ÇUKUROVA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Denetimsiz Makine Öğrenmesi ile Hasta Bekleme Sürelerinin Tahmini

Serhat DOĞAN

Endüstri Mühendisliği Anabilim Dalı

Ağustos 2025

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ONAYI

Denetimsiz Makine Öğrenmesi ile Hasta Bekleme Sürelerinin Tahmini

---

Serhat DOĞAN

Endüstri Mühendisliği Anabilim Dalı

Bu Yüksek Lisans Tezi 12/08/2025 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Değerlendirilmiş ve Oy Birliği / ~~Oy Çokluğu~~ ile Kabul Edilmiştir.

Jüri : Doç. Dr. Melik KOYUNCU (Danışman) ……

: Doç. Dr. Z. Figen ANTMEN ……

: Dr. Öğr. Üyesi Selin SARAÇ GÜLERYÜZ ……

Bu Tez Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Anabilim Dalında Hazırlanmıştır.
Tez No:

Prof. Dr. Sadık DİNÇER
Enstitü Müdürü

İÇİNDEKİLER

ÖZ I
ABSTRACT II
TEŞEKKÜR III
ÇİZELGELER DİZİNİ IV
ŞEKİLLER DİZİNİ V
SİMGELER VE KISALTMALAR VII
1. GİRİŞ 1
1.1. Sağlık, Sağlık Hizmetleri, Sağlık Hizmetlerinde Karar Verme 3
1.1.1. Sağlık 3
1.1.2. Hastalık 4
1.1.3. Hastane 4
1.1.4. Hizmet 6
1.1.5. Sağlık Hizmetleri 7
1.1.6. Sağlık Hizmetlerinde Karar Verme 12
1.2. Veri Madenciliği, Büyük Veri, Makine Öğrenmesi 14
1.2.1. Veri 14
1.2.2. Veritabanı 19
1.2.3. Veri Ambarları ve OLAP 19
1.2.4. Veri Madenciliği 20
1.2.5. Büyük Veri 24
1.2.6. Sağlık Sistemlerinde Büyük Veri 31
1.2.7. Büyük Veri Analizi Araçları 37
1.2.8. Makine Öğrenmesi 38
2. ÖNCEKİ ÇALIŞMALAR 41
3. MATERYAL VE METOT 43
3.1. Makine Öğrenmesi (Machine Learning-ML) 43
3.1.1. Denetimli (Gözetimli) Öğrenme 45
3.1.2. Denetimsiz (Gözetimsiz) Öğrenme 47
3.1.3. Pekiştirmeli (Takviyeli) Öğrenme 49
3.2. Veri Madenciliği ve Makine Öğrenmesi için Veri Ön İşleme 50
3.2.1. Veri Kalite Kriterleri 51
3.2.2. Verinin Temizlenmesi 53
3.2.3. Veri Dönüştürme 54
3.2.4. Verilerde Boyut İndirgeme İşlemleri 55
3.3. Makine Öğrenmesinde Algoritmalar 56

3.3.1. K-En Yakın Komşu Algoritması 57
3.3.2. Karar Ağaçları 58
3.3.3. Rastgele Orman 60
3.3.4. Lojistik Regresyon 60
3.3.5. Destek Vektör Makinesi 61
3.3.6. Naive Bayes 62
3.3.7. Lineer Regresyon 63
3.3.8. Çoklu Lineer Regresyon 63
3.3.9. Polinom Regresyon 64
3.3.10. K-Ortalamalı Kümeleme 65
3.3.11. Hiyerarşik Kümeleme 65
3.4. Karma Dağılımlar 66
3.4.1. Poisson Karma Dağılımı 68
3.4.2. Gamma Karma dağılımı 75
4. BULGULAR VE TARTIŞMA 83
5. SONUÇLAR VE ÖNERİLER 97
KAYNAKLAR 99
ÖZGEÇMİŞ 105
EKLER 107
EK-1: Acil Servis Veri Seti (bir kesit) 109

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Denetimsiz Makine Öğrenmesi ile Hasta Bekleme Sürelerinin Tahmini

Serhat DOĞAN

Danışman: Doç. Dr. Melik KOYUNCU

Endüstri Mühendisliği Anabilim Dalı

ÖZ

Veri, günümüz dünyasında siyasi, ekonomik, sosyal, bilimsel ve teknolojik ilerlemenin temel unsuru haline gelmiştir. Dijitalleşme sürecinin hızlanmasıyla birlikte, sağlık, eğitim, mühendislik, finans, kamu yönetimi gibi hemen her alanda büyük miktarda veri üretilmektedir. Üretilen bu veriler analiz edilip, anlamlı içgörüler çıkarıldığında, karar verme süreçlerinin vazgeçilmez bir bileşeni olarak kullanılmaktadır. Ham verinin anlamlı bilgiye dönüşebilmesi için istatistiksel yöntemler ve makine öğrenmesi algoritmaları ile işlenmesi gerekmektedir. Bu tez çalışmasında, veri, büyük veri, veri madenciliği, makine öğrenmesi, makine öğrenmesi türleri gibi kavramlar ve metodojiler, sağlık, sağlık hizmetleri, sağlık hizmetlerinde karar verme gibi kavramlar incelenmiştir. Sağlık sistemleri araştırmanın çerçevesi olarak belirlenmiş ve büyük veri ile makine öğrenmesi tekniklerinin sağlık hizmetlerinde kullanım alanları incelenmiştir. Sağlık hizmetlerinde makine öğrenmesi yöntemlerinin kullanımına yönelik olarak, bir hastanenin acil servisine başvuran hastaların bekleme süreleri makine öğrenmesinin denetimsiz öğrenme yöntemlerinden karma dağılım metodu ile analiz edilmiştir. Çalşımada Gamma karma dağılımı ve Poisson karma dağılımı modelleri kullanılarak hastaların farklı branşlarda bekleme sürelerine ilişkin kümeleme analizi yapılmıştır. Hasta bekleme sürelerinin istatistiksel dağılımlarının branş bazında farklılık gösterdiği ve bu verilerin tekil bir istatistiksel dağılım yerine karma istatistiksel dağılımlar ile daha uygun biçimde temsil edildiği gözlemlenmiştir.

Anahtar Kelimeler: Veri, Büyük Veri, Karma Dağılım, Kümeleme, Makine Öğrenmesi

CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Predicting Patient Waiting Times with Unsupervised Machine Learning

---

Serhat DOĞAN

Advisor: Assoc. Prof. Dr. Melik KOYUNCU

Department of Industrial Engineering

ABSTRACT

Data has become a fundamental element of political, economic, social, scientific, and technological progress in today's world. With the acceleration of digitalization, vast amounts of data are being generated in almost every field, including healthcare, education, engineering, finance, and public administration. Once this data is analyzed and meaningful insights are extracted, it becomes an indispensable component of decision-making processes. To transform raw data into meaningful information, it must be processed using statistical methods and machine learning algorithms. This thesis examines concepts and methodologies such as data, big data, data mining, machine learning, types of machine learning, and concepts such as health, healthcare services, and decision-making in healthcare. Healthcare systems were chosen as the research framework, and the application of big data and machine learning techniques in healthcare was examined. To address the use of machine learning methods in healthcare, the waiting times of patients presenting to a hospital emergency department were analyzed using the mixture distribution method, an unsupervised learning method in machine learning. Cluster analysis was conducted on patient waiting times across different branches using the Gamma mixture distribution and Poisson mixture distribution models. It has been observed that the statistical distributions of patient waiting times vary by branch and that these data are more appropriately represented by mixture statistical distributions rather than a single statistical distribution.

Keywords: Data, Big Data, Mixture Distribution, Clustering, Machine Learning

TEŞEKKÜR

Yüksek Lisans eğitimim boyunca ve bu tezin oluşturulması sürecinde bilgi ve deneyimiyle bana rehberlik eden, yol gösteren değerli danışman hocam Sayın Doç. Dr. Melik KOYUNCU' ya kişisel gelişimime ve yazmış bulunduğum teze olan tüm katkı ve yardımları için çok teşekkür ederim.

Tez savunmamda değerlendirme jüri üyeleri olarak bulunan, yaptıkları geribildirimler ile tezin iyileştirilmesinde önemli katkıları olan sayın hocalarım Doç. Dr. Z. Figen ANTMEN ve Dr. Öğr. Üyesi Selin SARAÇ GÜLERYÜZ' e ayrıca teşekkür ederim.

Tez hazırlama süreci boyunca verdiği pratik bilgilerle ufkumu genişleten ve sürecin iyileşmesine katkı sağlayan sayın hocam Arş. Gör. Dr. Olcay KALAN' a teşekkür ederim.

Araştırma Görevlisi olarak çalışmaya başladığım günden itibaren her zaman bana destek olan, mesleğe adaptasyon sürecimi kolaylaştıran sayın hocalarım Arş. Gör. Dr. Melek IŞIK, Arş. Gör. Dr. Fatma Şeyma YÜKSEL, Arş. Gör. Dr. Selin ÇABUK, Arş. Gör. Enver Can DORAN'a teşekkür ederim.

Tez sürecinin heyecanını benimle birlikte yaşayan, süreci takip eden arkadaşlarıma da teşekkür ederim.

Manevi olarak verdiği destekle her zaman moral ve motivasyon kaynağım olan halam Öznur KESGİN' e teşekkür ederim.

Bugünlere gelmemde beni yetiştiren, her zaman yanımda bulunan ve en büyük destekçilerim olan annem Selma DOĞAN' a, babam Ahmet DOĞAN' a, abilerim Ufuk DOĞAN ve Selçuk DOĞAN' a, ablam Hatice DOĞAN' a şükranlarımı sunarım.

Biricik yeğenim Ahmet Çınar DOĞAN' a benim neşe ve enerji kaynağım olduğu için teşekkür ederim.

ÇİZELGELER DİZİNİ

Çizelge 2.1. Sorunun Tanımlanmasında Dikkate Alınması Gereken Bazı İpuçları 12
Çizelge 2.2. Sorun Çözme Sürecinde Kullanılabilecek Bazı Sorular 13
Çizelge 3.1. Naive Bayes Algoritmasının Zayıf ve Güçlü Yönleri..... 62
Çizelge 3.2. Lineer Regresyon Algoritmasının Zayıf ve Güçlü Yönleri..... 63
Çizelge 3.3. Polinom Regresyon Güçlü-Zayıf Yönler 64
Çizelge 4.1. Ki-kare ($\chi 2$) Uyum İyiliği Testi p-değerleri..... 84
Çizelge 4.2. Hastalık Türlerine Göre Gamma Karma Modeli Sonuçları 95
Çizelge 4.3. Hastalık Dışı Kişilerin Bekleme Süreleri..... 95
Çizelge 4.4. Gamma Karma Dağılım ile Gözlemlenen Verinin Karşılaştırılması 96
Çizelge 4.5. Poisson Karma Dağılım ile Gözlemlenen Verinin Karşılaştırılması 96

ŞEKİLLER DİZİNİ

Şekil 1.1. Veri-Bilgi-İlim-İrfan İlişkisi 15
Şekil 1.2. Mantıksal Özelliklerine Göre Verinin Sınıflandırılması 16
Şekil 1.3. Fiziksel Düzenine Göre Verinin Sınıflandırılması 18
Şekil 1.4. Yıllara Göre Büyük Veri Konulu Bildiriler (The evolution of big data as a research and scientific topic: Overview of the literature, 2012) 25
Şekil 1.5. Yıllara Göre Büyük Veri Konulu Konferans Bildirileri ve Makaleler (The evolution of big data as a research and scientific topic: Overview of the literature, 2012) 25
Şekil 1.6. Coğrafyaya Göre Büyük Veri Konulu Bildiriler (2012'ye kadar) (The evolution of big data as a research and scientific topic: Overview of the literature, 2012) 26
Şekil 1.7. Veri analizi için genel teknoloji adımları 37
Şekil 3.1. Makine Öğrenmesi Süreci 43
Şekil 3.2. Makine Öğrenmesi Tipleri 45
Şekil 3.3. Denetimli Öğreneme 46
Şekil 3.4. Denetimsiz Öğrenme 48
Şekil 3.5. Pekiştirmeli Öğrenme 50
Şekil 3.6. Girdi-Çıktı Kalitesi İlişkisi 51
Şekil 3.7. Makine Öğrenmesi Algoritmaları 56
Şekil 3.8. K-En Yakın Komşu Algoritması 57
Şekil 3.9. Dört boyutlu özellik uzayına sahip üç sınıftan oluşan bir karar ağacı yapısı 59
Şekil 3.10. Lojistik Regresyon 61
Şekil 3.11. Destek Vektör Makinesi 62
Şekil 3.12. Lineer Regresyon 63
Şekil 3.13. Polinom Regresyon 64
Şekil 3.14. K-Ortalamalı Kümeleme 65
Şekil 4.1. Makine Öğrenmesi Süreci 43
Şekil 4.2. Makine Öğrenmesi Tipleri 45
Şekil 4.3. Gözetimli/ Denetimli Öğreneme (Dutt S., Chandramouli S., Das K. A., 2018) 46
Şekil 4.4. Gözetimsiz/Denetimsiz Öğrenme (Dutt S., Chandramouli S., Das K. A., 2018) 48
Şekil 4.5. Mesafe Tabanlı Kümeleme (Dutt S., Chandramouli S., Das K. A., 2018)**Hata! Yer işareti tanımlanmamış.**
Şekil 4.6. Pekiştirmeli/Takviyeli Öğrenme (Dutt S., Chandramouli S., Das K. A., 2018) 50
Şekil 4.7. Garbage-In-Garbage-Out 51
Şekil 4.8. Makine Öğrenmesi Algoritmaları 56
Şekil 4.9. K-En Yakın Komşu Algoritması 57
Şekil 4.10. Karar Ağacı Türleri 59

Şekil 4.11. Logistik Regresyon 61
Şekil 4.12. Naive Bayes Algoritmasının Zayıf ve Güçlü Yönleri 62
Şekil 4.13. Lineer Regresyon 63
Şekil 4.14. Lineer Regresyon vs. Polinom Regresyon 64
Şekil 4.15. K-Ortalamalı Küme 65
Şekil 5.1. Kardiyoloji Hastaları Bekleme Süreleri 85
Şekil 5.2. Kardiyoloji Hastaları Bekleme Süreleri-Gamma Karma dağılımı 85
Şekil 5.3. Gastroenteroloji Hastaları Bekleme Süreleri 86
Şekil 5.4. Gastroenteroloji Hastaları Bekleme Süreleri- Gamma Karma dağılımı 87
Şekil 5.5. Genel Pratisyenlik Hastaları Bekleme Süreleri 88
Şekil 5.6. Genel Pratisyenlik Hastaları Bekleme Süreleri-Gamma Karma dağılımı 88
Şekil 5.7. Nöroloji Hastaları Bekleme Süreleri 89
Şekil 5.8. Nöroloji Hastaları Bekleme Süreleri-Gamma Karma dağılım 90
Şekil 5.9. Ortopedi Hastaları Bekleme Süreleri 91
Şekil 5.10. Ortopedi Hastaları Bekleme Süreleri-Gama Karma dağılımı 91
Şekil 5.11. Fizyoterapi Hastaları Bekleme Süreleri 92
Şekil 5.12. Fizyoterapi Hastaları Bekleme Süreleri- Gamma Karma dağılımı 93
Şekil 5.13. Böbrek Hastaları Bekleme Süreleri 94
Şekil 5.14. Böbrek Hastaları Bekleme Süreleri- Gamma Karma dağılımı**Hata! Yer işareti tanımlanmamış.**

SİMGELER VE KISALTMALAR

$\hat{\sigma}$: Popülasyon standart sapması tahmini

AIC : Akaike Bilgi Kriteri (Akaike Information Criterion)

BIC : Bayesci Bilgi Kriteri (Bayesian Information Criterion)

BT : Bilgisayarlı Tomografi

BV : Büyük Veri

BVA : Büyük Veri Analizi

DSÖ : Dünya Sağlık Örgütü

EHR : Elektronik Sağlık Kayıtları (Electronic Health Records)

EM : Beklenti Maksimizasyonu Yöntemi (Expectation Maximization)

ICL : Entegre Tamamlanmış Olabilirlik (Integrated Completed Likelihood)

KA : Karar Ağaçları

KNN : K En Yakın Komşu (K Nearest Neighbour)

LL : Logaritmik Olabilirlik (Log-Likelihood)

ML : Makine Öğrenmesi (Machine Learning)

OLAP : Çevrimiçi Analitik İşleme (On- Line Analytical Processing)

RO : Rastgele Orman

STK : Sivil Toplum Kuruluşları

SVM : Destek Vektör Makinesi (Support Vector Machine)

Vb : ve benzerleri

Yy : yüzyıl

$\hat{\mu}$: Popülasyon ortalaması tahmini

YZ : Yapay Zeka

1. GİRİŞ

Günümüzde veri, dünyanın yeni enerji kaynağı olarak kabul edilmektedir. Dijitalleşmenin etkisiyle, yaşamın her alanı büyük ölçekli veri yığınlarıyla çevrelenmiştir. Bu veri yığınları uygun bir biçimde düzenlenip, analiz edilebildiği takdirde, görülemeyen problemleri, ilişkileri, trendleri, problemlerin nedenlerini ortaya çıkarabilir ve uygun metotlar yardımıyla bu problemlere çözümler üretebilir.

Sağlık verisi milyonlarca insandan toplanmaktadır. Bu veriler sağlık sistemlerini geliştirebilmek için güçlü bir kaynak haline gelmiştir. Bu devasa veri kümesinden anlamlı ve faydalı bilgiler üretmek sağlık sistemlerinin gelişmesine olanak sağlar. Aksi halde veriler anlamlı bilgi üretmeye katkı sağlamadan sistemlerde gereksiz bir yoğunluğa yol açabilir.

Sağlık sistemlerinde anlamlandırılmış veriler sayesinde hekimler, araştırmacılar, ilaç üreticileri, mühendisler, kamu sağlığı kuruluşları, toplumlar, bireyler; her seviyede sağlık konuları hakkında daha iyi kararlar alabilirler.

Sağlık hizmetlerinde yapay zekanın potansiyeline yönelik önemli beklentiler bulunmaktadır. Hastalık gözetiminin daha iyi sağlanması, hastalıklarda erken teşhis sürecinin hızlanması, hastalık tanısı koymanın doğruluğunu yükseltme, yeni tedavi süreçlerinin ortaya çıkarılması, kişiselleştirilmiş tıp uygulamalarının yaygınlaşmasına katkı sunabilme gibi potansiyel barındırmaktadır (Fogel ve Kvedar, 2018).

Son yıllarda dünya genelinde artan yaşlı nüfus oranı ve bu nüfusun büyük çoğunluğunun en az iki kronik rahatsızlığı (çoklu hastalık) olması sağlık ve sosyal hizmet sunumunu daha da önemli hale getirmiştir. Çoklu hastalık; hastaneye yatış olasılığını, hastanede kalış süresini, sağlık hizmeti maliyetlerini artırmaktadır (Kingston ve ark., 2018). Tedavi, muayene ve cerrahi operasyonlara talep artmaktadır. Sağlık personelleri aynı zamanda gelişen iş akışlarını, protokolleri de olağan süreçlere uygulamaya çalışmaktadır. Sağlık hizmetlerinde yaşanan problemlere yapay zeka (YZ) uygulamaları çözüm olarak sunulmaktadır. Mevcut sağlık sistemlerindeki idari işlemlerin, prosedürlerin hızlandırılmasını, karar destek sistemleri aracılığıyla daha yüksek doğruluk oranında hastalık tespitinin yapılmasını, hasta bakımlarının iyileştirilmesini, risk ve belirsizlik altında karar almaya çalışan sağlık personellerinin karar alma süreçlerine destek sunmasını sağlamaktadır. Bu nedenle YZ uygulamaları giderek daha yaygın kullanılmaya çalışılmaktadır. Sağlık sistemleri, bilgisayar teknolojilerinin ve YZ metotlarının gelişmesiyle birlikte büyük veriye erişim ve kullanım fırsatlarını da sunmaktadır (Yanar, 2024).

Sağlık verilerinin artan erişilebilirliği ve büyük veri analitiği yöntemlerinin hızla gelişmesi, YZ'nin sağlık hizmetlerinde son zamanlarda başarılı bir şekilde uygulanmasını sağlamıştır. YZ karmaşık algoritmalar kullanarak büyük hacimli sağlık verilerinden öğrenme işlemini gerçekleştirmekte, içgörüler elde ederek, büyük miktarda veride gizli olan klinik bilgileri açığa çıkarmaktadır. Böylece klinik kararların alınabilmesine yardımcı olmaktadır (Jiang ve ark., 2017).

Sağlık kurumları rekabet edebilmek, karar alma süreçlerini optimize etmek, iş süreçlerini iyileştirmek için büyük veriden (BV) yararlanmak ve büyük veri analizi (BVA) yapabilecek teknolojilere yatırım yapmak zorundadırlar.

Son yıllarda BVA metotları geliştikçe, dünya sağlık yönetimi hastalık merkezli bir modelden, kişiselleştirilmiş tıbba doğru bir modele evrilmiştir. Kişiselleştirilmiş tıp modelini uygulayabilmek için de BV'nin yönetilebilmesi ve analiz edilebilmesi gerekmektedir (Yanar, 2024).

Sağlık verileri hastaneler, diğer sağlık kurumları, sigorta şirketleri ve ilgili kamu kurumları başlıcaları olmak üzere birçok kuruluş tarafından toplanmaktadır. Sağlık bilgi sistemlerinin misyonu büyük miktardaki verilerden faydalı bilgiler üretmektir (Koyuncugil ve Özgülbaş, 2009).

Sağlık hizmetleri yönetiminde, BV ve Makine Öğrenmesi (Machine Learning, ML) metotları özellikle sağlık kurumundaki kaynakların etkin kullanımı (personel, tıbbi ekipman,vb), hastane yataklarının optimum doluluk oranı ile kullanımı, hastalıkların teşhis edilip hasta bakım sürelerinin iyileştirilmesi, sağlık politikalarının oluşturulması gibi amaçlar için kullanılmaktadır (Koyuncugil ve Özgülbaş, 2009).

ML algoritmaları, hastaların tıbbi geçmişlerini inceleyerek yeniden hastaneye geliş durumunu ya da yeniden aynı hastalığa yakalanma olasılığını tahmin ederek klinik ve yönetimsel süreçler için karar destek sistemleri sunabilmektedir (Yanar, 2024).

Literatürde sağlık alanı da dahil olmak üzere farklı alanlarda makine öğrenmesi yöntemlerinin kullanıldığı birçok araştırma yer almaktadır. Bu yöntemlerden biri karma dağılımlar ile kümeleme yapılmasıdır. Anatomi, biyoinformatik, hücre biyolojisi, kronik hastalıklar, genetik, geriatri, bulaşıcı hastalıklar, görüntü işleme, ortopedi, farmakoloji, beslenme, jinekoloji, psikiyatri gibi alanlarda karma dağılımlar kullanılmıştır (Lee ve ark., 2001).

Sağlık kurumlarını ziyaret eden hastaların hastaneye gelişler arası süreleri, hastanede kalış süreleri rastgele gerçekleşen stokastik olaylardır ve bu veriler istatistiksel olarak çarpık bir yayılım göstermektedir. Bu nedenle bazı durumlarda saf bir istatistiksel dağılıma uymayan veriler oluşmaktadır. Hastaların durumunun önemine göre sağlık kurumlarında kalma süreleri uzun veya kısa olabilmektedir. Hastanenin operasyonel performansının kalitesini artırmak ve personel, ekipman, zaman gibi kısıtlamaları yönetebilmek için hastaların sistemde kalma sürelerinin doğru bir şekilde belirlenmesi ve tahmin edilmesi gerekmektedir. Çarpık dağılımlar istatistiksel analizlerde sıkça karşılaşılan bir durumdur. Bu veriler normal dağılıma uymadığı ve simetrik bir yapıya sahip olmadığı için, modellenmesi daha zordur. Bu noktada karma dağılımların önemi ortaya çıkmaktadır. Karma dağılımlar istatistiksel analizlerin doğruluğunu artırmak için sıklıkla tercih edilmektedir. Bu yöntemle hastaların sistemdeki durumları analiz edilebilir ve klinik karar verme aşamasında katkıda bulunulabilir (Güleryüz, 2023).

Bu çalışmada makine öğrenmesinde denetimsiz öğrenme yöntemleri arasında yer alan karma dağılımlar kullanılarak bir sağlık kurumunun acil servisini ziyaret eden hastaların bekleme sürelerine göre kümeleme analizi yapılması amaçlanmaktadır. Acil servislerde hasta yoğunluğu, sağlık personeli sayısı, operasyonel süreçler, ekipman kısıtları gibi etkenler nedeniyle hastaların bekleme süreleri farklılık göstermektedir. Bu farklılıklar çoğu zaman istatistiksel olarak çarpık dağılım özellikleri gösterdiğinden, tekil istatistiksel dağılımlarla temsil edilmesi zorlaşmaktadır. Karma dağılımlar, farklı alt gruplardan gelen verilerin birleşiminden oluşan heterojen yapıları modellemede güçlü bir araç olduğundan, acil servisteki hasta bekleme sürelerinin modellenmesinde uygun bir yöntem olarak değerlendirilmektedir.

Çalışmada; "Giriş" başlığında sağlık ile ilgili terimler, BV, veri madenciliği, ML kavramları, sağlık sistemlerinde BV kullanımıma değinilmiştir. Önceki çalışmalar kısmında literatür örneklerine yer verilmiş, materyal ve metot kısmında makine öğrenmesi yöntemlerinden ve karma dağılımlardan bahsedilmiş; bulgular ve tartışma kısmında uygulamada elde edilen sonuçlar açıklanmış; son olarak da sonuç ve öneriler kısmında daha sonraki araştırmacılar için önerilerden bahsedilmiştir.

1.1. Sağlık, Sağlık Hizmetleri, Sağlık Hizmetlerinde Karar Verme

Bu bölümde sağlık ile ilişkili kavramlar incelenmiştir.

1.1.1. Sağlık

Dünya Sağlık Örgütü (DSÖ)'a göre; sağlık, sadece hastalık ve sakatlığın yokluğu değil, tam bir fiziksel, zihinsel ve sosyal refah durumu olarak tanımlanır. (https://www.who.int/about/governance/constitution). Bu tanımda sosyal refah ve hastalık yokluğuna vurgu yapılmaktadır. Sosyal refahın sağlık tanımına dahil edilmesi konusunda herkes hemfikir değildir. Sosyal refahın, sağlık dışı yönleri, ekonomik ve politik koşullar vb. gibi yönleri de vardır.

Sağlık, bir bireyin işlev düzeyidir. Burada optimum işlev, toplumun fiziksel ve zihinsel refah standartlarıyla kıyaslanarak değerlendirilir. Bir bireyin yaşamında ne kadar iyi işlev gördüğü, fiziksel, zihinsel ve sosyal sağlık alanlarında algılanan refahı olarak tanımlanabilir. İşlevsellik bir kişinin önceden tanımlanmış bazı aktiviteleri gerçekleştirme becerisini ifade ederken, refah ise bireyin öznel duygularını ifade eder (Karimi ve Brazier, 2016).

Bireylerin içinde yaşadıkları kültür, standartlar, hedefler, beklentiler, endişeler gibi faktörlerle ilişkili olarak, yaşamdaki konumlarına ilişkin algıları farklılık göstermektedir. Bu sebeple sağlık kavramının tanımı çeşitli biçimlerde yapılabilmektedir.

Sağlıklı bir birey olmak; bireyin kendisini ruhsal açıdan mutlu hissetmesi, fiziksel açıdan gereksinimlerini diğer bireylere ihtiyacı olmadan karşılayabilme durumudur. Sağlıksız bir birey

olmak; bireyin ruhsal açıdan yetersiz ve mutsuz hissetmesi, fiziksel açıdan sakatlık hali gibi durumun içinde olmasıdır (Yanar, 2024).

Sağlık, bireylerin öncelikleri ile zaman faktörüne bağlı olarak da değişmekte, gelişmekte ve kapsamı derinleşmekte olan çok boyutlu dinamik bir kavramdır.

Davranış bilimcilere göre sağlık; bireylerin sosyal alanlarında diğer bireyler ile uyum içerisinde olması becerisi olarak tanımlanmıştır. Biyoloji açısından sağlık; bireylerin yaşamlarını sürdürebilmeleri için tüm bedensel işlevlerinin ve fonksiyonlarının uyum içerisinde çalışması olarak tanımlanmıştır. Felsefi açıdan sağlık; bireylerin kendi emel ve arzularını gerçekleştirebilmeleri için gerekli olan yaşama sahip olmaları olarak tanımlanmıştır (Yanar, 2024).

1.1.2. Hastalık

Hastalık; çağdan çağa ve toplumdan topluma değişkenlik gösteren dinamik bir kavramdır. Bir bireye hangi koşullarda ve ne zaman hastalık teşhisi konulabileceği net bir ifade değildir. Toplumsal açıdan hastalık; sağlık ve normallik standartlarındaki sapma durumudur. Fiziksel ve psikoloji normallerinin işleyişini bozan durumlardır (Yorulmaz ve Erdem, 2021).

Toplumların en ilkel halden, bilgi toplumuna dönüşüne kadar geçirdikleri değişim içerisinde sağlık ve hastalık olgusunda, hasta ve hekim arasındaki iletişimde değişiklikler meydana gelmiştir (Yorulmaz ve Erdem, 2021).

Türk Dil Kurumu'na göre hastalık; organizmada birtakım değişikliklerin meydana gelmesi ile sağlığın bozulması durumu, rahatsızlıktır (Türk Dil Kurumu, 2024).

Hastalık kavramı tıbbi olduğu kadar aynı zamanda sosyal ve kültürel alandır. Tıbbi açıdan hastalık; kişinin bedeninde yaşam fonksiyonlarını aksatacak semptomlar belirmesidir. Sosyal ve kültürel açıdan hastalık; bireyin içerisinde bulunduğu sosyo-ekonomik durumun niteliği , sosyo-kültürel çevre ile uyumu vb. niteliklerdir (Aytaç ve Kurtdaş, 2015; Yanar,2024).

1.1.3. Hastane

Hastaneler yalnızca bakımın yapıldığı yer değil, aynı zamanda bilginin de üretildiği mekanlardır. Hastane (hospital) kelimesi misafirperverlik (hospitality) kelimesinden türetilmiştir. Aynı zamanda barınak (spital), otel (hotel), darülaceze (hospice) ile de ilişkilendirilir (Turner, 2006).

Dünyada yaşanan değişimler ve gelişmeler sonucunda hastane olgusundaki istek ve beklentiler de değişmektedir. Dolayısıyla hastane kavramının tanımlanması da dönemsel farklılıklar göstermektedir.

Hastane genel olarak hastalıkların tedavi edilmesi amacı ile içinde hekim, hemşire, hasta bakıcı gibi sağlık personelleri barındıran, hastalıklara tanı konulan ve tedavi süreçlerinin uygulanabilmesi için gerekli bina ve binanın içinde yeterli sağlık ekipmanlarının bulunduğu kurumlardır. Aynı zamanda hastalıklar ile mücadele edilebilmesi için araştırmaların yapıldığı,

sağlık personellerinin eğitim alarak mesleki gelişimlerinin tamamlanmasını sağlayan kurumlardır (Yanar, 2024).

Sağlık hizmetleri ya da tıbbi uygulamalar insanlık tarihi kadar eskiye dayanmaktadır ve insan yaşamı ile paralel olarak gelişme göstermiştir. Eski uygarlıklarda hastaneler ibadethane olarak açılmış mekanlardır. Hastalıklardan kurtulmak için dini bilgiler, sihir ve büyü kullanılmıştır. Dolayısıyla ilk hastaneler dinsel kurumlar olarak değerlendirilebilir (Aslan ve Erdem, 2017).

Eski uygarlıklarda hekimlik; eğitimi alınan bir meslek grubu olmaktan ziyade, tıbbi bilgilerin babadan-oğula geçtiği ya da tıbba ilgi duyan araştırmacıların bitkilerle tedavi denemeleri yaparak icra ettiği bir alan olmuştur. Eski uygarlıklarda tıp eğitimi almak isteyenler hekimlerin evlerinde kalarak eğitim görmüş ve hocaların bilgi ve deneyimlerinden faydalanmışlardır. Zaman içerisinde hastalıkların yaygınlaşıp, hasta sayısının da artması ile tıp eğitiminin ve hasta bakımının yapılabilmesi için hastaların bir arada tutulması gerektiği yapıların inşa edilmesi zaruri bir durum olmuştur. Bu amaçla Antik Çağ'da kurulan ilk hastanelerden Asklepion hem tıp eğitimi vermiş hem de bu kurum hastane olarak hizmet sunmuştur. Hipokrat ve Galen gibi sağlık ilminin önde gelen bilim insanlarının yetiştiği bir tıp merkezi olmuştur (Aslan ve Erdem, 2017).

15. ve 16. yüzyıl (yy) 'da Rönesans döneminde tedavi ve müşhade benimsenmiştir. Bu dönemde farklı üniversitelerde tıp eğitimi verilmeye başlanmış, hastaların tedavi edilmesi için özel klinikler oluşturulmuştur. 17. yy sonlarına dek klinikteki çalışmalar hayvanlar üzerinde deney yapılarak ya da ölüler üzerinde gerçekleştirildiklerinden, hastalıkların tam olarak anlaşılması ve uygun tedavi yöntemlerinin farkına varılması zaman almıştır. 18. yy kliniklerin yenilenmesi ve hastanelerin gerçek anlamda tam olarak kullanıldığı dönem olmuştur. Bu dönemde (kulak-burun-boğaz, kadın hastalıkları, akıl sağlığı, salgınla mücadele vb.) gibi branş hastaneleri de ortaya çıkmıştır. 19. yy'da hekim ve diğer sağlık personelleri sayıca yetersiz kalmıştır. Hem tıp eğitimi hem de tedavi hizmeti verilen hastanelerin zayıflaması nedeniyle özel sağlık okulları oluşturulmuştur. Hekimlik ve sağlık personeli öğrencileri bu okullarda aldıkları eğitimlerinin uygulamasını hastanelerde yapmışlardır. Bu şekilde de kamu-özel sektör iş birliği ile tıp eğitiminin iyileştirilmesi amaçlanmıştır. 21. yy'da sağlık hizmetlerinde yaşanan teknolojik gelişmeler ve eğitimdeki iyileşmeler ile birlikte dünya genelinde kamu ve özel hastanelerin sayısı artış göstermiştir. Sağlık hizmetlerinde devletlere ekonomik yük olmasından dolayı da özel hastanelerin sayısı da son yıllarda artmaktadır (Aslan ve Erdem, 2017).

Sağlık hizmetlerinde yaşanan gelişmeler ile birlikte hastanelerin hizmet kapsamında da değişimler meydana gelmiştir. Hastaneler yalnızca tıbbi hizmet veren kurumlar değil aynı zamanda sosyal hizmet veren kurumlar olma özelliği kazanmışlardır.

Sağlık hizmeti alırken ve sağlık hizmeti aldıktan sonra sosyal hizmete ihtiyaç duyan bireylere tıbbi sosyal hizmet adı altında hizmet verilmektedir. Bu tür hizmetlerin en yoğun olduğu yerler hastanelerdir. Bu hizmetler; yoksul, engelli, şiddet mağduru, istismara uğramış olanlar,

mülteci-sığınmacılar, madde bağımlıları vb. dezavantajlı olan gruplara verilen hizmetleri kapsamaktadır (Keleşlioğlu ve ark., 2022).

Türkiye'de Hastaneler

Türkiye'nin hastane yapılarının altyapısı, temeli, Türkiye topraklarında daha önce hüküm sürmüş uygarlıklara dayanmaktadır. Bu coğrafyada farklı isimlerle (Şifahane, bimaristan, maristan, darüssıhha, darülafiye, me'menülistirahe, darüttıb, darülmerza, şifaiyye, bimarhane, tımarhane, darüşşifalar) anılan hastaneler çeşitli vakıflarca kurulmuştur. Selçuklu döneminden kalma bu yapılar kimsesizler, öğrenciler, yolcular ve tüccarlar gibi muhtaç insanlara sağlık hizmeti vermek amacıyla inşa edilmiştir (Acıduman, 2010).

1800'lü yıllardan sonra Osmanlı Devleti tarafından inşa ettirilen 36 hastanenin büyük bir kısmı yıkılsa da, Cumhuriyet dönemine ulaşan hastaneler de olmuştur. Bu hastaneler farklı padişahlar tarafından farklı amaçlarla (gazilerin tedavisi, salgın hastalıklarla mücadele, ordu ve donanma hastaneleri vb.) inşa edilmiştir (Aslan ve Erdem, 2017).

Türkiye tarihi boyunca hastanelerin büyük bir kısmı devlet tarafından yapılmış olsa da vakıflar tarafından da hastaneler kurularak sağlık hizmeti vermeye başlamıştır. Ayrıca üniversite hastaneleri ve özel hastaneler de son yıllarda sağlık sektöründe önemli pay sahibi olmuştur (Aslan ve Erdem, 2017).

1.1.4. Hizmet

İnsanların birlikte yaşamalarının doğal bir sonucu olarak yaşantımızın her aşamasında farklı biçimlerde hizmet kavramı ile karşılaşılmaktadır (Sayım ve Aydın, 2011).

Hizmet kelimesi; bireyler ve toplumlar için bir şeyler yapan, bir şeyler üretmeyen endüstriyel sektörü ifade etmek için kullanılmaktadır. Bu yorum, ekonomistlerin ekonomik faaliyetleri sınıflandırmak istemelerinden dolayı doğmuştur. Birleşik Krallık Ulusal İstatistik Ofisi; hizmet sektörlerini finans, ulaşım, perakende ve kişisel hizmetler olarak sınıflandırır (Johns, 1998).

Hizmet ayrıca toplumun ihtiyaçlarını karşılayan sağlık hizmeti ve kamu hizmeti gibi faaliyetleri ve bu faaliyetleri gerçekleştiren kuruluşları da kapsar. Kamu hizmetleri bürokratik kurallarla gelişmektedir ve endüstriyel hizmet sektöründen oldukça farklıdır (Johns, 1998).

Hizmet, insanların ihtiyaçlarını karşılamak amacıyla, belli bir fiyattan satışa sunulan, elle tutulamayan, koklanamayan, kolayca heba edilebilen, standartlaştırılamayan, sunulduğu kişiye yarar ve doyum hissi oluşturan soyut faaliyetlerin bütünüdür (Sayım ve Aydın, 2011).

Bireylerin ve kurumlarının mevcut gereksinimlerinin belirlenip, bu gereksinimlerin karşılanması amacıyla çözüm sunan kişi ya da kuruma değer fiyatı ödenerek ulaşılan, soyut nitelikte olan, çözüm sunulan kişi ya da kuruma yarar sağlayan faaliyetler de hizmet kapsamına girmektedir (Yanar, 2024)

Hizmetler genellikle " maddi olmayan faaliyetler" olarak tanımlanıp, çıktıları somut bir nesneden ziyade bir faaliyet olsa da bu ayrımın hizmet kavramını yeterince temsil etmediği bilinmektedir. Hizmet çıktılarının önemli bir somut bileşeni olabilir. Örneğin bir restoranın yiyecek, içecek servisi sağlaması, bir perakende faaliyetinde tedarik edilen malların maddi açıdan sınıflandırılması gibi faaliyetler somut bileşenler içermektedir (Johns, 1998).

Hizmetlerin özgün özellikleri şu şekilde özetlenebilir:

- İstisnalar haricinde, ölçülemezler.
- Stoklanamazlar. Sunulduğu anda tüketilir. Tekrarı olabilir fakat aynı hizmet yeniden sunulamaz.
- Tetkik edilemez ya da incelenemezler. Hizmetler gözlemlendikten sonra bazı sonuçlara ulaşılabilir. Hizmetlerle bütünleşen fiziksel şartlar ya da fiziksel ürünler belirli performans ya da çevre standartlarıyla incelenir.
- Hizmet sunulmadan, kalite değerlendirilemez.
- Yaşam süresi yoktur. Sadece hizmetlerin oluşturulması ve sunulması süresi vardır.
- Zaman boyutu vardır. Belli bir saatte başlar ve belli bir saatte biter.
- Nesne değil, performanstır. İnsan davranışı ile yönlendirilir. Bu durum hizmet sunan kişilerin uzman olmasını gerektirir.
- Talebe göre sunulur. İki türlü talepten söz edilebilir. Bunlar; sürekli talep (su, telefon, elektrik hizmetleri vb.) ve programlanmış talep (perakende hizmetler, doktor muayeneleri, bankalar, ulaşım vb.) (Sayım ve Aydın, 2011).

1.1.5. Sağlık Hizmetleri

Dünya Sağlık Örgütü tarafından kişilerde bedenen, ruhen ve sosyal yönden tam olarak iyi olma hali şeklinde ifade edilen sağlık, tüm ülkelerin temel insani bir hak olarak kabul ettiği, korunması ve güvence altına alınması gereken bir ihtiyaç olarak ifade edilebilmektedir (Karaçor, ve Arkan, 2012).

Sağlık hizmetleri; bireylerin ve toplumun sağlığını korumak, hastalıkları önlemek, teşhis etmek, erken tanı girişimlerinde bulunarak ciddi sorunlar teşkil edebilecek rahatsızlıkları önlemek, tedavi etmek, hastayı rehabilite ederek iyileştirmek gibi hizmetlerin bütününü kapsar. Bu hizmetler sağlık sisteminin temel bileşenleri olarak kabul edilir ve bireyin ve toplumun yaşam standartlarının kalitesini yükseltmeyi amaçlar. Yalnızca tedavi edici hizmetlerle sınırlı değildir; aynı zamanda koruyucu ve rehabilite edici özellikleri de kapsar (Yanar,2024)

Sağlık hizmetlerinin temel amaçlarından bir diğeri ise; önlemler alınmasına rağmen, karşılaşılabilecek sağlık sorunlarında, mümkün olan en kısa sürede, en uygun maliyetlerde ve etik ilkeler doğrultusunda bireylerin sağlık ihtiyaçlarını gidermektir (Karaçor ve Arkan, 2012).

Sağlık hizmetlerini genel olarak; kamu, özel sektör, sivil toplum kuruluşları (STK) sunmaktadır.

Sağlık alanında faaliyet gösteren STK'lar kısıtlı hizmet alan topluluklara tıbbi bakım, tedavi, rehabilitasyon ve eğitim hizmeti sunarak toplumun sağlık statüsünün geliştirilmesinde rol oynamaktadır. STK'lar, yerel yönetimler ve sağlık hizmeti sunan kuruluşlar ile iş birliği içerisinde çalışmaktadır. Ayrıca toplum sağlığının iyileştirilmesine yönelik politika ve programların savunulmasında da önemli rol oynamaktadırlar (Gülay ve ark., 2023).

Sağlık hizmeti:

- Tüketicilerin alınan hizmet hakkında tüm bilgilere sahip olmaması
- Hizmeti sunan kişilerin bilgi gücü
- Tüketicilerin akıl düzeyi yetersizliği nedeniyle uygun olmayan davranışlarda bulunması
- Tıp mesleğinin kuralları
- Ürün, kalite ve talepte belirsizlik
- Devlet müdahalesine açık olması
- Sağlık hizmet tüketiminin rastlantısal olması
- İkame edilemez ve ertelenemez olması
- Hizmetin boyutu ve kapsamının, hizmeti sunan hekimler tarafından belirlenmesi
- Pazarın tekelleşmeye müsait olması
- Fiyatlarla gerçek maliyetler arasındaki korelasyonun zayıf olması

gibi özellikleri ile diğer hizmet türlerinden ayrışmaktadır (Karaçor ve Abdullah, 2012).

Sağlık hizmetlerinde, sağlık personelleri için; gerekli olan araç-gereçler, ilaçlar ve diğer hizmetlere kolay bir şekilde erişebilirlik, hasta kişilerin kaliteli ve sürekli hizmet alabilmesi için oldukça önemlidir. Bunlara ilaveten sağlık hizmetlerinde kaynakların planlı ve verimli kullanılması, ekonomi ve süreklilik açısından oldukça değerli bir yaklaşımdır.

Sağlık Hizmetlerinde Sınıflandırma

Sağlık hizmetleri ilk bakışta bireysel hizmet olarak algılanabilmektedir fakat sağlık hizmetlerinin faaliyetleri şu şekilde sınıflandırılmaktadır:

- Koruyucu Sağlık Hizmetleri
- Tedavi Edici Sağlık Hizmetleri
- Rehabilitasyon Hizmetleri

Bu hizmetler sağlık faaliyetlerinin toplumsal hizmetler olduğunu göstermektedir (Tunçsiper ve Bakar, 2023).

Üç kategoride ifade edilen sağlık faaliyetlerinde öncelikli amaç; bireylerin sağlıklarının bozulmasından önce gerekli tedbirlerin alınması ve bireylerin sağlıklarının korunmasıdır. İkincil amaç; bireylerin mevcut hastalıklarına en erken sürede tanı konulmasıdır. Üçüncül amaç; bireylerin fiziksel ve ruhsal rehabilitasyonlarının yapılmasıdır. Öncelikli amaç sağlanabilirse; diğer hizmet aşamalarına olan gereksinimler de azalabilecektir (Yanar, 2024).

a. Koruyucu Sağlık Hizmetleri

Birinci basamakta sunulan sağlık hizmetleridir. Sadece hastalığa değil, aynı zamanda hastalığa da yol açan tüm risklere müdahale etmek ve bunlara engel olabilmeye yönelik plan oluşturma misyonu vardır. Bulaşıcı hastalıklarla mücadele, hijyen kurallarının topluma aktarılması, rehabilitasyon süreçlerinin hasta ve yakınları ile paylaşılması, topluma biyo-psiko-sosyal anlamda sağlık bilgilerinin verilmesi gibi hastalığı tedavi etmeden önce, hastalığı önlemeyi hedefleyen hizmetlerdir (Gençer ve ark., 2021).

Koruyucu sağlık hizmetleri genellikle, kişiye yönelik koruyucu sağlık hizmetleri ve çevreye yönelik koruyucu sağlık hizmetleri olmak üzere iki kategoride değerlendirilmektedir (Yanar, 2024).

a.1. Kişiye Yönelik Koruyucu Sağlık Hizmetleri

Bu hizmetler, doğrudan bireyin sağlıkla ilgili kararlarına ve yaşam biçimine müdahale ederek sağlık bilincini artırmayı hedeflemektedir. Bireyin; fiziksel, ruhsal ve sosyal sağlığını bir bütün olarak inceler. Sağlıklı yaşam alışkanlıklarını teşvik etmek amacıyla sunulan hizmetlerdir. Aşılama, bu hizmet kategorisinin en yaygın ve en etkili uygulamalarından biridir.

Kişiye yönelik koruyucu sağlık hizmetlerinin en önemli parçalarından olan, branşa özel kaliteli eğitim almış hekimler ve hemşireler, bu hizmetlerin uygulanabilmesini sağlamaktadır (Yanar, 2024).

a.2. Çevreye Yönelik Koruyucu Sağlık Hizmetleri

Bu hizmetler, insan sağlığına olumsuz etki edebilecek çevresel faktörlerin kontrol altına alınmasını ve normallikten sapan durumların, iyileştirilmesini hedeflemektedir. Çevresel faktörlerin sebep olduğu hastalıkların önlenmesinde kritik bir rolü vardır. Bu hizmetler, toplum sağlığını korumak ve iyileştirmek için geniş çaplı stratejiler ve politikalar vasıtasıyla uygulanmaktadır.

Halk sağlığının korunması için;

- Temiz suya erişebilmek ve güvenli, kalite standartları yüksek gıda üretimi ve tüketimi halk sağlığının korunmasında hayati bir öneme sahiptir.
- Hava kirliliği, küresel ölçekte en büyük çevre sorunlarından biridir. Sanayi tesislerinden çıkan emisyonlar, hava kirliliğinin başlıca kaynaklarından biridir. Bu nedenle emisyonların azaltılması için sıkı bir denetim ve etkili politikalar oluşturulması gerekmektedir. Araç emisyonları azaltılmalıdır. Alternatif ulaşım yöntemleri desteklenmelidir.
- Yeşil alanlar artırılmalıdır.
- Uygun atık yönetimi, çevre kirliliğini önler ve insan sağlığına zarar verebilecek maddelerin yayılımını engeller. Kimyasal atıklar, radyoaktif maddeler, biyomedikal atıklar vb. tehlikeli atık gruplarının iş sağlığı ve güvenliği ilkelerine uygun şekilde depolanması, taşınması ve bertaraf edilmesi, çevre ve insan sağlığının zarar görmemesi için hayati önem taşır.
- Bulaşıcı hastalıkların çevresel kontrolü halk sağlığının korunması için önem arz etmektedir (Yanar, 2024).

b. Tedavi Edici Sağlık Hizmetleri

Aile hekimleri ve ilgili sağlık personellerinin, koruyucu sağlık hizmetlerinin hedeflerini gerçekleştirmeye çalışmalarına rağmen, farklı sebeplerle tedavi edici sağlık hizmetlerine ihtiyaç olabilmektedir. Bu hizmetler genel olarak; devlet hastanelerinde, özel hastanelerde, aile polikliniklerinde, üniversite hastaneleri gibi kurumlarda gerçekleşmektedir

Bu aşamada bireylerin hastalık süreçlerinin başladığı söylenebilir. Hasta bireyler çeşitli sağlık kurumlarında tedavi görmektedirler. Bu hizmetler, tedavi sürecince kullanılacak ilaçların ve sağlık teçhizatlarının termin edilmesini de içermektedir.

Tedavi edici sağlık hizmetleri; hasta bireylere hastalık tanısı konması ve tedavi süreçlerinin gerçekleşmesinin yanı sıra hasta bireyin ihtiyaç duyduğu tüm ilaç ve medikal ekipmanların temin edilmesi sürecini de kapsar. Koruyucu sağlık hizmetlerini herhangi bir nedenle kullanamayarak rahatsızlanan hastalar, yataklı veya ayakta tedavi hizmeti sunan sağlık kurum ve kuruluşlarına başvurmaktadır. Bu hizmet basamağına gelen hastalara uygun tedavi yöntemleri kullanılarak, tekrar eski sağlıklarına kavuşmaları sağlanmaya çalışılmaktadır (Yanar, 2024).

c. Rehabilitasyon Hizmetleri

Hastalığın sebep olduğu yeti kaybı oranının azaltılması, fiziksel ve ruhsal bozukluğa yol açabilecek yan etkilerin hafifletilmesine yönelik rehabilitasyon çalışmalarını kapsar (Abay ve Çölgeçen, 2018).

Rehabilitasyon; bireyin hastalık veya yaralanma sonrasında yeniden fonksiyonlarını en yüksek düzeyde yerine getirebilme süreçleridir. Rehabilitasyon hizmetlerinin amacı; bireyin

toplum ile yeniden entegrasyon içerisinde olabilmesi için fiziksel, entelektüel ve duygusal becerilerini yeniden kazanmasını sağlamaktır (Gürhan ve Üstün, 2015).

Sağlık Hizmeti Veren Kurum ve Kuruluşların Kademeleri

Üç kategoride incelenebilmektedir: (https://antalyaism.saglik.gov.tr/TR-234669/saglik-hizmet-sunucularinin-basamaklandirilmasina-dair-yonetmelik.html).

1. Birinci Kademe Sağlık Hizmetleri
2. İkinci Kademe Sağlık Hizmetleri
3. Üçüncü Kademe Sağlık Hizmetleri

1. Birinci Kademe Sağlık Hizmetleri

Sağlık sistemlerinin ilk basamağıdır. Dolayısıyla; sağlık hizmetlerinin yerinde ve doğru zamanda verilmesi, oluşabilecek komplikasyonların önlenebilmesi açısından kritik bir öneme sahiptir (Yanar, 2024). Acil müdahale gerektirmeyen genel muayene alanlarıdır. Sağlık hizmetlerinin en büyük kısmını oluşturmaktadır. Sağlık sistemine ilk giriş noktasıdır ve bütçe harcamalarında önemli bir konumdadır. "Sağlık Ocağı" uygulaması Türkiye'nin birinci basamak sağlık hizmetleri örgütlenmesinin temelini oluşturmuştur (Akman ve Tarım, 2020).

2. İkinci Kademe Sağlık Hizmetleri

Birinci kademe sağlık hizmetlerini tamamlayıcı nitelikte olan sağlık kurum ve kuruluşlarıdır. Genel olarak bireylerde mevcut olan hastalıkların tanısının konulmasında ve tedavi sürecinin uygulanmasında, ileri düzey tıbbi teknolojilere ve yüksek düzeyde bilgi gereksinimine ihtiyaç duyulmamaktadır. Bireyler, hastalıklarının tanısının konulmasının ardından yataklı birimler kapsamında tedavi edilmektedir; ancak bu uygulamalar, ileri seviye tetkiklerin yapılması ve tedavinin uygulanması biçiminde değildir.

Birinci kademe sağlık hizmetlerinde hasta ya da yaralı bireyler, doktorların gerekli görmesi durumunda, ikinci kademe sağlık hizmetleri sunan hastanelere sevk edilebilmektedirler. Bu kişiler hastalıklarına ilişkin tanıların konulması amacıyla tetkik yaptırabilmektedirler.

İkinci kademe sağlık hizmeti veren kurum ve kuruluşlarda hastalıkların tedavileri yapılamıyorsa, hastalar özel dal hastanelerine sevk edilmektedir (Yanar, 2024).

3. Üçüncü Kademe Sağlık Hizmetleri

İleri tetkik ve özel tedavi gerektiren hastalıklar için yüksek teknolojiye sahip ekipman, araç- gereçlerinin bulunduğu, eğitim ve araştırma hizmetlerinin de verilebileceği altyapıya sahip, üst düzey hastanelerdir Alanında uzman sağlık personellerini de bünyesinde barındırmaktadır. (https://antalyaism.saglik.gov.tr/TR-234669/saglik-hizmet-sunucularinin-basamaklandirilmasina-dair-yonetmelik.html24).

Sağlık hizmetlerinden yararlanmak için hastaların birinci ve ikinci kademe sağlık hizmetleri sunan kurumlardan geçmiş olmaları gerekmektedir. Aksi takdirde yaşanabilecek yoğunluk, hizmet kalitesini düşürecektir (Yanar, 2024).

1.1.6. Sağlık Hizmetlerinde Karar Verme

Karar verme; sorun çözme sürecinde, alternatif çözümlerden bir tanesini seçme durumudur. Sorun çözme ve karar verme bilimsel süreç aşamalarından oluşmaktadır.

Tekrarlayan ve devrimsel bir süreç olarak ifade edilen sorun çözme süreci aşamaları şu şekildedir:

- Sorunun Tanımlanması: Sorun çözmenin en önemli aşaması; sorunun belirlenmesi aşamasıdır.

Çizelge 1.1' de sağlık sisteminde sorunun tanımlanmasına dair faydalanılabilecek konu başlıklarına yer verilmiştir.

Çizelge 1.1. Sorunun Tanımlanmasında Dikkate Alınması Gereken Bazı İpuçları (Abaan, S. (1996). Sorun Çözme ve Karar Verme. G. Uyer, G. Kocaman, S. Oktay, G. Argon, S. Abaan (Ed.), Hemşirelik Hizmetleri Yönetimi El Kitabı, (s.36-127) İstanbul: VKV.)

İşlerde yığılma	Araç gereç bulmak için zaman harcama
İşlerde kesintiler	Aletlerin olması gereken yerde olmaması
Sürekli daha fazla zaman isteme	Klinikteki sürtüşmeler
İnsan araç, gereç, yer, zaman yönünden kayıplar	Yürüme, eğilme, uzanma sonucu yorgunluk
Maliyette artış	"O" klinikte çalışmak istememe
Çalışanların sık sık rapor alması	Beklentileri karşılayamama
İşe gelmeme veya memnuniyetsizlik	Planlanan hedeflere ulaşamama
Etrafta sürekli kalabalık grupların olması	Sık sık ziyaretçilerle tartışma
İlaç hataları	Servis odasındaki dağınıklık
Hastaların tetkikler için uzun süre bekletilmeleri	Hastaların sık sık komplikasyonlarla geri gelmesi
Hastaların hastanede yatış sürelerinin uzaması	

- Bilgi Toplanması: Mümkün olduğunca fazla kaynaktan ve bulgularla ilgili görünenlerin tümü hakkında veri toplamak önemlidir. Veri toplama sürecinde olabildiğince çok soru sorup, bu soruların cevabına ulaşmak, asıl soruna ve sorunların sebeplerine ulaşmayı kolaylaştıracaktır.

  Çizelge 1.2'de veri toplama aşamasında sorulabilecek bazı sorulara yer verilmiştir.

Çizelge 1.2. Sorun Çözme Sürecinde Kullanılabilecek Bazı Sorular ([26] Abaan, S. (1996). Sorun Çözme ve Karar Verme. G. Uyer, G. Kocaman, S. Oktay, G. Argon, S. Abaan (Ed.), Hemşirelik Hizmetleri Yönetimi El Kitabı, (s.36-127) İstanbul: VKV)

Soru Tipleri	**Örnek Soru İfadeleri**
Kim	Kim bunun bir sorun olduğunu söylüyor? Soruna kim sebep oldu ya da oluyor? Kimler etkilendi?
Ne	Ne oldu ya da oluyor? Belirtiler ne? Elinizde bunun sorun olduğunu gösteren neler var? Başkaları için ne tür sonuçlar var? Sorunun ortaya çıkmasına neden olan koşullar nedir? İstenildiği gibi yürünmeyen nedir? Sorunu tanımlamak için ne tür bilgiye gereksinim var? Eldeki verilerden ne tür sonuçlar çıkarılabilir?
Ne zaman	Ne zaman oldu? İlk ne zaman fark edildi? Ne zamandan beri olduğu düşünülüyor?
Nerede	Sorun nerede? İlk nerede meydana geldi? Sorun neleri etkiliyor ya da etkileyecek?
Niçin	Bu niye bir sorundur? Neden oldu? Sorunun ortaya çıkmasını engellemek için neden bir şey yapılmadı? Neden birisi hemen sorunu görüp bir şeyler yapmadı? Neden şimdi bu soruna çözüm bulmak gerek?
Nasıl	Süreç nasıl yürütülebilir? Başkaları bu tür sorunları nasıl çözüyor? Bunun bir sorun olduğunu nasıl anlıyorsunuz?
Genel sorular	Sorunun çözüldüğünü nasıl anlayacaksınız? Beklenen durum/sonuç nedir? Soruna ilişkin farklı görüşler var mı? Bu sorulara yanıt bulmak için ne tür verilere ihtiyacımız var?

- Hedeflenen Sonuçların Tanımlanması: Soruna netlik kazandırdıktan sonra, yapılması düşünülen girişimler ve bu girişimlerin sonunda gerçekleşmesi hedeflenen bireysel ya da örgütsel sonuçlar belirlenmelidir.
- Çözümlerin Geliştirilmesi: Beklenen sonuçlara erişebilmek için mümkün olduğunca fazla çözüm önerileri geliştirilmelidir. Böylece en etkili çözüm yoluna karar verilmesini ve alternatif planın olası bir başarısızlık durumunda devreye alınmasını sağlamaktadır.
- Sonuçların Değerlendirilmesi: Bu aşamada karar verme tekniklerinden yararlanılarak geliştirilen çözüm seçenekleri, belirli hedefler ile uyum ve öncelik açısından, maliyet, etkililik, zaman, yasal ve etik yönden değerlendirilmektedir.
- Karar Verilmesi: En iyi çözüm yöntemine karar verilirken geçmiş deneyimlerden çıkarımlar yapılabilir ancak; sağlık sistemlerinde yaşanan hızlı değişimler, bulunan çözüm yöntemin güncelliğini yitirmesine neden olabilir. Karar vermede seçilen

yöntem, alınan kararın alternatif kararlar arasından seçildiğini net bir biçimde gösterebilmelidir.

- Çözümlerin Uygulanıp Değerlendirilmesi: Bu adımda alınan kararların uygulanması için aksiyon alınır. Karar alındıktan sonra, uygulanması için detaylıca plan yapılmalıdır. Planlama, görevlendirme ve koordinasyonun nasıl olacağı açıkça ifade edilmelidir. Planlanan faaliyetlerle ilgili olarak performans ölçütleri oluşturulmalıdır. İzleme ve kontrolün kimler tarafından ne sıklıkla rapor verileceği belirlenmelidir (Yorgancılar ve Özlük, 2022).

Karar verme, sorun çözme sürecinin bir aşamasıdır. Karar verme süreci, sorun çözme süreci ile benzer aşamalara sahiptir. Karar verme süreci aşamaları;

- Sorunun tanımlanması,
- Durumun analizi,
- Seçeneklerin belirlenmesi,
- En iyi seçeneğin seçilmesi,
- Kararın uygulanması,
- Sonucun değerlendirilmesi.

şeklindedir (Yorgancılar ve Özlük, 2022).

Sağlık hizmetlerinin etkili bir şekilde yapılabilmesi için, karar verme süreçleri kritik bir öneme sahiptir. Dijitalleşme ile beraber sağlık bilgi sistemleri önem kazanmıştır. Bu sistemlerin en önemli avantajları veri toplama, işleme ve kullanılması olarak söylenebilir. Karar Destek Sistemleri; kurumların sağlık harcamalarının takip edilmesi, hastalık yoğunluğu analizlerinin yapılması, demografik verilerin hızlıca incelenmesi vb. faaliyetler için kolaylık sağlamakta, bu analizler sonucunda alınacak kararların, daha isabetli olmasını sağlamaktadır (Yanar, 2024).

1.2. Veri Madenciliği, Büyük Veri, Makine Öğrenmesi

Bu bölümde günümüzde oldukça popüler olan, kavramsal olarak da birbirlerinin içine geçmiş durumdaki büyük veri, veri madenciliği, makine öğrenmesi terimleri ve bu teknolojilerle ilişkili olan temel kavramlar incelenmiştir.

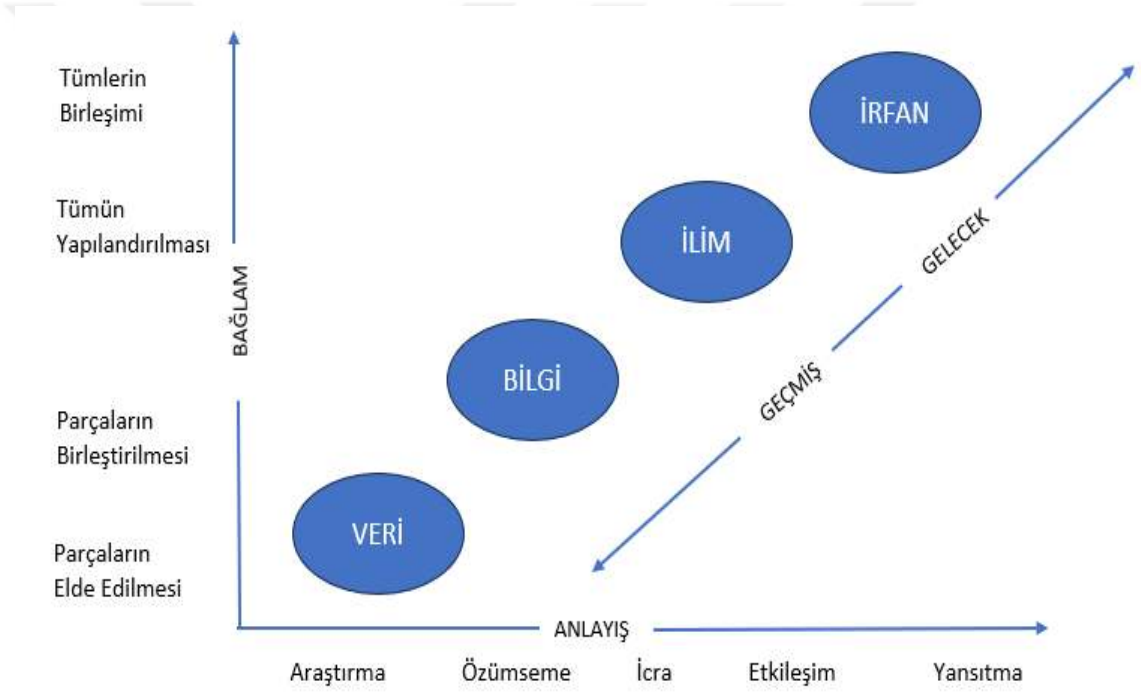
1.2.1. Veri

Nitel veya nicel değişken kümesinden oluşan bilgi parçacıklarıdır. Veri (data) ile bilgi (information) kavramları birbirlerinin yerine de kullanılabilmektedir; ancak bilgi, verilerin işlenmesiyle elde edilen nicel ya da nitel değerlerdir. Bu karmaşayı sonlandırmak için bilgisayar

bilimlerinde veri yerine genellikle ham veri ya da işlenmemiş veri kavramları kullanılmaktadır (Köse, 2018).

Günümüzde web, veritabanları, sosyal medya, algılayıcılar gibi birçok farklı alanda çok büyük boyutlarda ham veriler üretilmektedir. ML ve veri madenciliği gibi YZ'nin farklı alanları ham veriden bilgi elde edilmesi süresine katkıda bulunurlar. Ham veriden bilgiye dönüşüm süreci, ham veriler arasında fark edilemeyen, gizli kalmış ilişkilerin ortaya çıkarılması olarak da ifade edilebilir. Veriler, kendilerinden bilgi elde edilmediği sürece hiçbir anlam barındırmazlar (Oğuz, 2023).

Veriyi işleyerek bilgi, birden fazla bilgiyi birleştirerek ilim (knowledge), ilimleri bir araya getirerek de irfanı (wisdom) elde edebiliriz. Veri, bilgi, ilim, irfan arasındaki ilişki kavramsal bir model olarak literatürde yerini almıştır (Ackoff, 1989).



Şekil 1.1. Veri-Bilgi-İlim-İrfan İlişkisi

Şekil 1.1 'de ifade edilen ilişkiye göre, örneğin; bir hastanenin girişinde her bir hastanın kaydı esnasında veri tabanlarına aktarılanlara veri , gün içerisinde gelen hastalara dair hesaplanan değere, bilgi denilmektedir. Şekilde yukarı-sağa doğru çıkıldıkça, kliniklere, doktorlara, hastalıklara göre hasta sayısı, hastalara yapılan işlemlerin maliyeti gibi hesaplamalar yapılırsa, bu durum ilim olarak adlandırılabilir. Kaydedilen ham verilerden oluşturulabilecek tüm raporlar, ilim kapasitesini artırır, bütüne dair algı oluşturur. Türetilen bilgiler ve bu bilgilerden elde edilen ilim kapsayıcılığı arttıkça, bütünün algılanması o ölçüde gerçekçi olacaktır (Köse, 2018).

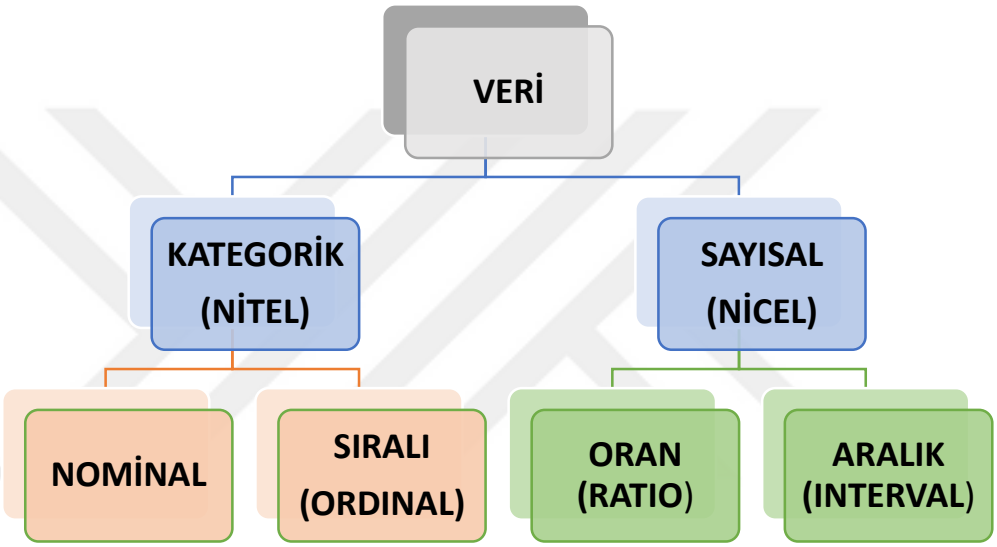
İlim kapasitesi ne kadar artarsa artsın, sadece geçmiş ile ilgili bilgi sunar. Geleceğe yönelik projeksiyonlarda bulunmak için irfan sahibi olunması gerekmektedir. İrfan sahibi olabilmenin bilgi

sistemlerindeki karşılığı, sisteme dair bir modelin oluşturulup, bu model aracılığıyla gelecekteki davranışın tahmin edilmesidir. Model oluşturma süreci de genellikle "öğrenme" olarak isimlendirilmektedir (Köse,2018).

Veri Sınıflandırması

Veri madenciliğinde karşılaşılan temel veri sınıflandırmaları şunlardır:

İstatistiksel Sınıflandırma: Verinin matematiksel ve mantıksal özelliklerine göre sınıflandırılmasıdır.



Şekil 1.2. Mantıksal Özelliklerine Göre Verinin Sınıflandırılması

- **Nominal:** Latince "nomen" kelimesinden türetilmiştir. Sayısal bir değere sahip olmayan verileri isimlendirmek ve niteliklere isimlendirilmiş değerler atamak için kullanılan, niceliksel olarak ifade edilemeyen veri türüdür. Veri kümesinde bulunan kişi, nesne, kurum adları ile kategorik veriler nominal veri örnekleridir. Örneğin; il adı, il plakası kodu , kan grubu, cinsiyet, milliyet gibi değişkenler nominal veriler arasında yer alır. Toplama, çıkarma, çarpma, bölme gibi matematiksel işlemler nominal veri üzerinde gerçekleştirilemez. Bu nedenle ortalama, varyans gibi istatistiksel fonksiyonlar nominal verilere uygulanamaz; ancak frekans değeri (Mod) belirlenebilir.
- **Sıralı (Ordinal):** Nominal verilerin özelliklerini taşımanın yanı sıra doğal olarak sıralanabilen veri türüdür. Niteliklerine adlandırılmış değerler atanabilir; ancak nominal verilerin aksine artan veya azalan bir dizi halinde düzenlenirler. Böylece bir değerin başka bir değerden büyük veya küçük olduğu yorumları yapılabilir. Müşteri memnuniyeti; "Çok Memnun", "Memnun", "Memnun değil", notlar;"A","B","C",

metalin sertliği;" Çok sert", "Sert", "Yumuşak" gibi değişkenler sıralı veriye örnek teşkil eder. İstatistiksel olarak frekans (Mod) değeri hesaplanabilir. Ayrıca sıralama yapılabildiği için medyan ve çeyrekler de hesaplanabilir; ancak ortalama hesaplanamaz.

- **Aralık (Interval):** Sıralamanın yanı sıra, değerler arasındaki kesin farkın da bilindiği sayısal veri türüdür. Santigrat derece sıcaklık verileri, tarih, saat, boy, kilo ve yaş aralık verilere örnek olarak verilebilir. Bu veri üzerinde toplama ve çıkarma yapılabilir. Bu nedenle istatistiksel hesaplamalarda merkezi eğilim ölçüleri olan ortalama, medyan, mod ile dağılım ölçülerinden olan standart sapma hesaplanabilir. Ancak; aralık verilerinde "mutlak sıfır" değeri bulunmamaktadır. Örneğin; "0 sıcaklık" veya "sıcaklık yok" şeklinde bir ifade yoktur. Bu nedenle aralık verileri için sadece toplama ve çıkarma işlemi yapılabilir. Oran hesaplanamaz. Yani 40°C'lik sıcaklığın 20°C'lik sıcaklıktan iki kat daha sıcak olduğu söylenemez.
- **Oran (Ratio):** Kesin değerler ile ölçülebilen sayısal veri türüdür. Oran verilerinde mutlak sıfır tanımlanabilir. Bu veri üzerinde toplama, çıkarma, çarpma, bölme işlemleri uygulanabilir. Merkezi eğilim ölçülerinden ortalama, medyan, mod ile dağılım ölçülerinden standart sapma ölçümleri yapılabilir. Vücut kitle endeksi, yoğunluk, ivme gibi değişkenler oran verilerine örnek teşkil eder.
- **Sürekli (Continuous):** Ardışık değerlerin elde edilebileceği, genellikle ölçümle elde edilen veri türüdür. Sıcaklık, boy, yükseklik, kilo gibi herhangi bir olası değer alabilirler.
- **Ayrık (Discrete):** Sayısal olmakla beraber, ardışık olması gerekmeyen, bir durumun sayısını ifade eden veri türüdür. Sonlu veya sayılabilir sonsuz sayıda değerler alabilir. Örneğin; hasta sayısı, ameliyat sayısı gibi. Yalnızca iki değer alabilen özel ayrık nitelik türleri ise örneğin; kadın/erkek, pozitif/negatif, evet/hayır gibi durumlardır.

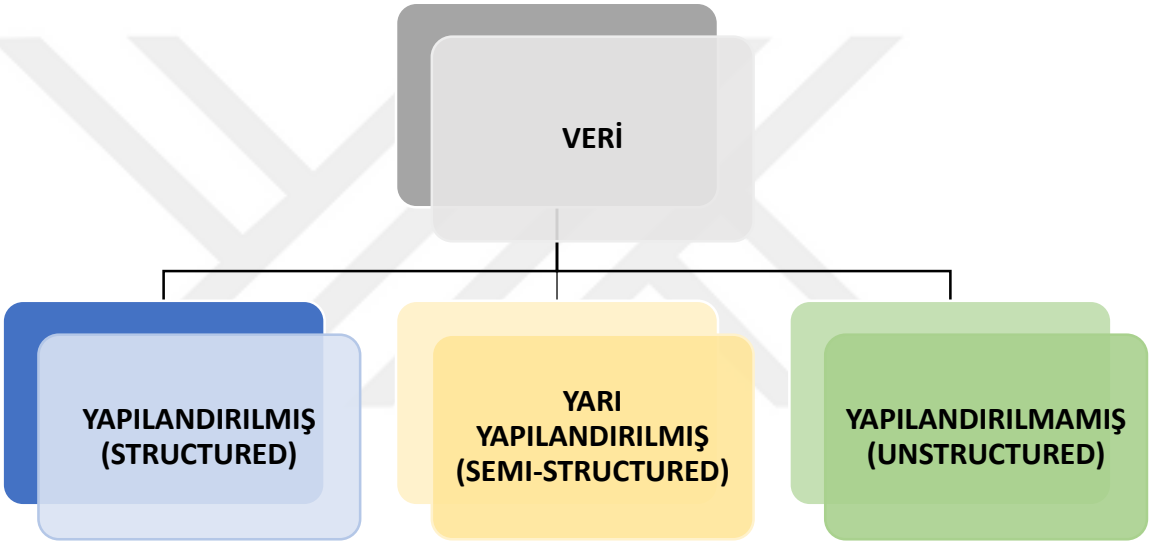
Ayrıca; veri madenciliğinde hesaplamalarda özel bir anlam ifade etmemesine rağmen, veri türleri nitel ve nicel olarak da gruplandırılır.

- **Nitel (Quality):** Ölçülemeyen bir nesnenin veya bilginin kalitesi hakkında bilgi verir. Veriye ait kategorik ve tanımlayıcı özelliklerdir. Örneğin; işçilerin performans kalitesinin "iyi, "kötü", "orta" olarak sınıflandırılması, işçilerin adı veya sicil numaraları bu gruba girmektedir.
- **Nicel (Quantity):** Veriye ait ölçülebilir ve sayısal olarak ifade edilebilen niteliklerdir. Sayısal verileri incelemek için en etkili iki yöntem; kutu grafiği (box-plot) ve histogram yöntemleridir (Dutt ve ark., 2018).

Veri madenciliği yöntemleri açısından iki tür veri (sayısal, kategorik) vardır. Tüm veri madenciliği yöntemleri, verilerdeki sayısal değerler vasıtasıyla bir hesaplama yapmakta ya da değerlerin kategorik olması halinde ilgili değerin veri kümesindeki frekansları ile ilgilenmektedir (Köse, 2018).

Yukarıda daha fazla veri türü tanımlanmıştır. Bu tanımlamalara ihtiyaç duyulmasının nedeni, veri madenciliği yöntemlerinin başarısını artırmak için veri ön işleme ile veri dönüştürme aşamasında yapılan işlemlerdir. Verinin türü ve bu türle ilişkili olarak kayıtlardaki çeşitlilik, kullanılan yöntemle beraber değerlendirilir ve veri madenciliği çalışmasının sonucunu direkt olarak etkiler.

Yapısal Sınıflandırma: Verinin fiziksel düzenine göre sınıflandırılmasıdır.



Şekil 1.3. Fiziksel Düzenine Göre Verinin Sınıflandırılması

- **Yapılandırılmış Veri:** Belirli bir biçime ve uzunluğa sahip , depolanması ve analiz edilmesi diğer veri türlerine göre daha kolay olan, yüksek düzeyde organize edilmiş veri türüdür. Bu özellik, verilerin kurumsal amaçlarla bilgi elde etmek için yapılacak sorgulara yanıt verebilecek şekilde düzenlendiği anlamına gelir. SQL, Access gibi ilişkisel veri tabanlarından basit ve anlaşılabilir arama-tarama algoritmaları ile sorgulama yapılarak ulaşılabilen verilerdir.
- **Yarı Yapılandırılmış Veri:** Sabit ve açık bir şema ile uyum sağlamayan, düzensiz veri türüdür. Bu özellik; yarı yapılandırılmış verilerin ilişkisel veri tabanı modelindeki gibi tablo odaklı olmadığını gösterir. Yarı yapılandırılmış veri modeli, birbiriyle ilişkili fakat farklı özelliklere sahip çeşitli kaynakların bir bütün halinde bir araya getirilmesini sağlar. E-mailler, XML dosyaları, CSV dosyaları, Web sayfaları örnek olarak gösterilebilir.

- **Yapılandırılmamış Veri:** Belirli bir yapısı olmayan veri türüdür. BV teknolojilerinin gelişmesi ile birlikte kurum ve kuruluşlardaki verilerde yaklaşık olarak %70-80 civarında oran ile yapılandırılmamış veri bulunduğu keşfedilmiştir. Metin , ses, video, resim, mesajlar, sosyal medya paylaşımları, açık uçlu soruların yer aldığı anketler, yapılandırılmamış veriye örnek olarak gösterilebilir.

Yapılandırılmış verilerin işlenmesi, depolanması ve manipüle edilmesi, yapılandırılmamış veri türlerine göre daha kolaydır. Bu nedenle verimli bir analiz yapabilmek için yapılandırılmamış verilerin, veri dönüştürme işlemleri ile yapılandırılmış hale getirilmesi gerekmektedir (Bansal ve ark., 2021; Hanig ve ark., 2010)

1.2.2. Veritabanı

Basitçe bir ifade ile; organize olmuş veri topluluğudur. Bu durumda verinin saklandığı ortamın profesyonel bir veritabanı yönetim sistemi (My SQL, Postgre SQL, MS SQL, Oracle gibi.) tarafından yönetilmesi gerekmez. Serbest bir metin dosyası, ses dosyası, video dosyası, MS Excel dosyası vb. ortamlar da eğer veri amaca uygun bir biçimde organize halde saklanıyorsa, veritabanı olarak değerlendirilir.

1.2.3. Veri Ambarları ve OLAP

Veri madenciliği çalışması yapabilmek için, veri ve veritabanı gerekmektedir. İşletmelerde kullanılan veritabanları doğrudan veri madenciliği işlemlerinde kullanılamaz. Bu veriler, veri madenciliği amacıyla kullanılabilmek için uygun hale getirilmelidir. Belirli bir döneme ait olan, yapılacak çalışmaya göre düzenlenmiş, birleştirilen, sabitlenen işlemlere ait veritabanlarına "veri ambarları" denilmektedir (Silahtaroğlu, 2020). Başka bir tanıma göre; bir veya daha fazla veri kaynağından elde edilen verilerin birbirine entegre şekilde toplanarak raporlama ve veri analizi amacıyla kullanıldığı veritabanlarına, "veri ambarı" denilmektedir (Köse, 2018).

Veri ambarından analiz ve raporlamaların oldukça hızlı bir şekilde yapılmasını sağlamaktadır.

Veri ambarları:

- Konu odaklıdırlar
- Bütünleşiktirler
- Belirli bir zaman dilimine aittir
- Geçici ve uçucu değildirler

OLAP (On- Line Analytical Processing); Veri ambarları üzerinde, çeşitli taktik ve stratejik konular hakkında karar vermeye yardımcı olabilecek veri analizi ve sorgulama işlemidir.

Sorgulama işlemi, işletmelerdeki tüm ilişkisel veritabanları üzerinde yapılabilir fakat; OLAP sorgulamaları klasik sorgulamalardan farklıdır. OLAP işlemleri, bilgisayar üzerinde akıl yürüterek yapılan işlemler olarak da tanımlanabilir (Silahtaroğlu, 2020).

İlişkisel veritabanlarıda günlük kayıt giriş çıkış işlemleri yapılmaktadır. Örnek olarak; bir hastane veritabanı verilebilir. Buradaki kayıtlar günlük muayene, depo giriş, satın alma vb. kayıtlardır. İlişkisel veritabanlarında sorgulamalar günlük veya haftalık muayene, depo doluluk oranı, en sık rastlanılan rahatsızlık vb. olacaktır. OLAP sorgulamaları, çarşamba günleri dermatoloji anabilim dalında muayene olan hasta sayısının yüz kişiyi aşma olasılığı, başak burcu müşterilerden nöroloji ana bilim dalını ziyaret eden hastaların üç ay içinde dahiliye ana bilim dalını ziyaret etme olasılığı gibi sorgulamalar yapmaktadır.

İlişkisel veritabanları kayıt personeli, veritabanı yöneticisi, sistem uzmanı gibi kişiler tarafından kullanırken, veri ambarları ve OLAP uygulamaları, analistler ve yöneticiler gibi bilgi üreten ve kullanan personellerce kullanılır.

1.2.4. Veri Madenciliği

Bilgisayarların yaşamın içerisine daha çok entegre olması ile birlikte artık her yaptığımız işlem sayısal ortamda kayıt altına alınmaya başlanmıştır.

Günümüzde veri toplama kaynakları ve bilişim teknolojileri oldukça gelişmiş ve yaygınlaşmıştır. Bu durum, kurumların ellerinde çok ciddi boyutlarda veritabanları olmasını sağlamıştır. Hastanelerde, belediyelerde, ya da ticarette, örneğin marketlerde, yaptığımız her işlem veritabanlarındaki yerini almaktadır. Bir alış-veriş merkezine girerken, yolda yürürken kameraya alınan görüntülerimiz, marketlerde yaptığımız alış-verişlerde alınan her bir ürün, iade ettiğimiz ürün ve o ürünle beraber alınan diğer ürünler vb. veriler, veri tabanlarında tutulmaktadır. Tüm bu veriler, veritabanlarından çıkarılmayı bekleyen değerli madenlerdir (Silahtaroğlu, 2020). Bu veritabanlarındaki verilerden, bilgi elde edilmek istenilmesi, veri madenciliği kavramının ortaya çıkıp, gelişmesini sağlamıştır.

Veri madenciliği, 1990'lardan beri, veri depolama araçları, barkod, RFID teknolojilerinin gelişmesi ile beraber kullanım alanları yayılmakta olan bir konudur. Her geçen gün daha da gelişmektedir. Kullanılan yer ve zamana göre farklı tanımlamaları yapılmıştır. Yaşanılan gelişmelere paralel olarak bugün yapılan bir tanımın, yarın için eksik yönleri olabilmektedir. En yaygın tanımlamalardan biri şöyledir:

"Veri madenciliği daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veritabanlarından elde edilmesi ve bu bilgilerin işletme kararları verirken kullanılmasıdır."

Bu tanımda dikkat edilmesi gereken önemli hususlardan birisi elde edilecek bilginin önceden bilinmiyor oluşudur. Bu bilinmeyiş, elde edilecek sonucum tahmin edilmemesi anlamına gelmektedir. Veri madenciliği; tahmin edilen, öngörülen ya da farklı yöntemler ile elde edilmiş

sonuçların ispatını yapmak üzere kullanılacak bir araç değildir. Bu gibi durumlarda veri madenciliği kullanmak hem ekonomik değil, hem de mantıklı değildir (Silahtaroğlu,2020).

Daha önceden bilinmeyen ya da tahmin edilemeyen bilgilere en ünlü örneklerden birisi bira-çocuk bezi örneğidir.

Bir perakende mağazalar zincirinin ürün satışları üzerinde yaptığı veri madenciliği sonucuna göre; bira ile çocuk bezi satışları arasında özellikle de cuma günleri güçlü bir korelasyon vardır. Çocuk bezi satın alan kişilerin büyük çoğunluğu, aynı zamanda bira da satın almaktadır (Cabena, 1998).

Veri madenciliği disiplinler arası bir konudur. Veri madenciliğinin, istatistik, veritabanı, karar destek sistemleri, makine öğrenmesi, gibi farklı konu ve disiplinlerle yakından ilişkisi vardır.

Veri madenciliği; büyük verilerde bulunan ilginç, keşfedilmemiş örüntüler ve ilişkilerin tespiti, yararlı bilgilerin elde edilmesi işlemidir. Veri madenciliği; istatistik ve YZ alanlarındaki araçları, veritabanı yönetimi ile bir araya getirerek, büyük verileri analiz edebilmek için kullanılır.

Veri madenciliği; bir madencilik çalışmasıdır. Klasik anlamda madencilik çalışmalarında olduğu gibi çok miktardaki bir kitle içerisinden gizli, az miktarda ve değerli olan "madenleri" ortaya çıkarıp, keşfetmeyi hedeflemektedir (Köse, 2018).

Veri Madenciliği ve İstatistik

Veri madenciliği; istatistik biliminden doğan ve makine öğrenmesi algoritmalarıyla kendisini geliştiren bir bilimdir. Bu anlamda; veri madenciliği, veri biliminin temel unsurlarından biridir.

En eski veri bilimi aracı olarak kabul edilebilecek istatistik biliminin çalışma alanı üç başlıkta incelenebilir:

1. **Tanımlayıcı İstatistik**: Ortalama, mod, medyan, standart sapma, dağılım istatistikleri gibi konu ve araçları içerir.
2. **Tahmine Yönelik İstatistik**: Regresyon, zaman serileri, olasılık ile tahmin gibi konuları kapsar.
3. **Testler:** Elde edilen bulguların doğruluğunu ölçmek için t-testi, ANOVA, Cronbach Alfa, Durbin Watson gibi test istatistiklerini kapsar (Silahtaroğlu, 2020).

Bu kapsamda veri madenciliği ile istatistik biliminin benzerliği şu şekilde özetlenebilir.

Tanımsal istatistik bölümü veri madenciliğinde kümeleme yöntemleri dahilinde kullanılır. Kümeleme modelinde ML denetimsiz öğrenme algoritmaları aracılığıyla daha önceden bilinemeyen gizli kümeler ortaya çıkarılır. K-ortalama, Bulanık C-ortalama, DBSCAN gibi ML algoritmaları kullanılabilir. Tanımsal istatistikten, kümeleme aşamasından sonra yararlanılır (Silahtaroğlu, 2020).

Tahmine yönelik istatistik bölümü veri madenciliğinde sınıflandırma modellerinde kullanılır. Bu tarz modellerde, veri madenciliği lojistik regresyon, bayes olasılık modelleri gibi yöntemleri istatistikle ortak kullandığı gibi, yapay sinir ağları, derin öğrenme, genetik algoritmalar gibi denetimsiz öğrenme algoritmalarını da kullanır (Silahtaroğlu, 2020).

Testler kısmında ise veri madenciliği, öğrenme prosesinin denetimini kullanmaktadır. Doğruluk, hassasiyet, duyarlılık gibi ölçütlerin haricinde, C-endeksi, Jaccard endeksi, Silhouette gibi endeks ve algoritmalar kullanmaktadır (Silahtaroğlu, 2020).

Veri madenciliği ve istatistik bilimin farkları şu şeklide özetlenebilir.

Veri madenciliğini, istatistik biliminden ayıran en önemli fark; istatistikte ana kitleye ait tüm verilerin elde edilememesi nedeniyle, ana kitleyi temsil edebilecek örneklem üzerinde çalışılıp, ana kitleye dair bilgi edinilmeye çalışılmaktadır. Veri madenciliğinde araştırmacının elinde çok fazla veri vardır. Bu devasa miktardaki verinin içerisindeki gizli bilgilerin ortaya çıkarılması, keşfedilmesi gerekmektedir. İşlem yapılan verinin kapsam ve miktarı iki bilim arasında fark gözetmektedir (Köse,2018).

Veriden bilgi elde etme konusunda uzunca bir süredir bilimsel bir çalışma alanı olsa da, analiz edilmek istenilen kitleye ilişkin verilerin elde edilmesindeki güçlükler, araştırmacıları daha az veri ile daha güvenilir analizler yapılabilmesini sağlayacak yeni metotlar geliştirmeye yönlendirmiştir. Bu sayede de istatistik, uzunca bir dönem gelişimine devam etmiş ve oldukça popüler bir bilim olarak bilim dünyasına katkı sağlamıştır. Veritabanı yönetim sistemlerinin gelişmesi, kablolu-kablosuz haberleşme teknolojilerinin daha büyük boyutta verinin transfer edilmesini sağlayabilmesi, bir konu hakkında verinin tamamına ulaşılabilmesinin önünü açmıştır. Bu noktada ise temel prensibi veriden bilgi elde etmek olan konvansiyonel istatistik yöntemleri yetersiz kalmış ve büyük veri kitleleri üzerinde çalışabilen veri madenciliği yöntemleri geliştirilmeye başlanmıştır (Köse, 2018).

Veri Madenciliğinin Uygulama Alanları

Veri madenciliğinin kullanılmasında sektör farkı olmamakla birlikte, geniş veri ambarlarının oluşturulmasına olanak veren, perakende, satış, sigortacılık, sağlık gibi alanlarda daha yaygın kullanılmaktadır.

- **Müşterilerin satın alma örüntülerinin belirlenmesi:** Müşterilerin herhangi bir ürünü aldıktan sonra istatistiksel olarak anlamlı bir periyotta başka bir ürün alma durumunu araştırır. Örneğin; Laptop ve mouse alanların %40'ı 2 hafta sonra laptop çantası alıyor olsun. İşletme açısından önemli bir bilgi olacaktır ancak; yukarıda veri madenciliği tanımlanırken, daha önceden bilinmeyen ya da tahmin edilemeyen bilgilerin elde edilmesi için kullanılır, ibaresinden bahsedilmiştir. Laptop aldıktan 15 gün sonra laptop çantasının alınması deneyimli bir işletmecinin rahatlıkla tahmin

edebileceği bir olgudur. Veri madenciliğinden beklenen durum bu değildir. Veri Madenciliği, geleneksel tekniklerin dışına çıkarak, Laptop satın alanların % 10' u 15 gün içerisinde mağazaya tekrar gelip elektronik cihazlar ve tamamlayıcı ürünlerinin dışında bir ürün alıyorlar gibi tahmin edilmesi zor bilgileri ortaya çıkarır.

- **Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması:** Müşterilerin yaş, medeni hali, eğitim durumu, cinsiyet vb. özelikleri ile satın aldıkları ürünler arasındaki ilişkiyi ortaya çıkarır.
- **Mecvut müşterilerin elde tutulması, yeni müşterilerin kazanılması:** Mevcut müşterilerin bağlılığı test edilir ve kaybedilmeye ya da kazanılmaya en yakın müşteriler belirlenerek, kişiselleştirilmiş kampanyalar sunulabilmektedir. Her sektörde bu araştırma yöntemi kullanılabilir. Bazı sektörler daha çok veri kaydı içerebildikleri için, daha avantajlı olabilmektedir. Örneğin; telekomünikasyon şirketleri bu araştırmayı çok sık yapmaktadırlar. Kullanıcının ne zaman hat değiştireceğini ön görmek, veri madencilerinin işidir.
- **Pazar sepeti analizi:** Müşterinin birden fazla ürün aldığında, kombin ürünler yapıp-yapmadığı veri madenciliği yöntem ve teknikleri ile belirlenebilir.
- **Risk yönetimi ve dolandırıcılık saptama:** Kredi kartı dolandırıcılığı, sigorta dolandırıcılığı, kara-para aklama, telefon dolandırıcılığı, üyelik abonelik dolandırıcılığı, bilgisayar sistemleri ve ağlarına girme gibi dolandırıcılık türleri teknolojinin ilerlemesi ile gelişmiştir. Dolandırıcılık tüm dünya endüstrisi için her yıl milyonlarca dolar kayba sebep olan büyük bir sorundur (Silahtaroğlu,2020) Master-Card International 1997' deki satışlarının %7,7 sinin dolandırıcılık faaliyetleri kapsamında gerçekleştiğini bildirmiştir (Frenkel, 2003). Müşterilerine kredi kartı veren finans kuruluşları dolandırıcılığı, bilgi keşfi, yapay zeka, veri madenciliği gibi yöntemleri kullanarak tespit edebilmektedir.
- **Biyoloji (DNA Sıra Analizi):** İnsanda yaklaşık olarak 100.000 adet gen olduğu bilinmektedir. Hastalıklara yol açan gen sırasını bulmak ve tanımlamak oldukça zor bir iştir. Veri madenciliği ile geliştirilen benzerlik arama yöntemleri DNA üzerinde analiz yapabilmeyi kolaylaştıran bir araçtır (Silahtaroğlu, 2020).
- **Tıp:** Ameliyat riski taşıyan fakat; ameliyat öncesinde hastanın gerçekten ameliyat ile tedavi edilmesi gerekliliğinin analizi veri madenciliği ile yapılır. Ayrıca; parmak izi tespiti, yüz tanıma ile kimlik tespiti gibi uygulamalar, YZ teknikleri, veri madenciliği için de geçerlidir.

1.2.5. Büyük Veri

Akıllı telefonlardan, TV'lerden, kredi kartlarından, sensörlerden, bilgisayarda kullanılan fare tıklamalarından, uçaklardan, trenlerden, gemilerden, sokaktaki kameralardan vb. birçok alandan veri oluşmaktadır. Her geçen gün üretilen verinin miktarı artmaktadır.

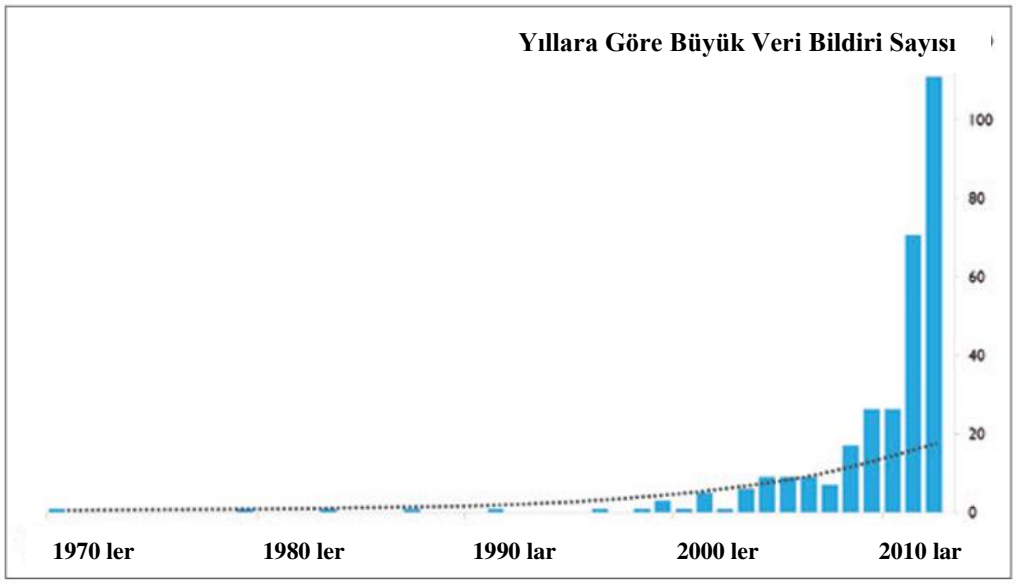
İnsanoğlu 1900'lü yılların sonu ile 2000'li yılların başında, tarihinde gördüğü en büyük değişimlerden birini, belki de en büyük değişimini yaşamıştır. Bu dönemde eğlence, ticaret, iletişim, sağlık, eğitim, sanat vb. gibi insan faaliyetleri sanal bir aleme taşınmıştır.

Günümüzde veri üstel bir hızla artarken, analiz edilebilen verinin oranı azalmaktadır. İnsanoğlu halen akan verinin %1' inden daha azını analiz edebilmektedir. Bu oran, gelişmiş ve gelişmekte olan ülkeler açısından çok farklılık göstermektedir (Gürsakal, 2017).

Büyük veri (Big Data) kavramı, ilk kez Francis X. Diebold tarafından ortaya atılmıştır. "Big Data Dynamic Factor Models for Macroeconomic Measurument and Forecasting" adlı bildiri, Ağustos 2000'de 8. Dünya Ekonometri Kongresi'nede Seattle'da sunulduktan sonra yayınlanmıştır. Bu bildiride Diebol, verilerin daha fazla toplandığını, veri depolama yöntemlerinin ve veri analiz süreçlerinin değişmek zorunda olduğunu beyan etmekte ve bazı öneriler sunmaktadır (Gürsakal, 2017).

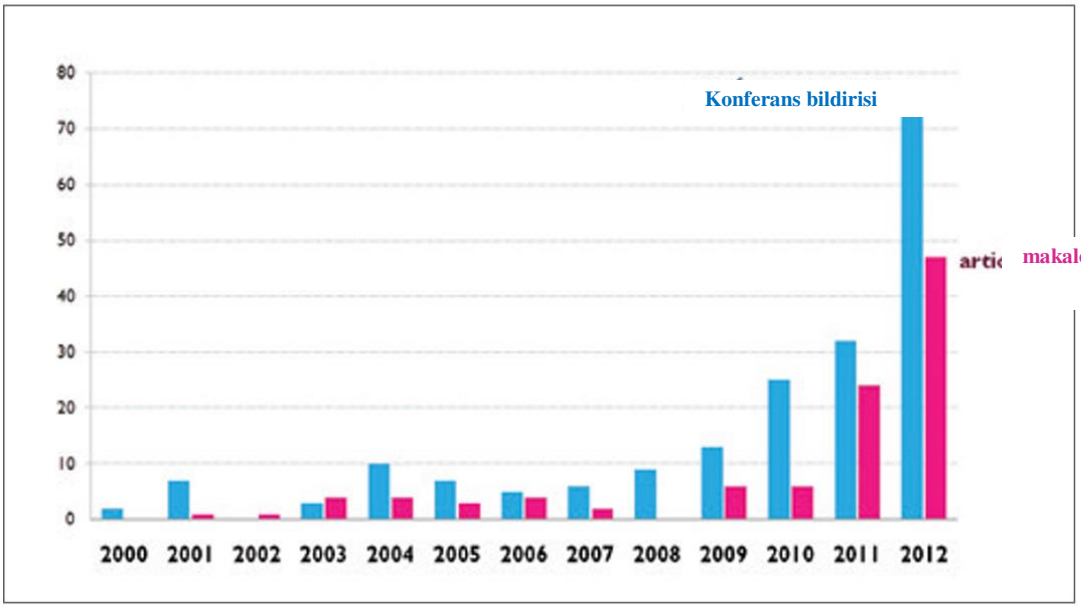
Büyük veri, analiz edilebilmesi için yeni yöntemler, yemi programlar, yeni bakış açıları gerektirmektedir. Analiz araçlarında yeni gelişmeler ortaya çıktıkça, bilimin yönetimi de verinin büyüklüğünden etkilenir hale gelmiştir. Diğer bir ifade ile büyük veriyi kullanabilen, analiz ederek, anlamlı bilgiler elde eden devletler ve kurumlar bilimi yöneterek, kendilerine ekonomik, sosyal, siyasi avantaj sağlayabileceklerdir. Gelecekte ekonomik rekabetin veri üzerinden olacağı, büyük verinin, büyük üstünlükler sağlayacağı yönünde çalışmalar da mevcuttur.

2012 yılı, gelişmiş ülkeler açısından "Büyük Veri Yılı" olarak anlamlandırılabilir. 2012 yılında World Economic Forum Davos toplantılarında, Büyük Veri Büyük Etki (Big Data Big Impact) başlığı ile bir rapor yayınlanmıştır. Çok sayıda dergi büyük veri ile ilgili özel sayı çıkarmıştır. ABD yönetimi 29 Mart 2012 tarihinde, büyük beri girişimini başlatarak, büyük veri hesaplamaları ile ilgili araştırma programları için 200 milyon dolarlık bütçe ayırmıştır. Ağustos 2012'de American Statistical Association ve Royal Statistical Society, ortak yayınları olan Significance dergisinde "Büyük Veri" özel sayısını çıkarmışlardır. American Mathematical Society, American Statistical Association, Mathematical Association of America ve Society for Industrial and Applied Mathematics; Nisan 2012'yi matematik, istatistik ve veri seti için farkındalık ayı olarak ilan etmiştir (Gürsakal, 2017).



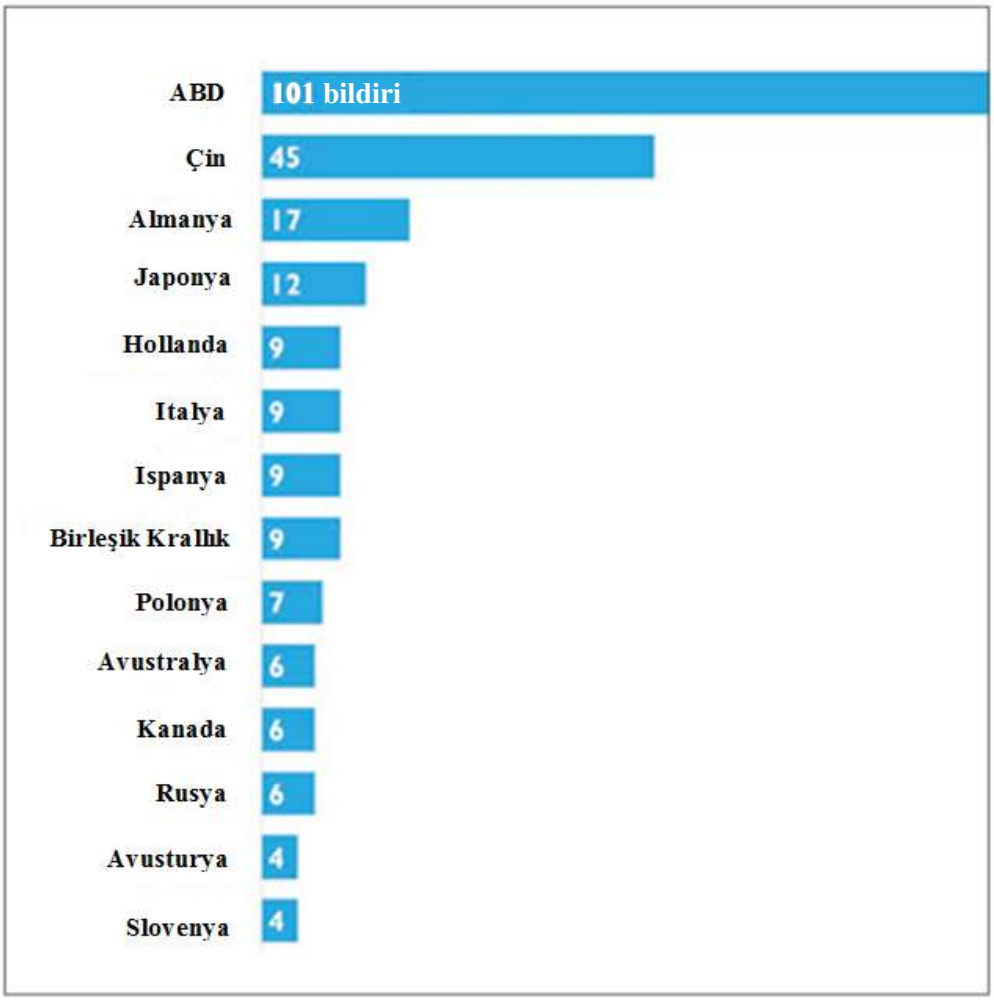
Şekil 1.4. Yıllara Göre Büyük Veri Konulu Bildiriler (The evolution of big data as a research and scientific topic: Overview of the literature, 2012)

Şekil 1.4' deki grafikte görülebileceği gibi, büyük veri konusundaki yayınlar 1970'lerde başlamaktadır ancak; araştırmacılar tarafından revaçta olması ilk olarak 2008 yılında gerçekleşmektedir. 2012 yılı ise büyük veri olgusunun iyice sahiplenildiğini göstermektedir.



Şekil 1.5. Yıllara Göre Büyük Veri Konulu Konferans Bildirileri ve Makaleler (The evolution of big data as a research and scientific topic: Overview of the literature, 2012)

Şekil 1.5'deki grafikte Büyük Veri konusunda yayınlanan konferans bildirileri ve makalelerin zaman içerisindeki değişimi gözlemlenebilmektedir. 2012 yılına gelindiğinde yaşanan net artış göze çarpmaktadır.



Şekil 1.6. Coğrafyaya Göre Büyük Veri Konulu Bildiriler (2012'ye kadar) (The evolution of big data as a research and scientific topic: Overview of the literature, 2012)

Şekil 1.6 'daki grafikte Büyük Veri konusunda 2012 yılına kadar ülkelerde sayısı incelendiğinde ABD'nin bariz üstünlüğü ve onu takip eden Çin'le birlikte toplam makale sayılarının grafikte yer alan diğer ülkelerde üretilen toplam makale sayısından daha fazla olduğu görülebilmektedir. Avrupa' da ise Almanya bu konuda öncü ülke konumundadır. ABD'nin bu alanda lider konumda olması, sektöre yön veren teknoloji şirketlerinin çoğunluğuna sahip olması, konu ile ilgili araştırma yapabilecek imkan düzeyleri elit seviyede olan üniversitelere sahip olması, araştırmacıların büyük veri ekosisteminin içinde olması gibi etkenler sayılabilir. Bu etkenler ABD'yi takip eden ülkelerde de bulunmaktadır.

Büyük veriyi kullanan şirketlere Google, Amazon, Microsoft, Facebook, Yahoo, Linkedin gibi şirketleri sayabiliriz. Bu şirketler;

- Kalabalığın gücünden faydalanmak
- BV merkezlerini kullanmak
- Bilişim ve sosyal ağlarla alakalı yeni teknolojiler üretmek

gibi ortak özelliklere sahiptir (Gürsakal, 2017).

BV; sadece büyük bir veri topluluğunu temsil etmez, aynı zamanda belirli bir amaç doğrultusunda bir araya getirilmiş verilerin analizini de içerir. BV; BV analitiği ya da BV ve analiz metotları olarak değerlendirilmektedir (Kızmaz ve Küçükçolak, 2021).

Büyük veri ifadesindeki "büyük" kelimesi sabit bir miktarı ifade etmemektedir. Verinin miktarı artarken, bu verilerin işlenmesine ilişkin metot ve teknolojiler de gelişmektedir. Bu nedenle BV, geleneksel araç ve gereçlerle işlenemeyecek kadar çok miktardaki veri kümelerini ifade etmektedir (Köse,2018).

BV; bugün ve gelecekte mevcut olan/olacak tüm dijital kaynaklardan üretilen geniş, karmaşık ve farklı kaynaklardan elde edilen verinin ciddi bir oranda artması ile meydana gelmektedir. Kaynaklara örnek olarak; sensör verisi, e-posta paylaşımları, GPS sinyalleri, imalat sensörleri, sosyal medya hesapları vb. verilebilir (Silahtaroğlu, 2020).

BV'yi tanımlayan V'ler anlamında ilk olarak 2001 yılında Douglas Laney "Data Management: Controlling Data Volume, Velocity, and Variety" adında bir bildiri yayınlamıştır. Bu bildiride üç boyutlu veri yönetiminin öneminden bahsetmiştir. Veri hacmi, veri hızı, veri çeşitliliğini daha etkili kullanabilen kurumların, daha fazla bilgi getirisi sağlayacağını ifade etmiştir (Kızmaz ve Küçükçolak, 2021).

- **Hacim (Volume) :** Büyük verinin isimlendirilmesinde temel alınan özelliğidir. Büyük veriyi değerli yapan en önemli kriterdir. Veri ne kadar çok olursa, içinde barındırdığı, açığa çıkarılmayı bekleyen gizli bilgi miktarı da aynı ölçüde artacaktır. Çokluk ile kastedilen, terabaytlar, peta baytlar düzeyidir.

(Silahtaroğlu, 2020) Dünyada üretilen veri hacminin büyüklüğü ile ilgili tahminlerde bulunan pek çok çalışma yapılmıştır. Bu çalışmalarda ortak olarak şu çıkarımlarda bulunulmuştur:

- ✓ 1980' den beri her 40 ayda bir kişi başına düşen veri miktarının 2 katına çıktığı görülmektedir.
- ✓ 1986 yılında üretilen verilerden sadece %1'i dijital veri iken, 2007'ye gelindiğinde bu oran %94'e ulaşmıştır.
- ✓ 2012 yılı itibariyle günlük 2,5 Exabyte ($2{,}5 \times 10^{18}$) veri üretilmektedir (Köse, 2018).

Sağlık sektörü; Elektronik Sağlık Kayıtları (Electronic Health Records ,EHR), Laboratuvar Bilgi Yönetim Sistemi, hastalık teşhis veya izleme cihazları, tedarik zincirleri, faturalandırma, eczane, sosyal medya gibi çeşitli kaynaklardan üretilen muazzam bir veri yığınına sahiptir (Bansal ve ark., 2021).

Büyük veri hacimlerini işleme, NoSQL veritabanları ve bulut depolama sistemleri gibi teknolojiler yardımıyla etkili bir şekilde gerçekleştirilebilir.

- **Hız (Velocity):** Verilerin hem üretim hızını hem de üretilen verilerin üzerinde hızlıca hareket edilerek verinin işlenmesini ifade etmektedir (Silahtaroğlu, 2020). Haberleşme teknolojisinin gelişmesi ile gerçek zamanlı veri akışı büyük oranda artmıştır. Gerçek zamanlı görüntü, ses ve veri akışı geleneksel yöntemlerle analiz edilememektedir. Örneğin; finansal ticaret platformları saniyede milyonlarca işlem gerçekleştirmektedir. Bu da yüksek hızla veri işleme hızı gerektirmektedir.

Tıbbi verilerin üretim hızı yüksektir ve özel olarak veri işleme ihtiyacı duymaktadırlar (Bansal ve ark, 2021).

- **Çeşitlilik (Variety):** BV'nin ses, görüntü, harita, metin vb. gibi değişik formatlardan meydana gelmesini ifade eden özelliktir. Verinin analiz edilmesindeki anlamı artırması nedeni ile kritik öneme sahip bir özelliktir. Çeşitlilik, verinin yapılandırılmış, yapılandırılmamış, yarı-yapılandırılmış türlerini ifade eder. BV'nin gerçek anlamda BV olarak tanımlanabilmesi için farklı veri türleri, değişik format ve kayıtları içermesi gerekmektedir (Silahtaroğlu, 2020).

Sağlık hizmetlerinde çeşitli kaynaklardan elde edilerek, çeşitli biçimlerde depolanır. Toplam sağlık verisinin yaklaşık olarak %80'i yapılandırılmamış verilerden oluşmaktadır. Örneğin; görüntüler, sinyaller, metinler gibi. Verilerin kalan %20'si yapılandırılmış verilerden oluşmaktadır. Örneğin; vücut sıcaklığı, kan basıncı, hasta demografisi gibi (Bansal ve ark., 2021).

Teknolojiye yön veren şirketler zamanla Douglas Lane'in 3V olarak temellerini attığı, BV özelliklerinin 5V'ye çıkmasına aracı olmuşlardır. Oracle şirketi Şubat 2013'de "Information Management and Big Data: A Reference Architecture" adlı yayınında bilgi yönetimi ve BV hakkında işletmelere yönelik birtakım değerlendirmeler ve önerilerde bulunmuştur (Kızmaz ve Küçükçolak, 2021). Değer (Value) ve Doğruluk (Veracity) kavramları literatürde yerini almıştır.

- **Doğruluk (Veracity):** Toplanan verinin sahte, yanlış vb. olmaması durumunu ifade eder. BV'nin önemli bir kısmı kullanıcı tarafından oluşturulan veriler oldukları için, doğruluk özelliğine dikkat edilmelidir. Değişime uğramış, tutarsız veriler ile analiz yapmak; zaman ve maliyet israfına yol açabileceği gibi, hatalı sonuçlar üretilmesine de sebep olacaktır (Silahtaroğlu, 2020). Veri doğruluğunun sağlanması için veri yönetim sürecinin her aşamasında kalite kontroller ve veri doğrulamalar sağlanmalıdır. Özellikle sağlık verilerinin doğru veriler olması hayati öneme sahiptir.

Yanlış ve eksik veriler, telafi edilemez sonuçlara sebebiyet verebilme potansiyeline sahiptir.

- **Değer (Value):** Verinin bu özelliği ile analiz edilmesi, kullanılması halinde kurumlar ve işletmeler için artı değerler sağlamaktadır. BV'nin hız ve çeşitlilik özellikleri dikkate alındığında, çoğu zaman elde bulunan kayıtların hepsinin değil, sadece belli bir bölümünün değerini kullanmak yeterli olabilmektedir. Örneğin; sağlık sektöründe hizmet veren ve sağlık ile ilgili stratejik kararlar alan bir kurumun; hekim bilgilerine, yoğun görülen hastalıklara, sık tüketilen ilaç türlerine, bölge, il, ilçe gibi ayrıntılar ile anlık olarak ulaşabilmesi büyük verinin "değer" özelliğidir (Silahtaroğlu, 2020; Köse, 2018).

Analiz yapmak için daha fazla veri toplanmasının nedenlerini şu şekilde açıklayabiliriz:

- ✓ Analize daha fazla veri eklemek, bir iş sorununa ilişkin etraflıca ve eksiksiz bakış açısı ile çözüm yöntemleri geliştirmeye olanak sağlar.
- ✓ Bir problemin farklı yönlerinden daha fazla veri eklendikçe, daha bütüncül bir bakış açısı elde edilir.
- ✓ Daha fazla verinin bir araya getirilmesi, verinin daha derin keşfedilmesine olanak sağlar.
- ✓ Daha fazla veri eklemek, veri bilimi problemlerinde doğruluğu artırmaya yardımcı olacaktır.
- ✓ Daha fazla veri kümesi birleştirilerek ve daha kapsamlı veriler oluşturularak, daha fazla boyuta sahip olunur.

Büyük Verinin Kullanım Alanları

BV uygulamalarının kullanılmasının önemli nedenlerinden bazıları; tüketici deneyimlerinde iyileşme sağlanması, iş süreçlerindeki maliyetlerin azaltılması, daha etkili pazarlama stratejilerinin oluşturulması gibi faaliyetlerdir. BV'nin başlıca uygulama alanları arasında bankacılık, sağlık hizmetleri, medya ve eğlence sektörü, üretim, eğitim, kamu hizmetleri, perakendecilik ve ticaret, sigortacılık, ulaşım, enerji sektörü gibi birçok alan bulunmaktadır (Aktan, 2018).

BVA yaparak gelirlerini artıran çeşitli sektörlerden çok sayıda şirket vardır.

- **Bankacılık Sektöründe Büyük Veri Uygulamaları:** BV kullanımı ile bankalar para dolaşımlarının detaylarını görebilmektedir. Tüketici davranışlarını daha iyi anlayabilmektedir. Bankalar ulusal ve uluslararası alanlarda müşteri davranışlarının analizi, risk yönetimi, finansal suçlarla mücadele etme gibi birçok alanda BV'den

faydalanmaktadırlar. Web sitesi kullanımı ve yapılan işlemlerin analizi ile potansiyel müşteri adaylarını müşterilere dönüştürmek ve çeşitli finansal ürünlerin kullanımına teşvik edilmesini sağlamak için de BV'den faydalanılmaktır. Böylece bankalar bireysel müşteri tercihlerine ilişkin profiller oluşturabilmektedirler.

- **Medya ve Eğlence Sektöründe Büyük Veri Uygulamaları:** Dünyadaki medya kuruluşları küreselleşen medya pazarındaki rekabette avantajı ele geçirmek, izleyicilere daha iyi bir içerik sunmak, müşteri incelemelerinden içgörü elde etmek, hedef kitlenin ilgi alanlarını ve tercihlerini tahmin etmek, hedef pazarlama kampanyalarını oluşturmak için BV'den faydalanmaktadırlar. BV; izleyicilerin beklentilerini, isteklerini kendileri dahi farkında olmadan tahmin ederek, içerik sunma imkanını sağlamaktadır. Haberleşme ve sosyalleşme aracı olan sosyal medya ile de işletme, kurum, kuruluş hakkında yapılan değerlendirmeler incelenerek müşteri memnuniyeti saptanmakta ve elde edilen verilerden anlamlı pazarlama bilgileri çıkarılmaktadır.
- **Kamu Hizmetlerinde Büyük Veri Uygulamaları:** Kamu kurum ve kuruluşlarında her gün petabaytlar seviyesinde veri üretilmektedir. Bu verilerin gerçek zamanlı analiz edilmesi ile, eğitim kalitesinin iyileştirilmesi, işsizlik oranlarının azaltılması, trafikteki canlı akış verisinin incelenerek trafik yoğunluğunun azaltılması ve trafik güvenliğine yönelik önlemlerin alınması, ambulans hizmetlerinin iyileştirilmesi, sosyal ve ekonomik politikaların geliştirilmesi vb. birçok alanda vatandaşlara katma değerli hizmet sunulabilme fırsatı olmaktadır. Bu hususta büyük veri, vatandaşlara sunulan hizmetin hızlı ve güvenilir olmasını sağlayarak, akıllı şehirlerin geliştirilmesinde önemli rol oynamaktadır. Ayrıca; belirli bir bölgedeki geçmişteki suç kayıtlarının incelenerek kriminal aktivitelerinin gerçekleşmeden önce tahmin edilip, güvenlik önlemlerinin alınması da kamu hizmetlerinde büyük veri uygulamalarına örnek olarak verilebilir. Devletlerin bu teknolojiye ne kadar adapte olup, faydalanabileceği konusu gelişmişlik düzeyleri de ilişkilidir.
- **Üretimde Büyük Veri Uygulamaları:** BV'nin coğrafi, metinsel, grafiksel unsurlarından bilgi çıkaran tahmin modellerinden yararlanılarak, üretim ve zamanında kaynak tahsisi alanında karar verme süreçleri desteklenmektedir. Akıllı üretim süreci ve ürün yaşam döngüsü gibi kavramlar büyük veri ile birlikte gelişmektedirler. Büyük veri analitiği ile akıllı üretim sistemlerinde aktif önleyici bakım uygulanabilmektedir. Üretim cihazlarındaki arızaları önceden tespit etmek için cihaz durum bildirimleri, cihaz olay kayıtları gibi gerçek zamanlı birçok cihaz bilgisi toplanmaktadır.
- **Perakendecilik ve Ticarette Büyük Veri Uygulamaları:** Perakendecilik sektöründe büyük veri müşteriler, ürünler, zaman, yer ve kanallar olmak üzere beş temel boyutta ele alınmaktadır. Bu boyutlar her müşteri ile ilişkili veri, ürün özellikleri ve stok

seviyeleri ile ilişkili veri, gerçek zamanlı veri akışı, coğrafi konum ve hedef verisi, tüm iletişim kanallarından elde edilen verileri kapsamaktadır. BV kullanımı ile alışveriş örüntülerinden elde edilen bilgiler, personel istihdam optimizasyonu ve müşteri ilişkileri devamlılığı için kullanılmaktadır.

- **Ulaşımda Büyük Veri Uygulamaları:** Bilgi ve iletişim teknolojilerinin yaygın olarak kullanıldığı akıllı ulaşım sistemlerinin gelişmesi ile birlikte, trafik alanında yol durumu, araç sayısı ve sürücü davranışları gibi trafik verileri büyük veriyi oluşturmaktadır. Bu verinin analiz edilmesi ile geliştirilen modeller akıllı ulaşım sistemlerinin gelişmesine katkı sağlamaktadır. BV, kamu sektörü ve özel sektör tarafından özgün amaçlar için de kullanılabilmektedir. Özel sektörde; ticaret ile uğraşanların mal transferinde araç rotasının optimizasyonunun sağlanarak rekabet avantajı elde edebilmelerine, yakıt ve zaman tasarrufu sağlayabilmelerine yönelik amaçlar için kullanılmaktadır. Kamu sektöründe; trafiği kontrol ederek yollarda tıkanıklığın önüne geçmek gibi koşulların iyileştirilmesi amacıyla kullanılmaktadır.
- **Enerji Sektöründe Büyük Veri Uygulamaları:** Enerji sektöründe sensörler, bulut bilişim teknolojileri ve kablosuz ağ iletişiminin yaygınlaşmasıyla birlikte büyük miktarda veri üretilmektedir. Enerji BV'si sadece, günün belirli periyotlarında bilgi toplayan akıllı sayaçlardan elde edilen verileri içermemektedir.. Hava durumu verisi, coğrafi bilgi sistemi gibi kaynaklardan gelen çok miktarda veriyi de kapsamaktadır. Bu verilerin analiz edilmesiyle daha verimli kaynak ve işgücü planlamasının yapılması sağlanmaktadır.
- **Telekomünikasyon Sektöründe Büyük Veri Uygulamaları:** Akıllı telefonların ve akıllı cihazların popülerliğinin artması ile beraber telekomünikasyon şirketlerinin faaliyetleri hızlıca büyümüştür. Bu şirketlerin ellerinde tuttukları BV sayesinde müşterilerine özel hizmetler geliştirebilmektedirler (Aktan, 2018).

1.2.6. Sağlık Sistemlerinde Büyük Veri

Sağlık hizmetlerinde BV; hastane, hekimler, sağlık personelleri, hasta ve tıbbi süreçlerin bütünüdür. Akıllı telefonlar, sensörler, hastalar, hastaneler, sağlık hizmeti sağlayıcıları, araştırmacılar vb. kişi, kurum ve kuruşlar. Büyük miktarda veri üretmektedir. İnsanların hayatlarını daha sağlıklı ve kolay hale getirmek için verilerin bulunması, verilerden elde edilen bilgilerin toplanması, analiz edilmesi ve yönetilmesi gerekmektedir. Bu sayede sadece yeni hastalıkları ve tedavileri anlamada değil, aynı zamanda hastalık semptomlarının daha erken aşamalarda tahmin edilmesi, gerçek zamanlı kararlar alınması da sağlanabilmektedir. Hastaların sağlığını desteklemek, tıbbı geliştirmek, maliyetleri düşürmek, sağlık hizmetlerinin kalitesini ve değerini artırmak için de BV ve BV analitiğinden faydalanılmaktadır (Asri ve ark., 2015).

Büyük sağlık verileri arasında, farklı kaynaklardan elde edilen, heterojen yapıda olan veriler bulunmaktadır. Sağlık hizmetlerinde üretilen büyük verinin başarı ile sistemle entegrasyonu, başta hastalar olmak üzere, tüm paydaşlar için fayda sağlamaktadır.

Sağlık sistemlerinde üretilip, kullanılan verilerin boyutu terabayt ve petabayt cinsindendir. Sağlık sistemleri; kan testleri, idrar testleri, tiroid fonksiyon testleri vb. kaynaklardan elde edilen sayısal veriler, radyoloji görüntüleri, bilgisayarlı tomografi (BT) taramaları, MR taramaları, röntgenler vb. kaynaklardan elde edilen görüntü verileri, metinsel veriler, kişisel bilgiler, tıbbi kayıtlar, genomik ve biyometrik sensör verileri vb. bilgilerini içermektedir. Bulut bilişimin kullanılması neticesinde bu karmaşık verileri işlemek, depolamak, kullanmak, yönetmek, analiz etmek mümkün hale gelmiştir (Agrawal ve ark., 2021; Altındiş ve Morkoç, 2018).

KPMG danışmanlık firmasının raporuna göre sağlık verilerinin hacmi 2013 yılında 150 eksabayt'a ulaşmıştır. Bu boyuttaki hacim içerisinde, yapılandırılmış, yarı-yapılandırılmış, yapılandırılmamış veri çeşitliliği nedeniyle analizde zorluklar teşkil etmektedir. Klinik verilerin (yapılandırılmış veriler) makineler tarafından işlenmesi, depolanması ve analiz edilmesi, yapılandırılmamış verileri analiz etmeye göre nispeten daha kolay bir süreçtir. Ancak sağlık sistemlerinde; doktor notları, reçeteler, tıbbi görüntüler vb. yarı-yapılandırılmış veya yapılandırılmamış sağlık verileri, yapılandırılmış verilere göre daha fazladır (Altındiş ve Morkoç, 2018).

Sağlık sistemlerinde depolanan veriler düzenli olarak güncellense bile her zaman doğru olmayabilir. Bu nedenle; gerçek zamanlı veri işlemeye dayalı olarak büyük verinin zamanında analiz edilmesi gerekmektedir. Hastaların yaşamı veya ölümü gerçek zamanlı verilere dayanmaktadır. Örneğin; enfeksiyonları önlemek ve mümkün olabildiğince en erken sürede tespit etmek, daha etkili kararlar alıp, hayat kurtarmak için büyük veri analitiğinin hızlıca uygulanması gerekmektedir (Altındiş ve Morkoç, 2018).

Veri analitiği süreçlerinin sonunda güvenilir sonuçlar elde edebilmek için, yüksek kaliteli verilere ihtiyaç vardır. Özellikle yapılandırılmamış verilerin doğru olması, veri kalitesi açısından ve veri analitiğinin etkili olması açısından önemlidir.

BV analitiği sağlık sistemlerinin gelişmesini sağlasa da, bu alanda BV kullanabilmenin bazı kısıtlamaları mevcuttur:

- Hastaneler, eczaneler, ilaç şirketleri, tıp merkezleri vb. kuruluşlardan gelen verilerin farklı formatlarda kaydı tutulmaktadır. Bu kuruluşların verileri kendilerine özel sistem ve ortamlarda bulunmaktadır. Bu kuruluşlar tarafından üretilen büyük miktardaki verileri kullanabilmek ve bu verileri yönetebilmek için, ortak bir veri ambarına ihtiyaç duyulmaktadır. Bu tür sistemler de büyük maliyetler gerektirmektedir.

- Toplanan veriler uygunsuz, standartlaştırılmamış ve yapılandırılmamış olabilmektedir. Sektörün bu tarz verileri anlamlandırabilmesi için ciddi bir çaba sarf etmesi gerekmektedir.
- Veri bilimi alanındaki şirketlerin; sağlık kuruluşlarını BV analitiğini kullanmaları için ikna edebilmeleri gerekmektedir.
- Toplanan verilerin kalitesi ile ilgili ciddi sorunlar nedeni ile, analiz sonuçlarında hatalar meydana gelebilmektedir (Asri ve ark., 2015).

Sağlık Sistemlerinde Büyük Veri Analitiği

BV analitiğini kullanan sağlık sektöründeki kurum ve kuruluşlar, hastalıkların erken teşhisi, tedavilerin zamanında yapılması, klinik operasyonlar, hasta profili analiti vb. amaçlar için BV'den faydalanmaktadırlar. Sağlık sistemlerinde BV analitiği; bilimsel kanıta dayalı karar almak için, sağlık verilerinden anlamlı içgörüler elde etmek üzere verilerin işlenip, analiz etmek için dönüştürülmesi işlemleridir. Karmaşıklık düzeyinin küçükten büyüğe doğru sıralanarak ifade edildiği üç analitik düzey mevcuttur:

1. **Tanımlayıcı Analitik (Descriptive Analytics):** Tablo ve grafik gibi istatistiksel araçlar yardımı ile geçmiş olayları özetleyen ve güncel olayları açıklayan raporların oluşturulmasıdır. Bu sayede hekimlere, hasta hakkındaki operasyonel veriler yardımı ile hastanın geçmişteki davranış kalıplarını inceleme ve anlaması sürecinde yardımcı olunur. Bu veriler mevcuttaki sorunları çözmek için kullanılabilir.
2. **Öngörücü Analitik (Predictive Analytics**): ML, veri madenciliği gibi yöntemlerde tanımlayıcı verilerin kullanılarak, geleceğin tahmin edilmesi ve analizlerin yapılmasıdır.
3. **Önerici Analitik (Prescriptive Analytics):** Sorunlara yönelik çeşitli olası alternatif çözüm arasından en uygun çözümün seçilmesidir. Gerçek zamanlı verilerin, veri madenciliği, ML, YZ yöntemleri kullanarak geçmiş veriler ile analiz edilmesidir. Uygulayıcılar, simülasyon ve optimizasyon tekniklerini kullanarak alınan kararların olası etkilerini belirleyebilir ve başarı ve başarısızlık senaryolarına karşı hazırlıklı olabilirler (Bansal ve ark., 2021).

Sağlık Sistemlerindeki Genel Tıbbi Veri Türleri

Hastaneler tarafından günlük olarak çeşitli tıbbi veriler üretilmektedir. Bu veriler, sağlık sektöründeki hastaların karmaşık hastalıklarının teşhisi ve hekimlere bilimsel araştırmalarında katkı sunması için gereklidir. Tıp bilimcilerinden bir grup, tıbbi verileri dört kategoride sınıflandırmıştır:

1. **Sayısal Veriler:** Hasta tarafından, ölçümler yapılarak oluşturulan ve sağlık sektöründe daha yaygın olarak kullanılan verilerdir. Örneğin; kan basıncı veya şeker seviyesinin ölçümleri, bir hastalığın şiddet seviyesinin 0 (hiç yok), 1 (düşük seviye), 2 (orta düzey), 3 (yüksek düzey) olarak tanımlanması, sistolik kan basıncının 70-120 mg arasında olması, diyastolik kan basıncının 90-80 mg arasında olması vb. gibi ayrık veya sürekli nitelikte olan verilerdir.
2. **Metinsel Veriler:** Alfabetik veriler içerir. Tıbbi verilerin çoğu, hastalıklarla ilgili net tanımlar içeren metinsel ifadelerdir. Bu tür veriler, hekimlerin hastalıkları teşhis edip, tedavi edebilmelerine katkı sunmaktadır. Örneğin; hastanın kalp rahatsızlığı olması muhtemel ise metinsel veriler, okuyucunun cümleleri değerlendirmesi için daha etkili olabilmektedir. ML yönteminde tıbbi metinsel veriler, doktorların reçeteli rapor sonuçları ile kullanılarak hastalığın tahmin edilmesi gibi amaçlar için kullanılmaktadır.
3. **Kategorik Veriler:** Kategori belirleyebilen metinsel veri türleridir. Tıbbi verilerin dağınık olduğunda bölünme ve kategorilere ayrılmasının etkili bir yöntemidir. Çoğunlukla, ML modellerinin kategorik tahminler yaptığı uygulamalarda kullanılır. Örneğin; hastalığın ciddiyet düzeyinin tahmin edilmesi vb. Kategorik veriler matematiksel ifadelerle temsil edilir. Örneğin; kan grubu 0Rh(+) olanlar; 1, A olanlar; 2 vb.
4. **Görüntüleme Verileri:** Hastalıkların teşhis edilmesi ve tedavi süreçlerinde oldukça önemli olan verilerdir. Görüntüleme verileri; hastalıkları otomatik olarak tanımak, ciddiyetini tahmin etmek için görüntü tanıma ve sınıflandırma modelleri geliştirmede etkili bir yöntem olan piksel tabanlı çok boyutlu verilerdir. Örneğin; X-ışın verileri, ultrason görüntü verileri, BT tarama verileri vb. Bu veriler, doktorların hasta bireyin enfekte bölgelerini tespit etmesine yardımcı olmaktadır (Agrawal ve ark., 2021).

Sağlık Sistemlerinde Büyük Veri Kaynakları

Sağlık hizmetlerinde büyük verinin beslendiği çok fazla sayıda kaynak vardır. Bu kaynaklar şu şekilde gruplandırılabilir:

- **Makinelerin Oluşturduğu Veriler:** Uzaktan algılayıcılar, akıllı sensörler ve sayaçlar, hayati bulgu sinyalleri sağlayan cihazlar vb. kaynakların ürettiği verilerdir.
- **Biyometrik Veriler:** Kişilerin parmak izleri, imzası, retina taramaları, kan basınçları, kalp çarpıntı hızları, nabızları gibi fiziksel özellikleri ile beraber röntgen ve diğer tıbbi görüntüleme cihazlarından elde edilen verilerdir.
- **İnsanların Oluşturduğu Veriler:** Sağlık sistemlerinde doğrudan insanların oluşturduğu verilerdir. Hasta durum belgeleri, laboratuvar sonuçları, hastane kabul

kayıtları, epikriz, taburcu özetleri, elektronik postalar vb. yarı-yapılandırılmış ve yapılandırılmamış verilerdir. Yapılandırılmış EHR'de bu grubun içerisinde yer almaktadır.

- **İşlem Verileri:** Hastaların sağlık talepleri, fatura kayıtlarından elde edilen veriler vb.bu gruba girebilir.
- **Davranış Verileri:** Sosyal medya platformlarından kullanıcıların konumları, sağlık durumları, duygusal ve sosyal etkileşimlerinin görüntülenmesi vb. gibi çok çeşitli veri tipinde üretilen verilerdir.
- **Epidemiyolojik Veriler:** Sağlık araştırmaları, hastalık kayıtları gibi konularda istatistiksel olarak tutulan verilerdir. Epidemiyolojik araştırma anlamında büyük veriler bir ülke içerisindeki veritabanları veya global olarak veritabanlarının birbirlerine entegre edilmesi ile oluşturulan büyük veri setleridir.
- **Güncel Hayata İlişkin Veriler:** Akıllı cihazlardan elde edilen, bireylerin günlük yaşamdaki faaliyetlerinin kaydının tutulduğu verilerdir. Örneğin; günlük atılan adım sayısı, harcanan kalori miktarı vb. Ayrıca; kilo değişimleri, beslenme alışkanlıkları vb.'de bu veri sınıfında yer almaktadır (Altındiş, ve Morkoç, 2018; Kurşun, 2021).

Sağlık Sistemlerinde Büyük Verinin Kullanım Alanları

Sağlık sistemlerinde oluşturulan BV ile diğer alanlarda oluşturulan BV karşılaştırıldığında, sağlık sistemlerinin kendine özgü özellikleri mevcuttur. Tıp alanında BV'ye ulaşmak zordur ve veriyi kullanmanın yasal olarak komplikasyonları vardır. Genel olarak sağlık hizmetlerinde BV; klinik uygulama ve araştırma, halk sağlığı informatiği, risk tahmini, biyoinformatik uygulamalar, giyilebilir cihazlar ve mobil sensörler, hastalıkların teşhis edilmesi ve tedavisi, bulaşıcı hastalıkların yayılması tahmini, tıbbi görüntüleme, ruh sağlığı değerlendirmesi gibi alanlarda kullanılmaktadır (Kurşun, 2021).

- **Klinik Uygulama ve Araştırma:** Teknolojinin gelişmesi ile geleneksel hasta tedavi yöntemleri yerini veri ile yönetilen süreçlerle desteklenen tedavi yöntemlerine bırakmıştır. Bilgisayar programları aracılığıyla üretilen karar destek sistemleri, klinik vakalarda karar vericilere destek sunmaktadır. BV; hastalık ve genetik verilerin toplanıp, analiz edilmesi ile hasta kişiye en uygun tedavi yöntemlerinin uygulanabilmesini sağlamaktadır. EHR klinik tıp için potansiyel bir BV kaynağıdır. BV, bilgiyi sürekli iyileştirme ve yapılabilecek yenilikler hususunda oldukça önemlidir. Ayrıca BV; sunduğu veri çeşitliliği sayesinde birbiri ile ilgisiz gibi görünen verileri bir araya toplayarak sağlık sistemlerinde bir içgörü oluşturabilme, değer yaratabilme potansiyeline sahiptir.

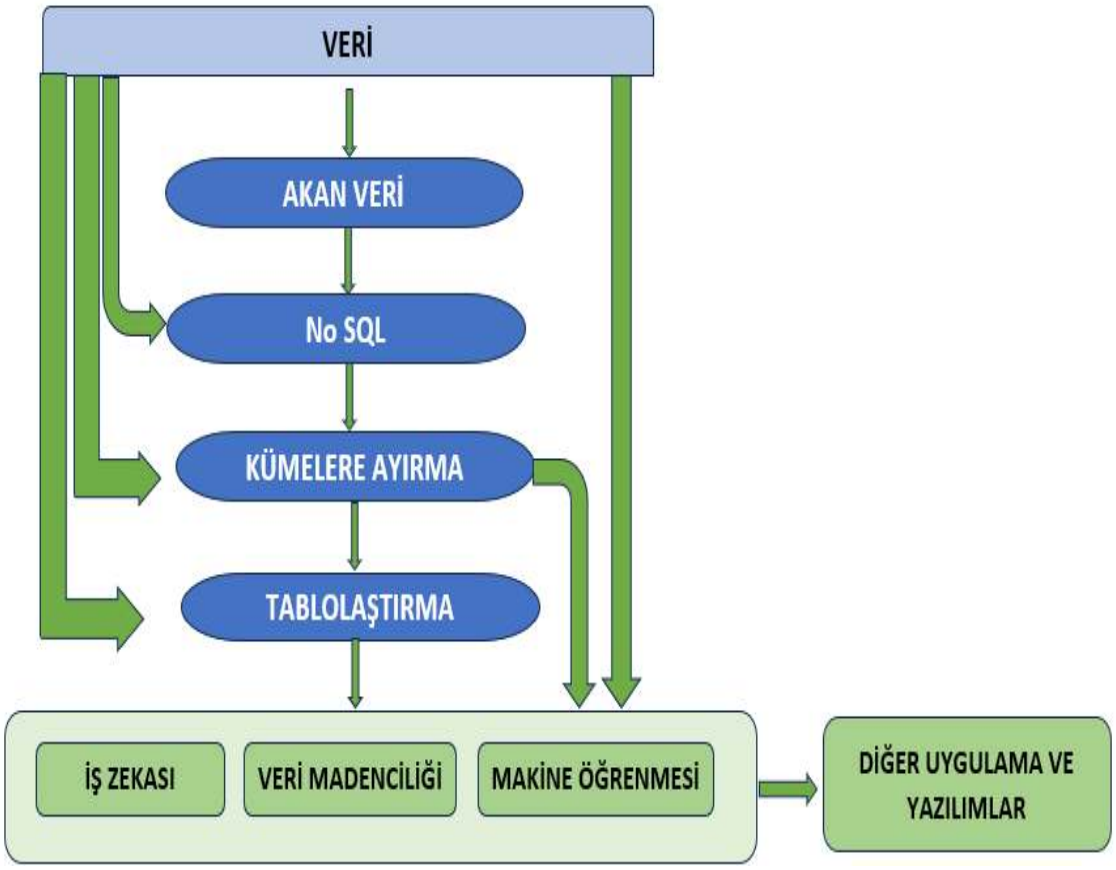
- **Halk Sağlığı İnformatiği:** EHR sistemlerinin kullanımı sağlık hizmetlerinde gün geçtikçe artmaktadır. Bu sayede her seviyede halk sağlığı hekimleri, araştırma yapmak, uygun tedavi ile hastaya müdahalede bulunmak, genel sağlığı iyileştirmek için oldukça fazla veri ve bilgiye sahip olabilmektedirler.
- **Sağlık Hizmetlerinde Risk Tahmini:** BV kavramının popüler hale gelmesinden sonra, sağlık hizmetlerinde risk tahminlerinde büyük veri araçlarının kullanılarak çeşitli analizler yapılmaya devam edilmektedir.
- **Biyoinformatik Uygulamalar Büyük** Veri, biyolojik bilgileri toplamada ve analiz etmede, araştırmacılara etkili veri işleme araçları ve veri havuzları sağlamaktadır. Hadoop ve MapReduce gibi büyük veri araçları yaygın olarak kullanılmaktadır.
- **Giyilebilir Cihazlar ve Mobil Sensörler:** Giyilebilir cihazlar, hastalar ile fiziksel etkileşimi sınırlandırarak veri toplamayı sağlayan teknolojilerdir. Bu teknoloji ile veriler uzaktan da değerlendirilerek, tüm hastane süreçlerinin iyileşmesine katkı sunulmuştur. Mobil sensörler ise "akıllı" çağının ürünleridir. Akıllı telefonlar, veri yakalamak için fazla sayıda gömülü sensörler içermektedir. Bu sensörler arasında; kameralar, ivme ölçerler, ışık sensörleri, yakınlık sensörleri, barometreler, termometreler, adım ölçerler vb. bulunmaktadır.
- **Hastalıkların Teşhisi ve Tedavisi:** Sağlık hizmetlerinde farklı hastanelerde farklı standartlar kullanılmakta ve tedavi görülen bir hastanedeki standartlar, farklı bir hastanenin sağlık standartları ile örtüşmeyebilmektedir. Bu nedenle farklı hastanelerden elde edilen tıbbi verilerin, bir standardizasyon işleminden geçmesi gerekmektedir. YZ, ML; hekimlere farklı hastalıkları teşhis etmesi, tedavi alternatifleri, tıbbi görüntüleri yorumlayabilmeleri için ipuçları sağlanması gibi konularda yardımcı olacaktır. Böylece sağlık hizmetlerinde standardizasyonun oluşması sürecine yeni teknolojiler katkı sunacaklardır.
- **Bulaşıcı Hastalıkların Tahmini:** Bulaşıcı hastalık; bir kişiden veya hayvandan başka bir kişiye, patojenlerinin enfekte olması ile ortaya çıkan klinik vakadır. Sosyal bir sorun olarak kabul edilmektedir. Salgın sırasında insanlar yoğun bir şekilde internette hastalık bilgisi aramaktadırlar. Bu toplu arama verileri, halk sağlığı araştırmalarında kullanılmaktadır. Bu veriler; hastalık dinamiklerini ölçmek, sağlık kurumu kaynaklarını verimli bir şekilde kullanabilmek, hastalığa hızlı bir şekilde müdahale etmek gibi faydalar sağlamaktadırlar.
- **Tıbbi Görüntüleme:** Görüntü veritabanı, görüntü raporlarını, patolojik, genetik ve diğer klinik verilerini içinde bulundurur. Genellikle görüntüler, görüntü içermeyen verilerden ayrılmaktadır. Görüntüleme bilgileri "meta veriler" olarak isimlendirilir.

Hastalıklı dokunun radyografik görüntüleri, BV kaynağı olabilmektedir. Bu veri kümelerinin işlenmesi ve anlamlı bilgiler çıkarılması gerekmektedir

- **Ruh Sağlığı Durumunun Değerlendirilmesi:** Sosyal medya platformları, kullanıcılardan elde ettiği büyük veriyi analiz ederek, kullanıcıların ruh hallerine yönelik çıkarımlarda bulunmakta, kullanıcıların ruh sağlığı hastalıklarına yakalanma riski faktörlerini belirlemeye çalışmaktadır (Kurşun, 2021).

1.2.7. Büyük Veri Analizi Araçları

Altyapısal olarak BV'nin saklanması ve işlenmesi için büyük donanımlara ihtiyaç duyulmaktadır. BV ile çalışırken, uygulamalarda performans sorunları ile de karşılaşılabilmektedir. Verilerin saklanmasında ve işlenmesinde dev sunucular, veritabanları ihtiyaçları karşılama görevi üstlense de bunların da yetersiz olduğu noktalar bulunmaktadır. Bu noktada ihtiyaçları gidermek için Hadoop, Hive, Impala, Spark gibi teknolojiler geliştirilmiştir.



Şekil 1.7. Veri analizi için genel teknoloji adımları

Şekil 1.7'de ifade edildiği üzere; eğer veriler standart kapasiteli bir bilgisayar yardımı ile işlenip, analiz edilebilecek boyutta ve bir tablo haline getirilmiş ise doğrudan iş zekası, veri madenciliği, ML gibi yazılım veya algoritmalar ile işlenebilir.

Veriler çok büyük boyutlarda değil fakat farklı yapı, tablo veya ortamda ise verilerin önce tablolaştırılmalı ve standart hale getirilip, analize hazır duruma getirilmelidir.

Veriler görece olarak büyük boyutta ve analiz edilmesi zaman alıyor ise veriler önce kümelere ve parçalara ayrılıp daha sonra standart hale getirilmelidir. Bu amaçla geliştirilen teknolojilerden bazıları MapReduce, APACHE SPARK, H2O, KAFKA, APACHE FLUME teknolojileridir.

Verilerin daha hızlı sorgulanması gerekiyorsa, klasik ilişkisel veritabanı mimarisi yerine NoSQL veritabanı mimarisi kullanılıp, verilerin JSON formatına dönüştürülmesi gerekebilir. Veritabanı NoSQL bir yapıya taşınırken; REDIS, MongoDB, Casandra gibi veritabanlarından biri verilerin analiz amacına göre tercih edilmelidir.

Veriler tek bir donanımın işleyemeyeceği kadar büyük ve sürekli akan bir veri ise, bu verileri birden fazla makineye dağıtıp işleyecek bir sistem kullanılmalıdır. Bunun için geliştirilen ve yaygın olarak kullanılan teknolojiler HADOOP veya APACHE STORM'dur. Bu teknolojiler Dağıtık Dosya Sistemi (Hadoop Distributed File System, HDFS) ile entegre biçimde çalışarak büyük boyutlu verilerin dağıtık ortamda depolanmasını ve işlenmesini sağlamaktadır. BV' nin işlenmesi için yüksek maliyetli olan süper bilgisayarların kullanımı yerine standart bilgisayarlardan oluşan bir ağ üzerinde bu sistemlerin kullanabilmesi prensibine sahiptirler (Silahtaroğlu G., 2020).

1.2.8. Makine Öğrenmesi

Bu kısımda genel hatları ile makine öğrenmesi kavramından bahsedilecektir. Makine öğrenmesi algoritmaları gibi teknik konulardan, bu tezin uygulama kısmında faydalanıldığı için, Materyal ve Metot kısmında da ayrıca bahsedilecektir.

Makine öğrenmesi, günümüzde neredeyse hayatın her alanında uygulama olanağı olan bir teknolojidir. Teknoloji ile ilgilenenlere teknoloji kitabı önerilmesi, küçük çocuklara oyuncak önerilmesi, borsa yatırımcılarına destek vermesi için piyasa tahmini yapması, bir onkoloğun hastasının tümörünün iyi huylu mu yoksa kötü huylu mu olduğunu anlamasına yardımcı olması, enerji tüketiminin optimize edilmesi vb. birçok alanda uygulama alanı bulan, hayatın hangi yönünden bakılırsa bakılsın, bir yaşam biçimi haline gelen teknolojidir (Dutt ve ark, 2018).

1999 yılında IBM'in Derin Mavi (Deep Blue) adını verdiği program, satranç oyununda dünya satranç şampiyonu olan Garry Kasparov'u yenerek bilim dünyasında bir dönüm noktasını oluşturmuştur. Bu başarı süper bilgisayarların ve yapay zekanın, insan düşüncesini simüle edebileceği bir geleceğin habercisi olmuştur (https://www.ibm.com/history/deep-blue).

Öğrenme kavramını iki gruba ayırabiliriz. Bunlar; insan öğrenmesi ve makine öğrenmesidir.

Günlük yaşamımızda sokakta yürümek, okul ödevini yapmak, şarkı söylemek gibi basit bir görev olan aktiviteleri gerçekleştirebilmenin yanında, bir roketin belirli bir yörüngeye sahip olabilmesi için hangi açıyla fırlatılması gerektiğine karar vermek gibi karmaşık bir görev olan

aktiviteleri de gerçekleştirmek gerekebilir. Bir görevi uygun bir şekilde icra etmek için göreve dair daha öncesinden az da olsa bilgi sahibi olunması gerekir. Daha fazla bilgi edinildikçe diğer bir deyişle daha fazla öğrendikçe görevler daha verimli biçimde yapılabilmektedir. Bu durumlar insan öğrenmesine örnek gösterilebilir.

Makine öğrenmesi kavramı, bilgisayar sistemi içeren bir makinenin belirli bir görevi yapıp deneyim kazanarak, gelecekte benzer görevleri yaparken performansını artırabildiği öğrenme durumudur. Geçmiş deneyimler ile anlatılmak istenilen belirli bir göreve ait geçmiş verilerdir. Bu veriler makineye bir kaynaktan gelen girdilerdir (Dutt ve ark, 2018).

Makine öğrenmesi yöntemi ile kesin karar verme yaklaşımı kullanışlı bir çözüm yöntemi değildir. İnsanlarda olduğu gibi makine öğrenmesi yöntemlerinde de sezgisel veya yaklaşma yöntemi ile sorunlar çözülmelidir. Bu yöntemle de doğru karar vermeme riski vardır çünkü; yapılan bazı varsayımlar gerçekte doğru olmayabilir. Ancak makinelerde olduğu gibi insanlar da sezgiye ve içgüdüye dayanarak karar aldığında aynı hataları yapabilir (Dutt ve ark, 2018).

Makine öğrenmesi; büyük miktarda veriyi işleyip analiz edebilmesi, çok boyutlu veri setlerindeki karmaşık yapıları tespit edebilmesi, aynı girdilerle hep aynı çıktıyı vererek tutarlı olması, tekrara bağlı görevlerin otomasyonu, tahmin yeteneği, yeni verilerle sürekli eğitilerek performansının artması vb. nedenlerle tercih edilen bir teknolojidir. İnsan beynini taklit etse de; insan zekasının yerini almak yerine, insanı güçlendiren ve destekleyen bir araç olarak değerlendirilmesi gerekir.

2. ÖNCEKİ ÇALIŞMALAR

Bu bölümde giriş kısmında bahsi geçen konuları da kapsayan, daha önce yapılmış ve literatüre katkı sunmuş çalışmalardan bahsedilecektir.

Makine öğrenmesi, YZ'nin bir alt dalı olup, bilgisayar sistemlerinin açıkça programlanmadan verilerden öğrenmesini sağlayan algoritma ve teknikler bütünüdür.

Makine öğrenmesi çok çeşitli alanlarda uygulanma fırsatı olan tekniklerdir. Sağlık sistemleri de bu alanlardan biridir. Makine öğrenmesi metotları uygulanarak sağlıkla ilgili çeşitli konularda çalışmalar yapılmıştır. Hastalık teşhisi, kronik rahatsızlıkların tedavi süreçlerinin iyileştirilmesi, hasta takibinin sağlanması gibi birçok alanda makine öğrenmesi tekniklerine başvurulmuştur.

Mujumdar ve Vaidehi (2019) çalışmasında diyabetin daha iyi sınıflandırılması için glikoz, vücut kitle indeksi, yaş, insülin vb. gibi düzenli faktörlerin yanı sıra diyabetten sorumlu birkaç dış faktörü de içeren bir diyabet tahmini sağlayan makine öğrenmesi modeli geliştirmişlerdir. Ahmed ve ark. (2020) 2016 yılından kalp hastalarına ait demografik verileri ve kan şekeri değeri, kolesterol değeri gibi tıbbi verileri de toplayıp; bir kişinin kalp hastalığı belirtisi içerip içermediğini makine öğrenmesi algoritmaları ile tahmin etmiştir Alves ve ark. (2021). Rastgele Orman (RO) sınıflandırıcısı ile COVID-19 vakalarını tahmin etmeye çalışmıştır. Karar Ağacı Açıklayıcısı ve Kriter Grafiği kullanılarak klinisyenlerin kan parametreleri arasındaki bağlantıyı genel olarak veya vaka bazında anlamalarına yardımcı olabilecek bireyler arasında bazı örüntüleri bulmaları mümkün hale gelmiştir.

Makine öğrenmesi algoritmaları, etkili bir şekilde öğrenmek ve genellenebilir modeller oluşturmak için büyük ve çeşitli veri kümelerine ihtiyaç duyar. Büyük verinin sağladığı hacim, ML modellerinin karmaşık örüntüleri daha doğru bir şekilde tanımasını mümkün kılmaktadır.

Büyük veri, geleneksel veri işleme yöntemlerinin başa çıkmakta yetersiz kaldığı, hacim, hız ve çeşitlilik gibi özelliklerle karakterize edilen devasa ve karmaşık veri kümeleridir. İlk olarak Douglas Laney (2001) tarafından tanımlanan 3V (Volume, Velocity, Variety) kavramına ek olarak, literatürde daha sonra Doğruluk (Veracity) ve Değer (Value) boyutları da eklenerek 5V modeli yaygınlaşmıştır.

Literatür, büyük veri ve makine öğrenmesi arasında karşılıklı yarar sağlanan bir ilişki olduğunu vurgulamaktadır:

Dal (2021) çalışmasında toplumda BVA ile entegre edilmiş mobil-sağlık (m-sağlık) uygulamasının kullanılmasını tüketicilerin benimseme davranışını incelemiştir.

Yanar (2024) çalışmasında karaciğer kanseri nedeni ile karaciğer nakli olan hastaların, kanserinin nüksetme durumunu tahmin edebilen makine öğrenmesi modellerini geliştirmiştir. Hastalara ait; yaş, cinsiyet, tümör çapı, invazyon, tümör volümü gibi değişkenleri veri setinde kullanmıştır.

Bulut (2023) çalışmasında BV bağlamında işlenen tüm sağlık bilgilerinin kişisel verilerle ilişkili olduğunu, BV'nin sağlık alanında kullanılması sürecinde daha fazla korunması gerektiğini ifade etmiştir. Sağlık sistemlerinde kullanılan veritabanlarındaki etik sorunları tanımlamaya ve bu sorunları meslek ahlakı yükümlülükleri ve insan hakları temelinde çözebilmek için öneriler geliştirmiştir.

Mumcu (2024) tez çalışmasında sağlık sektöründe büyük verinin konumu, kullanımı ve etkisini kapsamlı analiz etmek amacıyla literatürdeki uluslararası yayınların oluşturduğu veri yığınının bibliyometrik analizini yapmıştır.

Sağlık sistemlerinde talep tahmini amacıyla karma dağılım modellerinin kullanıldığı çeşitli çalışmalar literatürde yer almaktadır. Son dönemlerde araştırmalar, sağlık yönetiminde karar süreçlerini desteklemek için geliştirilen karma dağılım tabanlı yaklaşımların umut verici sonuçlar sunduğunu göstermektedir. Sağlık hizmetlerinden yararlanan bireylerin yaş, cinsiyet ve sosyo-ekonomik durum açısından farklılık göstermesi, bu tür modellerin heterojen veriyi temsil etmede güçlü bir yöntem olmasını sağlamaktadır. Karma dağılım modelleri, özellikle farklı çalışma sürelerine sahip personelin görev sürelerini modellemede etkili bir şekilde kullanılmıştır (Millard, 1988). Ayrıca, hastane kalış süresinin öngörülmesinde veriyi yatan hasta ve ayakta tedavi gören hasta gruplarına ayırmak, tahminlerin doğruluğunu artırmaktadır (Xiao ve ark., 1999).

Literatürde, sisteme gelişler arası süre ve sistemde kalış sürelerinin normal, negatif binom, üstel, Weibull, Poisson ve lognormal karma dağılımlarının kullanıldığı çalışmalara rastlanmıştır.

Güleryüz (2023) tez çalışmasında Çukurova Üniversitesi Balcalı Eğitim ve Araştırma Hastanesinin yoğun bakım ünitesi yatak kapasitesi problemini ele almıştır. Yoğun bakım ünitesindeki tüm hastalara bekleme süresi olmaksızın hizmet vermek için gereken yatak kapasitesinin yönetilmesi için girdi parametreleri olarak karma dağılımının kullanıldığı simülasyon tabanlı bir model oluşturmuştur.

Hastanede kalış süresini modellemek amacıyla gerçekleştirilen bir araştırmada, iki bileşenden oluşan Poisson karma modeli uygulanmış ve modelin istatistiksel açıdan güçlü bir uyum sağladığı rapor edilmiştir (Wang ve ark., 2002).

Amerika'da gerçekleştirilen bir araştırmada, hastaların hastaneye yatışlarından taburculuklarına kadar geçen sürenin, iki bileşenli üstel karma dağılım kullanılarak modellendiği belirtilmiştir (Cleary, 1991).

Bir hastanede gerçekleştirilen araştırmada, lösemi ve lenfoma hastaları arasında yüksek maliyet oluşturan ve uzun süre hastanede kalan hasta gruplarını belirlemek için Weibull karma dağılımından yararlanılmıştır (Quantin ve ark., 1999).

Başka bir çalışmada hastanede kalış süresi değişkeni, tanıya göre oluşturulan hasta grupları içerisinde sonlu karma dağılımlar aracılığıyla modellenmiştir. Bu çalışmada tek bir dağılım temelli karma model yerine, Gamma, Weibull ve lognormal gibi asimetrik dağılımların birlikte kullanıldığı bir karma model önerilmiş ve bu yaklaşımın geçerliliği ortaya konmuştur (Atienza ve ark., 2008)

3. MATERYAL VE METOT

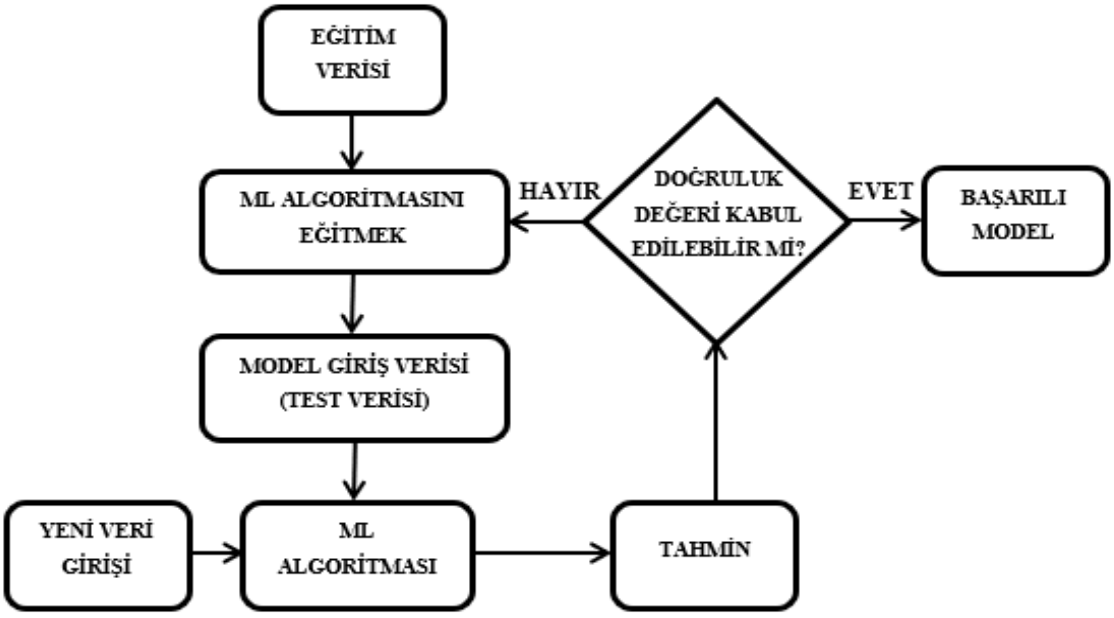
Bu bölümde tez çalışmasının uygulama aşamasında kullanılan verinin özelliklerinden ve bu veriyi analiz etmek için kullanılan yöntemlerden bahsedilecektir.

3.1. Makine Öğrenmesi (Machine Learning-ML)

Öğrenme genel olarak bir çalışma veya deneyim sırasında belirli bir beceri veya bilgi edinme sürecidir. Bir makinenin bu temel öğrenme eylemini yeniden icra edebilmesine makine öğrenimi denir (Saxena ve Chandra , 2021).

Makine öğrenmesi, sistemlerde ilgili verileri gözlemleyerek, bir görevdeki performansın otomatik olarak iyileşmesini sağlayan bir tekniktir. Bilgisayarlara açıkça programlama yapılmadan öğrenme yeteneği kazandırılmasıdır. Deterministik kurallar yerine, olasılık ve istatistikle yönlendirilen matematiksel modeller kullanılarak belirsizlikle mücadele etmek için kullanılır. Yapay zekanın alt dalıdır (Natarajan ve ark., 2017; Saxena ve Chandra, 2021).

Makine öğrenmesi, insanların karar vermek durumunda olduğu vakalarda geleceği tahmin etmek ya da, bilgiyi sınıflandırmak için karar destek sistemleri olarak yaygın biçimde kullanılan tekniklerdir. ML değişkenler arasındaki ilişkiyi açıklayan algoritmalara dayanan tekniklerdir. ML algoritmaları, geçmiş verilerden öğrendikleri örnekler üzerinde eğitilir ve geçmiş verileri değerlendirebilir. Sürekli eğitimden sonra, bir algoritma tahmin yapmak için verideki desenleri tanır. Bir öğrenme algoritması kendi kendine öğrenmek ve zamanla daha iyi performans sergilemek için verileri ve deneyimi kullanır. Süreç boyunca öğrenen algoritma daha iyi tahminler üretebilmek için sürekli kendini optimize eder (Natarajan ve ark., 2017; Bansal ve ark., 2021).



Şekil 3.1. Makine Öğrenmesi Süreci

Şekil 3.1; makine öğrenmesi algoritmalarının genel iş akışını ifade etmektedir. Makine öğrenmesindeki süreçler 4 grupta incelenebilir.

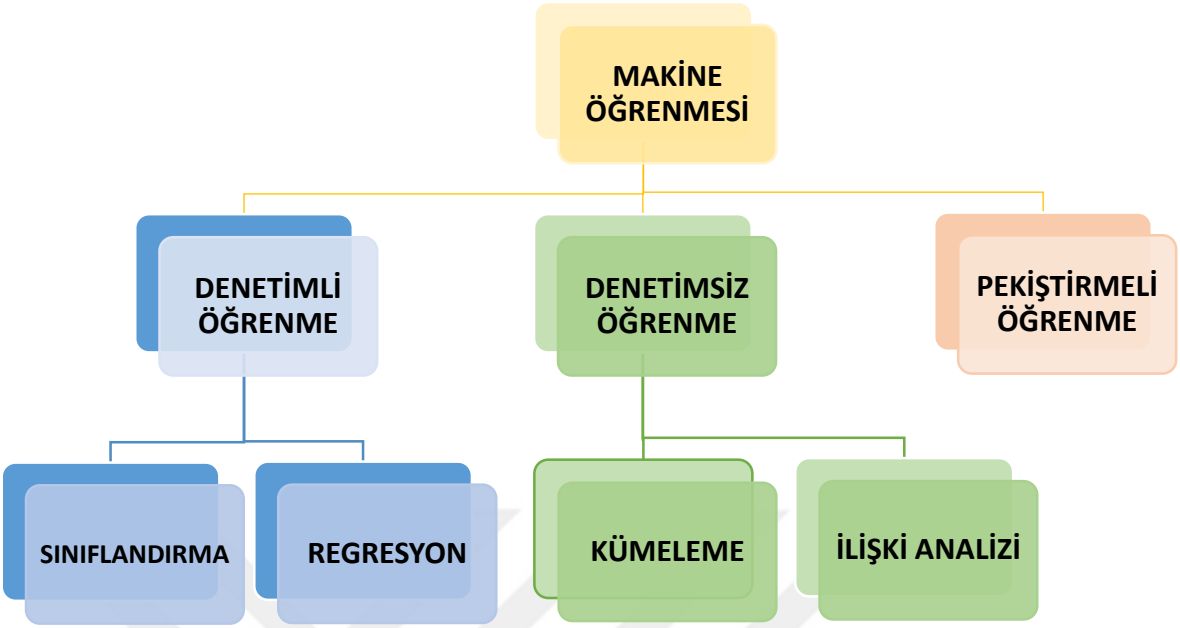
1. Veri toplama ve veri ön işleme
2. Veri kümelerinin oluşturulması
3. Eğitim
4. ML değerlendirmesi

Başlangıçta veriler farklı kaynaklardan elde edilir ve tek bir veri kümesinde birleştirilir. Ardından veriler, kullanılabilir bir veri kümesine dönüştürülmek amacıyla veri ön işlemeye tabii tutulur. İşlenen veriler; eğitim verisi kümesi, doğrulama verisi kümesi ve test verisi kümesi olarak ayrılırlar. Daha sonra eğitim kümesindeki veri, sınıflandırmada kullanılan özellikleri öğrenmesi için ML algoritmalarına verilir. Eğitim tamamlandıktan sonra model, kabul edilebilir düzeyde bireysel parametrenin olduğu doğrulama veri kümesi kullanılarak iyileştirilebilir. Parametreler belirlendikten sonra, test veri kümesi kullanılarak test yapılır. Sonuçlar kabul edilebilir ise ML algoritması tahmin yapmak için kullanılabilir. Aksi halde doğruluk değerini artırmak için eğitime devam edilmelidir (Bansal ve ark., 2021).

Makine öğrenmesi, sık sık insan müdahalesi gerektiren görevlerde ya da insanların çok etkili olduğu görevlere uygulanmamalıdır. Örneğin; hava trafik kontrolü, yoğun insan katılımı gerektiren çok karmaşık bir görev olması nedeni ile buradaki görevler makine öğrenmesi algoritmalarına devredilmemelidir. Aynı zamanda geleneksel programlama paradigmaları kullanılarak uygulanabilen çok basit görevler için de makine öğrenmesi kullanımının gereği yoktur. Örneğin; fiyat hesaplama motoru gibi formül tabanlı veya basit kural odaklı uygulamalar için makine öğrenmesi teknikleri kullanılmasına gerek yoktur. Makine öğrenmesi iş sürecinde aksaklıklar meydana geldiğinde kullanılmalıdır. Görevler optimize edilmişse makine öğrenmesi kullanımının gereği yoktur. Ayrıca eğitim verilerinin yeterli olmadığı durumlarda makine öğrenmesi etkili bir şekilde kullanılamaz. Çünkü; küçük eğitim verilerinde, kötü verinin sonuca etkisi kat ve kat daha kötü olmaktadır. Tahmin veya önerinin iyi kalitede olabilmesi için eğitim verilerinin de önemli miktarda ve kaliteli olması gerekmektedir (Dutt ve ark., 2018).

Makine öğrenmesi 3 geniş kategoriye ayrılabilir:

- Denetimli (Gözetimli) Öğrenme
- Denetimsiz (Gözetimsiz) Öğrenme
- Pekiştirmeli (Takviyeli) Öğrenme



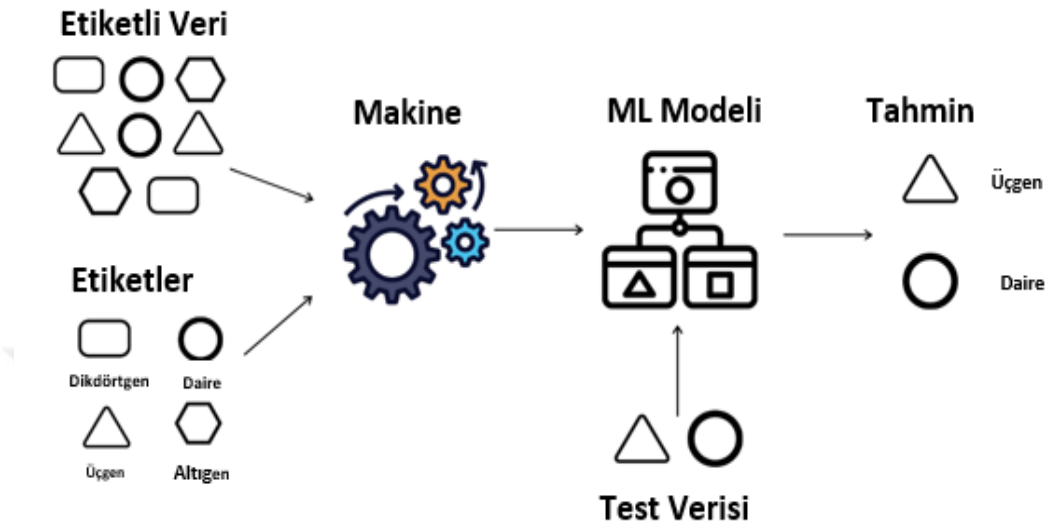
Şekil 3.2. Makine Öğrenmesi Tipleri

3.1.1. Denetimli (Gözetimli) Öğrenme

Hem girdi verilerinin hem de çıktı verilerinin sistemde yer aldığı makine öğrenmesi türüdür. Denetimli öğrenme türünde algoritma, etiketli verileri makinenin öğrenebileceği ve girdi değerler ile çıktı değerler arasında örüntüler geliştirebileceği şeklide eğiterek çalışır. Bu sayede gözetimli öğrenme algoritması, veri noktalarını kategorize edebilecek yöntemi veya sınıflandırabilecek örüntüyü bulur. Etiketlenmiş veri; veri örneklerine atanan, bilinen tanımları ifade eder. Örneğin; bir hastalık için farklı semptomlar gösteren 20 kişiye yapılan test sonucuna göre, her hasta pozitif ya da negatif değere sahip etiketlendirme ile temsil edilebilir. Böylece, etiketlenen veriler çıktı değerleri için bir şekil sağlar. Dolayısıyla denetimli öğrenme süreci ile makinenin örüntüyü öğrenip verileri sınıflandırması beklenir (Saxena ve Chandra, 2021).

Denetimli öğrenme temel olarak geçmiş bilgilerden öğrenir. Makinenin yürütmesi gereken görev hakkında bilgi sahibi olması gerekmektedir. Makine öğreniminin tanımı kapsamında bu bilgiler, deneyimdir. Örneğin; bir makine farklı nesnelerin görüntülerini girdi olarak alsın ve bu makinenin görevi, görüntüleri nesnenin şekline veya rengine göre ayırmak olsun. Eğer şekle göre ayıracaksa, yuvarlak şekilli nesnelerin görüntülerinin, üçgen şekilli nesnelerin görüntülerinden vb. ayrılması gerekir. Eğer ayrım renge göre yapılacaksa, mavi nesnelerin görüntülerinin, yeşil nesnelerin görüntülerinden vb. ayrılması gerekir. Bir makinenin, nesnenin yuvarlak mı yoksa üçgen mi olduğunu ya da nesnenin renginin mavi mi yoksa yeşil mi olduğunu ayırt edebilmesi için, temel bilgilerin makineye sağlanması, verilmesi yani öğretilmesi gerekir. Bu temel girdi ya da makine

öğrenmesi paradigmasındaki deneyim, eğitim verileri şeklinde verilir. Eğitim verileri, belirli bir görevle ilgili geçmiş verilerdir. Görüntü ayırma sorunu ile ilişkili olarak, eğitim verileri, bir dizi görüntünün farklı yönleri ve özelliklerine dair geçmiş verilerdir. Bu veriler görüntünün yuvarlak ya da üçgen vb. olduğu, mavi renk ya da yeşil renk vb. olduğuna dair etiketler içermektedir.



Şekil 3.3. Denetimli Öğreneme

Şekil 3.3' de ifade edildiği üzere; gözetimli öğrenmede, geçmiş bilgileri içeren etiketli eğitim verileri girdi olarak gelir. Eğitim verilerine dayanarak makine, test verilerindeki her kayıt için bir etiket atayabilecek tahmin modeli oluşturur.

Gözetimli öğrenmeye bazı örnekler şu şekildedir:

- Bir tümörün iyi huylu mu yoksa kötü huylu mu olduğunu tahmin etmek
- Emlak, hisse senedi fiyatlarını tahmin etmek
- E-postaları spam veya spam olmayan olarak sınıflandırma
- Bir oyunun sonucunu tahmin etmek
- Bankacılık işlemlerinde dolandırıcılık tespiti
- Doğal afet tahmini
- El yazısı tanıma
- Hava durumu tahmini
- Perakende sektöründe talep tahmini

Gözetimli öğrenme 2 biçime ayrılabilir:

1. **Sınıflandırma:** Önceden tanımlanmış sınıflardan gelen girdi verilerini sınıflandırır. Algoritmalar, etiketli verilerden kategorik çıktıyı tahmin etmeyi sağlar. Başka bir

ifade ile, eğitim verileri tarafından sağlanan bilgileri referans alarak, test verileri için kategorik tipte bir sınıfın tahmin edildiği denetimli öğrenmedir.

2. **Regresyon:** Değişkenler/özellikler arasındaki ilişkiyi bulur. Algoritmalar, girdi verildiğinde özellikler arasındaki ilişkiyi bularak çıktıyı tahmin etmeyi sağlar.

Yukarıdaki örneklerden, bir tümörün iyi huylu mu yoksa kötü huylu mu olduğunu tahmin etmek ve emlak, hisse senedi vb. fiyatlarını tahmin etmek incelendiğinde, iki problemin özünde aynı olmadığı görülür. Her ikisi de tahmin problemi olsa da ilk durumda bilinmeyen bir verinin hangi kategoriye veya sınıfa ait olduğu tahmin edilmeye çalışılırken; ikinci durumda bir sınıf değil, mutlak bir değer tahmin edilmeye çalışılmaktadır. Kategorik veya nominal bir değişkenin tahmin edilmeye çalışılması problemine sınıflandırma problemi, sayısal değerli bir değişkenin tahmin edilmeye çalışılması regresyon problemi biçimine girer.

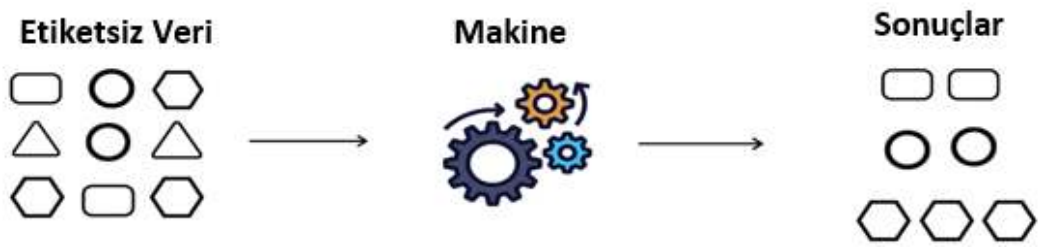
Denetimli makine öğrenimi, kendisini eğitmek için kullanılan veriler kadar iyidir. Eğitim verileri kalitesiz ise, tahmin sonucu da kesin olmaktan uzaklaşacaktır (Saxena ve Chandra, 2021; Dutt ve ark., 2018).

3.1.2. Denetimsiz (Gözetimsiz) Öğrenme

Makineye sadece giriş değerleri/verileri sağlanır. Sistemde sabit bir çıktı sağlanmaz. Bu nedenle gözetimsiz öğrenme algoritmaları etiketli verilere sahip değildir. Denetimsiz öğrenme algoritmaları girdi verilerindeki gizli örüntüyü bularak çıktı değerlerini tahmin eder. Denetimsiz öğrenme sürecinde makinelere etiketsiz bir veri girdisi sağlanır. Bu veriler, denetimsiz öğrenme algoritması tarafından veriler içerisinde bir örüntü varsaymak için kullanılır. Bu örüntü kullanılarak veri noktası örnekleri gruplandırılır. Benzer örüntü ve grupla eşleşen veriler çıktı değeri olarak tahmin edilir (Saxena ve Chandra, 2021).

Başka bir tanımlamaya göre denetimsiz öğrenme kavramı şu şekilde ifade edilebilir.

Denetimli öğrenmeden farklı olarak, denetimsiz öğrenmede öğrenme işleminin yapılabileceği etiketli bir eğitim verisi ve yapılacak tahmin yoktur. Denetimsiz öğrenmede amaç; bir veri kümesini girdi olarak alıp, veri öğeleri içerisinde veya veri kayıtları içerisinde doğal gruplamalar ya da kalıplar bulmaktır. Yani; veri kümesinde çıktının önceden tanımlanmadığı gizli örüntüleri keşfetmektir. Bundan dolayı; denetimsiz öğrenme genellikle tanımlayıcı model olarak isimlendirilir ve denetimsiz öğrenme süreci de bilgi keşfi veya kalıp keşfi olarak isimlendirilir (Dutt ve ark., 2018; Saxena, ve Chandra 2021). Şekil 4.4 ile anlatılanlar tasvir edilmeye çalışılmıştır.



Şekil 3.4. Denetimsiz Öğrenme

Denetimsiz öğrenme iki çeşittir:

1. **Kümeleme:** Girdi verilerinden kümeler oluşturmayı amaçlayan yöntemdir. Benzerlik gösteren veri noktaları bir araya gelerek kümeler oluşturur ve bu kümeler üzerinden tahminleme ve veri analizi yapılabilir. Başka bir deyişle, denetimsiz öğrenmenin temel türlerinden biri olan kümeleme, benzer nesneleri gruplandırmayı veya organize etmeyi hedefler. Bu nedenle; aynı kümeye ait nesneler birbirlerine belirli özellikleri açısından oldukça benzerken, farklı kümelerdeki nesneler birbirlerinden belirgin şekilde farklıdır. Şekil 3.4' de görsel olarak gösterildiği üzere, kümelemenin amacı etiketlenmemiş verilerden kümeler oluşturmaktır. Kümeleme işlemi için farklı benzerlik ölçütleri kullanılabilir. En yaygın ölçütlerden biri mesafedir. İki veri öğesi aralarındaki mesafe az ise aynı kümenin parçası kabul edilir. Aksine, veri öğeleri arasındaki mesafe yüksekse, genellikle aynı kümede yer almazlar. Bu yaklaşım mesafeye dayalı kümeleme olarak adlandırılır. Ayrıca; her bir kümenin merkez noktasını bulmaya dayalı teknikler ve veri dağılıma dayanan istatistiksel yöntemler de kümeleme analizinde kullanılmaktadır. Veri dağılımına dayanan istatistiksel yöntemlere karma dağılımlar örnek verilebilir. Karma dağılımlar, iki ya da daha fazla dağılımın kombinasyonuyla oluşan ve çarpık veri modellenmesinde kullanılan bir istatistiksel dağılım türüdür. Bu yaklaşımla aynı grup içerisindeki farklı kümeler başarılı bir şekilde temsil edilebilir. Karma dağılım modellemesi ile bir popülasyonun grup yapısı tanımlanabilmekte ve kümeleme analizinde kullanılan güçlü bir olasılık yaklaşımı sağlanmaktadır. Kümeleme analizinde varlıkların birkaç kümeye gruplandırılması karma dağılım aracılığıyla elde edilmektedir. Böylece küme içi benzerliğin en fazla, kümeler arası benzerliğin en az olduğu veri grupları elde edilmeye çalışılır. Karma dağılım temelli kümelenme yaklaşımları, gözlenen özellik verileri ve her bir birimle ilişkili değişkenler temelinde, temel popülasyonun grup yapısını önceden herhangi bir bilgiye ihtiyaç duymadan ortaya çıkarmak amacıyla uygulanır. Bu yöntem, denetimsiz sınıflandırma olarak adlandırılmaktadır. Karma dağılımlar ayrıca küme bileşen sayısının belirlenmesi ve sınıflandırma belirsizliğinin

değerlendirilmesinde de istatistiksel çıkarımlar sağlamaktadır. Gerçek veriler genellikle homojen olmayıp, heterojendir. Karma dağılım modelleri, farklı parametrelere sahip heterojen alt popülasyonların varlığını dikkate aldığından dolayı, veri temsilinde önemli rol oynamaktadır. (Güleryüz, 2023).

Sağlık sektöründe kümeleme, hastaları benzer komorbidite ya da risklere sahip anlamlı gruplara ayırmak için kullanılabilir. Karmaşık ya da nadir gelişen hastalıkların bakımında, farklı tedavi yöntemlerinin benzer durumlardaki diğer hastalar için işe yaraması durumunu, geçmiş hastanın mevcut hastaya öngörü değeri taşıyabilecek benzerlikte olup olmadığının analiz edilmesinde kullanılabilir. Mevcut nüfus sağlığı yönetimi yöntemlerinin uygulanabilmesi için geniş hasta gruplarının tespit edilmesinde kümeleme kullanılabilir (Natarajan ve ark., 2017; Uğuz, 2023).

2. **İlişki Analizi:** Algoritmaların, girdi verilerindeki kuralları bulup, verilerden tahminlerde bulunduğu ilişkidir. Örneğin; bir market sepetine geçmiş verilerden yola çıkarak A ürününü satın alan müşterilerin çoğunun B ve C ürünlerini de satın almasının ürün ilişkileri bağlamında incelenip aralarındaki bağlantının ortaya çıkarılmasıdır.

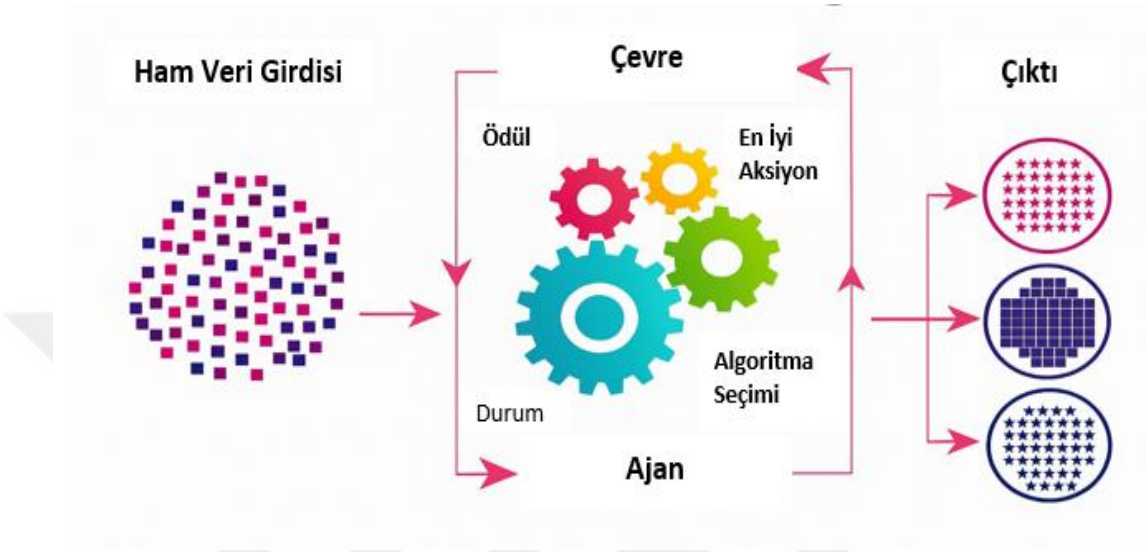
3.1.3. Pekiştirmeli (Takviyeli) Öğrenme

Bir YZ modelinin, verilerden öğrenmek yerine, bir çevre (environment) ile etkileşime girerek belirli bir görevi gerçekleştirmeyi öğrenmesini sağlayan bir ML yaklaşımıdır. Bu yöntem, verilen görevin başarımında doğru sonucu en üst düzeye çıkaracak optimal stratejinin bulunmasına odaklanır. Kararlar sıralı olarak verilir. Toplam sonuçlara giden yolda algoritma, attığı her adımda pozitif veya negatif bir çıktı (ödül/ceza) elde eder. Dolayısıyla nihai başarı, süreç boyunca karşılaşılan tüm negatif ve pozitif sonuçların toplamına bağlıdır. Algoritma sonucu en üst düzeye çıkararak, en iyi yolu bulmayı amaçlar.

Pekiştirmeli öğrenme süreci bebeklerin yürümeyi öğrenme süreci ile benzerlik gösterir. Bebekler başlangıçta yürümeyi bilmemelerine rağmen çevresindeki bireyleri gözlemleyerek adım atmak için bacaklarını sırayla kullanmaları gerektiğini kavrarlar. Yürümeye çalışırken bazen engele takılıp düşerler, bazen de engelli yollardan kaçınarak rahatça yürüyebilirler. Başarılı bir şekilde yürüyebildiklerinde ebeveynleri tarafından ödüllendirilirler. Başarısız olduklarında ödüllendirilmezler. Bebekler deneyimlerinden geribildirim alarak daha kolay yürüyebilir hale gelirler. Benzer şekilde; pekiştirmeli öğrenmede ajan (agent) belirli bir görevi yerine getirmeye çalışır. Bebek örneğinde gerçekleştirilmeye çalışılan eylem yürüme eylemidir. Bebek ajandır. Bebeğin yürümeye çalıştığı engelli yer çevredir. Bebek görevi yerine getirme performansını artırmaya çalışır. Bir alt görev başarı ile tamamlandığında ödül verilir. Bir alt görev doğru bir

şekilde yerine getirilmediğinde ise ödül verilmez. Bu pekiştirmeli öğrenme süreci makine tüm görevi yerine getirene kadar devam eder (Saxena, ve Chandra, 2021; Dutt ve ark., 2018).

Pekiştirmeli öğrenmenin bir başka güncel örneği ise sürücüsüz otomobillerdir. Farklı yol segmentlerindeki hız sınırı, trafik şartları, yol ve hava durumu gibi dikkate alınması gereken kritik bilgilerdir. Aracı başlatma/durdurma, aracın hızlanması/yavaşlaması, sağa/sola dönme gibi tamamlanması gereken görevlerdir.



Şekil 3.5. Pekiştirmeli Öğrenme

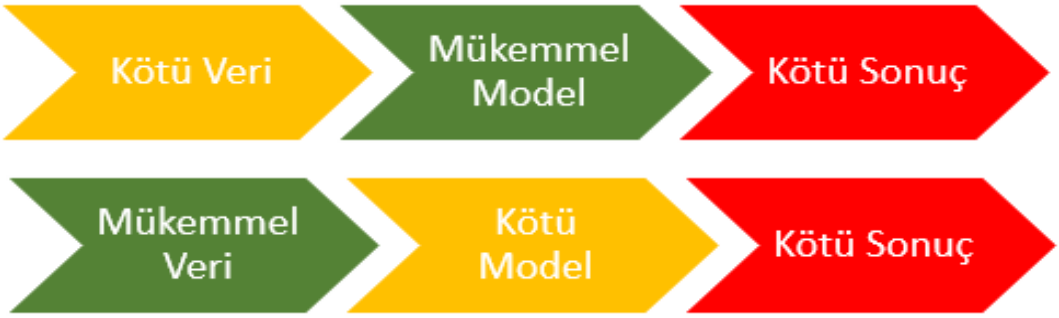
3.2. Veri Madenciliği ve Makine Öğrenmesi için Veri Ön İşleme

Veri setini hazırlama sürecidir. Makine öğrenmesinde modeller öğrenme işlemini verilerden gerçekleştirirler. Elde edilen herhangi bir veri düzenli değildir. Diğer bir ifade ile makine öğrenmesi modelleri için eğitilmeye hazır değillerdir. Bu veriler üzerinde veri ön işleme yapılması gerekir. Veri madenciliği açısından ele alındığında ise; veri madenciliğinin ortaya çıkmasının ve günümüzde oldukça popüler olmasının, üzerinde araştırmalar yapılmasının en önemli nedenlerinden biri, günümüzde büyük veritabanlarının erişilebilir olmasıdır. Farklı sektörlerden, farklı nedenlerle kaydı tutulan her türlü bilgi bilgisayar belleklerinde muhafaza edilmektedir. Ancak bu veritabanlarında tutulan bilgilerin tamamının gerçek ve doğru bilgiler olduğu %100 olarak garanti edilemez. Ayrıca; bu bilgilerin mevcut hali ile yapılacak olan çalışmaya hizmet edebileceği de kesin değildir. Bu bilgileri kullanabilmek, analizler yapabilmek, içgörüleri ortaya çıkarabilmek, öngörüler elde edebilmek için belirli işlemlerden geçirmek gerekir. Diğer bir ifade ile veriler üzerinde bir ön işleme çalışması yapılmalıdır. Çünkü gerçek hayatta üretilen veriler;

3. **Eksiktir**: Bazı kaynaklarda bazı özniteliklere ait veriler eksiktir ya da öznitelik başka verilerden elde edilen bilgileri içerdiği için anlamsal olarak kaynak verilere erişmek mümkün olmayabilir. Örneğin; doğum tarihi yerine yaş bilgisi, boy ve kilo yerine vücut kitle indeksinin kullanılması gibi.

4. **Gürültülüdür:** Özniteliğin özelliğine aykırı değerler ya da hatalı girilen değerler mevcuttur. Örneğin; bir kişinin doğum tarihinin 1283 olarak girilmesi gibi.
5. **Tutarsızdır:** Bütüncül bir bakış açısı ile ele alındığında, aynı varlığa ait farklı kayıtlar arasında ya da kayıt bazlı incelendiğinde, bir kaydın birkaç özniteliğindeki değerler arasında çelişkiler ve tutarsızlıklar vardır.
6. **Hatalıdır**: Verilerin bir kısmı gerçekten yanlış, anlamsız bilgiler içerebilir. Örneğin; ürün kodları hatalı girilmesi, bir bilginin birkaç farklı yere gereksiz bir şekilde girilerek aynı anlama gelebilecek birden fazla bilginin olması gibi.

Bu durumlar göz ardı edilip, gerekli veri ön işleme adımları yerine getirilmezse, kalitesiz bir veri kümesi üzerinde çalışılır ve kalitesiz sonuçlar elde edilir. Veri madenciliğinde sıkça dile getirilen "Çöp girerse, çöp çıkar" ifadesi, kalitesiz verinin sonuçları nasıl etkilediğini oldukça iyi açıklar.



Şekil 3.6. Girdi-Çıktı Kalitesi İlişkisi

Başarılı bir veri madenciliği ya da makine öğrenmesi çalışması için Şekil 3.6'da gösterildiği üzere hem veri (malzeme) hem de modelin (yöntem) kaliteli olması gerekmektedir (Köse. 2018; (Silahtaroğlu, 2020).

3.2.1. Veri Kalite Kriterleri

Eldeki verinin eksik, gürültülü, tutarsız, hatalı olabileceği kabul edildikten sonra, veri üzerinde temizlik yapılabilmesi için birtakım kalite kriterlerine ihtiyaç vardır:

- **Geçerlilik:** Verinin geçerli bir değer taşıdığının söylenebilmesi için veri şu koşullara uygun olmalıdır:
  - ✓ **Aralık:** Öznitelik için anlamlı olan bir aralık varsa veri bu aralıkta olmalıdır. Örneğin; yaş özniteliği için 320 değeri normal kabul edilemez.
  - ✓ **Tür:** Verinin, özniteliğin veri türüne uygun olması gerekir. Örneğin; tarih özniteliği GG.AA.YYYY formatında olabilir.

- ✓ **Zorunluluk:** Bazı özniteliklerin boş değer alması kabul edilebilir bir durum değildir. Örneğin; hastalık tanısı tahmini yapmak için geliştirilen bir modelde, eldeki veri setinde tanı özniteliğine sahip veri olmayan kayıtlar geçerli olamaz ve analize dahil edilmemelidirler.
- ✓ **Teklilik:** Bazı özniteliklerde yer alan değerlerin tüm veri kümesi içinde tekil olması gerekebilir. Örneğin; TC No vb.
- ✓ **Seçenek:** Bazı öznitelikler birden fazla değer içerebilir. Bu durumda ilgili özniteliğe ait veriler mutlaka tanımlanmış seçeneklerden biri olmalıdır. Örneğin; cinsiyet özniteliği için "erkek", "kadın", "belirtilmedi" tanımlı ise başka bir değer geçersizdir.
- ✓ **Harici anahtar:** Veri kaynağının ilişkisel bir veritabanı olması durumunda, öznitelik vektöründe başka tablolardan referans edilen öznitelikler yer alabilir. Örneğin; muayene veri setindeki "sevk eden hastane" gibi bir nitelik, veritabanında kayıtlı olan hastanelerden biri olmalıdır. Tanımsız bir hastane adı içeren kayıtlar geçersizdir.
- ✓ **Alanlar arası uygunluk:** Bazı durumlarda bir kayıtta yer alan niteliklerdeki değerler kendi aralarında tutarsız olabilir. Örneğin; muayene veri kümesinde yer alan "hastanın çıkış tarihi", "hastanın giriş tarihi" nden önce olamaz. Aksi halde kayıtlar geçersizdir.

- **Bütünlük:** Eldeki verinin, analiz yapılan dönem ya da konuya dair tüm kayıtları ve örnekleri barındırıyor olmasıdır. Örneğin; Yıl içerisindeki tüketici hareketlerini izleme konulu bir çalışmada 3 aylık market verisi sağlıklı bir analiz yapabilmek için yeterli olmayacaktır.
- **Tutarlılık:** Aynı konu, kişi, alan vb. ile ilgili elde bulunan birden fazla kaydın birbirleri ile tutarlı olması durumudur. Tutarsızlık durumunda doğru bilgi her zaman öğrenilemeyebilir ve çözümleme yapılamayabilir.
- **Tekdüzelik:** Aynı veriye ait ölçümlerin her zaman aynı tür ve yapıda kayıt altına alınmasıdır. Örneğin; kişinin boyunun bazı yerlerde cm, bazı yerlerde metre cinsinden kaydedilmesi geçerli bir durum olsa bile tekdüzelik kuralını ihlal ettiği için verinin kalitesini olumsuz etkiler.
- **Yoğunluk:** Analizde kullanılacak verinin çeşitliliği ve miktarının problem için yeterli düzeyde olmasıdır. Veri araştırma yapılacak konuyu temsil edecek düzeyde olmalıdır.
- **Eşsizlik:** Veri kümesindeki verilerin hepsinin birbirinden farklı olması, diğer bir ifade ile tekrar eden verinin olmaması durumudur. Kayıt tekrarının geçerli nedenlerinin olması eşsizlik prensibine zarar vermez ancak; aynı kaydın hatalı olarak tekrarı, sınıflandırma ve kümeleme gibi yöntemleri olumsuz etkiler.

- **Tamlık:** Bütünlük ve geçerlilikle ilişkili bir kalite kriteridir. Eldeki veriler geçerli ve tüm çeşitliliği içeren yeterli düzeyde veri içeriyorsa, veri kümesinin tam olduğu söylenebilir.
- **Doğruluk:** Tamlık, tutarlılık ve yoğunluk ile ilişkili bir kalite kriteridir. Bu koşullar sağlanıyorsa verinin doğru olduğu söylenebilir (Köse, 2018; Silahtaroğlu, 2020).

3.2.2. Verinin Temizlenmesi

Kirli veri olarak da adlandırılan kayıp ve gürültünün ortadan kaldırılmasıdır. Yanlış ve aşırı uçta olan verilerin ortadan kaldırılmasını da kapsar. Veri temizleme ifadesi, kirli verilerin veri kümesinden çıkarılması anlamına gelmemektedir. Veri kümesinden verilerin çıkarılması, veri temizleme aşamasında başka hiçbir çare kalınmadığında başvurulan yöntemdir. Mümkünse, öncelikli olarak belirlenen eksikliğin, hatanın telafi ve tamir edilmesine çalışılır. Veri temizleme için uygulanan temel yöntemler:

- Eksik verileri tamamlama
- Aykırı verileri düzeltme
- Tutarsız verileri düzeltme

Eksik verileri tamamlama için kullanılan temel yöntemler:

- **Kaydın tamamen silinmesi**: Özellikle sınıf niteliğinin eksik olduğu kayıt, tahmin edici bir analizde kullanılamaz. Bu nedenle kaydın kendisi direkt olarak silinir. Sınıf niteliği dışındaki niteliklerin eksik olması durumunda, veri eksikliğinin oranı incelenir. Örneğin; %30'dan fazla eksik veri içeren niteliklerin, nitelik vektöründen çıkarılması işlemi gerçekleştirilebilir.
- **Veri kümesinin ortalamasının alınması:** Veri kümesindeki ilgili özniteliğin ortalaması ya da en çok tekrar eden nominal değer, eksik olan kısma yazılmasıdır. Bu tür müdahalelerde konuya hakim olan uzman görüşü oldukça önemlidir.
- **Kümenin ortalamasının alınması:** Veri seti üzerinde kümeleme yapılarak, eksik veri yerine ilgili küme ortalaması ya da en çok tekrar eden nominal değerin yazılmasıdır.
- **Tahmin edilmesi:** Eksik verinin bir öğrenme algoritması ile tahmin edilmesidir.

Aykırı verileri düzeltmek için kullanılan temel yöntemler:

- **Sepetleme**: Öznitelik değerine göre sıralama yapılarak, ortalama, orta ve sınır sepetlerinin oluşturulmasıdır. Aykırı değerler yerine, oluşturulan sepetlerdeki değerlerin ortalama değeri yazılabilir.
- **Kümeleme**: Kayıtlar kümelendikten sonra kümenin dışında kalan kayıtların aykırı değer olarak kabul edilip, veri kümesinden çıkarılmasıdır.
- **Regresyon**: Aykırı olan verinin, regresyon yöntemi ile yeniden hesaplanmasıdır.

Tutarsız verileri düzeltmek için kullanılan temel yöntemler:

Uzman görüşü: Tutarsız veriler yerine ne yazılabileceği konusuna tek çözüm genellikle uzman görüşüne başvurmaktır. Ancak; her tutarsızlığın çözümü mümkün olmayabilir. Sağlıklı karar alınamıyorsa kayıt veri kümesinden çıkarılır (Köse, 2018; Silahtaroğlu, 2020).

3.2.3. Veri Dönüştürme

Veri madenciliği veya makine öğrenmesinde kullanılacak model, teknik ve algoritmalar belirli türdeki veriler ile çalışıp, bazı veri türlerinde ise çalışmamaktadırlar. Bazı algoritmalar sadece sayısal değerler ile çalışmakta iken, bazı algoritmalar kategorik değerler kullanırlar. Bazı algoritmalar ise sadece 0 ve 1'lerden oluşan değerler ile çalışmaktadırlar. Bu durumda elimizdeki verileri amacına uygun bir şekilde yeniden yapılandırmamız gerekmektedir. Veri dönüştürmeyi gerektiren durumları şu şekilde inceleyebiliriz:

- **Niteliklerdeki verilerin dağılımlarındaki farklılıklar:** Örneğin; mevcuttaki öznitelik vektöründeki "yaş" niteliğindeki değerler 0-120 arasında değişirken, "gelir" niteliğindeki değerler 22.000-100.000 arasında, "toplam varlık" niteliği ise 25.000-5.000.000 arasında değişiyor olabilir. Bu durumda niteliklerdeki sayısal değerlerin değişim aralığı birbirlerinden çok farklıdır. Bu durum, sayısal değerleri esas alarak hesaplama yapan algoritmaların, her nitelik için aynı hassasiyeti gösterememesine neden olur. Bu nedenle özniteliklerdeki değişimin sayısal miktarı açısından değil de oransal değişimler açısından incelenmelidir. Bu tür durumlarda nitelikteki verilerin normalizasyon işlemine tabii tutulması gerekmektedir. Yaygın olarak kullanılan normalizasyon yöntemleri:

  - ✓ Min-Max Normalizasyonu
  - ✓ Z-Skor Normalizasyonu
  - ✓ Ondalık Esaslı Normalizasyon

(Köse, 2018; Silahtaroğlu, 2020)

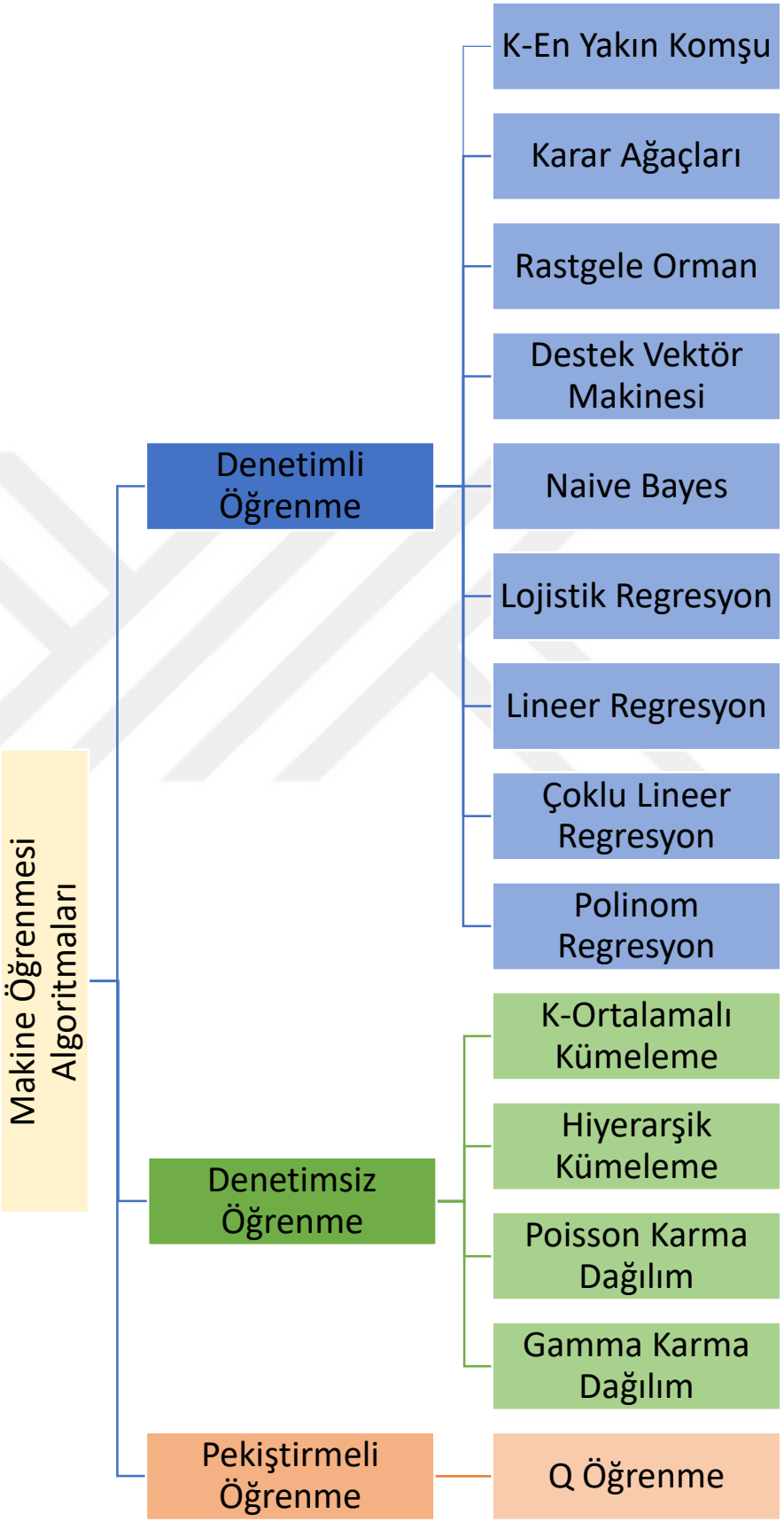
Mevcut niteliklerin değerlerinden elde edilecek yeni ve daha anlamlı niteliklerin olması: Bazı durumlarda, nitelik vektöründeki kimi nitelikler başka nitelikler ile birlikte yeni ve kendilerinden daha anlamlı, hesaplamaları kolaylaştırıcı niteliklerin elde edilmesine olanak sağlamaktadır. Örneğin; "boy" ve "kilo" nitelikleri kullanılarak, nitelik vektöründe yer almayan "vücut kitle indeksi" niteliği hesaplanarak, analizlerde diğer iki nitelikten daha anlamlı olabilmesi sağlanabilir. Bazı durumlarda, genelleştirme ihtiyacı olabilir. Örneğin "hastane adı" niteliği yerine başka verileri de dikkate alarak "hastane türü: {devlet, özel, üniversite}" gibi bir niteliğe dönüştürmek bazı durumlarda yararlı olabilir. Bazı durumlarda da nitelik vektöründeki kimi nitelikler başka niteliklerin değerleri üzerinde çalıştırılan toparlama fonksiyonları ile elde edilebilir. Örneğin; mevcutta sağlık işlemleri ile ilgili bir veri varken, nitelik vektöründe "yıllık ortalama muayene sayısı" gibi bir niteliğe ihtiyaç duyulabilir. Bu durumda sağlık işlemlerinden bu değerin hesaplanması ve yeni bir nitelik olarak nitelik vektörüne eklenmesi gerekmektedir. Yaygın olarak kullanılan normalizasyon yöntemleri:

- ✓ Yeni nitelik oluşturma
- ✓ Genelleştirme
- ✓ Toparlama (Köse, 2018) (Silahtaroğlu, 2020)

3.2.4. Verilerde Boyut İndirgeme İşlemleri

Normalizasyon dışında verilerin miktarlarının indirgenmesi gereken durumlar da oluşabilmektedir. Miktarın indirgenmesi, değişken sayısının azaltılmasıdır. Örneğin; gereksiz yere tutulan bir değişken var ise bu değişkenin kaldırılması işlemi de bir indirgeme işlemidir. Ancak; genellikle indirgeme işlemi birden çok değişkenin birleştirilerek tek bir değişkenle ifade edilmesi durumudur. Algoritmanın yapısı ve çıkacak sonuçların hassasiyeti açısından, belirli değişkenler birleştirilip, tek bir değişken olarak işleme girdirilebilirler (Köse, 2018; Silahtaroğlu, 2020).

3.3. Makine Öğrenmesinde Algoritmalar



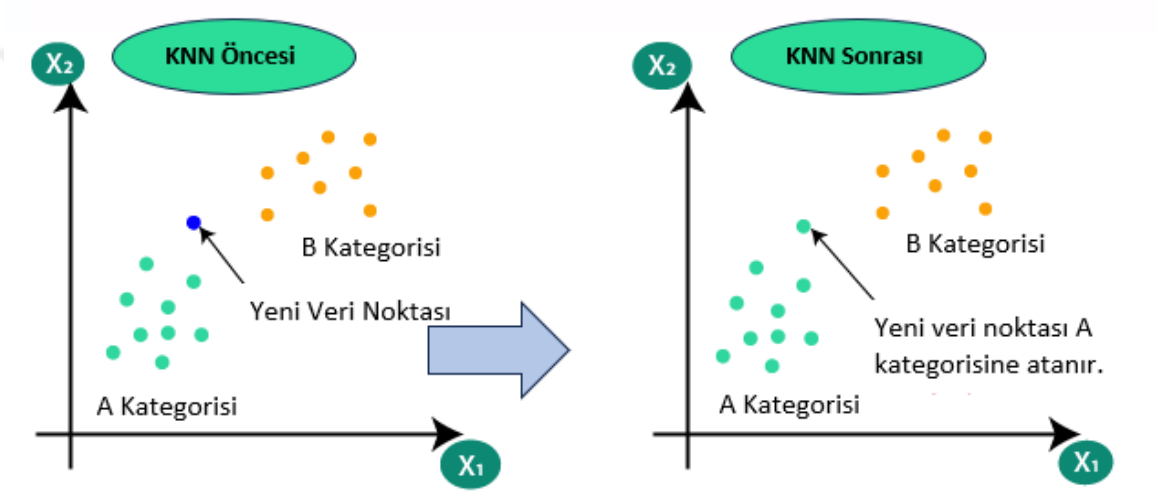
Şekil 3.7. Makine Öğrenmesi Algoritmaları

Bu bölümde Şekil 3.7' de yer alan algoritmalar incelenecektir.

3.3.1. K-En Yakın Komşu Algoritması (K Nearest Neighbour, KNN)

K En yakın komşu sınıflandırma ve regresyon problemleri için kullanılan bir denetimli makine öğrenmesi algoritmalarından biridir. KNN algoritması özellikle küçük boyutlu veri kümelerinde ve sınıflar arasında ayrımın belirgin olduğu durumlarda etkili çalışır. Yeni sınıflandırılacak olan veri örneği ile eğitim kümesindeki verilerin uzaklıklarını ölçerek en yakın K adet komşunun sınıf değerlerine göre, veriyi sınıflandırma işlemini gerçekleştirir.

Sınıflandırılacak veri ve eğitim verileri arasındaki mesafeyi ölçmek için Öklid, Manhattan ve Minkowski uzaklık ölçüm formülleri kullanılır. (Hacıbeyoğlu ve ark. 2023)



Şekil 3.8. K-En Yakın Komşu Algoritması

KNN iş akışı şu şekildedir:

1. Komşu sayısı K değerinin belirlenmesi
2. K adet komşunun mesafesinin hesaplanması
3. Hesaplanan mesafelere göre K en yakın komşunun seçilmesi
4. K komşulardan her kategorideki veri noktasının sayılması
5. Yeni veri noktalarının, komşu sayısı maksimum olan kategoride atanması

KNN algoritmasının avantajları:

- Basitlik ve doğruluk
- Parametre ayarının kolaylığı
- Örnek bazlı öğrenme: Veri setinin yapısına ve dağılımına dair önemli varsayımlar yapmaz, bu nedenle KNN esnek bir modeldir.

KNN algoritmasının dezavantajları:

- Boyutluluk problemi: Bir nokta ile tüm veri setindeki noktalar arasındaki mesafenin bulunması gerekir. Bu işlem de boyutluluk problemine sebep olur ve işlem çok uzun sürer.
- Verilere duyarlılık: Aykırı değerlerden çabuk etkilenir.
- Yüksek hesaplama maliyeti: Tek bir nokta için tüm veri seti taranıp, mesafeler bulunmalı, sıralanmalı, K değerine göre sınıflandırma yapılmalıdır. Bu işlemler veri sayısı arttıkça uzun sürebilen işlemlerdir.

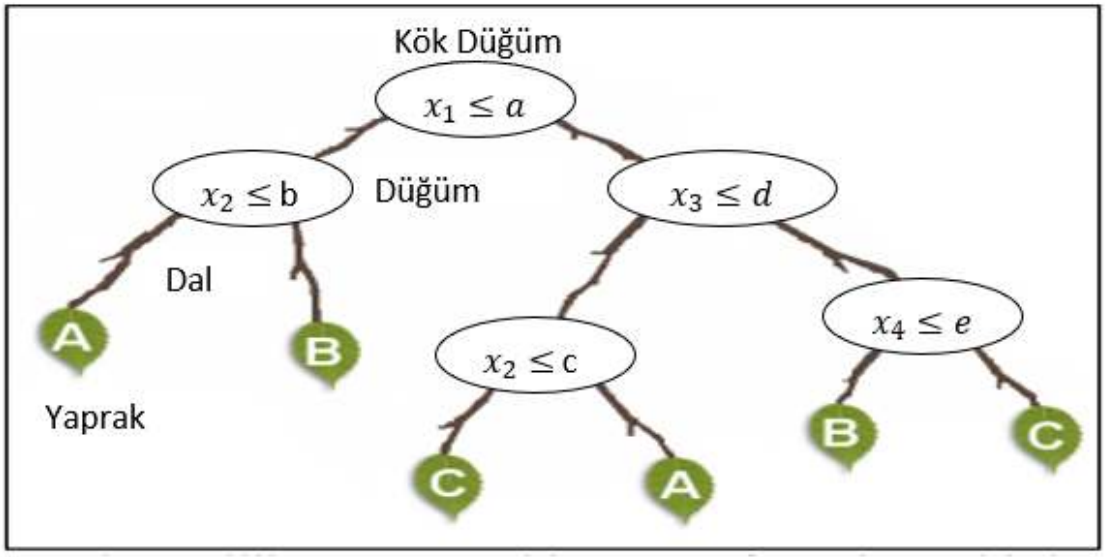
Şekil 4.9‘ da görüleceği üzere; K değeri seçildikten sonra model eğitilir ve veri noktalarının en yakın komşuları belirlenir. Söz konusu veri, komşu sayısının daha fazla olduğu kategoriye atanır.

3.3.2. Karar Ağaçları

Karar ağaçları (KA), bir sınıflandırma ve örüntü tanımlama algoritmasıdır. Bu yöntemin yaygın olarak kullanılmasının önemli nedenlerinden biri ağaç yapılarının oluşturulmasında kullanılan kuralların anlaşılabilir ve sade olmasıdır. KA sınıflandırma işleminin gerçekleştirilmesinde çok aşamalı ve ardışık bir yaklaşım kullanmaktadır.

Bir karar ağacı Şekil 4.10’ da görüleceği gibi düğüm, dal ve yaprak olmak üzere üç temel kısımdan oluşur. Bu ağaç yapısında her bir öznitelik bir düğüm tarafından temsil edilir. Ağaçta en son kısım yaprak, en üst kısım ise kök olarak adlandırılır. Kök ve yapraklar arasında kalan kısımlar ise dal olarak ifade edilir.

Eğitim verilerine ait özniteliklerden faydalanılarak bir karar ağacı oluşturulurken, verilere ilişkin bir dizi sorular sorulur ve elde edilen cevaplar ile hareket edilerek en kısa sürede sonuca gidilmeye çalışılır. Bu şekilde karar ağacı sorulara aldığı cevapları toplayarak karar kuralları oluşturur. Ağacın ilk düğümü olarak ifade edilen kök düğümde, verilerin sınıflandırılması ve ağaç yapısının oluşturulması için sorular sorulmaya başlanır ve dalları olmayan düğümler ya da yapraklar bulunana kadar bu işlem devam eder. Oluşturulan ağacın yeni bir veri seti için genelleme kabiliyetinin ölçülmesinde test veri seti kullanılır.



Şekil 3.9. Dört boyutlu özellik uzayına sahip üç sınıftan oluşan bir karar ağacı yapısı

KA oluşturulmasındaki en önemli adım, ağaçtaki dallanmanın hangi kriterlere göre yapılması gerektiğidir. Literatürde bu problemin çözümü için geliştirilen farklı yaklaşımlar mevcuttur. Bunların en önemlileri; bilgi kazancı ve bilgi kazanç oranı, Gini indeksi, Towing kuralı, Ki-kare olasılık tablo istatistiği gibi yaklaşımlardır.

- Gini; karışıklığı ölçen bir metriktir. Gini; karışıklığı ifade ettiğine göre karar ağaçlarında karışıklığın az olması istenir.
- Entropi; düzensizliği veya belirsizliği ölçer. Entropinin düşük olması, heterojenliğin az olması, homojenliğin yüksek olması yani bilgi kazancının yüksek olması istenir.

(Kavzaoğlu ve Çölkesen, 2010)

KA Avantajları:

- Anlaşılabilirlik
- Değişken Önemi
- Çok çıkışlı modellere uygunluk: KA; sınıflandırma ve regresyon gibi farklı problem türlerine uygun olarak kullanılabilir.
- Veri ön işleme ihtiyacının azalması: Genellikle veri setinin normalize edilmesi ya da ölçeklendirilmesi gibi karmaşık ön işleme adımlarına ihtiyaç duyulmaz. Ayrıca kategorik değişkenler ile doğrudan çalışabilir.

KA Dezavantajları:

- Aşırı uyum: Hiper-parametreler düzgün seçilmez ise veriyi ezberler.
- Duyarlılık: Veri kümesindeki küçük değişiklikler modelin yapısını önemli ölçüde etkileyebilmektedir. Aykırı değerlere karşı hassastırlar.
- Dengesiz veri kümeleri

3.3.3. Rastgele Orman

Karar ağaçlarının bir araya gelmesi ile ormanları oluşturulur. RO, ağaç tipi sınıflandırıcılar topluluğudur, ormandaki her ağaç bağımsız olarak örneklenen bir rasgele vektörün değerlerine ve aynı dağılıma bağlıdır

Sınıflandırma problemlerinde KA kullanılabildiği gibi rasgele ormanlar da kullanabilir. Karar ağaçları veri noktalarını belirli sınıflara ayırmak için kullanılır. Ardından orman içerisindeki tüm ağaçların tahminleri birleştirilerek en çok oy alan sınıfın tahmini RO sonucudur.

KA ile yapılabilen her şey rasgele ormanlarda da yapabilir. RO'nun çıkma sebebi; aşırı öğrenmenin önüne geçmek ve daha büyük veri setleri ile daha karmaşık problemleri çözebilmektir (Ekelik ve Altaş, 2019).

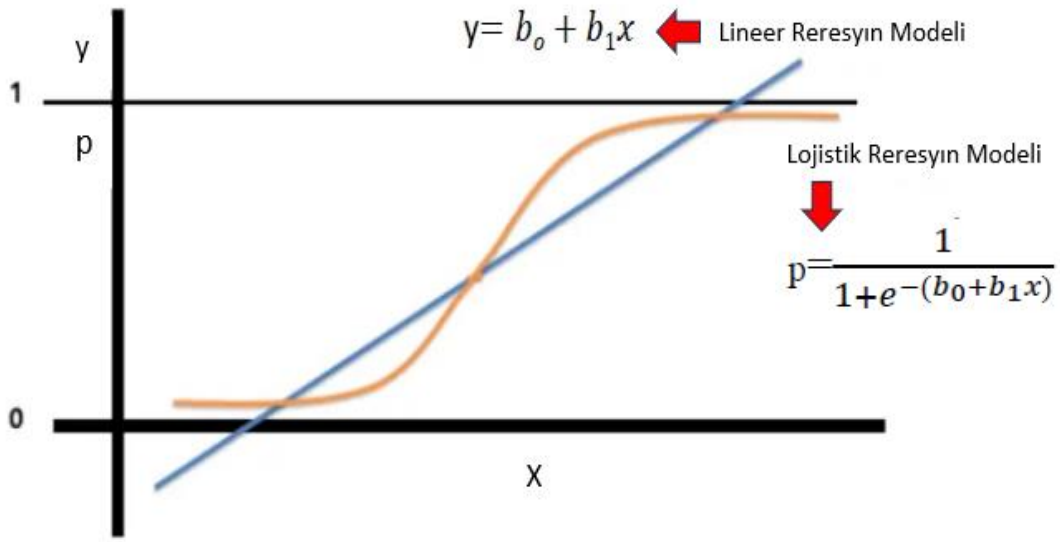
3.3.4. Lojistik Regresyon

Sınıflandırma problemlerini çözmek için kullanılan gözetimli ML öğrenme yöntemlerinden biridir.

Her ne kadar isminde "regresyon" kavramı yer alsa da, sınıflandırma problemlerini çözmek için geliştirilen bir algoritmadır. "Regresyon" ifadesinin kullanılmasının nedeni, algoritmanın lineer regresyon yaklaşımından türetilmiş olmasıdır. Lineer regresyon modeline lojistik sigmoid fonksiyonun eklenmesiyle elde edilen lojistik regresyon, çıktı değerlerini olasılık biçiminde yorumlanabilir hale getirir. Böylece lojistik regresyon, bağımlı değişkenin belirli sınıflara ait olma olasılığını tahmin ederek sınıflandırma problemlerini çözer; regresyon analizi ile doğrudan bir ilişkisi bulunmamaktadır.

Geleneksel makine öğrenmesi algoritmaları örneğin; müşterinin bir ürünü satın alıp almayacağını ikili bir çıktı (0 ya da 1) şeklinde tahmin edebilmektedir. Lojistik regresyon algoritması ise müşterinin bir ürünü satın alıp almama olasılığını tahmin eder.

Lojistik regresyon, derin öğrenmenin temeli olduğu için derin öğrenmede de çıktılar olasılıksal olarak ortaya çıkmaktadır.



Şekil 3.10. Lojistik Regresyon

İkili (binary) veriler, kategorik verilerin en yaygın biçimidir. Bu nedenle ikili verilerin modellenmesi için en sık kullanılan yöntem lojistik regresyondur.. Lojistik regresyon analizi, bağımlı değişkenin iki kategoriden oluştuğu durumlarda, bu değişkene ait değerleri bir veya birden fazla bağımsız (tahmin edici) değişken aracılığıyla tahmin etmeye yönelik istatistiksel yöntemdir (Bulut ve Eygü, 2020).

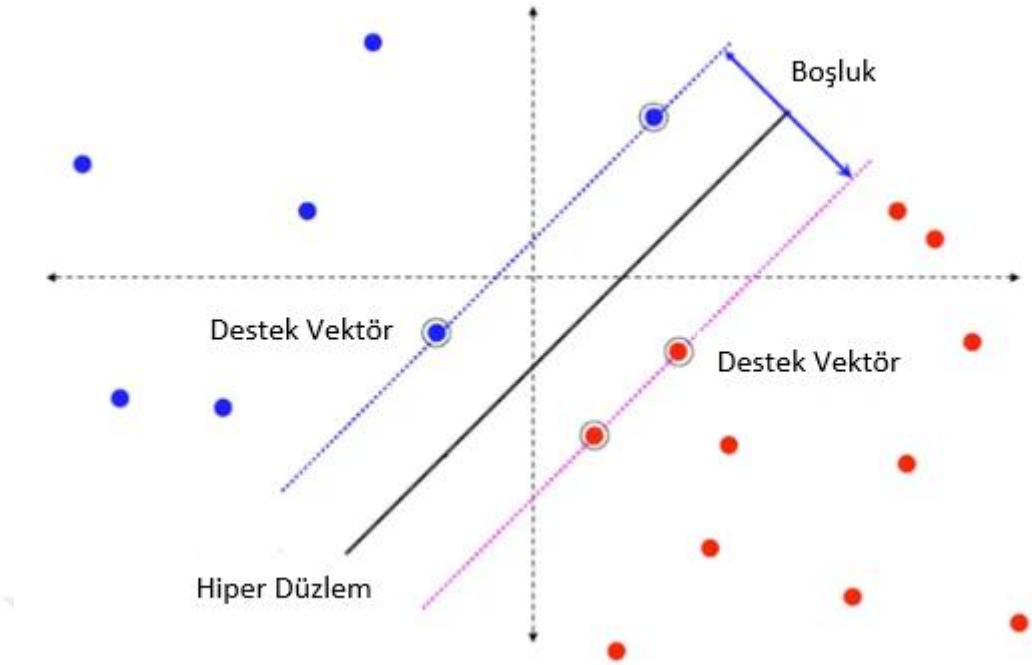
Tıp ve sağlık alanında; hastalıkların teşhis edilmesi, hastalığın riskinin belirlenmesi, tedavi başarısının tahmin edilmesi gibi problemlerin çözümünde;

Finans alanında; risk yönetimi, kredi riski değerlendirmesi, dolandırıcılık tespiti gibi finansal uygulamalarda,

İnsan kaynakları alanında: personel devir hızının tahmin edilmesi, performans değerlendirmesi gibi konularda, lojistik regresyon sıklıkla kullanılmaktadır.

3.3.5. Destek Vektör Makinesi

Destek Vektör Makinesi (Support Vector Machine, SVM), sınıflandırma ve regresyon problemlerini çözmek için kullanılan denetimli ML algoritmasıdır. İstatistiksel öğrenme teorisi ve yapısal risk minimizasyonu teknikleri üzerine kurulduğundan dolayı teorik altyapısı güçlü makine öğrenmesi tabanlı örüntü sınıflandırma tekniğidir. SVM' nin temel amacı eğitim verilerini, bilinen sınıf etiketleriyle ayırabilen çok boyutlu uzayda bir fonksiyon elde etmektir. Sınıf etiketleri genellikle pozitif ve negatif olarak adlandırılan Şekil 4.12 de görüleceği gibi, veri seti içerisinde en uygun ayıracı yani hiper-düzlemi bulabilmektir. (Harman, 2021)



Şekil 3.11. Destek Vektör Makinesi

3.3.6. Naive Bayes

Makine öğreniminde sınıflandırma ve olasılıklı tahminler yapmak için kullanılan olasılıksal bir yöntemdir. Bayes teoremine dayalı bir olasılık tabanlı sınıflandırma yöntemidir.

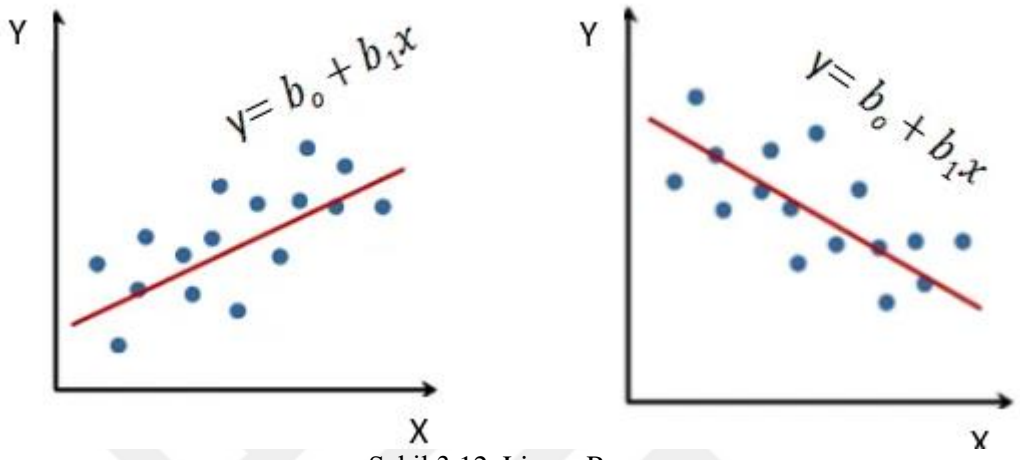
Naive Bayes algoritması sınıflandırma problemini çözmek için kullanılan makine öğrenmesinin gözetimli öğrenme algoritmalarından bir tanesidir. Tahmine dayalı modelleme yapmak için kullanılan algoritmalardan biridir. Bu yüzden, özellikle sinyal ve görüntü işleme alanlarında sıklıkla kullanılır. Bu algoritmada, bir sınıftaki belirli özelliklerin varlığının diğer herhangi bir özellik ile ilgili olmadığı varsayılır (Harman, 2021).

Çizelge 3.1. Naive Bayes Algoritmasının Zayıf ve Güçlü Yönleri

Özellikler	**Zayıf Yönler**	**Güçlü Yönler**
Hiper-parametreler	Daha az hiper-parametreye sahip olması nedeni ile daha az esneklik sağlar.	Basit, hızlı ve düşük veri gereksinimiyle yüksek performans sağlar.
Öznitelik Bağımsızlığı	Öznitelikler arasındaki bağımsızlık varsayımı gerçek dünyada her zaman geçerli değildir.	Metin sınıflandırma ve spam filtreleme gibi birçok uygulama alanında başarılıdır.
Hassasiyet	Genellikle çok karmaşık yapıları yakalayamaz. Bu nedenle yüksek hassasiyet gerektiren problemlerde performansı düşük olabilir.	Veri seti dengesizliğine ve gürültülü veriye karşı duyarlıdır.

3.3.7. Lineer Regresyon

Doğrusal regresyon, gözetimli öğrenmede kullanılan basit fakat etkili bir araçtır. Tek bağımsız değişken olduğu durumda doğrusal regresyon y=$(b_0 + b_1 x)$ denklemini çözmeye çalışır. Bağımsız değişkenlerin, bağımlı değişkene olan etkisini tahmin etmek için kullanılır (Hacıbeyoğlu ve ark., 2023).



Şekil 3.12. Lineer Regresyon

Pazarlama alanında; müşteri davranışlarını anlamada ve tahmin etmede kullanılabilir. Örneğin; reklam harcamaları ile satışlar arasındaki ilişkiyi incelemede kullanabilir.

Sağlık alanında; ilaç dozajlarının etkilerini belirlemek, hastalık risklerini tahmin etmek gibi alanlarda kullanılabilir.

Çizelge 3.2. Lineer Regresyon Algoritmasının Zayıf ve Güçlü Yönleri

Zayıf Yönler	**Güçlü Yönler**
Doğrusal olmayan ilişkileri yakalamakta zorlanabilir.	Basit ve anlaşılır modelleme
Aykırı değer hassasiyeti	Yüksek yorumlanabilirlik
	Kısa eğitim süresi

3.3.8. Çoklu Lineer Regresyon

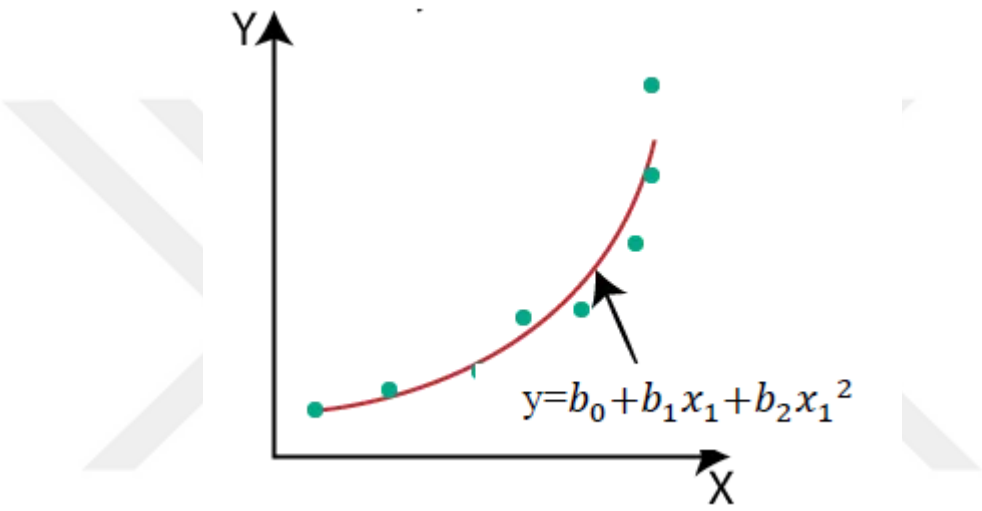
Günümüzde makine öğrenmesi problemlerinde çok sayıda bağımsız olduğu düşünülen değişken bulunmaktadır. Çok değişkenli doğrusal regresyon yöntemi ile bağımsız değişkenler ve bağımlı değişken y arasındaki ilişkiyi modellemeyi amaçlar. y = f $(x_1 x_2, \ldots x_n) = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$ denklemini çözmeye çalışır ve bu işlemi eğitim kümesindeki verileri kullanarak yapar. Bu denklemde a_i'ler , i=1,2,…,n olmak üzere doğrusal denklemin bilinmeyen değerleridir ve gerçek y değerleri ile modelin tahmin ettiği y değerleri arasındaki artık kareler toplamını en küçük değerde tutmayı hedefler (Hacıbeyoğlu ve ark., 2023).

Ekonomi ve finans alanında; hisse senedi fiyatlarının tahmininde, faiz oranları-enflasyon gibi ekonomik değişkenleri arasındaki ilişkiyi analiz etmek için kullanılabilir. Farklı ülkelerin ya

da firmaların tüketici davranışları, gelir tahmini gibi finansal analizlerde de kullanılabilir. Pazarlama, sağlık bilimleri, eğitim gibi alanlarda da kullanılırlar.

3.3.9. Polinom Regresyon

y bağlı değişkeni ile x serbest değişkeni arasındaki eğrisel ilişki; ikinci derece bir parabol, polinom, logaritmik ya da bir başka modelle ifade edilebilir. Lineer regresyonun genişletilmiş bir versiyonu olan polinom regresyon, doğrusal olmayan ilişkileri modellemek için kullanılır. Polinom regresyonda bağımsız değişkenlerin doğrusal birleşimi yerine polinom terimleri kullanılır. Denklemdeki derecenin büyüklüğüne göre eğri şekli değişmektedir (Erkan, 2002).



Şekil 3.13. Polinom Regresyon

Polinom derecesi modelin esnekliğini ve karmaşıklığını belirler. Bu nedenle polinom regresyonunda hiper-parametre seçimi önemlidir. n değeri yanlış seçilirse model, veri setine uygun kararlar ortaya çıkaramayacaktır.

Fizik alanında; yayın elastik davranışını modellemek için gerilme ve uzama arasındaki ilişkiyi belirlemede;

Finans ve ekonomi alanında; şirketin gelirinin zamana bağlı nasıl değiştiğini analiz etmede;

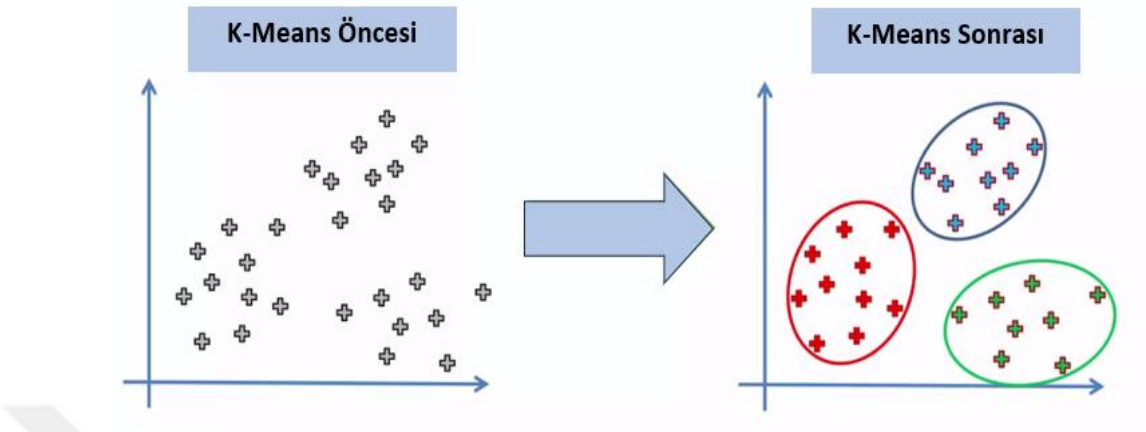
İklim bilimi alanında; iklim verilerinin analizi ve iklim modellerinin oluşturulmasında polinom regresyonu kullanabilir.

Çizelge 3.3. Polinom Regresyon Güçlü-Zayıf Yönler

Zayıf Yönler	**Güçlü Yönler**
Yüksek dereceli polinomlar aşırı uyuma yol açabilir.	Esneklik

3.3.10. K-Ortalamalı Kümeleme (K-Means Clustering)

K-ortalamalı kümeleme gözetimsiz öğrenme algoritmalarından biridir. Veri içerisindeki benzer örnekleri belirli sayıda kümelere gruplamak için kullanılır.



Şekil 3.14. K-Ortalamalı Kümeleme

K-Ortalamalı Kümeleme Algoritması adımları şunlardır:

Küme Merkezi Başlatma: İlk adımda belirlenen K sayısı kadar, veri seti içerisinden rastgele küme merkezi seçilmesidir.

Veri Noktalarını Kümelere Atama: Her bir veri noktası en yakın küme merkezine atanır.

Küme Merkezlerini Yeniden Hesaplama: Her bir küme merkezi, kendi kümesine yeni veriler atandıktan sonra, atanan veri noktalarının ortalamaları alınarak yeniden hesaplanır.

Yeniden Atama ve Güncelleme: Belirli bir iterasyon sayısı kadar veya kümelerde herhangi bir değişiklik olmayana kadar 2. ve 3. adım tekrarlanır (Günher ve ark., 2025).

Şekil 4.16'da K adet küme oluşturulmaktadır. K değeri bir hiper-parametre olup kaç adet küme oluşturulacağı en baştan belirlenmelidir. Şekil 3.14'de K-Means algoritmasından önce ve K-Means algoritmasından sonra olmak üzere 2 adet grafik yer almaktadır. İlk grafikteki veri noktalarının herhangi bir etiketi ve sınıfları yoktur. Bunları gruplayabilmek, kümeleyebilmek için k ortalamalı kümeleme algoritması kullanılır.

3.3.11. Hiyerarşik Kümeleme

Veri setindeki örnekleri bir hiyerarşik yapı içinde gruplamak için kullanılan bir gözetimsiz öğrenme yöntemidir. Kümelerin ayrı olarak ele alınıp aşamalı olarak bir küme içerisinde birleştirilmesi veya kümelerin bir ana küme olarak ele alınıp, aşamalı olarak alt kümelere ayrılması yöntemidir. Bu yöntemde; her veri noktası bir küme olarak başlar ve ardından benzerlikleri veya

uzaklıkları temel alarak bu kümeler birleştirilir veya bölünür. Bu işlem, bir ağaç benzeri hiyerarşi oluşturur. Hiyerarşik kümeleme iki ana türde gerçekleştirilebilir:

- **Birleştirici Hiyerarşik Küme:** Başlangıçta her bir veri noktası (gözlem) bağımsız bir küme olarak değerlendirilir. Ardından diğer veri noktaları ile olan benzerliklerini temel alarak bu kümeler bir araya getirilir ve tek bir küme olana kadar veya birleştirme kriterini sağlayan belirli bir küme sayısına ulaşıncaya kadar en yakın kümeler birleştirilir.
- **Ayrıştırıcı Hiyerarşik Kümeleme:** Başlangıçta bütün gözlemler tek bir küme olarak değerlendirilir. Daha sonra tekrarlı bir şekilde bütün gözlemler birbirinden bağımsız tek bir küme oluncaya kadar, her bir gözlem kümesi veya gözlem kendisine en uzak olan gözlem ya da gözlem kümesinden ayrılıp, yeni bir küme oluşturur. Başlangıçta tek bir küme olduğundan her adımda 1 veya daha fazla küme bölünür. Bu süreç bölme kriteri karşılanana veya belirli bir küme sayısına ulaşıncaya kadar devam eder (Yeşilbudak ve ark., 2010).

3.4. Karma Dağılımlar

Denetimsiz makine öğrenme yöntemlerinden bir tanesi de karma dağılımlardır. Karma dağılımlar iki veya daha fazla istatistiksel dağılımın belli oranlarda birleşmesinden meydana gelir. Bu dağılımlar aynı dağılımlar olabileceği gibi farklı dağılımlarda olabilir. Aynı türden dağılımların oluşturduğu karma dağılımlara saf, farklı türden dağılımların oluşturduğu dağılımlara katlama (convolute) karma dağılımlar adı verilir.

Gözlemlerin tek bir popülasyonda toplandığı varsayımı başka bir deyişle homojen olduğu varsayımı genellikle gerçekçi değildir ve veriyi doğru biçimde temsil etmek için yeterli değildir. Bu noktada, karma dağılım modelleri, farklı parametrelere sahip heterojen yapıda alt popülasyonların varlığını dikkate alan bir yaklaşım sunduklarından dolayı, veriyi temsil etmede önemli bir rol oynamaktadır.

Karma dağılım modelleri, sınıflandırma yöntemleri olarak da bilinmektedir. Yığılım analizi, ayrıştırma analizi ve görüntü analizi gibi istatistiksel teknikleri içermektedir. Farklı bileşen sayısından oluşan dağılımların ağırlıklı kombinasyonunu kullanarak, karmaşık verilerin modellenmesinde kullanılır. Bu dağılımlar, gerçek hayattaki karmaşık olayları daha iyi açıklamak için kullanılır. Bu nedenle, karma dağılım modelleri teorik ve pratik uygulamalarda önemi daha iyi kavranmakta ve kullanımı giderek artmaktadır.

Karma dağılımlarda olasılık yoğunluk fonksiyonu, karma dağılımı oluşturan dağılımların olasılık yoğunluk fonksiyonlarının ağırlıklı toplamı olarak ifade edilir. Her bir bileşenin ağırlığı, ait olduğu dağılımın karma dağılıma olan etkisini ifade eder. Olasılık yoğunluk fonksiyonu, farklı bileşenlerin etkileşimini ve ağırlıklarını dikkate alarak verinin dağılımını modellemeye çalışır.

Karma dağılımların kümülatif olasılık yoğunluk fonksiyonu, karma dağılımı oluşturan dağılımların kümülatif olasılık yoğunluk fonksiyonlarının ağırlıklı toplamı olarak ifade edilir ve veri içerisinde bir değerin belirli bir değeri aşma olasılığını ifade eder.

Karma dağılımlar, karma oranı (π_1, π_2, π_3), bileşen sayısı (k) ve dağılım parametreleri $f_1(x; \theta_1)$, $f_2(x; \theta_2)$…, $f_k(x; \theta_k)$ olmak üzere üç nitelik ile tanımlanır.

$$f(x) = \pi_1 f_1(x; \theta_1) + \pi_2 f_2(x; \theta_2) + \ldots + \pi_k f_k(x; \theta_k)$$

$$0 \leq \pi_i \leq 1 \qquad i = 1,2,\ldots,k$$

$$\sum_{i=1}^{K} \pi_i = 1$$

Karma dağılımların genel bir gösteriminde parametreler için ψ sembolünü kullanılır.

$$f_1(x; \psi) = \sum_{i=1}^{K} \pi_i \, f(x, \theta_i)$$

Bir veri setinin karma dağılım ile modellenebilmesi için:

1. Bileşen sayısı tahmin edilmelidir
2. Karma oranları belirlenmelidir
3. Dağılım parametreleri tahmin edilmelidir

Optimum bileşen sayısı (k)' yı bulmanın amacı; veriyi iyi ayrılmış kümelere bölmektir. Diğer bir deyişle, heterojen veriyi, homojenlik gösteren anlamlı parçalara bölmektir. Bileşen sayısı belirlenirken bilgi kriteri analizinin de yapılması gereklidir. Bu çalışmada en uygun bileşen sayısının belirlenmesinde Akaike Bilgi Kriteri (AIC) ve Bayesci Bilgi Kriteri (BIC) yöntemleri kullanılmıştır. En uygun karma dağılımı ve bileşen sayısı belirlendikten sonra karma dağılımın parametreleri tahmin edilmelidir (Güleryüz, 2023).

3.4.1. Karma Dağılımlarda Bileşen Sayısı Belirleme

Kümeleme, benzer veri noktalarını gruplara ayırmak için kullanılan bir veri madenciliği tekniğidir. Bir kümeleme algoritması, veri kümesindeki benzer veri noktalarını bir araya getirerek kümeler oluşturur Kümelerin sayısını belirlemek kimi durumlarda zorlayıcı olabilmektedir. Bu nedenle kümeleme algoritmaları genellikle verideki kümelerin sayısını otomatik olarak belirleyebilen özellikler içermektedirler.

Karma dağılımlarda bileşen sayısının belirlenmesi, kümeleme algoritmalarının veri kümesinde kaç kümenin var olduğunu saptamasına yönelik kullanılan bir yöntemdir. Bu yaklaşım, her bileşenin ayrı bir olasılık dağılımı ile temsil edildiğini varsaymaktadır. Farklı bileşen sayıları

için alternatif modeller kurulmakta ve her bir modelin veri ile uyumu istatistiksel ölçütler aracılığıyla değerlendirilmektedir. En uygun model, veri kümesine en yüksek uyumu sağlayan model olarak kabul edilmektedir. Bileşen sayısının belirlenmesindeki temel amaç, aynı küme içerisindeki gözlemler arasındaki benzerliği en yüksek düzeye çıkarırken, farklı kümeler arasındaki farklılığı en belirgin hale getirmektir.

3.4.2. Karma Dağılımlarda Parametre Tahmini

Karma dağılım modellerinde bileşen sayısı belirlendikten sonra, her bir bileşene ait parametreler tahmin edilmelidir.

En Çok Olabilirlik Yöntemi (Maximum Likelihood)

Parametrelerin tahmininde temel olarak En Çok Olabilirlik Yöntemi kullanılmaktadır. Karma dağılım modelinin olabilirlik fonksiyonu karmaşık olduğundan, bu tahminlerin hesaplanması zor olabilir. Bundan dolayı, En Çok Olabilirlik Yöntemi temelinde daha gelişmiş algoritmaların kullanılması, parametre tahminlemesi sürecini daha hızlı hale getirerek, verimliliği artırır.

Beklenti Maksimizasyonu Yöntemi (Expectation Maximization, EM)

EM algoritması, verilerin belirli bir olasılık dağılımlarının karışımından üretildiği varsayımına dayanan ve model yapısı ile parametrelerin tahmin edilmesinde kullanılan bir yöntemdir. Temelinde maksimum olabilirlik tahminine dayanan bu algoritma, kapalı formda çözümün elde edilemediği durumlarda uygun bir alternatif sunmakta ve ilgili olasılık fonksiyonu için yaklaşık bir çözüm geliştirmektedir. Bu bağlamda, EM algoritması özellikle karma dağılımların parametrelerinin tahmin edilmesinde yaygın olarak tercih edilen bir yöntemdir.

3.4.3. Poisson Karma Dağılımı

İstatistiksel modelleme aracı olarak Sonlu Karma Modeli (Finite Mixture Model-FMM) çeşitli rastgele olay türlerinin analiz edilmesinde yaygın olarak kullanılmakta ve karmaşık olasılık dağılımlarını etkili bir şekilde temsil edebilmektedir.

Karma dağılım modelleri; veri setinin farklı popülasyonların karışımından türetildiği ve karışım içindeki her bir temel bileşenin dağılımının bilinmediği özellikteki veriler için uygundur.

Matematiksel olarak FMM'ler genellikle karışım içindeki bireysel bileşenlerden elde edilen olasılık yoğunluk fonksiyonlarının ağırlıklı toplamı olarak ifade edilir. Karma dağılımlarından rastgele bir değişken x elde edildiği varsayıldığında, olasılık yoğunluk fonksiyonu f aşağıdaki denklemde ifade edildiği gibi tanımlanabilir:

$$f(x;\theta) = \sum_{k=1}^{K} w_k g_k(x, \lambda_k) \quad (1)$$

Buna Poisson karma dağılımı denir ve burada θ = (w1, λ1, . . . , wk, λk) olur. Denklem (2) aşağıdaki gibidir:

$$g_k(x, \lambda_k) = \frac{\lambda_k^x e^{\lambda_k}}{x!} \quad \text{x = 0, 1, 2, . . . ve k = 1, . . . , K} \quad (2)$$

Burada $g_k(x, \lambda_k)$, λ_k parametreli Poisson yoğunluk fonksiyonunu temsil eder. Ayrıca; ağırlıklar w_k Denklem (3)'de verilen aşağıdaki koşulu sağlamalıdır:

$$\sum_{k=1}^{K} w_k = 1 \quad (3)$$

Poisson Karma dağılımı için EM Algoritması

FMM analizinde optimum bileşen sayısını belirlemek için genellikle dört temel metot kullanılır:

1. Momentler Yöntemi
2. Minimum Mesafe Yaklaşımı
3. Maksimum Olabilirlik Tahmini
4. Bayes Çıkarımı

Bu yöntemler arasında en sık kullanılanı maksimum olabilirlik tahminidir. Denklem (1)' de tanımlandığı üzere, bir karışım yoğunluğundan alınan bir gözlem kümesi (x_1,…, x_n) değerlendirildiğinde, Poisson karma dağılımı için olabilirlik fonksiyonu ve log-olabilirlik fonksiyonu sırasıyla Denklem(4) ve Denklem(5)'te gösterildiği gibi türetilebilir. Denklem(4) aşağıdaki gibidir:

$$L(X;\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \omega_k g_k(x_i; \lambda_i) \quad (4)$$

Burada L(X; θ), θ parametreleri verildiğinde gözlenen verilerin olasılığını temsil eder. w_k her bir bileşenle ilişkili karma oranlarını ve $g_k(x_i; \lambda_k)$, λ_k parametreli x'inci gözlem için Poisson yoğunluk fonksiyonunu ifade eder. Denklem (5) şu şekildedir:

$$logL(X;\theta)\ L(\theta) \quad \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \omega_k g(x_i; \lambda_k)) \quad (5)$$

Burada $L(\theta)$, gözlemlenen verilerin logaritmik olabilirliğini temsil eder ve her gözlem için ağırlıklı Poisson yoğunluklarının toplamının logaritması alınır. Bu dönüşüm, olabilirlik fonksiyonundaki çarpımı bir toplama dönüştürür. Özellikle θ parametrelerini tahmin etmek için maksimizasyon prosedürleri uygulanırken, matematiksel olarak işlemleri kolaylaştırdığı için etkilidir. Olabilirlik ve logaritmik olabilirlik fonksiyonları, Poisson karma dağılımlarının istatistiksel analizinde temel bileşenlerdir. Maksimum olabilirlik tahmini (Maximum Likelihood Estimation-MLE) gibi yöntemlerle parametre tahmini için temel oluştururlar.

Poisson karma dağılımları için MLE bağlamında $\hat{\theta}$ tahmini, log-olasılık denkleminin kök dizisini temsil eder ve matematiksel olarak şu şekilde ifade edilir:

$$\frac{L(\theta)}{\partial\theta}=0 \tag{6}$$

Denklem (6), logaritmik olasılık fonksiyonunun θ parametrelerine göre eğimi olan puan fonksiyonunun, maksimum olasılık tahminlerinde sıfıra eşit olması gerektiğini ifade eder. Denklem (6)'dan türetilen puan fonksiyonları aşağıdaki gibidir:

$$\frac{\partial \mathrm{L}(\theta)}{\partial \lambda_k} = \sum_{i=1}^{n} \frac{\omega_k g'(x_i;\lambda_k)}{\sum_{k=1}^{K} \omega_k g_k(x_i;\lambda_k)} \qquad \text{k= 1,….,K} \tag{7}$$

$g'(x_i, \lambda_k)$ Poisson yoğunluk fonksiyonunun λ_k parametresine göre türevini temsil eder. Denklem (8) aşağıdaki gibidir:

$$\frac{\partial \mathrm{L}(\theta)}{\partial \omega_k} = \sum_{i=1}^{n} \frac{g_k(x_i;\lambda_k) - g_k(x_i;\lambda_K)}{\sum_{i=1}^{K} \omega_k g_k(x_i;\lambda_k)} \tag{8}$$

Denklem (8), log-olasılığın karıştırma oranlarındaki (w_k) değişikliklere olan duyarlılığını yansıtır. k'inci bileşenin genel karma olasılığına katkısını, diğer bileşenlerin katkılarıyla karşılaştırır.

Poisson karma dağılımlarının maksimum olabilirlik tahmini (MLE) için kapalı formda bir çözüm mevcut değildir. Bu yüzden, karma dağılımının parametrelerini tahmin etmek için Dempster ve arkadaşları tarafından geliştirilen beklenti-maksimizasyon (EM) algoritması kullanılır. EM algoritması, istatistiksel modellerde parametrelerin maksimum olabilirliğini veya maksimum bir sonraki tahminlerini belirlemek için kullanılan yinelemeli bir hesaplama yöntemidir. Bu yöntem, eksik verilerin incelendiği durumlarda, eksik bilgilerin temsili ile modelin ve olasılık fonksiyonunun basitleştirildiği varsayım olarak kabul edilebilir. Poisson dağılımı için EM algoritmasının ayrıntıları çok sayıda araştırmacı tarafından gösterilmiştir.

(x_1,…, x_n) Denklem (1)'de tanımlanan Poisson karma dağılımından alınan bağımsız ve özdeş dağılımlı bir rastgele örneği temsil etsin. f(x; θ)'nin olasılık fonksiyonu Denklem (4) ile ifade edilir. Hesaplama kolaylığı açısından aşağıdaki log olasılık fonksiyonunu (bkz. Denklem 9) kullanmak daha pratiktir:

$$logL(x;\theta) = \mathrm{L}(\theta) = log\prod_{i=1}^{n}\sum_{k=1}^{K}\omega_k g_k(x_i,\lambda_k) = \sum_{i=1}^{n} log\{\sum_{k=1}^{K}\omega_k g_k(x_i;\lambda_k)\} \quad (9)$$

Poisson karışım modeli kapsamında eksik veriler şu şekilde tanımlanabilir:

z_{ik} = 1, eğer i'inci gözlem i'inci Poisson dağılımından ise i = 1, . . . , n ve j = 1, . . . , k

z_{ik} = 0, aksi takdirde

zi = $(z_{i1},\ldots,z_{ik})^T$

Eksik veriler z_{ik}' e dahil edilerek, artırılan veriler şu şekilde tanımlanabilir:

y= $y_1,\ldots,y_n$), burada y_i= (x_i, $z_i{}^T$). Artırılmış log-olasılık fonksiyonu daha sonra Denklem (10) ile ifade edilir.

$$logL_a(\theta,y) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\{\omega_k g_k(x_i;\lambda_k)\} \quad (10)$$

Eksik veriler z_{ik} gözlemlenemediğinden, λ'nın MLE'si için kapalı formda bir çözüm mevcut değildir. EM algoritması, en uygun olasılık tahminini doğrudan elde etmek yerine, Beklenti (E) ve Maksimizasyon (M) adında iki ana adım kullanır: E adımında tahmini bir ortalama sağlamak için log-olasılık fonksiyonunun beklentisi eksik verilere göre hesaplanır.

EM algoritmasının E adımında gözlemlenen verilere ve mevcut parametre tahminlerine bağlı olarak, artırılmış log-olasılık fonksiyonunun beklenen değerinin hesaplanması gerekir. Bu beklenti aşağıdaki şekilde ifade edilir. E adımında, aşağıdaki Denklem (11)'in hesaplanması gerekir.

$$Q(\theta|\theta^{(r)}) = E\{logL_a(\theta;y)|x;\theta^{(r)}\} \quad (11)$$

Burada $\theta^{(r)}$ r'inci yinelemedeki parametre tahminlerini temsil eder. Daha sonra, eksik veri z_{ik} 'nin koşullu beklentisi olan hesaplanır. $logL_a(\theta;y)$'nin verileri eksik veri z_{ik}'nin doğrusal bir

fonksiyonu olduğundan beklenen değer $logL_a(\theta; y)$'nin yerine kullanılır. Beklenen bu değerler denklem (12)' de verilmiştir:

$$p_{ik}^{(r+1)} = E\left(z_{ik}\middle|x_i, \theta^{(r)}\right) = P\left(z_{ik} = 1\middle|x_i, \theta^{(r)}\right) = \frac{\omega_k^{(r)} g_k(x_i;\lambda_k^{(r)})}{\sum_{v=1}^{K} \omega_v g_v(x_i;\lambda_v^{(r)})} \quad (12)$$

Burada $p_{ik}{}^{(r+1)}$ i'inci gözlemin k'inci Poisson bileşenine ait olma olasılığını, mevcut parametre tahminleri $\theta^{(r)}$'ye göre temsil eder. Bu olasılıklar, parametre tahminlerini güncellemek için M adımında kullanılır ve algoritmanın yinelemeli yakınsamasını sağlar.

Poisson karma dağılımına uygulanan EM algoritmasının adımları şu şekildedir:

i. Başlangıç parametre değerleriyle başlanır $\hat{\lambda}_k{}^{(r)}$, $\hat{w}_k{}^{(r)}$(r=0), k=1,…,K ve yakınsamaya kadar aşağıdaki adımlar tekrarlanır.

ii. E-adım: Gözlem x_i'nin k'inci Poisson dağılımından gelme sınıflandırma olasılığı aşağıdaki tahmine göre bulunur:

$$p_{ik}^{(r+1)} = \frac{\omega_k^{(r)} g_k(x_i;\lambda_k^{(r)})}{\sum_{v=1}^{K} \omega_v g_v(x_i;\lambda_v^{(r)})} \quad (13)$$

Burada; i=1,…*n ve k* = 1,…,*K*

iii. M-adım: Poisson karma dağılım parametrelerini aşağıdaki gibi güncellenir:

$$\lambda_k^{(r+1)} = argmax \sum_{i=1}^{n} p_{ik}^{(r+1)} log\, g_k(x_i|\lambda_k) \quad (14)$$

$$\omega_k^{(r+1)} = \frac{\sum_{i=1}^{n} p_{ik}^{(r+1)}}{n} \quad (15)$$

EM algoritmasının E adımında, Poisson karma dağılımının olasılık yoğunluk fonksiyonu (Denklem (1)'de tanımlandığı gibi) $\hat{p}_{ik}$ değerini hesaplamak için Denklem (13)'ün yerine kullanılır. Elde edilen ifade Denklem (16)'da verilmiştir.

$$\hat{p}_{ik} = \frac{\hat{\omega}_k(\frac{e^{-\widehat{\lambda}_k}\hat{\lambda}_k^{x_i}}{x_i!})}{\sum_{v=1}^{K} \hat{\omega}_k(\frac{e^{-\widehat{\lambda}_v}\hat{\lambda}_v^{x_i}}{x_i!})} \quad (16)$$

Burada $\hat{p}_{ik}$, mevcut parametre tahminleri göz önüne alındığında i'inci gözlemin k'inci Poisson bileşenine ait olma olasılığını temsil eder.

EM algoritmasının M adımında, Denklemler (14) ve (15), kısıtlı bir optimizasyon problemini temsil eder ve Denklemler (17) ve (18) şeklinde ifade edilebilir. Bu optimizasyon problemi, Lagrange çarpanı (δ) kullanılarak Denklem (19)'a dönüştürülebilir.

$$\hat{\lambda}_k, \hat{\omega}_k = argmaxQ(\lambda_1, \dots, \lambda_k, \omega_1, \dots, \omega_K) \tag{17}$$

$$\sum_{k=1}^{K} \omega_k = 1 \tag{18}$$

$$\mathcal{L}(\lambda_1, \dots, \lambda_k, \omega_1, \dots, \omega_K, \delta) = Q(\lambda_1, \dots, \lambda_k, \omega_1, \dots, \omega_K) - \delta(\sum_{k=1}^{K} \omega_K) \tag{19}$$

Poisson karma dağılımının olasılık fonksiyonuna (Denklem (1)'den) olasılık ve log-olasılık fonksiyonları (Denklem (9)–(11)'de tanımlandığı gibi) dahil edilerek, Lagrangian, Denklem (20) ve (21)'de gösterildiği gibi açıkça ifade edilebilir.

$$\mathcal{L}(\lambda_1, \dots, \lambda_k, \omega_1, \dots, \omega_K, \delta) = \sum_{i=1}^{n} \sum_{k=1}^{K} [\hat{p}_{ik} log\omega_k + \hat{p}_{ik} log g_k(x_i; \lambda_k)] - \alpha(\sum_{k=1}^{K} \omega_k - 1) \tag{20}$$

$$\mathrm{L}(\lambda_1, \dots, \lambda_k, \omega_1, \dots, \omega_K, \delta) = \sum_{i=1}^{n} \sum_{k=1}^{K} [\hat{p}_{ik} log\omega_k + \hat{p}_{ik}(-\lambda_k + x_i log\lambda_k - logx_i!)] - \delta(\sum_{k=1}^{K} \omega_k - 1) \tag{21}$$

Denklem (21)'deki Lagrangian fonksiyonunun λ_k'ye göre kısmi türevi alınıp sıfıra eşitlendiğinde, Denklem (22) ve (23)'te gösterildiği gibi, Denklem (24)'te verilen $\hat{\lambda}_k$ değeri elde edilir.

$$\frac{\partial L}{\lambda_k} = \sum_{i=1}^{n} \hat{p}_{ik} \frac{\partial}{\lambda_k} (log\omega_k + x_i log\lambda_k - logx_i! - \lambda_k) \tag{22}$$

$$\frac{\partial L}{\lambda_k} = \sum_{i=1}^{n} \left[\hat{p}_{ik} \left(\frac{x_i}{\lambda_k} - 1\right)\right] = 0 \tag{23}$$

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} \hat{p}_{ik} x_i}{\sum_{i=1}^{n} \hat{p}_{ik}} \tag{24}$$

Aynı işlemleri yapıp ω_K ve $\propto_K$' ya göre kısmi türevleri alıp bunları sıfıra eşitleyerek ω_K için bir tahmin elde edilebilir.

$$\frac{\partial L}{\partial \omega_k} = \sum_{i=1}^{n} \hat{p}_{ik} \frac{\partial}{\omega_k}[(\log\omega_k + x_i log\lambda_k - log x_i! - \lambda_k)] - \delta(\sum_{k=1}^{K} \omega_k - 1)] \quad (25)$$

$$\frac{\partial L}{\partial \omega_k} = \sum_{i=1}^{n} \frac{\hat{p}_{ik}}{\omega_k} - \delta = 0 \; buda \; verir \; \; \widehat{\omega}_k = \frac{1}{\delta}\sum_{i=1}^{n} \hat{p}_{ik} \quad (26)$$

$$\frac{\partial L}{\partial \delta} = \sum_{k=1}^{K} \omega_k - 1 = 0 \quad (27)$$

Denklem (26)'nın Denklem (27)'de uygulanması ile, Denklem (28)'de gösterildiği gibi Lagrange çarpanı δ elde edilir.

$$\sum_{k=1}^{K} \frac{1}{\delta}\sum_{i=1}^{n} p_{ik} = 1 \Rightarrow \delta = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} \quad (28)$$

Böylece $\widehat{\omega}_k$ tahmini Denklem (29)'da gösterildiği gibi elde edilir.

$$\widehat{\omega}_k = \frac{\sum_{i=1}^{n} p_{ik}}{\sum_{i=1}^{n}\sum_{v=1}^{K} p_{iv}} \quad (29)$$

Poisson Karma dağılımının Ortalaması ve Varyansı

Önceki bölümde tanımlanan parametrelere sahip bir Poisson karma dağılımının ortalaması λ_{mix} o olarak gösterilirse, Denklem (30) kullanılarak hesaplanabilir.

$$\lambda_{mix} = \sum_{k=1}^{K} \omega_k \lambda_k \quad (30)$$

Herhangi bir dağılımın varyansı Denklem (31) kullanılarak hesaplanabilir.

$$Var(X) = E[X^2] - (E[X])^2 \quad (31)$$

Burada $E[X^2]$ rastgele değişkenin karelerinin beklenen değerini temsil ederken, $E[X]$ beklenen değerin kendisini ifade eder. Poisson karma dağılımı için, Denklem (31)'deki varyans formülü, Denklem (32)'de gösterildiği gibi parametrelerine uyarlanır.

$$\mathrm{Var(X)} = \mathrm{E}[X^2|\omega_k, \lambda_k] - (\mathrm{E[X}|\omega_k, \lambda_k])^2 \quad (32)$$

Cebirsel işlemlerden sonra, Denklem (33)'te ifade edildiği gibi Poisson dağılımının varyansı elde edilir.

$$\sigma_{mix}^2 = \sum_{k=1}^{K} \omega_k(\sigma_k^2 + \lambda_k^2) - \lambda_{mix}^2 \quad (33)$$

Poisson dağılımının ortalamasının varyansına eşit olması önemli bir özelliğidir. Bu nedenle önceki denklemde σ_k^2 ' λ_k ile değiştirilir ve Denklem (34)'teki şu ifade elde edilir:

$$\sigma_{mix}^2 = \sum_{k=1}^{K} \omega_k\,(\lambda_k + \lambda_k^2) - \lambda_{mix}^2 \quad (34)$$

Denklem (34), Poisson karma dağılımının varyansını hesaplamak için gereken son ifadeyi sunar.

3.4.4. Gamma Karma dağılımı

Gamma karma dağılımının olasılık yoğunluk fonksiyonu (PDF), Denklem(37)'de verildiği gibi ifade edilir.

$$f(x;\Theta) = \sum_{k=1}^{K} \pi_k g_k\,(x, \alpha_k, \beta_k) \quad (37)$$

Burada $\Theta = \{(\theta_1, \theta_2, \ldots, \theta_K,)\}$ karma dağılımının parametre uzayını temsil eder. Her bileşen $\{\theta_K,) = (\pi_k, \alpha_k, \beta_k)\}$ her k = 1, . . . , K için. α_k, k'ıncı bileşenin şekil parametresini belirtir. β_k k'ıncı bileşenin ölçek parametresini belirtir. π_k, k'ıncı bileşenin karışım oranını belirtir ve tüm k ve $\sum_{k=1}^{K} \pi_k = 1$ $\pi_k \geq 0$' ı sağlar.

Ek olarak, bu formülasyonda $g_k(x, \alpha_k, \beta_k)$ her bir gamma bileşenini , ilgili oran π_k ile ağırlıklandırarak Denklem (38)' de gösterildiği gibi ifade eder.

$$g(x, \alpha_k, \beta_k) = \frac{\beta_k^{\alpha_k} x^{\alpha_k - 1} e^{-\beta_k x}}{\Gamma(\alpha_k)} \qquad x \geq 0 \; ve \; k = 1, \ldots, K \quad (38)$$

Burada $\Gamma(\alpha_k)$ gamma fonksiyonudur ve Denklem (39)' da ifade edilmiştir.

$$\Gamma(\alpha_k) = \int_0^{\infty} t^{\alpha_k - 1} e^{-t} dt \qquad k = 1, \ldots, K \quad (39)$$

Gamma Karma Dağılımı için EM Algoritması

Gamma karma dağılımını izlediği varsayılan bir gözlem kümesi için, olasılık ve logaritmik olasılık fonksiyonları, Denklem (36)'da verilen yoğunluk fonksiyonu kullanılarak ifade edilebilir. Bunlar Denklem (40) ve (41)'de ifade edilmiştir.

$$L(X;\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k g_k(x_i;\alpha_k,\beta_k) \tag{40}$$

$$logL(X;\theta) = \mathrm{L}(\theta) = \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \pi_k g_k(x_i,\alpha_k,\beta_k)) \tag{41}$$

Maksimum olabilirlik tahmini (MLE) ile $\hat{\theta}$ tahmini Denklem (6)'da verilen logaritmik olabilirlik denkleminin köklerinden oluşan bir diziyi tanımlar. MLE için kapalı formda bir çözümün bulunmaması nedeniyle, gamma karma dağılımının parametreleri EM algoritması kullanılarak tahmin edilmektedir. Gamma karma dağılımları için EM algoritması çeşitli araştırmacılar tarafından ayrıntılı olarak açıklanmıştır.

Eksik veriler şu şekilde tanımlanabilir:

z_{ik} = 1, eğer i'inci gözlem k'inci Gamma dağılımından ise i = 1, . . . , n ve k = 1, . . . , K

z_{ik} = 0, aksi takdirde

$$z_i = (z_{i1}, \dots, z_{ik})^T \tag{42}$$

y=($y_1, \dots, y_n$) artırılmış veri olduğunu varsayalım. Burada, y_i=(x_i, z_i^T). Daha sonra artırılmış log-olasılık fonksiyonu Denklem(43)' de ifade edildiği gibi tanımlanabilir.

$$logL_a(\theta,y) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log\{\pi_k g_k(x_i,\alpha_k,\beta_k)\} \tag{43}$$

Eksik veriler z_{ik} gözlemlenebilir olmadığından θ için MLE'nin kapalı formda bir çözümü yoktur. EM algoritması, en uygun olasılık tahminini doğrudan belirlemek yerine beklenti (E) adımı ve maksimizasyon (M) adımı olmak üzere iki temel adım üzerinden çalışır. E-adımında, eksik verilerle ilgili olarak log-olasılık fonksiyonunun beklenen değeri hesaplanır ve ortalamanın tahmini elde edilir. E-adımında, daha önce Poisson karma dağılımı bağlamında gösterildiği gibi, Denklem (11)'i kullanarak $\left(\theta \middle| \theta^{(r)}\right)$'i hesaplamak gereklidir. Daha sonra, eksik verinin koşullu beklentisi olan z_{ik} hesaplanır. log$l_a(\theta,y)$'nin eksik veri z_{ik}'nin doğrusal bir fonksiyonu olduğu göz önüne alındığında, beklenen değer log$l_a(\theta,y)$'nin yerine kullanılır. Bu beklenen değerler Denklem (44)'de verilmiştir.

$$p_{ik}^{(r+1)} = E\left(z_{ik}\middle|x_{i,}\theta^{(r)}\right) = P\left(z_{ik}=1\middle|x_{i,}\,\theta^{(r)}\right) = \frac{\pi_k^{(r)} g_k(x_i;\,\alpha_k^{(r)},\beta_k^{(r)})}{\sum_{v=1}^{K}\pi_v g_v(x_i;\alpha_v^{(r)},\beta_v^{(r)})} \quad (44)$$

Burada, $p_{ik}^{(r+1)}$ mevcut parametre tahminleri $\theta^{(r)}$'ya dayanarak, i'inci gözlemin k'ıncı gamma bileşenine ait olma olasılığını temsil eder. Bu olasılıklar daha sonra, Poisson karışım modelindekine benzer bir süreci izleyerek parametre tahminlerini güncellemek için M adımında kullanılır.

Gamma karma dağılımına uygulanan EM algoritmasının adımları şu şekildedir:

i. (r = 0) k = 1, . . . , K başlangıç parametre değerleriyle başlanır ve yakınsamaya kadar aşağıdaki adımları tekrarlanır

ii. E-adım: Gözlem x_i'nin k^{th}ıncı gamma dağılımından gelme sınıflandırma olasılığı aşağıdaki tahmine göre bulunur:

$$p_{ik}^{(r+1)} = \frac{\pi_k^{(r)} g_k(x_i;\,\alpha_k^{(r)},\beta_k^{(r)})}{\sum_{v=1}^{K}\pi_v^{(r)} g_v(x_i;\alpha_v^{(r)},\beta_v^{(r)})} \quad (45)$$

Burada i = 1, . . . n ve k = 1, . . . , K

iii. M-adım: Gamma karma dağılım parametreleri aşağıdaki gibi güncellenir:

$$\beta_k^{(r+1)} = \frac{\alpha_k^{(r)}\sum_{i=1}^{n}|p_{ik}^{(r)}}{\sum_{i=1}^{n} x_i p_{ik}^{(r)}} \quad (46)$$

$$\alpha_k^{(r+1)} = \alpha_k^{(r)} + \varepsilon_m G_{\alpha_k}(X,\theta^{(r)}) \quad (47)$$

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^{n} p_{ik}^{(r)}}{n} \quad (48)$$

EM algoritmasının E adımında, gamma karma dağılımının olasılık yoğunluk fonksiyonu (Denklem (37)'te tanımlandığı gibi), Denklem (45)'in yerine kullanılarak, mevcut parametre tahminleri verildiğinde, i'inci gözlemin k'ıncı gamma bileşenine ait olma olasılığını temsil eden $\hat{p}_{ik}$ değeri hesaplanır.

EM algoritmasının M adımında, Denklemler (46)–(48), kısıtlı bir optimizasyon problemini temsil eder ve Denklem (49) ve (50) şeklinde ifade edilebilir.

$$\hat{\theta}_k, \hat{\pi}_k = argmaxQ(\theta_1, \dots, \theta_k, \pi_1, \dots, \pi_K) \quad (49)$$

$$\sum_{k=1}^{K} \pi_k = 1 \quad (50)$$

Bu optimizasyon problemi, Lagrange çarpanı (δ) kullanılarak Denklem (51)'a dönüştürülebilir.

$$L(\theta_1, \dots, \theta_k, \pi_1, \dots, \pi_k, \delta) = Q(\theta_1, \dots, \theta_K, \pi_1, \dots, \pi_k) - \delta(\sum_{k=1}^{K} \pi_k) \quad (51)$$

Olasılık ve log-olasılık fonksiyonlarının (Denklem (40) ve (41)'da tanımlandığı gibi) gamma karma dağılımının olasılık fonksiyonuna (Denklem (37)'den) dahil edilmesiyle, Lagrangian, Denklem (52) ve (53)'de verildiği gibi açıkça ifade edilebilir.

$$L(\theta_1, \dots, \theta_k, \pi_1, \dots, \pi_k, \delta) = \sum_{i=1}^{n} \sum_{k=1}^{K} [\hat{p}_{ik} log\pi_k + \hat{p}_{ik} log g_k(x_i; \alpha_k, \beta_k)] - \delta(\sum_{k=1}^{K} \pi_k - 1) \quad (52)$$

$$L(\theta_1, \dots, \theta_k, \pi_1, \dots, \pi_k, \delta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\hat{p}_{ik} log\pi_k + \hat{p}_{ik}((\alpha_k - 1) log x_i - \frac{x_i}{\beta_k} - \alpha_k \log(\beta_k) - \log(\Gamma(\alpha_k)))\right] - \delta(\sum_{k=1}^{K} \pi_k - 1) \quad (53)$$

Denklem (53)'deki Lagrangian fonksiyonunun π_k'ya göre kısmi türevi alınıp sıfıra eşitlendiğinde, aşağıdaki ilişki Denklem (54) türetilir.

$$\frac{\partial \mathrm{L}}{\partial \pi_k} = \sum_{i=1}^{n} \hat{p}_{ik} - \delta\pi_k = 0 \quad (54)$$

Her iki tarafın k üzerinden toplanması Denklem (55)'i verir.

$$\delta = -n \quad (55)$$

Bu sonuç denkleme geri koyup sadeleştirilerek, Denklem (56)' da verilen $\hat{\pi}_k$ için nihai hedefe ulaşılır.

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{ik} \quad (56)$$

Bu ifade, beklenti adımı sırasında elde edilen arka olasılıklar $\hat{p}_{ik}$'ye dayalı karışım oranı $\hat{\pi}_k$ için maksimum olabilirlik tahminini temsil eder. Lagrangian fonksiyonunun β_k'ya göre kısmi türevini alıp sıfıra eşitlediğimizde Denklem (57) elde edilir.

$$\frac{\partial \mathrm{L}}{\beta_k} = \sum_{i=1}^{n} \hat{p}_{ik} \left(\frac{x_i}{\beta_k^2} + \frac{\hat{p}_{ik}}{\alpha_k} \right) = 0 \tag{57}$$

Daha sonra β_k için çözüm yapılarak güncellenmiş parametre tahmini Denklem (58)'de ifade edildiği gibi elde edilir.

$$\beta_k = \frac{1}{\alpha_k} \frac{\sum_{i=1}^{n} \hat{p}_{ik} x_i}{\sum_{i=1}^{n} \hat{p}_{ik}} \tag{58}$$

Denklem (58), β_k'nın maksimum olabilirlik tahminini, $\hat{p}_{ik}$ arka olasılıkları, gözlemlenen veriler x_i ve şekil parametresi α_k cinsinden ifade eder. Lagrangian fonksiyonunun α_k'ya göre kısmi türevini alıp sıfıra eşitlediğimizde Denklem (59) elde edilir.

$$\frac{\partial L}{\partial \alpha_k} = \sum_{i=1}^{n} \hat{p}_{ik} \log(x_i) - \sum_{i=1}^{n} \hat{p}_{ik} \log \left(\frac{\sum_{i=1}^{n} \hat{p}_{ik} x_i}{\sum_{i=1}^{n} \hat{p}_{ik}} \right) + \sum_{i=1}^{n} \hat{p}_{ik} \log(\alpha_k) - \sum_{i=1}^{n} \hat{p}_{ik} \psi(\alpha_k) = 0 \tag{59}$$

Denklem (59)'un kapalı formda bir ifadesi olmadığından, α_k için en uygun güncelleme şeması mevcut değildir. Bu sınırlamayı gidermek için, EM algoritmasının yinelemeli ve gradyan tabanlı yapısı kullanılır. Her yinelemede, parametre güncellemeleri, sistem parametrelerine göre log-olasılık fonksiyonunun gradyanına pozitif bir projeksiyon sağlar. Buna göre, $\alpha_k{}'$nin güncellenmesi, ε_k ($\varepsilon_k \in (0, 1)$) adım boyutu kullanılarak gradyan yönünde gerçekleştirilebilir. Bu, Denklem (60) ve (61)'da tanımlandığı gibi $\alpha_k{}'$ için yinelemeli bir güncelleme şeması oluşturur ve bu da olasılık fonksiyonunun yerel maksimumuna doğru yakınsamasını sağlar.

$$\alpha_k^{(r+1)} = \alpha_k^{(r)} + \varepsilon_k G_{\alpha_k}(X, \theta^{(r)}) \tag{60}$$

$$G_{\alpha_k}\left(X, \theta^{(r)}\right) = \frac{\partial Q(X, \theta^{(r)})}{n \partial \alpha_k} = \frac{1}{n} \sum_{i=1}^{n} [\log(x_i) + \log\left(\beta_k^{(r)}\right) - \psi(\alpha_k^{(r)})] p_{ik}^{(r+1)} \tag{61}$$

Bu denklem, Denklem (62)'da tanımlanan digamma fonksiyonu $\psi(x)$'i içerir.

$$\psi(x) = \Gamma'(x)/\Gamma(x) \tag{62}$$

Bu fonksiyon, gama karma modeli için EM algoritmasında k'inci gama dağılımının şekil parametresi $\alpha_k{}'$nin tahmininde önemli rol oynar.

Denklem (62)'nin kapalı formda bir çözümü olmadığından, Denklem (63)'de verilen digamma fonksiyonu için, özellikle x ≥ 2 için iyi bir yaklaşım kullanılabilir.

$$\psi(x) \approx G(x) = \log\left(x - {}^{1}/_{2}\right) + \frac{1}{24\left(x-{}^{1}/_{2}\right)^{2}} \tag{63}$$

Denklem (63)'de sunulan yaklaşım kullanılarak, digamma fonksiyonu Denklem (64)'de gösterildiği gibi α_k'nin bir fonksiyonu olarak ifade edilebilir.

$$\psi\left(\alpha_k^{(r)}\right) \approx \log\left(\alpha_k^{(r)} - {}^{1}/_{2}\right) + \frac{1}{24\left(x-{}^{1}/_{2}\right)^{2}} \tag{64}$$

Gamma Karma Dağılımının Ortalaması ve Varyansı

Her bir bileşenin, karışım oranı π_k (k = 1, . . . , K) olan bir Gamma(α_k, β_k) dağılımını izlediği bir gama karma dağılımı bağlamında, ortalama $(\mu)_{mg}$ ve varyans(σ_{mg}^2) Denklem (65) ile gösterilebilir.

$$\mu_{mg} = \sum_{k=1}^{K} \pi_k(\alpha_k\beta_k) \tag{65}$$

Denklem (31)'deki varyans tanımı kullanılarak, gama dağılımları için varyans (σ_{mg}^2) Denklem (66) ile ifade edilebilir.

$$\sigma_{mg}^2 = \sum_{k=1}^{K} \pi_k(\alpha_k^2\,\beta_k + \alpha_k^2\beta_k^2) - \mu_{mg}^2 \tag{66}$$

Alternatif olarak, Denklem (67)'te gösterildiği gibi başka bir biçimde de sunulabilir.

$$\sigma_{mg}^2 = \sum_{k=1}^{K} \pi_k(\beta_k + 1)(\alpha_k^2\beta_k) - (\mu_{mg})^2 \tag{67}$$

Bu tez çalışmasında, farklı parametre yapılandırmalarına sahip çeşitli aday modelleri değerlendirmek için yinelemeli bir yaklaşım kullanılmıştır. Seçim süreci, AIC ve BIC değerlerinin

karşılaştırılmasını içerir ve en küçük kriter değerini veren model, verilen veri kümesini temsil etmek için en uygun seçenek olarak kabul edilmiştir.

Model Değerlendirmesi

Optimal modelin belirlenmesi için, model seçiminde yaygın olarak kullanılan istatistiksel kriterler olan Akaike Bilgi Kriteri (AIC), Bayesci Bilgi Kriteri (BIC), Logaritmik Olabilirlik (Log-likelihood, LL) kullanılır. Bu kriterlerin formüle edilmiş hali denklem (35) ve denklem (36) da ifade edilmiştir:

$$AIC = 2k - 2\ln(L) \tag{35}$$

$$BIC = \ln(n)k - 2\ln(L) \tag{36}$$

Burada k parametre sayısını, L olasılık fonksiyonunu ve n ise örneklem büyüklüğünü ifade eder.

Hem AIC hem de BIC, parametre sayısını dikkate alarak model karmaşıklığını hesaba dahil eden ceza terimleri içerir. Ancak BIC, AIC'ye kıyasla daha büyük bir ceza uygular ve bu da onu aşırı karmaşıklığa sahip modelleri seçerken daha tutucu hale getirir. AIC, verilere daha iyi uyum sağlayan modelleri tercih etme eğiliminde olsa da, özellikle büyük veri kümeleriyle çalışırken aşırı uyuma yol açabilir. Diğer taraftan, BIC daha güçlü bir ceza puanı ekleyerek bu riski azaltır. Böylece BIC, model karmaşıklığı ile uyum iyiliği arasında bir dengenin önemli olduğu durumlar için daha uygun hale getirilir.

Model seçiminde alternatifler arasından en uygun olanının seçilmesinde düşük AIC ve BIC değerlerine sahip olan model tercih edilir. Düşük AIC ve BIC değeri, daha iyi uyum sağlayan ve daha az karmaşık olan bir modeli temsil eder.

AIC, bir veri kümesi için birden fazla model varsa, her modelin kalitesini diğer modellerin kalitesiyle karşılaştırarak tahmin eder. Bu nedenle, AIC model seçimi için bir çözüm sunar. Bilgi teorisine dayalı olan AIC, modelin veriyle olan uyumuna ilişkin bilgileri temsil etmekte ve modelin uygunluğu ile karmaşıklığı arasındaki dengeyi değerlendirmek amacıyla kullanılmaktadır. Bununla birlikte, AIC mutlak anlamda bir modelin kalitesini ölçmez; yalnızca alternatif modellerle karşılaştırmalı olarak göreli bir değerlendirme sunar.

BIC, modeldeki parametre sayısını azaltarak aşırı uyuma karşı bir çözüm sunar (Güleryüz, 2023).

LL, istatistiksel modellemede parametre tahminlerinin değerlendirilmesinde temel bir ölçüt olarak kullanılan olabilirlik fonksiyonunun doğal logaritmasının alınmasıyla elde edilir. Olasılık değerlerinin genellikle çok küçük sayılarla ifade edilmesi nedeniyle, logaritmik dönüşüm hesaplamaları kolaylaştırmakta ve sayısal kararlılığı artırmaktadır. Ayrıca LL, maksimum

olabilirlik tahmininde (MLE) parametrelerin en uygun değerlerinin belirlenmesine imkân tanır. Bu bağlamda, LL değeri ne kadar büyükse, modelin gözlemlenen veriyi açıklama gücü de o ölçüde yüksek kabul edilmektedir.

Model seçiminde yalnızca LL incelenirse, daha fazla parametre içeren karmaşık modeller avantajlı görünür. Bu nedenle AIC ve BIC gibi ölçütler, LL' yi model karmaşıklığı ile değerlendirerek en uygun modelin seçimine olanak sağlar.

4. BULGULAR VE TARTIŞMA

Tez kapsamında verinin, BV'nin tanımlamaları yapılarak; BV kullanım alanları, sağlık sisteminde BV kullanım alanları, BV analitiği türleri, BV'nin ML algoritmalarını beslemesi ile ML algoritmaları ve algoritmaların kullanım alanları, denetimsiz öğrenme yöntemlerinden kümeleme, karma dağılımlar gibi konular üzerinde durulmuştur. Teorik açıklamaların üzerine uygulama, denetimsiz öğrenme yöntemlerinden olan karma dağılımlar ile yapılmıştır.

Uygulama yapılırken; R ve Statistica programlarından yararlanılmıştır. Sağlık sektöründeki bir kurumun acil servisini 2023 (Nisan- Aralık) ve 2024 (Ocak-Ekim) tarihleri arasında ziyaret eden kişilere ait, Kaggle platformu üzerinden kamu verisi olarak paylaşılan bir veri seti kullanılmıştır. Bu veri seti 9217 satır ve 12 sütundan oluşmaktadır. İçerisinde hasta kabul tarihi, hasta cinsiyeti, hastaların acil servise geldikten sonra sevk edildikleri branşlar, hastaların hastanede bekleme süreleri, hastaların hizmet memnuniyet puanları gibi bilgiler barındırmaktadır. (EK-1)

Veri setinde yer alan hastaların acil serviste bekleme süreleri ilk olarak Statistica programına aktarılmıştır. Bu programda hastaların bekleme sürelerinin histogram grafiği oluşturulup, üzerinde olasılık dağılım uygunluğu testleri yapılmıştır. Sürekli olasılık dağılımlarından Normal, Gamma, Lognormal, Üstel; kesikli olasılık dağılımlarından Poisson dağılımının veri ile uyumu Çizelge 4.1' de sonuçları gösterilen Ki-Kare testleri ve dağılım uygunluk analizi görselleri ile incelenmiştir. Elde edilen sonuçlar; bekleme sürelerinin genellikle tek bir istatistiksel dağılım ile temsil edilemeyeceğine, karma dağılımların kullanılması gerektiğine dair güçlü kanıtlar ortaya çıkarmıştır. Bu ilk izlenimden sonra veriler, karma dağılımların oluşturulması için R Studio ortamına aktarılmıştır.

Veri seti Excel programından R Studio ortamına aktarıldıktan sonra, veriye belirli ön işleme adımları uygulanmıştır. Öncelikle veri setindeki sütun isimleri incelenmiş ve her bir sütunda bulunan eksik değerlerin sayısı tespit edilmiştir. Daha sonra, eksik değerler (NA) ilgili sütunlardan çıkarılarak, her sütun eksiksiz verilerden oluşan bir listeye dönüştürülmüştür. Analiz kapsamında çalışılacak sütun, bu listeden seçilmiş ve x değişkenine atanarak vektör veri yapısına dönüştürülmüştür; böylece x, eksik değerlerden arındırılmış bir veri vektörü olarak elde edilmiştir.

Veri seti üzerinde, hastaların bekleme süreleri için makine öğrenmesi tekniklerinden denetimsiz öğrenme grubuma giren karma dağılımlı kümeleme analizi yapılmıştır. R Studio ortamında yazılan kodlar ile; Normal, Weibull, Gamma, Lognormal, Poisson dağılımları ayrı ayrı incelenmiş, her bir dağılım için en uygun bileşen sayısı belirlenmiştir. Daha sonra, belirlenen bileşen sayısı kullanılarak, seçilen dağılım üzerinde karma dağılım modeli tahmin edilmiştir. Modelin çıktısı bir görsel grafikle desteklenmiş ve bu sayede verinin hangi karma dağılım yapısına daha uygun olduğu görsel ve sayısal olarak değerlendirilmiştir.

Veri setinin hikayesine göre, hastalar bir hastanenin acil servisine geldikten sonra hastalıklarına göre ilgili uzmanın bulunduğu branşlara yönlendirilmektedirler. Hastaların hastanede

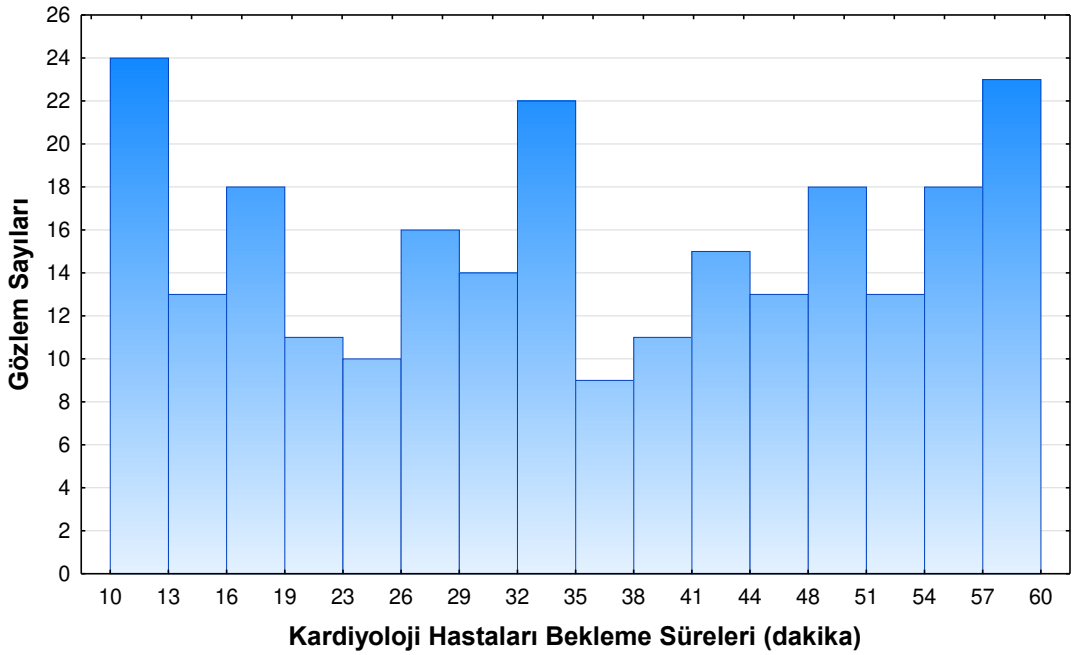
bekleme süreleri ve hastalık dışı durumlar olarak kabul edilen kişilere (tedaviyi reddeden, refakatçi vb.) ait bekleme süreleri için veriyi temsil eden istatistiksel dağılımlar belirlenmiştir. Veri setinde; bölüm yönlendirmesi (department referral) sütununda 8 grup yer almaktadır. Bunlar; kardiyoloji (cardiology), gastroenteroloji (gastroenterology), genel pratisyenlik (general practice), nöroloji (neurology), hastalık dışı (none), ortopedi (orthopedics), fizyoterapi (physiotherapy), böbrek (renal) gruplarıdır.

Bu tez çalışmasında her bir gözlem grubuna ait kişilerin bekleme süreleri ayrı ayrı incelenmiştir.

Çizelge 4.1. Ki-kare ($\chi 2$) Uyum İyiliği Testi p-değerleri

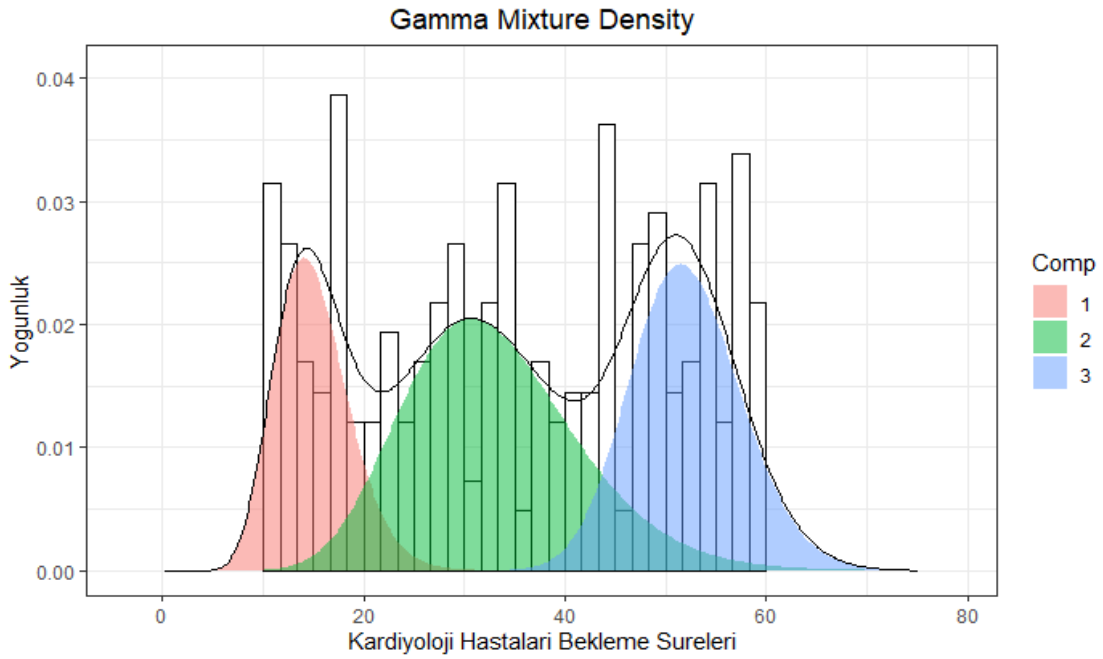
Veri/Dağılım	**Normal**	**Gamma**	**Lognormal**	**Üstel**	**Poisson**
Kardiyoloji B.	0,00	0,00	0,00	0,00	0,00
Gastroenteroloji B.	0,00	0,00	0,00	0,00	0,00
Genel Pratisyenlik B.	0,00	0,00	0,00	0,00	0,00
Nöroloji B.	0,00	0,00	0,00	0,00	0,00
Ortopedi B.	0,00	0,00	0,00	0,00	0,00
Fizyoterapi B.	0,00	0,00	0,00	0,00	0,00
Böbrek Bölümü	0,06655	0,00003	0,00	0,00	0,00
Hastalık Dışı B.	0,00	0,00	0,00	0,00	0,00

Hastaların bekleme sürelerinin saf dağılıma uygunluğu Ki-Kare testi ile incelenmiş ve sonuçlar Çizelge 4.1‘ de ifade edilmiştir. Bu bilgiye göre; böbrek hastalığı dışındaki grupların Ki-Kare test sonuçlarında $P < 0{,}05$ elde edilmiştir. Bu sonuç; gözlemlerin, teorik olasılık dağılımından gelen değerler ile örtüştüğü ve teorik olasılık dağılımı ile temsil edilebileceği hipotezinin reddedilmesi gerektiğini ifade etmektedir. Diğer bir deyişle; saf dağılımlar veriyi yeterince açıklayamamaktadır. Böbrek hastalarının acil serviste bekleme sürelerinin ise Normal dağılıma uygunluğu için yapılan Ki-Kare test sonucunda $P > 0{,}05$ elde edilmiştir. Bu sonuç; verinin teorik dağılım olan Normal dağılım ile temsil edilebileceği hipotezinin reddedilememesini ifade etmektedir. Diğer bir deyişle; Normal dağılım veriyi açıklayabilmektedir. Birden fazla bileşenli karma dağılım kullanılmasına gerek yoktur.



Şekil 4.1. Kardiyoloji Hastaları Bekleme Süreleri

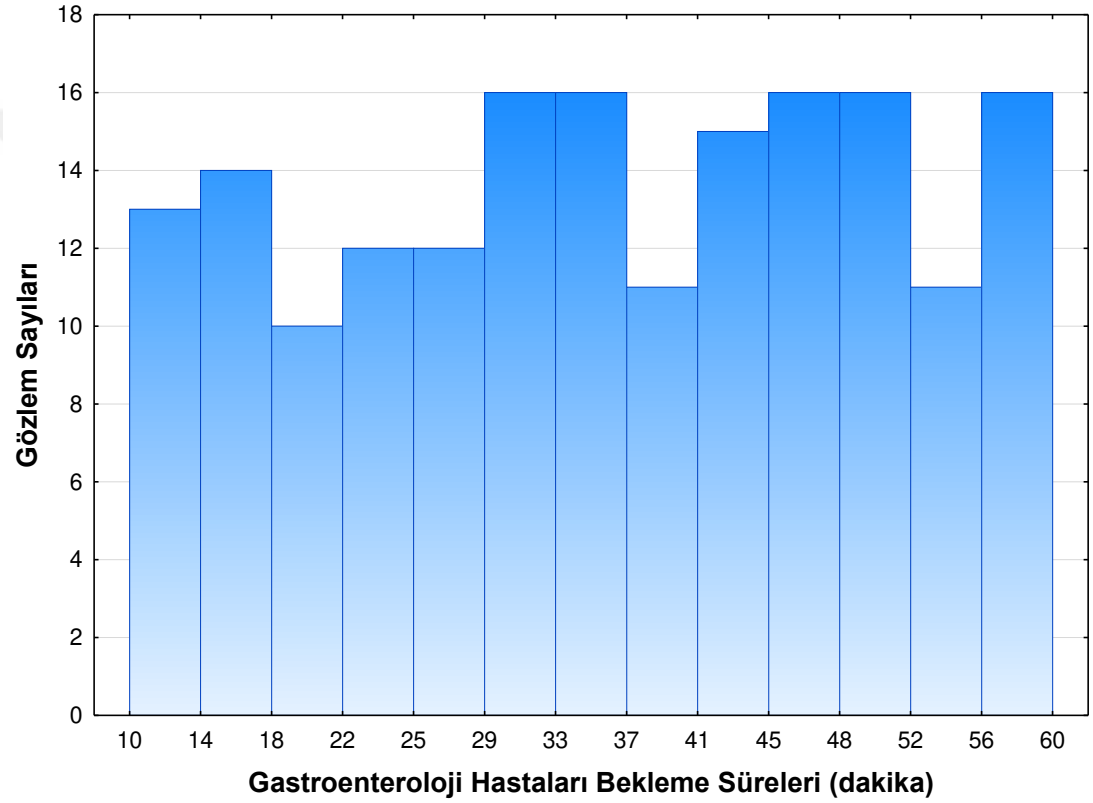
Kardiyoloji hastalarının bekleme sürelerinin histogram grafiğine göre; birden fazla tepecik bulunduğu gözlemlenebilir. 10-13 dakika, 32-35 dakika, 54-57 dakika aralarında bekleyen hastaların sayısı, diğer dakika aralarında bekleyen hastalara göre daha fazladır.



Şekil 4.2. Kardiyoloji Hastaları Bekleme Süreleri-Gamma Karma dağılımı

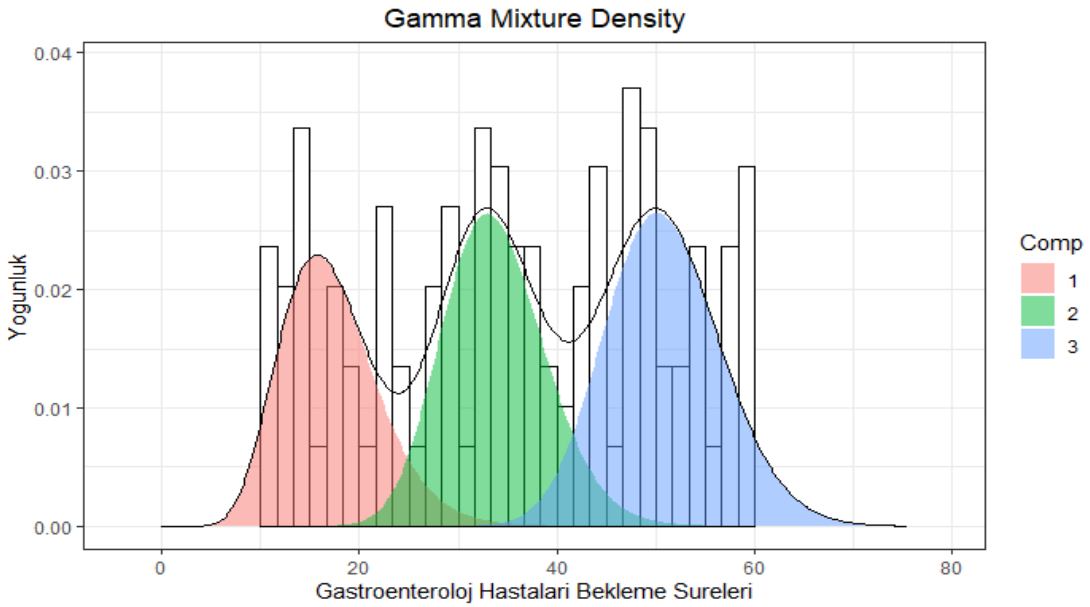
Farklı bileşen ve farklı modeller arasından AIC , BIC ve Log-likelihood değerleri kıyaslandığında, üç bileşenli Gamma karma dağılımının kardiyoloji hastalarının bekleme süresini en iyi şekilde temsil ettiği gözlemlenmiştir. Veri heterojendir; bu nedenle farklı alt grupları temsil

eden bir karma dağılım modeli uygulanmıştır. Modelde temsil edilen birinci bileşenin tahmini ortalaması ve standart sapması ($\hat{\mu}$ =14,99, $\hat{\sigma}$=3,71) olup , bu bileşen gözlemlerin %23' ünü ($\boldsymbol{\pi_1}$=0,23) temsil etmektedir. İkinci bileşende ($\hat{\mu}$ =32,99 , $\hat{\sigma}$=8,60) ve $\boldsymbol{\pi_2}$=0,43 olarak elde edilmiş, üçüncü bileşen için ($\hat{\mu}$=52,11, $\hat{\sigma}$=5,49) ve $\boldsymbol{\pi_3}$=0,34 olarak elde edilmiştir. Bu sonuçlar kardiyoloji hastalarının bekleme sürelerinin ağırlıklı olarak orta süreli grupta yoğunlaştığını göstermektedir. Veri tek bir dağılım ile açıklanamayacak kadar çeşitli alt gruplardan oluşmakta olup, Gamma karma dağılım modeli bazı histogram çubuklarında sapmalar olsa da veri heterojenliğini başarılı bir şekilde yansıtmaktadır.



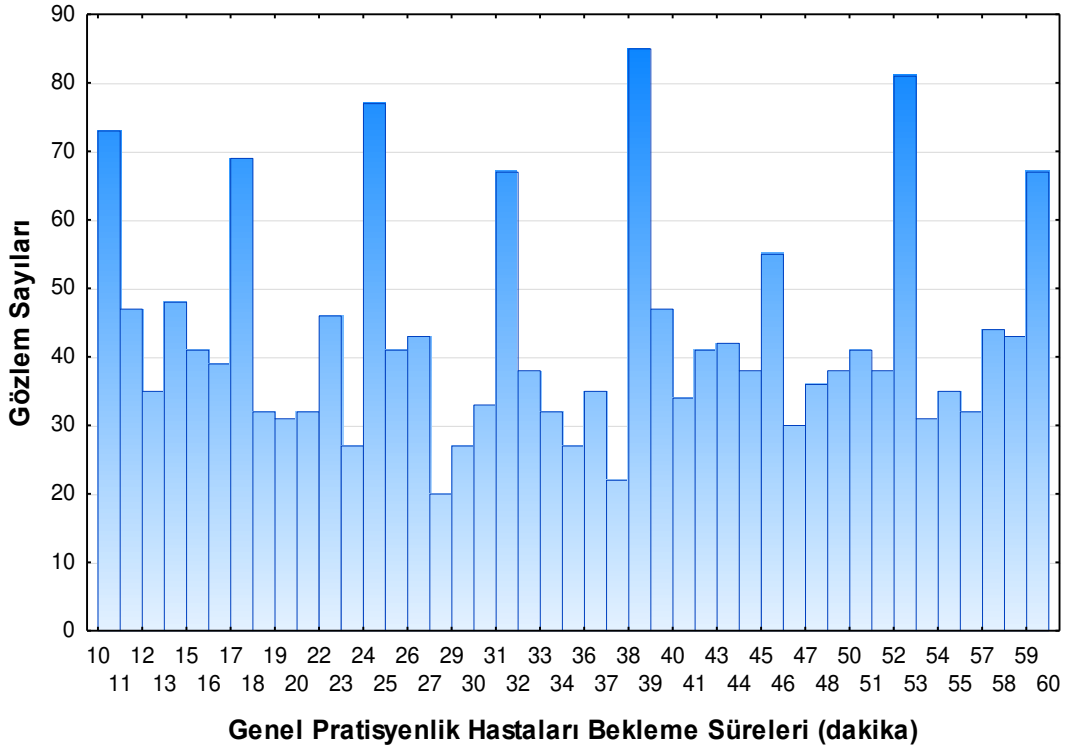
Şekil 4.3. Gastroenteroloji Hastaları Bekleme Süreleri

Gastroenteroloji hastalarının bekleme sürelerinin histogram grafiğine göre; birden fazla tepecik bulunduğu gözlemlenebilir. 29-37 dakika, 41-52 dakika, 56-60 dakika arasında bekleyen hasta sayısı, diğer dakika aralıklarında bekleyen hasta sayısından daha fazladır.



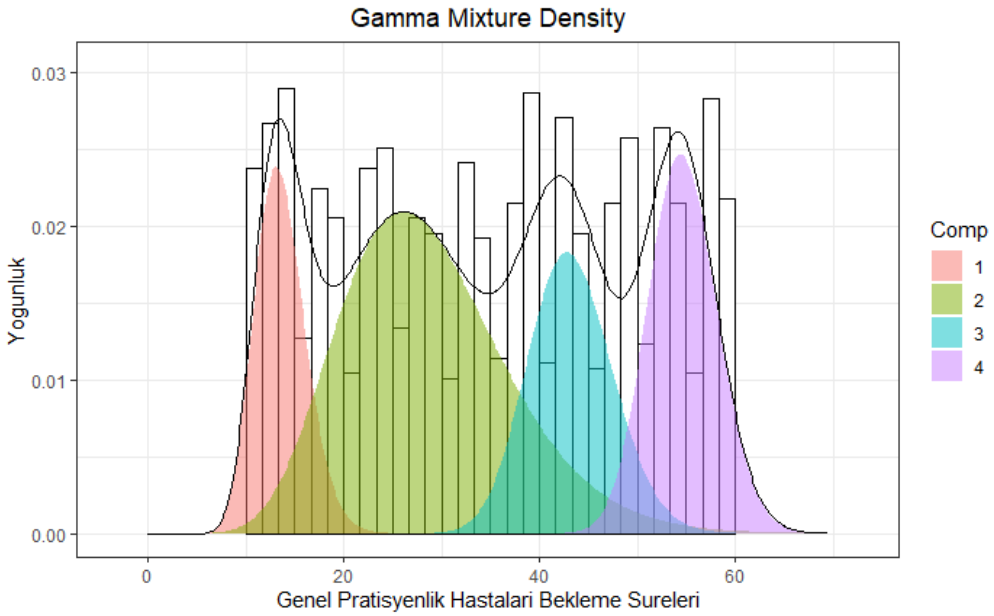
Şekil 4.4. Gastroenteroloji Hastaları Bekleme Süreleri- Gamma Karma dağılımı

Farklı bileşen ve farklı modeller arasından AIC , BIC ve Log-likelihood değerleri kıyaslandığında, üç bileşenli Gamma karma dağılımının gastroenteroloji hastalarının bekleme süresini en iyi şekilde temsil ettiği gözlemlenmiştir. Veri, heterojen biçimde dağılmıştır; bu nedenle, farklı alt gruplar karma dağılım modeli ile temsil edilebilmiştir. Modelde temsil edilen birinci bileşenin tahmini ortalaması ve standart sapması ($\hat{\mu}$ =17,12 $\hat{\sigma}$=4,93) olup , bu bileşen gözlemlerin %27' sini ($\boldsymbol{\pi}_1$=0,27) temsil etmektedir. İkinci bileşende ($\hat{\mu}$ =33,77 , $\hat{\sigma}$=5,18) ve $\boldsymbol{\pi}_2$=0,33 olarak elde edilmiş, üçüncü bileşen için ($\hat{\mu}$=50,86, $\hat{\sigma}$=5,87) ve $\boldsymbol{\pi}_3$=0,38 sonucu elde edilmiştir. Bu sonuçlar kardiyoloji hastalarının bekleme sürelerinin ağırlıklı olarak yüksek süreli grupta yoğunlaştığını göstermektedir. Veri tek bir dağılım ile açıklanamayacak kadar çeşitli alt gruplardan oluşmakta olup, Gamma karma dağılım modeli, bazı histogram çubuklarında sapmalar olsa da veri heterojenliğini başarılı bir şekilde yansıtmaktadır.



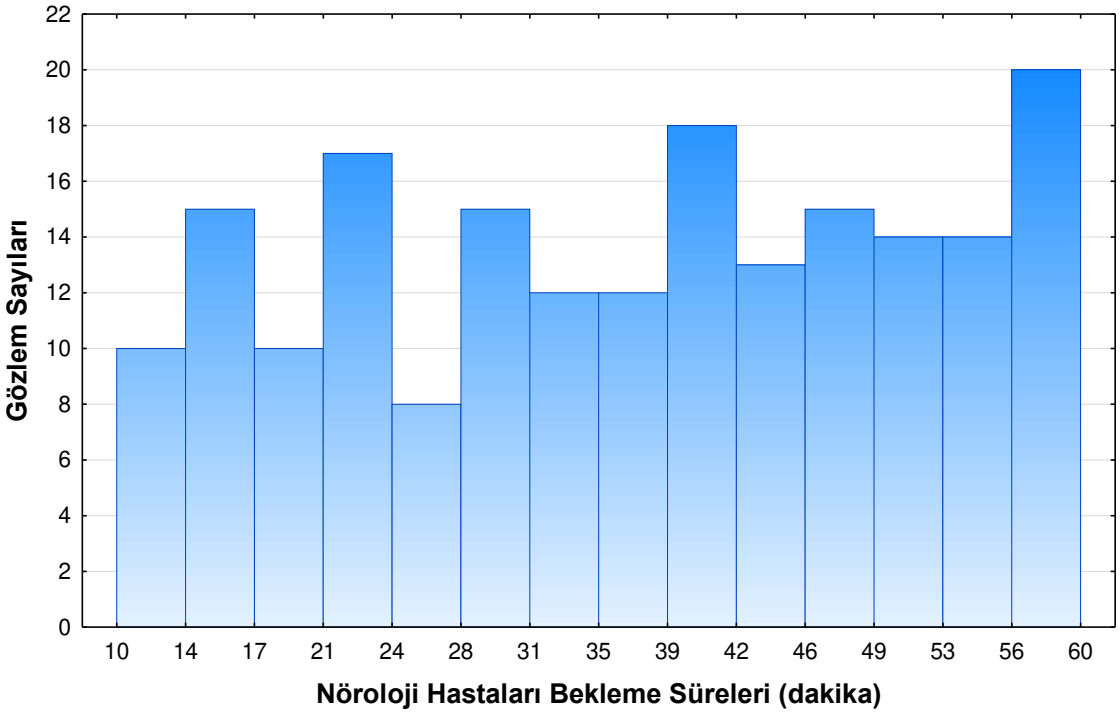
Şekil 4.5. Genel Pratisyenlik Hastaları Bekleme Süreleri

Genel Pratisyenlik hastalarının bekleme sürelerinin histogram grafiğine göre; birden fazla tepecik bulunduğu gözlemlenebilir. 10-11 dakika, 17-18 dakika, 24-25 dakika, 38-39 dakika, 52-53 dakika aralığında bekleyen hasta sayısı, diğer dakika aralıklarında bekleyen hasta sayısından daha fazladır.



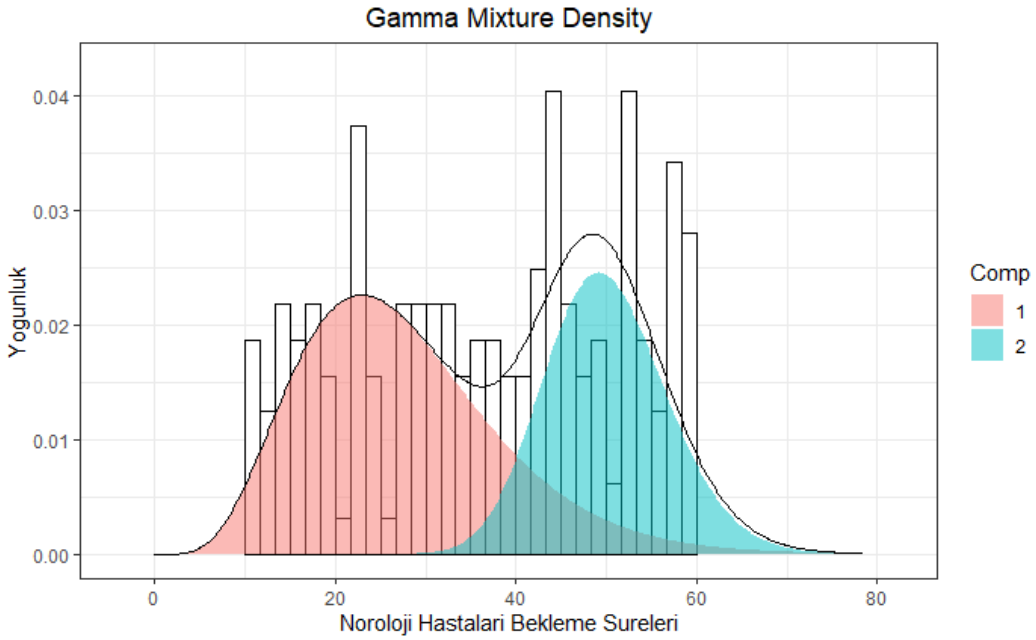
Şekil 4.6. Genel Pratisyenlik Hastaları Bekleme Süreleri-Gamma Karma dağılımı

Farklı bileşen ve farklı modeller arasından AIC , BIC ve Log-likelihood değerleri kıyaslandığında, dört bileşenli Gamma karma dağılımının genel pratisyenlik hastalarının bekleme süresini en iyi şekilde temsil ettiği gözlemlenmiştir. Veri, heterojen biçimde dağılmıştır; bu nedenle, farklı alt gruplar karma dağılım modeli ile temsil edilebilmiştir. Modelde temsil edilen birinci bileşenin tahmini ortalaması ve standart sapması ($\hat{\mu}$ =13,59 $\hat{\sigma}$=2,59) olup , bu bileşen gözlemlerin %15' ini ($\boldsymbol{\pi}_1$=0,15) temsil etmektedir. İkinci bileşenin değerleri ($\hat{\mu}$ =28,64 , $\hat{\sigma}$=8,53) ve $\boldsymbol{\pi}_2$=0,43 olarak elde edilmiş, üçüncü bileşen için ($\hat{\mu}$=43,26, $\hat{\sigma}$=5,87) ve $\boldsymbol{\pi}_3$=0,19 olarak elde edilmiştir. Bu sonuçlar genel pratisyenlik hastalarının bekleme sürelerinin ağırlıklı olarak düşük değer sayılabilecek, ikinci en küçük bekleme süresine sahip grupta yoğunlaştığı, verinin %60' a yakınının düşük bekleme süresinden oluştuğu çıkarımının yapılmasını sağlamaktadır. Veri tek bir dağılım ile açıklanamayacak kadar çeşitli alt gruplardan oluşmakta olup, Gamma karma dağılım modeli, veri heterojenliğini başarılı bir şekilde yansıtmaktadır.



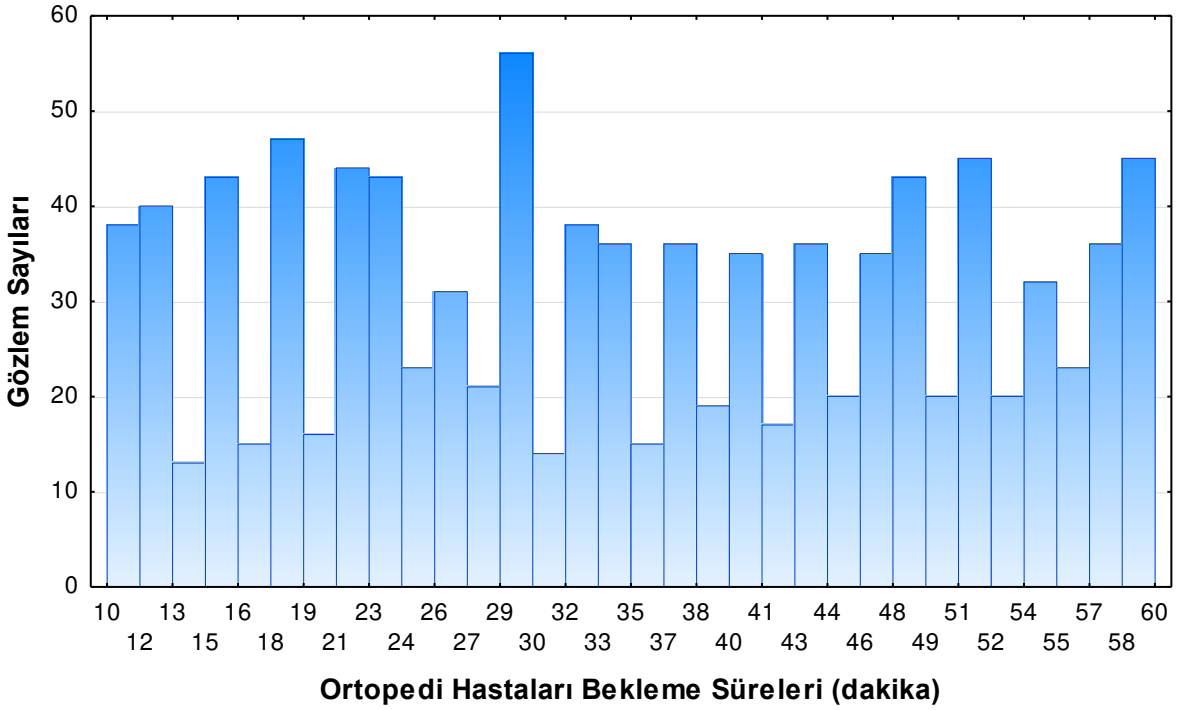
Şekil 4.7. Nöroloji Hastaları Bekleme Süreleri

Nöroloji hastalarının bekleme sürelerinin histogram grafiğine göre; birden fazla tepecik bulunduğu gözlemlenebilir. 21-24 dakika, 39-42 dakika, 56-60 dakika arasında bekleyen hasta sayısı, diğer dakika aralıklarında bekleyen hasta sayısından fazladır.



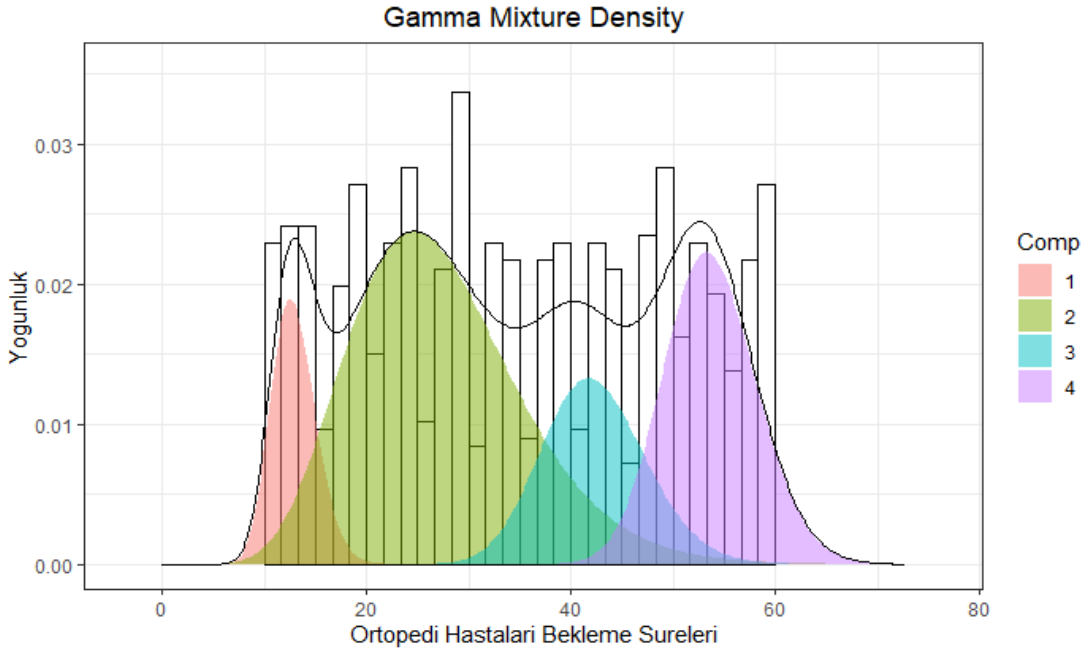
Şekil 4.8. Nöroloji Hastaları Bekleme Süreleri-Gamma Karma dağılım

Farklı bileşen ve farklı modeller arasından kıyaslandığında iki bileşenli Gamma karma dağılımının nöroloji hastalarının bekleme süresini en iyi şekilde temsil ettiği gözlemlenmiştir. Farklı bileşen ve farklı modeller arasından AIC , BIC ve Log-likelihood değerleri kıyaslandığında, iki bileşenli Gamma karma dağılımının nöroloji hastalarının bekleme süresini en iyi şekilde temsil ettiği gözlemlenmiştir. Veri, heterojen biçimde dağılmıştır; bu nedenle, farklı alt gruplar karma dağılım modeli ile temsil edilebilmiştir. Modelde temsil edilen birinci bileşenin tahmini ortalaması ve standart sapması ($\hat{\mu}$ =27,44 $\hat{\sigma}$=11,17) olup , bu bileşen gözlemlerin %58' ini ($\boldsymbol{\pi}_1$=0,59) temsil etmektedir. İkinci bileşende ($\hat{\mu}$ =50,13 , $\hat{\sigma}$=6,75) ve $\boldsymbol{\pi}_2$=0,41 olarak elde edilmiştir. Bu sonuçlar nöroloji hastalarının ağırlıklı olarak düşük bekleme sürelerinde yoğunlaştığını ve verinin heterojenliğini gösterir. Histogram çubuklarında sapmalar mevcut olsa da modelin başarılı biçimde veriyi temsil ettiği yorumu yapılabilir.



Şekil 4.9. Ortopedi Hastaları Bekleme Süreleri

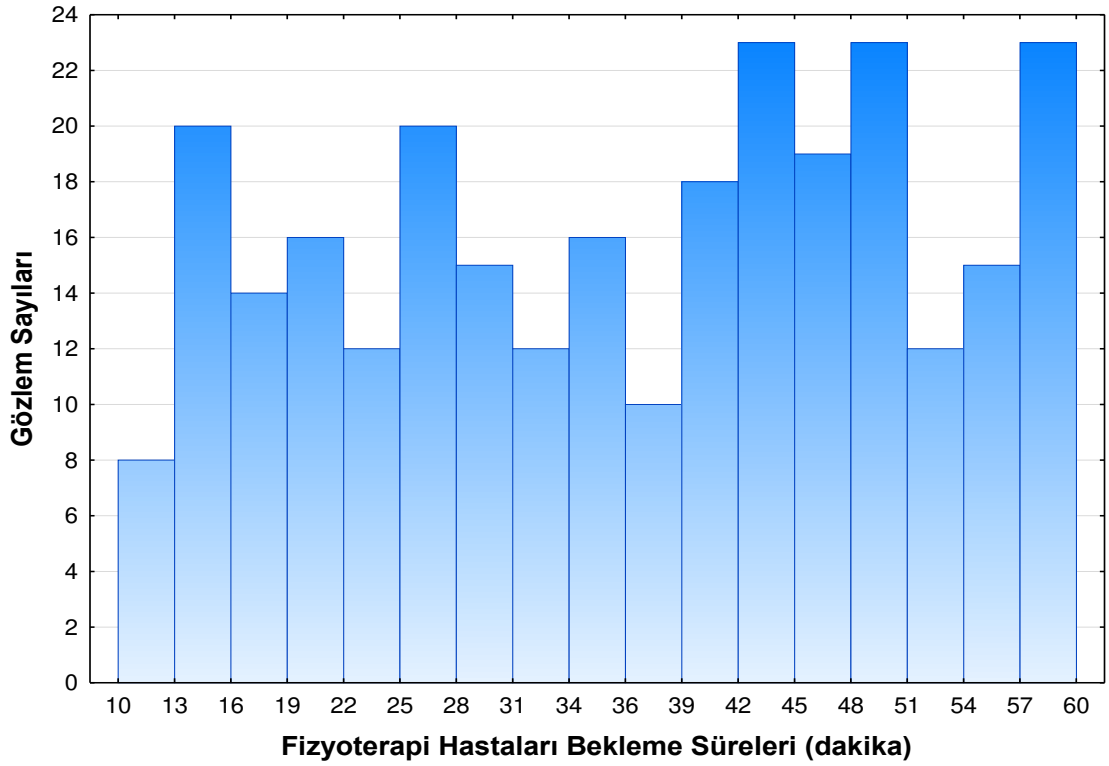
Ortopedi hastalarının bekleme sürelerinin histogram grafiğine göre; birden fazla tepecik bulunduğu gözlemlenebilir.



Şekil 4.10. Ortopedi Hastaları Bekleme Süreleri-Gamma Karma dağılımı

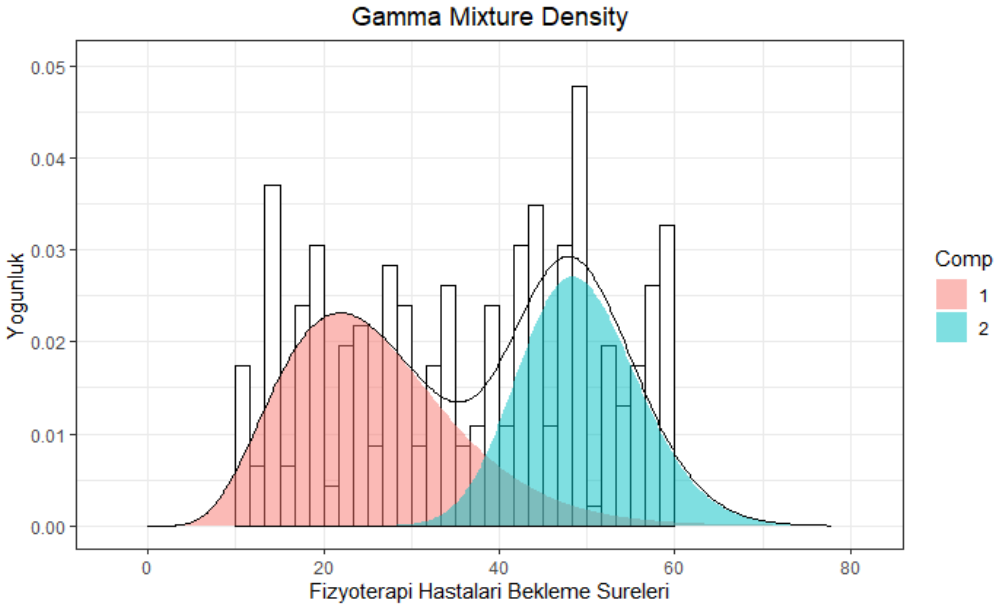
Farklı bileşen ve farklı modeller arasından AIC , BIC ve Log-likelihood değerleri kıyaslandığında, dört bileşenli Gamma karma dağılımının ortopedi hastalarının bekleme süresini

en iyi şekilde temsil ettiği gözlemlenmiştir. Veri, heterojen biçimde dağılmıştır; bu nedenle, farklı alt gruplar karma dağılım modeli ile temsil edilebilmiştir. Modelde temsil edilen birinci bileşenin tahmini ortalaması ve standart sapması ($\hat{\mu}$ =12,92 $\hat{\sigma}$=2,18) olup , bu bileşen gözlemlerin %10' unu ($\boldsymbol{\pi}_1$=0,10) temsil etmektedir. İkinci bileşende ($\hat{\mu}$ =27,20 , $\hat{\sigma}$=8,34) ve $\boldsymbol{\pi}_2$=0,47 olarak elde edilmiş, üçüncü bileşende ($\hat{\mu}$=42,41, $\hat{\sigma}$=5,06) ve $\boldsymbol{\pi}_3$=0,16 olarak ve dördüncü bileşende ($\hat{\mu}$=53,70, $\hat{\sigma}$=4,53) değerleri elde edilmiştir. Bu sonuçlar ile ortopedi hastalarının bekleme sürelerinin ağırlıklı olarak düşük değer sayılabilecek, ikinci en küçük bekleme süresine sahip grupta yoğunlaştığını, en küçük bekleme süresine sahip iki grubun ağırlıkları toplamının %60 a yakın değer aldığı gözlemlenebilmektedir. Veri tek bir dağılım ile açıklanamayacak kadar çeşitli alt gruplardan oluşmakta olup, Gamma karma dağılım modeli, veri heterojenliğini başarılı bir şekilde yansıtmaktadır.



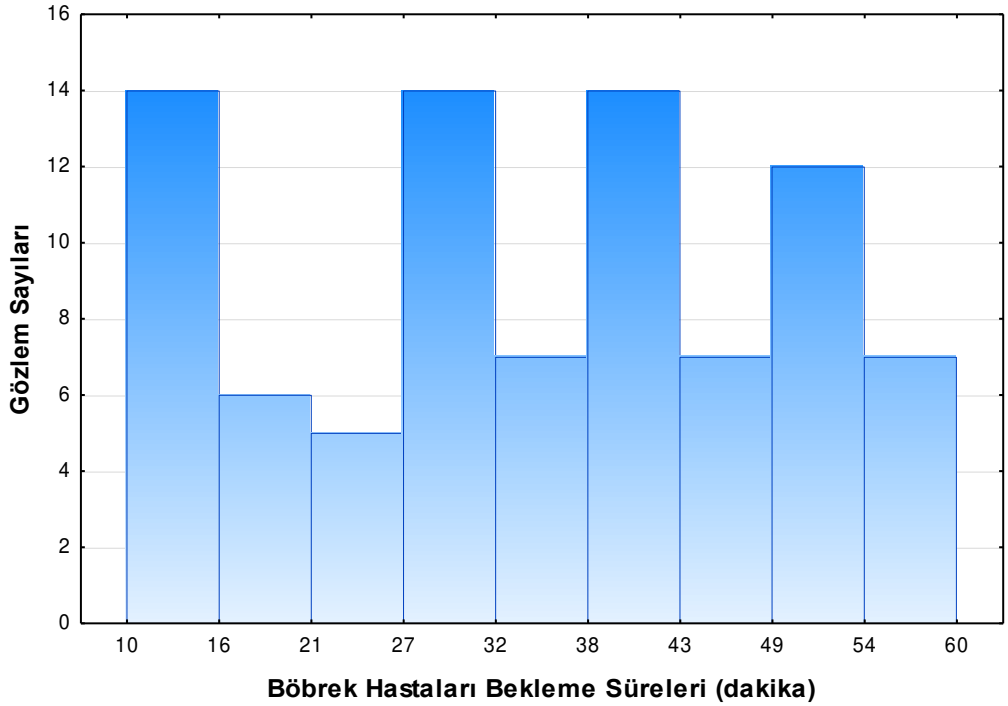
Şekil 4.11. Fizyoterapi Hastaları Bekleme Süreleri

Fizyoterapi hastalarının bekleme sürelerinin histogram grafiğine göre; birden fazla tepecik bulunduğu gözlemlenebilir.13-16 dakika, 25-28 dakika, 42-51 dakika, 57-60 dakika bekleyen hasta sayısı, diğer dakika aralıklarında bekleyen hasta sayısından fazladır.



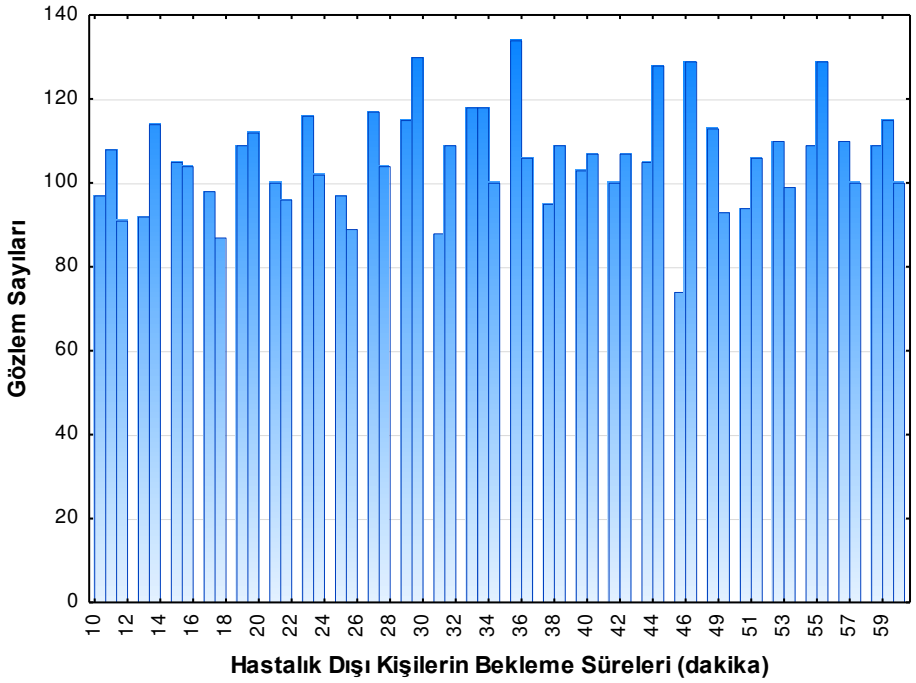
Şekil 4.12. Fizyoterapi Hastaları Bekleme Süreleri- Gamma Karma dağılımı

Farklı bileşen ve farklı modeller arasından AIC , BIC ve Log-likelihood değerleri kıyaslandığında, iki bileşenli Gamma karma dağılımının fizyoterapi hastalarının bekleme süresini en iyi şekilde temsil ettiği gözlemlenmiştir. Veri, heterojen biçimde dağılmıştır; bu nedenle, farklı alt gruplar karma dağılım modeli ile temsil edilebilmiştir. Modelde temsil edilen birinci bileşenin tahmini ortalaması ve standart sapması ($\hat{\mu}$ =25,74 $\hat{\sigma}$=9,97) olup , bu bileşen gözlemlerin %54' ünü ($\boldsymbol{\pi}_1$=0,54) temsil etmektedir. İkinci bileşenin tahmini ortalaması ve standart sapması ($\hat{\mu}$ =49,35 $\hat{\sigma}$=6,81) olup, bu bileşen gözlemlerin %46' sını ($\boldsymbol{\pi}_1$=0,46) temsil etmektedir. Fizyoterapi hastalarının ağırlıklı olarak düşük bekleme süreleri sonunda acil servisten ayrıldıkları çıkarımı yapılabilir. Şekil 4.12 incelendiğinde, sapan histogram çubukları olsa da iki bileşenli Gamma dağılımının veriyi iyi bir şekilde temsil ettiği yorumunda bulunulabilir.



Şekil 4.13. Böbrek Hastaları Bekleme Süreleri

Böbrek hastalarının bekleme sürelerinin saf dağılımlara uygunluğunun kontrolü için yapılan Ki-Kare testi sonucunda $P > 0{,}05$ elde edilmiştir. Veri seti Normal dağılım ile temsil edilebilmektedir. Bu sonuç, verilerin homojen bir dağılımdan geldiğini ifade eder.



Şekil 4.14 Hastalık Dışı Kişilerin Bekleme Süreleri

Hastalık dışı kişilerin bekleme sürelerinin histogram grafiğine göre; histogram çubukları arasında boşluk olduğu ve birden fazla tepecik olduğu gözlemlenmektedir. Bu görüntü karma dağılımlardan Poisson karma dağılımı kullanılarak analiz yapılmasına öncüllük etmiştir. Bekleme süreleri dakika cinsinden tam sayı olarak kayıt altına alındığı için kesikli dağılımlar için kullanılan Poisson dağılımları kullanılmasında herhangi bir sakınca görülmemiştir.

Çizelge 4.2. Hastalık Türlerine Göre Gamma Karma Modeli Sonuçları

		Oranlar				**Gamma Karma**				**Seçim Kriterleri**		
Veri	**k**	**Dağılım Parametreleri**										
		π1	**π2**	**π3**	**π4**	**(α1;β1)**	**(α2;β2)**	**(α3;β3)**	**(α4;β4)**	**AIC**	**BIC**	**LL**
Kardiyoloji B.	3	0,23	0,43	0,34	-	(16,25;1,08)	(14,69;0,44)	(90;1,72)	-	1995,83	2023,93	-989,91
Gastroenteroloji B.	3	0,27	0,34	0,39	-	(12,04;0,7)	(42,5;1,25)	(75;1,47)	-	1432,84	1458,3	-708,42
Genel Pratisyen B.	4	0,15	0,43	0,20	0,22	(27,5;2,02)	(11,26;0,39)	(100;2,31)	(240;4,38)	14611,29	14671,98	-7294,6
Nöroloji B.	2	0,59	0,41	-	-	(6,03;0,21)	(55;1,09)	-	-	1559	1575,31	-774,5
Ortopedi B.	4	0,10	0,48	0,17	0,25	(35;2,7)	(10,64;0,39)	(70;1,65)	(140;2,6)	7929,79	7983,72	-3953,9
Fizyoterapi B.	2	0,54	0,46	-	-	(6,65;0,25)	(52,5;1,06)	-	-	2213,34	2231,44	-1101,6
Böbrek Bölümü	2	0,59	0,41	-	-	(5,02;0,19)	(45;0,95)	-	-	699,67	711,94	-344,83

Çizelge 4.2 alternatif Gamma karma dağılımları arasından en iyi sonucu veren modellerin parametreleri ve model performans değerlerini içermektedir. Yukarıda Gamma karma dağılımı grafikleri yorumlanırken Çizelge 4.2'deki değerler kullanılmıştır.

k değeri, karma dağılımlardaki bileşen sayısını ifade etmektedir. π değerleri, bileşenlerin karma modeldeki bulunma oranını ifade etmektedir. α ve β değerleri Gamma dağılımının şekil parametreleridir. AIC, BIC, LL değerleri aynı veri üzerinde kullanılan farklı modeller arasında kıyaslama yapmak için kullanılmaktadır. Tek başlarına bir anlam ifade etmezler. AIC ve BIC değerleri küçüldükçe model iyileşmekte, Log-likelihood (LL) değeri sıfıra yaklaştıkça model iyileşmektedir. En iyi modeli bulabilmek için alternatifler karşılaştırılmalıdır.

Çizelge 4.3. Hastalık Dışı Kişilerin Bekleme Süreleri

		Oranlar				**Poisson Karma**				**Seçim Kriterleri**		
Veri	**k**	**Dağılım Parametreleri**										
		π1	**π2**	**π3**	**π4**	**(λ_1)**	**(λ_2)**	**(λ_3)**	**(λ_4)**	**AIC**	**BIC**	**ICL**
Hastalık Dışı	4	0,23	0,43	0,25	0,09	(34,15)	(50,04)	(24,30)	(15,16)	43170,73	43216,89	45768,20

Çizelge 4.3 alternatif Poisson karma dağılımları arasından en iyi sonucu veren modelin parametre ve model performans değerlerini içermektedir. Poisson karma dağılımına uyan "hastalık dışı durumlar" verisine uygulanan modelin sonuçlarını yansıtmaktadır.

k değeri, karma dağılımdaki bileşen sayışını ifade etmektedir. π değerleri, bileşenlerin karma modeldeki bulunma oranını ifade etmektedir. λ. değeri ortalama ve varyans değerini temsil etmektedir. AIC ve BIC modelin uyumunu ve karmaşıklığını değerlendirir. Değerleri küçüldükçe modelin kalitesi artmaktadır. ICL kümeleme yorumları için kullanılır. Değeri arttıkça, kümeleme yapısını daha net gösterir. En kalabalık grup π2 = %43 ile bekleme sürelerinin ortalaması (λ_1) =50 olan hastaları temsil ediyor. En seyrek grup π4= %9 ile bekleme sürelerinin ortalaması (λ_4) 15 olan düşük değerleri temsil ediyor.

Çizelge 4.4. Gamma Karma Dağılım ile Gözlemlenen Verinin Karşılaştırılması

Veri	Gözlemlenen Veri		Gamma Karma Dağılım Verisi	
	Ortalama	Standart Sapma	Ortalama	Standart Sapma
Kardiyoloji B.	35,35	15,47	35,61	15,11
Gastroenteroloji B.	35,83	14,66	36,10	14,63
Genel Pratisyenlik B.	34,91	14,98	35,04	14,96
Nöroloji B.	36,80	14,66	37,68	14,62
Ortopedi B.	34,98	14,82	35,06	14,79
Fizyoterapi B.	36,57	14,62	37,10	14,47
Böbrek Bölümü	34,70	14,42	35,09	14,44

Çizelge 4.5. Poisson Karma Dağılım ile Gözlemlenen Verinin Karşılaştırılması

Veri	Gözlemlenen Veri		Poisson Karma Dağılım Verisi	
	Ortalama	Standart Sapma	Ortalama	Standart Sapma
Hastalık Dışı	35,29	14,62	34,66	13,87

Çizelge 4.2 ve Çizelge 4.3' de parametreleri verilen karma dağılım modelleri ile elde edilen verilerin, Çizelge 4.4 ve Çizelge 4.5' de gösterilen gözlemlenen bekleme süreleri ile karşılaştırılması amacıyla ortalama ve standart sapma değerleri incelenmiştir. Verilerin aynı istatistiksel dağılımından gelip gelmediği Mann-Whitney U karşılaştırma testleri yapılarak kontrol edilmiştir. Tüm testlerde $P > 0{,}05$ bulunması, parametreleri verilen karma dağılımlardan üretilen değerlerle, gözlemlenen gerçek değerlerin aynı dağılımdan geldiğini göstermektedir.

5. SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında bir hastanenin acil servisine başvuran hastaların bekleme süreleri analiz edilmiştir. İlk olarak hastaların bekleme sürelerinin histogram grafikleri çıkarılarak verinin dağılımı hakkında bir öngörü elde edilmeye çalışılmıştır. Öngörünün netleşmesi açısından hastalık grupları üzerinde tekil teorik dağılımlar için ayrı ayrı Ki-Kare testi yapılmıştır. Histogram grafiğindeki görsel ifade ve Ki-Kare testi sonuçları, hastalık gruplarından yalnızca böbrek hastalarının bekleme sürelerinin Normal dağılım ile temsil edilebildiğini ortaya çıkarmıştır. Acil servisi ziyaret eden diğer gruplara, Gamma ve Poisson karma dağılımları kullanılarak kümeleme analizi yapılmıştır. Teorik karma dağılımından elde edilen ortalama ve standart sapma verileri ile gözlemlerden elde edilen değerlerle çok benzer sonuçlar ortaya sunmuştur. Mann-Whitney U testi yapılarak bu benzerli istatistiksel olarak doğrulanmıştır. Elde edilen bulgular, karma dağılımların heterojen yapıyı başarıyla temsil ettiğini ve farklı hasta gruplarının bekleme sürelerinin daha doğru tahmin edilebildiğini göstermiştir.

Bekleme süreleri yalnızca istatistiksel bir analiz konusu değil, aynı zamanda sağlık hizmetlerinde operasyonel verimliliğin ve hasta memnuniyetinin en önemli göstergelerinden biridir. Uzayan bekleme süreleri, hem hizmet kalitesini düşürmekte hem de hasta memnuniyetini olumsuz yönde etkilemektedir. Bu nedenle, karma dağılım temelli makine öğrenmesi algoritmalarının karar destek sistemlerine entegre edilmesi; acil servis yoğunluğunun önceden öngörülmesini ve buna bağlı olarak personel, ekipman ve yatak planlamasının daha etkin yapılmasını mümkün kılacaktır.

Çalışmada bekleme sürelerinin branşlar arasında anlamlı farklılıklar gösterdiği ortaya konulmuştur. Bu nedenle sağlık kurumlarının kaynak dağılımı yaparken branş bazlı analizler gerçekleştirmesi ve yoğun branşlara ek kaynak tahsis etmesi önerilmektedir. Ayrıca, yaş, cinsiyet ve haftanın günleri gibi faktörlere göre de bekleme sürelerinin incelenmesi, daha ayrıntılı analiz imkânı sunacaktır.

Analizler standart donanıma sahip bir bilgisayar ile yürütülmüş ve veri büyüklüğü arttıkça sistem performansının yavaşladığı gözlenmiştir. Daha büyük veri setleri ile çalışacak araştırmacılara, bulut tabanlı çözümler veya dağıtılmış sistemlerden yararlanmaları önerilmektedir. Ayrıca yalnızca sayısal bekleme süresi verilerinin değil; elektronik sağlık kayıtları, görüntü verileri, reçeteler ve laboratuvar sonuçları gibi farklı veri türlerinin de analizlere dâhil edilmesi, modelin kapsamını ve öngörü gücünü artıracaktır.

Gelecek çalışmalar açısından, karma dağılımların yalnızca acil servis bekleme sürelerinde değil, sağlık sisteminin diğer süreçlerinde de uygulanabilirliği üzerinde durulabilir. Örneğin, ameliyathane planlaması, yoğun bakım yatak yönetimi, polikliniklerde hasta yoğunluğu tahmini gibi kritik alanlarda bu modellerin kullanılması mümkündür. Ayrıca, klasik istatistiksel modellerin yanı sıra, Bayesyen yaklaşımlar, derin öğrenme algoritmaları ve simülasyon tabanlı yöntemlerle

karşılaştırmalı çalışmalar yapılması, literatüre önemli katkılar sunacaktır. Özellikle anlık veri akışıyla çalışan gerçek zamanlı karar destek sistemlerine karma dağılım modellerinin entegre edilmesi, bekleme sürelerinin dinamik biçimde tahmin edilmesini sağlayarak sağlık kurumlarına operasyonel esneklik kazandıracaktır.

Sonuç olarak, bu çalışma, acil servislerde bekleme sürelerinin modellenmesinde karma dağılımların hem güçlü hem de uygulanabilir bir yöntem olduğunu ortaya koymuştur. Bulgular, yalnızca teorik literatüre katkı sunmakla kalmamış, aynı zamanda sağlık kurumlarının operasyonel süreçlerini geliştirmeye yönelik somut öneriler de ortaya koymuştur. Bekleme sürelerinin doğru bir şekilde tahmin edilmesi, hasta memnuniyetini artırmanın yanı sıra, sağlık hizmetlerinde kaliteyi ve verimliliği yükseltmeye yönelik stratejilerin geliştirilmesinde de önemli bir araç olarak değerlendirilebilir.

KAYNAKLAR

Abay R. A., Çölgeçen Y., (2018). Psikiyatrik sosyal hizmet- koruyucu, tedavi edici ve rehabilite edici ruh sağlığı Alanında Sosyal Çalışmacıların Rolü, s 2147-2185.

Acıduman, A. (2010). Darüşşifalar bağlamında kitabeler, vakıf kayıtları ve tıp tarihi açısından önemleri-anadolu selçuklu darüşşifaları özelinde, Ankara Üniversitesi Tıp Fakültesi Mecmuası, 63(1): 9-15.

Ackoff, R. (1989). From data to wisdom, Journal of Applied Systems Analysis, s. 3-9

Agrawal R., Chatterjee M. J., Kumar A., Rathore S. P., Le D., (2021). Machine learning for healthcare handling and managing data, CRC Press.

Ahmed, H., Younis, E.M.G., Hendawi, A. and Ali, A.A. (2020). Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation ComputerSystems, 111, 714–722.

Akman E., Tarım M, (2020). Türkiye ve İngiltere sağlık sistemleri: birinci basamak sağlık hizmetleri karşılaştırması, Uluslararası Sağlık Yönetimi ve Stratejileri Araştırmaları Dergisi, s. 303-316

Aktan. E, (2018). Büyük veri: uygulama alanları, Analitiği ve Güvenlik Boyutu, s.1-22.

Altındiş S., Morkoç K. İ., (2018). Sağlık Hizmetlerinde Büyük Veri, s.257-271.

Alves, M.A., Castro, G.Z., Oliveira, B.A.S., Ferreira, L.A., Ramírez, J.A., Silva, R. And Guimarães, F.G. (2021). Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. Computers in Biology and Medicine, 132, 104-335.

Aslan S., Erdem R., (2017). Hastanelerin Tarihsel Gelişimi, Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü Dergisi. 2017/2, 27, 7-21.

Asri H., Mousannif H., Moatassime A.H., Noel T., (2015). Big data in healthcare: Challenges and Opportunities, Institute of Electrical and Electronics Engineers IEEE.

Atienza, N., Garcia-Heras, J., Munoz-Pichardo, J., ve Villa, R., (2008). An application of mixture distributions in modelization of length of hospital stay. Statistics in Medicine, 27: 1403-1420.

Aytaç, Ö. ve Kurtdaş, M.Ç. (2015). Sağlık-hastalığın toplumsal kökenleri ve sağlık sosyolojisi. Fırat Üniversitesi Sosyal Bilimler Dergisi, 25, 231-250.

Bansal H., Balusamy B., Poongodi T., KP K.F., (2021). Machine learning and analytics in healthcare systems principles and applications, CRC Press.

Bulut F. (2023). Sağlıkta büyük veri: ulusal düzenlemeler ve veri kayıt sistemlerinin tıp etiği açısından incelenmesi, Bursa Uludağ Üniversitesi Sağlık Bilimleri Enstitüsü, Tıp Fakültesi Tıp Tarihi ve Etik Anabilim Dalı, Doktora Tezi.

Bulut M., Eygü H. (2020). İş kazalarının lojistik regresyon yöntemi ile incelenmesi: bayburt ili örneği, Uluslararası Toplum Araştırmaları Dergisi, 10, 4956-4974.

Cabena Peter ve ark., (1998). Discovering data mining: from concept to implementation, USA, International Business Machines Corporation, s.12.

Cleary, P. G., 1991. Variations in length of stay and outcomes for six medical and surgical conditions in Massachusetts and California. JAMA, 266: 73-85.

Dal.Ö. (2021). Sağlık hizmetlerinde büyük veri: mobil sağlık uygulamalarının kullanımını etkileyen faktörlerin genişletilmiş teknoloji kabul modeli ile incelenmesi, Beykent Üniversitesi Lisansüstü Eğitim Enstitüsü, Doktora Tezi.

Dempster, A.P.; Laird, N.M.. (1977). Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 39, 1–22.

Dutt S., Chandramouli S., Das K. A., (2018). Machine learning, Pearson Press.

Eberendu, C., A., (2016). Unstructured Data: an overview of the data of big data, International Journal of Computer Trends and Technology (IJCTT), s.46-50.

Ekelik H., Altaş D. (2019). Dijital reklam verilerinden yararlanarak potansiyel konut alıcılarının rastgele orman yöntemiyle sınıflandırılması, İktisat Araştırmaları Dergisi, s-28-45

Erkan N. (2002) Regresyon Analizi ve Ormancılıkta Kullanımı, İstanbul Üniversitesi Orman Fakültesi Dergisi, s-56-76

Fogel, A.L. and Kvedar, J.C. (2018). Artificial intelligence powers digital medicine. NPJ Digital Medicine, 1, 1–4.

Frenkel Brian, (2003). Using Artificial Intelligence to Detect Fraud in Credit Cards.

G. Uyer, G. Kocaman, S. Oktay, G. Argon, S. Abaan (1996). Sorun Çözme ve Karar Verme. (Ed Hemşirelik Hizmetleri Yönetimi El Kitabı içinde, (s.36-127) İstanbul: VKV

Gençer C. Ç., Er F., Barut B., Kara Y., (2021). Koruyucu Sağlık Hizmetlerinin Sunumunda Sosyal Hizmet Mesleğinin Önemi, s 1125- 1142

Gülay, A., Kabataş,Y., & Attilla, İ. (2023). Türkiye ve İngiltere'de sağlık alanında faaliyet gösteren sivil toplum kuruluşlarının karşılaştırılması. Alanya Akademik Bakış, 7(3), Sayfa No.1425-1446

Güleryüz S. S., (2023). Karma dağılımların hizmet sektöründe uygulanması, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi

Günher E., Fidan M., Akbayır Ö., (2025). SOM ve K-ortalama kümeleme algoritmaları kullanarak vagon tamire tutma verilerinin incelenmesi, Demiryolu Mühendisliği Araştırma Makalesi, s-168-177

Gürhan N., Üstün B., (2015). Rehabilitasyon hizmetleri, s-46-51

Gürsakal, N. (2017). Büyük veri, Dora Basın-Yayın Dağıtım Ltd.

Hacıbeyoğlu, M., Çelik M., Çiçek E. Ö. (2023). K En yakın komşu algoritması ile binalarda enerji verimliliği tahmini, Necmettin Erbakan Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, s-65-74

Halevi, Gali and Moed, Henk F. Dr. (2012). "The evolution of big data as a research and scientific topic: Overview of the literature," Research Trends: 1(30), Article 2.

Hänig, C., Schierle, M., & Trabold, D. (2010). Comparison of structured vs. unstructured data for industrial quality analysis. In Proceedings of The World Congress on Engineering and Computer Science.

Harman, G. (2021). Destek vektör makineleri ve naive bayes sınıflandırma algoritmalarını kullanarak diabetes mellitus tahmini, Avrupa Bilim ve Teknoloji Dergisi, s-7-13

https://antalyaism.saglik.gov.tr/TR-234669/saglik-hizmet-sunucularinin-basamaklandirilmasina-dair-yonetmelik.html

https://www.ibm.com/history/deep-blue

https://www.who.int/about/governance/constitution

Jiang, F., Jiang, Y., Zhi, H., Dong, Y. and Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2, 230–243.

Karaçor S., Arkan A., (2012). Sağlık kuruluşlarında pazarlama: Sağlık pazarlama karması unsurlarının hasta/müşteri açısından önemi üzerine bir araştırma, s 90-118

Karimi M., Brazier J. (2016). Health, health-related quality of life, and quality of life: What is the difference? 645-649.

Kavzaoğlu T., Çölkesen İ. (2010). Karar ağaçları ile uydu görüntülerinin sınıflandırılması: Kocaeli örneği, Harita Teknolojileri Elektronik Dergisi, s-36-45

Keleşlioğlu K., Arslan D, Aba G., (2022). Hastanelerde tıbbi sosyal hizmet uygulamaları: Türkiye ve seçili ülkelerde mevcut durum, Türkiye Sosyal Hizmet Araştırmaları Dergisi, s 100-112

Kızmaz, H.M. ve Küçükçolak A.R. (2021). Büyük veri teriminin kökeni ve büyük verinin V'leri, s.19-31

Kingston, A., Robinson, L., Booth, H., Knapp, M. and Jagger, C. (2018). Projections of multimorbidity in the older population in England to 2035: Estimates from the population ageing and care simulation model, 47, 374–380.

Koyuncugil, A. ve Özgülbaş, N. (2009). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. Bilişim Teknolojileri Dergisi, 2(2), 21-32.

Köse İ. (2018). Veri madenciliği teori uygulama ve felsefesi, Papatya Bilim Üniversite Yayıncılığı

Kurşun A., (2021), Büyük veri ve sağlık hizmetlerinde büyük veri işleme araçları, Hacettepe Sağlık İdaresi Dergisi, s.921-940

Lee, A. H., Ng, S. A., Yau, K. K. (2001). Determinants of maternity length of stay: a Gamma mixture risk-adjusted model, Health Care Management Science, 4(4): 449-455.

Meyer, A., Zverinski, D., Pfahringer, B., Kempfert, J., Kuehne, T., Sündermann, S., Eickhoff, C., (2018). Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respiratory Medicine, 6(12): 905-914. doi:10.1016/S2213-2600(18)30300-X

Millard, P., (1988). Geriatric medicine: A new method of measuring bed usage a theory for planning. Yüksek Lisans Tezi, Londra Üniversitesi, Londra

Mujumdar, A. and Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms.Procedia Computer Science, 165, 292–299.

Natarajan P., Frenzel C. J., Smaltz H. D., (2017). Demystifying big data and machine learning for healthcare, CRC Press

Nick Johns, (1998). What is this thing called service? 958-973

Quantin, C., Sauleau, E., Bolard, P., Mousson, C., Kerkri, M., Lecomte Brunet, P., Moreau, T. Ve Dusserre, L., (1999). Modeling of high-cost patient distribution within renal failure diagnosis related group, Journal of Clinical Epidemiology, 52(3): 251-258. doi:10.1016/S0895-4356(98)00164-4

Saxena A., Chandra S., (2021). Artificial intelligence and machine learning in healthcare, Springer Press

Sayım F., Aydın V., (2011). Hizmet Sektörü Özellikleri Ve Sistematik Olmayan Risklerin Sektör Menkul Kıymetleri İle Etkileşimine Dair Teorik Bir Çalışma

Silahtaroğlu G. (2020). Bilgisayar Bilimleri Veri Madenciliği Kavram ve Algoritmaları, Papatya Bilim Akademik Yayınevi

Statista. (2016, Ekim). Share of big data and business analytics revenues worldwide in 2016, by Industry.

Tunçsiper Ç., Bakar A., (2023). Sağlık ekonomisi çerçevesinde sağlık harcamaları: türkiye örneği, s 20-28

Turner B S. (2006). Hospital, 573-579.

Türk Dil Kurumu. (2024). Hastalık. Erişim 14.08.2024, https://sozluk.gov.tr/.

Uğuz S.,(2023). Makine Öğrenmesi teorik yönleri ve Python uygulamaları yapay zeka ekolü, Nobel Akademik Yayıncılık Eğitim Danışmanlık Tic.Ltd.Şti

Wang, K., Yau, K., ve Lee, A., (2002). A hierarchical poisson mixture regression model to analyse maternity length of hospital stay. Statistics in Medicine, 21(23): 3639-3654.

Xiao, J., Lee, A., ve Vemuri, S.,1999. Mixture distribution analysis of length of hospital stay for efficient funding. Socio-Economic Planning Sciences, 33: 39-59.

Yanar.G. (2024). Sağlık hizmeti sunumunda büyük veri ve makine öğrenmesi algoritma uygulamaları. Sağlık Bilimleri Üniversitesi Hamidiye Sağlık Bilimleri Enstitüsü, Doktora Tezi

Yeşilbudak M., Kahraman T., H., Karacan H. (2010). Veri madenciliğinde nesne yönelimli birleştirici hiyerarşik kümeleme modeli, Gazi Üniv. Müh. Mim. Fak. Der., s-26- 39

Yorgancılar, F.E. ve Özlük, B. (2022). Hemşirelik hizmetlerinde yönetsel sorun çözme ve karar verme üzerine bir derleme, Genel Sağlık Bilimleri Dergisi, 4, 68-80.

Yorulmaz, R. ve Erdem, R. (2021). Sağlıklı yaşam üzerine kavramsal bir çerçeve. Uluslararası Sağlık Yönetimi ve Stratejileri Araştırma Dergisi, 7, 57-74.

ÖZGEÇMİŞ

Serhat DOĞAN, 2014 yılında Seyhan Rotary Anadolu Lisesi'nden mezun oldu. Lisans eğitimini Çukurova Üniversitesi Endüstri Mühendisliği Bölümü'nde Ocak 2022'de tamamladı. Ocak 2023-Ağustos 2023 tarihleri arasında metal sanayi sektöründe silo üretimi yapan bir firmada Planlama Uzmanı olarak görev yaptı. Ağustos 2023-Aralık 2023 tarihleri arasında yine metal sanayi sektöründe silo üretimi yapan bir firmada Proje Sorumlusu olarak çalıştı. Ekim 2024 itibarıyla bir devlet üniversitesinde Araştırma Görevlisi olarak akademik kariyerine devam etmektedir.

EKLER

EK-1: Acil Servis Veri Seti (bir kesit)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Patient Id	Patient Admission Date	Patient First Inital	Patient Last Name	Patient Gender	Patient Age	Patient Race	Department Referral	Patient Admission Flag	Patient Satisfaction Score	Patient Waittime	Patients CM
2	766-20-3662	1.04.2023 01:13	N	Sparsholt	M	60	Two or More Races	Cardiology	TRUE	1	35	0
3	701-64-7880	1.04.2023 01:21	L	Brothwell	M	18	Two or More Races	None	TRUE	3	40	0
4	259-51-9518	1.04.2023 02:45	T	Rolingson	M	62	Asian	None	FALSE		12	0
5	125-52-4944	1.04.2023 04:34	N	Sommerscales	M	79	White	General Practice	FALSE		21	0
6	628-45-6492	1.04.2023 06:07	L	McCarrison	F	6	White	None	TRUE		52	0
7	576-82-1180	1.04.2023 07:16	X	Bassford	M	69	Asian	General Practice	TRUE		50	0
8	874-57-8352	1.04.2023 09:49	V	Matyushenko	M	38	White	None	FALSE		50	0
9	170-60-7747	1.04.2023 13:25	T	Snowdon	M	31	White	General Practice	FALSE	0	56	0
10	161-03-2325	1.04.2023 13:31	Q	Everson	F	26	Declined to Identify	None	FALSE	8	44	0
11	428-40-4491	1.04.2023 14:07	J	Antonnikov	F	46	African American	None	TRUE		45	0
12	689-22-4635	1.04.2023 14:08	V	Game	M	53	African American	None	TRUE	7	46	0
13	389-14-6190	1.04.2023 15:10	E	Bygate	F	20	Asian	General Practice	FALSE		30	0
14	338-11-7529	1.04.2023 17:41	B	Grzelczak	F	38	African American	None	TRUE		54	0
15	868-10-5414	1.04.2023 18:16	R	Fulker	F	55	African American	None	TRUE		24	0
16	111-52-9197	1.04.2023 19:07	G	Hawkeswood	F	64	White	None	FALSE		59	0
17	889-96-0276	1.04.2023 22:10	Q	Purseglove	M	10	Two or More Races	Cardiology	FALSE		45	0
18	570-93-0122	1.04.2023 22:29	X	Olech	F	43	Declined to Identify	Orthopedics	FALSE		47	0
19	106-45-9457	1.04.2023 22:49	I	Nicol	F	71	Two or More Races	General Practice	TRUE		14	0
20	417-42-4910	1.04.2023 23:46	V	Clelle	F	79	Two or More Races	None	FALSE		45	0
21	771-03-7336	2.04.2023 00:43	E	Sessions	F	29	White	General Practice	FALSE	5	26	0
22	563-50-3688	2.04.2023 03:22	W	Reburn	F	26	White	None	TRUE		12	0
23	262-69-4482	2.04.2023 05:05	U	Llopis	M	38	Declined to Identify	Gastroenterology	TRUE		14	0
24	682-34-0219	2.04.2023 05:22	N	Trenam	M	38	Native American/Alaska Native	Gastroenterology	TRUE		36	0
25	248-21-1787	2.04.2023 09:15	C	Benez	F	60	White	Physiotherapy	TRUE		60	0
26	476-32-0713	2.04.2023 13:07	K	Bunhill	M	68	White	None	TRUE		24	0
27	372-25-7518	2.04.2023 15:52	K	Hambly	F	67	Declined to Identify	General Practice	TRUE		22	0
28	729-10-9426	2.04.2023 17:50	Z	Brotherhed	M	6	Asian	General Practice	FALSE		45	0
29	732-17-2371	2.04.2023 19:40	H	Halmkin	M	57	Declined to Identify	None	TRUE	10	33	0