TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
BİLGİ VE BELGE YÖNETİMİ
ANABİLİM DALI

BİLGİ YÖNETİMİNDE VERİ VE METİN MADENCİLİĞİ: BİR DİJİTAL İÇERİK ANALİZİ UYGULAMASI

Doktora Tezi

Levent KURT

Ankara-2023

TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
BİLGİ VE BELGE YÖNETİMİ
ANABİLİM DALI

BİLGİ YÖNETİMİNDE VERİ VE METİN MADENCİLİĞİ: BİR DİJİTAL İÇERİK ANALİZİ UYGULAMASI

Doktora Tezi

Levent KURT

Tez Danışmanı
Prof. Dr. Oya GÜRDAL

Ankara-2023

TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
BİLGİ VE BELGE YÖNETİMİ
ANABİLİM DALI

Levent KURT

BİLGİ YÖNETİMİNDE VERİ VE METİN MADENCİLİĞİ: BİR DİJİTAL İÇERİK ANALİZİ UYGULAMASI

Doktora Tezi

Tez Danışmanı: Prof. Dr. Oya GÜRDAL

Tez Jürisi Üyeleri

Adı ve Soyadı	İmzası
Prof. Dr. Oya GÜRDAL (Danışman)	.....
Prof. Dr. Sacit ARSLANTEKİN	.....
Prof. Dr. İnci BATMAZ	.....
Prof. Dr. Ümit KONYA	.....
Prof. Dr. Tülay OĞUZ	.....

Tez Sınavı Tarihi: 09/01/2023

TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Prof. Dr. Oya GÜRDAL danışmanlığında hazırladığım "Bilgi Yönetiminde Veri ve Metin Madenciliği: Bir Dijital İçerik Analizi Uygulaması" adlı doktora tezimdeki bütün bilgilerin akademik kurallara ve etik davranış ilkelerine uygun olarak toplanıp sunulduğunu, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim (31/01/2023).

Levent KURT

Öyküş'e...

ÖNSÖZ

Verilerin işlenerek değerli bilgilerin ortaya çıkarılma süreçlerini konu edinen veri madenciliği kavramı, özellikle son yıllarda önemli bir uygulama alanı haline gelmiştir. Bilgisayar sistemlerinin giderek daha hızlı ve güçlü hale gelmesiyle birlikte, veri madenciliği süreçleri daha hızlı ve daha kapsamlı bir şekilde yapılabilmektedir. İçinde bulunduğumuz bu dönemde artık yüksek işlemci gücüne sahip donanımlar ve ileri düzeyde geliştirilen yazılımlar sayesinde, sensörler aracılığıyla elde edilen veriler anlık olarak işlenerek yüksek teknolojiye sahip otonom, yani kendi kendine sürüş yapabilen otomobiller geliştirebilmek mümkün hale gelmiştir. Tesla şirketi bu konuda öncü firmalardan biri olarak gösterilebilir. Apple şirketinin geliştirdiği Siri, Amazon'un geliştirdiği Alexa gibi yapay zeka asistanlarıyla iletişim kurmak, Midjourney gibi uygulamalarla kısa metin açıklamaları girerek özgün görseller tasarlatmak, Open AI'ın piyasaya sürdüğü ChatGPT uygulamasına soru sorarak neredeyse mükemmele yakın bir dil becerisiyle gerçek zamanlı bilgi edinmek, Boston Dynamics şirketinin ürettiği yürüyen, koşan, zıplayan robotlar görmek sıradanlaşmaya başlamıştır. İçinde bulunduğumuz bu dönemde bir yapay zeka patlaması yaşandığını ifade etmek yanlış olmayacaktır. Bu gelişim hızı düşünüldüğünde muhtemelen yakın gelecekte çok daha fazlasını yapabilecek yapay zekaya sahip uygulama ve araçlar göreceğiz. Heyecan verici olan bu teknolojilerin neredeyse hepsinin temelinde, veri işleme süreçlerinin olduğunu söylemek yanlış olmayacaktır. Yapay zeka uygulamaları, otonom araçlar ve robotlar, hepsi arka planda sürekli olarak verileri analiz ederek işlevlerini yerine getirmekteler. Eğer bilim ve teknolojideki gelişmelerle veri hızlı bir şekilde işlenemeseydi, bu teknolojilerin ortaya çıkması mümkün olmayacaktı.

Yukarıda bahsedilen heyecan verici teknolojileri sürekli takip etmeyi seven biri olarak bu tez konusunu seçmek benim için kaçınılmazdı. Bu çalışmayla, Bilgi ve Belge

Yönetimi alanında sınırlı sayıda bulunan veri madenciliği çalışmalarına katkıda bulunarak, kendi gelişim yolculuğumu devam ettirme fırsatı buldum. Başlangıçta sevgili eşim teşvik etmeseydi belki de doktora eğitimine hiç başlamayacaktım. Altı yılı aşan doktora sürecinde istisnasız verdiği destek çok önemliydi, kendisine çok teşekkür ediyorum. Benim için yeni bir alanda çalışmak başta ne kadar heyecanlı görünse de, öğrenmek uzun ve yorucuydu. Bu yolculukta bazen pes ettiğim anlarda bile, bana başından itibaren güvenip desteğini hiç esirgemeyen değerli hocam Prof. Dr. Oya Gürdal'a ne kadar teşekkür etsem azdır, Oya Hocam olmasa büyük olasılıkla bu tez de bitmeyecekti. Tezin istatistik altyapısının oluşturulmasında doğrudan katkısı olan, istatistiki yöntem ve teknikleri öğrenmemde bana kılavuzluk eden, konuyla ilgili attığım her adımda yanımda olan ODTÜ İstatistik Bölümü öğretim üyesi Prof. Dr. İnci Batmaz'a bana ayırdığı değerli zamanı ve verdiği emekler için sonsuz teşekkür etmek isterim. Tez izleme komitesinin değerli üyesi Prof. Dr. Sacit Arslantekin Hocama da, en başından itibaren verdiği destek ve geri bildirimler için teşekkürlerimi sunuyorum.

İÇİNDEKİLER

ÖNSÖZ ..... **İ**
TABLO LİSTESİ ..... **V**
ŞEKİL LİSTESİ ..... **Vİ**
1. BÖLÜM: GİRİŞ **1**
1.1. Araştırmanın Konusu ve Önemi 3
1.2. Araştırmanın Amacı 11
1.3. Araştırmanın Kapsamı 13
1.4. Araştırmanın Yöntemi 14
1.5. Araştırmanın Düzeni 17
1.6. Kaynaklar 17
2. BÖLÜM: KAVRAMSAL ÇERÇEVE **19**
2.1. Veri 19
2.2. Enformasyon 20
2.3. Bilgi 21
2.4. Bilgi Yönetimi 21
2.5. Büyük Veri 25
2.6. Metin Madenciliği 28
2.6.1. Parçalara Ayırma 30
2.6.2. Kök Çözümleme 30
2.6.3. Kelime Etiketleme 31
2.6.4. Etkisiz Kelimeler 31
2.6.5. Terim Ağırlıklandırma 32
2.6.6. Kelime Dizileri 32
2.7. Veri Madenciliği 33
2.7.1. Veri Madenciliğinin Amaçları 37
2.7.2. Veri Madenciliğinin Önemi 39
2.7.3. Veri Madenciliğinde Başarı 41
2.7.4. Veri Madenciliğinde Zorluklar 42
2.7.5. Veri Madenciliği Süreci 43
2.7.5.1. CRISP-DM 43
2.7.5.1.1. İşi Anlamak 44
2.7.5.1.2. Veriyi Anlamak 44
2.7.5.1.3. Veri Hazırlama 45
2.7.5.1.4. Modelleme 45
2.7.5.1.5. Değerlendirme 45
2.7.5.1.6. Uygulama 46
2.7.5.2. SEMMA 46
2.7.5.2.1. Örnekleme 46
2.7.5.2.2. Keşfetme 46
2.7.5.2.3. Dönüştürme 47
2.7.5.2.4. Modelleme 47
2.7.5.2.5. Değerlendirme 47
3. BÖLÜM: VERİ MADENCİLİĞİ YÖNTEM VE TEKNİKLERİ **48**
3.1. Sınıflama 50
3.1.1. Karar Ağaçları 52
3.1.2. K-En Yakın Komşu 56
3.1.3. Yapay Sinir Ağları 59
3.1.4. Destek Vektör Makineleri 63
3.1.5. Bayesyen Sınıflama 64
3.2. Kümeleme 68
3.2.1. K-Ortalamalar Algoritması 69

3.2.2. K-Medoids Algoritması 71
3.2.3. En Yakın Komşu Algoritması 74
3.2.4. En Uzak Komşu Algoritması 75
3.3. Birliktelik Kuralları 75
3.3.1. Apriori Algoritması 76
3.4. Regresyon 77
3.4.1. Doğrusal Regresyon 78
3.4.2. Doğrusal Olmayan Regresyon 81
3.4.3. Belirtme Katsayısı 83
3.5. Makine Öğrenimi 83
3.5.1. Denetimli Öğrenme 85
3.5.2. Denetimsiz Öğrenme 87
4. BÖLÜM: DİJİTAL İÇERİK PLATFORMU İÇİN VERİ MADENCİLİĞİ UYGULAMASI 89
4.1. Okunma Sayılarını Etkileyen Faktörler 89
4.1.1. Amaç 89
4.1.2. Yöntem 90
4.1.2.1. Veri Hazırlama 91
4.1.2.2. Metin Ön İşleme 94
4.1.2.3. Değişken Yaratma 96
4.1.2.4. Analiz 98
4.1.3. Bulgular 105
4.2. Okunma Sayısının Gelire Etkisi 106
4.2.1. Amaç 106
4.2.2. Yöntem 107
4.2.2.1. Veri Hazırlama 107
4.2.2.2. Analiz 108
4.2.3. Bulgular 118
4.2.4. Farklı Yöntemler 119
4.3. Bulguların Genel Analizi 122
4.4. Başlık Bazlı Dijital İçerik Analiz Süreci 123
5. BÖLÜM: SONUÇ VE ÖNERİLER 127
KAYNAKÇA 138
ÖZET 153
SUMMARY 154

TABLO LİSTESİ

Tablo 1. Sınıflama Yöntemi Örneği 52
Tablo 2. K-En Yakın Komşu Algoritması Veri Tablosu 58
Tablo 3. K-En Yakın Komşu Algoritması Sıralama Tablosu 58
Tablo 4. Bayesyen Sınıflama Örnek Veri Seti 66
Tablo 5. K-Medoids Algoritması Örneği 72
Tablo 6. K-Medoids Kümeleme-1 72
Tablo 7. K-Medoids Kümeleme-2 72
Tablo 8. K-Medoids Kümeleme-3 73
Tablo 9. K-Medoids Kümeleme-4 73
Tablo 10. Çoklu Doğrusal Regresyon Örnek Verileri 79
Tablo 11. Ham Veri Kümesi 92
Tablo 12. Google Analitik'ten Çekilen Veri 92
Tablo 13. Birleştirilmiş Ham Veri Kümesi 93
Tablo 14. Metin Ön İşleme Sonucu Elde Edilen Veri Kümesi 95
Tablo 15. Frekansa Göre İlk 5 Kelime ve Kelime Grubu 96
Tablo 16. Analiz İçin Oluşturulan Veri Kümesi 97
Tablo 17. Model Katsayıları Tablosu 105
Tablo 18. Google Analitik İlk 3 Gözlem 108
Tablo 19. Veri Kümesinin Son Hali 108
Tablo 20. Cochrane-Orcutt Yöntemi 120
Tablo 21. Fiyat İfadesi Geçen Değişkenler 129
Tablo 22. Aynı Köke Sahip Değişkenler 130

ŞEKİL LİSTESİ

Şekil 1. Bilgi Hiyerarşisi 22
Şekil 2. Parçalara Ayrıma (Tokenization) İşlemi 30
Şekil 3. CRISP-DM Konseptinin Aşamaları 43
Şekil 4. Karar Ağacı Örneği 53
Şekil 5. K-En Yakın Komşu Algoritması Örneği 57
Şekil 6. Bir Yapay Sinir Hücresinin Çalışma Örneği 61
Şekil 7. Bir Yapay Sinir Ağı Örneği 61
Şekil 8. Doğrusal Olarak Ayrılabilen Veri 63
Şekil 9. Doğrusal Olarak Ayrılamayan Veri 63
Şekil 10. Doğrusal Olmayan Sınıflama 64
Şekil 11. Kümeleme Örneği 69
Şekil 12. K-Ortalama Algoritması 71
Şekil 13. En Yakın Komşu Algoritması 74
Şekil 14. En Uzak Komşu Algoritması 75
Şekil 15. Denetimli Öğrenme 86
Şekil 16. Denetimsiz Öğrenme 87
Şekil 17. Standartlaştırılmış Artık Grafikleri-1 99
Şekil 18. Box-Cox Dönüşümü 100
Şekil 19. Standartlaştırılmış Artık Grafikleri-2 101
Şekil 20. Normal Olasılık Test Grafiği-1 102
Şekil 21. Standartlaştırılmış Artık Grafikleri-3 109
Şekil 22. Normal Olasılık Test Grafiği-2 110
Şekil 23. Normal Olasılık Test Grafiği-3 112
Şekil 24. Standartlaştırılmış Artık Grafikleri-4 113
Şekil 25. Normal Olasılık Test Grafiği-4 113
Şekil 26. Standartlaştırılmış Artık Grafikleri-5 116
Şekil 27. Normal Olasılık Test Grafiği-5 116
Şekil 28. Önerilen Modele Ait Regresyon Doğrusu 119
Şekil 29. Başlık Bazlı Dijital İçerik Analiz Süreci 124
Şekil 30. Model Çıktısı 126

1. BÖLÜM: GİRİŞ

İnsanoğlu dünyayı kontrol etmek için itici bir tutkuya sahiptir, öyle ki diğer her türlü canlının yeteneklerinin ötesinde farklı yollarla dünyayı anlamakta ve kontrol etmektedir. İnsanlar sadece istenen çıktıları elde etmek için olayların akışının nasıl değiştirileceğinin değil, aynı zamanda tüm tasarımı manipüle edip yeniden şekillendirir. İnsanlar doğal akışta olayların ve süreçlerin değişiminde basit ve karmaşık örüntüleri tanıma konusunda üstün yeteneklere sahiptirler. Sayılar ve harflerle sembolize edilen örüntüler; sosyal hayatta, hayvan davranışlarında, fizikte, zihinsel ve duygusal yaşamda ve aslında evren içerisinde insanların algılayabildiği her yerde ortaya çıkmaktadır. Bu sembollerin birbirleriyle çeşitli şekillerde ilişkilendirilerek anlamlı sonuçlar keşfedilmesiyle ele alınan problemlerin çözümü sonucu, dünya daha anlaşılabilir hale gelmektedir (Pyle, 2003, s. 3).

Eski çağlardan günümüze bilim insanları ve düşünürlerin, zihinlerindeki araştırma problemi ve bunun çözümü için, önceden üretilmiş bilgi kaynaklarını keşfetme ve bu kaynaklardan yararlanma sürecine gereksinim duyulmuştur. Bu durum bilginin ve bilimin birikimli (kümülatif) ilerlemesi ilkesinin ortaya çıkmasını sağlamıştır (Gürdal Tamdoğan, 2009, s. 153). Modern anlamıyla bilim, doğa araştırması anlamı içerir ve yukarıda bahsedilen ilkenin gücüyle de Francis Bacon'ın tümevarım yöntemi üzerine kurulmuştur. Bacon, bilginin sadece deney ve gözlemlerle üretilebileceğini ileri sürmekteydi. Tümevarım yöntemini fizik, kimya, biyoloji'de ve diğer tüm doğa ve uygulamalı bilimlerde uygulayan Brahe, Kopernik, Kepler, Galileo, Newton, Faraday ve Curie gibi bilim dehaları için daha çok gözlem daha çok bilim anlamına gelmekteydi. Bu bağlamda gözlem ve deneyler bilgi üretmenin yegane yolu olarak görülmüştür. Viktor Mayer-Schonberger ve Kenneth Cukier'a göre ise, büyük veriyle gelinen bugünkü noktada, bilgi

ve bilim üretiminde artık deney ve gözleme ek olarak veriye bakılmaktadır (Işıklı, 2014, s. 90).

Farklı bir açıdan bakıldığında bilim; fiziksel, biyolojik ve sosyal sistemleri tanımlamak için ilk prensip modellerini kullanmaya dayanmaktadır. Bu tür bir yaklaşım, Newton'un hareket yasaları veya Maxwell'in elektromanyetizmadaki denklemleri gibi temel bir bilimsel modelle, yani ilk prensip modelleriyle başlar ve daha sonra teknoloji ve mühendislik alanlarında çeşitli uygulamalar üzerine kurulur. Bu yaklaşımda deneysel veriler, temel ilkeleri doğrulamak ve doğrudan ölçmesi zor veya bazen imkansız olan parametreleri ortaya çıkartmak için kullanılır. Bununla birlikte birçok alanda temel ilkeler bilinmemekte veya incelenen sistemler matematiksel olarak biçimlendirilemeyecek kadar karmaşık olmaktadır. İlk prensip modellerinin olmadığı durumlarda bile, var olan veri, değişkenler arasında ilişkileri tahmin ederek işe yarar bilimsel sonuçlar çıkarmayı sağlamaktadır. Böylelikle günümüzde, ilk prensiplere dayalı klasik analiz ve modellerden, direkt olarak veriden uygun analizlerin yapılabildiği modellere doğru bir paradigma kaymasından söz edilebilir (Kantardzic, 2011, s. 3).

Dünya 1990'lı yıllardan bu yana çok büyük bir değişim yaşadı. Neredeyse tüm sektörler bu değişime ayak uydurarak yeni dijital bir dünya kuruldu. Bilgisayar-insan arasında yeni ilişkiler kuruldu ve neredeyse her gün bu ilişki daha da pekişiyor. Bu yaşananlara bilişim devrimi veya dijital dönüşüm gibi ne ad verilirse verilsin, artık içinde bulunduğumuz dünya eskiye göre çok farklı ve çok büyük etkilere sahip. Bilişim teknolojileri gelişmeye devam ettikçe; üretilen, depolanan ve analiz edilen veri miktarında da artış yaşanıyor. Bu gelişimin sonucunda "veri bilimi" alanı da gelişmeye devam ediyor. Bu alan başta istatistik ve bilgisayar bilimleri olmak üzere, mühendislikten fen bilimleri ve sosyal bilimlere kadar çok sayıda alanı ilgilendiriyor (Gürsakal, 2014, s. 2). Araştırma nesnesi bilgi olan kütüphanecilik ve enformasyon bilimi de söz konusu alanların başında geliyor kuşkusuz. Çağımızda veri olgusu, bilim ve teknolojinin

ilerlemesi bakımından değişiklikler geçirerek oldukça önemli hale gelmiştir. Öte yandan bu değişim, toplumların yaşam biçimini farklılaştırmış, iletişim ve bilişim teknolojilerinin gelişimiyle birlikte kendine has yeni bir yaşam düzeni oluşturmuştur.

İçinde bulunduğumuz veri çağında şirketler ve resmi kurumlar veri bilimine odaklanarak yeni fırsatlar yakalamaktadırlar. Yakın geleceğin en önemli teknoloji piyasasının (artan veri miktarının boyutu da düşünüldüğünde) veri analizi üzerinde oluşması ve gelecekte bu pazarın trilyon dolarları aşması beklenmektedir (Ünal, 2015, s. 34). Mevcut durum ve önceki paragraflarda bahsedilen düşünceler göz önüne alınarak değerlendirilecek olunursa; devletlerin ve şirketlerin bu kadar kapsamlı veri okyanusu içinde kaybolmaması, verinin şirketlere ve tüm organizasyonlara katma değer getirebilmesi ve veriler içerisinden hayatı kolaylaştıracak ve değer yaratacak anlamlı sonuçlar çıkarılabilmesi için veri madenciliği uygulamalarının arttırılması önem taşımaktadır.

1.1. Araştırmanın Konusu ve Önemi

Günümüzün rekabetçi ortamında tüm organizasyonlar için katma değerli ürünler ve hizmetler sunabilmek ve iş süreçlerini iyileştirebilmek hayati önem taşımaktadır. Bu çalışmanın konu edindiği dijital içerik üreten bir organizasyon için de bu durum değişmemektedir. Dijital içerik üretimi alanında okunma sayıları ve buna bağlı istatistikler, şirketin hayatına devam edebilmesi ve büyümesi için en önemli konular arasındadır. Eğer bir içerik platformu yeteri kadar okunmuyorsa, başka bir ifadeyle yeteri sayıda kullanıcıyı kendisine çekemiyorsa yaşamına daha fazla devam edemez. Bu tür organizasyonlar için ne kadar çok kullanıcı, o kadar çok büyüme anlamına gelmektedir. Öyle ki, veri analizleriyle ortaya çıkarılacak bilgilerin organizasyon içerisinde kullanımının, okunma istatistikleri üzerinde yapacağı çok küçük artışlar bile, toplam şirket değeri ve gelirleri üzerinde büyük etkiler yapabilme potansiyeline sahiptir. Bu

bağlamda bu çalışma, bilgi yönetimi kavramı kapsamında organizasyon içerisinde biriken ham verinin veri madenciliği uygulamalarıyla değerli bilgiye dönüşme sürecini konu edinmektedir. Çalışmada bu çerçevede, Bilgi ve Belge Yönetimi alanı içerisinde yer alan bilgi yönetimi (information management) kavramına odaklanılarak, tezin uygulama kısmında organizasyonlarda bilgi yönetiminin bir parçası olan veri madenciliği uygulamaları gerçekleştirilmiştir. Bu kapsamda öncelikle literatür taraması yapılmış, yurt içi ve yurt dışı bilgi yönetimi alanında yayınlanan veri madenciliği çalışmaları incelenmiştir.

İlk olarak Türkiye'de Bilgi ve Belge Yönetimi alanında yayın yapan önde gelen bilimsel dergiler incelenmiştir. Yayınlanan çalışmalar kronolojik sırayla verilmiştir. Takçı ve Soğukpınar (2002) tarafından Bilgi Dünyası dergisinde ele alınan "Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi" çalışmasında, kullanıcıların kütüphane web siteleri üzerindeki erişim örüntüleri ortaya çıkarılmaya çalışılmıştır. Araştırmada web madenciliği tekniklerinin kullanılmasıyla birtakım örüntüler elde edilmiş ve değerlendirmeler yapılmıştır. Sever ve Oğuz (2002) tarafından yine aynı dergide yayınlanan "Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorguları ve Algoritmalar" makalesinde, veritabanlarında bilgi keşfi kavramı çerçevesinde temel sorunlar üzerine değerlendirmeler yapılmış ve veri madenciliğinde kullanılan temel yöntem ve teknikler ele alınmıştır. Sever ve Oğuz (2003) tarafından araştırmanın ikinci kısmı "Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım: Kısım II- Eşleştirme Sorgularının Biçimsel Kavram Analizi ile Modellenmesi" adlı makale Bilgi Dünyası'nda yayınlanmıştır. Bu çalışmada ise, konunun detaylarına inilerek biçimsel kavram analizi ile eşleştirme kural çıkarımı arasında bir model önerilmiştir. Arslantekin (2003) tarafından Türk Kütüphaneciliği dergisinde "Veri Madenciliği ve Bilgi Merkezleri" adlı çalışma yayınlanmıştır. Makalede, bilgi merkezleri ve veri madenciliği kavramları arasındaki ilişki incelenmiş ve bilgi merkezlerindeki iş süreçlerinde potansiyel

olarak uygulanabilecek veri madenciliği uygulamaları hakkında değerlendirmeler yapılmıştır. Türk Kütüphaneciliği dergisinde Çalışkan ve Can (2018) tarafından yayınlanan bir diğer çalışma, "Türkçe Metinler Üzerine Yapılan Sayısal Üslup Araştırmalarını İnceleyen ve Benim Adım Kırmızı Çevirilerinin Aslına Olan Sadakatini Ölçen Bir Çalışma" adlı makaledir. Çalışmada, Orhan Pamuk'un Benim Adım Kırmızı adlı romanı ve çevirilerindeki üslup uyumu üzerine metin madenciliği teknikleri kullanılarak sayısal üslup analizi gerçekleştirilmiştir. Araştırma sonucunda istatistiksel olarak yazar üslubunun, çevirilerde devam ettirildiği belirlenmiştir. Bilgi Yönetimi dergisinde Aktan (2018) tarafından ele alınan "Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu" adlı çalışmada, büyük veri kavramı, uygulamaları ve süreçleri hakkında literatüre dayalı incelemeler sunulmuş ve değerlendirmeler yapılmıştır. Aynı dergide Yıldırım ve Özdemirci (2019) tarafından yayınlanan "Kurumlarda Örtük Bilginin Yapay Zekâ Destekli Tavsiye Sistemleri Aracılığıyla Ortaya Çıkarılması" adlı makalede, organizasyonlardaki örtük bilginin kullanıma sunulması amacıyla yapay zeka destekli sistemler üzerine bir derleme çalışması yapılmış ve organizasyonlarda tavsiye sistemlerinin kullanılabileceğine vurgu yapılmıştır. Cibaroğlu ve Yalçınkaya (2019) aynı dergide yayınlanan "Belge ve Arşiv Yönetimi Süreçlerinde Büyük Veri Analitiği ve Yapay Zeka Uygulamaları" adlı çalışmasıyla, yapay zeka ve büyük veri kavramlarını incelemiş ve bu teknolojilerin belge ve arşiv yönetimi süreçlerinde uygulanabilirliğine yönelik değerlendirmelerde bulunmuştur. Yıldırım (2019) tarafından Bilgi Dünyası'nda yayınlanan "Endüstri 4.0'a Kapsamlı Bir Bakış: 2011'den Bugüne" adlı çalışmada, endüstri 4.0 kavramı ele alınmış ve bu kavramla yakından ilgili olan, nesnelerin interneti, büyük veri, veri analitiği ve yapay zeka kavramlarına yönelik literatür çalışması ve değerlendirmeler yapılmıştır. Çakmak ve Eroğlu (2020) tarafından hazırlanan "Sosyal Medyada Kullanıcı Etkileşimi ve İçerik Kategorizasyonu: Ankara'daki Halk Kütüphanelerinin Facebook Gönderilerinin Analizi" adlı makale, Türk Kütüphaneciliği

dergisinde yayınlanmıştır. Araştırmada halk kütüphanelerinin Facebook gönderileri metin madenciliği teknikleriyle analiz edilmiş ve öneriler sunulmuştur. Bilgi Dünyası'nda Çakmak ve Tırnavalı (2020) tarafından ele alınan "Danışma Hizmetlerinin Değerlendirilmesi: Türkiye Büyük Millet Meclisi Kütüphanesi Danışma Hizmetlerine Yönelik Bir Araştırma" adlı makalede, TBMM Kütüphanesindeki kullanıcı talepleri toplanarak bir veri kümesi oluşturulmuş ve kullanıcılardan elde edilen talepler analiz edilerek bulgular yorumlanmış ve önerilere yer verilmiştir. Ankaralı ve Külcü (2020) tarafından ele alınan ve Bilgi Yönetimi dergisinde yayınlanan "RapidMiner ile Twitter Verilerinin Konu Modellemesi" adlı makalede, metin ve veri madenciliği teknikleri kullanılarak Twitter verileri üzerine bir analiz gerçekleştirilmiş ve konu modellemesi oluşturulmuştur. Yine aynı dergide Öztürk (2021) tarafından hazırlanan "Arşivler ve Yapay Zeka" adlı makalede, arşivcilik alanında yapay zekanın kullanımlarından örnekler verilmiş, yenilik ve gelişmeler değerlendirilmiştir. Öztürk ve Özel (2021) tarafından ele alınan, "Yapay Zeka ve Kütüphaneler" adlı çalışma Bilgi Dünyası dergisinde yayınlanmış ve kütüphanelerde yapay zeka kullanımı üzerine literatüre dayalı bir inceleme yapılmıştır. Konu hakkındaki farkındalık ve beklentilerin tespit edilmesi amacıyla kütüphanecilere anket çalışması yapılmış ve sonuçlar analiz edilerek yorumlanmıştır. Bilgi ve Belge Araştırmaları dergisinde Kaya (2021) tarafından ele alınan "Bir Araştırma Kaynağı Olarak Arşivlenen Sosyal Medya Verilerinin Kullanımı" adlı çalışmada, sosyal medya üzerinde oluşan büyük veri ele alınmış ve bununla ilgili olarak literatüre dayalı değerlendirmeler yapılmıştır. Efe (2022) "Yargısal ve Hukuki Süreçlerde Yapay Zekâ Kullanan Araçlar Üzerine Bir Araştırma" adlı Bilgi Yönetimi dergisinde yayınlanan makalede, hukuk otomasyonları ve yapay zeka tabanlı yönetim bilişim sistemleri araçları incelenmiş ve gelecekte bu alandaki beklentiler değerlendirilmiştir. Aynı dergide Eker (2022) tarafından hazırlanan "Sosyal Ağlar Büyük Veriden Nasıl Yararlanır: Facebook ve Twitter" adlı makalede, sosyal ağların sahip oldukları veriye nasıl ulaştıkları hakkında

incelemeler yapılmış ve bu ağların ne tür çalışmalarla kullanıcı verilerini oluşturduğu tespit edilmeye çalışılmıştır.

Bilgi ve Belge Yönetimi alanında yayın yapan bilimsel dergilere ek olarak, Türkiye'de Bilgi ve Belge Yönetimi bölümlerinde veri madenciliği başlığı altında değerlendirilebilecek beş tanesi doktora, iki tanesi yüksek lisans olmak üzere yedi adet lisansüstü tez çalışması tespit edilmiştir.

Doktora çalışmalarından ilki, Binici (2016) tarafından İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde hazırlanan "Kütüphane ve Bilgi Bilimi Çalışmalarında Dönemsel Konu Analizi" adlı tezdir. Çalışmada bilgi bilim ve kütüphane alanında yayınlanmış bilimsel yayınlar metin madenciliği teknikleri, ağ analizi ve kelime birliktelik yöntemleri kullanılarak analiz edilmiştir. Sonuç olarak bahsedilen alandaki çalışmalarda hangi temaların işlendiği ve dönemsel olarak hangi konulara rağbet gösterildiği ortaya konmaya çalışılmıştır. Doktora çalışmalarından diğeri, Taşkın (2017) tarafından Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde yapılan "İçerik Tabanlı Atıf Analizi Modeli Tasarımı: Türkçe Atıflar İçin Metin Kategorizasyonuna Dayalı Bir Uygulama" adlı tez çalışmasıdır. Araştırmada Türkçe atıflar için içerik tabanlı bir analiz modeli oluşturulmuştur. Çalışmada analiz için belirlenen makalelerin üst veri, kaynakça ve tam metin bilgileri doğal dil işleme süreçleriyle analiz edilmiş ve sınıflara ayrılmıştır. Sonrasında sınıflamanın başarısı test edilmiş ve bu başarı göz önüne alınarak içerik tabanlı atıf modeli önerilmiştir. Doktora çalışmalarından bir başkası, Hamde (2018) tarafından İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde hazırlanan "Kurumsal Belgelere (Metin Verilerine) Metin Madenciliği Tekniği ile Erişimin Değerlendirilmesi: Türk Özel Sektörüne Yönelik Bir İnceleme" adlı tezdir. Çalışmada, organizasyonlardaki belge yönetiminde metin madenciliği uygulamalarının genel kavramsal çerçevesi çizilmiş ve Türkiye'deki şirketlerde çalışan bir grup birey üzerinde, kurumsal belge yönetiminde

metin madenciliği uygulamaları hakkında anket çalışması yapılmıştır. Sonuç kısmında organizasyonların metin madenciliği ile performanslarının arttırılması ve kurumsal belgelere erişimin geliştirilmesi yönünde öneriler sunulmuştur. Bir diğer çalışma Çelik (2021) tarafından Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde yapılan "Türkçe Akademik Yayınlar İçin Yapısal Öz Çıkarım Sistemi" adlı tezdir. Araştırmada metin madenciliği teknikleri kullanılarak bilimsel makale özlerinin otomatik olarak yapısal yöntemlerle elde edilmesi amacıyla bir sistem geliştirilmiştir. Elde edilen sonuçlara göre çalışmada önerilen otomatik öz oluşturma sisteminin, klasik öz çıktılarına göre daha başarılı olduğu tespit edilmiştir. Geliştirilen otomatik yapısal öz hazırlama uygulamasının tüm Türkçe doküman grupları için uygulanabilir olduğu belirtilmektedir. Doktora çalışmalarından 2022 yılı verilerine göre sonuncusu ise Doğan (2022) tarafından Ankara Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde hazırlanan "Bilgi Merkezi ve Hizmetlerinde Veri Madenciliğinin Kullanılabilirliği: Üniversite Kütüphaneleri Örneği" adlı tez çalışmasıdır. Bu araştırmada üniversite kütüphanelerinde oluşan kullanıcı verilerinin veri madenciliği yöntemleriyle işlenip kütüphanelerin hizmet ve performansını arttırmada kullanılabilirliği üzerine analizler yapılmıştır. Araştırma sonucunda veri madenciliği teknikleri kullanılarak elde edilen sonuçların kütüphane hizmetleri için karar destek sistemlerinde kullanılabileceği ortaya çıkarılmıştır.

Yüksek lisans düzeyinde veri madenciliği hakkında Bilgi ve Belge Yönetimi Bölümü çerçevesinde yapılan çalışmaların ilki, Akçay (2014) tarafından İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde yapılan "Bilgi ve Belge Yönetiminde Veri Madenciliği" adlı tez çalışmasıdır. Çalışmada kavramsal olarak Bilgi ve Belge Yönetimi Bölümü ile veri madenciliği kavramı arasında ilişki kurulmuş ve konuyla ilgili kavramlar ve kullanım alanları hakkında teorik bilgiler verilmiştir. Yüksek lisans düzeyinde yapılan diğer çalışma ise, Doğan (2015) tarafından Ankara

Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi Bölümü'nde hazırlanan "Büyük Verinin Üniversite Kütüphaneleri Açısından Önemi ve Kütüphane Hizmetlerinde Kullanımı: Ankara Üniversitesi Kütüphaneleri Örneği" adlı tezdir. Çalışmada kütüphane hizmetlerinde büyük verinin önemi ve kullanımı üzerine anket çalışması yapılmıştır. Araştırmada kütüphanecilere ilk olarak bir anket yapılıp büyük veri kavramıyla ilgili farkındalık ölçümü yapılmış, sonrasında anket katılımcılarına konuyla ilgili bilgilendirici bir sunum gerçekleştirilmiştir. Son olarak aynı katılımcı grubuna sunum sonrası tekrar bir anket çalışması yapılarak büyük veri kavramı hakkında farkındalığın oluşup oluşmadığı gözlemlenmiştir. Sonuç olarak eğitim sunumunu izleyen katılımcıların bilgi düzeyinde anlamlı bir artış gözlemlendiği tespit edilmiştir.

Öte yandan yurt dışı literatür araştırması yapılmıştır. Bu çerçevede bilgi yönetimi kavramı kapsamında veri madenciliği alanında yapılan lisans üstü tezler incelenmiştir. ProQuest Dissertations & Theses[1] tez veritabanında veri madenciliği ve bilgi yönetimi anahtar kelimelerinin beraber bulunduğu, tüm zamanlarda tüm dillerde hazırlanmış 207 lisansüstü tez çalışması tespit edilmiştir. Bunlar içerisindeki bazı tezler incelenmiştir. Wageningen University'de, Nguyen (2019) tarafından hazırlanan "Value of Information in Agro-Food Logistics Management: A Case of the Dutch Floriculture Supply Chain Network" adlı çalışmada, tarım ürünlerinin lojistik yönetimine odaklanılmış ve lojistik süreçleri iyileştirmek için şirket içi bilgi kullanımı ele alınmıştır. University of Salford'da, Darwish (2019) tarafından hazırlanan "Towards Intelligent Diabetes Knowledge Management and Knowledge Discovery: A Data Mining Approach" adlı tez çalışmasında, sağlık alanında veri madenciliği uygulaması yapılmıştır. Verilerin işlenmesiyle elde edilen bilgiler çerçevesinde, Birleşik Krallık'ta diyabet hastalığıyla mücadelede bilgi yönetiminin; bilgiyi paylaşmak ve hastaların semptomları kendi

[1] Bu veritabanı, 1861 yılından itibaren tüm dünyada yazılan yüksek lisans ve doktora tezlerini kapsamaktadır ve içeriğinde 5 milyonun üzerinde kayıt bulunmaktadır.

kendilerine yönetmelerine yardımcı olmak için değerli bir yaklaşım olduğu gösterilmiştir. Universidade do Porto'da, Ferreira (2021) tarafından yazılan "Predictive Models for Shop Floor Optimization in the Agrofood Industry" adlı çalışma, gıda endüstrisinde yer alan bir şirket için veri madenciliğine dayalı bir optimizasyon projesini içermektedir. Çalışma sonucunda elde edilen model ile iş süreçlerinde geliştirme ve veriye dayalı karar vermeyi destekleyen organizasyondaki bilgi yönetimi süreçlerinde iyileşme sağlandığı belirtilmektedir. Yurt dışı literatürde tespit edilen bilgi yönetimi kapsamında yapılan veri madenciliği çalışmalarına genel olarak bakıldığında, ele alınan konuların tarım, lojistik, sağlık, inşaat, otomasyon, kimya, iletişim, finans, dil bilimleri ve biyoloji gibi çok sayıda farklı alanı kapsama aldığı görülmüştür.

Literatür tarama çalışmasını özetlemek gerekirse, Türkiye'de Bilgi ve Belge Yönetimi alanında öne çıkan önceki sayfalarda adı geçen dört dergide, veri madenciliği kavramı altında değerlendirilebilecek 19 çalışma tespit edilmiştir. Bunlardan altı tanesinin veri madenciliği analiz ve uygulamaları içerdiği, 13 tanesinin ise literatür incelemesi veya kavramsal düzeyde değerlendirme makalelerinden oluştuğu gözlemlenmiştir. YÖK Ulusal Tez Merkezi kullanılarak yapılan araştırmada Türkiye'de Bilgi ve Belge Yönetimi bölümlerinde önceki sayfalarda açıklanmaya çalışılan veri madenciliğiyle ilgili toplam yedi lisansüstü tez çalışmasına ulaşılmıştır. Bu çalışmalardan üçü veri madenciliğini kavramsal olarak ele alıp anket çalışmalarıyla desteklemiştir, diğerlerinin ise metin madenciliği teknik ve yöntemlerinin gerçekleştirildiği uygulama çalışmalarından olduğu gözlemlenmektedir. Veri madenciliği uygulaması içeren altı makaleden sadece birinin Bilgi ve Belge Yönetimi alanı dışında bir konuyu çalıştığı, tüm doktora ve yüksek lisans tez çalışmalarının ise Bilgi ve Belge Yönetimi alanı içerisinde yer alan konuları ele aldığı görülmektedir. Diğer taraftan, yurtdışında hemen hemen her alanda bilgi yönetimi kapsamında veri madenciliği çalışmalarının yapıldığı belirlenmiştir. Aslında tam olarak bu durum, bu tezin ortaya çıkış motivasyonlarından birini

oluşturmaktadır. Çünkü bilgi yönetimi olgusu çerçevesinde ele alınmaya çalışılan bu tez ile Türkiye'de Bilgi ve Belge Yönetimi kapsamında olan konuların ötesine gidilerek, veri madenciliğinin, organizasyonlarda bilgi yönetimi kavramı kapsamına giren; lojistikten iletişime, otomasyondan finansa ve web hizmetlerine kadar hemen hemen her konuda, Bilgi ve Belge Yönetimi Bölümü mezunları tarafından yapılabilirliğinin gösterilmesi ve tezin bu çerçevede uçtan uca kapsamlı bir veri madenciliği süreci ile örnek teşkil etmesi hedeflenmiştir. Tezde sadece sonuca ulaşmak değil, aynı zamanda araştırma sürecinin baştan sona kendisi de bu nedenle önemli görülmektedir. Bu sebepten dolayı tüm veri madenciliği süreçleri detaylı bir şekilde adım adım açıklanmaya çalışmıştır. Tezin, veri madenciliği hakkında sınırlı sayıda uygulama çalışmasının olduğu Bilgi ve Belge Yönetimi alanı literatürüne, önemli bir katkı sunacağı düşünülmektedir.

1.2. Araştırmanın Amacı

Tez kapsamında ele alınan veri madenciliği uygulaması, bilgi yönetimi bakış açısıyla ele alınmaya çalışılmıştır. Kavram olarak bilgi yönetimi, organizasyondaki iş süreçlerini daha etkili kılmak için katma değerli bilgi yaratmak üzere veri/enformasyon/bilgiyi elde etme, sistematik hale getirme ve paylaşma/yayma süreci olarak tanımlanabilir. Bilgi yönetimi kavramı tanımından da anlaşılacağı gibi, Bilgi ve Belge Yönetimi Bölümü mezunlarının veri madenciliği süreçlerine bilfiil dahil olması ve her alanda veri madenciliği uygulamalarının, bölüm mezunları tarafından da yapılabileceğinin gösterilmesi, tez çalışmasının önemli motivasyonlarından birini oluşturmaktadır.

Çalışmanın uçtan uca bir veri madenciliği uygulama örneği teşkil edebilmesi için iki farklı analiz yapılmıştır. Kullanılan uçtan uca kavramı, bir uygulama veya projenin en başından sonuna kadar tüm basamaklarının ele alınıp yapılan işlemlerin detaylarıyla açıklandığını vurgulamak için seçilmiştir. İlk analizdeki amaç; bu araştırmada ele alınan

probleme özel olarak farklı bir metin madenciliği yaklaşımı ile oluşturulan veri kümesi üzerinde analizler gerçekleştirilerek, içeriklerin başlık bazında okunma sayılarını etkileyen faktörleri (kelime ve kelime gruplarını) belirlemektir. Böylelikle içeriklerin okunma sayılarını iyileştirebilecek ve performansı arttırabilecek faktörler ortaya çıkarılmaya çalışılmıştır. Araştırmadaki ikinci analizin amacı ise, içeriklerin okunma sayılarının şirketin geliri üzerindeki etkisini ölçmektir. Araştırmanın her iki amacı için gerçekleştirilen analizler farklı gibi görünse de organizasyon açısından önemli ve birbirleriyle son derecede yakından ilişkilidir. Çünkü gelir ve okunma sayıları arasındaki ilişkinin saptanması ile birlikte, okunma sayılarının performansını arttırabilecek faktörlerin ortaya çıkarılmasının önemi daha da artmaktadır. Organizasyon, bu araştırma sonucu elde edilen okunma performansını arttırabilecek faktörleri, yeni içeriklerin hazırlanma sürecinde kullanarak, gelir üzerindeki etkiyi görebilir ve veri bilimine dayalı sürdürülebilir bir büyüme yakalayabilir. Çalışmada yukarıda tanımlanan iki amaca bağlı olarak aşağıda sunulan araştırma soruları belirlenmiştir:

- Veri ve metin madenciliği yöntemleri kullanılarak, başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime grupları) ortaya çıkarılabilir mi?
- Başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime grupları) nelerdir?
- Başlık metinlerindeki tek başına anlam ifade etmeyen etkisiz kelimeler (stop-word) çalışmanın ele aldığı bu problem için istatistiksel olarak anlamlı mıdır?
- Başlık metinlerindeki kelimeler, köklere indirgenmeden ekleriyle beraber analiz edildiğinde, istatistiksel olarak anlamlı sonuçlar ortaya çıkıyor mu?
- İçeriklerin okunma sayılarıyla şirket gelirleri arasında bir ilişki var mıdır?
- İçeriklerin okunma sayıları ve şirket gelirleri kullanılarak şirket gelirleri için veri madenciliği yöntem ve teknikleriyle bir tahmin modeli oluşturulabilir mi?

Bu çalışmada iki farklı analiz ile organizasyon içerisinde oluşan ham verilerin işlenerek değerli bilgilerin ortaya çıkarılması hedeflenmiştir. Bu bağlamda, araştırmada ele alınan iki farklı analiz, aşağıda sunulan iki hipoteze dayandırılmaktadır.

- İçerik başlıklarında kullanılan kelime ve kelime grupları içeriklerin okunma sayısına etki etmektedir.
- İçeriklerin okunma sayıları ile organizasyon gelirleri arasında bir ilişki bulunmaktadır.

1.3. Araştırmanın Kapsamı

Yayınlanan içeriklerin araştırmanın amacına uygun olduğu görüldüğünden, tezin uygulama bölümü için blokzincir (blockchain) teknolojileri ve kripto paralar hakkında dijital içerik üreten CoinTürk (https://coin-turk.com) platformu seçilmiştir. Mart 2014 tarihinden itibaren içerik üreten platformda, Eylül 2022 tarihine kadar toplamda 30.000'in üzerinde metin formatında içerik yayınlanmıştır. Platform genel olarak incelendiğinde içeriklerin; haber, analiz, uzman yorumları, son dakika gelişmeleri, makale ve rehber yazılar gibi çeşitli kategorilerde, özgün yazılar olduğu görülmektedir.

Çalışmanın 4.1. bölümünde sunulan birinci analizde gerekli ham veriyi toplamak için üç aylık bir dönem belirlenmiştir. Belirlenen 1 Mayıs 2022 – 31 Temmuz 2022 tarihleri arasında platform 2.743 içerik yayınlamıştır. Yayınlanan 2.743 içerik içerisinden basit rastgele örnekleme yöntemiyle 500 içerik seçilmiştir. Bu araştırma, seçilen 500 içeriğe ait başlıkların, veri ve metin madenciliği teknik ve yöntemleriyle incelenmesiyle içeriklerin okunma sayılarını etkileyen faktörlerin ortaya çıkarılmasına dayanmaktadır.

Çalışmanın 4.2. bölümünde yer alan ikinci analizde kullanılan veriyi oluşturmak için belirli bir dönem seçilmiştir. Araştırma için Ocak 2016 – Ekim 2021 tarihleri arasındaki şirket gelirleri (USD para biriminde) aylık bazda ilgili şirketten Excel dosyası olarak sağlanmıştır. Veri kümesinin 2016 yılının Ocak ayından itibaren başlamasının

nedeni, şirket gelirlerinin bu tarihten itibaren yazılı olarak kayda geçirilmesinden kaynaklanmaktadır. Ayrıca platformun sayfa gösterim (okunma) verileri de aynı tarih aralığı için aylık bazda sağlanmıştır. Bu araştırmada, elde edilen veriler üzerinde analizler gerçekleştirilerek şirketin gelirleriyle okunma sayıları arasındaki ilişki modellenmiştir.

1.4. Araştırmanın Yöntemi

Araştırmada ele alınan konuya ait kavramsal çerçevenin oluşturulabilmesi için belgesel tarama yöntemi kullanılmıştır. Belgesel tarama, araştırılan konu hakkında yayınlanmış makale, kitap veya video gibi kaynakların toplanması ve incelenmesi olarak ifade edilebilir (Karasar, 2022, s. 229). Bu bağlamda tez çalışmasının ilk yıllarında çok sayıda kitap ve makale incelenmiş ve her şeyden önce konunun teorik kısmı kavranmaya çalışılmıştır. İlk olarak ilgili kaynaklardan veri, enformasyon ve bilgi kavramları incelenmeye çalışmıştır. Bu kavramlar arasındaki ilişki üzerine okumalar yapılmış ve veri madenciliği ile bilgi yönetimi kavramları arasındaki etkileşim araştırılmıştır. Sonraki aşamada veri madenciliği kavramıyla ilgili kaynaklar detaylı bir şekilde incelenerek öğrenilmeye çalışılmıştır. Veri madenciliği kavramının ne olduğu, önemi, zorlukları, modelleme ve süreç yönetimi hakkında oldukça fazla sayıda okuma yapılmış ve tezin kavramsal çerçevesi oturtulmaya çalışılmıştır. Sonraki süreçlerde veri madenciliği yöntem ve teknikleri araştırılmış ve veri madenciliğinde en çok kullanılan teknikler ve algoritmalar incelenmiştir. Veri madenciliğinin kavramsal çerçevesi oluşturulduktan sonra tezde yapılması planlanan uygulama için veri madenciliğinin özel uygulama alanlarından biri olan metin madenciliği hakkında araştırma ve incelemeler yapılmıştır. Bu kapsamda metin madenciliği hakkında kavramsal çerçeve oluşturulmuş ve metin madenciliğinde en çok kullanılan teknikler ile birlikte tezin teorik kısmı şekillendirilmiştir.

Sonraki aşamada çalışmanın uygulama kısmına geçilmiştir. Çalışmadaki tüm veri madenciliği süreçleri için CRISP-DM yaklaşımı benimsenmiştir. Bu yaklaşımla ilgili detaylı bilgi, tezin 2.7.5.1. CRISP-DM bölümünde sunulmaktadır. CRISP-DM yaklaşımında süreç; işi anlamak, veriyi anlamak, veri hazırlama, modelleme, değerlendirme ve uygulama aşamalarından oluşmaktadır. Sürecin aşamaları döngüsel olarak ilerlemektedir ve uygulama, istenen sonuca ulaşılana kadar sürecin önceki basamaklarına geri dönüp tekrar aşamaları gözden geçirmek üzerine tasarlanmıştır.

Tez çalışmasının amaçlarına bağlı olarak uygulama kısmı iki farklı analiz olarak kurgulanmıştır. İlk analizde, içeriklerin başlık bazında okunma sayılarını etkileyen faktörler üzerine analizler yapılmıştır. Öncelikle uygulamanın gerçekleştirileceği platformun MySQL[2] veritabanından SQL[3] sorgusuyla ham veri çekmek için denemeler yapılmıştır. İncelemeler sonucu araştırmanın amacına uygun veri çekimi için sorgular oluşturulmuştur. Sorgular sonucu araştırmaya uygun, içeriklerin başlıklarına ait metin formatında ham veriler elde edilmiştir. Sonrasında Google Analitik[4] uygulamasıyla içeriklere ait okunma istatistikleri elde edilmiştir. Metin madenciliği kapsamına giren veri hazırlama, metin ön işleme ve değişken yaratma işlemleri için ise veri madenciliği uygulamalarında sıklıkla kullanılan Python[5] programlama dili üzerine çalışmalar ve incelemeler yapılmıştır. Tüm metin madenciliği, değişken yaratma ve veri işleme süreçleri için Anaconda Navigator[6] 1.9.12 sürümü içerisinde yer alan JupyterLab[7] 2.1.5

[2] Veritabanı kavramı temel olarak bir veya daha fazla ilişkili dosyada depolanan veriler topluluğudur. MySQL ise veritabanlarında verilerin güvenli bir şekilde saklanması, gerektiğinde sıralanarak sorgulanması ve filtrelenmesi vb. gibi işlemlerin yapılmasını sağlayan açık kaynak kodlu bir veritabanı yönetim sistemidir (Kofler, 2004, s. 4).
[3] SQL (Structured Query Language/Yapılandırılmış Sorgu Dili), veritabanı yönetim sistemleri tarafından yönetilen veritabanlarına kullanılan bir sorgulama dilidir (Melton ve Simon, 1993, s. 4).
[4] Google Analitik, web sitelerine yapılan kullanıcı ziyaretleri hakkında oturum, sayfa gösterimi ve oturum süresi gibi ayrıntılı istatistikler oluşturan ücretsiz bir uygulamadır (Plaza, 2011, s. 477).
[5] Python; modüller, dinamik yazım uygulamaları, yüksek seviyeli dinamik veri türleri ve sınıfları içeren, etkileşimli ve nesne yönelimli bir programlama dilidir (https://docs.python.org/3/faq/general.html#what-is-python).
[6] Anaconda Navigator, Python programları için yaratılan, açık kaynaklı araç ve veri kütüphanelerinin bulunduğu bir masaüstü kullanıcı arabirimidir (https://docs.anaconda.com/navigator/).
[7] JupyterLab, veri biliminde kullanılan web tabanlı ve etkileşimli bir geliştirme ortamıdır (https://jupyter.org/).

üzerinde, Python programlama dili ile kodlama yapılmış ve ayrıca, Numpy[8], Pandas[9], Matplotlib[10] ve Sklearn[11] veri kod kütüphaneleri kullanılmıştır. Metin formatındaki ham veriler metin madenciliği süreçleri sonucu Excel formatında yapılandırılmış biçime dönüştürülmüştür. Veriler hazırlandıktan sonraki aşamada ise, analizi gerçekleştirebilmek için bir istatistik yazılımı olan Minitab[12] programı öğrenilmiştir. Elde edilen içeriklere ait yapısal veriler Minitab programına yüklenmiştir. Bu program üzerinde, veri kümesi içerisinde yer alan her bir içeriğin okunma sayıları ile içeriğin başlığından (bu çalışmanın ele aldığı probleme özel) farklı bir yaklaşımla metin madenciliği teknikleriyle elde edilen değişkenler, bir çoklu doğrusal regresyon problemi olarak ele alınmıştır. Okunma sayıları bağımlı değişken olarak atanırken, metinlerden yaratılan diğer tüm değişkenler (kelime ve kelime grupları) bağımsız değişkenler olarak atanmış ve çoklu doğrusal regresyon modeli oluşturulmuştur.

İkinci analizde ise içeriklerin okunma sayılarının şirketin gelirleri (USD para biriminde) üzerindeki etkisi, basit doğrusal regresyon problemi olarak ele alınmıştır. Bu analizde kullanılan veriler ilgili şirketten Excel formatında aylık bazda sağlanmıştır. Aylık şirket gelirleri bağımlı değişken olarak atanırken, aylık okunma sayıları bağımsız değişken olarak atanıp basit doğrusal regresyon modeli Minitab programında oluşturulmuştur.

Her iki analiz için yapılan tüm işlemler tezin dördüncü bölümünde detaylarıyla adım adım açıklanmıştır. Elde edilen bulgular, "Sonuç ve Öneriler" bölümünde tartışılmış ve yorumlanmıştır.

[8] Numpy, Python'da bilimsel hesaplamalarda kullanılan, diziler üzerinde matematiksel ve mantıksal işlemlerin yanı sıra, sıralama ve seçme gibi işlemlerin yapılabildiği bir kütüphanedir (https://numpy.org/devdocs/user/whatisnumpy.html).
[9] Pandas, Python programlama dili üzerinde inşa edilen açık kaynaklı bir veri analiz ve işleme aracıdır (https://pandas.pydata.org/).
[10] Matplotlib, Pyhton programlama dilinde veri görselleştirme için kullanılan kapsamlı bir kütüphanedir (https://matplotlib.org/).
[11] Sklearn, veri analizleri için kullanılan, istatiksel ve matematiksel işlem ve algoritmaların çalıştırılabildiği Pyhton programlama dili için hazırlanmış bir kütüphanedir (https://scikit-learn.org).
[12] Minitab; verilerdeki örüntüleri ve eğilimleri keşfetmek, gizli ilişkileri ortaya çıkarmak ve görselleştirme, tahminleme yapmak gibi çok sayıda veri analiz işleminin yapılabildiği veri analiz aracıdır (https://www.minitab.com).

1.5. Araştırmanın Düzeni

Çalışma beş ana bölümden oluşmaktadır. İlk bölümde bir giriş yapıldıktan sonra araştırmanın konusu, önemi, amacı, kapsamı, yöntemi, düzeni ve kullanılan kaynaklar verilmiştir. İkinci bölümde veri madenciliğine giden yol olarak veri, enformasyon, bilgi ve bilgi yönetimi kavramı ve aralarındaki ilişki, sonrasında büyük veri, metin madenciliği ve veri madenciliği kavramları ele alınmıştır. Ek olarak, yine bu bölümde çalışmanın ana konusu olan veri madenciliği kavramının detaylarına inilmiştir. Üçüncü bölümde ise veri madenciliğinde kullanılan yöntem ve teknikler ile veri madenciliği algoritmalarıyla ilgili tanımlamalara ve açıklamalara yer verilmiştir. Çalışmanın dördüncü bölümünde ise iki farklı analiz yapılmıştır. İlk analizde ham verinin elde edilmesiyle başlayan süreçte, yapısal olmayan verilerin metin madenciliği teknikleriyle işlenmesi ve yapısal forma dönüştürülmesi, sonrasında probleme uygun veri madenciliği yöntem ve tekniklerin belirlenerek uçtan uca bir veri madenciliği uygulaması ele alınmıştır. Ayrıca ilk analiz için ele alınan problem özelinde "Başlık bazlı dijital içerik analiz süreci" geliştirilmiştir. İkinci analizde ise platformun okunma sayılarının gelir üzerindeki etkisi araştırılmış ve analiz edilmiştir. Bu analize yönelik gelir ve okunma sayıları arasında bir tahmin modeli oluşturulmuştur. Araştırmaya dair değerlendirme ve öneriler ise son bölüm olan beşinci bölümde paylaşılmıştır.

1.6. Kaynaklar

Araştırmanın konusuyla ilgili kavramsal çerçeveyi oluşturmak için birçok kaynaktan literatür taraması yapılmıştır. Bu kaynaklar aşağıda sıralanmıştır:

- Ankara Üniversitesi Kataloğu http://katalog.ankara.edu.tr/
- Bilkent Üniversitesi Kataloğu https://librarycatalog.bilkent.edu.tr/
- DergiPark https://dergipark.org.tr/
- Emerald http://www.emeraldinsight.com/

- Google Scholar https://scholar.google.com
- IEEEXplore Digital Library https://ieeexplore.ieee.org/
- Library Information Science and Technology Abstracts http://www.ebsco.com/products/research-databases/library-information-science-and-technology-abstracts
- Milli Kütüphane Kataloğu http://www.mkutup.gov.tr/
- ODTÜ Kütüphanesi Bütünleşik Arama https://eds.p.ebscohost.com/
- ODTÜ Kütüphanesi Kataloğu https://lib.metu.edu.tr/tr
- ProQuest Dissertations and Theses https://search.proquest.com/
- Scopus https://www.scopus.com/
- SpringerLink https://link.springer.com/
- Taylor & Francis Online Journals http://www.tandfonline.com/
- Web of Science http://webofknowledge.com/WOS
- Wiley Online Library https://onlinelibrary.wiley.com/
- YÖK Tez Veritabanı https://tez.yok.gov.tr/UlusalTezMerkezi/

2. BÖLÜM: KAVRAMSAL ÇERÇEVE

Bu bölümde çalışmanın konusuna uygun olarak kavramsal tanımlamalar yapılmış, teknik detaylar ve terimlere açıklık getirilmiştir.

2.1. Veri

"Latince tekil 'datum' sözcüğünün çoğulu olan veri (data); bir uslamlama, tartışma veya hesaplama dayanağı olarak kullanılan olgudan elde edilmiş bilgi anlamına gelir. Veri; kaydedilen, yeniden düzenlenebilen ve analiz edilebilen bilgi birimidir" (Işıklı, 2014, s. 92). Veri; analiz edilmemiş, ilişkilendirilmemiş ve ham haliyle anlam çıkaramayacağımız bir bilgidir. Veriler, bilgiye ulaşmada temel ham madde olarak görülebilecek çeşitli sembol, harf ve rakam gibi işaretlerle temsil edilen, ham ve işlenmemiş izlenimlerdir (Safran, 2015, s. 552). Veri en basit haliyle karmaşık ve kaotik yapıyı temsil eder. Temel amaç veriye bir düzen verip analiz ederek ondan anlam çıkarmak, yani bilgiye ulaşmaktır (Gürsakal, 2014, s. 35). Genellikle amacı bilgiye dönüştürülmek olan işlenmemiş ham malzemeye "veri" denir (Şentürk, 2006, s. 4).

Sayısal veya dijital verinin temeli 1 ve 0'lardan oluşan bit dizisi şeklindedir (Çölkesen, 2016, s. 44). Bu, enformasyon ve bilginin en küçük ham halini temsil eder. Günümüz bilgi çağında veri, bu tanımlara ek olarak, bir veya birden fazla karakteristik özelliklerine göre sütün ve satırlarla organize edilen değerler olarak değerlendirilebilir. Bilgi çağından çok önce ticaret ve hesaplamanın ortaya çıkmasıyla birlikte veri kavramı doğmuştur. Basitçe ifade etmek gerekirse, değerlerin rakamlara atanması, sonrasında bu rakamların parasal değere, nüfusa, vergilere veya takvimlere dönüşmesine yol açmıştır. Böylece insan zihni veri edinimi ve onu yönetmeyle ilgili bir uğraşa girmiştir (Krishnan, 2013, s. 2). Bilgisayar açısından veri türleri; yapılandırılmış veri, yapılandırılmamış veri ve yarı yapılandırılmış veri olarak üç gruba ayrılabilir.

Yapılandırılmış Veri (structured data); aynı formata, kurala ve aynı düzene sahiptir ve temel anlamda satır ve sütunlarla organize edilmiştir (Gürsakal, 2014, s. 54). Bir dosyada satır ve sütunlara ayrılmış son bir yıllık banka hesabı işlem geçmişi yapılandırılmış veriye örnek olarak gösterilebilir.

Yapılandırılmamış Veri (unstructured data) herhangi bir türde olabilir ve herhangi bir format, dizi veya kuralı izlemezler. Yapılandırılmamış veri örnekleri satır ve sütunlarla ifade edilemeyen metin, video, ses ve görüntü gibi içerikleri tanımlamaktadır (Gürsakal, 2014, s. 54).

Yarı Yapılandırılmış Veri ise (semi-structured data), ne ham veridir ne de katı bir şekilde geleneksel veritabanı sistemlerinde olduğu gibi yazılan verilerdir. Verilerin belirli bir yapısının olduğu, fakat hepsinin aynı yapıya sahip olmadığı durumlar yarı yapılandırılmış verilere örnek gösterilebilir (Abiteboul, 2005, s. 1).

2.2. Enformasyon

Temel olarak enformasyon (information), anlaşılabilir organize bilginin iletimidir. Daha geniş bir ifadeyle, bir ortam üzerinde anlaşılabilir ve iletilebilir bir biçimde yer alan veriler topluluğudur (Gürdal, 1991, s. 116). Bu bağlamda enformasyon, "bilginin iletilebilir bir biçimde organize edilmesi sürecinin ürünüdür" şeklinde ifade edilebilir (Gürdal, 1998, s. 35).

Veriler üzerinde çalışılıp çeşitli tekniklerle analizler yapılıp ilişkiler kurulduktan sonra, veri enformasyona dönüşmektedir (Safran, 2015, s. 552). Diğer bir ifadeyle enformasyon, verinin düzenlenmiş halidir (Tüzüntürk, 2010, s. 65). Enformasyon bunlara ek olarak, veriler arası ilişkiyi içerisinde barındırmaktadır (Gülseçen, 2014, s. 63).

2.3. Bilgi

Enformasyon kavramının bir sonraki basamağı olarak görülen bilgi (knowledge) kavramının çok sayıda farklı tanımlamaları vardır. Bunlardan bazıları aşağıda sıralanmıştır.

Bilgi kişinin beynindedir, özümsenmiştir ve öğrenme/deneyim yoluyla kazandıklarının toplamıdır. Alınan enformasyon değerliyse onu alan kişinin bilgi birikimi bütünleşir ve bilgi deposuna eklenir (Safran, 2015, s. 552). Bilgi, anlamlı bir ileti, gözlem ve sayısal/görsel raporlar toplamıdır (Arslantekin, 2003, s. 372). Anlamlı bir biçimde derlenen ve birleştirilen, verilerden oluşan gerçek değer "bilgi" olarak kabul edilir (Şentürk, 2006, s. 4). Türk Dil Kurumunun yaptığı bilgi tanımları ise şu şekildedir:

- İnsan aklının erebileceği olgu, gerçek ve ilkelerin bütünü, bili, malumat.
- Öğrenme, araştırma veya gözlem yolu ile elde edilen gerçek, malumat, vukuf.
- İnsan zekâsının çalışması sonucu ortaya çıkan düşünce ürünü, malumat, vukuf.
- Genel olarak ve ilk sezi durumunda zihnin kavradığı temel düşünceler.
- Kurallardan yararlanarak kişinin veriye yönelttiği anlam (Türk Dil Kurumu, 2020).

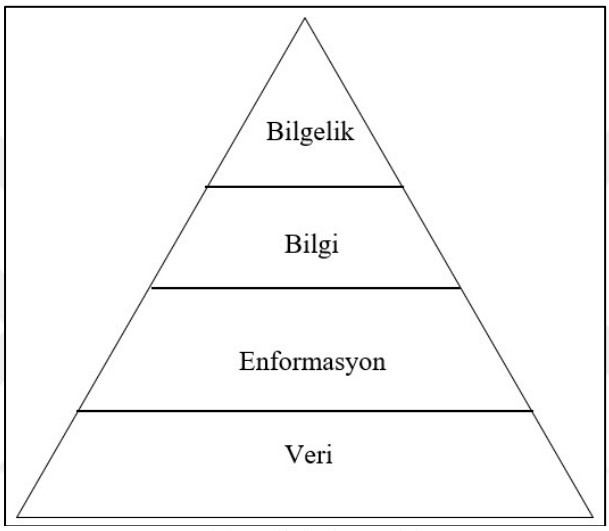
Bilgi, yeni olay ve durumları değerlendirmek üzere kullanılan, eyleme geçme kapasitesini arttıran, birleşik ve geçerli deneyimler, inançlar ve fikirlerdir (Alkan, 2017, s. 158). Bilgi o zamana kadar bildiklerimizi ya da inandıklarımızı değiştirir, cehaleti ve belirsizliği azaltır (Tonta, 2001, s. 199). Görülebileceği gibi bilgi kavramı hakkında çok sayıda farklı görüş ve tanım bulunmaktadır. Ele alınan alana göre de bazen bilginin tanımının değişebileceği de bilinmektedir.

2.4. Bilgi Yönetimi

Verinin elde edilmesi, hazırlanması, analiz edilip işlenmesi ve sonrasında anlamlı iletilebilir enformasyona dönüşme süreci tek bir olguya ulaşmak için yapılmaktadır, bu

olgu da tartışmasız bilgidir. Stratejik önem ve değere sahip bilgiye ulaşmak, veriyle ilgili tüm süreçlerde ana hedef olarak görülmektedir. Çünkü günümüzde bilgi istisnasız olarak kurumlar, şirketler ve organizasyonlar için en önemli unsur haline gelmiştir. Bu bağlamda değerlendirilecek olunursa, veriyle dolup taşan organizasyonların bilgiye ulaşma ve elde edilen bilgiyi kullanma çabası içerisinde olması kaçınılmazdır.

Şekil 1'de gösterildiği gibi veriden enformasyona, oradan da bilgiye ve bilgeliğe doğru giden bir hiyerarşi söz konusudur (Rowley, 2007, s. 164).



Şekil 1. Bilgi Hiyerarşisi

Veri, enformasyon, bilgi ve bilgelik hiyerarşisi (The data–information–knowledge–wisdom hierarchy – DIKW-), genellikle bilgi hiyerarşisi, enformasyon hiyerarşisi veya bilgi piramidi ismiyle de kullanılmaktadır (Rowley, 2007, s. 163). Bilgi hiyerarşisi; veri, enformasyon, bilgi ve bazen de bilgeliği, birbirleri arasında ilişkilendirmek ve bir varlığın (örneğin verinin) hiyerarşide daha yüksek bir seviyedeki varlığa (örneğin enformasyona) dönüştürülmesi ile ilgili süreçleri tanımlamak ve belirlemek için kullanılır (Rowley, 2007, s. 165). Literatürde bilgi hiyerarşisi ile ilgili geçmişten günümüze çok farklı görüşler yer almaktadır.

Kenneth Boulding, 1955 yıllında bu hiyerarşinin ilk formlarından biri olan, "Signals, Messages, Information, and Knowledge", "Sinyaller, Mesajlar, Enformasyon ve Bilgi" kavramlarından oluşan yaklaşımını çalışmasında sunmuştur (Wallace, 2017, s.

13). Russell L. Ackoff'un (1989) "From Data to Wisdom" adlı makalesinde geliştirdiği yaklaşımda, bilgelik (wisdom) kavramı piramidin tepesinde yer almıştır. Ackoff, bu yaklaşımını makalesinde bir görsel olarak sunmasa da bilgelik kavramının piramide girmesinde önemli bir rol oynamıştır. Singh ve Hetlevik (2017) ise; veri, enformasyon, bilgi ve bilgelikten sonra piramidin en tepesine "karar alma" (decision) kavramını eklemiş ve literatüre farklı bir bakış açısı getirmiştir. Literatürde bu ve benzeri birçok farklı yaklaşım bulunmaktadır.

Danny P. Wallace'a göre ise; veri, enformasyon, bilgi ve bazen de bilgelik kavramları arasındaki ilişki ve aralarındaki hiyerarşinin düzenlenmesi, uzun yıllardır bilgi bilim dilinin bir parçası olmuştur (Wallace, 2017, s. 13). Bilgi hiyerarşisi, bilgi yönetimi kavramının merkezinde yer alan bir olgudur. Bu nedenle de konu, bilginin temel doğası gereği; iletişim teorileri, kütüphane ve bilgi bilim, bilgi yönetimi, bilgi sistemleri ve bilişsel bilimler tarafından dikkatle incelenmiştir (Rowley, 2007, s. 165). Awad ve Ghaziri'den (2004) aktaran Gülseçen (2014)'e göre; bilgi yönetimi disiplinler arası bir iş modelidir ve ekonomi, psikoloji, enformasyon yönetimi gibi birçok disipline dayanarak insanları, teknolojiyi ve süreçleri içeren ölçümler ve değerlendirmeler gibi konulara odaklanır. Bilgi yönetimi için diğer yaygın tanımı ise organizasyon içerisinde bilgiye ihtiyacı olanlara ihtiyaç duydukları yerde, ihtiyaç duydukları zaman ve ihtiyaç duydukları şekilde organizasyonel performansı ve rekabet avantajını artırmak için bilgi sunmaktır (Vukasinov ve Draskovic, 2013, s. 70).

Bilgi yönetimi (information management) kavramı, organizasyonların rekabet avantajı için iş süreçlerini, sistemleri, yazılım, donanım, veri ve bilgi paylaşımını ve tüm bunların katma değerli varlıklara dönüştürülmesini içerir. Bilgi yönetimi günümüzün dijital ekonomisi için hayati önem taşımaktadır. Bilgi yönetiminin organizasyona faydaları şu şekilde sıralanabilir:

- Rekabet avantajı sağlar,
- Maliyeti düşürmeye ve organizasyonel verimliliği artırmaya yardımcı olur,
- Organizasyonun değişen iş ve müşteri gereksinimlerine tepki verecek ve bunlara uyum sağlayacak kadar esnek olmasını sağlar,
- İnovasyon ve yaratıcılığın geliştirilmesine katkı sağlar,
- Sürdürülebilir iş modelleri kurulmasına yardımcı olur,
- Çalışan üretkenliğini artırır ve yönetim kararlarını daha verimli hale getirebilir (Vukasinov ve Draskovic, 2013, s. 71).

Günümüz dijital çağında bilgi yönetimi, rekabet avantajı sağlayan önemli bir kurumsal kaynak haline gelmiştir. Birçok kuruluş çok büyük miktarlarda veri toplamakta ve saklamaktadır. Ancak bu verileri faydalı bilgilere dönüştürerek verilerde saklı olan değerli bilgiler ortaya çıkarılamazsa, veri toplamanın da bir anlamı kalmaz. Dolayısıyla bilgi yönetimi kavramı aslında bir veri kullanım sürecini ifade etmektedir. Veri madenciliğinin temeli ise, büyük veri kümelerinden faydalı bilgiler çıkarmak için uygulanan süreçlerdir. Bu nedenle veri madenciliği bilgi yönetiminin önemli bir parçasını oluşturmaktadır (Silwattananusarn ve Tuamsuk, 2012, s. 13).

Bilgi yönetiminin temel amacına bakıldığında; yeni teknolojilerin organizasyona uyarlanması, verimliliğin arttırılması, maliyetlerin düşürülmesi, rekabet gücünün yükseltilmesi ve isabetli kararlar ile işletmenin sağlıklı büyümesinin sağlanması olarak ifade edilmektedir (Barutçugil, 2002, s. 21). Veri madenciliği ise verileri uygulanabilir bilgilere dönüştürür (Tsiptsis ve Chorianopoulos, 2010, s. 3). Veri madenciliğinde temel amaç değerli bilgiye ulaşmaktır. Ham haliyle hiçbir anlam içermeyen veri, üzerinde çalışılıp analiz edildiğinde iletilebilir enformasyona dönüşür, sonrasında ise alınan bu enformasyon, bilgi olarak değerlenir.

Bilgi yönetimi ile veri madenciliği uygulamalarının; verimlilik, karar destekte kullanma, rekabet gücünü arttırma, maliyetlerin düşürülmesi gibi bir organizasyon

içerisindeki amaçları neredeyse aynıdır. Her ne kadar bilgi yönetimi kavramı içerisinde değerlendirilebilecek bilgi, sadece veri madenciliği sonucunda değil, tecrübe ve birikim gibi farklı kanallardan sağlansa da bilginin üretilmesi ve bilginin yönetilmesi şeklinde ifade edilebilecek bu iki kavram, birbirlerini tamamlayıcı unsurlar olarak yorumlanabilir ve birbiriyle yakın bir ilişkiye sahip olduklarından bahsedilebilir. Çünkü içerisinde veri, enformasyon ve bilginin yönetimini barındıran bilgi yönetimi alanının, veri çalışmaları ve veri biliminden ayrı düşünülmesi doğru bir yaklaşım olmayacaktır. Bu düşünceyi 2000'li yılların başlarında fark eden Sacit Arslantekin, 2003 yılında yazdığı "Veri Madenciliği ve Bilgi Merkezleri" adlı makalesinde, veri madenciliği ve bilgi merkezleri olarak kabul edilen kütüphaneler arasındaki ilişkiye dikkat çekmektedir. Makale, ülkemizdeki alan literatüründe veri madenciliği konusunu ele alan ilk çalışmalardan biri olarak değerlendirilebilir.

Gün geçtikçe artan veriye sahip organizasyonlar, rekabet, verimlilik, karar alma vb. süreçler için veri madenciliği ve analizler yapmak zorundadırlar. Analiz sonrası bilgisayar tarafından sunulan sonuç ise enformasyondur. Enformasyon, bireyler tarafından yorumlanıp kullanıldığında insana özgü bir kavram olan bilgiye, yani bir değere dönüşür. Bu noktada elde edilen bilgi, bilgi yönetimi kavramı çerçevesinde değerlendirilerek, organizasyona değer katmaktadır. Böylelikle doğal olarak bilgiye ve bilgeliğe ulaşmada hammadde olarak görülen veri ve verinin işlenmesiyle ilgili tüm çalışmaların ve uğraşların, bilgi yönetimi ve bilgi bilimiyle uğraşan alanlarla yakından ilgili olduğu ve hatta bu alanların önemli bir parçasını oluşturduğu şeklinde bir çıkarım ve değerlendirme yapılabilir.

2.5. Büyük Veri

Günümüzde her yer gittikçe artan veri ile dolup taşmakta ve görünüşe göre de bunun bir sonu gelmeyecek. Bilgisayarlar bir şeyleri kaydedip saklamayı çok

kolaylaştırdığından, artık elde var olan veri hakkında verilmesi gereken kararlar ertelenebiliyor ve çok daha fazla bellek ile neredeyse her şey saklanabiliyor. Üstelik içerisinde bulunduğumuz dijital çağda bireylerin yaptığı her şey ve her seçim bir veritabanına kaydediliyor (Witten, Frank ve Hall, 2011, s. 4). Küresel olarak tüm bu saklanan, depolanan ve kaydedilen veriler karşımıza büyük veri (big data) kavramını ortaya çıkarıyor. Literatürde büyük veri hakkında evrensel olarak kabul gören bir tanım olmadığından dolayı birçok farklı yaklaşım bulunmaktadır.

Büyük veri, farklı hızlarda değişkenlerle yaratılan, geleneksel teknolojiler, algoritmalar, yöntemler veya çözümlerle işlenemeyen, karmaşıklığın çeşitli derecelerinde var olan veri hacmi olarak düşünülebilir. Bu bağlamda değerlendirildiğinde yukarıda büyük veri olarak tanımlanan verilere ek olarak, nükleer santrallerden, x-ışını ve tarama cihazlarına, uçak motorlarından, dijital dünyadaki sosyal medya uygulamalarına kadar yaratılan her tür veri büyük veri olgusu kapsamına girmektedir. Büyük veri, verilerdeki tekrarlanan veya tek bir sefere mahsus örüntülerden kritik öngörüler kazanmak için kullanılan büyük miktardaki veriye erişim yeteneği olarak da tanımlanabilir (Krishnan, 2013, s. 5).

Büyük veri, başlangıçta büyük miktarlarda toplanan veriyi belirtmek için kullanılmıştır; fakat büyük verinin ne olduğuyla ilgili formüle edilmiş açık bir tanımlama olmadığından herhangi bir veri setinin aslında büyük veri olup olmadığına karar vermek mümkün değildir. Büyük verinin diğer bilimsel metotlarla karşılaştırıldığında büyük avantajlara sahip olduğu iddiası normal veri ile büyük veri arasındaki farklılıkların netleştirilmesini gerektirmiştir. Literatürde büyük veri kavramını formüle etmeye çalışan bazı araştırmalar bulunmaktadır; yaygın kullanılanlar şu şekildedir (Grandinetti, 2015, s. 6):

- 3V tanımı: Hacim (volume), çeşitlilik (variety) ve hız (velocity);
- Bir organizasyondaki tüm veriler;

- Yerel bilgisayar sistemi hafıza boyutuna kıyasla daha büyük veri setleri.

Bunlar arasında en popüler olan büyük veri tanımı 3V modelidir. Hacim, Çeşitlilik ve Hız olarak sıralanabilecek bu özellikler büyük verinin tanımlayıcıları olarak değerlendirilir. Hacim veri miktarını, çeşitlilik veri türlerinin sayısını ve hız ise veri işlem hızını ifade etmektedir. 3V modeline göre büyük veri yönetiminin zorlukları, yönetilecek veri miktarından ziyade bu üç özelliğin açıkça tanımlanamamasından kaynaklanır; çünkü hacim, çeşitlilik veya hız ile ne kastedildiği ya da farklı veri setlerini karşılaştırmak için bunların hangi normlara göre nasıl ölçüleceği açık değildir. Bu nedenle, belirli bir veri havuzunun gerçekten bir büyük veri setini temsil edip etmediğini belirlemek için bu tanımı kullanmak da mümkün görünmemektedir. 3V modelinin 4V ve 5V modellerine genişletilmesiyle ilgili öneriler bulunmasına karşın, bu genişletmeler de temel olarak 3V ile tanımlanan büyük veri tanımlama problemini çözmekte yetersiz kaldığı görülmektedir (Grandinetti, 2015, s. 6).

Bir organizasyondaki tüm verilerin büyük veri olarak değerlendirilebileceği diğer tanım ele alındığında; Walmart, Amazon, Facebook veya Google gibi büyük bir şirketteki tüm verilerin yanı sıra, köşedeki küçük dükkanda bulunan tüm verilerin de büyük veri olacağı anlamı ortaya çıkmaktadır. Bu bakış açısına rağmen, bu tanım pratik durumlarda kullanılabilir. Büyük veya küçük olsun, bir organizasyonda mevcut olan veriler üzerinde tüm olası analizler uygulanabilir. Bu durumda da veri seti boyutunun, mutlaka "büyük" olması gerekmediğinden, büyük veri terimi kendi içerisinde bir ikilem oluşturabilir. Öte yandan, dünyadaki tüm veriler eğer küçük veya büyük fark etmeksizin tüm organizasyonlara ait olarak değerlendirilirse, bunların tümü büyük veri olarak sınıflandırılacak ve sonuç olarak dünya üzerindeki verilerin neredeyse tamamı büyük veri olarak adlandırılacaktır. Bu bakış açısıyla da büyük veri kavramını tanımlamak mümkün görünmemektedir (Grandinetti, 2015, s. 7).

Yerel bilgisayar sistemi hafıza boyutuna kıyasla daha büyük veri setleri olarak ifade edilen diğer alternatif tanım, belki de ölçüm ve karşılaştırmanın temelini oluşturabilecek bir tanımlamaya en yakın olanıdır. Böylece büyük veri, boyut olarak tanımlanabilir (Grandinetti, 2015, s. 7). Fakat büyük veriyi sadece verinin büyük olmasıyla ilişkilendirmek de doğru değildir ve bu kavramı veri boyutu ile sınırlandırmamak gerekir (Özdoğan, 2016, s. 11).

Genel olarak değerlendirildiğinde ise, büyük veri (big data), yapılandırılmamış, satır ve sütunlar düzeyinde tertip edilmemiş verilerin, büyük potansiyeli ve bu potansiyelin kullanılmasıyla ilgili bir kavram olarak ifade edilebilir. Dünyadaki verilerin %90'ına yakın bölümünün bu tip yapılandırılmamış veri olması büyük veri olgusunun önemini ortaya koymaktadır (Gürsakal, 2014, s. 56). Tüm verilerin %90'ına kadar yaklaşan bu oran; sosyal medya paylaşımları, tüm metin, ses ve görüntü dosyaları, ağ günlükleri, kişisel bloglar, bu çalışmanın da konu edindiği dijital içerikler, web siteleri, fotoğraflar, grafikler, videolar, görseller, log dosyaları, yazılımlar/programlar, e-postalar, bir uygulama üzerindeki mesajlar ve kaydedilmiş telefon görüşmeleri ve kısa mesajlar, kısaca internet üzerinde olsun veya olmasın işlenebilir her türlü içeriği kapsamaktadır.

2.6. Metin Madenciliği

Bir önceki başlıkta bahsedilen büyük veri kavramından bilgi çıkarmanın en önemli yollarından biri metin madenciliğidir. Bu açıdan değerlendirildiğinde büyük veri kavramı ile metin madenciliği arasında organik bir ilişki olduğu söylenebilir. Genel olarak metin madenciliği kavramı ise, veri madenciliği kapsamında değerlendirilebilir.

Metin madenciliğinde amaç, metinden veriyi çıkartarak analiz edip bilgiye dönüştürmek ve çeşitli karar verme süreçleri için üretilen bilgiyi kullanmaktır. Metin madenciliği süreci ilk olarak yapılandırılmamış olan metinlerden analiz için uygun nicel verilere dönüştürme işlemiyle başlar ve veri kümesi hazırlandıktan sonra, çeşitli makine

öğrenimi veya istatistiksel analiz yöntemlerinden birinin veya birkaçının uygulanmasıyla devam eder. Sonuç olarak, analiz sonucu elde edilen bilgiler yorumlanarak üretilen bilgi kullanıma sunulur (Bach ve diğerleri, 2019, s. 1278).

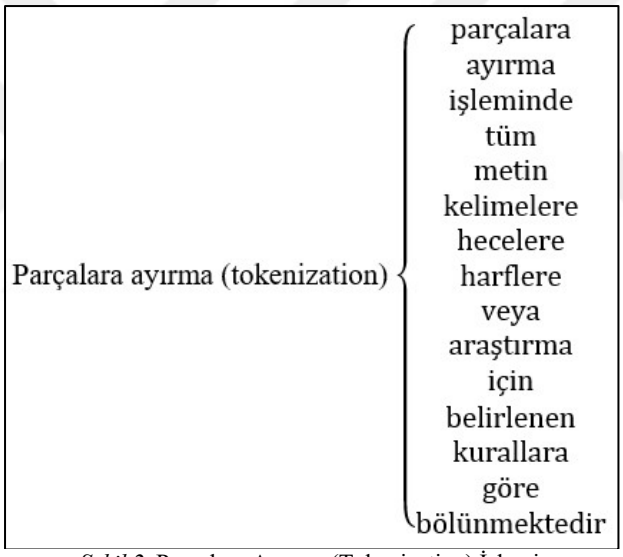
Metin madenciliği, yapılandırılmamış metinsel verilerden anlamlı bilgiler elde etmek amacıyla uygulanan yöntem ve tekniklerdir. Metin madenciliği, büyük veri kavramının da içerisinde yer alan tüm metin dosyaları, HTML dosyaları, sohbet mesajları ve e-postalar gibi yapılandırılmamış metinsel verilerden, faydalı modeller, eğilimler, kalıplar veya kurallar bulmaya odaklanır. Metin madenciliği, veri madenciliğinin içerisinde yer alan özel bir uygulama alanı olarak konumlandırılır (He, Zha ve Li, 2013, s. 464). Bir başka anlatımla metin madenciliği, daha önce bilinmeyen ve ortaya çıkarılması kolay olmayan metin verilerinden bilgi üretme işlemi olarak tanımlanmaktadır. Temel olarak metni, veri madenciliği sürecinde kullanılabilecek sayısal verilere dönüştürmeyi içerir (Zhang ve diğerleri, 2019, s. 239).

Bir metin madenciliği uygulamasında öncelikle, bir kaynaktan elde edilen metnin içerisinde yer alan özel karakterler ve noktalama işaretleri kaldırılarak eğer gerekliyse rakamların temizlenmesi gibi işlemler gerçekleştirilir, bu aşama metin ön işleme olarak değerlendirilebilir. Sonrasında metinlerden oluşan henüz yapılandırılmamış ham veri üzerinde kelime çıkarma/saydırma, ağırlıklandırma, kelime köklerine indirgeme, kelimelerin ikili veya üçlü kombinasyonlarını çıkartma, bazı problemlerin çözümü için kelimeleri hece veya harfe kadar indirgeme gibi işlemlerle analiz için uygun nicel/sayısal veriler yaratılır. Bu süreç, değişken yaratma veya özellik çıkartma olarak ifade edilebilir. Aslına bakılırsa bir araştırmanın metin madenciliği kısmı değişkenler yaratıldığında sona erer. Oluşturulan veri kümesiyle araştırma problemine uygun makine öğrenimi veya istatistiksel yöntem ve teknikler seçilerek veri madenciliği işlemleriyle araştırma sonuçlandırılır. Metin madenciliğinde kullanılan birçok teknik ve yöntem bulunmaktadır. Bunlardan en çok kullanılanları aşağıda açıklanmaya çalışılmıştır.

2.6.1. Parçalara Ayırma

Metin işlemenin ilk adımlarından biri olan parçalara ayırma (tokenization) işleminde tüm doküman cümlelere, kelimelere, hecelere, harflere veya araştırma için belirlenen kurallara göre bölünmektedir. Her bir cümle, kelime, hece veya harf, noktalama işaretleri veya boşluklarla belirlenen araştırma için seçilmiş kurallara göre parçalara ayrılır (Weiss, Indurkhya ve Zhang, 2010, s. 17).

Örnek olarak; "*parçalara ayırma işleminde tüm metin kelimelere, hecelere, harflere veya araştırma için belirlenen kurallara göre bölünmektedir*" cümlesi, kelime düzeyinde parçalara ayırma işlemine tabi tutulursa Şekil 2'de verilen liste elde edilir.



Şekil 2. Parçalara Ayrıma (Tokenization) İşlemi

Örnek verilen cümlede noktalama işaretleri ve boşluklar bir bölme aracı olarak kullanılarak kelime düzeyinde bir parçalara ayırma işlemi yapılmıştır. Bu yaklaşım metin madenciliğinde en çok kullanılan tekniklerden biridir.

2.6.2. Kök Çözümleme

Bir önceki başlıkta bahsedilen parçalara ayırma işleminden sonra, bir sonraki olası adım kelimelerin köklerini çözümleme (lemmatization) adımıdır. Bu adımın gerekli olup olmadığı araştırmaya konu olan probleme bağlıdır. Bu teknik, kelimelerin bağlamında

değerlendirilip köklerine indirgenmesi olarak tanımlanabilir (Weiss, Indurkhya ve Zhang, 2010, s. 18). Diğer bir kök bulma uygulaması olan gövdeleme (stemming) ise kelimedeki ön ekleri ve son ekleri atarak kelime köklerini bulmaktadır (Jivani, 2011, s. 1930).

Örnek olarak bir önceki başlıkta verilen cümlede geçen "*parçalara*" kelimesinin köküne inildiğinde kelime kökü "*parça*" olarak bulunur. Eğer veri kümesini oluşturan tüm metinlerde "*parça*" kökünden türetilmiş, "*parçaların*" veya "*parçalamak*" gibi aynı kökten türetilmiş kelimeler varsa, tüm bu kelimeler "*parça*" kökü altında birleştirilir.

2.6.3. Kelime Etiketleme

Kelime türlerini etiketleme (part-of-speech tagging) işlemi, kelimelerin bağlamına göre etiketlerle açıklanmasıdır. Bu etiketler, kelimeleri göründükleri cümlede oynadıkları role göre kategorilere ayırır ve kelimelerin anlamsal içeriği hakkında bilgi sağlar. En yaygın kullanılan etiket setleri; isim, fiil, sıfat, edat, bağlaç, zamir veya özel isimler vb. gibi etiketlerdir. Bu teknik, morfolojik analiz (morphological analysis) olarak da adlandırılmaktadır (Feldman ve Sanger, 2006, s. 60).

Örnek olarak; "*parçalara ayırma işleminde tüm metin kelimelere, hecelere, harflere veya araştırma için belirlenen kurallara göre bölünmektedir*" cümlesi ele alınsın. Bu cümlede yer alan "*veya*" sözcüğü bir bağlaçtır ve tek başına bir anlam ifade etmeyen bir kelimedir. Bu kelimenin bağlaç olarak etiketlendirilmesi bu tekniğe örnek olarak verilebilir.

2.6.4. Etkisiz Kelimeler

Etkisiz kelimeleri çıkarma (stop-word removal) işlemi, bir önceki başlıkta açıklanan kelime etiketleme işlemini takiben yapılmaktadır. Etiketlenen kelimeler içerisinden etkisiz kelimeler veri kümesi dışına alınmaktadır. Etkisiz kelimeler metin içeriğiyle ilgili olmayan dilbilgisel sözcüklerdir, bu nedenle daha fazla verim için

kaldırmaları gerekir. Etkisiz kelimeler kaldırıldıktan sonra kalan kelimeler genellikle isimler, fiiller ve sıfatlardır (Jo, 2019, s. 24).

Örnek olarak bir önceki başlıkta verilen cümle içerisinde geçen "*veya*" ifadesi bir bağlaç olarak etkisiz kelime grubuna girmektedir. Bu nedenle araştırmanın yapıldığı veri kümesinden kaldırılır.

2.6.5. Terim Ağırlıklandırma

Terim ağırlıklandırma (term weighting) işlemi, her kelimenin ilgili doküman, metin veya tüm veri kümesi içerisindeki ağırlığının önem derecesi olarak hesaplama ve atama sürecini ifade etmektedir (Jo, 2019, s. 26).

En yaygın kullanılan ağırlıklandırma tekniklerinden biri TF-IDF'dir (terim frekansı - ters metin frekansı / term frequency - inverse document frequency). TF-IDF, dokümanda bir kelimenin görünme sayısının, kelimenin görüldüğü doküman sayısının tersi ile çarpımıdır. Bu yöntemde, bir dokümanda sıklıkla geçen ve pek çok dokümanda yer almayan kelimeler bu nedenle önemli bir ağırlığa sahiptir (Berry ve Castellanos, 2008, s. 28).

2.6.6. Kelime Dizileri

Kelime dizileri (n-gram) tekniğinde, bir dokümanda bitişik şekilde yer alan kelime ve kelime gruplarının kaç defa geçtiği hesaplanmaktadır. Tek bir kelimenin sıklık (frekans) sayısının hesaplanması 1-gram (unigram), bitişik iki kelimenin sıklığının hesaplanması 2-gram (bigram) ve bitişik üç kelimenin sıklık sayısının hesaplanması ise 3-gram (trigram) olarak adlandırılmaktadır (Kumar ve Paul, 2016, s. 16).

Örnek olarak tekrar, "*parçalara ayırma işleminde tüm metin kelimelere, hecelere, harflere veya araştırma için belirlenen kurallara göre bölünmektedir*" cümlesi ele

alınsın. Örnek cümle, kelime dizileri (ngram) yöntemiyle çözümlenerek aşağıdaki gibi listeler elde edilmiştir.

1-gram: {parçalara, ayırma, işleminde, tüm, metin, kelimelere, hecelere, harflere, veya, araştırma, için, belirlenen, kurallara, göre, bölünmektedir}.

2-gram: {parçalara ayırma, ayırma işleminde, işleminde tüm, tüm metin, metin kelimelere, kelimelere hecelere, hecelere harflere, harflere veya, veya araştırma, araştırma için, için belirlenen, belirlenen kurallara, kurallara göre, göre bölünmektedir}.

3-gram: {parçalara ayırma işleminde, ayırma işleminde tüm, işleminde tüm metin, tüm metin kelimelere, metin kelimelere hecelere, kelimelere hecelere harflere, hecelere harflere veya, harflere veya araştırma, veya araştırma için, araştırma için belirlenen, için belirlenen kurallara, belirlenen kurallara göre, kurallara göre bölünmektedir}.

Yukarıdaki verilen örnek listelerde, liste içerisindeki virgül ile ayrılmış her bir kelime ve kelime grubu, ilgili metnin bir özelliğini oluşturmaktadır.

Metin madenciliği bu bölümde bahsedilen teknikleri kapsamına alan ve günümüzde doğal dil işleme, sesi metne veya metni sese dönüştürme, farklı diller arası otomatik tercüme görevleri gibi işlevleri yerine getirmek için sıklıkla başvurulan bir veri madenciliği uygulama alanı olarak önemini giderek arttırmaktadır.

2.7. Veri Madenciliği

Kavramsal olarak 1960'lı yıllarda temelleri atılan veri madenciliği kavramı, problemleri çözmek için bilgisayarlı veri analizinin kullanılmaya başlaması birlikte ortaya çıkmıştır. O yıllarda bilgisayar yardımıyla yapılan işlemlerde yeterli bir süre geçtiğinde istenilen bilgilere ulaşmanın mümkün olacağı tespit edilmiş ve geleneksel yöntemler yerine analizin, geliştirilen bilgisayar modelleri tarafından yapılması amaçlanmıştır (Köktürk, Ankaralı ve Sümbüloğlu, 2009, s. 21). 1960'lardan beri gerçekleştirilen dijital formda veri toplama, bilgisayarlar vasıtasıyla geriye dönük veri

analizine olanak sağlamıştır. Bağlantılı (ilişkisel) veritabanları veri analizinde dinamikliği sağlayan yapılandırılmış programlama dilleriyle (SQL) birlikte 1970'lerde ortaya çıkmıştır. 1980'lere gelindiğinde ise veri ambarları verilerin depolanması için kullanılmaya başlanmıştır. Böylece büyük miktarlardaki veriler üzerine çalışılırken ortaya çıkan zorlukları aşabilmek için veri madenciliği kavramı doğmuştur (Hornick, Venkayala ve Marcade, 2007, s. 6). Veri madenciliğinde odaklanılan nokta; hipotez, analiz, yorum ve deneysel tasarım için kullanılan modelleri, araçları ve algoritmaları da kapsayan çözümler geliştirmektir (O'Byrne, 2013, s. 35).

Veri madenciliği kavramının kökeni; istatistik, matematik ve bilgisayar bilimleri gibi farklı alanlara dayanmaktadır. Gerçekte ham veriden değerli bilgi çıkarımı insanoğlunun yüzlerce yıldır devam ettirdiği bir iştir. Fakat günümüzde bilgisayar teknolojilerindeki gelişmelerle birlikte, büyük miktarlardaki veri her yeri kuşatmıştır. Oluşan verinin çoğu ticari ve bilimsel amaçlarla veritabanlarında depolanmaktadır (Irmak, Köksal ve Asilkan, 2012, s. 101). Veri madenciliği kavramının dayandığı köklerden biri olan istatistik bilimi, değişkenler arasındaki ilişkileri ve veri setindeki örüntüleri bulmak amacıyla veriyi inceleyip analiz eden, veri madenciliğinden çok daha eski yıllarda ortaya çıkmış ve veri analiz etmede yüzyıllar boyunca kullanılmıştır (Emre ve Erol, 2017, s. 162). İstatistik, veriden öğrenme, belirsizliği ölçme, kontrol etme bilimidir. Böylelikle sosyal ve bilimsel ilerlemenin yönünü belirlemede önemli bir araç olmuştur (Davidian ve Louis, 2012, s. 12). İstatistik ve veri madenciliği kavramsal olarak farklı olsa da; veriden öğrenme, verinin bilgiye dönüştürülmesi, veriyi analiz etme, verinin anlamını çözme, belirsizlikleri ortadan kaldırma, olayı etkileyen faktörleri belirleme ve öngörüde bulunma gibi ortak özellikleri bulunmaktadır. Genel olarak bakıldığında bu iki kavram için odaklanılan temel nokta, veriden öğrenme ve veriyi bilgiye dönüştürmektir (Emre ve Erol, 2017, s. 163).

Veri bilimiyle ilgili Fayyad, Shapiro ve Smyth tarafından 1996 yılında kaleme alınan "*From data mining to knowledge discovery in databases"* ve "*The KDD process for extracting useful knowledge from volumes of data*" başlıklı makalelerde, temel olarak veri madenciliği ve ilgili kavramların ortaya çıkmaya başladığından ilk kez bahsedilmiştir.

Yukarıdaki paragrafta anılan çalışmanın üzerinden 25 yıldan fazla bir süre geçmesine rağmen bu konular güncelliğini hâlâ korumakla beraber, bugünün araştırmacıları için büyük öneme sahiptir. Veri madenciliğinin bilinirliği ve kullanımı 1990'lardan itibaren yüksek ölçüde artmış ve bu kavram giderek daha önem kazanmıştır (Emre ve Erol, 2017, s. 162).

Veri madenciliği kavramı için kelimelerin yalın anlamlarına bakılırsa, yeryüzünün kıymetli kaynaklarının açığa çıkarılması süreci olan madencilik olgusunun, veri kelimesi ile ilişkilendirilmesi, veri yığınları içerisinde ilk bakışta fark edilemeyen kıymetli bilgilerin bulunması sürecine işaret ettiği çıkarımı kolayca yapılabilir (Irmak, Köksal ve Asilkan, 2012, s. 102). Han, Kamber ve Pei'ye göre (2012) veri madenciliği, büyük miktarda veriden bilginin çıkarılması veya "madenciliği" ifade eder; fakat bu terim aslında tam karşılığı sağlamayabilir. Örneğin kaya veya kum içerisinden altın madenciliği yapılıyor; fakat buna kaya veya kum madenciliği yerine "altın madenciliği" denilmektedir. Bu nedenle, veri madenciliği kavramının "veriden bilginin çıkartılması" olarak adlandırılmasının uygun olmasına rağmen bu terim biraz fazla uzun bulunmuştur. Bunun yanında başlarda bilgi çıkarma (knowledge extraction), veri/örüntü analizi (data/pattern analysis), veri arkeolojisi (data archaeology) ve veri tarama (data dredging) gibi terimler kullanılmasına rağmen, veri madenciliği terimi kısa zamanda popüler olmuş ve yaygın olarak kullanılmaya başlanmıştır (Han, Kamber ve Pei, 2012, s. 6).

Sohrabi ve Akbari 2016'ya göre veri madenciliği; potansiyel kullanışlı bilgiyi ve anlamlı modelleri tanımlama sürecidir. Bunlara ek olarak veri madenciliği kavramıyla ilgili literatürdeki diğer tanımlar da incelenmiştir.

Frawley'den (1991) aktaran Erdem ve Özdağoğlu'na (2008) göre; veri madenciliği, elde edilen veri kümesinin, önceden belirlenmiş bir algoritma vasıtasıyla analize tabi tutularak, karar vericilere referans olabilecek gizli bilgileri veya ilişkileri ortaya çıkarmak amacıyla geliştirilmiş uygulama alanıdır (Erdem ve Özdağoğlu, 2008, s. 262). Veri madenciliği, verideki örüntüleri keşfetmek ve analiz etmek için bilgisayar kullanılarak uygulanan çeşitli tekniklere verilen isimdir. Veri örüntüleri kullanılarak, model oluşturma, sınıflama, farklı grupları tanımlama ya da durum veya vakaları kümeleme işlemleri yapılır. Veri madenciliği günümüz dünyasında işletmelerde yaygın bir şekilde kullanılmaktadır. Öte yandan sosyal bilimler ve diğer araştırma alanlarında da yaygınlaşmaya başlamıştır (Attewell ve Monaghan, 2015, s. 3).

Veri madenciliği kavramı, verileri analiz ederek bilgiye ulaşma, bilgiyi bir madenden değerli bir varlık gibi gün yüzüne çıkarma sürecidir. Başka bir anlatımla, büyük miktarlardaki veri içerisinden tanımlama veya tahminleme işlevlerini oluşturabilecek örüntülerin, çeşitli yöntem ve teknikler vasıtasıyla aranmasıdır. Diğer bir ifadeyle, var olan ham veriden önceden tahmin edilemeyecek örüntü kalıplarını bularak potansiyel olarak kullanılabilir bilginin üretilmesidir. Özetle veri madenciliği; verilerin analiz edilerek veri içerisinde saklı olan ilişkilerin, değişimlerin, örüntülerin ve desenlerin keşfedilmesi ve kullanıma sunulmasıdır (Pektaş, 2013, s. 100).

Veri madenciliği, farklı veri türlerinden yararlı bilgi üretmek için metodolojik araştırma ve teknik araçların kullanıldığı bir uygulama alanıdır. Veri madenciliğinde; karar ağaçları, yapay sinir ağları, makine öğrenimi ve veri görselleştirme gibi çok sayıda ileri teknik kullanılmaktadır (Capri, 2015, s. 5). Örneğin 100 bireyin 2 farklı özellik verisinin incelenmesi, 10^6 bireyin 10^2 özelliğinin incelenmesinden çok farklı bir durumu

ifade etmektedir. İlk durumda basit veri analitik araçları, analizin sonunda önemli bilgileri ortaya çıkartabilir ve dağılım grafikleri yoluyla kullanışlı indikatörler sunulabilir. Fakat ikinci durumda, işlem çarpıcı biçimde değişmektedir. Bu durumda basit veri araçlarının çoğu etkinliklerini kaybeder; çünkü 10^6 hacmindeki bir dağılım grafiği tek bir biçimsiz çizgi halinde görünecektir. Ayrıca 10^2 sayısı kadar özellik, ortaya çıkan bu çizgilerden binlerce üretecektir ve anlamlı bir sonuç çıkmayacaktır. Dolayısıyla basit veri araçları ve uygulamaları bu tür işlemler için oldukça zayıf ve kullanışsızdır (Azzalini, Walton ve Scarpa, 2012, s. 4). Bu nedenlerle veri kümelerinde insan gözüyle farkedilmesi mümkün olmayan örüntü ve ilişkiler ancak veri madenciliği süreçleriyle veriye uygun teknik ve yöntemlerin kullanılmasıyla ortaya çıkarılabilmektedir.

2.7.1. Veri Madenciliğinin Amaçları

Veri madenciliği uygulamasının birincil hedefleri, tahmin etmek ve tanımlama yapmaktır. Tahmin modelleri, ilgili değişkenlerin bilinmeyen veya gelecekteki değerlerini tahmin etmek için veri setindeki bazı değişkenleri kullanmayı içermektedir. Tanımlama modelleri ise veriler üzerinde yorumlanabilen kalıpları bulmaya odaklanmaktadır (Fayyad, Shapiro ve Smyth, 1996, s. 44). Veri madenciliğinde temel olarak tahmin veya tanımlama amacıyla hesaplamalı bir model oluşturulmaktadır (Anderson, 2013, s. 1). Amaca ulaşmak için istatistik ve makine öğrenimi kavramları önemli bir yere sahiptir (Hand, 1998, s. 114).

Veri madenciliği, verileri değerli bilgilere dönüştürür. Analiz edilecek veriler, iyi organize edilmiş veritabanlarında bulunabilir veya yapılandırılmamış çeşitli veri kaynaklarından elde edilebilir. Bir veri madenciliği işleminin birçok aşaması bulunmaktadır. Genellikle bir istatistik veya makine öğrenme algoritması uygulanmadan ve uygun bir model geliştirilmeden önce kapsamlı bir veri hazırlık süreci (veriyi kullanıma hazırlama) yapılmaktadır. Sonraki aşamada veri madenciliği işlemlerini

destekleyebilecek özel yazılım paketleri (veri madenciliği araçları) ile analiz yapmaya başlanabilir (Tsiptsis ve Chorianopoulos, 2010, s. 3). *Data mining: Concepts, models, methods, and algorithms* (Kantardzic, 2011) adlı kaynak diğer kaynak kitaplara benzer şekilde, veri madenciliğinin amaçlarını tahminleme ve tanımlama olmak üzere iki gruba ayırmaktadır.

Tanımlamada, karar vermede yol gösterebilecek örüntüleri bulmak amaçlanır (Al-Shawi, 2011, s. 224). Ayrıca karar vermeye rehberlik etmede kullanılabilecek modellerin tanımlanması yapılmaktadır (Köktürk, Ankaralı ve Sümbüloğlu, 2009, s. 22). "Belirli bir yaş aralığındaki evli erkekler ile belirli bir yaş aralığından daha küçük yaşta olan bekar bayanların alışverişlerindeki satın alma örüntülerinin benzerlik gösterip göstermediğinin ortaya çıkartılması tanımlama modeline örnek olarak verilebilir" (Çelik, Akçetin ve Gök, 2017, s. 21). Kümeleme ve Birliktelik Kuralları yöntemleri tanımlayıcı modeller kapsamında yer almakta ve öncelikle verilerdeki davranışları tanımlamak için kullanmaktadır (Catarci, 2010, s. 43).

Bounsaytip ve Rinta'dan (2001) aktaran Silahtaroğlu'na (2016) göre, tanımlayıcı bir model daha çok veriler arasındaki gizli kalmış ilişkiyi ortaya çıkartır ve tipik olarak şöyle bir sonuç elde ederler: "Çocuk bezi alan bir müşterinin, mama alma olasılığı diğerlerinden 3 kat fazladır." Ya da "60.000 TL ile 90.000 TL arasında geliri olan aileler iki ya da daha fazla otomobili olan ailelere, çocuksuz ve 40.000 ile 60.000 TL arasında gelir olan ailelerden daha çok benzerler."

Bireyler sürekli olarak çevrelerini gözlemleyerek öğrenmektedir. Aslına bakılırsa dünyada üzerindeki her şeyin gerçek bir veritabanından başka bir şey olmadığını söylemek yanlış olmayacaktır. Tahminleyici modeller verileri inceleyerek veriler arasındaki temel faktörleri birbirleriyle ilişkilendirerek gözlemleri tanımaya, isimlendirmeye ve tıpkı bir çocuğun öğrenme sürecinde çevresinde gördüğü şeyleri sınıflamaya çalışması gibi sınıflamaya çalışır. Örneğin bir bebek için ilk zamanlarda

cinsiyet kavramı henüz oluşmamıştır. Bebek büyüdükçe anne- baba, hala, dayı gibi kendi çevresindekileri ve başka kız ve erkek çocuklarını görür. Bunların tamamı bebek için bir veritabanıdır. Bu veritabanını analiz eden bebek, kadın ve erkek arasındaki temel ayrımları keşfeder. Sonrasında karşısına önceden hiç görmediği bir kız çocuğu çıktığında, daha önceki deneyimine dayanarak bunun bir kız olduğuna karar verir. Temelde yaptığı sınıflamaya dayalı bir tahmindir (Silahtaroğlu, 2016, s. 50).

Tahmin modellerinde, gelecekteki eğilimleri tahmin etmek amaçlanır (Al-Shawi, 2011, s. 224). Eğer tahmin edilecek değişken sayısal bir değişkense regresyon problemi, kategorik ise sınıflama problemi adını alır (Köktürk, Ankaralı ve Sümbüloğlu, 2009, s. 22). Örneğin sınıflama yöntemleri tahmin amaçlı kullanılıp tahminleyici modeller arasında yer almaktadır (Catarci, 2010, s. 43).

Bounsaytip ve Rinta'dan (2001) aktaran Silahtaroğlu'na (2016) göre, Tahminleyici modeller şu sorunun yanıtını arar: "Bu işlemde dolandırıcılık var mıdır?" ya da "Bu müşteriden ne kadar gelir elde edilebilir?"

Hand (1998) "Data mining: statistics and more?" adlı makalesinde tahminleyici modellerin genellikle ticari veriler üzerinde kullanıldığını, bilimsel verilerin ise genellikle tanımlayıcı modellerde kullanıldığını vurgulamaktadır.

2.7.2. Veri Madenciliğinin Önemi

Veri madenciliği şirketlerin ve devletlerin işleyişinde giderek daha önemli hale gelmiştir. Veri madenciliği uygulamalarının sayısı ve çeşitliliği istikrarlı bir şekilde artmaktadır. Kendi süreçlerine veri madenciliği uygulamalarını erken kabul eden bazı iş alanları; bankacılık, sigorta, perakende ve telekomünikasyon sektörleridir. Daha yakın zamanlarda ise eczacılıkta, sağlıkta ve her çeşit e-ticarette yaygın hale gelmiştir (Soares, 2008, s. 1).

Gelişen internet ve dijital depolama teknolojileri, veri hacminin astronomik olarak büyümesini sağladığından, faydalı bilgilerin elde edilmesi için bu verilerin analiz edilmesine ihtiyaç vardır. Ancak, veri madenciliği uygulamaları sadece terabayt veriye sahip büyük kurumlar için geçerli değildir. Veri madenciliğinin faydaları büyük finansal kuruluşlardan yerel otomobil galerilerine, milyonlarca müşterisi olan şirketlerden yüzlerce müşterisi olanlara kadar, veri analizi gereksinimlerine sahip hem büyük hem de küçük organizasyonlar tarafından kullanılabilir (Hornick, Venkayala ve Marcade, 2007, s. 25). Mevcut durumdaki verinin büyüklüğü ve gelecekteki durumu her ne kadar gözü korkutsa da veri madenciliği uygulamaları sadece organizasyonlara kâr amaçlı katma değer yaratmak için değil, aynı zamanda, sağlık, şehircilik, eğitim, savunma sanayi, kültür ve siyaset gibi kamu alanlarında karar verme süreçlerinde kullanılmaktadır.

Veri madenciliği uygulamaları verileri kullanarak organizasyonların geçmişten ders çıkarmalarını da sağlayabilir. Örneğin bir fabrikada üretim sürecini negatif etkileyen geçmiş eğilimleri tanımlamak için kullanılabilir. Böylece aynı üretim süreci daha etkin ve verimli hale gelebilir (Bassan ve Sarkar, 2014, s. 19). Konuya günümüzün iş dünyası açısından bakıldığında oldukça rekabetçi bir ortamdan bahsetmek gerekir. Rekabet nedeniyle ürün karlılıkları ve müşteri memnuniyeti çok değişkendir. Tüketiciler artık muadil ürün ve hizmetler hakkında eskiden olmadığı kadar daha fazla bilgiye sahiptir. Bu nedenle müşteri sadakatleri, ya müşteri deneyimi pozitif kaldığı sürece ya da daha iyi bir alternatif ortaya çıkana kadar devam eder. Internet ile birlikte bireylerin bilgiye ulaşmasının kolaylaşması, şirketlerin ürün ve hizmetlerini geliştirmesini daha önemli kılmaktadır (Hornick, Venkayala ve Marcade, 2007, s. 4).

Son yıllardaki bilgi teknolojilerindeki gelişimle birlikte artık veriler üretildiği anda kullanılmaktadır. Bu gelişim, yeni üretilen veriler üzerinde hızla tepki verip veri daha akarken işleme almayı olanaklı hale getirmiştir (Doğan ve Arslantekin, 2016, s. 24).

Gelecek yıllarda 5G gibi teknolojilerin başlatılmasıyla birlikte sayısallaştırılmış (dijital) iletişimin hızının şimdikinden onlarca hatta yüzlerce kat daha hızlı olacağı hesap edilmektedir. Bu yüksek hız beraberinde veri büyüklüğünün beklenenden daha hızlı bir şekilde artacağını düşündürmektedir.

2.7.3. Veri Madenciliğinde Başarı

Bir veri madenciliği uygulamasının başarısı, büyük ölçüde analistin enerjisi, bilgisi ve yaratıcılığına bağlıdır. Temelde, veri madenciliği bir bulmacayı çözmek gibidir. Bulmacanın ayrı parçaları, kendi başlarına karmaşık yapılar değildir. Bununla birlikte, toplu biçimde bir bütün olarak ele alındıklarında çok ayrıntılı sistemler oluşturabilirler (Kantardzic, 2011, s. 15).

Veri madenciliği projesinin başarısındaki ana faktörler şu şekilde sıralanabilir: Doğru, önemli ve gerçekçi hedefler, toplanan verinin kalitesi, organizasyondaki analistlerin ve alan uzmanlarının iş birliği, uygun veri madenciliği tekniklerinin kullanımı, üretilen bilginin entegrasyonun doğru yapılması ve sonraki uygulamalar için veri madenciliği uygulamasından elde edilen deneyimin geri bildirim olarak analiz edilmesi (Tufféry, 2011, s. 23).

Veri analizinde, veri sağlamadaki alan/uzmanlık bilgisi en önemli aşamalardan biri olarak kabul edilmektedir (Rokach ve Maimon, 2008, s. 1). Bir veri madenciliği analisti veri kümesi ile bir model inşa eder ve sonuç olarak düşük bir tahminlemeye sahip olabilir. Aynı veri kümesiyle çalışan diğer bir analist ise bir model tasarlayarak daha yüksek bir tahminleme gücüne ulaşabilir. Peki aynı veri kümesiyle çalışan bu iki analistin sonuçları neden farklıdır? Çoğu durumda bu fark veri hazırlama aşamasından kaynaklanmaktadır. Böyle bir durumda büyük olasılıkla ikinci analist verilerin sağlandığı uygulama alanıyla ilgili daha çok bilgiye sahiptir. Bu bilgi, var olan değişken setinin daha iyi ve verimli kullanılacağı anlamına gelmektedir. Veri madenciliği uygulamalarında

alan/uzmanlık bilgisinin yerine geçebilecek başka bir şey yoktur. Genellikle veri madenciliği projelerinde ya alanı bilen bir uzman projeye dahil edilmekte ya da analist o alanla ilgili detaylı araştırmalar yapmaktadır (Nisbet, Miner ve Elder, 2009, s. 30). Projeye dahil edilen alan uzmanı ile analistin, sorunların çözümü için birlikte çalıştığı birçok proje, başarıya daha hızlı ulaşmaktadır (Ghani ve Soares, 2010, s. 2).

Yukarıda bahsedilen nedenlerle bir problem üzerinde çalışan veri analistinin ele aldığı konuyla ilgili alan bilgisine sahip olmasının önemli olduğu ortaya çıkmaktadır. Alan bilgisine sahip bir analist, ele aldığı problem özelinde daha derinlemesine incelemeler yapabilir. Ayrıca uzmanlık bilgisine sahip olmayan bir analiste göre alan bilgisine sahip bir veri analisti, veri madenciliğinin en önemli aşamalarından biri olan veriyi anlama ve problemi çözecek kurguyu oluşturma süreçlerinde daha iyi çıkarımlar yapabilmektedir. Bu nedenlerden dolayı veri madenciliği süreci disiplinlerarası bir çalışmayı ve işbiliğini gerektirir.

2.7.4. Veri Madenciliğinde Zorluklar

Dünya üzerinde var olan bilginin her 20 ayda bir ikiye katlandığı hesap edilmektedir (Köktürk, Ankaralı ve Sümbüloğlu, 2009, s. 21). Veri artış hızı göz önüne alındığında, veriden doğru çıkarımlarda bulunmak giderek zorlaşmaktadır. Bu nedenle analizlerin daha doğru yapılması ve analizlerden yararlı sonuçlar çıkarılabilmesi için farklı veri madenciliği teknikleri günümüzde de geliştirilmeye devam etmektedir.

Geliştirilen bu teknikler uzun zamandır organizasyonlarda önemli sorunları çözmek için kullanılmaktadır; fakat veri madenciliği uygulamalarının karşılaştığı en büyük zorluklarından biri de hangi tür problemlere kolayca uygulanabileceğinin iyi bilinmemesidir. Bir veri madenciliği çalışması çoğu zaman belirli bir algoritma ve onun sayısız özelliğine derinlemesine girmektedir. Ancak veri madenciliğinde başarı elde etmek için doğru algoritmayı seçmekten çok daha ötesine ihtiyaç vardır. Belirli bir

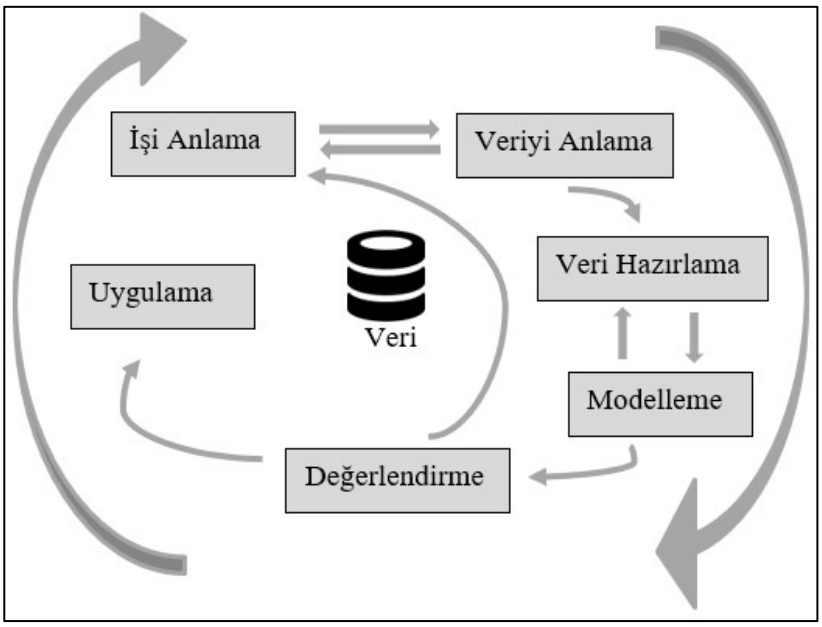
alandaki gereklilikleri ve bu sorunların çözümüne yönelik problemleri ve yaklaşım türlerini iyi anlamak ve belirlemek, veri madenciliği çalışmasındaki diğer önemli zorluklardır (Hornick, Venkayala ve Marcade, 2007, s. 25).

2.7.5. Veri Madenciliği Süreci

Veri madenciliği projelerinin yürütülmesini sağlayan yol haritaları olarak ifade edilebilecek en yaygın kullanılan iki yöntem bulunmaktadır: CRISP-DM ve SEMMA.

2.7.5.1. CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining / Veri Madenciliğinde Endüstriler Arası Standart İşleme) metodolojisi, veri madenciliği süreçlerini tanımlamak, belgelemek ve sürekli iyileştirmek için çok uluslu ve standartlara dayalı bir yaklaşımdır (De Ville, 2001, s. 37). CRISP-DM konsepti Daimler Chrysler, SPSS ve NCR tarafından 1996 yılında tasarlandı ve sonraki yıllar geliştirildi. CRISP-DM konseptinin amacı veri madenciliği projelerini daha hızlı ve daha ucuz bir şekilde gerçekleştirmektir. Bunun için veri madenciliği aşamalarını standartlaştırır, projeleri güvenilir ve yönetilebilir hale getirir (Hornick, Venkayala ve Marcade, 2007, s. 52).



Şekil 3. CRISP-DM Konseptinin Aşamaları

CRISP-DM yapısı veri madenciliği sürecinde altı farklı adımı tanımlar. CRISP-DM, Şekil 3'de görüleceği gibi; işi anlamak, veriyi anlamak, veri hazırlama, modelleme yapmak, değerlendirme ve uygulama aşamalarını barındırır (De Ville, 2001, s. 37). Bu bölümün devamında her bir madde ayrı ayrı açıklanmaya çalışılmıştır.

2.7.5.1.1. İşi Anlamak

İşi Anlamak (Business Understanding) olarak adlandırılan ilk aşamada, proje amaçlarını ve gereksinimlerini anlamaya ve daha sonra elde edilen bu bilgiyi veri madenciliği problem tanımına ve hedeflere ulaşmak için tasarlanmış bir ön plana dönüştürmeye odaklanılır (Chapman ve diğerleri, 2000, s. 10). İlk aşama diğerleri içerisinde en önemlisi ve çoğu zaman en zorlusudur. Problemi çözmek için neye ihtiyaç duyulduğunu net bir şekilde anlamadan ve sonuçların nasıl kullanılacağını daha işin başında bilmeden, beklentiler bulanık ve gerçekçi olmayacaktır. İlk aşamanın diğer bir yönü de mevcut insan, donanım, yazılım ve veri gibi kaynakları tanımlamayı içerir. Problem üzerinde çalışmak için hangi alanlardan ve teknik uzmanlardan yararlanılacağı, projeye uygun bilgisayar, yazılımlar ve ihtiyaç duyulan veri kaynaklarının nereden nasıl temin edileceği bu aşamada tespit edilir (Hornick, Venkayala ve Marcade, 2007, s. 53).

2.7.5.1.2. Veriyi Anlamak

Problemi ortaya koyduktan sonra, hangi verilerin mevcut olduğunu, verinin kalitesini ve problemi çözmeye uygun olup olmadığını belirlemek gerekir. Bu, Veriyi Anlama (Data Understanding) aşaması olarak ele alınır. Genellikle veriler daha iyi anlaşıldığında problem daha arıtılmış ve anlaşılır hale gelir. Verilerde eksiklik veya bozulma olabilir, bu tür veriler kirli olarak adlandırılır. Bu durum verilerin temizlenmesi ya da yeni verilerin temin edilmesi gibi yeni gereksinimler doğurabilir (Hornick, Venkayala ve Marcade, 2007, s. 55). Sonrasında verilere ilişkin ilk içgörüleri veya ilginç

altkümeleri keşfetme olanağı sağlayan etkinliklerle devam etmektedir (Chapman ve diğerleri, 2000, s. 10).

2.7.5.1.3. Veri Hazırlama

Üçüncü adım olan veri hazırlama aşaması, ham verilerden nihai veri kümesini oluşturmak için yapılması gereken tüm faaliyetleri kapsar ve bu faaliyetlerin amaca ulaşılması için birden çok defa yapılması muhtemeldir (Chapman ve diğerleri, 2000, s. 10). Veri hazırlamada (Data Preparation) amaç, ham veriden madencilik için uygun veri kümeleri üretmektir. Veri madenciliği sürecinde en başarılı sonuca ulaşmak için veri üzerinde çeşitli düzenleme, temizleme ve geliştirme bu aşamada yapılır (Hornick, Venkayala ve Marcade, 2007, s. 56).

2.7.5.1.4. Modelleme

Veri seti uygun bir şekilde hazırlandıktan sonra Modelleme (Modeling) aşaması başlar. Bu aşamada uygulayıcı, madencilik fonksiyonları için veri madenciliği araçları tarafından önceden tanımlanmış tüm ayarları özelleştirir, algoritmaları seçer ve modeli tasarlar. Belirli bir problem için birçok olası algoritma ve teknik bulunduğundan, uygulayıcılar hangisinin en iyi sonucu ürettiğini belirlemeye çalışır (Hornick, Venkayala ve Marcade, 2007, s. 57). Belirleme işlemi için veri formundaki özel gereksinimler değerlendirilir ve bu nedenle çoğu zaman bir önceki aşama olan veri hazırlama aşamasına dönmek gerekebilir (Chapman ve diğerleri, 2000, s. 10).

2.7.5.1.5. Değerlendirme

Değerlendirme (Evaluation) aşamasında, modelin iş hedeflerine uygun olup olmadığından emin olmak için değerlendirme ve gözden geçirme yapılmaktadır (Chapman ve diğerleri, 2000, s. 11). Bu aşamanın amacı, modelin uygulanıp uygulanmayacağına karar vermektir. Dolayısıyla bu aşama probleme getirilen çözümlerin

ihtiyaçları ne denli karşılayacağını değerlendiren, uzmanlık alan bilgisine dayanılan ve eleştirel düşünce içeren bir safhadır (Hornick, Venkayala ve Marcade, 2007, s. 58).

2.7.5.1.6. Uygulama

Uygulama (Deployment) aşamasında; kullanıcılar, bilgi işlem uzmanları ve müşteriler açısından veri madenciliği projesinin sonuçları değerlendirir ve bunun yanı sıra veriden çıkarılan bilgi, süreç ve madencilik deneyimine odaklanılır. Uygulama safhası; bir raporla veya uygulamanın bazı kısımlarında ya da iş sürecine entegre edilmiş bir çözümle neticelenebilir (Hornick, Venkayala ve Marcade, 2007, s. 59). Eğer uygun modelleri yaratmak için genel bir prosedür belirlenmişse, bu prosedür sonraki uygulamalarda kullanılmak için kayıt altına alınır (Chapman ve diğerleri, 2000, s. 28).

2.7.5.2. SEMMA

SEMMA süreci SAS Enstitüsü tarafından geliştirilen tescilli bir metodolojidir. SEMMA; Sample (Örnekleme), Explore (Keşfetme), Modify (Dönüştürme), Modelling (Modelleme) ve Assess (Değerlendireme) anlamına gelir (De Ville, 2001, s. 37). SEMMA temelde veri madenciliği projelerinin anlaşılmasını, organizasyonunu, geliştirilmesini ve sürdürülmesini sağlar (Shafique ve Qaiser, 2014, s. 220).

2.7.5.2.1. Örnekleme

İlk adım olan Örneklemede (Sample) veri setinin bir kısmı ayrılır (Makhabel, 2015, s. 36). Ayrılan kısım örnekleme için kullanılır; fakat veri dizisi örüntü çıkarımı için yeterli büyüklükte, yöntemlerin etkin bir şekilde kullanılmasını sağlayacak kadar da küçük olmalıdır (Akpınar, 2014, s. 76).

2.7.5.2.2. Keşfetme

Veri kümesini daha iyi anlamak için, bu adımda beklenmeyen eğilimler ve anomaliler tespit edilir (Makhabel, 2015, s. 36). Ayrıca Keşfetme (Explore) aşamasında

veri hakkında fikir kazanılarak keşif süreci daraltılabilir (Shafique ve Qaiser, 2014, s. 220).

2.7.5.2.3. Dönüştürme

Model inşa sürecine odaklanmak için değişkenler yaratılır, seçilir ve dönüştürülür (Makhabel, 2015, s. 36). Dönüştürme (Modify) aşaması temelde değişkenler üzerinde bahsedilen işlemlerde kullanılan yöntemleri içermektedir (Akpınar, 2014, s. 76).

2.7.5.2.4. Modelleme

Modelleme (Modelling) aşamasında veri setinin üzerinde çeşitli veri madenciliği teknikleri uygulanır (Akpınar, 2014, s. 76). Böylelikle istenilen sonucu elde etmek için olası modeller oluşturulur (Makhabel, 2015, s. 36).

2.7.5.2.5. Değerlendirme

Değerlendirme (Assess) SEMMA sürecinin beşinci ve son aşamasıdır. Bu aşamada bulguların güvenilirliği ve yararlılığının değerlendirilmesine ve performansa odaklanılmaktadır (Shafique ve Qaiser, 2014, s. 220). Temelde bu aşamada veri madenciliği sürecinden elde edilen bulgular değerlendirilir (Makhabel, 2015, s. 36).

Veri madenciliğinde uygulama sürecinin nasıl yürütüleceği ele alınan problemin özelliklerine göre belirlenmektedir. Problem için en uygun süreç yönetimi en başta belirlenir. Sonrasında süreç basamakları adım adım gerçekleştirilerek yeterli sonuç alınıncaya kadar adımlar döngüsel olarak işletilir. Bu tezin uygulama çalışmasında, dördüncü bölümde bahsedildiği üzere CRISP-DM yaklaşımı benimsenmiştir.

3. BÖLÜM: VERİ MADENCİLİĞİ YÖNTEM VE TEKNİKLERİ

Veri madenciliği yöntemleri; büyük veri yığınlarında gizli olan örüntüleri ve ilişkileri ortaya çıkartma amacıyla istatistiksel ve hesaplamalı algoritmaları ve yapay zeka algoritmalarını bir arada içermektedir (Dolgun, Özdemir ve Oğuz, 2009, s. 49). Veri madenciliğinde genellikle veri setinin tekil birimlerden oluştuğu varsayılmaktadır. Başka bir deyişle, çalışılan alana bağlı olarak bu birimler bir bankanın müşterilerinden, moleküler bileşiklere veya bir kütüphanedeki kitaplara kadar her şey olabilir. Her birim için veri setinde o birimi tanımlayan, örneğin bir banka hesap sahibi için isim, adres, hesap bilgileri gibi farklı karakteristikler ve tanımlayıcı bilgiler bulunmaktadır. Tanımlayıcı bilgiler değerlendirildiğinde ortak özellikler hesaba katılarak alt grup ve kümeler oluşturulabilir. Örneğin 18 yaşından küçük bireyler belirlenip bir alt grup oluşturulabilir. Bu alt grupların içerisinde onları birbirlerinden ayıran karakteristikler keşfedilebilir. Banka müşterileri örneğinden devam edilirse, "18 yaşın altındaki grup ortalamada negatif bir bakiyeye sahiptir" gibi çıkarımlar yapılabilir. Sonuç olarak bu da, bir müşterinin yaşı ile hesap bakiyesi arasında bir ilişki olduğunu ortaya koymaktadır. Veri madenciliği yöntemleri vasıtasıyla bu örnekteki gibi bilgi keşfine odaklanılmaktadır (Knobbe, 2006, s. 2).

Geleneksel olarak veri madenciliği yöntemleri üç ana başlık altında toplanmaktadır, bunlar; Sınıflama (Classification), Kümeleme (Clustering) ve Birliktelik Kuralları (Association Rules) yaklaşımlarıdır (Wankhade, Patka ve Thool, 2013; Özkan, 2016; Altunkaynak, 2017; Silahtaroğlu, 2016; Savaş, Topaloğlu ve Yılmaz, 2012; Köktürk, Ankaralı ve Sümbüloğlu, 2009; Akpınar, 2014). Öte yandan en çok kullanılan veri madenciliği yöntem ve teknikleri arasında regresyon ve makine öğrenimi uygulamaları da bulunmaktadır.

Tezin bu bölümü için literatürde binlerce yöntem ve algoritma bulunduğundan çalışmada hangi algoritmaların yer alacağını belirlemek için literatür taraması yapılmıştır.

Bu bağlamda Web of Science veritabanı kullanılarak veri madenciliği alanındaki kaynak kitaplar belirlenmeye çalışılmış ve kaynak kitaplarda çoğunlukla yer alan yöntem ve algoritmalar çalışmanın teori kısmına dahil edilmiştir. Bu çerçevede Web of Science veritabanının "Book Citation Index - Science (BKCI-S)" ve "Book Citation Index - Social Science & Humatities (BKCI-SSH)" dizinlerinde "data mining" konu alanı içerisinde yer alan, başlığında "data mining" söz dizimi geçen, dili İngilizce olan ve 2005-2019 aralığında yayınlanan kitap künyelerine ulaşılmıştır. Sorgu aşağıdaki şekilde formülize edilmiştir:

(TS=(data mining) AND TI=("data mining")) AND LANGUAGE: (English) Indexes=BKCI-S, BKCI-SSH Timespan=2005-2019

16.11.2019 tarihinde yapılan sorgu neticesinde 806 kayıt, bu zamana kadar aldığı toplam atıf sayılarına göre sıralanmış ve en fazla atıf alan 10 kaynak kitap seçilmiştir. Seçilen kaynak kitaplar gözden geçirilerek, hepsinde çoğunlukta olan veri madenciliği yöntem ve algoritmaları çalışmanın teori kısmına dahil edilmiştir. Atıf sayılarına göre seçilen ilk 10 kaynak şu şekildedir:

- Data Mining: Concepts and Techniques, 3rd Edition (Han, J.; Kamber, M.; Pei, J.) Aldığı atıf sayısı: 3500
- Data Mining: Practical Machine Learning Tools and Techniques, 3rd (Witten, I. H.; Frank, E.; Hall, M. A.) Aldığı atıf sayısı: 2305
- Data Mining and Knowledge Discovery Handbook, Second Edition (Maimon, O; Rokach, L) Aldığı atıf sayısı: 214
- Data Mining for Imbalanced Datasets: An Overview (Chawla, Nitesh V.) Aldığı atıf sayısı: 203
- Privacy-Preserving Data Mining: Models and Algorithms (Aggarwal, CC; Yu, PS) Aldığı atıf sayısı: 187

- Geographic Data Mining and Knowledge Discovery An Overview (Miller, Harvey J.; Han, Jiawei) Aldığı atıf sayısı: 171
- Top Ten Algorithms in Data Mining (Wu, X; Kumar, V) Aldığı atıf sayısı: 171
- Data Mining with Rattle and R: The Art of Excavationg Data for Knowledge Discovery (Williams, Graham) Aldığı atıf sayısı: 151
- Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Second Edition (Liu, B) Aldığı atıf sayısı: 151
- Data Preprocessing in Data Mining (Garcia, S; Luengo, J; Herrera, F) Aldığı atıf sayısı: 147

Veri madenciliği ve veri bilimi çok geniş kapsamlı ve birçok disiplini içerisine alan bir uygulama alanıdır. Her bir veri madenciliği yöntemi veya tekniği farklı ele alınış senaryoları, algoritmaları veya formülleriyle yüzlerce sayfa içerik ile açıklanabilir. Bu nedenle çalışmada veri madenciliği yöntem, teknik ve algoritmaları temel düzeyde ele alınmıştır.

3.1. Sınıflama

Sınıflama (Classification), en çok bilinen veri madenciliği yöntemlerinden biridir. Sınıflamada amaç sınıfları tahmin etmektedir; havanın bir sonraki gün nasıl olacağı ya da bir kutuda ne kadar mavi top olduğunun tahmin edilmesi bir sınıflama işlevidir (Silahtaroğlu, 2016, s. 67). Sınıflama en temel haliyle, bağımsız değişken değerleri belli iken bağımlı değişken (hedef değişken) sınıf değerini tahmin etmektir (Altunkaynak, 2017, s. 18). Sınıflama, bazı yaygın ilişkilere göre sınıf veya grupların içerisine örneklerin dağıtılması olarak düşünülebilir. Örneğin e-posta uygulamalarındaki spam (gereksiz ileti) filtresi, gelen iletinin uygun olup olmadığına karar veren bir sınıflandırıcıdır (Bifet, 2010, s. 13).

Sınıflama; kampanyaya katılım, müşteri segmentasyonu, kredi analizleri ve hasta sonuçlarını analiz etme gibi geniş bir aralıkta problemleri ele alır (Hornick, Venkayala ve Marcade, 2007, s. 88). Ayrıca risk yönetimi, resim ve örüntü tanıma, dolandırıcılık tespiti, kalite kontrol çalışmaları, reklamcılık ve pazarlama gibi geniş bir uygulama alanına sahiptir. Sınıflama temelde her bir durumu bir kategoriye atama işlemidir. Her durum, karakteristikleri belirlenen bir dizi özellik içerir. Sınıflamada, sınıfların ayrı ayrı özelliklerini ve karakteristiklerini tanımlayan bir model geliştirilir (MacLennan, Crivat ve Tang, 2009, s. 6).

Sınıflamada örüntüleri ortaya çıkarmak için belirli bir süreç izlenir. Öncelikle veri setinin bir kısmı kullanılarak sınıflama kurallarının oluşturulması yani eğitim aşaması ve sonrasında test etme süreci uygulanır. Daha sonra elde edilen kurallar tüm veriye uygulanarak sonuçlar analiz edilir (Özkan, 2016, s. 32).

Sınıflamanın matematiksel olarak tanımı şöyledir: $D = \{t_1, t_2, \ldots t_n\}$ bir veri kümesi olsun ve her bir t_i , i=1,…,n bir gözlemi temsil etsin. $C = \{C_1, C_2, \ldots C_m\}$ ise m adet sınıftan oluşan sınıfları temsil etsin. Bu durumda $f: D \rightarrow C$ ve her bir t_i bir sınıfa dahil olur. Artık her C_j ayrı bir sınıftır ve her sınıf kendisine ait gözlemleri içermektedir. Başka bir ifadeyle, $C_j = \{ t_i / f(t_i) = C_j, 1 \leq i \leq n, ve\ t_i \in D \}$'dir (Silahtaroğlu, 2016, s. 67).

Sınıflamada, örnek olarak verilen Tablo 1'deki gelir düzeyi gibi üç sınıfa veya kategoriye ayrılabilecek bir hedef kategorik değişken bulunmaktadır. Bunlar; yüksek gelir, orta gelir ve düşük gelirdir. Sınıflama veri madenciliği modelleri, her biri hedef değişken hakkında bilgi ve ayrıca bir girdi veya tahmin değişkenleri seti içeren büyük bir kayıt kümesini incelemektedir. Araştırmacının, belirtilen veri setinde bulunmayan, yeni bireylerin gelir grubunu, yaş, cinsiyet ve meslek gibi o bireyle ilişkili diğer özelliklerine

göre sınıflamak istediği varsayılmaktadır. Bu, veri madenciliği yöntem ve tekniklerine uygun bir sınıflama örneğidir (Larose, 2005, s. 10).

Tablo 1. Sınıflama Yöntemi Örneği

Yaş	**Cinsiyet**	**İş**	**Gelir Düzeyi**
47	K	Yazılım Mühendisi	Yüksek
28	E	Pazarlama Danışmanı	Orta
35	E	İşsiz	Düşük
…	…	…	…

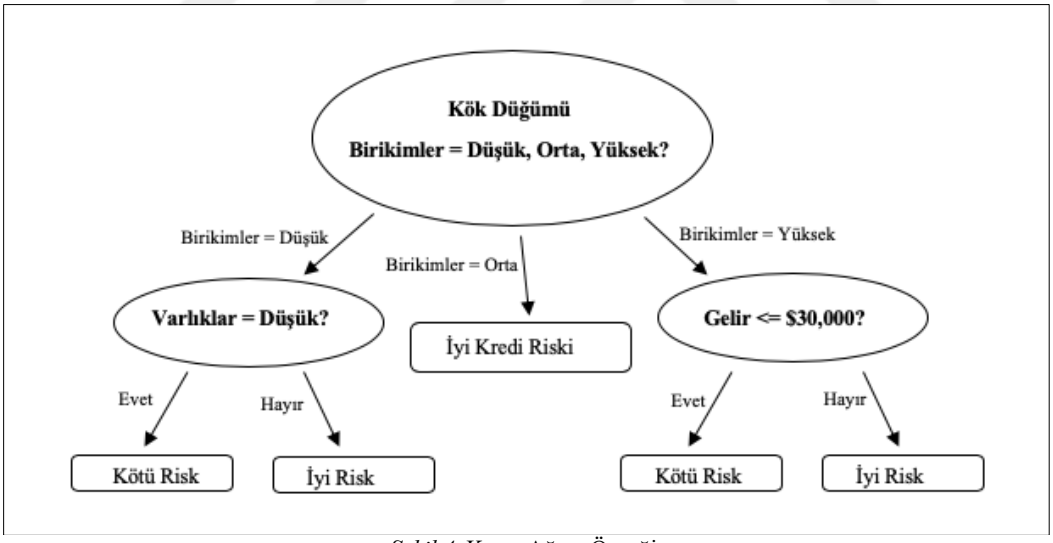
Bu örnekteki sınıflama algoritmasının kabaca şu şekilde ilerlemesi gerekmektedir. İlk olarak hem tahmin edici değişkenleri hem de (zaten sınıflandırılmış) hedef değişkeni (gelir düzeyini) içeren veri kümesi analiz edilir. Bu şekilde, algoritma hangi değişken kombinasyonlarının hangi gelir grubuyla ilişkili olduğunu öğrenmeye çalışır. Örneğin, yaşlı kadınlar, yüksek gelir grubuyla ilişkili olabilir. Bu veri kümesine eğitim kümesi denmektedir. Ardından algoritma, gelir grubuyla ilgili bilgi bulunmayan yeni kayıtlara (test veri setine) bakmaktadır. Eğitim setindeki sınıflamalara ve öğrenilmiş bilgiye dayanarak, algoritma yeni kayıtlara sınıflamalar atar ve gelir düzeylerini tahmin etmeye çalışır. Örneğin, 60 yaşında bir kadın profesör, yüksek gelir grubuna dahil olabilir (Larose, 2005, s. 10). Bu örnek temel bir sınıflama problemidir, farklı problemlere göre uygulanabilecek çok sayıda sınıflama algoritması bulunmaktadır.

3.1.1. Karar Ağaçları

Karar almada kullanılan en klasik yöntemlerden biri olan karar ağaçlarında (decision tree), verilen bir kararın önceden belirlenmiş olasılıklar dahilinde, başka karar noktalarını tetiklediği ardışık karar verme süreçlerini hesaplamak amaçlanır (Akpınar, 2014, s. 200). Karar ağaçları sınıflandırıcı algoritmalardır. En basit versiyonlarında, ağaçtaki her iç düğüm bir değişkenin özelliği üzerinde bir test içerir, bir düğümden gelen her dal testin olası bir sonucuna karşılık gelir ve her yaprak bir sınıf tahmini

barındırmaktadır (Bifet, 2010, s. 14). Başka bir deyişle, bir karar ağacı, ağaç benzeri bir karar grafiği ve bunun olası çıktıları olarak da tanımlanabilir. Bu ağaçtaki en yüksek (ilk) düğüm, kararın verildiği kök düğümdür. Ağacın dallarındaki her düğümde, giriş veya öznitelik değişkenleri üzerinde test yapılmaktadır. Düğümü izleyen her dal testin sonucuna gitmektedir (Hooman ve diğerleri, 2016, s. 375).

Diğer tanımlamalara bakılacak olunursa, karar ağaçları yapısı gereği dallarla birbirlerine bağlanmış, kök düğümden yaprak düğümlere kadar uzanan bir karar düğümleri topluluğudur. Karar ağacı diyagramının en üstüne yerleştirilen kök düğümünden başlayarak, tüm nitelikler karar düğümlerinde test edilir ve her bir çıktı bir dalla sonuçlanmaktadır. Her dal daha sonra başka bir karar düğümüne veya sonlandırıcı bir yaprak düğümüne yol açmaktadır. Şekil 4'te basit bir karar ağacının örneği sunulmaktadır (Larose, 2005, s. 165).



Şekil 4. Karar Ağacı Örneği

Şekil 4'te Larose (2005) tarafından oluşturulan karar ağacı için hedeflenen değişken, kredi riskidir. Bu değişken, potansiyel müşterilerin iyi veya kötü kredi riski taşıyıp taşımadığını ifade etmekte ve müşterilerin iyi veya kötü kredi riskine sahip olup olmadığını anlamak için sınıflandırılmaktadır. Tahmin değişkeni olan "birikimler", düşük, orta ve yüksek olarak, alt dallardaki değişkenlerden ilki "varlıklar", düşük veya

düşük değil olarak ve diğer değişken ise gelirin yıllık $30.000 altında olup olmamasını ifade etmektedir (Larose, 2005, s. 166).

Örnek detaylı incelendiğinde; en üstte yer alan Birikimler adlı kök düğümü üç farklı değişkene ayrılmıştır ve üç farklı dal oluşmuştur; düşük, orta ve yüksek. Sonrasında birikim açısından düşük değere sahip olanlar sınıfında varlıkların düşük olup olmadığı sorgulanmaktadır. Eğer müşterinin varlıkları düşükse bu potansiyel müşteriler "kötü risk" sınıfına dahil olmaktadır. Diğer yandan aynı grup içerisinde varlıklar düşük değilse, bu gruptakiler "iyi risk" olarak derecelendirilmektedir. Birikimleri orta seviyede olanlar "iyi kredi riski" sınıfına girmektedir. Birikim seviyeleri yüksek olan son grup ise, yıllık gelir "$30.000 altı mı değil mi?" açısından irdelenmektedir. $30.000'a eşit ve altında olanlar "kötü risk", üstünde olanlar ise "iyi risk" gurubuna girmektedir. Bu şekilde bir analizle bir banka müşterileri için risk gruplarını belirleyebilmekte ve çıkan sonuçlara göre en iyi kararları alabilmektedirler.

Bu örnekteki gibi veri setlerinin küçük olduğu durumlarda bu şekilde kullanımlar mümkündür; fakat veri setlerinin çok daha büyük olduğu uygulamalarda bu yaklaşımı elle uygulamalı gerçekleştirmek mümkün değildir. Bu durumlarda makine öğrenimi kapsamında geliştirilen karar ağacı öğrenimi algoritmaları kullanılması uygun olabilir. Bu yöntemler verinin büyük miktarlarda olduğu uygulamalarda kullanılmak üzere tasarlanmıştır (Akpınar, 2014, s. 204). Karar ağaçlarında yaygın kullanılan ve birbirleriyle aralarında farklılıklar olsa da ciddi benzerlikler bulunan çok sayıda algoritma bulunmaktadır. Bunlar arasında en çok kullanılan algoritmaların temel karar ağacı özelliklerinden ayrılan yönleri kısaca aşağıda ele alınmıştır.

Veri madenciliği süreçlerinde sıklıkla kullanılan ID3 yönteminde, bölen özniteliğin seçiminde entropi kullanılmaktadır. Daha öncesinde bölen olarak uygulanmamış öznitelikler ayırt edilerek ağacın yapraklarına doğru homojenleşmesi amaçlanır (Akpınar, 2014, s. 220). ID3'te herhangi bir budama prosedürü uygulanmaz ve

sayısal öznitelikler veya eksik değerler işlenmez (Rokach ve Maimon, 2008, s. 71). ID3 enformasyon teorisi yaklaşımına dayalı bir algoritmadır (Mannan, 2012, s. 2234). ID3'ün uzantısı olarak kabul edilen C4.5 algoritmasında ise en iyi bölen öznitelik seçiminde entropi ve enformasyon kazancı teknikleri kullanılır (Akpınar, 2014, s. 221). Bir başka algoritma olan C4.5 algoritmasında sayısal öznitelikler işlenebilir ve örnek sayısı belirli bir eşiğin altında olduğunda bölme sona erer. (Rokach ve Maimon, 2008, s. 71). CHAID (CHi-squared Automatic Interaction Detection / Ki-kare Otomatik Etkileşim Çıkarma) algoritmasının ise ilk uygulamaları tıbbi araştırmalar için olsa da özellikle pazarlama alanında tüketici gruplarının seçiminde kullanılmıştır. CHAID; ikiden daha fazla dala ayrılabilmesi, bütün ölçü skalaları ile çalışılabilir olması ve gözlemlere ağırlık ve frekans değerlerinin atanabilmesi gibi önemli avantajlar sunar. Ayrıca gürültülü veya kirli verilerle de çalışılabilmesi önemli esnekliğidir (Akpınar, 2014, s. 223). CHAID verideki eksik değerleri geçerli bir kategori olarak ele alır ve bu algoritmada budama gerçekleşmez (Rokach ve Maimon, 2008, s. 72). CART (Classification and Regression Trees / Sınıflama ve Regresyon Ağaçları) algoritması klasik karar ağacı öğrenme algoritmaları içerisinde önemli bir yere sahiptir (Akpınar, 2014, s. 225). İsmi sınıflama ve regresyon problemlerinden ileri gelen CART ağaçları, sınıflamaya çok benzerdir; fakat sınıf etiketleri yerine düğümlere gerçek değerler atanmıştır (Grabczewski, 2014, s. 16). Kök düğümden başlayarak her seviyede iki alt diziye ayrılan ve ikili olarak büyüyen CART algoritmasının amacı, sonraki düzeylerde daha homojen hale getirilmiş düğümleri elde edebilmektir. CART algoritması derinliğe fazla büyüme eğilimi gösterir ve bölen değişkenlerinin birden fazla kullanılmasına da izin verildiği için sonuçlar karar ağacı üzerinde verimli bir şekilde izlenemeyebilir (Akpınar, 2014, s. 225). CART'ın önemli bir özelliği regresyon ağaçları üretebilmesidir. Regresyon ağaçlarında yapraklar, bir sınıf değil gerçek bir rakamı tahmin eder. Ayrıca olasılık dağılımı yapılabilmesini sağlamaktadır (Rokach ve Maimon, 2008, s. 72). QUEST (Quick, Unbiased, Efficient

Statistical Tree / Hızlı, Tarafsız, Verimli İstatistik Ağacı) ise diğer benzer yöntemlerin çok kategorili bağımsız değişkenlerde yaşadığı hataları önlemeyi ve hesaplamaların daha hızlı yapılmasını amaçlayan bir karar ağacı öğrenimi algoritmasıdır (Akpınar, 2014, s. 227). QUEST sıklıkla istatistiksel ağaçlar olarak da tanımlanabilir; çünkü ağacı oluşturma ve geliştirmede istatistiksel araçlara güçlü bir şekilde bağlıdırlar (Grabczewski, 2014, s. 23).

3.1.2. K-En Yakın Komşu

K-en yakın komşu algoritması, sınıfları belli olan bir örnek kümedeki değerlerden yararlanarak, yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılır. Bu yöntem, örnek kümedeki gözlemlerin her birinin, belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en yakın gözlemin seçilmesi esasına dayanmaktadır. Uzaklıkların hesaplanmasında, i ve j noktaları için aşağıdaki Öklid uzaklık formülü kullanılabilir (Özkan, 2016, s. 141). Bu yöntemde kullanılan birçok mesafe ölçüm formülü bulunmaktadır. Öklid uzaklık ölçüsü en yaygın kullanılanlardan biridir (Silahtaroğlu, 2016, s. 119).

$$d(i,j) = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$$

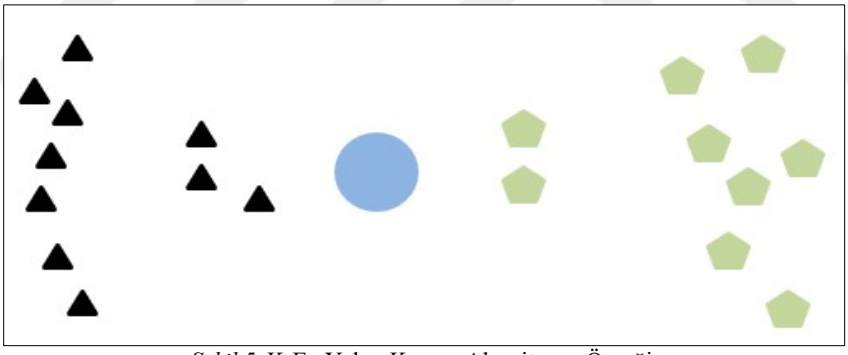
K-en yakın komşu algoritmasında aşağıdaki adımlar uygulanır:

- Verilen bir noktaya en yakın komşuların sayısı olan k parametresi belirlenir.
- Bu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanır.
- Hesaplanan uzaklıklar sıralanır ve en küçük olan k tanesi seçilir.
- Seçilenlerin hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir.

- Seçilen kategori, tahmin edilecek gözlem değerinin kategorisi olarak kabul edilir (Özkan, 2016, s. 142).

Beyer'dan (1999) aktaran Silahtaroğlu'na göre, bu yöntem coğrafi bilgi sistemlerinde yaygın olarak kullanılmaktadır. Belirlenen bir noktaya en yakın şehir, istasyon vs. belirlenmesi aslında k-en yakın komşu algoritmasının temelini oluşturur (Silahtaroğlu, 2016, s. 118).

Örneğin, Şekil 5'teki ortada yer alan yuvarlak simge, hangi tarafa daha yakındır? Yanıt, verilen ***k*** değerine göre değişecektir. ***k***=2 olarak belirlendiğinde ortadaki yuvarlak nesne, sağ taraftaki beşgen nesnelere daha yakındır. Çünkü sınıflandırılan simgeye en yakın iki nesne beşgendir. ***k***=3 için ise yanıt sol taraftaki üçgen nesneler olacaktır (Silahtaroğlu, 2016, s. 118). Özetle, ele alınan probleme göre ***k*** değeri sonucu değiştirmektedir.



Şekil 5. K-En Yakın Komşu Algoritması Örneği

Örnekle açıklanacak olunursa; Tablo 2'deki veriler kullanılarak X = (8,4) vektörüne ilişkin Y ataması k-en yakın komşu algoritmasına göre k=3 alınarak uygulanmak isteniyor. Uzaklık Öklid formülüne göre hesaplanacaktır (Altunkaynak, 2017, s. 111).

Tablo 2. K-En Yakın Komşu Algoritması Veri Tablosu

i	X_{1i}	X_{2i}	Y_i
1	2	4	kötü
2	3	6	iyi
3	3	4	iyi
4	4	10	kötü
5	5	8	kötü
6	6	3	iyi
7	7	9	iyi
8	9	7	kötü
9	11	7	kötü
10	10	2	kötü

Başlangıçta X_0 = (8,4) vektörünün diğer gözlemlere olan uzaklıklarının hesaplanması gerekir. X_0'ın ilk gözleme olan uzaklığı; $d_{01} = \sqrt{(2-8)^2-(4-4)^2}$ = 6.00 şeklinde bulunur. Diğer gözlem noktaları için de hesaplamalar yapıldığında uzaklıklar ve sıralama Tablo 3'te gösterildiği gibi olacaktır (Altunkaynak, 2017, s. 112).

Tablo 3. K-En Yakın Komşu Algoritması Sıralama Tablosu

i	X_{1i}	X_{2i}	Y_i	d_{0i}	Sıra
1	2	4	kötü	6.00	9
2	3	6	iyi	5.39	8
3	3	4	iyi	5.00	6
4	4	10	kötü	7.21	10
5	5	8	kötü	5.00	5
6	6	3	iyi	2.24	1
7	7	9	iyi	5.10	7
8	9	7	kötü	3.16	3
9	11	7	kötü	4.24	4
10	10	2	kötü	2.83	2

Tablo 3'te görüleceği gibi en küçük 6, 8 ve 10 olmak üzere 3 uzaklığa ait gözlemlerden 8. ve 10. gözlemlerin sınıf değeri "kötü", 1 tanesinin yani 6. gözlemin sınıf değeri ise "iyi"dir. Sonuç olarak X_0 = (8,4) vektörünün sınıf değeri "kötü" olarak belirlenir (Altunkaynak, 2017, s. 112). K-en yakın komşu algoritmasının kavramsal olarak basit olmasına rağmen, gerçek hayatta sayısız uygulamada çok iyi performanslar gösterebilir. Örneğin bir bireyin yürüme ve atlama gibi günlük yaşam aktivitelerinin, biyo-sinyaller ile ayrım yapılarak hesaplanabilmesi ve düşmenin otomatik olarak algılanması şekilde farklı birçok uygulama bu algoritma üzerine kurulabilir (Liu ve diğerleri, 2016, s. 15).

3.1.3. Yapay Sinir Ağları

Yapay sinir ağları (artificial neural network), insan beyni fonksiyonlarını taklit etmek için inşa edilmiş doğrusal olmayan bir istatistiksel model olarak tanımlanabilir. Belirsiz veri ilişkilerini modellemek için güçlü bir araçtır (Hooman ve diğerleri, 2016, s. 376). Bir yapay sinir ağı, belirli yollar vasıtasıyla insan zihni gibi davranan bir örüntü sınıflayıcısıdır. Yapay sinir ağları, belirli örüntüleri ve büyük veri setlerindeki ilişkileri ortaya çıkartabilir (Cox, 2005, s. 51).

Yapay sinir ağları uygulamaları bilgisayar bilimleri, matematik ve biyolojinin birleştiği yüksek derecede disiplinler arası bir alandır. Deneysel verilerin analizinde kilit rol oynayan bu alan; nöronların ve sinir sisteminin matematiksel modelinin çıkartılması ve ayrıca geleneksel bilgisayarlarda bilgi işleme sürecinin sıralı ve algoritmik yöntemi ile biyolojik sistemlerdeki bilgi işleme sürecinin paralel ve birleştirici yolu arasındaki tutarsızlıkların sonucu olarak ortaya çıkmıştır (Baerns ve Holeňa, 2009, s. 80).

Yapay sinir ağları veri analizinde tespit, tanımlama, örüntüleri sınıflama ve kanıt bulma açısından son 20 yıldır kullanılan muhtemelen en güçlü araçlardan biridir. Her ne kadar yapay sinir ağları 1950'lerin sonlarında tasarlansa da 1980'lerin sonlarına kadar uygun yazılımların geliştirilememesi ve bilgisayarların güçlü işlemcilere sahip olmamasından dolayı kullanılamadı. 1990'lı yıllara girildiğinde ise ticari yapay sinir ağ araçları ve uygulamaları geliştirildi ve finans, pazarlama, tıp ve üretim alanlarında kullanımı yaygınlaştı. İlk ve belki de en başarılı uygulama alanlarından biri ise kredi kartı dolandırıcılıklarını tespit etmek oldu. Günümüzde ise yapay sinir ağları giderek artan karmaşık gerçek dünya problemlerine uygulanmaktadır (Mena, 2003, s. 9).

Yapay sinir ağları; genelleme yeteneği, sağlamlık, açık bir problem tanımının bulunmaması, büyük miktarda veriyi idare edebilme, daha az istatistiksel varsayımlara

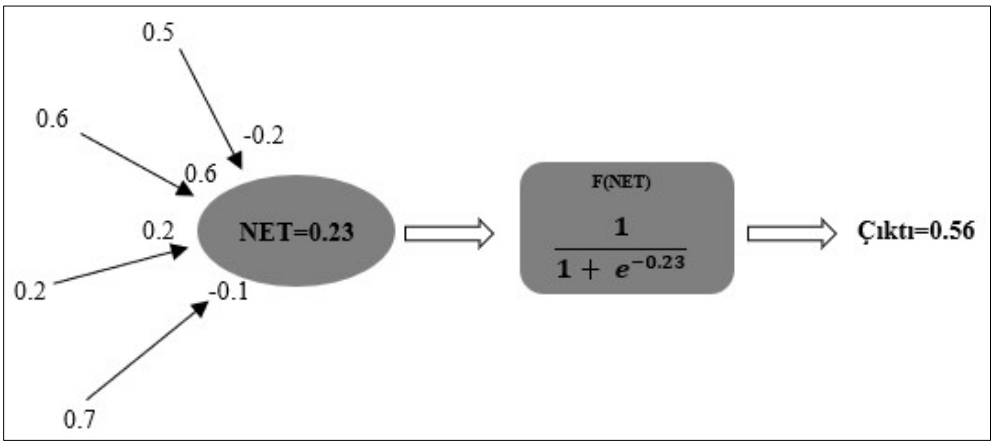
ihtiyaç duyulması gibi özellikleri nedeniyle tercih edilmektedir (Hooman ve diğerleri, 2016, s. 376):

Yapay sinir ağları, yapay sinir hücrelerinden oluşur. Hücrelerin birbirlerine bağlaması yapay sinir ağını oluşturmaktadır. Yapay sinir hücrelerinin 5 temel elemanı bulunmakladır. Girdiler, bir hücreye dış dünyadan, başka hücrelerden veya kendi kendisinden gelen bilgilerdir. Ağırlıklar, hücreye gelen bilginin önemini ve hücre üzerindeki etkisini gösterir. Toplama fonksiyonu, bu fonksiyon ağırlıklı toplamı bularak bir hücreye gelen net girdiyi hesaplar. Bunun için farklı fonksiyonlar kullanılsa da en yaygın olanı ağırlıklı toplamı hesaplamaktır. Aktivasyon fonksiyonu, bu fonksiyonda hücreye gelen net girdi işlenerek hücrenin bu girdiye karşılık üreteceği çıktı belirlenir. Diğer fonksiyonda olduğu gibi aktivasyon fonksiyonunda da çıktıyı hesaplamak için Lineer, Step, Sinus gibi farklı formüller kullanılır ve en uygun fonksiyon, denemeler yapılarak bulunabilir. $F(NET) = \frac{1}{1+ e^{-\mathrm{NET}}}$ formülü ile gösterilen Sigmoid fonksiyonu en yaygın kullanılanıdır. NET ile hücreye gelen NET girdi değeri ifade edilir ve toplama fonksiyonu kullanılarak hesaplama yapılır. Hücrenin çıktısı ise aktivasyon fonksiyonu tarafından hesaplanan çıktı değeridir (Öztemel, 2016, s. 48).

Örnekle açıklanacak olunursa; bir hücreye gelen bilgiler ve ağırlıklar Şekil 6'da verildiği gibidir. Görüleceği gibi 4 girdisi ve 4 ağırlık değeri vardır. Hücreye gelen NET bilgi, ağırlıklı toplam hesaplanarak şu şekilde alınır:

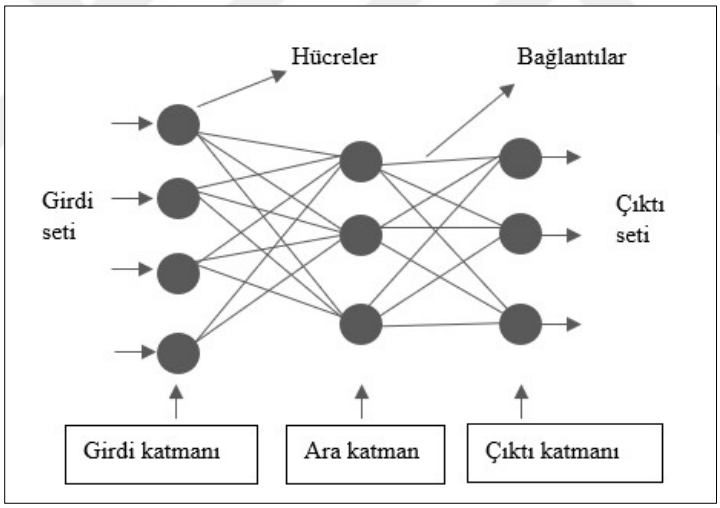
NET = 0.5*(-0.2) + 0.6*0.6 + 0.2*0.2 + 0.7*(-0.1)
NET =-0.1 + 0.36 + 0.04 – 0.07
NET = 0.23

Hücrenin çıktısı (Ç) sigmoid fonksiyonuna göre hesaplandığında; Ç = 1/ (1+$e^{-0.23}$)'den Ç = 0.56 bulunur. Ağın üzerindeki hücre çıktılarının hesaplanması sonucu Şekil 6'da gösterildiği gibi girdilere karşılık çıktıların nasıl üretildiği görülebilir (Öztemel, 2016, s. 52).



Şekil 6. Bir Yapay Sinir Hücresinin Çalışma Örneği

Yapay sinir hücreleri bir araya gelerek yapay sinir ağını oluştururlar. Sinir hücrelerinin bir araya gelmesi rastgele olmaz. Genel olarak hücreler Şekil 7'de gösterildiği gibi üç katman halinde ve her katman içinde paralel olarak bir araya gelerek ağı oluştururlar (Öztemel, 2016, s. 52).



Şekil 7. Bir Yapay Sinir Ağı Örneği

Girdi katmanı, bu katmandaki hücreler dış dünyadan bilgileri alarak ara katmanlara transfer etmekle sorumludurlar. Bazı ağlarda girdi katmanında herhangi bir bilgi işleme olmaz. Ara katmanlar, girdi katmanından gelen bilgiler ara katmanlarda işlenerek çıktı katmanına gönderilir. Bir ağ için birden fazla ara katman olabilir. Çıktı katmanı, bu katmandaki hücreler ara katmandan gelen bilgileri işleyerek ağın girdi katmanından sunulan girdi seti (örnek) için üretmesi gereken çıktıyı üretirler. Üretilen çıktı dış dünyaya gönderilir (Öztemel, 2016, s. 52).

Yapay sinir ağlarının çalışma mantığı şu şekilde özetlenebilir. Yapay sinir hücreleri dışarıdan gelen bilgileri bir toplama fonksiyonu ile toplar ve aktivasyon fonksiyonundan geçirerek çıktıyı üretip ağın bağlantıları, yani ağırlık değerleri üzerinden diğer hücrelere gönderir. Hücreler birbirlerine paralel olarak girdi, ara ve çıktı olmak üzere üç katman halinde bir araya gelerek bir ağ oluştururlar. Girdi katmanından elde edilen bilgiler ağın ağırlık değerleri kullanılarak ara katmanlarda işlenerek çıktı katmanına gönderilirler. Ağın doğru çıktıları üretebilmesi için ağın eğitilmesi ve ağırlıkların doğru değerlere sahip olması gerekir. Başlangıçta rastgele atanan değerler daha sonra eğitim sırasında her örnek ağa gösterildiğinde ağın öğrenme kurallarına göre değiştirilir. Bir sonraki aşamada başka bir örnek ağa gösterilerek ağırlıklar yine değiştirilir ve bu durum ağ eğitim setindeki örneklerin tamamı için doğru çıktılar üretinceye kadar tekrarlanır. İşlem tamamlandıktan sonra test setindeki örnekler ağa gösterilerek doğru cevaplar verilip verilmediği kontrol edilir, eğer doğru cevaplar alınıyorsa ağ eğitilmiş kabul edilmektedir. Tüm kararlar ağırlıklar kullanılarak verilir ve bundan dolayı ağın zekasının bu ağırlıklarda saklandığı söylenebilir (Öztemel, 2016, s. 57).

Regresyon gibi diğer geleneksel istatistiksel yöntemlerden farklı olarak yapay sinir ağları tahmin modelini oluştururken görece küçük eğitim örnek veri setleriyle çalışabilecek yeteneğe sahiptirler. Bu, yapay sinir ağlarıyla ilgili anahtar konsepttir. Yapay sinirlerin tıpkı bir çocuk veya evcil hayvan gibi eğitilmeleri gereklidir. Eğitim, örnek veri setlerinin bir yapay sinir ağ algoritması tarafından işlenmesi anlamına gelir. Çoğu durumda yapay ağ algoritması, örüntüler öğrenilene kadar binlerce kez işlemleri geriye dönüp tekrarlar. Böylece benzer profiller veya davranışlar keşfedilir ve büyük miktardaki verilere uygulanarak bir sınıflama yapılabilir (Mena, 2003, s. 10).

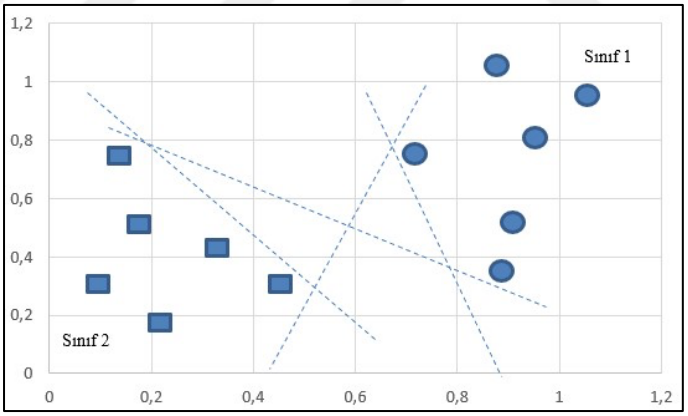
Veri madenciliğinde yapay sinir ağları, içerisinde bulunduğumuz dönemde çok sayıda uygulama için kullanılmaktadır. Sosyal medyaya eklenen fotoğraflardaki yüzleri

tanıma işlemlerinden, arama motorlarına ve çoğu yapay zeka uygulamalarına kadar geniş bir yelpazede başarılı kullanımları bulunmaktadır.

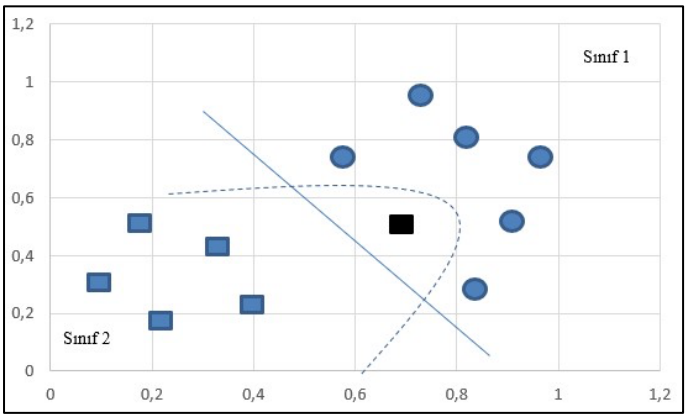
3.1.4. Destek Vektör Makineleri

Veri madenciliğinde sınıflama problemlerinde kullanılan bir yöntem olan destek vektör makineleri (support vector machines), sınıflamayı doğrusal ya da doğrusal olmayan bir fonksiyon kullanılarak yerine getirmektedir (Özkan, 2016, s. 169). Destek vektör makineleri genel olarak; el yazısı tanıma, zaman serisi analizi, konuşma tanıma ve kanser teşhisi gibi birçok alanda uygulanmaktadır (Akpınar, 2014, s. 268).

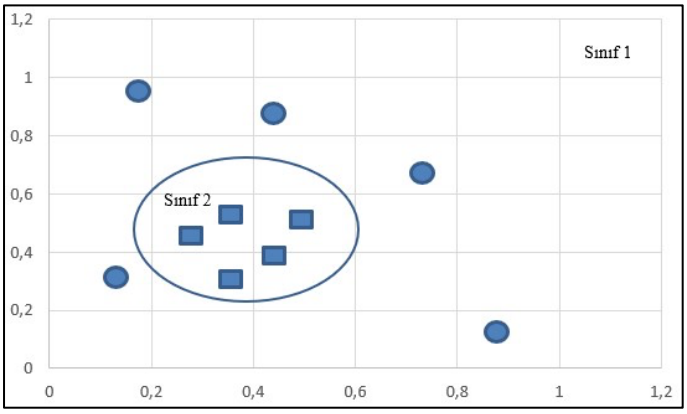
Destek vektör makineleri veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanmaktadır. Destek vektör makineleri yöntemi ile iki boyutlu düzlemde doğrusal ve doğrusal olmayan sınıflama, Şekil 8, 9 ve 10'daki gibi yapılmaktadır (Özkan, 2016, s. 169).



Şekil 8. Doğrusal Olarak Ayrılabilen Veri



Şekil 9. Doğrusal Olarak Ayrılamayan Veri



Şekil 10. Doğrusal Olmayan Sınıflama

Yukarıdaki örneklerde görülebileceği gibi Sınıf 1 ve Sınıf 2 olarak belirlenen iki sınıflı bir öğrenme görevinde amaç, eğitim setindeki iki sınıf üyeleri arasında ayrım yapmak için en iyi sınıflamayı bulmaktadır. En iyi sınıflama fonksiyonu için ölçüm geometrik olarak gerçekleştirilebilir (Wu ve diğerleri, 2008, s. 10).

Şekil 8'de gösterilen örnekte iki sınıfa ayrılmak istenen veri kümesi doğrusal bir fonksiyon çizilerek ayrılabilmektedir. Şekil 9'da iki sınıfın doğrusal bir fonksiyonla ayrılamadığı ve sınıf öğelerinin iç içe geçtiği örnek verilmiştir. Şekil 10'da ise doğrusal olmayan fakat sınıflama yapılabilen bir örnek gösterilmektedir. Bunun gibi veri kümeleri için destek vektör makineleri kullanılarak verilerin sınıflanması mümkün olmaktadır.

3.1.5. Bayesyen Sınıflama

Bayesyen sınıflama (bayes classifiers) tekniği, elde var olan hali hazırda sınıflanmış verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine girme olasılığını hesaplayan bir yöntemdir. Bayesyen kuralına dayalı geliştirilmiş algoritma ve sınıflama teknikleri bu adla anılmaktadır (Silahtaroğlu, 2016, s. 97). Bayesyen sınıflamasında öğrenme süreci, açık bir kural bulmak yerine olasılıkları tahmin eder. Bu yaklaşımın avantajı, veri seti büyük olduğunda ve belirli olasılıkları tahmin etme yöntemi tutarlı olduğunda uygulayıcının en az hataya ulaşmasıdır. Bayesyen metodolojisi tıp, genetik ve işletme gibi birçok farklı alanda geniş bir uygulama alanına sahiptir. Çıktısının basitliği ve doğal yorumu onu birçok uygulama için uygun kılar, ancak bazı durumlarda

dezavantajları vardır, çünkü çoğu durumda Bayesyen modelleri belirli veri türleri için naif veya çok basit olma eğilimindedir (Browne ve Berry, 2006, s. 16).

Bayesyen, basitliği ve düşük işlem maliyeti ile bilinen bir sınıflama algoritmasıdır (Bifet, 2010, s. 13). Bu yöntem, metin dokümanlarını sınıflamak için bilinen en başarılı algoritmalar arasındadır. Verilen dokümanda her bir kategorinin olasılıklarını tahmin etmek için eğitim setini kullanır (Motoda, 2002, s. 5). Bayes teoremi şu şekilde ifade edilebilir:

$$P(C_1 \mid x_i) = \frac{P(x_i \mid C_1)P(C_1)}{P(x_i \mid C_1)P(h_1) + P(x_i \mid C_2)P(C_2)}$$

C_1 ve C_2 olarak iki ayrı sınıfın olduğu kabul edilmiştir. $P(C_1 \mid x_i)$ x_i 'nin C_1sınıfında olma olasılığını ifade etmektedir. $P(x_i)$, x_i değerinin veri setindeki bulunma olasılığıdır. $P(C_1)$ ve $P(C_2)$ ise C_1 ve C_2 sınıflarının veri setinde bulunma olasılığını ifade etmektedir. Eğer *m* adet sınıf (hipotez) olduğu düşünülürse bu durumda kural;

$$P(x_i) = \sum_{j=1}^{m} P(x_i \mid C_j)P(C_j)$$

şeklinde olacak ve x_i 'nin C_1 sınıfında olma olasılığı aşağıdaki bağıntı ile hesaplanacaktır:

$$P(C_1 \mid x_i) = \frac{P(x_i \mid C_1)P(C_1)}{P(x_i)}$$

Algoritma ilk olarak öğrenme kümesinde her sınıfın öğrenme kümesi içinde bulunma olasılığını, yani $P(C_j)$ değerini hesaplar. Daha sonra, x_i'ler hesaplanarak $P(x_i)$ değeri bulunur. Benzer şekilde her bir sınıfta, her bir x_i değerinin olasılığı $P(x_i \mid C_j)$, C_j'ler içinde x_i lerin sayılmasıyla elde edilir (Silahtaroğlu, 2016, s. 98).

Tablo 4. Bayesyen Sınıflama Örnek Veri Seti

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Orta	Yaşlı	Erkek	Evet
2	İlk	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Orta	Orta	Erkek	Evet
5	İlk	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlk	Genç	Kadın	Hayır
8	Orta	Orta	Kadın	Evet

Örnekle açıklanacak olunursa; Tablo 4'teki eğitim verilerini kullanarak, eğitimi "Yüksek" (x_1), yaşı "Orta" (x_2) ve cinsiyeti "Erkek" (x_3) olan bir örneğin kabul durumunun "Evet" veya "Hayır" olarak hangi sınıfa ait olduğu belirlenmek isteniyor.

(C_1) : Kabul = Evet
(C_2) : Kabul = Hayır

Olmak üzere $P(\mathrm{X}\,|\,C_1)P(C_1)$ ve $P(X\,|\,C_2)P(C_2)$ ifadelerinin hesaplanması gerekmektedir.

a) $P(\mathrm{X}\,|\,C_1)P(C_1)$ için;

$$P(X_1\,|\,C_1) = P(Eğitim\ =\ Yüksek\ |\ \mathrm{Kabul}\ =\ \mathrm{Evet}) = \frac{1}{5}$$

→ *Açıklama: "Evet" sınıfındaki eğitimi "Yüksek" olanların oranı.*

$$P(X_2\,|\,C_1) = P(Yaş\ =\ Orta\,|\ \mathrm{Kabul}\ =\ \mathrm{Evet}) = \frac{3}{5}$$

→ *Açıklama: "Evet" sınıfındaki Yaşı "orta" olanların oranı.*

$$P(X_3\,|\,C_1) = P(Cinsiyet\ =\ Erkek\ |\ \mathrm{Kabul}\ =\ \mathrm{Evet}) = \frac{3}{5}$$

→ *Açıklama: "Evet" sınıfındaki Cinsiyeti "erkek" olanların oranı.*

$$P(\mathrm{X}\,|\,C_1) = P(X\ |\ \mathrm{Kabul}\ =\ \mathrm{Evet}) = \left(\frac{1}{5}\right)\left(\frac{3}{5}\right)\left(\frac{3}{5}\right) = \frac{9}{125}$$

$P(\mathrm{Kabul}\ =\ \mathrm{Evet})$ olasılığı ise şu şekilde hesaplanır:

$$P(C_1) = P(\mathrm{Kabul}\ =\ \mathrm{Evet}) = \frac{5}{8}$$

→ *Açıklama: Tüm veriler içerisindeki "Evet" sınıfına dahil olanların oranı.*

Sonuç olarak;

$$P(\mathrm{X}\,|\,C_1)P(C_1) = P(\mathrm{X}\,|\,\mathrm{Kabul} = \mathrm{Evet})P(\mathrm{Kabul} = \mathrm{Evet})$$

$= \left(\frac{9}{125}\right)\left(\frac{5}{8}\right) = 0{,}045$ elde edilmiş olur.

b) $P(\mathrm{X}\,|\,C_2)P(C_2)$ için;

$$P(X_1\,|\,C_2) = P(Eğitim = Yüksek\,|\,\mathrm{Kabul} = \mathrm{Hayır}) = \frac{1}{3}$$

→ *Açıklama: "Hayır" sınıfındaki eğitimi "Yüksek" olanların oranı.*

$$P(X_2\,|\,C_2) = P(Yaş = Orta\,|\,\mathrm{Kabul} = \mathrm{Hayır}) = \frac{1}{3}$$

→ *Açıklama: "Hayır" sınıfındaki Yaşı "orta" olanların oranı.*

$$P(X_3\,|\,C_2) = P(Cinsiyet = Erkek\,|\,\mathrm{Kabul} = \mathrm{Hayır}) = \frac{1}{3}$$

→ *Açıklama: "Hayır" sınıfındaki Cinsiyeti "erkek" olanların oranı.*

$$P(\mathrm{X}\,|\,C_2) = P(X\,|\,\mathrm{Kabul} = \mathrm{Hayır}) = \left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right) = \frac{1}{27}$$

$P(\mathrm{Kabul} = \mathrm{Hayır})$ olasılığı ise şu şekilde hesaplanır:

$$P(C_2) = P(\mathrm{Kabul} = \mathrm{Hayır}) = \frac{3}{8}$$

→ *Açıklama: Tüm veriler içerisindeki "Hayır" sınıfına dahil olanların oranı.*

Sonuç olarak;

$$P(\mathrm{X}\,|\,C_2)P(C_2) = P(\mathrm{X}\,|\,\mathrm{Kabul} = \mathrm{Hayır})P(\mathrm{Kabul} = \mathrm{Hayır})$$

$= \left(\frac{1}{27}\right)\left(\frac{3}{8}\right) = 0{,}014$ elde edilmiş olur.

O halde büyük olan değer örneğin sınıfını verecektir. C_1 = 0,045 ve C_2 = 0,014 ise örneğin sınıfı C_1 yani "Evet" olacaktır (Özkan, 2016, s. 160).

Yukarıda verilen örnekte veri kümesi içerisinde önceden tanımlanmamış, eğitimi "Yüksek", yaşı "Orta" ve cinsiyeti "Erkek" olan bir örneğin kabul durumu tespit edilmeye çalışılmıştır. Bayesyen sınıflama tekniği ile yapılan hesaplamalar sonucu verilen örneğin kabul durumu "Evet" olarak hesaplanmıştır. Bu örnekte görülebileceği üzere veri madenciliği amaçlarına uygun olarak var olan verilerin işlenip örüntülerin keşfedilmesiyle, veri kümesinde olmayan bir örneğe ait bir tahmin yapılarak örneğin hangi sınıfa girmesi gerektiği belirlenmiştir.

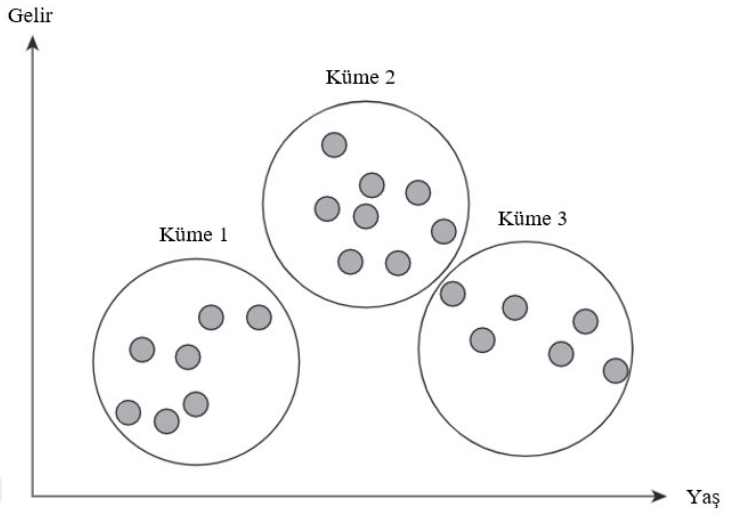
3.2. Kümeleme

Kümeleme (Clustering) uygulamaları, veriyi anlamlı altsınıflara gruplayan veri madenciliği tekniğidir (Catlett, 2013, s. 193). Kümeleme, bir veri seti içerisindeki benzerlikleri ayırmaya ve çoklu kümeler arasındaki ilişkileri bulmaya çalışır (Cox, 2005, s. 47). Kümeleme algoritmalarının amacı, bir obje veya noktalar kümesi olarak verilen veri setini, mesafe veya olasılık ölçümlemeleriyle gruplara bölmektir. Aynı gruptaki üyeler mesafe, benzerlik veya diğer ölçümler açısından birbirlerine daha yakındır. Diğer bir deyişle, küme içi benzerlik en üst düzeye çıkartılırken, küme dışı benzerlikler en aza indirgenmeye çalışılır (Makhabel, 2015, s. 176).

İlk bakışta kümeleme problemleri sınıflama ile aynı görülebilir. Ancak ikisi arasındaki temel fark kümelemenin denetimsiz doğasıdır. Sınıflama, özel bir girdi eğitim seti veya nesnelerin yerleştirildiği önceden tanımlanmış bir dizi sınıf içerirken, kümeleme veriden anlamlı sınıflar türetmeye çalışır (Browne ve Berry, 2006, s. 99). Kümeleme; gen ve protein analizinde, ürün gruplandırmasında, cinse göre tasnif etmede ve metin madenciliğinde kullanılmaktadır (Hornick, Venkayala ve Marcade, 2007, s. 97). Ayrıca uydu görüntülerini parçalara bölme, gürültü filtreleme, aykırı değer tespiti ve biyoinformatik verilerin kümelenmesi gibi çeşitli bilimsel alanlarda da kullanılmaktadır (Catlett, 2013, s. 193). Bu uygulamaların dışında kümeleme yaygın olarak verilerin hazırlanmasında da kullanılmaktadır. Özellikle günümüzde doğal dil işleme ve metin madenciliği gibi uygulamalar için kümeleme yöntemleriyle veriler hazırlanmaktadır.

Bir örnekle açıklanacak olunursa; Şekil 11'de "Yaş" ve "Gelir" olarak tanımlanmış iki farklı karakteristik (özellik) bulunmaktadır. Kümeleme algoritmaları, veri kümesini bu iki karakteristiği temel alan üç bölüme ayırmıştır. Küme 1, düşük gelirli genç bireylerden oluşmaktadır. Küme 2, yüksek gelire sahip orta yaştaki müşterileri içermektedir. Küme 3 ise, nispeten düşük geliri olan bir grup yaşlı bireydir (MacLennan,

Crivat ve Tang, 2009, s. 6). Bu örnekte bir kümeleme algoritması kullanılarak veri kümesinin üç farklı gruba ayrılması gösterilmektedir.



Şekil 11. Kümeleme Örneği

Kümeleme yönteminin temel özellikleri şu şekilde sıralanabilir: Ortaya çıkan kümeler çoğu durumda nitelikli kümeler değildir. Çoğu ölçüm mesafe bazlı yapıldığından kümelerin şekli küreseldir. Her grup bir ortalamayı ifade eder ve her bölüm bir kümeyi temsil eder. Kümeler en küçükten en büyük veri setlerine kadar uygulanabilir (Makhabel, 2015, s. 179).

Son 20 yıla kadar kümeleme algoritmalarını hiyerarşik ve bölümleyici olarak iki başlık altında sınıflamak mümkün olmaktaydı; fakat gelişen bilgisayar teknolojileri ve büyüyen veri hacmi sonucunda bu sınıflama son derece yetersiz kalmıştır. Son dönemde; yoğunluk temelli, ızgara temelli, olasılık temelli, grafik bazlı ve alt uzay gibi başlıklar altında toplanabilecek 100'ün üzerinde çeşitli kümeleme algoritması geliştirilmiştir (Akpınar, 2014, s. 297). Çalışmada hepsinin listelenmesi mümkün olmadığı için yaygın kullanılan kümeleme algoritmaları incelenmiştir.

3.2.1. K-Ortalamalar Algoritması

K-Ortalamalar (k-means) algoritması gözlem noktalarını belirli sayıda kümeye ayırmada kullanılan bir kümeleme yöntemidir. K-Ortalamalar yöntemi ile her gözlem

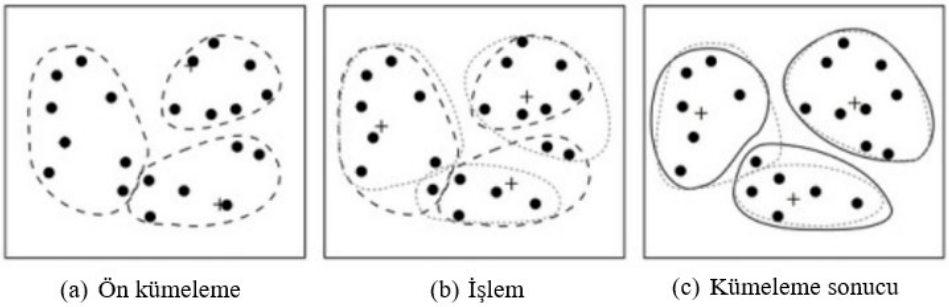
noktasının bir kümeye yakınlığı veya uzaklığı belirlenir ve noktalar en yakın küme merkezine atanır. Her atama işleminden sonra küme merkezleri yeninden hesaplanmaktadır. Gözlemler arasındaki uzaklık hesaplamalarında yaygın kullanılan uzaklık hesaplama yöntemi aşağıda verilen Öklid Uzaklık formülüdür (Şentürk, 2006, s. 41). Bunun dışında uzaklık hesaplamalarında kullanılan; Minkowski Uzaklığı, Manhattan Uzaklığı, Mahalanobis D^2 Uzaklığı ve Pearson Uzaklık ölçüsü gibi yöntemler bulunmaktadır.

$$d(i,j) = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$$

K-Ortalamalar algoritması büyük veri setleriyle çalışmak için etkilidir. K-Ortalama algoritması, verinin en özel yollarda bölündüğü, bir objenin yalnızca bir grup ya da kümeye ait olabileceği kümeleme yönteminin bir çeşididir. Algoritma temelde her seferinde optimum değeri ve en uygun çözümleri bulmaya çalışır (Makhabel, 2015, s. 179). Bunun için sürekli olarak kümeleri yeniler ve en uygun çözüme ulaşılana kadar döngüsel olarak bu süreç devam eder (Silahtaroğlu, 2016, s. 186). Veri madenciliğinde en çok kullanılan kümeleme algoritmalarından biri olan bu algoritma, hiyerarşik olmayan kümeleme yöntemi sınıfına dahildir (Ching ve Kwok-Po Ng, 2003, s. 27). K-Ortalama algoritmasının uygulama adımları şu şekilde özetlenebilir:

1. *k* değeri belirlenir → veri kaç kümeye bölünmek isteniyorsa bu değer (*k*) başlangıçta analist tarafından belirlenir.
2. *k* adet gözlem birer küme olarak seçilir.
3. Geriye kalan *k* gözlem ile *n-k* gözlem arasındaki uzaklıklar hesaplanır.
4. *n-k* gözlem sırayla en yakın kümeye atanır ve her atamadan sonra küme merkezi yeniden hesaplanır.

5. *n* tane gözlem tekrar küme merkezlerine göre uzaklıkları hesaplanarak atama işlemi baştan yapılır.
6. Aynı sonuç üst üste iki kez elde edilene kadar 5. adım tekrarlanır ve sonuca ulaşılır (Altunkaynak, 2017, s. 140).



Şekil 12. K-Ortalama Algoritması

Algoritma adımları Şekil 12'de açıklanmaya çalışılırsa; 1 ve 2. adımlar "Ön Kümeleme" safhasını, 3, 4 ve 5. adımlar "İşlem" aşamasını ve 6. adım ise "Kümeleme Sonucunu" ifade etmektedir (Makhabel, 2015, s. 179).

3.2.2. K-Medoids Algoritması

K-medoids algoritması, k-ortalama algoritmasının istisnai verilerde oluşan zayıf yanlarını gidermek amacıyla geliştirilmiştir. Algoritma, k-ortalama algoritması gibi küme elemanlarının ortalamasını almak yerine, kümenin temsil edileceği noktayı bulmak için kümenin en merkezi noktasındaki elemanı küme merkezi olarak atamaktadır (Çelik, Akçetin ve Gök, 2017, s. 153). Kümedeki diğer elemanları temsil eden bu elemana medoid adı verilmektedir (Akpınar, 2014, s. 319). Medoid elemanını veya gözlemini tespit etmek için önce küme merkezi hesaplanır daha sonra merkez ile diğer değerler arasındaki uzaklık Öklid uzaklık ölçeğiyle hesaplanır ve merkeze en yakın eleman medoid olarak seçilir (Altunkaynak, 2017, s. 128). Böylece veri setindeki sıra dışı veya gürültülü elemanlara karşı hassaslık azaltılmaya çalışılır (Akpınar, 2014, s. 319).

Örnek ile açıklanacak olunursa; Tablo 5'te verilen gözlemler k-medoids yöntemine göre kümelere ayrılacaktır (Altunkaynak, 2017, s. 144).

Tablo 5. K-Medoids Algoritması Örneği

i	X_{1i}	X_{2i}
1	4	10
2	10	2
3	3	4
4	2	4
5	5	8
6	6	3
7	7	9
8	9	7
9	11	7
10	3	6

"Analist tarafından 1. ve 2. gözlem birer küme (medoid) olarak Tablo 6'daki gibi alınmaktadır.

Tablo 6. K-Medoids Kümeleme-1

Küme 1		**Küme 2**	
Gözlem	Medoid	Gözlem	Medoid
1	$\binom{4}{10}$	2	$\binom{10}{2}$

3. gözlemin Küme 1 ve Küme 2'ye olan uzaklıkları Öklid uzaklık formülü ile hesaplanırsa;

Küme 1 için; $\sqrt{(3-4)^2+(4-10)^2} = 6.08$

Küme 2 için; $\sqrt{(3-10)^2+(4-2)^2} = 7.28$

Bu durumda 3. gözlem Küme 1'e atanır. Yeni küme yapısı Tablo 7'deki gibi olacaktır.

Tablo 7. K-Medoids Kümeleme-2

Küme 1		**Küme 2**	
Gözlem	Medoid	Gözlem	Medoid
1,3	$medoid\left[\binom{4}{10},\binom{3}{4}\right] = \binom{4}{10}$	2	$\binom{10}{2}$

Küme 1'de 2 gözlem olduğundan herhangi biri medoid olarak alınabilir. 4. gözlemin kümelere olan uzaklıkları Öklid uzaklık formülü ile hesaplanırsa;

Küme 1 için; $\sqrt{(2-4)^2+(4-10)^2}$ = 6.32

Küme 2 için; $\sqrt{(2-10)^2+(4-2)^2}$ = 8.25

Bu durumda 4. gözlem Küme 1'e atanır. Yeni küme yapısı Tablo 8'deki gibi olacaktır.

Tablo 8. K-Medoids Kümeleme-3

Küme 1		**Küme 2**	
Gözlem	Merkez	Gözlem	Merkez
1,3,4	$merkez = \frac{1}{3}\left[\binom{4}{10}+\binom{3}{4}+\binom{2}{4}\right] = \binom{3}{6}$ $medoid\left[\binom{4}{10},\binom{3}{4},\binom{2}{4}\right] = \binom{3}{4}$	2	$\binom{10}{2}$

Küme 1'in ortalaması $\binom{3}{6}$'dır. Bu noktaya en yakın gözlem $\binom{3}{4}$ vektörüne sahip gözlem olduğundan bu gözlem medoid olarak seçilir. Bu şekilde hesaplamalar her gözlem değeri için yapıldığında sonuç olarak Tablo 9'da verildiği gibi küme yapısına ulaşılmaktadır" (Altunkaynak, 2017, s. 144).

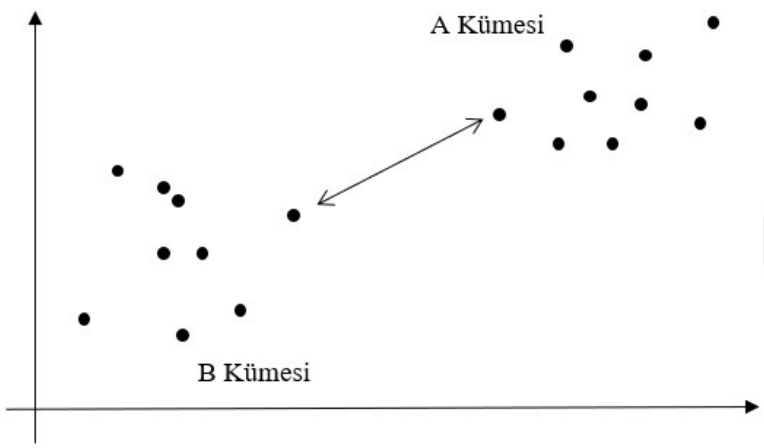
Tablo 9. K-Medoids Kümeleme-4

Küme 1		**Küme 2**	
Gözlem	Merkez	Gözlem	Medoid
1,3,4,5,7,10	$\binom{4.00}{6.83}$	2,6,8,9	$\binom{9.00}{4.75}$

K-medoids algoritmasına göre işlemler adım adım yapılarak nihai değerlere ulaşılmaktadır. Yukarıdaki örnekte veriyi başlangıçta belirlenen iki kümeye ayırma girdisine göre, 10 gözlemden oluşan veri, iki kümeye ayrılmıştır. 1, 3, 4, 5, 7 ve 10. gözlemler Küme 1 olarak gruplanırken, 2, 6, 8 ve 9. gözlemler Küme 2 olarak gruplandırılmıştır.

3.2.3. En Yakın Komşu Algoritması

Tekli bağlantı kümeleme yöntemi olarak da adlandırılan en yakın komşu (nearest neighbors) algoritmasında, başlangıçta tüm gözlem değerlerinin her biri bir küme olarak değerlendirilir. Sonrasında bu kümeler birleştirilerek yeni kümeler elde edilir. En yakın komşu algoritması uygulanırken ilk aşamada gözlemler arasındaki mesafeler, önceki bölümlerde bahsedilen Öklid Uzaklık formülü ile hesaplanır. Uzaklıklar göz önüne alınarak her seferinde en yakın, yani en kısa mesafedeki kümeler belirlenir ve birleştirilerek yeni kümeler oluşturulur. İşlemler tekrar edilerek her seferinde yeni oluşan kümeler üzerinden hesaplama yapılarak istenen sonuca ulaşılır (Özkan, 2016, s. 197).



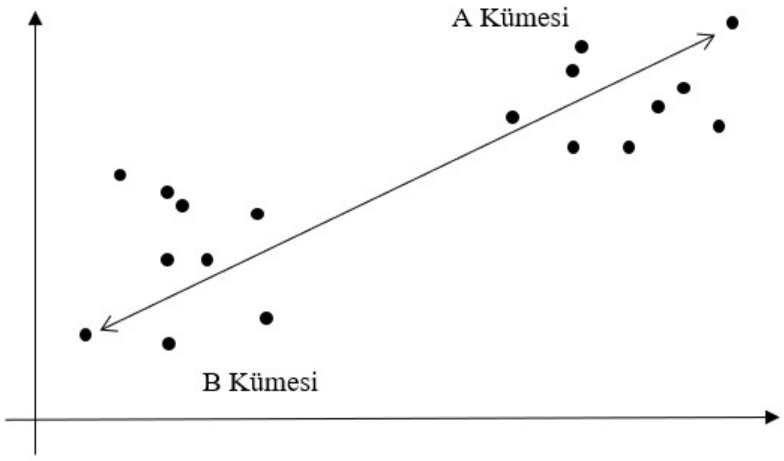
Şekil 13. En Yakın Komşu Algoritması

En yakın komşu algoritmasında iki kümenin birbirine en yakın elemanları arasındaki Şekil 13'te ok işareti ile gösterilen uzaklık, A ve B kümesinin birbirine olan uzaklığı olarak değerlendirilmektedir (Özkan, 2016, s. 197).

Yukarıda sözü edilen örnekte en yakın komşu algoritmasında bir elemanın en yakınındaki başka bir elemanla küme oluşturması uygulanmış ve veri, Şekil 13'te görülebileceği üzere A Kümesi ve B Kümesi olarak ikiye ayrılmıştır.

3.2.4. En Uzak Komşu Algoritması

Tam bağlantı kümeleme yöntemi olarak da adlandırılan en uzak komşu (the farthest neighbor algorithm) algoritması, en yakın komşu algoritmasına çok benzemektedir (Özkan, 2016, s. 203). Bu yöntemde, en yakın komşu algoritmasında olduğu gibi kümelerin elemanları arasındaki en yakın mesafe yerine en büyük uzaklık dikkate alınmaktadır (Akpınar, 2014, s. 307).



Şekil 14. En Uzak Komşu Algoritması

En uzak komşu algoritmasında iki kümenin birbirine en uzak elemanları arasındaki Şekil 14'de ok işareti ile gösterilen uzaklık A ve B kümesinin birbirine olan uzaklığı olarak değerlendirilmektedir (Özkan, 2016, s. 203).

En uzak komşu algoritmasında kümeleme yapılırken iki nokta arasındaki en uzak mesafe dikkate alınmaktadır. Yukarıda sözü edilen örnekte, en uzak komşu algoritması uygulanmış ve veri, Şekil 14'te görülebileceği üzere A Kümesi ve B Kümesi olarak ikiye ayrılmıştır.

3.3. Birliktelik Kuralları

Pazar sepet analizi olarak da adlandırılan birliktelik kuralları temel olarak bir birliktelik problemi satış işlemini analiz eder ve sıklıkla aynı anda satın alınan ürünleri tespit etmeye çalışır (MacLennan, Crivat ve Tang, 2009, s. 7).

Birliktelik kuralları; satışta, pazarlamada, katalog tasarımında, mağaza düzenini tasarlamada ve diğer karar verme süreçlerinde yaygın olarak kullanılır. Birliktelik temelde pazar sepet analizi için kullanılan bir madencilik fonksiyonudur ve bir gruptaki ürünler arasındaki ilişkileri ve korelasyonları analiz etmeye yarar. Bir marketten süt ve yumurta satın alan insanların %90'ının aynı alışveriş içerisinde ekmek de satın aldığının tespit edilmesi bu analize bir örnektir. Birliktelik algoritmaları sağlanan kurallar çerçevesinde verilen veri setinde sıklıkla beraber ilişkilendirilen ürünleri tespit etmeye yarar. Bu yöntemde en sık kullanılan algoritma apriori algoritmasıdır (Hornick, Venkayala ve Marcade, 2007, s. 93).

3.3.1. Apriori Algoritması

Birliktelik açısından her bir bilgi parçası bir ürün olarak değerlendirilir. Birliktelik analizinin iki amacı bulunmaktadır: Birincisi, sıklıkla beraber kullanılan ürünleri tespit etmek, ikincisi ise bu birlikteliklerle alakalı kuralları belirlemektir (MacLennan, Crivat ve Tang, 2009, s. 7).

Apriori uygulanırken sıklıkla geçen kümeleri ortaya çıkarmak için veritabanını çok kez taramak gerekmektedir. İlk incelemede minimum destek eşiğini aşan bir elemanlı sık geçen veriler bulunur. Ortaya çıkartılan veriler sonraki incelemelerde aday veriler olarak adlandırılan potansiyel sıklıkla kullanılan verileri üretmek için kullanılır. Aday verilerin destek değerleri hesaplanır. Minimum eşik değerini aşan veriler o geçişte bulunan sıklıkla geçen veriler olur. Sonrasında ortaya çıkarılan sıklıkla geçen veriler bir sonraki geçiş için tekrar aday veriler olurlar. Bu uygulama yeni bir sıklıkla geçen veri olmayıncaya kadar devam ettirilir (Argüden ve Erşahin, 2008, s. 56).

Apriori algoritmasında “Destek” ve “Güven” olmak üzere iki ölçütten faydalanılmaktadır. Destek ölçütü ürünler arasındaki ilişkinin tüm alışverişler içerisinde tekrarlanma oranını belirler. Güven ölçütü ise A ürününü alan müşterilerin B ürününü de

alma olasılığını ifade etmektedir. Destek ve Güven ölçütü aşağıdaki gibi ifade edilebilir (Özkan, 2016, s. 217).

$$destek(A \rightarrow B) = \frac{sayı(A,B)}{N}$$

$$güven(A \rightarrow B) = \frac{sayı(A,B)}{sayı(A)}$$

sayı(*A*,*B*) destek sayısı *A* ve *B* ürünlerinin birlikte olduğu alışveriş sayısını ve N ise tüm alışverişlerin adetini göstermektedir. Hesaplama yapılırken bu değerleri karşılaştırmak için eşik değerlere ihtiyaç duyulmaktadır. Bunlar analizin başında analist tarafından belirlenir ve hesaplanan destek ve güven ölçütlerinin belirlenen eşik değerlerden büyük olması beklenir (Özkan, 2016, s. 217).

Örneğin bir markete giden müşteriler için apriori algoritması uygulandığında şu şekilde bir sonuca ulaşılması beklenmektedir: Bu müşteri et ve peynir alırsa ekmek de alır.

3.4. Regresyon

Regresyon (regression), bağımlı değişkeni tahmin etmek için bağımsız değişkenlerin kullanılmasıdır (Zhao, 2012, s. 41). Regresyon sınıflama işlemine benzer, regresyonda bir sınıfı tanımlayan karakteristikler ve öznitelikler aramak yerine, sayısal değerleri belirlemek için örüntüler bulmak amaçlanır. Basit doğrusal hat oluşturma regresyona bir örnektir. Gelişmiş regresyon modelleri sayısal girdileri olduğu kadar kategorik girdileri de işleyebilir (MacLennan, Crivat ve Tang, 2009, s. 8). Regresyon; finans, mühendislik, çevre, tıp, tarım, ekonomi, askeri ve pazarlama alanlarını kapsayan çok geniş aralıktaki problemlere uygulanabilir (Refaat, 2007, s. 15).

Regresyon, değişkenler arasındaki ilişkilerin araştırılması ve ortaya çıkarılması için kullanılan istatistiksel bir araç olarak da tanımlanmaktadır. Genellikle araştırmacı bir

değişkenin diğeri üzerindeki nedensel etkisini, örneğin bir fiyat artışının talep üzerindeki etkisini belirlemeye çalışır. Bu tür araştırma konularının incelenmesi için, hedef (bağımlı) değişken üzerinde etkisinin olduğu düşünülen bağımsız değişkenler veri kümesinde bir araya getirilir ve bağımsız değişkenlerin hedef değişken veya değişkenlerin üzerindeki nicel etkisini ortaya çıkartmak için regresyon kullanılır. Ayrıca regresyon analizinde tipik olarak tahmin edilen ilişkilerin "istatistiksel önemini", yani gerçek ilişkinin tahmin edilen ilişkiye ne kadar yakın olduğuna dair güven aralığı da değerlendirilmektedir (Sykes, 1993, s. 1).

Regresyon analizi farklı amaçlarla kullanılabilir. Analiz sonucunda oluşturulan regresyon denklemi ve katsayılar ile hedef değişkene ait tahminleme yapılabilir. Öte taraftan bağımsız ve bağımlı değişkenler arasındaki ilişkileri ortaya çıkarmak için de regresyon kullanılabilir (Maulud ve Abdulazeez, 2020, s. 140). Regresyon modelleri genel olarak doğrusal ve doğrusal olmayan regresyon olarak ikiye ayrılmaktadır.

3.4.1. Doğrusal Regresyon

Doğrusal regresyon (linear regression) analizi, temelde iki ya da daha çok değişken arasındaki ilişkiyi bulmada kullanılan yaygın bir yöntemdir. Çıktı (bağımlı değişken) hakkında bilgi sahibi olmak için tek bir bağımsız değişken kullanılıyorsa "basit doğrusal regresyon", birden çok bağımsız değişken kullanılıyorsa "çoklu doğrusal regresyon" olarak adlandırılır. Regresyon analizi ile değişkenler arasındaki ilişkinin varlığı ve eğer var ise ilişkinin gücü hakkında bilgi edinilebilir. Regresyon analizi için buğday verimi ve gübre arasındaki ilişki veya gelir düzeyi ile tüketim harcamaları arasındaki ilişkinin tespiti gibi uygulamalar örneklendirilebilir (Pektaş, 2013, s. 116).

Basit ve çoklu doğrusal regresyon denklemleri aşağıdaki gibidir (Silahtaroğlu, 2016, s. 103):

Basit Doğrusal Regresyon→ $y = b_0 + b_1 x + e$

Çoklu Doğrusal Regresyon→ $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots + b_i x_i + e$

Örnek ile açıklanacak olunursa; Tablo 10'da verilen beden değişkeni *y* (bağımlı değişken/hedef değişken), bağımsız değişkenler x_1 , x_2 , x_3 cinsiyet, kilo ve boy olacaktır. Çoklu doğrusal regresyon denkleminde b_0 , b_1 , b_2 *ve* b_3 değerleri en küçük kareler yöntemiyle hesaplanır ve denklemde ilgili yerlerde kullanılıp hedef değişken hakkında tahminleme/yorumlama yapılabilir (Silahtaroğlu, 2016, s. 104).

Tablo 10. Çoklu Doğrusal Regresyon Örnek Verileri

Cinsiyet	**Kilo**	**Boy**	**Beden**
Kız	65	1.70	Küçük
Erkek	85	1.84	Büyük
Erkek	84	1.81	Orta
Kız	70	1.65	Küçük

Doğrusal regresyon analizinde, bağımlı değişken değerleri (*y*) ile tahmini değerler ($\hat{y}$) arasındaki farkın kareler toplamını en az yapmak amaçlanır, bunun için en küçük kareler yöntemi kullanılır (Çokluk, 2010, s. 1358).

En küçük kareler yöntemi doğrusal regresyon analizinde parametrelerin tahmini için kullanılan en yaygın yöntemlerden biridir. Regresyon denkleminde yer alan b_0 ve b_1 katsayılarının örneklemden elde edilen tahmin değerleri $\hat{\beta}_0$ ve $\hat{\beta}_1$ olarak değerlendirildiğinde, basit doğrusal regresyon denklemi aşağıdaki biçimdedir (Alma ve Vupa, 2008, s. 221).

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}, \quad i = 1,2, \ldots, n$$

Yukarıdaki eşitlikte yer alan $\hat{\beta}_0$ ve $\hat{\beta}_1$ katsayılarının değerlerini bulmak için en küçük kareler yöntemi ile hata karelerinin toplamını en küçük yapacak değerin hesaplanması gerekmektedir. Hatalar gerçek Y_i ile uydurulan $\hat{Y}_i$ değerleri arasındaki farklardır. Hata terimleri aşağıdaki şekilde ifade edilebilir (Montgomery, Peck ve Vining, 2013, s. 15).

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

Hesaplanan hatalar pozitif, negatif veya sıfır değerlerinde olabilirken, toplamları aşağıdaki denklemde görülebileceği üzere sıfıra eşit olmalıdır (Albayrak, 2006, s. 230).

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0$$

En küçük kareler yöntemiyle β_0 ve β_1 katsayılarının uydurulan değerleri olan $\hat{\beta}_0$ ve $\hat{\beta}_1$ katsayılarının farkı en küçük olacak şekilde aşağıdaki gibi belirlenir (Alma ve Vupa, 2008, s. 221).

$$\text{en küçük} \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2 = \text{ en küçük} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Regresyon katsayılarını elde edebilmek için aşağıdaki eşitlikte $\hat{\beta}_0$ ve $\hat{\beta}_1$ 'e göre kısmi türevler alınıp sıfıra eşitlenir (Montgomery, Peck ve Vining, 2013, s. 13).

$$\sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i}))^2 = L$$

Böylece aşağıdaki gibi en küçük kareler normal denklemleri elde edilir (Alma ve Vupa, 2008, s. 222).

$$\sum_{i=1}^{n} Y_i = \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_{1i} Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}{}^2$$

En küçük kareler normal denklemleri üzerinden gerekli hesaplamalar yapılarak aşağıdaki gibi denklemlerin çözümü yapılabilir ve $\hat{\beta}_0$ ve $\hat{\beta}_1$ katsayılarına ulaşılabilir (Albayrak, 2006, s. 230; Montgomery, Peck ve Vining, 2013, s. 14).

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{1i} - \bar{X})^2}$$

Sonuç olarak hesaplanan $\hat{\beta}_0$ ve $\hat{\beta}_1$ katsayılarıyla birlikte bağımlı (hedef) değişkene dair tahminleme ve yorumlama yapılabilir.

Basit veya çoklu doğrusal regresyon analizlerinde uydurulan modelin geçerli olabilmesi için bazı varsayımlar bulunmaktadır. Uydurulan modelde varsayımların sağlanamaması model ile ilgili yapılacak her türlü yorumun şüphe ile karşılanmasına sebep olur. Bu nedenle her bir varsayım için gerekli testlerin yapılması ve varsayımların kontrol edilmesi geçerli bir model yaratmak için çok önemli görülmektedir. Bu varsayımlar; hataların normal dağılımı, hata varyanslarının sabit olması, hatalar arasında otokorelasyonun bulunmaması, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin doğrusal olması ve bağımsız değişkenler arasında çoklu doğrusal bağlantı probleminin olmaması şeklinde ifade edilebilir (Tunalı ve Batmaz, 2000, s. 309).

Çalışmanın dördüncü bölümünde basit ve çoklu doğrusal regresyon analizleri yapılmıştır. Bu kısımda açıklanan teorik anlatım, uygulama ve analiz bölümü olan dördüncü bölümde pratik düzlemde çalışılmıştır.

3.4.2. Doğrusal Olmayan Regresyon

Doğrusal regresyon analizleri, çoğu problemin çözümüne uygun olan geniş ve esnek bir yapı sağlamaktadır (Montgomery, Peck ve Vining, 2013, s. 390). Genellikle değişkenler arasındaki ilişkiler tanımlanırken en basit modeller tercih edilir. Doğrusal modeller ekseriyetle ilişkileri doğrusal olmayanlara göre daha basit tanımlarlar. Bu nedenle doğrusal olmayan fonksiyonların birçoğu, dönüşümler yapılarak doğrusal hale getirilebilir ve böylece en küçük kareler yöntemi ile parametrelerin tahmini analitik

olarak doğrudan hesaplanabilir. Örnek olarak aşağıdaki Cobb-Douglas üretim fonksiyonu üzerinde bir dönüşüm uygulanacaktır. Cobb-Douglas eşitliğinde Q, üretimi; A, teknolojik katsayıyı; E, emeği; S, sermayeyi; e, hata terimini; B ve C ise modelde tahmin edilecek parametreleri ifade etmektedir (Albayrak, 2006, s. 285).

$$Q = A \text{ x } E^{B} \text{x } S^{C} \text{ x } e$$

Cobb-Douglas eşitliğinin logaritmik dönüşümü yapıldığında aşağıdaki gibi doğrusal bir fonksiyon elde edilebilir. Bu fonksiyon doğrusal olduğu için en küçük kareler yöntemi ile modelin B ve C parametreleri artık tahmin edilebilir (Albayrak, 2006, s. 286).

$$\text{LnQ} = \text{In A} + \text{B In E} + \text{C In S} + \text{In e}$$

Örnekteki gibi bir dönüşüm ile problemin doğrusal model olarak ele alınması istenilen bir durum olabilir; fakat bazı problemleri doğrusal modele dönüştürmek imkansızdır. Bu durumlarda doğrusal olmayan regresyon modelleri kullanılır.

Mühendislik ve fen bilimlerinde değişkenlerin bilinen bir doğrusal olmayan fonksiyon ile ilişkili oldukları çok sayıda örnek bulunmaktadır. Bu durum araştırmacıları doğrusal olmayan modellere götürmektedir. Bu modellere en küçük kareler yöntemi uygulandığında, oluşan denklemler doğrusal değildir ve çözümü zordur. Bu nedenle doğrusal olmayan modellerde genel olarak iterasyona dayalı teknikler ile en küçük kareler toplamı minimize edilmeye çalışılır (Montgomery, Peck ve Vining, 2013, s. 390).

Doğrusal olmayan (nonlinear regression) modeller doğrusal modeller gibi basit bir formda genel olarak aşağıdaki eşitlikte verildiği gibi gösterilebilirler (Keskintürk ve Şahin, 2009, s. 169).

$$y = f(x, \theta) + \varepsilon_i$$

Eşitlikteki θ, bilinmeyen parametrelerin p x 1 vektörü; ε, $E(e) = 0$ ve $\text{Var}(e) = \sigma^2$ olmak üzere hata terimi; $f(x, \theta)$ ise doğrusal olmayan regresyon için beklenti fonksiyonudur (Keskintürk ve Şahin, 2009, s. 169).

Doğrusal olmayan regresyon modellerinin parametrelerinin tahmini için çoğunlukla doğrusal olmayan en küçük kareler yöntemi kullanılmaktadır. Doğrusal olmayan regresyon analizinde, doğrusal regresyonda olduğu gibi hata kareler toplamını en küçük yapan değerler seçilmektedir. Fakat doğrusal modellerde olduğu gibi, doğrusal olmayan modelin parametreleri denklem üzerinde hesaplanarak elde edilememektedir. Doğrusal olmayan regresyon modellerinin hata karelerini en küçük yapan değerler iteratif (yineleme) bir yaklaşımla hesaplanabilmektedir (Albayrak, 2006, s. 287). Böylece hata karelerini en küçük yapan parametreler belirlenir ve bağımlı değişkene dair tahmin ve yorum yapılabilir.

3.4.3. Belirtme Katsayısı

Belirtme katsayısı, *x* bağımsız değişkeni/değişkenleri tarafından açıklanabilen *y* bağımlı değişkenindeki değişkenliğin oranı olarak ifade edilebilir. Belirtme katsayısı R^2 şeklinde gösterilmektedir. R^2 sıfır ile bir arasında bir değer alır ve hesaplanan bire yakın R^2 değeri, *y*'deki değişkenliğin büyük bir kısmının, oluşturulan regresyon modeliyle açıklanabildiğini göstermektedir (Montgomery, Peck ve Vining, 2013, s. 36).

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

Belirtme katsayısı olan R^2 yukarıdaki formülle hesaplanmaktadır. Formülde SS_{err} hata kareler toplamı (HKT) ve SS_{tot} ise genel kareler toplamını (GKT) ifade etmektedir (Alma ve Vupa, 2008, s. 222). Tezin dördüncü bölümü olan uygulama kısmında analiz sonuçları içerisinde yer alan R^2 değerleri, uydurulan modelin açıklama gücü olarak değerlendirilmiş ve yorumlanmıştır.

3.5. Makine Öğrenimi

Makine öğrenimi (machine learning) bilgisayarların verilere dayanarak yeni davranışlar öğrenmelerini sağlayan bir bilgisayar bilimi dalıdır. Amaç, insan eğitimi

yerine bilgisayarın tekrarlanan deneyimlerden öğrenilen davranışları göstermesini sağlayan algoritmalar tasarlamaktır (Hand, 1998, s. 114). Algoritmalar uygulama sürecinde değerli çıkarımlar yapabilmektedir (Atan, 2016, s. 143). Makine öğrenimi, yapay zekanın gelişmesi için gerekli olmakla birlikte çok sayıda günlük işlem için de kullanılmaktadır (Hand, 1998, s. 118). Makine öğrenimi, regresyon, sınıflama veya kümeleme gibi işlemleri yaparak toplanan verilerin organize edilmesine yardımcı olur. Böylece veriyi etkin bir şekilde değerlendirir ve kullanışlı öngörüler sağlar (Voulgaris, 2014, s. 121).

Makine öğrenimi başka bir ifadeyle, bilgisayarların veriye dayanarak nasıl öğrenebileceklerini araştırır. Ana araştırma alanı, oluşturulan modellerin, karmaşık kalıpları tanımayı ve verilere dayanarak belirli kararlar almak için kendi kendine öğrenmesini sağlamaktır. Makine öğrenmesinde tipik olarak eğitim veri setinden öğrenilen kalıplar ve kurallar, tüm veri setinde kullanılarak çıkarımlar yapılmaktadır (Han, Kamber ve Pei, 2012, s. 24). Makine öğrenimi modelleri; karar ağaçları, yapay sinir ağları ve birliktelik kuralları analizi gibi teknikleri ve algoritmaları içermektedir (Tsiptsis ve Chorianopoulos, 2010, s. 61).

Bilişim teknolojilerindeki son yıllarda yaşanan hızlı gelişmelerle birlikte, makine öğrenimi algoritmalarında hızlı bir artış yaşanmış hem verimlilik hem de ölçeklenebilirlik açısından analitik yetenekleri genişletilmiştir. Depolanan verilerin sürekli artan hacmiyle karşı karşıya kalan veri analistleri, potansiyel zaman ve boyut sınırlamalarının üstesinden gelebilecek daha hızlı algoritmalar aramaya başladığında, makine öğrenme modelleri, büyük miktarda veriyi makul bir sürede analiz etme ihtiyacına cevap olarak geliştirilmiştir. Ayrıca, istatistiksel modellerin belirli varsayım ve kısıtlamalarının üstesinden gelmek ve yeni ortaya çıkan işletme problemlerine çözüm sağlamak için yeni algoritmalar geliştirilmiştir (Tsiptsis ve Chorianopoulos, 2010, s. 61).

Makine öğrenimi akademik ve endüstriyel araştırmalarda geniş yer tutan bir bilgisayar bilimi alanıdır. Veri madenciliği, makine öğreniminin yanı sıra diğer disiplinlerden de faydalanırken, makine öğrenimi robot bilimi gibi veri madenciliği dışındaki uygulama alanlarında da kullanılmaktadır (Agnellutti, 2014, s. 127).

Makine öğrenimi algoritmalarının önemli bir avantajı, daha az kesin veri varsayımlarına sahip olmalarıdır. Bu nedenle, modelin inşasında teknik yönler hakkında az deneyime sahip olan analistler için kullanımı daha kolay ve basittir. Genellikle, bir model inşa ederken istatistiksel algoritmalar daha fazla çaba gerektirebilir (Tsiptsis ve Chorianopoulos, 2010, s. 62).

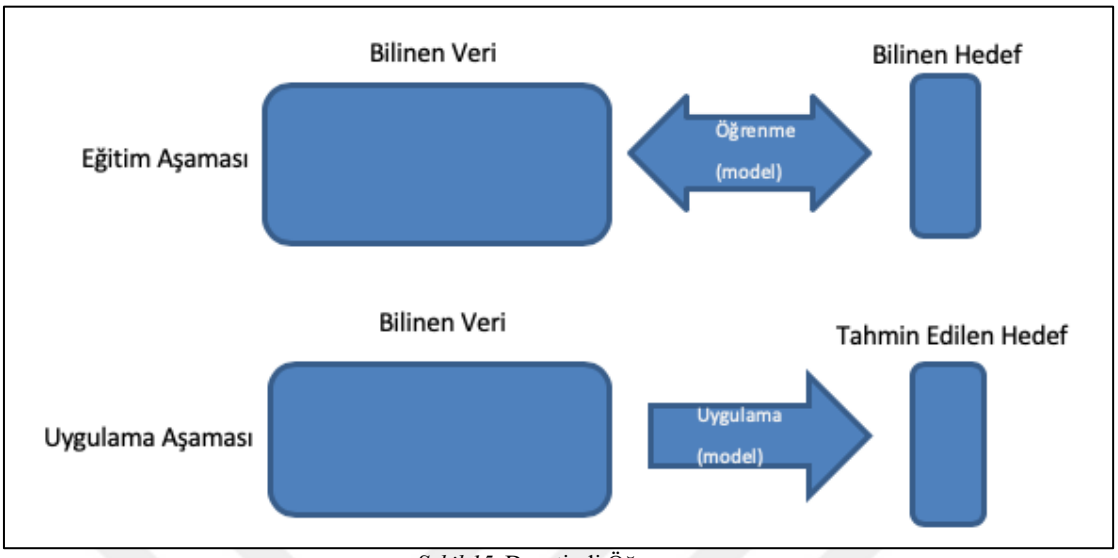
Data mining techniques in CRM (Tsiptsis ve Chorianopoulos, 2010); *A practical guide to data mining for business and industry* (Ahlemeyer-Stubbe ve Coleman, 2014); *Data mining in finance: Advances in relational and hybrid methods* (Kovalerchuk ve Vityaev, 2002) adlı temel kaynaklarda, makine öğrenimli veri madenciliği modelleri denetimli öğrenme (supervised learning) ve denetimsiz öğrenme (unsupervised learning) olarak iki temel gruba ayrılmıştır.

Her iki modelde de veri, veri seti veya girdi (inputs) gereklidir. Veri seti, eğitim ve uygulama aşaması olarak ikiye ayrılmaktadır. Eğitimde kullanılan veriler toplam veri setindeki örüntüleri öğrenmek için madencilik aktivitesinde kullanılmaktadır. Sonraki aşamada da elde edilen bulgular ve öğrenilen bilgi uygulama için ayrılan veri setine uygulanmaktadır (Ahlemeyer-Stubbe ve Coleman, 2014, s. 17).

3.5.1. Denetimli Öğrenme

Denetimli öğrenme (supervised learning), temel olarak sınıflama ile eş anlamlıdır. Öğrenmedeki denetim, eğitim veri setindeki etiketli örneklerden gelmektedir. Örneğin, bir dizi el yazısı ile yazılmış posta kodu görüntüsü olduğu varsayılsın. Örnekteki görüntüye karşılık gelen makineyle okunabilen çevirilerin bir kısmı eğitimde kullanılır

ve genel olarak modelin öğrenmesini denetler, denetimli öğrenme kavramı da buradan gelmektedir (Han, Kamber ve Pei, 2012, s. 25).



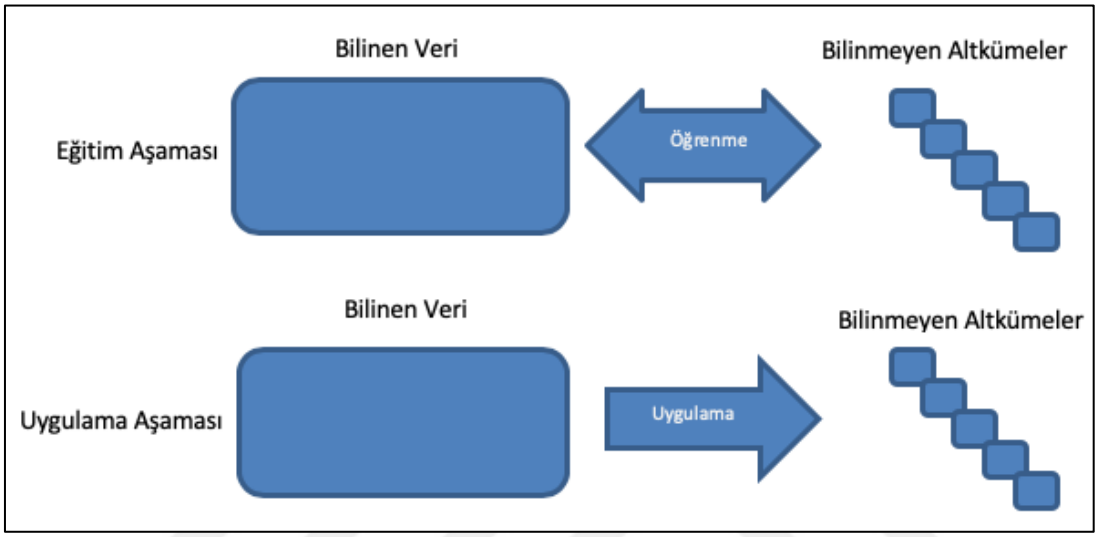
Şekil 15. Denetimli Öğrenme

Şekil 15'teki denetimli öğrenme akış şemasında da görüleceği gibi denetimli modellerde amaç, bir olayı tahmin etmek veya sürekli devam eden öz nitelikleri/karakteristikleri kestirebilmektir (Ahlemeyer-Stubbe ve Coleman, 2014, s. 16). Bu modelde girdi ve hedeflenen değer (çıktı) bulunmaktadır. Giriş verileri çıkış değerlerinin tahmin fonksiyonlarını tanımlamak için model tarafından kullanılır. Girdiler fonksiyonun X bölümü olarak, hedeflenen değer yani çıkış değeri de Y bölümü olarak düşünülebilir. Model, giriş verilerini hedeflenen değer üzerindeki etkilerini ortaya çıkartmak için kullanır. Örüntü tanıma, hedeflenen değer tarafından denetlenir. Giriş ve çıkış değerleri arasında ilişkiler kurulur ve çıkış değerlerinin öngörü oluşturması sağlanır (Tsiptsis ve Chorianopoulos, 2010, s. 3).

Denetimli öğrenme yöntemlerindeki temel amaç, veri kümesi içerisinde yer alan tüm değişkenlere ait değerlerin birtakım süreçler sonucu analiz edilerek, örüntülerin ortaya çıkarılmasıdır. Denetimli öğrenmede başlangıçta değişkenlere ait gözlem değerleri bilinmektedir ve analiz sonucu ortaya çıkan model ile değişkenlere ait veri kümesi içerisinde daha önce olmayan gözlem değerleri hakkında tahminleme yapılabilmektedir.

3.5.2. Denetimsiz Öğrenme

Denetimsiz öğrenme (unsupervised learning) modellerinde ise denetimli modelin tam tersi olarak çıkış değeri bulunmamakta ve sadece giriş verileri yer almaktadır. Denetimsiz modellerde amaç; giriş veri setindeki veri örüntülerini (data patterns) ortaya çıkartmak ve tanımlamaktır.



Şekil 16. Denetimsiz Öğrenme

Denetimsiz öğrenme temelde kümeleme ile eşanlamlıdır. Şekil 16'da görülebileceği gibi Girdi örnekleri sınıf etiketli olmadığından öğrenme süreci denetlenemez (Ahlemeyer-Stubbe ve Coleman, 2014, s. 17). Genellikle, veri içindeki grupları keşfetmek için kümeleme kullanılmaktadır. Örneğin, denetimsiz öğrenme modeli için girdi olarak el yazısı ile yazılmış rakamlardan oluşan bir görüntü kullansın. Burada 10 veri kümesi bulduğunu varsayılıyor. Bu kümeler, sırasıyla, 0 ila 9 arasındaki 10 ayrı basamağa karşılık geliyor. Ancak eğitim verileri etiketlenmediğinden, öğrenilen model bize bulunan kümelerin semantik anlamını söyleyememektedir (Han, Kamber ve Pei, 2012, s. 25).

Denetimsiz öğrenme yöntemlerinde veri kümesi bağımlı değişken tarafından veya sınıf etiketleri tarafından denetlenmez. Veriler, analiz için seçilmiş denetimsiz öğrenme algoritmalarına göre, birbirlerine olan uzaklık veya yakınlık bakımından farklı ele alış

yöntemleriyle gruplandırılarak kümelere ayrılır. Böylece denetimsiz öğrenme amacına uygun olarak veriler kümeler vasıtasıyla tanımlanır.

4. BÖLÜM: DİJİTAL İÇERİK PLATFORMU İÇİN VERİ MADENCİLİĞİ UYGULAMASI

Bu çalışmada dijital bir içerik platformu (CoinTürk - https://coin-turk.com) için iki farklı analizle veri madenciliği uygulaması tasarlanmıştır. Bölüm 4.1.'de detayları sunulan analizlerden ilki, seçilen içerik platformu için içeriklerin başlık bazında okunma sayılarını etkileyen faktörlerin (kelimelerin) ortaya çıkarılmasını içermektedir. Bölüm 4.2.'de açıklanan ikinci analiz ise okunma sayılarıyla, ilgili organizasyonun gelirleri arasındaki ilişkinin ortaya çıkarılmasını kapsamaktadır. Her iki analiz de içeriklerin okunma sayıları ardındaki gizli örüntüleri ortaya çıkardığı için birbirleriyle yakından ilişkilidir. Bu bağlamda iki ayrı analiz yapılmış ve elde edilen bulgular "Sonuç ve Öneriler" bölümünde detaylı bir şekilde yorumlanmıştır.

4.1. Okunma Sayılarını Etkileyen Faktörler

Yayınlanan içeriklerin araştırmanın amacına uygun olduğu görüldüğünden, analizler için blokzincir (blockchain) teknolojileri ve kripto paralar hakkında dijital içerik üreten CoinTürk (https://coin-turk.com/) platformu seçilmiştir. Mart 2014 tarihinden itibaren içerik üreten platformda Eylül 2022 tarihine kadar toplamda 30.000'in üzerinde içerik yayınlanmıştır. Platform genel olarak incelendiğinde içeriklerin; haber, analiz, uzman yorumları, son dakika gelişmeleri, makale ve rehber yazılar gibi çeşitli kategorilerde, özgün yazılar olduğu görülmektedir. Bu analizde platformun yayınladığı dijital içerikler başlık bazında incelenmiş ve içerik başlıklarının analiz edilmesiyle okuma sayıları üzerinde anlamlı bir etkiye sahip olan kelime ve kelime grupları araştırılmıştır.

4.1.1. Amaç

Bu araştırmada ele alınan probleme özel olarak farklı bir metin madenciliği yaklaşımı ile oluşturulan veri kümesi üzerinde analizler gerçekleştirilerek başlık bazında okunma sayılarını etkileyen faktörlerin ortaya çıkarılması amaçlanmıştır. Böylelikle

içeriklerin okunma sayılarını iyileştirebilecek ve performansı arttırabilecek faktörler belirlenmiştir. Bu analizin amacına uygun olarak aşağıdaki araştırma soruları oluşturulmuştur.

- Veri ve metin madenciliği yöntemleri kullanılarak başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime grupları) ortaya çıkarılabilir mi?
- Başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime grupları) nelerdir?
- Başlık metinlerindeki tek başına anlam ifade etmeyen etkisiz kelimeler (stop-word) çalışmanın ele aldığı bu problem için istatistiksel olarak anlamlı mıdır?
- Başlık metinlerindeki kelimeler, köklere indirgenmeden ekleriyle beraber analiz edildiğinde, istatistiksel olarak anlamlı sonuçlar ortaya çıkıyor mu?

Bu kısımdaki analizin hedeflerine uygun olarak aşağıdaki hipotez oluşturulmuştur.

- İçerik başlıklarında kullanılan kelime ve kelime grupları içeriklerin okunma sayısına etki etmektedir.

4.1.2. Yöntem

Analiz için gerekli ham veriyi toplamak için üç aylık bir dönem belirlenmiştir. 1 Mayıs 2022 – 31 Temmuz 2022 tarihleri arasında platformun yayınladığı içerikler çalışmanın veri kümesini oluşturmaktadır. Belirlenen üç aylık dönemde platform toplamda 2.743 içerik yayınlamıştır ve çalışmanın veri kümesini 2.743 içerik oluşturmaktadır. Bu araştırma, uygulama için seçilen platformun, belirlenen üç aylık dönemde yayınladığı 2.743 içeriğe ait başlıkların, veri ve metin madenciliği teknik ve yöntemleriyle incelenmesine dayanmaktadır.

Öncelikle analiz için seçilen platformun veritabanından ham veri, başka bir ifadeyle, platformda yayınlanan içeriklerin başlıkları, metin olarak sağlanmıştır. Ayrıca aynı içeriklere ait okunma istatistikleri de elde edilmiştir. Sonraki aşamada metin madenciliği kapsamına giren veri hazırlama, metin ön işleme ve değişken yaratma işlemleri yapılmıştır. Metin içerikli ham veriler metin madenciliği süreçleri sonucu yapısal forma dönüştürülmüştür. Veriler hazırlandıktan sonraki aşamada ise, elde edilen içeriklere ait yapısal veriler, istatistik yazılımına yüklenmiştir. Veri kümesi içerisinde yer alan her bir içeriğin okunma sayıları ile içeriğin başlığından farklı bir yaklaşımla metin madenciliği teknikleriyle elde edilen değişkenler, bir çoklu doğrusal regresyon problemi olarak analiz edilmiştir. Bu analizle ilgili gerçekleştirilen tüm veri madenciliği süreçleri, 4.1.2. yöntem bölümü kapsamında 4.1.2.1. veri hazırlama, 4.1.2.2. metin ön işleme, 4.1.2.3. değişken yaratma ve 4.1.2.4. analiz kısımlarında adım adım açıklanarak detaylar sunulmuştur.

4.1.2.1. Veri Hazırlama

Verilerin hazırlanması aşamasında iki farklı kaynaktan ham veriler elde edilmiştir. İlk olarak platformun veritabanından 2.743 içeriğe ait URL (içeriğe ait bağlantı linki) ve içerik başlığı olmak üzere veri çekilmiştir. Veritabanı yönetim sistemi olan MySQL[13] üzerinden ham verinin elde edilebilmesi için bazı işlemler yapılmıştır. Öncelikle veritabanına bağlantı kurulduktan sonra, SQL [14] sorgusu ile tarih aralığı girilerek "wp_posts" tablosunda "select" işlemi yapılıp veri topluca çekilmiştir. Başlık için post_title sütunu ve URL için post_name sütunu tanımlanmıştır. Hazırlanan verilerin csv formatında çıktısı alınıp Excel dosyası olarak kaydedilmiştir. Aşağıda bu işlemler için oluşturulan SQL sorgusu görülebilir.

[13] Veritabanı kavramı temel olarak bir veya daha fazla ilişkili dosyada depolanan sıralı veriler topluluğudur. MySQL ise veritabanlarında verilerin güvenli bir şekilde saklanması, gerektiğinde sıralanarak sorgulanması ve filtrelenmesi vb. gibi işlemlerin yapılmasını sağlayan bir veritabanı yönetim sistemidir (Kofler, 2004, s. 4).
[14] SQL (Structured Query Language/Yapılandırılmış Sorgu Dili), veritabanı yönetim sistemleri tarafından yönetilen veritabanlarında kullanılan bir sorgulama dilidir (Melton ve Simon, 1993, s. 4).

```
SELECT post_name, post_title FROM wp_posts WHERE post_name NOT LIKE '%autosave%' AND
post_name NOT LIKE "%revision%" AND post_content != "" AND (post_date_gmt BETWEEN '2022-
05-01 00:00:00' AND '2022-08-01 00:00:00');
```

İşlemler sonrası elde edilen ham veri kümesi için ilk üç satıra ait örnek veriler Tablo 11'de gösterilmektedir.

Tablo 11. Ham Veri Kümesi

Post_Name (URL)	**Post_Title (Başlık)**
/muskin-twitterda-spam-bot-cozumu-doge	Elon Musk'ın Twitter'da Spam Bot Çözümleri: Çare Dogecoin'de (DOGE) mi?
/piyasadaki-dusus-ivmesi-devam-ediyor-bitcoin-btc-ve-altcoinlerde-son-durum	Piyasadaki Düşüş İvmesi Devam Ediyor: Bitcoin (BTC) ve Altcoin'lerde Son Durum!
/coinshares-raporu-kripto-yatirim-fonlarinda-buyuk-olcekli-cikis	CoinShares Raporu: Kripto Yatırım Fonlarında Büyük Ölçekli Çıkış!

Verilerin hazırlanmasının ikinci aşamasında, sunucudan elde edilen 2.743 içeriğe ait oturum istatistiklerini sağlamak için çalışma yapılmıştır. Oturum istatistiği içeriklere tıklayarak kullanıcıların platformu kaç kez ziyaret ettiğini göstermektedir. Bu çalışmada oturum istatistiği ilgili içeriklerin okunma sayılarını ifade etmektedir. Belirtilen istatistiğe ulaşabilmek için platformun Google Analitik [15] uygulaması kullanılmıştır. Google Analitik uygulaması üzerinde 1 Mayıs 2022 – 31 Temmuz 2022 tarih aralıkları girilerek Davranış menüsü altında yer alan Site İçeriği ve ardından Tüm Sayfalar kısmındaki raporlara ulaşılarak sayfa (URL) ve oturum sayılarını da içeren dosya Excel formatında indirilmiştir. Tablo 12'de Google Analitik'ten çekilen dosyaya ait ilk 3 kayıt örnek olarak verilmiştir. Görüleceği gibi her bir içeriğe ait dört farklı istatistik yer almaktadır. Bu araştırma için sadece oturum istatistiği gerektiğinden, diğer sütunlar veri kümesinden silinmiştir.

Tablo 12. Google Analitik'ten Çekilen Veri

Sayfa (URL)	**Sayfa Görüntüleme**	**Sayfada Geçirilen Süre**	**Oturum**	**Hemen Çıkış Yüzdesi**
/unlu-analist-bitcoinde-yuzde-400-yukselis-icin-tarih-verdi	101.086	00:00:55	41.933	%0,45
/sicak-gelisme-binance-mujdeyi-verdi	67.245	00:00:39	25.590	%0,46
/kim-ne-kadar-alacak-terra-ust-ve-luna-airdrop-detaylari-netlesti	54.370	00:01:34	14.259	%0,20

[15] Google Analitik, web sitelerine yapılan kullanıcı ziyaretleri hakkında oturum, sayfa gösterimi ve oturum süresi gibi ayrıntılı istatistikler oluşturan ücretsiz bir uygulamadır (Plaza, 2011, s. 477).

İndirilen dosyada 15.000 içeriğe ait veri bulunmaktadır. Bunun nedeni, platforma gelen okuyucuların sadece belirtilen tarih aralıklarında yayınlanan 2.743 içeriği değil, tüm tarihlerde yayınlanmış diğer yazıları da okumasından kaynaklanmaktadır ve son derece doğal bir durumdur. Bu aşamada, 15.000 içerik içerisinden belirlenen 2.743 içeriğin oturum istatistiklerine ulaşmak için çalışma yapılmıştır. Mevcut durumda Google Analitik'ten çekilmiş, URL ve oturum istatistiklerinden oluşan 15.000 kayıt ve platformun veritabanından elde edilmiş, içerisinde URL ve başlıkların bulunduğu 2.743 kayıtlık iki farklı dosya bulunmaktadır. Amaç 2.743 kaydın oturum istatistiklerine erişmek olduğu için iki dosya önce tek bir Excel dosyası olarak birleştirilmiştir. Sonrasında 15.000 kayıt üzerinde =DÜŞEYARA formülü kullanılarak her iki veri kümesine ait URL'ler ilişkilendirilmiş ve 2.723 içeriğe ait oturum sayılarına erişilmiştir. 20 adet içeriğe ait oturum istatistiklerine Google Analitik kaynaklı teknik nedenlerden dolayı erişilememiştir. İstatistiklerine erişilen 2.723 adet satırdan oluşan veri kümesinde URL verileri silinerek sadece başlık ve oturum istatistikleri kalacak şekilde düzenlenmiş ve kaydedilmiştir. Veri hazırlama aşamaları sonucunda elde edilen ham veri kümesine ait üç gözlem Tablo 13'de örnek olarak verilmiştir.

Tablo 13. Birleştirilmiş Ham Veri Kümesi

Başlık	**Oturum**
Terra (LUNA) Krizi: Güney Kore Kripto Para Borsaları Harekete Geçti!	690
Analist Son Noktayı Koydu: Bitcoin'de Boğa Koşusunun Başlaması İçin Bu Şart!	9.230
Kripto Para Firmalarından Dev Anlaşmalar!	1.052

2.723 satırdan oluşan Excel dosyası üzerinde =RASGDİZİ formülü kullanılmış ve rastgele 500 kayıt, basit rastgele örnekleme yöntemiyle seçilmiştir.

Basit rastgele örnekleme (simple random sampling) yöntemi, araştırma için elde edilen çalışma evreni içerisindeki gözlemlerin eşit oranda ve rastgele seçilme olasılığına sahip olduğu yansız bir örnekleme türüdür. Araştırmayı tüm veya gereğinden büyük örneklem verisiyle uygulamak, araştırmaya dahil edilen değişkenlerin kontrollerini güçleştirebilmektedir. Bilimsel araştırmalarda temel amaç, çok veri ile analiz

gerçekleştirmek yerine, geçerli ve tutarlı verilerle çalışmaktır (Karasar, 2022, s. 148 ve 151). Bu çalışmada uygulanan doğrusal regresyon yöntemini kapsayan ve veri kümesinin rastgele dağıldığını geçerli sayan parametrik modellerde, rastgele dağılıma sahip olmayan büyük miktarda veriler için gerekli doğrusal regresyon varsayımları sağlanamamaktadır (Batmaz ve diğerleri, 2017, s. 814; Batmaz, Karagöz ve Serdar, 2017, s. 1898). Bu sebeple literatürde, genel olarak iki yaklaşım kullanılmaktadır: Dağıtık sistemler ve örnekleme (Bifet, 2013, s. 16). İşlemci gücü ve zaman kısıtlılığı hesaba katılarak örnekleme yaklaşımı tercih edilebilir (Batmaz ve diğerleri, 2017, s. 814; Altınok, Karagöz ve Batmaz, 2021, s. 373). Diğer taraftan veriler, herhangi bir varsayıma ihtiyaç duymayan parametrik olmayan modellerle de modellenebilir. Fakat bu çalışmada parametrik modellerle çalışılmak istenildiğinden, bu araştırma için basit rastgele örnekleme yöntemi kullanılmıştır. Çünkü çalışma, metin madenciliği yöntem ve tekniklerinin uygulanmasıyla birlikte yüzlerce değişken içerdiğinden, 2.723 gözlemden 500 tanesi (500 adet başlık) rastgele seçilerek değişken sayıları kontrol altına alınmak istenmiş, tutarlı ve geçerli bir veri kümesi üzerinde çalışmak amaçlanmıştır. Tüm bu işlemler sonucunda oluşturulan 500 gözlemden oluşan Excel dosyası, bu araştırmanın veri kümesini oluşturmaktadır.

4.1.2.2. Metin Ön İşleme

Metin ön işleme ve değişken yaratma işlemleri için Anaconda Navigator'de[16] yer alan JupyterLab[17] üzerinde, Python[18] programlama dili kullanılmıştır.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.feature_extraction.text import TfidfVectorizer
```

[16] Anaconda Navigator, Python programları için yaratılan, açık kaynaklı araç ve veri kütüphanelerinin bulunduğu bir masaüstü grafik kullanıcı arabirimidir (https://docs.anaconda.com/navigator/).
[17] JupyterLab, veri biliminde kullanılan web tabanlı ve etkileşimli bir geliştirme ortamıdır (https://jupyter.org/).
[18] Python; modüller, dinamik yazım uygulamaları, yüksek seviyeli dinamik veri türleri ve sınıfları içeren etkileşimli ve nesne yönelimli bir programlama dilidir (https://docs.python.org/3/faq/general.html#what-is-python).

İşlemler için yukarıdaki kod parçacığında verilen Numpy [19], Pandas [20], Matplotlib[21] ve Sklearn[22] veri kod kütüphaneleri yüklenmiştir. Sonrasında veri hazırlama aşamasında oluşturulan dosyanın veri girişi aşağıdaki kod parçacığı yardımıyla yapılmıştır.

```
ct = pd.read_excel("data/random_500.xlsx")
df = ct.copy()
```

Veri girişi yapıldıktan sonra, bu çalışmanın 2.6.'da yer alan metin madenciliği bölümünde de bahsedildiği gibi, veri kümesinde bir birliktelik sağlamak amacıyla tüm karakterler küçük harfe dönüştürülmüştür. Ayrıca veri kümesi içerisindeki tüm noktalama işaretleri kaldırılmış ve rakamlar silinmiştir. Bu paragrafta bahsedilen metin ön işleme süreçleri aşağıdaki kod ile gerçekleştirilmiştir.

```
df['baslik'] = df['baslik'].apply(lambda x: " ".join(x.lower() for x in x.split()))
df['baslik'] = df['baslik'].str.replace('[^\w\s]','')
df['baslik'] = df['baslik'].str.replace('\d','')
```

Dönüşüm işlemlerinden sonra Tablo 14'te gösterildiği gibi veri kümesi bir sonraki aşama için hazır hale getirilmiştir. Dosya iki sütun ve 500 satırdan oluşmaktadır.

Metin ön işleme süreçleri sonucu veri kümesinde yer alan tüm kelimeler küçük harfe dönüştürüldüğünden, bu aşamadan sonraki oluşturulan tüm tablolarda kelimeler küçük harf ile gösterilmektedir.

Tablo 14. Metin Ön İşleme Sonucu Elde Edilen Veri Kümesi

Başlık	**Oturum**
terra luna krizi güney kore kripto para borsaları harekete geçti	690
analist son noktayı koydu bitcoinde boğa koşusunun başlaması için bu şart	9230
kripto para firmalarından dev anlaşmalar	1052

[19] Numpy, Python'da bilimsel hesaplamalarda kullanılan, diziler üzerinde matematiksel ve mantıksal işlemlerin yanı sıra sıralama ve seçme gibi işlemlerin yapılabildiği bir kütüphanedir (https://numpy.org/devdocs/user/whatisnumpy.html).
[20] Pandas, Python programlama dili üzerinde inşa edilen açık kaynaklı bir veri analiz ve işleme aracıdır (https://pandas.pydata.org/).
[21] Matplotlib, Pyhton programlama dilinde veri görselleştirme için kullanılan kapsamlı bir kütüphanedir (https://matplotlib.org/).
[22] Sklearn veri analizleri için kullanılan, istatiksel ve matematiksel işlem ve algoritmaların çalıştırılabildiği Pyhton programlama dili için hazırlanmış bir kütüphanedir (https://scikit-learn.org).

4.1.2.3. Değişken Yaratma

Değişken yaratma işlemleri için Sklearn içerisinde yer alan feature_extraction.text ve TfidfVectorizer komutuyla ngram yöntemi kullanılmıştır. Tezin 2.6.6.'da yer alan kelime dizileri kısmında ngram yönteminin detayları açıklanmıştır. Aşağıdaki kod parçacığı kullanılarak veri kümesi içerisinde yer alan başlıklarda geçen kelimeler 1-gram (unigram), 2-gram (bigram) ve 3-gram (trigram) olarak terim sıklık sayıları (frekans) çıkarılmıştır.

```
degiskenler = TfidfVectorizer(binary=False, norm=None,
            use_idf=False, smooth_idf=False,
            min_df=2, max_features=None, ngram_range=(1, 3))
```

Değişken oluşturma işlemi sonucunda 7.616 benzersiz kelime ve kelime grubu (değişken) frekans değerleriyle beraber listelenmiştir. Bu çalışmanın da ele aldığı parametrik [23] modellerde doğrusal regresyon varsayımlarının sağlanabilmesi ve istatistiksel olarak anlamlı ve tutarlı bir analiz ortaya çıkarabilmek için değişken sayısı kontrol altında tutulmak istenmiştir. Çünkü ortaya çıkan 7.616 değişken ile istatistiksel olarak geçerli bir parametrik model kurmak; zaman kısıtları, bellek ve işlemci gücü nedeniyle mümkün olmamaktadır. Bu nedenlerle liste içerisinden frekans sayısı iki ve üzeri olan 949 kelime ve kelime grubu analize dahil edilmiştir. Değişken sayısını kontrol altında tutabilmek ve tutarlı bir analiz gerçekleştirebilmek için bu şekilde bir sınırlama yapılması uygun bulunmuştur. Tablo 15'te frekansa göre ilk beş kelime ve kelime grubu gösterilmektedir.

Tablo 15. Frekansa Göre İlk 5 Kelime ve Kelime Grubu

Kelimeler	**Frekans**
kripto	120
bu	111
bitcoin	85
para	78
kripto para	76

[23] Veri kümesinin rastgele dağıldığı esasına dayanan ve varsayımlar içeren modellerdir (Batmaz ve diğerleri, 2017, s. 814; Batmaz, Karagöz ve Serdar, 2017, s. 1898).

Metin madenciliği çalışmalarında genellikle köke indirgeme (lemmatization) ve etkisiz kelimeleri çıkarma (stop-word removal) işlemleri yapılmaktadır. Tezin 2.6.2. ve 2.6.4. bölümlerinde açıklanan bu tipik metin madenciliği işlemleri bu çalışmada uygulanmamıştır. Bu araştırmada farklı bir yaklaşımla, kelimeler olduğu gibi kullanılmış, köke indirgeme yapılmadan ve etkisiz kelime olarak değerlendirilebilecek kelimeler çıkarılmadan metinler çözümlenmiştir. Çünkü bu çalışmada başlık bazında okunma sayısını etkileyen faktörler araştırıldığından her bir kelime olduğu haliyle önemli görülmektedir. Her kelimenin kullanıcı üzerindeki etkisinin farklı olacağı düşünülerek, kelimeler olduğu gibi bırakılmış ve tek başına anlamsız gibi görünen kelimeler veri kümesi dışına alınmayarak farklı bir yaklaşım uygulanmıştır. Çalışmanın ele aldığı problem özelinde uygulanan bu farklı yaklaşımla, başlıklarda kelimelerin nasıl kullanıldığı ve tek başına bir anlam ifade etmiyor gibi görünen kelimelerin, başka kelimelerle bir arada kullanılarak anlamlı sonuçlar verip vermeyeceği de araştırılmak istenmiştir. Farklı bir yaklaşımla ele alınan bu yöntem, analiz sonuçlarıyla beraber "Sonuç ve Öneriler" bölümünde detaylı bir şekilde örneklendirilerek açıklanmıştır. Değişken yaratma işlemleri sonucunda elde edilen 949 kelime ve kelime grubu, aşağıdaki kod yardımıyla vektör haline getirilerek veri kümesinde değişken olarak atanmıştır.

```
df1 = pd.DataFrame(degiskenler.fit_transform(df["baslik"]).toarray(), columns=
degiskenler.get_feature_names())
```

Yaratılan dosyada değişkenler oluşturulduğu ve değişken işlemleri tamamlandığı için metinsel ifadeler barındıran başlık sütunu silinmiştir. Tablo 16'da analiz öncesi veri kümesinin son hali örnek olarak gösterilmektedir.

Tablo 16. Analiz İçin Oluşturulan Veri Kümesi

oturum	**abd**	**ada**	**btc**	**kripto**	**önemli**	**örnek**	**ünlü**	**şimdi**	**…**
841	0	0	0	1	0	0	0	1	…
528	1	0	0	0	0	0	0	0	…
2.469	0	0	1	0	1	0	0	0	…
925	0	1	0	0	0	0	1	0	…
761	0	0	1	1	0	1	1	0	…

Veri kümesinde yer alan kelime ve kelime gruplarından hangilerinin oturum istatistiğini anlamlı bir şekilde etkilediğini ortaya çıkarma amacına uygun olarak veri kümesi, Tablo 16'da gösterildiği gibi hazır hale getirilmiştir.

4.1.2.4. Analiz

Araştırmada, başlıkta geçen kelime ve kelime gruplarının okunma sayısı üzerindeki etkisini ortaya çıkarmak amaçlandığından, bir içeriğe ait "oturum" istatistiği bağımlı değişken olarak ve oluşturulan 949 değişken ise bağımsız değişkenler olarak atanmıştır. 949 bağımsız değişken içerisinden modele anlamlı katkısı olabilecek değişkenleri belirleyebilmek için adımsal (stepwise) regresyon yöntemi uygulanmıştır.

Adımsal (stepwise) regresyon yönteminde potansiyel bağımsız değişkenler içerisinden modele en iyi katkı sunan değişkenler kümesi seçilmektedir (Tabachnick ve Fidell, 2019, s. 117). Bu araştırmada da uygulanan çoklu doğrusal regresyon denklemi bu çalışmanın 3.4.1. bölümünde açıklandığı gibi $y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_p x_{pj} + \varepsilon_{ji}$ eşitliği ile gösterilmektedir (Albayrak, 2006, s. 225).

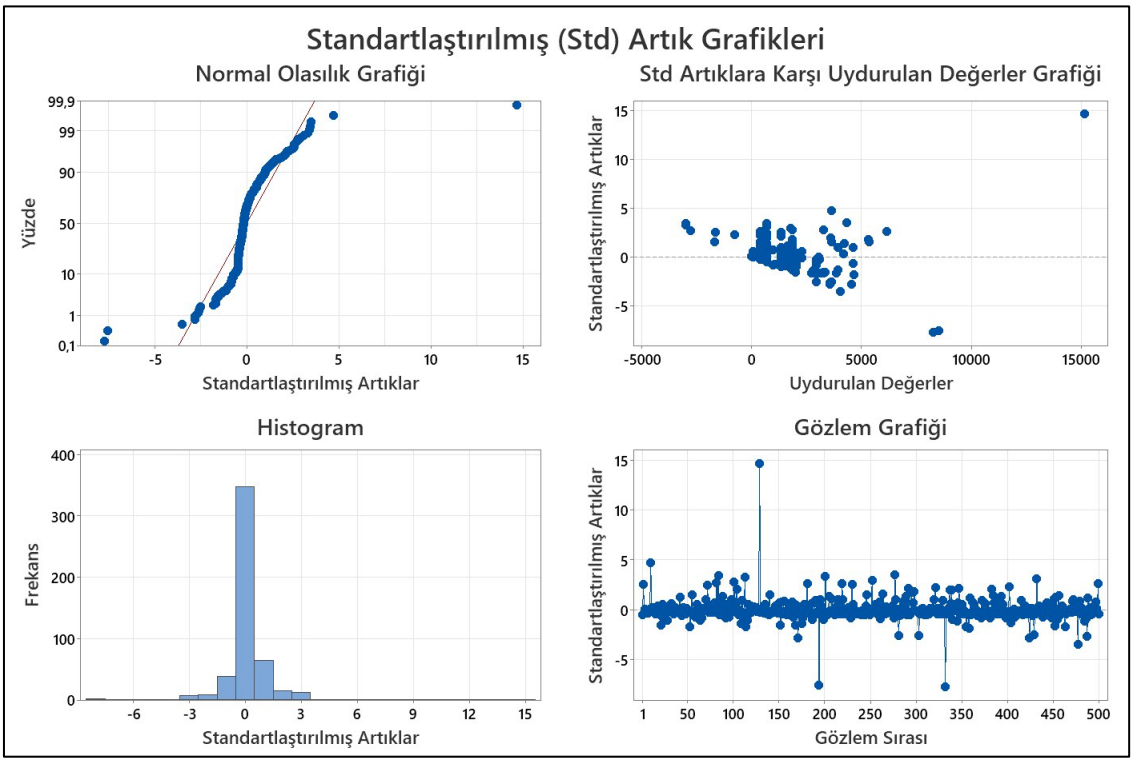
Adımsal regresyon yöntemi kullanılırken modele giriş için α (alpha) değeri 0,025, çıkış için 0,10 değerleri belirlenmiş ve %97,5 güven aralığı (p-value < 0,025) ile Minitab[24] istatistik yazılımında model kurulmuştur.

Doğrusal regresyon analizinde, bölüm 3.4.1.'de sunulduğu gibi, hataların normal dağılımı, hata varyanslarının sabit olması, hatalar arasında otokorelasyonun bulunmaması, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin doğrusal olması ve bağımsız değişkenler arasında çoklu doğrusal bağlantı probleminin olmaması şeklinde ifade edilebilecek beş temel varsayım bulunmaktadır. Uydurulan modelde bu varsayımların sağlanamaması, model ile ilgili yapılacak her türlü yorumun istatistiki

[24] Minitab; verilerdeki örüntüleri ve eğilimleri keşfetmek, gizli ilişkileri ortaya çıkarmak ve görselleştirme tahminleme yapmak gibi çok sayıda veri analiz işleminin yapılabildiği istatistik ve veri analiz aracıdır (https://www.minitab.com).

geçerliliğinin sorgulanmasına sebep olmaktadır. Bu nedenle her bir varsayım için gerekli testlerin yapılması ve varsayımların kontrol edilmesi geçerli bir model yaratmak için çok önemli görülmektedir (Tunalı ve Batmaz, 2003, s. 441).

Kurulan modelde ortaya çıkan artıklarla[25] (residual) ilgili grafikler Şekil 17'de gösterilmektedir. Normal olasılık grafiğinde referans doğrusundan sapmalar ve std artıklara karşı uydurulan değerler grafiğinde sabit olmayan varyans gözlemlenmektedir. Bu problemler, bazı doğrusal regresyon varsayımlarının sağlanmadığına işaret etmektedir.



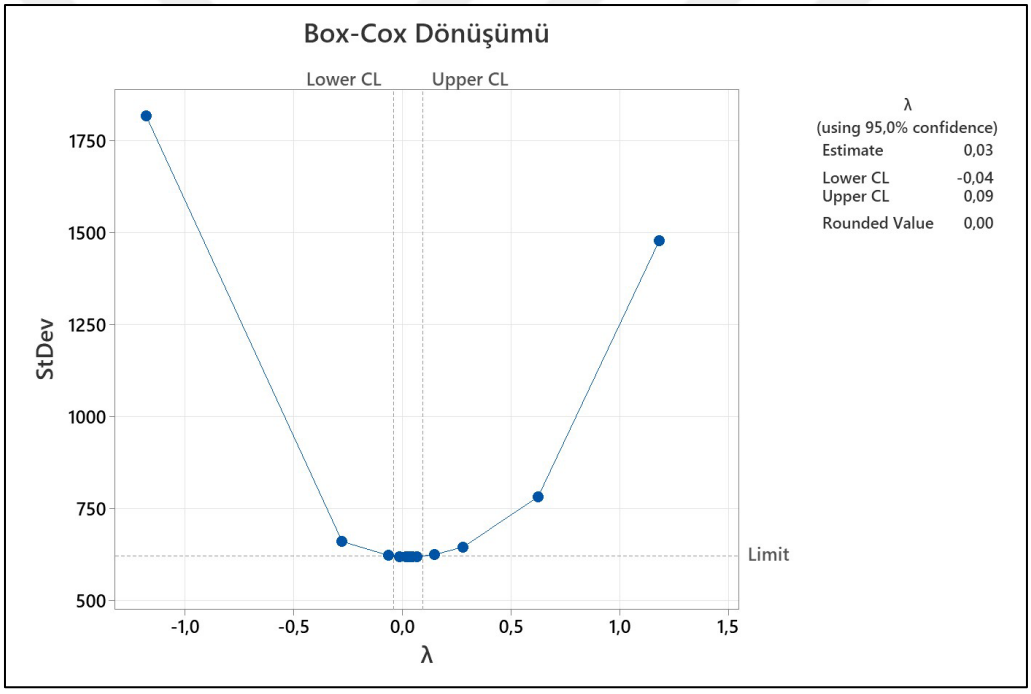
Şekil 17. Standartlaştırılmış Artık Grafikleri-1

Geçerli bir model oluşturabilmek için doğrusal regresyon varsayımlarının tamamının sağlanması amaçlanmıştır. Bu nedenle istatistiki olarak daha tutarlı sonuçlara ulaşabilmek adına oturum bağımlı değişkenine Box-Cox dönüşümü uygulanmıştır.

Veri dönüşümleri, bir dağılımın normalliğini iyileştirmek veya hata varyansını sabitlemek gibi birçok işleve hizmet edebilen yaygın olarak kullanılan yöntemlerdir. Box-

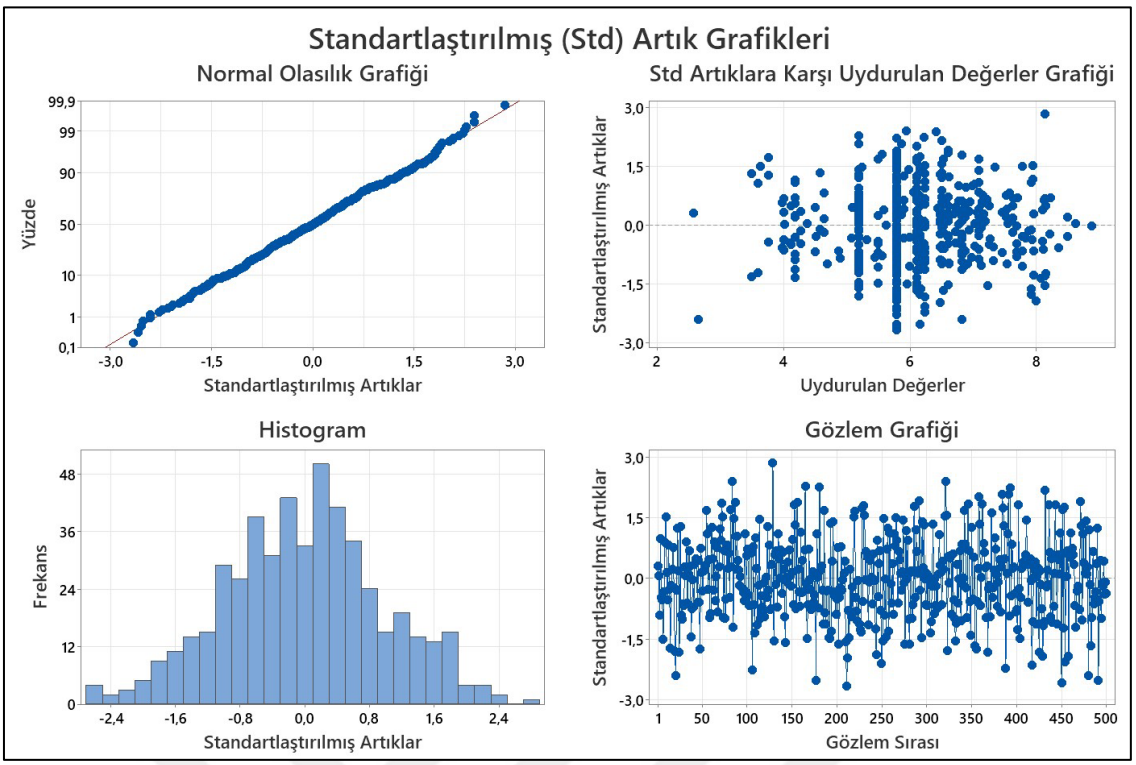
[25] Artıklar, modelin tahmin ettiği değerlerle gözlem değerleri arasındaki farklardır (Nobre ve Singer, 2007, s. 864).

Cox dönüşümü, en iyi seçeneği belirlemek için rastgele farklı dönüşümleri deneme ihtiyacını ortadan kaldırarak bağımlı değişkeni en uygun (optimal) şekilde normalleştirecek bir dönüşüm türüdür (Osborne, 2010, s. 2). En uygun dönüşümü bulabilmek için ise bir lambda değeri hesaplanmaktadır. Lambda değeri 1 ise dönüşüme gerek yoktur, -1 ise karşıt dönüşüm, -0,5 veya 0,5 değerleri için karekök dönüşümü ve 0,0 değeri için ise e tabanında logaritmik dönüşüm yapılmaktadır (Box ve Cox, 1964, s. 214). Şekil 18'de Box-Cox dönüşümü için en uygun dönüşüm hesaplanmıştır. %95 güven aralığında lambda değeri (estimate value) 0,03 bulunmuştur. Yuvarlanan değer (rounded value) 0,00 olduğundan e tabanında logaritmik dönüşüm yapılmıştır.



Şekil 18. Box-Cox Dönüşümü

Dönüşüm işlemi yapılmış ve adımsal regresyon yöntemi tekrar uygulanmıştır. Modele giriş için α (alpha) değeri 0,025, çıkış için 0,10 değerleri belirlenmiş ve %97,5 güven aralığı (p-value < 0,025) ile model oluşturulmuştur. Analiz sonucunda istatistiksel olarak anlamlı bulunan 43 bağımsız değişken belirlenmiştir.

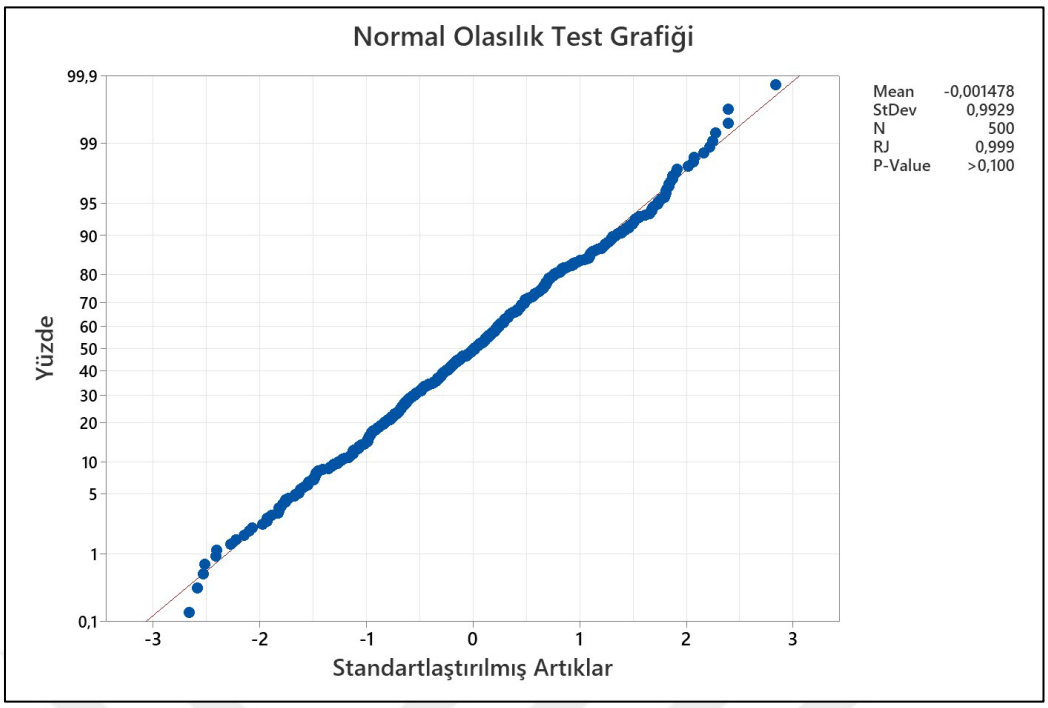


Şekil 19. Standartlaştırılmış Artık Grafikleri-2

Şekil 19'da yer alan normal olasılık grafiğinde standartlaştırılmış hataların referans doğrusunu takip ettiği gözlemlenmektedir. Ayrıca std artıklara karşı uydurulan değerler grafiğinde iyileşme gözlemlenmektedir. Yapılan Box-Cox dönüşümünün veri kümesine uygun olduğu görülmektedir. Grafikler uygun olmasına rağmen emin olmak ve tutarlı bir model yaratabilmek için tüm doğrusal regresyon varsayımları test edilmiştir.

Ryan-Joiner (RJ) normallik testi, hataların normal dağılıp dağılmadığını ölçen bir testtir. Bu testte korelasyon katsayısı bire yakınsa verilerin normal dağıldığı yorumu yapılabilir (Nosakhare ve Bright, 2017, s. 53). Hatalar normal dağılır varsayımını kontrol etmek için RJ testi yapılmıştır. Bu test için boş ve alternatif hipotez oluşturulmuştur.

- H0: Hatalar normal dağılmaktadır.
- H1: Hatalar normal dağılmamaktadır.



Şekil 20. Normal Olasılık Test Grafiği-1

Şekil 20'de gösterilen normal olasılık test grafiğindeki RJ test sonucuna göre *p*-değeri 0,10 olarak hesaplanmıştır. Bu değer α=0,05 değerinden büyük olduğu için hataların normal dağıldığını belirten H0 hipotezi reddedilmez. Sonuç olarak hatalar normal dağılır varsayımı sağlanmaktadır.

Bir diğer varsayım olan hata varyansları sabittir varsayımı için Şekil 19'da yer alan std artıklara karşı uydurulan değerler grafiği incelenmiştir. Bu grafikte hata varyanslarının eşit veya sabit olmadığına dair bir kanıt gözlenmemiştir. Sonuç olarak bu varsayımın sağlandığı şeklinde bir değerlendirme yapılmıştır.

Hatalar arasında otokorelasyon yoktur varsayımını test etmek için Durbin-Watson (DW) istatistiği yöntemi kullanılmıştır.

Durbin-Watson (DW) istatistiği sıfır ila dört aralığında ($0 < DW < 4$) bir değer ile ifade edilmektedir. Hesaplama sonucu bulunan sıfıra yakın bir değer pozitif otokorelasyon anlamına gelirken, ikiye yakın bir değer otokorelasyonun bulunmadığına, dörde yakın hesaplanan bir değer ise negatif otokorelasyonun olduğuna işaret etmektedir (Kanji, 2006, s. 169). Otokorelasyonun varlığı hakkında tespit yapılırken şu şekilde

değerlendirmelerde bulunmak mümkündür: d < dL ise pozitif otokorelasyon, dL < d < dU ise çıkarımda bulunmak için yeterli kanıt yoktur, dU < d < 4-dU ise otokorelasyon yoktur, 4-dU < d < 4-dL ise çıkarımda bulunmak için yeterli kanıt yoktur ve 4-dL < d ise negatif otokorelasyon bulunmaktadır (Uysal ve Günay, 2001, s. 279). Analizde hatalar arasında otokorelasyon yoktur varsayımını test etmek için oluşturulan hipotezler şu şekilde belirlenmiştir:

- Hatalar arasında otokorelasyon yoktur.
- Hatalar arasında negatif otokorelasyon vardır.

Analiz sonucu hesaplanan DW istatistiği (d) için 2,02 değeri bulunmuştur. Modelde n=500 ve k=43 (n gözlem sayısı, k ise bağımsız değişken sayısını ifade etmektedir) olduğundan Durbin-Watson üst ve alt kritik değerler tablosundan[26] veri kümesindeki değerlere en yakın bulunan dL= 1,34 ve dU= 1,93 değerleri değerlendirmeye alınmıştır. DW istatistiği 1,93 < 2,02 < 2,07 şeklinde hesaplandığında dU < d < 4-dU koşulu sağlandığından dolayı H0 hipotezi reddedilmez. Sonuç olarak hatalar bağımsızdır ve aralarında otokorelasyon bulunmamaktadır.

Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusaldır varsayımı incelenmiştir. Şekil 19'daki std artıklara karşı uydurulan değerler grafiğine bakıldığında herhangi bir patern görülmemektedir. Modelde uyum sorunu gözlemlenmemiştir.

Bağımsız değişkenler arasında çoklu doğrusal bağlantı problemi yoktur varsayımını değerlendirmek için model çıktıları arasında yer alan VIF (Variance Inflation Factors) değerleri incelenmiştir.

VIF, her bir bağımsız değişkenin bağımsızlık derecesini ölçmek için kullanılan bir yöntemdir. Yaygın olarak beşten büyük olan VIF değerleri modelde bir sorun

[26] Durbin-Watson üst ve alt kritik değerleri için Savin ve White'ın (1977) Durbin-Watson istatistiği tablosu kullanılmıştır.

olabileceğine işaret etmektedir (Craney ve Surles, 2002, s. 392). Modelde tüm değişkenlerin VIF değerlerinin bire yakın olduğu görülmektedir. Dolayısıyla bu varsayım sağlanmaktadır. Ayrıca veri kümesindeki aykırı ve etkili gözlemler değerlendirilmiştir.

Analiz sonucu hesaplanan standartlaştırılmış artıklar tablosunda -3 ila 3 değerleri arası dışında kalan değerler potansiyel aykırı gözlem olarak değerlendirilebilir (Montgomery, Peck ve Vining, 2013, s. 131). Modelde -3 ila 3 değerleri arası dışında kalan aykırı gözlem bulunmamaktadır.

Etkili gözlemleri tespit etmek için hii (kaldıraç noktası), Cook Distance ve DFFITS olmak üzere üç farklı istatistik kullanılmıştır.

Kaldıraç noktasına ait sınır değer 2*p/n formülü ile bulunmaktadır (n gözlem sayısı, p ise sabit katsayı dahil bağımsız değişkenlerin sayısı). Cook Distance (D) hesaplanırken D_i > 1 koşulunu sağlayan gözlemlerin etkili olduğu değerlendirilebilir. Etkinlik ölçülerinden bir diğeri olan DFFITS, |DFFITS| > 1 koşulu ile saptanabilmektedir. Belirtilen her üç yöntemin herhangi ikisinden eşik değeri geçen gözlemler etkili gözlemler olarak yorumlanabilir (Montgomery, Peck ve Vining, 2013, s. 212, 215 ve 217). Kaldıraç noktasına ilişkin sınır değer 2*44/500 =0,176 olarak hesaplanmıştır. Sınır değeri geçen 88 adet gözlem tespit edilmiştir. Hesaplanan Cook Distance tablosunda D_i > 1 koşulunu sağlayan gözlem bulunamamıştır. DFFITS çıktılarına bakıldığında ise sınır değeri geçen 12 adet gözlem tespit edilmiştir.

Analiz sonucunda kurulan nihai model için, tüm doğrusal regresyon varsayımları sağlanmıştır. Veri kümesinde aykırı değer bulunmamakla birlikte; hii, Cook Distance ve DFFITS etkinlik ölçülerinden herhangi ikisinde eşik değerleri geçen 12 adet etkili gözlem saptanmıştır. Tespit edilen etkili gözlemler üzerine ek araştırmalar yapılmıştır. Veri sağlama/toplama ve veri hazırlama aşamalarında 12 gözlem için herhangi bir yazım hatası yapılıp yapılmadığına bakılmış ve herhangi bir hataya rastlanmamıştır. 12 gözlemin

oturum istatistikleri üzerinde veri sağlama aşamasında diğer gözlemlere göre anormal bir duruma rastlanmadığından, veri kümesinde veri kaybına yol açmamak için 12 etkili gözlem veri kümesi dışına alınmamıştır. Analiz, 500 satırlık başlangıçta elde edilen veri kümesi ile sonlandırılmış ve önerilen model kullanıma uygun bulunmuştur. Bir sonraki bölüm olan "Bulgular" kısmında analiz sonucu elde edilen önerilen modele yer verilmiştir.

4.1.3. Bulgular

Araştırma sonucunda elde edilen bilgilere göre, modelde yer alan 43 değişkenden 27 tanesinin olumlu, 16 tanesinin ise okunma sayıları üzerinde olumsuz etkiye sahip olduğu ortaya çıkarılmıştır. Araştırmada düzeltilmiş R^2 değeri %52,34 olarak bulunmuştur, yani bağımlı değişken olan oturum sayısındaki değişikliğin %52,34'ü, model içerisinde yer alan 43 bağımsız değişken tarafından açıklanmaktadır. Modelin R^2 $(pred/tahmin)$ değeri ise %47,32'dir.

Tablo 17. Model Katsayıları Tablosu

Değişkenler	**Katsayılar**	***p*-değeri**		**Değişkenler**	**Katsayılar**	***p*-değeri**
Constant (sabit)	5,7950	0,000		kripto paraları	1,344	0,014
altcoin	0,714	0,000		kripto paraların	1,837	0,006
altcoinin fiyatı	1,782	0,007		kritik fiyat	-0,913	0,019
arrows	-1,670	0,012		mutlu	-1,610	0,017
açıkladı	0,983	0,000		nasıl	-0,920	0,018
açıklama geldi	1,902	0,004		neden düşüyor	1,469	0,000
bakış bitcoin	-2,592	0,000		nft	-1,608	0,000
balinaları	1,520	0,001		nftler	-2,194	0,000
binance	0,830	0,002		paylaştı	1,644	0,005
bitcoin	0,368	0,003		ralli	1,432	0,000
boğa	1,834	0,000		ralliye	1,443	0,003
bu	0,313	0,003		rekor	1,985	0,000
coinex	-2,297	0,001		shib	0,924	0,002
davasında yeni	2,502	0,001		stablecoin	-2,152	0,001
detaylar	2,330	0,000		tahmini	1,986	0,000
farklı	1,982	0,003		terra	1,290	0,000
gelişme amerikadan	2,265	0,001		usdc	-1,811	0,006
gelişme binance	1,513	0,012		ve	0,439	0,001
geçen	-1,644	0,020		veri	-1,104	0,024
gözler	1,575	0,018		yatırım şirketi	1,812	0,006
kaç	-2,033	0,000		yeni bir	-2,510	0,000
kripto	-0,604	0,000		yeni gelişme	-1,420	0,016

Tablo 17'de modelde yer alan 43 değişkene ait *p*-değerleri ve katsayılar gösterilmektedir. Bu tabloda her bir değişkenin modele yaptığı katkı görülebilir. Tabloda sunulan pozitif katsayı değerlerine sahip değişkenlerin okunma sayılarına olumlu katkı yaptığı, negatif katsayı değerlerine sahip değişkenlerin ise okunma sayılarını olumsuz etkilediği söylenebilir.

Bu analiz sonucu elde edilen modele ilişkin yorumlama ve tartışma, tezin beşinci bölümü olan "Sonuç ve Öneriler" bölümünde yapılmıştır.

4.2. Okunma Sayısının Gelire Etkisi

Tez çalışmasının ikinci uygulaması olan bu analizde, uygulama için seçilen platformun (CoinTürk / https://coin-turk.com/) yayınladığı haber, analiz, uzman yorumları, son dakika gelişmeleri, makale ve rehber yazılar gibi içeriklerinin okunma sayılarının, şirket gelirleri üzerindeki etkisi araştırılmıştır.

4.2.1. Amaç

Bu analizin temel amacı, içeriklerin okunma sayılarının şirketin geliri üzerindeki etkisini ölçmektedir. Bu bağlamda şirket gelirleri ve okunma sayıları aylık bazda ele alınıp incelenmiş ve bir tahmin modeli oluşturulmuştur.

Bu analizdeki hedeflere uygun olarak aşağıdaki araştırma soruları belirlenmiştir.

- İçeriklerin okunma sayılarıyla şirket gelirleri arasında bir ilişki var mıdır?
- İçeriklerin okunma sayıları ve şirket gelirleri kullanılarak şirket gelirleri için veri madenciliği yöntem ve teknikleriyle bir tahmin modeli oluşturulabilir mi?

Bu kısımdaki analizin hedeflerine uygun olarak aşağıdaki hipotez oluşturulmuştur.

- İçeriklerin okunma sayıları ile organizasyon gelirleri arasında bir ilişki bulunmaktadır.

4.2.2. Yöntem

Bu analizde kullanılan veriyi oluşturmak için belirli bir dönem seçilmiştir. Araştırma için Ocak 2016 – Ekim 2021 tarihleri arasındaki şirket gelirleri (USD para biriminde) aylık bazda ilgili şirketten Excel dosyası olarak sağlanmıştır. Veri kümesinin 2016 yılının Ocak ayından itibaren başlamasının nedeni, şirket gelirlerinin bu tarihten itibaren yazılı olarak kayda geçirilmesinden kaynaklanmaktadır. Ayrıca platformun sayfa gösterim (okunma) verileri de aynı tarih aralığı için aylık bazda sağlanmıştır.

Sağlanan her iki dosya Excel formatında birleştirilerek tek bir dosyaya dönüştürülmüştür. Sonrasında gelir değişkeni y bağımlı değişken olarak ve sayfa görüntüleme değişkeni ise x bağımsız değişken olarak atanmış ve basit doğrusal regresyon modeli, istatistik programında uygulanmıştır.

Bu analizle ilgili gerçekleştirilen tüm veri madenciliği süreçleri, 4.2.2. yöntem bölümü kapsamında yer alan 4.2.2.1. veri hazırlama ve 4.2.2.2. analiz kısımlarında adım adım açıklanarak detaylar sunulmuştur.

4.2.2.1. Veri Hazırlama

Bu araştırmanın veri kümesini oluşturmak için aylık bazdaki platformun sayfa görüntüleme istatistikleri Google Analitik uygulamasından elde edilmiştir. Sayfa görüntüleme istatistiği platformun yayınladığı tüm içeriklerin belirlenen tarih aralığında kaç defa okunduğunu göstermektedir.

Belirtilen istatistiğe ulaşabilmek için Google Analitik uygulaması üzerinde, 1 Ocak 2016 – 31 Ekim 2021 tarih aralığı girilerek Davranış menüsü altında yer alan Site İçeriği ve ardından Tüm Sayfalar kısmındaki raporlara ulaşılarak metrik olarak aylık baz seçilip rapor Excel formatında indirilmiştir.

Tablo 18'de Google Analitik'ten çekilen dosyaya ait ilk üç kayıt örnek olarak verilmiştir.

Tablo 18. Google Analitik İlk 3 Gözlem

Ay Dizini	Sayfa Görüntüleme Sayısı
Ocak 2016	27.615
Şubat 2016	41.334
Mart 2016	38.045

Sonraki aşamada, şirketten aylık bazda sağlanan gelir dosyası ve Google Analitik uygulamasından yine aylık bazda çekilen sayfa görüntüleme istatistikleri, tek bir dosyada birleştirilmiş ve ay dizini kaldırılmıştır. Dosyanın son hali Tablo 19'da örnek olarak gösterilmiştir.

Tablo 19. Veri Kümesinin Son Hali

Gelir	Sayfa Görüntüleme Sayısı
1529	27.615
1930	41.334
1472	38.045

Veri kümesi bir bağımlı bir bağımsız olmak üzere iki değişkenden ve 70 satırlık elde edilen verilerden oluşmaktadır.

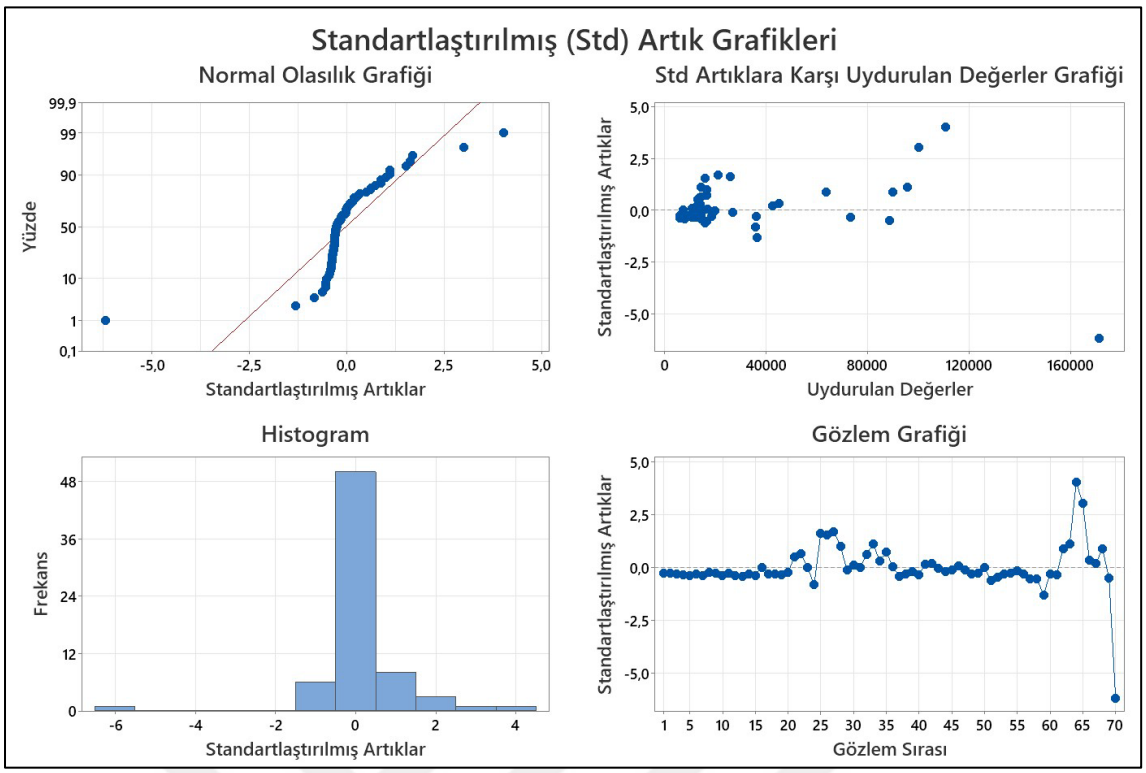
4.2.2.2. Analiz

Bu analizde sayfa görüntüleme (okunma) sayılarının gelir üzerindeki etkisi araştırılmak istenmektedir. Gelir değişkeni *y* bağımlı değişken olarak ve sayfa görüntüleme değişkeni ise *x* bağımsız değişken olarak atanmış ve %95 güven aralığı (*p*-value < 0,05) ile basit doğrusal regresyon modeli Minitab istatistik programında kurulmuştur.

Basit doğrusal regresyon denklemi $y = \beta_0 + \beta_1 x + \varepsilon$ şeklinde gösterilmektedir (Yan ve Su, 2009, s. 2).

Modele ilişkin katsayılar ve regresyon eşitliği şu şekilde hesaplanmıştır:

$$gelir = 5450 + 0.02043 * sayfa_goruntuleme$$

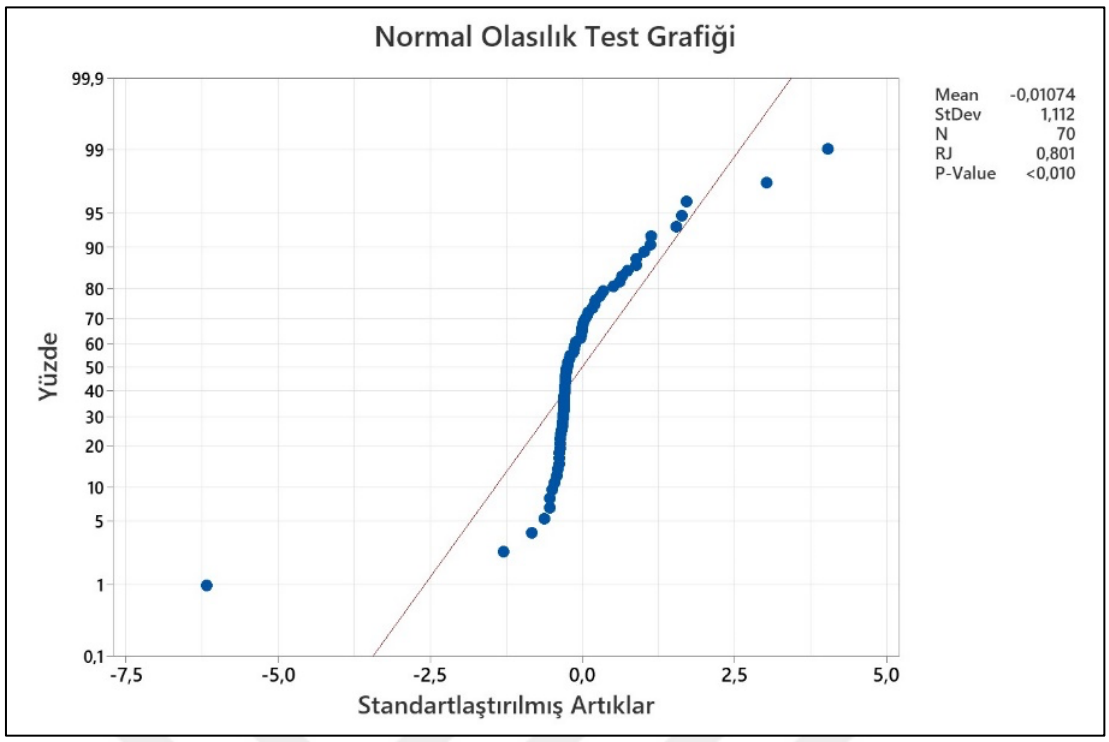


Şekil 21. Standartlaştırılmış Artık Grafikleri-3

Şekil 21'deki normal olasılık grafiği incelendiğinde, hataların normal dağılımı ilgili problemler olabileceği gözlemlenmektedir. Bu bağlamda tüm doğrusal regresyon varsayımları test edilmiştir.

Hatalar normal dağılır varsayımını kontrol etmek için RJ testi yapılmıştır. Bu test için boş ve alternatif hipotez oluşturulmuştur.

- H0: Hatalar normal dağılmaktadır.
- H1: Hatalar normal dağılmamaktadır.



Şekil 22. Normal Olasılık Test Grafiği-2

Şekil 22'de gösterilen normal olasılık test grafiğindeki RJ test sonucuna göre *p*-değeri 0,010 olarak hesaplanmıştır. Bu değer α=0,05 değerinden küçük olduğu için hataların normal dağıldığını belirten H0 hipotezi reddedilir. Sonuç olarak hatalar normal dağılır varsayımı sağlanamamaktadır.

Bir diğer varsayım olan hata varyansları sabittir varsayımı için Şekil 21'deki std artıklara karşı uydurulan değerler grafiği incelenmiştir. Bu grafikte hata varyanslarının eşit veya sabit olmadığı gözlemlenmektedir. Sonuç olarak bu varsayımın sağlanamadığı şeklinde bir değerlendirme yapılmıştır.

Hatalar arasında otokorelasyon yoktur varsayımını test etmek için Durbin-Watson (DW) istatistiği hesaplanmıştır. Bu varsayımı test etmek için oluşturulan hipotezler şu şekilde belirlenmiştir:

- Hatalar arasında otokorelasyon yoktur.
- Hatalar arasında pozitif otokorelasyon vardır.

Analiz sonucu hesaplanan DW istatistiği (d) için 0,77 değeri bulunmuştur. Modelde n=70 ve k=1 olduğundan Durbin-Watson tablosundan veri kümesindeki değerlere en yakın bulunan dL= 1,42 ve dU= 1,48 değerleri değerlendirmeye alınmıştır. DW test istatistiği (0,77) dL=1.42'dan küçük olduğundan ve d < dL koşulu sağlandığından dolayı H0 hipotezi reddedilir. Sonuç olarak pozitif otokorelasyon vardır.

Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusaldır varsayımı için veri kümesinde tekrarlanan gözlemlerin olmaması sebebiyle test gerçekleştirilememiştir. Ancak Şekil 21'de std artıklara karşı uydurulan değerler grafiğine bakıldığında, modelde uyum sorunu olabileceği şeklinde değerlendirme yapılmıştır.

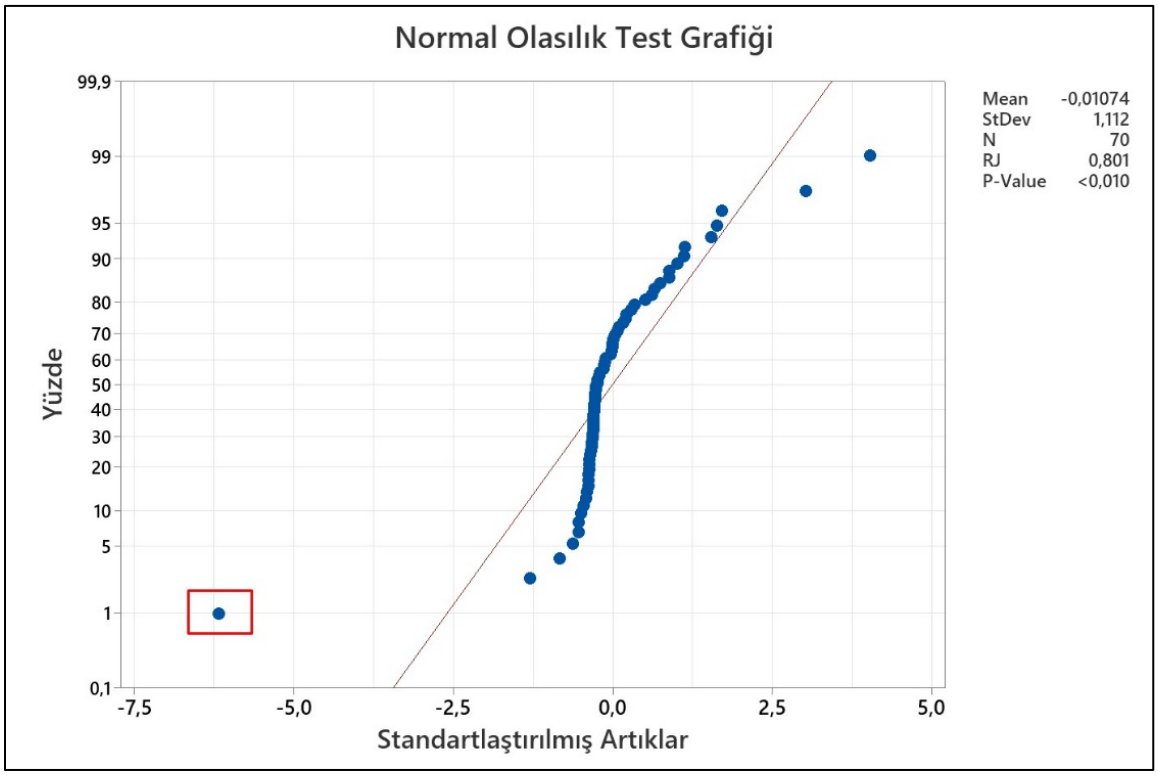
Standartlaştırılmış artıklar incelendiğinde modelde -3 ila 3 değerleri arası dışında kalan 64, 65 ve 70. gözlemlerin değerleri yüksek bulunmuştur. Gözlem 64, 65 ve 70 aykırı gözlem olarak değerlendirilebilir.

Kaldıraç noktasına ilişkin sınır değer 2*2/70 =0,057 olarak hesaplanmıştır. Sınır değeri geçen 62, 63, 64, 65, 69 ve 70. gözlemler kaldıraç noktaları olarak değerlendirilebilir. Hesaplanan Cook Distance değerleri incelendiğinde $D_i > 1$ koşulunu sağlayan 64 ve 70. gözlemler etkili gözlemler olarak değerlendirilebilir. DFFITS sınır değeri geçen 64, 65 ve 70. gözlemler etkili gözlem olarak değerlendirilebilir.

Yapılan analiz ve varsayım testleri sonucu, hataların normal dağılmadığı gözlemlenmiştir ve hatalar normal dağılır varsayımı sağlanamamıştır. Modelde Durbin-Watson istatistiğine göre pozitif otokorelasyon sorunu vardır.

Standartlaştırılmış hatalara bakıldığında ise 64, 65 ve 70. gözlemler aykırı gözlem olarak değerlendirilmiştir. Modeldeki etkili gözlemleri tespit etmek için hii, Cook Distance ve DFFITS olmak üzere üç farklı istatistik uygulanmıştır. Üç farklı istatistiğin en az herhangi ikisinde eşik değerleri geçen 64, 65 ve 70. gözlemler etkili gözlemler

olarak değerlendirilmiştir. Ayrıca Şekil 23'teki normal olasılık grafiği incelendiğinde kare içerisinde gösterilen 70. gözlemin referans doğrusundan büyük bir sapma yaptığı görülmektedir. Bu nedenle 70. gözlemin veri kümesi dışına alınması yönünde bir karar verilmiştir.

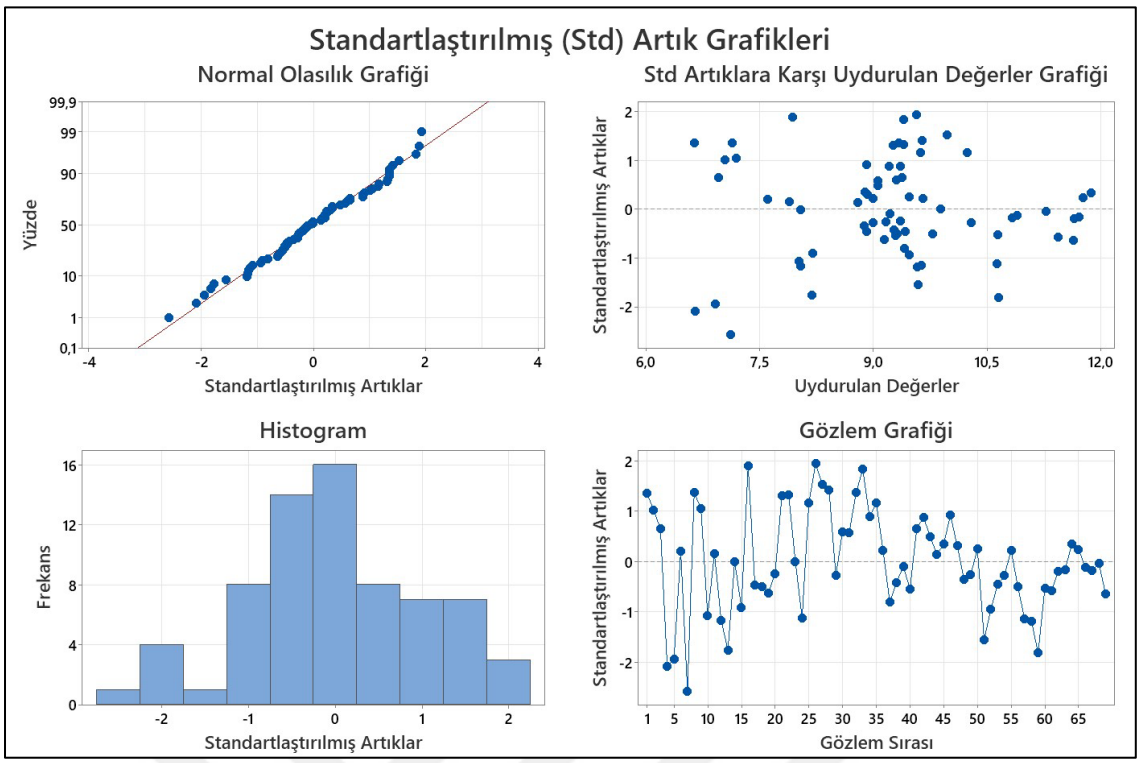


Şekil 23. Normal Olasılık Test Grafiği-3

Modeldeki varsayım problemlerini çözmek için *y* ve *x* değişkenleri üzerinde Natural log (log base e) bazlı dönüşüm yapılmıştır. 70. gözlem veri kümesi dışına alınıp dönüşüm yapıldıktan sonra model yeniden kurulmuştur.

Yeni modele ilişkin katsayılar ve regresyon eşitliği şu şekilde hesaplanmıştır:

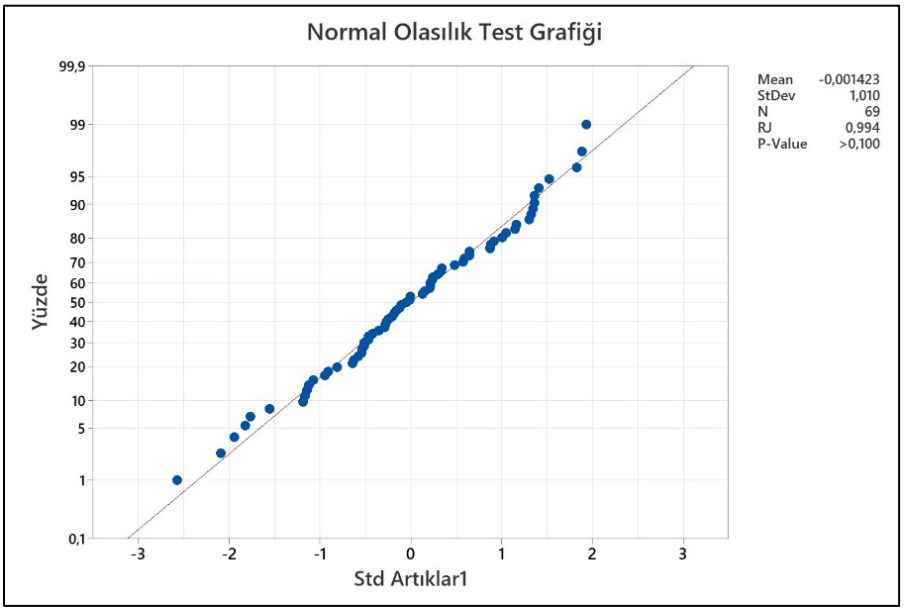
$$log_y = -3{,}588 + 1{,}0001 * log_x1$$



Şekil 24. Standartlaştırılmış Artık Grafikleri-4

Yeni model için tüm varsayımlar tekrar kontrol edilmiştir. Hatalar normal dağılır varsayımını kontrol etmek için RJ testi yapılmıştır. Bu test için boş ve alternatif hipotez oluşturulmuştur.

- H0: Hatalar normal dağılmaktadır.
- H1: Hatalar normal dağılmamaktadır.



Şekil 25. Normal Olasılık Test Grafiği-4

Şekil 25'te gösterilen normal olasılık test grafiğindeki RJ test sonucuna göre *p*-değeri 0,10 olarak hesaplanmıştır. Bu değer α=0,05 değerinden büyük olduğu için hataların normal dağıldığını belirten H0 hipotezi reddedilmez. Sonuç olarak hatalar normal dağılır varsayımı sağlanmaktadır.

Bir diğer varsayım olan hata varyansları sabittir varsayımı için Şekil 24'te std artıklara karşı uydurulan değerler grafiği incelenmiştir. Grafik incelendiğinde hatalar eşit (ya da sabit) varyansa sahiptir şeklindeki ifadeyi yanlışlayacak bir kanıt bulunamamıştır. Bundan emin olmak için bir önceki normallik varsayımı sağlandığından hataların varyansları sabittir varsayımı test etmek için Bartlett testi yapılmıştır, fakat veri kümesinde tekrarlanan gözlemlerin olmaması nedeniyle bu test hesaplanamamıştır. Sonuç olarak bu varsayımın sağlandığı şeklinde bir değerlendirme yapılmıştır.

Hatalar arasında otokorelasyon yoktur varsayımını test etmek için Durbin-Watson (DW) istatistiği hesaplanmıştır. Bu varsayımı test etmek için oluşturulan hipotezler şu şekilde belirlenmiştir:

- Hatalar arasında otokorelasyon yoktur.
- Hatalar arasında pozitif otokorelasyon vardır.

Analiz sonucu hesaplanan DW istatistiği (d) için 1,314 değeri bulunmuştur. Modelde n=69 ve k=1 olduğundan Durbin-Watson tablosundan veri kümesindeki değerlere en yakın bulunan dL= 1,42 ve dU= 1,48 değerleri değerlendirmeye alınmıştır. DW test istatistiği (1,314) dL=1.42 değerinden küçük olduğundan ve d < dL koşulu sağlandığından dolayı H0 hipotezi reddedilir. Sonuç olarak hatalar arasında pozitif otokorelasyon sorunu devam etmektedir.

Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusaldır varsayımı için Şekil 24'te std artıklara karşı uydurulan değerler grafiğine bakıldığında modelde uyum sorunu gözlemlenmemiştir.

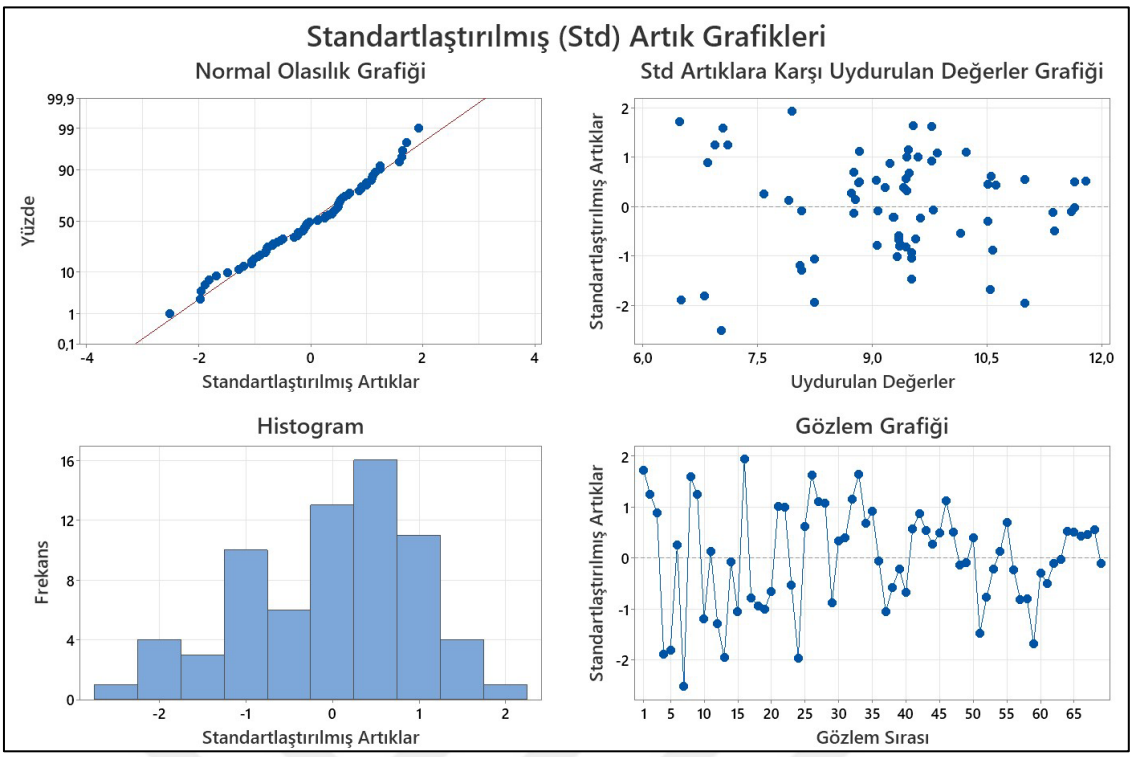
Standartlaştırılmış artıklar incelendiğinde modelde -3 ila 3 değerleri arası dışında kalan aykırı gözlem bulunmamaktadır.

Kaldıraç noktasına ilişkin sınır değer 2*2/69 =0,057 olarak hesaplanmıştır. Sınır değeri geçen 1, 2, 3, 4, 5, 62, 63, 64, 65 ve 69. gözlemler kaldıraç noktaları olarak değerlendirilebilir. Hesaplanan Cook Distance değerleri incelendiğinde $D_i > 1$ koşulunu sağlayan hiçbir gözlem bulunamamıştır. DFFITS sınır değeri geçen 4, 5 ve 7. gözlemler etkili gözlem olarak değerlendirilebilir.

Yapılan iyileştirme çalışmaları ve testler sonucunda varsayımlar içerisinden sadece otokorelasyon (pozitif otokorelasyon) sorunun devam ettiği görülmektedir. Ayrıca modelde hii, Cook ve DFFITS değerlerinden en az herhangi ikisinde eşik değerleri geçen 4 ve 5. gözlem olmak üzere iki adet etkili gözlem bulunmaktadır.

Bu araştırmadaki veri kümesi aylık bazda bir zaman dizisi içerdiğinden dolayı pozitif otokorelasyon problemini çözmek için veri kümesine zaman değişkeni eklenmiştir. Model; log_gelir (*y*), log_sayfa_goruntuleme (*x1*) ve zaman (*x2*) değişkenleri ile 69 gözlem kullanılarak yeninden uygulanmıştır. Uygulanan modelde pozitif otokorelasyon sorunun halen devam ettiği gözlemlenmiştir. Bir sonraki modelde diğer değişkenler aynı kalmak üzere zaman değişkeninin karesi alınarak model tekrar uygulanmıştır. Bu modelde de pozitif otokorelasyon sorununun olduğu tespit edilmiştir. Çözüm için zaman değişkeninin küpü $(zaman^3)$ alınarak tekrar model kurulmuştur. Uygulanan modelde sonuçlar şu şekildedir:

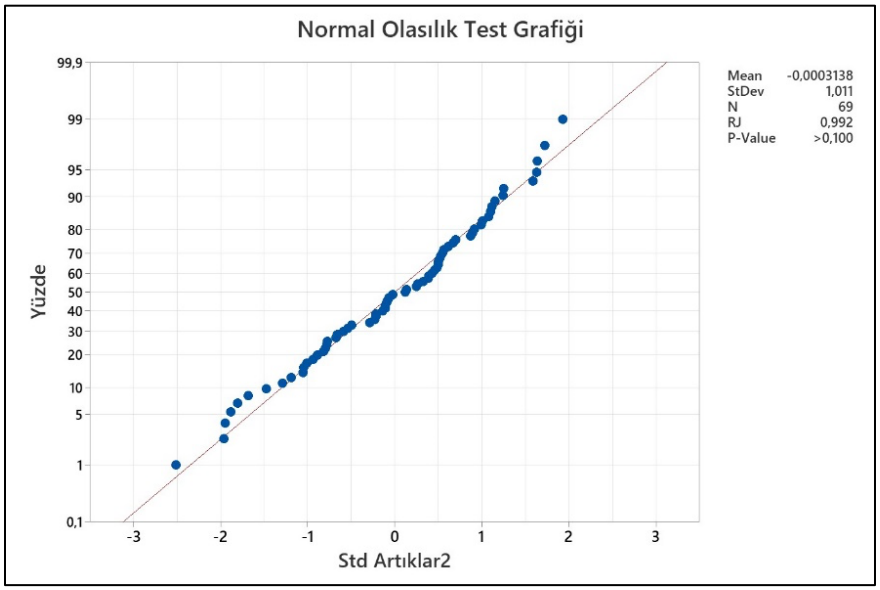
$$log_gelir = -5{,}142 + 1{,}1374 * log_sayfa_goruntuleme - 0{,}000002 * zaman^3$$



Şekil 26. Standartlaştırılmış Artık Grafikleri-5

Yeni model için tüm varsayımlar tekrar kontrol edilmiştir. Hatalar normal dağılır varsayımını kontrol etmek için RJ testi yapılmıştır. Bu test için boş ve alternatif hipotez oluşturulmuştur.

- H0: Hatalar normal dağılmaktadır.
- H1: Hatalar normal dağılmamaktadır.



Şekil 27. Normal Olasılık Test Grafiği-5

Şekil 27'de gösterilen normal olasılık test grafiğindeki RJ test sonucuna göre *p*-değeri 0,10 olarak hesaplanmıştır. Bu değer α=0,05 değerinden büyük olduğu için hataların normal dağıldığını belirten H0 hipotezi reddedilmez. Sonuç olarak hatalar normal dağılır varsayımı sağlanmaktadır.

Bir diğer varsayım olan hata varyansları sabittir varsayımı için Şekil 26'daki std artıklara karşı uydurulan değerler grafiği incelenmiştir. Grafik incelendiğinde, hatalar eşit (ya da sabit) varyansa sahiptir, şeklindeki ifadeyi yanlışlayacak bir kanıt bulunamamıştır. Sonuç olarak bu varsayımın sağlandığı şeklinde bir değerlendirme yapılmıştır.

Hatalar arasında otokorelasyon yoktur varsayımını test etmek için Durbin-Watson (DW) istatistiği hesaplanmıştır. Bu varsayımı test etmek için oluşturulan hipotezler şu şekilde belirlenmiştir:

- Hatalar arasında otokorelasyon yoktur.
- Hatalar arasında pozitif otokorelasyon vardır.

Analiz sonucu hesaplanan DW istatistiği (d) için 1,515 değeri bulunmuştur. Modelde n=69 ve k=2 olduğundan Durbin-Watson tablosundan veri kümesindeki değerlere en yakın bulunan dL= 1,400 ve dU= 1,514 değerleri değerlendirmeye alınmıştır. 1.514 < 1,515 < 2.486 hesaplanarak, dU < d < 4-dU koşulu sağlandığından dolayı H0 hipotezi reddedilmez. Sonuç olarak hatalar arasında otokorelasyon yoktur.

Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusaldır varsayımı için veri kümesinde tekrarlanan gözlemlerin olmaması sebebiyle test gerçekleştirilememiştir. Ancak Şekil 26'da std artıklara karşı uydurulan değerler grafiğine bakıldığında modelde uyum sorunu olmadığı görülmektedir.

Standartlaştırılmış artıklar incelendiğinde modelde -3 ila 3 değerleri arası dışında kalan aykırı gözlem bulunmamaktadır.

Kaldıraç noktasına ilişkin sınır değeri geçen 1, 2, 3, 4 , 5, 7, 8, 23, 24, 25, 29, 61, 62, 63, 64, 65, 66, 67, 68 ve 69. gözlemler kaldıraç noktaları olarak değerlendirilebilir. Hesaplanan Cook Distance değerleri incelendiğinde $D_i > 1$ koşulunu sağlayan hiçbir gözlem bulunamamıştır. DFFITS için sınır değeri geçen 1, 4, 5, 7 ve 24. gözlemler etkili gözlem olarak değerlendirilebilir.

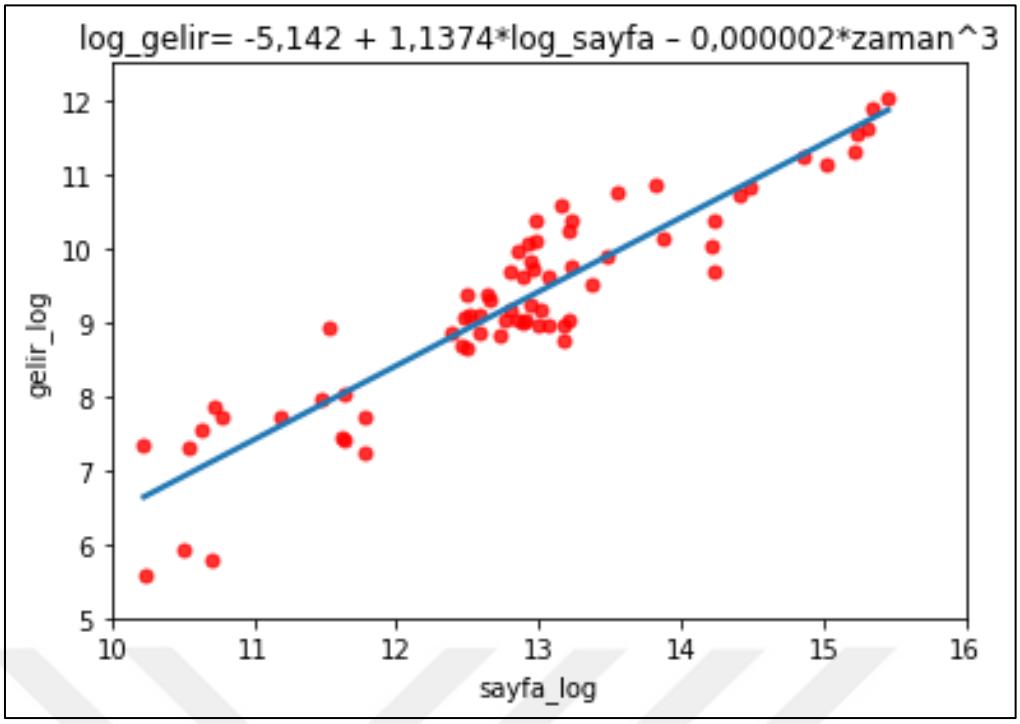
4.2.3. Bulgular

Veri kümesi üzerinde yapılan dönüşüm ve iyileştirme çalışmaları sonucunda, önerilen model için tüm doğrusal regresyon varsayımları sağlanmaktadır. Veri kümesinde aykırı değer olmamakla birlikte, herhangi iki etkinlik ölçüsünden en az iki tanesinde sınır değeri aşan 1, 4, 5, 7 ve 24. gözlemler olmak üzere, modelde beş adet etkili olarak değerlendirilebilecek gözlem bulunmaktadır. Yapılan araştırma ve değerlendirme sonucunda bu gözlemlerde veri sağlama aşamalarında herhangi bir anormallik bulunamamıştır. Veri kümesinde herhangi bir bilgi kaybına yol açmamak için saptanan beş etkili gözlem araştırma dışına alınmamıştır. Araştırma 69 gözlem ile sonuçlandırılmıştır.

Aşağıda analiz sonucu elde edilen model denklemi verilmiştir. Modelin R^2 değeri %86,61, düzeltilmiş R^2 değeri ise %86,20'dir. Önerilen model denklemi:

$$log_gelir = -5{,}142 + 1{,}1374 * log_sayfa_goruntuleme - 0{,}000002 * zaman^3$$

Önerilen modele ait regresyon doğrusu Şekil 28'deki grafikte gösterilmiştir. Gösterilen regresyon doğrusu, gelir ve sayfa görüntüleme değişkenleri arasındaki doğrusal ilişkinin fonksiyonel şeklini göstermektedir.



Şekil 28. Önerilen Modele Ait Regresyon Doğrusu

Önerilen modele ilişkin tartışma, tezin beşinci bölümü olan "Sonuç ve Öneriler" bölümünde ele alınmıştır.

4.2.4. Farklı Yöntemler

Gelir ve sayfa görüntüleme arasındaki ilişki, farklı yöntemler kullanılarak da açıklanmaya çalışılmıştır. Bunun nedeni, bir araştırmanın birden çok model ile açıklanabildiğini göstermek ve organizasyonun farklı yaklaşımlar içerisinden kendine en uygun ve kullanılabilir olan modeli seçmesi için alternatifler sunma amacı taşımaktadır. Ayrıca zaman serisi içeren veri kümelerinde otokorelasyon sorunları olduğu bilinmektedir. Bu analizdeki veri kümesi bir zaman serisi içerdiğinden dolayı, otokorelasyon sorunu farklı yöntemlerle ele alınarak, sorunun alternatif yollarla da çözülebileceği gösterilmeye çalışılmıştır. Bu farklı yöntemlerin ilkinde, önceki analizde olduğu gibi gelir değişkeni bağımlı, sayfa görüntüleme değişkeni ise bağımsız değişken olarak atanmış ve 70 satırlık veri üzerinde model kurulmuştur. Kurulan modelde doğrusal regresyon varsayımları sağlanamamış ve her iki değişkende de Natural log (log base e)

bazlı dönüşüm yapılmıştır. Yapılan dönüşüm sonucu otokorelasyon varsayımı dışında tüm varsayımlar sağlanmıştır. Otokorelasyon sorunun çözümü için farklı bir yaklaşım olarak Cochrane-Orcutt yöntemi uygulanmıştır. Cochrane-Orcutt yöntemi, model hataları arasında gözlemlenen otokorelasyon sorununu çözmek için Cochrane ve Orcutt (1949) tarafından geliştirilen değişkenleri dönüştürme yöntemidir. Aşağıdaki formül ile hesaplanmaktadır.

$$y'_t = y_t - \rho y_{t-1}$$

$$x'_t = x_t - \rho x_{t-1}$$

Formüldeki ρ'yu bulmak için y_t'nin x_t ile doğrusal regresyonundan artıkların elde edilmesi ve daha sonra elde edilen artıkların, başka bir ifadeyle e_t'nin, e_{t-1} ile regresyona sokulması ile belirlenebilir (Montgomery, Peck ve Vining, 2013, s. 482).

e_t ile e_{t-1}'in regresyona sokulması sonucunda 0.348 değeri hesaplanmıştır. Bu değer yukarıda belirtilen formülde yerine konulup y' ve x' değerleri her bir gözlem için hesaplanmış ve veri setine eklenmiştir. Örnek olarak Tablo 20'deki ikinci gözlem için:

$$y'_t = 3{,}28 - (0{,}348 * 3{,}18)$$

Bu eşitlikten y'_t (Tablo 20'de *log_y_prime* olarak gösterilmektedir) 2,17 olarak hesaplanır. Tabloda *log_y_prime* sütununun ikinci gözlemi ile bu değer karşılaştırılabilir (rakam yuvarlanmıştır). Veri setinin son hali Tablo 20'de verilmektedir.

Tablo 20. Cochrane-Orcutt Yöntemi

Sıra	log_y	log_x1	RESI	lag_resi	log_y_prime	log_x1_prime
1	3,184407	10,2261143	0,28086194	*	*	*
2	3,285557	10,6294406	0,21069625	0,280861943	2,177383504	7,0707528
3	3,167908	10,5465249	0,12826576	0,210696257	2,024533866	6,8474795
4	2,423246	10,2416726	-0,4869081	0,128265764	1,320813956	6,5714819
5	2,569374	10,5100687	-0,4547831	-0,48690813	1,726084345	6,9459666
6	3,348694	11,1902792	0,03561327	-0,45478311	2,45455207	7,5327752
7	2,510545	10,7064528	-0,5970274	0,035613273	1,345199432	6,8122356
8	3,407731	10,7261706	0,29178298	-0,59702745	2,534061064	7,0003250
9	3,359266	10,7758473	0,22221793	0,291782983	2,173375871	7,0431399
10	3,238799	11,6115845	-0,2532345	0,222217938	2,069773937	7,8615897

Dönüştürülen değişkenlerin lag'ları (bir önceki satırları) alındığı için yıldız (*) ile gösterilen 1. gözlem analize dahil edilememiştir. Dönüşüm yapıldıktan sonra model 69 gözlem ile uygulanmıştır. Uygulanan modelde sonuç olarak tüm doğrusal regresyon varsayımları sağlanmıştır. Diğer taraftan modelde üç adet etkili gözlem tespit edilmiştir. Bu gözlemler veri kümesi dışına alınıp 66 gözlem ile uygulama tekrarlanmıştır. Gerçekleştirilen varsayım testleri sonucunda tüm varsayımlar sağlanmış, fakat veri setinde yeniden iki adet etkili gözlem tespit edilmiştir. Belirlenen iki etkili gözlem veri kümesi dışına alınıp model 64 gözlem ile tekrar kurulmuştur.

Sonuç olarak son modelde tüm varsayımlar sağlanmış ve veri kümesinde aykırı ve etkili gözlem kalmamıştır. Bu yaklaşım aşağıdaki önerilen regresyon denklemi ile sonuçlandırılmıştır. Modelin R^2 değeri %72,35, düzeltilmiş R^2 değeri ise %71,90'dır. Önerilen model denklemi:

$$log_y_prime = -0{,}592 + 0{,}3856 * log_x1_prime$$

Son olarak gelir ve sayfa görüntüleme araştırması için farklı bir yöntem daha uygulanmıştır. İlk olarak model kurmak için gelir değişkeni bağımlı, sayfa görüntüleme ise bağımsız değişken olarak atanmış ve 70 satırlık veri ile model kurulmuştur. Modelde doğrusal regresyon varsayımlarının sağlanamadığı görülmüş ve değişkenler üzerinde Natural log (log base e) dönüşümü uygulanıp tekrar model kurulmuştur. Bu modelde de otokorelasyon varsayımı sağlanamadığından bir önceki yöntemde olduğu gibi Cochrane-Orcutt yöntemi uygulanarak bağımlı değişken üzerinde dönüşüm yapılmıştır. Bu yöntemde ilk satırda bulunan veri bir önceki yaklaşımda bahsedildiği gibi kaybedildiğinden 69 gözlem ile uygulama yapılmıştır. Modelde tüm doğrusal regresyon varsayımları sağlanmıştır ve üç adet etkili gözlem saptanmıştır. Bu yöntemin son aşamasında etkili gözlemleri çıkarmadan zaman değişkeni eklenerek model yeniden kurulmuştur. Veri seti aylık bazda verilerden oluştuğundan dolayı bir zaman serisi

içermektedir. Farklı bir yaklaşımla en uygun çözümü bulmak için zaman değişkeni veri setine eklenmiş ve nihai model kurulmuştur. Son modelde tüm varsayımlar sağlanmış olmakla birlikte dört adet etkili gözlem bulunmaktadır. Bu yöntem aşağıdaki önerilen model ile sonuçlandırılmıştır. Modelin R^2 değeri %74, düzeltilmiş R^2 değeri ise %73,21'dir. Önerilen model denklemi:

$$log_y_prime = -2{,}292 + 1{,}000 * log_x1_prime - 0{,}0021 * zaman$$

4.3. Bulguların Genel Analizi

Bu çalışmada okunma sayılarını etkileyen faktörler ve okunma sayılarının gelire etkisi olarak kurgulanan iki farklı analiz gerçekleştirilmiştir.

Okunma sayılarını etkileyen faktörler analizinde, bir sonraki bölümde detayları verilen "Başlık bazlı dijital içerik analiz süreci" oluşturulmuştur. Bu süreç ile içeriklere ait başlık metinlerinden, metin madenciliği teknikleri kullanılarak 949 faktör (kelime ve kelime grubu) elde edilmiş ve analiz sonucunda okunma sayılarını etkileyen kelime ve kelime grupları ortaya çıkarılmıştır. Analiz sonucu istatistiksel olarak anlamlı olan 43 farklı kelime ve kelime grubunun 27 tanesinin içerik başlıklarında kullanılmasının okunma sayılarına olumlu etki yaptığı, 16 tanesinin ise olumsuz etki yaptığı ortaya çıkarılmıştır.

Diğer analizde ise platformun genel okunma sayılarının şirket gelirleri üzerindeki etkisi araştırılmıştır. Analiz sonucu şirket gelirleri ve okunma sayıları arasında yüksek derecede neden sonuca dayanan bir ilişki tespit edilmiştir. Platformun genel olarak okunma sayıları arttığında şirket gelirleri de artmaktadır. Bu bağlamda bu iki değişken arasında bir tahmin modeli kurulmuştur. Oluşturulan tahmin modelinin düzeltilmiş R^2 değeri %86,20 olarak hesaplanmıştır. Bu yüksek R^2 değeri ile geleceğe dönük şirket gelirleri hakkında bir tahminleme yapabilmek mümkündür. Ayrıca bu analizde doğrusal regresyon varsayımlarının sağlanabilmesi için farklı yöntemler kullanılarak farklı tahmin

modelleri oluşturulmuştur. Farklı tahmin yöntemleri oluşturulmasındaki amaç ise, bir problemin birden fazla model ile açıklanabileceğini göstermek ve problem için alternatif modeller üretmektir.

Her iki analize ait değerlendirme, yorum ve öneriler son bölüm olan "Sonuç ve Öneriler" kısmında yapılmıştır.

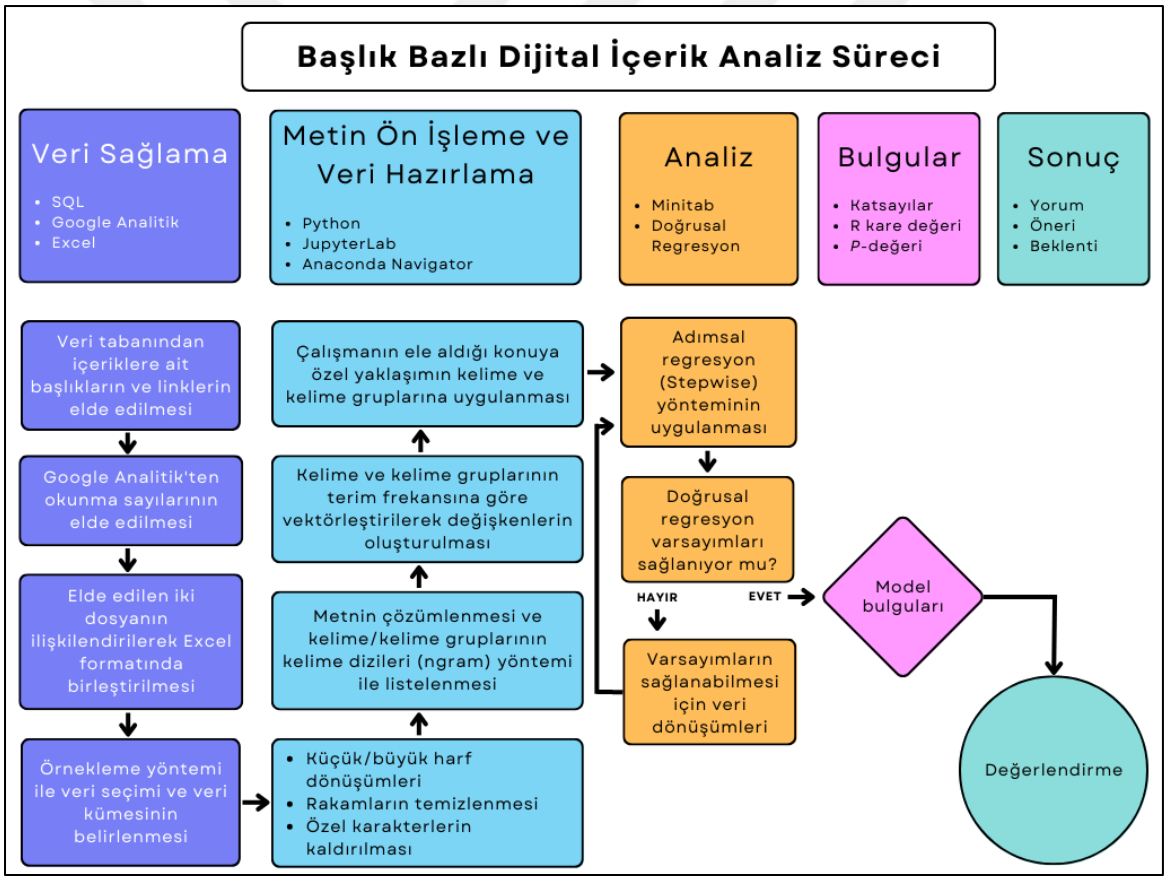
4.4. Başlık Bazlı Dijital İçerik Analiz Süreci

Veri madenciliği modelleri; faydalı veri kalıplarını tanımlamak, davranışları anlamak ve tahmin etmek için kullanılabilecek bir dizi kural ve denklemden oluşmaktadır (Tsiptsis ve Chorianopoulos, 2010, s. 3). Veri madenciliğinde model oluşturma özellikle de uygulamalı bilimlerde en önemli konular arasındadır (Spendler, 2010, s. 159). Veri madenciliği ve modelleme süreci boyunca veri analistinin takdir yetkisi çok önemlidir; çünkü literatürde bazı kural ve öneriler olmasına rağmen, çok az sayıda kesinlik bulunmaktadır. Karar verme aşamasında; özellikle de kabul edilebilir hataların doğası, genelleştirebilirlik ve modeldeki bazı konseptleri anlamak çok önemlidir. Analistin alan uzmanlığı bu süreçte doğru ve akıllı seçimler yapmak için şarttır (McCue, 2007, s. 6). En son geliştirilen veri madenciliği algoritmalarının her veri seti için her zaman en iyi olduğu söylenemez. Her veri seti kendi içerisinde benzersiz özellikler barındırdığından çok sayıda farklı algoritma denenerek ilgili veri setine uygun en iyi modeli veren algoritmayı belirlemek ve onun üzerine yoğunlaşmak/optimizasyon yapmak en doğru yaklaşım olarak değerlendirilebilir.

Bu çalışmada, okunma sayılarını etkileyen faktörlerin ortaya çıkarılabilmesi için başlık bazlı dijital içerik analiz süreci geliştirilmiştir. Süreç, dijital içerik başlıklarının, metin madenciliği teknikleriyle çözümlenmesiyle yapısal verilerin elde edilmesi ve elde edilen verilerin analiz edilerek dijital içeriklerin okunma sayılarını etkileyen faktörlerin (kelime ve kelime gruplarının) ortaya çıkarılmasına dayanmaktadır.

Araştırma blokzincir teknolojileri ve kripto paralar hakkında dijital içerik üreten bir platform için yapılmıştır. Araştırma sonucu oluşturulan başlık bazlı dijital içerik analiz süreci, farklı konularda dijital içerik üreten platformlar için de uygulanabilir. Çünkü süreç, içeriklerin konularından ve semantik (anlamsal) bağlantılarından bağımsız olarak, içerik başlıklarını salt kelime bazlı ele almaktadır. Bu yönüyle oluşturulan süreç, dijital içerik üreten tüm platformlar için uygundur.

Şekil 29'da, bu çalışmada oluşturulan "Başlık bazlı dijital içerik analiz süreci" sunulmuştur. Sürecin uygulama basamakları Şekil 29'da gösterildiği gibi adım adım ele alınarak bir yol haritası oluşturulmaya çalışılmıştır.



Şekil 29. Başlık Bazlı Dijital İçerik Analiz Süreci

Başlık bazlı dijital içerik analiz süreci, verilerin sağlanmasından sonuç kısmına kadar beş farklı aşamadan oluşmaktadır.

Sürecin ilk aşaması verilerin sağlanması aşamasıdır. Bu aşama, ilgili platformun veritabanından ve Google Analitik uygulamasından ham verilerin elde edilmesi süreci ile başlamaktadır. Sonrasında Excel formatında veriler ilişkilendirilip tek bir dosyada birleştirilerek örnek veri kümesi seçimi yapılmaktadır. İkinci aşamada, belirlenen veri kümesi üzerinde metinlerdeki küçük büyük harf dönüşümü, rakamların ve özel karakterlerin kaldırılması gibi metin ön işleme uygulamaları yapılmaktadır. Sonrasında metnin çözümlenmesi ile değişkenlerin yaratılarak verinin hazırlanması işlemleri ele alınmaktadır. Üçüncü aşama ise analiz kısmıdır. Analiz aşamasında çoklu doğrusal regresyon tekniklerinden biri olan adımsal regresyon uygulanır ve doğrusal regresyon varsayımları test edilir. Doğrusal regresyon varsayımları sağlanıyorsa bir sonraki aşamaya geçilebilir, fakat sağlanamıyorsa gerekli veri dönüşümleri yapılarak varsayımların doğrulanması gerekmektedir. Bu döngü varsayımlar sağlanıncaya kadar devam eder. Dördüncü aşama ise bulguların ortaya çıkarılması kısmıdır. Doğrusal regresyon varsayımları sağlanmış bir model için ortaya çıkan değişkenler, katsayılar, p-değerleri ve R^2 değerleri analizin bulguları olarak kayıt altına alınır. Son aşama ise sonuç kısmıdır. Bu aşamada elde edilen bulgular değerlendirilerek yorumlanır ve modele ait beklentiler ve öneriler sunulur.

Bu araştırmada Şekil 29'da sunulan sürecin uygulanması sonucunda, Şekil 30'da görselleştirilen kelimeler ortaya çıkarılmıştır. Araştırma sonucunda elde edilen içeriklerin başlıklarında geçen 43 adet kelime ve kelime grubunun, içeriklerin okunma sayısı üzerinde istatistiksel olarak anlamlı bir etkisinin olduğu ve okunma sayısındaki değişimin %52,34'ünün model içerisinde yer alan bu kelimeler tarafından açıklanabildiği tespit edilmiştir.



Şekil 30. Model Çıktısı

Ortaya çıkarılan bilgiler, ilgili organizasyona katma değer sağlayacak yeni iş süreçleri yaratma potansiyeline sahiptir. Hangi kelimelerin kullanılmasıyla içeriklerin okunma sayılarının arttığı, hangi kelimelerde ise düştüğü, oluşturulan modelde net bir şekilde ortaya konulduğundan, keşfedilen kelime ve kelime grupları, platformun oluşturacağı yeni içerikler için belirleyici olacaktır. İlgili platform bu bilgilerle, içerik yayınlama stratejilerini yeniden yapılandırarak performansı daha yüksek içerikler oluşturabilecektir.

Bir sonraki bölüm olan “Sonuç ve Öneriler” bölümünde araştırmaya dair yorum, değerlendirme ve beklentiler irdelenmiştir.

5. BÖLÜM: SONUÇ VE ÖNERİLER

Özellikle son yirmi yılda önemi giderek artan bir kavram haline gelen veri madenciliği, ham verilerin çeşitli süreçlerle işlenerek değerli bilgilerin elde edilmesini ifade etmektedir. Onlarca yıldır kendi veritabanlarında veri depolayan hemen hemen her organizasyon, biriken verileri işleyerek elde edilen yeni bilgilerle iş süreçlerini iyileştirme, kullanıcı deneyimini arttırma, verimlilik sağlama veya karar destek sistemlerinde kullanma gibi çok sayıda konuda geliştirme yapabilmektedir. Veri madenciliği konusu, özellikle bilişim sektöründe yer alan her organizasyon için günümüzde yapılması zorunlu bir iş sürecine dönüşmüştür. Bu tez çalışmasında da veri madenciliği iş süreçlerine uygun, araştırmanın temel hedefleri doğrultusunda dört farklı çıktı elde edilmiştir.

İlk olarak, araştırma için seçilen platformun ham verileri kullanılarak gerçekleştirilen veri madenciliği süreçlerinde, platformun yayınladığı içeriklerde başlık bazında okunma sayılarını etkileyen faktörler ortaya çıkarılmıştır. Ayrıca bu araştırma için "Başlık bazlı dijital içerik analiz süreci" geliştirilmiştir. İkinci olarak, içeriklerin başlıkları için ele alınan bir önceki cümlede bahsedilen analiz için, seçilen probleme özel olarak farklı bir metin madenciliği yaklaşımı uygulanmıştır. Üçüncü olarak, içeriklerin okunma sayıları ile şirket gelirleri arasındaki ilişki ortaya çıkarılmış ve bir tahmin modeli oluşturulmuştur. Son olarak, çalışmada veri madenciliği kavramı bilgi yönetimi çerçevesinde ele alınmış ve gelecekte Bilgi ve Belge Yönetimi Bölümü öğrenci veya mezunlarından veri madenciliği çalışması yapmak isteyenler için uçtan uca örnek uygulama ile motivasyon oluşturmak amaçlanmıştır. Bu çıktıların her biri bu bölümde değerlendirilmiş ve yorumlanmıştır.

Tez çalışmasının birinci ve ikinci çıktısı aynı analiz üzerinde yer aldığı için beraber değerlendirilmiştir. Ayrıca "Veri ve metin madenciliği yöntemleri kullanılarak başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime

grupları) ortaya çıkarılabilir mi?" ve "Başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime grupları) nelerdir?" araştırma soruları aynı paralelde değerlendirilerek cevaplandırılmıştır. Ek olarak bu analiz için "İçerik başlıklarında kullanılan kelime ve kelime grupları içeriklerin okunma sayılarına etki etmektedir" hipotezi sınanmıştır.

Okunma sayılarını etkileyen faktörler üzerine yapılan araştırma sonucunda elde edilen, içeriklerin başlıklarında geçen 43 adet kelime ve kelime grubunun, içeriklerin okunma sayısı üzerinde istatistiksel olarak anlamlı bir etkisinin olduğu ortaya çıkarılmıştır. Bu sonuç, "İçerik başlıklarında kullanılan kelime ve kelime grupları içeriklerin okunma sayılarına etki etmektedir" hipotezinin doğrulandığını göstermektedir. Okunma sayısındaki değişikliğin %52,34'ü, model içerisinde yer alan 43 kelime ve kelime grubu tarafından açıklanmaktadır. Bu sonuçla, "Veri ve metin madenciliği yöntemleri kullanılarak başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime grupları) ortaya çıkarılabilir mi?" ve "Başlık bazında içeriklerin okunma sayılarını etkileyen faktörler (kelime ve kelime grupları) nelerdir?" sorularına cevap bulunmuştur.

Bu analiz sonucunda oldukça değerli bilgiler elde edilmiştir. Bu bilgiler, veri madenciliğinin gerçekleştirildiği ilgili organizasyona katma değer sağlayacak yeni iş süreçleri yaratma potansiyeline sahiptir. Bu analiz sonucunda geliştirilen "Başlık bazlı dijital içerik analiz süreci" bölüm 4.4.'te detaylarıyla sunulmuştur. Analiz sonucu geliştirilen model içerisinde yer alan tüm kelime ve kelime grupları, oluşturulacak yeni içerikler için yön verici ve ışık tutucu olacaktır. Çünkü hangi kelimelerin başlıklarda kullanılmasıyla içeriklerin okunma sayılarının arttığı, hangi kelimelerde ise düştüğü, oluşturulan modelde net bir şekilde ortaya konmaktadır. Organizasyon bu bilgilerle, içerik yayınlama stratejilerini yeniden yapılandırarak performansı daha yüksek içerikler oluşturabilir ve böylelikle şirket hedefleri dahilinde daha yüksek okunma sayılarına

ulaşabilir. Bu araştırmada kullanılan veri kümesi üç aylık bir dönemi kapsadığından, elde edilen sonuçlar bir müddet sonra geçerliliğini yitirebilir. Fakat bu çalışmada ortaya konan başlık bazlı dijital içerik analiz sürecinin, üç aylık dönemlerde tekrarlanarak ortaya çıkarılacak yeni kelime ve kelime gruplarıyla şirket için veri bilimine dayalı sürdürülebilir bir büyüme sağlanabilir.

Bu analizin 4.1.2.3. değişken yaratma bölümünde, kelimeleri köke indirgeme (lemmatization) ve etkisiz kelimeleri çıkarma (stop-word removal) işlemlerinin genellikle metin madenciliği süreçlerinde yapıldığından; fakat bu çalışmada bu işlemlerin yapılmadığından ve farklı bir yaklaşım uygulandığından söz edilmişti. Bu çerçevede, "Başlık metinlerindeki tek başına anlam ifade etmeyen etkisiz kelimeler (stop-word) çalışmanın ele aldığı bu problem için istatistiksel olarak anlamlı mıdır?" ve "Başlık metinlerindeki kelimeler, köklere indirgenmeden ekleriyle beraber analiz edildiğinde, istatistiksel olarak anlamlı sonuçlar ortaya çıkıyor mu?" araştırma soruları geliştirilmiştir. Yukarıdaki cümlede bahsedilen iki araştırma sorusuna cevap bulabilmek için köke indirgeme ve etkisiz kelimeleri çıkarma işlemlerinin bu çalışmanın ele aldığı problem için neden uygulanmadığı, analiz sonucu elde edilen bilgiler ışığında açıklanmaya çalışılmıştır. Bu bağlamda, modelde içerisinde "*fiyat*" kelimesi bulunan değişkenler incelenmiştir. Model çıktılarına bakıldığında Tablo 21'de görüleceği gibi içerisinde "*fiyat*" kelimesi geçen iki farklı değişken tespit edilmiştir.

Tablo 21. Fiyat İfadesi Geçen Değişkenler

Değişkenler	**Katsayılar**
kritik fiyat	-0,913
altcoinin fiyatı	1,782

Bu iki değişken, kendi içerisinde değerli çıktılar barındırmaktadır. "*kritik fiyat*" kelime grubunda, her iki kelimenin de herhangi bir ek almadan kullanıldığı görülmektedir. Diğer kelime grubu olan "*altcoinin fiyatı*" dizisinde ise her iki kelimenin de ekler aldığı görülmektedir. "*fiyat*" kelimesi her iki değişken için farklı kullanımlara

sahiptir. İçerisinde herhangi bir ek kullanılmadan kelime kökleriyle oluşturulan "*kritik fiyat*" kelime grubu, okunma sayısını -0,913 katsayı ile negatif etkilerken, her iki kelimenin de ek alarak oluşturulduğu "*altcoinin fiyatı*" kelime dizisi ise, okunma sayısına 1,782 katsayı ile olumlu katkı yapmaktadır. Eğer çalışma kelime köklerine indirgenip kelime grupları sadece "*fiyat*" olarak veri kümesine dahil edilseydi, bu değişkenlerin barındırdığı ilişki ortaya çıkarılamayacaktı.

Diğer bir örnek ise Tablo 22'de verilen kökleri aynı olan "*açıkladı*" ve "*açıklama geldi*" ifadeleridir. Aynı köke sahip olan bu kelimelerin başlıklardaki farklı kullanımlarının, farklı sonuçlar verdiği görülmektedir. Bu bağlamda, bu analizde ele alınan problem için köke indirgeme işleminin uyumlu olmadığı görülmüştür. Çünkü analiz sonucunda sadece kelime köklerinden oluşan bir grup kelime ve kelime grubu tespit edilseydi, bu kelimelerin başlık içerisindeki bağlamları hakkında bilgi edinmek mümkün olmayacaktı. Dahası, sadece köklerden oluşan, okunma üzerinde anlamlı etkiye sahip bu kelimeleri, gerçek kullanımda uygulamaya koymak da olanaksız olacaktı. Zira sadece kelime köklerini yan yana getirerek anlamlı başlıklar oluşturmak mümkün değildir. Anlamlı başlıklar oluşturmak için Türkçe dil bilgisine uygun yapım ekleri ve çekim eklerine ihtiyaç vardır.

Tablo 22. Aynı Köke Sahip Değişkenler

Değişkenler	**Katsayılar**
açıkladı	0,983
açıklama geldi	1,902

Dijital içerik yayınlarında başlıklarda seçilen kelimelerin her biri ve hatta kelimelerin sıralaması dahi, okuyucuların ilgisini çekmesi bakımından önemli bir yer tutmaktadır. Aynı konu ve içeriğe ait olarak üretilen iki farklı başlık, okuyucularda farklı tepkilere ve aksiyonlara yol açabilir. Bu nedenle bu çalışmada ele alınan probleme özel, başlık oluştururken hangi kelimelerin hangi eklerle ve hangi diğer kelimelerle birlikte kullanılacağının, önemli bir ayrıntı olduğu ortaya çıkmaktadır.

Model değişkenleri içerisinde yer alan "*bu*" ifadesi diğer bir örnek olarak ele alınacak olunursa, ifadenin modele 0,313 katsayıyla olumlu bir katkı yaptığı görülmektedir. "*bu*" ifadesi genellikle metin madenciliği çalışmalarında tek başına bir anlam ifade etmeyen bir kelime olarak etiketlenip veri kümesi dışına alınmaktadır. Ama bu çalışmanın konu aldığı problemde "*bu*" kelimesinin model için önemli olduğu ortaya çıkarılmıştır. Araştırmaya daha fazla açıklık getirmek için ortaya çıkan teorik bilgi ile pratikteki uygulama karşılaştırılmış ve tutarlı bir analize ulaşabilmek için ek incelemeler yapılmıştır. Örnek olarak veri kümesinde "*bu*" ifadesi geçen üç başlık incelenmiştir, bunlar: "Altcoinlerde Büyük Düşüş Henüz Yaşanmadı Mı? Bu Tarihe Dikkat", "Tehlike Çanları Çalıyor Bu Altcoin Terra (LUNA) İle Aynı Kaderi Paylaşabilir" ve "Bitcoin Ne Zaman Yükselecek? Bu Seviyeye Dikkat! BTC Son Dakika 2022". Üç örnek başlığın ortak özelliğine bakıldığında, hepsinde yer alan "*bu*" ifadesinin kendinden sonra gelen kelime için bir gizlilik kattığı görülmektedir. Birinci başlıkta "*bu altcoin*", ikinci başlıkta "*bu tarihe*" ve üçüncü başlıkta ise "*bu seviyeye*" şeklindeki kullanımların üstü örtülü bilgiler içererek okuyucuda ilgi ve merak uyandırdığı şeklinde değerlendirme yapmak yerinde olacaktır. Okuyucu bu başlıklardan birini gördüğünde eğer konu ilgisini çekiyorsa, verilen bu üç örnek başlık için, hangi altcoin'den bahsedildiği, hangi tarihe dikkat edileceği veya hangi seviyenin söz konusu olduğu gibi bilgilere, ancak içeriğe tıklayarak ulaşabilmektedir. Bu nedenle okuyucu öğrenmek istediği bilgiye ulaşmak için içeriğe tıklayacak ve okumak isteyecektir. Tüm bu sebeplerle, "*bu*" ifadesi tek başına bir anlam ifade etmiyor gibi görünse de, bu çalışmada ele alınan problem için "*bu*" ifadesinin önemli olduğu ortaya çıkmıştır. Eğer çalışmada "*bu*" ifadesi, etkisi olmayan eleman olarak etiketlenip analiz dışına alınsaydı "*bu*" kelimesinin önemi keşfedilemeyecekti.

Modelde 0,439 katsayıyla yer alan "*ve*" ifadesi, "*bu*" ifadesi ile aynı şekilde tek başına bir anlam ifade etmeyen kelime grubuna (stop-word) girmektedir. Konuya daha fazla yorum getirmek amacıyla içerisinde "*ve*" ifadesi geçen, veri kümesi içerisinde yer

alan üç farklı başlık incelenmiştir. Bu başlıklar; "Yatırımcılar Dikkat: Bitcoin ve Ethereum İçin Kritik Seviyeler!", "Piyasaya Güncel Bakış: Zilliqa (ZIL) ve Loopring (LRC) Son Durum!" ve "BTC, ETH, AVAX ve SOL Fiyat Grafikleri: Popüler Kriptolarda Son Durum!" başlıklarıdır. Bu başlıkların ortak özelliklerine bakıldığında, en az iki farklı kripto para hakkında bilgiler içeren başlıklar olduğu görülmektedir. Bu başlıklarda en az iki, en çok dört olmak üzere farklı kripto paralar "*ve*" bağlacıyla tek bir içerikte kombine edilmiştir. Farklı kripto paraların tek bir içerikte kombine edilmesi, içeriği zenginleştirmiş ve okuyucuda bu yazıda diğer içeriklere göre daha fazla bilgi alabileceği kanısı uyandırmıştır. Böylelikle "*ve*" bağlacının farklı kripto paraları kombine etmek için kullanılmasının önemli olduğu ortaya çıkarılmıştır.

Elde edilen bu sonuçlarla, "Başlık metinlerindeki tek başına anlam ifade etmeyen etkisiz kelimeler (stop-word) çalışmanın ele aldığı bu problem için istatistiksel olarak anlamlı mıdır?" ve "Başlık metinlerindeki kelimeler, köklere indirgenmeden ekleriyle beraber analiz edildiğinde, istatistiksel olarak anlamlı sonuçlar ortaya çıkıyor mu?" araştırma sorularına net olarak cevap verilmiştir.

Bu açıklamalar ışığında, çalışmada konu edinilen problem özelinde, metinlerdeki kelimelerin hiçbir değişiklik yapılmadan ele alınmasını ve tek başına anlam ifade etmiyor gibi görünen kelimelerin veri kümesinden çıkarılmamasını içeren bu farklı yaklaşımın, geçerli ve güvenilir olduğu ortaya konmuştur. Bu çalışmanın konu edindiği benzeri veri ve metin madenciliği uygulamalarında, bu tezde gerçekleştirilen yaklaşımın uygulanabilirliği gösterilmiştir.

Tez çalışmasının üçüncü çıktısı olarak içeriklerin aylık okunma sayıları ile şirketin aylık gelirleri arasındaki ilişki ele alınmıştır. Bu çerçevede araştırma sorularından diğer iki tanesi olan, "İçeriklerin okunma sayılarıyla şirket gelirleri arasında bir ilişki var mıdır?" ve "İçeriklerin okunma sayıları ve şirket gelirleri kullanılarak şirket gelirleri için

veri madenciliği yöntem ve teknikleriyle bir tahmin modeli oluşturulabilir mi?" soruları cevaplandırılmıştır. Ayrıca bu analizde "İçeriklerin okunma sayıları ile organizasyon gelirleri arasında bir ilişki bulunmaktadır" hipotezi sınanmıştır.

Bu analiz için ele alınan problemde, aylık bazda yaklaşık beş yıllık şirket gelirleri ve şirketin işlettiği web platformunun yine aylık bazda okunma sayıları arasındaki ilişki modellenmiştir. Bu bağlamda bir modelleme yapılmış ve bir tahmin modeli oluşturulmuştur. Bu analizde farklı yöntemler kullanılarak birden fazla model elde edilmiştir. Bunun nedeni, bir araştırmanın birden çok model ile açıklanabildiğini gösterme ve organizasyonun farklı yöntemler içerisinden kendine en uygun ve kullanılabilir olan modeli seçmesi için alternatifler sunma amacı taşımaktadır. Bu araştırmada elde edilen sonuçlara göre, uygulanan farklı yöntemler içerisinden düzeltilmiş R^2 değeri en yüksek olan model kullanıma uygun olabilir. Aşağıda denklemi verilen önerilen model için %86,20 oranında düzeltilmiş R^2 değeri sağlanmıştır. Bu yüksek düzeltilmiş R^2 oranına sahip model ile gelecek aylar veya yıllar için şirket gelirlerini başarılı bir şekilde tahmin edebilmek mümkündür. Elde edilen bu sonuçla, "İçeriklerin okunma sayıları ile organizasyon gelirleri arasında bir ilişki bulunmaktadır" hipotezi doğrulanmıştır. Önerilen model (logaritmik ölçekte) şu eşitliği vermektedir.

$$gelir = -5{,}142 + 1{,}1374 * okunma\ sayısı - 0{,}000002 * zaman^3$$

Zaman değişkeninin katsayısı çok küçük olduğundan göz ardı edilebilir derecede küçük öneme sahiptir. Bu nedenle bu analiz için nihai model aşağıdaki şekilde kullanıma uygun bulunmuştur.

$$gelir = -5{,}142 + 1{,}1374 * okunma\ sayısı$$

Modelde *okunma sayısı* yukarıda verilen eşitlikteki yerine konulduğunda *gelir* tahmin edilebilir. Bu model ile şirket, gelecek aylarda oluşabilecek gelir akışını tahmin

ederek yeni stratejiler uygulayabilir veya operasyon, yönetim ve planlama konularında stratejik değişiklikler yapabilir. Ayrıca bu model, şirketin ana iş alanı olan içerik oluşturma ile ilgili yapmayı planladığı her türlü değişiklik veya geliştirme için karar destek sistemlerinde kullanılabilir ve böylece şirketin rekabet gücü artabilir. Öte yandan elde edilen bu model, gelecekte yaşanabilecek şirket devri veya hisse satışı gibi konularda, şirket değerlemesinin daha doğru yapılabilmesi için önemli bir bilgi vermektedir. Sektörün büyüme hızı da hesaba katılarak bu model ile yapılacak değerlemeler, daha doğru sonuçlara ulaşabilmek için etkili olabilir. Model, şirket içerisindeki karar vericilere yardımcı olarak şirket için tüm süreçlerde daha doğru ve hızlı kararlar alınmasını sağlayabilir.

Tez çalışmasının son çıktısı olarak, veri madenciliği kavramı ile Bilgi ve Belge Yönetimi alanı içerisinde yer alan bilgi yönetimi kavramı arasındaki bağlantı ele alınmıştır. Bu bağlamda tez çalışmasında, Bilgi ve Belge Yönetimi Bölümü mezun veya öğrencisi olup veri madenciliği alanında çalışma yapmak isteyenler için bir motivasyon oluşturmak amaçlanmıştır. Bu motivasyonu sağlamak için çalışmada veri madenciliği hakkında kavramsal bir altyapı verilmiş ve kapsamlı bir uygulama bölümü hazırlanmıştır.

Çalışmanın uygulama bölümü için seçilen organizasyon içerisinde oluşan ham veri, metin madenciliği ve veri madenciliği süreçleriyle işlenerek ilgili organizasyon için son derece değerli bilgilere ulaşılmıştır. Ham verilerin veritabanından elde edilmesiyle başlayan süreç, metin ön işleme, değişken işlemleri ve analizlerle sonuçlandırılmıştır. Uygulama bölümü, baştan sona bir veri madenciliği uygulamasının tüm adımlarını kapsama aldığından, tezin adı da çalışmanın bu amacına uygun olarak "Bilgi Yönetiminde Veri ve Metin Madenciliği: Bir Dijital İçerik Analizi Uygulaması" şeklinde belirlenmiştir.

Bilgi ve Belge Yönetimi Bölümü'nde yapılan bu doktora tez çalışmasıyla, Bilgi ve Belge Yönetimi Bölümü öğrencilerine ve mezunlarına veri madenciliği kavramı hakkında bir farkındalık oluşturmak amaçlanmıştır. Çalışmada, veri madenciliğinin son yıllarda öne çıkan bir alan olduğu ve Bilgi ve Belge Yönetimi alanının da geri kalmayarak veri madenciliği uygulamaları yapması gerektiği hakkında bir savunma yapılmamaktadır. Bunun ötesinde tezde, veri madenciliği ile Bilgi ve Belge Yönetimi alanı arasında güçlü ve organik bir ilişki olduğu ve bu ilişki hakkında farkındalığın arttırılması üzerine düşünceler vurgulanmaktadır. Bu farkındalıkla çalışmada, potansiyel iş fırsatlarının bölüm öğrencileri veya mezunları tarafından keşfedilmesi ve veri madenciliği alanına ilgi duyan öğrenci ve mezunların, kendilerini veri madenciliği kavramı hakkında geliştirerek bu sektörde istihdam edilmelerinin yolunun açılabileceği düşünülmektedir.

Bu zamana kadar Bilgi ve Belge Yönetimi alanında yapılan veri madenciliği çalışmalarının (1.1. araştırmanın konusu ve önemi kısmında bahsedildiği gibi) özellikle ülkemizde konu edindiği alanların, kütüphaneler, arşivler, belge yönetim sistemleri ve atıf analizleri gibi çalışmalar olduğu görülmektedir. Oysaki veri, enformasyon, bilgi ve bunlarla ilgili diğer kavramlar ile sürekli temas halinde olan Bilgi ve Belge Yönetimi Bölüm öğrencilerinin/mezunlarının, gerekli nitelikleri kazanarak bilgi yönetimi kavramı çerçevesinde pazarlama, eğitim, sağlık, savunma sanayi veya şehircilik örnekleri gibi diğer tüm alanlarda proje ekiplerine dahil olarak, veri madenciliği projelerine katılmalarının önünde bir engel bulunmamaktadır. Bu bağlamda değerlendirildiğinde, hali hazırda bilgi yönetimi kavramının, sektör fark etmeksizin tüm organizasyonlar için gerekli bir süreç olduğu ve bilgi yönetimi ve veri madenciliği kavramlarının organik olarak birbirlerine bağlı olduğu unutulmamalıdır.

Elbette Bilgi ve Belge Yönetimi Bölümlerinde verilen ders içerikleri ve öğrencilerin/mezunların bu alanda kendilerini geliştirme arzuları, yukarıda bahsedilenlerin sağlanabilmesi için önemli faktörlerdir. Türkiye'de eğitim veren Bilgi ve

Belge Yönetimi Bölümlerinin lisans ders programlarına bakıldığında, veri madenciliğine yönelik ders içeriklerinin programlarda yer almaya başladığı görülmektedir. Bu derslerden veri madenciliği kavramı çerçevesinde değerlendirilebilecek olanlar şu şekilde sıralanabilir: Bilgi yönetimi ve matematik, programlama ve algoritmalar, istatistik, metin analitiği, veritabanı yönetim sistemleri, büyük veri ve veri madenciliği dersleridir. Son yıllarda Bilgi ve Belge Yönetimi Bölümlerinin ders içeriklerinin veri madenciliği kavramı çerçevesinde geliştiği görülmektedir. Bu kapsamda bölümler, derslerde teorik anlatımın yanında uygulamalı olarak da veri madenciliği projeleri ile öğrencileri hazırlayabilirler. Öğrenciler dönem boyunca çeşitli veri madenciliği projeleri uygulayarak kendilerini geliştirebilmeliler. Bölümde verilen derslere ek olarak, bu tez çalışmasında da uygulanan tüm veri ve metin madenciliği süreçlerinin doğru bir şekilde yürütülebilmesi için veri madenciliği analistlerinin bazı niteliklere sahip olması gerektiği düşünülmektedir.

Neredeyse tüm ham verilerin veritabanlarında olduğu düşünüldüğünde MySQL gibi veritabanı yönetim sistemlerinin ve sorgular oluşturup verilerin çekilebilmesi için SQL dilinin, gerekli olmamakla birlikte, en azından başlangıç düzeyinde bilinmesi önemli bir avantajdır. Verilerin düzenlemesi ve bazı basit/orta düzeyde veri hazırlama işlemlerinin yapılabilmesi için en az orta seviyede Excel programı hakkında bilgi sahibi olunması gerekmektedir. Çünkü ham veriler genellikle Excel veya muadil dosya formatlarında yer almaktadır. Verilerin işlenmesi, temizlenmesi, değişken işlemleri, veri görselleştirme ve gerektiğinde makine öğrenimi algoritmalarının çalıştırılabilmesi için veri biliminde oldukça fazla kullanılan Python veya R gibi programlama dillerinin bilinmesi, gerekli olmamakla birlikte önemli bir avantaj ve esneklik sağlamaktadır. Eğer bu sağlanamıyorsa, çok sayıda veri madenciliği yöntem ve tekniği içerisinde barındıran Knime veya RapidMiner gibi veri işlemeye olanak sağlayan yazılımların orta veya ileri seviyede bilinmesi gerekmektedir. Bunlara ek olarak istatistiksel test ve analizlerin yapılabildiği Minitab veya SPSS gibi istatistik yazılımlarında orta veya ileri seviyede

çalışabilmek gerekmektedir. Son olarak ise sınıflama, kümeleme veya regresyon gibi veri madenciliği algoritma, yöntem ve tekniklerini öğrenebilmek ve uygulayabilmek için istatistik ve matematik alanlarında disiplinlerarası çalışmaların yapılması gerektiği düşünülmektedir.

Sonuç olarak bu tez çalışmasında, seçilen örnek bir organizasyon için veri ve metin madenciliği analizleri yapılmış ve organizasyon için potansiyel katma değer yaratabilecek değerli bilgiler ortaya çıkarılmıştır. Bu analizler yapılırken farklı bir metin madenciliği yaklaşımı denenmiş ve elde edilen sonuçlar başarılı bulunmuştur. Bu çerçevede, başlık bazlı dijital içerik analiz süreci geliştirilmiş ve bu sürecin benzer veri madenciliği problemleri için kullanılabileceği gösterilmiştir. Bununla birlikte tüm veri ve metin madenciliği süreçlerinin her adımı detaylarıyla açıklanmaya çalışılmış ve Bilgi ve Belge Yönetimi Bölümü öğrenci ve mezunları için uçtan uca örnek bir çalışma olması amaçlanmıştır. Tüm bu çalışmalarla, Bilgi ve Belge Yönetimi Bölümü öğrenci ve mezunlarında, veri madenciliği kavramı hakkındaki farkındalığın artmasıyla birlikte, potansiyel iş fırsatlarının keşfedilmesinin yolunun açılabileceği umulmaktadır.

KAYNAKÇA

Abiteboul, S. (2005). *Querying Semi-Structured Data*. Berlin: Springer. Erişim adresi: https://link.springer.com/chapter/10.1007/3-540-62222-5_33

Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16(1), 3-9.

Agnellutti, C. (2014). *Big Data: An Exploration of Opportunities, Values, and Privacy Issues*. New York: Nova Science Publishers.

Ahlemeyer-Stubbe, A. ve Coleman, S. (2014). *A Practical Guide to Data Mining for Business and Industry*. Chichester, West Sussex: John Wiley and Sons.

Akçay, A. (2014). *Bilgi ve Belge Yönetiminde Veri Madenciliği*. Yayımlanmamış Yüksek Lisans Tezi, İstanbul Üniversitesi, İstanbul. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Akpınar, H. (2014). *Data: Veri Madenciliği Veri Analizi*. İstanbul: Papatya Yayıncılık.

Aktan, E. (2018). Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu. *Bilgi Yönetimi*, 1(1), 1-22. Erişim adresi: https://dergipark.org.tr/tr/pub/by/issue/37228/403010

Albayrak, A.S. (2006). *Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın.

Alkan, N. (2017). Bilgi Yönetimi Süreçlerini Geliştirmede Enformasyon Sistemlerinin Kullanılması: Çalışma Ortamı, İş Özellikleri ve İş-Teknoloji Uyumunun Rolü. *Türk Kütüphaneciliği*, 31(1), 155-167. Erişim adresi: https://dergipark.org.tr/tr/pub/tk/issue/48649/618508

Alma, Ö. G. ve Vupa, Ö. (2008). Regresyon Analizinde Kullanılan En Küçük Kareler ve En Küçük Medyan Kareler Yöntemlerinin Karşılaştırılması. *SDÜ Fen Edebiyat Fakültesi Fen Dergisi*, 3(2), 219-229. Erişim adresi: https://dergipark.org.tr/tr/pub/sdufeffd/issue/11268/134658

Al-Shawi, A. (2011). Data Mining Techniques for Information Security Applications. *Wıres Computational Statistics*, 3(3), 221- 229. Erişim adresi: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.161

Altınok, G., Karagöz, P., ve Batmaz, İ. (2021). Learning to Rank by Using Multivariate Adaptive Regression Splines and Conic Multivariate Adaptive Regression Splines.

Computational Intelligence, 37(1), 371-408. Erişim adresi: https://onlinelibrary.wiley.com/doi/abs/10.1111/coin.12413

Altunkaynak, B. (2017). *Veri Madenciliği Yöntemleri ve R Uygulamaları: Kavramlar, Modeller, Algoritmalar*. Ankara: Seçkin Yayıncılık.

Anaconda Navigator. (07.11.2022). *Anaconda Documentation*. Erişim adresi: https://docs.anaconda.com/navigator/

Anderson, R.K. (2013). *Visual Data Mining: The Vismíner Approach*. Chichester, West Sussex: Wiley.

Ankaralı, E. ve Külcü, Ö. (2020). Rapidminer ile Twitter Verilerinin Konu Modellemesi. *Bilgi Yönetimi*, 3(1), 1-10. Erişim adresi: https://dergipark.org.tr/tr/pub/by/issue/54402/641878

Argüden, Y. ve Erşahin, B. (2008). *Veri Madenciliği: Veriden Bilgiye, Masraftan Değere.* İstanbul: ARGE Danışmanlık Yayınları.

Arslantekin, S. (2003). Veri Madenciliği ve Bilgi Merkezleri. *Türk Kütüphaneciliği*, 17(4), 369-380. Erişim adresi: http://www.tk.org.tr/index.php/TK/article/view/303/295

Atan, S. (2016). Veri, Büyük Veri ve İşletmecilik. *Balıkesir University Journal of Social Sciences Institute*, *19*(35), 137-153. Erişim adresi: https://dergipark.org.tr/tr/pub/baunsobed/issue/50168/645312

Attewell, P.A. ve Monaghan, D.B. (2015). *Data Mining for The Social Sciences: An Introduction*. California: University of California Press. Erişim adresi: https://www.jstor.org/stable/10.1525/j.ctt13x1gcg

Azzalini, A., Walton, G. ve Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*. Oxford: Oxford University Press. Erişim adresi: https://onlinelibrary.wiley.com/doi/10.1111/insr.12011_23

Bach, M. P., Krstic, Z., Seljan, S. ve Turulja, L. (2019). Text Mining for Big Data Analysis in Financial Sector: A Literature Review. *Sustainability*, 11(5), 1277-1304. Erişim adresi: https://www.mdpi.com/2071-1050/11/5/1277

Baerns, M. ve Holeňa, M. (2009). *Combinatorial Development of Solid Catalytic Materials: Design of High-Throughput Experiments, Data Analysis, Data Mining*. London: Imperial College Press. Erişim adresi: https://www.worldscientific.com/worldscibooks/10.1142/p620

Barutçugil, İ. (2002). *Bilgi Yönetimi*. İstanbul: Kariyer Yayınları.

Bassan, A.S. ve Sarkar, D. (2014). *Mastering SQL Server 2014 Data Mining*. Birmingham: Packt Publishing.

Batmaz, İ., Danışoğlu, S., Kartal-Koç, E., ve C. Yazıcı. (2017). A Data Mining Application to Deposit Pricing: Main Determinants and Prediction Models. *Applied Soft Computing*, 60, 808-819. Erişim adresi: https://www.sciencedirect.com/science/article/abs/pii/S1568494617304672?via%3Dihub

Batmaz, İ., Karagöz, P. ve G. Serdar. (2017). A Comparative Study on Learning to Rank with Computational Methods. *2017 IEEE International Conference on Big Data (IEEE Big Data 2017),* Boston, USA, 1898-1906. Erişim adresi: https://ieeexplore.ieee.org/document/8258135

Berry, M.W. ve Castellanos, M. (2008). *Survey of Text Mining II: Clustering, Classification, and Retrieval*. London: Springer-Verlag.

Bifet, A. (2010). *Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams*. Amsterdam: IOS Press.

Bifet, A. (2013). Mining Big Data in Real Time. *Informatica*, 37(1), 15-20. Erişim adresi: https://www.informatica.si/index.php/informatica/article/view/428/432

Binici, K. (2016). *Kütüphane ve Bilgi Bilimi Çalışmalarında Dönemsel Konu Analizi*. Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi, İstanbul. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Box, G.E.P. ve Cox, D.R. (1964). An Analysis of Transformations. *Journal of The Royal Statistical Society, Series B (Methodological),* 26(2), 211-252. Erişim adresi: https://www.jstor.org/stable/2984418

Browne, M. ve Berry, M.W. (2006). *Lecture Notes in Data Mining*. Singapore: World Scientific. Erişim adresi: https://www.worldscientific.com/worldscibooks/10.1142/6103#t=aboutBook

Capri, H.L. (2015). *Data Mining: Principles, Applications and Emerging Challenges*. New York: Nova Science Publishers.

Catarci, T. (2010). *User-Centered Data Management.* California: Morgan ve Claypool.

Catlett, C. (2013). *Cloud Computing and Big Data*. Amsterdam: IOS Press.

Chapman, P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer, C. ve Wirth, R. (2000). *CRISP-DM 1.0: Step-By-Step Data Mining Guide*. SPSS Inc. Erişim adresi: http://www.statoo.com/CRISP-DM.pdf

Ching, W.K. ve Kwok-Po Ng, M. (2003). *Advances in Data Mining and Modeling*. Singapore: World Scientific. Erişim adresi: https://www.worldscientific.com/worldscibooks/10.1142/5260#t=aboutBook

Cibaroğlu, M. O. ve Yalçınkaya, B. (2019). Belge ve Arşiv Yönetimi Süreçlerinde Büyük Veri Analitiği ve Yapay Zeka Uygulamaları. *Bilgi Yönetimi*, 2(1), 44-58. Erişim adresi: https://dergipark.org.tr/tr/pub/by/issue/45460/570634

Cochrane, D. ve Orcutt, G. H. (1949). Application of Least Squares Regression to Relationships Containing Auto- Correlated Error Terms. *Journal of The American Statistical Association*, 44(245), 32-61. Erişim adresi: https://www.jstor.org/stable/2280349

Cox, E. (2005). *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*. Amsterdam: Morgan Kaufmann. Erişim adresi: https://www.sciencedirect.com/book/9780121942755/fuzzy-modeling-and-genetic-algorithms-for-data-mining-and-exploration

Craney, T A. ve Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, 14(3), 391-403. Erişim adresi: https://www.tandfonline.com/doi/abs/10.1081/QEN-120001878

Çakmak, T. ve Eroğlu, Ş. (2020). Sosyal Medyada Kullanıcı Etkileşimi ve İçerik Kategorizasyonu: Ankara'daki Halk Kütüphanelerinin Facebook Gönderilerinin Analizi. *Türk Kütüphaneciliği*, 34(2), 160-186. Erişim adresi: https://dergipark.org.tr/en/pub/tk/article/706882

Çakmak, T. ve Tırnavalı, S. (2020). Danışma Hizmetlerinin Değerlendirilmesi: Türkiye Büyük Millet Meclisi Kütüphanesi Danışma Hizmetlerine Yönelik Bir Araştırma. *Bilgi Dünyası*, 21(1), 7-33. Erişim adresi: https://bd.org.tr/index.php/bd/article/view/540

Çalışkan, S. ve Can, F. (2018). Türkçe Metinler Üzerine Yapılan Sayısal Üslup Araştırmalarını İnceleyen ve Benim Adım Kırmızı Çevirilerinin Aslına Olan Sadakatini Ölçen Bir Çalışma. *Türk Kütüphaneciliği*, 32(4), 251-286. Erişim adresi: http://www.tk.org.tr/index.php/TK/article/view/2967

Çelik, A.E.Ö. (2021). *Türkçe Akademik Yayınlar İçin Yapısal Öz Çıkarım Sistemi*. Yayımlanmamış Doktora Tezi, Hacettepe Üniversitesi, Ankara. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Çelik, U. Akçetin, E. ve Gök, M. (2017). *Rapid Miner ile Veri Madenciliği*. İstanbul: Pusula.

Çokluk, Ö. (2010). Lojistik Regresyon Analizi: Kavram ve Uygulama. *Kuram ve Uygulamada Eğitim Bilimleri,* 10(3), 1357-1407. Erişim adresi: https://www.idealonline.com.tr/IdealOnline/lookAtPublications/paperDetail.xhtml?uId=2144

Çölkesen, T.R. (2016). *Algoritma Geliştirme ve Veri Yapıları*. İstanbul: Papatya Yayıncılık.

Darwish, F. (2019). *Towards Intelligent Diabetes Knowledge Management and Knowledge Discovery: A Data Mining Approach*. Yayımlanmamış Doktora Tezi, University of Salford, Manchester. Erişim adresi: https://search.proquest.com/

Davidian, M. ve Louis, T.A. (2012). Why Statistics? *Science*, 336(6077), 12-12. Erişim adresi: https://www.science.org/doi/10.1126/science.1218685

De Ville, B. (2001). *Microsoft Data Mining: Integrated Business Intelligence for E-Commerce and Knowledge Management*. Boston: Digital Press.

Doğan, K. (2022). *Bilgi Merkezi ve Hizmetlerinde Veri Madenciliğinin Kullanilabilirliği: Üniversite Kütüphaneleri Örneği.* Yayımlanmamış Doktora Tezi, Ankara Üniversitesi, Ankara. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Doğan, K. (2015). *Büyük Verinin Üniversite Kütüphaneleri Açısından Önemi ve Kütüphane Hizmetlerinde Kullanımı: Ankara Üniversitesi Kütüphaneleri Örneği*. Yayımlanmamış Yüksek Lisans Tezi, Ankara Üniversitesi, Ankara. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Doğan, K. ve Arslantekin, S. (2016). Büyük Veri: Önemi, Yapısı ve Günümüzdeki Durum. *DTCF Dergisi*, 56(1), 15-36. Erişim adresi: https://dergipark.org.tr/tr/pub/dtcfdergisi/issue/66793/1044418

Dolgun, M.Ö., Özdemir, T.G. ve Oğuz, D. (2009). Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği. *İstatistikçiler Dergisi*, 2(2), 48-58. Erişim adresi: https://dergipark.org.tr/tr/pub/jssa/issue/10041/123866

Efe, A. (2022). Yargısal ve Hukuki Süreçlerde Yapay Zeka Kullanan Araçlar Üzerine Bir Araştırma. *Bilgi Yönetimi*, 5(1), 92-117. Erişim adresi: https://dergipark.org.tr/tr/pub/by/issue/69559/971873

Eker, H.S. (2022). Sosyal Ağlar Büyük Veriden Nasıl Yararlanır: Facebook ve Twitter. *Bilgi Yönetimi*, 5(1), 118-130. Erişim adresi: https://dergipark.org.tr/en/pub/by/article/983553

Emre, İ.E. ve Erol, Ç.S. (2017). Veri Analizinde İstatistik mi Veri Madenciliği mi? *International Journal of Informatics Technologies*, 10(2), 161-167. Erişim adresi: https://dergipark.org.tr/tr/pub/gazibtd/issue/28916/309297

Erdem, S. ve Özdağoğlu, G. (2008). Ege Bölgesi'ndeki Bir Araştırma ve Uygulama Hastanesinin Acil Hasta Verilerinin Veri Madenciliği ile Analiz Edilmesi. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 9(2), 261-270. Erişim adresi: https://earsiv.anadolu.edu.tr/xmlui/handle/11421/1645

Fayyad, U., Shapiro, G.P. ve Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-53. Erişim adresi: https://ojs.aaai.org//index.php/aimagazine/article/view/1230

Fayyad, U., Shapiro, G.P. ve Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39(11), 27-34. Erişim adresi: https://dl.acm.org/doi/10.1145/240455.240464

Feldman, R. ve Sanger, J. (2006). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press. Erişim adresi: https://www.cambridge.org/core/books/text-mining-handbook/0634B1DF14259CB43FCCF28972AE4382

Ferreira, I. M. M. P. (2021). *Predictive Models for Shop Floor Optimization in The Agrofood Industry*. Yayımlanmamış Doktora Tezi, Universidade Do Porto, Porto. Erişim adresi: https://search.proquest.com/

Ghani, R. ve Soares, C. (2010). *Data Mining for Business Applications*. Amsterdam: IOS Press.

Grabczewski, K. (2014). *Meta-Learning in Decision Tree Induction*. Newyork: Springer International Publishing. Erişim adresi: https://link.springer.com/book/10.1007/978-3-319-00960-5

Grandinetti, L. (2015). *Big Data and High-Performance Computing*. Amsterdam: IOS Press.

Gülseçen, S. (2014). Bir Değer Olarak Bilgi ve Bilginin Yönetimi. *Türk Kütüphaneciliği*, 28(1), 62-68. Erişim adresi: https://dergipark.org.tr/tr/pub/tk/issue/48826/621979

Gürdal, O. (1991). Enformasyon: Terminolojik Yaklaşım. *Türk Kütüphaneciliği,* 5(3), 116-121. Erişim adresi: http://www.tk.org.tr/index.php/TK/article/view/1241

Gürdal, O. (1998). Tekstil Endüstrisinin Enformasyon Gereksinimleri: Metodolojik Yaklaşım. *Türk Kütüphaneciliği,* 12(1), 33-46. Erişim adresi: http://www.tk.org.tr/index.php/TK/article/view/729

Gürdal Tamdoğan, O. (2009). Enformasyon Zincirinde Bilgi Erişim Sistemleri, Bilgi Erişim Sürecinde Kütüphane Kurumu ve Diğer Bilgi Merkezleri. *Türk Kütüphaneciliği*, 23(1), 151-168. Erişim adresi: http://www.tk.org.tr/index.php/tk/article/view/463

Gürsakal, N. (2014). *Büyük Veri*. Bursa: Dora.

Hamde, M.A. (2018). *Kurumsal Belgelere (Metin Verilerine) Metin Madenciliği Tekniği ile Erişimin Değerlendirilmesi: Türk Özel Sektörüne Yönelik Bir İnceleme*. Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi, İstanbul. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Han, J., Kamber, M. ve Pei, J. (2012). *Data Mining: Concepts and Techniques*. Burlington, MA: Morgan Kaufmann. Erişim adresi: https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician*, 52(2), 112-118. Erişim adresi: https://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480549

He, W., Zha, S. ve Li, L. (2013). Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry. *International Journal of Information Management*, 33(3), 464-472. Erişim adresi: https://www.sciencedirect.com/science/article/abs/pii/S0268401213000030

Hooman, A., Marthandan, G., Yusoff, W.W., Omid, M. ve Karamizadeh, S. (2016). Statistical and Data Mining Methods in Credit Scoring. *Journal of Developing Areas*, 50(5), 371-381. Erişim adresi: https://www.jstor.org/stable/26415601

Hornick, M.F., Venkayala, S. ve Marcadé, E. (2007). *Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for Architecture, Design, and Implementation*. Amsterdam: Morgan Kaufmann.

Irmak, S., Köksal, C.D. ve Asilkan, Ö. (2012). Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleri ile Tahmin Edilmesi. *Uluslararası Alanya İşletme Fakültesi Dergisi,* 4(1), 101-114. Erişim adresi: https://dergipark.org.tr/en/pub/uaifd/issue/21593/231863

Işıklı, Ş. (2014). Büyük Veri, Epistemoloji ve Etik Tartışmalar. *AJIT-E: Online Academic Journal of Information Technology*, *5*(17), 89-122. Erişim adresi: https://dergipark.org.tr/tr/pub/ajit-e/issue/54437/740935

Jivani, A. G. (2011). A Comparative Study of Stemming Algorithms. *International Journal of Computer Applications in Technology,* 2(6), 1930-1938. Erişim adresi: https://kenbenoit.net/assets/courses/tcd2014qta/readings/Jivani_ijcta2011020632.pdf

Jo, T. (2019). *Text Mining: Concepts, Implementation, and Big Data Challenge*. Springer. Erişim adresi: https://link.springer.com/book/10.1007/978-3-319-91815-0

Jupyterlab. (08.11.2022). *Jupyter*. Erişim adresi: https://jupyter.org

Kanji, G. K. (2006). *100 Statistical Tests*. California: SAGE. Erişim adresi: https://link.springer.com/article/10.1007/s00362-008-0145-0

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, N.J.: John Wiley.

Karasar, N. (2022). *Bilimsel Araştırma Yöntemi: Kavramlar İlkeler Teknikler*. Ankara: Nobel.

Kaya, A. (2021). Bir Araştırma Kaynağı Olarak Arşivlenen Sosyal Medya Verilerinin Kullanımı. *Bilgi ve Belge Araştırmaları*, 16, 49-79. Erişim adresi: https://dergipark.org.tr/tr/pub/bel/issue/67331/1008002

Keskintürk, T. ve Şahin, S. (2009). Doğrusal Olmayan Regresyon Analizinde Gerçek Değer Kodlamali Genetik Algoritma. *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi*, 8(15), 167-178. Erişim adresi: https://ticaret.edu.tr/uploads/kutuphane/dergi/s15/167-178.pdf

Knobbe, A.J. (2006). *Multi-Relational Data Mining*. Amsterdam: IOS Press.

Kofler, M. (2004). *What Is MySQL? In: The Definitive Guide to MySQL*. Berkeley: Apress. Erişim adresi: https://link.springer.com/chapter/10.1007/978-1-4302-0669-9_1

Kovalerchuk, B. ve Vityaev, E. (2002). *Data Mining in Finance: Advances in Relational and Hybrid Methods*. New York: Kluwer Academic Publishers.

Köktürk, F., Ankarali, H. ve Sümbüloğlu, V. (2009). Veri Madenciliği Yöntemlerine Genel Bakış. *Türkiye Klinikleri Biyoistatistik Dergisi,* 1(1): 20-25. Erişim adresi: https://www.turkiyeklinikleri.com/article/tr-veri-madenciligi-yontemlerine-genel-bakis-53374.html

Krishnan, K. (2013). *Data Warehousing in The Age of Big Data*. Amsterdam: Morgan Kaufmann. Erişim adresi: https://www.sciencedirect.com/book/9780124058910/data-warehousing-in-the-age-of-big-data

Kumar, A. ve Paul, A. (2016). *Mastering Text Mining with R*. Birmingham: Packt Publishing.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, N.J.: Wiley-Interscience. Erişim adresi: https://onlinelibrary.wiley.com/doi/book/10.1002/0471687545

Liu, S., Mcgree, J., Ge, Z. ve Xie, Y. (2016). *Computational and Statistical Methods for Analyzing Big Data with Applications.* London: Elsevier.

Maclennan, J., Crivat, B. ve Tang, Z. (2009). *Data Mining with Microsoft Sql Server 2008*. Indianapolis, IN: Wiley.

Makhabel, B. (2015). *Learning Data Mining with R*. Birmingham: Packt Publishing.

Mannan, S. (2012). *Less' Loss Prevention in the Process Industries: Hazard Identification, Assessment and Control*. Texas: Elsevier. Erişim adresi: https://www.sciencedirect.com/book/9780123971890/lees-loss-prevention-in-the-process-industries

Matplotlib. (09.11.2022). *Matplotlib*. Erişim adresi: https://matplotlib.org

Maulud, D. A. ve Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(4), 140 –147. Erişim adresi: https://jastt.org/index.php/jasttpath/article/view/57

Mccue, C. (2007). *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*. Amsterdam: Butterworth-Heinemann.

Melton, J., ve Simon, A. R. (1993). *Understanding the New SQL: A Complete Guide*. California: Morgan Kaufmann.

Mena, J. (2003). *Investigative Data Mining for Security and Criminal Detection*. Amsterdam: Butterworth-Heinemann.

Minitab. (07.11.2022). *Minitab Documents*. Erişim adresi: https://www.minitab.com/content/dam/www/en/uploadedfiles/documents/brochures/Minitab-Brochure_EN.pdf

Montgomery, D. C., Peck, E. A. ve Vining, G. G. (2013). *Doğrusal Regresyon Analizine Giriş*. Ankara: Nobel.

Motoda, H. (2002). *Active Mining: New Directions of Data Mining*. Amsterdam: IOS Press.

Nguyen, Q.V. (2019). *Value of Information in Agro-Food Logistics Management: A Case of The Dutch Floriculture Supply Chain Network*. Yayımlanmamış Doktora Tezi, Wageningen University, Wageningen. Erişim adresi: https://search.proquest.com/

Nisbet, R., Miner, G. ve Elder, J.F. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Amsterdam: Academic Press. Erişim adresi: https://www.elsevier.com/books/handbook-of-statistical-analysis-and-data-mining-applications/nisbet/978-0-12-374765-5

Nobre, J.S. ve Singer, J.D.M. (2007). Residual Analysis for Linear Mixed Models. *Biometrical Journal*, 49(6), 863–875. Erişim adresi: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200610341

Nosakhare, U.H. ve Bright, A.F. (2017). Evaluation of Techniques for Univariate Normality Test Using Monte Carlo Simulation. *American Journal of Theoretical and Applied Statistics*, 6(5-1), 51-61. Erişim adresi: https://www.sciencepublishinggroup.com/journal/paperinfo?journalid=146&doi=10.11648/j.ajtas.s.2017060501.18

NumPy. (08.11.2022). *What is NumPy?* Erişim adresi: https://numpy.org/devdocs/user/whatisnumpy.html

O'Byrne, S. (2013). *Data Security, Data Mining, and Data Management: Technologies and Challenges*. New York: Nova Science Publishers.

Osborne, J. (2010). Improving Your Data Transformations: Applying the Box-Cox Transformation. *Practical Assessment, Research, and Evaluation*, 15(1), 1-9. Erişim adresi: https://scholarworks.umass.edu/pare/vol15/iss1/12/

Özdoğan, O. (2016). *Büyük Veri Denizi: Veri Yönetimi Hakkında Her Şey*. Ankara: Elma Yayınevi.

Özkan, Y. (2016). *Veri Madenciliği Yöntemleri*. İstanbul: Papatya Yayıncılık.

Öztemel, E. (2016). *Yapay Sinir Ağları*. İstanbul: Papatya Yayıncılık.

Öztürk, F. ve Özel, N. (2021). Yapay Zeka ve Kütüphaneler. *Bilgi Dünyası*, 22(2), 351-386. Erişim adresi: https://bd.org.tr/index.php/bd/article/view/648

Öztürk, H. (2021). Arşivler ve Yapay Zeka. *Bilgi Yönetimi*, 4(2), 283-300. Erişim adresi: https://dergipark.org.tr/tr/pub/by/issue/64615/987197

Pandas. (09.11.2022). *Pandas*. Erişim adresi: https://pandas.pydata.org

Pektaş, A.O. (2013). *SPSS ile Veri Madenciliği*. İstanbul: Dikeyeksen.

Plaza, B. (2011). Google Analytics for Measuring Website Performance. *Tourism Management*, 32 (3), 477-481. Erişim adresi: https://www.sciencedirect.com/science/article/abs/pii/S0261517710000622

Pyle, D. (2003). *Business Modeling and Data Mining*. Amsterdam: Morgan Kaufmann. Erişim adresi: https://www.elsevier.com/books/business-modeling-and-data-mining/pyle/978-1-55860-653-1

Python. (07.11.2022). *General Python FAQ*. Erişim adresi: https://docs.python.org/3/faq/general.html#what-is-python

Refaat, M. (2007). *Data Preparation for Data Mining Using SAS*. Amsterdam: Morgan Kaufmann. Erişim adresi: https://www.sciencedirect.com/book/9780123735775/data-preparation-for-data-mining-using-sas

Rokach, L. ve Maimon, O. Z. (2008). *Data Mining with Decision Trees: Theory and Applications*. Singapore: World Scientific. Erişim adresi: https://www.worldscientific.com/worldscibooks/10.1142/6604#t=aboutBook

Rowley, J. (2007). The Wisdom Hierarchy: Representations of The DIKW Hierarchy. *Journal of Information Science*, 33(2), 163–180. Erişim adresi: https://journals.sagepub.com/doi/10.1177/0165551506070706

Safran, B. (2015). Bilgi Yönetilebilir mi? Kavramsal Bir Yaklaşım. *Türk Kütüphaneciliği.* 29(3), 550-565. Erişim adresi: https://dergipark.org.tr/tr/pub/tk/issue/48741/620093

Savaş, S., Topaloğlu, N. ve Yılmaz, M. (2012). Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 11(21), 1-23. Erişim adresi: https://dergipark.org.tr/tr/pub/ticaretfbd/issue/21361/229131

Savin, N. E. ve White, K. J. (1977). The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. *Econometrica*, 45(8), 1989–1996. Erişim adresi: https://www.jstor.org/stable/1914122

Sever, H., ve Oğuz, B. (2002). Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorguları ve Algoritmalar. *Bilgi Dünyası*, 3(2), 173-204. Erişim adresi: https://bd.org.tr/index.php/bd/article/view/517

Sever, H., ve Oğuz, B. (2003). Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım: Kısım II- Eşleştirme Sorgularının Biçimsel Kavram Analizi ile Modellenmesi. *Bilgi Dünyası*, 4(1), 15-44. Erişim adresi: https://bd.org.tr/index.php/bd/article/view/509

Shafique, U. ve Qaiser, H. (2014). Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12(1), 217-222. Erişim adresi: http://www.ijisr.issr-journals.org/abstract.php?article=IJISR-14-281-04

Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları*. Papatyabilim: İstanbul.

Silwattananusarn, T. ve Tuamsuk, K. (2012). Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining and Knowledge Management Process (IJDKP)*, 2(5), 13-24. Erişim adresi: https://aircconline.com/ijdkp/V2N5/2512ijdkp02.pdf

Singh, M. ve Hetlevik, S. (2017). Data–Information–Knowledge Hierarchy Based Decision Support System for Risk Based Inspection Analysis. *International Journal of System Assurance Engineering and Management*, 8(2), 1588–1595. Erişim adresi: https://link.springer.com/article/10.1007/s13198-017-0631-7

Sklearn. (09.11.2022). *Sklearn*. Erişim adresi: https://scikit-learn.org

Soares, C. (2008). *Applications of Data Mining in E-Business and Finance*. Amsterdam: IOS Press.

Sohrabi, M.K. ve Akbari, S. (2016). A Comprehensive Study on The Effects of Using Data Mining Techniques to Predict Tie Strength. *Computers in Human Behavior*, 60, 534-54. Erişim adresi: https://www.sciencedirect.com/science/article/abs/pii/S0747563216301558

Spendler, L.I. (2010). *Data Mining and Management*. New York: Nova Science Publishers.

Sykes, A. O. (1993). An Introduction to Regression Analysis. *Law and Economics Working Papers*, 20, 1-33. Erişim adresi: https://chicagounbound.uchicago.edu/law_and_economics/51/

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

Tabachnick, B. G. ve Fidell, L. S. (2019). *Using Multivariate Statistics*, Boston: Pearson.

Takçı, H. ve Soğukpınar, İ. (2002). Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi. *Bilgi Dünyası*, 3(1), 12-26. Erişim adresi: https://bd.org.tr/index.php/bd/article/view/522

Taşkın, Z. (2017). *İçerik Tabanlı Atıf Analizi Modeli Tasarımı: Türkçe Atıflar İçin Metin Kategorizasyonuna Dayalı Bir Uygulama*. Yayımlanmamış Doktora Tezi, Hacettepe Üniversitesi, Ankara. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Tonta, Y. (2001). Bilgi Erişim Sorunu. *21. Yüzyıla Girerken Enformasyon Olgusu Ulusal Sempozyum Bildirileri, 19-20 Nisan 2001, Hatay*. Erişim adresi: http://eprints.rclis.org/9468/

Tsiptsis, K. ve Chorianopoulos, A. (2010). *Data Mining Techniques in CRM: Inside Customer Segmentation*. Hoboken, N.J.: Wiley.

Tufféry, S. (2011). *Data Mining and Statistics for Decision Making*. Hoboken, NJ.: Wiley. Erişim adresi: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470979174

Tunalı, S. ve Batmaz, İ. (2003). A Metamodeling Methodology Involving Both Qualitative and Quantitative Input Factors. *European Journal of Operational Research*, 150(2), 437–450. Erişim adresi: https://www.sciencedirect.com/science/article/abs/pii/S0377221702005210

Tunalı, S. ve Batmaz, İ. (2000). Dealing with the Least Squares Regression Assumptions in Simulation Metamodeling. *Computers & Industrial Engineering*, 38(2), 307–

320. Erişim adresi: https://www.sciencedirect.com/science/article/abs/pii/S0360835200000462

Türk Dil Kurumu. (2020). *Bilgi* [Güncel Web Sözlük]. Erişim adresi: https://sozluk.gov.tr/

Tüzüntürk, S. (2010). Veri Madenciliği ve İstatistik. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 29(1), 65-90. Erişim adresi: https://acikerisim.uludag.edu.tr/handle/11452/17947

Uysal, M. ve Günay, S. (2001). Durbin-Watson Ölçütüne Göre Kararsızlık Bölgesinde Bulunan Negatif Otokorelasyon İçin Bazı Testler. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 2(2), 277-284. Erişim adresi: https://kutuphane.dogus.edu.tr/mvt/pdf.php?pdf=0003791&lng=0

Ünal, F. (2015). *Büyük Veri ve Semantik*. İstanbul: Abaküs.

Voulgaris, Z. (2014). *Data Scientist: The Definitive Guide to Becoming a Data Scientist*. Basking Ridge, NJ: Technics Publications.

Vukasinov, T. ve Draskovic, N. (2013). *Knowledge Management: Technology, Applications and Impact*. New York: Nova Science Publishers.

Wallace, D. P. (2007). *Knowledge Management: Historical and Cross-Disciplinary Themes*. Westport: Libraries Unlimited.

Wankhade, K. Patka, S. ve Thool, R. (2013). An Overview of Intrusion Detection Based On Data Mining Techniques. *2013 International Conference On Communication Systems and Network Technologies, Gwalior*, 626-629. Massachusetts: IEEE Computer Society. Erişim adresi: https://dl.acm.org/doi/10.1109/CSNT.2013.134

Weiss, S.M., Indurkhya, N. ve Zhang, T. (2010). *Fundamentals of Predictive Text Mining*. London: Springer-Verlag London Limited.

Witten, I.H., Frank, E. ve Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann. Erişim adresi: https://www.sciencedirect.com/book/9780123748560/data-mining-practical-machine-learning-tools-and-techniques

Wu, X., Kumar, V., Quinlan, J.R. ve Diğerleri. (2008). *Top Ten Algorithms in Data Mining.* London: Springer. Erişim adresi: https://link.springer.com/article/10.1007/s10115-007-0114-2

Yan, X. ve Su, X. G. (2009). *Linear Regression Analysis Theory and Computing*. New York: World Scientific Publishing.

Yıldırım, B. F. ve Özdemirci, F. (2019). Kurumlarda Örtük Bilginin Yapay Zekâ Destekli Tavsiye Sistemleri Aracılığıyla Ortaya Çıkarılması. *Bilgi Yönetimi*, 2(1), 34-43. Erişim adresi: https://dergipark.org.tr/tr/pub/by/issue/45460/544239

Yıldırım, Y. (2019). Endüstri 4.0'a Kapsamlı Bir Bakış: 2011'den Bugüne. *Bilgi Dünyası*, 20(2), 217-249. Erişim adresi: https://bd.org.tr/index.php/bd/article/view/6

Zhang, F., Fleyeh, H., Wang, X. ve Lu, M. (2019). Construction Site Accident Analysis Using Text Mining and Natural Language Processing Techniques. *Automation In Construction*, 99, 238-248. Erişim adresi: https://www.sciencedirect.com/science/article/abs/pii/S0926580518306137

Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies*. Burlington: Academic Press. Erişim adresi: https://www.sciencedirect.com/book/9780123969637/r-and-data-mining

ÖZET

KURT, Levent

Bilgi Yönetiminde Veri ve Metin Madenciliği: Bir Dijital İçerik Analizi Uygulaması

Doktora Tezi

Danışman: Prof. Dr. Oya GÜRDAL

Bu çalışmada, organizasyonlarda bilgi yönetimi kavramı kapsamında dijital içerik yayınlayan bir organizasyondan elde dilen ham veriler işlenerek ilgili organizasyon için iki farklı analiz yapılmış ve analizler sonucu değerli bilgiler üretilmiştir. Araştırmanın ilk analizinde, yayınlanan içeriklerin okunma sayılarını etkileyen faktörleri ortaya çıkarmak amacıyla, içerik metinleri farklı bir metin madenciliği yaklaşımıyla çözümlenmiş ve çoklu doğrusal regresyon analizi uygulanarak okunma sayılarını etkileyen faktörler ortaya çıkarılmıştır. Ayrıca bu analizde ele alınan problem özelinde uygulanan farklı metin madenciliği yaklaşımının istatistiksel olarak geçerli olduğu gösterilmiştir. Ek olarak, bu araştırma için başlık bazlı dijital içerik analiz süreci geliştirilmiştir. İkinci analizde ise, okunma sayılarının ilgili organizasyonun gelirleri üzerindeki etkisi araştırılmıştır. Bu amaçla, aylık bazda şirket gelirleri ve okunma sayıları, basit doğrusal regresyon problemi olarak ele alınmış ve bir tahmin modeli oluşturulmuştur. Son olarak çalışmada, ham verinin elde edilmesi, verilerin işlenerek değişkenlerin yaratılması, sonrasında analizlerin yapılması ve sonucun yorumlanmasını da içeren tüm süreçler tüm detaylarıyla açıklanmaya çalışılarak bu tezin bilgi yönetimi alanında uçtan uca bir veri madenciliği uygulama örneği olması hedeflenmiştir.

Anahtar Kelimeler: “veri madenciliği”, “metin madenciliği”, “bilgi yönetimi”, “çoklu doğrusal regresyon”, “dijital içerik analizi”.

SUMMARY

KURT, Levent

Data and Text Mining in Information Management: A Digital Content Analysis Application

PhD Thesis

Advisor: Prof. Dr. Oya GÜRDAL

In this study, the raw data collected from an organization that publishes digital content was processed within the context of information management in the organization. Two separate analyses were carried out for the organization and valuable information was obtained as a result. In the first analysis, with the aim of finding out the factors affecting the reading rates for the published content, texts were analyzed with a particular text mining method and the factors determining the reading rates were revealed with the multiple linear regression analysis. Moreover, it was demonstrated that the particular text mining approach applied to the problem in this analysis is statistically valid. Additionally, a title based digital content analysis process was developed for this research. In the second analysis, the effect of the reading rates on the organization's income was evaluated. For this purpose, monthly company revenue and the reading counts were considered as a simple linear regression problem and an estimation model was established. Finally, an attempt was made to explain all the processes, from the acquisition of the raw data to the creation of variables with data processing, as well as the analyses and the interpretation of the results. Thus, the thesis aims to be an example for the application of end-to-end data mining in the field of information management.

Keywords: “data mining”, “text mining”, “information management”, “multiple linear regression”, “digital content analysis”.