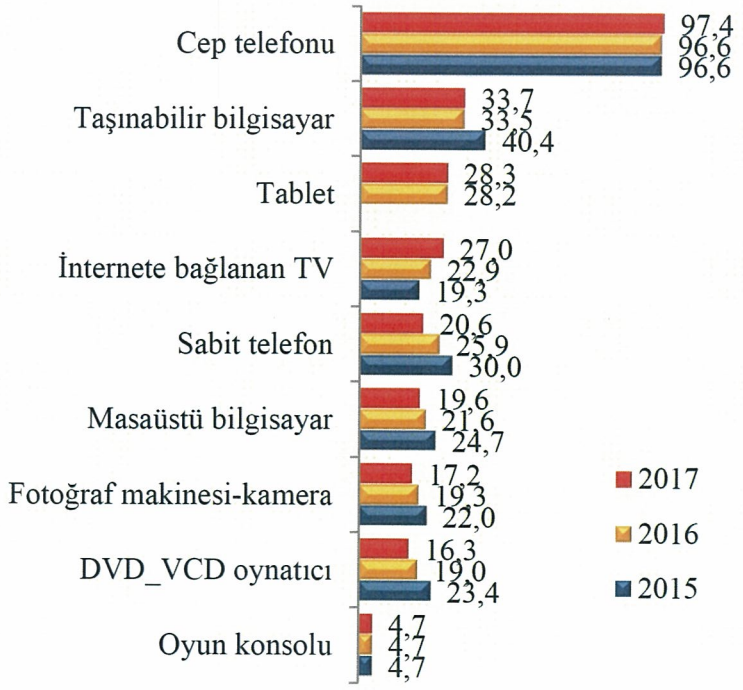
Şekil 1. Hanede İnternet Erişim Durumu

2015 yılında herhangi birinin bilgisayar kullandığı hanelerin %65,2'si internet erişimine sahipken, 2016 hanelerin %71,7'si internet erişimine sahip ve 2017 yılında hanelerin %77,4'ü internet erişimine sahiptir. 2015 yılından 2017 yılına kadar geçen süre hanelerin internet erişim durumlarında artış gözlemlenmiştir.



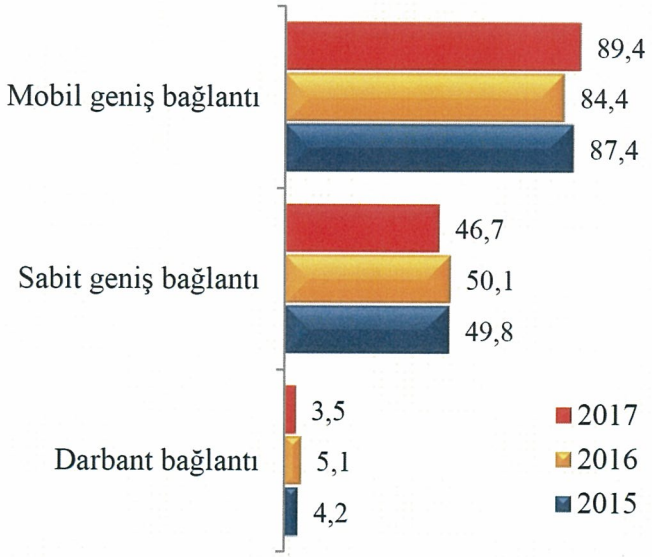
Şekil 10. Hanede İnternet Kullanım Durumu

2015 yılında internet erişimine sahip hanelerin %94'ü interneti kullanırken, 2016 yılında hanelerin %95,2'si ve 2017 yılında hanelerin %96,3'ü internet erişimine sahiptir ve interneti kullanmaktadır. 2015 yılından 2017 yılına kadar geçen sürede hanelerin internet kullanma oranlarında artış olmuştur.



Şekil 11. Hanede Bulunan Bilişim Ekipmanları

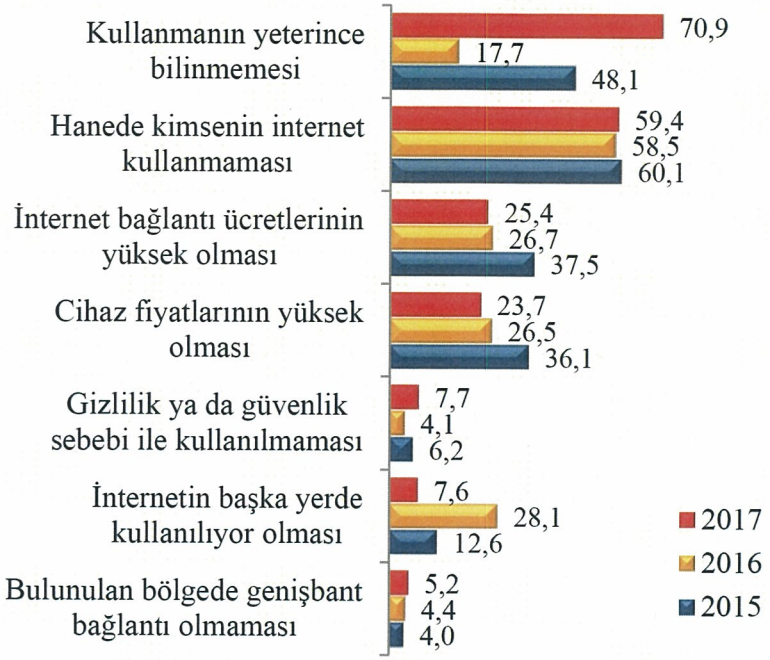
2015 yılında evinde internet kullanan 9847 hane, 2016 yılında 11276 hane ve 2017 yılında 12781 hane bulunmaktadır. Bu bilgiler doğrultusunda, 2015 yılında 9847 hanenin %96,6'sında cep telefonu, %40,4'ünde taşınabilir bilgisayar, %19,3'ünde internete bağlanan TV, %30'unda sabit telefon, %24,7'sinde masaüstü bilgisayar, %22'sinde fotoğraf makinesi veya kamera, %23,4'ünde DVD-VCD oynatıcı ve %4,7'sinde oyun konsolu bulunmaktadır. 2016 yılında 11.276 hanenin %96,6'sında cep telefonu, %33,5'inde taşınabilir bilgisayar, %28,2'sinde tablet, %22,9'unda internete bağlanan TV, %25,9'unda sabit telefon, %21,6'sında masaüstü bilgisayar, %19,3'ünde fotoğraf makinesi veya kamera, %19'unda DVD-VCD oynatıcı ve %4,7'sinde oyun konsolu bulunmaktadır. 2016 yılında 12.781 hanenin %97,4'ünde cep telefonu, %33,7'sinde taşınabilir bilgisayar, %28,3'ünde tablet, %27'sinde internete bağlanan TV, %20,6'sında sabit telefon, %19,6'sında masaüstü bilgisayar, %17,2'sinde fotoğraf makinesi veya kamera, %16,3'ünde DVD-VCD oynatıcı ve %4,7'sinde oyun konsolu bulunmaktadır.



Şekil 2. Evde Kullanılan İnternet Bağlantı Türleri

"Evde kullanılan internet bağlantı türü" sorusu evinde internet bağlantısı bulunan ve bilişim ekipmanlarından en az birine sahip hanelere sorulan bir sorudur. 2015 yılında evinde internet bağlantısı bulunan ve bilişim ekipmanlarından en az birine sahip 6424 hane bulunurken, 2016 yılında 8081 hane ve 2017 yılında 9888 hane bulunmaktadır. 2015 yılında 6.424 hanenin %87'sinde mobil geniş bağlantı, %49,8'inde sabit geniş bağlantı ve %4,2'sinde darbant bağlantı bulunmaktadır. 2016

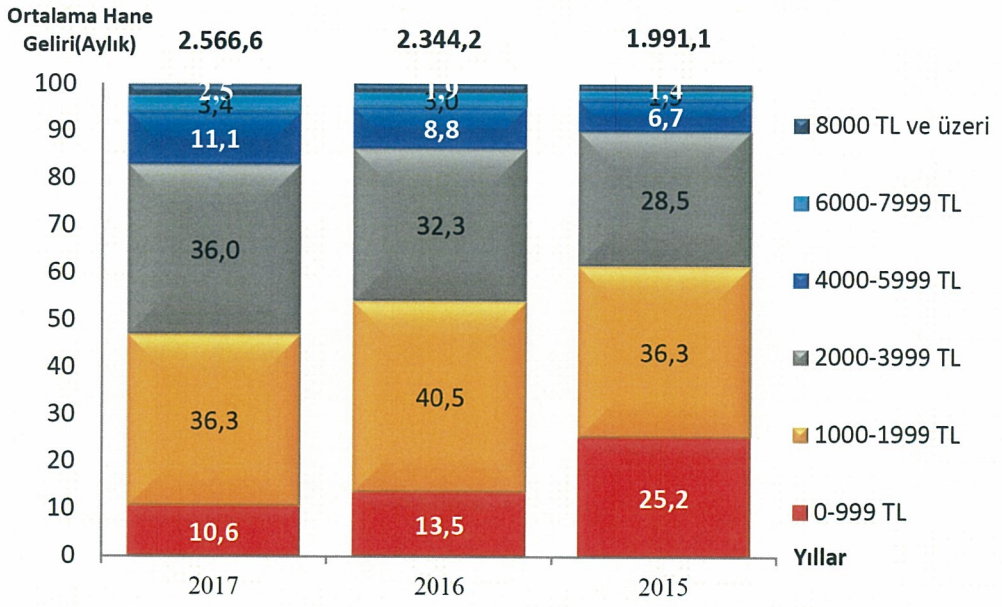
yılında 8.081 hanenin %84,4'ünde mobil geniş bağlantı, %50,1'inde sabit geniş bağlantı ve %5,1'inde darbant bağlantı bulunmaktadır. 2017 yılında 9888 hanenin %89,4'ünde mobil geniş bağlantı, %46,7'sinde sabit geniş bağlantı ve %3,5'inde darbant bağlantı bulunmaktadır.



Şekil 13. Evden İnternete Bağlanmama Nedenleri

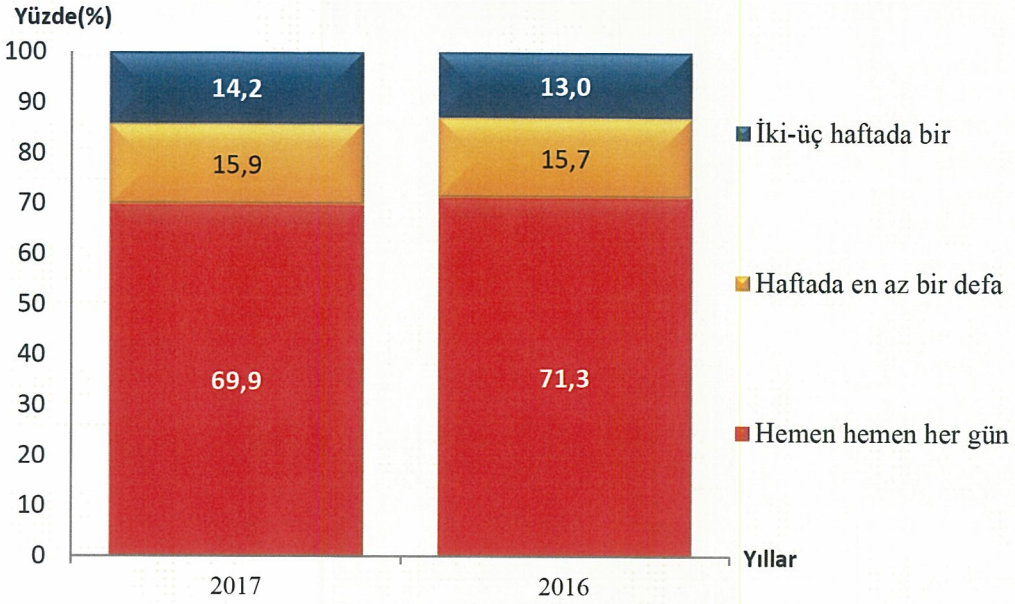
"Evden internete bağlanmama nedenleri" sorusu, evinde internet erişimi bulunduğu halde interneti kullanmayan hanelere sorulan bir sorudur. Buna göre, 2015 yılında evinde internet erişimi olduğu halde interneti kullanmayan 3265 hane bulunmaktadır ve bu hanelerin %48,1'i internet kullanımının yeterince bilinmemesi, %60,1'i hanede interneti kimsenin kullanmaması, %37,5'i internet bağlantı ücretlerinin yüksek olması, %36,1'i cihaz fiyatlarının yüksek olması, %6,2'si gizlilik ya da güvenlik sebebiyle kullanılmaması, %12,6'sı internetin başka yerde kullanılması ve %4'ü bulunulan bölgede genişbant bağlantı olmamasını internet kullanmamaya sebep göstermiştir. 2016 yılında evinde internet erişimi olduğu halde interneti kullanmayan 3140 hane bulunmaktadır ve bu hanelerin %17,7'si internet kullanımının yeterince bilinmemesi, %58,5'i hanede interneti kimsenin kullanmaması, %26,7'si internet bağlantı ücretlerinin yüksek olması, %26,5'i cihaz fiyatlarının yüksek olması, %4,1'i

gizlilik ya da güvenlik sebebiyle kullanılmaması, %28,1'i internetin başka yerde kullanılması ve %4,4'ü bulunulan bölgede genişbant bağlantı olmamasını internet kullanmamaya sebep göstermiştir. 2017 yılında evinde internet erişimi olduğu halde interneti kullanmayan 2865 hane bulunmaktadır ve bu hanelerin %70,9'u internet kullanımının yeterince bilinmemesi, %59,4'ü hanede interneti kimsenin kullanmaması, %25,4'ü internet bağlantı ücretlerinin yüksek olması, %23,7'si cihaz fiyatlarının yüksek olması, %7,7'si gizlilik ya da güvenlik sebebiyle kullanılmaması, %7,6'sı internetin başka yerde kullanılması ve %5,2'si bulunulan bölgede genişbant bağlantı olmamasını internet kullanmamaya sebep göstermiştir. Yıllara göre karşılaştırma yapıldığında internet kullanımının yeterince bilinmemesi 2017 yılında daha yüksek bir oranda internet kullanmama sebebi olarak gösterilmiştir.



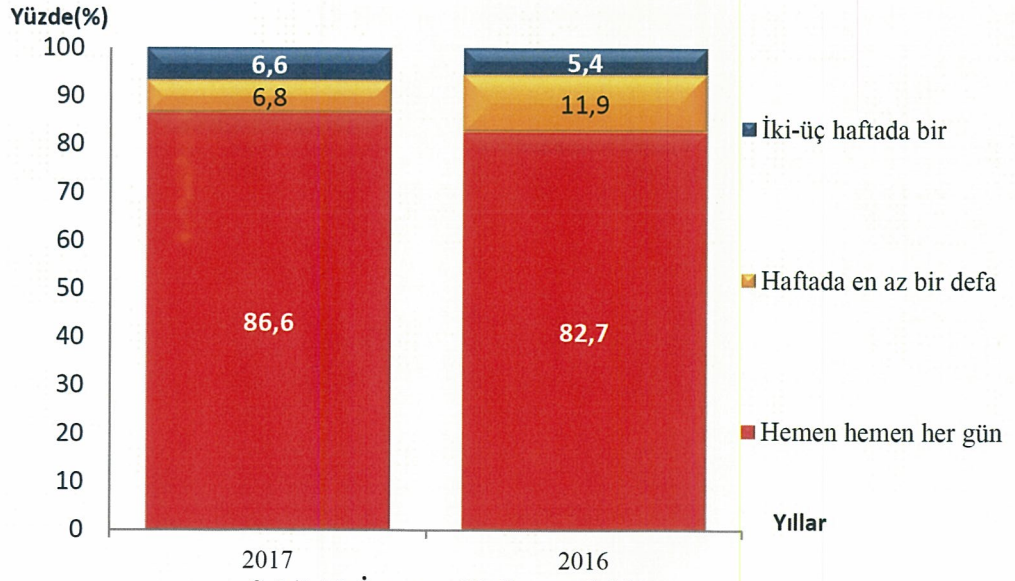
Şekil 14. Hanelerin Aylık Net Ortalama Geliri

"Hanelerin aylık net geliri" sorusu cevaplı tüm hanelere sorulmaktadır. 2015 yılında 9847 hane, 2016 yılında 11276 hane, 2017 yılında 12781 cevaplı hane bulunmaktadır. Hanelerin aylık net gelirlerine bakıldığında, 2015 yılında hanelerin aylık ortalama geliri 1991,06 TL iken, 2016 yılında aylık ortalama gelir 2344,24 TL ve 2017 yılında aylık ortalama gelir 2566,63 TL'dir. 2017 yılında katılımcıların gelirlerinin diğer yıllara oranla aylık gelirlerinde artış olduğu gözlemlenmektedir.



Şekil 15. Bilgisayar Kullanım Sıklığı

Bilgisayar kullanım sıklığında, araştırma yapılmadan son üç ay içinde bilgisayar kullanan katılımcılar baz alınmıştır. 2016 yılında son üç ay içinde bilgisayar kullanan birey sayısı 10311 iken 2017 yılında son üç ay içinde bilgisayar kullanan birey sayısı 12088'dir. 2016 yılında katılımcıların %71,3'ü hemen hemen her gün bilgisayar kullanırken, %15,7'si haftada en az bir defa, %13'ü de iki-üç haftada bir bilgisayar kullanmaktadır. 2017 yılında katılımcıların %69,9'u hemen hemen her gün bilgisayar kullanırken, %15,9'u haftada en az bir defa, %14,2'si iki-üç haftada bir bilgisayar kullanmaktadır.

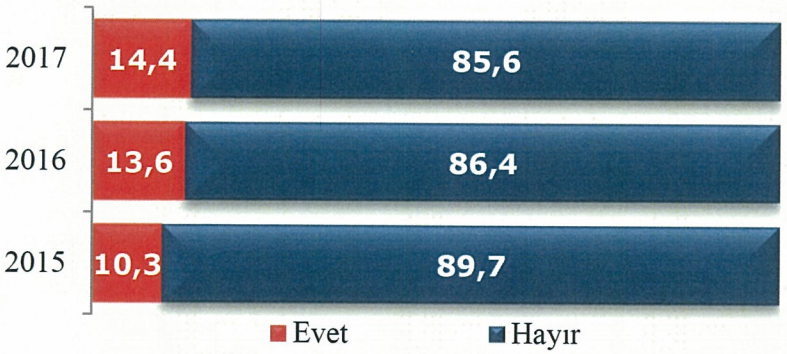


Şekil 16. İnternet Kullanım Sıklığı

2016 yılında son üç ay içinde internet kullanan birey sayısı 13510 iken, 2017 yılında son üç ay içinde internet kullanan birey sayısı 17901'dir. 2016 yılında son üç ay içinde internet kullanan bireylerin %82,7'si hemen hemen her gün internet kullanırken bu oran 2017 yılında %86,6 olarak artış göstermiştir. 2016 yılında haftada en az bir defa internet kullanan bireyler %11,9 oranındayken, 2017 yılında %6,8 oranına gerilemiştir. 2016 yılında bireylerin %5,4'ü iki-üç haftada bir internet kullanırken, 2017 yılında bireylerin %6,6'sı iki-üç haftada bir internet kullanmaktadır.

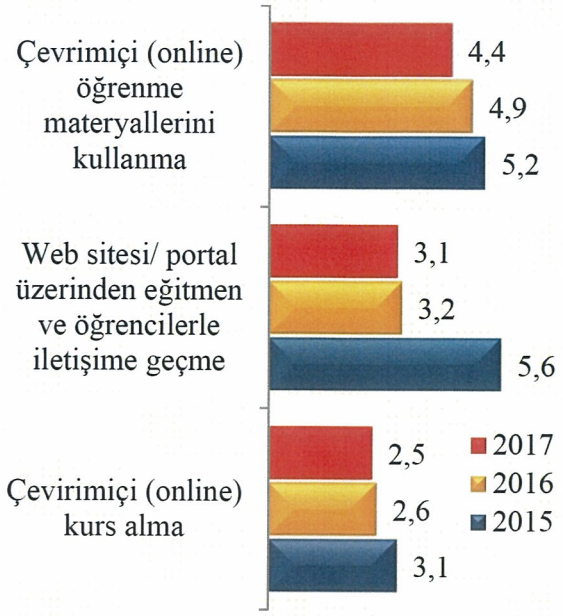
Şekil 17. Kişisel Amaçlarla İnternette Yapılan Faaliyetler (Son Üç Ay)

"Son üç ay içerisinde kişisel amaçla internette yapılan faaliyetler sorusuna", son üç ay içerisinde internet kullanan bireyler cevap vermiştir. 2015 yılında 10918 bireyin %79,8'i sosyal medya üzerinden profil oluşturma, fotoğraf ve içerik paylaşımı yaparken, 2016 yılında 13510 bireyin %81,2'si sosyal medya üzerinden profil oluşturma, fotoğraf ve içerik paylaşımı yapmakta ve 2017 yılında 17901 bireyin %82,6'sı sosyal medya üzerinden profil oluşturma, fotoğraf ve içerik paylaşımı yapmaktadır. 2017 yılında katılımcıların %69,4'ü sağlıkla ilgili bilgi araması yaparken, %67,2'si çevrimiçi haber sitelerini ziyaret etme, %64,7'si mal ve hizmetler hakkında bilgi alma, %61,3'ü internet üzerinden telefon/video görüşmesi, %44,7'si e-posta gönderme/alma, %33,6'sı internet bankacılığı, %17,8'i mal veya hizmet satışı amacıyla internet kullanmaktadır. 2017 yılında, diğer yıllara oranla yapılan faaliyet oranlarına genel olarak artış gözlemlenmektedir.



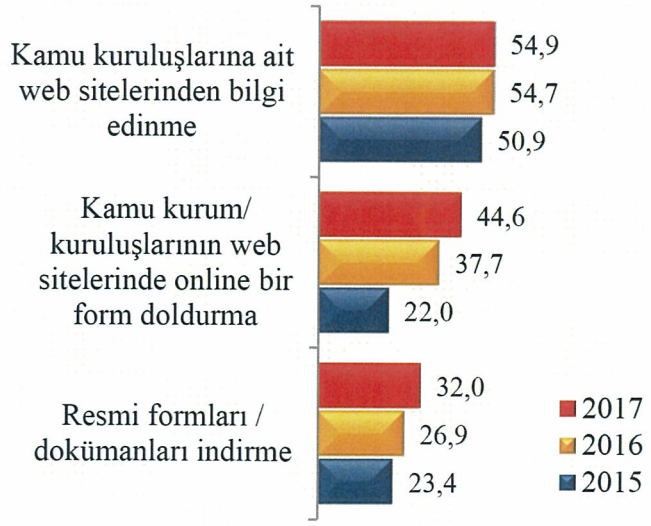
Şekil 18. İnternet Üzerinden Bulut Depo Kullanımı

"Resim, müzik, video veya dosya gibi dökümanların kişisel amaçlarla internette depolama alanı kullanımı" sorusuna son üç ay içinde internet kullanan bireyler cevap vermiştir. Buna göre 2015 yılında 10918 bireyin %10,3'ü internet üzerinden bulut depo kullanımı yaparken, 2016 yılında 13510 bireyin %13,6'sı internet üzerinden bulut depo kullanımı yapmakta ve 2017 yılında 17901 bireyin %14,4'ü internet üzerinden bulut depo kullanımı yapmaktadır. 2017 yılında internet üzerinden bulut depo kullanımı diğer yıllara oranlara daha yüksektir.



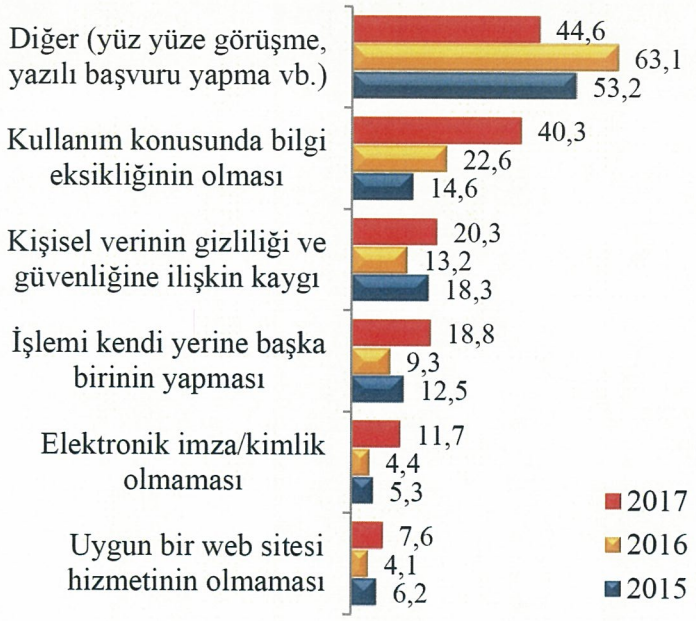
Şekil 19. İnternet Üzerinden Yapılan Öğrenim Faaliyetleri

2015 yılında 16-74 yaş arası 22977 bireye, 2016 yılında 16-74 yaş arası 13510 bireye ve 2017 yılında 16-74 yaş arası 17901 bireye "İnternet üzerinden yapılan öğrenim faaliyetleri" sorusu sorulmuştur. Buna göre, 2017 yılında 17901 bireyin %4,4'ü çevrimiçi öğrenme materyalleri kullanırken, %3,1'i web sitesi/portal üzerinden eğitmen ve öğrencilerle iletişime geçmekte ve %2,5'i çevrimiçi kurs almaktadır. 2017 yılında internet üzerinden yapılan öğrenim faaliyetlerinde diğer yıllara oranla düşüş yaşanmıştır.



Şekil 3. Kamu Kurum/Kuruluşları İle İletişimde Bulunma

"Kamu kurum/kuruluşları ile iletişimde bulunma sorusu" son üç ay içerisinde veya son üç ay ile bir yıl arasında internet kullanımı yapan bireylere sorulmuştur. 2015 yılında 11381 birey, 2016 yılında 13789 birey ve 2017 yılında 11137 birey baz alınmıştır. 2017 yılında 11.137 katılımcının %54,9'u kamu kuruluşlarına ait web sitelerinden bilgi edinirken, %44,6'sı kamu kurum/kuruluşlarının web sitelerinde online form doldururken, %32'si resmi formları ya da dökümanları indirmektedir. 2017 yılı kamu kurum/kuruluşları ile iletişimde bulunma oranları 2016 ve 2015 yıllarına oranla artış göstermiştir.



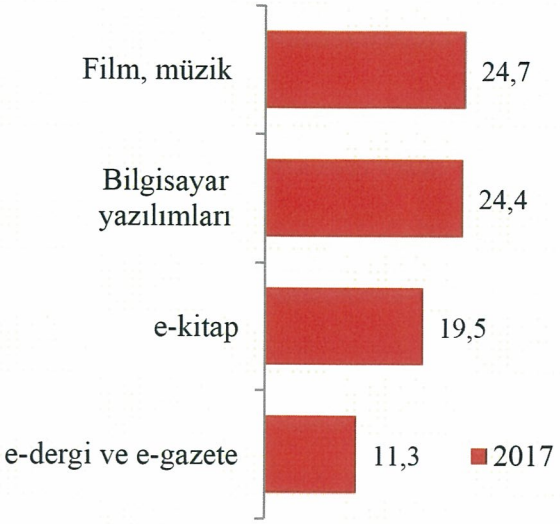
Şekil 21. Kamu Kurum/Kuruluşlarının Web Siteleri Üzerinden Form Göndermeme Nedenleri

"Kamu kurum/kuruluşlarının web siteleri üzerinden doldurulmuş form göndermeme nedenleri" sorusu, 18. Şekilde form göndermeyen bireylere sorulmuştur. Buna göre, 2015 yılında 513 katılımcı, 2016 yılında 1830 katılımcı ve 2017 yılında 5824 katılımcı kamu kurum/kuruluşlarına internet üzerinden doldurulmuş form göndermemiştir. 2017 yılında katılımcıların %44,6'sı yüz yüze görüşme vb. cevabı verirken, %40,3'ü kullanım konusunda bilgi eksikliğinin olduğunu, %20,3'ü kişisel verinin gizliliği ve güvenliğine ilişkin kaygı şeklinde cevaplar vermiştir. Kamu kurum/kuruluşlarının web siteleri üzerinden doldurulmuş form göndermeme sebepleri arasında, 2017 yılında kullanım konusunda bilgi eksikliğinin olması diğer yıllara oranlara daha yüksek bir oranla cevaplanmıştır.

Şekil 22. İnternet Üzerinden Alınan Mal veya Hizmet Türleri

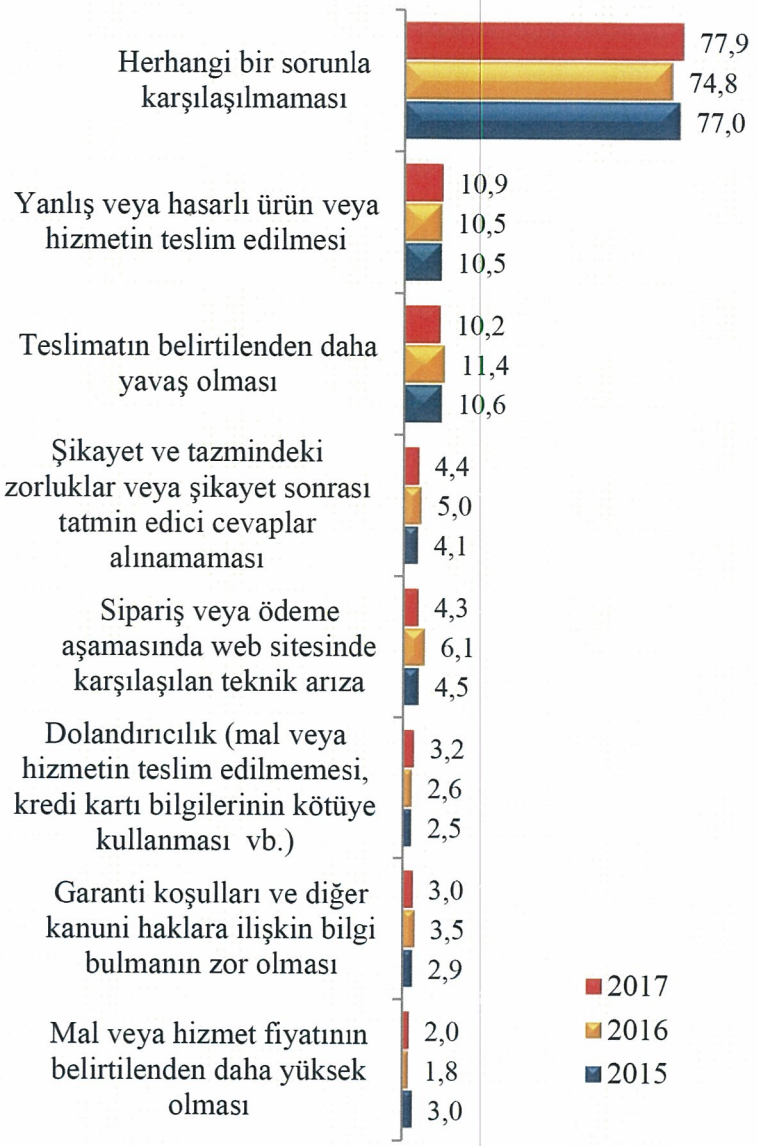
"İnternet üzerinden alınan mal veya hizmet türleri" sorusu son üç ay içerisinde veya üç ay ile bir yıl arasında internet üzerinden mal veya hizmet satın alan ya da sipariş veren bireylere sorulmuştur. 2015 yılında 3249 birey, 2016 yılında 4054 birey ve 2017 yılında 5530 birey baz alınmıştır. Buna göre 2015 yılında bireylerin %58,5'i, 2016 yılında bireylerin %60'ı ve 2017 yılında bireylerin %62,5'i giyim, spor malzemelerini internet üzerinden satın almıştır. 2017 yılında ikinci sırada %25,3 ile ev eşyası, %21,9 oranı ile üçüncü sırada seyahat ile ilgili işlemler yer almıştır. 2016 yılında ikinci sırada %28,4 oranı ile ilgili işlemler, üçüncü sırada %26,4 ile ev eşyası

yer almıştır. 2015 ve 2017 yıllarında internet üzerinden alınan mal veya hizmet türleri sırası değişiklik göstermemiştir.



Şekil 23. İnternet Üzerinden İndirilen ya da Satın Alınan Mal veya Hizmet Türleri

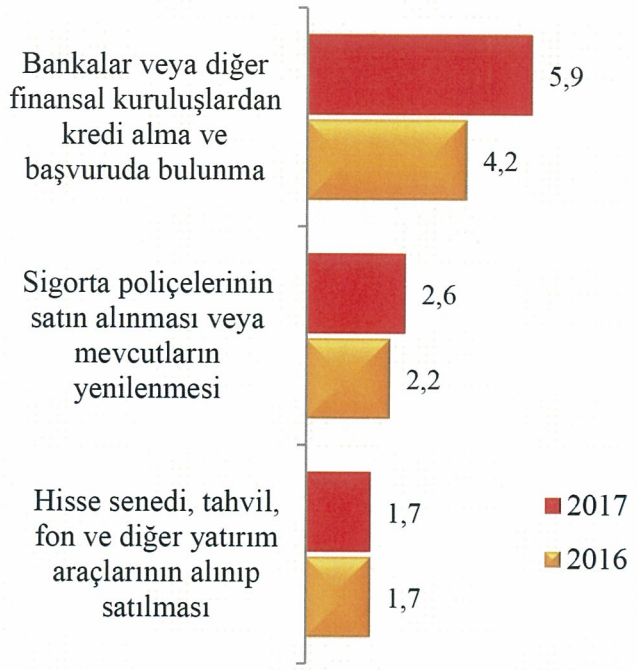
"İnternet üzerinden indirme ya da internet üzerinden alınan mal veya hizmet türleri sorusu", internet üzerinden film, müzik, kitap, gazete, oyun yazılımı, diğer bilgisayar yazılımı ve yazılım güncellemelerini internet üzerinden alan bireylere sorulmuştur. Buna göre, 2017 yılında 1136 bireyin %24,7'si film, müzik alırken, %24,4'ü bilgisayar yazılımları, %19,5'i e-kitap ve %11,3'ü e-dergi ve e-gazete gibi mal veya hizmet türlerini internet üzerinden indirme yolu ile ya da internet üzerinden satın almaktadır.



Şekil 24. İnternet Üzerinden Mal/Hizmet Alım/Sipariş Sırasında Karşılaşılan Sorunlar

"İnternet Üzerinden Mal ve Hizmet Siparişi Alırken/Verirken Karşılaşılan Sorunlar" sorusu internet üzerinden son üç ay içerisinde ya da son üç ay ile bir yıl arasında internet üzerinden mal veya hizmet satın alan ya da sipariş veren bireylere sorulmuştur. 2015 yılında 3249 birey, 2016 yılında 4053 birey ve 2017 yılında 5541 birey baz alınmıştır. Buna göre, 2015 yılında bireylerin %77,0'ı, 2016 yılında %74,8'i ve 2017 yılında %77,9'u herhangi bir sorunla karşılaşmamıştır. 2015 ve 2016 yıllarında bireylerin %10,5'i, 2017 yılında bireylerin %10,9'u yanlış hasarlı ürün veya ürünün teslim edilmemesi sorunlarıyla karşılaşmıştır. 2015 yılında bireylerin %10,6'sı, 2016 yılında bireylerin %11,4'ü ve 2017 yılında bireylerin %10,2'si teslimatın belirtilenden daha yavaş olması sorunu ile karşılaşmıştır. 2017

yılında bireylerin %4,4'ü şikayet ve tazmindeki zorluklar veya şikayet sonrası tatmin edici cevaplar alınamaması sorunu yaşarken, %4,3'ü sipariş veya ödeme aşamasında web sitesinde karşılaşılan teknik arıza, %3,2'si dolandırıcılık, %3'ü garanti koşulları ve diğer kanuni haklara ilişkin bilgi bulmanın zor olması, %2'si mal veya hizmet fiyatının belirtilenden daha yüksek olması ve %1,6'sı yurt dışı menşeli web sitelerinden ülkeye mal ve hizmet satılmaması sorunlarını yaşamaktadır.

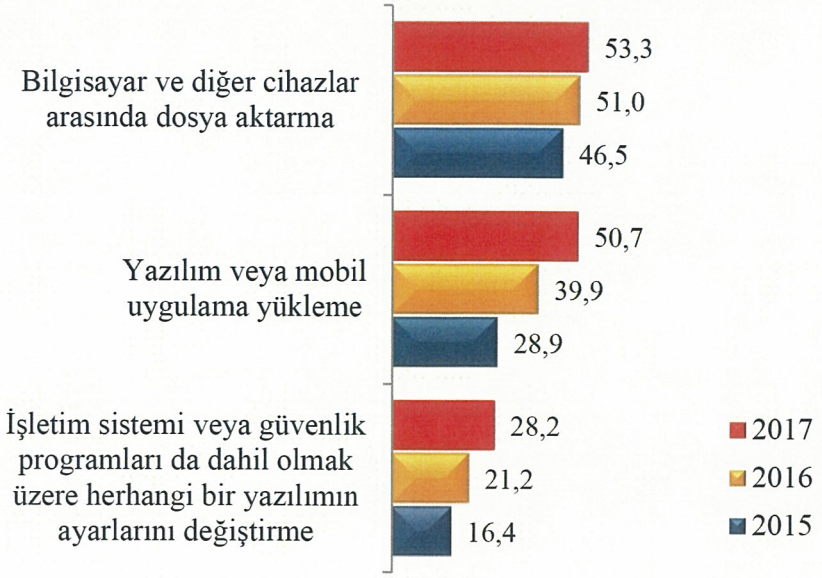


Şekil 25. İnternet Üzerinden Gerçekleştirilen Finansal İşlemler

"Son 12 ay içerisinde internet üzerinden gerçekleştirilen finansal işlemler sorusunda" 2016 yılında 13788 birey ve 2017 yılında 18222 birey baz alınmıştır. Buna göre, 2016 yılında bireylerin %4,2'si, 2017 yılında da bireylerin %5,9'u bankalardan veya diğer finansal kuruluşlardan kredi alma ve başvuruda bulunmuştur. 2016 yılında bireylerin %2,2'si sigorta poliçelerinin satın alınması veya mevcutlarının yenilenmesi işlemlerini internet üzerinden gerçekleştirirken, 2017 yılında bireylerin %2,6'sı sigorta poliçelerinin satın alınması veya mevcutlarının yenilenmesi işlemlerini internet üzerinden gerçekleştirmiştir. 2016 ve 2017 yıllarında bireylerin %1,7'si hisse senedi, tahvil, fon ve diğer yatırım araçlarının alınıp satılması işlemlerini internet üzerinden gerçekleştirmiştir.

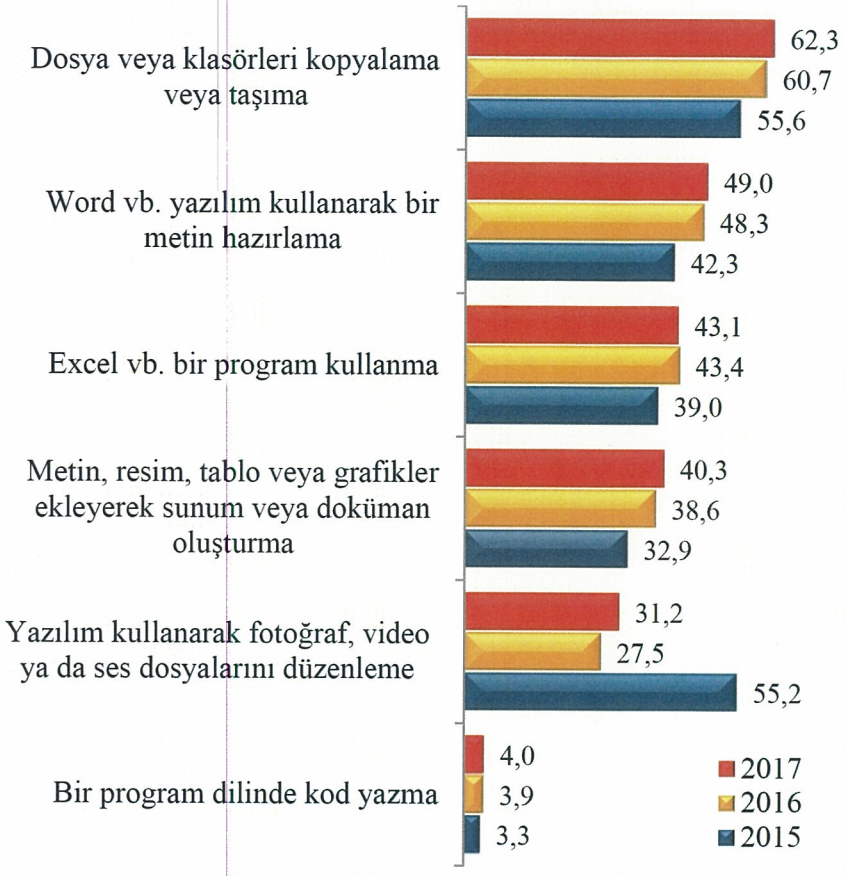
Şekil 26. İnternet Üzerinden Alışveriş Yapmama Nedenleri

"İnternet üzerinden alışveriş yapmama" nedenleri sorusu, bir yıldan uzun zamandır ya da hiç internet üzerinden mal veya satın almamış olan bireylere sorulmuştur. Soruya verilen cevaplar doğrultusunda; 2015 yılında 8605 birey, 2017 yılında da 13023 birey baz alınmıştır. Buna göre; 2015 yılında katılımcıların %81,1'i ürünü yerinde görerek alma, alışveriş yapılan dükkana bağlılık ve alışkanlıklar cevabı vermişken, 2017 yılında katılımcıların %73,3'ü ürünü yerinde görerek alma, alışveriş yapılan dükkana bağlılık ve alışkanlıklar cevabı vermiştir. 2015 yılında katılımcıların %45'i ödemede gizlilik ya da güvenlik kaygıları nedeniyle internet üzerinden alışveriş yapmazken, 2017 yılında bu oran %54,4'e yükselmiştir. 2017 yılında katılımcıların %29,6'sı bilgi ve beceri eksikliği, %28'i ürün teslimi alma ve şikayet ile ilgili sorun giderme konusunda güvensizlik, %20,7'si internet üzerinden sipariş edilen malların teslim problemi ve %16'sı internet üzerinden ödeme olanağı veren kredi kartı olmaması sebebiyle internet üzerinden alışveriş yapmamıştır. 2015 ve 2017 yılları arasında yüksek oranda değişiklik gözlemlenmemiştir.



Şekil 27. Bilgisayar ya da Mobil Cihazla Yapılan İşlemler

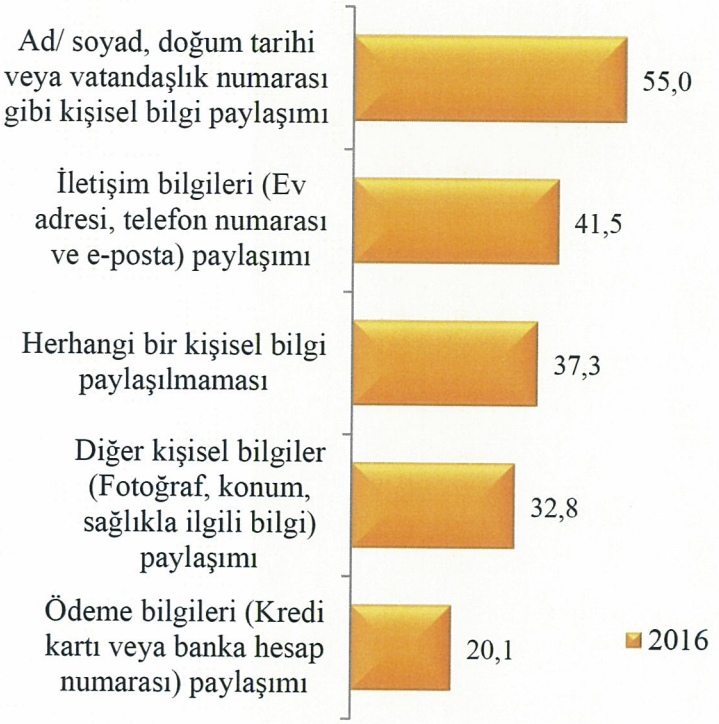
"Bilgisayar ya da mobil cihaz ile yapılan işlemler" sorusu son 12 ay içerisinde internet kullanan bireylere sorulmuştur. 2015 yılında 11381 birey, 2016 yılında 13789 birey ve 2017 yılında 18222 birey baz alınmıştır. 2015 yılında katılımcıların %46,5'i bilgisayar ve diğer cihazlar arasında aktarma, %28,9'u yazılım veya mobil uygulama yükleme, %16,4'ü ise işletim sistemi veya güvenlik programları dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmiştir. 2016 yılında katılımcıların %51'i bilgisayar ve diğer cihazlar arasında aktarma, %39,9'u yazılım veya mobil uygulama yükleme, %21,2'si ise işletim sistemi veya güvenlik programları dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmiştir. 2017 yılında katılımcıların %53,3'ü bilgisayar ve diğer cihazlar arasında aktarma, %50,7'si yazılım veya mobil uygulama yükleme, %28,2'si ise işletim sistemi veya güvenlik programları dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmiştir.



Şekil 28. Yazılım İle İlgili Yapılan Faaliyetler

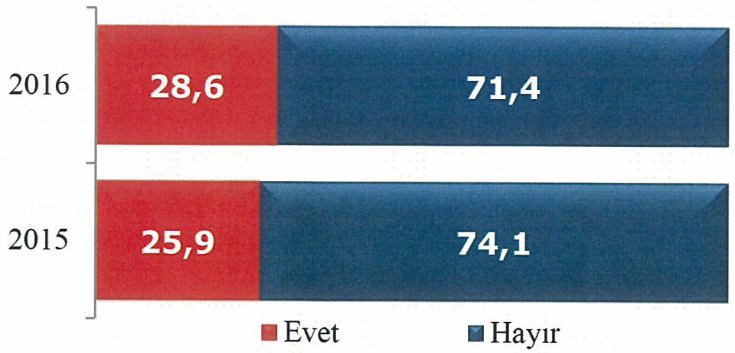
"Yazılım ile ilgili yapılan faaliyetler" sorusu, son 12 ayda internet kullanımı yapan bireylere sorulmuştur. 2015 yılında katılımcıların %55,6'sı dosya veya klasörleri kopyalama veya taşıma, %42,3'ü word vb. yazılım kullanarak metin hazırlama, %39'u excel vb. bir program kullanma, %32,9'u metin, resim, tablo ya da grafikler ekleyerek sunum ya da doküman oluşturma, %55,2'si yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme, %3,3'ü ise bir program dilinde kod yazmıştır. 2016 yılında katılımcıların %60,7'si dosya veya klasörleri kopyalama veya taşıma, %48,3'ü word vb. yazılım kullanarak metin hazırlama, %43,4'ü excel vb. bir program kullanma, %38,6'sı metin, resim, tablo ya da grafikler ekleyerek sunum ya da doküman oluşturma, %27,5'i yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme, %3,9'u ise bir program dilinde kod yazmıştır. 2017 yılında katılımcıların %62,3'ü dosya veya klasörleri kopyalama veya taşıma, %49'u word vb. yazılım kullanarak metin hazırlama, %43,1'i excel vb. bir program kullanma, %40,3'ü metin, resim, tablo ya da grafikler ekleyerek sunum ya da doküman oluşturma, %31,2'si yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme, %4'ü ise bir program dilinde kod yazmıştır. Yıllara göre en büyük

oransal değişim 2015 ve 2017 yılları arasında "yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme" seçeneğine verilen cevaplarda geçekleşmiştir. 2015 yılında katılımcıların %55,2'si yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlerken, 2017 yılında katılımcıların %31,2'si yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemiştir.



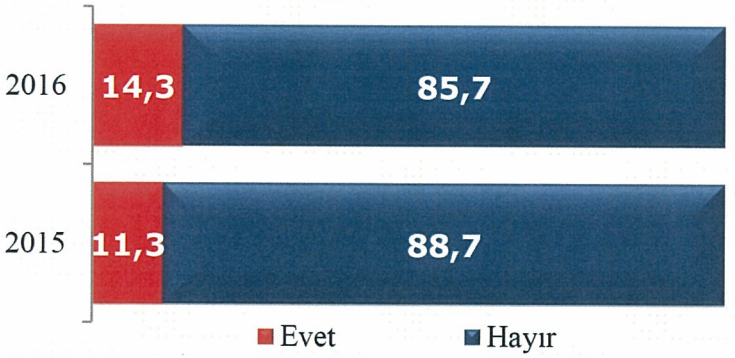
Şekil 29. İnternette Paylaşılan Kişisel Bilgiler

2016 yılında son 12 ayda internet kullanımı yapan 13789 bireyin %55'i ad/soyad, doğum tarihi veya vatandaşlık numarası gibi kişisel bilgi paylaşımı, %41,5'i iletişim bilgileri paylaşımı, %32,8'i fotoğraf, konum, sağlık ile ilgili paylaşım, %20,1'i ödeme bilgileri gibi kişisel bilgileri internet üzerinden gerçekleştirirken, bireylerin %37,3'ü internet üzerinden herhangi bir kişisel bilgi paylaşımında bulunmamıştır.



Şekil 30. Çerezler Hakkında Bilgi Sahibi Olma Durumu

Bireylere "çerezler hakkında bilgi sahibi olma" sorusu 2015 ve 2016 yıllarında sorulmuştur. 2015 yılında 11381 birey ve 2016 yılında 13789 birey baz alınmıştır. Buna göre, 2015 yılında 11381 bireyin %25,9'u çerezler hakkında bilgi sahibi iken, 2016 yılında 13789 bireyin %28,6'sı çerezler hakkında bilgi sahibidir. 2016 yılında çerezler hakkında bilgi sahibi olma oranı 2015 yıla oranla %2,7 artış göstermiştir.



Şekil 31. İnternet Tarayıcı Ayarlarında Çerezlerin Devre Dışı Bırakılması Durumu

Bireylere "internet tarayıcı ayarlarının çerezlerden korunmak ya da çerezleri devre dışı bırakmak için değiştirilmesi" sorusu 2015 ve 2016 yıllarında sorulmuştur. 2015 yılında 11381 birey ve 2016 yılında 13789 birey baz alınmıştır. Buna göre, 2015 yılında 11381 bireyin %11,3'ü internet ayarlarını çerezleri devre dışı bırakmak için değiştirirken, 2016 yılında 13789 bireyin %14,3'ü internet ayarlarını çerezleri devre dışı bırakmak için değiştirmektedir. 2016 yılında internet ayarlarını çerezleri devre dışı bırakmak için değiştirme oranı 2015 yıla oranla %3 artış göstermiştir.

7.3. CART Algoritması İle Oluşturulan Sınıflandırma Ağaçları ve Bulgular

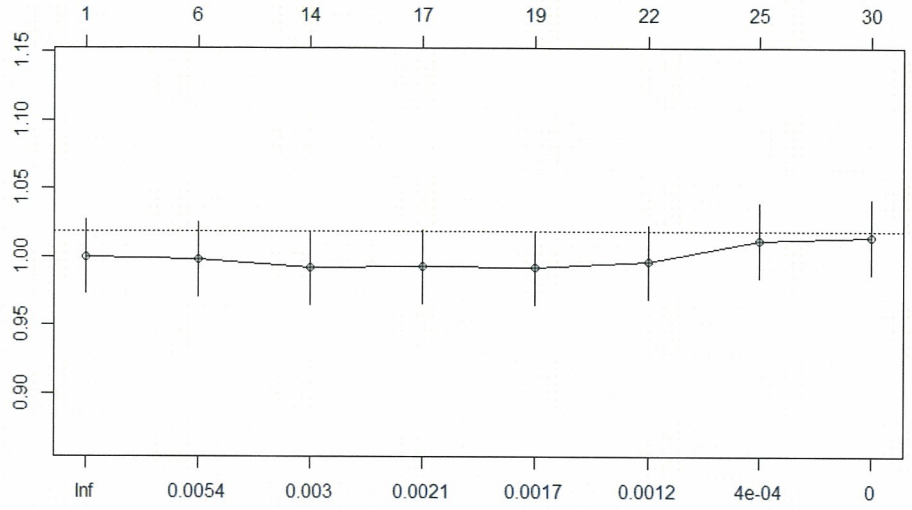
CART algoritması ile sınıflandırma ağacı ile model oluşturabilmek için, veri setinin sırasıyla %60'ı, %70'i, %80'i ve %90'ı ile eğitim verileri oluşturulmuştur. CART algoritması ile elde edilen sınıflandırma ağaçlarında gini ayırma kriteri kullanılmıştır. Sınıflandırma ağaçları, "R programlama dili" arayüzü olan RStudio 3.4.4 sürümü kullanılarak oluşturulmuştur. Uygulamada kullanılan değişkenler ve açıklamaları Tablo 2'de gösterilmiştir.

Tablo 2. Uygulamada Kullanılan Değişkenler ve Açıklamaları

bulut_depo_kullanimi	Resim, müzik, video veya dosya gibi dokümanların kişisel amaçlarla internette depolama alanı kullanımı (Google Drive, Dropbox, Windows Skydrive, iCloud, Amazon Cloud Drive vb.)	1= Evet 2= Hayır
yas	Bireylerin yaşı	16-74 yaş arası bireyler
cinsiyet	Bireylerin cinsiyeti	1= Erkek 2= Kadın
egitim_durumu	Bireylerin eğitim durumları	1= Bir okul bitirmedi 2= İlkokul 3= Lise 4= Üniversite 5= Yüksek lisans 6= Doktora
tasinabilir_bilgisayar	Son üç ay içinde ev ve işyeri dışında İnternete bağlanmak için kullanılan taşınabilir bilgisayar	1= Evet 2= Hayır
eposta_gonderme_alma	Kişisel amaçla (iş dışında, özel) internette e-posta gönderme/alma faaliyeti	1= Evet 2= Hayır
web_siteye_icerik_yukleme	Kendi oluşturduğunuz metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme	1= Evet 2= Hayır
cihazlar_arasi_dosya_aktarma	Bilgisayar ve diğer cihazlar arasında dosya aktarma	1= Evet 2= Hayır
yazilim_mobiluygulama_yukleme	Yazılım veya mobil uygulama (application) yükleme	1= Evet 2= Hayır
sistem_degistirme	İşletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirme	1= Evet 2= Hayır
yazilim_foto_video	Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme	1= Evet 2= Hayır
yazilim_kod_yazma	Bir program dilinde kod yazma	1= Evet 2= Hayır

(Kaynak: TÜİK, 2017)

Eğitim verileriyle oluşturulan sınıflandırma ağaçlarında optimum ağaca karar verebilmek için sınıflandırma oranlarına bakılması gerekmektedir. Uygun ağaca karar verme aşamasında hatalı sınıflandırma oranı en düşük olan ağaç, doğru sınıflandırma oranının en yüksek olduğu ağaçtır. Bu nedenle, çeşitli eğitim verileriyle analiz yapmak daha doğru sonuçlara ulaşmayı sağlayacaktır.



Şekil 32. 1. Sınıflandırma Ağacı Budama Kararı

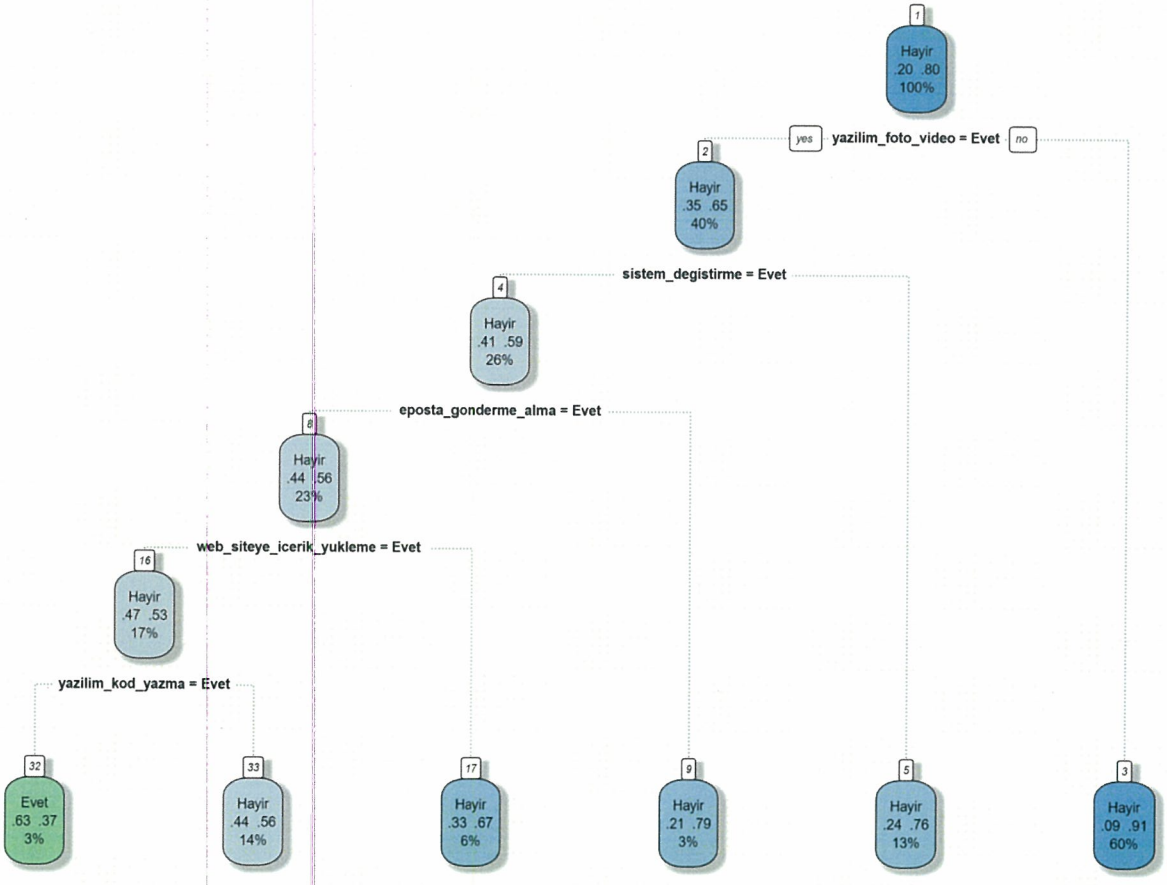
1. sınıflandırma ağacında, veri setinin %60'ı modeli oluşturmak, %40'ı da modeli test etmek için kullanılmıştır. 1. sınıflandırma ağacına ait budama kararı Şekil 32'de gösterilmiştir.

Karar ağaçlarında budama yaparken dikkat edilecek nokta, budama sonrası elde edilecek ağacın budama öncesi ağacı temsil edebilmesidir. Gini ayırma kriteri katsayısı olarak 0.0054 kullanıldığında, optimum ağaç elde edilecektir. 0.003 ve 0.0021 gini ayırma kriteri katsayıları kullanıldığında hatalı sınıflandırma oranı daha yüksek çıkacaktır. Budanmış sınıflandırma ağacında, etkili olan değişkenler ve hatalı sınıflandırma oranı Tablo 3'te gösterilmiştir.

Tablo 3. 1. Sınıflandırma Ağacı Modeli

```
Classification tree:
rpart(formula = bulut_depo_kullanimi ~ ., data = dataset_1, method = "class",
    control = r.ctrl)

Variables actually used in tree construction:
[1] eposta_gonderme_alma     sistem_degistirme        web_siteye_icerik_yukleme
[4] yazilim_foto_video       yazilim_kod_yazma

Root node error: 1106/5665 = 0.19523

n= 5665

        CP nsplit rel error  xerror     xstd
1 0.0081374      0   1.00000 1.00000 0.026975
2 0.0057000      5   0.95931 0.99819 0.026956
```

Gini ayırma kriteri olarak, 0.0054 katsayısının kullanıldığı 1. Sınıflandırma ağacının hatalı sınıflandırma oranı %19,5 olarak belirlenmiştir.



Şekil 33. CART Algoritması İle Oluşturulan 1. Sınıflandırma Ağacı

Şekil 33'te gösterilen ve CART algoritması ile oluşturulan 1. sınıflandırma ağacı incelendiğinde, internetten bulut depolama kullanımında birinci dereceden etkili bağımsız değişken yazılım kullanarak fotoğraf, video ya da ses dosyalarını

düzenleme, ikinci dereceden etkili bağımsız değişken işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılım ayarlarını değiştirme, üçüncü dereceden etkili bağımsız değişken kişisel amaçla internetten e-posta gönderme ve alma faaliyeti, dördüncü dereceden etkili bağımsız değişken bireyler tarafından oluşturulan metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme, beşinci dereceden etkili bağımsız değişken ise bir program dilinde kod yazma faaliyeti gibidir. Sınıflandırma ağacında, internetten bulut depolama kullanımı bağımlı değişkeni aile düğümünü ayıran ilk soru yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme sorusudur. Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleyenler 2 numaralı sol düğüme, yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenler ise 3 numaralı sağ düğümde sınıflandırılmıştır. Genel analiz yapıldığında, bireylerin %20'si internetten bulut depolama alanı kullanırken, %80'i internetten bulut depolama alanı kullanmamaktadır. Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleyenlerin %35'i internetten bulut depolama kullanırken, yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenlerin %9'u internetten bulut depolama kullanımı yapmaktadır.

3 numaralı sağ düğüm terminal düğümdür ve 2 numaralı düğümden itibaren sınıflandırma devam etmektedir. 2 numaralı düğümü saflaştırmak için bireylere, işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirip değiştirmedikleri sorulmuştur. Soruya verilen cevaplara uygun olarak 2. düğüm, 4 numaralı ve 5 numaralı iki alt düğüme ayrılmıştır. Düğüm 4'te işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştiren bireylerin %41'i internetten bulut depo kullanımı yaparken, düğüm 5'te ise işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmeyen bireylerin %24'ü internetten bulut depo kullanımı yapmaktadır.

Düğüm 4'ü oluşturan bireyler, kişisel amaçla internette e-posta alma ya da gönderme faaliyeti yapan bireyler (düğüm 8) ve kişisel amaçla internette e-posta alma ya da gönderme faaliyeti yapmayan bireyler (düğüm 9) olmak üzere iki alt düğüme ayrılmıştır. Kişisel amaçla internette e-posta alma ya da gönderme faaliyeti yapan bireylerin %44'ü internetten bulut depolama kullanımı yaparken, kişisel amaçla

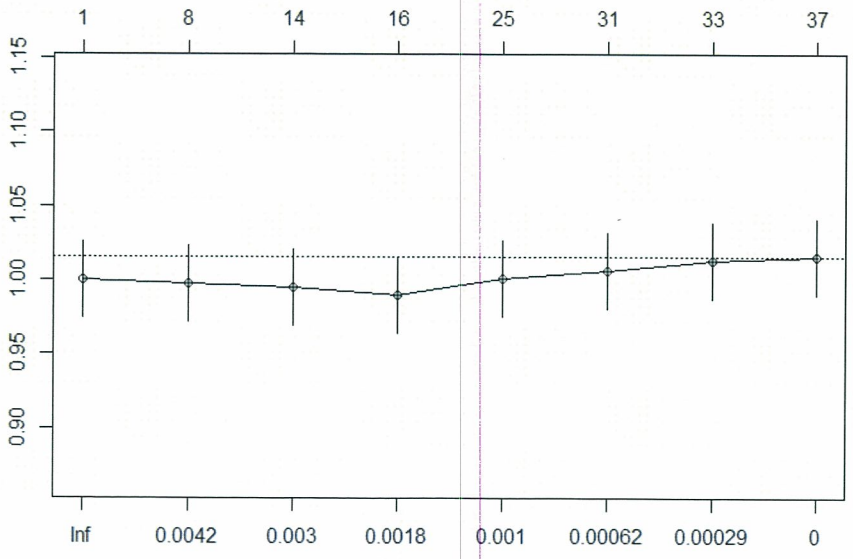
internette e-posta alma ya da gönderme faaliyeti yapmayan bireylerin %21'i internetten bulut depolama yapmaktadır.

Düğüm 8'i oluşturan bireyler; metin, fotoğraf, müzik, video, yazılım gibi içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenler (düğüm 16) ve yüklemeyenler (düğüm 17) olmak üzere iki alt düğüme ayrılmıştır. Metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenlerin %47'si internetten bulut depolama kullanırken, yüklemeyenlerin %33'ü internetten bulut depolama kullanımı yapmaktadır.

Düğüm 16'yı oluşturan bireyler, programlama dilinde kod yazan bireyler (düğüm 32) ve programlama dilinde kod yazmayan bireyler (düğüm 33) olmak üzere iki alt düğüme ayrılmıştır. Bir programlama dilinde kod yazanların %63'ü internetten bulut depolama yaparken, bir programlama dilinde kod yazmayanların %44'ü internetten bulut depolama yapmaktadır.

1. sınıflandırma ağacı ile, internetten bulut depolama kullanımında en önemli bağımsız değişkenin yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme olduğu sonucuna varılmıştır. Aynı zamanda, internette yapılan faaliyetlerin internetten bulut depolama alanı kullanımını arttırdığı gözlemlenmiştir.

2. sınıflandırma ağacında, veri setinin %70'i modeli oluşturmak, %30'u da modeli test etmek için kullanılmıştır. 2. sınıflandırma ağacına ait budama kararı Şekil 34'te gösterilmiştir.



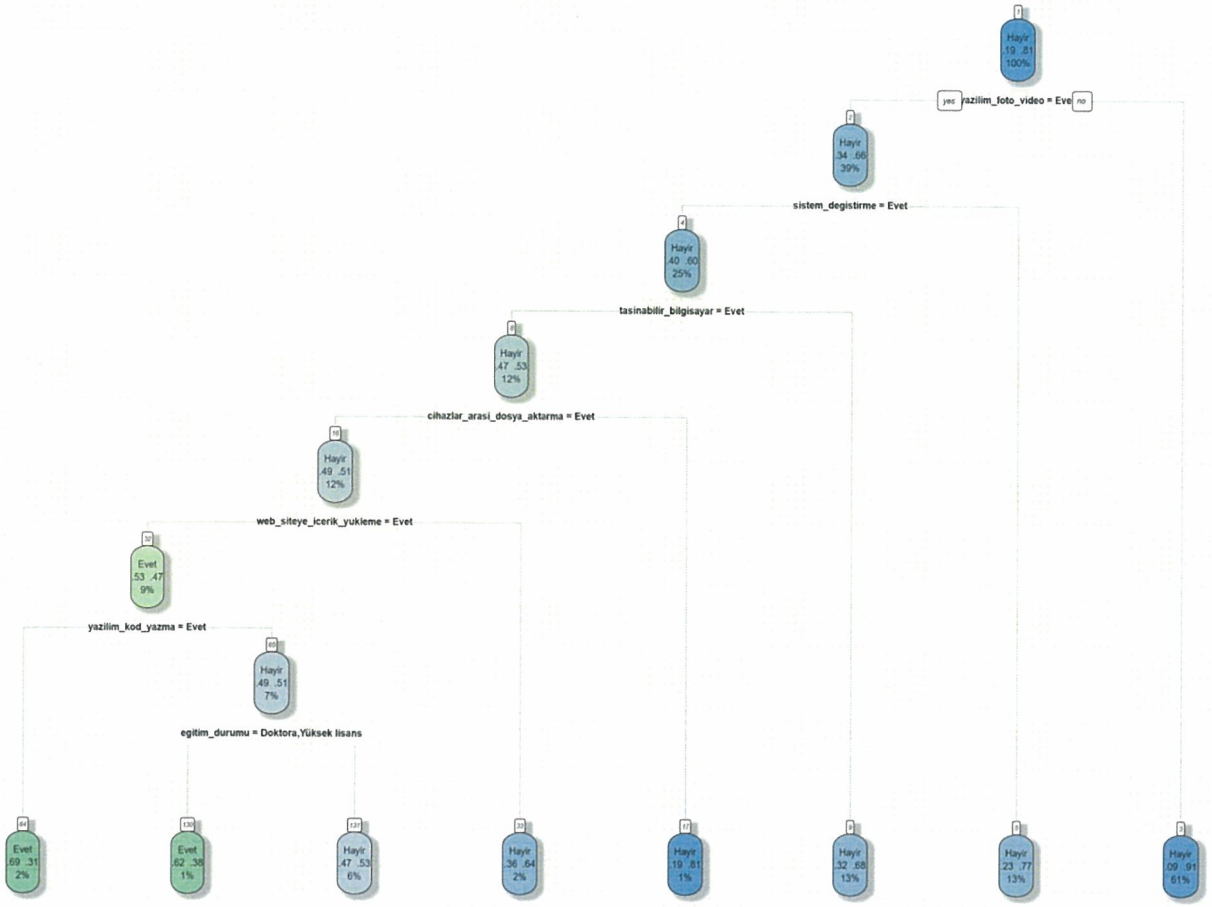
Şekil 34. 2. Sınıflandırma Ağacı Budama Kararı

Gini ayırma kriteri katsayısı olarak 0.0042 kullanıldığında, optimum ağaç elde edilecektir. 0.003 ve 0.0018 gini ayırma kriteri katsayıları kullanıldığında hatalı sınıflandırma oranı daha yüksek çıkacaktır. Budanmış sınıflandırma ağacında, etkili olan değişkenler ve hatalı sınıflandırma oranı Tablo 4'te gösterilmiştir.

Tablo 4. 2. Sınıflandırma Ağacı Modeli

```
Classification tree:
rpart(formula = bulut_depo_kullanimi ~ ., data = dataset_2, method = "class",
    control = r.ctrl)

Variables actually used in tree construction:
[1] cihazlar_arasi_dosya_aktarma egitim_durumu
[3] sistem_degistirme            tasinabilir_bilgisayar
[5] web_siteye_icerik_yukleme    yazilim_foto_video
[7] yazilim_kod_yazma

Root node error: 1229/6609 = 0.18596

n= 6609

        CP nsplit rel error  xerror     xstd
1 0.005533      0   1.00000 1.00000 0.025736
2 0.004200      7   0.95037 0.99756 0.025712
```

Gini ayırma kriteri olarak, 0.0042 katsayısının kullanıldığı 2. Sınıflandırma ağacının hatalı sınıflandırma oranı %18,5 olarak belirlenmiştir.



Şekil 35. CART Algoritması İle Oluşturulan 2. Sınıflandırma Ağacı

Şekil 35'te gösterilen ve CART algoritması ile oluşturulan 2. sınıflandırma ağacı incelendiğinde, internetten bulut depolama kullanımında birinci dereceden etkili bağımsız değişken yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme, ikinci dereceden etkili bağımsız değişken işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirme, üçüncü dereceden etkili bağımsız değişken hanede bilişim ekipmanı olarak taşınabilir bilgisayar bulunması, dördüncü dereceden etkili bağımsız değişken bilgisayar ve diğer cihazlar arasında dosya aktarma, beşinci dereceden etkili bağımsız değişken metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme, altıncı dereceden etkili bağımsız değişken bir programlama dilinde kod yazma, yedinci dereceden etkili bağımsız değişken ise eğitim durumudur. Bireylerin %19'u internetten bulut depolama kullanımı yaparken, %81'i internetten bulut depolama kullanımı yapmamaktadır. Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleyenler 2 numaralı sol düğüme, yazılım

kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenler ise 3 numaralı sağ düğümde sınıflandırılmıştır. Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleyenlerin %34'ü internetten bulut depolama kullanımı yaparken, yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenlerin %9'u internetten bulut depolama kullanımı yapmaktadır.

3 numaralı sağ düğüm, terminal düğümdür ve 2 numaralı düğümden itibaren sınıflandırma devam etmektedir. 2 numaralı düğümü saflaştırmak için bireylere, işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirip değiştirmedikleri sorulmuştur. Soruya verilen cevaplara uygun olarak 2. düğüm, 4 numaralı ve 5 numaralı iki alt düğüme ayrılmıştır. Düğüm 4'te işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştiren bireylerin %40'ı internetten bulut depo kullanımı yaparken, düğüm 5'te ise işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmeyen bireylerin %32'si internetten bulut depo kullanımı yapmaktadır.

Düğüm 4, hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olan bireyler (düğüm 8) ve hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olmayan bireyler (düğüm 9) olmak üzere iki alt düğüme ayrılmıştır. Hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olan bireylerin %47'si internetten bulut depolama kullanımı yaparken, hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olmayan bireylerin %32'si internetten bulut depolama yapmaktadır.

Düğüm 8, bilgisayar ve diğer cihazlar arasında dosya aktaran bireyler (düğüm 16) ile bilgisayar ve diğer cihazlar arasında dosya aktarmayan bireyler (düğüm 17) olmak üzere iki alt düğüme ayrılmıştır. Bilgisayar ve diğer cihazlar arasında dosya aktaran bireylerin %49'u internette bulut depolama kullanırken, bilgisayar ve diğer cihazlar arasında dosya aktarmayan bireylerin %36'sı internetten bulut depolama kullanmaktadır.

Düğüm 16; metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenler (düğüm 32) ve metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yüklemeyenler (düğüm 33) olmak üzere iki alt düğüme ayrılmıştır. Metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenlerin %53'ü

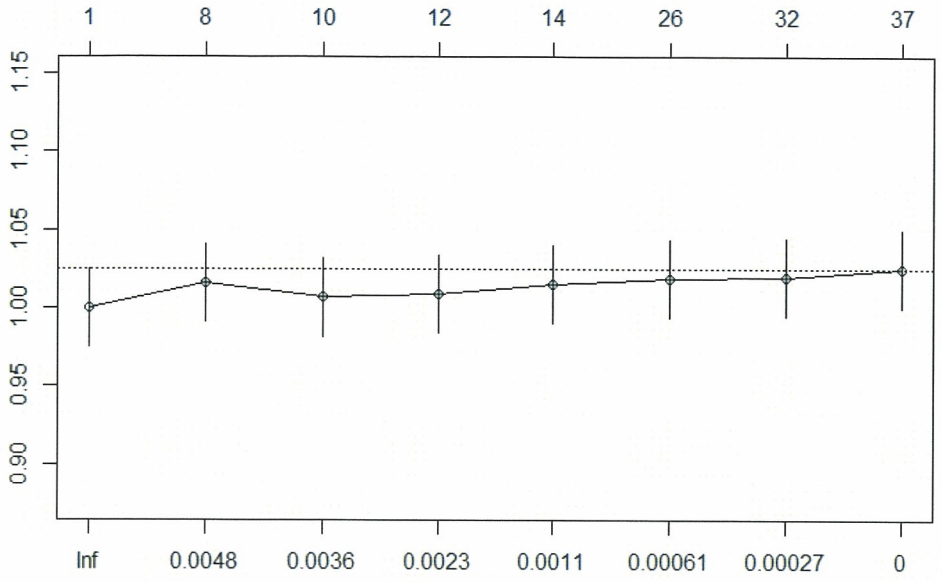
internette bulut depolama kullanırken, metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yüklemeyenlerin %36'sı internetten bulut depolama kullanımı yapmaktadır.

Düğüm 32, programlama dilinde kod yazan bireyler (düğüm 64) ve programlama dilinde kod yazmayan bireyler (düğüm 65) olmak üzere iki alt düğüme ayrılmıştır. Bir programlama dilinde kod yazan bireylerin %69'u internetten bulut depolama kullanımı yaparken, programlama dilinde kod yazmayan bireylerin %49'u bulut depolama kullanımı yapmaktadır.

Düğüm 65, eğitim durumu yüksek lisans ve doktora olan bireyler (düğüm 130) ile eğitim durumu yüksek lisans ve doktora olmayan bireyler (düğüm 131) olmak üzere iki alt düğüme ayrılmıştır. Eğitim durumu yüksek lisans ve doktora olan bireylerin %62'si internetten bulut depolama kullanımı yaparken, eğitim durumu yüksek lisans ve doktora olmayan bireylerin %47'si internetten bulut depolama kullanımı yapmaktadır.

CART algoritması ile oluşturulan 2. sınıflandırma ağacında, hanede taşınabilir bilgisayara sahip bireylerin bağımsız değişkenlerde yer alan eylemleri yapma olasılığı yüksek olduğundan, doğru orantılı olarak internette bulut depolama kullanım yüzdelerinde de artış gözlemlenmiştir.

3. sınıflandırma ağacında, veri setinin %80'i modeli oluşturmak, %20'si de modeli test etmek için kullanılmıştır. 3. sınıflandırma ağacına ait budama kararı Şekil 37'de gösterilmiştir.



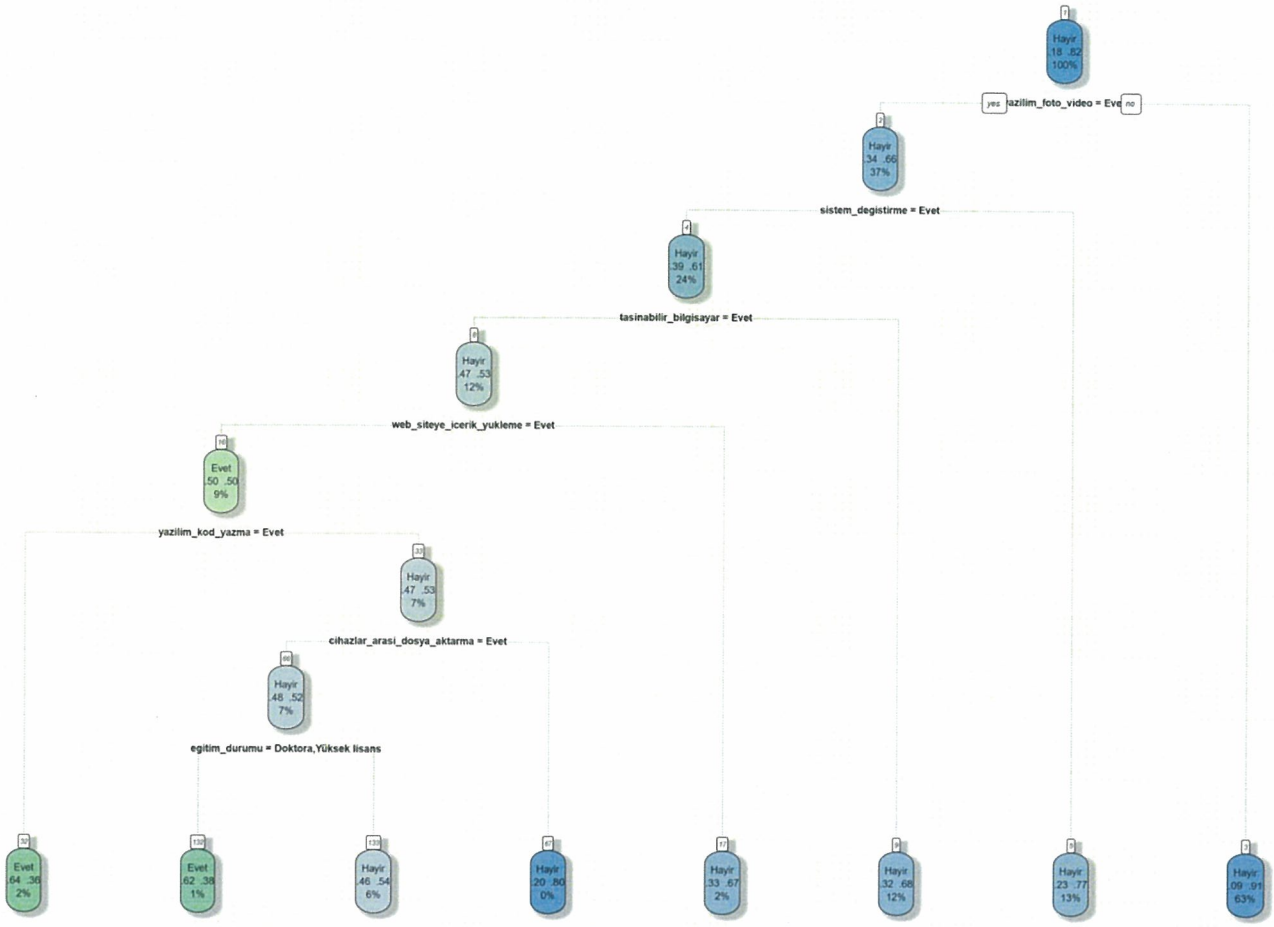
Şekil 36. 3. Sınıflandırma Ağacına Ait Budama Kararı

Gini ayırma kriteri katsayısı olarak 0.0048 kullanıldığında, optimum ağaç elde edilecektir. 0.0036 ve 0.0023 gini ayırma kriteri katsayıları kullanıldığında hatalı sınıflandırma oranı daha yüksek çıkacaktır. Budanmış sınıflandırma ağacında, etkili olan değişkenler ve hatalı sınıflandırma oranı Tablo 5'te gösterilmiştir.

Tablo 5. 3. Sınıflandırma Ağacı Modeli

```
Classification tree:
rpart(formula = bulut_depo_kullanimi ~ ., data = dataset_3, method = "class",
    control = r.ctrl)

Variables actually used in tree construction:
[1] cihazlar_arasi_dosya_aktarma egitim_durumu                sistem_degistirme
[4] tasinabilir_bilgisayar       web_siteye_icerik_yukleme    yazilim_foto_video
[7] yazilim_kod_yazma

Root node error: 1330/7479 = 0.17783

n= 7479

        CP nsplit rel error  xerror     xstd
1 0.006015      0   1.00000 1.00000 0.024863
2 0.004800      7   0.95789 0.99774 0.024841
```

Gini ayırma kriteri olarak, 0.0048 katsayısının kullanıldığı 3. Sınıflandırma ağacının hatalı sınıflandırma oranı %17,8 olarak belirlenmiştir.



Şekil 37. CART Algoritması İle Oluşturulan 3. Sınıflandırma Ağacı

Şekil 37'de gösterilen ve CART algoritmasıyla oluşturulan 3. sınıflandırma ağacı incelendiğinde, internetten bulut depolama kullanımında birinci dereceden etkili bağımsız değişken yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme, ikinci dereceden etkili bağımsız değişken işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirme, üçüncü dereceden etkili bağımsız değişken hanede bilişim ekipmanı olarak taşınabilir bilgisayar bulunması, dördüncü dereceden etkili bağımsız değişken metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme, beşinci dereceden etkili bağımsız değişken bir programlama dilinde kod yazma, altıncı dereceden etkili bağımsız değişken bilgisayar ve diğer cihazlar arasında dosya aktarma, yedinci dereceden etkili bağımsız değişken ise eğitim durumudur. Bireylerin %18'i internetten bulut depolama kullanımı yaparken, %82'si internetten bulut depolama kullanımı yapmamaktadır. Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleyenler 2 numaralı sol düğüme, yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenler ise 3 numaralı sağ düğümde

sınıflandırılmıştır. Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleyenlerin %34'ü internetten bulut depolama kullanımı yaparken, yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenlerin %9'u internetten bulut depolama kullanımı yapmaktadır.

3 numaralı sağ düğüm terminal düğümdür ve 2 numaralı düğümden itibaren sınıflandırma devam etmektedir. 2 numaralı düğümü saflaştırmak için bireylere, işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirip değiştirmedikleri sorulmuştur. Soruya verilen cevaplara uygun olarak 2. düğüm, 4 numaralı ve 5 numaralı iki alt düğüme ayrılmıştır. Düğüm 4'te işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştiren bireylerin %39'u internetten bulut depo kullanımı yaparken, düğüm 5'te ise işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmeyen bireylerin %23'ü internetten bulut depo kullanımı yapmaktadır.

Düğüm 4, hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olan bireyler (düğüm 8) ve hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olmayan bireyler (düğüm 9) olmak üzere iki alt düğüme ayrılmıştır. Hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olan bireylerin %47'si internetten bulut depolama kullanımı yaparken, hanesinde bilişim ekipmanı olarak taşınabilir bilgisayarı olmayan bireylerin %32'si internetten bulut depolama yapmaktadır.

Düğüm 8; metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenler (düğüm 16) ve metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yüklemeyenler (düğüm 17) olmak üzere iki alt düğüme ayrılmıştır. Metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenlerin %50'si internette bulut depolama kullanırken, metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yüklemeyenlerin %33'ü internetten bulut depolama kullanımı yapmaktadır.

Düğüm 16, programlama dilinde kod yazan bireyler (düğüm 32) ve programlama dilinde kod yazmayan bireyler (düğüm 33) olmak üzere iki alt düğüme ayrılmıştır. Bir programlama dilinde kod yazan bireylerin %64'ü internetten bulut depolama

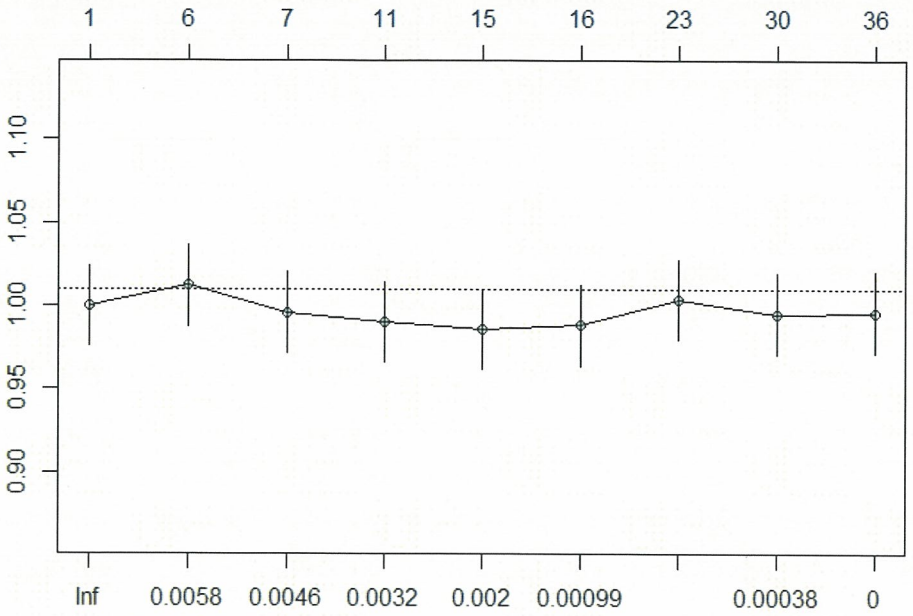
kullanımı yaparken, programlama dilinde kod yazmayan bireylerin %47'si bulut depolama kullanımı yapmaktadır.

Düğüm 33, bilgisayar ve diğer cihazlar arasında dosya aktaran bireyler (düğüm 66) ile bilgisayar ve diğer cihazlar arasında dosya aktarmayan bireyler (düğüm 67) olmak üzere iki alt düğüme ayrılmıştır. Bilgisayar ve diğer cihazlar arasında dosya aktaran bireylerin %48'i internette bulut depolama kullanırken, bilgisayar ve diğer cihazlar arasında dosya aktarmayan bireylerin %20'si internetten bulut depolama kullanmaktadır.

Düğüm 66, eğitim durumu yüksek lisans ve doktora olan bireyler (düğüm 130) ile eğitim durumu yüksek lisans ve doktora olmayan bireyler (düğüm 131) olmak üzere iki alt düğüme ayrılmıştır. Eğitim durumu yüksek lisans ve doktora olan bireylerin %62'si internetten bulut depolama kullanımı yaparken, eğitim durumu yüksek lisans ve doktora olmayan bireylerin %46'sı internetten bulut depolama kullanımı yapmaktadır.

CART algoritması ile oluşturulan 3. sınıflandırma ağacında metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşma, bir programlama dilinde kod yazımı ile eğitim durumu internette bulut depo kullanımını arttıran bağımsız değişkenlerdir.

4. sınıflandırma ağacında, veri setinin %90'ı modeli oluşturmak, %10'u da modeli test etmek için kullanılmıştır. 4. sınıflandırma ağacına ait budama kararı Şekil 38'de gösterilmiştir.



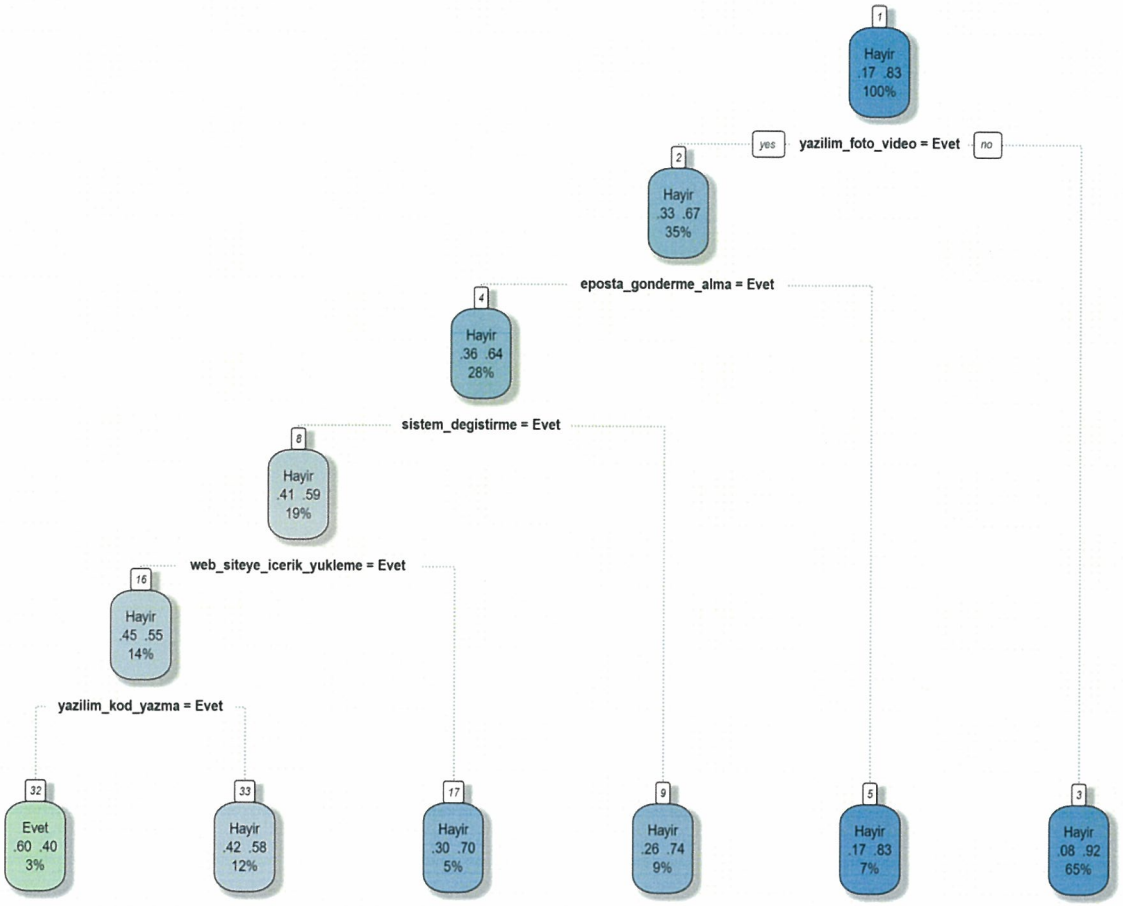
Şekil 38. 4. Sınıflandırma Ağacına Ait Budama Kararı

Gini ayırma kriteri katsayısı olarak 0.0058 kullanıldığında, optimum ağaç elde edilecektir. 0.0046 ve 0.0032 gini ayırma kriteri katsayıları kullanıldığında hatalı sınıflandırma oranı daha yüksek çıkacaktır. Budanmış sınıflandırma ağacında, etkili olan değişkenler ve hatalı sınıflandırma oranı Tablo 6'da gösterilmiştir.

Tablo 6. 4. Sınıflandırma Ağacı Modeli

```
Classification tree:
rpart(formula = bulut_depo_kullanimi ~ ., data = dataset_4, method = "class",
    control = r.ctrl)

Variables actually used in tree construction:
[1] eposta_gonderme_alma      sistem_degistirme
[3] web_siteye_icerik_yukleme yazilim_foto_video
[5] yazilim_kod_yazma

Root node error: 1422/8499 = 0.16731

n= 8499

        CP nsplit rel error xerror     xstd
1 0.0059072      0   1.00000  1.000 0.024199
2 0.0058000      5   0.97046  1.012 0.024314
```

Gini ayırma kriteri olarak, 0.0058 katsayısının kullanıldığı 4. sınıflandırma ağacının hatalı sınıflandırma oranı %16,7 olarak belirlenmiştir.



Şekil 39. CART Algoritması İle Oluşturulan 4. Sınıflandırma Ağacı

Şekil 39'da gösterilen ve CART algoritmasıyla oluşturulan 4. sınıflandırma ağacı incelendiğinde, internetten bulut depolama kullanımında birinci dereceden etkili bağımsız değişken yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme, ikinci dereceden etkili bağımsız değişken kişisel amaçlarla internet üzerinden e-posta gönderme/alma faaliyeti, üçüncü dereceden etkili bağımsız değişken işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirme, dördüncü dereceden etkili bağımsız değişken metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme, beşinci dereceden etkili bağımsız değişken bir programlama dilinde kod yazma değişkenidir. Bireylerin %17'si internetten bulut depolama kullanımı yaparken, %83'ü internetten bulut depolama kullanımı yapmamaktadır. Yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleyenler 2 numaralı sol düğüme, yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenler ise 3 numaralı sağ düğümde sınıflandırılmıştır. Yazılım kullanarak fotoğraf, video ya

da ses dosyalarını düzenleyenlerin %33'ü internetten bulut depolama kullanımı yaparken, yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenlemeyenlerin %8'i internetten bulut depolama kullanımı yapmaktadır.

3 numaralı sağ düğüm terminal düğümdür ve 2 numaralı düğümden itibaren sınıflandırma devam etmektedir. 2 numaralı düğümü saflaştırmak için bireylere, kişisel amaçlarla internet üzerinden e-posta gönderme/alma yapıp yapmadıkları sorulmuştur. Soruya verilen cevaplara uygun olarak 2. düğüm, 4 numaralı ve 5 numaralı iki alt düğüme ayrılmıştır. Düğüm 4'te kişisel amaçlarla internet üzerinden e-posta gönderme/alma faaliyetlerini gerçekleştiren bireylerin %36'sı internetten bulut depo kullanımı yaparken, düğüm 5'te ise kişisel amaçlarla internet üzerinden e-posta gönderme/alma faaliyeti gerçekleştirmeyen bireylerin %17'si internetten bulut depo kullanımı yapmaktadır.

Düğüm 4, işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştiren bireyler (düğüm 8) ve işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmeyen bireyler (düğüm 9) olmak üzere iki alt düğüme ayrılmıştır. İşletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştiren bireylerin %41'i internetten bulut depolama kullanımı yaparken, işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirmeyen bireylerin %26'sı internetten bulut depolama yapmaktadır.

Düğüm 8; metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenler (düğüm 16) ve metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yüklemeyenler (düğüm 17) olmak üzere iki alt düğüme ayrılmıştır. Metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleyenlerin %45'i internette bulut depolama kullanırken, metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yüklemeyenlerin %30'u internetten bulut depolama kullanımı yapmaktadır.

Düğüm 16, programlama dilinde kod yazan bireyler (düğüm 32) ve programlama dilinde kod yazmayan bireyler (düğüm 33) olmak üzere iki alt düğüme ayrılmıştır. Bir programlama dilinde kod yazan bireylerin %60'ı internetten bulut depolama

kullanımı yaparken, programlama dilinde kod yazmayan bireylerin %42'si bulut depolama kullanımı yapmaktadır.

CART algoritması ile oluşturulan 4. sınıflandırma ağacında metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşma ve bir programlama dilinde kod yazımı internette bulut depo kullanımını arttıran bağımsız değişkenlerdir.

7.4. Sınıflandırma Ağaçlarından Elde Edilen Sınıflandırma Kuralları

Tablo 7. Sınıflandırma Ağacı Karar Kurallarından Bazıları

Sınıf	Kural	Destek Oranı
1	Eğer yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme EVET ise ve işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirme =EVET	%40
2	Eğer işletim sistemi veya güvenlik programları da dahil olmak üzere herhangi bir yazılımın ayarlarını değiştirme EVET ise ve kişisel amaçla (iş dışında, özel) internette e-posta gönderme/alma faaliyeti = EVET	%26
3	Eğer kişisel amaçla (iş dışında, özel) internette e-posta gönderme/alma faaliyeti EVET ise ve metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme = EVET	%23
4	Eğer metin, fotoğraf, müzik, video, yazılım vb. içerikleri herhangi bir web sitesine paylaşmak üzere yükleme EVET ise ve bir program dilinde kod yazma = EVET	%17
5	Eğer hanede herhangi birinde taşınabilir bilgisayar bulunma durumu EVET ise ve bilgisayarlar arası ya da cihazlar arası dosya aktarma = EVET	%12
6	Eğer bir program dilinde kod yazma EVET ise ve eğitim durumu yüksek lisans ya da doktora =EVET	%7

Sınıflandırma ağacı sonucu oluşan karar kuralları Tablo 7'de gösterilmiştir. karar kuralları ağacın genel şekli hakkında kolaylıkla bilgi sahibi olmamızı sağlamaktadır.

8. SONUÇ

Büyük verilerin modellenmesinde veri madenciliği ve makine öğrenimi algoritmalarının kullanımı gün geçtikçe artış göstermektedir. Büyük verinin işlenebilmesi için yapay zeka popüler hale gelmiş bununla birlikte makine öğrenimiyle alakalı birçok çalışma yapılmaya başlanmıştır.

Makine öğrenimi, bilgisayar algoritmalarının gelişimine olanak sağlamış ve son zamanlarda özellikle veri bilimciler makine öğrenimi algoritmalarını kullanmaya özen gösterir hale gelmişlerdir. Makine öğreniminde karar ağaçları hem sınıflandırma hem de tahminleme yapmak için kullanılmaktadır. Karar ağaçları hem varsayımlara dayanmaması hem de yorumlanma açısından kolay olması nedeniyle çalışmalarda sıkça kullanıldığı görülmektedir.

Bu çalışmada; CART algoritmasının uygulanma nedeni, büyük veri setlerine uygulanma ve yorumlanma kolaylığı açısından kullanılan algoritma olmasıdır. Çalışmada, Türkiye İstatistik Kurumu tarafından gerçekleştirilen "Hanehalkı Bilişim Teknolojileri Kullanımı Anketi" 2017 verileri doğrultusunda sınıflandırma ağaçları oluşturulmuştur.

Çalışmanın uygulama bölümünde 2015, 2016 ve 2017 yılları verileri ile analizler yapılarak haneler ve bireylerin bilişim teknolojileri ile ilgili davranışları incelenmiştir. Daha sonra, R programlama dili arayüzü olan RStudio 3.4.4 sürümü ile sınıflandırma ağaçları oluşturulmuştur.

Ağaçların yapısı hakkında genel bir yorum yapılacak olursa, ağaçların oluşumunda 1. dereceden etkili değişken olan yazılım kullanarak fotoğraf, video ya da ses dosyalarını düzenleme değişkeni dört ağaçta da yüksek oranla etkili olmuştur. Genellikle internet üzerinden bulut depolama kullanımı üzerinde etkili olan değişkenlerin, yazılım ile ilgili faaliyetler olduğunu söylemek mümkündür.

Analizde yer alan 1. sınıflandırma ağacı, %60 eğitim verisi ile modellenerek ve %40 test verisi ile test edilerek, 2. sınıflandırma ağacı, %70 eğitim verisi ile modellenerek ve %30 test verisi ile test edilerek, 3. sınıflandırma ağacı, %80 eğitim verisi ile modellenerek ve %20 test verisi ile test edilerek oluşturulmuştur. 4. sınıflandırma ağacı, %90 eğitim verisi ile modellenerek ve %10 test verisi ile test edilerek

oluşturulmuştur. Farklı eğitim verileriyle analiz yapılmasının nedeni, düğümlerde minimum hata ile oluşmuş ağacın bulunabilmesini kolaylaştırmak içindir.

Eğitim verisindeki artış, sınıflandırma ağaçlarında yer alan düğüm sayılarını da arttırmakta ve ağaç karmaşık bir yapıya bürünmektedir. Ağacın daha kolay anlaşılabilmesi için analizler sonucu oluşan budama kararı tablosundan yararlanılarak ağaçların budanması sağlanmıştır. Budamada dikkat edilecek nokta, tüm ağaçlarda sonsuza yakın olan ayırma kriterinin kullanılmasıdır.

Sınıflandırma ağaçları incelendiğinde, eğitim verisinin oranının artış göstermesine karşılık internetten bulut depolama kullanımında birinci dereceden ve ikinci dereceden etkili bağımsız değişkenlerin farklılık göstermediği sonucuna varılmıştır.

Diğer taraftan, ağaçlara ait hatalı sınıflandırma oranları Tablo 8'de gösterilmiştir.

Tablo 8. Hatalı Sınıflandırma Oranları

1. Sınıflandırma Ağacı	2. Sınıflandırma Ağacı	3. Sınıflandırma Ağacı	4. Sınıflandırma Ağacı
%19,5	%18,5	%17,8	%16,7

Ağaçların incelenmesi ve hangi ağacın seçileceği konusunda literatürde net bir bilgi yer almamaktadır. Bu çalışmada, hangi ağacın seçileceğinin belirlenmesinde hatalı sınıflandırma oranından faydalanılmıştır. 1. sınıflandırma ağacının hatalı sınıflandırma oranı %19,5, 2. sınıflandırma ağacının hatalı sınıflandırma oranı %18,5, 3. sınıflandırma ağacının hatalı sınıflandırma oranı %17,8 ve 4. Sınıflandırma ağacının hatalı sınıflandırma oranı %16,7'dir. Eğitim verisinin %90 olarak alındığı ağaçta, hatalı sınıflandırma oranının düşük olduğu görülmektedir. Ancak, optimum ağacın seçiminde eğitim verisi yerine hatalı sınıflandırma oranını kriter olarak almak daha doğru sonuçlar elde edilmesini sağlayacaktır.

KAYNAKLAR

Aitkenhead, M. J. A., 2008. Co-Evolving Decision Tree Classification, Expert Systems with Applications, 34 (1): 1-14.

Akşahan, R., Keskin, İ., 2015. Sığırlarda Besi Sonu Canlı Ağırlığını Etkileyen Bazı Vücut Ölçülerinin Regresyon Ağacı Yöntemi ile Belirlenmesi. cilt 2, sayı 1, 53-59.

Aktürk, D., Bayramoğlu, Z., Savran, F., 2012. Sınıflandırma ve Regresyon Ağacı Yönteminin Örnek Veri Seti İle Uygulanması. 817-823.

Alan, M. A., 2014. Karar Ağaçlarıyla Öğrenci Verilerinin Sınıflandırılması. Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, cilt:28, sayı:4, 101-112.

Alan, M. A., Dündar, S., 2017. Yatırım Teşvik Verilerinin Veri Madenciliği ile Analizi. Kırıkkale Üniversitesi Sosyal Bilimler Dergisi, cilt: 7, sayı: 2, 119-130.

Albayrak, A. S., Yılmaz, Ş. K., 2009. " Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama-Data Mining: Decision Tree Algorithms and An Application On ISE Data". Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14 (1), 31-52.

Altunkaynak, B., 2017. Veri Madenciliği Yöntemleri ve R Uygulamaları. Seçkin Yayıncılık, 16.

Altunkaynak, B., 2017. Veri Madenciliği Yöntemleri ve R Uygulamaları. Seçkin Yayıncılık, 22.

Atalay, M., Çelik, E., 2017. Büyük Veri Analizinde Yapay Zeka ve Makine Öğrenmesi Uygulamaları. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 161.

Atasever, Ü. H., Özkan, C., 2012. Arazi Örtüsünün Belirlenmesinde Torbalama-Karar Ağaçları Yönteminin Kullanımı, IV. Uzaktan Algılama ve Cografi Bilgi Sistemleri Sempozyumu, Zonguldak.

Avcı, M. A., Altay, N. O., 2014. Finansal Krizlerin Öngörüşünde Regresyon Ağaçları Modeli: Gelişmekte Olan Ülkelere Yönelik Bir Analizi. Uluslararası İktisadi ve İdari İncelemeler Dergisi, 12; 191-212.

Ayık Y. Z., Özdemir, A., Yavuz, U., 2007. Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi, Sosyal Bilimler Enstitüsü Dergisi, cilt 10, no. 2, 441-454.

Ayşe Oğuzlar, 2004. CART Analizi İle Hanehalkı İşgücü Anketi Sonuçlarının Özetlenmesi. Uludağ Üniversitesi İktisadi ve İdari Bilimler Dergisi.

Baschab, J., Piot, J., 2007. The Executive's Guide to Information Technology (2 ed.): John Wiley & Sons, Inc., Hoboken, New Jersey.

Bounsaythip, Catherine & Esa Rinta-Runsala, 2001. "Overview of Data Mining For Customer Behavior Modeling". VTT Information Technology Research Report, Version: 1, 21.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2): 123-140.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J., 1984. Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks.

Chang, L.Y., Wang, H.W., 2006. "Analysis of traffic injury severity: an application of non-parametric classification tree techniques". Accident Analysis and Prevention 38, 1019-1027.

Chien, C. F., Chen. L. F., 2008. "Data Mining to Improve Personnel Selection and Enhance Human Capital: A Case Study in High-Technology Industry," Expert Systems with Applications, vol. 34, 280-290.

Çalış, A., Kayapınar, S., Çetinyokuş, T., 2014. Veri Madenciliğinde Karar Ağacı Algoritmaları İle Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama. Endüstri Mühendisliği Dergisi, Cilt:25, Sayı 3-4, 2-19.

Çelikten, Ö., Karacan, H., 2013. Constructing a Decision-Prediction Mechanism for Dynamic Data. Marmara Üniversitesi Fen Bilimleri Dergisi, 25(3).

Dener, M., Dörterler, M., Orman A., 2009. Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama. XI. Akademik Bilişim Konferansı Bildirileri, Harran Üniversitesi.

Dolgun vd., 2009. Veri Madenciliğinde Yapısal Olmayan Verilerin Analizi: Metin ve Web Madenciliği, İstatistikçiler Dergisi 2, 48-58.

Dunham, M. H., 2003. Data Introductory and Advanced Topics. New Jersey: Pearson Education, Inc.

Emel, G. G., Taşkın, Ç., 2005. Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması. Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 6 (2), 221-239.

Ersöz, T., Özseven, T., 2015. KOBİ'lerin Finansal Sorunlarını Etkileyen Faktörlerin CRT Karar Ağacı ile Modellenmesi. Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 15(1).

Fayyad, U., Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. American Association for Artifical Intelligence, vol. 17., 37-54.

Güneri, Ö. İ., Aydın, D., 2017. Grup Üyelerini Belirlemede İstatistiksel Sınıflandırma Yöntemleri: Karşılaştırmalı Bir Çalışma. sayı: 9, 45-67.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1998. Multivariate Data Analysis. 5. Baskı, Prentice Hall, New Jersey.

Han, J., Kamber, M. and Pei, J., 2006. Data Mining: Concepts and Techniques. Morgan Kauffman.

Han, Jiawei, Kamber&Micheline, 2006. Data Mining: Concepts and Techniques, Second Edition. Morgan Kaufmann Publications, San Francisco.

J. Gehrke, 2003., "Decision Trees", The Handbook of Data Mining", Editör: Nong Ye, Lawrence Erlbaum Associates Publishers, London, 149-175.

Jarošík, V., 2011. CART and related methods. In: Simberloff D, Rejmánek M. Encyclopaedia of Biological Invasions. University of California Press, Berkeley and Los Angeles, 104–108.

Kalaycı, Ş., 2016. SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri, 349.

Kavzoğlu T., Şahin E. K., Çölkesen İ., 2010, CBS Tabanlı Çok Kriterli Karar Analizi Yöntemiyle Heyelan Duyarlılık Haritasının Üretilmesi: Trabzon İli Örneği, 3. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu, 11-13 Ekim, Gebze.

Kaya, Y., Ertuğrul, Ö. F., Tekin, R., 2012. Epileptik EEG İşaretlerinin Sınıflandırılmasında Karar Kuralları ve Karar Ağaçlarının Kullanılması, Batman Üniversitesi Yaşam Bilimleri Dergisi, 403-413.

Kayri, M., 2014. Karar Ağaçları. Lisansüstü Ders Notları.

Kayri, M., Boysan M., 2008. Assesment of Relation Between Cognitive Vulnerability and Depression's Level By Using Classification and Regression Tree Analysis. Hacettepe University Journal of Education, vol. 34, 168-177.

Kayri, M., Boysan, M., 2008. Bilişsel Yatkınlık ile Depresyon Düzeyleri İlişkisinin Sınıflandırma ve Regresyon Ağacı Analizi ile İncelenmesi. cilt:34, sayı:34, 168-177.

Kayri, M., Gökdaş, İ., 2006. Karışımlı Model Analiz Tekniğinin Eğitim Bilimleri Araştırmalarında Uygulanabilirliği Üzerine Bir Araştırma Örneği. Kuram ve Uygulamada Eğitim Bilimleri Dergisi, 6(3), 753-778.

Keskin, S., Ankaralı, H., Noyan, T., Kamacı, M., 2007. Çok Değişkenli Varyans Analizinde Gruplar Arasındaki Farkın Tespiti: Bir Uygulama. Türkiye Klin., Tıp Bil., Der., 27(6), 838-845.

Kothari R., Dong, M., 2001. "Decision Trees for Classification: A Review and Some New Results", Pattern Recognition: From Classical to Modern Approaches, Editör: S. K. Pal, A. Pal, World Scientific, New Jersey.

Kotsiantis, S.B., 2013. "Decision Trees: A Recent Overview", Artificial Intelligence Review, 39(4), 261-283.

Kuhn, M. & Johnson, K., 2013. Applied Predictive Modeling. New York. Springer.

Kurt I., Ture M., Kurum A. T., 2008. Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease. Expert Systems with Applications, 34: 366-374.

Kuyucu, Y. E., 2012. Lojistik Regresyon Analizi (LRA), Yapay Sinir Ağları (YSA) ve Sınıflandırma ve Regresyon Ağaçları (C&RT) Yöntemlerinin Karşılaştırılması ve Tıp Alanında Bir Uygulama. Gaziosmanpaşa Üniversitesi Sağlık Bilimleri Enstitüsü, (Yüksek Lisans Tezi, basılmamış), Tokat.

Küçükoğlu, O., 2010. Veri Madenciliği Yöntemlerinin Hayvancılıkta Kullanımı. Yüksek Lisans Tezi, Çanakkale Onsekiz Mart Üniversitesi, Fen Bilimleri Enstitüsü, Çanakkale.

Li J., Fu, A. W., Fahey, P., 2009. "Efficient discovery of risk patterns in medical data." Artificial Intelligence in Medicine, 45: 77-89.

Loh, W. Y., & Shih, Y. S., 1997. Split selection methods for classification trees. Statistica Sinica, 7, 815-840.

Orekici Temel, G., Çamdeviren, H., Yazıcı, A. C., 2005. "Regresyon Modellerine Alternatif Bir Yaklaşım: MARS", VIII. Biyoistatistik Sözlü Bildiriler, Bursa, 105-123.

Özkan Y., 2013. "Veri Madenciliği Yöntemleri", Papatya Yayıncılık Eğitim.

Özkan, K., 2012. Sınıflandırma ve Regresyon Ağacı Tekniği (SRAT) ile Ekolojik Verinin Modellenmesi. Süleyman Demirel Üniversitesi Orman Fakültesi Dergisi, sayı: 13, 1-4.

Pang-Ning Tan, Steinbach, M., Kumar, V., Introduction to Data Mining. (USA: Pearson Education, 2006), 2.

Papazoglou, M. P., Traverso, P., Dustdar, S., & Leymann, F., 2008. SERVICEORIENTED COMPUTING: A RESEARCH ROADMAP. International Journal of Cooperative Information Systems, 17(2), 223-255.

Pehlivan, G., 2006. Chaid Analizi ve Bir Uygulama Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Quinlan J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kauffman Publishers Inc. San Francisco, CA, USA.

Quinlan J.R., 1996. "Improved use of continuous attributes in C4.5" Journal of Artificial Intelligence Research, 4: 77-90.

Quinlan, J. R., 1986. Induction of Decision Trees. Machine Learning, 81-106.

Ratner, B., 2000. Chaid for interpreting a logistic regression model. Journal of Targeting, Measurement and Analysis of Marketing, 4, 16-29.

Satıcı, Ö., Akkuş, Z., Alp, A., 2009. Tıp Fakültesi Öğretim Elemanlarının Teknolojiye Ilişkin Tutumlarının Chaid Analizi ile İncelenmesi. Dicle Tıp Dergisi, Sayı:36, 267-274.

Sezer, E. A., Bozkır, A. S., Yağız, S., & Gökçeoğlu, C., 2010. Karar Ağacı Derinliğinin CART Algoritmasında Kestirim Kapasitesine Etkisi: Bir Tünel Açma Makinesinin İlerleme Hızı Üzerinde Uygulama. Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu.

Shahnaz, F., 2006. Decision Tree Based Algorithms. Michael W. Berry (Ed.), Lecture Notes in Data Mining, USA: World Scientific Publisher.

Silahtaroğlu, G., 2008. Kavram ve Algoritmalarıyla Temel Veri Madenciliği. Papatya Yayıncılık Eğitim, İstanbul, 33: 45-47.

SPSS Inc. (1998). AnswerTree user's guide.Chicago: SPSS Inc.

Sultan, N. A., 2010. Reaching for the "cloud": How SMEs can manage. International Journal of Information Management- Sultan, 2010.

Şatır, E., Azboy, F., Aydın, A., Arslan, H., Hacıefendioğlu, Ş., 2016. Veri İndirgeme ve Sınıflandırma Teknikleri ile Glokom Hastalığı Teşhisi. El-Cezeri Fen ve Mühendislik Dergisi, cilt: 3, no. 3, 485-497.

Takma, Ç., Gevrekçi, Y., Karahan, A. E., Atıl, H., Çevik, M., 2017. Yumurta Verimi Üzerine Bazı Özelliklerin Etkisinin Regresyon Ağacı ile Belirlenmesi. Ege Üniversitesi Ziraat Fakültesi Dergisi.

Temel G,O., Çamdeviren, H., Akkuş, Z., 2005. Sınıflama Ağaçları Yardımıyla Restless Legs Syndrome (RLS) Hastalarına Tanı Koyma. İnönü Üniversitesi Tıp Fakültesi Dergisi; 12(2): 111-117

Temel, G. O., 2004. Sınıflandırma ve Regresyon Ağaçları. Mersin Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı Yüksek Lisans Tezi.

TÜİK, 2017. http://www.TÜİK.gov.tr

Uysal, İ., Bilen, M., Ulukuş, S., 2014. Twoing Algoritması İle Sınıflandırma: Kalp Hastalığı Uygulaması. Akademik Bilişim Konferansı Bildirileri, Mersin.

Üngüren, E., Doğan, H., 2010. Beş Yıldızlı Konaklama İşletmelerinde Çalışanların İş Tatmin Düzeylerinin CHAID Analiz Yöntemiyle Değerlendirilmesi. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, cilt 11, sayı 2, 39-52.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd. Edition. Elsevier Inc, Amsterdam.

ÖZGEÇMİŞ

Adı Soyadı : Güner Gözde TEKSİN

Doğum Yeri ve Yılı : Çankaya/Ankara, 23/07/1993

Yabancı Dili : İngilizce

E-posta : ggteksin@gmail.com

Eğitim Durumu

Lise : Bakırköy Anadolu Lisesi, 2011

Lisans Bölümü : İstanbul Ticaret Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik

Yüksek Lisans : İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı

Mesleki Deneyim

TC. Dışişleri Bakanlığı Stratejik Araştırmalar Merkezi, Stratejik Araştırmalar Stajyeri	2014
TC. Enerji ve Tabii Kaynaklar Bakanlığı, Enerji İstatistikleri Stajyeri	2015
Batik Pazarlama Planlama Uzman Yardımcısı	2017-2018
Akademetre Araştırma ve Stratejik Planlama Proje Sorumlusu	2018- Devam Ediyor

Yayınları

Teksin, G.G., Turanli, M., 2018, Classification of Household Users Using Information Technologies Based on C5.0 Algorithm, Eurasian Journal of Business and Management, vol. 6, No: 2, pp. 33-47.