



T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİNDEN KARAR AĞAÇLARI VE BİR UYGULAMA

Nuriye TOPYILDIZ

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

Kasım-2019
KONYA
Her Hakkı Saklıdır

ÖZET

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ SINIFLANDIRMA YÖNTEMLERİNDEN KARAR AĞAÇLARI VE BİR UYGULAMA

Nuriye TOPYILDIZ

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Yunus AKDOĞAN

2019, 48 Sayfa

Jüri
Dr. Öğr. Üyesi Yunus AKDOĞAN
Doç. Dr. Neslihan İYİT
Dr. Öğr. Üyesi İlkay ALTINDAĞ

Bilişim teknolojisi çok büyük miktarlardaki veriyi saklamayı mümkün kılmaktadır. Biriken veri üzerinde analizler yapmak amacıyla çeşitli matematiksel ve istatistiksel yöntemler kullanılabilmektedir. Fakat veri miktarı arttıkça sorunlar ortaya çıkmaktadır. Bu veriyi yönetmek ve veriyi çözümleyerek "yararlı bilgiye" erişilmesini sağlayabilmek için veri madenciliği yöntemi geliştirilmiştir. Veri madenciliği, en basit tanımıyla, toplanan veriyi bilgiye dönüştürme tekniklerinin bütünüdür. Madencilik teriminin kullanılmasının nedeni ise, büyük miktarda verinin içerisinden bilgiye dönüşecek uygun veriyi arama işleminin maden arama işiyle benzerlik göstermesidir. Bu çalışmanın önemi; günümüzde oldukça önemli ve de ihtiyaç duyulan bir alan haline gelen veri madenciliğinin sınıflandırma yöntemlerinden olan karar ağaçlarını derinlemesine ele alıyor olmasıdır. Karar ağaçlarında, değişken tiplerine göre farklı algoritmalar kullanılabilmektedir. C5.0, CART ve CHAID en popüler algoritmalardır. Bu çalışmada da, C5.0, CART ve CHAID algoritmaları derinlemesine incelenmiştir. The UCI Machine Learning Repository veritabanından temin edilen mantar veri setinde yer alan her bir gözlem, veri madenciliği sınıflandırma yöntemlerinden karar ağaçları algoritmaları kullanılarak zehirli veya yenilebilir şeklinde sınıflandırılmaya çalışılmıştır.

Anahtar Kelimeler: C5.0, CART, CHAID, Karar ağaçları, Veri madenciliği

ABSTRACT

MS THESIS

DATA MINING CLASSIFICATION METHOD DECISION TREES AND AN APPLICATION

Nuriye TOPYILDIZ

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER
IN STATISTICS

Advisor: Asst. Prof. Dr. Yunus AKDOĞAN

2019, 48 Pages

Jury
Asst. Prof. Dr. Yunus AKDOĞAN
Assoc. Prof. Neslihan İYİT
Asst. Prof. Dr. İlkay ALTINDAĞ

Information technology makes it possible to store huge amounts of data. Various mathematical and statistical methods can be used to analyze the accumulated data. However, as the amount of data increases, problems arise. Data mining method has been developed to manage this data and provide accessibility to "useful information" via analyzing the data. Data mining, in its simplest definition, is a collection of techniques for converting collected data into information. The reason for using the term mining is that the process of searching for the appropriate data to be transformed into information from a large amount of data is similar to that of mining. The importance of this study; the fact that data mining, which has become a very important and needed field in today's world, covers the decision trees which is one of the classification methods in depth. Different algorithms can be used in decision trees according to type of variables. C5.0, CART and CHAID are the most popular algorithms. In this study, C5.0, CART and CHAID algorithms are examined in depth. Each observation in the mushroom data set obtained from The UCI Machine Learning Repository database has been tried to be classified as poisonous or edible by using decision trees algorithms which are data mining classification methods.

Keywords: C5.0, CART, CHAID, Decision trees, Data mining

ÖNSÖZ

Öncelikle, çalışmamın her aşamasında, kıymetli zamanını ayırıp sabırla benden desteğini bir an bile esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren saygıdeğer tez danışmanım Sayın Dr. Öğr. Üyesi Yunus AKDOĞAN'a teşekkürü bir borç bilip şükranlarımı sunarım.

Bu zorlu süreçte her zaman yanımda olan, çalışmamın stresli, zor günlerinde anlayış ve yardımını hiçbir zaman eksik etmeyen değerli eşime, eğitim öğretim hayatım boyunca hep yanımda olup destek veren sevgili ağabeyime, hayatları boyunca evlatlarını her şeyin üstünde tutan, fedakârca ve karşılıksız iyilikleri ile her daim sonuna kadar evlatlarına destek olan ve haklarını hiçbir zaman ödeyemeyeceğim kıymetli anneme ve babama şükranlarımı sunarım.

Nuriye TOPYILDIZ
KONYA-2019

İÇİNDEKİLER

ÖZET ..... **iv**

ABSTRACT ..... **v**

ÖNSÖZ ..... **vi**

İÇİNDEKİLER ..... **vii**

1. GİRİŞ **1**

2. VERİ MADENCİLİĞİ VE TARİHSEL GELİŞİMİ **4**

2.1. Veri Madenciliğinin Tarihi Gelişimi 4
2.2. Veri Madenciliği 7

3. KAYNAK ARAŞTIRMASI **12**

4. MATERYAL VE YÖNTEM **15**

4.1. Veri Madenciliği Yöntemleri 20
4.1.1. Sınıflandırma ve regresyon 22
4.1.1.1. C5.0 algoritması 28
4.1.1.2. CART algoritması 30
4.1.1.3. CHAID algoritması 31

5. UYGULAMA **34**
5.1. C5.0 Algoritması ile Elde Edilen Sonuçlar 34
5.2. CART Algoritması ile Elde Edilen Sonuçlar 36
5.3. CHAID Algoritması ile Elde Edilen Sonuçlar 38

6. SONUÇLAR VE ÖNERİLER **43**

KAYNAKLAR **45**

ÖZGEÇMİŞ **48**

1. GİRİŞ

Amerika Birleşik Devletleri (ABD) merkezli olan International Data Corporation (IDC) market araştırmaları şirketinin dünya çapında yaptığı araştırmanın sonuçlarına göre 2013 ila 2020 yılları arasında üretilecek olan veri tutarı 44 zettabayt olarak kestirilmektedir ve bu miktar, 2013 yılına kadar üretilen bütün verinin 10 katı büyüklüğündedir (Altun, 2017). Verinin kapasitesi yüksek dijital ortamlarda saklanmaya ve işlenmeye başlanmasıyla birlikte, veriyi saklamakta ulaşmakta kolaylaşmıştır. Ancak, bu gelişmeler gerçekleşirken bilim insanlarının, analistlerin, mühendislerin yani toplanan veriyi işleyecek ve bilgiyi elde edecek kişilerin sayısı artış göstermemektedir (Vahaplar ve İnceoğlu, 2001).

Bilişim teknolojisi çok büyük miktarlardaki veriyi saklamaya yeterli olabilmektedir. Biriken veri üzerinde analizler yapmak amacıyla da çeşitli matematiksel ve istatistiksel yöntemler kullanılabilmektedir. Fakat veri miktarı arttıkça sorunlar ortaya çıkmaktadır. İşte bu veriyi yönetmek ve veriyi analiz ederek "yararlı bilgiye" erişilmesini sağlayabilmek için veri madenciliği yöntemi geliştirilmiştir. Veri madenciliği, en basit tanımıyla, toplanan veriyi bilgiye dönüştürme tekniklerinin bütünüdür. Madencilik teriminin kullanılmasının nedeni ise, büyük miktarda verinin içerisinden bilgiye dönüşecek uygun veriyi arama işleminin maden arama işiyle benzerlik göstermesidir. Veri madenciliği aslında klasik istatistiksel yöntemlere oldukça fazla benzemektedir. Fakat, klasik istatistiksel metotlar yeterince derlenmiş ve genellikle özet veriler üzerinde uygulanmaktadır ve kayıtlar binlerce olabilmektedir. Ancak, veri madenciliğinde, içerisinde çok fazla değişken bulunan milyonlarca hatta milyarlarca veri ile çalışılmaktadır (Özkan, 2008).

Veri madenciliği; anlamlı ilişkiler ve de kurallar bulabilmek için büyük boyutlu verinin analizi ve keşfi, verinin yararlı özetinin ortaya çıkarılması, veriden anlamlı bilgi, güvenilir karar ve neticelere varılması için verinin dönüşümü ve verinin analizi ile veri kaynağındaki ilişki ve benzerlikleri bulmak için kullanılan tekniklerin bütünüdür (Ergün, 2008). Veri madenciliği, bir süreçtir. Teoride, veri madenciliği, bilgi keşfi sürecinin adımlarından birisidir. Pratikte ise bilgi keşfi ile veri madenciliği anlamdaş olarak kullanılmaktadır. Veri madenciliğinin işlevleri; adli analiz (forensic analysis), keşif (discovery) ve tahmini modelleme (predictive modeling) şeklinde üç sınıf altında toplanmaktadır.

Bugün karar verme prosesine gereksinim duyulan pek çok alanda veri madenciliği yöntemleri yaygın biçimde kullanılmaktadır. Borsa, sigortacılık, bankacılık, telekomünikasyon, pazarlama, perakendecilik, mühendislik, biyoloji, istihbarat, tıp, genetik, kriminoloji, endüstri vb. pek çok dalda başarılı uygulamaları görülmektedir (Savaş ve ark., 2012).

Son 20 yıldır ABD'de türlü veri madenciliği yöntemlerinin gizli dinlemeden, vergi kaçakçılıklarının meydana çıkarılmasına varana kadar türlü uygulamalarda kullanıldığı bilinmektedir. Çalışmalar incelendiğinde, veri madenciliği yöntemlerinin en yaygın olarak kullanıldığı alanlar biyoloji, tıp ve genetiktir (Akpınar, 2000).

Veri madenciliğinin temeli, ilk sayısal bilgisayar olan "ENIAC'a (Electrical Numerical Integrator And Calculator)" dayanmaktadır. 1946 yılında geliştirilmiş olan ve günümüzde kullandığımız bilgisayarların atası sayılabilen ENIAC, ABD'li bilim insanları J. Presper Eckert ile John Mauchly tarafından, II. Dünya Savaşı'nda ABD ordusunun kullanması için geliştirilmiştir. Bu "ilk" bilgisayar, 30 ton ağırlığında, 170 m^2 'lik alanı kaplayan bir cihazdır (Öğüt, 2014).

Veri madenciliği yöntemlerini denetimli (supervised) ve de denetimsiz (unsupervised) şeklinde iki ana kategoriye ayırabilmek mümkündür. Denetimli veri madenciliği yöntemleri, veriden bilgi ve sonuç elde etmeye binaen kullanılmaktadır. Denetimsiz veri madenciliği yöntemleri ise, eldeki veriyi anlama, tanıma ve keşfetme amacıyla kullanılan ve de sonraki aşamalarda uygulanacak metotlar için fikir vermeyi amaçlayan yöntemlerdir.

Veri madenciliğinde kullanılmakta olan modeller; tahmin edici (predictive) modeller (sınıflandırma, zaman serileri, vd.) ve tanımlayıcı (descriptive) modellerdir (kümeleme, birliktelik analizi, vd.). Veri madenciliği modellerini üç temel başlıkta incelemek mümkündür. Sınıflandırma ve regresyon; karar ağaçları, bayesyen sınıflandırma, yapay sinir ağları, vd. algoritmalar kullanılır. Kümeleme; k-ortalamalar, kohonen ağlar, en yakın komşu, iki aşamalı kümeleme, en uzak komşu, vd. algoritmalar kullanılır. Birliktelik kuralları; apriori, apriori tid, vd. algoritmalar kullanılır (Akpınar, 2000).

Nobel ödüllü Dr.Penzias 1999 yılında vermiş olduğu röportajında veri madenciliğinin ehemmiyetini şu sözlerle belirtmektedir; "Veri madenciliği oldukça önemli bir konuma gelecektir ve firmalar müşterileriyle ilgili hiçbir bilgiyi atamayacaklardır, çünkü bu bilgiler oldukça değerli olacaktır. Eğer siz müşterilerinizin

bilgilerini değerlendirmiyorsanız, firmanız sektörünüzün dışında kalacaktır.” (Annoni ve ark., 2009).

Bu çalışmanın önemi; günümüzde kendisi oldukça önemli ve de ihtiyaç duyulan bir alan haline gelen veri madenciliğinin sınıflandırma yöntemlerinden olan karar ağaçlarını derinlemesine ele alıyor olmasıdır.

Çalışmanın uygulama kısmındaysa, internet üzerinden erişim sağlanan ücretsiz, açık kaynak olan The UCI Machine Learning Repository (UCI Makine Öğrenimi Deposu, https://archive.ics.uci.edu/ml/index.php) veritabanından temin edilen mantar veri seti analiz edilmiştir (Lincoff, 1981). Veri setinde yer alan her bir gözleme (mantara) ilişkin 22 bağımsız değişkenin değeriyle mantarların zehirli veya yenilebilir olduğu verisi kullanılarak mantarlar veri madenciliği sınıflandırma yöntemlerinden karar ağaçları algoritmaları ile zehirli veya yenilebilir şeklinde sınıflandırılmaya çalışılmıştır. Karar ağaçları algoritmalarından yaygın olarak kullanılan C5.0, CART ve CHAID algoritması çalışmada tercih edilmiştir.

Çalışma, giriş bölümü ile birlikte altı bölümden oluşmaktadır. İkinci bölümde; veri madenciliği ve tarihsel gelişimi ayrıntılı olarak ele alınmıştır. Üçüncü bölümde, literatür taramasına yer verilmiştir. Dördüncü bölümde, çalışmada kullanılan materyaller ve yöntemler geniş bir şekilde açıklanmıştır. Beşinci bölümde analiz sonuçları verilerek yorumlanmış ve altıncı bölüm olan sonuç ve öneriler bölümünde ise değerlendirmelerde bulunulmuştur.

2. VERİ MADENCİLİĞİ VE TARİHSEL GELİŞİMİ

Bu bölümde, veri madenciliğinin tarihsel gelişimine, tanımına, önemine, kullanıldığı alanlara, işlevlerine, bilgi keşfi veya veri madenciliği sürecinin aşamalarına, diğer disiplinlerle ilişkisine ve veri madenciliğinde karşılaşılan problemlere yer verilmektedir.

2.1. Veri Madenciliğinin Tarihi Gelişimi

Veri madenciliğinin kökenleri, ilk nümerik bilgisayar olan "ENIAC'a (Electrical Numerical Integrator And Calculator)" dayanmaktadır. 1946 'da geliştirilmiş olan ve bugün kullanılan bilgisayarların atası sayılabilen ENIAC, ABD'li bilim insanları J. Presper Eckert ile John Mauchly tarafından, II. Dünya Savaşı'nda ABD ordusunun kullanması için geliştirilmiştir. Bu "ilk" bilgisayar, 30 ton ağırlığında, 170 m^2 'lik alanı kaplayan bir cihazdır (Öğüt, 2014). 1950'li yıllarda kullanılmaya başlayan ilk bilgisayarlar sayım için kullanılmaya başlanmıştır (Savaş ve ark., 2012).

Veri madenciliği kavramı, 1960'larda bilgisayarların veri analizi yapmak için kullanılmasıyla ortaya çıkmış ve veri tabanı, verilerin depolanması kavramları teknoloji dünyasındaki yerini almıştır. O dönemlerde, bilgisayar aracılığıyla, yeteri kadar tarama yapıldığında, istenilen veriye ulaşmanın muhtemel olacağı kabul edilmiştir. Bu işleme ilk başlarda veri madenciliği ismi değil, veri taraması (data dredging) ya da veri yakalaması (data fishing) şeklinde isimler verilmiştir (Öğüt, 2014). 1960'ların sonunda, bilim insanları basit kuralları öğrenebilen bilgisayarlar geliştirmişlerdir. Minsky ve Papert, bugün sinir ağları diye bilinmekte olan perseptronların, çok basit kuralları öğrenebileceğini ortaya koymuşlardır (Savaş ve ark., 2012).

1970'li yıllarda "ilişkisel veri modelleri", buna bağlı olarak "ilişkisel veritabanı yönetim sistemleri (ilişkisel VTYS)" uygulamaları kullanılmaya başlanmıştır. Uzmanlar bu sistemle beraber, basit kuralları kullanan bilgisayar sistemleri geliştirmişler ve çok basit manada makine öğrenimini gerçekleştirmişlerdir. 1980'lerde yaygınlaşan veri tabanı yönetim sistemleri mühendislik, vb. alanlarda kullanılmaya başlanmıştır. Yine bu yıllarda firmalar, müşterileri, rakip şirketleri ve arz edilen ürünleri ile ilgili verilerden meydana gelen veri tabanlarını oluşturmuşlardır. Oluşturulan bu veri tabanları içerisinde çok fazla miktarda veri bulunmaktadır. Bu veri tabanlarına SQL veri tabanı dili gibi benzeri veri tabanı dilleri kullanılarak ulaşılabilmektedir. 1990'larda toplanan veri miktarının katlanarak artmasıyla, faydalı bilgiye kolaylıkla nasıl ulaşılabileceği

düşünülmeye başlanmıştır. Bu itibarla, çalışmalara ve yayınlara başlanmıştır. Veri madenciliği ismi de yine o yıllarda bilgisayar mühendisleri tarafından ortaya atılmıştır. 1989 yılında yapılan "Knowledge Discovery from Databases / KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu Toplantısı" nın ve 1991 yılında yapılan KDD (IJCAI)-89'un sonuç bildirgesi niteliğindeki "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesinin "KDD (Knowledge Discovery and Data Mining)" ile ilgili başlıca tanım ve kavramları ortaya koymasıyla süreç ivmelenerek hızlanmıştır (Savaş ve ark., 2012). Bilim insanları, bugün kendisi bir çalışma alanı olan veri madenciliğine, türlü yaklaşımlar getirmeye başlamışlardır. Bu yaklaşımların temelinde; istatistik, veritabanları, makine öğrenimi, otomasyon ve araştırma gibi farklı disiplinler ve kavramlar vardır (Öğüt, 2014).

Veri madenciliği için ilk yazılım 1992 yılında geliştirilmiştir. 1995 yılında, "Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansı (KDD-95)" yapılmıştır. 2000'lerde, veri madenciliği sürekli gelişmiş neredeyse bütün alanlara uygulanır olmuştur. Elde edilen sonuçlardan sağlanan faydalar ortaya çıktıkça bu alana duyulan ilgi giderek artmıştır (Savaş ve ark., 2012).

Veri madenciliğinin tarihsel gelişimi yıllara göre; 1950'ler, ilk bilgisayarların (sayımlar için) kullanılmaya başlanması; 1960'lar, veritabanı ve verilerin depolanması kavramının ortaya çıkması; 1970'ler, ilişkisel veri modeli ve ilişkisel veritabanı yönetim sistemleri (VTYS) uygulamalarının kullanılması; 1980'ler, ilişkisel VTYS yaygınlaşması ve bilimsel alanlarda uygulanmaya başlanması; 1990'lar, veri madenciliği alanında ilk yazılımın geliştirilmesi ve 2000'ler, veri madenciliğinin tüm alanlarda uygulamalarının yaygınlaşması şeklinde özetlenebilir. Veri madenciliğinin gelişme süreci Tablo 2.1 'de de belirtilmektedir.

Tablo 2.1. Veri madenciliğinin gelişimi (Aldana, 2000)

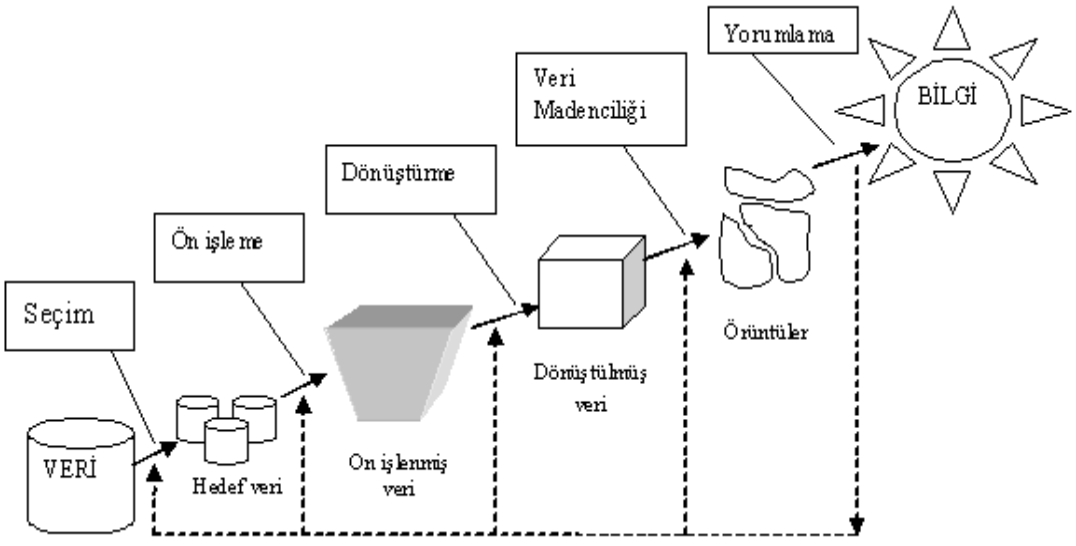
Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960'lar)	Benim toplam kârım geçen beş yılda ne kadardı?	Bilgisayarlar, Bantlar (Teypler), Diskler	IBM, CDC	Geriye dönük, statik veri dağıtımı
Veri Erişimi (1980'ler)	İngiltere'de geçen Mart ayında birim satışlar ne kadardı?	İlişkisel veritabanı, SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	İngiltere'de geçen Mart ayında birim satışlar ne kadardı? Boston'a inceleyin.	OLAP, Çok boyutlu veritabanı sistemleri, Veri ambarları	Pilot, Comshare, Arbor, Cognos, Microstrategy	Çoklu düzeylerde geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir, neden?	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veritabanları	Pilot, Lockheed, IBM, SGL, çok sayıda yeni şirket (yeni oluşan sanayi)	Geleceğe dönük, proaktif enformasyon dağıtımı

2.2. Veri Madenciliği

Piatetsky-Shapiro tarafından, veri madenciliği; "verilerden, daha önceden bilinmeyen, zımnî, muhtemelen faydalı enformasyonun monoton olmayan bir süreçte çıkartılması işlemi" şeklinde adlandırılmaktadır (Akpınar, 2000). Veri madenciliği, veri içerisindeki ilişkilerin, örüntülerin, kuralların, değişimlerin ve düzensizliklerin ve de istatistiksel olarak önem arz eden yapıların yarı otomatik biçimde keşfedilmesidir. Veriler arasındaki ilişki ve kurallar bilgisayar aracılığıyla belirlenmektedir. Buradaki amaç, kolaylıkla bilgiye veya görsel sonuçlara çevrilebilecek modellerin elde edilmesidir. Bu bağlamda, veri madenciliğinin merkezinde insan olmakla birlikte, bazen insan ve bilgisayar arayüzü birleşmektedir (Vahaplar ve İnceoğlu, 2001).

Veri madenciliği, veriyi toplayan kişi tarafından anlaşılacak ve faydalı olacak şekilde, çoğu zaman büyük boyutlu veri setlerinden, daha önceden bilinmeyen ilişkilerin ve bilgilerin tespit edilmesi ve verinin metin olarak özetlenmesi amacını taşımaktadır. Veri madenciliği çalışması neticesinde tespit edilen ilişkiler ve veriden elde edilen metinler çoğunlukla model ya da örüntü olarak isimlendirilmektedir. Veri madenciliği tek başına çözüme ulaştıran bir yöntem olmayıp karar verme sürecini destekleyen ve hızlandıran, problemi çözmek için ihtiyaç duyulan bilgileri elde etmeyi sağlayan bir yöntemdir.

Veri madenciliği, bir süreçtir. Teoride, bilgi keşfi sürecinin adımlarından birisi de veri madenciliğidir. Pratikte ise bilgi keşfi ve veri madenciliği anlamdaş olarak kullanılmaktadır. Bilgi keşfi prosesinde, veri madenciliği adımı Şekil 2.1 'de gösterilmektedir.



Şekil 2.1. Bilgi keşfi sürecinde veri madenciliğinin yeri (Koyuncugil ve Özgülbaş, 2009)

Bilgi keşfi sürecinin aşamaları sırasıyla aşağıdaki gibi açıklanabilir (Albayrak ve Koltan Yılmaz, 2009).

Araştırma probleminin tanımlanması aşaması; araştırmanın amacı çerçevesinde mevcut durum değerlendirilmesini, problem için uygulanacak veri madenciliği yönteminin amacını ve planlama sürecini kapsamaktadır.

Önişleme aşamasında; veri toplamakla başlayan süreç, benzer veri setlerini bir araya getirmek, veri niteliklerini keşfetmek ve gizli bilgileri sınıflandırmayla devam etmektedir.

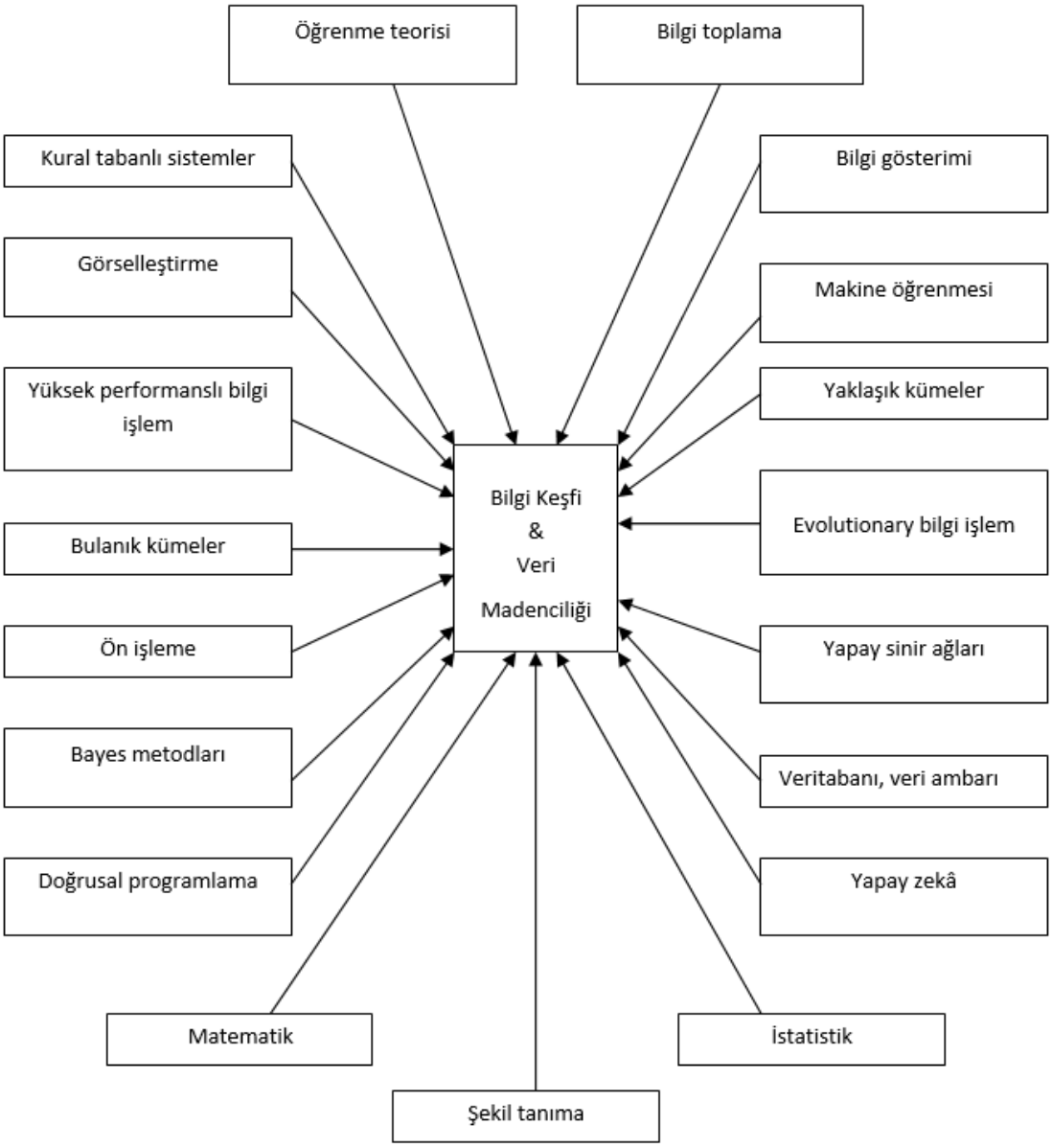
Veri dönüştürme aşaması; veri dönüştürme ve modelleme için veri temizleme sürecini kapsamaktadır.

Veri madenciliği aşaması; uygun modelleme yönteminin belirlenmesi, test modelin üretimi, modeli iyileştirme ve tahmin aşamalarını kapsamaktadır. Her yöntem her veri tipi için uygun olmadığından, bazen özel tanımlamalar yapılması gerekebilmektedir. Böyle bir durumda üçüncü aşama olan veri dönüştürme (veri hazırlama) aşamasına geri dönülmektedir.

Yorumlama aşamasında; sonuçlar değerlendirilerek veri madenciliği süreci gözden geçirilmekte ve sürecin nasıl devam edeceği yorumlanmaktadır.

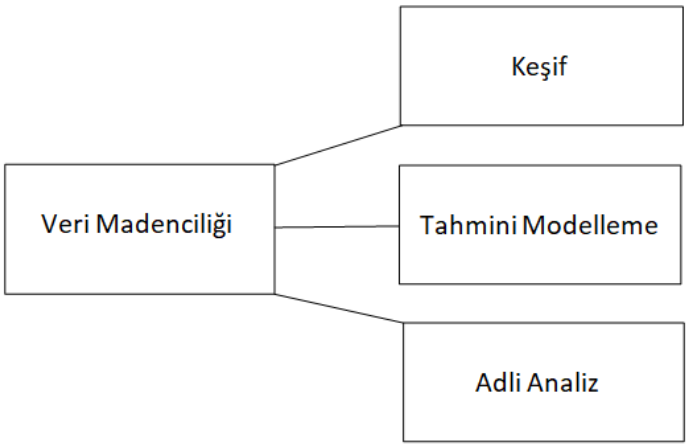
Uygulama aşamasında; veri madenciliği yöntemleriyle ortaya çıkarılan bilgiler, problemlerin çözümünde kullanılmaktadır. Son olarak, araştırma raporunun yazılması ve çalışmanın gözden geçirilmesiyle süreç tamamlanmaktadır (Albayrak ve Koltan Yılmaz, 2009).

Şekil 2.2 ’de, veri madenciliğine etki eden disiplinler belirtilmiştir. Günümüzde kendisi ayrı çalışma alanı olan veri madenciliği, uzun yıllar kısıtlı amaçlar için değişik disiplinlerde kullanılmıştır. Veri madenciliği yöntemleri, ilgili disiplinler tarafından bulunarak kullanılmıştır. Ancak, söz konusu disiplinlerde yapılan araştırmalar kapsamında verinin miktarı küçük boyutlu olup sınırlıdır. Veri madenciliğinde ise içerisinde çok fazla değişken bulunan milyonlarca hatta milyarlarca veri ile ilgilenilmektedir (Kalıpsız ve ark., 2018).



Şekil 2.2. Veri madenciliğine etki eden disiplinler (Kalıpsız ve ark., 2018).

Veri madenciliğinin işlevleri; keşif, adli analiz ve tahmini modelleme şeklinde üç sınıf altında toplanmaktadır (Vahaplar ve İnceoğlu, 2001).



Şekil 2.3. Veri madenciliğinin işlevleri (Vahaplar ve İnceoğlu, 2001)

Keşif (discovery), ne olabileceği konusunda herhangi bir düşünceye sahip değilken toplanmış veri seti içinden gizli desenleri arama ve bulma işlemidir. Burada önemli olan, bulunacak desenin ve bu desenden çıkarılan bilginin zenginliği ve kalitesidir. Örneğin; herhangi bir ülkenin nüfus kayıtları düşünülsün. Kullanıcı, veri tabanına "Mesleği bankacılık olanların yaş ortalaması nedir?" sorusunu sorabilir. Sistemin bu soruya cevabının "47" olduğu varsayılsın. Sistem bu cevapla birlikte bir analist gibi kurallar ortaya koymaya çalışacaktır. Örneğin; "Eğer mesleği sporcu ise yaşı %71 kesinlikle 30 'dan küçüktür." kuralı ortaya konulmuş ise veri tabanından mesleği sporcu olan 100 kişi seçildiğinde, bunlardan 71 'inin yaşının 30 'dan küçük olacağı anlaşılır. Benzer olarak, "Eğer mesleği sporcu ise yaşı %97 kesinlikle 60 'dan küçüktür." kuralı da çıkarabilir. Bu da, seçilen 100 sporcudan en az 97 sporcunun 60 yaşından küçük olacağını belirtir (Vahaplar ve İnceoğlu, 2001).

Tahmini modelleme (predictive modeling), veri setinden çıkarılan desenlerle geleceğe dair tahminlerde bulunmak için kullanılmaktadır. Tahmini model, kullanıcının birtakım alan bilgilerini hakim olmasa bile kaydetmesine izin verir. Sistem, bu eksikleri, önceki kayıtlara bakarak tahmin ederek doldurur. Keşif, veri setindeki desenleri bulmak iken, tahmini modelleme, bu desenlerden yeni veri nesnelerini keşfetmek amacıyla yararlanır. Aynı örnek dikkate alınacak olursa, mesleği sporcu olan bir kişinin yaşı artık yaklaşık olarak tahmin edilebilmektedir. Kayıtlarda mesleği sporcu olan ancak yaşı bilinmeyen birisi söylendiğinde, yaşının %71 kesinlikle 30 'dan küçük, %97 kesinlikle

60 ’dan küçük olduğu tahmini çıkarılabilir. Burada keşif, genelleme yapılacak bir bilginin bulunmasına yardımcı olurken tahmini modelleme, daha yeni ve özgül bilgileri tahmin etmekte kullanılır (Vahaplar ve İnceoğlu, 2001).

Adli analiz (forensic analysis), sıra dışı gözlemleri tespit etmek amacıyla çıkarılmış desenleri inceleme ve uygulama işlemidir. Olağan dışı gözlemi bulmak için öncelikle sıradan kısmın tespit edilmesi gerekir. Örneğe göre, yaşı 60 ’ın üstünde olup hâlâ spor yapan %3 ’lük dilimin olduğu biliniyor ancak sebebi bilinmiyor. Bunlar sıra dışı gözlem olarak sayılmaktadırlar. Bu gözlemlerin kimisi normalden daha sağlıklı bireyler veya yaşla alakası olmayan spor dallarıyla (golf, vs.) uğraşan bireyler olabilirler. Belki de, bu veri tabanında yanlış bilgilerin bulunabileceğini de gösteriyor olabilir. Anlaşılacağı üzere, adli analiz, keşif ile aranılan genel bilginin aksine, sıra dışı durumları araştırmaktadır (Vahaplar ve İnceoğlu, 2001).

Günümüzde karar verme süreci olan çoğu alanda veri madenciliği uygulamaları kullanılmaktadır. Bankacılık, borsa, sigortacılık, telekomünikasyon, pazarlama, perakendecilik, sağlık, biyoloji, genetik, kriminoloji, mühendislik, istihbarat, vb. birçok alanda başarılı şekilde uygulanmaktadır (Savaş ve ark., 2012).

Veri madenciliğinde karşılaşılan sorunlar ise; veri tabanlarının boyutu (küçük örneklemler için geliştirilmiş olan veri tabanları, büyük örneklemlerde dikkatsiz kullanıldığında hatalı tahminlere yol açılmaktadır), veri tabanlarındaki bilginin sınırlı olması, veri tabanında aykırı veri olması (veri girişi/veri toplama aşamasında meydana gelen sistem dışı yanlışlara gürültü denilmektedir ve gürültülü veri ile yapılacak olan tahminlerin doğruluğu azalmaktadır, bu nedenle gürültülü veriler temizlenmelidir) ve veritabanlarında eksik veri bulunması şeklinde sıralanabilmektedir (Atılgan, 2011).

3. KAYNAK ARAŞTIRMASI

Bu bölümde, çalışma için yapılmış olan literatür taramasına yer verilmiştir. Yapılan tarama sonucunda, veri madenciliğinin ve karar ağaçları yönteminin birçok farklı alanda, farklı amaçlar için kullanılabileceği gösterilmektedir.

Albayrak ve Koltan Yılmaz (2009), İstanbul Menkul Kıymetler Borsası 100 endeksinde sınai ve hizmet sektörlerinde etkinlik gösteren 173 ticari müessesenin 2004–2006 yıllarına ilişkin senelik ekonomik indikatörlerinden faydalanarak veri madenciliği yöntemlerinden birisi olan karar ağaçları tekniğini uygulamıştır. Seçilen finansal indikatörlere göre sınai ve hizmet sektörlerinde etkinlik gösteren şirketleri ayıran en önemli değişkenler tespit edilmiştir.

Altan ve ark. (2015), çalışmasında genel sağlık durumlarını etkileyen temel faktörleri belirlemeye çalışmıştır. Bu maksatla, Orta Doğu Teknik Üniversitesi'nde, Temel İngilizce Bölümü öğrencilerine anket uygulamış ve elde edilen veriye karar ağacı algoritmalarından biri olan otomatik ki-kare sınıflandırma algoritması (CHAID) uygulamıştır. Analiz sonuçlarına göre, açıklanan değişken olan genel sağlık durumu değişkenini açıklayan en önemli bağımsız değişken "önceki yıl sağlık durumu" değişkeni olmuştur. Diğerleri; "yaşam standardı", "bir sağlık kuruluşuna gitme sıklığı" ve son olarak "cinsiyet" ile "kişilik türü" değişkenleri olarak saptanmıştır.

Gökçen Alıç (2014) araştırmasında, Ankara ilinde, bir kamu hastanesinde 206 akut pankreatit hastasına ilişkin verileri, "IBM PASW (Predictive Analytics Software) Modeler 14.0" da analiz ederek hastaların mortalite (ölüm) risklerini tespit etmeye çalışmıştır. Pek çok sınıflandırıcı karar ağacı yöntemini denemiş, en iyi performansı C5.0 karar ağacı yöntemiyle elde etmiştir. Çalışmada elde edilen karar ağacı kurallarının, hastalara uygulanacak tedavi metotlarının tayin edilmesi ve doğru tedavinin ivedilikle öngörülmesi açısından, doktorlara önemli karar desteği sağlayacağı belirtilmektedir.

Kavzoğlu ve Çölkesen (2010), karar ağaçlarının sınıflandırma performansını "Landsat ETM+" uydu görüntüsü kullanarak detaylı bir şekilde analiz etmiştir. Metodun sınıflandırma performansı yaygın bir kullanıma sahip maksimum olabilirlik metodunun performansı ile karşılaştırılmıştır. Araştırmada kullanılan veri için sınıflandırma metotlarının performansları istatistiksel açıdan Z testiyle analiz edilmiştir. Analiz sonuçlarına göre, uydu görüntülerinin sınıflandırılmasında karar ağaçlarının etkin bir metot olduğu belirlenmiştir.

Ersöz ve Özseven (2015) araştırmasında, küçük-orta büyüklükteki işletmelerin finansal problemlerini etkileyen etkenlerin, sınıflayıcı modeller olan karar ağaçları ile modellenmesini amaçlamıştır. Araştırmada, nitel metotlardan yüz yüze anket metodu uygulanmıştır. Anket sonuçları ışığında firmaların pek çoğunun aile şirketi olduğu, çeşitli finansal problemlerinin olduğu ve finansal problemlerin çözümleri konusunda yeteri kadar araştırma yapmadıkları neticesine ulaşılmıştır.

Avşar ve Yalçın (2015) çalışmasında; öğrencilerin, "Uluslararası Öğrenci Değerlendirme Programı" olan PISA araştırma projesinin, PISA 2009 okuma okuryazarlığı anketinden seçilen maddelerden elde etmiş oldukları okuma becerileri başarı skorlarını açıklayan ailesel faktörlerin belirlenmesini amaçlamıştır. İlişkisel tarama modelinde olan çalışma, Ankara'da 10'uncu sınıf seviyesinde öğrenim gören 170 öğrenciyle velilerine uygulanmıştır. Veri seti, PISA 2009 araştırmasında okuma okuryazarlığı anketinden seçilen maddelere öğrencilerin vermiş olduğu yanıtlardan ve veli anketlerinden elde edilen bilgilerden meydana gelmektedir. Öğrencilerin okuma becerilerini ortaya koyan değişkenler karar ağaçları algoritmalarından CHAID ile tespit edilmiştir. Araştırmanın sonuçlarına göre, öğrencinin okulundaki öğretmenlerinin çoğunun alanında yeterli, kendisini işine adayan kişiler olduğunu düşünmekte olan velilerin çocuklarının okuma başarılarının diğerlerinden daha yüksek olduğu tespit edilmiştir.

Irmak ve Ercan (2017) çalışmasının maksadı, hanehalkı bütçe anketi verisiyle Türkiye hanehalkının zeytinyağı tüketimi durumunun ve tüketim düzeyinin veri madenciliği metotlarından karar ağaçları aracılığıyla tespit edilmesidir. Çalışmada, zeytinyağı tüketimini etkileyen en önemli değişkenin maddi durum olduğu belirlenmiş, dört farklı gelir kümesi için gelir arttıkça zeytinyağı tüketen hane oranının yükseldiği tespit edilmiştir. Diğer bir önemli belirleyici faktör hanehalkı reisinin yaşı, eğitim seviyesi ile hanehalkı büyüklüğü olarak belirlenmiştir. Gelir düzeyi düşük hanelerde zeytinyağı tüketiminde hanehalkı büyüklüğünün tesirli olduğu ve büyüklük yükseldikçe zeytinyağı tüketen hanelerin oranının düştüğü belirlenmiştir.

Akbal ve ark. (2017) çalışmasında, telefon dolandırıcılığı verisinin analiz edilmesini sağlamıştır. Kişilerin yaşı, mesleği, yaşadığı yer gibi bireye özgü olan özelliklerine göre, telefon dolandırıcılığı verisi karar ağaçları algoritmalarından C4.5 algoritması ile sınıflandırılmış ve elde edilen sonuçlar değerlendirilmiştir.

Emel ve Taşkın (2005) çalışmasında, bir perakendeci firmanın müşterilerine göre kişiye özel satış hareketlerini ihtiva eden veri tabanından faydalanarak ayrıntılı ve

göreli ölçüm neticelerini kapsayan satış analizi yapmıştır. Kurulacak satış tahmin modeli için CART/C&RT (Classification And Regression Tree) karar ağacı algoritmasından faydalanılmaktadır. Bu modelden, detaylı satış değerleri ile müşterilerin satın alma tutumu profilleri birlikte elde edilebileceğinden, firmanın pazarlama stratejilerinin etkinliğinin de arttırılabileceği belirtilmektedir.

4. MATERYAL VE YÖNTEM

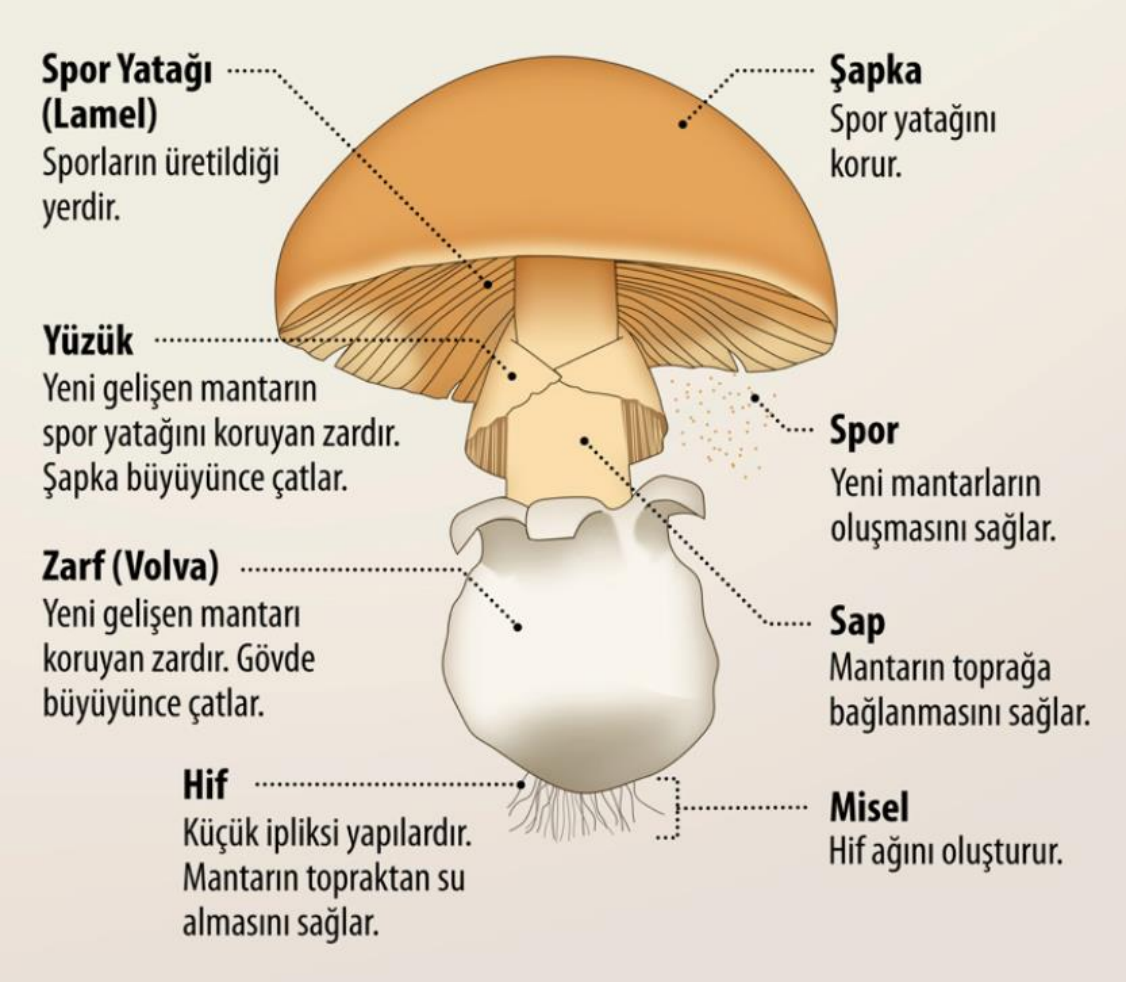
Bu çalışmada, konusuyla ilgili çeşitli kitaplar, e-dergiler, internet siteleri, yurtiçi ve yurtdışında yapılmış tez çalışmaları ve makalelerden yararlanılmaktadır. Araştırmada, SPSS programının bir modülü olan Clementine 12.0 ve MS Office 2013 paket programlarından faydalanılmaktadır. Intel® Core™ i5-8250U CPU @ 1.60 GHz 1.80 GHz işlemcili, 8 GB RAM donanımına sahip bilgisayar kullanılmaktadır.

Çalışmada, internet üzerinden erişim sağlanan ücretsiz, açık kaynak olan The UCI Machine Learning Repository (UCI Makine Öğrenimi Deposu, https://archive.ics.uci.edu/ml/index.php) veritabanından temin edilen mantar veri seti kullanılmıştır. Veri setinin menşei Audubon Society Field Guide'dan Kuzey Amerikan Mantarlarına (1981) hazırlanan mantar kayıtlarıdır. Donörü ise Jeff Schlimmer'dır (Lincoff, 1981). Veri setinde, 8124 adet gözlem yer almaktadır. Bağımsız değişken sayısı 22 olup toplam değişken sayısı 23'tür. Veri setinde yer alan her bir gözleme (mantara) ilişkin 22 bağımsız değişkenin değeri ile mantarların zehirli ya da yenilebilir olduğu verisi kullanılarak mantarlar veri madenciliği sınıflandırma yöntemlerinden karar ağaçları algoritmaları ile zehirli veya yenilebilir şeklinde sınıflandırılmaya çalışılmıştır. Tablo 4.1'de, veri setinde yer alan bağımlı ve bağımsız değişkenler ile onların açıklamalarına yer verilmiştir. Ayrıca, Şekil 4.1 ila Şekil 4.7'ye kadar verilen tüm şekillerde ise değişkenlere ilişkin olarak mantarların özellikleri gösterilmektedir.

Tablo 4.1. Değişkenlere ilişkin açıklamalar

Bağımlı/Bağımsız Değişken Adı		Değişkene ilişkin bilgiler ve açıklamalar
Bağımlı Değişken	Zehirliliği	zehirli: p, yenilebilir: e
Bağımsız Değişken	Şapka şekli	çan: b, konik: c, dışbükey/konveks: x, düz: f, topuzlu/memeli: k, içe çökük: s
	Şapka yüzeyi	lifli: f, yivli: g, pullu: y, pürüzsüz: s
	Şapka rengi	kahverengi: n, ten rengi: b, tarçın: c, gri: g, sarı: y, yeşil: r, pembe: p, mor: u, kırmızı: e, beyaz: w
	Çürükler?	çürükler: t, hayır: f
	Koku	badem: a, anason: l, kreozot/katran ruhu: c, küf: m, balık: y, bozuk/iğrenç: f, yok: n, mayhoş/acı: p, baharatlı: s
	Solungaç (lamel) bağlantısı	bitişik: a, inişli: d, serbest: f, çentikli: n
	Solungaç aralığı	yakın: c, kalabalık: w, uzak: d
	Solungaç boyutu	geniş/kalın: b, dar/ince: n
	Solungaç rengi	siyah: k, kahverengi: n, ten rengi: b, çikolata: h, gri: g, yeşil: r, turuncu: o, pembe: p, mor: u, kırmızı: e, beyaz: w, sarı: y
	Sap şekli	genişleyen: e, daralan: t
	Sap kökü	soğanlı: b, kulüp: c, bardaklı: u, eşit: e, rizomorf/rizoid/köksü: z, köklü: r, eksik: ?
	Sap yüzeyi-Yüzüğün üstünde	lifli: f, pullu: y, ipeksi: k, pürüzsüz: s
	Sap yüzeyi-Yüzüğün altında	lifli: f, pullu: y, ipeksi: k, pürüzsüz: s
	Sap rengi-Yüzüğün üstünde	kahverengi: n, ten rengi: b, tarçın: c, gri: g, turuncu: o, pembe: p, kırmızı: e, beyaz: w, sarı: y
	Sap rengi-Yüzüğün altında	kahverengi: n, ten rengi: b, tarçın: c, gri: g, turuncu: o, pembe: p, kırmızı: e, beyaz: w, sarı: y
	Peçe tipi	parçalı: p, bütün: u
	Peçe rengi	kahverengi: n, turuncu: o, beyaz: w, sarı: y
	Yüzük sayısı	yok: n, bir: o, iki: t
	Yüzük tipi	örümcek ağı: c, kaybolan: e, yayılan: f, büyük: l, yok: n, kolye: p, kaplayan/kılıflı: s, kuşak/zon: z
	Spor izi rengi	siyah: k, kahverengi: n, ten rengi: b, çikolata: h, yeşil: r, turuncu: o, mor: u, beyaz: w, sarı: y
	Popülasyon	bol: a, kümelenmiş: c, çok sayıda: n, dağınık: s, birkaç: v, tek: y
	Habitat	çimler/otlar: g, yapraklar: l, çayırlar: m, yollar: p, kent: u, atık: w, ormanlar: d

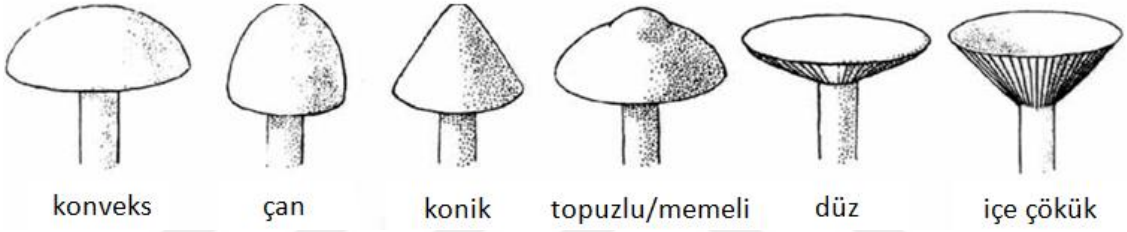
Mantarlar, biyosferin, yani selülozun en bol bulunan biyomolekülü üzerinde yetişen küçük orman ürünlerinin önemli bileşenleridir. Mantarlar, çıplak gözle görülebilen ve elle toplanabilecek kadar belirgin olabilen farklı bir meyve veren gövdeye sahip makro mantarlar (büyük mantarlar/şapkalı mantarlar) olarak kabul edilmektedir. Mantarın sadece meyve veren gövdesi görülürken mantarın geri kalan kısmı miselyum olarak yeraltında kalmaktadır (Wani ve ark., 2010). Şekil 4.1'de şapkalı mantarın bölümleri gösterilmektedir (Okatan, 2019).



Şekil 4.1. Şapkalı mantarın bölümleri (Okatan, 2019)

Jeolojik olarak, mantarlar, insanın ortaya çıkmasından daha önce, fosil kayıtlarından anlaşıldığı üzere, dünyada varlığını sürdürmüştür. Buradan hareketle, insanın bir yiyecek toplayıcısı ve avcı olduğu zamanlarda, mantarları yiyecek olarak kullanma olasılığı vardır. Mantarlar ise, temel ekolojik rollerinin yanı sıra gıda ve ilaç olarak kullanılabilecekleri gibi günümüzde başka faydalar da sunmaktadır. Mantarlar,

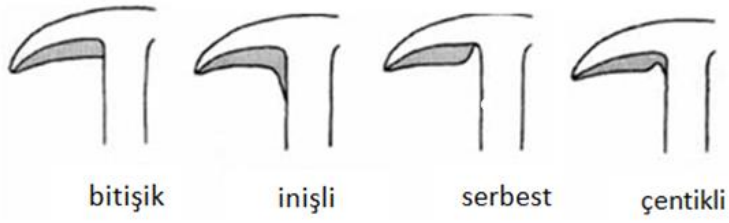
dünyanın en büyük keşfedilmemiş beslenme ve geleceğin lezzetli yiyecek kaynaklarını temsil etmektedir. Mantarlar kanser, kolesterol düşürme, stres, uykusuzluk, astım, alerji ve diyabete karşı etkili bulunmuştur. Yüksek miktarda protein bulundurduğundan, protein açısından yetersiz beslenme boşluğunun kapatılması için de kullanılabilirler. Fonksiyonel gıdalar olarak mantarlar, tablet şeklinde bağışıklığı arttırmak için besin takviyesi olarak da kullanılmaktadır. Düşük nişasta içeriği ve düşük kolesterol nedeniyle, diyabet ve kalp hastaları için uygundur. Mantardaki demirin üçte biri kullanılabilir formdadır. Polisakkarit içerikleri, antikanser ilaç olarak kullanılır. Mantarlardaki biyolojik olarak aktif bileşikler, antifungal, antibakteriyel, antioksidan ve antiviral özelliklere sahiptir ve ayrıca böcek öldürücüler olarak da kullanılmıştır (Wani, ve ark., 2010).



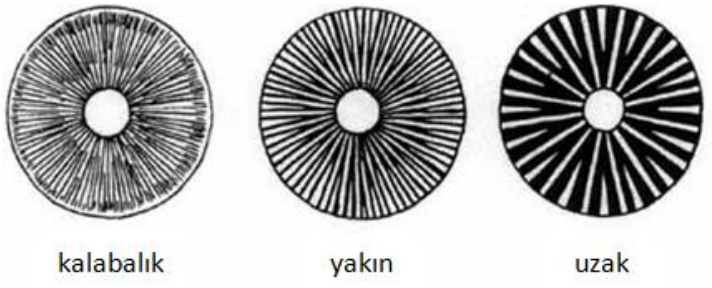
Şekil 4.2. Mantar şapka şekilleri (Bossenmaier, 1997)



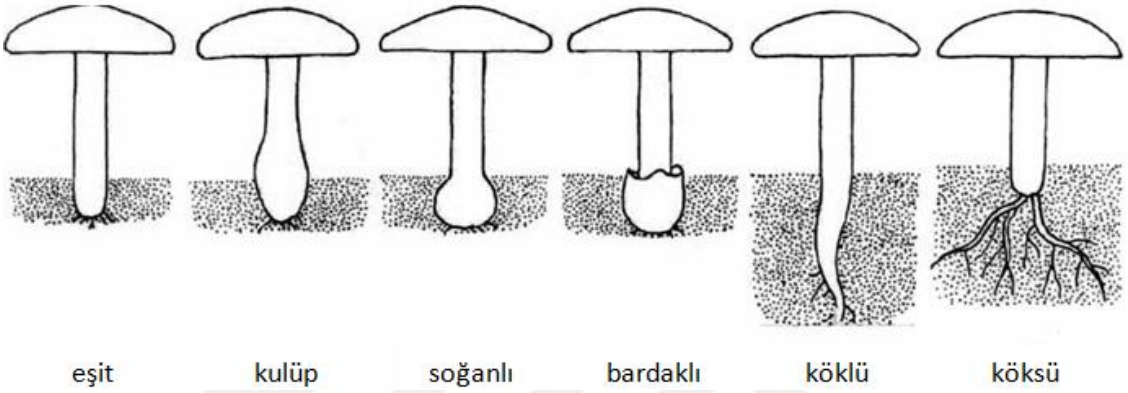
Şekil 4.3. Mantar şapka yüzeyleri (Bossenmaier, 1997)



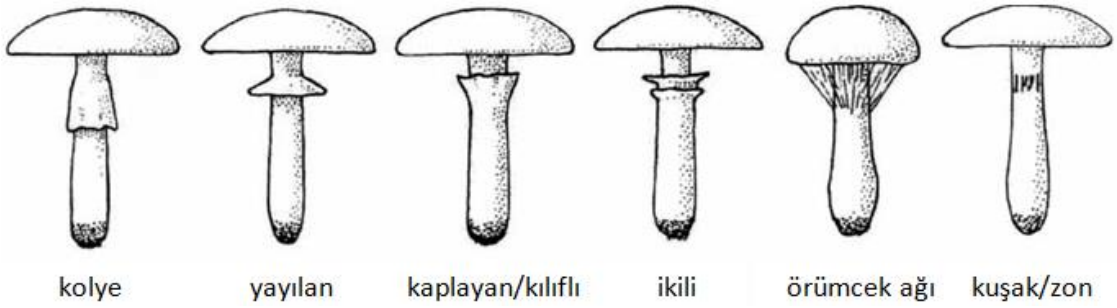
Şekil 4.4. Mantar sapı ile solungaç (lamel) arasındaki bağlantı (Bossenmaier, 1997)



Şekil 4.5. Solungaç (lamel) aralığı (Bossenmaier, 1997)



Şekil 4.6. Mantar sap kökü çeşitleri (Bossenmaier, 1997)



Şekil 4.7. Mantar yüzük tipi çeşitleri (Bossenmaier, 1997)

Türkiye'de, zehirli ve zehirsiz pek çok şapkalı mantar türü bulunmaktadır. Mantarların zehirli olup olmadığını anlamaksa hiç kolay değildir. Çünkü aynı ortamda yaşayabilen mantarlar şekillerine, kokularına ve renklerine göre kolayca ayırt edilememektedir. Zehirli mantarlar insanlarda farklı etkilere sebep olabilmektedir. Kimisinin etkileri tedavi edilebilirken kimisi ise ölümcül olabilmektedir (Okatan, 2019).

Bu nedenle, mantarların tüm bu faydaları karşısında, ölüme götürebilen zehirli etkisi de göz önünde bulundurularak bu çalışmada, mantarların zehirli olup olmadığını belirleyebilmek için en önemli olabilecek fenotipik özellikleri ile veri setinde gözlem

değerleri yer alan 22 bağımsız değişkene göre sınıflandırma (zehirli/yenilebilir) kuralları veri madenciliği yöntemleri ile araştırılmıştır.

4.1. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri, denetimsiz ve denetimli şeklinde iki kategoriye ayrılmaktadır. Veri madenciliğinde iyi belirlenmiş bir hedef varsa denetimli ifadesi kullanılmaktadır. Denetimli veri madenciliği yöntemleri, veriden anlamlı bilgi ve sonuç çıkarma ihtiyacına yönelik olarak kullanılmaktadır. Elde edilmek arzu edilen sonuç için anlamlı bir betimleme yapılmamışsa ve bir belirsizlik varsa denetimsiz kavramı kullanılmaktadır. Denetimsiz veri madenciliği yöntemleri, veriyi anlama, tanıma ve keşfetme amacıyla kullanılan ve daha sonra uygulanacak metotlar için fikir vermesi maksadıyla kullanılan yöntemlerdir.

Denetimli (supervised) veri madenciliği metotları; karar ağaçları (decision trees), regresyon modelleri (regression models), k-en yakın komşuluk (k-nearest-neighbor), sinir ağları (neural networks), k-ortalamalar kümeleme, kural çıkarımı (rule induction) olarak sınıflandırılabilir.

Denetimsiz (unsupervised) veri madenciliği metotları ise; aşamalı kümeleme (hierarchical clustering) ve kendi kendini düzenleyen haritalar şeklinde sınıflandırılabilmektedir.

Ayrıca, diğer veri madenciliği yöntemlerinden en bilinenleri ise kohonen ağları, genetik algoritmalar, bulanık mantığa dayalı yöntemler, faktör analizi, temel bileşenler analizi, bayesci ağlar, diskriminant analizi, pürüzlü (rough) küme teorisine dayanan metotlar olarak sıralanabilmektedir.

Veri madenciliği alanında kullanılan modeller; tahmin edici (predictive) modeller (sınıflandırma, zaman serileri, vd.) ve tanımlayıcı (descriptive) modellerdir (kümeleme, birliktelik kuralları, vd.) (Akpınar, 2000).

Tahmin edici modellerde maksat; mevcut veriyi kullanarak geleceğe yönelik olarak kestirimler yapabilmek olup bu amaç doğrultusunda sınıflama ve regresyon modelleri kullanılmaktadır. Kestirime dayalı modellerle; kredi başvurularını risk gruplarına göre ayırma, şirketle çalışmayı bırakacak olan müşterileri tahmin etme gibi çalışmalar yapılmaktadır. Örneğin; bir banka geçmiş dönemlerde kullandırmış olduğu krediler ve müşteriler hakkında tüm verilerini saklamış olsun. Bu verilerde, bağımsız değişken kredi kullanan müşterilerin nitelikleri, açıklanan değişken kredinin geri ödenip ödenmediği olacaktır. Bu verilere uygun olarak kurulacak olan modelle, müşteri

özelliklerine göre, daha sonraki kredi taleplerinde, kullandırılacak kredinin geri ödenip ödenmeyeceğinin tahmini yapılabilmektedir (Akpınar, 2000).

Tanımlayıcı modellerde maksat; mevcut verideki gizli bağlantıları, kümeleri ve veriyi niteleyebilecek özellikleri ortaya çıkarmak olup bu amaç için kümeleme, birliktelik kuralları ile ardışık zamanlı örüntü modelleri kullanılmaktadır. Tanımlayıcı modellerle; en iyi müşterilerim kimlerdir, hangi ürünler beraber satılıyor, hangi müşteri kümelerinin alışveriş alışkanlıkları benzerdir gibi sorulara cevap bulunmaktadır. Örneğin; Y_1/Y_2 aralığında gelire ve bir veya daha fazla otomobile sahip olan çocuklu aileler ile hiç çocuğa sahip olmayan ve geliri Y_1/Y_2 aralığından daha az olan ailelerin satın alma davranışlarının birbirine benzerlik gösterip göstermediğinin tespit edilmesi tanımlayıcı modeller ile olmaktadır (Akpınar, 2000).

Veri madenciliği yöntemleri aslında birer istatistik uygulamasıdır ve yapılan çalışmalarda, istatistik temel roldedir. Klasik istatistiksel uygulamalar ile veri madenciliğini birbirinden ayıran temel nokta veri kümesinin büyüklüğüdür. İstatistikte birkaç yüz veya birkaç bin veri büyük bir veri kümesi iken veri madenciliğinde milyon veya milyarlık veri setleriyle çalışılmaktadır. Veri madenciliği ile istatistiksel analizlerin karşılaştırması Tablo 4.2 ’de verilmektedir.

Tablo 4.2. Veri madenciliği ile istatistiksel analizin karşılaştırılması (Koyuncugil ve Özgülbaş, 2009)

İstatistiksel Analiz	Veri Madenciliği
İstatistikçiler genellikle bir hipotez ile başlarlar.	Veri madenciliği hipoteze ihtiyaç duymaz.
Hipotezlerini eşleştirmek için kendi eşitliklerini geliştirmek zorundadırlar.	Veri madenciliği algoritmaları eşitlikleri otomatik olarak geliştirir.
İstatistiksel analizler sadece sayısal veriyi kullanır.	Veri madenciliği sadece sayısal veriyi değil, farklı tiplerde veriyi de kullanır (örneğin; metin, ses).
Kirli veriyi analizleri sırasında bulur ve filtre eder.	Veri madenciliği temiz veriye dayanır.
İstatistikçiler kendi sonuçlarını yorumlar ve bu sonuçları yöneticilere iletirler.	Veri madenciliğinin sonuçlarını yorumlamak kolay değildir. Sonuçlarını analiz etmede, yorumlamada ve bulguları yöneticilere iletmede mutlaka istatistikçiye ihtiyaç duyulmaktadır.

Veri madenciliği modellerini üç temel başlık altında toplamak mümkündür. Sınıflandırma ve regresyon; k-en yakın komşu, karar ağaçları, bayesyen sınıflandırma, vd. algoritmalar kullanılır. Kümeleme; en yakın komşu, iki aşamalı kümeleme, en uzak komşu, kohonen ağlar, k-ortalamalar, vd. algoritmalar kullanılır. Birliktelik kuralları; apriori, apriori tid, vd. algoritmalar kullanılır. Tahmin edici modeller, sınıflandırma ve regresyon modelleridir; kümeleme ile birliktelik kuralları, tanımlayıcı modellerdir (Akpınar, 2000).

Bu çalışmada, veri madenciliği yöntemlerinden sınıflandırma ve regresyon modellerinden karar ağaçları algoritmaları derinlemesine incelenecektir.

4.1.1. Sınıflandırma ve regresyon

Sınıflama, veri madenciliğinde çokça kullanılan ve en bilinen tekniklerden birisi olup veri tabanlarındaki saklı örüntüleri ortaya çıkartmakta kullanılmaktadır. Sınıfı belirlenmiş mevcut veriden yararlanılarak sınıfı belirlenmemiş olan verinin sınıfını tahmin etmekte ya da önceden sınıflandırılmış olan verilerin doğru şekilde sınıflandırılıp sınıflandırılmadığının belirlenmesi ve eğer yanlış sınıflandırma varsa verinin doğru gruba atanması için kullanılmaktadır. Sınıflandırma, kategorik değerleri

tahmin etmekte kullanılmakta iken regresyon, süreklilik gösteren değerleri tahmin etmekte kullanılmaktadır. Açıklayıcı değişken olarak belirtilen tahmin edici değişkenler; açıklanan değişken olarak belirlenen tahmin edilecek olan değişken değerini saptayacak ağırlıklandırmaları içerecek açıklayıcı değişkenler birleşimidir. Örneğin; bir firmanın müşterilerinin belirli özellikleri baz alınarak sınıflandırılması yapılabilir. Bir alışveriş merkezinin yöneticileri, ortalama alışveriş tutarından fazla alışveriş yapan müşterileri "varlıklı", diğerlerini "yoksul" olarak birbirinden ayırt edebilir. Sınıflama ile "profil çıkarımı" yapıldığı söylenebilir. 40 yaş üstü, ekonomik durumu iyi bayanların oy verirken yoğunlukla belirli bir partiyi tercih etmesi bir profil örneği olarak verilebilir. İki tür sınıflandırma söz konusudur. Denetimli (supervised) sınıflandırmada, sınıfların sayısı ile hangi nesnenin hangi sınıfa ait olduğu bilinmektedir. Denetimsiz (unsupervised) sınıflandırmada ise hangi nesnenin hangi sınıfa ait olduğu ve genelde sınıf sayısı bilinmemektedir.

Sınıflandırma, çok çeşitli alanlarda kullanılabilmektedir. Örneğin; finansal pazardaki eğilimleri sınıflandırmak amacıyla kullanılmaktadır. Bir bankanın kredi taleplerinin kullandırım değerlendirilmesinde sınıflandırma algoritmaları kullanılabilmektedir.

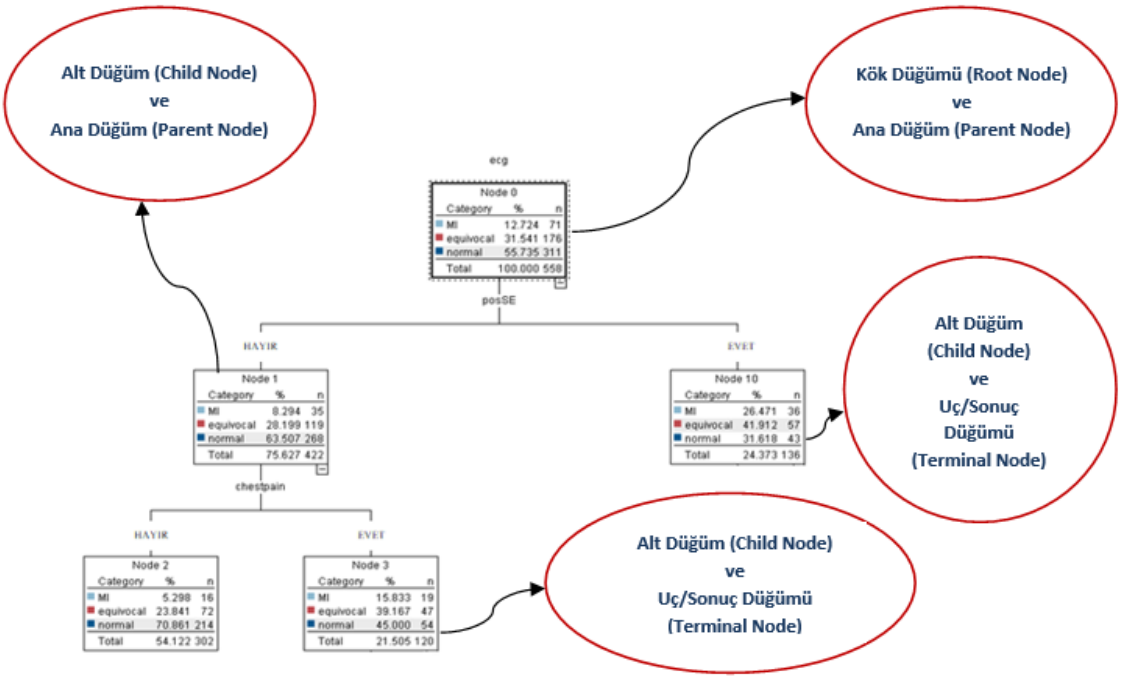
Sınıflandırmada kullanılan en temel yöntemler; karar ağaçları, mesafeye ve istatistiğe dayalı (bayesyen sınıflandırma, vd.) sınıflandırmadır.

Karar ağaçları; veri madenciliğinde elde edilişinin ve yorumlanmasının zor olmaması, veri tabanı sistemleri ile kolay entegre olabilmesi ve güvenilirliğinin yüksek olması nedenleriyle sınıflama yöntemleri arasında en çok kullanıma sahip yöntemdir. Karar ağaçlarının en büyük faydası, öğrenilen kuralın anlaşılır şekilde ifade edilebilmesidir (Akpınar, 2000).

Araştırmada, veri madenciliğinin sınıflandırma yöntemlerinden olan karar ağaçları (decision trees) kullanılacaktır. Karar ağaçları, denetimli (supervised) veri madenciliği yöntemlerinden olup sınıflandırma ve regresyon için kullanılmaktadır. Sınıflama, veri madenciliğinde çokça kullanılan ve en bilinen tekniklerden birisi olup veri tabanlarındaki saklı örüntüleri ortaya çıkartmakta kullanılmaktadır. Sınıfı tanımlanmış mevcut veriden faydalanarak sınıfı belli olmayan verinin sınıfını tahmin etmekte ya da önceden sınıflandırılmış verinin doğru sınıflandırılıp sınıflandırılmadığının belirlenmesi ve eğer yanlış sınıflandırma var ise verinin doğru gruba atanması için kullanılmaktadır. Sınıflandırma, kategorik değerleri tahmin etmekte

kullanılmakta iken regresyon, süreklilik gösteren değerleri tahmin etmekte kullanılmaktadır.

Karar ağacı; ağaç görünümünde, tahmin edici bir model olarak kullanılan bir tekniktir. Karar ağacı, testi belirten düğümlerden, testteki değerleri gösteren dallardan, sınıfı belirten yapraklardan meydana gelen akış şeması şeklinde ağaç görünümünde bir yapıdır. Karar ağacındaki en üstteki düğüm, "kök düğümü" dür. En son yapı "yaprak", kök düğümü ile yapraklar arasında kalan yapılar "dal" olarak isimlendirilmektedir. Karar ağaçları; nicel, nitel, kesikli ve sürekli tüm değişken tiplerine uygulanabilen algoritmaları, ağaç şeklinde görsel desteği ve SQL sorgusuna kolayca dönüştürülebilir yapısı ile en popüler sınıflandırma yöntemlerinden birisidir. Karar ağaçlarında, geliştirilen algoritmaların hemen hemen tümünde, temel olarak entropi veya gini indeksinden yararlanılarak veritabanı mümkün olduğunca eşit parçalara bölünmekte ve böylece "alt düğüm/yaprak" olarak adlandırılan sınıflara ulaşılmaktadır. Şekil 4.8 'de, karar ağacı örneği verilmiş olup kök düğümü ve alt düğümler/yapraklar örnek üzerinde belirtilmiştir.



Şekil 4.8. Karar ağacı örneği

Karar ağaçlarında, değişken tiplerine göre farklı algoritmalar kullanılabilmektedir. C5.0, CART ve CHAID en popüler algoritmalardır.

C5.0 algoritması, düğümlerden çıkan çoklu dallarla ağaç oluşturmaktadır. Dalların sayısı tahmin edici değişkenlerin kategori sayısına eşit olmaktadır. Tek bir sınıflayıcıda birden fazla karar ağacını birleştirmektedir. Ayırma işlemi için bilgi kazancını kullanmaktadır.

CART algoritmasında, gini indeksine dayalı ikili bölme işlemi mevcuttur. Dallanma kriteri olarak gini indeksinden yararlanılan bu teknikte; son olmayan her düğümde, iki tane dal yer almaktadır. CART algoritması hem nominal hem de sayısal değerler ile çalışabilmektedir.

Karar ağaçları algoritmalarından bir diğeri olan CHAID analizi ise ilk kez Kass tarafından uygulanmıştır hemen ardından, Jay Magidson tarafından geliştirilerek günümüzdeki haline getirilmiştir (Albayrak ve Koltan Yılmaz, 2009; Altan ve ark., 2015). Bu algoritmada, bağımsız değişkenler belirlenirken önem testleri uygulanmakta ve açıklanan değişkeni en iyi açıklayabilecek bağımsız değişkenler ayrıntılı alt kümelere bölünmektedir. Kullanılan istatistiksel test, açıklanan değişkenin ölçüm düzeyine bağlı olmaktadır. Eğer açıklanan değişken sürekliyse F testi, kategorikse ki-kare testi kullanılmaktadır. CHAID analizi ki-kare metriği aracılığıyla ilişki seviyesine göre farklılıkla karşılaşılan kümeleri birbirinden ayrı sınıflamakta, karar ağacının yaprakları, ikili değil de, verideki değişik yapı sayısı kadar dallanmaktadır.

Karar ağacı analizlerinin, günümüzde yaygın olarak, yapılan araştırmalarda elemanların/gözlemlerin belirli bir sınıfın üyesi olup olmayacağının belirlenmesinde, gözlemlenen hadiselerin yüksek/orta/düşük risk grubu gibi farklı kategorilere ayrılmasında, parametrik modellerde kullanılmak üzere çok sayıda değişkenden en önemli değişkenlerin seçilmesinde, gelecekte meydana gelebilecek olayların tahmin edilebilmesini sağlayan kuralların oluşturulmasında, sadece belirlenmiş alt gruplara has olan ilişkilerin belirlenmesinde, kategorilerin birleştirilebilmesinde, sürekli değişkenlerin kesikli değişkenlere dönüştürülebilmesinde kullanıldığı görülmektedir. Örneğin; karar ağaçları ile gümrüklerden giren kişi profili, gümrüklerden giren şirket profili, kaçakçılık profili, risk profili gibi profiller elde edilebilmektedir.

Karar ağaçlarında, geliştirilen algoritmaların hemen hemen tümünde, temel olarak entropi veya gini indeksinden yararlanılarak veritabanı mümkün olduğunca eşit parçalara bölünmekte ve böylece düğüm/yaprak olarak adlandırılan sınıflara ulaşılmaktadır.

Karar ağaçları çoğu zaman karmaşık bir görünüme sahip olabilmektedir. Ancak, elde edilen karar ağacının sade ve kolay anlaşılır olması istenen özellikler arasındadır.

Bu nedenle de; bir karar ağacında, alt ağaçlardan biri atılarak yerine bir yaprak yerleştirilebilir. Bu şekilde yapılan işleme "karar ağacının budanması" adı verilmektedir (Özkan, 2008). Ancak, bu yerleştirme işleminde alt ağacın yerine konulacak olan yaprağın hatasının, alt ağacın hatasından az olması gerekmektedir. Bu durumda budama gerçekleştirilebilmektedir (Silahtaroğlu, 2008). Eğer, tek bir yaprağı bulunan alt ağacın ya da ağacın en çok yararlanılan dalının budanması, beklenen hata rasyosunu düşürecek ise ağaç budanır. Böylece, alt dallardaki hata oranı azaldığında bütün ağaç için hata oranı azalacağından, budama işlemi sonucunda, hata oranının minimuma indirildiği bir ağaç elde edilmiş olmaktadır (Kavzoğlu ve Çölkesen, 2010). Budama işlemi yapılmasının bir diğer nedeni de, karar ağaçlarının maksadının veri keşfetmek değil, veriyi basit sınıflandırma modeli oluşturmak için kullanmasıdır (Altıkardeş ve ark., 2012).

Sınıflandırma analizlerinde kurulan modellerin doğruluk derecelerinin belirlenmesinde ise faydalı bir araç olarak risk matrisi kullanılmaktadır. Bu matrisin satırlarında gerçek, sütunlarında ise tahmini sınıflandırma değerleri yer almaktadır (Bkz. Tablo 4.3). Bir başka önemli değerlendirme ölçütüyse modelin anlaşılabilir olmasıdır yani karmaşık olmamasıdır (Akpınar, 2000).

Tablo 4.3. Risk matrisi

	Tahmini	
Gerçek	**A sınıfı**	**B sınıfı**
A sınıfı	AA	AB
B sınıfı	BA	BB

$$Hata\ Oranı = \frac{BA + AB}{AA + BB + BA + AB} \tag{4.1}$$

$$Doğruluk\ Oranı\ (Accuracy) = \frac{AA + BB}{AA + BB + BA + AB} \tag{4.2}$$

$$Doğruluk\ Oranı = 1 - Hata\ Oranı \tag{4.3}$$

Risk matrisinin sütun yüzdesine duyarlılık (precision), satır yüzdesine anma (recall) denilmektedir. Bu ölçütler de, kurulan modelin değerlendirilmesinde ve modellerin karşılaştırılmasında kullanılmaktadır (Atılgan, 2011).

$$Duyarlılık = \frac{AA}{AA + BA} \tag{4.4}$$

$$Anma = \frac{AA}{AA + AB} \tag{4.5}$$

Kurulan modellerin karşılaştırılmasında kullanılan bir diğer önemli ölçüt ise lift (kaldıraç) oranı ve grafiğidir. Lift oranı, bir modelin sağladığı yararın değerlendirilmesinde kullanılan önemli yardımcıdır (Akpınar, 2000).

$$Lift = \frac{AA/(AA + BA)}{(AA + AB)/N} \qquad ,N: veri\ kümesindeki\ eleman\ sayısı \tag{4.6}$$

Kurulan bir modelin doğruluk miktarı ne kadar yüksek miktarda olursa olsun, yaşanılan gerçek dünyayı tam manasıyla modellediğini garanti edebilmek olası değildir. Yapılan testler neticesinde geçerli bir modelin doğru olmayışındaki başlıca sebepler, modelin kuruluş aşamasında kabul edilen sayıltılar ve modelde kullanılan verinin doğru olmamasıdır. Örneğin, bir modelin kurulması esnasında varsayılan enflasyon oranının zamanla değişmesi, bireyin satın alma tutumunu belirgin bir biçimde etkileyecektir (Akpınar, 2000).

Zaman içerisinde tüm sistemlerin özelliklerinde ve üretmiş oldukları veride meydana gelen değişiklikler, kurulmuş olan modellerin devamlı olarak izlenmesini ve eğer gerekiyorsa tekrar düzenlenmesini gerektirecektir (Akpınar, 2000).

4.1.1.1. C5.0 algoritması

C5.0 algoritması, düğümden çıkan çoklu dallarla ağaç oluşturmaktadır. Dalların sayısı bağımsız değişkenlerin kategori sayısına eşittir. Tek sınıflayıcıda birden fazla karar ağacını birleştirmektedir. Ayırma işlemi için bilgi kazancını kullanmaktadır.

C5.0 algoritması hem sayısal (kesikli ve sürekli) hem de nominal (kategorik) değerler üzerinde çalışabilmektedir. Bu algoritmada; hedef sınır değerlerini içeren nitelik belirlenir ve bu niteliğin kümesi için "entropi" hesaplanır. Entropi; sistemdeki belirsizliğin ölçüsüne verilen addır. S bir kaynak olmak üzere, bu kaynağın $\{m_1, m_2, \dots, m_n\}$ şeklinde n mesaj üretebildiği varsayılsın. Bütün mesajlar birbirinden bağımsız olarak üretilmekte ve m_i mesajlarının üretilme olasılıkları p_i 'dir. $P = \{p_1, p_2, \dots, p_n\}$ olasılık dağılımına sahip olan mesajları meydana getiren S kaynağının entropisi $H(S)$ aşağıdaki gibidir (Özkan, 2008):

$$H(S) = -\sum_{i=1}^{n} p_i log_2(p_i) \tag{4.7}$$

Veritabanından eğitim için elde edilen bir kayıt kümesi ele alınsın. Eğitim kümesinin sınıf niteliğinin sahip olacağı değerlere göre $\{C_1, C_2, \dots, C_k\}$ olmak üzere k sınıfa ayrıldığı varsayılsın. Bu sınıflara ilişkin olarak ortalama bilgi miktarına gereksinim olabilir. Burada T sınıf değerlerini içermekte olan küme için P_T sınıfların olasılık dağılımıdır ve de şu şekilde hesaplanmaktadır:

$$P_T = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \dots, \frac{|C_k|}{|T|}\right) \tag{4.8}$$

$|C_i|$ ifadesi, C_i kümesindeki elemanların sayısını vermektedir. Şöyle ki, $p_1 = \frac{|C_1|}{|T|}$ olasılığını ifade etmektedir. Öyleyse; T için ortalama bilgi miktarı, bir diğer deyişle entropi,

$$H(T) = -\sum_{i=1}^{n} p_i log_2(p_i) \tag{4.9}$$

olarak ifade edilir. Hedef niteliğini belirten *T*, hedef niteliği olmayan (sınıf niteliği olmayan) bir *X* niteliğinin değerine bağlı olarak T_1, T_2, ... , T_n alt kümelerine ayrılırsa, *T*'nin bir elemanının sınıfını tespit etmek için gerekli bilgi, T_i 'nin bir elemanının sınıfının tespit edilmesinde ihtiyaç duyulan bilginin ağırlıklı ortalaması olarak kabul edilir. Bu tanıma bağlı olarak *T*'nin bir elemanının sınıfını belirlemek için ihtiyaç duyulan bilgi şu şekilde hesaplanır (Özkan, 2008):

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i) \tag{4.10}$$

T veritabanını *X* testine göre bölmekle sahip olunan bilgileri ölçmek için "kazanç ölçütü" ismi verilen bir değere başvurulur. Bu ölçüt aşağıdaki gibi tanımlanmaktadır:

$$Kazanç(X,T) = H(T) - H(X,T) \tag{4.11}$$

Burada, ayırma işlemi gerçekleştirilirken *Kazanç(X,T)* değerini maximize etme (ençoklama) hedeflenir. Maksimum bilgi kazancını sağlayan, bir başka deyişle kazancı maksimize edebilecek *X* testi seçilir (Özkan, 2008).

$$Bilgi\ Bölümleme\ (X,T) = -\sum_{i=1}^{k} \frac{|T_i|}{|T|} log_2 \frac{|T_i|}{|T|} \tag{4.12}$$

$$Kazanç\ Oranı\ (X,T) = \frac{Kazanç\ (X,T)}{Bilgi\ Bölümleme\ (X,T)} \tag{4.13}$$

Burada kazanç oranı, bilgi kazancını normalize etmektedir. Kazanç oranı, dal sayısının kısıtlanmasını sağlamakta, bir nitelikten dallanan dal sayısını belirli bir dengede tutarak kazancı optimize etmeye çalışmaktadır (Gökçen Alıç, 2014).

4.1.1.2. CART algoritması

1984 yılında, Breiman, Friedman, Olshen, Stone'un geliştirdiği CART (Classification & Regression Trees) algoritmasında, her adımda, ilgili grubun kendisinden daha da homojen iki alt gruba bölünmesi sağlanmaktadır (Akpınar, 2000). CART algoritmasında, gini indeksine dayalı ikili bölme işlemi mevcuttur. Dallanma kriteri olarak gini indeksinden yararlanılan bu teknikte; son olmayan her düğümde, iki tane dal bulunmaktadır. CART algoritması hem nominal hem de sayısal değişkenlerle çalışabilmektedir. Bağımlı ve açıklayıcı değişkenlerin nominal, ordinal ya da sürekli olduğu durumlarda uygulanan bir yaklaşımdır. CART algoritması, sınıflandırma ve regresyonu destekleyen bir yapıdadır.

Gini indeksi ile geliştirilen algoritma; ikili bölünmelere dayalı sınıflandırma metodudur. Bu algoritma, nitelik değerlerinin sağ ve solda olmak üzere iki bölüme ayrılmasına dayanmaktadır. Gini algoritması aşağıda belirtildiği üzere uygulanmaktadır (Özkan, 2008):

a) Her nitelik değeri ikili olacak bir biçimde gruplanır. Bu yolla elde edilen sağ ve sol bölünmelere karşılık gelen sınıf değerleri gruplandırılır.

b) Her bir nitelikle ilgili sol ve sağ taraftaki bölünmeler için $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanır.

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{sol}|} \right)^2 \tag{4.14}$$

$$Gini_{sağ} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{sağ}|} \right)^2 \tag{4.15}$$

Bu bağıntılarda yer alan ifadelerin anlamları ise şu şekildedir:

k	: Sınıfların sayısı
T	: Bir düğümdeki örnekler
$\|T_{sol}\|$	: Sol taraftaki örneklerin sayısı
$\|T_{sağ}\|$	: Sağ taraftaki örneklerin sayısı
L_i	: Sol tarafta *i* kategorisindeki örneklerin sayısı
R_i	: Sağ tarafta *i* kategorisindeki örneklerin sayısı

c) Her bir j niteliği için, n eğitim kümesindeki satır sayısı olmak üzere aşağıdaki eşitliğin değeri hesaplanır:

$$Gini_j = \frac{1}{n}\left(|T_{sol}|Gini_{sol} + |T_{sağ}|Gini_{sağ}\right) \quad (4.16)$$

d) Her bir j niteliği için hesaplanan $Gini_j$ değerleri arasından en küçük olanı seçilir ve bölünme bu nitelik üzerinden gerçekleştirilir.

e) En baştaki (a) adımına dönülerek işlemlere devam edilir.

4.1.1.3. CHAID algoritması

Karar ağaçları algoritmalarından CHAID (Chi-squared Automatic Interaction Detection) ilk defa 1980 yılında Kass tarafından uygulanmış Jay Magidson tarafınca geliştirilerek bugünkü haline getirilmiştir (Albayrak ve Koltan Yılmaz, 2009; Altan ve ark., 2015). Açıklanan değişken nominal, ordinal kategorik veya sürekli; açıklayıcı değişkenler ise kategorik, sürekli ya da nominal kategorik olduğunda kullanılabilen bir algoritmadır (Doğan ve Özdamar, 2003). Bu algoritmada, bağımsız değişkenler belirlenirken önem testleri uygulanmakta ve açıklanan değişkeni en iyi açıklayabilecek bağımsız değişkenler ayrıntılı alt kümelere bölünmektedir. Kullanılan istatistiksel test, açıklanan değişkenin ölçüm düzeyine bağlı olmaktadır. Eğer açıklanan değişken sürekliyse F testi, kategorikse ki-kare testi kullanılmaktadır. CHAID analizi ki-kare metriği aracılığıyla ilişki seviyesine göre farklılık rastlanan grupları birbirinden ayrı sınıflandırmakta ve karar ağacının yaprakları, ikili değil de, verideki farklı yapı sayısı kadar dallanmaktadır (Altan ve ark., 2015).

Genel anlamda CHAID yönteminin algoritmasını açıklamak gerekirse; açıklanan değişken $d \geq 2$ kategoriye, analizde yer alan açıklayıcı değişken ise $c \geq 2$ kategoriye sahip olsun. Analizde, bağımsız değişkenin benzer kategorileri birleştirilerek $c * d$ boyutlu çapraz tablodan, $j * d$ boyutlu en önemli indirgenmiş tablo oluşturulur. $j * d$ tablosu oluşturulurken $c * d$ tablosu için χ^2 istatistiği olan $T_j^{(i)}$ istatistiği hesaplanır $(j = 2, 3, \dots, c;\ i'$nin aralığı bağımsız değişkenin türüne bağlı). Eğer $T_j^{(*)} = max_i T_j^{(i)}$ en iyi $j * d$ tablosu için χ^2 istatistiği ise en önemli $T_j^{(*)}$ seçilmiş olur. Algoritmanın tüm basamakları adım adım aşağıda belirtilmektedir (Kass, 1980):

Adım 1: Her bir açıklayıcı değişken için açıklanan değişkenin kategorileri ile açıklayıcı değişkenin kategorileri arasında çapraz tablo elde edilir ve 2'nci ve 3'üncü adım uygulanır.

Adım 2: $2 * d$ alt tablosunda açıklayıcı değişkene ilişkin anlamlılığı en düşük çıkan kategori çiftleri bulunur ve birleşmeleri anlamlı olan iki kategori birleştirilir. Bu birleşim tek bir kategori olarak düşünülür. Bu basamak, açıklayıcı değişkenin kendi içerisindeki birleşmeleri anlamsız oluncaya dek devam eder.

Adım 3: Üç veya daha fazla sayıda orijinal kategori içeren birleşik kategorilerin her birisi için birleşmenin yeniden çözümlendiği en mühim iki bölünme bulunur. Eğer ki anlamlılık kritik bir değerin altında ise bölünme tamamlanarak ikinci basamağa dönülür.

Adım 4: Optimum düzeyde bir araya getirilip birleştirilen açıklayıcı değişkenlerin her birisinin anlamlılığı hesaplanır ve en önemli olan ayrılır. Eğer ki anlamlılık bir kritik değerden büyük ise seçilen açıklayıcı değişkenin birleştirilen kategorilerine bakılarak veri alt gruplara bölünür.

Adım 5: Henüz analiz edilmemiş data için birinci adıma dönülür (Kass, 1980).

Her açıklayıcı değişken için kendi içerisinde kategorileri anlamlı bir şekilde birleştirilerek en iyi bölünme bulunduktan sonra, açıklanan değişkene göre çapraz tablo oluşturulur. Bundan sonra, χ^2 ve Bonferroni p değeri hesaplanır. Açıklayıcı değişkenler birbiriyle karşılaştırılıp en küçük p değerine sahip olan açıklayıcı değişkenin kategorilerine göre veri alt gruplara ayrılır (Doğan ve Özdamar, 2003).

Bonferroni yaklaşımı, her bir grubun ortalama vektörlerinin genel ortalama vektöründen farkları bulunduktan sonra bu farkların sıfır olup olmadığını araştırmaya dayanmaktadır. Genel ortalama vektörü $\bar{x}$ ve her grubun i'inci değişkene göre ortalama vektörleri $\bar{x}_{.g}$ aşağıdaki gibi gösterilmektedir;

$$\bar{x} = \begin{bmatrix} \bar{x}_{1.} \\ \bar{x}_{2.} \\ \cdot \\ \cdot \\ \bar{x}_{p.} \end{bmatrix} \quad \bar{x}_{.1} = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{21} \\ \cdot \\ \cdot \\ \bar{x}_{p1} \end{bmatrix} \quad \bar{x}_{.2} = \begin{bmatrix} \bar{x}_{12} \\ \bar{x}_{22} \\ \cdot \\ \cdot \\ \bar{x}_{p2} \end{bmatrix} \quad \dots \quad \bar{x}_{.g} = \begin{bmatrix} \bar{x}_{1g} \\ \bar{x}_{2g} \\ \cdot \\ \cdot \\ \bar{x}_{pg} \end{bmatrix} \tag{4.17}$$

Her bir grubun ortalama vektörünün, genel ortalama vektöründen farkları değişkenlere göre aşağıdaki gibi belirlenmektedir;

$$d_1 = \bar{x}_{.1} - \bar{x} = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{21} \\ \cdot \\ \cdot \\ \bar{x}_{p1} \end{bmatrix} - \begin{bmatrix} \bar{x}_{1.} \\ \bar{x}_{2.} \\ \cdot \\ \cdot \\ \bar{x}_{p.} \end{bmatrix} \quad \ldots \quad d_g = \bar{x}_{.g} - \bar{x} = \begin{bmatrix} \bar{x}_{1g} \\ \bar{x}_{2g} \\ \cdot \\ \cdot \\ \bar{x}_{pg} \end{bmatrix} - \begin{bmatrix} \bar{x}_{1.} \\ \bar{x}_{2.} \\ \cdot \\ \cdot \\ \bar{x}_{p.} \end{bmatrix} \tag{4.18}$$

k'ıncı grup ile *ℓ*'inci grup *i*'inci değişken ortalamaları arasındaki ortalama farkları arasındaki $1 - \alpha$ güven aralığı aşağıdaki biçimde hesaplanmaktadır.

$$(d_{ki} - d_{\ell i}) = (\bar{x}_{ki} - \bar{x}_{\ell i}) \pm t\left(\frac{\alpha}{pg(g-1)}, (N-g)\right)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{w_{ii}}{N-g}} \tag{4.19}$$

Burada, $N = n_1 + n_2 + \ldots + n_g$, p değişken sayısı, g grup sayısı ve w_{ii}, W matrisinin köşegen elemanlarıdır. W matrisi, gruplar içi değişimi gösterir ve;

$$W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \tag{4.20}$$

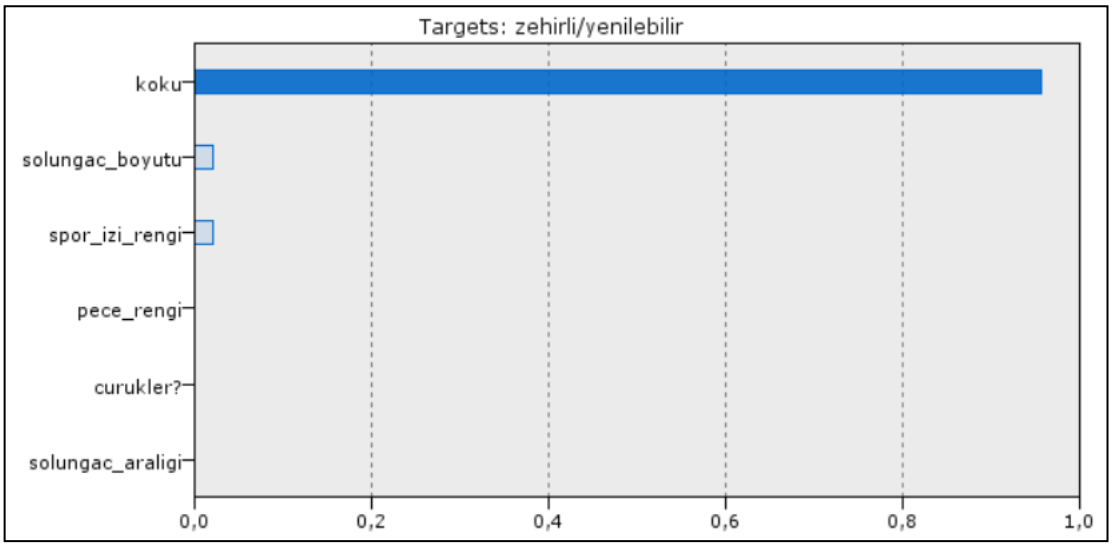
"g = grup sayısı", "n_i = i'inci gruptaki veri sayısı" olacak biçimde hesaplanır. Her bir değişken için gruplar ikişerli olacak şekilde dikkate alınmaktadır ve denklem (4.19) kullanılarak *i*'inci değişken için elde edilen aralığın sıfır değerini içerip içermediği kontrol edilir. Eğer sıfır değeri belirlenen aralıkta yer alıyor ise ilgili gruplar arasında anlamlı bir farklılık olmadığı, aksi takdirde grupların farklı oldukları biçiminde yorumlanmaktadır (Doğan, 2003).

5. UYGULAMA

Çalışmanın uygulama aşamasında, The UCI Machine Learning Repository veritabanından temin edilen mantar veri setinde yer alan 8124 adet gözlem değeri analiz edilmiştir. Veri setinde yer alan her bir gözleme (mantara) ilişkin 22 bağımsız değişkenin değeriyle mantarların zehirli ya da yenilebilir olduğu verisi kullanılarak mantarlar veri madenciliği sınıflandırma yöntemlerinden karar ağaçları algoritmaları ile zehirli veya yenilebilir şeklinde sınıflandırılmaya çalışılmıştır. Ayrıca, mantarların zehirli olup olmadığının tespitine etki eden fenotipik özellikler belirlenmeye çalışılmıştır. Burada, karar ağaçları algoritmalarından en yaygın kullanıma sahip olan C5.0, CART ve CHAID algoritmaları kullanılmış olup analizler SPSS programının bir modülü olan Clementine 12.0 aracılığıyla yapılmıştır. Çalışma kapsamında incelenen değişkenler ve düzeyleri daha önce Tablo 4.1'de verilmiştir.

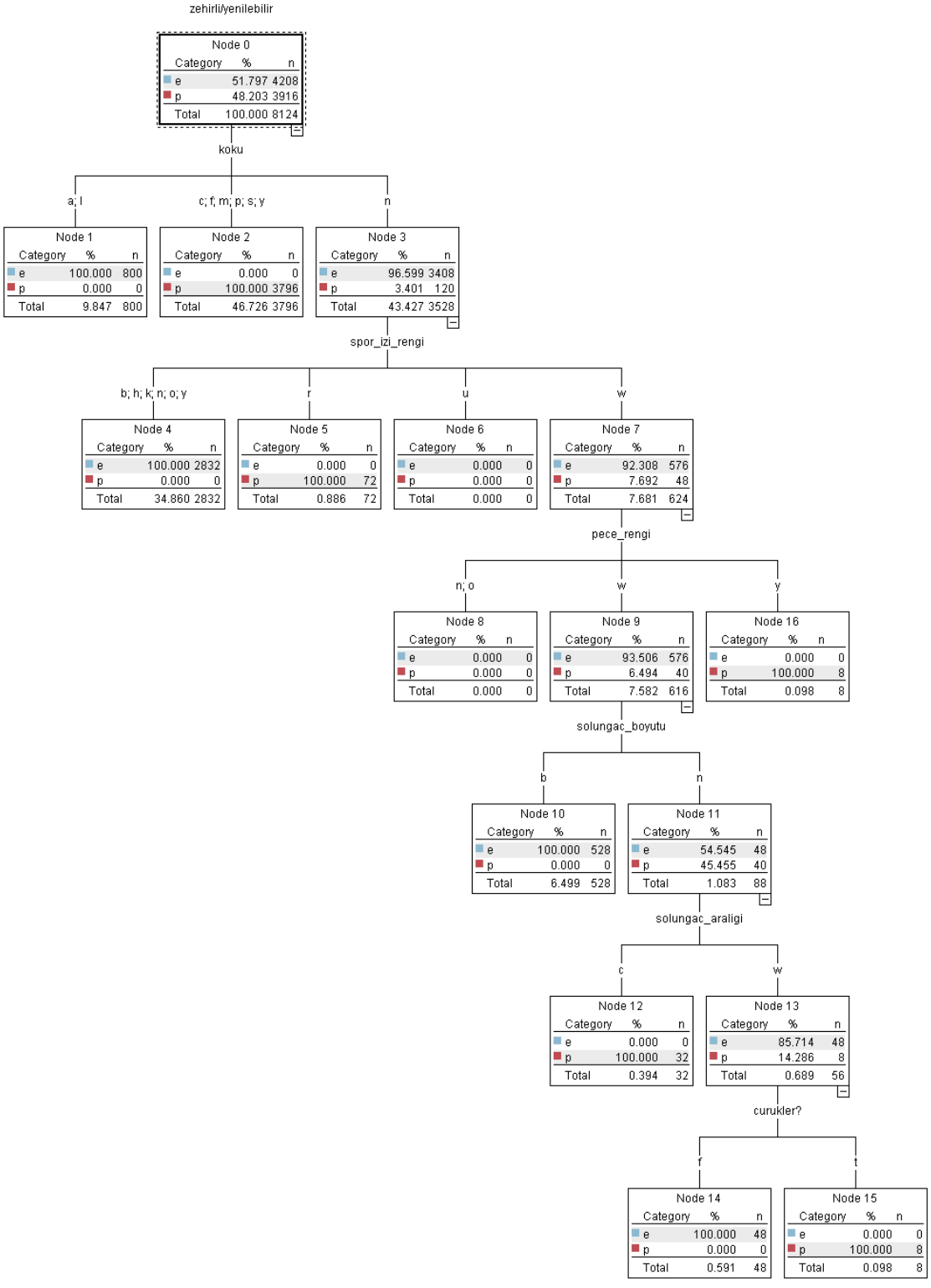
5.1. C5.0 Algoritması ile Elde Edilen Sonuçlar

C5.0 modeline göre mantarın zehirli olup olmadığının tespitinde en önemli faktör mantarın kokusudur. Bunu sırasıyla mantarın solungaç boyutu ve mantarın spor izi rengi izlemektedir. Değişkenlerin önem sırası Şekil 5.1'de gösterilmiştir.



Şekil 5.1. Mantarın zehirli olup olmadığının tespitine etki eden değişkenlerin önem sırası

Şekil 5.2'de, C5.0 modeline göre elde edilen karar ağacı görülmektedir. Her bir düğümde, bölünmenin gerçekleştiği dallara göre, zehirli/yenilebilir mantarların oranları görülebilmektedir.



Şekil 5.2. C5.0 modeliyle elde edilen karar ağacı

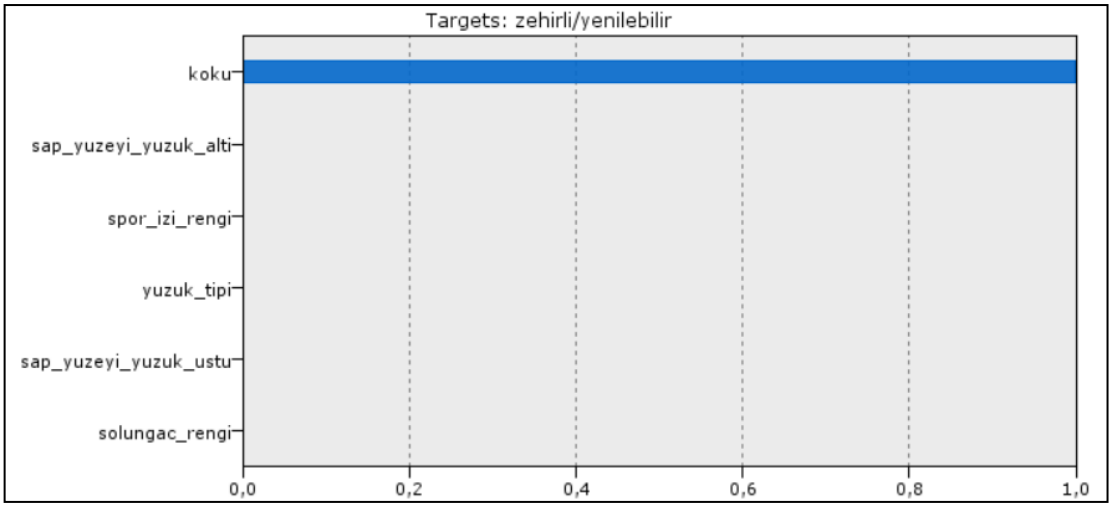
Tablo 5.1. C5.0 modelinin tahminine ilişkin anma (recall) değerleri

Gerçek Sınıf	Modelden Tahmin Edilen Sınıf	
	yenilebilir	**zehirli**
yenilebilir	100,0	0
zehirli	0	100,0

Anma değeri, bağımlı değişkenin herhangi bir düzeyi için modelin, düzeyin ne kadarını doğru, ne kadarını yanlış tahmin ettiğini göstermektedir. Tablo 5.1 incelendiğinde, C5.0 modelinin bağımlı değişkenin her iki düzeyi içinde %100 doğru tahmin yaptığı görülmektedir.

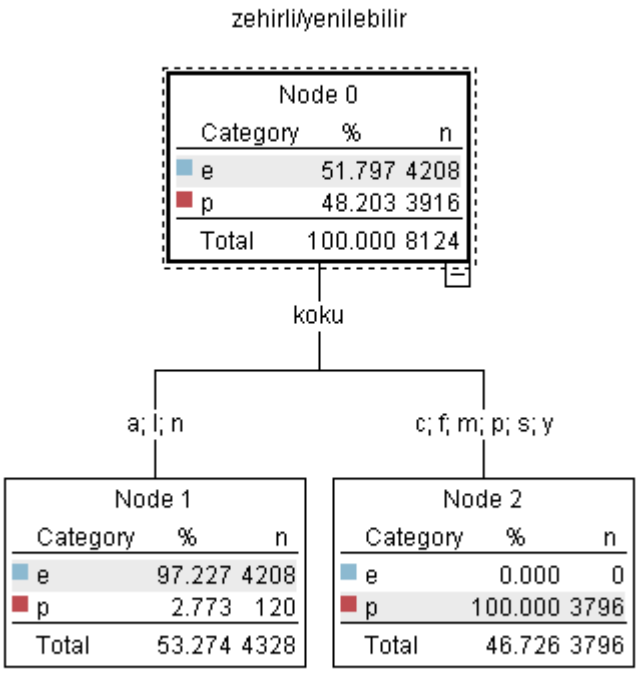
5.2. CART Algoritması ile Elde Edilen Sonuçlar

CART modeline göre mantarın zehirli olup olmadığının tespitinde en önemli faktör mantarın kokusudur. Değişkenlerin önem sırası Şekil 5.3'te gösterilmiştir.



Şekil 5.3. Mantarın zehirli olup olmadığının tespitine etki eden değişkenlerin önem sırası

Şekil 5.4'te, CART modeline göre elde edilen karar ağacı görülmektedir. Her bir düğümde, bölünmenin gerçekleştiği dallara göre, zehirli/yenilebilir mantarların oranları görülebilmektedir.



Şekil 5.4. CART modeliyle elde edilen karar ağacı

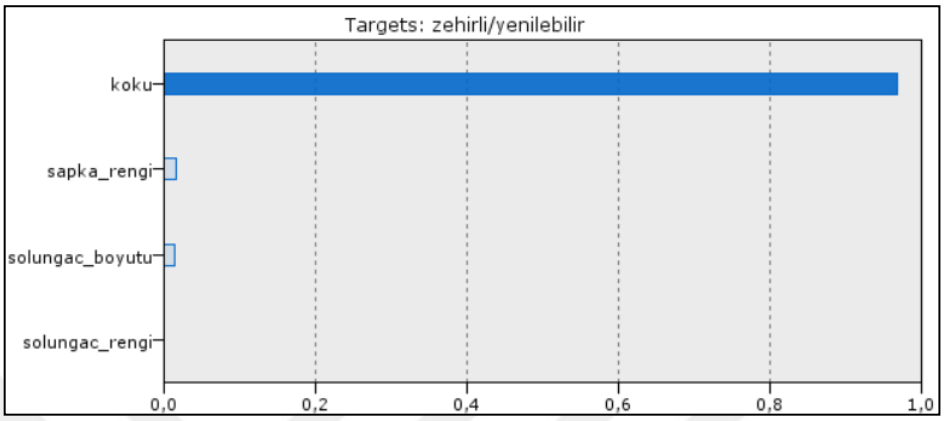
Tablo 5.2. CART modelinin tahminine ilişkin anma (recall) değerleri

Gerçek Sınıf	**Modelden Tahmin Edilen Sınıf**	
	yenilebilir	**zehirli**
yenilebilir	100,0	0
zehirli	3,06	96,94

Tablo 5.2 incelendiğinde, CART modelinin, bağımlı değişkenin "yenilebilir" sınıfı için %100 doğru tahmin yaparken "zehirli" sınıfı için %96,94'ünü doğru olarak tahmin ettiği görülmektedir.

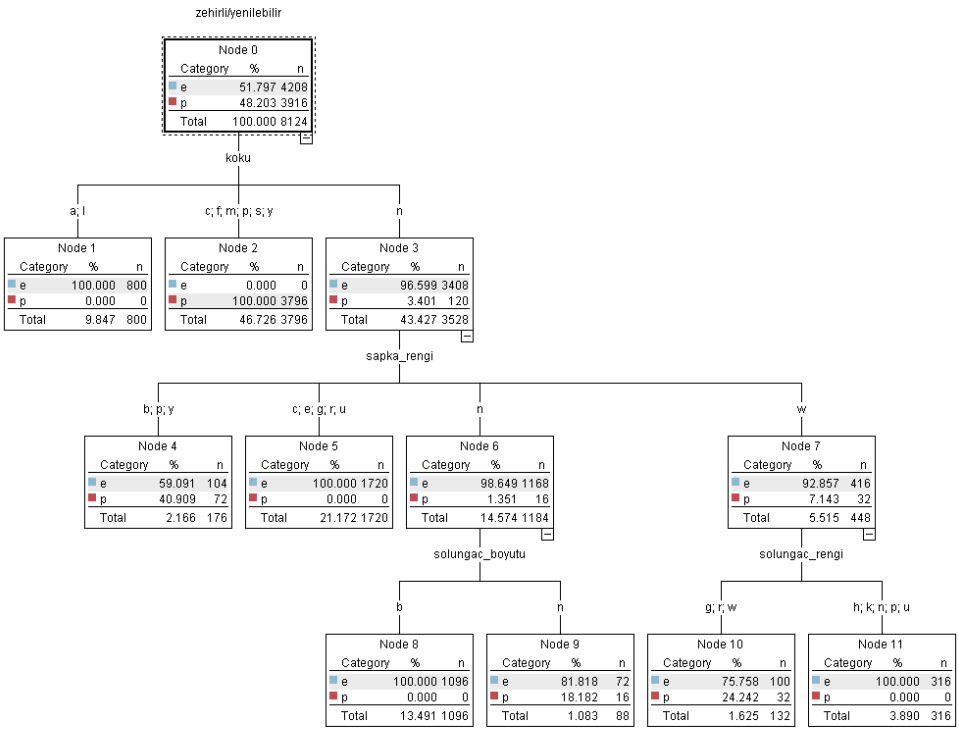
5.3. CHAID Algoritması ile Elde Edilen Sonuçlar

CHAID modeline göre mantarın zehirli olup olmadığının tespitinde en önemli faktör mantarın kokusudur. Bunu sırasıyla mantarın şapka rengi ve mantarın solungaç boyutu izlemektedir. Değişkenlerin önem sırası Şekil 5.5'te gösterilmiştir.



Şekil 5.5. Mantarın zehirli olup olmadığının tespitine etki eden değişkenlerin önem sırası

Şekil 5.6'da, CHAID modeline göre elde edilen karar ağacı görülmektedir. Her bir düğümde, bölünmenin gerçekleştiği dallara göre, zehirli/yenilebilir mantarların oranları görülebilmektedir.



Şekil 5.6. CHAID modeliyle elde edilen karar ağacı

Tablo 5.3. CHAID modelinin tahminine ilişkin anma (recall) değerleri

Gerçek Sınıf	Modelden Tahmin Edilen Sınıf	
	yenilebilir	**zehirli**
yenilebilir	100,0	0
zehirli	3,06	96,94

Tablo 5.3 incelendiğinde, CHAID modelinin, bağımlı değişkenin "yenilebilir" sınıfı için %100 doğru tahmin yaparken "zehirli" sınıfı için %96,94'ünü doğru olarak tahmin ettiği görülmektedir. Bu doğru tahmin oranları CART modeli aynı düzeyde olup C5.0 modelinin doğru tahmin yüzdeleri ile mukayese edildiklerinde "zehirli" sınıfı için daha düşük kalmaktadır. Zira, C5.0 modeli ile bağımlı değişkenin her iki düzeyi içinde %100 doğru tahminler yapılmıştır. Burada, tüm modeller birbirine yakın sonuçlar vermektedir. Ancak, en iyi model en yüksek doğru tahmin oranlarıyla C5.0 modeli olmaktadır.

Mantar verisi için en iyi modeli sunan C5.0 karar ağacı algoritması ile elde edilen kurallar listesi aşağıda verilmektedir. "Yenilebilir" sınıfı için 4 adet kural, "zehirli" sınıfı için de 5 adet kural olmak üzere toplam 9 adet kural listelenmiştir.

"Yenilebilir" sınıfı için C5.0 kuralları:

Kural 1:

```
if koku in [ "a" "l" ]
 then e
```

Eğer mantar badem (a) veya anason (l) kokulu ise bu durumda kural mantarın yenilebilir olacağını göstermektedir.

Kural 2:

```
if koku in [ "n" ]
 and spor_izi_rengi in [ "b" "h" "k" "n" "o" "y" ]
 then e
```

Eğer mantarın kokusu yok (n) ise ve spor izi rengi ten rengi (b), çikolata (h), siyah (k), kahverengi (n), turuncu (o) veya sarı (y) renklerinden herhangi biri ise bu durumda kural mantarın yenilebilir olacağını göstermektedir.

Kural 3:

```
if koku in [ "n" ]
 and spor_izi_rengi in [ "w" ]
 and pece_rengi in [ "w" ]
 and solungac_boyutu = b
 then e
```

Eğer mantarın kokusu yok (n) ise ve spor izi rengi beyaz (w) ise ve peçe rengi beyaz (w) ise aynı zamanda solungaç boyutu kalın (b) ise bu durumda kural mantarın yenilebilir olacağını göstermektedir.

Kural 4:

```
if koku in [ "n" ]
 and spor_izi_rengi in [ "w" ]
 and pece_rengi in [ "w" ]
 and solungac_boyutu = n
 and solungac_araligi = w
 and curukler? = f
 then e
```

Eğer mantarın kokusu yok (n) ise ve spor izi rengi beyaz (w) ise ve peçe rengi beyaz (w) ise aynı zamanda solungaç boyutu dar (n) ise ve solungaç aralığı kalabalık (w) ise ve de çürükler yok (f) ise bu durumda kural mantarın yenilebilir olacağını göstermektedir.

“Zehirli” sınıfı için C5.0 kuralları:

Kural 1:

```
if koku in [ "c" "f" "m" "p" "s" "y" ]
 then p
```

Eğer mantar kreozot/katran ruhu (c), bozuk/iğrenç (f), küf (m), mayhoş/acı (p), baharatlı (s) veya balık (y) kokulu ise bu durumda kural mantarın zehirli olacağını göstermektedir.

Kural 2:

```
if koku in [ "n" ]
 and spor_izi_rengi in [ "r" ]
 then p
```

Eğer mantarın kokusu yok (n) ise ve spor izi rengi yeşil (r) ise bu durumda kural mantarın zehirli olacağını göstermektedir.

Kural 3:

```
if koku in [ "n" ]
 and spor_izi_rengi in [ "w" ]
 and pece_rengi in [ "w" ]
 and solungac_boyutu = n
 and solungac_araligi = c
 then p
```

Eğer mantarın kokusu yok (n) ise ve spor izi rengi beyaz (w) ise ve peçe rengi beyaz (w) ise aynı zamanda solungaç boyutu dar (n) ise ve solungaç aralığı da yakın (c) ise bu durumda kural mantarın zehirli olacağını göstermektedir.

Kural 4:

```
if koku in [ "n" ]
 and spor_izi_rengi in [ "w" ]
 and pece_rengi in [ "w" ]
 and solungac_boyutu = n
 and solungac_araligi = w
 and curukler? = t
 then p
```

Eğer mantarın kokusu yok (n) ise ve spor izi rengi beyaz (w) ise ve peçe rengi beyaz (w) ise aynı zamanda solungaç boyutu dar (n) ise ve solungaç aralığı kalabalık (w) ise ve de çürükler var (t) ise bu durumda kural mantarın zehirli olacağını göstermektedir.

Kural 5:

```
if koku in [ "n" ]
 and spor_izi_rengi in [ "w" ]
 and pece_rengi in [ "y" ]
 then p
```

Eğer mantarın kokusu yok (n) ise ve spor izi rengi beyaz (w) ise ve peçe rengi sarı (y) ise bu durumda kural mantarın zehirli olacağını göstermektedir.

6. SONUÇLAR VE ÖNERİLER

Büyük miktarda veri denildiğinde, çağrıştırmış olduğu ilk kavram "Veri Madenciliği" dir. Veri madenciliği, birçok analiz yöntemini kullanarak veri içerisinde örüntü ve ilişkileri keşfedip bunları geçerli tahminler yapmak için kullanan bir süreçtir. Veri madenciliğinin amacı, geçmiş faaliyetlerin analizini temel alarak gelecekteki davranışların tahminine yönelik karar verme modelleri oluşturmaktır. Veri madenciliği, 1990'lı yıllardan bu yana, büyük miktarda veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgileri açığa çıkarmak ve stratejik karar desteği sağlamak amacıyla kullanılmaktadır (Koyuncugil ve Özgülbaş, 2009).

Veri madenciliği aslında klasik istatistiksel yöntemlere oldukça fazla benzemektedir (Özkan, 2008). İstatistik, verilerin analizinde, birçok problemi çözmede önemli bir araçtır. Fakat, bazı durumlarda kullanımı sınırlıdır. Böyle durumlarda, yapay sinir ağları, genetik algoritma, karar ağaçları gibi akıllı veri analiz yöntemlerini içeren veri madenciliğine ihtiyaç duyulmaktadır. Fakat, yöntemlerden bazıları kara kutu yaklaşımları şeklindedir. Farklı bir deyişle, açıklama getirmeden iyi tahmin yaparlar. Bu da, o yöntemlerin zayıf yönünü oluşturmaktadır. Karar ağacı yöntemi ise bu zayıflığın söz konusu olmadığı veri madenciliği yöntemlerindendir (Emel ve Taşkın, 2005).

Veri madenciliği denildiğinde, akla gelen ilk yöntemlerden olan karar ağaçları, sınıflandırma yöntemlerindendir ve C5.0, CART ve CHAID en popüler algoritmalardır (Koyuncugil ve Özgülbaş, 2009).

Çalışmanın uygulama aşamasında da görüldüğü üzere, veri setinden karar ağaçları oluşturulduktan sonra, karar ağacında kök düğümden yapraklara doğru gidilerek kurallar (if - then) anlaşılır bir biçimde yazılabilmektedir. Analizlerde, C5.0, CART ve CHAID karar ağacı algoritmalarından yararlanılmıştır.

Analiz sonuçlarına göre, CHAID ve CART modellerinin, bağımlı değişkenin "yenilebilir" sınıfı için %100 doğru tahmin yaptığı tespit edilirken "zehirli" sınıfı için %96,94'ünü doğru olarak tahmin edebildiği belirlenmiştir. Bu doğru tahmin yüzdeleri C5.0 modelinin doğru tahmin yüzdeleri ile mukayese edildiğinde, "zehirli" sınıfı için CART ve CHAID modellerinin doğru tahmin yüzdeleri düşük kalmıştır. Zira, C5.0 modeli ile bağımlı değişkenin her iki sınıfı (zehirli/yenilebilir) içinde %100 doğru tahminler yapılmıştır. Analizlerde, tüm modeller birbirine yakın ve iyi sonuçlar vermiştir denilebilmektedir. Ancak, ilgilenilen veri seti için en iyi model, en yüksek doğru tahmin yüzdeleriyle C5.0 modeli olmaktadır.

C5.0 modeline göre mantarın zehirli olup olmadığının tespitinde en önemli değişkenin mantarın kokusu olduğu ortaya çıkmıştır. Bunu sırasıyla mantarın solungaç boyutu ve mantarın spor izi rengi izlemiştir.

Mantar verisi için en iyi modeli sunan C5.0 karar ağacı algoritmasıyla ortaya çıkarılan kurallar listesine göre; "yenilebilir" sınıfı için 4 adet kural, "zehirli" sınıfı için de 5 adet kural olmak üzere toplam 9 adet kural elde edilmiştir.

Doğada mevcut olan binlerce şapkalı mantar türünün bir kısmı zehirli, bir kısmı ise zehirsiz özellik göstermesine rağmen, bu mantarlar birbirlerine fenotipik (koku, spor izi rengi, solungaç boyutu, vb.) açıdan oldukça benzemektedirler. Bu sebeple, zehirli ve zehirsiz mantarların, toplanma aşamasında, benzer fenotipik özellikleri nedeniyle karıştırılması sonucu zaman zaman mantar zehirlenmesi vakaları yaşanabilmektedir. Çalışmaya bu açıdan yaklaşıldığında, her ne kadar yapılan analizler neticesinde "yenilebilir" sınıfı için 4 adet kural (örneğin; badem ve anason kokulu mantarların yenilebilir olması gibi) tespit edilmiş olsa da, zehirlenme vakalarının ölüme kadar gidebildiği önemle dikkate alındığında, bu çalışmanın sonuçlarının deneysel çalışmalarla desteklenmesi ve ayrıca, daha büyük ve daha farklı veri setleriyle geliştirilmesi gerektiği değerlendirilmektedir.

KAYNAKLAR

Akbal, E., Doğan, Ş. ve Varol, N., 2017, Karar Ağaçları ile Telefon Dolandırıcılığı Verilerinin Analizi, *Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi,* 29 (1), 171-177.

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

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

Aldana, W., A. (2000), Data Mining Industry: Emerging Trends and New Opportunities", Yüksek Lisans Tezi, *Massachusetts Institute of Technology*, USA, 11.

Alpaydın, E., 2000, Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri [online], https://www.cmpe.boun.edu.tr/~ethem/files/papers/veri-maden_2k-notlar.doc [Ziyaret Tarihi: 03.03.2018].

Altan, Ş., Atan, M. ve Kızılkaya, S., 2015, Genel Sağlık Durumunu Etkileyen Faktörlerin CHAID Analizi Yöntemi ile İncelenmesi, ODTÜ Örneği, *Journal of New World Sciences Academy-Social Sciences*, 10, 92-106.

Altıkardeş, A., Erdal, H., Baba, F. ve Fak, A., S., 2012, ABPM Ölçümü Olmaksızın Karar Ağaçları Algoritması ile Non-Dipper/Dipper Öngörüsü, *Tıp Bilişimi '12, IX. Ulusal Tıp Bilişimi Kongresi*, Antalya, 77-86.

Altun, M., (2017), Veri Madenciliği ve Uygulama Alanları", Doktora Semineri Raporu, *Akdeniz Üniversitesi EYTEPE ABD Doktora Programı, Eğitim Bilimleri Bölümü,* Antalya, 2-27.

Annoni, P., Ferrari, P. ve Salini, S., 2009, A Data Mining Approach to Advance Knowledge in Public Government: Profiling Households. In Rahman, H. (Ed.), *Social and Political Implications of Data Mining: Knowledge Management in E-Government*, Information Science Reference, New York, 107.

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

Avşar, A., Ş. ve Yalçın, S., 2015, Öğrencilerin Okuma Başarılarını Açıklayan Ailesel Değişkenlerin CHAID Analizi ile Belirlenmesi, *Eğitim ve Bilim,* 40 (179), 1-9.

Bossenmaier, E., F., (1997), Mushrooms of the Boreal Forest [online], Lone Pine Publishing, https://www.usask.ca/biology/fungi/glossary.html, [Ziyaret Tarihi: 10 Kasım 2019].

Doğan, İ., 2003, Holştayn Irkı İneklerde Süt Verimine Etki Eden Faktörlerin Chaid Analizi ile İncelenmesi, *Ankara Üniversitesi Veterinerlik Fakültesi Dergisi*, 50, 65-70.

Doğan, N. ve Özdamar, K., 2003, Chaid Analizi ve Aile Planlaması ile İlgili Bir Uygulama, *T. Klinik Tıp Bilimleri*, 23, 392-398.

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

Ergün, E., (2008), Ürün Kategorileri Arasındaki Satış ilişkisinin Birliktelik Kuralları ve Kümeleme Analizi ile Belirlenmesi ve Perakende Sektöründe Bir Uygulama, Doktora Tezi, *Afyon Kocatepe Üniversitesi Sosyal Bilimler Enstitüsü*, Afyonkarahisar, 39-40.

Ersöz, T. ve Özseven, T., 2015, Kobi'lerin Finansal Sorunlarını Etkileyen Faktörlerin CRT Karar Ağacı ile Modellenmesi, *AİBÜ Sosyal Bilimler Enstitüsü Dergisi,* 15 (1), 1-20.

Gökçen Alıç, Z., H., (2014), Akut Pankreatit Hastalarının Mortalite Risklerinin Karar Ağacı Yöntemi ile Belirlenmesi, Yüksek Lisans Tezi, *Gazi Üniversitesi Bilişim Enstitüsü,* Ankara, 15-16.

Irmak, S. ve Ercan, U., 2017, Karar Ağaçları Kullanılarak Türkiye Hanehalkı Zeytinyağı Tüketimi Görünümünün Belirlenmesi, *Uluslararası Yönetim İktisat ve İşletme Dergisi,* 13 (3), 553-564.

Kalıpsız, O., Güven, A. ve Bozkurt, Ö., Ö., 2018, Veri Madenciliğinin Geleceği [online], http://www.bilgis.org/download/veri-madenciliginin-gelecegi.doc [Ziyaret Tarihi: 01.03.2018].

Kass, G.,V., 1980, An Exploratory Technique for İnvestigating Large Quantities of Categorical Data, *Applied Statis*tics, 29, 119-127.

Kavzoğlu, T. ve Çölkesen, İ., 2010, Karar Ağaçları ile Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örneği, *Harita Teknolojileri Elektronik Dergisi,* 2, 36-45.

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

Lincoff, G. H. (1981). The Audubon Society Field Guide to North American Mushrooms [online], New York, http://archive.ics.uci.edu/ml/datasets/Mushroom [Ziyaret Tarihi: 29 Ekim 2019].

Okatan, A., (2019), Türkiye'nin Zehirli Mantarları [online], Tübitak Bilim Genç, http://www.bilimgenc.tubitak.gov.tr/makale/turkiyenin-zehirli-mantarlari, [Ziyaret Tarihi: 10 Kasım 2019].

Öğüt, S., 2014, Veri Madenciliği Kavramı ve Gelişim Süreci [online], http://sertacogut.com/blog/?p=6 [Ziyaret Tarihi: 02.03.2018].

Özkan, Y., 2008, Veri Madenciliği Yöntemleri, *Papatya Yayıncılık Eğitim*, İstanbul, 37-106.

Savaş, S., Topaloğlu, N. ve Yılmaz, M., 2012, Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 11, 1-23.

Silahtaroğlu, G., 2008, Kavram ve Algoritmalarıyla Temel Veri Madenciliği, *Papatya Yayıncılık Eğitim*, İstanbul, 56.

Vahaplar, A. ve İnceoğlu, M., M., 2001, Veri Madenciliği ve Elektronik Ticaret, *VII. Türkiye'de İnternet Konferansları*, Harbiye, İstanbul, 1-7.

Wani, B., A., Bodha, R., H. and Wani, A., H., (2010), Nutritional and medicinal importance of mushrooms, *Journal of Medicinal Plants Research*, 4(24), 2598-2604.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Nuriye TOPYILDIZ
Uyruğu : Türkiye Cumhuriyeti
Doğum Yeri ve Tarihi : Ankara, 1990
Telefon : -
e-mail : nuriye.topyildiz@meb.gov.tr

EĞİTİM

Derece		Adı, İlçe, İl	Bitirme Yılı
Lise	:	**Çağrıbey Anadolu Lisesi/Ankara**	**2008**
Üniversite	:	**Hacettepe Üniversitesi, İstatistik Bölümü/Ankara**	**2012**
Yüksek Lisans	:	**Gazi Üniversitesi, Ekonometri A.B.D./Ankara**	**2019**
Yüksek Lisans	:	**Selçuk Üniversitesi, İstatistik A.B.D./Konya**	**Devam ediyor**

İŞ DENEYİMLERİ

Yıl	Kurum	Görevi
2014 - 2015	**Sosyal Güvenlik Kurumu**	**Sosyal Güvenlik Denetmen Yardımcısı**
2015 - 2019	**Millî Eğitim Bakanlığı**	**Millî Eğitim Uzman Yardımcısı**
2019 - …	**Millî Eğitim Bakanlığı**	**Millî Eğitim Uzmanı**
2018 - …	**Millî Eğitim Bakanlığı**	**Bakan Yardımcısı Özel Kalem Müdürü**

UZMANLIK ALANI : İstatistik, Veri Madenciliği, Ekonometri

YABANCI DİLLER : İngilizce

YAYINLAR : Topyıldız, N. (2018). *Cumhuriyet Dönemi Boyunca Eğitim Harcamalarının İktisadi Büyüme ile İlişkisinin Ekonometrik Analizi*. Millî Eğitim Bakanlığı, Millî Eğitim Uzmanlık Tezi, Ankara.

Topyıldız, N. (2019). *Doğrusal Olmayan Modellerle Türkiye'de Turizm Sektörünün Yakınsama Hipotezinin Sınanması.* Gazi Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, Ekonometri Bilim Dalı, Ankara.