T.C.
BİTLİS EREN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

R PROGRAMI İLE METİN MADENCİLİĞİ ÜZERİNE UYGULAMA

YÜKSEK LİSANS TEZİ

MEHMET ŞULAN

DANIŞMAN
DOÇ. DR. FAHRETTİN ÖZBEY

ARALIK 2023
BİTLİS

T.C.
BİTLİS EREN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

R PROGRAMI İLE METİN MADENCİLİĞİ ÜZERİNE UYGULAMA

YÜKSEK LİSANS TEZİ

MEHMET ŞULAN
ORCID: 0000-0001-5964-3340

DANIŞMAN
DOÇ. DR. FAHRETTİN ÖZBEY

ARALIK 2023
BİTLİS

T.C.

BİTLİS EREN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ÇALIŞMASI

ETİK BEYANI

Lisansüstü Eğitim Enstitüsü İstatistik Anabilim Dalı Yüksek Lisans öğrencisiyim. Hazırlamış olduğum "R PROGRAMI İLE METİN MADENCİLİĞİ ÜZERİNE UYGULAMA" başlıklı tezde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi; tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu; tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi; kullanılan verilerde herhangi bir değişiklik yapmadığımı; bu tezde sunduğum çalışmanın özgün olduğunu bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim. 19/12/2023.

Mehmet ŞULAN

T.C.

BİTLİS EREN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TEZ YAZIM KILAVUZU UYGUNLUK BEYANI

"R PROGRAMI İLE METİN MADENCİLİĞİ ÜZERİNE UYGULAMA" başlıklı yüksek lisans tezi Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü Tez Yazım Kılavuzuna uygun olarak hazırlanmıştır. 19/12/2023.

Tezi Hazırlayan	**Danışman**
İmza	İmza
Mehmet ŞULAN	Doç. Dr. Fahrettin ÖZBEY

İstatistik Anabilim Dalı Başkanı

İmza

Doç. Dr. Hamit MİRTAGİOĞLU

T.C.

BİTLİS EREN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TEZ ONAY SAYFASI

Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü İstatistik Anabilim Dalı öğrencisi Mehmet ŞULAN tarafından hazırlanan "R PROGRAMI İLE METİN MADENCİLİĞİ ÜZERİNE UYGULAMA" adlı Yüksek Lisans Tezi ile ilgili tez savunma sınavı, 19/12/2023 tarihinde yapılmış ve tezin oy birliği ile kabul edilmesine karar verilmiştir.

JÜRİ: **İMZA**

Danışman: Doç. Dr. Fahrettin ÖZBEY .....
(Bitlis Eren Üniversitesi)

Üye: Doç. Dr. Hamit MİRTAGİOĞLU .....
(Bitlis Eren Üniversitesi)

Üye: Doç. Dr. Gökhan GÖKDERE .....
(Fırat Üniversitesi)

Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü Yönetim Kurulunun …… tarih ve …… sayılı kararıyla jüri tarafından kabul edilmiş bu çalışmanın yüksek lisans olarak kabulü onaylanmıştır.

…./…../2023

Doç. Dr. Mehmet Bakır ŞENGÜL

Enstitü Müdürü

T.C.

Bitlis Eren Üniversitesi Lisansüstü Eğitim Enstitüsü

İstatistik Anabilim Dalı

R PROGRAMI İLE METİN MADENCİLİĞİ ÜZERİNE UYGULAMA

Yüksek Lisans Tezi

Mehmet ŞULAN

Danışman: Doç. Dr. Fahrettin ÖZBEY

Aralık 2023

ÖZET

Veri madenciliği yöntemi sayısal özellikli veya sayısal bir biçimde temsil edilen veri setlerinde kullanılan verileri istatistiksel bir şekilde analiz ederek sonuca ulaşmayı hedefler. Metin madenciliği ise metinlerin analizi ile yapısal olmayan, word, pdf formatlarındaki metin dosyaları ile sosyal medya paylaşımları ve blog dosyaları gibi elektronik metin ve belge yığınları arasından daha önceden keşfedilmeyen ve potansiyel olarak kullanılabilir verileri yapısal ve düzenli bir şekilde elde etme işlemidir. Metin madenciliği istatistiksel olarak metin üzerinden sonuca ulaşmayı hedeflemektedir.

Metin halindeki bir veriden bilgi çıkarımının yapılabilmesi için öncelikle bazı işlemlerin gerçekleştirilmesi ve yapısal şekilde bulunmayan metinsel verilerin dönüştürülerek yapısal bir hale çevrilmesi gereklidir. Daha sonra yapısal hale dönüştürülen metinsel verilerin veri madenciliği yöntemlerinin uygulanılabileceği hale çevrilmiş olunur. Metinlerin toplanması ile başlayan bu süreç, toplanan metinlerin bazı veri ön işleme ve dönüştürme sürecinden sonra istatistik ve veri madenciliği yöntemlerinin kullanılmasıyla anlamlı bilgiye ulaşılır.

Tez çalışmasında veri madenciliği ve metin madenciliği uygulama ve yöntemlerinden bahsedilip YÖK Akademi ve Dergipark Uygulamasından elde edilen veriler R programı kullanılarak ilk etapta veri ve metin madenciliği yöntemleriyle gerekli analizler yapıldıktan sonra kelime bulutları oluşturulmuştur.

Anahtar Kelimeler: Veri madenciliği, Metin madenciliği, R programı, Dergipark uygulaması.

Republic of Türkiye

Bitlis Eren University Graduate School

Department of Statistics

AN APPLICATION ON TEXT MINING WITH THE R PROGRAM

Master'sThesis

Mehmet ŞULAN

Supervisor: Assoc. Prof. Dr. Fahrettin ÖZBEY

December 2023

ABSTRACT

Data mining aims to analyze data in a statistically significant manner, particularly in datasets featuring numerical attributes or those represented in a numerical format. On the other hand, text mining involves the structured and systematic extraction of previously undiscovered and potentially valuable information from electronic text and document collections, such as textual files in formats like word or pdf, social media posts, and blog entries. Text mining seeks to derive meaningful results from text through statistical analysis.

To extract information from textual data, certain processes must first be executed, and non-structurally formatted textual data needs to be transformed into a structured form. Subsequently, the text data transformed into a structured format becomes amenable to the application of data mining methods. Commencing with the collection of texts, this process culminates in meaningful information through the utilization of statistical and data mining methods after subjecting the collected texts to some data preprocessing and transformation procedures.

In the thesis study, data mining and text mining applications and methods are discussed, and data obtained from YÖK Academy and Dergipark Application are subjected to initial analyses using R programming with data and text mining methods. Following the necessary analyses, word clouds were generated.

Keywords: Data mining, Text Mining, R Program, Dergipark Application.

TEŞEKKÜR

Yüksek lisans eğitimi süresince her zaman yanımda olan ve gerekli bilgi ve deneyimlerini benimle paylaşarak tezimi hazırlamamda yardımlarını esirgemeyen kıymetli danışman hocam Doç. Dr. Fahrettin ÖZBEY hocama şükran ve minnettarlığımı sunarım.

Eğitim ve öğretimin her alanında benden yardımlarını esirgemeyen aileme ve gerekli durumlarda teknik desteği sağlayan okul arkadaşım ve dostum olan Nami KURŞUN'a ve her zaman gerekli desteği sağlayan Faruk BİLGİÇ'e sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

ÖZET……I

ABSTRACT……II

TEŞEKKÜR……..III

İÇİNDEKİLER……IV

ÇİZELGELER DİZİNİ……...VII

ŞEKİLLER DİZİNİ……...VIII

1. GİRİŞ……1

1.1. Veri Madenciliğin Literatürdeki Tanımları……...2

1.2. Veri Madenciliği Sürecinin Tarihsel Gelişimi……3

1.3. Veri ve Metin Madenciliği……5

1.4. Metin Madenciliği……6

1.5. Literatür Taraması……...7

2. MATERYAL VE YÖNTEM……10

2.1. Veri Madenciliğinde Kullanılan Yöntemler……10

2.1.1. Tahmin Edici Yöntemler……11

2.1.1.1. Sınıflama Yöntemleri……....……11

2.1.1.1.1. Karar Ağaçları……...11

2.1.1.1.2.Yapay Sinir Ağları……..13

2.1.1.1.3.Bayes Sınıflandırması……15

2.1.1.1.4. Karar Destek Sistemleri……17

2.1.1.1.5. K-en Yakın Komşu……18

2.1.1.1.6. Genetik Algoritmalar……19

2.1.1.1.7. Zaman Serileri Analizi……..20

2.1.1.2. Regresyon Analizi…….....20

2.1.2. Tanımlayıcı Yöntemler……21

2.1.2.1. Kümeleme Yöntemleri……..21

2.1.2.1.1.Bölme Yöntemleri……...22

2.1.2.1.2. Hiyerarşik Yöntemler……22

2.1.2.1.3. Yoğunluk Tabanlı Yöntemler……...23

2.1.2.1.4. Izgara Tabanlı Yöntemler……23

2.1.2.1.5. Model Tabanlı Yöntemler……23

2.1.2.2. Birliktelik Kuralları……..24

2.2. Metin Madenciliğinde Kullanılan Yöntemler ve Uygulamalar……..24

2.2.1. Metin Madenciliğinde Kullanılan Yöntemler……...24

2.2.1.1. Terim Bazlı Yöntem……..24

2.2.1.2. Cümle Tabanlı Yöntem……25

2.2.1.3. Konsepte Dayalı Yöntem……25

2.2.1.4. Model Taksonomi Yöntemi……26

2.2.2. Metin Madenciliği Uygulamaları……...26

2.2.2.1.Bilgi Getirimi/Erişimi……27

2.2.2.2. Metin Kümeleme……..28

2.2.2.2.1 Metin Belgelerinin Kümelenmesi İçin Kullanılan Algoritmalar……28

2.2.2.3. Metin Sınıflandırma……29

2.2.2.4. Web Madenciliği……...30

2.2.2.5. Bilgi Çıkarımı……31

2.2.2.6. Doğal Dil İşleme……31

2.2.2.7. Kavram Çıkarımı……32

2.3. R Programı……..33

2.4. Dergipark Uygulaması……...34

3. BULGULAR VE TARTIŞMA ……36

3.1. Veri Toplama Süreci……..36

3.2. Veri analizi……..37

3.3. Metin Analizi……...44

4. SONUÇ VE ÖNERİLER……..57

KAYNAKLAR…..…….59

ÇİZELGELER DİZİNİ

Çizelge 3.1. Ünvan dağılımına göre toplam akademisyen sayıları…….37

Çizelge 3.2. Yayını bulunan ve bulunmayan akademisyen sayılar…….38

Çizelge 3.3. Ünvan dağılımına göre dergipark yayın sayıları…….38

Çizelge 3.4. Bilim alanı dağılımına göre toplam akademisyen sayıları……..39

Çizelge 3.5. Bilim alanı dağılımına göre ünvan sayıları…….40

Çizelge 3.6. Bilim alanı dağılımına göre yayını bulunan akademisyen sayıları ve dergipark yayın sayıları……..41

Çizelge 3.7. Bilim alanı ve ünvan dağılımına göre dergipark yayın sayıları……..42

Çizelge 3.8. Dergipark yayın sayısına göre en yüksek ilk 6 üniversite ve bilim alanına göre dağılımı……...44

Çizelge 3.9. Biyoloji bilim alanında en çok kullanılan 100 kelime……45

Çizelge 3.10. Fizik bilim alanında en çok kullanılan 100 kelime……...47

Çizelge 3.11. Kimya bilim alanında en çok kullanılan 100 kelime……49

Çizelge 3.12. İstatistik bilim alanında en çok kullanılan 100 kelime……..51

Çizelge 3.13. Matematik bilim alanında en çok kullanılan 100 kelime……..53

Çizelge 3.14. Moleküler biyoloji ve genetik bilim alanında en çok kullanılan 100 kelime……..55

ŞEKİLLER DİZİNİ

Şekil 1.1. Veri madenciliğinin tarihsel gelişimi……4

Şekil 1.2. Metin madenciliğinin işlenmesi……7

Şekil 2.1. Veri madenciliğinde kullanılan yöntemler……...10

Şekil 2.2. Karar ağacının yapısı……...11

Şekil 2.3. Yapay sinir hücresi……..14

Şekil 2.4. Karar destek sisteminin yapısı……17

Şekil 2.5. Metin madenciliği uygulamaları……..…27

Şekil 2.6. Bilgi çıkarma……...31

Şekil 3.1. Veri ve metin analiz süreci……..36

Şekil 3.2. Ünvan ve dergipark yayın sayılarının grafiksel gösterimi……..39

Şekil 3.3. Bilim alanı dağılımına göre toplam akademisyen sayılarının grafiksel gösterimi……..40

Şekil 3.4. Bilim alanı dağılımına göre ünvan sayılarının grafiksel gösterimi……41

Şekil 3.5. Bilim alanlarına göre akademisyen ve yayın sayılarının grafiksel gösterimi……..42

Şekil 3.6. Bilim alanı, akademisyen sayısı ve yayın sayısının grafiksel gösterimi…….43

Şekil 3.7. Dergipark yayın sayısına göre ilk altı üniversite……43

Şekil 3.8. Biyoloji bilim alanı kelime bulutu……..46

Şekil 3.9. Fizik bilim alanı kelime bulutu……...…48

Şekil 3.10. Kimya bilim alanı kelime bulutu……...50

Şekil 3.11. İstatistik bilim alanı kelime bulutu……52

Şekil 3.12. Matematik bilim alanı kelime bulutu……54

Şekil 3.13. Moleküler biyoloji ve genetik bilim alanı kelime bulutu……..56

1. GİRİŞ

Çağdaş dünyada metin, bilgi alışverişi için kullanılan en yaygın araçtır. Zaman içinde teknolojinin gelişmesiyle beraber metinlerin kullanımında da hızlı bir yükseliş meydana gelir. Bunun sonucunda metin belgelerinin saklanması için bilgisayarların kullanımı artar. Böylece bilgisayarlarda belgeler biçiminde önemli miktarda veri depolanır. Bilgisayarlarda depolanan bu veriler yapılandırılmış, yarı yapılandırılmış ve yapılandırılmamış formlardan herhangi biri olabilmektedir. Veri tabanlarında depolanan veriler ise yapılandırılmış veri kümelerine bir örnektir. Yarı yapılandırılmış ve yapılandırılmamış veri kümelerine örnek olarak HTML dosyaları, e-postalar ve tam metin belgeleri gösterilebilir. Metin Madenciliği de yapılandırılmamış metin belgelerinden gizli, yararlı ve ilginç kalıpları keşfetme süreci olarak tanımlanır (Sumathy & Chidambaram, 2013: 29).

Yapılandırılmamış metinden bilgi alma ve yararlı kalıpları çıkarmak, özel işleme yöntemleri ve algoritmalar gerektiren çok büyük bilgiler içerdiğinden dolayı metin madenciliği karmaşık bir yapıya sahiptir. Veri madenciliğinin uzantısı veya yapılandırılmış veri tabanlarından bilgi keşfi olarak da görülebilir. Düzenli metin madenciliği ile veri madenciliği arasındaki fark ise metin madenciliğinde kalıpların, gerçeklerin yapılandırılmış veri tabanlarından değil de doğal dil metninden çıkarılma işlemidir (Tan, 2000: 1; Hearst, 2003; Sumathy & Chidambaram, 2013: 29).

Wep aramasında da aşina olduğumuzdan farklı olan ve geniş teorik yaklaşımlar ve yöntemler alanını içeren metin madenciliği temelde giriş bilgisi olarak metni kullanır. Metin madenciliği, veri madenciliği, dil bilim, bilgisayar bilimi ve hesaplamalı istatistik gibi disiplinler arası bir alan olarak da kabul edilir. Metin madenciliği alanında klasik uygulamalar, belge kümeleme ve belge sınıflandırmadır ve bu iki yaklaşım, veri madenciliği topluluğundan gelmektedir. Her iki yaklaşımda da metin terim frekanslarına dayalı yapılandırılmış bir formata dönüştürülür ve ardından standart veri madenciliği teknikleri uygulanır. Belge kümeleme alanında yaygın olarak kullanılan uygulamalar arasında haber makalelerinin veya bilgi servisi belgelerinin gruplandırılması yer alırken, metin sınıflandırma yöntemleri örneğin e-posta filtreleri ve iş kütüphanelerindeki belgelerin otomatik olarak etiketlenmesinde kullanılmaktadır (Feinerer, Hornik & Meyer, 2008; Hearst, 2003).

Özellikle kümeleme bağlamında, belirli mesafe ölçüleri önemli bir rol oynamaktadır. Burada, kullanıcının yapısı belirsiz bir sorgusu, öncelikle yapılandırılmış bir formata dönüştürülür ve ardından bir veri tabanından gelen metinlerle eşleştirilir. Sonraki aşamada ise, veri tabanının oluşturulması için yapılandırılmamış giriş verilerinin bilgi kalitesi ve yapısal gereksinimleri karşılamak üzere normalleştirilmesi gerekmektedir. Bu süreç genellikle dil bilimsel ayrıştırma adımlarını içermektedir (Feinerer, Hornik & Meyer, 2008).

Çeşitli alanlarda da daha yenilikçi metin madenciliği yöntemleri kullanılmakta ve yeni metotlar geliştirilmektedir. Örneğin, dil bilimsel stilometri alanında, bir yazarın belirli bir metni yazma olasılığı, yazarın yazım stili analiz edilerek hesaplanmaktadır. Ayrıca, arama motoru kullanıcı davranışlarının kayıtlarından elde edilen bilgilerle, arama motorlarında belge sıralamalarının öğrenilmesi için metin madenciliği yöntemleri kullanılmaktadır (Feinerer, Hornik & Meyer, 2008).

1.1. Veri Madenciliğin Literatürdeki Tanımları

Veri madenciliği temel olarak veri setleri arasında yer alan desen, örüntülerin ya da düzen verilerinin analiz edilmesi ve yazılım tekniklerinin uygulanmasıyla ilişkilendirilir. Teorik olarak da veri madenciliği bilgi keşfinin bir aşaması olarak kabul edilir fakat gerçekte ikisi aynı anlamda kullanılır (Lee & Stolfo, 1998).

Veri madenciliğinin temel amacı veri depolarında saklanılan birden fazla çeşitli verilere dayanılarak önceden ortaya çıkarılmamış bilgilerin meydana getirilmesidir. Bu bilgiler de karar vermeye yardımcı olmak ve eylem planlarını gerçekleştirme ve ayrıca yüksek miktardaki veri içerisinden gelecek ile ilgili tahminler yapma imkânı sağlayan ilişki ve kuralların aranması işlemidir. Veri madenciliği veriler içerisindeki ilişkilerin, kuralların, desenlerin, değişimlerin, düzensizliklerin ve ayrıca önemli yapıların istatistiksel olarak yarı otomatik şekilde keşfedilme işlemidir. Aslında veri madenciliği bilgi keşif sürecinin başlaması şeklinde kabul edilir (Dener, Dörterler & Orman, 2009: 788). Bu genel kabulle beraber bilginin hızlı ve dinamik bir şekilde veri yığınları arasından kullanılabilir verinin ortaya çıkarılmasıyla verinin analiz edilip yorumlanma işlemi yapılır.

Veri madenciliği, yüksek miktardaki veri setlerinden değişkenler arasındaki önemli bağlantıları araştıran ve farklı disiplinlerce kullanılan yöntemdir (Erdoğan & Timor, 2005: 53).

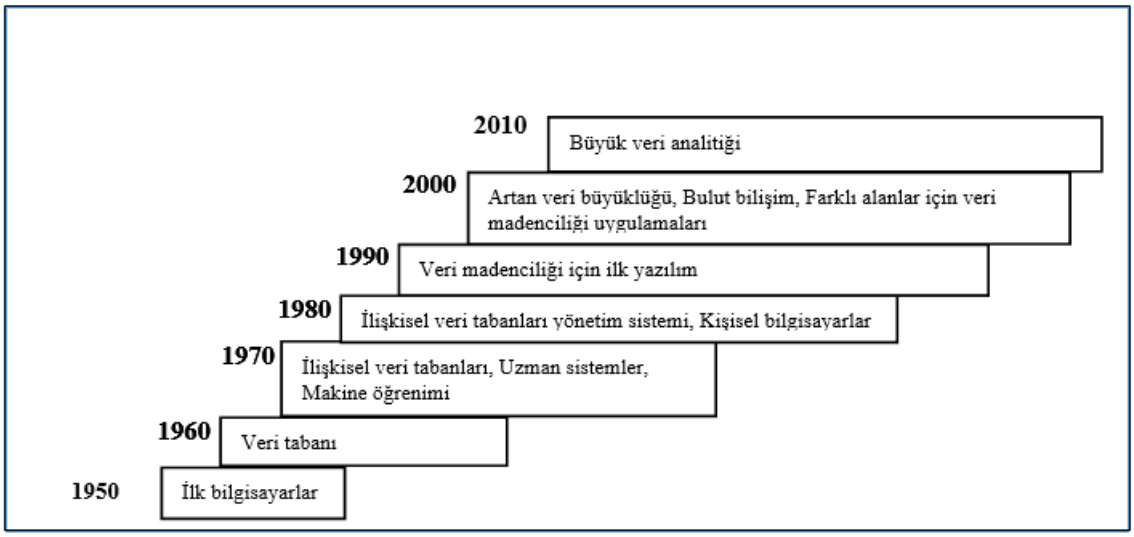
Veri madenciliği, toplanılan verilerden anlamı olan bilgileri bularak veri içinde gizli kalmış birtakım eğilimleri ve örüntüleri tespit ederek ve birden fazla değişkenler arasında bulunan ilişkileri bularak kararların verilmesi için kullanılan bir yöntemdir. Veri madenciliği makine öğrenimi, yapay zekâ, istatistik ve birçok değişik alanda geliştirilmiş teknik ve yöntemlerin kullanımıdır (Seyrek & Ata, 2010: 71).

Veri madenciliğinde daha önceden bilinmeyen geçerliliği olan ve uygulanabilen veriler, veri yığınları arasından hızlı ve dinamik bir süreçle elde etme işlemi olarak tanımlanır (Baykal, 2006: 96).

Veri madenciliği, büyük miktardaki veri yığınları içerisinde bulunan anlamlı fakat ortaya çıkarılmamış bilginin, kuralların veya örüntülerin otomatik ya da yarı otomatik yöntemler yardımı ile ortaya çıkarılması işlemidir. Özellikle büyük miktardaki veri yığınları arasından anlamı olan bilgileri bulmak tek bir aşamayla ve kolayca yapılamamaktadır. Buna ancak belli bir çerçeve ile yaklaşmak ve kademeli bir şekilde ilerlemek gerekir. Günümüze kadar birbirini hızlı bir şekilde takip eden bu süreç ile ifade edilen veri madenciliği günümüzde ise bir döngü şeklinde ilerlemektedir. Anlamlı bilginin tek seferde bulunması artık yeterli gelmemektedir. Yeni verilerin artmasıyla anlamı olan bilginin ortaya çıkarılması için sürecin yeniden başlatılması ve döngünün devam ettirilmesi gerekir (Takçı, 2020: 39-40).

1.2. Veri Madenciliği Sürecinin Tarihsel Gelişimi

Veri madenciliği sürecinin tarihsel gelişimi Şekil 1.1'de gösterildiği özere ABD ordusu için 1950'lerde, 2. Dünya Savaşı sırasında geliştirilen ilk bilgisayar olan ENIAC ile başlamıştır. Geliştirilen bu bilgisayarla birlikte veri madenciliğinin teknikleri üzerinde çalışan matematikçiler, bilgisayar bilimleri ve mantıkta makina öğrenimi ve yapay zekâ alanlarını ortaya çıkarmaya başladılar. İlk etapta karmaşık hesaplamaların gerçekleştirilmesi için kullanılan bilgisayarlar ilerleyen süreçlerde de veri depolama amacı ile kullanılmaya başlandı. Verilerin depolanması ile birlikte 1960'larda veri tabanı ortaya çıktı. İstatistikçiler bu sayede yeni bir algoritmayı keşfettiler. Veri madenciliğinin ilk adımları olan bu algoritmalar en büyük olabilirlik kestirimi ve regresyon analizidir (Şaylan, 2013: 3; Akgün & Özek, 2020: 199).



Şekil 1.1. Veri madenciliğinin tarihsel gelişimi (Akgün & Özek, 2020: 199)

1970'lere gelindiğinde ise veri tabanlarının düzenlenmesi ve fazla bilginin depolanması ihtiyacı veri modellemeyi ortaya çıkarır. Deneyimlerin ve örneklerin problem çözmede kullanılmasıyla aynı yıllarda bilgisayarları programlayan makine öğrenmesi gelişimine başlanır. Bu sayede veri tabanındaki bilginin keşfi ile ilk adımları atılmış olup bununla birlikte büyük veri tabanları için veri ambarları geliştirilir. 1990'da istatistik, veri tabanı ve makine öğrenmesi disiplinlerinin birlikte çalışmasıyla veri madenciliğinin ilk yazılımı geliştirilmeye başlanır. Aynı zamanda yeni teknolojilerin ortaya çıkmasıyla beraber veri madenciliği yaygın olarak kullanılmaya başlanmıştır. 2000'li yıllara gelindiğinde ise video, ses, yazı gibi farklı türlerden verilerin bulut ortamlarında depolanmaya başlanmasıyla ilerleyen yıllarda da bu tür verilere ek olarak sürekli bir şekilde artan ve değişen hızda konum bilgisi, log dosyaları, sosyal medya paylaşımları, bloglar da dâhil oldu. Ve ayrıca hız, çeşitlilik ve hacim bakımından artan günümüz verilerinde Büyük Veri Analitiği yaklaşımının ortaya çıkmasına neden oldu (Şaylan, 2013: 3; Akgün & Özek, 2020: 199).

Öte yandan verilerin biriktirilmesi ile bu veriler içerisinden elde edilebilecek önemli bilgilerin gelecek için de tahmin fırsatını sunması ve ortaya çıkabilecek bazı olumsuz durumlarda önceden tedbir alınabilme olanağını tanıması veri madenciliğini önemli kılmış ve bunun yanında kullanım alanlarını da yaygınlaştırmıştır (Akgün & Özek, 2020: 199).

1.3. Veri ve Metin Madenciliği

Veri madenciliği makine öğrenimi, veri tabanı yönetim sistemleri ve istatistik gibi birden fazla disiplin yöntemleri kullanılarak daha önce ortaya çıkarılmamış ve ilk bakışta görülmeyen bilgilerin ortaya çıkarılması için kullanılabilen bir veri analiz yöntemidir. Metin madenciliği de veri madenciliği disiplini içerisinde bulunan ve bu disiplinin alt bir dalı olarak isimlendirilerek metinsel dokümanlar arasından ortaya çıkarılmamış bilgileri bilgisayar sistemleri sayesinde ortaya çıkartılmasında kullanılan bir bilim dalıdır (Çalış, Gazdağı & Yıldız, 2013: 2-3).

Veri madenciliği yöntemi, çoğunlukla sayısal özellikli veya sayısal bir biçimde temsil edilen veri setlerinde kullanılan yöntemler bütünü için kullanılırken metin madenciliği, başta internet olmak üzere elektronik ortamda yaygınlaşan sözcük içeren yapılandırılmış ya da yarı yapılandırılmış ortamların ayrıntılı incelenmesinde tercih edilir. Sözcüksel örüntülerin keşfi başta olmak üzere metin birimlerinin yapısal ve anlamsal görünümleriyle ilgilenilmekle birlikte (Tahiroğlu, 2021: 320) metin madenciliği çalışmalarında çoğunlukla istatistiksel bir şekilde metin üzerinden sonuçlara ulaşmayı hedeflemektedir(Seker, 2015: 31).

Büyük veriler yapılandırılmış ya da yapılandırılmamış veri şeklinde olabilir. Yapılandırılmış verilerde satırlardaki sütun sayıları sabit iken yapılandırılmamış verilerde ise her bir satırda farklı sayıda sütun bulunabilir. Veri madenciliğinde çoğunlukla yapılandırılmış veriler ile çalışılırken metin madenciliğinde ise yapılandırılmamış veri ile çalışılmaktadır. Büyük veriler üzerinde metin madenciliği uygulaması geliştirilebilmesi için seçim, aktarım, barınma, işlem ve çıktı aşamalarının oluşturulması gerekir (Zontul & Aydın, 2017: 104).

Metin madenciliği, doğal dil ile yazılmış büyük metinler arasından ilişkili bilgileri bulup titizlikle ortaya çıkaran ve ilginç ilişkiler, söz dizimsel korelasyon veya semantik ilişki arayışında olan bir veri madenciliği yöntemidir(Toprak, 2018: 12). Bu yöntem veri madenciliği alt dalı olmakla beraber kullandığı araç ve girdiler oldukça farklı bir şekildedir. Veri madenciliği girdileri çoğunlukla veri tabanları ya da veri dosyalarında bulunan tablo şeklindeki verilerden oluşurken metin madenciliği girdisi ise tablo formatında olmayan dosyalardır. Bunlar web sayfaları, HTML, PDF ya da Word dokümanlarından oluşmaktadır. Başka bir deyişle metin madenciliği, bir metin içerisinde daha önceden bilinmeyen kavramların analiz edilerek bulunması işlemidir (Atan, 2020:

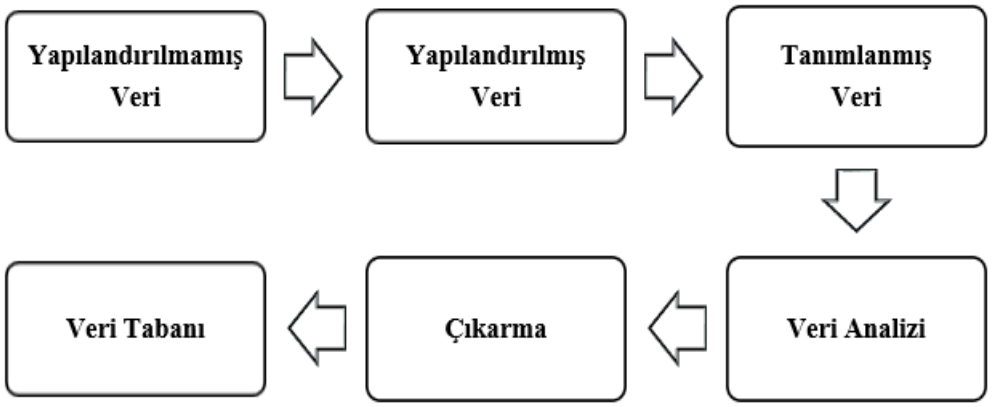
224). Bu işlem ile metin madenciliğinde üretilen anlamlı bilgilerin kullanılması ile kurum ve kuruluşların faydalanacağı çeşitli sonuçlara ulaşılabilir (Kılınç, Borandağ, Yücalar, Tunalı, Şimşek & Özçift, 2016: 90).

1.4. Metin Madenciliği

Metin, bilgi depolamanın en doğal şekli olduğundan dolayı metin madenciliğinin veri madenciliğinden daha yüksek bir ticari potansiyele sahip olduğu düşünülmektedir. Bununla birlikte, metin madenciliği doğası gereği yapılandırılmamış ve bulanık metin verileriyle uğraşmayı içerdiğinden (veri madenciliğinden) çok daha karmaşık bir yapıya sahiptir (Tan, 2000; 1). Metin madenciliği, birbirinden farklı yazılı belgelerden bilgileri otomatik bir şekilde çıkararak daha önceden bilinmeyen bilgilerin bilgisayar tarafından keşfedilme işlemidir (Hearst, 2003: 1). Veri madenciliğinin bir alt türü olmakla birlikte word formatında bulunan metin dosyaları, pdf, sosyal medya paylaşımları ve bloglar gibi yapısal halde bulunmayan metinlerden oluşan veri tabanlarındaki anlamlı bilgiyi elde etmek amacıyla kullanılmaktadır. Bu amaç doğrultusunda metin içinde bulunan önemli, anlaşılabilen ve kullanılabilir örüntülerin belirlenip analiz edilme işlemidir (Ünlü, 2022: 6).

Metinlerin analizi ile anlamlı bilgilere ulaşılabilmesi için öncelikle verilerin ön işlem ve özellik çıkartımı gibi bazı aşamaları uygulaması gereklidir. Yapılan bu adımlardan sonra yapısal halde bulunmayan veriler, metin madenciliğinin kullanılacağı ve bilgisayarlar tarafından işleneceği yapısal bir biçime dönüştürülerek büyük miktardaki veriler arasında bulunan değerli bilgiler keşfedilmiş olunur (Kılınç vd., 2016: 90).

Metin halindeki bir veriden bilgi çıkarımının yapabilmesi için öncelikle bazı işlemlerin yapılması gerekmektedir. Yapılan bu işlemlerle birlikte yapısal bir halde bulunmayan metinsel veriler yapısal bir hale dönüştürülür. Böylece metinsel veriler veri madenciliği tekniklerinin uygulanabileceği uygun formata dönüştürülür(Kılınç vd., 2015: 3). Metin belgelerinin toplanmasıyla başlayan bu süreç, toplanan metinlerin birtakım veri ön işleme ve dönüştürme süreçlerinden geçirilmesini içeren, istatistiksel ve veri madenciliği yöntemleri aracılığıyla anlamlı bilgiye ulaşmayı hedefleyen, akış şeması Şekil 1.2'de sunulan altı adımlık bir süreçtir (Ünlü, 2022: 6).



Şekil 1.2. Metin madenciliğinin işlenmesi (Dang & Ahmad, 2014: 22)

Metin Madenciliği Adımları

A- Yapılandırılmamış verilerden bilgi toplanması

B- Alınan bu bilgileri yapılandırılmış verilere dönüştürülme işlemi

C- Yapılandırılmış verilerden modelin tanımlanması

D- Modelin analiz edilmesi

E- Değerli bilgilerin ayıklanması

F- Veri tabanında saklanılması (Dang & Ahmad, 2014: 22) olarak altı adımdan oluşmaktadır.

Veri madenciliğinin bir parçası olarak düşünülen metin madenciliği aslında alışılagelen veri madenciliğinden oldukça farklıdır. Temel farklılık ise metin madenciliği yönteminde örüntülerin olay tabanlı veri tabanlarından değil de doğal dil metinlerinden çıkarılması işlemidir. Basit bir şekilde tanımlayacak olursak metni veri kaynağı şeklinde kabul eden metin madenciliği, veri madenciliği alanının bir çalışmasıdır. Metin üzerinden yapısallaştırılmış veri elde etmeyi amaçlamaktadır (Balıkesir Üniversitesi, 2023).

1.5. Literatür Taraması

Kılınç vd., (2016) yaptıkları bilimsel makale tasnifi çalışmasında 50 farklı dergiden 2000 adet makale özet bilgilerini toplayıp bu özetlerden oluşturdukları veri seti üzerinden metin madenciliği çalışması gerçekleştirmişlerdir. Elde ettikleri veri setini değerlendirmek için de R dili ve R Studio aracını ve K-en algoritmasını kullanmışlar. Oluşturdukları veri setinin sınıflandırma bilgisini gizleyerek K-en sınıflandırması

yaparak doğruluk performans oranlarını elde ettiklerinin belirtmişler. Gerçekleştirdikleri deneysel çalışmanın sonucunda %96.67 oranında doğruluk (ACC) değerini bulduklarını ve yayınların hangi sınıfa ait olduğunu tespit etmişlerdir. K-en algoritmasının doğruluk oranlarını karşılaştırmak içinde J48 algoritmaları ve Naive Bayes algoritmasını da kullandıklarını ve metin madenciliği alanında K-en algoritması oluşturulan veri seti üzerinde kullanıldığında, karşılaştırdıkları diğer algoritmalara göre daha iyi sonuç alındıklarını belirtmişlerdir.

Sütçü vd., (2022) yaptıkları çalışmada metin madenciliği ve içerik analiz yöntemiyle twitter uygulamasında aşı karşıtı olan paylaşımları incelemişlerdir. Twitter API anahtarı kullanarak toplam 5000 tweet özerinden yaptıkları veri hazırlama ve ön işlem aşamalarından sonrasında kalan 1258 tweet'i incelemişler. Kullanıcıların tweetlerin çoğunda maske, plandemi ve biontech yan etki gibi ifadelerin sıklıkla tekrar edilmiştir. Daha sonra yaptıkları analizler sonucunda aşı karşıtlığına dair söylemlerin aşının bir biyolojik silah olduğu, aşının yan etkilerinin zararlı ve riskli olduğu ve pandeminin gerçek olmadığı gibi söylemlerden verilerin kümelendiğini gözlemlemişlerdir.

Buldu (2019) da yaptığı çalışmada metin madenciliği yöntemiyle Türkiye'deki otomotiv firmalarına yönelik şikâyetleri incelemiştir. O, çalışmasında en çok satışı olan 15 otomotiv markasına ait 51599 şikâyet vakasına yönelik verileri internet kazıma işlemiyle elde ettiğini, doğal dil işleme ve metin madenciliği analizi süreçleriyle çözümlediğini belirtmiştir. Metin madenciliği çalışmasındaki frekans analizine ait olan bulguların şikâyetlerin yoğun olarak ilgili olduğu konuları gösterdiğini açıklayan Buldu, şikâyet miktarlarının da markalara göre kategorilere ayrıldığında sıfır satış miktarlarından çok ikinci el satış miktarı ile paralellik gösterdiğini belirtmiştir.

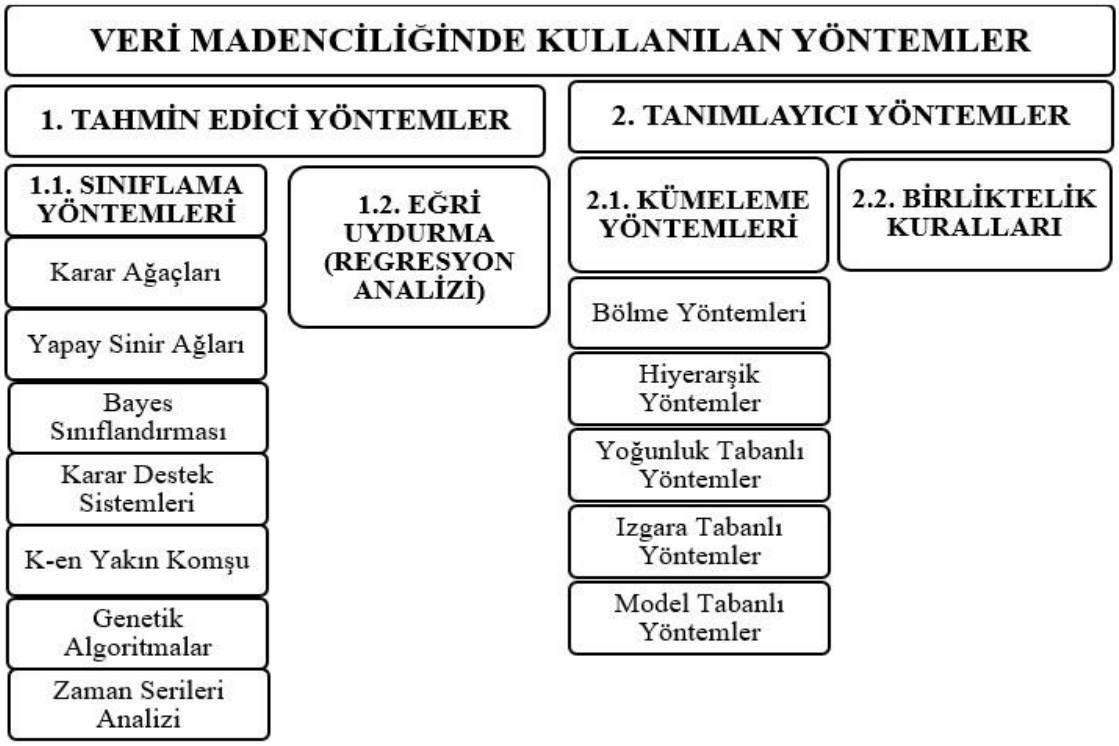
Ünlü (2022) de yaptığı çalışmada metin madenciliği yöntemi içerisinde yer alan terim frekansı ağırlıklandırma tekniğini Türkiye'de ve yurtdışında düzenlenen ve başlığında "Veri Bilimi (Data Science)" ifadesi bulunup kongre dili İngilizce olan kongrelerde sunarak bildiri kitaplarında basılan 1863 bildiri özetini değerlendirmiştir. İncelenen bildiri özetleriyle kongrelerde sunulan bildirilerde sıklıkla tercih edilen veri analiz yöntemlerinin neler olduğu, yıllara göre değişim olup olmadığı ve kongrenin yapıldığı yere göre sıklıkla kullanılan yöntemlerin neler olduğu sunulmuştur. Ünlü yaptığı analiz sonucunda en sık geçen kelime öbeklerini incelediğinde Türkiye'de ve yurt dışında düzenlenen kongrelerin bildiri özetlerinde ilk altı sırada geçen kelime öbeklerinin

aynı olduğunu saptamıştır. İlk altı sıradaki kelime öbeklerinin aynı olması, kongrelerde çalışılan konuların benzer olduğu sonucunu ortaya çıkarmıştır.

2. MATERYAL VE YÖNTEM

2.1. Veri Madenciliğinde Kullanılan Yöntemler

Veri madenciliği yöntemlerinde tahmin edici modeller, sonuçları belli olan verilerde tahminde bulunma süreci iken; tanımlayıcı modeller, karar verme ve rehberlik aşamasında tahminlerde bulunma sürecidir. Tahmin edici modellerde karar ağaçları, Bayes sınıflandırması, yapay sinir ağları, karar destek sistemleri, zaman serileri analizi, k-en yakın komşu ve genetik algoritmalarının yanı sıra sınıflama ve regresyon analizinin de içerisinde bulunduğu istatiksel tahmin tekniklerini içermektedir. Tanımlayıcı modeller ise ardışık zaman örüntüleri ve birliktelik kurallarının da bulunduğu kümeleme analizi ve ilişki analizi yer almaktadır (Artsın, 2020: 345). Aşağıdaki Şekil 2.1. veri madenciliğinde kullanılan yöntemler verilmiştir.



Şekil 2.1. Veri madenciliğinde kullanılan yöntemler

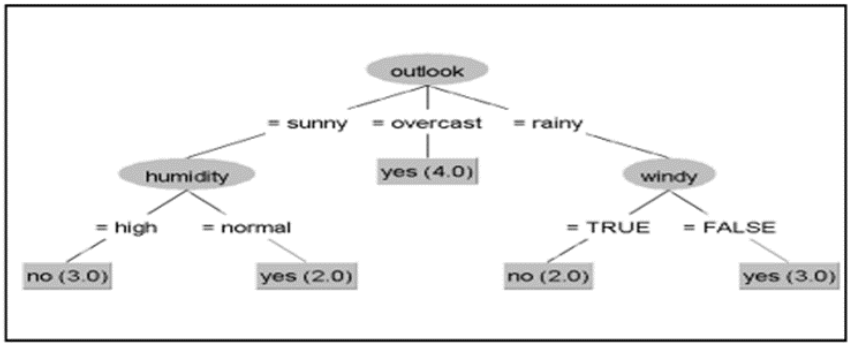
2.1.1. Tahmin Edici Yöntemler

2.1.1.1. Sınıflama Yöntemleri

Sınıflama yöntemi, veri setlerinin içerdiği verilerin ortak özelliklerine göre belirli sınıflara ayrılmasının sağlanmasıdır. Bu amaçla birçok çeşitli algoritma geliştirilmiştir (Özkan & Erol, 2015). Başlıca bu algoritmalar, karar ağaçları, K-en yakın komşu, yapay sinir ağları, karar destek sistemleri, Bayes sınıflandırması, zaman serileri analizi ve genetik algoritmalardır.

2.1.1.1.1. Karar Ağaçları

Önemli sınıflandırma araçlarından olan karar ağaçları öğrenme algoritmasının kullanıldığı basit bir yöntemdir. Karar ağaçları kararları göstermenin yanında kararların açıklamasını da içermektedir. Karar ağacını oluşturan eğitim süreci tümevarım olarak yapılır. Ağaç tümevarım yönteminde öz bilgi keşfi yapılması en yaygın yöntemler arasındadır. Sınıflama ya da tahmin için kullanılabilecek ağaç benzeri örüntüleri keşfetmek içinde uygulanan bir yöntemdir (Emel & Taşkın, 2005: 225). Sinir ağlarında olduğu gibi diğer metodolojiler de sınıflandırma işleminde kullanılmasına rağmen, karar ağaçlarının rahat yorumlanması ve anlaşılabilir olması açısından karar vericiler için önemli bir avantajdır (Chien & Chen, 2008). Şekil 2.2. de bir karar ağacının yapısı gösterilmiştir.



Şekil 2.2. Karar ağacının yapısı

Karar ağacı tekniğinin kullanılması ilk olarak verinin sınıflanması, ikinci olarak ise sınıflama ve öğrenme olmak üzere iki aşamalı bir işlemin uygulanmasıdır. Öğrenme aşamasında model oluşturmak amacı ile sınıflama algoritması tarafından daha önceden bilinen bir eğitim verisinin analizi yapılır. Yapılan analizle modelin öğrenilmesi karar ağacı veya sınıflama kuralları olarak gösterilir. İkinci aşama olan sınıflama basamağında

da karar ağacının doğruluğunu veya sınıflama kurallarını belirlemek için test verisi kullanılır. Kuralların doğruluk oranı kabul edilir orandaysa yeni verilerin sınıflanması kullanılabilir. Eğitim verisinde bulunan hangi alanın hangi sırada kullanılıp ağacın oluşturulacağı önceden saptanmalıdır. Bu amaç için de entropi ölçümü en yaygın kullanılan ölçümdür. Entropi ölçümünün fazla kullanılması durumunda ortaya çıkan sonuçlar da o kadar belirsiz ve kararsızdır. Bu yüzden entropi ölçüsü karar ağacının kökünde en az olduğu alanlarda kullanılır. Bir Ak alanının entropi ölçüsünü bulan denklem şöyledir (Özekeş & Çamurcu, 2002).

$$E(C \setminus A_k) = \sum_{j=1}^{M_k} p(a_k, j) x \left[-\sum_{i=1}^{N} p(c_i \setminus a_k, j) \log_2 p(c_i \setminus a_k, j) \right] \quad (2.1)$$

Bu denklemde;

$E(C \setminus A_k) = A_k$ Alanın sınıflama özelliğinin Entropi ölçüsü,

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

$p(c_i \setminus a_k, j) = a_k$ alanı j. değerinde iken sınıf değerinin c_i olma olasılığı,

$M_k = a_k$ alanın içerdiği değerlerin sayısı; $j = 1, 2, ..., M_k$,

N= farklı sınıfların sayısı;1,2,...,N,

K=alanların sayısı; k= 1,2,...,K.

Eğer bir S kümesindeki elemanlar, kategorik olarak $C_1, C_2, C_3, \ldots, C_i$ sınıflarına ayrıştırılırsa, S kümesinde bulunan bir elemanın sınıfını belirlenmek için gerekli olan bilgiler şu denklemle hesaplanabilir:

$$I(S) = -(p_1 \log_2(p_1) + p_2 \log_2(p_2) + ... + p_i \log_2(p_i)) \quad (2.2)$$

Bu denklemde p_i, C_i sınıfına ayırma olasılığıdır.

Entropi denklemi bu şekilde de ifade edilebilir:

$$E(A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} x I(S_i) \quad (2.3)$$

Bu durumda A alanı kullanılarak yapılacak dallanma işleminde, bilgi kazancı şu denklemle hesaplanılır:

$$\text{Kazanç}(A) = I(S) - E(A) \quad (2.4)$$

Başka bir anlatımla kazanç(A), A alanının değerini bilmekten kaynaklanan entropideki azalmadır. Karar ağaçlarında kullanılabilen birden fazla algoritma bulunmaktadır. Başlıca algoritmalar LADTre, CART, CHAID, C4.5, C5.0, ID3 algoritmaları örnek olarak gösterilebilmektedir (Çalış, Kayapınar & Çetinyokuş, 2014: 6).

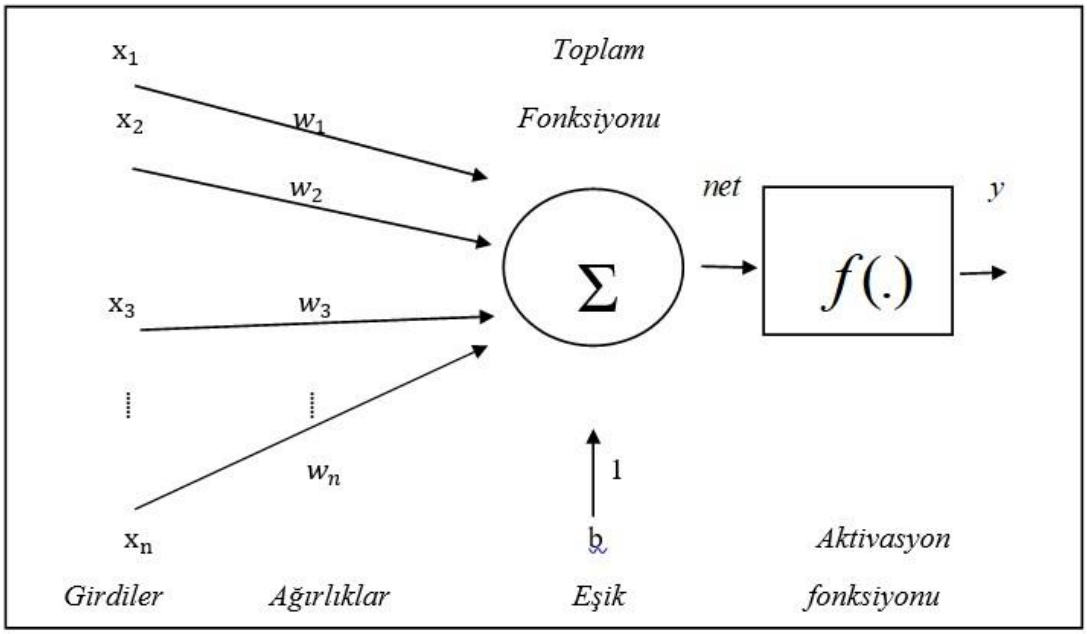
Karar ağaçlarının çok fazla kullanılmasının diğer bir nedeni ise maliyetinin düşük olması, anlaşılması, yorumlanması ve veri tabanlarıyla entegrasyon kolaylığı, güvenilirliklerinin iyi olmasından ötürü en yaygın kullanılan sınıflandırma tekniklerindendir (Çalış, Kayapınar & Çetinyokuş, 2014: 5).

2.1.1.1.2.Yapay Sinir Ağları

YSA (Yapay sinir ağları), insan beyninin öğrenme yolunu taklit edip beynin hatırlama, öğrenme, genelleme yapma yoluyla topladığı verilerden yeni veri üretebilme gibi temel işlevlerin gerçekleştirilebildiği bilgisayar yazılımlarıdır (Öztürk & Şahin, 2018: 27). İnsan beyni birçok nöronun (sinir hücresinin) bir araya gelmesi ile oluşmuştur. Bir yapay sinir ağı ise nöron adı verilen birden fazla işlem biriminden oluşmaktadır. Genellikle nöronlar katman adı verilen mantıksal gruplar arasında yer almaktadır (Kutlu & Badur, 2009: 28). Bu mantıksal gruplar yapay sinir ağını oluşturur. Önceleri tıp alanında başlayan çalışmalar yapay sinir ağlarının gelişmesiyle matematik, fizik, elektrik ve bilgisayar alanlarında da kullanılmaya başlanmıştır (Gültekin, 2017: 42).

Yapay sinir ağları; yazılan karakteri tanıma, resim işleme, ses tanıma, yüz tanıma ve robot kontrol sistemlerinde çok sık olarak kullanılmaktadır (Ergezer, Dikmen & Özdemir, 2003: 15).

Şekil 2.3 de görüleceği üzere en basit yapay sinir hücresinde girdiler, ağırlıklar, eşik, aktivasyon fonksiyonu ve çıkış olmak üzere beş temel bileşenden oluşmaktadır.



Şekil 2.3. Yapay sinir hücresi (Kaynar & Taştan, 2009: 163)

Burada girdiler $(x_1, x_2, x_3, \dots, x_n)$, başka hücrelerden veya dış ortamdan hücre içine giren bilgilerdir. Burada ağın öğrenmesi için istenen örnekler (kimin tarafından?) belirlenir. Ağırlıklar $(w_1, w_2, w_3, \dots, w_n)$ ise girdi kümesi ya da kendisinden önceki bir tabakadaki başka bir işlem elemanının bu işlem elemanı üzerindeki etkisini ifade eden değerlerdir. Her bir girdi, o girdiyi işlem elemanına bağlayan ağırlık değeriyle çarpıldıktan sonra toplam fonksiyonu aracılığıyla birleştirilir. Toplam fonksiyonu Denklem 2.5'te verildiği şekildedir.

$$net = \sum_{i=1}^{n} w_i x_i + b \qquad (2.5)$$

Toplam fonksiyonu sonucunda elde edilen değer doğrusal veya doğrusal olmayan türevlenebilir bir transfer fonksiyonundan geçirilerek işlem elemanının çıktısı hesaplanır.

$$y = f(net) = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.6}$$

Yapay sinir ağlarının birden fazla çeşitli ağ yapıları ve modelleri mevcuttur. Günümüzde değişik alanlarda ve belirli amaçlarla kullanılmaya uygun çok sayıda yapay sinir ağı modelleri (MLP, Adaline, Perceptron, Hopfield, LVQ, ART, SOM, Recurrent vb.) geliştirilmiştir. Bu ağ yapıları arasında en çok kullanım alanına sahip olan çok katmanlı ileri beslemeli yapay sinir ağlarıdır (Kaynar & Taştan, 2009: 163-164).

2.1.1.1.3.Bayes Sınıflandırması

Veri madenciliğinde Naive Bayes sınıflandırıcıları veride bulunan tüm değişkenlerin birbirinden bağımsız ve eşit derecede önemli olduğu varsayılarak analizi gerçekleştiren ve varsayımlarıyla Bayes teoreminin uygulanmasına dayanan basit olasılıksal sınıflandırıcılar ailesidir(Kale, 2018: 62). Naive Bayes sınıflandırması yapılırken sisteme belli bir seviyede öğretilmiş veri sunulur. Öğretim için sunulan bu verilerin kesinlikle bir kategorisi veya sınıfı bulunması gerekir. Öğretilen veriler üzerinden yapılan olasılık işlemleri ile sisteme sunulan yeni test verileri daha önceden elde edilen olasılık değerlerine göre işletilir. Bu verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılmaktadır. Tabii ki öğretilmiş olan veri sayısı ne kadar fazla olursa test verisinin gerçek kategorisini tespit etmek o kadar kesin olabilmektedir. Bunun yanında Naive Bayes sınıflandırma yönteminin de birden fazla kullanım alanı vardır. Ama önemli olan neyin sınıflandırıldığı değil de daha çok nasıl sınıflandırılma yapıldığıdır. Öğretilecek veriler, ikili ya da metin türünde veriler olabilmektedir. Burada veri tipinin ne olduğu değil de bu veriler arasında nasıl bir oransal ilişkinin kurulduğu önem kazanmaktadır. Kullanılan yöntemin matematiksel bir şekilde gösterimi aşağıda verilmiştir (Taşcı & Şamlı, 2020: 91).

Naive Bayes sınıflandırması Bayes teoremine dayanmaktadır. $y_j, j \in \{1,\ldots,J\}$ olmak üzere J tane sınıftan birini temsil eden ayrık bir değişkendir. X özelliği, m adet

özellikten oluşan $X=(x_1,x_2,...,x_m)$ özellik vektörü ile ifade edilmektedir. Olası y_j değeri için Bayes teoremine göre sonsal olasılık $p(y_j \mid X)$ Denklem 2.7'de belirtilmektedir;

$$P(y_j \mid X)=\frac{P(X \mid y_j)P(y_j)}{P(X)} \tag{2.7}$$

Y'de hedef sınıfının tahmininde tüm özellikler için koşullu olasılıkların çarpımı hesaplandığında Naive Bayes sınıflandırıcı için Denklem 2.8'deki temel eşitlik oluşur.

$$Y' \leftarrow \arg\max_{y_j} \frac{P(y_j)\prod_{i=1}^{m} P(X=x_i \mid y_j)}{\sum_{j=1}^{j} P(y_j)\prod_{i=1}^{m} P(X=x_i \mid y_j)} \tag{2.8}$$

Y' hedef sınıfının tahmini için kullanılan Naive Bayes sınıflandırıcıda paydanın tek sınıfa bağımlı olmaması ve tüm hesaplamalarda paydanın ortak olarak işleme katılması nedeniyle denklem formülü 2.9'daki gibi sadeleştirilebilir.

$$Y' \leftarrow \arg\max_{y_j} P(Y=y_j)\prod_{i=1}^{m} P(X=x_i \mid Y=y_i) \tag{2.9}$$

$$P(y_j)=\frac{|N_j|}{|N|} \tag{2.10}$$

Denklem 2.10'a göre, sınıf etiketindeki, eğitim niteliklerinin toplam sayısı $|N|, y_j$ sınıfındaki eğitim niteliklerinin sayısı $|N_j|$. Her bir sınıf için Denklem 2.9 uygulanıp olasılıklar hesaplandıktan sonra sonuç değerleri içerisinde en büyük olasılığa sahip olan sınıf hedef sınıf olarak seçilir (Arpacı & Kalıpsız, 2018: 3-4).

2.1.1.1.4. Karar Destek Sistemleri

KDS kavramı aslında YBS'nin eksikliğinden doğmuş olup karar verme sürecinde yönetime destek vermek için hedeflenen bilginin üretilmesi, sunulması ve kullanıcı etkileşimli yazılım ve donanım vasıtalarının bütünleşik kümesinden oluşan etkileşimli bilgi sistemleridir (Çetinyokuş & Gökçen, 2002: 44). Veri miktarlarının artması ile birlikte karar sayısını ve karar süreçlerini de artırmakta ve buna bağlı olarak bir karmaşıklığın ortaya çıkmasına neden olabilmektedir. Bu nedenle karar vericiler çok yoğun veriler ile karar analizleri yapmak durumunda kalmaktadır. Karar vericilerin daha verimli, etkin ve hızlı bir şekilde bu analizleri gerçekleştirebilmesi için Karar Destek Sistemleri (KDS) olarak adlandırılan araçlar araştırmacılar tarafından geliştirilmiştir (Bilgilioğlu, 2018: 5). Bu araçlar karar vericinin yerine geçmekten ziyade onun kararlarını destekleyen, yarı-yapısal ve yapısal olmayan problemlerin de çözümü için yardımcı olan etkileşimli sistemlerdir (Yıldız, Dağdeviren & Çetinyokuş, 2008: 241).

Karar Destek Sistemleri; veri tabanı, kullanıcı ara yüzü, karar destek sistemleri modeli ve karar destek sistemleri ağ yapısı olmak üzere dört ana bileşenden oluşurken karar verme işlemi de üç aşamadan oluşmaktadır. İlk aşamada bir problemin çözülmesi gerekir; yani bir karar verme durumu ortaya çıkar. İkinci aşamada, mümkün olabilen hareket tarzları belirlenir. Üçüncü aşamada ise belirlenen çeşitli hareket tarzları arasından seçim yapılır. Bu durumda doğru verilere sahip olunması karar destek sisteminin planlanması, uygulanması ve kontrolünde de önemli adımlar oluşturur. Genel olarak karar verme süreci bir sürecin sonucu ya da birimi olsa da başka bir sürecin de başlangıcı olarak da kabul edilir. Veri analizinin de sunulabilmesi için modellere ihtiyaç duyulmaktadır. Şekil 2.4'te genel olarak KDS'nin yapısı oluşturulmuştur (Özsever, Gençoğlu & Erginel, 2009: 52; Sütçü, 1995: 6).

Şekil 2.4. Karar destek sisteminin yapısı (Özsever, Gençoğlu & Erginel, 2009: 53)

Karar Destek Sistemlerinin Yararları;

1-Geliştirilebilir(geliştirme olanaklarını içsel olarak barındıran)

2-Özel amaçlı veri analizi ve karar modelleme yeteneğine sahip

3-Geleceği planlamaya yönelik

4-Düzensiz planlanmış zaman aralıklarında kullanılan sistemler olarak tanımlanmıştır (Mutlu, 1989: 37).

2.1.1.1.5. K-en Yakın Komşu

Denetimli öğrenme yöntemleri arasında yer alan, sınıflama ve regresyon ayağında kullanılabilen K-en Yakın Komşu algoritması çok yönlü bir algoritmadır. Bu algoritmayı basit bir şekilde tanımlarsak sınıfı bilinmeyen bir veri seti eğitim setinde bulunan başka verilerle karşılaştırılıp ona göre bir uzaklık ölçümü yapılır. Daha sonra da hesaplanan uzaklığa göre henüz bir sınıfa atanamamış veriye en yakın olan sınıf belirlenir ve bu sınıfa aktarılır (Dilki & Başar, 2020: 227). K-en algoritması, büyük eğitim setlerinin varlığında, oldukça etkin sonuçlar verebilmekte ve uygulanabilme kolaylığı, basit bir matematiksel temele sahip olması pek çok farklı alanda tahmin modelleri kullanımını yaygınlaştırmıştır (Altunkaynak, Başakın & Kartal, 2020: 1550; Taşcı & Onan, 2016: 3).

K-en algoritması, sınıf kategorisi bilinen eğitim veri setindeki örneklerin uzaklık ölçüsüne dayalı olan örnek-tabanlı sınıflamada p adet nitelik ile tanımlanan n tane $x_i^T = (x_{11},...,x_{1p}),...,x_n^T = (x_{n1},...,x_{np})$ veri örneğinin olduğunu varsayalım. Her örnek, p boyutlu vektör uzayında bir noktayı temsil etmektedir. i. ve j. örnek noktaları arasındaki uzaklık $d(x_i, x_j)$ ile gösterilir ve niteliklerin hepsi nümerik olduğunda genellikle Denklem 2.11'de verilen euclidean uzaklığı ile hesaplanır:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - y_{jk})^2} \tag{2.11}$$

Yeni bir örneğin hangi sınıfa ait olacağını belirlemek için bu noktadan eğitim veri setinde bulunan tüm noktalara olan uzaklığı hesaplanır. Daha sonra hesaplanan uzaklıklar küçükten büyüğe doğru sıralanır. Önceden belirlenen K komşu sayısı parametresi dikkate alınıp K adet en yakın uzaklığa sahip olan noktalar belirlenir. Belirlenen K tane sınıf kategorisi bilinen eğitim veri noktaları içerisinde yeni gözlemin sınıfını belirlemek için genellikle ağırlıklı oylama yöntemleri ya da çoğunluk oylamasına başvurulur. Ağırlıklı oylama yönteminde yeni örnek noktasına daha yakın olan komşuların sınıflama yaparken

söz hakkının yüksek olması amaçlanır. Çoğunluk oylamasına göre de yeni örneğin sınıf ataması yapılırken K adet komşu arasında en çok tekrar eden sınıf seçilir. Eğitim veri setindeki K tane $x_m; m = 1,...,K$ en yakın komşunun oyu, yeni örnek noktası x_z 'ye olan mesafe ile ters orantılı olduğu Denklem 2.12'de verilmiştir:

$$oy(x_m) = \begin{cases} \infty, & d(x_{m,} x_z = 0 ise \\ \dfrac{1}{d(x_m, x_z)}, & diğer\ durumda \end{cases} \tag{2.12}$$

Denklem 2.12'de yer alan tüm oylar toplanır ve yeni örneğin sınıfı, en yüksek ağırlıklı oylamaya sahip olan sınıf kategorisine atanır. K-en algoritmasında genellikle $K = \sqrt{n}$ ile seçilir veya çapraz doğrulama yöntemi ile K parametresi deneme yapılarak bulunmuş olunur (Kemalbay & Alkış, 2020: 558-559).

2.1.1.1.6. Genetik Algoritmalar

Evrimden ilham alan genetik algoritmalar bir hesaplama modelleri ailesidir. Bu algoritmalar, belirli bir problemin potansiyel çözümünü kromozom benzeri basit bir veri yapısı üzerinde kodlar ve bu yapılara, kritik bilgileri korumak için rekombinasyon operatörleri uygular. Burada canlı sistemlerde bulunan genetik şifre tekniğinin kullanılarak sezgisel bir şekilde en iyi çözüme ya da en iyi çözüme yakın olabilecek bir sonuç bulmayı amaçlar. Bu mantık, doğal seçime yani güçlü bireyin hayatta kalma olasılığının yüksek olmasına dayanır. Bu yöntemle evrim sonucu hayatta kalan birey en iyi sonuç olarak alınır. Genetik algoritmaların uygulandığı problemlerin aralığı oldukça geniş olmasına rağmen, genetik algoritmalar genellikle fonksiyon optimize edici olarak görülür. Genetik algoritmanın uygulanması, (tipik olarak rasgele) bir kromozom popülasyonu ile başlar. Daha sonra bu yapılar değerlendirilir ve üreme fırsatları, hedef soruna daha iyi bir çözümü temsil eden bu kromozomlara göre daha fazla üreme şansı verilecek şekilde tahsis edilir. Bir çözümün iyiliği tipik olarak mevcut nüfusa göre tanımlanır (Mathew, 2012: 1; Altay, 2007: 9).

Temelde genetik algoritmaların potansiyel çözümleri içeren bir popülasyon üzerinde, daha iyi bir sonuç vermeye uygun olan çözümlerin yaşaması prensibine dayalı

olarak çalışır. Her yeni nesilde, problem için uygunluk seviyelerine göre bireylerin seçimi ve çoğaltılması prosesleri ile en iyi sonuca olan yakınlık artar (Vural, 2005: 6).

Genetik algoritmalar, değişik planlama teknikleri ile bir fonksiyonun optimizasyonu veya ardışık değerlerin tespitini içine alan birçok problem tipi içinde çözüm arama yöntemi olmakla birlikte kullanım alanları, gezgin satıcı problemi, otomatik programlama ve bilgi sistemleri, finans pazarlama, mekanik öğrenme, montaj hattı dengeleme problemi, çizelgeleme problemi, tesis yerleşim problemi, optimizasyon, sistem güvenilirliği problemi gibi yerlerde kullanılmaktadır (Daş, Türkoğlu & Poyraz, 2006: 68; Güracar, 2019).

2.1.1.1.7. Zaman Serileri Analizi

Zaman serileri analizi zamanın periyodik noktalarında, bir cevap değişkenin gözlemlenmesi yolu ile verilerin bir araya getirilmesi olarak adlandırılır (Özek, Daş & Akpolat, 2010: 101). Fakat her seri bir zaman serisi değildir. Bir serinin zaman serisi olabilmesi için zamana bağlı bir durum olması gerekir. Elde bulunan verilerden de en az bir tanesi zamana bağlı olmalıdır (Seker, 2015: 23).

Zaman serisi veri tabanı, bir zaman dilimi içerisinde tekrarlayan ölçümlerden elde edilen değerleri içermektedir. Bu değerler tipik olarak eşit zaman aralıklarında ölçülür (günlük, haftalık ve yıllık gibi). Çeşitli değişkenlerle düzenlenmiş zaman serileri için de özel tahmin teknikleri geliştirilmiştir. Ekonomi ve işletme, zaman serisi analizlerinin kullanıldığı en önemli alanlardandır (Irmak, Köksal & Asilkan, 2012: 105).

Genellikle bir zaman serisi şu amaçlar için de analize tabi tutulmaktadır: açıklama (açıklayıcı değikenler), tanımlama (grafiksel, sayısal, özellik), tahmin (gelecek yılki satış rakamları) ve kontrol (ekonomide imalat yönteminin kalitesi). Zamana göre değişen bir ya da daha çok olayın incelenmesinde de zaman serileri kullanılmaktadır (Asilkan, 2008: 46).

2.1.1.2. Regresyon Analizi

Regresyon analizi bir ölçüt değişkeniyle bir veya birden daha fazla sayıdaki tahmin değişkenleri arasında bulunan ilişkiyi sayısal hale dönüştürmek için kullanılan istatistiksel analiz metodudur. Regresyon analizi gerçekte farklı ilişkilerin niteliğini saptamayı hedefler. Basit regresyon analizinde bir değişken tahmin değişkeni olarak kullanılırken çoklu regresyon analizinde iki veya daha fazla değişken tahmin değişkenleri

olarak kullanılabilir. Hedef her tahminde ölçüt değişkenindeki toplam değişmeye olan katkısının saptanması ve tahminde doğrusal kombinasyonunun değerinden hareketle ölçüt değerinin tahmin edilmesidir (Özmen & Şen, 2013: 837).

Regresyon analizinde bağımlı değişken olan Y ve bağımsız değişken olan $X_k(i = 1, \dots, k)$ değişkenler arasındaki ilişkileri sebep-sonuç olarak matematiksel bir model halinde ortaya koyar. Y bağımlı değişkeni ile X bağımsız değişkeni arasındaki ilişkiyi tahmin etmek üzere kurulan modele basit doğrusal regresyon modeli denir. Çoklu doğrusal regresyon modelinde ise Y bağımlı ve $X_k(i = 1, \dots, k)$ bağımsız değişkenleri arasındaki ilişkiyi tahmin etmek üzere kurulan modeldir (Tüzüntürk, 2010: 81).

Basit doğrusal regresyon modeli şöyle gösterilebilir:

$$Y=a+\beta X+u$$

Burada Y bağımlı değişken iken X bağımsız değişkendir. Tahmin edilecek olan katsayılar a ve β 'dır. Hata terimi ise u 'dur. Model tahmin edildiğinde tahmin edilen katsayılar $\hat{a}$ ve $\hat{\beta}$ olarak gösterilebilir. Katsayıların tahmin edilmesinde alternatif yaklaşımlar bulunmasına karşın en sık kullanılan, olağan en küçük kareler tahmin yöntemidir (Tüzüntürk, 2010: 81).

2.1.2. Tanımlayıcı Yöntemler

2.1.2.1. Kümeleme Yöntemleri

Birden fazla kümeleme yönteminin bulunmasının temel nedeni "küme" kavramının tam olarak tanımlanmamış olmasıdır. Kümeleme modellerindeki temel hedef birbirlerine çok benzeyen üyelerin özellikleri bakımından birbirinden farklı şekilde olan kümelerin bulunmasıdır. Veri tabanlarındaki kayıtların bu farklı kümelere bölünmesi işlemidir. Kümeleme analiz işleminde veri tabanlarındaki kayıtların hangi kümelere ayrılacağı ya da kümelemenin hangi değişken özelliklerine göre yapılacağını konunun uzmanı olan bir kişi tarafından yapılabildiği gibi bunu yapabilen geliştirilmiş bilgisayar programları da bulunmaktadır (Ayık, Özdemir & Yavuz, 2010: 447).

Literatürde birden fazla kümeleme algoritması bulunmaktadır. Veri tipine ve kullanım amacına bağlı olarak kümeleme yöntemi seçimi yapılır. Başlıca kümeleme yöntemleri şu şekilde sınıflandırılabilir (Özekes, 2003: 71).

2.1.2.1.1.Bölme Yöntemleri

Bölme yöntemlerinde k oluşturulacak küme sayısı ve n veri tabanında bulunan nesne sayısı şeklinde kabul edilmektedir. Bölme algoritması da n sayıdaki nesneyi, k sayıda kümeye böler (k < n). Kümeler tarafsız bölme kriteri olarak isimlendirilen bir kritere uygun şekilde oluşturulduğu için aynı kümede bulunan nesneler birbirlerine benzerken, başka kümedeki nesneler birbirinden farklıdır (Özekes, 2003: 72). Bölme yöntemlerinde işlem sırası şu şekildedir:

• İlk etap başlangıç için kullanılacak olan küme merkezlerinin seçimi gelişigüzel bir şekilde yapılır.

• Daha sonra nesnelerin, belirlenmiş olan kümelerin merkezlerine olan uzaklıklarına bakılarak yeni küme merkezleri oluşturulur.

• Yapılan bu işlemler birbirlerinden farklı olarak ve kendi içerisinde homojen ve aralarında benzerlik bulunmayacak bir şekilde k adet küme oluşturuluncaya kadar devam edilir. En iyi şekilde bilinen ve en çok kullanılan bölümleme yöntemleri k-medoids, k-ortalamalar(k-means) yöntemi ve bunların çeşitleridir (Darakçı, 2011: 38-39).

2.1.2.1.2. Hiyerarşik Yöntemler

Hiyerarşik kümeleme yöntemlerinde kümeler bir ana küme olarak ele alınır. Daha sonra aşamalı bir şekilde içerdiği alt kümelere ayrılması ya da ayrı ayrı ele alınan kümelerin aşamalı olarak bir küme biçiminde birleştirilmesi esasına dayanır. Bu yöntemler, örnekleri yukarıdan aşağıya veya aşıdan yukarıya yinelemeli olarak bölümleyerek kümleri oluşturur. Bu yöntemler aşağıdaki gibi alt gruplara ayrılabilir:

Toplayıcı Hiyerarşik Kümeleme- Her nesne başlangıçta kendi kümesini temsil eder. Ardından istenen küme yapısı elde edilene kadar kümeler art arda bir araya getirilir.

Bölücü hiyerarşik kümeleme- Başlangıçta tüm nesneler bir kümeye aittir. Daha sonra küme, sırasıyla kendi alt kümelerine bölünen alt kümelere bölünür. Bu süreç istenilen küme yapısı elde edilene kadar devam eder (Yeşilbudak, Kahraman & Karacan, 2011: 29; Rokach & Maimon, 2005: 330-331)

2.1.2.1.3. Yoğunluk Tabanlı Yöntemler

Yoğunluk dayalı yöntemlerde her kümeye ait noktaların belirli bir olasılık dağılımından çekildiğini ve verilerin genel dağılımının birkaç dağılımın karışımı olduğu varsayılır. Bir yoğunluk fonksiyonu aracılığıyla nesnelerin doğal dağılımını tespit ederek bir eşik yoğunluğunu aşan bölgeleri küme olarak adlandırırlar. Yoğunluk dayalı yöntemler tek taramayla sonuca ulaşma avantajları, gürültü ve istisnalardan etkilenmeme ve düzgün şekilli olmayan kümeleri bulma başarısıyla en başarılı kümeleme metotları arasında bulunmaktadır (Bilgin, 2003: 34).

2.1.2.1.4. Izgara Tabanlı Yöntemler

Izgara tabanlı kümeler yöntemi, bir kümedeki tüm nesneleri ızgaralar olarak bilinen bir dizi kare hücreye eşler. Izgara tabanlı kümeleme, tipik olarak veriler yerine ızgaranın boyutuna bağlı olan hızlı bir işlem süresine sahiptir. Izgara hücrelerinin yoğun olduğu noktalara göre küme sayılarının belirlenebilmesi için veri nesnesi olmayan ızgara hücreleriyle küme merkezleri arasındaki uzaklığın hesaplanması bu teknik içerisinde bulunan algoritmaların hızlı bir şekilde çalışmasını sağlamaktadır. Bu durum ızgara tabanlı tekniklerin en büyük avantajı olarak kabul edilir. Izgara Tabanlı Kümeleme tekniklerinin en etkili algoritmalardan olmasına rağmen, bu tekniklerin dezavantajı olarak kabul edilen durum ise algoritmaların etkisi önceden belirlenmiş ızgaraların büyüklüğünden ve belirli hücrelerin eşik değerinden ciddi bir şekilde etkilenmesidir (Shah & Nair, 2015: 3; Akın, 2008: 116).

Izgara yoğunluğu tabanlı algoritmalar, kullanıcıların bir ızgara boyutu veya yoğunluk eşiği belirlenmesini gerektirir. Izgara tabanlı kümeleme algoritmaları STING, Wave Cluster ve CLIQUE'dir. Izgara hücrelerinde depolanmış olan bilgiyi istatistiksel olarak açıklayan STING algoritmasıdır. WaveCluster algoritması da dalga dönüşüm (wavelettransform) yöntemi kullanarak nesneleri kümeler ve CLIQUE algoritmasıda çok boyutlu olan bir veri uzayında kümeleme için hem yoğunluğa dayalı hem de ızgara tabanlı yaklaşımı temsil eden algoritmadır (Shah & Nair, 2015: 3; Akın, 2008: 116).

2.1.2.1.5. Model Tabanlı Yöntemler

Model tabanlı metotlarda elde bulunan verilerin bir matematiksel modelle ifade edilmeye çalışır. Bu metotlar verilerin belirli bazı olasılık teorilerinin karışımından oluşan bir mantıkla veri uzayına yerleştiklerini farz ederler. Model tabanlı metotlar yaklaşımı yapay zekâ ve istatistik yaklaşımını kullanırlar (Bırtıl, 2011: 47).

2.1.2.2. Birliktelik Kuralları

Birliktelik Analizi veri tabanında çoğunlukla birlikte gerçekleşen ya da aynı süre içerisinde gerçekleşen durumları bulmak üzere kullanılan bir yöntemdir. Bundan dolayı da tanımlayıcı bir modeldir. Rastgele bir ürünün alınmasında bu ürünün yanında baka ürünün de satın alınması bir birliktelik kuralını vermektedir. Ürün ve bu ürünlerin birlikte alınmaları söz konusu olunca daha çok pazarlama alanında satış ve reklam stratejileri geliştirmede ve perakendecilik sektöründe sıkça kullanılmaktadır(Akgün, 2012: 70). Market sepeti analizi, birliktelik analizinin kullanıldığı en tipik örnektir. Bu analizde müşterilerin satın alma alışkanlıklarını analiz etmek için yaptıkları alışverişlerde ürünler arasındaki birliktelikleri bularak analiz yapılır. Bu analizlerin yapılmasıyla müşterilerin hangi ürünleri beraber aldıkları bilgisi ortaya çıkar. Daha sonra da market yöneticileri bu bilgiler ışığında daha iyi satış stratejileri geliştirebilmektedirler(Gürgen, 2008: 18). Birliktelik analizi "sepet analizi " olarak da tanımlanmaktadır (Bilekdemir, 2010: 21).

Birliktelik analizinin matematiksel modeli Imielinski, Swami ve Angrawal tarafından1993 yılında ortaya çıkarılmıştır. Bu modelde, $I = \{\dot{I}_1, \dot{I}_2, ..., \dot{I}_M\}$ kümesine "ürünler" adı verilmektedir. TID her harekete ait olan tek belirteçtir. T ürünlerin her bir hareketini, D ise veri bütünlüğündeki tüm hareketleri simgeler.

Birliktelik kuralını şu şekilde tanımlayabiliriz;

A_i ve B_j bu ifadede yer alan yapılan iş yada nesnelerdir. Genellikle bu kural "$A_1, A_2, ..., A_m$" iş ya da nesneleri meydana getirdiğinde, sık olarak "$B_1, B_2, ..., B_n$" ise iş veya nesnelerinin aynı olay veya hareket içinde yer aldığını belirtir (Öztürk & Tanrısevdi, 2017: 137/138).

2.2. Metin Madenciliğinde Kullanılan Yöntemler ve Uygulamalar

2.2.1. Metin Madenciliğinde Kullanılan Yöntemler

Geleneksel olarak metin madenciliği problemini çözmek için geliştirilen birden fazla teknik bulunmaktadır. Bu kullanıcının gereksinimlerine göre ilgili bilgilerin erişimine bağlıdır. Bilgi erişimine göre temel olarak kullanılan dört yöntem vardır:

2.2.1.1. Terim Bazlı Yöntem

Belgedeki terim, anlamsal anlamı olan kelimedir. Keşfedilen birçok terimin anlamsal anlamı, kullanıcıların ne istediğini yanıtlamak için belirsizdir. Bilgi alma, bu

zorluğu çözmek için birçok terime dayalı yöntem sağladı. Terim bazlı yöntemler, çok anlamlılık ve anlamlılık sorunlarından bağımsızdır. Çok anlamlılık, bir kelimenin birden fazla anlamı olduğu anlamına gelir ve eş anlamlılık, aynı anlama gelen birden fazla kelimenin bulunmasıdır. Terim Bazlı yöntemde belge, terim bazında analiz edilir ve verimli hesaplama performansının yanı sıra terim ağırlıklandırma için uygun teori avantajlarına sahiptir. Bu teknikler bilgi alma ve makine öğrenimi toplulukları tarafından ortaya çıkmıştır (Gaikwad, Chaugule & Patil, 2014: 43).

2.2.1.2. Cümle Tabanlı Yöntem

İfade, bilgi gibi daha fazla anlam taşır ve daha az belirsizdir. İfadeye dayalı yöntemde, ifadeler tek tek terimlere göre daha az belirsiz ve daha ayırt edici olduğundan belge tümce bazında analiz edilir. Göz korkutucu performansın olası nedenleri arasında şunlar sayılabilir:

a- İfadeler, terimlere göre daha düşük istatistiksel özelliklere sahiptir.

b- Görülme sıklıkları düşüktür ve

c- Aralarında çok sayıda gereksiz ve gürültülü ifadeler vardır (Gaikwad, Chaugule & Patil, 2014: 43).

2.2.1.3. Konsepte Dayalı Yöntem

Kavram temelli terimler cümle ve belge düzeyinde incelenir. Metin Madenciliği teknikleri kelime veya deyimin istatistiksel analizine dayanır. Sıklık teriminin istatistiksel analizi, belgesiz kelimenin önemini yakalar. Aynı belgede iki terim aynı frekansa sahip olabilir, ancak anlam, bir terimin diğer terimin sağladığı anlamdan daha uygun bir şekilde katkıda bulunur. Metin anlamını yakalayan terimlere daha fazla önem verilmeli, böylece kavram temelli yeni bir madencilik ortaya konulmalıdır (Gaikwad, Chaugule & Patil, 2014: 43).

Bu model üç bileşen içermektedir. İlk bileşen, cümlelerin anlamsal yapısını analiz eder. İkinci bileşen, semantik yapıları açıklamak için kavramsal bir ontolojik grafik (COG) oluşturur ve son bileşen, standart vektör uzayı modelini kullanarak özellik vektörleri oluşturmak için ilk iki bileşene dayalı olarak en iyi kavramları çıkarır. Kavrama dayalı model, bir cümlenin anlamını tanımlayan anlamlı terimler ile önemli olmayan terimler arasında etkili bir şekilde ayrım yapabilir. Kavram tabanlı model genellikle doğal dil işleme tekniklerine dayanır. Gösterimi optimize etmek ve gürültüyü ve belirsizliği

ortadan kaldırmak için sorgu kavramlarına özellik seçimi uygulanır (Gaikwad, Chaugule & Patil, 2014: 43).

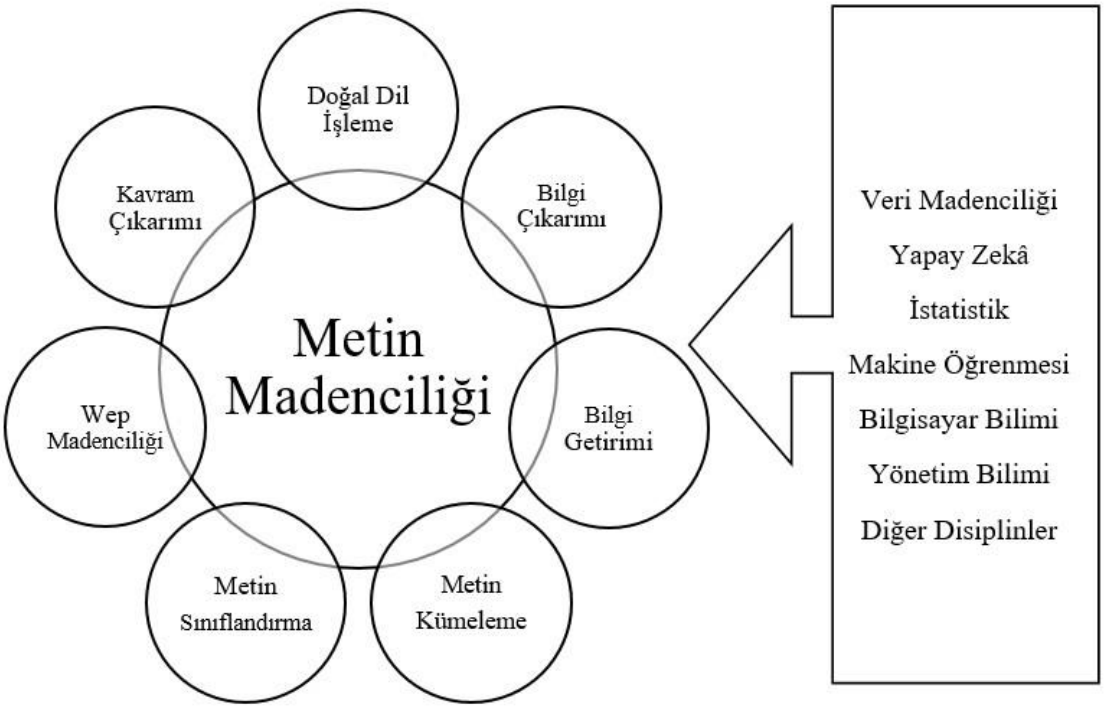
2.2.1.4. Model Taksonomi Yöntemi

Model madenciliği, uzun yıllardır veri madenciliği toplulukları tarafından kapsamlı bir şekilde çalışılmaktadır. Örüntü taksonomisi yönteminde dokümanlar örüntü bazında analiz edilir. Kalıplar is-a ilişkisi kullanılarak taksonomide yapılandırılabilir. Örüntüler, birliktelik kuralı madenciliği, sık öğe seti madenciliği, sıralı örüntü madenciliği ve kapalı örüntü madenciliği gibi veri madenciliği teknikleri ile keşfedilebilir. Metin madenciliği alanında keşfedilen bilginin (kalıpların) kullanımı zor ve etkisizdir, çünkü yüksek özgürlüğe sahip bazı yararlı uzun kalıplar destekten yoksundur(yani, düşük frekans problemi). Tüm sık kullanılan kısa kalıplar yararlı değildir, bu nedenle kalıpların yanlış yorumlanması olarak bilinir ve etkisiz performansa yol açar (Gaikwad, Chaugule & Patil, 2014: 43).

Araştırma çalışmalarında, metin madenciliği için düşük frekans ve yanlış yorumlama problemlerinin üstesinden gelmek için etkili bir örüntü keşif tekniği önerilmiştir. Örüntü tabanlı teknik, desen dağıtma ve desen geliştirme olmak üzere iki süreç kullanır. Bu teknik, metin belgelerinde keşfedilen kalıpları geliştirir. Deneysel sonuçlar, örüntü tabanlı modelin yalnızca diğer saf veri madenciliği tabanlı yöntemlerden ve kavram tabanlı modelden değil, aynı zamanda terim tabanlı modellerden daha performans ortaya çıkardığını göstermektedir (Gaikwad, Chaugule & Patil, 2014: 43).

2.2.2. Metin Madenciliği Uygulamaları

Metin madenciliği uygulamaları yedi başlık altında incelenmektedir. Bu uygulama alanları birbirlerinden farklı olmakla birlikte aynı zamanda birbirleriyle ilişkilidir. Şekil 2.5'te metin madenciliği uygulamaları verilmiş olup bu uygulamalardan yararlanan çeşitli disiplinler yer almaktadır (Yıldız & Ağdeniz, 2018: 292).



Şekil 2.5. Metin madenciliği uygulamaları (Yıldız & Ağdeniz, 2018: 292)

2.2.2.1.Bilgi Getirimi/Erişimi

Bilgi Erişimi, belgelerin kullanıcı tarafından talep edilen belirli bilgileri yoğunlaştırmak veya çıkarmak için işlendiği belge alımının bir uzantısı olarak kabul edilir. Burada bilgilerin temsili, saklanması ve erişilmesi için kullanılan ve işlenen bilgilerin çoğunlukla metinsel belgeler, gazeteler ve kullanıcı istek veya sorgularına göre veri tabanlarından alınan kitap ve gazeteler biçimindeki yöntemler olarak tanımlanmaktadır (Sumathy & Chidambaram, 2013: 30).

En iyi bilinen bilgi erişimi (IR) sistemleri ve World Wide Web'de verilen bir dizi kelimeyle ilişkili olan belgeleri tanıyan Google arama motorlarıdır. Bu kullanıcı için çok önemli olan yararlı bilgileri bulmak için geri gönderilen belgelerin işlendiği ve belge alımının bir uzantısı olarak ölçülür. Bu nedenle, belge alınımı, kullanıcı tarafından yöneltilen sorguya odaklanan bir metin özetleme aşaması veya bir bilgi çıkarımı aşaması takip eder. Daha geniş anlamda IR, bilgi alımından bilgi alımına kadar tüm bilgi işleme aralığıyla ilgilenir. İlk otomatik indeksleme girişimlerinin 1975'te yapıldığı nispeten eski bir araştırma alanıdır. World Wide Web'in büyümesi ve klas arama motorlarına olan ihtiyacın artmasıyla beraber ilgi alanı haline gelir (Dang & Ahmad, 2014: 23).

2.2.2.2. Metin Kümeleme

Metinlerin kümelenmesin de belge koleksiyonlarında bulunan belgelerin benzerlik oranlarına bağlı olarak kümelere ayrıştırılma işlemidir. Kümeleme işleminin yapılabilmesi için bir küme içerisinde bulunan belgelerin çoğunlukla birbirine yakın konularda olması beklenir. Kümeleme Hipotezinde(cluster hypothesis) ise birbiriyle ilişkili olan belgeler ilişkisizlere oranla birbirine daha çok benzemektedir. Bu hipoteze dayanılarak belgelerin kümelenmesi sayesinde geniş belge koleksiyonları içerisinde dolaşım (navigasyon) ve arama işlemlerinin kolaylaştırılması ve istenilen bilgilere daha hızlı bir şekilde ulaşılması hedeflenmektedir (Tunalı, 2011: 11).

2.2.2.2.1 Metin Belgelerinin Kümelenmesi İçin Kullanılan Algoritmalar

Denetimsiz sınıflandırma veya kümeleme, yapılandırılmış veya yapılandırılmamış verileri kümelemek için kullanılan temel veri madenciliği tekniklerinden birdir. Çeşitli kümeleme yöntemleri bulunmaktadır. Bu yöntemler aşağıdaki gibi sınıflandırılabilir (Amine vd., 2007: 29):

Hiyerarşik Yöntemler: Bu yöntemler, dendrogram adı verilen sınıfların hiyerarşik bir ağacını oluştururlar. Ağacı oluşturmanın iki yolu vardır: Belgeden başlamak ya da tüm belgelerin veya metin korpusunun kümesinden başlamak.

- Belgelerden başlarken her belge başlangıçta kendi sınıfına konulur. Ardından en benzer sınıflar birleştirilerek tek bir sınıf oluşturulur. Bu işlem, belirli bir sonlandırma koşulu sağlanana kadar tekrarlanır. Bu yönteme "benzer grupların birleştirilmesi" veya "yükselen hiyerarşik kümeleme" denir.
- Tüm belge setiyle (veya korpusla) başlandığında, yöntem "farklı grupların bölünmesi" veya "alçalan hiyerarşik kümeleme" olarak adlandırılır. Bu sürecin başlangıcında, tüm belgeleri içeren tek bir sınıf vardır. Bir sonraki iterasyonda sınıf iki alt sınıfa bölünür. Süreç, sonlandırma koşulu sağlanana kadar devam eder. İki belge arasındaki benzerlik ise belgeler arasındaki mesafeye dayanır (Amine vd., 2007: 29).

Bölümleme Yöntemleri: Bu yöntemlere düz "kümeleme" de denir. En bilinen yöntemler, K-medoidler yöntemi, dinamik bulutlar yöntemi ve K-means veya mobil merkezler yöntemidir. Örneğin, K-means yönteminde sınıfların sayısı önceden ayarlanmıştır. Belgenin vektörü ile bu sınıfın merkezi arasındaki mesafe, vektör ile diğer

sınıfların merkezleri arasındaki mesafelere kıyasla en küçükse bir belge bir sınıfa konur (Amine vd., 2007: 29).

Yoğunluk Tabanlı Yöntemler: Bu yöntem, yakınlık yoğunluğu belirli bir sınırı aştığı sürece nesneleri gruplamayı içerir. Gruplar veya sınıflar, seyrek yoğun alanlar tarafından ayrılır. Bir nokta (belge vektörü), komşularının sayısı belirli bir eşiği aşıyorsa yoğundur ve bir nokta, belirli bir sabit değerden daha düşük bir mesafede ise diğer bir noktaya yakındır. Bir grup veya sınıfın keşfi iki aşamada gerçekleşir: Birincisi rastgele yoğun bir nokta seçin, ikincisi bu noktadan başlayarak yoğunluk eşiğine göre ulaşılabilir tüm noktalardan bir grup veya sınıf oluşturun (Amine vd., 2007: 29).

Izgara tabanlı Yöntemler: Veri alanının bir ızgara oluşturan çok boyutlu hücrelere bölünmesidir. Izgaradaki noktalar veri öğelerini temsil eder ve yakın hücreleri mesafe açısından gruplandırır. Sınıflar yeterli veri içeren (yoğun) hücrelerin birleştirilmesiyle oluşturulur. Giderek daha yüksek çözünürlüğe sahip çeşitli düzeylerde ızgaralar kullanılır (Amine vd., 2007: 29).

Model Tabanlı Yöntemler: Model tabanlı yöntemlerden biri kavramsal yaklaşımdır. Bu yaklaşımda, veriye özgü bir kavramsal hiyerarşi bulunur, burada bir kavram bir çifti ifade eder (niyet, genişleme), niyet vektörlere ortak olan en büyük özellik kümesidir ve genişleme özellikleri paylaşan vektörlerin en büyük kümesidir. Diğer bir model tabanlı yöntem ise Kohonen ağları yöntemidir, ayrıca öz-düzenleyen haritalar (SOM) olarak da adlandırılır. Bu ilginç sinirsel yöntem, elde edilen sınıfları genellikle bir harita üzerinde topolojik olarak düzenler, genellikle bir düzlem üzerinde (örneğin, iki boyutlu) gerçekleşir (Amine vd., 2007: 29).

2.2.2.3. Metin Sınıflandırma

Metin sınıflandırmasında daha önceden oluşturulan ve belirlenmiş olan sınıflardan faydalanılarak, incelenecek olan belgenin bir metnin veya bir cümlenin bu belirlenen sınıflardan hangisine dâhil edileceğini otomatik bir şekilde hesaplanmasıdır. Başka bir anlatımla metin sınıflandırmasında ele alınmakta olan bir metne, bu metnin içeriğine göre ve önceden belirlenmiş olan yöntemlerle kategori ya da etiket atama işlemi olarak tanımlama yapılabilir (Şahin & Chouseınoglou, 2019: 2). Metin sınıflandırma işlemleri temelde 5 aşamadan meydana gelmektedir:

1- Veri kümesinin oluşturulması

2- Verilerin önişleme aşaması

- Dönüştürme

- Tarama ve işaretleme

- Durak kelimelerin çıkarılması

- Kök bulma

3- Özellik seçimi

4- Vektör uzay model seçimi

Sınıflandırma (Göker & Tekedere; 2017: 293).

2.2.2.4. Web Madenciliği

Web madenciliği, web kayıt dosyalarında ihtiyaç duyulan yararlı bilgilerin çıkarılması ve değerlendirilmesi işlemi olarak tanımlanır. Internetten bilgi keşfi ve bilgi çıkarımı işlemleri, web madenciliğinin önemli bir alanı arasındadır. İnternette bulunan verilere sürekli bir şekilde yeni bilgilerin eklenmesi, değişmesi ve güncellenmesinden dolayı da webden bilgi çıkarımı işleminde karşılaşılan bir zorluk haline gelmektedir. Webden bilgi çıkarımı, normal metin tabanlı dokümanlara göre daha zor olmasının sebebi web sayfalarının bu dinamik yapısından dolayıdır (Daş, Türkoğlu & Poyraz, 2006: 68). Bunun için de web madenciliği işlemleri kullanarak yapısal bir halde olmayan web verileri yapısal bir hale dönüştürülme işlemi temel olarak üç alt başlık altında bir araya getirilebilir:

- Web yapı madenciliği: İnternetin temel yapısını oluşturan web siteleri, web sayfaları arası veya web sayfasında bulunan bağlantılar arasındaki ilişkileri inceler (Dolgun, Özdemir & Oğuz, 2009: 51/52).

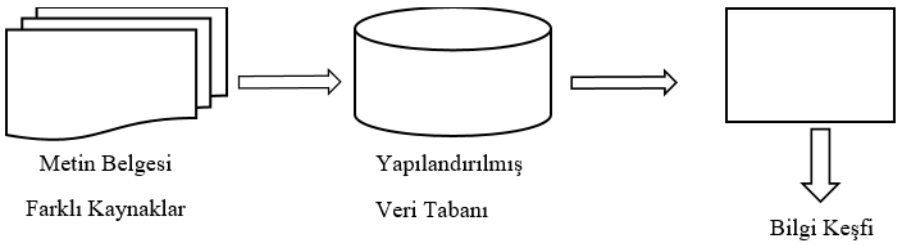
-Web içerik madenciliği: Burada web sayfalarının içerikleri incelenerek kullanışlı bilgilerin çıkarımı sağlanır. Bu teknik kullanılarak web sayfalarının başlıkları içerisinde geçen kelimeler, resimler ya da müzik dosyaları incelenir ve daha sonra bulunan içeriklere göre web siteleri belirli kümelere ya da sınıflara ayırabilir (Dolgun, Özdemir & Oğuz, 2009: 51-52).

-Web kullanım madenciliği: Bu yöntemde web sunucularında tutulan kullanıcı erişim kayıtları incelenerek anlamlı ve faydalı kalıplar bulunabilir. Bu yöntemlerin

uygulanmasıyla web sitelerini ziyaret eden kişilerin davranış ve tutumları da belirlenebilmektedir (Dolgun, Özdemir, & Oğuz, 2009: 51-52).

2.2.2.5. Bilgi Çıkarımı

Bilgi çıkarımında yapılandırılmamış doğal dil belgesinden otomatik olarak yapılandırılmış veri elde etme işlemi olarak tanımlanır. Bu işlem genellikle bir veya birden fazla şablon olarak üzerinde çalışılan ve ileri aşamada çıkarım işlemini yönlendirecek bilginin genel biçimini tanımlamayı kapsama tekniğidir (Hamde, 2018: 35). Belki de günümüzde kullanılan bu teknik metin madenciliği uygulamaları arasındaki en önemli tekniktir. Bu teknik sayesinde sistemler yapısal halde olmayan metinleri veri tabanı tablosuna aktarabilecek bir formata dönüştürebilmektedir (Yıldız & Ağdeniz, 2018: 294).



Şekil 2.6. Bilgi çıkarma (Gaikwad, Chaugule & Patil, 2014: 44)

2.2.2.6. Doğal Dil İşleme

Doğal dil işleme (NLP) olarak da bilinen hesaplamalı dil bilim, bilgisayar biliminin insan dili içeriğini öğrenmek, anlamak ve üretmek için hesaplama tekniklerini kullanmakla ilgili bir alt alanıdır (Hirşehberg & Manning, 2015; 261). Doğal dil işleme yapılandırılmamış metinsel bilgilerin otomatik olarak işlenmesi ve analizi ile ilgilenir. NLP araştırmasının bir yönü, tipik olarak metinlerde bulunan sözcüklerin işlenmesini içeren istatistiksel tekniklere dayanır. Başka bir yönü ise ontolojiler, taksonomiler ve dilsel kural tabanları gibi bilgi kaynaklarından yararlanan kural tabanlı teknikleri kullanır (Popowich, 2005: 59). Metinsel dokümanlardan bilgi elde edilmesinde kullanılan NLP metotlarının temel amacı yapısal olmayan yani kesin formatlarla tanımlanamayan yapılardan bir çıkarım yapmaktır (Çalış, Gazdağı &Yıldız, 2013: 3).

Doğal dil işleme, daha çok yapay zekâ altındaki dil bilim bilgisine dayalı çalışmaları kapsamaktadır. Metin madenciliği çalışmaları ise daha çok istatistiksel olarak

metin üzerinden sonuçlara ulaşmayı amaçlamaktadır (Kahya, 2021: 11). İstatistiksel insan dili işleme sistemleri, istenen veya istenmeyen ilişkileri ve bağımlılıkları örnekleyen eğitim materyali koleksiyonları gerektirir. Sistemin müteakip modifikasyonu daha sonra sistemin bir dereceye kadar yeniden eğitilmesini gerekli kılar. Kural tabanlı teknikler, eğitim materyali gerektirmek yerine, çevrimiçi sözlükler, yerleşik dil teorileri biçiminde bilgi gerektirir ve mevcut sınıflandırma sistemlerini veya taksonomik çerçeveleri kullanabilirler. NLP uygulamaları bu tekniklerden herhangi birini veya her ikisini kullanabilir ve hangi tekniğin kullanılacağına karar vermek genellikle eğitim materyallerinin, dış kaynakların mevcudiyetine ve sonuçta ortaya çıkan uygulamada gereken gerçek metin analizi görevlerine bağlıdır (Popowich, 2005: 59).

2.2.2.7. Kavram Çıkarımı

Kavram çıkarımı, veri madenciliği uygulamaları arasında yer alan kavram madenciliğinin bir alt dalıdır. Kavram çıkarma çalışmalarında oldukça yüksek karmaşıklığa sahip olan terim tabanlı metin madenciliği çalışmalarının performansını artırmayı hedeflemektedir. Kavram çıkarmanın birinci aşaması dokümanları temsil eden ve belli aralıktaki terimlerin belirlenmesi işlemidir. Bu terimler daha sonra gruplandırılarak kavramlaşma yolunda önemli bir işlemi yerine getirir. Terimlerden kavramlara geçiş işlemlerinde de terimlerin benzerlik seviyelerine göre kümelenmesi ve daha sonra bu kümeler uzman tarafından etiketlenmesiyle yapılabilen bir çalışmadır. Buradaki aşamanın başarısı, seçilecek yöntemler ve veri setinin kalitesiyle yakından ilişkilidir (Balkan & Takçı, 2010: 1).

Kavram madenciliği metnin altındaki anlamla ilgilenmekle birlikte kullandığımız her kelime birden çok anlama sahip olabilmektedir. Bu karmaşıklığı ortadan kaldırabilmek için kavramı kullanırız. Kavram madenciliği dört aşama ile yapılır (Balkan & Takçı, 2010:1).

a- Söz dizimsel Analiz: Verilmiş olan dil bilgisi kuralları çerçevesinde kelimelerden oluşmuş bir metnin analiz edebilmesi için parçalara ayırmayı hedefler.

b- Anlamsal Analiz: Büyük doküman kümelerinden kavramları ifade edebilecek yapıları belirlemeyi amaçlar ve genellikle dokümanlar için daha önceden belirlenmiş anlamsal bilgileri kullanmaz.

c- İlişkilerin belirlenmesi: Anlamsal analiz ile elde edilen kavramların birbirleriyle olan ilişkilerini belirlemeyi amaçlar. Bu aşamada da kavram haritası madenciliği kullanılır.

d- Tasnif: Elde edilen kavramlar ve bu kavramlar arasındaki ilişkileri kullanarak dokümanların dizinlenmesini ve sınıflandırılmasını sağlar.

Bu dört aşamanın sonucunda sunulan başarılı değerlendirme çalışması sistemin kavramları ve ilişkileri öğrendiğini gösterir ve bu kavramlar dokümanları dizinlemede kullanılabilmektedir (Balkan & Takçı, 2010: 1).

2.3. R Programı

R arkasında çok geniş bir desteği olan açık kaynaklı ve birçok analizi yapma imkânını bünyesinde barındıran kütüphaneleriyle birçok kullanıcıya faydalı çözümler sunan (Atabay, Çizel & Ajanovic, 2019: 1136) R programlama dilinin temeli Jhon Chambers ve diğerleri tarafından 1976'da Bell Laboratuvarlarında S bilgisayar diline dayandırılmıştır. 1993'te de Yeni Zelanda Auckland Üniversitesinden Robert Gentleman ve Ross Ihaka, dili deneyerek bir uygulama geliştirirler ve buna R adını verirler. 1995 yılında da açık kaynak haline getirilerek dünyanın her yerinden binlerce insanın gelişmesine katkıda bulunurlar (Braun & Morduch, 2021: 3).

R ortamın esnek bir istatistiksel analiz aracı olarak farklı formatlardaki verilerin aktarılabilmesiyle ve bu veriler üzerinde istenilen değişikliklerin yapılmasıyla uygun istatistiksel yöntemlerin kullanılmasını sağlamaktadır. Aynı zamanda bu ortam klasik istatistik testleri, sınıflandırma, kümeleme, zaman serileri analizleri gibi büyük veriler üzerinde de analizler gerçekleştirme imkânı sunar. R'nin programlama dili birçok istatistik hesaplama dilinden daha kuvvetli bir nesneye yönelik programlama kabiliyetine de sahiptir (Sarıkaya, Doğan & Aktaş, 2017: 94/95).

Metin madenciliği ve metin analizi için de R programının kullanımı birçok avantaj sağlamaktadır. Çoğu programlama dilinin aksine R programı platformlar arası bir program olup açık kaynaklı ve ücretsizdir. Özellikle istatistiksel analizlerin yapılması için tasarlanmıştır. Bu da veri bilimi uygulamaları için de onu uygun kılmaktadır. R programının hızla bir şekilde talep görmesinin diğer bir sebebi ise R terminolojisinde "paket" ismi ile kullanım alanı geniş olan yazılım kütüphanelerinin bulunmasıdır. Ayrıca her paket de temel R dilini ve çekirdek paketlerin işlevselliğini arttırır. Veriler ve fonksiyonlara ek olarak çoğunlukla paketin kullanımını gösteren örnek formunda

formüller bulunmaktadır. R programıyla metin analizleri yapmanın en temel avantajlarından biri de farklı paketler arasında geçiş yapmanın veya bunları bir araya getirmenin genellikle mümkün ve nispeten kolay olabilmesidir (İçöz, 2021: 29).

R programlama dili ile SPSS ve minitab gibi programlarla gerçekleştirilen her türlü analiz uygulanabildiği gibi ilgili kütüphanelerin kullanılması ile de makine öğrenimi ve veri madenciliği alanları kullanılabilir. R programlama dili ―ggplot‖ ve seaborn gibi başarılı görselleştirme kütüphanelerinin yardımı ile raporlama imkânını da sunabilmektedir (Atan, 2020: 230).

2.4. Dergipark Uygulaması

TÜBİTAK-ULAKBİM tarafından ilk olarak bir "veri tabanı oluşturma projesi" olma özelliği ile ortaya çıkarılan ve daha sonraki aşamalarda genişletilerek sürekli bir gelişime kavuşturulan DergiPark uygulamasının temelleri 1993 yılına kadar uzanır. İlk etaptaki temel hedefi bibliyometrik analiz yapmak için gerekli olan bilgi kanalını açmak ve üniversiteler ile TÜBİTAK arasında bir veri yolu inşa etmek olarak aktarılmıştır. Daha sonra 2013 yılında belirtilen amaçlar doğrultusunda TÜBİTAK ULAKBİLİM tarafından kurulan DergiPark projesi, 2017 yılına kadar OJS üzerinden hizmet vermiştir. Daha sonra kullanıcı sayısının ve DergiPark'ta yayınlanan dergilerin hızlı bir şekilde artmasından dolayı OJS yeterli gelmemiştir. 2017 yılından itibaren ULAKBİM Dergi Sistemleri (UDS) adında, kullanıcı sayısı ve dergi artışından etkilenmeyecek şekilde çalışabilen yerli ve yeni bir dergi yönetim sistemi geliştirilerek hizmete sunulmuştur (Tamer, Övgün & Yalçıntaş, 2020: 101; Aslan, 2019: 206).

Daha sonra temel amacı sabit kalmakla birlikte (TR Dizin'e ve OpenAIRE'e4 temiz veri sağlanması ile devam etmekte) çeşitlenmiş dijital akademik ortamının geleceği açık erişimde olabileceğini düşünülerek açık erişim dergilere dönük bir altyapı sağlanması hedeflenmiştir. Bu bağlamda 2018 yılında YÖK ve TR Dizin ile entegre sistemler geliştirilmiştir. Aslında DergiPark, TR Dizin veya DOAJ gibi bir dizinleme sistemi olmayıp dergilerin ulusal ve uluslararası dizinlere girebilmesi için standart bir alt yapı hizmeti sunmaktadır. DergiPark'ın TR Dizinle entegrasyonu; dizine kolay başvuru, TR Dizin'e temiz veri aktarımı ve TR Dizin ile uyumlu gelişme süreçlerini içermektedir. DergiPark projesi, Türkiye'deki tüm akademik dergileri kapsayabilecek biçimde planlanmıştır (Tamer, Övgün & Yalçıntaş, 2020: 101; Aslan, 2019: 206).

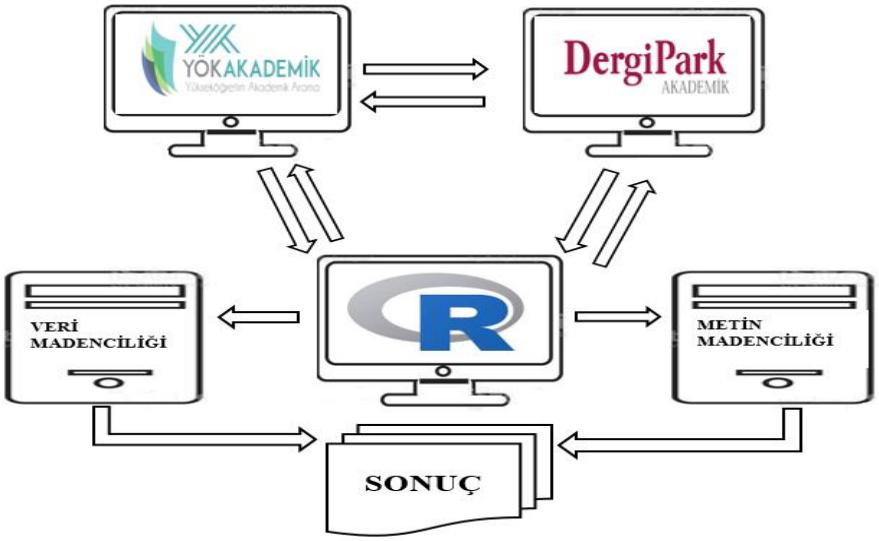
Dergipark’ın temel amacı akademik süreli yayıncılığın Türkiye'de kaliteli ve standartlara uygun olarak gelişmesini sağlayarak ulusal akademik dergilerin tüm dünyada görünürlüğünü ve kullanımını artırmak ve TR Dizin ve Ulusal Atıf Dizini için ölçülebilir temiz veri sağlamaktır (Dergipark, 2023).

3. BULGULAR VE TARTIŞMA

3.1. Veri Toplama Süreci

Yapılan bu tez çalışmasında 10.01.2023 ve 15.04.2023 tarihleri arasında YÖK Akademik web sayfasında bulunan fen bilimleri ve matematik temel alanında yer alan akademisyenlerin dergipark uygulaması üzerindeki yayınları araştırılmıştır. Bahsedilen tarihler arasında ve sonrasında akademik personel alımlarının gerçekleştirilmesi ve kadro değişikliklerinden dolayı güncel verilerle farklılık gösterebilmektedir.

İlk olarak 126 devlet üniversitesinden fen bilimleri ve matematik alanında görev yapan 11412 akademisyenin kadrolarının bulunduğu üniversite ve uzman oldukları bilim dallarına ait veriler YÖK Akademik web sayfasından elde edilmiştir. Daha sonra elde edilen veriler ile R programı kullanılarak dergipark uygulaması üzerinden akademisyenlere ait makale sayılarına ve makale özetlerine ulaşılmıştır. Bu süreci gösteren Şekil 3.1.'de verilmiştir.



Şekil 3.1. Veri ve metin analiz süreci

Yapılan çalışmalar ve analizler sonucunda veriler elde edildikten sonra R programlama dili ile veri madenciliği yöntemleri kullanılarak dağınık halde bulunan veriler ilk etapta kümele işlemleri yapılıp aynı kategori ve aynı sınıfta bulunan veriler bir araya getirilmiştir.

YÖK akademik web sayfasında bulunan fen bilimleri ve matematik temel alanında çalışmalarını sürdüren 11412 akademisyene ait bilgilere ulaşılmıştır. 7657 akademisyene ait dergipark uygulamasında 35774 yayına ulaşılırken, 3755 akademisyene ait herhangi bir yayına ulaşılmamıştır.

Daha sonra toplam akademisyen sayıları ve yayını bulunan akademisyen unvan dağılımı, bilim alanına göre dağılımı ve yayın sayılarına göre dağılım çizelgeler halinde verilerek yorumlanmıştır.

Üniversite bazındaki sıralamalarda devlet üniversiteleri isimleri için tesadüfi olarak Üni1, Üni2, … , Üni126 şeklinde kodlanma işlemleri yapılmıştır. Kodlama işleminden sonra dergipark uygulamasında en çok yayını bulunan ilk altı üniversite seçilmiştir. Seçilen bu üniversitelere bilim alanları bazında yayın sayılarına göre sıralama işlemleri yapılmıştır.

Son olarak dergipark uygulamasından elde edilen 35774 makaleye ait özetler, metin madenciliği yöntemleri kullanılarak bilim alanları bazında kelime frekansları ve kelime bulutları uluşturulmuştur.

3.2. Veri analizi

Fen bilimleri ve matematik temel alanında görev yapan 11412 akademisyenin yer aldığı veri kümesinin unvana göre dağılımı Çizelge 3.1. gösterilmiştir. Çizelgeye bakıldığı zaman 3724 ile profesör unvanı ilk sırada yer alırken ikinci sırada 2786'yla doktor öğretim üyesi yer almaktadır. Üçüncü sırada ise 2119'la doçent, dördüncü sırada 1396 ile öğretim görevlisi ve son sırada 1387'yle araştırma görevlisinin bulunduğu görülmektedir.

Çizelge 3.1. Ünvan dağılımına göre toplam akademisyen sayıları

Ünvan	Akademisyen Sayısı
Profesör	3724
Doçent	2119
Doktor Öğretim Üyesi	2786
Öğretim Görevlisi	1396
Araştırma Görevlisi	1387

11412 akademisyenin yer aldığı veri kümesinde 7657 akademisyenin dergipark uygulamasında yayımlanan yayınlarına ulaşılırken 3755 akademisyenin yayınına

ulaşılmamıştır. Dergipark uygulamasında yayını bulunan ve bulunmayan akademisyen sayılarının ünvana göre dağılımı Çizelge 3.2. de verilmiştir.

Çizelge 3.2. Yayını bulunan ve bulunmayan akademisyen sayıları

Ünvan	Yayını Bulunan	Yayını Bulunmayan
Profesör	2706	1018
Doçent	1745	374
Doktor Öğretim Üyesi	2073	713
Öğretim Görevlisi	584	812
Araştırma Görevlisi	549	838

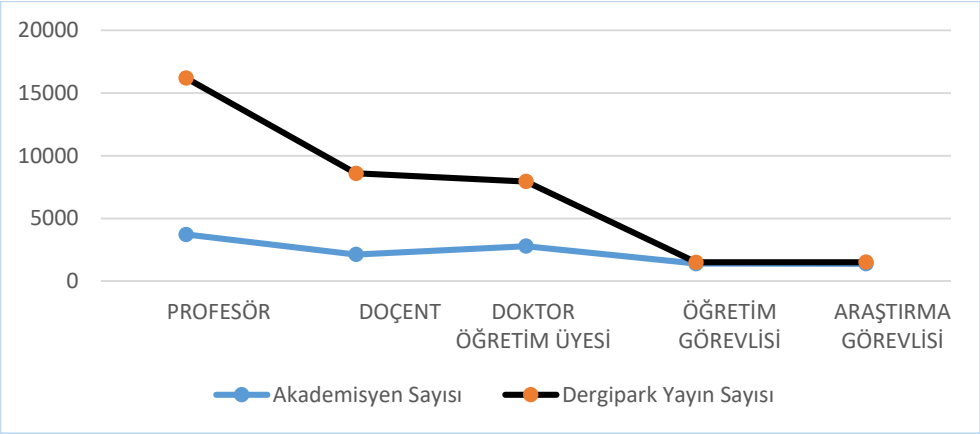
Çizelge 3.3'te dergipark uygulamasında yayımlanan 35774 yayının ünvanlara göre dağılımı verilmiştir.

Çizelge 3.3. Ünvan dağılımına göre dergipark yayın sayıları

Ünvan	Dergipark Yayın Sayısı
Profesör	16213
Doçent	8603
Doktor Öğretim Üyesi	7942
Öğretim Görevlisi	1503
Araştırma Görevlisi	1514

Çizelgeye göre ilk sırada 16212 yayın sayısı ile profesör ünvanı yer alırken, 8603 yayın sayısı ile doçent unvanı ikinci sırada yer almaktadır. Üçüncü sırada 7942 yayın sayısı ile doktor öğretim üyesi, dördüncü sırada 1514 yayın sayısı ile araştırma görevlisi ve beşinci sırada 1503 yayın sayısı ile öğretim görevlisi yer almaktadır.

Ünvan ve dergipark yayın sayılarının karşılaştırmalı grafiksel gösterimi Şekil 3.2'de verilmiştir.



Şekil 3.2. Ünvan ve dergipark yayın sayılarının grafiksel gösterimi

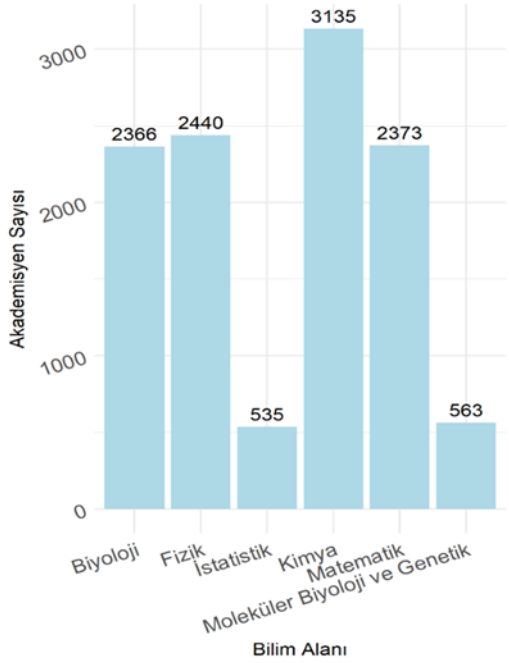
Grafiğe bakıldığı zaman bazı istisnai durumlar hariç yayın sayılarının akademik derece kademe ilerlemesine bağlı olarak yükseldiği görülmektedir.

11412 akademisyenin bilim alanları dağılımına göre akademisyen sayıları Çizelge 3.4.'te verilmiştir. 3135 ile en yüksek akademisyenin görev aldığı bilim alanı olan kimya ilk sırada yer alırken, 2440 akademisyenin bulunduğu fizik alanı ikinci sırada yer almaktadır. Üçüncü sırada ise 2373 ile matematik alanı bulunmakta ve dördüncü sırada da 2366 akademisyenle biyoloji alanı bulunmaktadır. Beşinci sırada 563 akademisyenle moleküler biyoloji ve genetik gelmekte ve son sırada 535 kişi ile istatistik yer almaktadır.

Çizelge 3.4. Bilim alanı dağılımına göre toplam akademisyen sayıları

Bilim Alanı	Akademisyen Sayısı
Biyoloji	2366
Fizik	2440
İstatistik	535
Kimya	3135
Matematik	2373
Moleküler Biyoloji ve Genetik	563

Şekil 3.3'te Bilim alanı dağılımına göre toplam akademisyen sayılarının grafiksel dağılımı verilmiştir.



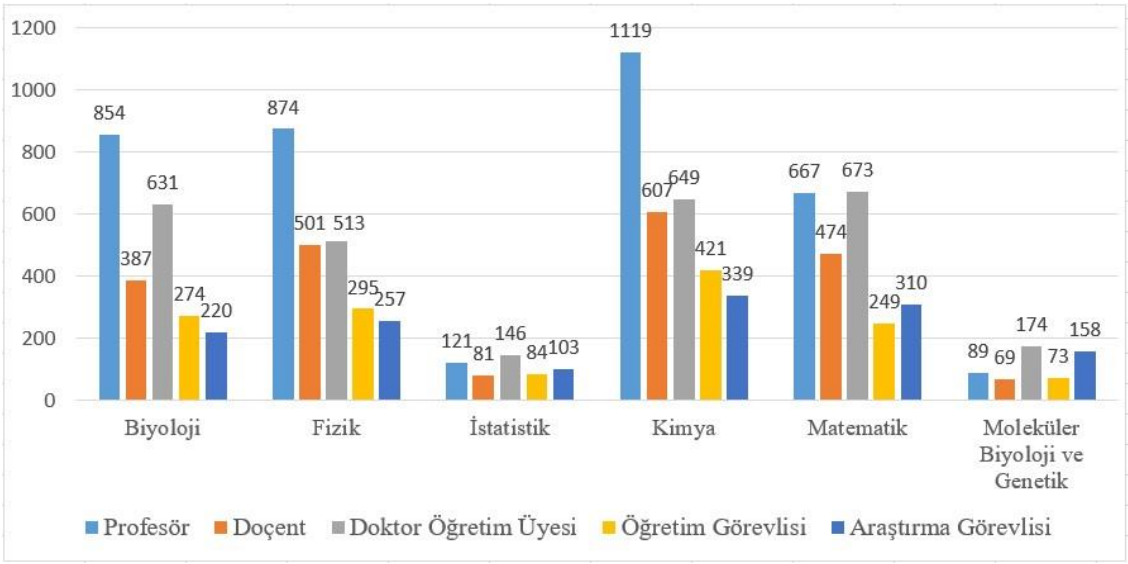
Şekil 3.3. Bilim alanı dağılımına göre toplam akademisyen sayılarının grafiksel gösterimi

11412 akademisyenin bilim alanı ve unvan dağılımı Çizelge 3.5'te verilmiştir. Çizelgeye bakıldığı zaman profesör unvanının en yüksek olduğu bilim alanı 1119 akademisyenin görev aldığı kimya, doçent alanında ise 607 kişi ile matematik bölümü gelmektedir. Doktora öğretim üyesinin en çok görev aldığı alan 673 akademisyen ile matematiktir. Öğretim görevlisi ünvanında 421 akademisyen ile kimya ve araştırma görevlisi ünvanında da 339 akademisyenle yine kimya alanı başta gelmektedir.

Çizelge 3.5. Bilim alanı dağılımına göre ünvan sayıları

Bilim Alanı	Profesör	Doçent	Doktor Öğretim Üyesi	Öğretim Görevlisi	Araştırma Görevlisi
Biyoloji	854	387	631	274	220
Fizik	874	501	513	295	257
İstatistik	121	81	146	84	103
Kimya	1119	607	649	421	339
Matematik	667	474	673	249	310
Moleküler Biyoloji ve Genetik	89	69	174	73	158

Bilim alanı dağılımına göre ünvan sayılarının grafiksel dağılımı Şekil 3.4'te gösterilmiştir.



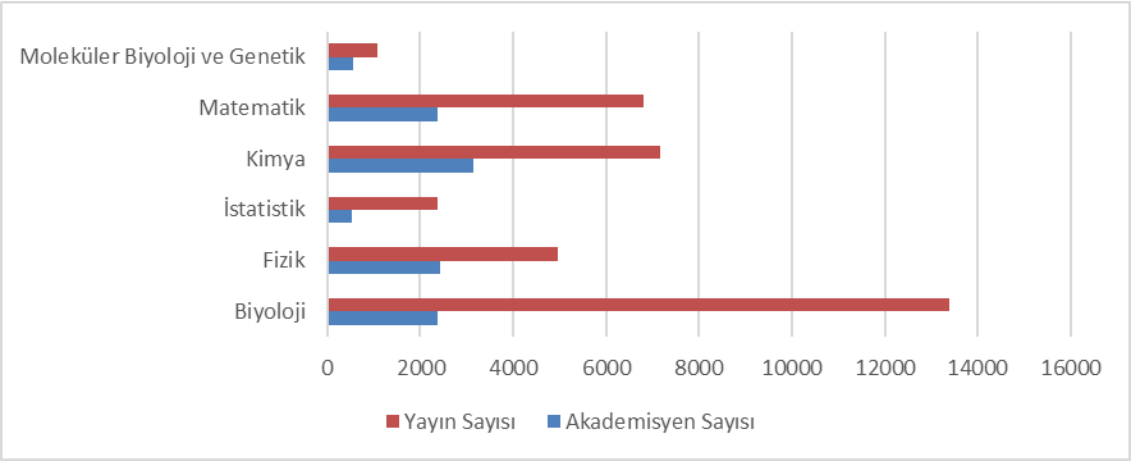
Şekil 3.4. Bilim alanı dağılımına göre ünvan sayılarının grafiksel gösterimi

Bilim alanı dağılımına göre dergipark uygulamasında yayını bulunan 7657 akademisyenin yayın sayılarının dağılımı Çizelge 3.6.'da verilmiştir.

Çizelge 3.6. Bilim alanı dağılımına göre yayını bulunan akademisyen sayıları ve dergipark yayın sayıları

Bilim Alanı	Akademisyen Sayısı	Dergipark Yayın Sayısı
Biyoloji	1944	13396
Fizik	1410	4950
İstatistik	418	2370
Kimya	1988	7170
Matematik	1598	6814
Moleküler Biyoloji ve Genetik	299	1074

Bilim alanı dağılımına göre akademisyen sayısı en yüksek 1988 ile kimya alanında olmasına rağmen bu alanda yayınlanan yayın sayısı 7170 ile ikinci sırada kalmıştır. Buna karşın 1944 akademisyen sayısı ile ikinci sırada yer alan biyoloji alanı 13396 yayın sayısı ile birinci sırada yer almaktadır. 1598 akademisyen sayısı ve 6814 yayın ile matematik alanı üçüncü sırada bulunmakta ve 1410 akademisyen ile dördüncü sırada yer alan fizik temel alanı 4950 yayın ile dördüncü sırada yer almaktadır. İstatistik temel alanı 418 akademisyen ve 2370 yayın ile beşinci sırada, moleküler biyoloji ve genetik alanı 299 akademisyen ve 1074 yayınla son sırada yer almaktadır.



Şekil 3.5. Bilim alanlarına göre akademisyen ve yayın sayılarının grafiksel gösterimi

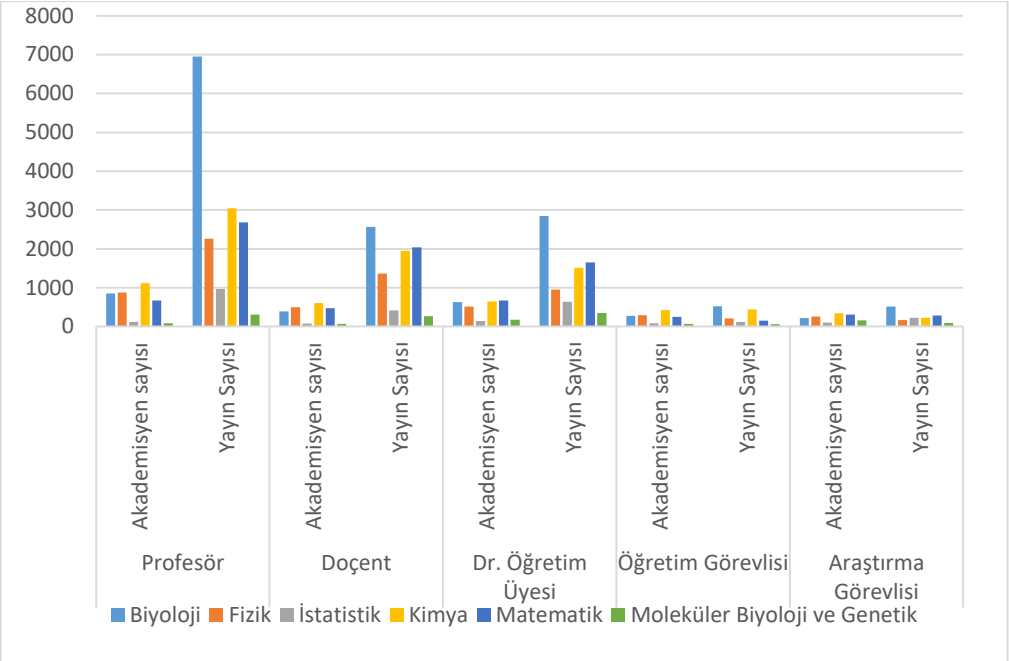
Şekil 3.5. incelendiğinde yayın sayısının akademisyen sayısı ile orantılı olmadığı ve bunun temel nedeni olarak da bazı temel alanlarda çalışma alanın dar olduğu ve bunun yayın sayısını etkilediğini düşünülmektedir.

Çizelge 3.7.'de 7657 akademisyene ait 35774 yayının bilim alanı ve unvan dağılımına göre sayıları verilmiştir.

Çizelge 3.7. Bilim alanı ve ünvan dağılımına göre dergipark yayın sayıları

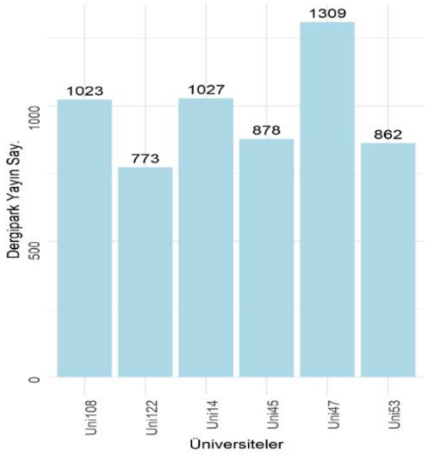
Bilim Alanı	**Profesör**	**Doçent**	**Doktor Öğretim Üyesi**	**Öğretim Görevlisi**	**Araştırma Görevlisi**	**Toplam**
Biyoloji	6948	2562	2846	523	517	13396
Fizik	2261	1364	949	212	164	4950
İstatistik	971	417	634	120	228	2370
Kimya	3045	1951	1508	440	226	7170
Matematik	2680	2042	1654	151	287	6814
Moleküler Biyoloji ve Genetik	307	267	351	57	92	1074

Şekilde 3.6. da bilim alanı, akademisyen sayısı ve yayın sayısının grafiksel olarak gösterimi verilmiştir. Grafikte de görüldüğü gibi ünvan ve yayın sayısının birbiri ile orantılığı olduğu görülmektedir.



Şekil 3.6. Bilim alanı, akademisyen sayısı ve yayın sayısının grafiksel gösterimi

Şekil 3.7. da dergipark yayın sayısına göre ilk altı üniversite verilmiştir. Burada 1309 yayın ile Uni47 ilk sırada yer alırken ikinci sırada 1027 yayın ile Uni14 üçüncü sırada 1023 yayın ile Uni108 yer almaktadır. 878 yayın ile Uni45 dördüncü sırada iken 862 yayın ile Uni53 beşinci sırada ve son olarak 773 yayın ile Uni122 son sırada yer almaktadır.



Şekil 3.7. Dergipark yayın sayısına göre ilk altı üniversite

Çizelge 3.8'de ise ilk altı sırada yer alan üniversitelerin bilim alanı dağılımına göre Dergiparkta yayımlanan yayın sayıları verilmiştir. Burada 6 bilim alanına karşılık en yüksek yayına sahip olan 6 üniversite seçilmiştir.

Çizelge 3.8. Dergipark yayın sayısına göre en yüksek ilk 6 üniversite ve bilim alanına göre dağılımı

Bilim Alanı	UNİ14	UNİ45	UNİ47	UNİ53	UNİ108	UNİ122
Biyoloji	427	311	539	304	554	363
Fizik	127	161	206	26	160	38
İstatistik	173	52	199	309	127	13
Kimya	140	175	165	171	95	220
Matematik	145	164	195	48	87	96
Moleküler Biyoloji ve Genetik	15	15	2	4	0	43
Toplam Yayın Sayıları	1027	878	1306	862	1023	773

Çizelge 3.8. incelendiğinde biyoloji temel alanında en yüksek yayına sahip olan Uni108'in 554 yayın ile ilk sırada yer almaktadır. Fizik alanına bakıldığı zaman 206 yayın ile Uni47 ilk sırada bulunur. İstatistik alanında ise 309 yayın ile Uni53 ilk sırada gelmektedir. Kimya alanında 220 yayın ile ilk sırada yer alan Uni122 ayrıca moleküler biyoloji ve genetik alanında da 43 yayın ile ilk sırada yer almaktadır. Matematik alanında ise 195 yayın ile Uni47 ilk sırada yer almaktadır.

Yapılan sıralamalar sonucunda:

- Biyoloji alanında Uni108
- Fizik alanında Uni47
- İstatistik alanında Uni53
- Kimya alanında Uni122
- Matematik alanında Uni47
- Moleküler biyoloji ve genetik alanında Uni122 yer almaktadır.

3.3. Metin Analizi

Toplamda 13396 biyoloji, 4950 fizik, 7170 kimya, 2370 istatistik, 6814 matematik ve 1074 moleküler biyoloji ve genetik bilim alanında olmak üzere toplamda 35774 makale özeti R programı kullanılarak dergipark uygulamasından çekilmiştir. Daha sonra çekilen makale özetlerinin analiz işlemleri yapılmıştır. Analiz işlemlerinde makale özetlerinde en çok kullanılan ilk 100 kelimeye ait kelime frekans tabloları oluşturulmuştur. Tablo işlemlerinden sonra makale özetlerinde en çok kelimeler ile R programında wordcloud komutu kullanılarak kelime bulutları oluşturulmuştur.

Çizelge 3.9. da biyoloji bilim alanına ait kelime frekansları verilmiştir. Çizelge incelendiğinde makalelerde en sık kullanılan 100 kelime içerisinde en yüksek frekans değerine sahip olan ilk beş kelime "species (n=2884)", "bitki (n=2310)", "takson(n=2121)", "su (n=1954)", "asit (n=1542)" olduğu görülmektedir.

İlk beş kelimenin her bir makalede kullanılma oranı hesaplandığında ise "species=0,21", "bitki=0,17", "takson=0,15", "su=0,14", "asit=0,11" olarak çıkmaktadır.

Biyoloji bilim alanında en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\bar{X} = 775{,}05$ çıkmaktadır.

Çizelge 3.9. Biyoloji bilim alanında en çok kullanılan 100 kelime

Kelime	**n**	**Kelime**	**n**	**Kelime**	**n**	**Kelime**	**n**
species	2884	extract	849	significant	659	alanı	546
bitki	2310	biyolojik	836	erkek	653	content	543
takson	2121	control	830	anadolu	646	balık	537
su	1954	coli	812	türünün	638	familya	534
asit	1542	yayılış	812	saat	637	grup	533
activity	1538	subsp	804	aktiviteleri	625	ekolojik	532
antioksidan	1419	activities	795	besin	624	arazi	531
metal	1209	aktivite	789	taksonların	614	anatomik	524
dna	1128	aureus	781	kayıt	613	gövde	517
hücre	1112	türleri	759	fenolik	612	bacteria	515
protein	1080	endemik	758	antimicrobial	603	çevre	511
türün	1040	türlerinin	752	insan	602	kromozom	509
yaprak	1013	gölü	747	boy	599	gen	503
extracts	1009	vitro	731	izole	598	moleküler	497
taxa	1000	etki	730	gıda	595	uzunluğu	491
yağ	988	amp	709	akdeniz	587	dişi	477
acid	987	bölgesi	685	bakteri	584	ağırlık	474
antimikrobiyal	965	plants	685	pozitif	573	büyüme	467
plant	943	bitkilerin	681	madde	569	meyve	464
türlerin	928	grubu	676	highest	565	uçucu	462
morfolojik	921	enzim	670	cins	558	kültür	458
antioxidant	884	bacillus	666	growth	558	tanesi	456
kök	868	toprak	666	stress	556	genetik	454
aktivitesi	857	polen	664	anti	554	bitkiler	451
cells	855	cell	662	staphylococcus	547	identified	446

Şekil 3.8. de biyoloji bilim alanına ait kelime bulutu verilmiştir. Kelime bulutunda kelime tekrar sıklığına bağlı olarak "species", "bitki", "takson", "su", "asit" kelimelerinin

daha büyük yazıldığı ve merkezde yer aldığı, diğer kelimelerin de kullanılma oranına bağlı olarak merkezden uzaklaşarak küçüldüğü görülmektedir.

Şekil 3.8. Biyoloji bilim alanı kelime bulutu

Çizelge 3.10. da fizik bilim alanına ait kelime frekansları verilmiştir. Çizelge incelendiğinde makalelerde en sık kullanılan 100 kelime içerisinde en yüksek frekans değerine sahip olan ilk beş kelime "enerji (n=1081)", "ince (n=833)", "optik (n=777)", "temperature (n=671)", "kristal (n=600) olduğu görülmektedir.

İlk beş kelimenin her bir makalede kullanılma oranı hesaplandığında ise "eneji=0,21", "ince=0,16", "optik=0,15", "temperature=0,13", "kristal=0,12 olarak çıkmaktadır.

Fizik bilim alanında en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması ise $\bar{X} = 302{,}39$ çıkmaktadır.

Çizelge 3.10. Fizik bilim alanında en çok kullanılan 100 kelime

Kelime	**n**	**Kelime**	**n**	**Kelime**	**n**	**Kelime**	**n**
enerji	1081	surface	349	phase	263	functional	193
ince	833	elektronik	345	radiation	262	teorisi	193
optik	777	electron	341	enerjisi	261	malzeme	192
temperature	671	doz	330	spektroskopisi	256	alloys	188
kristal	600	nükleer	320	gamma	255	aktivite	187
elektron	598	cam	318	filmler	254	diffraction	182
ışını	579	soğurma	317	oksit	251	reaksiyon	182
manyetik	578	spin	306	electronic	247	thermal	182
structure	575	elektriksel	302	termal	247	type	182
filmlerin	561	yoğunluk	300	dalga	241	fonksiyonu	181
film	495	bant	296	state	240	solar	181
sıcaklık	449	radon	295	crystal	234	energies	179
metal	433	magnetic	281	yoğunluğu	233	proton	176
moleküler	408	elektrik	277	su	232	nuclear	175
experimental	394	taramalı	276	layer	224	reaction	167
gama	388	fiziksel	275	dielektrik	217	voltage	166
density	384	mikroskobu	274	tio	216	dielectric	165
ray	379	faz	273	enerjileri	214	hidrojen	163
radyasyon	375	potansiyel	272	mekanik	211	elastik	159
optical	372	structural	272	molecule	211	homo	158
band	370	field	271	sistemin	210	voltaj	157
kütle	369	nano	271	electrical	206	simulation	155
molecular	369	ışık	270	atomik	197	current	154
kuantum	359	atom	266	gas	196	organik	152
katkılı	356	chemical	264	quantum	195	saf	151

Şekil 3.9'da fizik bilim alanına ait kelime bulutu verilmiştir. Kelime bulutunda kelime tekrar sıklığına bağlı olarak "enerji", "ince", "optik", "temperature", "kristal" kelimelerinin daha büyük yazıldığı ve merkezde yer aldığı, diğer kelimelerinde kullanılma oranına bağlı olarak merkezden uzaklaşarak küçüldüğü görülmektedir.

Şekil 3.9. Fizik bilim alanı kelime bulutu

Çizelge 3.11'de kimya bilim alanına ait kelime frekansları verilmiştir. Çizelge incelendiğinde makalelerde en sık kullanılan 100 kelime içerisinde en yüksek frekans değerine sahip olan ilk beş kelime "metal (n=1497)", "asit (n=1413)", "activity (n=1347)", "antioksidan (n=1097)", "synthesized (n=1046) olduğu görülmektedir.

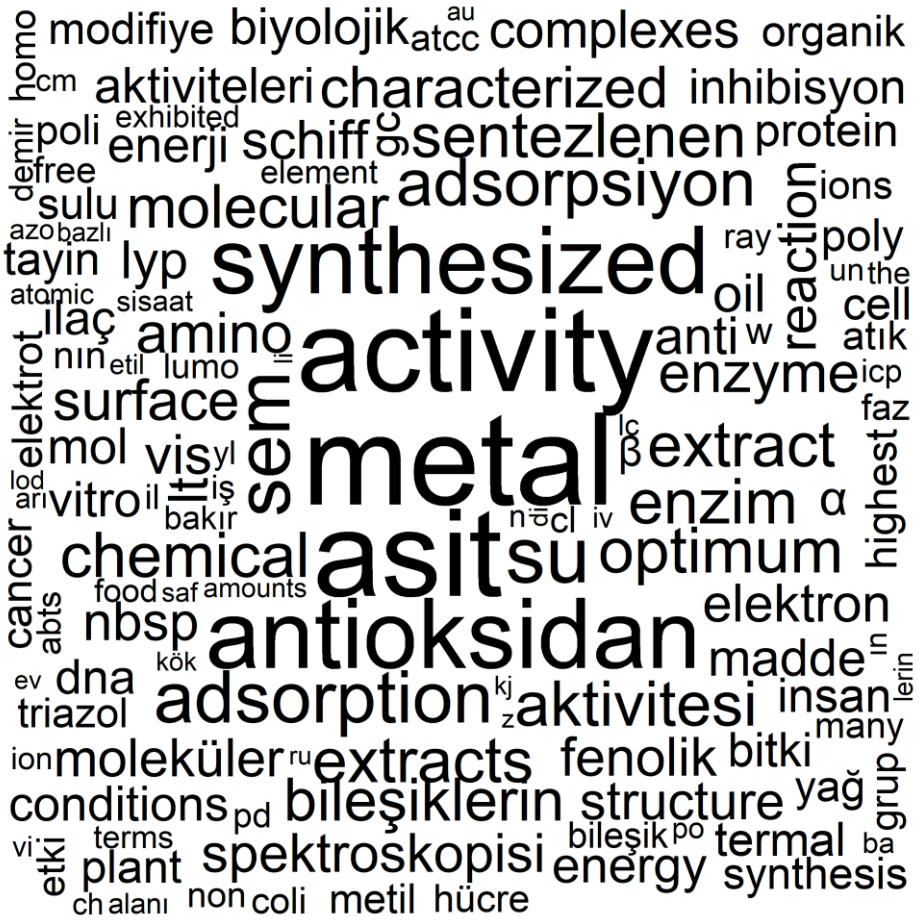
İlk beş kelimenin her bir makalede kullanılma oranı hesaplandığında ise "metal=0,20", "asit=0,19", "activity=0,18", "antioksidan=0,15", "synthesized=0,14 olarak çıkmaktadır.

Kimya bilim alanında en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\bar{x} =$474,05 çıkmaktadır.

Çizelge 3.11. Kimya bilim alanında en çok kullanılan 100 kelime

Kelime	n	Kelime	n	Kelime	n	Kelime	n
metal	1497	moleküler	495	plant	405	triazol	338
asit	1413	elektron	490	tayin	405	taramalı	337
activity	1347	lyp	490	nanoparticles	399	radikal	335
antioksidan	1097	structure	488	cell	396	ligand	334
synthesized	1046	complexes	486	poly	393	content	333
su	854	anti	484	modifiye	383	inhibition	332
sem	818	madde	477	sulu	382	karbon	330
adsorption	759	biyolojik	467	electron	379	atık	326
optimum	638	energy	463	synthesis	379	bileşikler	326
enzim	634	mol	458	yağ	377	ions	324
chemical	628	aktiviteleri	451	derivatives	374	phenolic	324
aktivitesi	627	enerji	447	organik	370	standart	324
extracts	627	temperature	446	sentezlendi	370	drug	321
sentezlenen	617	vitro	443	highest	369	human	321
molecular	593	inhibisyon	441	cancer	368	infrared	319
bileşiklerin	588	conditions	436	complex	364	aqueous	318
characterized	562	ilaç	431	antimikrobiyal	362	techniques	318
amino	553	spectroscopy	431	groups	357	grup	316
enzyme	551	dna	430	sıcaklık	357	etki	314
surface	549	bitki	423	thermal	357	capacity	309
reaction	511	insan	423	grubu	354	langmuir	309
fenolik	510	experimental	421	elektrot	352	determination	304
schiff	508	termal	421	higher	349	hücre	302
spektroskopisi	508	protein	416	poli	347	novel	301
oil	496	elektrokimyasal	406	metil	338	uçucu	299

Şekil 3.10'da kimya bilim alanına ait kelime bulutu verilmiştir. Kelime bulutunda kelime tekrar sıklığına bağlı olarak "metal", "asit", "activity", "antioksidan", "synthesized" kelimelerinin daha büyük yazıldığı ve merkezde yer aldığı, diğer kelimelerinde kullanılma oranına bağlı olarak merkezden uzaklaşarak küçüldüğü görülmektedir.



Şekil 3.10 Kimya bilim alanı kelime bulutu

Çizelge 3.12'de istatistik bilim alanına ait kelime frekansları verilmiştir. Çizelge incelendiğinde makalelerde en sık kullanılan 100 kelime içerisinde en yüksek frekans değerine sahip olan ilk beş kelime "regresyon (n=623)", "proposed (n=455)", "estimators (n=428)", "models (n=401)", "simulation (n=366) olduğu görülmektedir.

İlk beş kelimenin her bir makalede kullanılma oranı hesaplandığında ise "regresyon=0,26", "proposed=0,19", "estimators=0,18", "models=0,16", "simolation=0,15" olarak çıkmaktadır.

İstatistik bilim alanında en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\bar{x} = 169{,}36$ çıkmaktadır.

Çizelge 3.12. İstatistik bilim alanında en çok kullanılan 100 kelime

Kelime	n	Kelime	n	Kelime	n	Kelime	n
regresyon	623	dağılım	190	probability	141	fonksiyonu	117
proposed	455	covid	184	yöntemler	139	functions	117
estimators	428	approach	183	least	135	performans	117
models	401	değişkenler	183	series	135	sampling	117
simulation	366	kareler	183	ilişki	132	küme	116
performance	305	simülasyon	182	sosyal	132	analizinde	114
function	278	gerçek	181	çoklu	132	bulgular	114
parameter	260	linear	179	etkileyen	131	sınıflandırma	114
likelihood	258	robust	177	kümeleme	131	uygulanmıştır	113
maximum	255	karar	176	lojistik	131	örnek	113
risk	247	tests	174	olabilirlik	131	dağılımının	112
problem	242	fuzzy	172	non	129	bağımsız	111
mean	238	önerilen	164	square	128	poisson	110
error	233	statistical	161	tahmini	128	değişkenli	109
real	215	ekonomik	159	many	127	log	109
type	214	modelleri	156	yaşam	127	performances	109
hata	213	random	155	örnekleme	126	ridge	109
variables	211	statistics	153	modeller	124	better	108
carlo	209	parametre	150	dağılımı	121	confidence	108
monte	209	algorithm	149	etki	121	density	108
weibull	205	countries	148	bayes	120	exponential	108
estimator	197	compare	147	estimates	118	generalized	105
değişken	195	klasik	145	information	118	amaçlanmıştır	103
terms	194	variable	145	aim	117	power	102
distributions	192	considered	141	copula	117	algoritması	101

Şekil 3.11'de istatistik bilim alanına ait kelime bulutu verilmiştir. Kelime bulutunda kelime tekrar sıklığına bağlı olarak "regresyon", "proposed", "estimators", "models", "simulation" kelimelerinin daha büyük yazıldığı ve merkezde yer aldığı, diğer kelimelerinde kullanılma oranına bağlı olarak merkezden uzaklaşarak küçüldüğü görülmektedir.

Şekil 3.11. İstatistik bilim alanı kelime bulutu

Çizelge 3.13'te matematik bilim alanına ait kelime frekansları verilmiştir. Çizelge incelendiğinde makalelerde en sık kullanılan 100 kelime içerisinde en yüksek frekans değerine sahip olan ilk beş kelime "space (n=1414)", "function (n=1184)", "equation (n=1127)", "solutions (n=960)", "soft (n=836) olduğu görülmektedir.

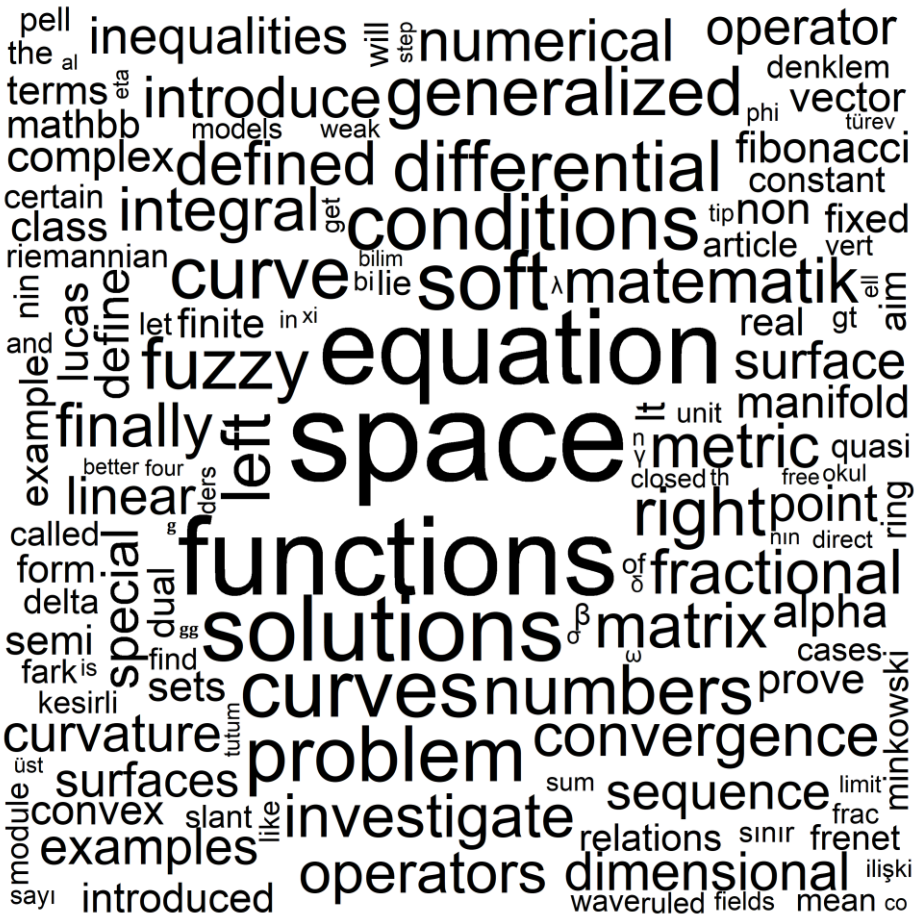
İlk beş kelimenin her bir makalede kullanılma oranı hesaplandığında ise "space=0,20", "function=0,17", "equation=0,16", "solutions=0,14", "soft=0,12" olarak çıkmaktadır.

Matematik bilim alanında en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\bar{x} = 434{,}08$ çıkmaktadır.

Çizelge 3.13. Matematik bilim alanında en çok kullanılan 100 kelime

Kelime	n	Kelime	n	Kelime	n	Kelime	n
space	1414	operators	538	prove	340	main	290
functions	1184	introduce	529	class	338	euclidean	286
equation	1127	dimensional	477	vector	337	modules	284
solutions	960	linear	469	theorem	333	stability	284
soft	836	surface	462	convex	332	proposed	280
curves	823	curvature	460	difference	332	dual	279
conditions	791	sequence	454	diferansiyel	329	points	279
problem	781	inequalities	450	semi	328	relations	275
curve	736	special	449	boundary	323	manifolds	270
fuzzy	705	examples	445	finite	319	general	262
differential	698	operator	444	introduced	319	product	262
numbers	673	point	442	mathcal	318	second	260
generalized	666	moreover	412	statistical	311	constant	254
right	662	surfaces	412	terms	311	many	252
left	653	define	409	fixed	310	ring	252
matematik	606	alpha	397	topological	309	article	250
matrix	603	problems	395	makalede	307	verilmiştir	250
fractional	585	manifold	380	mathbb	307	consider	245
metric	585	sequences	374	example	306	delta	244
integral	573	complex	367	frame	306	initial	243
defined	572	nonlinear	363	concept	302	matrices	242
finally	572	presented	358	known	299	frenet	241
investigate	567	fibonacci	356	related	297	shown	241
convergence	560	lucas	341	real	295	minkowski	238
numerical	546	sets	341	polynomials	294	theorems	236

Şekil 3.12. de matematik bilim alanına ait kelime bulutu verilmiştir. Kelime bulutunda kelime tekrar sıklığına bağlı olarak “space”, “function”, “equation”, “solutions”, “soft” kelimelerinin daha büyük yazıldığı ve merkezde yer aldığı, diğer kelimelerin de kullanılma oranına bağlı olarak merkezden uzaklaşarak küçüldüğü görülmektedir.



Şekil 3.12. Matematik bilim alanı kelime bulutu

Çizelge 3.14'te moleküler biyoloji ve genetik bilim alanına ait kelime frekansları verilmiştir. Çizelge incelendiğinde makalelerde en sık kullanılan 100 kelime içerisinde en yüksek frekans değerine sahip olan ilk beş kelime "dna (n=358)", "hücre (n=255)", "gen (n=248)", "protein (n=221)", "cancer (n=218) olduğu görülmektedir.

İlk beş kelimenin her bir makalede kullanılma oranı hesaplandığında ise "dna=0,33", "hücre=0,23", "gen=0,23", "protein=0,20", "cancer=0,20" olarak çıkmaktadır.

Moleküler biyoloji ve genetik bilim alanında en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\bar{X} = 93{,}27$ çıkmaktadır.

Çizelge 3.14. Moleküler biyoloji ve genetik bilim alanında en çok kullanılan 100 kelime

Kelime	n	Kelime	n	Kelime	n	Kelime	n
dna	358	antioxidant	110	filogenetik	72	higher	57
hücre	255	acid	107	turkey	70	saat	57
gen	248	coli	106	etki	69	staphylococcus	57
protein	221	extracts	106	proteins	69	chromosome	56
cancer	218	su	105	identified	68	potansiyel	56
species	194	biyolojik	101	enzyme	66	aktiviteleri	55
activity	189	activities	97	patients	66	bitkilerin	55
bitki	182	genetic	93	antimicrobial	65	conclusion	55
kanser	171	aureus	92	lines	65	detected	55
moleküler	160	bacteria	91	biological	64	sitotoksik	55
genetik	155	enzim	91	ilaç	64	strains	55
human	147	bacillus	89	including	64	gıda	54
asit	146	tür	89	kök	64	amino	53
insan	144	hücrelerinde	88	metal	64	izolatların	53
antioksidan	141	control	86	significant	64	products	53
expression	139	grup	82	adamts	62	tedavisinde	53
genes	137	direnç	80	diseases	62	üretimi	53
izole	135	isolates	79	grubu	60	breast	52
vitro	134	groups	77	real	60	kültür	52
antimikrobiyal	121	growth	77	assay	58	yaprak	52
aktivite	120	plants	77	bölgesi	58	antibakteriyel	51
molecular	118	pozitif	77	development	58	geni	51
anti	114	extract	76	research	58	health	51
aktivitesi	113	kanseri	76	stres	58	morfolojik	51
plant	113	conditions	74	etanol	57	sequence	51

Şekil 3.13'te moleküler biyoloji ve genetik bilim alanına ait kelime bulutu verilmiştir. Kelime bulutunda kelime tekrar sıklığına bağlı olarak "dna", "hücre", "gen", "protein", "cancer" kelimelerinin daha büyük yazıldığı ve merkezde yer aldığı, diğer kelimelerin de kullanılma oranına bağlı olarak merkezden uzaklaşarak küçüldüğü görülmektedir.

Şekil 3.13. Moleküler biyoloji ve genetik bilim alanı kelime bulutu

4. SONUÇ ve ÖNERİLER

Bu tez çalışmasında, Türkiye'deki devlet üniversitelerinin fen bilimleri ve matematik temel alanında görev yapan akademisyenlerin dergipark uygulamasında yayımlanan makaleleri R programı ile veri ve metin madenciliği yöntemleri kullanılarak analizler yapılmış ve kelime frekanslarıyla kelime bulutları oluşturulmuştur.

Toplamda 126 devlet üniversitesinden fen bilimleri ve matematik alanında görev yapan 11412 akademisyenin dergipark uygulamasında yayımlanan makaleleri araştırılmıştır. 3755 akademisyene ait herhangi bir yayına ulaşılmazken 7657 akademisyene ait ise 35774 yayına ulaşılmıştır.

Bilim alanları dağılımına göre dergipark uygulamasında yayını bulunan akademisyen ve yayın sayıları şöyledir:

Biyoloji: Biyoloji bilim alanında 1944 akademisyene ait 13396 yayına ulaşılmıştır. Ortalama yayın sayısı 6,8 ve makale özetlerinde en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\overline{X} = 775{,}05$ çıkmıştır.

Fizik: Fizik bilim alanında 1410 akademisyene ait 4950 yayına ulaşılmıştır. Ortalama yayın sayısı 3,5 ve makale özetlerinde en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\overline{X}=302{,}39$ çıkmıştır.

Kimya: Kimya bilim alanında 1988 akademisyene ait 7170 yayına ulaşılmıştır. Ortalama yayın sayısı 3,6 ve makale özetlerinde en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\overline{X}=474{,}05$ çıkmıştır.

İstatistik: İstatistik bilim alanında 418 akademisyene ait 2370 yayına ulaşılmıştır. Ortalama yayın sayısı 5,6 ve makale özetlerinde en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\overline{X}=169{,}36$ çıkmıştır.

Matematik: Matematik bilim alanında 1598 akademisyene ait 6814 yayına ulaşılmıştır. Ortalama yayın sayısı 4,2 ve makale özetlerinde en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\overline{X}=434{,}08$ çıkmıştır.

Moleküler biyoloji ve genetik: Moleküler biyoloji ve genetik bilim alanında 299 akademisyene ait 1074 yayına ulaşılmıştır. Ortalama yayın sayısı 3,5 ve makale

özetlerinde en çok kullanılan ilk 100 kelimenin kelime frekansları toplamının aritmetik ortalaması $\overline{X}$=93,27 çıkmıştır.

Çalışmada birden fazla akademisyenin aynı ismi ve soy ismi taşıdığını gözlemlenmiştir. Bunun da araştırmacılar tarafından akademisyenlere ait yayınlara yapılacak olan atıfların karıştırılmasına neden olabileceği düşünülmektedir.

Dergipark uygulaması bünyesinde sadece anlaşmalı dergi yayınlarının bulunması, çalışmada birtakım eksikliklere neden olmuştur.

Yapılan analizler sonucunda yayın sayılarının akademisyen sayıları ile orantılı olmadığı görülmüştür. Bunun temel nedeni ise bazı bilim alanlarının kaynak ve konu bakımından geniş olmasından dolayı yayın yapma imkânının fazla olmasıdır. Bu da araştırmacılara yüksek yayın imkânı tanımasına karşın bazı temel alanların konu bakımından dar olmasından kaynaklı olarak yayınların düşük seviyede kalmasına neden olmuştur. Bu durum akademik derece kademe ilerlemeye, akademik teşvik puanlamaya ve akademik ödeme teşviklerine etki etmektedir. Bunun için akademik derece kademe ilerlemesinin, akademik teşvik puanlamasının ve akademik ödeme teşviklerinin bilim alanı bazında olması gerektiği düşünülmektedir.

KAYNAKLAR

Akgün, A. (2012). *Seyahat Acentalarında Veri Madenciliği: Antalya Bölgesinde Bir Uygulama*[Yayınlanmış Yüksek Lisans Tezi]. Akdeniz Üniversitesi.

Akgün, K., & Özek, M.B. (2020). Eğitsel Veri Madenciliği Yöntemi İle İlgili Yapılmış Çalışmaların İncelenmesi: *İçerik Analizi. Uluslararası Eğitim Bilim Ve Teknoloji Dergisi, 6*(3), 197-2013. Doi:10.47714/uebt.753526

Akın, Y.K. (2008). *Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi*[Yayınlanmış Doktora Tezi]. Marmara Üniversitesi.

Altay, A. (2007*). Genetik Algoritma ve Bir Uygulama*[Yayınlanmış Yüksek Lisans Tezi]. İstanbul Teknik Üniversitesi.

Altunkaynak, A., Başakın, E.E., & Kartal, E. (2020). Dalgacık K-en Yakın Komşuluk Yöntemi İle Hava Kirliliği Tahmini. *Mühendislik Fakültesi Dergisi, 25*(3), 1547-1556.

Amine, A., Elberrichi, Z., Simonet, M., & Malki, M. (2008). Evaluation and Comparison of Concept Based and n-grams Based Text Clustering Using SOM. *INFOCOMP Journal of Computer Science*, *7*(1), 27-35.

Arpacı, S.A., & Kalıpsız, O. (2018). Yazılım Hata Sınıflandırmasında Farklı Naive Bayes Tekniklerin Kıyaslanması. *Mühendislik Bilimleri Dergisi, 7*(1), 1-13. Doi:10.28948/ngumuh.383709

Artsın, M. (2020). Bir Metin Madenciliği Uygulaması: Vosvıewer. *Eskişehir Teknik Üniversitesi Bilim ve Teknoloji Dergisi B-Teorik Bilimler, 8*(2), 344-354.

Asilkan, Ö. (2008). *Veri Madenciliği Kullanılarak İkinci El Otomobil azarında Fiyat Tahmini*[Yayınlanmış Doktora Tezi]. Akdeniz Üniversitesi.

Aslan, A. (2019). DergiPark Akademik. *Acta Medica Alanya*, *3*(3), 205-206. Doi: 10.30565/medalanya.632462

Atabay, E., Çizel, B., & Ajanovic, E. (2019). Akıllı Şehir Araştırmalarının R Programı ile Bibliometrik Analizi.*20. Ulusal Turizm Kongresi*, 1130-1137.

Atan, S. (2020). Metin Madenciliği: İmkanlar, Yöntemler ve Kısıtlar. Mehmet Akif Ersoy Üniversitesi *Sosyal Bilimler Enstitüsü Dergisi*, (31), 220-239. Doi: 10.20875/makusobed.476524.

Atan, S. (2020). Metin Madenciliği: İmkanlar, Yöntemler ve Kısıtlar. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 31, 220-239. Doi: 10.20875/makusobed.476524.

Ayık, Y.Z., Özdemir, A., & Yavuz, U. (2010). Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10*(2), 441-454.

Balıkesir Üniversitesi. (2023, Haziran). Veri Madenciliği. http://kergun.baun.edu.tr/veri_madenciligi_hafta11.pdf

Balkan, K., & Takçı, H. (2010). Kavram Çıkarma Çalışmalarında Terim Benzerliklerinin Bulunması. *Electrical, Electronics and Computer Engineering National Conference*, 1-5.

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

Bırtıl, F.S. (2011). *Kız Meslek Lisesi Öğrencilerinin Akademik Başarısızlık Nedenlerinin Veri Madenciliği Tekniği ile Analizi*[Yayınlanmış Yüksek Lisans Tezi]. Afyon Kocatepe Üniversitesi.

Bilekdemir, G. (2010). *Veri Madenciliği Tekniklerini Kullanarak Üretim Süresi Tahmini ve Bir Uygulama*[Yayınlanmış Yüksek Lisans Tezi]. Dokuz Eylül Üniversitesi.

Bilgilioğlu, S.S. (2018). *Makine Öğrenmesi Teknikleri ile Mekansal Karar Destek Sistemlerinin Geliştirilmesi: Aksaray İli Örneği*[Yayınlanmış Doktora Tezi] Aksaray Üniversitesi.

Bilgin, T.T. (2003). *Veri Madenciliğinde Kümeleme Analizi Yöntemi Uygulaması*[Yayınlanmış Yüksek Lisans Tezi]. Marmara Üniversitesi.

Braun, W.J., & Murdoch, D.J. (2021). *A First Course In Statistical Programming With R. University Printing House,* Cambridege.

Buldu, B. (2019). Türkiye'deki Otomotiv Firmalarına Yönelik Şikayetlerin Metin Madenciliği Yöntemiyle İncelenmesi. *VI Yıldız Uluslararası Sosyal Bilimler Kongresi*, 1002-1012

Chien, C.F., & Chen, L.F. (2008). Data Miningto Improve Personnel Selectionand Enhance Human Capital: A Case Study in High-TechnologyIndustry. *Expert Systemswith Applications,* 34, 280-290.

Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014). Veri Madenciliğinde Karar Ağacı Algoritmaları İle Bilgisayar Ve İnternet Güvenliği Üzerine Bir Uygulama. *Endüstri Mühendisliği Dergisi*, *25*(3/4), 2-19.

Çalış, K., Gazdağı, O., & Yıldız, O. (2013). Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti. *Bilişim Teknolojileri Dergisi, 6*(1), 1-7.

Çetinyokuş, T., & Gökçen, H. (2002). Borsada Göstergelerle Teknik Analiz için Bir Karar Destek Sistemi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 17*(1), 43-58.

Dang, S., & Ahmad, P.H. (2014). TextMining: Techniquesandits Application. *IJETI International Journal of Engineering and Technology Innovations, 1*(4), 22-25.

Darakçı, H.Ç. (2011). *Kümeleme Analizi Kullanılarak Benzin İstasyonlarının Operasyonel Değerlendirilmesi*[Yayınlanmış Yüksek Lisans Tezi]. Maltepe Üniversitesi.

Daş, R., Türkoğlu, İ., & Poyraz, M. (2006). Genetik Algoritma Yöntemiyle İnternet Erişim Kayıtlarından Bilgi Çıkarılması. *Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 10*(2), 67-72.

Dener, M., Dörterler, M., Orman, A. (2009). Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama. *XI. Akademik Bilişim Konferansı Bildirileri*, 787-796.

Dergipark. (2023, Haziran). Amaç. https://dergipark.org.tr/tr/pub/page/about

Dilki, G., & Başar, Ö.D. (2020). İşletmelerin İflas Tahmininde K-en Yakın Komşu Algoritması Üzerinden Uzaklık Ölçütlerinin Karşılaştırılması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 19*(38), 224-233.

Dolgun, M.Ö., Özdemir, T.G., &Oğuz, D. (2009). Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği. *İstatistikçiler Dergisi,* 2, 48-58.

Emel, G.G., & Taşkın, Ç. (2005). Veri Madenciliğinde Karar Ağaçları Ve Bir Satış Analizi Uygulaması. *Eskişehir Osmangazi Üniversitesi, Sosyal Bilimler Dergisi, 6*(2), 224-239.

Erdoğan, Ş.Z., & Timor, M. (2005). A Data Mining Application in a Student Database. *Journal of Aeronauticsand Space Technologies, 2*(2), 53-57.

Ergezer, H., Dikmen, M., & Özdemir, E. (2003). Yapay Sinir Ağları ve Tanıma Sistemleri. *Pivolka, 2*(6), 14-17.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *J. Stat. Soft., 25*(5), 1–54. DOI:10.18637/jss.v025.i05

Gaikwad, S.V., Chaugule, A., & Patil, P. (2014). Text Mining Methodsand Techniques. International *Journal of Computer Applications, 85*(17), 42-45.

Göker, H., & Tekedere, H. (2017). Fatih Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi. *Bilişim Teknolojileri Dergisi, 10*(3), 291-299. Doi: 10.17671/gazibtd.331041

Gültekin, S.U. (2017). Veri Madenciliği: Yapay Sinir Ağı ve Doğrusal Regresyon Yöntemleri ile Fiyat Tahmini[Yayınlanmış Yüksek Lisans Tezi]. Pamukkale Üniversitesi.

Güracar, B. (2019). Genetik Algoritmalar. http://kergun.baun.edu.tr/20172018Guz/YZ_Sunumlar/Genetik_Algoritmalar_Busra_Guracar.pdf

Gürgen, G. (2008). *Birliktelik Kuralları ile Sepet Analizi ve Uygulaması*[Yayınlanmış Yüksek Lisans Tezi]. Marmara Üniversitesi.

Hamde, M.A. (2018). *Kurumsal Belgelere (Metin Verilerine) Metin Madenciliği Tekniği ile Erişimin Değerlendirilmesi: Türk Özel Sektörüne Yönelik Bir İnceleme*[Yayınlanmış Doktora Tezi]. İstanbul Üniversitesi.

Hearst, M. (2003). What Is Text Mining?, SIMS, UC Berkeley.

Hirschberg, J., & Manning, C. D. (2015). Advances in Natural Language Processing. *Science, 349*(6245), 261-266.

Irmak, S., Köksal, C.D., & Asilkan, Ö. (2012). Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleri ile Tahmin Edilmesi. *Uluslararası Alanya İşletme Fakültesi Dergisi, 4*(1), 101-114.

İçöz, E. (2021). *Covid-19 Pandemi Sürecinde Milli Eğitim Bakanı'nın Twitter Mesajlarının Metin Madenciliği Yöntemiyle İncelenmesi*[Yayınlanmış Yüksek Lisans Tezi]. Akdeniz Üniversitesi.

Kahya, A.N. (2021). Wikipedia'daki Verilere Metin Madenciliği Yöntemlerinin Uygulanması. *ESTUDAM Bilişim Dergisi,* 2(1) , 1-14.

Kale, B. (2018). Veri Madenciliği Sınıflandırma Algoritmaları ile e-posta Önemliliğinin Belirlenmesi[Yayınlanmış Yüksek Lisans Tezi]. Çukurova Üniversitesi.

Kaynar, O., & Taştan, S. (2009). Zaman Serisi Analizinde Mlp Yapay Sinir Ağları Ve Arıma Modelinin Karşılaştırılması. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* (33), 161-172.

Kemalbay, G., & Alkış, B.N. (2020). Borsa Endeks Hareket Yönünün Çoklu Lojistik Regresyon ve K-en Yakın Komşu Algoritması İle Tahmini. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 27*(4), 556-569. Doi:10.5505/pajes.2020.57383.

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M., & Özçift, A. (2016). KNN Algoritması ve R Dili İle Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi. *Marmara Fen Bilimleri Dergisi, 28(*3), 89-94. 2016. Doi: 10.7240/mufbed.69674.

Kılınç, D., Bozyiğit, F., Özçift, A., Yücelar, F., & Borandağ, E. (2015). Metin Madenciliği Kullanarak Yazılım Kullanımına Dair Bulguların Elde Edilmesi. *Ulusal Yazılım Mühendisliği Sempozyumu (UYMS),* 257-265.

Kutlu, B., & Badur, B. (2009). Yapay Sinir Ağları İle Borsa Endeksi Tahmini. *Yönetim Dergisi, 20*(63), 25-40.

Lee, W., & Stolfo, J. (1998). Data Mining Approaches ForIntrusion Detection. 7 Th Usenıx Security Symposium San Antonio Teksas Abd.

Mathew, T. V. (2012). Genetic Algorithm. *Report submitted at IIT Bombay*, 1-15.

Mutlu, M.E. (1989). *Karar Destek Sistemleri ve Öğrenci İşlerinde Bir Uygulama*[Yayınlanmış Yüksek Lisans Tezi]. Anadolu Üniversitesi.

Özek, M.B., Daş, B., & Akpolat, Z.H. (2010). Zaman Serisi Tahmininde TYP-2 Bulanık Mantık Tabanlı Veri Madenciliği Uygulaması. *Technological Applied Sciences, 5*(2), 100-109.

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

Özekes, S., & Çamurcu, Y.(2002). Veri Madenciliğinde Sınıflama ve Kestirim Uygulaması. *Marmara Üniversitesi Fen Bilimleri Dergisi*, 18, 1-17.

Özkan, Y. & Erol, Ç. (2015). *Biyoenformatik DNA Mikrodizi Veri Madenciliği*. Papatya Yayıncılık.

Özmen, S., & Şen, B.(2013). Veri Madenciliğinde Regresyon Yöntemleri İle Doğalgaz Sektöründe Talep–Tüketim Analizi. *XV. Akademik Bilişim Konferansı Bildirileri*, 835-839.

Özsever, Ç., Gençoğlu, T., & Erginel, N. (2009). İşgücü Verimlilik Takibi İçin Sistem Tasarımı Ve Karar Destek Modelinin Geliştirilmesi. *Dumlupınar Üniversitesi Fen Bilimleri Dergisi*, 18, 45-58.

Öztürk, G., & Tanrısevdi, A.(2017). Uluslararası Kruvaziyer Ziyaretçilerine Ait Özelliklerin Birliktelik Kuralı Modeli İle Analizi. *Uluslararası İktisadi ve İdari Bilimler Dergisi, 3*(1), 131-148.

Öztürk, K., & Şahin, M.E. (2018). Yapay Sinir Ağları ve Yapay Zekâ'ya Genel Bir Bakış. *Takvim-i Vekayi, 6*(2), 25-36.

Popowich, F. (2005). Using Text Mining and Natural Language Processing for Health Care Claims Processing. *ACM SIGKDD Explorations Newsletter*, *7*(1), 59-66.

Rokach, L., & Maimon, O. (2005). *Clusterıng Methods. Data Mınıng And Knowledge Dıscovery Handbook*, 321-352. DOI: 10.1007/0-387-25465-X_15

Sarıkaya, U., Doğan, B., &Aktaş, A. (2017). R ile Sosyal Ağ Madenciliği. *Marmara Fen Bilimleri Dergisi*, 3, 94-101. Doi: 10.7240/marufbd.330585

Seker, S.E. (2015). Zaman Serisi Analizi (Time Series Analysis).*Yönetim Bilişim Sistemleri Ansiklopedisi,* 2(4), 23-31.

Seker, S.E. (2015). Metin Madenciliği (TextMining). *YBS Ansiklopedi*, 2(3), 30-32.

Seyrek, İ.H., & Ata, H.A. (2010). Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik Ölçümü. BDDK *Bankacılık ve Finansal Piyasalar Dergisi, 4*(2), 67-84.

Shah, M., & Nair, S. (2015). A Survey of Data Mining Clustering Algorithms. *International Journal of Computer Applications, 128*(1), 1-5.

Sumathy, K.L. & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues – An Overview. *International Journal of Computer Applications, 80*(4), 29-32.

Sütcü, K., Tekerek, B., & Özler G. (2022). Aşı Karşıtı Twitter Paylaşımlarının Metin Madenciliği ve İçerik Analizi Yöntemiyle İncelenmesi. *Hacettepe Sağlık İdaresi Dergisi, 25*(4): 827-838

Sütçü, C.S. (1995). *İstatistiksel Veri Sistemleri ve Basın Sektöründe Bir Karar Destek Sistemi Uygulaması*[Yayınlanmış Doktora Tezi]. Marmara Üniversitesi.

Şahin, İ., & Chouseınoglou, O. (2019).Metin Madenciliği, Makine Ve Derin Öğrenme Algoritmaları İle Wep Sayfalarının Sınıflandırılması. *Yönetim Bilişim Sistemleri Dergisi, 5*(2), 29-43.

Şaylan, Ç.A. (2013). *Böbrek Nakli Geçirmiş Hastalarda Akıllı Yöntem Tabanlı Yeni Öznitelik Seçme Algoritması Geliştirilmesi*[Yayınlanmış Yüksek Lisans Tezi]. Kadir Has Üniversitesi.

Tahiroğlu, B.T. (2021). 21. Metin Madenciliği Açısından Dede Korkut Kitabı Söz Varlığının Bazı Özellikleri. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, 23, 319-338. Doi: 10.29000/rumelide.948370.

Takçı, H. (2020). *Teori ve Uygulamada Veri Madenciliği*. Nobel Yayıncılık.

Tamer, H.Y., Övgün, B., & Yalçıntaş, A. (2020). Akademik Büyük Veri ve Bilimsel Bilgi Üretimi: Dergipark Örneği. *Ankara Üniversitesi Sosyal Bilimler Dergisi, 11*(1), 93-110. Doi:10.33537/sobild.2020.11.1.10

Tan, Ah-h. (2000). Text Mining: The State of The Art and The Challenges. Kent Ridge Digital Labs.

Taşcı, E., & Onan, A. (2016). K-en Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi. *Akademik Bilişim, 1*(1), 4-18.

Taşçı, M.E., & Şamlı, R. (2020). Veri Madenciliği İle Kalp Hastalığı Teşhisi. *Avrupa Bilim ve Teknoloji Dergisi*(Özel Sayı), 88-95. Doi: 10.31590/ejosat.araconf12.

Toprak, K. (2018). Metin Madenciliği Yöntemleri Kullanarak İllere Göre Haber Analizi[Yayınlanmış Yüksek Lisans Tezi]. Karadeniz Teknik Üniversitesi.

Tunalı, V. (2011). *Metin Madenciliği İçin İyi İyileştirilmiş Bir Kümeleme Yapısının Tasarımı ve Uygulanması*[Yayınlanmış Doktora Tezi]. Marmara Üniversitesi.

Tüzüntürk, S. (2010). Veri madenciliği ve istatistik. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 29*(1), 65-90.

Ünlü, H.K. (2022). Başlığında "Data Scıence" İfadesi Geçen Uluslararası Kongrelerde Sunulan Bildiri Özetlerinin Metin Madenciliği Yöntemleri İle İncelenmesi. *Nicel Bilimler Dergisi, 4*(1), 1-21. Doi:0.51541/nicel.1075225.

Vural, M. (2005). *Genetik Algoritma Yöntemi ile Toplu Üretim Planlama*[Yayınlanmış Yüksek Lisans Tezi]. İstanbul Teknik Üniversitesi.

Yeşilbudak, M., Kahraman, H., & Karacan, H. (2011). Veri Madenciliğinde Nesne Yönelimli Birleştirici Hiyerarşik Kümeleme Modeli. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 26*(1), 27-39.

Yıldız, B., & Ağdeniz, Ş. (2018). Muhasebede Analiz Yöntemi Olarak Metin Madenciliği. *Muhasebe Bilim Dünyası Dergisi, 20*(2), 286-315. Doi: 10.31460/mbdd.9746.

Yıldız, O., Dağdeviren, M., & Çetinyokuş, T. (2008). İşgören Performansının Değerlendirilmesi için Bir Karar Destek Sistemi Ve Uygulaması. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 23*(1), 239-248.

Zontul, M., & Aydın, G. (2017). NoSQL Veri Tabanları Üzerinde Bir Metin Madenciliği Uygulaması. *Altınbaş Üniversitesi Mühendislik Sistemleri Ve Mimarlık Dergisi, 1*(1), 103-113.