T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
İŞLETME BİLİM DALI

SAĞLIK İŞLETMELERİNDE VERİ MADENCİLİĞE İLE HASTA PROFİLLERİNİN TESPİTİ İÇİN BİR SİMÜLASYON UYGULAMASI

MUHAMMET ALİ KAYA

YÜKSEK LİSANS

DANIŞMAN:
DR. ÖĞR. ÜYESİ ÜMRAN MÜNİRE KAHRAMAN

KONYA-2024

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Sosyal Bilimler Enstitüsü

Bilimsel Etik Sayfası

Öğrencinin	Adı Soyadı	MUHAMMET ALİ KAYA		
	Numarası	22811101938		
	Ana Bilim / Bilim Dalı	İşletme / İşletme		
	Programı	Tezli Yüksek Lisans	X	
		Doktora		
	Tezin Adı	Sağlık İşletmelerinde Veri Madenciliği ile Hasta Profillerinin Tespiti İçin Bir Simülasyon Uygulaması		

Bu tezin hazırlanmasında bilimsel etiğe ve akademik kurallara özenle riayet edildiğini, tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada başkalarının eserlerinden yararlanılması durumunda bilimsel kurallara uygun olarak atıf yapıldığını bildiririm.

Öğrencinin Adı Soyadı
İmzası

Muhammet Ali KAYA

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Sosyal Bilimler Enstitüsü

ÖZET

Öğrencinin	Adı Soyadı	MUHAMMET ALİ KAYA		
	Numarası	22811101938		
	Ana Bilim / Bilim Dalı	İşletme / İşletme		
	Programı	Tezli Yüksek Lisans	X	
		Doktora		
	Tez Danışmanı	Dr. Öğr. Üyesi Ümran Münire KAHRAMAN		
	Tezin Adı	Sağlık İşletmelerinde Veri Madenciliği ile Hasta Profillerinin Tespiti İçin Bir Simülasyon Uygulaması		

Bu çalışma, sağlık verilerinin analizine yönelik olarak veri madenciliği tekniklerinden K-Means kümeleme algoritmasını kullanarak hasta profillemesi yapmayı amaçlamaktadır. Sağlık hizmetlerinin daha verimli sunulabilmesi ve hasta yönetiminin iyileştirilmesi hedefiyle, hastaların demografik ve sağlık durumlarına göre farklı segmentlere ayrılması gerçekleştirilmiştir. Araştırma kapsamında, yaş, vücut kitle indeksi (BMI), hipertansiyon, kalp hastalığı ve sigara içme durumu gibi sağlık göstergeleri temel alınarak her bir hasta profili belirgin özellikleriyle analiz edilmiştir. Elde edilen bulgular, farklı hasta grupları için kişiselleştirilmiş tedavi ve önleyici stratejilerin geliştirilmesi açısından önemli bir temel sunmaktadır. Özellikle yüksek risk grubunda yer alan yaşlı ve kronik hastalık geçmişine sahip bireylerin tanımlanması, bu kişilerin sağlık hizmetlerinden daha etkin faydalanmasını sağlamayı amaçlamaktadır. Bu segmentasyon sayesinde sağlık hizmetlerinin daha etkili bir şekilde yönlendirilmesi ve hastane kaynaklarının daha verimli kullanılması öngörülmektedir. Sonuç olarak, bu çalışma, hasta profillemesinin sağlık sektörü açısından hem bireysel sağlık yönetiminde hem de genel sağlık politikalarının şekillendirilmesinde değerli bir araç olduğunu göstermektedir.

Anahtar Kelimeler: Veri Madenciliği, Hasta Profilleme, K-Means

T.C.
NECMETTİN ERBAKAN ÜNİVERSİTESİ
Sosyal Bilimler Enstitüsü

ABSTRACT

Author's	Name and Surname	MUHAMMET ALİ KAYA		
	Student Number	22811101938		
	Department	Business Administration		
	Study Programme	Master's Degree (M.A.)	X	
		Doctoral Degree (Ph.D.)		
	Supervisor	Assistant Professor Ümran Münire KAHRAMAN		
	Title of the Thesis/Dissertation	A Simulation Application for Determining Patient Profiles with Data Mining in Healthcare Facilities		

This study aims to apply K-Means clustering, a data mining technique, to develop patient profiles through the analysis of health data. With the goal of enhancing healthcare delivery and improving patient management, patients are segmented based on demographic and health status indicators. Key health indicators such as age, body mass index (BMI), hypertension, heart disease, and smoking status were used to categorize and analyze distinct patient profiles. The findings provide a crucial foundation for developing personalized treatment and preventive strategies tailored to various patient groups. In particular, the identification of high-risk individuals—older adults and those with chronic health conditions—seeks to improve access to healthcare services for these populations. This segmentation enables more targeted healthcare provision and efficient use of hospital resources. In conclusion, this study demonstrates that patient profiling is a valuable tool in the healthcare sector, offering benefits both in individual health management and in shaping broader healthcare policies.

Keywords: Data Mining, Patient Profiling, K-Means

İÇİNDEKİLER

İKİNCİ BÖLÜM

HASTA SİMÜLASYONUNUN GERÇEKLEŞTİRİLMESİ

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Bilgisayar sistemlerinde depolanan veriler, işlenmeden bir anlam veya değer taşımamaktadır. Bu ham veriler, kurumlar ve sektörler açısından yalnızca işlemden geçtiklerinde anlam kazanarak zamandan tasarruf, maliyetin azaltılması gibi faydalar sağlar. Ham veriler doğrudan karar alma süreçlerinde yeterli olmaz; özellikle geçmişte yaşanan olumsuz deneyimlerden çıkarılacak dersler ve kayıpların önlenmesi gibi durumlar için işlenmiş verilere ihtiyaç duyulur. Burada önemli olan, geçmiş olaylara dair saklı bilgileri açığa çıkarmak ve geleceğe yönelik önleyici modeller kurarak olası riskleri en aza indirebilmektir. Bu çerçevede, veri madenciliği, geniş veri kümelerinde saklı kalan desenleri ve eğilimleri keşfetme amacıyla kullanılan bir yöntem olarak tanımlanabilir. Veri madenciliği teknikleriyle veriler anlamlı hale getirilir ve karar süreçlerinde kullanılabilecek stratejik bilgiye dönüştürülür (Savaş ve ark, 2012: 2).

Bilgi keşfi ve veri madenciliği, farklı disiplinleri bir araya getirerek verilerden anlamlı ve kullanışlı bilgiler elde etmeye odaklanır. Başlangıçta yalnızca belirli alanlardan profesyonellerin ilgi gösterdiği bu alana, günümüzde iş dünyası, bankacılık ve sağlık sektörü gibi birçok farklı sektör de yönelmeye başlamıştır. Özellikle sağlık alanında verilerin büyük miktarda olması ve insan sağlığı açısından taşıdığı önem, veri madenciliğinin bu sektördeki yerini daha da önemli hale getirmiştir. Tıbbi veri madenciliği, verilerin heterojen yapısı, etik ve hukuki sorumluluklar, hasta gizliliğini koruma zorunluluğu ve sosyal yönleriyle diğer sektörlerden farklı zorluklar sunar. Sağlık verilerinin analizinde yalnızca istatistiksel metotlar değil, aynı zamanda bu özel durumları dikkate alan yöntemler kullanılmaktadır. Veri madenciliğinde ilk aşama olarak, veriyi tanımak amacıyla tanımlayıcı istatistiklerin elde edilmesi, grafiklerle görselleştirme ve değişkenler arasındaki olası ilişkilerin incelenmesi tercih edilir. İstatistik bilimi, sınıflama ve yapısal analizi sorunlarını çözmede bu süreçlerin merkezinde yer alır. Ayrıca, bilgisayar teknolojilerindeki ilerlemeler, veri madenciliğinde yeni tekniklerin geliştirilmesini kolaylaştırmış ve çözümlerin daha erişilebilir hale gelmesine katkıda bulunmuştur (Köktürk ve ark, 2009: 20).

Sağlık bilgi teknolojilerindeki gelişmeler, veri tabanları ve veri ambarlarında büyük miktarda veriyi depolamayı kolaylaştırsa da bu verilerin anlamlı bilgiye

dönüştürülmesi konusunda sınırlamalar devam etmektedir. Sağlık sektörü açısından bu durumun başlıca nedeni, verilerin karar alıcılar tarafından doğru zamanda ve anlaşılabilir bir formatta erişilebilir olmamasıdır. Ayrıca, veriyi bilgiye dönüştüren süreçlerdeki mekanizmaların eksikliği ve istatistiksel analizlerin karmaşık veri kümelerinde yetersiz kalması, yeni analiz araç ve tekniklerine olan ihtiyacı artırmıştır. Bu bağlamda, veri madenciliği son yıllarda sağlık sektöründe önemli bir çözüm olarak benimsenmiştir. Veri madenciliği, sağlık veri tabanları veya diğer veri depolarındaki gizli bilgilere ulaşmak için kullanılan bir süreçtir. Bu süreçte veriler işlenerek eksik, gereksiz ve gürültülü verilerden arındırılır. Verilerin temizlenmesi ve öngörüler için hazır hale getirilmesi sağlanır ki bu süreç, sağlık alanındaki karar alma süreçlerinde etkin bir bilgi kaynağı oluşturur. Bununla birlikte, anlamlı sonuçlar elde etmek için veriler arasındaki ilişkilerin doğru anlaşılması önemlidir. Bu nedenle, veri görselleştirme teknikleri veri madenciliğinde önemli bir rol oynar; verinin görselleştirilmesi, karar vericilere veri hakkında genel bir bakış açısı sağlayarak analiz süreçlerini kolaylaştırır ve daha doğru tahmin ve kararların alınmasını destekler (Avcı ve Çınaroğlu, 2015: 55).

Sağlık sektörü, her hastanın muayene, laboratuvar işlemleri ve tedavi süreçlerinin kaydedilerek büyük miktarda verinin biriktiği bir alandır. Doktorlar, hastanın belirtilerine ve şikayetlerine dayanarak bir ön tanı koyarlar. Ancak bu ön tanının doğruluğu, tedavi sürecinin etkin planlanmasında büyük önem taşır; çünkü yanlış bir ön tanı, hastaların gereksiz testlere tabi tutulmasına ve bu nedenle zaman ve maliyet açısından hem hasta hem de kamu kaynaklarına gereksiz yük getirilmesine neden olabilir. Bu gibi durumların önüne geçmek amacıyla, ön tanının doğruluğunu arttıracak ve tanıya ulaşma sürecini hızlandıracak çeşitli testler ve analiz yöntemleri geliştirilmiştir. Bu testler, doktorların tanı koyarken daha isabetli kararlar almasını sağlayarak hem hasta memnuniyetini artırmakta hem de sağlık kaynaklarının daha verimli kullanılmasına katkıda bulunmaktadır (Talan, 2016: 2).

Sağlık verileri genellikle heterojen bir yapıdadır; hastaların demografik bilgileri, sağlık geçmişi ve tıbbi test sonuçları gibi çok sayıda veri farklı kaynaklardan toplanır. Bu verilerin doğru bir şekilde analiz edilmesi, sağlık hizmetlerinin kalitesini

artırmanın yanı sıra kaynakların etkin kullanımını sağlamaktadır. Veri madenciliği teknikleri, bu büyük ve karmaşık veri setlerinden anlamlı bilgiler çıkarmada önemli bir rol oynar. Sağlık sektöründe veri madenciliği uygulamaları, büyük veri analizi, kişiselleştirilmiş tedavi önerileri, risk gruplarının belirlenmesi gibi birçok alanda kullanılmaktadır.

Bu araştırmanın hedefi, veri madenciliği yöntemlerinden K-Means kümeleme algoritmasını kullanarak sağlık verileri üzerinde hasta profillemesi gerçekleştirmektir. Hasta profillemesi, sağlık hizmetlerini kişiselleştirmede önemli bir araç olup, hastaların sağlık durumlarına göre belirli gruplara ayrılmasını sağlar. Bu çalışmada, hastaların yaş, vücut kitle indeksi (VKI), hipertansiyon durumu, kalp hastalığı ve sigara kullanımı gibi faktörler üzerinden segmentlere ayrılması hedeflenmiştir. Elde edilen sonuçlar, farklı hasta grupları için kişiselleştirilmiş tedavi ve önleme stratejilerinin geliştirilmesi ve kaynakların etkin kullanımına yönelik bilgiler sağlayacaktır.

BİRİNCİ BÖLÜM

1.1. Veri Madenciliği Kavramı

Veri madenciliği, büyük yoğunluktaki veri setlerinden anlaşılabilir desenler, ilişkiler ve gizli bilgileri keşfetme sürecidir. Bu süreç, veriyi analiz etmek, kategorilere ayırmak, gruplandırmak ve öngörüde bulunmak için çeşitli istatistiksel ve yapay zeka tekniklerini içerir. Veri madenciliğinin amacı, verilerdeki karmaşık ve gizli örüntüleri ortaya çıkararak, işletmeler ve diğer alanlar için stratejik kararlar almayı kolaylaştırmaktır.

Veri madenciliği, özellikle veri yoğun sektörlerde (örneğin sağlık, finans, pazarlama) karar destek sistemleri geliştirmede kullanılır. Veriler, belirli bir problem veya karar süreci için hazır hale getirildikten sonra analiz edilir, böylece kullanıcılar bu bilgileri gelecekteki olayları öngörmek veya mevcut sorunları çözmek için kullanır.

Veri madenciliği, büyük ve karmaşık veri kümelerindeki ilişki ve örüntülerin ortaya çıkarılmasını amaçlayan bir bilgi keşfi sürecidir. Bu süreç, yalnızca verilerin depolanması veya saklanması gibi bir işlem olarak düşünülmemeli, aksine verilerden anlamlı çıkarımlar elde etmeyi hedefleyen analitik bir yaklaşım olarak değerlendirilmelidir (Paolo, 2023:2).

Veri madenciliğinde farklı tanımlara yer verilmektedir. Amerikan Pazarlama Birliği (AMA)'ya göre veri madenciliğinin tanımı şöyledir, ‘‘Veri kullanımında yeni verilerin olması ve verilerin önemli ve faydalı bilgilerinin bulunabilmesi için analizinin yapım aşaması. Bu aşamada elde edilmesi gereken zor örüntülerin bulunabilmesi için matematiksel yöntemlerin kullanılabilmesini içerir.’’ Gartner'a göre ise de veri madenciliğin tanımı AMA'dan farklıdır. Gartner'ın tanımında veri madenciliği şu şekildedir, ‘‘Büyük miktardaki verilerin içinden bu verilerin belli anlamlar taşıması için buradan matematiksel analizleri yapılarak elemelerin sonunda buradan anlamlı bağlantılar çıkarılması, örüntüler ve trendler bulma aşamasıdır. Örüntü tanıma teknolojisi kullanan veri madenciliği bunun yanında istatistiksel ve matematiksel yöntemleri de kullanmaktadır.’’ (Akküçük, 2011: 17).

Klasik istatistiksel yöntemler ile veri madenciliğinin arasında aslında çok benzerlik bulunmaktadır. Fakat klasik istatistiksel yöntemler de çalışmanın daha verimli olması için verilerin yeterli anlamda düzenlenmesi gerekmektedir ve genelde özet verilerin üzerinde çalıştırılabilir. Önemli detayların dahi binlerce kaydı olabilir. Veri madenciliği bu klasik istatistiksel yöntemlerden farklı olarak milyarlarca kayıtla ilgilenebilmektedir. Veri miktarı arttığı vakit, bazı özel çözümleme algoritmaların geliştirilmesi gerekmiş ve bundan sonra verinin depolandığı yapıların örneğin veri ambarı şeklinde yeni düzenlemeye gitme gereksinimi doğmuştur (Özkan, 2013: 38).

1.1.1. Veri Madenciliğinin Kullanım Amacı

Kararların isabet oranı, karar vericilerin sahip olduğu bilgi ve deneyim kadar, bilginin güvenilirlik ve kapsamlılık düzeyine de bağlıdır. Bu bağlamda, "bilgi" günümüzde mal ve hizmetlerin yanı sıra üçüncü bir üretim faktörü olarak kabul edilmektedir. Güvenilir bilgilere ulaşmak ise verilerin etkin bir biçimde toplanması, saklanması, analiz edilmesi ve anlamlandırılması sürecini içerir. Karar vericiler, doğru sonuçlar elde etmek amacıyla mümkün olduğunca fazla veri depolamaya ve bu verileri sağlam analiz teknikleri ile desteklemeye önem verirler. Bu yaklaşım, kararların doğruluğunu artırarak hem organizasyonların hem de süreçlerin verimliliğini ve işlevselliğini yükseltmektedir (Argüden ve Erşahin,2008).

Veri madenciliği, yoğun rekabetin yaşandığı pazarlama sektöründe, kârlılığı artırma ve pazar payını büyütme amacı güden işletmeler için kritik bir role sahiptir. Hangi müşterilerin hangi ürünleri ne zaman tercih edeceği, tedarikçilerini değiştirme potansiyeli taşıyan müşteri grupları ve bu müşterilerin elde tutulması veya geri kazanılması için uygulanabilecek stratejiler gibi soruların cevapları, devasa veri yığınları içinde saklıdır. Aynı şekilde, ürünlerin değerini düşüren etkenleri belirlemek gibi stratejik analizler de bu veri yığınlarının derinlemesine incelenmesini gerektirir ve veri madenciliği çözümlerine duyulan ihtiyacı artırır. Veri madenciliği sayesinde şirketler, daha önce farkında olmadıkları bilgileri ortaya çıkararak karar alma süreçlerini daha etkili hale getirebilir. Bu teknikler ile işletmeler maliyetleri düşürebilir, gelirlerini ve operasyonel etkinliklerini artırabilir, yeni iş fırsatları yakalayabilir, yenilikçi ürün ve hizmetler geliştirebilir, emek yoğun süreçleri

otomasyona geçirebilir, dolandırıcılığı saptayabilir ve müşteri deneyimini geliştirebilir. Bu şekilde şirketler, pazarda rekabet avantajını güçlendirerek daha sağlam bir konum elde ederler.

1990'lı yıllardan bu yana, büyük veri kümelerinde saklı, değerli ve anlamlı bilgileri açığa çıkarmak ve stratejik karar destek sistemlerine katkı sağlamak amacıyla benimsenen veri madenciliği, yalnızca çeşitli sorunlara çözümler üretmekle kalmamış; aynı zamanda sağlık verilerinin analizine özgün ve yenilikçi bir bakış açısı kazandırmıştır. Sağlık sektöründe veri madenciliğinin kullanım alanları hızla genişlemiş ve bu yaklaşım, sektörde kritik bir analiz aracı olarak ön plana çıkmıştır. Veri odaklı karar süreçlerini destekleyen veri madenciliği, sağlık verilerinin anlamlandırılması, klinik karar destek sistemlerinin geliştirilmesi ve hasta bakımının iyileştirilmesi gibi birçok alanda önemli katkılar sağlamayı sürdürmektedir.

1.2. Veri Madenciliği Süreci

Veri madenciliği, veri tabanlarında bilgi keşfinin bir parçası olarak geniş kapsamlı bir yaklaşımla ele alınmakta ve modern analiz tekniklerinin ötesinde derinlemesine bir bilgi edinme süreci sunmaktadır. Araştırmaya dayalı ve döngüsel bir yapıya sahip bu süreçte, yalnızca mevcut bilgilerin analiz edilmesiyle yetinilmez; aynı zamanda yeni hipotezlerin üretilmesi ve bu hipotezlerin veri üzerindeki yansımalarının araştırılması hedeflenir. Bu durum, araştırmacının farklı bakış açıları geliştirerek alternatif sorular sormasını ve veriye dair yeni açılımlar getirmesini sağlar.

Veri madenciliğinin en verimli sonuçları verebilmesi için, araştırmacının konu hakkındaki bilgi birikimi ve tecrübesiyle analiz sürecine katılması son derece önemlidir. Araştırmacının deneyimi, analiz sırasında keşfedilen kalıpların anlamlandırılmasına ve doğru soruların sorulmasına olanak tanır. Ayrıca, veri madenciliği sürecinde analiz ve veri ön işleme gibi aşamalar arasında sürekli bir etkileşim vardır; veri analizine yönelik yapılan her keşif, veri ön işlemeyi yeniden gözden geçirmeyi ve veriyi daha etkin hazırlamayı gerektirebilir. Bu döngüsel yapı, veri madenciliği sürecine hem esneklik hem de yenilikçi bir bakış açısı kazandırır, çünkü süreç boyunca keşfedilen her bilgi kırıntısı, süreci bir adım ileriye taşır ve bilgi

keşfi sürekli olarak zenginleşir. Bu açıdan bakıldığında, veri madenciliği yalnızca teknik bir analiz aracı değil, bilgiye ulaşmanın dinamik bir yoludur.

Veri madenciliği, iş ihtiyacının belirlenmesinden geliştirilmiş modelin pratik uygulamaya geçirilmesine kadar uzanan kapsamlı bir süreci ifade eder. Bu sürecin önemli bir diğer yönü ise yinelemeli yapısının bulunmasıdır. Veri madenciliği uygulamalarında çoğunlukla birden fazla model oluşturulmakta ve en verimli, en doğru modeli elde etmek hedeflenmektedir. Ancak, en uygun model elde edildiğinde dahi bu model yeterli bulunmazsa süreç, veri kümesine yeni kayıtların eklenmesi veya mevcut kayıtlardaki eksikliklerin ve sorunların giderilmesi yoluyla tekrar edilir. Bu adımlar, modelin sürekli iyileştirilmesi ve uyarlanabilir bir sonuç elde edilmesi için kritik önem taşır (Keskin, 2013: 13).

Veri madenciliği, bilgi keşfi ve veri tabanlarında bilgi keşfi terimleri, bazı araştırmacılar tarafından sıklıkla birbirine karıştırılmaktadır. Çoğu araştırmacı ve uygulayıcı, veri madenciliği ve bilgi keşfi kavramlarını eş anlamlı olarak kullanma eğilimindedir. Ancak, veri madenciliği aslında bilgi keşfi sürecinin yalnızca bir aşamasını temsil etmektedir. Veri tabanlarında bilgi keşfi, genel olarak, verilerden anlamlı, faydalı, yenilikçi ve belirli bir değere sahip örüntülerin sistematik bir şekilde ortaya çıkarılması süreci olarak tanımlanabilir (Cios vd., 2007: 10).

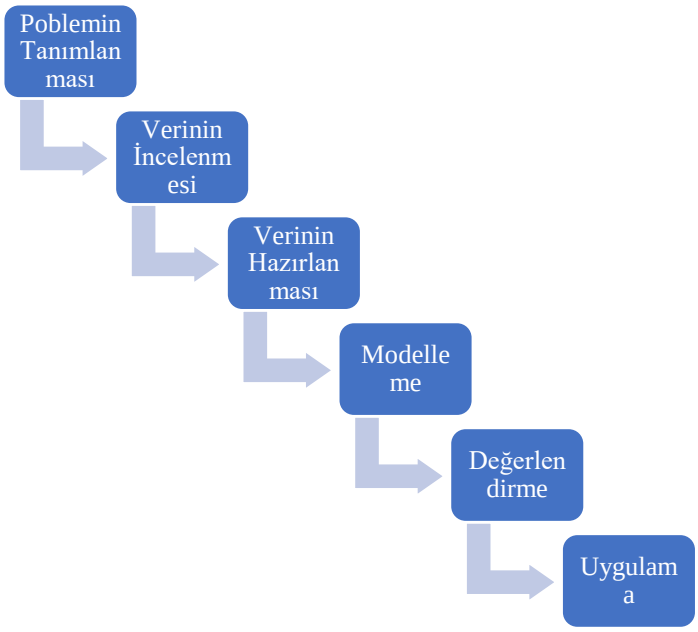
Veri madenciliği sürecine dair daha ayrıntılı bir tanımlama, uluslararası alanda kabul gören CRISP-DM (CRoss Industry Standard Process for Data Mining) modeli ile yapılmaktadır. CRISP-DM, veri madenciliği projelerinin daha hızlı, verimli ve düşük maliyetle gerçekleştirilmesi amacıyla geliştirilmiş bir standarttır ve süreci altı ana aşamada yapılandırmaktadır. Bu model, hem endüstriyel hem de akademik projelerde geniş bir uygulama alanı bularak veri madenciliğinde sistematik bir yol haritası sunar (Argüden ve Erşahin, 2008: 20).

Veri madenciliği, kısaca verilerden gizli bilgilerin ortaya çıkarılması süreci olarak tanımlanabilir. Bu sürecin etkin bir şekilde yürütülebilmesi için her bir aşamanın titizlikle takip edilmesi esastır; çünkü her aşama, bir sonraki aşamaya girdi sağlayarak

sürecin ilerlemesine katkıda bulunur. Başka bir deyişle, her aşama bir öncekine bağımlıdır ve bu bağlantı, sürecin bütüncül bir yapı olarak değerlendirilmesini sağlar. Veri madenciliği sürecinde en çok çaba ve zaman gerektiren kısım, verilerin hazırlanması aşamasıdır. Bu aşamayı takiben karar probleminin tanımlanması, veri analizinin gerçekleştirilmesi ve elde edilen sonuçların yorumlanması gibi adımlar gelir (Gürsoy, 2009: 30).

Veri madenciliği, belirli adımları takip eden sistematik bir süreç olarak tanımlanır. Bu süreç, veriden anlamlı bilgi ve desenlerin keşfedilmesi için yapılandırılmış aşamalardan oluşur ve her aşama, sürecin başarıyla tamamlanabilmesi için kritik bir rol oynar. Veri madenciliği süreci Şekil 1.2.'de görülmektedir. Belirtilen bu adımlar maddeler halinde yazılmak istenirse (Özkan, 2008);

1. Problemin Tanımlanması
2. Verilerin İncelenmesi
3. Verilerin Hazırlanması
4. Modellemenin Yapılması
5. Değerlendirme ve Analiz
6. Uygulama, olarak yazılabilir.



Şekil 1.2. Veri madenciliği süreci

1.2.1. Problemin Tanımlanması

Problemin tanımlanması aşamasında, veri madenciliği uygulayıcısının ilk olarak işletmenin ulaşmayı amaçladığı hedefleri açık bir şekilde belirlemesi gerekmektedir. Analizi gerçekleştiren uzman, veri madenciliği sürecinin sonuçlarını doğrudan etkileyebilecek temel faktörleri belirlemeye odaklanır. Veri madenciliği projesinin başarısı, projeye yönelik titiz bir planlamaya, ayrıca spesifik, gerçekleştirilebilir ve ölçülebilir hedeflerin belirlenmesine bağlıdır. Bu hedeflerin açık ve ulaşılabilir olması, projenin yönünü ve etkili sonuçların elde edilmesini sağlar (Gürsoy, 2009: 32).

Büyük veri yığınlarından farklı türde bilgilerin keşfedilmesi mümkündür ve bu keşiflerin kullanıcıya anlaşılır bir biçimde sunulması, veri madenciliği süreçlerinin geniş bir kullanıcı kitlesi tarafından etkin şekilde kullanılabilmesi için kritik öneme sahiptir. Bu bağlamda, veri madenciliği uygulamalarının kullanıcı dostu grafik arayüzler veya yüksek seviyeli programlama dilleriyle desteklenmesi, uzman olmayan kullanıcıların da sürece katılımını kolaylaştırmaktadır. Aynı zamanda, keşif sistemlerinin anlamlı ve anlaşılır bilgi sunum tekniklerine sahip olması, süreçlerin etkinliğini artırmaktadır. Veri tabanında keşfedilebilecek bilgilere ilişkin belirsizlikler nedeniyle, yüksek seviyeli veri madenciliği sorguları, ilginç ve potansiyel olarak değerli izlerin tespit edilmesi için güçlü bir araç olarak kullanılabilir. Bu yöntem, daha derinlemesine analizler için rehberlik sağlayabilir. Ayrıca, kullanıcıların analiz süreçlerini interaktif bir şekilde yönetebilmesi, odak noktalarını esnek biçimde değiştirebilmesi ve sonuçları detaylı olarak inceleyebilmesi, veri madenciliği süreçlerinin etkinliğini artıran önemli bir unsurdur.

1.2.2. Verinin İncelenmesi

İkinci aşama, ilk verilerin toplanması ve mevcut verilerin model oluşturma amacıyla uygunluğunun değerlendirilmesini içermektedir. Bu adımda, veri kalite ve yeterliliği üzerine detaylı bir değerlendirme yapılır; bu değerlendirme, model oluşturma sürecinde ihtiyaç duyulan veri türlerinin belirlenmesi ve mevcut kayıt sayısının yeterliliğinin analiz edilmesini kapsar. Bu düşünce süreci, verinin amaca uygun, güvenilir ve kapsamlı olmasını sağlamak için temel bir ön hazırlık niteliği taşır (Argüden ve Erşahin, 2008: 21).

İşletme hedeflerinin belirlenmesi ve proje planının oluşturulmasının ardından, veri inceleme aşamasına geçilir. Bu aşama, proje için gerekli veri ihtiyaçlarının belirlenmesi, verilerin toplanması, verilerin tanımlanması ve veri kalitesinin değerlendirilmesi gibi temel adımları kapsar. Bu süreç sonunda, verinin genel özelliklerini anlamak için özet istatistikler oluşturulabilir ve veri kümesindeki örüntülerin belirlenmesi amacıyla kümeleme analizi gibi yöntemler uygulanabilir. Veri seçimi, çoğunlukla farklı kaynaklardan ve çeşitlilik gösteren veri türlerinden yapılır; işletme uygulamaları için ise demografik, sosyografik ve işlemsel veriler sıkça tercih edilmektedir. Veri türleri genel olarak nicel ve nitel olarak iki kategoriye ayrılır. Nicel veriler, kesikli veya sürekli olarak sınıflandırılabilirken, nitel veriler nominal veya ordinal olabilir. Nominal veriler, kategorik grupları sıralama olmadan temsil eder; örneğin, cinsiyet verisi "kadın" ve "erkek" gibi değerlerle ifade edilir. Ordinal veriler ise belirli bir sıralamaya sahip kategorik verilerdir; örneğin, müşteri kredi kartı derecelendirmesi için iyi, vasat ve kötü gibi sıralı değerler kullanılır. Nicel veriler, olasılık dağılımları aracılığıyla çeşitli istatistiksel analizlere tabi tutulabilirken, nitel veriler frekans dağılımları kullanılarak, sayısal kodlama yoluyla incelenir ve yorumlanır. Bu veri ayrımı, analiz sürecinde doğru metodolojinin belirlenmesi açısından oldukça önemlidir (Irmak, 2009: 18).

1.2.3. Verinin Hazırlanması

Modelin oluşturulması için gerekli bilgilerin hazırlandığı bu aşamada, veriler üzerinde çeşitli istatistiksel ve görsel analizler yapılır. Dağılım ölçüleri olarak toplam, maksimum ve minimum değerler incelenirken, aritmetik ve ağırlıklı ortalama gibi cebirsel ölçüler kullanılır; ayrıca serpilme ve dağılım diyagramları gibi grafiksel araçlar ile veri yapısı hakkında genel bilgi edinilir. Bu analizler sayesinde, veri setinde eksik, hatalı veya gürültülü bilgiler olup olmadığı belirlenebilir. Eksik değerler için, eksik kaydı göz ardı etme, eksik değerleri global bir sabit ile doldurma veya o değişkenin ortalama değeri ile eksik değeri tamamlama gibi stratejiler uygulanabilir. Gürültülü verilere müdahale etmek için ise regresyon gibi teknikler kullanılarak veriyi belirli bir fonksiyonel kalıba oturtma yöntemi tercih edilir. Bu işlemler, modelleme

sürecinde verinin tutarlılığını sağlamak amacıyla önemli bir adımı temsil eder (Çelik, 2009: 9).

Bu aşama, başlangıç verilerinin analiz ve modelleme çalışmalarına sağlam bir temel oluşturacak şekilde nihai verilere dönüştürülmesini içerir. Bu süreçte, veri temizleme, dönüştürme ve yapılandırma adımları uygulanarak ham verilerden hedeflenen analitik amaçlara uygun, işlenmiş bir veri kümesi elde edilir (Argüden ve Erşahin, 2008: 22)

1. **Veri Seti Tanımlama:** Bu aşamada, modelin geliştirileceği problem doğrultusunda gerekli veri seti ve bu verilerin elde edileceği kaynaklar belirlenir. Hedef soruya yönelik en uygun verilerin seçimiyle sürecin temeli oluşturulur.
2. **Veriyi Seçme:** Analiz için kullanılacak verilerin seçilmesi sürecidir. Bu aşamada verinin projenin hedefleriyle olan ilişkisi, veri kalitesi ve teknik sınırlamalar göz önünde bulundurularak veri seçimi yapılır.
3. **Veriyi Temizleme:** Gürültülü veya tutarsız verileri ayıklayarak veri kalitesini artırma adımıdır. Bu aşamada, yanlış girilen veya istisnai olan veriler süreçten çıkarılarak daha güvenilir bir veri seti oluşturulması hedeflenir.
4. **Veriyi Kurma:** Modelleme sürecinde kullanılabilir değişken setleri oluşturmak amacıyla, mevcut veri değişkenlerinde gerekli modifikasyonlar yapılır. Bu sayede, modelin performansını artıracak şekilde veri yapılandırılır.
5. **Veri Birleştirme:** Veri madenciliği için farklı kaynaklardan toplanan verilerdeki uyumsuzluklar giderilir. Bu adımda, çeşitli kaynaklardan gelen veriler olabildiğince uyumlu hale getirilerek tek bir veri tabanında toplanır.
6. **Veri Formatlama:** Son aşamada, veri seti oluşturulduktan sonra seçilen modele uygun biçimde format düzenlemeleri yapılır. Anlam kaybı olmaksızın gerçekleştirilen bu formatlama, verinin modele uyum sağlamasını hedefler.

1.2.4. Modelin Kurulması

Modelleme aşaması, veri madenciliği yazılımının yardımıyla uygun analiz tekniklerinin uygulanarak farklı senaryolar için sonuçların elde edilmesini sağlar. Bu sürecin başlangıcında, veriyi anlamak ve içindeki örüntüleri keşfetmek için genellikle kümeleme analizi ve veri görselleştirme yöntemleri kullanılır. Verinin özelliklerine ve hedefe bağlı olarak, sonraki adımlarda daha çeşitli modelleme tekniklerine başvurulur. Eğer modelin amacı veriyi önceden tanımlanmış kategorilere ayırmak ise, diskriminant analizi gibi yöntemler tercih edilebilir. Öngörüleme hedefleniyorsa ve değişkenler sürekli bir yapı gösteriyorsa regresyon analizi uygun bir seçenektir; sürekli olmayan değişkenler içinse lojistik regresyon öne çıkar. Tahmin ve sınıflandırma amaçlarına yönelik olarak yapay sinir ağları oldukça etkili ve yaygın kullanılan bir modelleme tekniğidir. Ayrıca, veriyi sınıflandırma işlemlerinde karar ağaçları da etkin bir alternatif sunarak, veri içindeki karar mekanizmalarını daha anlaşılır kılmada yardımcı olabilir (Irmak, 2009: 20).

Benzer veri madenciliği problemlerinde birden fazla çözüm tekniği uygulanabilir ve bazı teknikler, veri üzerinde belirli özellikler veya özel koşulların sağlanmasını gerektirir. Bundan dolayı, veri hazırlama ve modelleme aşamaları, en uygun modele ulaşana kadar döngüsel bir biçimde tekrarlanır. Bu süreçte yapılan her deneme, verinin yapısı ve modelin performansına göre analiz edilerek gerekli düzenlemeler yapılır ve süreç yeniden başlatılır. Bu döngüsel yaklaşım, modelin esneklik kazanmasını ve hedeflenen en doğru çözüme ulaşmak için sürekli uyarlanmasını sağlar. (Argüden ve Erşahin, 2008: 23);

1. **Model Tekniğini Seçmek:** Bu aşama, veri madenciliği sürecinde kullanılacak fonksiyon ve algoritmanın amaca en uygun şekilde belirlenmesini içerir. Modelin gereksinimlerine ve veri yapısına göre doğru teknik seçilerek analiz süreci başlatılır.
2. **Model Test Tasarımı Yapmak:** Modeli çalıştırarak sonuçları elde etmeye başlamadan önce, modelin kalitesini ve doğruluğunu test etmek önemlidir. Bu aşamada, modelin performansını değerlendirmek için uygun test yöntemleri geliştirilir. Örneğin, sınıflandırma gibi öngörü amaçlı fonksiyonlarda, hata

oranları modelin kalite göstergesi olarak kullanılabilir. Bu test süreci, modelin güvenilirliğini ve amaca uygunluğunu ölçmek için kritik bir adımdır.

3. **Modeli Kurmak:** Bu aşama, seçilen algoritma, yöntem veya tekniğin hazırlanmış veri üzerinde çalıştırılmasını içerir. Kurulan model, doğruluğu ve geçerliliği onaylandıktan sonra, bağımsız bir uygulama olarak kullanılabileceği gibi, başka bir uygulamanın bir alt bileşeni olarak da işlev görebilir. Modelin başarılı bir şekilde kurulması, veri madenciliği sürecinin temel çıktılarından biridir ve uygulama süreçlerine doğrudan katkı sağlar.
4. **Modeli Değerlendirmek:** Bu aşama, modelin başarı kriterlerine, önceki deneyimlere ve test sonuçlarına göre teknik olarak değerlendirilmesini içerir. Burada amaç, tüm projeyi değil, yalnızca modelin teknik performansını ve amaca uygunluğunu analiz etmektir. Değerlendirme süreci, modelin belirlenen hedeflere ne kadar ulaştığını ve beklenen performansı sağlayıp sağlamadığını belirlemek için kritik öneme sahiptir.

1.2.5. Değerlendirme

Model sonuçlarının değerlendirilmesi, başlangıç aşamasında belirlenen iş hedefleri doğrultusunda yapılmalıdır. Veri madenciliği, analistin iş hedeflerine ulaşma sürecinde daha derin bir anlayış kazanmasını sağlayan iteratif bir süreçtir. Bu aşamada kullanılan çeşitli görselleştirme, istatistik ve yapay zeka araçları, verideki yeni ilişkileri ortaya çıkararak işletme faaliyetlerine dair daha kapsamlı bir içgörü sağlar. Yorumlama ve değerlendirme aşaması, veri madenciliği sürecinin temel bir bileşenidir; elde edilen sonuçların anlamlandırılması ve iş değeri olarak özümsenmesi bu noktada gerçekleşir. Bu aşamada dikkat edilmesi gereken iki önemli konu vardır. Birincisi, veri madenciliği sürecinde elde edilen bilgilerin işletme için nasıl bir değer ifade edeceğinin tanımlanmasıdır. İkincisi ise, sonuçların etkili bir şekilde sunulabilmesi için hangi görselleştirme araçlarının ve tekniklerinin kullanılacağına karar verilmesidir. Bu yaklaşımlar, sonuçların iş birimleri tarafından anlaşılır ve kullanılabilir hale getirilmesini sağlar (Irmak, 2009: 22).

Bu aşamada, kurulmuş olan modelin nihai sunumdan önce kapsamlı bir değerlendirmeye tabi tutulması hedeflenir. Bu süreçte, modelin iş hedefleri ile uyumlu

olup olmadığı titizlikle incelenir ve modelin performansının hedeflere ulaşma kapasitesini sağladığına dair güvence sağlanır. Yoğun değerlendirme aşaması, modelin işletme amaçlarına uygunluğunu doğrulamak ve gerekli görüldüğü durumlarda iyileştirmeler yaparak modeli daha güçlü hale getirmek için kritik bir adımdır (Argüden ve Erşahin, 2008: 24-25);

1. **Sonuçları Değerlendirmek:** Ön değerlendirme aşamaları, modelin geçerliliği ve uygunluğu hakkında bilgi sağlarken, bu aşamada modelin iş hedeflerini ne derece karşıladığına odaklanılır. Bu değerlendirme, modelin işletme amaçlarına katkı sağlama potansiyelini ve hedeflenen sonuçlara ne kadar yaklaştığını belirlemek için yapılır.
2. **Süreci Değerlendirmek:** Bu aşama, kalite güvence sürecini içerir. Modelin iş hedeflerini karşılama açısından yeterli olduğu doğrulandıktan sonra, modelin doğru bir şekilde kurulup kurulmadığı, yalnızca mevcut verilere mi dayandığı veya gelecekte eklenebilecek farklı veri kaynaklarının neler olabileceği gibi unsurlar değerlendirilir. Bu süreç, modelin uzun vadede sürdürülebilirliği ve esnekliği açısından önemli bir analiz sağlar.
3. **Gelecek Adımları Planlamak:** Bu aşamada, projenin mevcut durumu ve hedeflere ulaşıp ulaşmadığı değerlendirilir; ek çalışmaların gerekip gerekmediği analiz edilir. Devam edilmesi gereken adımlar nelerdir, bu adımlar için bütçe imkanları var mı ve projeye devam etmek mantıklı ise hangi noktadan ilerlenmelidir gibi stratejik konular üzerinde düşünülür. Bu planlama süreci, projenin verimliliğini ve sürdürülebilirliğini sağlamak için kritik kararların alındığı aşamadır.

1.2.6. Uygulama

Veri madenciliği çalışmaları, proje hedefleri ile uyumlu yeni bilgiler ortaya çıkarır ve bu bilgiler, yönetimin iş ortamına dair yeni bir anlayış geliştirmesi ve buna yönelik stratejik kararlar alabilmesi için kritik bir temel sağlar. Veri madenciliği sonucunda elde edilen bilgilerin değişim sürecinde izlenmesi de oldukça önemlidir; çünkü veri toplama sırasında geçerli olan bir durum, süreç ilerledikçe değişime uğrayabilir ve elde edilen bilgiler geçerliliğini yitirebilir. Bu nedenle, odak alanının ve temel göstergelerin

uygulama süreci boyunca sürekli olarak izlenmesi, bilgilerin güncelliğini ve doğruluğunu korumak için gereklidir (Irmak, 2009: 23).

Modelin tamamlanması, projenin nihai sonucu olarak değerlendirilmemelidir. Modelin asıl amacı, veriler hakkında daha derin bir anlayış sunmak olsa bile, elde edilen bilgilerin kullanılabilir bir şekilde organize edilmesi ve sunulması önemlidir. Bu sunum genellikle, gerçek verilerden seçilen örneklerin modelin sonuçlarını temsil edecek şekilde görselleştirilmesi veya raporlanması biçiminde yapılır. Bu adım, modelin çıktılarının anlaşılır ve karar vericiler için eyleme geçirilebilir hale getirilmesini sağlar. (Argüden ve Erşahin, 2008: 25);

1. **Yayma Planını Oluşturmak**: Bu aşama, model sonuçlarının değerlendirilmesinin ardından, sonuçların ilgili paydaşlara nasıl iletileceğine dair bir yayma stratejisi oluşturmayı içerir. Strateji, elde edilen bilgilerin doğru kişilere, uygun formatlarda ve zamanlamayla ulaştırılmasını sağlamak için planlanır. Bu plan, karar vericilerin sonuçları etkili bir şekilde kullanabilmeleri için kritik bir rol oynar.
2. **Bakımını ve Takibini Yapmak**: Sistem özelliklerinde ve üretilen verilerde zamanla ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerektiğinde yeniden yapılandırılmasını gerekli hale getirir. Bu adımda, modelin doğruluğunu ve güncelliğini korumak amacıyla düzenli bakım stratejileri planlanır. Böylece, uzun süre yanlış veya güncelliğini yitirmiş verilerle çalışmanın önüne geçilerek modelin işletmeye sağladığı değer sürdürülebilir hale getirilir.
3. **Final Raporlarını Hazırlamak**: Bu aşama, yapılan çalışmanın başkaları tarafından tekrarlanabilmesini sağlamak ve sonuçları karar vericilere net bir şekilde sunmak amacıyla hazırlanan nihai raporu içerir. Final raporu, çalışmanın üçüncü taraflarca denetlenebilmesi ve doğruluğunun güvence altına alınması açısından kritik öneme sahiptir. Raporda, izlenen adımlar, kullanılan yöntemler, elde edilen sonuçlar ve çıkarımlar detaylandırılarak projenin şeffaflığı ve güvenilirliği sağlanır.

4. **Projeyi Değerlendirmek**: Bu aşama, proje kapsamında yapılan çalışmalara dayalı olarak alınan kararların ve elde edilen sonuçların belirli bir zaman dilimi sonunda başlangıçtaki beklentilerle karşılaştırılmasını içerir. Bu değerlendirme, projenin uzun vadeli etkinliğini analiz ederek, sonuçların hedeflerle uyumlu olup olmadığını belirler. Gerektiğinde, projede yenileme veya iyileştirme çalışmaları yapılarak elde edilen bilgilerin güncelliği ve doğruluğu sağlanır. Bu süreç, projenin sürdürülebilir başarısını garanti altına almak açısından önemlidir.

1.3. Veri Madenciliği ile İlgili Diğer Disiplinler

Hastanelerin tıbbi cihazlara gereksinimleri vardır. Ancak alınacak olan tıbbi cihazların hastane yapım aşamasında ya da bu cihazların yerleştirilmesi düşünüldüğü aşamada bu cihazlar için uygun koşulların tesis edilmesi ve en üst verimlilikte işlevselliğinin korunması gerekmektedir. Hastane için alınan tıbbi cihazların yerleşimlerinin önceden planlarının yapılamamasında, bulunduğu koşullarda değişiklik yapmak hayli güç ve büyük masraflar gerektirmektedir. Bu konuda en büyük faktör alt yapıların çalışmasının yeterli manada incelenememesi ve hastane koşullarına uygun tıbbi cihazların tespit edilmemesindendir. Bu aşamada hastaneye alınacak tıbbi cihazların önceden araştırılması için ölçülü fizibilite çalışmaları yapmak ortaya çıkabilecek sorunları büyük ölçüde azaltacaktır. Çalışmaların bu büyüklükte oluşmasında disiplinler arası iş birliğine gidilmesi gerekliliğini ortaya çıkarır (Soylular, 2006: 37).

Şekil 1.3. Veri Madenciliği ve Diğer Disiplinler

Veri madenciliği, makine öğrenmesi, örüntü tanıma, veri tabanı teknolojileri, istatistik, yapay zeka uzman sistemler ve veri görselleştirme gibi disiplinlerin kesişim noktasında doğmuş ve bu çok yönlü yapı üzerine gelişimini sürdürmektedir. Bu alanlar arasındaki etkileşim, veri madenciliğini güçlü bir bilgi keşfi aracı haline getirir. Bu yapının temsili, Şekil 1.3'te görüldüğü üzere, veri madenciliğinin çok disiplinli doğasını ve bu alanlardan beslenerek ortaya çıkan kapsamlı yeteneklerini sembolize eder.

Disiplinler arası bir yapıya sahip olan veri madenciliği, özellikle veri tabanı sistemleri, istatistik, matematik, makine öğrenmesi, görselleme ve bilişim bilimleri gibi alanlarda geniş bir uygulama alanına sahiptir. Veri madenciliğinin diğer istatistiksel yöntemlerden farkı, verinin tümünü analiz sürecine dahil etmesidir. Bu özellik, geleneksel yöntemlerle analiz edilen küçük ve sınırlı veri kümeleri yerine daha büyük, bağımsız veri kümelerinin kullanılmasına olanak tanır ve böylece daha kapsamlı ve ayrıntılı değerlendirmeler yapılmasını sağlar. Bu durum, veri madenciliğini hem derinlemesine analizler hem de keşif temelli bilgi çıkarımı için son derece etkili bir araç haline getirir (Alan, 2012: 166).

Makine öğrenmesi, örüntü tanıma ve istatistik disiplinleri, veri madenciliğinde örüntü keşfetme sürecine katkı sağlarken; yapay zekâ teknolojileri, keşfedilen örüntülerin yorumlanması aşamasında rol oynamaktadır. Veri tabanı teknolojileri ise, mevcut verilerin depolanması, filtrelenmesi, temizlenmesi ve sorgulanması işlemlerinde kullanılırken; veri görselleştirme, sonuçların raporlanması ve insan zihni tarafından anlamlandırılabilir sembollere dönüştürülmesi sürecinde önemli bir destek sağlamaktadır (Aslan, 2008: 64).

Veri Tabanı Sistemleri: Veri tabanı sistemlerinin özellikleri bilgisayar ortamlarında tutulan büyük veri kaynaklarında veya kümelerinde daha çok kullanılan bir araç oluşudur. Veri tabanı ‘‘bir veya birçok uygulamada kullanılmak üzere tekrar edilen gereksiz verilerin bir düzen içinde bilgisayarların belleklerinde depolanan birbiriyle bağlantılı veri topluluğu’’ olaraktan tanımlaması mümkündür. Veri madenciliği programlarında kullanılmak üzere onlarca kuruluşa ait veri tabanlarında

kullanılması sağlanmıştır. Kuruluşların bu uygulamaları kullanarak geleneksel istatistik yöntemlerine gereksinimleri azaldığı gibi aynı zamanda gerekli fizibilite çalışmalarının yapılmasının da kolaylığından faydalanmaktadır. Bu çalışmalar için kullanılan veri madenciliği sistemleri kuruluşları gereksiz bilgi birikimlerinden uzak tuttuğu gibi gereksiz düzenlemeler ve maddi yüklerini de bir anlamda düşürmüştür.

İstatistik: Verinin yapısının keşfi için istatistik ve veri madenciliği araştırmalarda verilerin yapısını keşfetmeyi hedefleyen iki disiplindir. Veri madenciliği ile istatistik arsında konuların örtüşmesinden dolayı veri madenciliği istatistiğin bir alt dalı gibi düşünülmesi doğaldır. Ancak veri madenciliğini istatistikten ayıran bir konu vardır ki buda makine öğrenimi gibi birçok alanda ilişkisinin oluşudur. Aynı zamanda istatistik veri madenciliği gibi çok disiplinler arası değildir.

Makine Öğrenimi: Aslında bilgisayarları yapay zekaya çeviren veya başka bir deyimle bilgisayarların olayları öğrenmesine yardımcı bir teknolojidir. Çoğunlukla olayları örnekler kullanarak sisteme girdi çıktılar yaparak aralarındaki ilişkiler öğrenilir. Sisteme yapılan girdi çıktılar neticesinde problemler çözülebilir veya birbirine benzeyen olaylar üzerinden güçlü tahminler ya da yorumlamalar yapılar çalışır. Makine öğrenimini bir bütün olarak ele alacak olursak ‘‘bilgisayar teknolojilerinin yaşanmış olaylardan ve becerilerden örneklemeler yaparak tecrübeler elde etmesi ve ileri süreçlerde yaşanması mümkün olaylardan kazandığı tecrübe ve kazanımların bilgisayar teknolojilerinin karar verebilmesini ve karşılaşabileceği problemlere çözümler sunabilmesidir’’ Şeklinde tanımlama yapılabilir.

Veri Görselleştirme: Veri madenciliğinde, verilerin girdi ve çıktılarının anlaşılır ve kullanıcı dostu olması büyük önem taşır. Bu bağlamda, görselleştirme teknikleri veri dağılımlarını, kümeleri ve uç değerlerin etkisini daha çekici ve etkili bir şekilde sunmak için kullanılır. Bu amaç doğrultusunda bilgisayar grafiklerinde, veri dağılım haritaları, eğriler, üç boyutlu şekiller ve yüzeyler gibi çeşitli görselleştirme araçlarından yararlanılır. Bu teknikler, karmaşık veri yapılarının görsel olarak sadeleştirilmesini sağlayarak veriyi daha anlaşılır hale getirir ve analistlere

derinlemesine bir analiz yapma imkanı sunar. Bu sayede veri madenciliği sürecinde, verinin bütüncül bir şekilde yorumlanması kolaylaşır.

Diğer Disiplinler: Veri madenciliği veri görselleştirme, makine öğrenimi, istatistik, veri tabanı sistemlerinden oluşan disiplinlerin yanında veri ambarı, uzman sistemler, yapay sinir ağları ve genetik algoritmalarla da ilişki içindedir.

1.4. Veri Ambarları

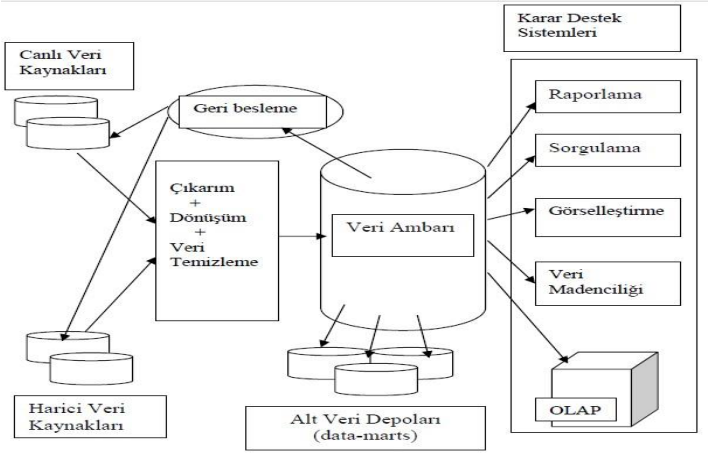
Uluslararası ve modern kuruluşlar pazarda ki gelişmelerden ve değişimlerden etkilenmektedirler. Bu değişimlere hızlı cevap verebilmek adına büyük baskı altındadırlar. Doğru ve etkileyici kararlar alabilmek için kuruluşların her türlü bilgilere hızlı ulaşması gerekmektedir. Yapılan araştırmalarda kuruluşların her geçen gün inanılmaz boyutlarda verilerinden çoğaldığını göstermektedir. Günümüzde teknolojinin gelişmesiyle yeni yöntemler aracılığıyla veriler ne kadar büyük olursa olsun kısa sürelerde toplanabilmekte ve analizleri yapılabilmektedir. Verilerin miktarları çoğaldıkça verilerden anlamlı bilgilerin ortaya çıkması daha zor ve daha karmaşık bir hal almıştır. Bundan dolayı veri ambarcılığı bu karmaşıklıkların giderilebilmesi adına yeni çözüm yollarından bir tanesidir. Veri ambarı bu karmaşıklıkları giderme adına karar veri destek sistemlerinde kullanılabilecek verileri saklayabilen bir yöntemdir. Veri ambarları son kullanıcıya sorgu yapabilme, raporlar oluşturabilme, analizler yapabilme ortamları sağlamıştır (Gürsoy, 2009: 3).

Veri ambarları, organizasyonların büyük ve karmaşık veri setlerini etkin bir şekilde depolamak, yönetmek ve ihtiyaç duyulduğunda hızlı erişim sağlamak amacıyla tasarlanmış özel bilişim sistemleridir. Bu sistemlerin temel işlevi, farklı kaynaklardan gelen verileri entegre ederek iş zekası ve analitik uygulamalarına yönelik anlamlı bilgiler sunmaktır. Günümüzde işletmelerin veri ambarlarında sakladığı verilerin hacmi ve çeşitliliği hızla artmakta, bu da analitik uygulamalar için verilerin hazırlanması ve hızlı erişim sağlanmasında yeni zorlukları beraberinde getirmektedir (Rao ve ark, 2019).

Veri ambarları bütün operasyonel işlemleri en basit ve alt düzeylerdeki verileri ne kadar inebilir, etkili araştırmalar çıkartabilmesi için özel modelleri ve tarihsel derinliği bulunan veri depolama sistematiği olarak tanımlanabilmektedir (Yaralıoğlu, 2004: 165). Başka bir ifade ile veri ambarı ilişkili verilerin sorgulandığı, analizlerin yapıldığı birleşik bir depodur (Şentürk, 2006: 8).

Veri ambarları coğrafi bilişimden sağlık sektörüne, fabrikadaki üretimden işletmelerde pazarlama bölümüne kadar, yıllar sonrasına tahmin yapmada sonuçlar çıkarmada ve sektörlerin yönetim stratejilerinin belirlemede kullandığı bir sistemdir. Veri ambarları sistemlerinin pahalı maliyetleri olsa bile sonuç olarak kazancının maliyet hesabı yapıldığında sistemin getirisinin sektörlere büyük kazançlar ortaya çıkarmıştır. Son yıllarda daha yaygın kullanılmaya başlanılan veri ambarı sistemi günlük kullanılan veri tabanlarının birleştirilerek işlemeye daha uygun bir özetini saklamayı amaçlar (Usgurlu, 2007).

Karar destek sistemi olan veri ambarları canlı sistem veya canlı sistemler üzerinden ve dahası farklı farklı kaynaklardan belli bir periyotta beslenen ve bunların analizlerini ortaya çıkartan veri kümeleridir. Verilerin aktarılma periyotları işletmelerin ihtiyaçları doğrultusunda farklılıklar göstermektedir. Veri aktarımı esnasında veri ambarı verilerin tekrar eden verilerden olmamasına aktarmalar üzerinden güncelleme ve silme gibi işlemlerin uygulanmasında verilerin güvenilirlikleri bir anlamsal bütünlükleri için gereklidir (Gazi, 2007: 6).



Şekil 1.4. Veri Ambarı Mimarisi

Şekil 1.4. te görüldüğü gibi, veri ambarı tek bir canlı sistemi kaynak olarak kullanmaz; bunun yerine, çeşitli kaynaklardan beslenir ve bu kaynaklarla arasında bir "geri besleme" mekanizması oluşturur. Bu sayede, kaynak sistemler veri ambarından gelen taleplere yanıt verebilme kapasitesine sahip olur. Geri besleme mekanizması, veri ambarının kaynak sistemlerle etkileşimli bir yapıda çalışmasını sağlar ve böylelikle sürekli güncel ve talep odaklı verilerin sağlanmasına olanak tanır (Gazi, 2007: 7).

Veri ambarı, herhangi bir kuruluşun ya da işletmenin farklı farklı birimlerinden elde edilen verilerin arasından analize uygun olan değerli veri kaynaklarının daha sonra analizlerde kullanılmak maksadıyla sistem veri tabanından başka ortamlarda birleştirmesinden ortaya çıkan büyük çaplı veri deposudur. İşletim sistemleri de depolanmış olan verilerin içlerinden alınması gereken önemli verilerin ayıklanarak temizlenmesi ve karar verme sistemlerinde hizmet edecek düzeyde hazırlanması, istenildiği üzere saklanabilmesi, farklı yazılımlar yardımı ile veriyi erişebilmesi belirleyicileri ilişkilerini arayıp bulması, işlemlerin tamamını ihtiva eden aktiviteler zinciridir. Veri ambarı kullanıldığı zaman günlük rutin işletimsel görevlerle meşgul olmak yerine veri tabanına gerek duyulmadan yani veri tabanı hiç kullanılmadan, analiz işlemleri başka ortamda kolay bir şekilde hızlı ve doğru biçimde yapılır. (Alan, 2012: 168).

Herhangi bir bilim dalı üzerinde çalışabilmek için verilerini iyi tanıyabilmek için gerekmektedir. Tıbbi veriler üzerinde çalışmak için aynı şekilde bu verilerin iyi tanınması ile mümkündür. Tıbbi verilerde uzmanların yorumlaması verilerin daha çok işlevselliği adına önemlidir. Tıp biliminde belli bir standart yoktur. Bundan dolayı veri ambarı oluşturulması tıp bilimindeki standartların arasında uyumsuzluk da söz konusu olduğu için uzmanların desteğiyle ancak veri ambarı zor da olsa oluşturulabilmektedir. Farklı standartlar arası kodlama yöntemleri bulunmaktadır. Tıp bilimindeki bazı terimlerin geneli itibariyle karışık ve birbirine benzemesi sebebiyle veri ambarı yapım aşamasını zorlaştırmak da aynı zamanda yukarıda da belirtildiği üzere uzman yardımına ihtiyaç duyulmakta ve daha sonra veri ambarı oluşturulmaktadır. Tıp bilimindeki veri, çoğunlukla başka kaynaklar arasında toplanmaya çalışılmaktadır.

Örneğin hastaneler de aynı kişiye ait olsa bile laboratuvar verileri ve teşhis bilgileri birbirlerinden farklı kaynaklar olsa da farklı veri ambarlarında farklı şekil ve düzenlerde tutulmaktadır (Turgut, 2012: 42).

Veri ambarlarında en baştaki özellik işletmelerin başlıca amaçlarına ya da konularına yönelik olan verileri değerlendirerek bu konuları öncelikle tutmasıdır. Veri ambarının işletmelerde konuya yönelik olmasındaki anlam, veri ambarının işletmelerdeki önceliğinin daha çok yüksek seviyeli varlıkların önemine değinmiş olmalarıdır. Veri ambarlarının ortamdaki dikkat çeken tarafı ise bütünleşik durumda görülmesidir. Bunun herhangi bir istisnası veya değişkenliği olmaz. Veri ambarlarında bütün verilerin, veri zamanının belirli bir vaktine aittir. Veri ambarlarında ki verilerin temel çalışma yöntemi, işlemsel sistemdeki verilerden bir hayli başkadır. İşlemsel yöntemlerde elde edilen veriler o an var olduğu değeri gösterir. İşlemsel sistemlerde bir veriye ulaşıldığı zaman genellikle veriniz o zaman ki önemi ile değerlendirilir. İşlemsel veride zaman dağılmış olabilir. Fakat o zaman, örneğin 90 gün olmuş olabilir. Veri ambarı farklı bir karakteristiği ise, veri ambarı da bulunan verilerin yalnızca okunabilir bir düzende bulunmasıdır. Veri ambarındaki veri yönetiminin ihtiyaçlarına cevap sunulmak üzere düzenlediği işlemleri tabi tutulmaz, yani güncelleme yapılamaz ve silinemez (Özkan, 2008: 29).

Veri tabanları veri ambarlarına göre temel de bazı karakteristik farklılıklar göstermektedir. Veri ambarları önemli veya ilişkili verileri toplarken veri tabanları ise önemli ya da önemsiz, ilişkili ya da ilişkisiz verilerin tamamına sahiptir. Veri ambarlarının tamamında tarihsellik bulunur iken veri tabanları ise online çalıştıkları için hem geçmiş verileri hem de anlık verileri toplamaktadır. Bu anlamda veri tabanı ve veri ambarları özellikleri aşağıdaki Tablo 1.4.'de görüldüğü gibi özetlenebilir.

Tablo 1.4. Veri Ambarı ve Veri Tabanı Özellikleri

Veri Ambarı	**Veri Tabanı**
•Metadatalardan (knowledge) oluşur. •Üst Yönetime hitap eden karar destek sistemleridir. •Son kullanıcı sayısı azdır. (<100) •Off-Line çalışır, anlık değil geçmiş bilgilerle işlem yapar. •Uzun süreçler sonucunda analizler yapılabilir. •Tarihsel verilerden (metadata) oluşur.	•Verilerden (data) oluşur. •Organizasyonun her aşamasında veriye ulaşılır. •Son kullanıcı sayısı fazladır. (>1000) •On-Line çalışır. •Sorgularla istenilen sonuçlara anında ulaşılır. •Güncel ve eski verileri bir arada barındırır.

Veri ambarlarının bazı eksiklikleri şunlardır; veri ambarlarında güncelleme yapıldığı zaman bazı gereksiz bilgiler de içerebilir. Veriler eksik olabilir ve bunun sonucunda tüm sorulara cevap bulamayabilir. Veri ambarlarının yararları ise şunlardır; Yapılan sorgular veri tabanlarından cevaplandığı için sorgulama performansları yüksektir. Konular birbirinden güzel ayrıştırılır. Tarihsel veriler madenlenebilir, sorgulanabilir analiz edilebilir tarihsel veriler bulunmaktadır. Yerel süreçlerde aksama ve duraklamanın olduğu düşünülerek ve veri ambarı güncellemelerini bu takribi duraksama dönemlerinde yapar. Karmaşık sorgular veri ambarlarında yapılırken OLTP sorguları ise kaynak sistemlerden çalışır. Bu şekilde veri ambarındaki işleyiş ile kaynaktaki yerel işleyişler karışmaz. Veri ambarlarında bilgi bağımsız kaynaklar üzerinden yapılır. Veri, veri ambarlarına aktarılır. Bu şekilde değişiklik yapılabilir, not alınabilir, özetlenebilir, silinebilir vs. özgün veri güvendedir (Şentürk, 2011: 4-5).

1.4.1. Veri Ambarının Özellikleri

Veri ambarları, çok boyutlu modelleme ile depolama amaçlı kullanılan ve farklı kaynaklardan gelen verilerin bir araya getirildiği yapılardır. Bu ambarlar, özellikle zaman serileri ve trend analizi gibi tarihsel bilgiye ihtiyaç duyan analizleri desteklemek için tasarlanmıştır. Veri ambarında bulunan veriler, sık değişime tabi tutulmaz; bunun yerine, belirli aralıklarla periyodik güncellemeler yapılır. Bu yapı,

veri ambarlarının geniş kapsamlı ve geçmişe dönük analizler için istikrarlı ve güvenilir bir bilgi kaynağı olarak işlev görmesini sağlar (Gürsoy, 2009: 8).

Veri ambarlarının taşıması gereken dört temel özellik bulunmaktadır:

1.4.1.1. Verinin Zamana Bağlı Olması

Verinin zamana bağlı olması, veri ambarındaki verinin belirli bir zaman dilimiyle ilişkilendirilmiş olmasını ifade eder. Operasyonel sistemlerde veriler, erişildiği anda güncel ve geçerli olmalıdır; ancak işlemler sürekli olarak güncellendiğinden, bu veriler birkaç saniye içinde geçerliliğini kaybedebilir. Bu nedenle operasyonel sistemlerdeki veri, anlık eksiksizlik ve güncellik gerektirir. Veri ambarlarındaki veriler ise, belirli bir an için eksiksiz ve anlamlı olmalıdır, ancak bu güncellemelerin anlık olması gerekmez. Zamana bağlı veri yapısı, veri ambarlarında zaman serileri analizlerinin yapılabilmesine olanak tanır, bu sayede geçmişe yönelik trendler incelenebilir. Veri ambarları genellikle 3 ila 10 yıllık bir zaman dilimini kapsayan verileri barındırırken, operasyonel sistemlerde tutulan veri 60 ila 90 günlük kısa dönem bilgileri içerir. Bu fark, veri ambarlarını uzun vadeli analizler için ideal kılarken, operasyonel sistemlerin güncel işlemler için uygunluğunu sağlar (Gürsoy, 2009: 9).

1.4.1.2. Verinin Kalıcı Olması

Veri ambarına veri yükleme işlemi belirli aralıklarla gerçekleştirilir ve kullanıcılar, bu işlem tamamlandıktan sonra verilere erişebilir. Veri ambarına aktarılan veriler, depolandıktan sonra herhangi bir değişikliğe uğramaz; bu nedenle, güncel veriye ihtiyaç duyan kullanıcılar operasyonel veri tabanlarını tercih etmelidir. Operasyonel veri tabanları, kayıtların sürekli olarak güncellendiği, yeni verilerin eklendiği veya mevcut verilerin silinip düzenlendiği dinamik sistemlerdir. Buna karşın, veri ambarına aktarılmış veri, değişmez bir yapıya sahiptir ve yalnızca yeni bir veri yüklemesi yapıldığında güncellenir. Veri ambarına aktarıldıktan sonra hatalı olduğu tespit edilen veriler, öncelikle operasyonel veri tabanında düzeltilmeli ve sonrasında güncellenmiş haliyle yeniden veri ambarına yüklenmelidir. Bu yapı, veri ambarlarının güvenilir ve istikrarlı bir bilgi kaynağı olarak kullanılmasını sağlar (Gürsoy, 2009: 9).

1.4.1.3. Veri Ambarının Konuya Yönelik Olması

Veri ambarları, özellikle uzun vadeli stratejik kararlar almak için tasarlanmıştır ve müşteri, tedarikçi, ürün veya etkinlik gibi konulara odaklanarak organize edilir. Operasyonel veri tabanları ise daha çok günlük, anlık işlemler için bilgi sağlar. Örneğin, operasyonel veri tabanı, belirli bir zaman diliminde yapılan satışların ayrıntılı kayıtlarını içerir ve bu kayıtlar, müşterilerin satın aldığı ürünleri detaylı bir biçimde tanımlar. Veri ambarları ise bu tür operasyonel bilgilere değil, daha çok karar alma süreçlerini desteklemek için gerekli olan geniş ve uzun dönemli verilere odaklanır. Bu ayrım, veri ambarını stratejik analiz ve raporlama ihtiyaçlarına yönelik ideal bir kaynak haline getirirken, operasyonel veri tabanı anlık iş süreçlerini destekler (Gürsoy, 2009: 9-10).

1.4.1.4. Verinin Entegre Edilmiş Olması

Veri ambarına operasyonel veya harici kaynaklardan veri aktarımı sırasında, veriler entegre edilir ve farklı kaynaklardan gelen bilgiler tutarlı bir formatta birleştirilir. Örneğin, bir kaynakta cinsiyet "E/K" olarak, diğer bir kaynakta ise "0/1" olarak tanımlanmışsa, veri ambarına aktarılmadan önce bu farklılıklar giderilerek tek bir kodlama standardına dönüştürülmelidir. Aynı şekilde, verilerin adlandırılmasında veya ölçü birimlerinde de tutarlı bir kodlama biçimi uygulanmalıdır. Bu entegrasyon sürecinde, farklı kaynaklardan gelen verilerdeki uyumsuzlukların ve tutarsızlıkların da düzeltilmesi gerekir, böylece veri ambarında bütünlüklü ve tutarlı bir bilgi altyapısı sağlanır (Gürsoy, 2009: 10).

1.4.2. Veri Ambarının Amaçları

Veri ambarının temel amacı, farklı kaynaklardan toplanan büyük miktarda veriyi bir araya getirerek, analiz ve raporlama süreçlerini desteklemektir. Veri ambarı, işletmelerin geçmiş verilere dayalı olarak stratejik kararlar almasına olanak tanır. Bu sistemler, veri entegrasyonu, tutarlılık, doğruluk ve güvenilirlik sağlayarak, kullanıcıların veriyi daha kolay anlamasına ve işlemesine yardımcı olur. Ayrıca, geçmiş trendleri analiz etmek, performans değerlendirmesi yapmak ve geleceğe yönelik tahminlerde bulunmak için kapsamlı bir altyapı sunar. Veri ambarı,

işletmelerin rekabet avantajı elde etmesine ve iş süreçlerini optimize etmesine önemli katkılarda bulunur.

1.5. Veri Madenciliğinin Etkileyen Sebepler

Asıl olarak veri madenciliği ve ilerleyişini beş ana etken etkilemektedir. Bunlar (Akpınar, 2000: 1-3):

1. **Veri**: Veri madenciliğinin ilerleyişinde ilk ve en başta gelen etkendir. Son 20 senenin sayısal verilerin de önemli oranda artışı gözlenmiştir. Buna müteakip veri madenciliğindeki gelişmeler geçen zamanda hızlı arttığı gözlenmiştir. Verilerin bu hızlı artışına karşı verilerle ilgili araştırmalar yapan bilim adamları, Mühendisler ve istatistikçiler aynı hızla artmamıştır. Veri analiz yöntemleri, problem çözümleri ve yöntemlerinin geliştirilmesine bağlı olmaktadır.
2. **Donanım**: Veri madenciliği rakamsal ve istatistiksel anlamlı büyük veri kümelerine yoğun işlemler yapması için planlamıştır. Son yıllarda gelişen teknolojik veri yükü kapasiteleri ve hızla gelişen işlemcilerin, son üç beş seneye kadar veri madenciliği yapılamayacak boyutta büyük beri ambarlarının çalışmasına imkan sunmuştur.
3. **Bilgisayar ağları**: En yeni internet teknolojileri, çok daha ileri hızlarda veri aktarımına izin vermektedir. Online teknolojilerin yardımıyla, düzensiz veri tabanlarına erişmek, verilerin faydalı kısımlarını ortaya çıkarabilmek amacıyla araştırmalarını yapmak ve başka algoritma sistemlerini kullanmakla mümkün olmaktadır. 2000'li yıllardan daha öncesinde bilgisayar ağları ve teknolojik gelişmelerin bugünkü teknolojilere nazaran daha eski olması sebebiyle 2000li yıllardan önce bugünkü teknolojileri hayal etmek mümkün değil iken günümüz teknolojileri veri kaynaklarına gelişimi kolaylaştırarak yeni analizleri ve istatistiksel araştırmaların üstünde veri tabanlarının oluşturarak işletmelere ve kuruluşları bilgisayar ağları üzerinden maliyet yüklerini hafifletecek yeni imkanlar sağlamıştır.
4. **Bilimsel hesaplamalar**: Son yıllarda bilim insanları ve mühendislerin, simülasyonu bilimin gücü olarak görmektedirler. Veri madenciliği ve bilgi

keşfi; teorileri, deneyleri ve simülasyonları birbirine entegre edecek mühim bir rol almaktadırlar.

5. **Ticari eğilimler**: Son zamanlarda sektörler arası rekabetin giderek artmakta ve şirketlerin bu ortamda kendilerini koruyabilmeleri gerekmektedir. Bu rekabet ortamında şirketlerin hızlı hareket etmesi, çok iyi kaliteli hizmetler sunabilmesi, onların yanında da maliyeti en aza düşürerek ve insan gücünün en az kullanması gerekmektedir. Bu anlamda veri madenciliği müşterilerin ve müşterilerin yaptıkları Etkinliklerin ortaya çıkarttığı fırsatları daha rahat analiz edebilmekte ve bu riskleri daha kolay görebilmektedir. Sektörlerin pazarda oluşan rekabet piyasasını iyi analiz ederek doğru adımlar atabilmek adına veri madenciliğinde yeterli analizleri yaparak şirketlerin büyüme eğrisini çok daha yukarılara taşıması mümkün olacaktır. Veri madenciliğinin maliyet hesaplamaları her ne kadar yüksek görünse de bu mali hesaplar pazarda doğan risklerle baş edebilmek adına önemli bir etkendir. Ve en önemlisi de veri madenciliğinin getirileri götürülerinden fazladır. Son olarak veri madenciliğinin ticari ilimler son yıllarda hızla artmakta ve sektörlerde yaygın olarak kullanılmaktadır.

1.6. Veri Madenciliğinin Tarihçesi

İnsanlar geçmişten bu zamana kadar elde ettikleri verilerin analizlerini yapıp verilerden malumat almaya çalışmışlardır. Bunun için her dönemde çeşitli Yollar kullanılmıştır. Bu yollar bilginin aktarımını sağlamıştır.

Veri madenciliğinin kökeni, ilk bilgisayar olan ENIAC'a (Electrical Numerical Integrator And Calculator) kadar uzanabilir. 1946 yılında ABD'li bilim insanları John Mauchly ve J. Presper tarafından geliştirilen bu teknoloji, başlangıçta II. Dünya Savaşı sırasında ABD'nin askeri yapısı için planlanmıştı. Verilerin etkin kullanımı, verilerin depolanmaya başlamasıyla önem kazanmış ve ilk dönemlerde bu sistem, veri hesaplamaya yönelik bilgisayarlar olarak kullanılmıştır. Kullanıcı ihtiyaçları doğrultusunda verilerin toplanıp düzenlenmesi amacıyla bu sistemler veri depolama işlemlerinde de kullanılmaya başlanmıştır. Veri madenciliği, özellikle 1960'larda bilgisayarların veri analizi problemlerini çözmek için kullanılmasıyla önem kazandı.

1990’larda ise “veri madenciliği” terimi bilgisayar mühendisleri tarafından literatüre kazandırıldı. Başlangıçta bu kavram; istatistik, makine öğrenimi, veri tabanları, otomasyon, pazarlama ve araştırma gibi farklı disiplinleri içermekteydi. Zamanla, istatistik verilerin değerlendirilmesi ve analizi alanında hizmet veren yöntemler olarak daha spesifik bir anlam kazanmıştır (Öğüt, 2002: 7).

1950'lerde ilerletilen yöntemler neticesinde devamlı gelişen veri madenciliği, son yıllarda birçok veriye hızlı bir şekilde erişebilmeyi mümkün kılarak yaşantımızı kolaylaştırmaktadır. Aynı zamanda onlarca meslek gruplarının da yükleri hafifletilmiştir. Bu yöntemler sebebiyle çalışmaları en başta 1950’li senelerde başlanmış, mantık ve bilgisayar alanı konularda çalışmalar yapmak suretiyle yapay zekâ ve makina öğrenme hususunda mühim ilerleme kazanmışlardır. O senelerde bilgisayarlar sayım için kullanılmaya başlaması veri madenciliğinin ortaya çıkmasına zemin hazırlayan olaydır. 1960’lar da istatistikçiler, regresyon analizi, en büyük olabilirlik kestirim, sinir ağları vb. gibi yeni algoritmalar ve etkili yöntemler üzerinde çalışmışlardır. Sistemlerdeki bu önemli ilerlemelere eş olarak veri tabanı işlemleri artarak ilerlemiş çok sayıda metin dokümanlarının depolanmasına imkân sağlamıştır. Bundan sonra veri tabanı ve veri depolama kavramları dünyada teknolojik kavram olarak tanınmıştır. 1960’lı senelerin sonlarına varıldığında bilim insanları kolay öğrenilebilen bilgisayarlar geliştirilmişlerdir. Böylece verilerin toplanması veri tabanlarının oluşturulması bağlamında hızlı ilerlemeler doğmuştur. Veri tabanı yönetim sisteminin ilk temelleri atılmış bulunmaktadır. 1970'lere gelindiğinde ise veri tabanı yönetim sistemleri işlevsellik kazanmış bulunmakta, uygulanabilir programlar ortaya çıkarak bilim insanlarının kolay kaideleri bulmada sistemlerin geliştirilmesine öncülük etmişlerdir. 1980'lerde ise ilişkisel veri tabanı yönetim sistemleri artık birçok alanda kullanılmaya başlanmış, şirketler, müşterileri, rakipleri ve hizmetleri ile ilgili verilerden derlenen veri tabanları kullanılmaya başlanmıştır. Bu şirketler tarafından kullanılabilen veri tabanları verileri önceki yıllarda tutulan veri depolarından çok daha ileride yeni daha büyük veri tabanlarına ulaştırmıştır. Buradaki veri tabanları gelişen yeni teknolojinin sunduğu imkanlar ile daha büyük depolama alanlarına sahip ve oradaki verileri veri tabanı sorgulama dilleri sayısında erişilebilmektedir. 1980'lerde en çok kullanılan veri tabanı sorgulama dili SQL’dir.

1990'lara gelindiğinde verilerinde çoğalması sebebiyle artık verilerden işe yarayan faydalı bilgiler edilebilen verilere nasıl ulaşılacağı düşünülmeye başlanmış ve bunun hakkında detaylı araştırmalar yapılarak yeni çalışmalara gidilmiştir. 1989'da düzenlenen KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve ardından 1991 yılında yayımlanan, KDD (IJCAI)-89'un sonuç bildirgesi niteliğindeki "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesi, KDD (Knowledge Discovery and Data Mining) ile ilgili temel tanım ve kavramların ortaya konulmasını sağlamış ve bu alandaki gelişmeleri hızlandırmıştır. Bu sürecin sonucunda, 1992 yılında veri madenciliğine yönelik ilk yazılım geliştirilmiştir (Savaş ve ark, 2012: 5).

2000'li yıllara gelindiğinde artık veri madenciliği sürekli tüm sektörlerde kullanılmaya başlanmıştır. Veri madenciliği sistemini getirileri ve faydaları tespit edilmiş ve şirketlere menfi fayda sağladığı görülmüş olup kullanım alanları her geçen gün daha da artmıştır. 2000li yıllardan sonra internetin de kullanımındaki artıştan dolayı internet üzerinden paylaşımlar çoğalmış buradan edinilen bilgi birikimleri toplandığında artık hayal edilemez bir şekilde bilgi depolarına ulaşılmıştır. Bu sebepten veri madenciliği sistemi de artık yeni yöntemlerle kendini geliştirmek adına yeni arayışlara gitmiştir.

Artık son yıllarda verirler çok daha farklı veri tabanlarını üzerinde depolanabilmektedir. Depolama mimarisi veri ambarı kavramının, dolayısı ile veri tutarlılaştırma, veri entegrasyonu ve OLAP işlemlerinin de oluşmasını sağlamıştır. OLAP sayesinde veri analizi, özetleme, birleştirme ve entegrasyon bileşenleri ile çok boyutlu bir şekilde yapılabilir. Bütün bunların yanında sınıflandırma, gruplama, verinin özelliğinin zamanla farklılaştığını izleme gibi detayları inceleyebilmek için ek olarak veri inceleme araçların gereklidir. Bunun nedeni ise donanımlı depolama teknolojilerinde inanılmaz ilerlemelerin veri zengini lakin bilgi yoksunu bir hal ortaya çıkmıştır. Büyük miktardaki verinin ve büyük arşivlerin olduğu böyle bir zamanda karar vericiler sezgilerine göre davranmaktadırlar. Ancak en önemli olan önce büyük bilgi kaynağından değerli olan bildiği veri sistemlerinden ihtiyaç duyulur zamanda alabilmektir. Evet bu devasa veri

kaynaklarından elde edilmesi gereken bilgileri aşağı çıkartacak olan veri madenciliği ve veri araçlardır. (Kocabaş, 2010: 6).

Veri madenciliği, öğrenme yöntemlerini iş ve bilimsel veri setlerine uygulayarak anlamlı bilgi çıkarma sürecidir. İstatistik, yapay zekâ ve makine öğrenmesi gibi disiplinlerin gelişimiyle ortaya çıkan veri madenciliği, eldeki veriden öğrenme yoluyla gizli bilgileri ve örüntüleri keşfetmeyi ve bu bilgilerden ileriye dönük tahminlerde bulunmayı amaçlayan modern bir bilim dalı olarak tanımlanır. İş dünyasında ve bilimsel araştırmalarda, yoğun ve büyük veri kümelerinden normalde elde edilmesi çok zor olan bilgileri çıkarma becerisi sayesinde her geçen gün daha fazla kabul görmektedir. Bu yetenek, veri madenciliğini stratejik ve analitik karar destek araçlarının merkezine yerleştirmektedir.

1.7. Veri Madenciliğinin Uygulama Alanları

Veri madenciliğinde çok daha yeterli uygulama alanları mevcuttur. Müşterilerin satın alma yaklaşımları, alışkanlıkların tespit edilmesi, internet sayfalarına erişimlerinin araştırılması, ilaçların ne şekilde yan etki gösterdiğini ortaya çıkarılması, hastalıklar ve şikayetler arasındaki bağlantı, bankacılık sektöründe müşteri davranışları, kredi borcunu ödemeyi yönelik elde edilen tahminler gibi pek çok alanda karar noktasında veri madenciliğinin öngörüleri bizlere yol göstermektedir (Ertuğrul ve ark, 2012: 98).

Veri madenciliği konusunda da kararlar alınma da yardımcı olarak iş kararlarının optimizasyonunda, var olan müşteri potansiyelinden nasıl daha fazla kazanılacağına aynı zamanda müşterilerin aldığı ürünlerden ve hizmetlerden nasıl daha çok memnun olmaları konusunda yardımcı olmaktadır. Veri madenciliği uygulamada yapılan araştırmalar sektörler arası işletmeden işletmeye, firmadan firmaya göre farklılıklar göstermektedir. Çok sık kullanılan bazı veri kaynakları; satış kayıtları, çağrı destek kayıtları, müşterilere ait demokratik veriler, firmanın internet sitesine ait ziyaretçilerin kayıtları ve buna benzer diğer kaynaklardır. (Demirel, 2008: 35).

Veri madenciliğinin temel amacı, anlamlı bilgiler elde ederek bu bilgileri eyleme dönüştürülebilir kararlar için kullanmaktır. Veri madenciliğinin odaklandığı ana kütle genellikle mevcut veya potansiyel müşterilerdir. Bu bağlamda, müşterilerin profillerini oluşturmak, satın alma eğilimlerini analiz etmek, bir ürün ya da hizmeti kabul etme veya reddetme olasılıklarını tahmin etmek veri madenciliği ile hedeflenen ana amaçlardandır. Elde edilen bu tahminler ve içgörüler, strateji belirleme sürecini destekler ve işletmenin gelecekte alacağı çeşitli kararlar için güçlü bir temel sağlar (Özmen, 2001: 3).

Firmalar, veri madenciliği ile (Demirel, 2008: 36):

- Çeşitli kaynaklardan toplanan verilere ait gizli eğilimleri ortaya çıkarmak.
- Büyük veri yığınlarından kritik ve değerli bilgileri elde etmek.
- Stratejik karar alma süreçlerini daha etkin hale getirir.
- Maliyetleri azaltır.
- Başarı performansını ve kârlılığı artırır.

Tablo 1.7.'da 2003 yılında, veri madenciliğinin sektörel bazda kullanımına dair yapılan bir araştırmanın sonuçları yer almaktadır.

CRM/ Müşteri Analizleri (41)	% 38.3
Bankacılık (34)	% 31.8
Dolandırıcılık Tespiti (21)	% 19.6
Finans (18)	% 16.8
Doğrudan Pazarlama (15)	% 14.0
Diğer (14)	% 13.1
Yatırırım/Borsa kararları (14)	% 13.1
Kredi kartı Skorlama (14)	% 13.1
Telekomünikasyon13)	% 12.1
Perekandecilik(13)	% 12.1
Reklam (13)	% 12.1
Biyoteknoloji/Genetik (12)	% 11.2
Bilim (11)	% 10.3
Sigortacılık (11)	% 10.3
Sağlık (10)	% 9.3
İmalat (9)	% 8.4
E-ticaret (8)	% 7.5
Web Kullanım Madenciliği(8)	% 7.5
Sosyal Politikalar/Anket Analizi(8)	% 7.5
Tıp/Farmakoloji (8)	% 7.5
Güvenlik/Anti-terör (6)	% 5.6
Web içerik madenciliği (6)	% 5.6
Kamu/Askeri uygulamalar (4)	% 3.7
Seyahat (3)	% 2.8
Junk e-posta / Anti-spam tespiti (3)	% 2.8
Eğlence /Müzik (3)	% 2.8
Sosyal Ağlar(2)	% 1.9

Tablo 1.7. Veri madenciliğinin uygulandığı alanların dağılımı

Veri madenciliği uygulamaları, son yıllarda dünya piyasasındaki değişen ekonomik koşulların rekabeti artırmasıyla birlikte geniş bir yelpazede kullanılmaya başlanmıştır. Bu uygulamalar, özellikle pazarlama alanında öne çıkmakla birlikte tıp, finans, astronomi, spor, lojistik, sigorta, trafik yönetimi, biyoloji, güvenlik, meteoroloji, tedarik zinciri yönetimi, milli güvenlik ve ulaşım gibi birçok alanda kendine yer bulmuştur. Veri madenciliğinin bu geniş çaplı kullanım alanları, çeşitli sektörlerde karar destek süreçlerini güçlendirerek daha verimli ve stratejik işleyişler sunmaktadır (Akpınar, 2000: 4).

Günümüzde yaygın olarak kullanıldığı alanlar şunlardır (Baykal, 2006: 97):

Pazarlama Alanı ile ilgili olarak;

· Müşterilerin satın alma davranışlarının örüntülerinin belirlenmesi ve segmentlere ayrılması,

· Demografik özellikler arasındaki ilişkilerin analiz edilmesi,

· Pazarlama kampanyalarının etkili bir şekilde tasarlanması,

· Mevcut müşteri bağlılığını artırmaya yönelik stratejilerin geliştirilmesi,

· Çapraz satış ve pazar sepeti satış analizlerinin yapılması,

· Müşteri değerlemesi ve müşteri ilişkileri yönetimi,

· Çeşitli müşteri analizleri ve satış tahminleri.

Bankacılık Alanı ile ilgili olarak;

· Farklı finansal veriler arasındaki saklı ilişkilerin bulunması,

· Kredi kartı usulsüzlükleri ve dolandırıcılıklarının tespiti,

· Müşteri gruplaması ve kredi değerlendirmesi,

· Kanun dışılık tespitinde,

· Risk analizi ve yönetiminde kullanımı.

Sigortacılık Alanı ile ilgili olarak;

· Yeni poliçe talebinde bulunabilecek potansiyel müşterilerin öngörülmesi,

· Riskli müşteri gruplarının ve sigorta dolandırıcılığının tespiti.

Perakendecilik Alanı ile ilgili olarak;

· Satış noktası verilerinin analizi ve alışveriş sepeti çözümleri,

· Tedarik zinciri yönetimi ve mağaza yerleşim düzeninin iyileştirilmesi.

Borsa Alanı ile ilgili olarak;

· Hisse senedi fiyatlarının tahmini,

· Genel piyasa analizi ve alım-satım stratejilerinin optimize edilmesi.

Telekomünikasyon Alanı ile ilgili olarak;

· Kalite iyileştirme analizleri ve hat yoğunluğu tahminleri.

Tıp ve Sağlık Alanı ile ilgili olarak;

· Test sonuçlarının öngörülmesi ve ürün geliştirme,

· Tıbbi teşhis ve tedavi süreçlerinin belirlenmesi.

Endüstri Alanı ile ilgili olarak;

· Kalite kontrol analizleri ve üretim süreçlerinde optimizasyon.

Bilim ve Mühendislik Alanı ile ilgili olarak;

· Ampirik veri modelleri oluşturma, yeni virüslerin keşfi ve sınıflandırılması,

· Gen haritası analizi, genetik hastalıkların ve kanserli hücrelerin tespiti,

· Gezegen yüzey yapılarının incelenmesi ve yeni galaksilerin keşfi.

Diğer Uygulama Alanları,

· Parmak izi veya yüz tanıma sistemleri ile kimlik doğrulama,

· Kanser hastalarının kemoterapiye yanıt olasılıklarının tahmini

· Bankalarda kredi kartı dolandırıcılığının tespiti,

· DNA araştırmaları ile hastalıklara sebep olan genlerin sıralanması,

· Süpermarketlerde müşterilerin ilgi gösterdiği ürünlerin belirlenmesi,

· Çip üretiminde kusurların veri madenciliği ile azaltılması,

· Telefon hatlarındaki gürültü ve parazitlerin giderilmesi.

Veri madenciliğinin kullanmak açısından sağlık sektörü en önemli sektörlerden biridir. Doğru ve vaktinde karar almanın hastanın sağlığı üzerinde çok büyük etkisinin oldu tartışılamaz bir gerçektir. Klinik ve hastanelerde oluşan araştırma ve hastane verileri hastaların risk faktörünün önceden belirlemesi ile ona uygun tedavi yöntemleri uygulanabilmesinde ya da tedavi süreçlerinde yaşanan bu sorunlarının giderilmesinde ve ilaç tedavilerinin kullanımında karşılaşılan yan etkileri tespit edilmesinde veri madenciliği uygulaması kullanılabilir. Ayrıca hastane ortamlarında hastane yönetimine de hastane birimlerinin çalışma başarıları, kaynak kullanımlarında, hastaların hastalık eğilimlerini tahmin etmede de veri madencilik uygulaması kullanılmaktadır (Erkuş, 2015: 18).

1.8. Veri Kaynakları

Veri Madencilik sistemi altı basamaktan oluşan bir süreci takip etmektedir. Bu süreç hangi bilgileri ihtiyacın olduğunun belirlenmesi ile başlamıştır. Verilerin hangi kaynaktan elde edilmesi gerektiğini ikinci basamakta karar verilir. Verileri farklı kaynaklardan bilgi edinmek maksadıyla olabilir. Toplama işlemi, verilerin farklı farklı kaynaklardan toplanmasıyla tek kaynakta birleştirilmesidir. Gereksinimi duyulan veriler şirket içi kaynaklardan, veri ambarlarında ya da hazır veri sunan bazı internet sitelerinden bulunabilmektedir. Örneğin hisse senetleri ile ilgili bazı araştırmalar yapılacaksa, bu araştırmalar Borsa İstanbul borsasının emin kaynaklarından faydalanılabilir (Gürsoy, 2012: 4).

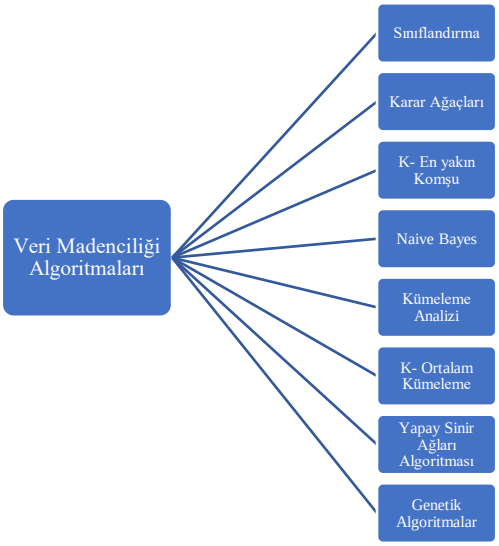
Veri kaynakları aşağıda yer almaktadır;

• Veri ambarları

• Gelişmiş veri tabanları ve bilgi depoları (uzamsal, metin, çoklu ortam, heterojen veri tabanları, zamansal veriler, internet tabanlı veriler, nesne tabanlı, nesne ilişkisel)

• İlişkisel veri tabanları (Farboudi, 2009: 29).

1.9. Veri Madenciliği Algoritmaları



Şekil 1.9. Veri madenciliği Algoritmaları

1.9.1. Sınıflandırma Algoritmaları

Sınıflandırma, yeni bir şeyin kalitesini araştırmak ve bu şeyi önceden tanımlanmış bir kümeye aktarmaktır. En önemli olanın da burada her bir kümenin daha önceden net bir şekilde belirlenmiş olması gerekir. Bu kümelemeye veya sınıflandırmaya örnek verecek olursak kredi kartı başvurularının düşük, orta ve riskli sınıflar olmak üzere ayırmaktır (Özkan, 2013: 41).

Veri madenciliğinin en temel işlevlerinden biri de kategorik sonuçları tahmin etmek için kullanılmasıdır. Modeli geliştirebilmek için lise sonuçları önceden bilinen durumları ve bu durumlarla ilgili olan işlevlerin alabileceği değerlerdir. Bu değerler ‘‘eğitim verisi’’ olarak da isimlendirilir. Elde edilmesi gereken tahminler ise ‘‘müşteri %80 olasılık ile kampanyaya olumlu cevap verecek’’ şeklindedir tahmin ile birlikte verilir. Sonuçlar, ‘‘Kesin Tercih Eder, Tercih Etmez Yanıt Vermez, Tercih Eder Yanıt Vermez’’ gibi çeşitli kategorilerde olabileceği gibi, ‘‘Hizmeti Bırakmaz-Hizmeti Bırakır’’ şeklinde iki yönlü de olabilir. Bu tür deneme sınıfları modeli de doğruluğu en olası şekliyle belirlemek amacıyla kullanılır. Çoğunlukla verilen veri kümesi öğrenme ve deneme kümesi olmak üzere iki gruba ayrılır. Öğrenme kümesi modeli oluşturulma maksadı ile deneme kümesi ise modelinin doğruluğunu teyit amacıyla kullanılır. Buna bir örnek verecek olursak otomobil satıcı şirket geçmişte satış yaptığı müşterilerin analizlerini yaparak yukarıda belirtildiği üzere bu iki kural şayet bulunursa genç kadınların okuduğu dergilere küçük model araçların reklamını verir. (Argüden ve Erşahin, 2008: 37).

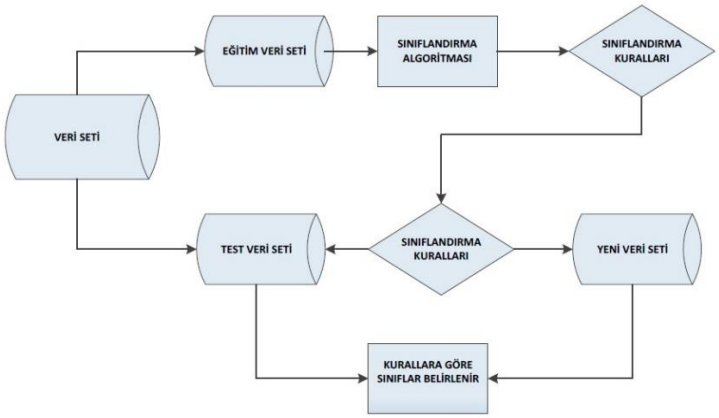
1.9.2. Karar Ağaçları Algoritması

Karar ağacı, karar düğümleri, dallar ve yapraklardan oluşan bir yapıdır (Han, 2000). Karar düğümü, gerçekleştirilecek testi temsil eder ve bu testin sonucuna bağlı olarak veri kaybı olmaksızın ağacın dallara ayrılmasını sağlar. Her düğümde test işlemi ve dallara ayrılma süreçleri sıralı bir şekilde gerçekleşir ve bu ayrılmalar, üst seviyedeki kararların sonuçlarına bağlıdır. Ağacın her dalı, sınıflama işlemini tamamlamak üzere yapılandırılmıştır. Eğer bir dalın ucunda sınıflama işlemi tamamlanamazsa, bu durumda yeni bir karar düğümü oluşur. Ancak, dalın sonunda belirli bir sınıf tanımlanırsa, o noktada yaprak yer alır. Bu yaprak, veri kümesindeki

hedef sınıflardan birini temsil eder. Karar ağacı işlemi kök düğümünden başlayarak yukarıdan aşağıya doğru ilerler ve yaprağa ulaşana kadar ardışık düğümleri takip ederek sonlanır (Özekes, 2003:66).

Karar ağacı tekniğini kullanarak verinin sınıflanması iki basamaklı bir işlemdir (Han, 2000). İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır (Özekes, 2003:66).

Karar ağaçları bir olasılık tekniği değildir. Çoğunlukla sınıflandırma, kümeleme bir tahmin biçimlerinde kullanılır ve sorunlarla alakalı araştırma alanlarının alt grupları için kullanılır. Karar ağaçları öğretici örnekteki veriyi sorgulayan algoritma aracılığıyla oluşturulur veya şirketin uzmanları tarafından oluşturulur. Karar ağaçları oluşturma yöntemlerine bağlı olarak birbirlerinden ayrılırlar (Silahtaroğlu, 2013: 21).

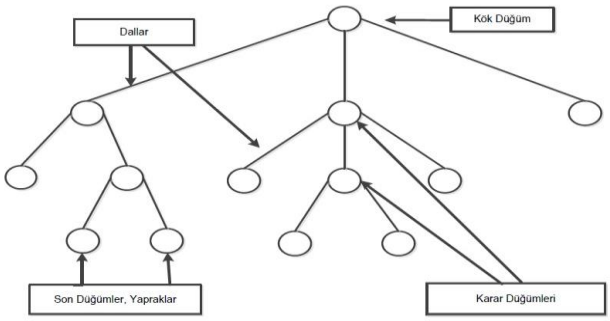


Şekil 1.9.2. Veri Sınıflandırma İşlemi

Karar ağaçları algoritması, veri setini eğitim ve test veri seti olarak ikiye ayırarak çalışır. Eğitim verileri öğrenme sürecinde sınıflandırma kurallarının oluşturulması için kullanılır. Test verileri ise bu kuralların doğruluğunu test etmek ve algoritmanın

başarısını değerlendirmek için ayrılmıştır. Karar ağaçları algoritması, test verisini veya yeni verileri oluşturulan bu kurallar doğrultusunda sınıflandırmaya imkan tanır. Şekil 1.9.2.'de veri setinin eğitim ve test seti olarak ikiye ayrılmasını göstermektedir. Eğitim veri seti üzerinde sınıflandırma algoritmaları uygulanarak sınıflandırma kuralları çıkarılır. Daha sonra bu kurallar, test veri setine veya yeni bir veri setine uygulanarak, veriler belirlenen kurallara göre sınıflandırılır (Pala, 2013: 8-9).

Bir karar ağacı, kök, iç ve yaprak düğümlerden oluşur. Kökten yapraklara doğru hiyerarşik olarak yapılandırılan ağaç, yukarıdan aşağıya doğru oluşturulur. En üstteki düğüm olan kök düğüm, sınıflandırmanın başlangıç noktasıdır. İç düğümler, algoritmalar yardımıyla en iyi kararları verecek şekilde ayrılır ve dallanır. Yaprak düğümler ise veri setindeki grupları temsil eden sınıf etiketlerini, yani kategorik özellikleri içerir. Karar ağaçlarının ağaç yapısının görselleştirilmesi ve kurallarının yorumlanması oldukça kolaydır, bu da onları anlaşılır ve kullanışlı bir sınıflandırma modeli haline getirir. Şekil 1.9.3.'te karar ağacı yapısını detaylı bir şekilde göstermektedir (Liu, 2012).



Şekil 1.9.3. Karar Ağacı Yapısı

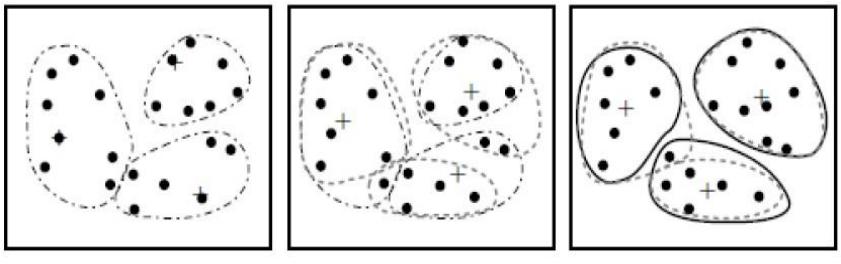
Şekil 1.9.3'teki karar ağacı yapısı incelendiğinde, akış diyagramı biçiminde bir yapı olduğu görülür. Bu yapı, bir alandaki kararları belirten karar düğümlerinden, karar değerlerini gösteren dallardan ve sınıfları temsil eden yaprak düğümlerinden oluşur. Ağaç yapısındaki ilk düğüm, sınıflandırma sürecinin başlangıç noktası olan kök düğümdür. Karar ağacı ve buna bağlı karar kuralları oluşturulduktan sonra, sınıfı bilinmeyen yeni bir veri örneği, karar ağacı üzerinde yer alan dallar aracılığıyla nitelik değerleri üzerinden test edilerek sınıflandırılmaktadır. Bu süreç, her bir niteliğin

ağaçtaki ilgili dallarda değerlendirildiği ve nihai olarak en uygun sınıfın belirlendiği bir sınıflandırma adımını içermektedir. Verinin hangi sınıfta olduğu, kök düğümden yaprak düğüme doğru ilerleyerek kolayca belirlenir, böylece sınıflandırma işlemi tamamlanmış olur. Bu yapı, karar ağacının hem anlaşılmasını hem de yorumlanmasını kolaylaştıran bir sistem sunar (Pala, 2013: 10).

1.9.3. K-En Yakın Komşu

Sınıflandırma problemlerini çözmede kullanılan denetimli öğrenme yöntemleri arasında K en yakın komşu yöntemi bulunmaktadır. Yöntem; sınıflandırılması olacak verilerin öğrenme kümesindeki normal verileri benzerliklerini hesaplayarak en yakın olduğu varsayılan k verisinin ortalamasıyla, belirlenmiş eşik değerlere göre sınıflara atar. Sınıf özelliklerinin önceden net olarak belirlenmiş olması, yöntemin doğruluğu ve etkinliğini artırmak için kritik bir öneme sahiptir. Bu sınıflandırma yönteminin performansını etkileyen faktörler arasında **k** değeri, eşik değeri, öğrenme ve benzerlik ölçümü kümesinde yeterli sayıda normal davranış örneğinin bulunması yer alır. Bu kriterler, yöntemin doğruluğunu artırarak sınıflandırma sürecinin daha güvenilir ve verimli olmasını sağlar .

Şekil 1.9.3'te k=3 durumu için gerçekleştirilen kümeleme işlemi görselleştirilmiştir. Her bir kümenin merkez noktası, verilerin gruplanma eğilimini göstermek amacıyla '+' simgesiyle işaretlenmiştir. Bu sembol, küme içindeki verilerin ortalama değerini temsil etmektedir ve kümeler arasındaki mesafeyi görsel olarak vurgulamaktadır.



Şekil 1.9.3. K-En Yakın Komşu Algoritması ile Kümeleme

K-En Yakın Komşu (K-NN) Algoritması, gözlem değerlerinden oluşan bir veri kümesi için aşağıdaki adımları içerir (Özkan, 2013: 118):

1- **K Değerinin Belirlenmesi:** İlk olarak, sınıflandırmada kaç komşunun dikkate alınacağını belirlemek için **k** değeri seçilir. Bu değer algoritmanın performansını doğrudan etkiler.
2- **Uzaklık Ölçüm Yönteminin Seçimi:** Her veri noktası ile sınıflandırılacak örnek arasındaki mesafeyi hesaplamak için uygun bir uzaklık ölçümü (örneğin, Öklidyen mesafe) seçilir.
3- **Uzaklık Hesaplama:** Sınıflandırılacak veri noktası ile eğitim kümesindeki her veri noktası arasındaki mesafeler hesaplanır.
4- **K En Yakın Komşunun Seçilmesi:** Hesaplanan uzaklıklara göre en yakın **k** komşu seçilir.
5- **Sınıf Belirleme:** Seçilen k komşunun çoğunluk sınıfı, sınıflandırılacak veri noktasının sınıfı olarak atanır.
6- **Sonuçların Değerlendirilmesi:** Algoritmanın performansı, seçilen k değeri ve uzaklık ölçüm yöntemi gibi faktörlere göre değerlendirilir ve gerekirse ayarlamalar yapılır.

1.9.4. Naive Bayes

Naive Bayes sınıflandırıcıları, Bayes teoremine dayalı olarak çalışan ve olasılıkları kullanarak sınıf tahmini yapan temel sınıflandırma algoritmalarından biridir. Bu yöntem, basitliğiyle öne çıkmasına rağmen, özellikle yüksek boyutlu veri setlerinde etkili bir performans sergileyebilir. Çoğu zaman karmaşık sınıflandırma yöntemlerine kıyasla daha iyi sonuçlar verebilir ve bu nedenle genellikle yeni sınıflandırma tekniklerinin performans değerlendirmesinde bir referans noktası olarak kullanılır (Pukar,2022:9)

Naive Bayes algoritması, her kriterin sonuca olan etkisini tahmin etmek için olasılık hesaplamalarını temel alır. Bunu bir örnekle açıklayacak olursak; elimizde bir tenis maçının oynanıp oynanmayacağına dair bilgi olduğunu varsayalım. Bu bilgilere ek olarak, maçın oynanacağı günün hava durumu, sıcaklık, nem ve rüzgâr gibi çevresel koşullar da kaydedilmiş olsun. Bu durumda, algoritma, örneğin hava rüzgârlıysa tenis maçının oynanmayacağı şeklinde bir karar verebilir. Ancak bu karar, veri madenciliği ile tüm faktörlerin etkisinin hesaba katıldığı olasılıksal bir çıkarıma dayanır. Veri

madenciliğinde, tüm bu çevresel kriterler bir araya getirilerek sistem eğitilir. Örneğin, geçmiş verilerden faydalanarak sistem "bugün hava güneşli, sıcak, nemli ve rüzgârsız" gibi bir bilgiyle beslendiğinde, Naive Bayes algoritması daha önce gerçekleşmiş durumlara dayanarak tenis maçının oynanma veya oynanmama olasılığını hesaplar ve en muhtemel sonucu bize tahmin olarak sunar. Bu şekilde, her kriterin etkisi ayrı ayrı analiz edilerek daha doğru bir sonuç elde edilir (Ayık ve ark, 2007: 446).

1.9.5. Kümeleme Analizi

Kümeleme çözümlemesi, araştırma konusu olan nesne veya bireylerin, aralarındaki benzerliklere göre gruplandırılmasını sağlar. Bu yöntemle nesne veya bireyler belirli kriterlere göre kümelere ayrılırken, kümelerin içindeki homojenlik en üst düzeyde tutulur ve kümeler arasındaki farklılık (heterojenlik) vurgulanır. Kümedeki bireyler kendi aralarında benzer özellikler gösterirken, farklı kümelerdeki bireyler arasında belirgin farklılıklar bulunmaktadır. Benzerlik ve homojenlik kavramları, yapılan analizde kullanılan yönteme ve analiz amacına göre değişiklik gösterebilir. Bu bağlamda, analiz amacına uygun olarak farklı sayıda ve özellikte kümeler oluşturulabilir. Kümeleme analizi, birçok farklı işlevi yerine getiren çeşitli yöntemlerden oluşur ve analiz amacı doğrultusunda farklı kümeleme teknikleri uygulanabilir. Ayrıca, değişkenlerin farklı ölçüm birimleri ve ölçüm teknikleri kullanılarak değerlendirilebilmesi nedeniyle, birimlerin benzerliklerini belirlemek için de çeşitli benzerlik ölçüleri kullanılır. Bu esneklik, kümeleme analizini geniş bir uygulama alanına sahip kılar (Gürsoy, 2009: 57).

1.9.6. K-Ortalama Kümeleme Algoritması

K-ortalama algoritması, veri kümesini kkk adet merkez kullanarak temsil etmeyi amaçlayan bir kümeleme yöntemidir. Algoritmada merkez noktalarına başlangıç değerleri rastgele olarak atanır ve bu merkezlerin güncellenmesi için iki farklı yöntem uygulanır.

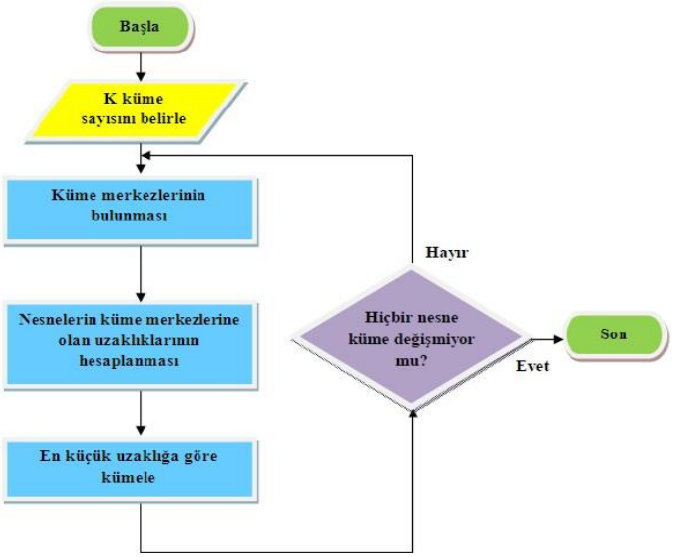
1. **Ortalama Bazlı Güncelleme**: İlk yöntemde, giriş kümesindeki her bir örneğin hangi merkeze en yakın olduğu hesaplanır. Aynı merkeze yakın olan örneklerin ortalaması alınarak merkezin konumu güncellenir. Bu işlem, durma koşulu

karşılanana kadar (örneğin, merkezlerin yer değişim miktarı belirli bir eşik değerin altına indiğinde) tekrarlanır.

2. **Örnek Bazlı Güncelleme**: Yöntemin ikincisi ise giriş kümesinden tek bir örnek seçilerek ve örneğin bu merkezlere olan uzaklıkları hesaplanır. En yakın merkez belirlenir ve bu merkezin değeri, seçilen örneğin verileriyle güncellenir. Bu güncelleme sırasında merkez ile örnek arasındaki mesafe her adımda azalan bir öğrenme katsayısı ile çarpılarak uygulanır. Bu sayede ilk aşamalarda merkezlerin konumları hızlıca değişirken, zamanla bu değişiklik azalır ve merkezler daha kararlı bir konuma yaklaşır. Her örnek için bu işlem durma koşulu sağlanıncaya kadar tekrarlanır.

Örnekler her adımda aynı sırada işlenebileceği gibi, rastgele bir sırada da ele alınabilir. Bu esneklik, algoritmanın hem performansını artırır hem de yerel optimumlara takılma olasılığını azaltır (Keskin, 2013: 36).

K-Ortalama kümeleme algoritmasının çalışmasını özetleyen akış diyagramı Şekil1.9.6.gösterilmiştir.



Şekil 1.9.6. K-Ortalama kümeleme algoritmasının akış diyagramı

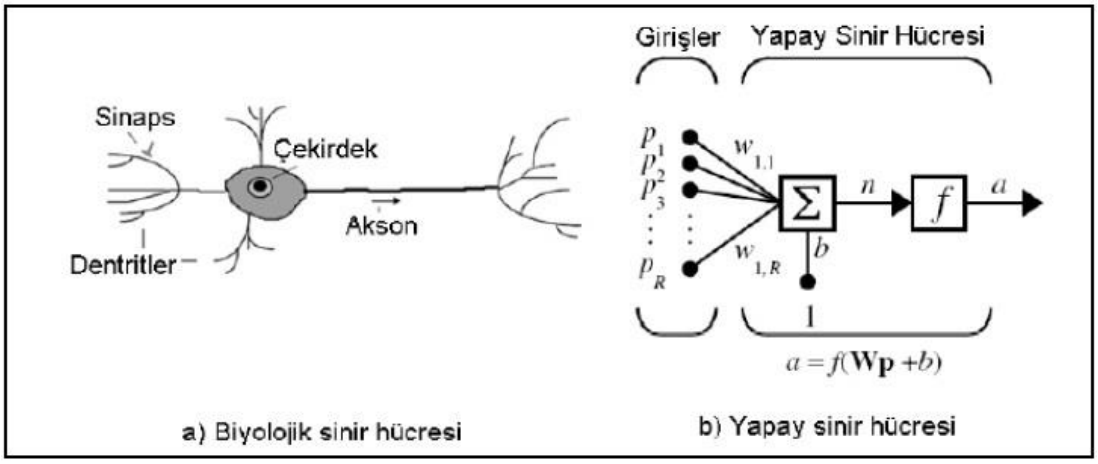
1.9.7. Yapay Sinir Ağları Algoritması

Yapay sinir ağları, insanı ya da insan beynine çok fazla benzeyen özellikleri itibariyle de öğrenme yoluyla bilgiler üretebilen, yeni bilgiler edinebilmek ve yeni araştırmalar yapabilmek için de herhangi bir yardıma ihtiyaç duymadan çalışabilen bilgisayar sistemleridir. Bu çalışma sistemini başka bir bilgisayar programında yapmak mümkün değildir ve zaten oldukça da zordur. Bu sebeple, yapay sinir ağlarının, programlama yoluyla çözülmesi imkânsız veya son derece zor olan durumlar için tasarlanmış, adaptif bilgi işleme alanında uzmanlaşan bir bilgisayar bilimi dalı olduğu ifade edilebilir (Gürsoy, 2009: 45).

Yapay sinir ağları, insan beyninden ilham alınarak geliştirilmiş, her biri kendi belleğine sahip ve ağırlıklı bağlantılar aracılığıyla birbirine bağlanan işlem elemanlarından oluşan paralel ve dağıtılmış bilgi işleme yapılarıdır. İnsan sinir ağlarını taklit eden bu yapılar, bilgisayar programları olarak karşımıza çıkar. Yapılan tanımlarda bazı ortak özellikler öne çıkar; en başta, yapay sinir ağları, hiyerarşik olarak birbirine bağlı ve paralel çalışabilen yapay sinir hücreleri, yani işlem elemanlarından oluşmasıdır. Bu elemanlar, birbirleriyle belirli bağlantılar aracılığıyla ilişkilendirilir ve her bağlantıya belirli bir ağırlık değeri atanır. Yapay sinir ağlarında bilgi, öğrenme yoluyla elde edilir ve süreç elemanlarının bağlantı değerlerinde saklanır, bu da dağıtık bir hafıza yapısının varlığına işaret eder. Proses elemanlarının bağlanmasıyla oluşan yapı, yapay sinir ağı olarak adlandırılır. Bu teknik, veri tabanlarındaki örüntüleri keşfetmek ve sınıflandırma ile tahmin amacıyla genelleştirmeler yapmak için kullanılır. Yapay sinir ağları algoritmaları, sayısal verilerle çalışır ve bu veriler üzerinden öğrenme süreci gerçekleştirilir (Albayrak, 2008: 73).

Yapay sinir ağlarının güçlü yönleri (Tüzüntürk, 2010: 79):

- Gürültü içeren büyük veri kümelerinde başarılı sonuçlar elde edilmesini sağlar.
- Hem sayısal hem de kategorik veriler üzerinde tahmin yapabilme yeteneğine sahiptir.
- Zaman faktörünün önemli olduğu veri kümesi analizlerinde kullanılabilir.
- Çeşitli alanlara kolayca uyarlanabilir ve esneklik gösterir.



Şekil 1.9.7. Biyolojik Sinir Hücresi İle Yapay Sinir Hücresi Karşılaştırması

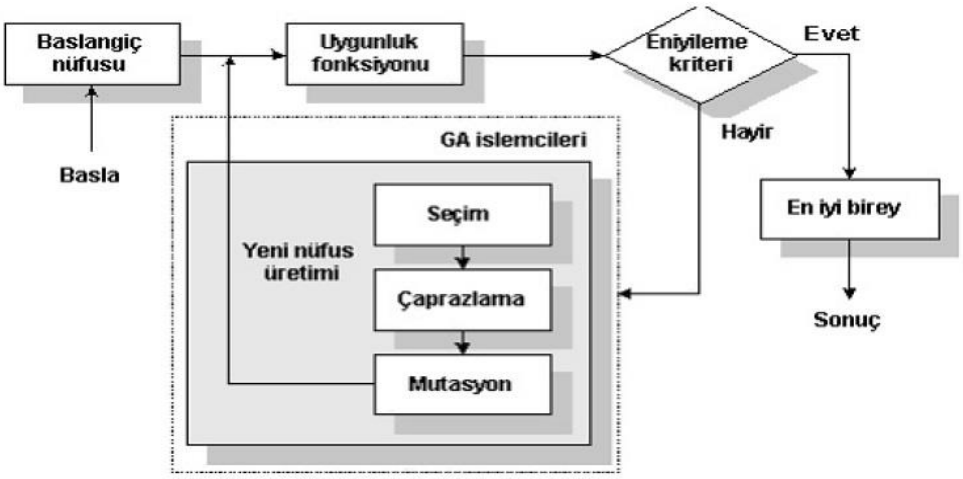
1.9.8. Genetik Algoritmalar

Genetik algoritmalar, doğal seçilim prensiplerine dayanan arama ve optimizasyon yöntemi olarak geliştirilmiştir. Bu algoritmaların temel prensipleri, John Holland tarafından ortaya konulmuştur ve sonrasında genetik algoritmalar üzerine birçok bilimsel araştırma yapılmış, uluslararası konferanslarda teorik ve uygulamalı yönleri kapsamlı olarak tartışılmıştır. Genetik algoritmalar, makine öğrenmesi, çizelgeleme, hücresel üretim, tasarım ve fonksiyon optimizasyonu gibi çeşitli alanlarda etkili uygulamalar sunmaktadır. Genetik algoritmalar, geleneksel optimizasyon yöntemlerinden farklı olarak, doğrudan parametreler üzerinde değil, bu parametrelerin kodlanmış biçimleri üzerinde çalışır. Genetik algoritmalar, olasılık kurallarını temel alarak çalışır, yalnızca amaç fonksiyonuna ihtiyaç duyar ve çözüm uzayının tamamını değil, yalnızca belirli bir bölümünü tarar. Bu, algoritmanın etkin bir arama gerçekleştirmesini ve çok daha kısa sürede çözüme ulaşmasını sağlar. Bir diğer önemli avantajı, çözümlerden oluşan bir popülasyonu eş zamanlı olarak inceleyebilmesi ve böylelikle yalnızca yerel en iyi çözümlere takılmadan daha kapsamlı bir optimizasyon sağlamasıdır (Emel ve Taşkın, 2002: 130).

Genetik algoritmalar, evrimsel hesaplama tekniğinin bir parçası olarak yapay zekanın hızla büyüyen bir alanını oluşturur. Darwin'in "doğada en iyinin hayatta kalması" ilkesinden esinlenerek geliştirilen genetik algoritmalar, geniş bir veri kümesinden belirli bir veri öbeğini bulmak amacıyla kullanılan güçlü bir arama yöntemidir. Bu algoritma, geleneksel yöntemlerle çözülmesi zor veya imkânsız olan

problemlerin çözümünde tercih edilir. Genetik algoritmaların temel yaklaşımı, çözüm sürecini bir evrimsel süreç gibi ele almak ve problemi sanal bir evrimden geçirmekten geçer. Bu yöntem, birçok farklı alanda başarılı bir şekilde kullanılmaktadır. Örneğin, Türkiye İkinci Futbol Ligi B Kategorisi'ndeki 51 takımın üç ayrı gruba dağıtılması probleminde genetik algoritma yöntemi uygulanmıştır. Bu uygulamada, takımlar arasındaki mesafelerin minimize edilerek yolculuk maliyetlerinin ve yol yorgunluğunun azaltılması, ayrıca sübjektif ayrımların önüne geçilmesi amaçlanmıştır. 2004-2005 futbol sezonu için yapılan bu çalışmanın sonucunda, genetik algoritmaların önerdiği planın Türkiye Futbol Federasyonu'nun mevcut plana kıyasla %10,5 daha az seyahat gerektirdiği tespit edilmiştir. Bu örnek, genetik algoritmaların optimizasyon süreçlerinde etkinliğini ortaya koymaktadır (Gürsoy, 2009: 47).

Genetik algoritmaların genel olarak kullanılan yapısı bu adımlardan oluşmaktadır. İşlemlerin sıralaması Şekil 1.9.8.'de gösterilmiştir.



Şekil 1.9.8. Genetik algoritmanın adımları

Genetik algoritmalar, karmaşık matematiksel işlemlerden ziyade yalnızca giriş ve çıkış bilgilerine ihtiyaç duyan bir yöntemdir. Bu nedenle, genellikle karmaşık problemlerin optimizasyonu için veri madenciliğinde sıkça tercih edilir. Genetik algoritmalar, tek bir çözüm üretmekle kalmaz; çeşitli çözümler üreterek parametrelerin en uygun sonuçlarını bulmaya çalışır. Bu sayede, birden fazla arama noktası üzerinden çok yönlü bir arama gerçekleştirir. Üretilen çözümler önceden tahmin edilemez niteliktedir ve bu da algoritmanın adaptasyon yeteneğini güçlendirir. Algoritma

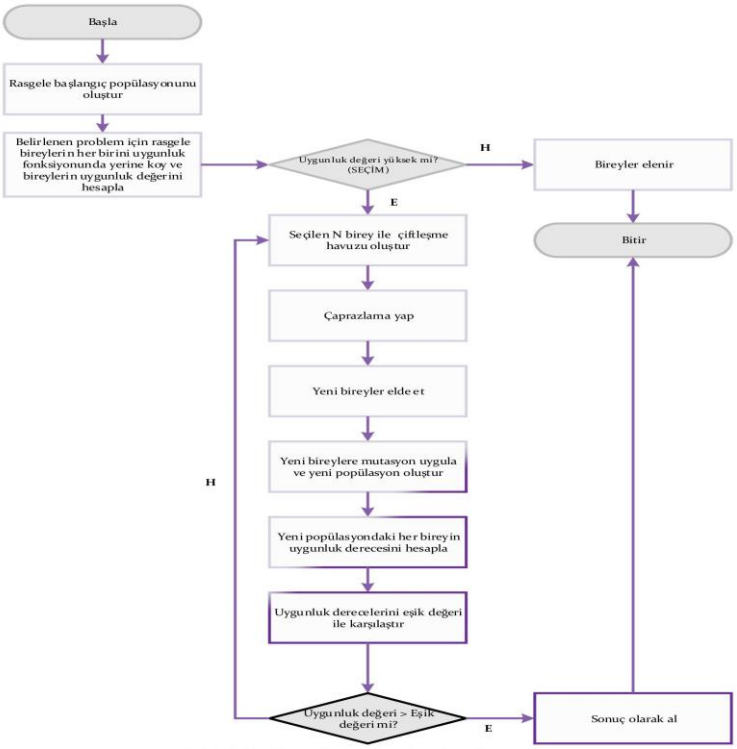
çalışırken, arama uzayındaki bireylerin uygunluk indekslerini hesaplamak için olasılıksal hesaplamalar yapar. Genetik algoritmaların avantajı, matematiksel modelleme ile doğrudan ilgilenmemesi, bunun yerine olasılık tabanlı bir yaklaşımla optimal çözümleri aramasıdır. Bu özellik, algoritmanın geniş bir uygulama yelpazesinde esnek ve etkili bir optimizasyon yöntemi olarak öne çıkmasını sağlar (Tuğ, 2005: 4).

Genetik algoritmalar, doğal seleksiyon ilkesini taklit eden bir bilgisayar uygulamasıdır ve evrim sürecinin bilgisayar ortamında modellenmiş halidir. Bu algoritmanın temel amacı, sorunları çözerek aşamalı sistemleri simüle etmektir. Genetik algoritmaların işleyişi şu adımlardan oluşur (Gülten ve Doğan, 2008: 12-16):

1. **Başlangıç Popülasyonu Oluşturmak:** Rastgele veya belirli kurallara göre başlangıç çözümlerinden oluşan bir popülasyon oluşturulur. Bu popülasyon, problemin çözüm alanındaki bireyleri temsil eder.
2. **Seçim:** En yüksek uygunluk değerine sahip bireyler seçilir ve bir sonraki nesli oluşturmak için kullanılır. Seçilim süreci, en uygun çözümleri bir araya getirerek daha iyi çözümler üretmeyi hedefler.
3. **Çaprazlama:** Seçilen bireyler arasında gen değişimi yapılarak yeni bireyler (çocuklar) oluşturulur. Bu işlem, ebeveyn çözümlerinin özelliklerini birleştirerek yeni çözümler üretir.
4. **Mutasyon:** Yeni bireylerin bazı genlerinde rastgele değişiklikler yapılır. Mutasyon, çözüm alanının tamamının keşfedilmesine olanak sağlar ve algoritmanın yerel optimumlara takılmasını önler.

Yeni popülasyondaki bireylerin uygunluk değerleri hesaplanır ve bu değerlere göre seçilim yapılır. Genetik algoritmada, uygunluk değeri belirlenen eşik değerin altında kalan bireyler, genellikle problem için en iyi çözümleri üretme potansiyeline sahip bireyler olarak kabul edilir ve sonraki nesillerde daha fazla kullanılmak üzere seçilirler. Bu süreç, algoritmanın adım adım daha iyi çözümler bulmasını sağlar. Şekil 1.9.8.'de genetik algoritmaların akış diyagramını göstermekte olup, algoritmanın başlangıç popülasyonunun oluşturulmasından itibaren uygunluk değerlendirmesi, seçilim, çaprazlama, mutasyon ve yeni popülasyonun oluşturulması gibi temel adımları içeren

genel işleyişini ortaya koyar. Bu akış diyagramı, algoritmanın tekrarlayan süreçlerle daha iyi çözümler elde etmesini sağlayan döngüsel yapısını özetler.

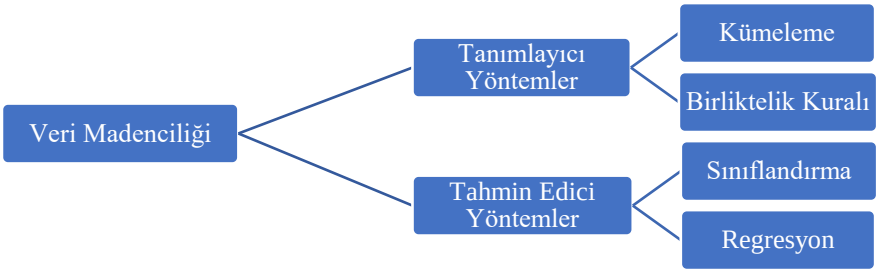


Şekil 1.9.8. Genetik algoritmalar akış diyagramı

1.10. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri tanımlayıcı ve tahmin edici modeller olmak üzere iki temel unsur altında incelenir. Tahmin edici modellerde, sonuçları belli olan verilerden yola çıkarak bir model oluşturulur ve bu model, sonuçları henüz bilinmeyen veri kümelerine uygulanarak tahmin yapılmasını sağlar. Bu modeller, özellikle gelecekteki olayları veya eğilimleri öngörmek için tercih edilir. Örneğin, kredi risk değerlendirmesi, satış tahminleri ve müşteri davranışlarının öngörülmesi gibi alanlarda tahmin edici modeller yaygın olarak kullanılmaktadır. Tanımlayıcı modellerin amacı ise veri kümesi içinde gizli kalmış örüntüleri, ilişkileri ve yapıları keşfetmektir. Bu modeller sayesinde veri içindeki anlamlı ilişkiler bulunarak veriyi daha iyi anlamak ve kategorize etmek mümkün hale gelir. Kümeleme, ilişkilendirme kuralları ve pazar sepeti analizi gibi teknikler tanımlayıcı modellemeye örnek teşkil eder ve bu tür analizler, veriyi gruplamak, veri kümeleri arasında bağlantılar bulmak veya ilişkiler ağını ortaya çıkarmak amacıyla kullanılır. Veri madenciliğinde öngörücü modeller,

geleceğe dair tahminler sağlarken, tanımlayıcı modeller mevcut veriyi derinlemesine analiz ederek verinin yapısını ve ilişkilerini anlamaya katkıda bulunur. Bu iki model türü, veri madenciliğinin geniş bir uygulama alanında anlamlı ve stratejik bilgiler üretmesini sağlar (Kaya ve Köymen, 2008: 161).



Şekil 1.10. Veri Madenciliği Yöntemleri

Veri tabanında mevcut durumun genel yapısını ortaya çıkarmaya yönelik yöntemler, veri madenciliğinde tanımlayıcı modelleme yöntemlerini öne çıkarır. Bu modelleme yaklaşımı, eldeki verilerden örüntüler tanımlayarak karar destek süreçlerine katkıda bulunur. Örneğin, gelir düzeyi X-Y aralığında olan, evi ve arabası bulunan, çocukları okul çağında olan ailelerin satın alma davranışlarının; çocukları olmayan ve geliri X-Y aralığından düşük olan ailelerle benzerlik gösterdiğini belirlemek, tanımlayıcı modelleme yöntemlerine bir örnektir. Bu tür analizler, farklı müşteri grupları arasındaki örüntüleri tanımlayarak, pazarlama stratejilerinde kullanılabilecek anlamlı bilgiler sunar (Farboudi, 2009: 46).

Tahmin edici modellerde veriler, geleceğe yönelik öngörülerde bulunmak ve sonuçlar hakkında çıkarımlar yapmak amacıyla kullanılır. Bu modellerde, sonuçları bilinen verilerden yola çıkılarak bir model geliştirilir ve oluşturulan bu model, sonuçları henüz bilinmeyen veri kümelerinde tahmin yapmaya yönelik uygulanır. Örneğin, bir banka geçmişte vermiş olduğu kredilere dair müşterilerin özellikleri, kredi tutarları ve geri ödeme bilgilerini içeren verilere sahip olabilir. Bu verilerde bağımsız değişkenler, kredi alan müşterinin özellikleri olurken, bağımlı değişken kredinin geri ödenip ödenmediği bilgisidir. Bu verilere dayanarak kurulan bir model, gelecekte yapılacak kredi başvurularında müşterinin özelliklerine göre kredinin geri ödenme olasılığını tahmin etmek için kullanılabilir. Tahmin edici modeller, bu tür

örneklerle işletmelere risk yönetimi, müşteri analizi ve stratejik karar alma süreçlerinde değerli bir araç sunar (Farboudi, 2009: 46).

1.10.1. Sınıflandırma

Sınıflandırma, veri madenciliğinin en temel işlevlerinden biri olup, kategorik sonuçları olasılıkla öngörmek için kullanılır. Bu modelin oluşturulabilmesi için önceden tahmin edilen durumlara ve bu durumlara karşılık gelen işlevsel değerlere ihtiyaç vardır; bu verilere "eğitim verisi" adı verilir. Sınıflandırma sonuçları, belirli bir olasılıkla tahmin edilir ve örneğin "Müşteri %80 olasılıkla bu kampanyaya olumlu yanıt verecek" gibi sunulur. Sonuçlar, iki seçenekli olarak "Hizmeti Devam Ettirir-Hizmeti Sonlandırır" şeklinde olabileceği gibi, "Mutlaka Tercih Eder-Tercih Eder-Cevap Vermez-Tercih Etmez-Asla Tercih Etmez" gibi çeşitli alternatifleri de içerebilir. Modelin doğruluğunu belirlemek için bir "deneme kümesi" kullanılır. Verilen veri kümesi genellikle "öğrenme kümesi" ve "deneme kümesi" olmak üzere ikiye ayrılır. Öğrenme kümesi, modeli oluşturmak için, deneme kümesi ise modelin doğruluğunu test etmek için kullanılır. Örneğin, bir otomobil satış firması geçmiş müşteri davranışlarını analiz ederek belirli sınıflandırma kuralları keşfederse, genç kadınların okuduğu bir dergide küçük model otomobillerinin reklamını yapmayı tercih edebilir. Bu tür sınıflandırma analizleri, pazarlama stratejilerini özelleştirmeye ve hedef kitlenin ihtiyaçlarına daha uygun kararlar almaya yardımcı olur (Argüden ve Erşahin, 2008: 37).

Veri madenciliğinde sınıflandırma, neredeyse her alanda kullanılabilen çok yönlü bir yöntemdir. Hangi sınıfa ait olduğu bilinmeyen veya tahmin edilemeyen kayıtların sınıf sınırlarını belirlemek için sınıflandırma algoritmalarına ihtiyaç duyulur. Bu nedenle sınıflandırma, oranlayıcı bir model olarak kabul edilir. Sınıflandırmanın temel amacı, verilerin içerdiği ortak özellikleri kullanarak bu verileri farklı sınıflara ayrıştırmaktır. Sınıflandırma algoritmaları, bu hedef doğrultusunda verileri analiz eder ve bir sınıflandırma modeli oluşturarak her bir veriyi uygun sınıfa yerleştirmeye çalışır (Özkan, 2013: 45).

Sınıflandırma, ürün ve müşteri özelliklerini analiz ederek bu iki değişkenin en uygun biçimde eşleştirilmesine olanak tanır. Bu şekilde, müşteri için en uygun ürün ve ürün için en uygun müşteri belirlenebilir. Örneğin, bir otomobil satıcısı, geçmiş müşteri davranışlarını analiz ederek "genç kadınlar küçük otomobil satın alır; yaşlı ve varlıklı erkekler ise büyük ve lüks otomobil tercih eder" gibi bir sınıflandırma kuralı geliştirebilir. Bu durumda, genç kadınların okuduğu bir dergide küçük bir otomobil modelinin reklamını yapmak mantıklı olacaktır. Bu yaklaşım, sınıflandırmanın pazarlama stratejilerini müşteri özelliklerine göre uyarlayarak daha etkili hale getirmesine olanak tanır (Alpaydın, 2000: 2).

1.10.2. Regresyon

Tekrar eden değerlerin tahmini için kullanılan regresyon analizi, girdiler ile çıktılar arasındaki ilişkiyi tanımlayan bir model oluşturarak en iyi tahmini yapmayı amaçlar. Bu yöntemde, tahmin edilecek sonuç "bağımlı değişken" olarak adlandırılırken, girdiler "bağımsız değişkenler" olarak bilinir. Sonucun alacağı değerler genellikle bir güven aralığı içinde sunulur ve bu aralık, tahminin olası değişkenliğini gösterir. Girdiler, çözülecek probleme bağlı olarak tek bir değişkenden veya birden fazla değişkenden oluşabilir. Örneğin, bir inşaat firması, konut satışlarının bulunduğu bölgedeki toplam gelirle bağlantılı olabileceğini düşünebilir. Bu durumda, bölgesel gelire dayalı bir model geliştirerek, gelirin değişimine göre satabileceği konut sayısını tahmin edebilir. Fakat gerçek yaşamda çoğu tahmin probleminin doğruluğunu artırmak için birden fazla girdiye ihtiyaç duyulur. Burada önemli olan, her bir girdinin sonuca sağladığı katkıyı doğru değerlendirmektir. Sonuca katkısı sınırlı olan girdileri modelden çıkarmak, daha etkin ve doğru bir model oluşturulmasına yardımcı olabilir. Bu şekilde, regresyon analizi doğru ve verimli tahminler yapmada güçlü bir araç haline gelir (Argüden ve Erşahin, 2008: 38).

Regresyon analizi, bağımlı bir değişken ile bir veya birden fazla bağımsız değişken arasındaki ilişkiyi matematiksel bir fonksiyonla ifade eder ve bu fonksiyon aracılığıyla bağımlı değişkenin gelecekteki değeri tahmin edilebilir. Örneğin, spor araba aksesuarlarının satış hacmi, önceki ay satılan spor arabaların sayısına göre öngörülebilir. Aynı şekilde, bir iş yerindeki çalışanların ücretlerini etkileyen eğitim

seviyesi, cinsiyet, yaş, ve deneyim gibi faktörlerin etkisini incelemek için de regresyon analizi kullanılabilir (Orhunbilge, 2002: 9).

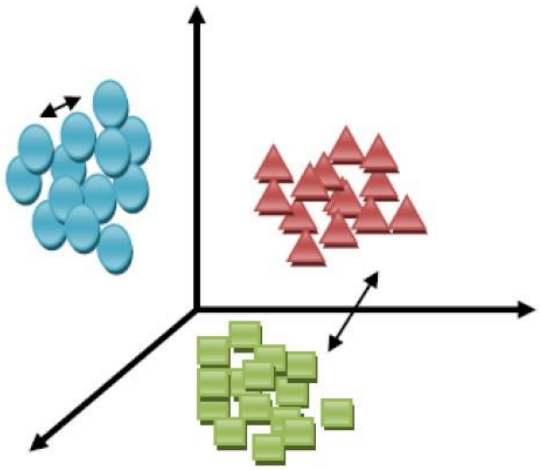
1.10.3. Kümeleme

Veriler arasındaki benzerlikler göz önüne alınarak ayrıştırılma işlemlerine kümeleme denir. Bu özellikleri göz önüne alındığında birçok alanda kümeleme işlemleri kullanılabilir (Özkan, 2013: 47). Kümeleme analizinde, veriler incelenerek birbirine benzeyenler aynı kümeye, benzemeyenler ise farklı kümelere ayrılır. Bu sınıflama sayesinde veriler, araştırmacı için daha uygun ve kullanışlı özet bilgilere dönüştürülür.

Kümeleme işlemlerim genelde bölümleme sorunlarını çözmek için kullanılır. Kümelemenin asıl amacı ise; büyük veriler için tanımlayıcı olmaları, araştırma yapılacak bir çokluğu azaltmak, büyük veri depolarında tanımlayıcı belirgin verileri belirlemek ve kümelerin dışında kalan istisnai durumları da tanımlamak denilebilir. En başta verilerin hangi kümeleri ayrılacağı belli olmayacaktır. Böyle durumlarda hangi değişken özelliklere göre kimlere ayrılacağı belirlenir ve konunun uzmanı olan kişiler tarafından kümelerin neler olabileceği tahmin edilmektedir. Kümeleme ve sınıflandırma işlemleri arasında bazı önemli farklar bulunmaktadır. Bu kümeleme ve sınıflandırma arasındaki farklar ise kümelerinin önceden belirlenmiş bir özelliğinin bulunmamasıdır. Sınıflandırmada ise girdiler tanımlı ve geçmişte aldıkları değerler ve temel modelleri ile oluşturulur. Kümeleme işlevlerinde önceden örnekler ve tanımlanmış girdiler yoktur. Kendi içlerindeki benzerliklerine göre gruplandırılır. Kümelerin belirlenmesinde uzmanlar benzerliği tanımlayacak boyutlar ve özellikler modelini oluştururlar (Argüden ve Erşahin, 2008: 40).

Kümeleme modellerinin temel amacı, üyeleri birbirine oldukça benzeyen ancak özellikleri diğer gruplardan farklı olan kümeler oluşturarak veri tabanındaki kayıtları bu kümelere ayırmaktır. Kümeleme analizinde, veri tabanındaki kayıtların nasıl gruplandırılacağı veya kümeleme işleminin hangi özelliklere göre gerçekleştirileceği, konu uzmanının rehberliğinde kararlaştırılabilir. Aynı zamanda, geliştirilen bilgisayar programları da bu ayrıştırmayı otomatik olarak gerçekleştirebilir. Bu süreç, veriyi daha

anlamlı gruplara ayırarak analizlerde kullanıma hazır hale getirir ve farklı özelliklere sahip kayıtların bir arada incelenmesini sağlar (Ayık ve ark, 2007: 447).



Şekil 1.10.3. Verilerin kümelenmesi; Küme içi ve kümeler arası uzaklıklar

Veri kümeleme analizinde benzerlik, iki veri arasındaki mesafenin ölçülmesi ve değerlendirilmesi yoluyla belirlenir. Bu değerlendirme, iki verinin veri tabanındaki diğer verilere kıyasla ne kadar yakın veya benzer olduğunu gösterdiği gibi, belirli kısıtlar veya eşik değerleri çerçevesinde de yapılabilir. Kümeleme analizinde, nesneler belirgin özelliklerine göre gruplara ayrılır; benzer nesneler aynı küme içinde yer alırken, farklı özelliklere sahip nesneler farklı kümelere ayrılır. Başarılı bir kümeleme analizi sonucu, bir geometrik çizimde nesneler kendi kümeleri içinde birbirine yakın, farklı kümeler ise birbirinden uzak konumda görülür. Kümeleme analizi, birçok disiplinde yaygın olarak kullanılmaktadır; istatistik, bilgisayar bilimleri ve matematik bu alanların başında gelir. Bilgisayar bilimlerinde ses, karakter ve görüntü tanıma ile makine öğrenmesi çalışmaları sıklıkla kümeleme analizine dayanır. Ayrıca, web sitelerinin içerik benzerliklerine göre gruplandırılması ve benzer grupların birbirleriyle ilişkilendirilmesi, arama sonuçlarında daha alakalı ve ilişkilendirilmiş sonuçların sunulmasına olanak tanımaktadır. Bu yönleriyle kümeleme analizi, farklı veri türlerini anlamlandırmada ve analiz etmede güçlü bir araç sunar (Elmas, 2014: 14).

1.10.4. Birliktelik Kuralları

Veri tabanında bulunan kayıtlar arasındaki ilişkileri inceleyerek, hangi olayların birbiriyle uyum içinde gerçekleşme olasılığı olduğunu belirlemeye çalışan veri madenciliği yöntemlerinden biri de birliktelik kurallarıdır. Bu kurallar, özellikle pazarlama alanında geniş uygulama alanı bulur ve "pazar sepet analizleri" olarak bilinen yöntemlere temel oluşturur. Bu tür analizlerle, müşterilerin alışveriş alışkanlıkları ve satın alma eğilimleri belirlenir. Pazar sepet analizleri sayesinde, bir müşterinin belirli bir ürünü satın aldığında, sepetine başka hangi ürünleri ekleme olasılığının yüksek olduğu tespit edilir. Bu bilgi, pazarlama stratejilerinin geliştirilmesinde kullanılarak çapraz satış fırsatlarını artırır ve müşteri davranışlarını daha iyi anlamaya olanak tanır (Özkan, 2013: 49).

Birliktelik analizi, makro veri kümeleri içindeki farklı veriler arasında ilişkileri ve uyumlu birliktelikleri keşfetme işlemidir. Bu analiz, belirli bir veri kümesinde sıkça bir arada görülen özellik değerlerine ait ilişkisel kuralları ortaya çıkarır. Elde edilen sonuçlar, "birliktelik kuralları" olarak sunulur ve özellikle şirketlerin karar alma süreçlerini daha verimli hale getirir. Bu kurallar, hangi veri öğelerinin birlikte ortaya çıktığını belirleyerek pazarlama stratejileri, stok yönetimi ve müşteri alışkanlıklarının analizinde kritik bilgiler sunar (Argüden ve Erşahin, 2008: 41).

Müşterinin tek bir alışverişte veya ardışık alışverişlerde hangi ürün veya hizmetleri satın almaya istekli olduğunu tespit etmek, satışları artırmanın etkili yöntemlerinden biridir. Müşteri satın alma eğilimlerini tanımlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama alanında "pazar sepeti analizi" olarak bilinen veri madenciliği tekniğiyle yaygın bir şekilde kullanılır. Bu analiz, müşterinin bir ürünü satın aldığında yanına hangi ürünleri ekleme olasılığının yüksek olduğunu ortaya koyarak çapraz satış stratejilerini destekler. Bunun yanı sıra, birliktelik kuralları ve ardışık örüntüler sadece pazarlama alanında değil; aynı zamanda tıp, finans gibi birbirine bağlı olayların analiz edildiği birçok alanda da önemli bilgiler sağlar. Bu teknikler, ilişkili verilerin tespit edilmesiyle sağlıkta teşhis süreçlerini, finansal risk analizlerini ve diğer önemli karar destek süreçlerini daha güçlü hale getirir (Ayık ve ark, 2007: 447).

Birliktelik kuralı analizi, gözlem değerleri arasındaki ilişkileri koşullu olasılıklar temelinde özetler ve yalnızca uygulayıcı tarafından belirlenen başarı oranını aşan kuralları sıralar. Bu analiz, hesaplama yapısı sayesinde hızlı sonuç verir ve çok büyük veri setlerine kolayca uygulanabilir; bu özellikleri, birliktelik kuralı analizini ticari veri tabanlarında giderek daha popüler bir veri madenciliği aracı haline getirmiştir. Özellikle geniş veri kümelerinde verimli çalışabilmesi, pazarlama ve perakende alanlarında tercih edilmesine katkıda bulunur (Koyuncugil ve Özgülbaş, 2009: 27).

1.11. Veri Madenciliği İşlevler

Veri madenciliği iki ana işlevden oluşur: tanımlama ve tahmin. Tahmin edici işlevler, sonuçları bilinen verilerden hareketle, henüz sonuçları bilinmeyen veriler için öngörülerde bulunabilecek modellerin oluşturulmasını sağlar. Örneğin, bir hastanede belirli bir hastalığa yönelik bir veri setinin olduğunu düşünelim. Veri madenciliği teknikleriyle, geçmiş vakalardan toplanan tıbbi veriler ve hasta durum bilgileri kullanılarak bu hastalığa dair bir tahmin modeli oluşturulabilir. Bu model, hastaneye yeni başvuran bir hastanın testler sonucu elde edilen tıbbi verileri üzerinden, ilgili hastalığa dair bir tahmin yapılmasını mümkün kılar. Bu sayede, geçmiş verilerden öğrenilen bilgiler doğrultusunda yeni vakalar için olasılıklar tahmin edilerek erken müdahale veya uygun tedavi süreçlerine destek sağlanabilir.

Tanımlama fonksiyonlarının amacı, belirli bir hedefin tahmin edilmesi değil, veri setinde bulunan veriler arasındaki ilişkilerin, bağlantıların ve davranış kalıplarının keşfedilmesidir. Bu fonksiyonlar, mevcut verileri yorumlayarak belirli davranış biçimlerini tespit etmeyi ve bu davranışı sergileyen alt veri setlerinin özelliklerini tanımlamayı hedefler. Tanımlama işlemi sayesinde, belirli bir tanıma uyan tekrar eden faaliyetlerde veya tanımı bilinen yeni bir verinin mevcut yapıya dahil edilmesinde nasıl bir yol izlenmesi gerektiği konusunda karar almak kolaylaşır. Bu süreç, verinin yapısını anlamak, kategoriler belirlemek ve veri kümeleri arasındaki ilişkileri analiz ederek stratejik kararları desteklemek için temel bir rol oynar (Argüden ve Erşahin, 2008: 39).

Veri madenciliği, işlemsel açıdan üç ana sınıfa ayrılabilir: keşif işlemleri, tahmini modeller ve analizler. Keşif işlemleri, belirli bir hipotez olmadan, veri tabanında gizlenmiş desenleri arama sürecini ifade eder. Kullanıcılar büyük ve karmaşık veri tabanlarında düşünülmesi zor olan, öngörülmeyen birçok gizli deseni keşfetme fırsatına sahip olur. Bu süreçte asıl hedef, veri tabanındaki desenlerin zenginliği ve bu desenlerden elde edilebilecek bilgilerin kalitesidir. Tahmini modellemede ise veri tabanında çıkarılan desenler geleceğe yönelik tahminler yapmak amacıyla kullanılır. Bu model, kullanıcının bazı alan bilgilerini bilmemesi durumunda bile tahmin yoluyla verileri doldurmasına olanak tanır. Sistem, mevcut kayıtlara dayanarak eksik bilgileri tahmin eder. Keşif işlemleri verideki desenleri bulmaya odaklanırken, tahmini modelleme bu desenleri yeni veri nesnelerini tahmin etmek için kullanır. Adli analiz, normal veya sıradan olmayan veri elemanlarını belirlemek amacıyla çıkarılan desenleri kullanır. Anormal veya sıra dışı unsurları tespit etmek için önce verinin genel ve sıradan yapısını tanımlamak gereklidir. Bu üç veri madenciliği işlevi, gizli desenleri keşfetmek, geleceğe yönelik tahminlerde bulunmak ve sıra dışı verileri belirlemek için temel yaklaşımlar sağlar (Ayık ve ark, 2007: 448).

1.12. Veri Madenciliği ve OLAP

İlişkisel veri tabanlarının yaygınlaşması ve ardından Veri Ambarlarının gelişmesi, verilere hızlı erişim ve çok boyutlu analiz ihtiyacını ortaya çıkarmış; bu da bilim insanlarını ve yazılım firmalarını yeni veri işleme yapıları geliştirmeye yöneltmiştir. Bu bağlamda geliştirilen Çevrimiçi Analitik İşleme (OLAP) teknolojisi, ilişkisel veri tabanları gibi bilimsel temellere değil, daha çok OLAP ürünleri geliştiren firmaların desteklediği bir teknoloji olarak öne çıkmıştır. Bu nedenle, ilişkisel veri tabanları ve veri ambarları üzerine kapsamlı akademik çalışmalar bulunmasına rağmen, OLAP hakkında genellikle ürün belgeleri ve şirket tanıtım yazıları yer almaktadır. OLAP terimi, ilk kez 1993 yılında Dr. E.F. Codd'un belirlediği kurallar doğrultusunda kullanılmış ve o tarihten itibaren iş zekası alanında çok boyutlu veri analizi için temel bir yapı olarak önem kazanmıştır (Gürsoy, 2009: 11-12).

OLAP (Çevrimiçi Analitik İşleme) terimi, veri tabanı veya veri ambarlarındaki verilerin analizinde kullanılan çeşitli sorgu odaklı analiz türlerini tanımlamak için

kullanılır. OLAP, veriyi farklı boyutlar veya perspektiflerden analiz etme ve görüntüleme imkanı sağlar. OLAP'ın, veri ambarlarındaki özetlenmiş veriyi çoklu ve dinamik görünümde sunabilme yeteneği, veri madenciliği için güçlü bir temel oluşturur. Bu nedenle, OLAP ve veri madenciliği genellikle birbirini tamamlayan araçlar olarak kabul edilir. OLAP, hareket işlemeye değil, sorgulama ve raporlama işlemlerine yönelik optimize edilmiş bir veri tabanı teknolojisidir ve iş zekası uygulamaları için güçlü bir sorgulama altyapısı sağlar. OLAP verileri, tablolar yerine küp biçiminde çok boyutlu yapılarla hiyerarşik olarak düzenlenir. Bu çok boyutlu yapılar, verilere hızlı erişim sağlayarak karmaşık veri analizlerini kolaylaştırır. Örneğin, iş zekası uygulamalarında özet tablo veya grafik raporları, tüm ülke veya bölgedeki toplam satışları görüntülemenin yanı sıra, belirli bölgelerdeki yüksek veya düşük satış verilerini de ayrıntılı olarak analiz edebilir. Bu özellik, karar vericilerin veriye daha hızlı ve derinlemesine ulaşarak stratejik analiz yapmalarına olanak tanır (Şık, 2014: 8).

Veri ambarları, modern iş zekası (Business Intelligence, BI) ve analitik uygulamalar için entegre, temiz ve tutarlı veri sağlamasıyla kritik bir bileşen olmaya devam etmektedir. Veri ambarlarında baskın olan veri gösterimi, önemli iş olaylarını (gerçekleri) çok boyutlu bir yapıda, granülerlik düzeylerinin hiyerarşik olarak düzenlendiği bir modelle sunan çok boyutlu modeldir. Bu modeller, verilerin farklı perspektiflerden analiz edilmesine olanak tanıyan çevrimiçi analitik işleme (Online Analytical Processing, OLAP) yöntemiyle küpler üzerinde analitik sorguların geliştirilmesini ve yürütülmesini mümkün kılar. Bu yaklaşım, kullanıcıların verileri çok yönlü bir şekilde inceleyerek stratejik karar alma süreçlerini desteklemelerine olanak tanır (Kovacic, 2022).

OLAP, ilişkisel veri tabanlarındaki verilere hızlı erişim sağlayarak çok boyutlu analiz ihtiyaçlarını karşılamak için geliştirilmiş bir teknolojidir. Bu yönüyle kavramsal olarak veri madenciliğine benzetilse de, aralarındaki farklar uzmanlar için bazen kafa karıştırıcı olabilir. OLAP, veri tabanları üzerinde stratejik kararlar almaya yardımcı olan analiz ve sorgu işlemlerine odaklanır ve genellikle tümdengelimsel bir yaklaşım izler. Geleneksel sorgulama ve raporlama araçları "Ne?" sorusuna yanıt ararken,

OLAP bir adım daha ileri giderek “Niçin?” sorusunu da yanıtlamaya çalışır. Ancak, incelenmesi gereken değişken ve parametre sayısı arttığında, OLAP ile etkili hipotezler üretmek ve doğrulamak daha zor hale gelir. Veri madenciliği ise OLAP'tan farklı olarak, tahmin edilemeyen ve gözle görülmeyen örüntü ve ilişkileri keşfetmeye yönelik bir süreçtir ve veriye tümevarımsal bir bakış açısı getirir. Veri madenciliği, hipotezler oluşturmak yerine, doğrudan veri üzerinden örüntüleri ve ilişkileri ortaya çıkarmaya odaklanır. Bu açıdan, veri madenciliği bilinmeyen ilişkileri keşfetmek için veriyi analiz ederek tahminsel modeller oluştururken, OLAP stratejik sorgulama ile var olan hipotezleri doğrulamayı amaçlar. Böylece OLAP, veri analizinde tümdengelimsel bir yaklaşım sunarken, veri madenciliği tümevarımsal bir yöntem izler (Kocabaş, 2010: 11).

OLAP veri tabanının sahip olduğu özellikler şunlardır (Şık, 2014: 9):

- **Çok yönlü inceleme**: Veriyi farklı açılardan analiz edebilme yeteneği sunar.
- **Şeffaflık**: Kullanıcıların veriyi daha açık ve anlaşılır şekilde görmesini sağlar.
- **Erişilebilirlik**: Verilere hızlı ve kolay erişim sağlar.
- **Her seviyede performans**: Farklı sorgu seviyelerinde aynı yüksek performansı sunar.
- **İstemci-sunucu yapısı**: İstemci-sunucu tabanlı bir mimariyi destekler.
- **Sınırsız çapraz raporlama**: Farklı boyut ve kategoriler arasında sınırsız raporlama olanağı sunar.
- **Otomatik veri ayarı**: En alt seviyedeki verilerin otomatik olarak güncellenmesini sağlar.
- **Boyutlandırılabilirlik**: Farklı koşullara uygun boyutlandırma yapılabilir.
- **Çoklu kullanıcı desteği**: Aynı anda birden fazla kullanıcıya hizmet verebilir.
- **Veri değiştirilebilirliği**: Tüm seviyelerde verilerin düzenlenebilir olması.
- **Esnek raporlama**: Kullanıcılara ihtiyaçlarına göre farklı rapor formatları sunar.

- **Boyut ve gruplamalarda sınırsızlık**: Veriler için boyut ve grup sayısında herhangi bir sınır olmadan işlem yapma olanağı tanır.

1.13. Veri Madenciliğinde Sorunlar

Son yıllarda birçok kuruluş, birkaç milyon kaydı aşan ve anlık internet veri akışıyla sürekli büyüyen çok büyük veri tabanlarına sahip hale gelmiştir. Bu devasa veri yığınları, büyük fırsatlar sunduğu kadar ciddi zorluklar da beraberinde getirir. Veri madenciliği sistemleri, işlenmemiş veriyi barındırma eğiliminde olan, dinamik, tamamlanmamış, gürültü içeren ve büyük ölçekli veri tabanlarına dayanır. Ancak, veri madenciliği işlemleri için kullanılan veri tabanlarının eksiksiz, geniş, net ve analiz konusu ile uyumlu olmaması durumunda sorunlar ortaya çıkabilir. Veri yetersizliği, ilgisiz bilgi veya konuyla uyumsuz veriler, analiz süreçlerinde doğru sonuçlara ulaşmayı zorlaştırır ve veri madenciliği süreçlerinin verimliliğini düşürebilir (Ertuğrul ve ark, 2013: 99).

Günümüzde veri madenciliği sistemlerinin karşılaştığı sorunlar şu şekildedir:

1.13.1. Veri Tabanı Boyutu

Büyük veri kümeleri genellikle eksik, kirli ve hatalı veri noktalarını içerir; bu tür hatalardan tamamen arındırılmış veri kümeleri nadiren görülür. Veri kümesinin boyutu, analiz sürecinde zorluklar yaratırken, standart istatistiksel yöntemlerde sıkça karşılaşılmayan bazı özelliklerin ortaya çıkmasına neden olabilir. Veri madenciliğinde kullanılan veriler, çoğunlukla doğrudan veri madenciliği amacıyla değil, farklı amaçlar doğrultusunda toplanmaktadır. Bunun aksine, istatistiksel çalışmalarda veriler genellikle belirli sorulara yanıt bulmak amacıyla toplanır ve analiz edilir. İstatistik bilimi, deney tasarımı ve alan araştırması gibi alt disiplinleri kapsar; bu disiplinler, veri toplamanın en uygun yöntemleri konusunda önemli rehberlik sağlar (Oğuzlar, 2003:13).

Veri tabanlarının boyutları hızla artmakta ve bu durum, veri işleme yöntemlerini önemli ölçüde etkilemektedir. İlk olarak yalnızca birkaç yüz kayıtla çalışan makine

öğrenimi algoritmaları, günümüzde milyonlarca kaydı içeren büyük veri setlerini işlemek durumundadır. Bu büyük veri setleri, gözlemlenen desenlerin doğruluğunu artırmak için faydalı bir kaynak sunarken, aynı zamanda çıkarılabilecek örüntülerin sayısını da önemli ölçüde artırarak analiz süreçlerini karmaşık hale getirmektedir. Veri madenciliği sistemleri için büyük ölçekli veri tabanlarıyla başa çıkmak, işleme, depolama ve analiz süreçlerinde ciddi altyapı ve kaynak gereksinimlerini beraberinde getirir. Bu durum, algoritmaların etkinliğini ve performansını doğrudan etkileyerek büyük veri ile çalışmayı stratejik bir zorluk haline getirmektedir.

1.13.2. Gürültülü Veri

Gürültülü veri, veri girişlerinde ya da toplanması aşamasında oluşabilen sistem dışı hatalardır. Şayet veri kümesi gürültülü ise ihmal etmeli ve anlamsız veriyi tanımalıdır. Son zamanlarda kullanılan ticari ilişkisel veri tabanlarında veri girişi sırasında oluşan hatalar otomatik giderilerek bu yaşanan sıkıntıyı bir nebze azaltmıştır. Hatalı gürültülü verileri ise dünya veri tabanlarında ciddi problemler oluşturmaktadır (Çalışkan, 2006).

Verilerdeki gürültü, ölçülmüş bir özelliğin rastgele hatası veya varyansı olarak tanımlanır. Gürültü miktarına bağlı olarak, bu durum bilgi keşfi sürecinde önemli bir problem haline gelebilir ve sürecin doğruluğunu tehlikeye atabilir. Verilerdeki gürültünün etkisi, veri giriş aşamasında anormallikleri tespit etmek için özelliklere kısıtlar uygulanarak önlenebilir. Gürültü oluşmuşsa, bu durum elle kontrol, ambarlama veya kümeleme gibi yöntemlerle, önceden belirlenmiş kısıtlar kullanılarak giderilebilir (Cios ve diğerleri, 2007: 40).

Eğer veri kümesi gürültü içeriyorsa, sistemin bozuk veriyi tanıyıp ihmal etmesi önemlidir. Quinlan, gürültünün sınıflama performansı üzerindeki etkisini incelemek için bir dizi deney gerçekleştirmiştir. Bu deneylerin sonuçları, etiketli öğrenme sürecinde etiket üzerindeki gürültünün öğrenme algoritmasının performansını doğrudan olumsuz etkilediğini göstermiştir. Bununla birlikte, eğitim kümesinde yer alan nesnelerin özellikleri veya nitelikleri üzerindeki gürültünün, %10'luk bir seviyeye kadar temizlenebileceği bulunmuştur. Bu durum, belirli bir seviyeye kadar olan gürültünün ayıklanmasının sınıflandırma performansını koruyabileceğini ortaya

koyar. Ancak gürültü oranının daha yüksek seviyelere ulaşması, modelin doğruluğunu ve güvenilirliğini düşürebilir, bu nedenle veri temizleme adımları büyük önem taşır (Quinlan, 1986: 85).

1.13.3. Eksik Veri

Eksik veri (missing data), bir veri setinde belirli bir değişkene ilişkin olması beklenen bilginin mevcut olmaması durumunu ifade eder. Veri analizi ve modelleme süreçlerinde, eksik veriler yaygın bir sorun teşkil eder ve genellikle verinin toplanması, ölçülmesi ya da kaydedilmesi sırasında meydana gelir.

Evrendeki her nesnenin ayrıntılı olarak tanımlandığı ve alabileceği değerler kümesinin sabit olduğu varsayılırsa, sınıflama işlemi bu nesnelerin alt kümeleri kullanılarak basit bir şekilde gerçekleştirilebilirdi. Ancak, veriler çoğunlukla kurumların ihtiyaçlarına göre düzenlenip toplandığı için, mevcut veri seti gerçek dünyayı tam anlamıyla yansıtmayabilir. Örneğin, bir hastalığın tanısını koymak için sadece yaşlı bireylerin belirtilerini içeren bir veri kümesi kullanılarak kurallar geliştirilirse, bu kuralları kullanarak bir çocuğa tanı koymak yanıltıcı olabilir. Bu tür durumlarda, bilgi keşfi modellerinin belirli bir güvenlik derecesinde tahmini kararlar alabilmesi gereklidir. Bu, sınıflandırma sürecinin belirsizlik içeren veya eksik verilere dayanarak karar almasını sağlayarak modeli daha esnek hale getirir (Luba, 1994).

1.13.4. Boş Değerler

Veri kümelerinde, bir niteliğin bilinmeyen veya uygulanamaz bir değere sahip olduğunu göstermek için null değerler kullanılır. Null değerler için çeşitli çözümler uygulanabilir. Bunlardan biri, null değer içeren kayıtların tamamen ihmal edilmesidir; bu, analizde eksik verinin dikkate alınmaması anlamına gelir. Alternatif olarak, null değerlerin yerini olacak olası bir değer atanabilir. Bu atanan değer, ilgili niteliğin en fazla frekansa sahip değeri, ortalama değer, varsayılan bir değer veya null değere en yakın olabilecek bir başka değer olabilir. Bu yöntemler, eksik verinin veri analizine olan etkisini azaltmak ve daha tutarlı sonuçlar elde etmek için kullanılır (Quinlan, 1986: 93).

Bir veri tabanında boş değerler, birincil anahtarda bulunmayan herhangi bir niteliğin değeri olarak karşımıza çıkabilir. Tanım olarak, boş değer kendisi de dahil hiçbir değere eşit olmayan bir değerdir. Bu durum, genellikle niteliğin bilinmeyen veya uygulanamaz bir değere sahip olduğunu gösterir ve ilişkisel veri tabanlarında oldukça yaygındır. İlişkisel veri tabanı modeline göre, bir ilişkideki tüm çoklular (tuple'lar), niteliğin değeri boş bile olsa, aynı sayıda niteliğe sahip olmalıdır. Bu standart, veri tabanının yapısal bütünlüğünü korumaya yardımcı olur ve boş değerlerin veri analiz süreçlerinde dikkate alınmasına yönelik çözümler geliştirir (Çalışkan, 2006).

1.13.5. Artık Veri

Örneklem kümesi, çözümlenmekte olan probleme uygun olmayan veya gereksiz nitelikler barındırabilir; bu gereksiz verilere "artık veri" denir. Bu tür nitelikleri elemek için kullanılan algoritmalar, özellik seçimi olarak adlandırılır. Özellik seçimi, yalnızca hedef bağlamı tanımlamak için gerekli ve yeterli niteliklerden oluşan küçük bir alt küme oluşturmayı hedefleyen bir işlemdir. Bu yöntem, yalnızca arama uzayını daraltmakla kalmaz, aynı zamanda sınıflama işleminin doğruluğunu ve kalitesini de artırır. Veri kümesinin problemi çözmek için gereksiz veya konu dışı nitelikler içermesi, veri işleme sürecinde sık karşılaşılan bir durumdur ve özellik seçimi bu soruna etkili bir çözüm sunar (Almuallim ve diğerleri, 1991).

1.13.6. Dinamik Veri

Sürekli değişken içerikli veri tabanlarıdır. Kurumsal çevrim-içi veri tabanları da bunlara örnek gösterilebilir. Veri tabanındaki şeyin sürekli değişmesi VM işletmelerinde uygulanabilmesi büyük oranda zorlaştırıcı sıkıntılar ortaya çıkartmaktadır (Özbay, 2007: 46):

- Veri madenciliği ile elde edilen örüntülerin, sürekli değişen verilere uyum sağlayıp sağlamadığını belirlemek zorluk yaratır. Bu örüntülerin hangi güncel veriyi ifade ettiğini tespit etmek ve zaman içinde ortaya çıkan değişiklikleri eski sonuçlarla karşılaştırarak farklılıkları belirlemek önemli bir çabadır. Ek olarak, gerekli yerlerin güncellenmesi de zaman alan ve karmaşık bir süreç olabilir. Bu durum, veri madenciliği analizlerinin doğruluğunu ve geçerliliğini korumak için sürekli bakım ve yenileme

gerektirir ve dinamik veri ortamlarında ciddi bir operasyonel zorluk oluşturur.

- Veri madenciliği algoritmalarının sağlıklı çalışabilmesi için, analiz sürecinde veriler üzerine okuma kilidi konulduğunda, bu verilerin başka uygulamalar tarafından değiştirilmemesi büyük önem taşır. Ancak, okuma kilidi altındaki veriler başka işlemler tarafından güncellenemediği için, veri tabanı güncelleme işlemlerinde sorunlar ortaya çıkabilir. Bu durum, özellikle veri tabanının aktif olarak kullanıldığı ve sık güncelleme gerektiren ortamlarda performans düşüşüne ve veri tutarlılığıyla ilgili zorluklara yol açabilir.
- Veri madenciliği algoritmalarının ve çevrim içi veri tabanı uygulamalarının eş zamanlı uygulanması, ciddi performans düşüşlerine yol açmaktadır.

1.14. Veri Madenciliğinde Karar Ağaçları

Karar ağacı, "Olası tüm eylem yönlerini, bu eylemleri etkileyebilecek tüm faktörleri ve her bir potansiyel sonucu veri doğrultusunda analiz eden; çizgi, kare, daire gibi geometrik semboller kullanarak düzenleyen ve böylece karar verilecek problemi anlamada kolaylık sağlayan bir yapı" olarak tanımlanabilir. Bu tanıma göre, karar ağacının farklı eylem seçeneklerini, bu eylemleri etkileyebilecek olası faktörleri ve bu faktörlerin olası sonuçlarını içerdiği söylenebilir. Karar ağaçları, bu yönleriyle hem karar verme sürecini görselleştirir hem de kullanıcıların karmaşık problemleri daha anlaşılır hale getirmelerine yardımcı olur (Gürsoy, 2009: 42).

Veri madenciliği yöntemleri, denetimsiz ve denetimli olmak üzere iki ana kategoriye ayrılır. Denetimsiz yöntemler, veri setinde herhangi bir hedef belirlemeden desenleri ve yapıları keşfetmek için kullanılır. Bu yöntemler, veriyi anlamak ve tanımlamak amacıyla analizin başlangıç aşamalarında tercih edilir ve sonraki süreçler için yol gösterici bilgiler sunar.

Denetimli yöntemler ise belirli bir hedefe yönelik analiz yapmak ve veriden anlamlı sonuçlar elde etmek için geliştirilmiştir. Bu yöntemler arasında, tahminsel

modelleme açısından önemli bir yere sahip olan karar ağaçları öne çıkar. Karar ağaçları, ağaç yapısında bir sınıflandırma modeli sunar. Her dal, belirli bir sınıflandırma sorusunu temsil ederken, yapraklar veri setinin belirli sınıflarına karşılık gelir. Bu yapı, karar ağaçlarının hem kolay anlaşılabilir hem de kullanıcı dostu olmasını sağlar. Bunun yanı sıra, karar ağaçları yalnızca sınıflandırma ve tahmin için değil, aynı zamanda veri setini keşfetmek ve iş problemlerini anlamak için de kullanılabilir. Ağacın dallarındaki tahmin ediciler ve değerler detaylı bir şekilde incelenerek veri hakkında önemli içgörüler elde edilebilir.

Karar ağaçlarının temel avantajlarından biri, veriyi sistematik bir şekilde sınıflandırma yeteneğidir. Bununla birlikte, ağacı aşırı detaylandırarak her bir dal ve yaprağı genişletmek, modelin karmaşıklığını artırabilir ve "aşırı öğrenme" sorununa yol açabilir. Bu durum, modelin genelleme kapasitesini düşürür ve gereksiz karmaşıklık yaratır. Bu nedenle, budama işlemi uygulanarak, karar ağacının gereksiz detaylardan arındırılması ve daha verimli bir yapı sunması sağlanır. Budama işlemi hem modelin performansını artırır hem de analizi daha sade ve etkili bir hale getirir.

Karar ağaçları, sınıflandırma amaçlı uygulamalarda sıkça tercih edilen tekniklerden biridir ve hem tahmin edici hem de tanımlayıcı özellikleriyle veri madenciliği projelerinde geniş bir kullanım alanına sahiptir (Gürsoy, 2009);

- Karar ağaçları, karmaşık modeller kadar yüksek maliyetli olmadığı için pratik bir sınıflama çözümü sunar.
- Yapısal olarak anlaşılır oldukları için, kullanıcıların sonuçları yorumlamaları oldukça basittir.
- Veri tabanı sistemleriyle kolayca entegre edilebilir, böylece büyük veri kümeleriyle verimli bir şekilde çalışabilir.
- Sınıflandırmada sağladıkları yüksek doğruluk oranı, güvenilir bir çözüm sunar.

Karar ağaçları, dallar, karar düğümleri ve yapraklardan oluşur. Karar düğümü, yapılacak testi belirtir; bu testin sonucu, ağacın veri kaybı olmadan dallara ayrılmasını sağlar. Her düğümde sırasıyla gerçekleşen test ve dallara ayrılma işlemleri, bir üst

seviyedeki ayrımlara bağlı olarak ilerler. Ağacın her bir dalı, sınıflama işlemini tamamlamaya adaydır; ancak dal sonunda sınıflama gerçekleşemiyorsa yeni bir karar düğümü oluşur. Eğer dalın sonunda belirli bir sınıf tanımlanabiliyorsa, bu dal yaprak ile sonlanır ve yaprak, veride belirlenmek istenen sınıfı temsil eder. Karar ağacı süreci, kök düğümden başlayarak yukarıdan aşağıya doğru ilerler ve yaprak düğümlere ulaşana kadar devam eder. Karar ağacı ile veriyi sınıflandırmak için iki aşamalı bir süreç izlenir. İlk aşama olan öğrenme aşamasında, model oluşturmak amacıyla sınıflama algoritması, önceden bilinen eğitim verisini analiz eder. Bu analiz sonucunda elde edilen model, sınıflama kuralları veya karar ağacı biçiminde temsil edilir. İkinci aşama olan sınıflama aşamasında, test verisi sınıflama kurallarının veya karar ağacının doğruluğunu değerlendirmek için kullanılır. Doğruluk kabul edilebilir seviyedeyse, model yeni verilerin sınıflandırılması için uygulanır. Bu iki aşamalı süreç, karar ağacının veriyi sistematik ve güvenilir bir biçimde sınıflandırmasını sağlar (Özekes, 2003: 66).

Karar ağacı temelli analizler, çok çeşitli alanlarda geniş bir kullanım alanına sahiptir (Aslan, 2008: 78). Bu alanlar şunlardır:

- **Belirli bir sınıfın potansiyel üyelerinin belirlenmesi**: Karar ağaçları, belirli bir kategoriye ait olması muhtemel olan öğeleri tanımlamak için kullanılır.
- **Risk gruplarının sınıflandırılması**: Farklı vakaları yüksek, orta veya düşük risk grupları gibi kategorilere ayırmada etkili bir tekniktir.
- **Gelecekteki olaylar için tahmin kuralları oluşturulması**: Karar ağaçları, gelecekte gerçekleşebilecek olayları tahmin etmek amacıyla kural geliştirme işlemlerinde kullanılır.
- **Parametrik modelleme için veri seçimi**: Çok sayıda değişken ve veri kümesi içinden analiz için en faydalı olan verileri seçerek parametrik modellerin oluşturulmasına katkı sağlar.
- **Belirli alt gruplara özgü ilişkilerin tanımlanması**: Yalnızca belirli gruplarda geçerli olan ilişkileri ortaya çıkarmak için karar ağaçları kullanılır.
- **Kategorilerin bir araya getirilmesi ve sürekli değişkenlerin kesikli değerlere dönüştürülmesi:** Karar ağaçları, kategorileri bir araya getirerek

veya sürekli değişkenleri kesikli hale dönüştürerek veri yapısını daha işlenebilir hale getirir.

Karar ağaçları, temel olarak bir tahmin tekniğidir ve sınıflandırma, kümeleme ile tahmin modellerinde yaygın olarak kullanılır. Bu yöntem, ilgili araştırma alanını alt gruplara ayırarak sorunu analiz etmek amacıyla tercih edilir. Karar ağaçları, eğitici örnekteki veriyi sınayan bir algoritma yardımıyla veya ilgili sektör uzmanının bilgisine dayalı olarak oluşturulabilir. Karar ağaçları, kullanılan oluşturma tekniğine bağlı olarak farklı türlerde ortaya çıkabilir ve her bir tür, veri setine ve amaca göre değişiklik gösterebilir. Bu esneklik, karar ağaçlarını çeşitli veri madenciliği projelerinde etkili bir araç haline getirir (Silahtaroğlu, 2013: 23).

Karar ağacı algoritması, ağacın kökünde hangi değişkenle test yapılacağını belirleyerek yukarıdan aşağıya doğru yapı oluşturur. Her bir düğümde, değişkenler test edilerek eğitim örneklerinin sınıflandırmasına karar verilir. Bu süreçte, her değişken, eğitim örnekleri üzerinde istatistiksel testler uygulanarak değerlendirilir. En iyi sonuç veren değişken, kök düğümde test için kullanılır. Her bir düğümde oluşturulacak dalların sayısı ise seçilen değişkenin alabileceği değerlerin sayısına bağlı olarak değişiklik gösterir. Bu yapı, karar ağacının esnekliğini sağlayarak sınıflandırmayı adım adım daha ayrıntılı hale getirir (Kök ve Kuloğlu, 2005).

Karar ağaçları, kurgulama, yorumlama ve veri tabanlarıyla entegrasyonunun kolay olması sebebiyle en yaygın kullanılan öngörü yöntemleri ve sınıflandırma teknikleri arasında yer alır. Bu tekniklerin bir diğer tercih edilme sebebi de güvenilirliklerinin yüksek olmasıdır. Karar ağaçlarının temel amacı, veriyi bağımlı değişkende farklılıkları maksimize edecek şekilde, sıralı bir biçimde parçalarına, yani farklı gruplara ayırmaktır. Bu nedenle sınıflandırma ağacı olarak da anılırlar. İstatistiksel yöntemler veya yapay sinir ağları gibi tekniklerde, veriden öğrenilen fonksiyonların insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zorken, karar ağaçlarında ağaç yapısı tamamlandıktan sonra kökten yaprağa inerek kolaylıkla kurallar yazılabilir. Bu kurallar, ilgili konuda uzman bir karar vericiye sunularak sonucun anlamlı olup olmadığı değerlendirilir. Karar ağacı, başka bir teknik

kullanılacak bile olsa önceden yapılacak kısa bir çalışma ile önemli değişkenler ve yaklaşık kurallar konusunda karar vericiye değerli bilgi sunabilir (Argüden ve Erşahin, 2008: 47).

Karar ağaçları ve karar kuralları, gerçek yaşamda birçok sınıflandırma problemine güçlü çözümler sunan etkili veri madenciliği yöntemleri arasında yer alır. Veriden sınıflandırıcılar oluşturmak için en etkin yöntemlerden biri olan karar ağaçları, mantıksal bir gösterim sunar ve yaygın bir kullanım alanına sahiptir. Özellikle makine öğrenimi ve uygulamalı istatistik gibi alanlarda tanımlanmış birçok karar ağacı türetme (endüksiyon) algoritması mevcuttur. Bu algoritmalar, bir girdi-çıktı örneklem setinden karar ağaçları yapılandıran denetimli öğrenme yöntemleridir. Tipik bir karar ağacı öğrenme sistemi, yukarıdan aşağıya doğru ilerleyen ve arama uzayının bir bölümünde çözüm arayan bir strateji izler. Karar ağacı, özniteliklerin test edildiği düğümlerden oluşur; her bir düğüm, belirli bir öznitelik üzerindeki testi temsil eder. Düğümlerin altındaki dallar ise, test edilen özniteliğin tüm olası sonuçlarına karşılık gelir. Bu yapı, karmaşık verilerden anlam çıkarma ve sınıflandırma işlemlerini basit ve anlaşılır hale getirerek karar ağaçlarını sınıflandırma problemlerinde güçlü bir çözüm aracı yapar (Irmak, 2009: 36).

Karar ağacı, karar düğümleri, kök, dallar ve yapraklardan oluşur. Kök, veri alanları içinde en önemli bileşen olarak ağacın başlangıç noktasını oluşturur. Karar düğümü, yapılacak testi temsil eder ve test sonucunda veri, ağacın dallarına ayrılır. Bu ayrılmalar her düğümde ardışık olarak gerçekleşir; ağacın her dalı, sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın sonunda sınıflama tamamlanamazsa, bu noktada yeni bir karar düğümü oluşturulur. Ancak, dalın sonunda belirli bir sınıf elde edilebiliyorsa, bu dal bir yaprak ile sonlanır. Yaprak, veri üzerinde belirlenmek istenen sınıfı ifade eder. Karar ağacı süreci, kök düğümden başlayarak yukarıdan aşağıya yapraklara ulaşana kadar ardışık düğümleri takip ederek ilerler ve veriyi sıralı bir şekilde sınıflandırır. Bu yapı, karar ağaçlarını anlaşılır ve etkili bir sınıflama aracı haline getirir (Özekes, 2003: 72).

Karar vericinin karar aldığı her bir nokta, karar noktası olarak adlandırılır ve ağaç üzerinde kare ile temsil edilir. Bu kareden çıkan dallar, karar vericinin olası stratejilerine karşılık gelir. Karar vericinin kontrolü dışında kalan olaylar ise olay düğüm noktaları olarak adlandırılır ve daire biçiminde gösterilir. Daire şeklindeki olay düğüm noktalarından çıkan her dal, bir olayı ve bu olayın gerçekleşme olasılığını simgeler; bu bilgi dal üzerinde belirtilir. Olay düğüm noktalarından çıkan dallar, karar vericiyi başka bir düğüm noktasına veya bir karar noktasına yönlendirebilir. Karar problemlerinin karar ağacı yöntemi ile çözümünde, sondan başa doğru bir hesaplama yaklaşımı benimsenir. Bu yaklaşımda, karar ağacının uç noktalarından başlayarak geriye doğru ilerlenir. Bu süreçte her karşılaşılan karar noktasında o düğümün beklenen değeri hesaplanır ve bu değerden en iyi olanı düğümün üzerine yazılır. Bu şekilde ilerleyerek başlangıç noktasına ulaşıldığında, çözüm işlemi tamamlanmış olur. Bu yöntem, karar vericinin her bir aşamada en iyi seçeneği belirleyebilmesine olanak tanır (Gürsoy, 2009: 73).

1.15. Veri Madenciliğinin Kullanım Alanları

Veri madenciliği, henüz yeni bir disiplin olmasına rağmen, birçok alanı kapsayan disiplinler arası bir yaklaşımdır (Chris, 2002:488).

Teknolojik gelişmelerin etkisiyle, kurumlarda veri miktarının artmasına çözüm olarak veri tabanı yönetim sistemleri geliştirilmiş ve bu sistemler verilerin toplanması, saklanması ve işlenmesini kolaylaştırarak veri işleme maliyetlerini düşürmüştür. Aynı zamanda, veri analizinde yeni yöntemlerin ortaya çıkması, veri madenciliği uygulamalarına olan ilgiyi artırmıştır (Sang, 2001: 205).

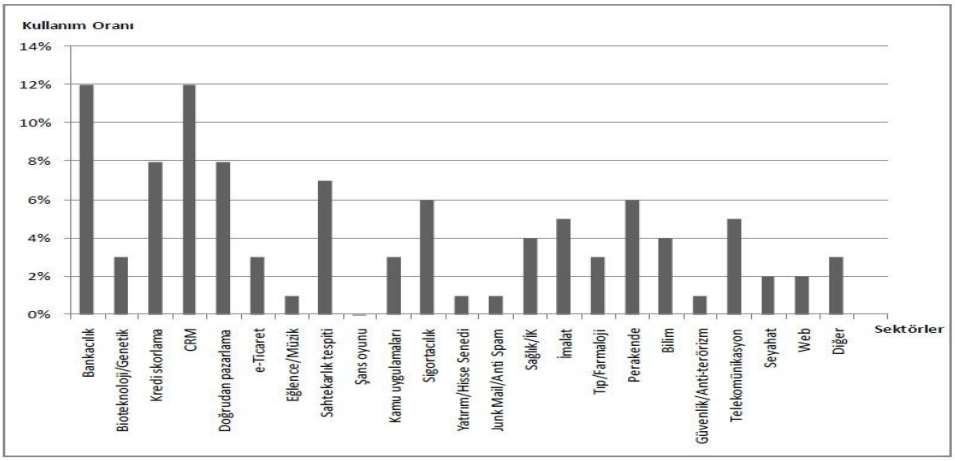
Bu ilgi, veri madenciliğinin uygulama alanlarının her geçen gün genişlemesine yol açmaktadır. Özel olarak bir disiplin için geliştirilmemiş olan veri madenciliği, verilerin toplandığı, saklandığı ve analiz edildiği tüm alanlarda uygulanabilir ve geniş bir kullanım alanına sahiptir (Akpınar, 2000: 4).

Veri madenciliği, işletmelerin iş kararlarını optimize etmelerine, mevcut müşterilerden daha fazla kazanç sağlamalarına, yeni müşteri kazanımlarını

artırmalarına ve sundukları mal veya hizmetlerde müşteri memnuniyetini yükseltmelerine yardımcı olur. Veri madenciliği uygulamalarında analiz edilen veriler, işletmeden işletmeye ve sektörden sektöre değişiklik gösterir. Sıkça kullanılan veri kaynakları arasında satış kayıtları, çağrı merkezi destek kayıtları, müşterilere ait demografik bilgiler ve firmanın internet sitesine ait ziyaretçi kayıtları gibi veri türleri bulunur. Bu veri kaynakları, işletmelerin daha bilinçli ve hedef odaklı kararlar almasına olanak tanır (Demirel, 2008: 35).

Veri madenciliğinin temel amacı, anlamlı bilgileri ortaya çıkararak bu bilgileri karar almayı ve eyleme geçmeyi desteklemek için kullanmaktır. Bu süreçte veri madenciliği, mevcut veya potansiyel müşteriler gibi ana kütlelere odaklanır. Müşterilerin profillerini, satın alma eğilimlerini ve bir ürünü ya da hizmeti kabul etme veya reddetme olasılıklarını tahmin etmek, veri madenciliğinin başlıca hedeflerindendir. Elde edilen bu tahminler, strateji belirleme sürecinde çeşitli kararlar almayı sağlar ve işletmelerin pazarlama, müşteri yönetimi ve ürün geliştirme gibi alanlarda daha bilinçli adımlar atmasına katkıda bulunur (Özmen, 2001: 3).

Veri madenciliği, son yıllarda dünya piyasalarında değişen ekonomik koşullar ve rekabetin artması sonucu birçok alanda yaygın olarak kullanılmaya başlanmıştır. Bu alanların başında pazarlama gelmekle birlikte; astronomi, biyoloji, finans, sigorta, tıp, güvenlik, milli güvenlik, spor, trafik, ulaşım, lojistik, tedarik zinciri ve meteoroloji gibi birçok farklı sektör de veri madenciliği uygulamalarından yararlanmaktadır (Akpınar, 2000: 4). Şekil 1.15'te, veri madenciliğinin çeşitli sektörlerdeki kullanım oranlarını özetleyen bir tablo gösterilmektedir. Bu tablo, veri madenciliğinin sektörel yaygınlığını ve bu alanlardaki etkisini genel bir bakışla sunar.



Şekil 1.15. Veri madenciliği uygulama alanları

Veri madenciliği, çeşitli sektörlerde farklı amaçlarla geniş bir uygulama alanına sahiptir (Akman, 2010: 10):

- **Pazarlama**: Müşterilerin demografik özellikleri arasındaki ilişkilerin tespiti, satın alma eğilimlerinin belirlenmesi, pazarlama kampanyalarının planlanması, mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması için pazarlama stratejilerinin geliştirilmesi, pazar sepeti ve çapraz satış analizleri, müşteri ilişkileri yönetimi ve satış tahminlerinde kullanılır.
- **Bankacılık**: Finansal göstergeler arasındaki gizli korelasyonların keşfi, kredi kartı dolandırıcılığının tespiti, kredi taleplerinin değerlendirilmesi, usulsüzlük tespiti, risk analizleri ve risk yönetimi gibi kritik alanlarda veri madenciliğinden faydalanılır.
- **Sigortacılık**: Yeni poliçe taleplerinin tahmini, sigorta dolandırıcılığının tespiti ve riskli müşteri profillerinin belirlenmesi için kullanılır.
- **Perakendecilik**: Satış noktası veri analizleri, alışveriş sepeti analizleri, tedarik ve mağaza yerleşim optimizasyonu gibi uygulamalarda veri madenciliği önem taşır.
- **Telekomünikasyon**: Kalite iyileştirme analizleri, hat yoğunluk tahminleri ve telefon dolandırıcılığının tespiti gibi alanlarda kullanılır.
- **Sağlık ve Farmakoloji**: İlaç geliştirme, hastalıkların teşhisi ve tedavi sürecinin planlanması gibi medikal uygulamalarda veri madenciliğinden yararlanılır.

- **Endüstri**: Kalite kontrol analizleri ile lojistik ve üretim süreçlerinin optimizasyonunda veri madenciliği kullanımı yaygındır.
- **Bilişim ve Mühendislik**: İnternet işlemlerinde dolandırıcılık tespiti, bilgisayar sistemlerine yetkisiz girişlerin belirlenmesi, parmak izi ve yüz tanıma gibi kimlik doğrulama sistemleri ve yapay zeka uygulamalarında veri madenciliği kritik bir rol oynar.

İKİNCİ BÖLÜM

2.1. Veri Madenciliği ve Sağlık Sektörü İlişkisi

Son dönemde bilgiye en fazla ihtiyaç duyulan alanların başında sağlık ve tıp gelmektedir. Özellikle veri modelleri, standartlar ve kodlama sistemlerinin kullanımı sayesinde, hastaneler ve sağlık merkezlerinde kullanılan bilgi sistemlerinde dikkate değer ilerlemeler kaydedilmiştir. Bu gelişmeler, bilgi keşfi gereksinimini ortaya çıkarırken, aynı zamanda verilerin güvenle saklanması ve yönetilmesi gibi çeşitli ihtiyaçların karşılanmasını da sağlamıştır.

Veri madenciliği, sağlık ve tıp alanında yalnızca hizmet kalitesinin artırılmasına katkı sağlamakla kalmayıp, aynı zamanda sağlık alanındaki büyük veri tabanlarında yer alan değerli bilgilerin gün yüzüne çıkarılmasına da olanak tanımaktadır. Sağlık sektörü, bu bilgi keşfi sürecinden büyük fayda sağlayarak, hem klinik karar destek süreçlerini güçlendirmiş hem de hasta bakımını daha verimli hale getirme yönünde önemli adımlar atmıştır.

Tıp alanında son yıllarda kaydedilen teknolojik gelişmeler, biyomedikal elektronik alanında önemli ilerlemelere zemin hazırlamış ve biyomedikal verilerin yanı sıra çok çeşitli kavramsal verilerin toplanıp bir arada değerlendirilmesine olanak tanımıştır. İnsan algısının bazı ince detayları gözden kaçırma olasılığına karşı, fark edilmesi güç bilgilerin belirlenmesi ve hastalıkların erken teşhisi ile hızlı müdahaleler veya tedavi değişikliklerinin yapılabilmesi veri madenciliği yaklaşımlarıyla daha mümkün hale gelmiştir. Bu yenilikler yalnızca biyolojik konularla sınırlı kalmayıp, hastane yönetimi, kurum performansı ve klinik süreçlerin analizinde de veri madenciliğinin etkili bir araç olarak kullanılmasını sağlamıştır. Dolayısıyla tıpta veri madenciliğine yalnızca biyolojik bir bakış açısıyla değil, aynı zamanda klinik uygulamalar ve yönetim süreçleri açısından da geniş bir perspektiften yaklaşmak gerekmektedir. Bu, sağlık hizmetlerinin hem tanısal doğruluğunu hem de operasyonel verimliliğini artırmada veri madenciliğinin kritik rol oynadığını ortaya koymaktadır.

Tıp alanında veri madenciliğine yönelik ilk uygulama, 1854 yılında Dr. John Snow tarafından Londra'da kağıt ve kalemle gerçekleştirilmiştir. Londra'da meydana gelen kolera salgını sırasında Snow, bir harita üzerine koleradan ölen kişilerin konumlarını işaretleyerek ilk kümeleme analizini gerçekleştirmiştir. Bu çalışma sayesinde ölümlerin belirli bölgelerde yoğunlaştığını fark etmiş ve salgının kaynağının bu bölgelerdeki su pompaları olduğunu tespit etmiştir. Snow'un bu analizi, su pompalarının kapatılması gibi halk sağlığını koruyucu önlemlerin alınmasını sağlayarak salgının sona erdirilmesine büyük katkıda bulunmuştur. Bu çalışma, tıp alanında veri analizinin ve veri madenciliğinin öncülerinden biri olarak kabul edilmektedir (Erkuş, 2015: 19).

Sağlık sektörü, bilgi içeriği ve yapısında en hızlı değişen alanlardan biridir. Sağlık hizmetlerinin doğru, hızlı, yüksek kalitede ve ihtiyaçlara cevap verecek şekilde sunulabilmesi için, sağlık profesyonellerinin en güncel ve doğru bilgiye ulaşması, bu bilgiyi de karar destek sistemleri aracılığıyla kullanabilmesi son derece önemlidir. Veri madenciliği, büyük veri yığınları içinde gizli kalmış, değerli ve kullanılabilir bilgileri ortaya çıkararak stratejik karar desteği sağlayan bir yöntemdir. Bu yöntem, verilerin analizine dayanarak karar verme modelleri oluşturur ve sağlık hizmetlerinin sunumunda, sağlık kurumlarının yönetiminde ve sağlık politikalarının geliştirilmesinde etkili bir karar destek aracı olarak kullanılabilir. Veri madenciliği, sağlık profesyonellerinin en optimal kararları almasına katkıda bulunarak sağlık hizmetlerinde kalite ve verimliliği artırır (Koyuncugil ve Özgülbaş, 2009: 24).

Veri madenciliği, sağlık alanında en çok kullanılan ve önemli sonuçlar elde edilen yöntemlerden biridir. Sağlık alanında veri madenciliği uygulamaları; ilaç geliştirme, ilaçların etkilerinin değerlendirilmesi, hastaların test sonuçlarının öngörülmesi, hastalıkların önceden teşhis edilmesi ve tedavi edilmesi gibi kritik süreçlerde büyük önem taşır (Özbay, 2015: 265). Hastanelerin veri tabanları, yararlı ve gereksiz birçok bilgi içerir ve bu bilgi yığını içinden değerli bilgilere ulaşmanın yolu veri madenciliğidir. Veri madenciliği ile mevcut veriler sınıflandırılarak gereksiz bilgilerden arındırılır ve istenilen bilgilere etkin bir şekilde ulaşılır. Hastaneler, veri madenciliği uygulamaları sayesinde belirli hastalık gruplarına uygulanan tedavi

yöntemlerini ve ilaç kullanımını analiz ederek tedavi sürelerinin kısalmasına katkı sağlamakta ve ilaçların tedavi süreçlerine olan etkilerini belirleyerek daha etkili tedavi stratejileri geliştirebilmektedir. Bu hem hasta bakım kalitesini artırır hem de sağlık hizmetlerinin verimliliğini yükseltir.

İlaçlarda tıbbın önemli araştırma konularından biridir. İlaçların onaylanmadan önce faydalarının risklerinden daha çok olması koşulu göz önün de bulundurulur. Bazı ilaçlar piyasaya sürüldükten sonra risklerinin çok fazla görülmesi nedeniyle kaldırılmaktadır. İlaç etkileri analizi, ilaç üretimi ve geliştirilmesi gibi konularda da veri madenciliği tekniklerinden sık sık faydalanılmaktadır. Yapılan bir çalışmada antipsikotik ilaçların kalp kası hastalıkları üzerine etkileri yapay sinir ağları yöntemi ile araştırılmış ve bulunan sonuçlara göre klozapin dışındaki antipsikotik ilaçların miyokardi ve kardiyomiyopati ile önemli derecede ilişkili olduğu belirlenmiştir (Köktürk ve ark, 2009: 24).

Tıp alanında bilginin kullanım şekli değiştikçe, sağlık bakım hizmetini sağlayan profesyoneller üzerinde de önemli etkiler yaratmıştır. Sağlık bakım hizmetlerinin verilmesinde bilgisayarların kullanımı, bilgi paylaşımını, ekip yaklaşımını ve veri ile bilgi temelli uygulamaları yaygınlaştırmıştır. Bilgisayarlar, hasta bakım hizmetlerini destekleme ve sağlık hizmetlerinin kalitesini değerlendirme gibi doğrudan sağlık bakım hizmetlerinin sunulmasında kullanıldığı gibi; teşhis koyma, tedavi süreçlerinin yönetimi, planlama ve tıbbi araştırmalar gibi yönetsel ve akademik alanlarda da giderek daha fazla yer bulmaktadır. Tıp alanında bulunan mevcut veriler oldukça büyük bir hacme sahip olup hayati öneme sahiptir. Hastane bilgi sistemleri, bu verileri düzenli olarak saklayarak tıbbi verilerin verimli şekilde kullanılmasını sağlar. Bu sistemlerdeki verilere veri madenciliği teknikleri uygulandığında ise sağlık uzmanları, hastane yönetimi ve hastalar için daha kaliteli bir hizmet sunulmasına katkı sağlanmaktadır. Veri madenciliği, hastane bilgi sistemlerinden veya diğer tıbbi veri toplama sistemlerinden alınan bilgileri analiz ederek sağlık hizmetlerinin optimizasyonuna olanak tanır ve bilgi temelli karar alma süreçlerini destekler. Bu sayede, hastalar daha iyi bakım alırken, uzmanlar ve yönetim daha stratejik kararlar alabilirler (Albayrak, 2008: 81).

Tıbbi verilerin etkin veri madenciliği uygulamaları için kullanılması, bu verilerin kendine özgü özelliklerinin dikkatle ele alınmasını gerektirir (Aslan, 2008: 68):

1. **Görüntüleme Verileri**: Çok sayıda tıbbi yordam, tanı aracı olarak görüntülemeye dayanır. Bu nedenle, görüntülerden oluşan veri tabanlarında veri madenciliği gerçekleştirmek için özel yöntemler geliştirilmelidir. Sayısal veri tabanlarından farklı olarak bu tür veriler, analiz ve yöntem açısından daha karmaşıktır.
2. **Heterojenlik**: Tıbbi veri tabanları genellikle heterojendir. Örneğin, bir organın görüntüsü çoğunlukla hekimin yorumu, klinik izlenimler veya tanı gibi diğer bilgilerle bir arada bulunur. Bu karmaşık verilerin analizi için özel araçlar ve yüksek kapasiteli veri depolama çözümleri gereklidir.
3. **Serbest Metin Kullanımı**: Hekimler, görüntüler, sinyaller veya klinik bilgilerle ilgili açıklamalarını çoğunlukla standartlaştırılması zor olan serbest metinlerde ifade eder. Aynı hastalığı tanımlarken bile farklı terminoloji kullanılır, bu da veri analizini zorlaştırır.
4. **Veri Sahipliği**: Tıbbi verilerin sahibinin kim olduğu sorusu belirsizdir. Hastalar mı, hekimler mi, yoksa sigorta kurumları mı bu verilerin sahipliğini üstlenir? Heterojen veri tabanlarında saklanan büyük miktarda tıbbi verinin sahipliği netleştirilememiştir.
5. **Hukuki Sorunlar**: Hekimlere veya sağlık hizmeti sağlayıcılarına yönelik davalar veri paylaşımında engeller oluşturabilir. Hekimler, özellikle gereksiz testler nedeniyle açılan davalardan çekindiğinden, verilerini analiz yapacak kişi veya kurumlara aktarmakta isteksiz davranabilir.
6. **Gizlilik ve Güvenlik**: Tıbbi veriler elektronik ortamda aktarılırken güvenlik riski taşır. Bu nedenle, kurum içindeki veri aktarımlarında bile verilerin dikkatle şifrelenmesi gerekmektedir.
7. **Matematiksel Yapı Yetersizliği**: Tıptaki temel veri yapıları, fiziksel bilimlerde olduğu gibi matematiksel olarak net bir şekilde karakterize edilmez. Veri madencileri için kümeleme, regresyon veya dizi çözümlemeleri gibi standart analiz yapılarını kullanmak zordur.

8. **Hasta Merkezli Veri Toplama**: Tıp, öncelikli olarak insan sağlığını koruma amacı taşır; araştırma ve bilgi toplama ikincil öneme sahiptir. Bilgi toplanması veya bazı bilgilerin toplanmaması, doğrudan hasta yararı ile ilişkilidir.

Bu özellikler, tıbbi verilerin analiz edilmesini zorlaştırmakla birlikte, doğru yöntemlerin seçilmesiyle sağlık hizmetlerinde veri madenciliğinin etkin kullanımını sağlama potansiyelini de barındırır.

2.2. Veri Madenciliğinin Tıp ve Sağlık Hizmetlerinde Kullanımı

Hastane Bilgi Yönetim Sistemleri (HBYS), hasta bilgileri, hastalık detayları, laboratuvar sonuçları, tedavi süreçleri ve tedaviye ilişkin kritik verilerin düzenli ve sistematik bir şekilde kaydedilmesini sağlamaktadır. Bu sistemler, hastanelerin işleyişini belirli bir düzene oturtarak bilgi akışının etkin bir şekilde yönetilmesine olanak tanır.

Hastane veri tabanlarında büyük miktarda faydalı ve gereksiz bilgi birikmektedir. Bu verilerden anlamlı sonuçlar elde edebilmek için veri madenciliği teknikleri devreye girmektedir. Veri madenciliği, veri tabanındaki bilgileri sınıflandırır, gereksiz olanları ayıklar ve ihtiyaç duyulan bilgilere ulaşmayı kolaylaştırır. Bu sayede, hastaneler veri tabanlarında gerçekleştirdikleri analizlerle, örneğin belirli hastalık gruplarına uygulanan tedavi yöntemlerini ve ilaçların etkinliğini değerlendirebilmektedir. Bu analizler, tedavi sürelerinin kısaltılması, ilaç kullanımının tedavi üzerindeki etkisinin tespit edilmesi ve hasta bakım kalitesinin artırılması gibi değerli sonuçlar sağlamaktadır.

Adrew Kusiak ve ekibi, akciğerdeki tümörün iyi huylu olup olmadığını belirlemeye yönelik bir karar destek çalışması gerçekleştirmiştir. Amerika Birleşik Devletleri'nde istatistiklere göre 160.000'den fazla akciğer kanseri vakası görülmekte ve bu vakaların %90'ı ölümle sonuçlanmaktadır. Bu durum, tümörlerin erken ve doğru teşhis edilmesini son derece önemli kılmaktadır. Noninvaziv testlerle yapılan teşhislerin doğruluk oranı %40–60 arasında değişmekte olup, kesin sonuç almak isteyen hastalar genellikle biyopsi gibi invaziv yöntemlere başvurmaktadır. Ancak biyopsi hem yüksek maliyetli hem de riskli bir işlemdir. Farklı zamanlarda ve farklı

kliniklerde toplanan invaziv test verileri üzerinde yapılan veri madenciliği çalışmaları ise teşhis doğruluğunu %100 seviyesine çıkarmıştır (Kusiak, 2000: 103-107).

Bir diğer çalışma ise Kore Tıbbi Sigorta Kurumu tarafından oluşturulan bir veri tabanı üzerinde yüksek tansiyon ile ilgili gerçekleştirilmiştir. Çalışma, 1998 yılına ait 127.886 kayıttan oluşan bir veri kümesi üzerinde yürütülmüştür. İlk aşamada yüksek tansiyonu olan 9.103 kayıt analiz edilmiş, ardından aynı sayıda yüksek tansiyonu olmayan kayıt incelenmiştir. Veriler, 13.689 kayıttan oluşan öğrenme seti ve 4.588 kayıttan oluşan test setine ayrılarak model eğitimi yapılmıştır. Bu süreçte CHAD, C4.5 ve C5.0 gibi karar ağacı algoritmaları kullanılmıştır. Sonuç olarak, yüksek tansiyon tahmininde etkili faktörlerin vücut kitle indeksi , idrar proteini, kan şekeri ve kolesterol düzeyleri olduğu tespit edilmiştir. Buna karşılık, yaşam tarzına ilişkin faktörler (diyet, tuz tüketimi, alkol, tütün gibi) tahmin açısından etkisiz bulunmuş ve yalnızca yaş faktörünün grafiksel analizlerde anlamlı bir etkiye sahip olduğu belirlenmiştir (Chae, 2001 : 168-171).

Veri madenciliği, sağlık alanında önemli bir yer tutmakta ve çeşitli kuruluşlarda farklı uygulama alanları bulmaktadır. Örneğin, ilaç sektöründe faaliyet gösteren Vysis**,** protein analizini gerçekleştirmek için yapay sinir ağları algoritmasını kullanarak veri madenciliği yapmaktadır. Rochester Kanser Merkezi ise araştırmalarında KnowledgeSEEKER adlı karar ağacı tekniğini uygulayarak verilerden anlamlı sonuçlar elde etmektedir. California Hastanesi de veri madenciliği amacıyla “Information Discovery” adlı bir yazılımı kullanmaktadır. Bu yazılımı kullanan bir doktor, bu program sayesinde hastalarını fiziksel testlere gerek duymadan teşhis edebildiğini belirtmiştir. Bu örnekler, veri madenciliğinin sağlık alanında teşhis, tedavi ve araştırma süreçlerinde nasıl etkili bir araç haline geldiğini göstermektedir (Doğan, 2007: 41).

Sağlık alanında veri madenciliği kullanılarak yapılan çalışmalardan biri, kadınlarda göğüs kanseri riskini incelemeye odaklanmıştır. Bu çalışmada, genetik algoritma tekniği uygulanarak, göğüs kanserinde erken teşhis olanaklarının artırılması ve doğru teşhis oranlarının yükseltilmesi sağlanmıştır. Bu veri madenciliği tekniği

sayesinde, hastalığın erken teşhis edilmesi mümkün olmuş ve sonuçların doğruluğu kanıtlanmıştır. Tıp alanında yapılan bir başka araştırmada ise genetik özellikler ve çevresel faktörlerin obezite ve diyabet gibi çok etmenli hastalıklar üzerindeki etkileri incelenmiştir. Bu çalışma, Lille Biyolojik Enstitüsü'nden elde edilen geniş kapsamlı veriler kullanılarak gerçekleştirilmiştir. Veri miktarının çok büyük olması nedeniyle, keşifsel (heuristic) bir yaklaşım benimsenmiş, bu sayede veri madenciliği ile karmaşık ilişkiler arasında anlamlı örüntüler elde edilmiştir. Bu tür çalışmalar, sağlık alanında hastalıkların daha iyi anlaşılmasını ve önleyici tedbirlerin geliştirilmesini sağlamaktadır (Gündoğdu, 2007: 38).

Tıp ve sağlık hizmetlerinde veri madenciliği uygulamalarına dair çeşitli çalışmalar, sağlık profesyonellerine önemli katkılar sunmaktadır (Savaş ve ark., 2012: 13):

1. **Barış Aksoy (2009)**: Aksoy, dekompresyon hastalığı üzerine veri madenciliği uygulaması gerçekleştirmiştir. Bu çalışmada k-ortalama, COBWEB ve EM gibi farklı kümeleme algoritmaları kullanılarak, Dalgıçların Acil Durum Ağı'ndan elde edilen dalış yaralanması verileri analiz edilmiştir. Dekompresyon hastalığına dair belirti ve bulgular kümeleme yöntemleriyle sınıflandırılmış ve sonuçlar klasik sınıflandırma yöntemleri ile karşılaştırılmıştır. Ayrıca, teşhise katkı sağlayabilecek birliktelik kuralları (association rules) elde edilmiştir. Sonuçlar, kümeleme yöntemleriyle elde edilen sınıfların hiyerarşik bir yapıda olduğu ve klasik sınıflandırma yöntemleriyle uyum gösterdiğini ortaya koymuştur.
2. **Pınar Yıldırım, Mahmut Uludağ ve Abdülkadir Görür (2008)**: Bu çalışmada, hastane bilgi sistemlerindeki veri madenciliği uygulamalarının çeşitli boyutları incelenmiştir. Sağlık kurumlarında veri madenciliği kullanılarak veri analizi yapılmasının, hasta bakım kalitesini ve hastane süreçlerinin verimliliğini artırabileceği vurgulanmıştır.
3. **Şengül Doğan ve İbrahim Türkoğlu (2008)**: Bu çalışma, kan biyokimya parametrelerini kullanarak demir eksikliği anemisi teşhisinde karar destek sistemi oluşturmuştur. Sistem, karar ağaçları yapısına dayanan veri madenciliği teknikleriyle geliştirilmiş ve biyokimya parametrelerinden serum demiri, serum demir bağlama kapasitesi (SDBK) ve ferritin enzimleri temel

belirleyiciler olarak kullanılmıştır. Sistem, anemi teşhisi için 96 hasta verisini değerlendirmiş ve sonuçları doktor kararları ile tamamen uyum göstermiştir.

4. **Mustafa Danacı, Mete Çelik ve A. Erhan Akkaya (2010)**: Meme kanseri üzerine yapılan bu çalışmada, Xcyt örüntü tanıma programı ile doku verileri analiz edilmiş, Weka programı kullanılarak meme kanseri hücrelerinin teşhis ve tahmini yapılmıştır. Bu çalışmada kadınlar arasında en sık rastlanan kanser türlerinden olan meme kanserine dair çeşitli veriler kullanılarak tahmin doğruluğu yüksek sonuçlara ulaşılmıştır.

Bu çalışmalar, veri madenciliği tekniklerinin sağlık sektöründe tanı, teşhis ve karar destek sistemlerinin geliştirilmesine nasıl katkı sağladığını ortaya koymaktadır.

Sun ve arkadaşlarının (2018) çalışmalarında, tıbbi kayıtlarda yer alan hasta verilerinin veri madenciliği yöntemleri ile analiz edilmesi ele alınmıştır. Çalışmada, mevcut sistem olarak kullanılan Elektronik Tıbbi Kayıtların (EMR), veri analiz süreçlerini zorlaştırdığı ifade edilmiş ve bu nedenle öncelikli olarak verilerin işlenmesi gerektiği vurgulanmıştır. Araştırmanın ilk aşamasında, Tıbbi Karar Destek Sistemi'nin (MDSS) doktorların tanı ve tedavi süreçlerinde hız kazanmalarına katkı sağlayacağı öngörülmüş ve bu sistemin Çin'de bir hastanede hala deneysel ve teorik araştırma aşamasında olduğu belirtilmiştir. İkinci olarak, mobil sağlık sistemlerinin hastane personelinin çalışma koşullarını büyük ölçüde esnekleştireceği ve fiziksel kaliteyi artıracağı değerlendirilmiştir. Son olarak, hastaların ilaç reaksiyonlarını hızlı bir şekilde tespit edebilecek bir veri altyapısının, özellikle salgın hastalıkların yaygın olduğu bölgelerde önemli bir rol oynayacağı ifade edilmiştir. Bu altyapının, olumsuz ilaç olaylarının daha düşük maliyetle tespit edilmesine olanak sağlayacağı sonucuna ulaşılmıştır (Wencheng, 2018).

Mellor ve arkadaşları (2018) tarafından yapılan çalışmada, odyoloji bölümüne başvuran hastaların işitme cihazı kullanımında dikkate alınan değişkenler kümeleme analizi yöntemi ile incelenmiştir. Çalışmada, işitme cihazlarının veri tabanlarına erişim sürecinde ticari gizlilik ve hasta mahremiyetinin korunmasına öncelik verilmiş, bu doğrultuda gerçek verilerin alt kümeleri oluşturulmuştur. Analiz sürecinde, kullanılan

cihaz türü ve hasta profili ile uyumlu bir model geliştirilmiş ve bu sayede daha işlevsel ve faydalı raporlar oluşturulması sağlanmıştır (Mellor, 2018).

Fayez (2018), koroner kalp hastalığının teşhis süreçlerini ve maliyetlerini iyileştirmeyi amaçlayan çalışmasında, üç farklı veri madenciliği algoritmasını kullanarak yüksek doğruluk oranlarına ulaşmıştır. Koroner kalp hastalığına sahip hastalara ait verilerin analizi sonucunda, SVM algoritmasıyla %58, Random Forest algoritmasıyla %99 ve Cleveland algoritmasıyla %94 doğruluk oranları elde edilmiştir. Bu analiz sonuçlarına dayanarak, daha hızlı ve kesin sonuçlar üretebilecek bir sistem tasarımı gerçekleştirilmiştir (Fayez, 2018: 1-54).

Sebik ve Bülbül (2018), Dünya Sağlık Örgütü verilerini kullanarak akciğer kanserinin erken teşhisine yönelik bir çalışma gerçekleştirmiştir. Çalışmada, WEKA veri madenciliği yazılımı kullanılarak tercih edilme oranı yüksek on farklı algoritma (Naive Bayes, BayesNet, Lojistik Regresyon, Multilayer Perceptron, KStar, Bagging, OneR, ZeroR, J48 ve Random Tree) test edilmiştir. Algoritmalar, kanser teşhisi konulan hastalardaki temel öznitelikler üzerinden doğruluk, kesinlik, duyarlılık ve F-ölçütü metrikleri açısından karşılaştırılmış ve en başarılı algoritma olarak Naive Bayes belirlenmiştir. Çalışmanın sonucunda, bu veri setinin kullanımı ile hastalığın teşhis süresinin kısaltılması ve erken tanı tedavisinde daha başarılı sonuçlar elde edilmesinin mümkün olduğu öne sürülmüştür (Sebik, 2018:1-7).

Teknolojinin hızlı ilerlemesi, özellikle tıp ve sağlık alanlarında veri artışını hızlandırmış ve bu verilerin analizi, kaliteli sağlık hizmetleri sunulmasında büyük bir önem kazanmıştır. Sağlık alanında veri madenciliğinin kullanımı bu bağlamda giderek artmaktadır; çünkü büyük veri setlerinin analiz edilmesi, hasta bakımı ve tıbbi araştırmalar açısından değerli içgörüler sağlamaktadır. Sağlık alanında veri biriktirme ve saklama süreçlerine olanak tanıyan temel uygulama alanları (Çarklı, 2010: 30) şu şekildedir:

- **Görüntüleme**: Tıbbi görüntüler (örneğin MR, tomografi) analiz edilerek hastalıkların belirlenmesinde kullanılır.

- **Teşhis Koyma**: Hastalıkların erken teşhisine yönelik veri biriktirme ve analiz çalışmaları yapılır.
- **Prognoz**: Hastalığın ilerleyişine dair öngörülerde bulunulması sağlanır.
- **Terapi**: Tedavi yöntemlerinin bireysel hasta verileriyle optimize edilmesi amaçlanır.
- **Hastalık Evrelerinin Kontrolü**: Hastalıkların farklı evrelerindeki değişimlerin izlenmesine yardımcı olur.
- **Biyomedikal Analizler**: Tıbbi cihazlardan elde edilen verilerin analiz edilmesi ile biyomedikal süreçlerin değerlendirilmesi yapılır.
- **Biyolojik Analizler**: Genetik ve moleküler düzeyde verilerin analiz edilmesiyle hastalık risk faktörleri belirlenir.
- **Epidemiyolojik Çalışmalar**: Hastalıkların toplum üzerindeki etkilerini analiz etmek için büyük veri setleri incelenir.
- **Hastane Yönetimi**: Hasta verilerinin ve hastane kaynaklarının etkili yönetimi sağlanır.
- **Tıbbi Yönerge ve Eğitim**: Tıbbi veriler, sağlık çalışanları için eğitici içerikler sağlamak amacıyla kullanılır.

Tıbbi veriler, yapıları gereği karmaşık, belirsiz ve doğrusal olmayan özellikler sergiler. Sağlık sektöründe bilgi sistemlerinin yoğun kullanımıyla birlikte, büyük miktarda veri birikmiş ve bu durum veri işleme süreçlerini oldukça zorlaştırmıştır. Ancak, bu veriler üzerinde veri madenciliği teknikleri kullanıldığında, karmaşık yapılar düzenlenebilir, anlamlı ilişkiler ortaya çıkarılabilir ve tahmin edilebilir sonuçlar elde edilebilir. Veri madenciliği, özellikle teşhis, tedavi ve sağlık yönetimi süreçlerinde derin analizlere olanak tanıyarak önemli bir rol oynar. Örneğin, bir hastalığa dair çok sayıda semptom bulunması teşhis sürecini güçleştirebilir. Veri madenciliği algoritmaları, bu semptomların sistematik bir şekilde analiz edilmesini sağlayarak, teşhis sürecini daha hızlı ve etkin hale getirir. Bu sayede, elde edilen veriye dayalı bilgiler sağlık uzmanlarının daha isabetli ve hızlı kararlar almasına yardımcı olur ve sağlık hizmetlerinin kalitesini artırır.

2.3. Veri Madenciliğinin Tıp ve Sağlık Hizmetlerinde Uygulama Alanları

Teşhis sürecinde, bir hastalığa ait çok sayıda semptomun varlığı, doğru teşhis koymayı zorlaştırıcı bir unsur haline gelebilir. Ancak, veri madenciliği algoritmalarının bu semptomların analizinde kullanılması, teşhis sürecini büyük ölçüde kolaylaştırır. Bu algoritmalar sayesinde, hekimler, çok sayıda semptom arasından en ayırt edici olanları belirleyebilir ve böylelikle daha net ve güvenilir sonuçlara ulaşabilir. Tıbbi verilerin bu şekilde sistematik olarak analiz edilmesi, sağlık profesyonellerinin daha doğru ve hızlı kararlar almasını destekler ve tedavi süreçlerinin etkinliğini artırır. Sonuç olarak, veri madenciliği, teşhis koyma sürecindeki karmaşıklığı azaltarak tıbbi karar destek sistemlerinde önemli bir rol oynar (Köktürk ve ark, 2009: 21).

Verma ve arkadaşlarının (2019) gerçekleştirdiği çalışmada, cilt hastalıkları (sedef hastalığı, seboreik dermatit, liken planus, pityriasis rosea, kronik dermatit) beş farklı veri madenciliği yöntemi ile analiz edilerek bir bilgi sistemi geliştirilmesi amaçlanmıştır. Çalışmada kullanılan yöntemler arasında Sınıflandırma ve Regresyon Ağacı (CART), Rastgele Orman, Karar Ağacı, Destek Vektör Makineleri (SVM) ve Degrade Artırıcı Karar Ağacı (GBDT) yer almaktadır. Araştırma, 35 cilt hastalığı değişkenini içeren bir veri kümesi ve 360 örneklem üzerinde gerçekleştirilmiştir. Girdi verileri üç kez değiştirilerek tüm yöntemler yeniden test edilmiş ve en yüksek doğruluk oranı %98,64 olarak elde edilmiştir. Bu bulgular, eritemaskuamöz hastalık veri kümesinin sınıflandırma problemleri için oldukça uygun bir yapıya sahip olduğunu ortaya koymaktadır.

Sağlık sistemi politikalarının ve yönetsel kararların temeli, güvenilir ve işlenmiş verilere dayanır. Sağlık politika ve kararlarının amaçlarına uygun, etkin ve etkili olabilmesi, güvenilir, güncel ve doğru verilere bağlıdır. Bu kapsamda, sağlık bilgi sistemlerinin temel amacı, büyük miktardaki tıbbi veriyi anlamlandırarak faydalı bilgiye dönüştürmektir. Elde edilen bu bilgiler, hasta düzeyinde daha kaliteli sağlık hizmeti sunumunu, sağlık kurumlarının daha etkili yönetimini, kaynakların verimli kullanılmasını ve sağlık politikalarının şekillendirilmesini destekler. Tıbbi veriler, hastaneler, sağlık kurumları, sigorta şirketleri ve ilgili kamu kurumları gibi çeşitli

kuruluşlar tarafından toplanmaktadır. Bu tür verilerden elde edilen bilgiler, sağlık sisteminin her düzeyinde etkili kararlar almak için kritik bir rol oynamaktadır. Özellikle, bu bilgiler, hasta bakımını iyileştirmek, kaynakların en iyi şekilde dağıtımını sağlamak ve sağlık alanında bilinçli politikalar geliştirmek için kullanılır (Koyuncugil ve Özgülbaş, 2009: 28).

Tıp alanında bulunan mevcut veri miktarı oldukça fazladır ve bu veriler hayati öneme sahiptir. Sağlık alanında yapılan pek çok veri madenciliği araştırmasında, hastaların elektronik tıbbi kayıtları ile idari işlemleri belgeleyen veriler analiz edilmekte, bu verilerden yararlanılarak farklı öngörüler yapılabilmektedir (Kudyba, 2004: 146-163).

Bu öngörüler şu şekilde sıralanabilir:

- Belirli bir hastalığa sahip bireylerin ortak özelliklerinin tahmin edilmesi,
- Tıbbi tedavi sonrası hastaların durumlarının öngörülmesi,
- Hastane maliyetlerinin hesaplanması,
- Ölüm oranları ve salgın hastalıkların yayılma potansiyelinin tahmin edilmesi.

Veri madenciliğinin, tıp ve sağlık hizmetleri alanında sunduğu diğer kullanımlar ise şunlardır:

- **Tıbbi Teşhis**: Hastalıkların daha doğru ve hızlı teşhisi için veri madenciliği uygulamaları kullanılır.
- **Uygun Tedavi Sürecinin Belirlenmesi**: Hastalar için en etkili tedavi sürecinin belirlenmesinde fayda sağlar.
- **Test Sonuçlarının Tahmini**: Yapılan testler sonucunda elde edilen veriler üzerinden, belirli tıbbi sonuçlar önceden tahmin edilebilir.
- **Ürün ve Hizmet Geliştirme**: Sağlık sektöründeki ürün ve hizmetlerin iyileştirilmesine yönelik veriler sunarak sağlık hizmet kalitesini artırır.

Ülkemizdeki sağlık sektöründe hem kamu hem de özel sektördeki sağlık profesyonellerine kararlarında destek amaçlı perspektifler yaratacak veri madenciliği çözümlerine örnekler aşağıda yer almaktadır (Koyuncugil ve Özgülbaş, 2009: 28-30);

- **Veri Ambarı Oluşturma**: Sağlık kuruluşlarında toplanan büyük miktardaki veri içinde, temiz ve analiz edilebilir verilere hızlı erişim sağlamak önemli bir ihtiyaçtır. Özellikle klinik ve demografik bilgilerin düzenlenmesi zaman ve maliyet açısından zorluklar yaratmaktadır.
- **Elektronik Hasta Dosyalarının Oluşturulması**: Hastaların tıbbi geçmişleri, tanı ve tedavi süreçleri, laboratuvar sonuçları gibi verilerin tek bir dosyada toplanması, zamana dayalı olarak düzenlenmesi sağlık hizmeti sunumunda kritik önem taşır.
- **Veri Kalitesi Sorunlarının Giderilmesi**: Sağlık verilerinde karşılaşılan kayıp veri, tutarsızlıklar ve aşırı uç değerler gibi sorunların çözümü için çeşitli teknikler kullanılmaktadır. Sorunlu kayıtların silinmesi veya ortalama gibi sabit değerler atanması gibi yaklaşımlar bu kapsamda ele alınmaktadır.
- **Kronik Hastalıklar İçin Erken Uyarı Sistemleri**: Kronik hastalıkların artan sıklığı ve getirdiği mali yük göz önüne alındığında, erken teşhis ve önleyici çözümler geliştirilmesi önem kazanmaktadır.
- **Laboratuvar Testleri ve Hata Tespiti**: Sağlık hizmetlerinde ortaya çıkan hata ve suistimallerin ayrımı yapılmalı, riskler en aza indirilmeli ve buna uygun önlemler alınmalıdır. Bu da hasta güvenliği için önemli bir adımdır.
- **Klinik Karar Destek Sistemleri**: Hekimlerin tanı ve tedavi sürecinde destek alabileceği veri ambarı tabanlı sistemlerin oluşturulması gereklidir. Bu sistemler hekimlere veri, çözüm önerileri ve risk analizlerini otomatik olarak sunarak karar alma sürecini destekler.
- **Hasta Odaklı Sağlık Hizmeti ve Kalite Yönetimi**: Sağlık hizmeti kalitesi hasta memnuniyeti ile ölçülmektedir. Bu nedenle kalite göstergeleri hasta gruplarının özelliklerine, sigorta durumuna ve klinik hizmetlere göre değerlendirilmelidir.

- **Hizmet Sunumunun Optimizasyonu**: Ulusal ve bölgesel düzeyde, hizmet sunum bileşenlerinin en uygun şekilde planlanması, kaynak kullanımını etkin hale getirmektedir.
- **Suistimallerin ve Fatura Yolsuzluklarının Önlenmesi**: Sağlık alanında yaygın olan fatura yolsuzluğu ve yeşil kart suistimali gibi uygulamalar büyük ekonomik kayıplara yol açmaktadır. Bu nedenle otomasyona dayalı gözetim sistemleri, insan kaynaklı denetimlerin yerine geçerek suistimallerin önlenmesine katkı sunar.
- **Maliyet Optimizasyonu**: Sağlık hizmeti maliyetlerini kontrol altına alabilmek için maliyete etki eden faktörlerin belirlenmesi ve maliyet düşürücü çözümler üretilmesi önemlidir.
- **Finansal Performans ve Risk Yönetimi**: Sağlık kurumlarında finansal performansın izlenmesi ve risklerin erken belirlenmesi için finansal uyarı sistemleri kullanılabilir. Bu sistemler finansal krizlerin önüne geçilmesinde etkin rol oynar.
- **Yönetsel Karar Destek Sistemleri**: Sağlık yöneticileri, sağlık kuruluşlarının daha etkili, verimli ve kaliteyi koruyarak yönetilmesi amacıyla veri tabanlı karar destek sistemlerine ihtiyaç duymaktadır.

2.4. Hasta Profillerinin Çıkartılması İçin Bir Simülasyon

2.4.1. Araştırmanın Sınırları

Bu çalışma, veri madenciliği yöntemlerinden K-Means algoritması kullanılarak hasta profillemesi üzerine odaklanılmıştır. Araştırma kapsamında aşağıdaki sınırlar ve kısıtlamalar belirlenmiştir.

1. **Veri Kaynakları ve Kapsam:** Çalışmada kullanılan veri seti, belirli bir popülasyonu temsil eden veri kümesi ile sınırlıdır. Bu durum, farklı popülasyonlar için genelleme yapmayı kısıtlayabilir.
2. **Metodolojik Sınırlamalar:** Araştırmada yalnızca K-Means kümeleme algoritması kullanılmış olup, diğer kümeleme teknikleri veya makine öğrenmesi yöntemleri incelenmemiştir. Alternatif yöntemlerin sonuçlar üzerindeki etkisi değerlendirilmemiştir.
3. **Örneklem Sınırlamaları:** Çalışmada analiz edilen veri, [örneğin yaş grubu, coğrafi bölge veya sağlık durumu] gibi belirli demografik ve sağlık göstergeleri ile sınırlandırılmıştır. Bu, sonuçların genelleştirilebilirliğini sınırlayabilir.
4. **Zaman ve Kaynak Kısıtlamaları:** Araştırma, belirli bir süre ve kaynaklarla gerçekleştirilmiştir. Daha geniş bir analiz veya daha detaylı bir çalışma için ek zamana ve kaynağa ihtiyaç duyulabilir.
5. **Veri Kalitesi ve Eksiklikler:** Veri setinde eksik veya yanlış değerler bulunma ihtimali, analizin doğruluğunu etkileyebilecek bir faktördür. Veri temizleme süreci bu sorunları minimize etmeye çalışsa da tüm eksikliklerin giderildiği garanti edilemez.

2.4.2. Araştırmanın Amacı

Bu çalışma, sağlık verileri üzerinde K-Means kümeleme algoritmasını kullanarak hasta profillemesi yapmayı amaçlamaktadır. Hastaların yaş, vücut kitle indeksi (BMI), şeker seviyesi, hipertansiyon, kalp hastalığı ve sigara içme durumu gibi temel sağlık göstergeleri üzerinden segmentlere ayrılması, her hasta grubunun belirgin özelliklerinin analiz edilmesine olanak sağlar. Çalışmanın nihai amacı, bu segmentasyona dayanarak farklı hasta grupları için kişiselleştirilmiş tedavi ve önleme stratejileri geliştirilmesidir.

2.4.3.Problem Tanımı

Sağlık sektöründe hasta verilerinin büyük veri kümeleri içerisinde kaybolması, kritik sağlık sorunlarının erken teşhis ve yönetimini zorlaştırmaktadır. Hastaların kişiselleştirilmiş tedavi planlarına ulaşabilmesi için verilerin doğru bir şekilde analiz edilmesi ve risk gruplarının belirlenmesi önem arz etmektedir. Bu tezde, veri madenciliği tekniklerinden biri olan K-Means algoritması ile hasta segmentasyonu gerçekleştirilmiş, farklı sağlık riskleri taşıyan gruplar belirlenmiştir.

2.4.4. Veri Seti

Sağlık sektöründe verilerin gizliliği ve hasta mahremiyeti en önemli önceliklerden biri olduğu için, etik kurallar ve yasal düzenlemelere uygun hareket edilmesi gerekmektedir. Bu bağlamda, gerçek hasta verilerinin korunması amacıyla simülasyon tabanlı bir veri seti hazırlanmıştır. Bu simülasyon verileri, analizlerin ve modellemelerin yürütülmesine olanak tanırken, kişisel bilgilerin gizliliğini koruma ve etik standartlara uygunluğu sağlama amacı taşımaktadır. Bu yöntem, sağlık sektöründeki verilerin güvenli bir şekilde kullanılmasını sağlayan önemli bir yaklaşımdır.

Bu çalışmada kullanılan veri seti, çeşitli demografik ve klinik bilgileri içermektedir. Veri seti aşağıdaki değişkenlerden oluşmaktadır:

- **Yaş:** Hasta yaş aralığı. (1 – 82)
- **Hipertansiyon:** Hipertansiyon var (1) veya yok (0).
- **Kalp Hastalığı:** Kalp hastalığı var (1) veya yok (0).
- **Vücut Kitle İndeksi (VKİ):** (10,3 – 97,6).
- **Şeker Seviyesi:** (55,12 – 271,74).
- **Sigara İçme Durumu:** Sigara içiyor (1) veya içmiyor (0).

Referans Aralıkları

Vücut Kitle İndexsi

- 18.5 ve altı: Zayıf.
- 18.5- 24.9: Normal kilolu.
- 25.0- 29.9: Fazla kilolu.
- 30.0- 34.9: Obez (1. derece obezite)
- 35.0- 39.9: Aşırı obez (2. derece obezite)
- 40 ve üstü: Morbid obez (3. derece obezite)

Açlık Kan Şekeri

- **Normal:** 70-99 mg/dL
- **Prediyabet :** 100-125 mg/dL
- **Diyabet:** ≥ 126 mg/dL

2.4.5. K-Means Kümeleme Algoritması

Bu çalışmada K-Means algoritması kullanılmıştır. K-Means, verileri k adet küme merkezine (centroid) en yakın olacak şekilde bölerek gözlemleri kümeler. Elbow yöntemi kullanılarak optimal küme sayısı belirlenmiştir. Bu yönteme göre 4 küme en uygun sonuçları vermiştir. Veriler standartlaştırıldıktan sonra K-Means algoritması uygulanmıştır. Çalışma platformu olarak visual studio code seçilmiş dil olarak python dili kullanılmıştır.

2.4.5.1. Analiz Adımları

Analizde Kullanılan Kütüphaneler ve Metodlar

Bu çalışmada Python programlama dilinde, veri analizi ve makine öğrenmesi alanında yaygın olarak kullanılan kütüphaneler ile çeşitli sınıflar ve metodlar kullanılmıştır. Aşağıda, bu kütüphanelerin kullanımı detaylandırılmaktadır:

1. Pandas

Pandas kütüphanesi, veri işleme ve analiz işlemlerinde yaygın olarak kullanılan bir kütüphanedir. Veriler, DataFrame veri yapısında tutulmuş ve aşağıdaki metodlar kullanılmıştır:

- **read_csv():** CSV formatında saklanan veri setini okuma ve bir DataFrame yapısına dönüştürme amacıyla kullanılmıştır.
- **isnull():** Veri setinde eksik (NaN) değerlerin tespit edilmesi amacıyla kullanılmıştır.
- **sum()**:Eksik verilerin sütunlar bazında toplamını hesaplamak için kullanılmıştır.
- **fillna()**: Eksik verilerin, ilgili sütunun ortalaması ile doldurulmasında kullanılmıştır.
- **select_dtypes():** Veri setindeki belirli veri tipine sahip sütunların seçilmesi amacıyla kullanılmıştır.
- **groupby():** Verinin belirli sütunlara göre gruplanması amacıyla kullanılmıştır.
- **mean():** Gruplandırılmış verinin ortalamalarını hesaplamak için kullanılmıştır.

2. Scikit-Learn (sklearn)

Scikit-learn kütüphanesi, makine öğrenmesi algoritmalarının uygulanmasında kullanılmıştır. Çalışmada aşağıdaki sınıflar ve metodlar tercih edilmiştir:

Sınıflar:

- **StandardScaler:** Veriyi standartlaştırmak amacıyla kullanılmıştır. Verinin ortalaması 0 ve standart sapması 1 olacak şekilde dönüştürülmesi sağlanmıştır.
- **KMeans:** K-Means algoritması kullanılarak veri kümeleme işlemleri gerçekleştirilmiştir. Veriyi belirli sayıda kümelere ayırmak için kullanılmıştır.
- **PCA (Principal Component Analysis):** Temel Bileşen Analizi kullanılarak veri setinin boyutları azaltılmış ve iki boyutta görselleştirilmiştir.

Metodlar:

- **fit_transform():** Veriyi standartlaştırmak için kullanılmış ve modele uygun hale getirilmiştir.
- **fit():** K-Means modelini veriye uygulayarak eğitme işlemi gerçekleştirilmiştir.

- **inertia_:** K-Means algoritması için toplam kare hatası (SSE) hesaplanmıştır.
- **labels_:** Kümelere atanmış etiketler alınmıştır.

3. Matplotlib

Matplotlib kütüphanesi, görselleştirme işlemlerinde kullanılmıştır. Çalışmada kullanılan grafikleri oluşturmak amacıyla aşağıdaki metodlar tercih edilmiştir:

- **figure():** Yeni bir grafik figürü oluşturmak için kullanılmıştır.
- **plot():** Dirsek yöntemi ile K-Means kümeleme analizinde kullanılan grafiği oluşturmak için kullanılmıştır.
- **xlabel(), ylabel(), title():** Grafikte eksen isimleri ve başlık eklemek için kullanılmıştır.
- **grid():** Grafikte daha iyi bir görselleştirme için ızgara eklemek amacıyla kullanılmıştır.
- **show():** Oluşturulan grafikleri görüntülemek için kullanılmıştır.
- **subplot():** Çoklu grafikler oluşturmak için kullanılmıştır.
- **suptitle():** Grafiklerin üstünde başlık eklemek için kullanılmıştır.
- **tight_layout():** Grafiklerin düzgün hizalanması için kullanılmıştır.

4. Seaborn

Seaborn kütüphanesi, daha gelişmiş ve estetik görselleştirmeler için kullanılmıştır. Aşağıdaki metodlar çalışmada kullanılmıştır:

- **scatterplot():** PCA ile boyutları azaltılmış verinin K-Means sonuçlarına göre dağılım grafiği oluşturulmuştur.
- **histplot():** Her bir kümedeki özelliklerin dağılımı histogram grafikleri ile görselleştirilmiştir.

5. Mlxtend

Mlxtend kütüphanesi, veri madenciliği tekniklerinin uygulanmasında kullanılmıştır. Özellikle sık görülen itemset'leri ve ilişkilendirme kurallarını belirlemek amacıyla aşağıdaki sınıflar ve metodlar kullanılmıştır:

- **Apriori:** Sık itemset'leri bulmak amacıyla Apriori algoritması uygulanmıştır.
- **Association_rules():** Apriori algoritmasından elde edilen itemset'ler ile ilişkilendirme kuralları çıkarılmıştır.

2.4.5.2.Veri Hazırlama Süreci

Bu çalışmada, veri hazırlama süreci kapsamında eksik verilerin temizlenmesi ve sayısal değişkenlerin standartlaştırılması işlemleri gerçekleştirilmiştir. Bu iki aşama, veri madenciliği ve makine öğrenmesi süreçlerinde veri setinin kalitesini artırmak ve modelin performansını optimize etmek amacıyla uygulanmıştır.

Veri setlerinde eksik verilerin bulunması, analiz ve modelleme süreçlerinde yanlı sonuçlara yol açabileceği için bu verilerin uygun yöntemlerle ele alınması gerekmektedir. Bu çalışmada, her bir sütunda eksik veri olup olmadığı kontrol edilmiştir. Eksik veri içeren sütunlar tespit edildikten sonra, bu sütunlardaki eksik gözlemler ortalama ile doldurma (mean imputation) yöntemi kullanılarak tamamlanmıştır. Ortalama ile doldurma işlemi, eksik veri oranının düşük olduğu durumlarda güvenilir bir yöntem olarak kabul edilmektedir, çünkü veri setindeki genel eğilimleri bozmadan eksik gözlemleri tamamlamayı mümkün kılar. Böylece, veri setindeki boş gözlemler, ilgili sütunun ortalama değeri ile doldurularak analizdeki eksik veri sorunları giderilmiştir.

Kümeleme algoritmaları gibi mesafe tabanlı makine öğrenmesi yöntemlerinde, değişkenlerin farklı ölçeklerde olması algoritmanın performansını olumsuz etkileyebilir. Özellikle K-Ortalama (K-Means) gibi algoritmalarda, büyük ölçekli değişkenler, küçük ölçekli değişkenlere kıyasla mesafe hesaplamalarında daha fazla etki yaratabilir. Bu nedenle, sayısal değişkenlerin ölçeklenmesi büyük önem taşımaktadır. Çalışmamızda, sadece sayısal değişkenler seçilerek standartlaştırma işlemi gerçekleştirilmiştir. Bu işlem, Z-puanı normalizasyonu yöntemi ile yapılmış

olup, her bir değişkenin ortalaması 0, standart sapması ise 1 olacak şekilde ölçeklendirilmiştir. Bunun için StandardScaler fonksiyonu kullanılmıştır. Standartlaştırma işlemi, değişkenlerin oransal farklılıklarını ortadan kaldırarak modelin tüm değişkenlere eşit ağırlık vermesini sağlar. Standartlaştırma işlemi sonrası veri seti, her bir özelliğin ortalaması 0, standart sapması 1 olacak şekilde yeniden ölçeklendirilmiştir. Bu işlem, modelin daha doğru ve güvenilir sonuçlar vermesini sağlayarak, analiz sürecine katkıda bulunmuştur.

```python
import pandas as pd
from sklearn.preprocessing import StandardScaler

file_path = 'C:/Users/P1296/Desktop/veri_seti.csv'
data = pd.read_csv(file_path, delimiter=';')

print("Eksik Veriler:")
print(data.isnull().sum())

for column in data.columns:
    if data[column].isnull().sum() > 0:
        data[column].fillna(data[column].mean(), inplace=True)

print("\nEksik Veriler Temizlendikten Sonra:")
print(data)

sayisal_sutunlar = data.select_dtypes(include=['int64', 'float64']).columns

scaler = StandardScaler()
veri_scaled = scaler.fit_transform(data[sayisal_sutunlar])

veri_scaled_df = pd.DataFrame(veri_scaled, columns=sayisal_sutunlar)

print("\nStandartlaştırılmış Veri:")
print(veri_scaled_df)
```

Şekil 2.5.1. Veri Hazırlama

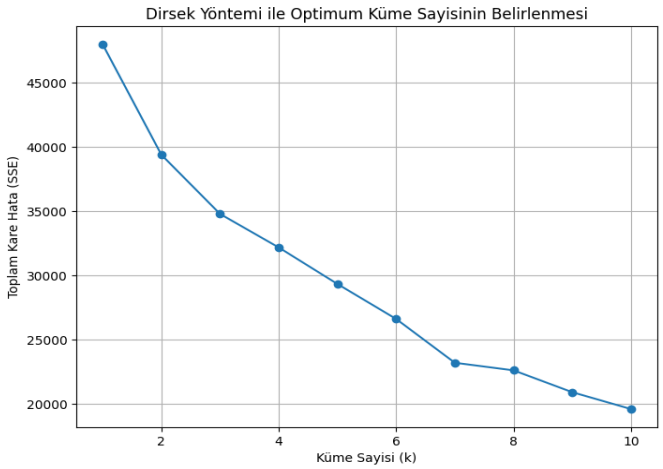
2.4.6. K-Means Algoritması ile Kümeleme

2.4.6.1.Elbow Yöntemi ile En Uygun K Değerinin Belirlenmesi

K-Means algoritmasında K değeri (küme sayısı) genellikle kullanıcı tarafından önceden belirlenir; ancak veri yapısına uygun en iyi K değerinin seçilmesi modelin performansı açısından büyük önem taşır. Bu çalışmada, en uygun küme sayısının belirlenmesi amacıyla Elbow Yöntemi kullanılmıştır. Elbow yöntemi, farklı K değerleri için K-Means algoritmasını çalıştırarak her bir K değeri için Hata Kareler Toplamı (Sum of Squared Errors, SSE) değerlerini hesaplar. SSE, veri noktalarının kendi küme merkezlerine olan uzaklıklarının kareleri toplamı olarak ifade edilir.

```python
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

file_path = 'C:/Users/P1296/Desktop/veri_seti.csv'
data = pd.read_csv(file_path, delimiter=';')

scaler = StandardScaler()
veri_scaled = scaler.fit_transform(data)

sse = []
k_range = range(1, 11)

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=0)
    kmeans.fit(veri_scaled)
    sse.append(kmeans.inertia_)

plt.figure(figsize=(8, 6))
plt.plot(k_range, sse, marker='o')
plt.xlabel('Küme Sayısı (k)')
plt.ylabel('Toplam Kare Hata (SSE)')
plt.title('Dirsek Yöntemi ile Optimum Küme Sayısının Belirlenmesi')
plt.grid(True)
plt.show()
```

Şekil 2.3.3.1. Elbow Metodu Kodu



Şekil 2.3.3.2. Elbow Sayısı

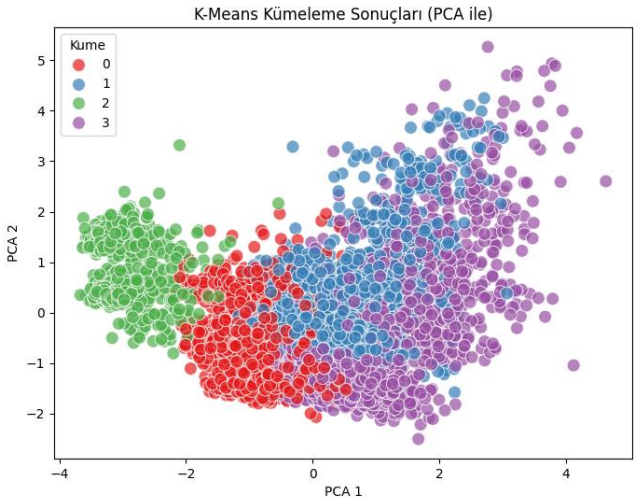
Şekil 2.3.3.2'deki grafikte SSE'nin K değerine karşı çizimi sonucu, belirli bir noktada SSE'deki azalma hızının yavaşladığı bir "dirsek" noktası gözlemlenir. Bu noktaya elbow (dirsek) adı verilir ve grafikte bu dirsek noktası, en uygun küme sayısının işaretçisi olarak kabul edilir. Bu çalışmada, Elbow Yöntemi kullanılarak en uygun K değeri 4 olarak belirlenmiş ve bu değere göre kümelenme işlemi optimize edilmiştir.

2.4.6.2. K-Means Algoritması ile Kümeleme

Veri setinin hazırlanmasının ardından, gözlemlerin kümelere ayrılması amacıyla K-Means kümeleme algoritması uygulanmıştır. K-Means, kullanıcı tarafından belirlenen k sayıda küme oluşturarak gözlemleri bu kümelere atayan bir kümeleme yöntemidir. Bu çalışmada, algoritma n_clusters = 4 parametresi ile çalıştırılmıştır, bu da veri setinin dört ayrı küme altında gruplandırılacağı anlamına gelmektedir. K-Means algoritmasının temel işleyişi, her veri noktasını, kendisine en yakın küme merkezi (centroid) ile ilişkilendirerek kümelere dağıtmaktır. Bu süreçte her bir veri noktası, minimum mesafe ilkesine göre bir küme merkezine atandığından, gözlemler arasında benzerlik temelinde ayrımlar yapılır. Çalışma sonuçları, veri setine eklenen "Küme" sütununda her bir gözlemin hangi kümeye ait olduğunu gösteren etiketler aracılığıyla temsil edilmiştir. Böylece, veri setindeki tüm gözlemler dört ana kümeye ayrılarak sınıflandırılmıştır.

```python
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns


file_path = 'C:/Users/P1296/Desktop/veri_seti.csv'
data = pd.read_csv(file_path, delimiter=';')

scaler = StandardScaler()
veri_scaled = scaler.fit_transform(data)

kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(veri_scaled)

data['Kume'] = kmeans.labels_

pca = PCA(n_components=2)
pca_components = pca.fit_transform(veri_scaled)

data['PCA1'] = pca_components[:, 0]
data['PCA2'] = pca_components[:, 1]

plt.figure(figsize=(8, 6))
sns.scatterplot(x='PCA1', y='PCA2', hue='Kume', data=data, palette='Set1', s=100, alpha=0.7)
plt.title('K-Means Kümeleme Sonuçları (PCA ile)')
plt.xlabel('PCA 1')
plt.ylabel('PCA 2')
plt.show()
```

Şekil 2.5.3.1. K-Means Kümeleme Kodu



Şekil 2.5.3.2. K-Means Kümeleme Sonuçları

2.4.6.3. Grafik Açıklaması

Yüksek boyutlu veriyi görselleştirmek amacıyla, iki bileşenli bir Ana Bileşen Analizi (PCA) uygulanmış ve veri iki boyutta temsil edilmiştir. Grafik, bu iki bileşen üzerinde farklı kümeleri renk kodlamasıyla göstererek küme dağılımını sunmaktadır. Görselleştirmede kullanılan renkler, kümelerin birbirine olan mesafelerini ve kümeler arasındaki ayrışmaları daha net bir şekilde göstermektedir. Özellikle, kümeler arasında belirgin sınırların varlığı, algoritmanın gözlemleri başarılı bir şekilde farklı gruplara ayırdığını ortaya koymaktadır. Her renk belirli bir kümeyi temsil etmektedir:

- **Küme 0 (Kırmızı)**: Veri noktaları ağırlıklı olarak negatif PCA1 ve PCA2 eksenlerinde yer almakta.
- **Küme 1 (Mavi)**: Kırmızı kümeden biraz daha sağda, pozitif PCA1 ve hafif negatif PCA2 ekseninde kümelenmiş.
- **Küme 2 (Yeşil)**: Grafiğin sol tarafında, negatif PCA1 ekseninde yer alıyor ve PCA2'de ise daha yaygın.
- **Küme 3 (Mor)**: Grafiğin sağ tarafında, pozitif PCA1 ekseninde yoğunlaşmış ve diğer kümelerden biraz daha ayrık bir biçimde bulunmakta.

2.4.6.4. Kümelerin Yorumlanması

Kümeleme algoritması sonucunda, her kümenin bazı özellikler açısından homojen olduğu varsayılır. Grafik, kümeler arasındaki ayrışmayı görselleştirmekte ve kümelerin nasıl dağıldığını anlamamıza yardımcı olmaktadır.

- **Küme 0 ve Küme 1**: Bu iki küme, PCA1 ekseni üzerinde yakın olmakla birlikte PCA2 ekseninde daha belirgin bir ayrışma göstermekte. Küme 0 daha sıkışık ve daha merkezi bir dağılım sergilerken, Küme 1'in yayılımı daha fazla.
- **Küme 2**: Küme 0 ve Küme 1'den oldukça ayrık. Veri noktalarının çoğu sol tarafta ve grafiğin kenarlarına yayılmış durumda. Bu küme, diğer kümelere kıyasla daha spesifik bir profil sunabilir.
- **Küme 3**: Küme 3'ün noktaları grafiğin sağ tarafında yoğunlaşmış durumda ve PCA1 ekseninde daha yüksek değerlere sahip. Bu kümenin diğer kümelerden farklı bir demografik ya da sağlık profilini temsil ettiği söylenebilir.

2.4.6.5. Her Küme İçin Ortalama Hasta Profillerinin Oluşturulması

```
kolonlar = ['Yas', 'Hipertansiyon', 'Kalp', 'Seker_Seviyesi', 'Vucut_Kitle_Indexsi', 'Sigara_Icme_Durumu']

kume_ortalama_filtered = data.groupby('Kume')[kolonlar].mean()

print(kume_ortalama_filtered)
```

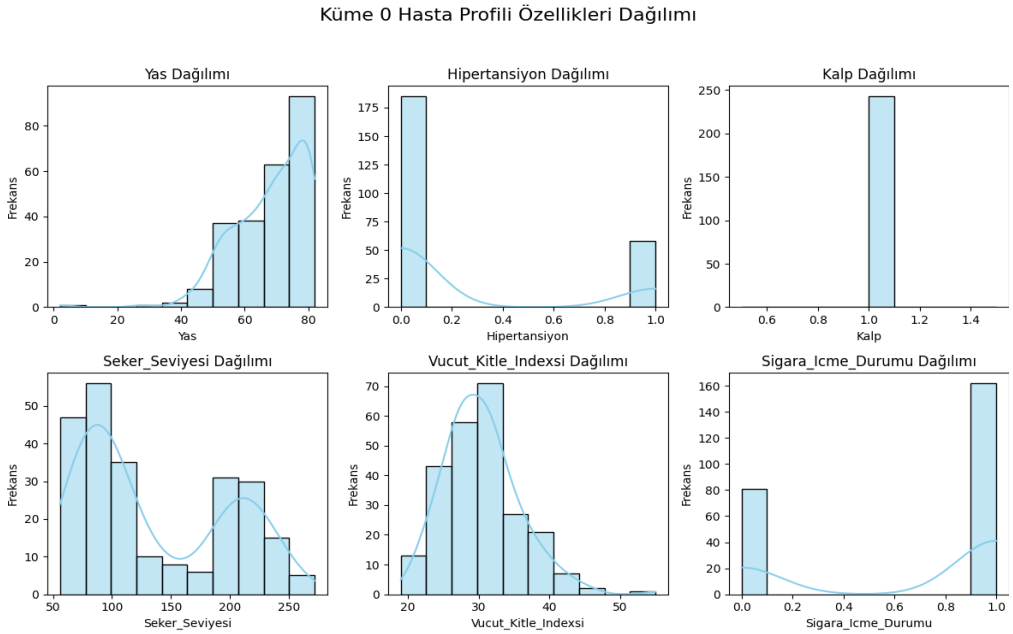
Şekil 2.5.4.1. Her Küme İçin Ortalama Hasta Profili

Kume	Yas	Hipertansiyon	Kalp	Seker_Seviyesi	Vucut_Kitle_Indexsi	Sigara_Icme_Durumu
0	68.275720	0.238683	1.0	135.374486	30.366255	0.666667
1	37.407807	0.000000	0.0	92.578073	27.603405	1.000000
2	61.547445	0.573723	0.0	164.635036	34.363504	0.550365
3	42.002051	0.000000	0.0	94.190704	29.145591	0.000000

Şekil 2.5.4.2. Her Küme İçin Ortalama Hasta Profil Sonuçları

```python
import seaborn as sns
import matplotlib.pyplot as plt

features = ['Yas', 'Hipertansiyon', 'Kalp', 'Seker_Seviyesi', 'Vucut_Kitle_Indexsi', 'Sigara_Icme_Durumu']

kume_0 = data[data['Kume'] == 0]

plt.figure(figsize=(12, 8))

for i, feature in enumerate(features, 1):
    plt.subplot(2, 3, i)
    sns.histplot(kume_0[feature], kde=True, bins=10, color='skyblue')
    plt.title(f'{feature} Dağılımı')
    plt.xlabel(f'{feature}')
    plt.ylabel('Frekans')

plt.suptitle('Küme 0 Hasta Profili Özellikleri Dağılımı', fontsize=16)
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()
```

Şekil 2.5.4.3. Küme 0 İçin Hasta Profil Dağılım Kodu



Şekil 2.5.4.4. Küme 0 İçin Hasta Profili Özellikleri Dağılımı

Küme 0

Bu kümenin demografik ve sağlık profili, yaşlı bireylerden oluşan yüksek risk grubunu temsil etmektedir. Hipertansiyon, kalp hastalığı, yüksek kan şekeri, obezite ve sigara kullanımı gibi ciddi sağlık riskleri, bu kümedeki bireylerin kronik hastalık komplikasyonlarına yatkın olduklarını göstermektedir. Bu grupta uygulanabilecek önleyici ve tedavi edici müdahaleler oldukça önemlidir. Kümenin özellikleri ve bu bağlamda alınabilecek önlemler aşağıda detaylandırılmıştır:

1. **Yaş:** Bu kümenin ortalama yaşı 68 olup, bireylerin çoğunlukla yaşlılardan oluştuğunu göstermektedir. Yaşlı bireylerde kronik hastalıkların ve buna bağlı komplikasyonların daha sık görülmesi, bu grupta düzenli sağlık kontrollerinin yapılmasını ve hastaların yakından izlenmesini gerektirir. Yaş faktörü, bu bireylerin sağlık durumlarının korunması ve komplikasyon risklerinin azaltılması için önemli bir etkendir.
2. **Hipertansiyon:** Bu kümede %23.8 oranında hipertansiyon görülmektedir. Hipertansiyon, yaşlı bireylerde ciddi kardiyovasküler olaylar ve böbrek fonksiyon bozuklukları gibi komplikasyonlara yol açabilir. Bu nedenle, kan basıncının düzenli olarak izlenmesi ve gerekli durumlarda medikal tedavi veya yaşam tarzı değişiklikleri ile kontrol altına alınması önemlidir. Hipertansiyon yönetimi, bu grubun kardiyovasküler sağlığını korumada kilit bir rol oynar.
3. **Kalp Rahatsızlığı:** Tüm bireylerin (%100) kalp hastalığına sahip olması, bu kümenin sağlık açısından en yüksek riskli gruplardan biri olduğunu göstermektedir. Kalp rahatsızlıkları, bu grupta düzenli tedavi, takip ve yaşam tarzı değişikliklerini zorunlu kılar. Kardiyovasküler sağlık açısından bu grubun ilaç uyumunun sağlanması, sigara gibi risk faktörlerinin azaltılması ve fiziksel aktivitenin (sağlık durumuna göre) teşvik edilmesi önemlidir. Bu grup, aynı zamanda ileri tedavi seçeneklerinin göz önünde bulundurulması gereken bir risk profiline sahiptir.
4. **Şeker Seviyesi:** Bu kümede ortalama kan şekeri seviyesi 135 mg/dL olup, diyabet veya prediyabet durumu göstergesi olarak değerlendirilebilir. Diyabetin kontrol altına alınması, kalp rahatsızlığı olan yaşlı bireyler için kritik önem taşır, çünkü yüksek kan şekeri seviyeleri, kardiyovasküler komplikasyon riskini daha da artırabilir. Bu nedenle, diyabet yönetimi için düzenli kan şekeri kontrolleri yapılmalı, beslenme düzeni gözden geçirilmeli ve gerektiğinde ilaç tedavisi uygulanmalıdır.
5. **Vücut Kitle İndeksi (VKİ):** Kümedeki bireylerin ortalama VKİ değeri 30.36 olup, obez kategorisine girmektedir. Obezite, bu grupta diyabet, hipertansiyon ve kalp hastalığı gibi mevcut sağlık sorunlarını daha da kötüleştirebilir. Yaşlı bireyler için güvenli ve kontrollü kilo kaybı stratejileri geliştirilmesi, obeziteye bağlı komplikasyonları azaltmaya yardımcı olabilir. Kilo yönetimi

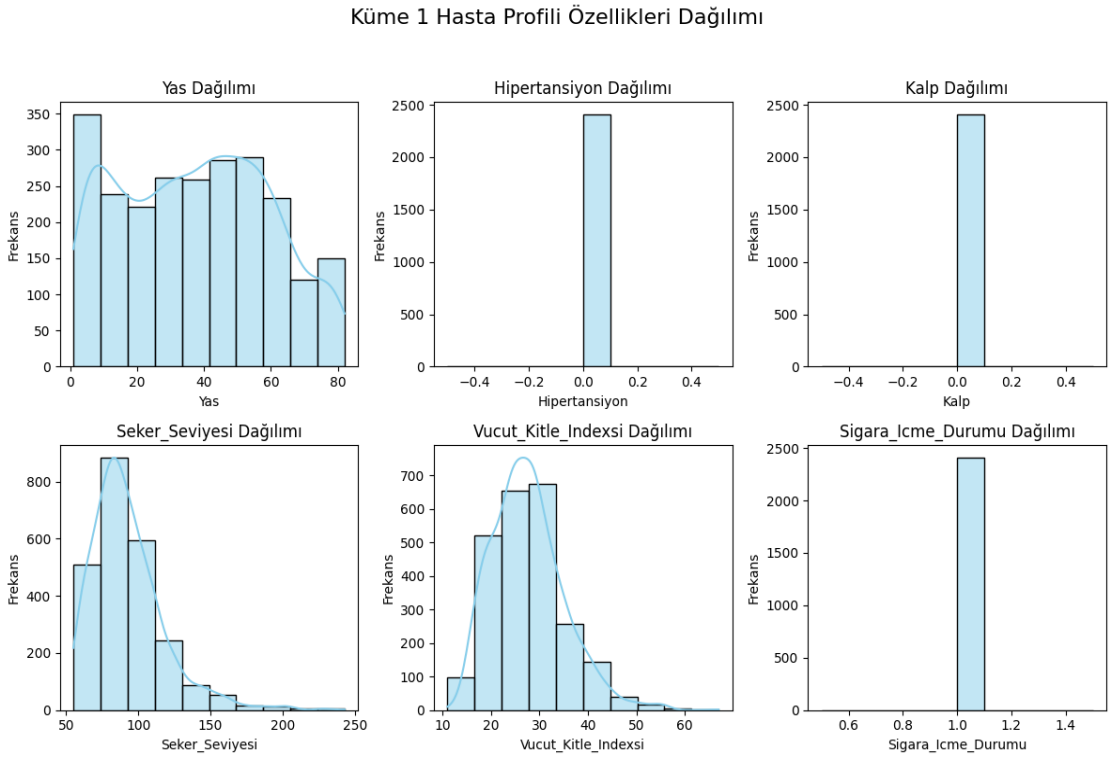
programları, bu grup için daha sağlıklı bir yaşam tarzı geliştirilmesinde önemli bir adım olacaktır.

6. **Sigara İçme Durumu:** Hastaların yaklaşık %66'sı sigara kullanmaktadır. Sigara kullanımı, özellikle kalp rahatsızlığı olan yaşlı bireylerde ciddi komplikasyonlara yol açabilir. Sigara bırakma programlarının bu grupta öncelikli olarak uygulanması, bireylerin yaşam kalitesini artırabilir ve komplikasyon risklerini azaltabilir. Bu nedenle, sigara bırakma konusunda destekleyici programlar ve davranışsal müdahaleler acilen devreye alınmalıdır.

Küme 0 Değerlendirme: Bu küme, yaş, kalp rahatsızlığı, hipertansiyon, yüksek kan şekeri ve obezite gibi birçok ciddi sağlık riskine sahip yaşlı bireylerden oluşmaktadır. Bu risk faktörleri göz önüne alındığında, bu grubun sağlık durumunun düzenli olarak izlenmesi, ilaç uyumunun sağlanması, sağlıklı yaşam tarzı alışkanlıklarının teşvik edilmesi ve gerekirse ileri tedavi seçeneklerinin değerlendirilmesi kritik önem taşır. Bu tür bir bütüncül yaklaşım, bu bireylerin yaşam kalitesini iyileştirebilir ve sağlık maliyetlerini azaltarak sağlık sistemine katkıda bulunabilir.

```
import seaborn as sns
import matplotlib.pyplot as plt

features = ['Yas', 'Hipertansiyon', 'Kalp', 'Seker_Seviyesi', 'Vucut_Kitle_Indexsi', 'Sigara_Icme_Durumu']

kume_1 = data[data['Kume'] == 1]

plt.figure(figsize=(12, 8))

for i, feature in enumerate(features, 1):
    plt.subplot(2, 3, i)
    sns.histplot(kume_1[feature], kde=True, bins=10, color='skyblue')
    plt.title(f'{feature} Dağılımı')
    plt.xlabel(f'{feature}')
    plt.ylabel('Frekans')

plt.suptitle('Küme 1 Hasta Profili Özellikleri Dağılımı', fontsize=16)
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()
```

Şekil 2.5.4.5. Küme 1 İçin Hasta Profili Dağılım Kodu



Şekil 2.5.4.6. Küme 1 İçin Hasta Profili Özellikleri Dağılımı

Küme 1

Bu kümenin demografik ve sağlık profili, genç yaş grubundaki bireyler arasında potansiyel risk faktörlerini ortaya koymaktadır. Yaş, kan basıncı, kalp rahatsızlığı ve kan şekeri düzeyleri açısından sağlıklı bir tablo sunan bu küme, özellikle yüksek sigara içme oranı ile dikkat çekmektedir. Aşağıda, bu kümenin özellikleri ve önerilen müdahaleler detaylı olarak değerlendirilmiştir:

1. **Yaş:** Bu kümenin ortalama yaşı 37 olup, bireyler genel olarak daha genç yaş grubunda yer almaktadır. Genç yaş grubundaki bireylerde kronik hastalık riskleri genellikle daha düşük olsa da, mevcut yaşam tarzı faktörleri ileriki yaşlarda sağlık üzerinde olumsuz etkiler yaratabilir. Dolayısıyla, risk faktörlerinin erken dönemde belirlenmesi ve kontrol altına alınması, ileride karşılaşılabilecek sağlık sorunlarının önlenmesi açısından önem taşır.
2. **Hipertansiyon:** Kümede hipertansiyon vakalarının bulunmaması, bu bireylerin kardiyovasküler sağlık açısından olumlu bir profile sahip

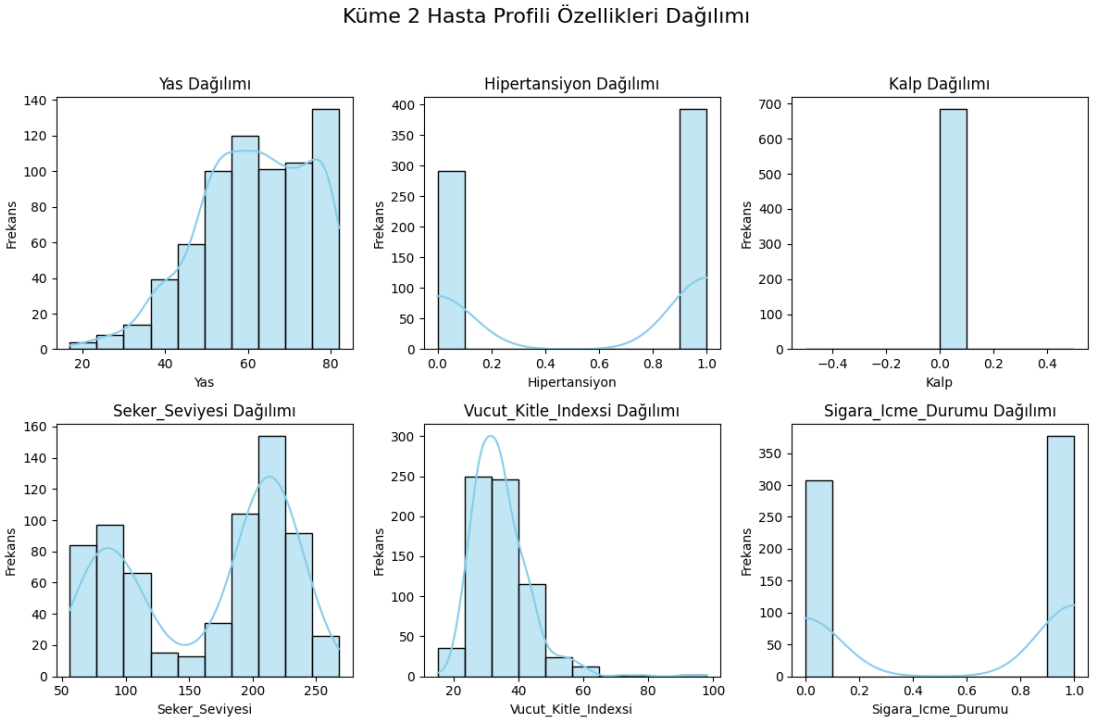
olduklarını göstermektedir. Ancak, yüksek sigara kullanımı gibi yaşam tarzı faktörleri uzun vadede kan basıncı ve kardiyovasküler sağlık üzerinde olumsuz etkiler yaratabilir. Bu nedenle, düzenli tansiyon kontrolleri, olası hipertansiyon gelişim riskini izlemek açısından yararlı olacaktır.

3. **Kalp Rahatsızlığı:** Bu kümede mevcut kalp hastalığı olan bireylerin bulunmaması, genç yaş grubu ve genel sağlık profiliyle uyumludur. Ancak, sigara içme oranının %100 olması, kardiyovasküler riskin uzun vadede artabileceğine işaret eder. Bu durumda, kalp sağlığını korumak için sigara kullanımının azaltılması veya tamamen bırakılması teşvik edilmelidir.
4. **Şeker Seviyesi:** Kümedeki ortalama kan şekeri seviyesi 92 mg/dL olup, normal sınırlar içinde yer almaktadır. Bu durum, bu bireylerin diyabet riski taşımadığını göstermektedir. Ancak, ilerleyen yaşlarda kan şekeri seviyelerinin değişebileceği göz önüne alınarak düzenli kan şekeri kontrolü yapılması önerilmektedir.
5. **Vücut Kitle İndeksi (VKİ):** Kümedeki bireylerin ortalama VKİ değeri 27.06 olup, bu grup hafif fazla kilolu olarak kabul edilmektedir. Hafif kilolu olmalarına rağmen, bu bireylerin obeziteye ilerlemelerini önlemek amacıyla sağlıklı yaşam tarzı müdahaleleri yapılması yararlı olacaktır. Dengeli beslenme ve düzenli fiziksel aktivite programları, ilerleyen yaşlarda obeziteye bağlı risklerin azaltılmasında etkili bir rol oynayabilir.
6. **Sigara İçme Durumu:** Bu kümede %100 oranında sigara içimi dikkat çekmektedir. Sigara kullanımı, başta kardiyovasküler hastalıklar olmak üzere, çeşitli sağlık sorunlarına yol açabilir. Genç yaş grubunda bu alışkanlığın devam etmesi, ilerleyen yaşlarda ciddi sağlık riskleri oluşturacaktır. Bu nedenle, sigara bırakma programları ve davranışsal müdahaleler bu grupta acilen uygulanmalıdır. Sigara bırakma programlarına erişim sağlanarak, bireylerin sigarayı bırakmaları teşvik edilmelidir.

Küme 1 Değerlendirme: Bu küme, genç yaş ve normal kan şekeri seviyeleri gibi olumlu sağlık göstergelerine sahip olmakla birlikte, %100 oranında sigara kullanımı ve hafif fazla kilolu olmaları nedeniyle gelecekteki sağlık riskleri açısından dikkatle izlenmelidir. Bu grupta erken yaşta başlatılacak sigara bırakma programları ve sağlıklı yaşam tarzı teşvikleri, ilerleyen yıllarda oluşabilecek kronik hastalık risklerini azaltmada kritik bir rol oynayacaktır. Erken dönemde yapılacak bu tür müdahaleler, uzun vadede sağlık maliyetlerini düşürmeye katkı sağlayacak ve bireylerin yaşam kalitesini artıracaktır.

```python
import seaborn as sns
import matplotlib.pyplot as plt

features = ['Yas', 'Hipertansiyon', 'Kalp', 'Seker_Seviyesi', 'Vucut_Kitle_Indexsi', 'Sigara_Icme_Durumu']

kume_2 = data[data['Kume'] == 2]

plt.figure(figsize=(12, 8))

for i, feature in enumerate(features, 1):
    plt.subplot(2, 3, i)
    sns.histplot(kume_2[feature], kde=True, bins=10, color='skyblue')
    plt.title(f'{feature} Dağılımı')
    plt.xlabel(f'{feature}')
    plt.ylabel('Frekans')

plt.suptitle('Küme 2 Hasta Profili Özellikleri Dağılımı', fontsize=16)
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()
```

Şekil 2.5.4.7. Küme 2 İçin Hasta Profili Dağılım Kodu



Şekil 2.5.4.8. Küme 2 İçin Hasta Profili Özellikleri Dağılımı

Küme 2

Bu kümenin demografik ve sağlık profili, belirgin risk faktörlerini ve sağlık açısından önem arz eden parametreleri içermektedir. Yaş, hipertansiyon, yüksek kan şekeri düzeyleri, obezite ve sigara kullanımı gibi özellikler, bu bireylerin çeşitli kronik hastalıklara ve komplikasyonlara yatkınlığını artırmaktadır. Bu faktörler aşağıdaki şekilde analiz edilebilir:

1. **Yaş:** Kümedeki bireylerin ortalama yaşı 61 olup, çoğunlukla yaşlı bireylerden oluşmaktadır. İleri yaş grubundaki bireylerin kronik hastalık geliştirme riskinin daha yüksek olduğu bilinmektedir. Özellikle kardiyovasküler hastalıklar, diyabet ve hipertansiyon gibi hastalıkların görülme sıklığı bu yaş grubunda artmaktadır. Bu demografik faktör, kümedeki bireylerin düzenli sağlık kontrolleri ile yakından izlenmesi gerektiğine işaret eder.
2. **Hipertansiyon:** Kümede %57.3 gibi yüksek bir oranda hipertansiyon görülmektedir. Hipertansiyon, kalp hastalıkları ve inme gibi ciddi komplikasyonlara yol açabilen, ancak düzenli takip ve uygun tedavi ile kontrol altına alınabilen bir risk faktörüdür. Bu oran, hipertansiyon yönetimi ve kan

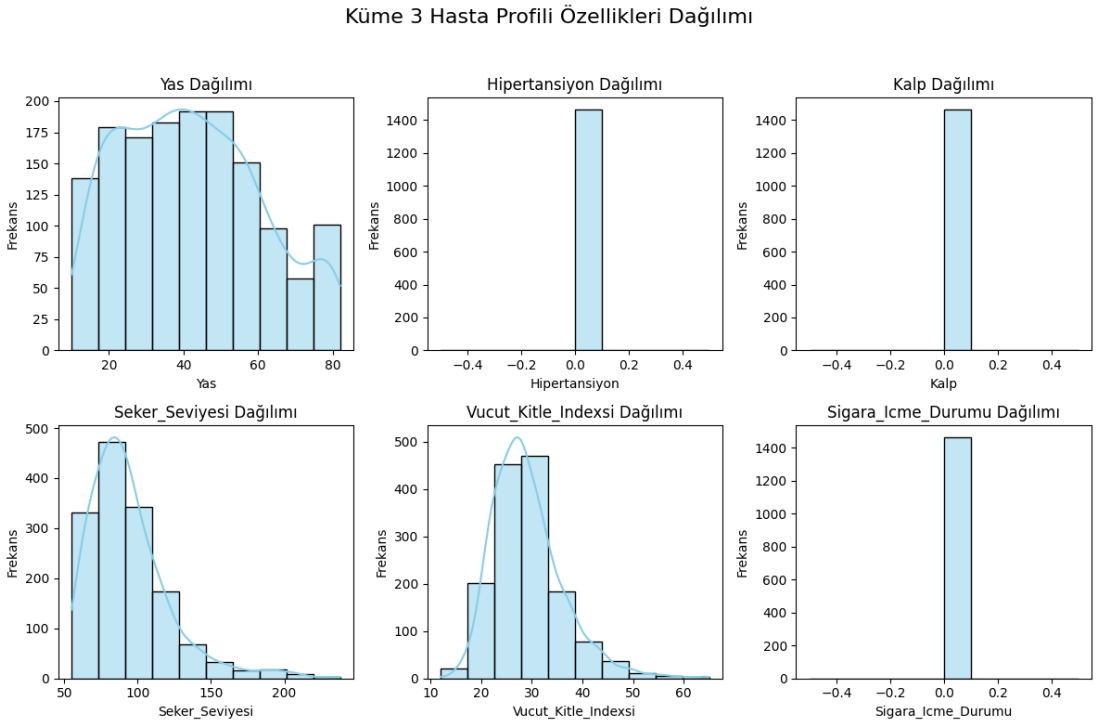
basıncı kontrolünün bu grupta kritik bir gereklilik olduğunu göstermektedir. Ayrıca, hipertansiyonun uzun vadede sağlık hizmetleri üzerindeki yükünü azaltmak için düzenli izlem ve önleyici tedavi stratejileri bu grup için önemli olabilir.

3. **Kalp Rahatsızlığı:** Kümede mevcut kalp rahatsızlığı olan bireylerin bulunmaması, kalp hastalığının henüz bu grupta belirgin bir sorun teşkil etmediğini göstermektedir. Ancak, hipertansiyon ve yüksek yaş gibi kalp rahatsızlığı risk faktörlerinin mevcut olması, gelecekte bu grupta kardiyovasküler olayların gelişme olasılığını artırmaktadır. Dolayısıyla, kardiyovasküler risklerin erken tanı ve önleyici tedavi stratejileri ile kontrol altına alınması, bu grubun kalp sağlığını korumak için önemlidir.
4. **Şeker Seviyesi:** Kümenin ortalama kan şekeri seviyesi 164 mg/dL olarak saptanmış olup, bu değer yüksek kan şekeri sınırlarının üzerindedir. Bu durum, bireylerin ciddi diyabet riski taşıdığını ve diyabet yönetimi açısından acil önlemler alınması gerektiğini göstermektedir. Diyabet riski, bu bireylerin uzun vadeli sağlık sonuçlarını olumsuz yönde etkileyebilir ve diğer komplikasyonlara yol açabilir. Bu bağlamda, diyabet taraması, diyet ve egzersiz gibi yaşam tarzı değişikliklerinin teşvik edilmesi, bu grubun diyabet riskini azaltmada etkili bir strateji olarak değerlendirilebilir.
5. **Vücut Kitle İndeksi (VKİ):** Kümenin ortalama VKİ değeri 34.36 olup, bireylerin obezite sınırında bulunduğunu göstermektedir. Obezite, diyabet, hipertansiyon ve çeşitli kardiyovasküler rahatsızlıkların riskini artıran bir faktördür. Bu yüksek VKİ değeri, kümedeki bireylerin obezite ile ilişkili komplikasyonlardan korunması için kişiselleştirilmiş beslenme ve fiziksel aktivite planları geliştirilmesinin önemini vurgulamaktadır.
6. **Sigara İçme Durumu:** Bireylerin %55.3'ü sigara içmektedir. Sigara kullanımı, özellikle yaşlı ve hipertansiyon sorunu olan bireylerde kardiyovasküler hastalık riskini artırır. Sigara bırakma programlarının bu grup için öncelikli bir halk sağlığı müdahalesi olarak uygulanması, sağlık sonuçlarının iyileştirilmesi açısından oldukça faydalı olacaktır.

Küme 2 Değerlendirme: Bu küme, özellikle diyabet, hipertansiyon ve obezite gibi kronik durumlar açısından yüksek risk taşımaktadır. Mevcut risk faktörlerinin etkili bir şekilde yönetilmesi, bu bireylerin sağlık durumlarını iyileştirmede ve komplikasyonları önlemede önemli bir rol oynayacaktır. Bu bağlamda, düzenli taramalar, bireyselleştirilmiş sağlık müdahaleleri ve yaşam tarzı değişiklikleri gibi stratejiler, sağlık hizmetlerinin bu grupta daha etkin kullanılmasına ve uzun vadeli sağlık maliyetlerinin azaltılmasına katkı sağlar.

```python
import seaborn as sns
import matplotlib.pyplot as plt

features = ['Yas', 'Hipertansiyon', 'Kalp', 'Seker_Seviyesi', 'Vucut_Kitle_Indexsi', 'Sigara_Icme_Durumu']

kume_3 = data[data['Kume'] == 3]

plt.figure(figsize=(12, 8))

for i, feature in enumerate(features, 1):
    plt.subplot(2, 3, i)
    sns.histplot(kume_3[feature], kde=True, bins=10, color='skyblue')
    plt.title(f'{feature} Dağ□l□m□')
    plt.xlabel(f'{feature}')
    plt.ylabel('Frekans')

plt.suptitle('Küme 3 Hasta Profili Özellikleri Dağ□l□m□', fontsize=16)
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()
```

Şekil 2.5.4.9. Küme 3 İçin Hasta Profili Dağılım Kodu



Şekil 2.5.4.10. Küme 3 İçin Hasta Profili Özellikleri Dağılımı

Küme 3

Bu kümenin demografik ve sağlık profili, orta yaş grubunda sağlıklı yaşam tarzı göstergelerine sahip bireylerden oluştuğunu göstermektedir. Kalp rahatsızlığı, hipertansiyon ve sigara kullanımının olmaması, bu grubun genel sağlık durumu açısından avantajlı bir profil sergilediğini ortaya koymaktadır. Ancak, sınırda fazla kilolu olmaları, uzun vadeli sağlık risklerini göz ardı etmemek gerektiğine işaret eder. Bu küme aşağıdaki şekilde analiz edilebilir:

1. **Yaş**: Kümenin ortalama yaşı 42 olup, bireyler nispeten orta yaş grubunda yer almaktadır. Bu yaş grubunda, yaşam tarzı faktörleri ileri yaşlarda sağlık risklerinin oluşmaması için kritik bir rol oynar. Orta yaş, bireylerin sağlıklarını koruma ve gelecekteki riskleri azaltma açısından farkındalığın arttığı bir dönemdir. Dolayısıyla, bu grubun sağlıklı yaşam tarzı alışkanlıklarını sürdürmesi önerilir.
2. **Hipertansiyon**: Kümede hipertansiyon vakalarının olmaması, bu bireylerin kan basıncının kontrol altında olduğunu ve kardiyovasküler sağlık açısından avantajlı bir durumda olduklarını göstermektedir. Hipertansiyon olmaması, kardiyovasküler hastalık riskini azaltıcı bir faktör olarak değerlendirilmelidir. Ancak yaş ilerledikçe hipertansiyon riski artabileceğinden, düzenli kontrollerin yapılması önem taşır.
3. **Kalp Rahatsızlığı:** Kalp hastalığı olmayan bireylerden oluşan bu küme, genç-orta yaş aralığında sağlıklı bir kalp profilini yansıtmaktadır. Mevcut sağlık profili göz önünde bulundurulduğunda, kalp rahatsızlığı riski düşük görünmektedir. Bu sağlıklı durumun korunması için bireylerin fiziksel aktivite düzeylerinin sürdürülmesi ve dengeli bir diyetle desteklenmesi önemlidir.
4. **Şeker Seviyesi:** Kümenin ortalama kan şekeri seviyesi 94 mg/dL olup, normal sınırlar içerisindedir. Bu durum, bireylerin diyabet açısından düşük risk taşıdığını göstermektedir. Ancak, yaş ilerledikçe kan şekeri seviyelerinde artış olabileceğinden, kan şekeri düzeylerinin düzenli olarak izlenmesi tavsiye edilir. Diyabet riskini düşük tutmak için sağlıklı beslenme alışkanlıklarının teşvik edilmesi de bu grupta önemlidir.

5. **Vücut Kitle İndeksi (VKİ):** Kümenin ortalama VKİ değeri 29.14 olup, sınırda fazla kilolu kategorisine girmektedir. Bu bireyler obezite sınıfına henüz girmemiş olmakla birlikte, kilo kontrolünün yapılması uzun vadede obezite ve buna bağlı sağlık sorunlarının önlenmesi açısından önem taşır. Orta yaşta kilo kontrolünün sağlanması, ilerleyen yaşlarda obeziteye bağlı hastalıkların gelişme riskini azaltacaktır. Bu nedenle, bireylerin dengeli bir diyet ve düzenli fiziksel aktivite ile kilo yönetimini sürdürmeleri önerilir.
6. **Sigara İçme Durumu:** Bu kümede sigara kullanan birey bulunmaması, genel sağlık profili açısından olumlu bir faktördür. Sigara kullanmamak, kardiyovasküler hastalık, solunum yolu hastalıkları ve çeşitli kanser türleri gibi birçok ciddi sağlık sorunu riskini azaltır. Bu durum, kümenin sağlıklı bir yaşam tarzını sürdürdüğünü ve sağlık açısından olumlu bir örnek teşkil ettiğini göstermektedir.

Küme 3 Değerlendirme: Bu küme, orta yaş grubunda sağlıklı bir profil sergilemekle birlikte, sınırda fazla kilolu olmaları dikkat çekmektedir. Kalp rahatsızlığı, hipertansiyon ve sigara kullanımının olmaması, bu grubun düşük kardiyovasküler risk taşıdığını göstermektedir. Ancak, fazla kilolu olmaları, uzun vadede obeziteye bağlı komplikasyonların gelişme olasılığını artırabileceğinden, bu bireylerin kilo yönetimine odaklanmaları önemlidir. Dengeli beslenme ve düzenli fiziksel aktivitenin teşvik edilmesi, bu kümedeki bireylerin uzun vadeli sağlıklarını koruma ve komplikasyon riskini azaltma açısından faydalı olacaktır. Bu profil, sağlık hizmetlerinin kaynaklarını daha etkin kullanma imkanı sunarken, bireylerin yaşam kalitesini artırma fırsatı sağlamaktadır.

Genel Değerlendirme: Sağlık Riskleri ve Profilleme

Veri madenciliği ile yapılan bu kümeleme analizi, sağlık kurumlarında farklı hasta gruplarının sağlık ihtiyaçlarına göre maliyetleri ve kaynak kullanımını optimize etme fırsatları sunar. Her bir kümenin taşıdığı riskler ve sağlık durumları, özelleştirilmiş müdahaleler gerektirmekte olup, bu müdahaleler sayesinde sağlık kurumu kaynaklarının etkin bir şekilde yönetilmesi mümkündür. Bu analiz, sağlık kurumlarında önleyici sağlık hizmetleri, tedaviye yönelik kaynak tahsisi ve uzun vadeli maliyet yönetimi stratejilerini destekler.

Kümeleme analizi sonucunda elde edilen dört hasta grubu, yaş, sağlık durumu, yaşam tarzı ve diğer faktörlere göre farklı sağlık riskleri taşımaktadır. Bu kümeler arasında yaşa dayalı farklılıklar dikkat çekici olup, her grubun yaşına özgü sağlık riskleri ve yaşam tarzı alışkanlıkları belirleyici olmuştur.

1-Yaş Faktörü ve Kronik Hastalıklar

Kümeleme analizi, yaşa dayalı sağlık risklerini belirlemekte ve yaşlı bireylerin, özellikle kronik hastalık yönetimi açısından hastane kaynakları üzerinde yoğun bir yük oluşturduğunu göstermektedir.

- **Küme 0 ve Küme 2:** Bu kümelerde yaşlı bireylerin yüksek oranda yer alması ve kronik hastalıkların (kalp hastalığı, hipertansiyon, diyabet, obezite) yaygın olması, bu gruplara yönelik düzenli izlem ve tedavi gereksinimini ortaya koyar. Bu gruplar, hastane kaynakları üzerinde yüksek maliyetli tedavi ve bakım hizmetleri talep eder. Sağlık kurumları, bu grupların izlenmesi ve yönetimi için kaynaklarını yoğunlaştırarak uzun vadeli maliyetleri düşürebilir. Kronik hastalık yönetimi stratejilerinin önceliklendirilmesi, ileriye dönük komplikasyonların önlenmesine ve hasta başına düşen maliyetin azalmasına katkıda bulunur. Aynı zamanda, bu bireylere yönelik bakım programları geliştirilerek, yatak doluluk oranları ve kaynak talebi dengeli bir şekilde yönetilebilir.

- **Küme 1 ve Küme 3:** Bu gruplar, daha genç ve orta yaş gruplarını içerdiği için şu anda yüksek maliyetli tedavi gerektirmemektedir. Ancak, Küme 1'deki bireylerin tamamının sigara kullanması, ilerleyen yaşlarda ciddi sağlık sorunları ve buna bağlı tedavi maliyetlerini artırabilir. Bu nedenle, sigara bırakma programlarının devreye alınması, gelecekte sağlık kurumları kaynaklarının üzerindeki yükü hafifletir. Küme 3 ise nispeten sağlıklı bireylerden oluştuğu için önleyici sağlık hizmetlerine yönelik düşük maliyetli programlarla uzun vadeli sağlık sorunlarının önüne geçer. Bu strateji, hastane kaynaklarının etkin kullanılmasını sağlar.

2-Obezite ve Diyabet Riski

Obezite ve diyabet riski taşıyan hastalar, sağlık hizmetlerinde sürekli bakım ve izlem gerektiren hasta profilleri oluşturur. Hastane kaynaklarının obezite ve diyabet yönetimine yönlendirilmesi, bu gruplardaki komplikasyonların azaltılmasını ve ileri tedavi maliyetlerinin önlenmesini sağlar.

- **Küme 2:** Obezite ve diyabet riski en yüksek olan gruptur. Bu grup için obezite ve diyabet yönetimi stratejilerinin uygulanması, hastane maliyetlerini azaltmak adına kritik öneme sahiptir. Bu bireyler için bireyselleştirilmiş kilo yönetimi programları, diyet danışmanlığı ve diyabet kontrolüne yönelik sürekli bakım, hastane maliyetlerini azaltıcı bir rol oynayabilir. Bu tür hizmetler, diyabet ve obeziteye bağlı komplikasyonları engelleyerek, bu grupta uzun vadeli sağlık maliyetlerinin düşürülmesine katkıda bulunur.

- **Küme 0:** Bu grupta da obezite oranları yüksektir ve kalp hastalığı ile hipertansiyon gibi risk faktörleriyle birleştiğinde, bu bireyler hastane kaynakları açısından önemli bir yük oluşturur. Obezite yönetiminin yapılması, hastane kaynaklarının ileri seviyedeki kardiyovasküler hastalık tedavisine ayrılmasını azaltır. Bu durum, hem hastane maliyetlerini kontrol altına alır hem de yatak kapasitesinin daha verimli kullanılmasına olanak tanır.

3-Sigara Kullanımı ve Uzun Vadeli Riskler

Sigara kullanımı, uzun vadede ciddi sağlık sorunlarına yol açarak hastane kaynaklarını önemli ölçüde tüketen bir risk faktörüdür. Sigara kaynaklı hastalıkların önlenmesi, hastanenin maliyetlerini ve kaynak kullanımını doğrudan etkileyebilir.

- **Küme 1 ve Küme 2:** Bu gruplarda yüksek oranda sigara kullanımı görülmektedir. Sigara kullanımı, akciğer hastalıkları, kalp hastalıkları ve çeşitli kanser türlerinin oluşumuna yol açarak hastane kaynaklarını uzun vadede yüksek maliyetli tedavilere yönlendirebilir. Bu nedenle, hastane yönetimi, sigara bırakma programlarına yatırım yaparak, gelecekte ortaya çıkabilecek kronik hastalıkları önlemeyi hedeflemelidir. Böylece, sigaraya bağlı hastalıkların tedavi maliyetleri düşürülerek kaynaklar daha verimli bir şekilde kullanılabilir. Bu gruplara yönelik önleyici müdahaleler, uzun vadeli maliyet yönetimi açısından etkili olacaktır.
- **Küme 0:** %66 oranında sigara kullanımının mevcut olduğu bu grupta, sigara bırakma programlarının aciliyeti vurgulanmalıdır. Kalp rahatsızlığı olan bu bireylerde sigara kullanımı, sağlık durumunu daha da kötüleştirerek tedavi maliyetlerini artırır. Sigara bırakma programlarına yapılacak yatırım, özellikle bu yüksek risk grubunda ileri evre sağlık sorunlarının önlenmesine katkıda bulunur ve kaynak kullanımını optimize eder.

4-Genç ve Sağlıklı Gruplar

Sağlık risklerinin düşük olduğu gruplar, hastane maliyetleri açısından önemli bir fırsat sunar. Genç ve sağlıklı bireylerde, sağlık sorunlarının önlenmesi odaklı düşük maliyetli programlar, hastanenin uzun vadeli kaynak kullanımını optimize eder.

- **Küme 3:** Bu grupta henüz ciddi sağlık sorunları görülmemekle birlikte, yaş ilerledikçe risk faktörleri artabilir. Bu nedenle, düşük maliyetli önleyici sağlık hizmetleriyle bu bireylerin sağlıklı durumlarını korumaları desteklenebilir. Sağlıklı bireyler için sağlıklı yaşam tarzı teşvik programları, uzun vadede hastane kaynaklarının daha etkin kullanılmasını sağlar. Bu strateji, sağlık sorunları henüz oluşmadan hastane kaynaklarının yükünü azaltır ve maliyetleri kontrol altına alır.

- **Küme 1:** Genç yaşlarına rağmen tamamı sigara kullanan bu grup, ilerleyen yaşlarda kronik hastalıklar geliştirme potansiyeline sahiptir. Sigara bırakma müdahaleleri, gelecekte hastane kaynaklarını tüketebilecek sağlık sorunlarını önceden engelleyerek maliyetleri düşürür. Bu gruba yönelik erken müdahale stratejileri, hastanenin uzun vadeli maliyet yönetimi ve kaynak kullanımı açısından önemlidir.

Hastane İşletmeleri İçin Stratejik Öneriler

1. **Kronik Hastalık Yönetimine Odaklanma:** Yaşlı ve kronik hastalıkları olan hastalara yönelik kronik hastalık yönetimi programlarının güçlendirilmesi, uzun vadeli komplikasyonları ve hastane maliyetlerini azaltacaktır. Kronik hastalıkların takibi için entegre bakım modelleri geliştirilerek, yatak doluluk oranları ve ileri tedavi ihtiyaçları minimize edilebilir.
2. **Önleyici Sağlık Hizmetlerine Yatırım Yapma:** Genç ve orta yaş gruplarında, sağlıklı yaşam tarzı teşvikleri, kilo yönetimi programları ve sigara bırakma destekleri gibi önleyici sağlık hizmetlerine yapılacak yatırımlar, hastanelerinin uzun vadeli kaynaklarını koruyarak maliyetleri düşürür. Bu tür programlar, hastaneye başvuruların önlenmesine ve kaynakların sürdürülebilir şekilde kullanılmasına olanak tanır.
3. **Sigara Bırakma Programlarına Ağırlık Verme:** Sigara kullanımı olan gruplarda sigara bırakma programları uygulayarak, hastane kaynaklarının ileri düzey akciğer ve kalp hastalıkları tedavisine yönlendirilmesi azaltılabilir. Bu programlara yapılacak yatırım, sigara kullanımının önüne geçilerek, kaynakların daha verimli bir şekilde kullanılmasını sağlar.
4. **Obezite ve Diyabet Yönetimi İçin Kaynak Ayırma:** Obezite ve diyabet riski yüksek gruplar için bireyselleştirilmiş kilo yönetimi ve diyabet izleme programlarına kaynak ayrılması, uzun vadede hastane kaynaklarının daha etkin kullanılmasını ve ileri aşamalardaki komplikasyonların önlenmesini sağlar.

SONUÇ

Bu çalışma, veri madenciliği teknikleri ve K-Means kümeleme algoritması kullanarak hastaların demografik ve sağlık durumlarına göre farklı gruplara ayrılmasını sağlamış ve hasta profillemesinin sağlık hizmetlerine sunduğu katkıları gözler önüne sermiştir. Yapılan analiz sonucunda elde edilen dört farklı hasta profili, yaş, sağlık durumu ve yaşam tarzı faktörlerine dayalı olarak belirlenmiştir. Bu profilleme çalışması, her bir kümenin özgün sağlık ihtiyaçlarına göre özelleştirilmiş müdahale stratejilerinin geliştirilmesine olanak tanımaktadır. Böylece, sağlık hizmetlerinin hedeflenmiş ve daha etkili bir şekilde sunulması sağlanarak, klinik sonuçların iyileştirilmesi ve kaynak kullanımının optimize edilmesi hedeflenmiştir.

1. Yaşlı ve Kronik Hastalık Riski Yüksek Gruplar

Kümeleme analizi sonucunda, özellikle Küme 0 ve Küme 2'de yer alan yaşlı bireylerin, kalp rahatsızlıkları, hipertansiyon, diyabet ve obezite gibi kronik hastalıklarla karşı karşıya olduğu görülmüştür. Bu gruplar, hastane kaynakları üzerinde en yoğun yükü oluşturan ve sürekli bakım gerektiren bireyleri temsil etmektedir. Kronik hastalıklara sahip bu bireylerde sağlık yönetiminin temel hedefleri:

- **Erken teşhis ve izlem:** Kardiyovasküler hastalıklar ve diyabet gibi kronik hastalıklar açısından yüksek risk taşıyan bu bireylerde erken teşhis ve düzenli izlem, komplikasyonların önlenmesine yardımcı olacaktır.
- **Önleyici sağlık eğitimi:** Yaşlı bireylerin sağlık bilincini artıracak ve yaşam tarzı değişikliklerini teşvik edecek eğitim programları, hastaneye başvuruların ve uzun vadeli sağlık maliyetlerinin azalmasına katkı sağlayacaktır.

Bu grupta yapılacak izlem ve bakım programlarına yönelik yatırımlar, hastane kaynaklarının daha verimli kullanılmasını ve komplikasyonların önlenmesi yoluyla sağlık sisteminin sürdürülebilirliğini artıracaktır.

2. Genç ve Orta Yaş Sağlıklı Gruplar

Küme 1 ve Küme 3 ise daha genç ve orta yaş bireylerden oluşan, şu anda önemli sağlık sorunları yaşamayan grupları içermektedir. Ancak bu gruplarda özellikle sigara kullanımı ve kilo kontrolü gibi yaşam tarzı faktörleri uzun vadede sağlık risklerini artırabilir. Bu gruplar için sağlık hizmetlerinde şu stratejiler ön plana çıkmaktadır:

- **Koruyucu sağlık hizmetleri ve yaşam tarzı müdahaleleri:** Sigara bırakma programları, sağlıklı beslenme ve düzenli fiziksel aktivitenin teşvik edilmesi gibi koruyucu sağlık önlemleri, uzun vadede bu grupların sağlık durumlarını koruyarak hastane kaynaklarının yükünü hafifletecektir.
- **Düzenli sağlık taramaları:** Bu bireylerin yaş ilerledikçe kronik hastalık geliştirme riskleri dikkate alınarak düzenli sağlık taramalarının yapılması, potansiyel sağlık sorunlarının erken tespit edilmesini sağlayacaktır.

Bu gruptaki bireyler için düşük maliyetli önleyici programların uygulanması, sağlık harcamalarının uzun vadede azalmasına katkı sağlar ve sağlık hizmetlerinin sürdürülebilirliğini destekler.

3. Sigara Kullanımı ve Uzun Vadeli Riskler

Çalışma, özellikle Küme 1 ve Küme 2'de yer alan bireylerin yüksek oranda sigara kullanması nedeniyle uzun vadeli sağlık riskleri taşıdığını ortaya koymaktadır. Sigara kullanımı, akciğer hastalıkları, kalp hastalıkları ve kanser gibi ciddi sağlık sorunlarının riskini artırarak hastane kaynaklarını ileri düzey tedavi gereksinimlerine yönlendirme potansiyeline sahiptir. Bu durumda:

- **Sigara bırakma programlarının uygulanması:** Bu gruplarda sigara kullanımını azaltmaya yönelik davranışsal müdahaleler, sağlık sonuçlarını iyileştirmenin yanı sıra, ileride ortaya çıkabilecek yüksek maliyetli tedavi ihtiyaçlarını önlemek için önemlidir.
- **Sigara kaynaklı risklerin izlenmesi ve yönetimi:** Bu bireylerde sigaraya bağlı sağlık sorunlarının önceden tespiti, sağlık sisteminin daha planlı ve maliyet-etkin bir şekilde yönetilmesine katkıda bulunur.

Sigara bırakma müdahaleleri, yalnızca bireylerin sağlıklarını korumakla kalmaz, aynı zamanda hastane kaynaklarının etkin yönetimi ve uzun vadeli maliyet tasarrufu sağlamak açısından da stratejik bir yatırım alanıdır.

4. Sağlıklı Gruplar İçin Koruyucu Sağlık Hizmetleri

Çalışmanın sonuçlarına göre, Küme 3'teki nispeten sağlıklı bireyler, şu anda belirgin sağlık sorunlarına sahip olmamakla birlikte yaş ilerledikçe risk faktörleri taşıma olasılıkları artacaktır. Bu bireylerde sağlıklı yaşam alışkanlıklarının sürdürülmesi, uzun vadeli sağlık sonuçlarının iyileştirilmesine yardımcı olacaktır. Bu grup için önerilen stratejiler:

- **Sağlıklı yaşam tarzı teşvik programları**: Sağlıklı bireylerin, sigara kullanımı gibi olumsuz alışkanlıklardan kaçınarak yaşam tarzlarını sürdürebilmeleri için teşvik edilmeleri gerekmektedir.
- **Düzenli sağlık taramaları ve risk değerlendirmeleri:** Bu gruptaki bireylerde sağlık riskleri henüz ortaya çıkmadığı için düşük maliyetli ve rutin sağlık taramaları, hastalık gelişme olasılığını en aza indirir.

Sağlık sistemi açısından, bu grupta önleyici sağlık hizmetlerinin yaygınlaştırılması, hastane kaynaklarının uzun vadede korunmasını sağlar ve sağlık harcamalarının azalmasına katkıda bulunur.

Sağlık Yönetimi ve Kaynak Kullanımının Optimizasyonu

Bu çalışmanın bulguları, hasta profillemesinin sağlık hizmetlerinin etkin ve hedeflenmiş bir şekilde sunulması açısından önemli katkılar sağladığını göstermektedir. Farklı hasta gruplarının belirlenmesi ve her grubun ihtiyaçlarına yönelik müdahaleler geliştirilmesi, hastane yönetiminin kaynakları daha verimli kullanmasına olanak tanır. Özellikle:

- **Kronik hastalık yönetimi**: Yüksek riskli bireylerin önceden tespit edilmesi ve sürekli bakım gereksinimlerinin optimize edilmesi, sağlık maliyetlerini düşürerek kaynakların etkin kullanımını sağlar.
- **Koruyucu ve önleyici hizmetlerin yaygınlaştırılması:** Sağlıklı bireylerde risk faktörlerinin yönetimi ve sağlıklı yaşam teşvikleri, hastane maliyetlerini uzun vadede azaltarak sağlık sisteminin sürdürülebilirliğini destekler.

Bu çalışma, veri madenciliği tekniklerinin kullanımıyla elde edilen hasta profillemesinin, hastaların sağlık risklerini anlamak ve bu risklere yönelik uygun müdahaleleri geliştirmek için güçlü bir araç olduğunu göstermiştir. Yaş, yaşam tarzı alışkanlıkları (sigara kullanımı, kilo kontrolü) ve kronik hastalıklar (diyabet, hipertansiyon, kalp hastalıkları) gibi faktörler hasta gruplarını belirlemede önemli rol oynamaktadır. Elde edilen bulgular, hem bireysel sağlık yönetimi hem de sağlık politikalarının şekillendirilmesi açısından değerlidir. Gelecekte yapılacak çalışmalar, daha geniş veri setleri ve gelişmiş analiz yöntemleri ile bu profillemenin sağlık sistemine katkılarını artıracak ve hasta merkezli sağlık politikalarının geliştirilmesine yardımcı olacaktır.

KAYNAKÇA

Akküçük U, (2011). *Veri Madenciliği - Kümeleme ve Sınıflama Algoritmaları*, 1. Basım, Yalın Yayıncılık, İstanbul.

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

Alan M, (2012). *Veri Madenciliği Ve Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama*. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 33, 165-174.

Albayrak M, (2008). *EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci İle Tespiti*. Doktora Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

Almuallim, H., Dietterich, T.G., *Learning with Many Irrelevant Features*, in the 9th National Conference on Artificial Intelligence, USA, 547–552, 1991

Alpaydın E., (2000). *Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*, Bilişim 2000 Eğitim Semineri, 1-10.

Argüden Y, Erşahin B, (2008). *Veri madenciliği veriden bilgiye masraftan değere*. ARGE danışmanlık.

Aslan Evren, (2008). *EMG İşaretlerinin İncelenmesi Ve Veri Madenciliği Uygulaması*. Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

Avcı K, Çınaroğlu S, (2015). *Hasta Yatışlarının Majör Tanı Sınıflaması Bakımından Görsel Veri Madenciliği Yöntemi Kullanılarak Sınıflandırılması*. Uluslararası Hakemli Akademik Spor Sağlık Ve Tıp Bilimleri Dergisi, 14, 110-123.

Ayık YZ, Özdemir A, Yavuz U, (2007). *Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi*. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10, 441-454.

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

Chae YM, Seung Hee Ho, Kyoung Won Cho, Dong Ha Lee, Sun Ha Ji , *Data Minig approach to policy analysis in a healt insurance domain*, 2001, s:168-171.

Chris R, Jyun-Chen W, David C. (2002) Yen, *Data Mining Techniques for Customer Relationship Management*, Technology in Society, 4, 2002, 488

Cios K., Pedrycz, Witold, Swiniars ki, Roman W. and Kurgan, Lukasz A. (2007). *Data Mining Knowledge Discovery Aapproach* , Eespringer Science Businesss Media,USA .

Çalışkan B., *Veri Madenciliği ve Müşteri İlişkileri Yönetimi*, Bilişim Teorisi Ders Notları, 2006

Çarklı B, (2010*). Sağlık Sektöründe Apriori Algoritması İle Bir Veri Madenciliği Uygulaması*. Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

Çerkezi M, (2013). *Veri Madenciliği Yöntemlerini Kullanarak Diyabetik Retinopati Hastalığının Teşhisi*. Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya.

Demirel B, (2008). *Meme Kanseri Tedavi Yöntemlerinin Veri Madenciliği İle Belirlenmesi*. Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Isparta.

Doğan Ş, (2007). *Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi*. Yüksek Lisans Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ.

Elmas F, (2014). *Kalp Krizi Riskinin Bir Veri Madenciliği Uygulaması İle Analizi*. Yüksek Lisans Tezi, Muğla Sıtkı Koçman Üniversitesi Fen Bilimleri Enstitüsü, Muğla.

Emel GG, Taşkın Ç, (2002). *Genetik algoritmalar ve uygulama alanları*, Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 1:129-152.

Erkuş S, (2015*). Veri Madenciliği Yöntemleri İle Kardiyovasküler Hastalık Tahmini Yapılması*. Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Ertuğrul İ, Organ A, Şavlı A, (2013). *Veri Madenciliği Uygulamasına İlişkin PAÜ Hastanesinde Hasta Profilinin Belirlenmesi*. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 19, 97-103.

Farboudi S, (2009). *Tıp Bilişiminde İstatistiksel Veri Madenciliği*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Fayez MA., (2018). "*Diagnoses of Coronary Heart Disease (Chd) Using Data Mining Techniques Based on Classification*". Ulusal tez merkezi, 520268: 1-54

Gazi V, (2007). *Veri Madenciliğinde Duyarlılık*. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Gülten A, Doğan Ş, (2008). *Genetik Algoritmalar Yönteminin Biyomedikal Veriler Üzerindeki Uygulamaları*. DAUM, 7, 12-16.

Gündoğdu ÖE, (2007). *Veri Madenciliğinde Genetik Algoritmalar*, Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli.

Gürsoy TŞ, (2012). *Uygulamalı Veri Madenciliği ve Sektörel Analizler (3. Basım).* Ankara: Pegem Akademi.

Gürsoy TŞ, (2009). *Veri Madenciliği ve Bilgi Keşfi (1. Basım)*. Ankara: Pegem Akademi.

Han, J., Pei, J., Yın, Y. (2000), *Mining Frequent Patterns Without Candidate Generation,* In Proceedings of the 2000 ACMSIGMOD International Conference on Management of Data (SIGMOD(00), Dallas, TX, pp:1–12, 2000

Irmak S, (2009). *Veri Madenciliği Yöntemleri ile Sağlık Sektörü Veri Tabanlarında Bilgi Keşfi: Tanımlayıcı Ve Kestirimci Model Uygulamaları*. Doktora Tezi, Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü, Antalya.

Kaya H, Köymen K, (2008). *Veri Madenciliği Kavramı ve Uygulama Alanları*. Doğu Anadolu Bölgesi Araştırma ve Uygulama Dergisi, 6, 159-164.

Kovacic I, (2022). *OLAP Patterns: A pattern-based approach to multidimensional data analysis.* Data & Knowledge Engineering, 138, 2022, 1.

Keskin M, (2013). *Spor Ve Yaşam Merkezleri Üzerine Veri Madenciliği Çalışması*. Yüksek Lisans Tezi, Okan Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Kocabaş F, (2010). *Veri Madenciliği Süreci ve Gerçek Bir Veri Seti Üzerinde Uygulanması*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

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

Kök B, Kuloğlu N, (2005). *Sollama Esnasında Taşıt ve Yol ile İlgili Faktörlerin Karar Ağacı Yöntemi ile İrdelenmesi*, Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 21, 180-188.

Köktürk F, Ankaralı H, Sümbüloğlu V, (2009). *Veri Madenciliği Yöntemlerine Genel Bakış*. Türkiye Klinikleri J Biostat, 1, 5-20.

Kudyba, S. (2004), *Managing Data Mining*, CyberTech Publishing, 146-163.

Kusiak A, K.H. Kernstine, J.A.Kern, K.A.McLaughlin and T.L.Tseng, *Medical and Engineering Case Studies*, May, 2000, s:103-107.

Liu, Z.; Du, Z.; Chen, H.; Zou, C. (2012) *Large Area Land Cover Classıfıcatıon Wıth Landsat Etm+ Images Based On Decısıon Tree*, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXII ISPRS Congress, Melbourne, Australia.

Luba T., Lasocki, R. (1994). *On Unknown Attribute Values in Functional Dependencies*, T.Y. Lin ed., The International Workshop on Rough Sets and Soft Computing, San Jose, California, 490-497, 1994

Mellor, J., Stone, Michael A. and Keane, J., (2018), *Application of Data Mining to a Large Hearing-Aid Manufacturer's Dataset to Identify Possible Benefits for Clinicians, Manufacturers, and Users, University of Manchester*, Kilburn Building, Oxford Road, Manchester, M13 9PL, UK, (2018)

Oğuzlar, A. (2003). *Veri Önişleme*. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21 (Temmuz-Aralık), 67-76.

Orhunbilge N, (2002). *Uygulamalı Regresyon ve Korelasyon Analizi (1. Basım)*. İ.Ü. İşletme Fakültesi Yayınları.

Öğüt S, (2002). *Veri Madenciliği Kavramı Ve Gelişim Süreci.* Yeditepe Üniversitesi, İstanbul.

Özbay E, (2007). *Finans Sektöründe Veri Madenciliği İle Dolandırıcılık Tespiti*. Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

Özbay Ö, (2015). *Veri Madenciliği Kavramı Ve Eğitimde Veri Madenciliği Uygulamaları*. Uluslararası Eğitim Bilimleri Dergisi, 5, 262-272.

Özekes, S, (2003). *Veri Madenciliği Modelleri ve Uygulama Alanları*, İstanbul Ticaret Üniversitesi Dergisi, Cilt:2, Sayı:3, s:66, 2003

Özkan Y, (2013*). Veri Madenciliği Yöntemleri (2. Basım).* İstanbul: Papatya Yayıncılık.

Özkan Y, (2008). *Veri Madenciliği Yöntemleri, 1. Baskı*, Papatya Yayıncılık Eğitim, İstanbul.

Özmen Ş, (2001). *İş Hayatı Veri Madenciliği İle İstatistik Uygulamalarını Yeniden Keşfediyor*. Marmara Üniversitesi İİBF İngilizce İşletme Bölümü.

Pala T, (2013). *Tıbbi Karar Destek Sisteminin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi.* Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Paolo G, (2003).*Applied Data Mining Statistical Methods for Business and Industry*, West Sussex, Willey 2003, s. 2

Rao, B. Jogeswara, Prof. M. Surendra Prasad Babu and Dr. S. Hanumanth Sastry. (2019). *A New Methodology to Perform Big Data Analytics on Business Warehouse Data*

Sang P., Selwyn P., Michael S, (2001) *Dynamic Rule Refinement in Knowledgebased Data Mining Systems*, Decision Support Systems, 31, 205

Savaş S, Topaloğlu N, Yılmaz M, (2012). *Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri.* İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21, 1-23.

Sebik, N.B.; Bülbül, H.İ. (2018). *Investigation of Seccess on Lung Cancer Data Set of Data Mining Models*, TÜBAV Bilim, 11 (3) ,1-7, (2018).

Silahtaroğlu G, (2013). *Veri Madenciliği Kavram ve Algoritmaları (2. Basım).* İstanbul: Papatya Yayıncılık.

Soylular B, (2006). *Hastanelerde Biyomedikal Klinik Mühendislik Hizmetlerinin Tıbbi Cihaz Kullanıcıları Ve Yöneticiler Bazında Değerlendirilmesi Ve DEÜ Hastanesi Uygulaması*. Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal İlimler Enstitüsü, İzmir.

Şentürk, A, (2006). *Veri Madenciliği: Kavram ve Teknikler*, 1. Basım, Ekin Kitabevi, Bursa.

Şentürk ZK, (2011). *Veri Madenciliği İle Kanser Tanısı*. Yüksek Lisans Tezi, Düzce Üniversitesi Fen Bilimleri Enstitüsü, Düzce.

Şık MŞ, (2014). *Veri Madenciliği Ve Kanserin Erken Teşhisinde Kullanımı*. Yüksel Lisans Tezi, İnönü Üniversitesi Sosyal Bilimler Enstitüsü, Malatya.

Talan Mİ, (2016*). Veri Madenciliği İle Karpal Tünel Sendromuna Yönelik Ön Tanı Destek Ve Hasta Takip Sisteminin Geliştirilmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara.

Tuğ E, (2005). *Genetik Algoritmalar ile Tıbbi Veri Madenciliği*, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Konya.

Turgut H, (2012). *Veri Madenciliği Süreci Kullanılarak Alzheimer Hastalığı Teşhisine Yönelik Bir Uygulama*. Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Isparta.

Tüzüntürk S, (2010). *Veri Madenciliği ve İstatistik*. Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 29, 65-90.

Usgurlu, Ü.B. (2007).Veritabanı, *Veri Madenciliği, Veri Ambarı, Veri Pazarı*, Başkent Üniversitesi Bilgisayar Mühendisliği Yönetim Bilişim Sistemleri Dönem Projesi

Verma, A.K., Pal, S. and Kumar, S. (2019). *Classification of Skin Disease Using Ensemble Data Mining Techniques*. Asian Pac J Cancer Prev.

Yaralıoğlu K, (2004). *Uygulamada Karar Destek Yöntemleri,* 1. Baskı, İlkem Ofset, İzmir.

Quinlan JR, (1986). *Induction of decision trees, Machine Learning*, 1: 81-106.

Wencheng, Sun, Cai, Zhiping, Li, Yangyang, Liu, Fang and Fang, Shenggun; (2018) Wang, Guoyan, *Data Processing and Text Mining Technologies on Electronic Medical Records: A Review*, College of Computer, National University of Defense Technology, Changsha 410073, China, (2018)