T.C.
PAMUKKALE ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI

DOKTORA TEZİ

ÇEVRİMİÇİ İŞ İLANLARININ VERİ VE METİN MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ: BİLGİ VE İLETİŞİM SEKTÖRÜ ÖRNEĞİ

Hazırlayan
Mustafa Onur KAÇAROĞLU

Danışman
Prof. Dr. Arzu ORGAN

Temmuz 2023
DENİZLİ

ÇEVRİMİÇİ İŞ İLANLARININ VERİ VE METİN MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ: BİLGİ VE İLETİŞİM SEKTÖRÜ ÖRNEĞİ

Pamukkale Üniversitesi
Sosyal Bilimler Enstitüsü
Doktora Tezi
İşletme Anabilim Dalı
Genel İşletme Programı

Mustafa Onur KAÇAROĞLU

Danışman: Prof. Dr. Arzu ORGAN

Temmuz 2023
DENİZLİ

BİLİMSEL ETİK SAYFASI

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bilimsel etiğe ve akademik kurallara özenle riayet edildiğini; bu araştırmanın doğrudan birincil ürünü olmayan bulguların, verilerin ve materyallerin bilimsel etiğe uygun olarak kaynak gösterildiğini ve alıntı yapılan çalışmalara atıfta bulunulduğunu beyan ederim.

Mustafa Onur KAÇAROĞLU

ÖN SÖZ

Bu tezin hazırlanması sürecinde desteklerini hiçbir zaman esirgemeyen ve sürekli yol gösteren başta danışman hocam Prof. Dr. Arzu ORGAN olmak üzere jüride yer alan Prof. Dr. Oğuz KARADENİZ, Dr. Öğr. Üyesi Fatmana ŞENTÜRK, Doç. Dr. Engin ÇAKIR ve Doç. Dr. Gülşah Sezen AKAR hocalarıma sonsuz teşekkürlerimi, sunarım. Tezimi okuyup fikirlerini paylaşan ve tezin yazım sürecinde bana sürekli destek olan sayın hocam Prof. Dr. Hülya KABAKÇI KARADENİZ' e de ayrıca teşekkür ederim.

Doktora sürecinde bana destek olan, doktora eğitimine beraber başladığım dostlarım Dr. İbrahim BUDAK, Dr. Günay KILIÇ ve Erkan TURHAN' a teşekkür ederim.

Eğitim hayatımın hem yüksek lisans hem de doktora döneminde bana destek olan sevgili eşim Gülşen KAÇAROĞLU' na göstermiş olduğu sabır, anlayış ve bana olan güveni için sonsuz teşekkür ederim. Ayrıca bu süreçte varlıkları ile hayatımın motivasyon kaynağı olan ve zaman zaman istemeden de olsa beraber geçireceğimiz vakitlerini çaldığım oğullarım Mete ve Tuna' ya da teşekkürlerimi sunarım. İyi ki varsınız.

ÖZET

ÇEVRİMİÇİ İŞ İLANLARININ VERİ VE METİN MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ: BİLGİ VE İLETİŞİM SEKTÖRÜ ÖRNEĞİ

KAÇAROĞLU, Mustafa Onur
Doktora Tezi
İşletme A.B.D.
Genel İşletme Programı
Tez Danışmanı: Prof. Dr. Arzu ORGAN

Temmuz 2023, X + 112 sayfa

Son yıllarda internetin hızla gelişmesi ve web tabanlı uygulamaların kullanılması ile veri miktarı hızlı bir şekilde artmaktadır. Artan verilerin iş dünyası için kullanışlı hale getirilmesi, veri ve metin madenciliği yöntemlerinin etkin kullanılması ile mümkündür. Veri madenciliği büyük veri yapıları içerisinden çeşitli algoritmalar yardımıyla anlamlı ve kullanışlı bilgiler elde etme sürecidir. Metin madenciliği, veri madenciliği ile ilişkili ya da veri madenciliğinin alt dalı olarak görülse de esasen ayrı bir disiplindir. Metin madenciliği, her türlü metin içeren ifadenin çeşitli algoritmalar yardımıyla işlenip analiz edilmesi ve sonucunda katma değer yaratacak çeşitli bilgilerin açığa çıkarılması sürecidir. İşletmeler metin ve veri madenciliği uygulamalarını kullanarak rekabetçi ortamda kendilerine avantajlar yaratabilirler. Bu çalışmada hem iş gören hem iş veren hem de eğitim kurumları için yararlı bilgilerin açığa çıkarılması amaçlanmıştır. Bunun için, veri ve metin madenciliği yöntemleri kullanılarak, Bilgi ve İletişim Sektörüne ait çevrim içi iş ilanları analiz edilip iş gücü piyasasına ilişkin tespitler yapılmıştır. Veriler, bir çevrimiçi iş ilanı sitesinden elde edilmiştir. Elde edilen veriler, Rastgele Orman (Random Forest) Karar Ağacı Algoritması ve Gizli Dirichlet Ayrımı yöntemleri ile analiz edilmiştir. Sonuç olarak iş ilanlarında yer alan nitelikler kullanılarak yapılan analizlere elde edilen konular için iş tanımları oluşturulmuştur. Bu iş tanımlarının yakın gelecekte iş ilanlarında yer alması muhtemeldir. Bununla birlikte teknik becerilerin son derece önemli olduğu bilgi ve iletişim sektörü için ikna, iletişim, motivasyon, liderlik, sorumluluk, esneklik, karar verme, tutarlılık gibi hassas becerilerin de baskın konular içerisinde ağırlıklı olarak yer aldığı saptanmıştır.

Anahtar Kelimeler: Veri Madenciliği, Metin Madenciliği, Konu Modelleme, Karar Ağaçları, Çevrimiçi İş İlanları, Bilgi ve İletişim Sektörü İş İlanları

ABSTRACT

ANALYSIS OF ONLINE JOBS POSTINGS WITH DATA AND TEXT MINING METHODS: EXAMPLE OF INFORMATION AND COMMUNICATION SECTOR

KAÇAROĞLU, Mustafa Onur
Doctoral Thesis
Business Administration Department
PhD. In Business Administration Programme
Adviser of Thesis: Prof. Dr. Arzu ORGAN

July 2023, X + 112 pages

In recent years, the amount of data has been increasing rapidly with the rapid development of the internet and the use of web-based applications. It is possible to make the increasing data useful for the business world with the effective use of data and text mining methods. Data mining is the process of obtaining meaningful and useful information from large data structures with the help of various algorithms. Although text mining has been associated with data mining or seen as a sub-branch of data mining, it is essentially a separate discipline. Text mining is the process of processing and analysing all kinds of text-containing expressions with the help of various algorithms and revealing various information that will create added value as a result. Companies may create advantages for themselves in the competitive environment by using text and data mining applications. In this study, it has been aimed to reveal useful information for all employees, employers and educational institutions. To achieve this, online job postings belonging to the Information and Communication Sector have been analysed by using data and text mining methods and determinations have been made regarding the labour market. Data have been obtained from an online job posting site. The obtained data have been analysed by Random Forest Decision Tree Algorithm and Hidden Dirichlet Discrimination methods. As a result, job descriptions were created for the subjects obtained from the analyzes made using the qualifications in the job postings. These job descriptions are likely to appear in job postings in the near future. In addition, it is determined that sensitive skills such as persuasion, communication, motivation, leadership, responsibility, flexibility, decision-making and consistency are among the dominant subjects for the information and communication sector, where technical skills are extremely important.

Keywords: Data Mining, Text Mining, Topic Modelling, Decision Trees, Online Job Postings, Information and Communication Industry Job Postings

İÇİNDEKİLER

BİLGİ VE İLETİŞİM SEKTÖRÜNDEKİ ÇEVRİMİÇİ İŞ İLANLARININ VERİ VE METİN MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGE VE KISALTMALAR DİZİNİ

BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies
CART	Classification and Regression Tree
CHAID	Chi-Squared Automatic Interaction Detector
CLARA	Clustering Large Application
CLUCDUH	Clustering Categorical Data Using Hierarchies
CRISP-DM	Cross Industry Standard Process for Data Mining
CURE	Clustering Using REpresentatives
DENCLUE	DENsity Based CLUstERing
DMME	Data Mining Methodology for Engineering Applications
GARP	Genetic Algorithm Rule-Set Production
GDA	Gizli Dirichlet Ayrımı
İŞKUR	Türkiye İş Kurumu
KDD	Knowledge and Data Discovery
KNN	K-Nearest Neighbors
NACE	Nomenclature des Activités Économiques dans la Communauté Européenne
OLAP	Çevrim İçi Analitik Süreçler
pLSA	Probabilistic Latent Semantic Analysis
QUEST	Quick Unbiased Efficient Statistical Tree
SEMMA	sample, Explore, Modify, Model, Evaluate
STING	STatistical Information Grid-based Method
TÜİK	Türkiye İstatistik Kurumu
VTBK	Veri Tabanlarında Bilgi Keşfi
YBS	Yönetim Bilişim Sistemleri
YSA	Yapay Sinir Ağları

GİRİŞ

Günümüzde teknolojinin gelişmesi sektörel ihtiyaçları değiştirmekte ve bu durum, iş gören adaylarından beklentilerini etkilemektedir. İş gören adayları, becerilerini teknolojik gelişmeler doğrultusunda geliştirmek durumundadır. İş görenler kariyer planlarının yaparken teknolojik değişimleri ve sektörel beklentileri dikkate almak zorundadır. Eğitimli genç iş gücü, iş bulmada güçlükler yaşamaktadır. İş gören adaylarının sektör beklentileri doğrultusunda kendi yeteneklerini de keşfetmek suretiyle kariyer planlarını yapmaları ve özellikle becerilerini bu yönde geliştirmeleri gerekmektedir. İş gücü piyasasında çevrim içi iş ilanları sektörel beklentilerin anlaşılmasında son derece önemlidir. Özellikle her gün yayınlanan binlerce iş ilanı bu beklentiler konusunda bizlere yol gösterebilecek çok özel bilgileri barındırmaktadır. Aynı şekilde işletmeler de kariyer sitelerinde ilan vermek suretiyle hem mevcut iş gören açıklarını tamamlamaya çalışmakta hem de geniş bir aday havuzu oluşturmayı amaçlamaktadır. İşletmelerin aday havuzlarının genişliği onları potansiyel iş gören açıklarının olduğu dönemlerde daha hızlı bir şekilde bu açıklarını doldurmalarını sağlayabilmektedir.

İşletmeler boş pozisyonlarını, iş gören adaylarına farklı yollardan duyurmaktadırlar. Ülkemizde iş ilanları çoğunlukla, kariyer siteleri, İŞKUR duyuruları, işletmeye ait web siteleri veya insan kaynakları danışmanlık firmaları vasıtasıyla yayınlanmaktadır. Yayınlanan iş ilanları büyük miktarda verinin ortaya çıkmasını sağlamaktadır. İş gören adayları ve işletmeler için bu çok fazla verinin anlamlı ve yararlı hale getirilmesi gerekmektedir. Bunun için veri ve metin madenciliği teknikleri en uygun araçtır.

Veri ve metin madenciliği yöntemleri artan verilerin analiz ederek kullanışlı bilgiler elde edilmesi için önemli imkanlar sunar. Büyük veri analitiğinde metin madenciliği son derece güçlü bir araçtır. Yapılandırılmamış metinsel verilerin gücünden yararlanmak için bu veriler, analiz edilerek yeni bilgiler elde edilebilir ve verilerde gizli olan önemli kalıplar ve korelasyonlar belirlenebilir (Hassani vd., 2019: 1). Veri ve metin madenciliği yöntemleri çok sayıda verinin analiz edilmesini ve daha kısa zamanda daha doğru yorumlanmasını sağlayabilir.

Günümüzde teknolojik gelişmeler ve dijitalleşme ile veri tabanları daha verimli kullanılır hale gelmiştir. Sürekli artan bir şekilde verilerin biriktirilmesi, bu verilerin barındırdığı kullanışlı bilgilerin açığa çıkarılması ihtiyacını doğurmuştur. Özellikle son yıllarda bu verilerden, anlamlı bilgiler elde edilmesini sağlayan, matematiksel ve istatistiksel yöntemlerin kullanıldığı veri madenciliği teknikleri geliştirilmiştir. Veri madenciliği, tanımlayıcı ve tahmin edici modeller ile verilerin analiz edilmesini kolaylaştırmaktadır. Veri madenciliği yöntemlerinden karar ağacı algoritmaları sınıflandırma amacıyla kullanılmaktadır (Timor ve Yüzbaşı Künç, 2021: 2). Karar ağaçları veri setinde yer alan verilerin sınıflandırılması için sistematik bir ağaç yapısı oluşturmak suretiyle çalışmaktadır. Karar ağacı algoritmaları basit ve anlaşılır kural yapılarına sahip oldukları sıklıkla uygulama alanı bulan önemli araçlardır (Çelik vd., 2022: 568).

Metin madenciliği de veri madenciliği ile ilişkili ya da veri madenciliğinin alt dalı olarak görülse de esasen ayrı bir disiplindir. Akıllı Metin Analizi, Metin Veri Madenciliği veya Metinde Bilgi Keşfi olarak da bilinen Metin Madenciliği, önemli ve ilgi çekici bilgilerin yapılandırılmamış metinlerden çeşitli algoritmalar yardımıyla çıkarılması sürecini ifade eder. Metin madenciliğinin finans, iş, klinik, biyoloji, biyomedikal gibi çeşitli alanlarda çeşitli uygulamalarını bulmak mümkündür (Ferreira-Mello vd., 2019: 2). Bilgisayar bilimleri, veri bilimi, matematik ve dilbilimin kesiştiği bir alan olan metin madenciliği, yalnızca büyük metin külliyatını verimli bir şekilde analiz etmeyi değil, aynı zamanda bunu şeffaf ve tekrarlanabilir bir şekilde yapmaya da olanak sağlayan yöntemleri içermektedir (Antons vd., 2020: 330). Metin madenciliğinin uygulamalarının amacı, metinsel belgelerin (e-postalar, incelemeler, düz metinler, web sayfaları, raporlar ve resmî belgeler) oluşturduğu büyük veri yığınları içerisinden çeşitli karar verme türleri için yararlı bilgiler ortaya çıkarmaktır. Metin madenciliği ayrıca, analiz, görselleştirme (haritalar, çizelgeler, zihin haritaları aracılığıyla), veri tabanları veya ambarlardaki yapılandırılmış verilerle entegrasyon, makine öğrenimi vb. için kullanılabilen dilbilimsel, istatistiksel ve makine öğrenimi tekniklerini de kapsamaktadır (Bach vd., 2019: 2).

Metin madenciliği alanında kullanılan önemli yöntemlerden biri de Konu Modelleme yöntemidir. Konu modelleme, veri madenciliği, gizli veri keşfi ve veriler ile metin belgeleri arasındaki ilişkileri bulmak için metin madenciliğinde en önemli tekniklerden biridir (Jelodar vd. 2019: 15169). Konu modelleme, metin belgelerinde

kelime gruplarının oluşturduğu konuları bulmak için kullanılan denetimsiz bir uygulamadır. Burada ortaya çıkan konular, sık sık birlikte kullanılan ve genellikle ortak bir konu ile ilgili kelimelerden oluşmaktadır. Bu nedenle, önceden tanımlanmış kelime kümesi, belgenin tamamını en iyi şekilde tanımlayabilecek kelime grubu olarak nitelendirilebilir. Konu modelleme, metin verilerinden oluşan büyük veri yığınlarını anlama ve özetleme imkânı verir. En bilinen konu modelleme yöntemlerinden bir tanesi Gizli Dirichlet Ayrımı (GDA) yöntemidir. GDA ayrık veri koleksiyonları için olasılıksal bir model olarak tanımlanabilir (Blei, 2003: 1014). Gizli Dirichlet Ayrımı, her bir metin belgesinin farklı konular içeren bir konu koleksiyonu olarak kabul edildiği ve bu koleksiyondaki her kelimenin, konulardan biri ile ilişkili olduğu bir konu modelleme yöntemidir. Dolayısıyla, Gizli Dirichlet Ayrımı yöntemi, bir metin belgesinin analizi esnasında, o metin belgesini esas alarak her bir konuyu o grubu en iyi açıklayan bir dizi kelimeyi tespit etmek suretiyle konu grupları oluşturur.

Bu çalışma üç bölümden oluşmaktadır. Birinci bölümde veri madenciliği ile ilgili kavramsal çerçeve dahilinde tanımlar, teknikler, kullanım alanları ve literatürdeki çalışmalara yer verilmiştir. Bu kapsamda veri madenciliği süreç modelleri, veri madenciliği yöntemleri olan sınıflandırma, kümeleme, birliktelik kuralları analizi ile ilgili algoritmalar anlatılmıştır.

İkinci bölümde metin madenciliği ile ilgili kavramsal çerçeve, literatürdeki çalışmalar, tanımlar, metin madenciliği süreci ve metin madenciliğinin kullanım alanları ayrıntılı bir şekilde ele alınmıştır. Çeşitli görselleştirme uygulamalarına ait yöntemler de bu bölümde anlatılmıştır. Ayrıca çalışmada kullanılan konu modelleme yöntemi olan Gizli Dirichlet Ayrımı yöntemi de detaylı bir şekilde açıklanmıştır.

Üçüncü bölümde ise araştırmanın uygulama kısmı yer almaktadır. Çalışmada çevrim içi iş ilanlarının veri ve metin madenciliği yöntemleri ile analiz edilmesi amaçlanmıştır. Bu çalışma sonucunda, kariyer planlama, iş gücü planlaması veya eğitim kurumlarının müfredatlarının oluşturulmasına katkı yapacak yeni bilgiler iş ilanlarının oluşturduğu büyük veri yığını içerisinden çıkarılabilir. Bu gibi platformlarda gizli kalan potansiyel bilgilere ulaşmak için en önemli kariyer sitelerinden birisi olan Kariyer.net web sitesinde yer alan iş ilanları bu çalışma kapsamında değerlendirilmiştir. Bir yıl boyunca elde edilen Bilgi ve İletişim Sektörü iş ilanları çeşitli ön işlemlerden geçtikten sonra analizler yapılmıştır. Elde veriler Türkiye İstatistik Kurumu sınıflama sunucusunda

yer alan NACE Rev.2-Altılı Ekonomik Faaliyet Sınıflaması esasına göre sınıflandırılmış ve bilgi ve iletişim sektörü için yayınlanan iş ilanları analizlere dahil edilmiştir. Çalışmada öncelikle karar ağacı analizi yapılmıştır. Bunun için Rastgele Orman algoritmasından yararlanılmıştır. Sonrasında ilanların iş tanımı kısımlarında yer alan metinler Gizli Dirichlet Ayrımı algoritması kullanılarak konu modelleme analizine tabi tutulmuştur. Sonrasında bir yıllık veri üçer aylık dört döneme ayrılmış ve analiz her üç aylık dönemi kapsayan verilere uygulanarak dönemler arasındaki farklılıklar tespit edilmeye çalışılmıştır. Bilgi ve iletişim sektörü iş ilanları tecrübeli/tecrübesiz, lisans mezunu/ön lisans mezunu şeklinde ayrımlar gözetilerek analizler yapılmıştır. Son olarak da konu modelleme, ilanların geneline uygulanmış ve bilgi ve iletişim sektörü için yayınlanan iş ilanlarında ön plana çıkan konuların genel durumu ortaya konulmuştur. Çalışma sonucunda, bilgi ve iletişim sektörünün iş gören adaylarından beklentileri ve yakın gelecekte sektörün hangi tür becerileri talep edeceği konusunda bulgular elde edilmiştir. Çalışma bu yönüyle özellikle iş gören adaylarının kariyer planlarını yapmaları ve eğitim kurumlarının müfredatlarını düzenlemeleri konusunda bazı aydınlatıcı bilgiler içermektedir.

BİRİNCİ BÖLÜM

VERİ MADENCİLİĞİ

Günümüzde verinin artan değeri, onu daha iyi anlama ihtiyacını doğurmuştur. Bu amaçla istatistik, matematik ve bilgisayar bilimleri gibi alanlarda veri analizi ön plana çıkmış ve veri madenciliğinin gelişmesine bu bağlamda hızlanmıştır. Bu bölümde veri madenciliği ile ilgili kavramsal çerçeveye yer verilmiştir.

1.1. Veri Madenciliği Kavramı

Veri madenciliği 1950'lerde istatistiksel analizlerde bilgisayar kullanımıyla başlayan ve günümüzde çok çeşitli alanlarda bir hizmet çözümü olarak yazılım araçlarının geliştirilmesine öncülük eden bir disiplindir (Mikut ve Reischl, 2011: 431).

Veri madenciliği, matematik, istatistik, yapay zekâ, makine öğrenimi ve veri tabanı araştırmalarında güçlü köklere sahip olan uzun bir geçmişi vardır. Bu alandaki gelişmelere, 1950' lerin başlarında istatistiksel analizler için bilgisayar kullanımıyla başlayan ve günümüzde hizmet çözümü olarak çok çeşitli bağımsız, istemci/sunucu ve web tabanlı yazılımlara yol açan ilgili yazılım araçlarının geliştirilmesi eşlik etmiştir.

Tüm dünyadaki veri miktarı her 20 ayda 2 katına çıktığı tahmin edilmektedir. Buna bağlı olarak veri tabanlarının büyüklükleri ve sayısı da artmaktadır. İş dünyası ve gündelik yaşantımızda yer alan faaliyetler; örneğin telefon görüşmeleri, kredi kartı harcamaları, bankacılık işlemleri, tıbbi testler ve muayeneler sürekli artan bir veri üretimi sağlamaktadır. Bu tip işlemlerden elde edilen veriler bilgisayarlara veya çeşitli depolama ünitelerine kaydedilmektedir (Frawley vd., 1992: 57).

Veri madenciliği, tıbbi teşhis ve tanıma, finansal piyasalar, pazarlama araştırmaları, kredi kartı yolsuzluklarının tespiti, televizyon izleyicilerinin analizi gibi çok farklı alanlarda kullanılmaktadır (Saritas ve Yasar, 2019: 88).

Veri madenciliği konusunda çeşitli tanımlar yapılsa da basit olarak "büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir" şeklinde tanımlanmaktadır. Bu şekilde veriler arası ilişkiler ortaya koymak ve geleceğe dair çıkarımlarda bulunmak mümkündür. Veri madenciliğini, bir kurumda üretilen tüm verilerin belirli yöntemler kullanılarak var olan ya da gelecekte ortaya çıkma potansiyeline sahip gizli kalmış bilgilerin ortaya çıkarılması süreci olarak değerlendirebiliriz. Dolayısıyla veri

madenciliği uygulamaları işletmelerin karar destek sistemleri için önemli bir yere sahiptir (Özkan, 2020: 12).

Veri madenciliği "büyük veri tabanlarından gizli bilgilerin, önceden tahmin edilmeyen örüntülerin ve yeni kuralların keşfedilmesi" şeklinde tanımlanabilir. Bilgisayar teknolojilerinin gelişip günlük hayatta yoğun bir şekilde kullanılması sonucunda günümüzde büyük miktarlarda veri birikimi meydana gelmeye başlamıştır. Bu durum biriken büyük miktarda verinin analiz edilmesi gerekliliğini ortaya çıkarmıştır. Bu doğrultuda yapılan istatistiksel analiz ve modeller, yapay zekâ uygulamaları gelecek adına sağlıklı tahminler yapılmasına olanak yaratmışlardır. Bu şekilde uygulamaları içeriğinde barındıran veri madenciliği, veri yığınları içinden anlamlı içerikler üretmek adına son derece kullanışlı olabilir (Şimşek Gürsoy, 2011: 3).

Veri madenciliği sürecinde; kümeleme, veri özetleme, sınıflama kurallarının öğrenilmesi, bağımlılık ağlarının bulunması, değişkenlik analizi ve anomali tespiti gibi farklı teknikler kullanılarak önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınları içerisinden elde edilmesi olarak tanımlanabilir (Baykal, 2006:96).

Veri madenciliği büyük hacimli veri kümeleri içinden bilgi çıkarma sürecidir. Başka bir deyişle, veri madenciliği, büyük gözlemsel veri tabanlarına vurgu yaparak, büyük ve karmaşık verilerdeki beklenmedik kalıpları, değerli yapıları ve ilginç ilişkileri keşfetme bilimidir (veya sanatıdır). Hızlı bilgisayar ve ucuz depolama süpermarket satın alma kalıplarından bankacılığa, hisse senedi ticaretinden tıbbi teşhise kadar birçok konuda faydalı bilgiler çıkarmayı kolaylaştırır. Veri madenciliği, kurumsal hedeflere ulaşmak ve araştırmacıların olayların nedenlerini belirlemek ve geleceği tahmin etmek için basit veri sorgularının ve raporlamanın ötesine geçmesine izin vermek için yaygın olarak kullanılmaktadır (Ganesh, 2002: 1).

Veri madenciliği, büyük veri tabanlarından önceden beklenmeyen ya da tahmin edilmeyen bilgilerin ortaya çıkarılması ve bunların karar vermeye zemin oluşturmasını içeren çok aşamalı bir sürectir. Veri madenciliği araçları, veriler içinden kalıpları algılar ve onlardan ilişkilendirmeler ve örüntüler çıkarır. Çıkarılan bilgiler daha sonra veri kayıtları veya veri tabanları arasındaki ilişkileri tanımlanmasıyla tahmin veya sınıflandırma modellerine uygulanabilir. Bu örüntüler ve kurallar karar vermeye rehberlik eder ve daha sonra bu kararların etkileri ile ilgili tahminde bulunabilirler.

Veri madenciliği kavramının üzerinde uzlaşılmış net bir tanımı olmamakla beraber aşağıda birkaç farklı tanım örnek olarak verilmiştir:

Veri madenciliği, büyük veri setlerinde ilginç, beklenmedik veya değerli yapıların keşfedilmesi sürecidir (Hand, 2007: 1).

Veri madenciliği, büyük veri tabanlarında gizli kalmış ve yararlı olma potansiyeli taşıyan bilgilerin açığa çıkarılması ve veri tabanlarındaki ilişkisel yapıların ortaya konulup aralarındaki korelasyonları bulma imkânı veren önemli bir araçtır (Agarwal, 2013: 203).

Veri madenciliği, verilerdeki kalıpları ve korelasyonları keşfetmek için çeşitli algoritmalar kullanan gelişmiş bir veri arama olanağı olarak tanımlanabilir. Veri madenciliği, kurumsal veri ambarlarında gömülü bilgileri ("veri külçeleri") veya ziyaretçilerin bir web sitesine bıraktığı bilgileri bulur ve çıkarır. Söz konusu bilgilerin çoğu, verilerin anlaşılmasında ve kullanılmasında iyileştirmelere yol açabilir. Veri madenciliği yaklaşımı, istatistikler, çevrimiçi analitik süreçler (OLAP), elektronik tablolar ve temel veri erişimi gibi diğer veri analizi tekniklerini tamamlayıcı niteliktedir. Basit bir ifadeyle, veri madenciliği, verilerde anlam bulmanın başka bir yoludur (Rygielski, 2002: 485).

Amerika Birleşik Devletleri Kongresinde, yönetimin uyguladığı veri madenciliği faaliyetlerini kongreye raporlamasına yönelik olarak verilen Veri Madenciliği Raporlama Kanunu 2003 önerisinde veri madenciliği, bir veya daha fazla veri tabanının, sorgulanması, araştırılması ve analizi olarak tanımlanmıştır. Ayrıca Amerika Birleşik Devletleri Genel Muhasebe Bürosunun 2004 Mayıs ayında yayınladığı raporda veri madenciliği "veri tabanı teknolojilerinin, istatistiksel analiz ve modelleme gibi tekniklerin verideki gizli kalıpların ve gizli ilişkilerin ortaya çıkarılması ve bunlardan gelecekteki olaylar ile ilgili çıkarımlar üretmeye yarayacak bazı kuralların belirlenmesi ve uygulanmasıdır" şeklinde tanımlanmıştır (Şentürk, 2006: 3).

1.2. Veri Madenciliği Alanındaki Çalışmalar

Veri madenciliği verinin yoğun bir şekilde ortaya çıktığı ve bununla birlikte veri tabanı oluşan her durumda kullanılabilmektedir (Can vd., 2012: 1). Aşağıda çeşitli alanlarda veri madenciliği ile yapılan çalışmalardan örnekler verilmiştir.

Uysal vd. (2014), yaptıkları çalışmada kalp hastalığı verilerini analiz ederlerken veri madenciliği yöntemleri kullanmışlardır. Verilerin analizi için TWOING sınıflandırma algoritmasından yararlanmışlardır. Analiz sonucunda kalp hastalığı veri seti sınıflandırılmış ve bir dizi karar kuralı oluşturulmuştur. Sonuç olarak kalp hastalığı belirtileri ortaya konulmuştur. Yaş, cinsiyet, maksimum kalp hızı, göğüs ağrısı tipi, açlık kan şekeri, büyük damarlar, talasemi, anjine bağlı depresyon gibi durumların kalp hastalığı ile ilişkisinin derecesi belirlenmiştir. Sınıflandırma algoritması vasıtasıyla oluşturulan karar kurallarına göre göğüs ağrısı, büyük damarlar ve ST eğiminin kalp hastalığının en önemli belirtisi olduğu ortaya çıkmıştır.

TÜİK verilerinden yararlanılarak yapılan başka bir çalışmada mutluluk düzeyini etkileyen faktörler belirlenmeye çalışılmıştır. Bu çalışmada Türkiye İstatistik Kurumu (TÜİK) tarafından 2015 yılında 9397 kişiye uygulanan yaşam doyumu anketinin verileri kullanılmıştır. Araştırmanın amacı, sınıflandırma ağaçları tekniği kullanılarak kişilerin mutluluk düzeyini etkileyen faktörleri belirlemektir. Sınıflandırma ağaçları, Sınıflandırma ve Regresyon Ağacı (CART) ve Ki-Kare Otomatik Etkileşim Tespiti (CHAID) algoritmaları kullanılarak oluşturulmuştur. Oluşturulan ağaçlar karşılaştırmalı olarak yorumlanmıştır (Kalkan ve Yücel, 2017).

Düzdar ve Temür (2017), yaptıkları çalışmada 2006 yılında AB'ye aday ve üye ülkelerden 29 ülkeye ait Enflasyon, gayri safi milli hâsıla, internet kullanım, işsizlik, ömür boyu eğitim göstergeleri, ithalat ve ihracata ait verilerini kullanarak ekonomik benzerlikleri tespit etmeye yönelik kümeleme analizi yapmışlardır. Bunun için veri madenciliğinde kullanılan açık kaynak kodlu yazılımlarından bazıları olan Orange, Knime, Weka ve RapidMiner gibi programları uygulayarak analizleri yapmışlar ve sonuçları karşılaştırmışlardır.

Gazioğlu ve Şeker (2017), veri madenciliği, makine öğrenmesi ve veri bilimi tekniklerini kullanarak, herhangi bir kişinin tweetlerini inceleyerek kişinin girişimcilik potansiyeli hakkında bilgi toplamayı amaçladıkları bir çalışma yapmışlardır. Sosyal medya kullanan kişilerin, kendi beyanlarını esas alarak, başarılı girişimciler ile görece daha az başarılı sayılabilecek girişimcilerin tweetlerini inceledikleri çalışmalarında başarılı girişimcilerin ayırt edici özelliklerini çıkarıp bunları öğrenen bir yapay zekâ tasarlamayı amaçlamışlardır.

Karamaşa ve Erdoğan (2018), bayram ve tatil dönemlerinde yaşanan trafik kazalarının artması sonucu bu kazaları önleme adına, Türkiye' de, 2009 – 2011 yılları arasında bayram gidiş ve dönüşlerini kapsayan dönemlerde trafik polisinin yetki bölgelerinde meydana gelen trafik kazalarını veri madenciliği yöntemlerinden olan birliktelik kuralları ile inceleyerek bu kazaların ortak özelliklerini gösteren kurallarını ortaya koymaya çalışmışlardır.

Hassani vd. (2018), gelecekteki araştırma ve geliştirme çalışmalarına bir yön vermek amacıyla, bankacılık sektöründeki veri madenciliği uygulamalarını ve araştırmalarını 2013 yılından itibaren kapsamlı bir şekilde incelemişlerdir. Araştırmada, bankacılık alanındaki veri madenciliği uygulamaları ve veri kaynakları kapsamlı bir şekilde analiz edilerek bir referans noktası oluşturulmaya çalışmışlardır. Bununla birlikte büyük veri analitiği ile ilgili engel ve tehditler de çalışmanın diğer çıktıları olarak görülmüştür.

Yang vd. (2018) yaptıkları çalışmada doğru ve güvenilir klinik tanı koymak için sınıflandırma algoritmalarından ID3 için geliştirilmiş bir versiyonunu önermişlerdir. Tıp alanında yapılan deneylerde geliştirilmiş ID3 algoritmasının diğer sınıflandırma algoritmalarından (J48, Decision Stump ve Random Tree) daha doğru ve güvenilir bir performans gösterdiği tespit edilmiştir.

Darabi vd. (2019) yaptıkları çalışmalarında yağış, eğim, eğri sayısı, nehre olan mesafe, kanala olan mesafe, yeraltı suyuna olan derinlik, arazi kullanımı ve yükseklik gibi sel ve taşkına neden faktörleri Genetic Algorithm Rule-Set Production (GARP) ve Quick Unbiased Efficient Statistical Tree (QUEST) tespit etmeye çalışmışlardır.

Huber vd. (2019), endüstriyel uygulamalarda veri analitiği yürütmek için fiili standart olan CRISP-DM metodolojisine eleştirel bir yaklaşımla farklı bir metodolojisi ortaya koymuşlardır. CRISP-DM modelinde üretim senaryoları için veri toplama aşaması belirtilmemesi bu çalışmanın temelini oluşturmuştur. Sonuç olarak bu modelin bir uzantısı olan, üretim alanındaki veri analitiği için bir iletişim ve planlama temeli sunan DMME (Data Mining Methodology for Engineering Applications) modelini bir vaka çalışması ile önerilmiştir.

Coşkun ve Bülbül (2019) yaptıkları çalışmada Türkiye İstatistik Kurumu'nun 2016 yılı Hane halkı Bilişim Teknolojileri Kullanım Anketi verilerini kullanarak hane

halkının internet hizmetine sahip olma durumu incelemişlerdir. CHAID, C5.0, C&RT ve QUEST gibi karar ağacı algoritmaları karşılaştırmak suretiyle yaptıkları çalışmada hanelerinde internet hizmeti kullananların aboneliklerini etkileyen en önemli faktörlerin hane cep telefonu sahipliği, hane bilgisayar kullanımı, hanede yaşayan 25 altı bireyler, hane tablet sahipliği, hane dizüstü bilgisayar sahipliği, hane halkı büyüklüğü, hane halkı reisinin yaşı, hane smart tv sahipliği, hane gelir grubu gibi değişkenlerin olduğu görülmüştür.

Aydemir ve Yavuz (2019), Türkiye'de faaliyet gösteren bir eczanenin bir yıllık ilaç satış verilerini düzenlenmiş ve birliktelik analizi ile incelemişlerdir. Mevsimsel değişkenliklerin incelendiği analizler sonucu en çok birlikte satılan ilaçların belirlenmesi amaçlanmış aynı zamanda satışlara eczanenin kuruluş yeri, sağlık kuruluşlarına yakınlık, zaman ve salgın hastalık gibi birçok değişkenden etkilendiği tespit edilmiştir. Çalışma sonucunda bölgedeki hastalıklar hakkında fikir sahibi olunabilecek bulgular elde etmişlerdir.

Pradipta vd. (2019), C4.5 karar ağacı algoritması ile üniversite öğrencilerinin zamanında veya zamanından daha geç mezun olmalarını incelemişlerdir. Çalışmada kategorik değişkenler doğrultusunda oluşan öğrenci profillerinin mezuniyet durumu ile olan ilişkileri ortaya konulmaya çalışılmıştır.

Abu Saa vd. (2019) yaptıkları araştırmada öğrencilerin performansını etkileyen faktörleri ve bu faktörleri belirlemek için uygulanan en yaygın veri madenciliği tekniklerini belirlemeye çalışmışlardır. Çalışmada 2009'dan 2018'e kadar toplam 420 araştırma makalesinden 36'sı sistematik bir literatür taraması yaklaşımı uygulanarak eleştirel olarak incelenmiş ve analiz edilmiştir. Sonuçlar, en yaygın faktörlerin öğrencilerin önceki notları ve sınıf performansı, öğrencilerin e-öğrenme etkinliği, öğrencilerin demografisi ve öğrencilerin sosyal bilgileri olmak üzere dört ana kategori altında toplandığını göstermiştir. Bununla birlikte, sonuçlar öğrencilerin faktörlerini tahmin etmek ve sınıflandırmak için kullanılan en yaygın veri madenciliği tekniklerinin karar ağaçları, Naïve Bayes sınıflandırıcıları ve yapay sinir ağları olduğunu ortaya koymuştur.

Gayathri vd. (2020)' de yaptıkları çalışmada, Hindistan' da bulunan Tamilnadu Krallığı' ndaki hava kirliliğinin tahmini için veri madenciliği yöntemlerine

başvurmuşlardır. Bu amaçla Tamilnadu Krallığı' nın hava kirliliği istatistiklerini içeren veri setlerini toplayıp Id3 algoritması yardımıyla bir tahmin problemi oluşturmuşlar ve sonuçta uygun bir tahmin yöntemi elde etmişlerdir.

Çöllü vd. (2020), 2016 – 2018 aralığında Borsa İstanbul' da işlem gören dokuma, giyim eşyası ve deri sektöründe faaliyet gösteren firmaların finansal başarısızlıklarını etkileyen finansal oranların tespiti ve finansal başarısızlığın tahmininde çeşitli veri madenciliği yöntemlerinin güçlerinin belirlenmesi amacıyla bir çalışma yapmışlardır. Bu amaçla ilgili sektörlerde faaliyette bulunan 20 firmanın 3 yıllık finansal verileri CHAID, Exh-CHAID, CART ve QUEST gibi karar ağacı yöntemleri kullanılarak analiz edilmiştir.

Saeed vd. (2020), elektrik kullanımında fatura dolandırıcılığının tespiti için bir çalışma yapmışlardır. Bunun için Pakistan' da faaliyette bulunan Multan Electric Power Company firmasının müşterileri arasında araştırma yürütülmüş olup C5.0 karar ağacı algoritmasından yararlanılmıştır.

Saura (2021), dijital pazarlamada veri bilimi ile ilgili kullanımların, yöntemlerin ve performans metriklerinin karşılaştırmalı bir analizini yapmıştır. Bu amaçla son 10 yılı kapsayan kapsamlı bir literatür incelesi yapılmıştır. Sonuç olarak, veri bilimlerinin dijital pazarlamaya yönelik ana uygulamalarına bütünsel bir genel bakış açısı ve yenilikçi veri madenciliği ve bilgi keşif tekniklerinin oluşturulmasıyla ilgili öneriler sunulmuştur.

1.3. Veri Madenciliği Süreç Modelleri

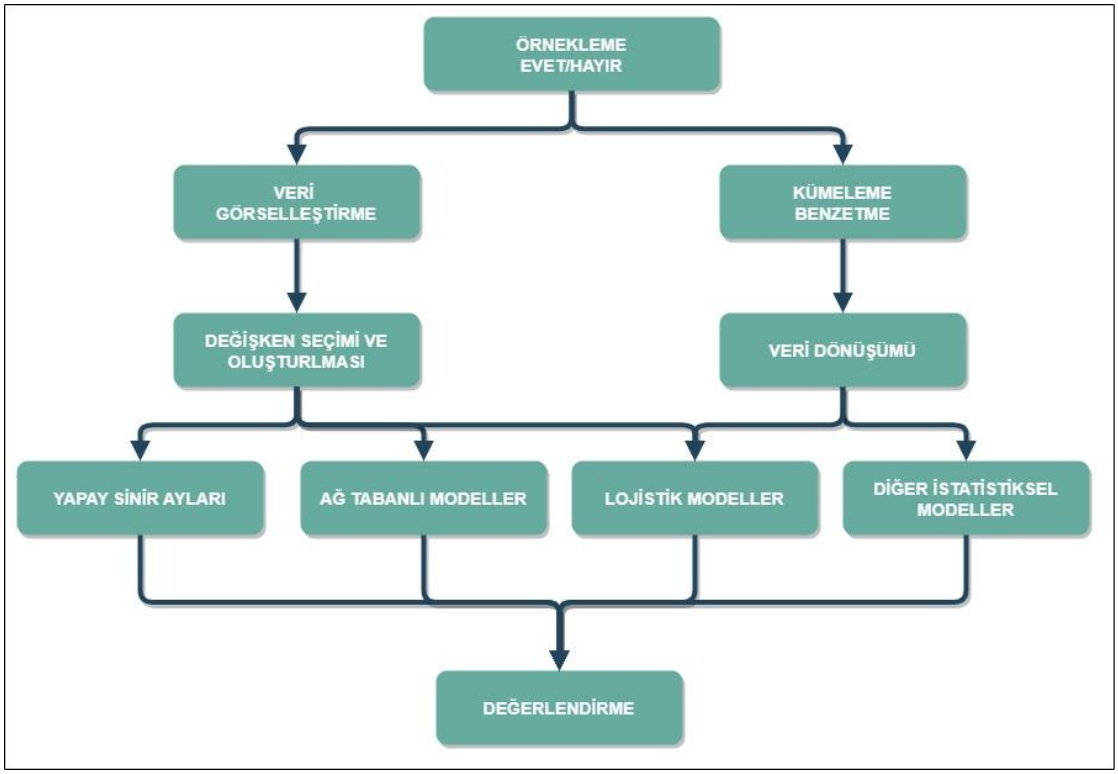
Veri Madenciliği, veri yığınlarından anlamlı bilgiler elde etmek adına bu veriler içerisinden çeşitli örüntüler ve eğilimleri tespit etmek isteyen, değişkenler arasındaki ilişkileri ortaya koyan ve karar vermeye önemli katkı yapan bir süreçtir. Fakat bu süreç sadece birtakım yöntemlerden ziyade veri toplama, temizleme, model kurma ve test etme, uygulama gibi alt süreçleri de içeren aşamalı bir yapıdır (Seyrek ve Ata, 2010: 71).

Başarılı bir veri madenciliği projesi için verilerde eksik, tutarsız veya hatalı verilerin az sayıda olması önemlidir. Doğru yöntemlerin seçilmesi de önemli bilgiler elde etmek adına bir diğer önemli konudur. Sağlıklı verilerin doğru yöntemlerle analiz edilmesi kalitesiz sonuçları minimize eder. Bunlardan ötürü veri madenciliği için çeşitli süreç modelleri geliştirilmiştir. En çok kullanılan süreç modelleri SEMMA, KDD ve CRISP-DM şeklindedir. Aşağıda bu süreç modellerinin açıklamalarına yer verilmiştir.

1.3.1. SEMMA Modeli

SEMMA süreç modeli, veri madenciliği projelerinin anlaşılmasını, düzenlenmesini, geliştirilmesini ve sürdürülmesini sağlar. Ayrıca işletme sorunlarının çözümü ve hedeflere ulaşmada için çözümler üretmeye yardımcı olur (Shafique ve Qaiser, 2014: 220).

SEMMA, örnekleme (Sample), keşfetme (Explore), düzenleme (Modify), modelleme (Model) ve değerlendirme (Evaluate) kelimelerinin baş harflerinden oluşan bir kısaltmadır. Kısaltmayı oluşturan kelimeler aynı zamanda modelin süreçlerini de sırasıyla içermektedir. Bir teknoloji firması olan SAS tarafından geliştirilmiş, uygulayıcıların çeşitli veri madenciliği projelerinde kullanabilecekleri standart bir süreç modelidir (Şeker, 2018: 14). Şekil 1' de SEMMA modeline ilişkin hiyerarşik gösterimi bulunmaktadır.



Şekil 1. SEMMA Modeli (Şeker, 2018)

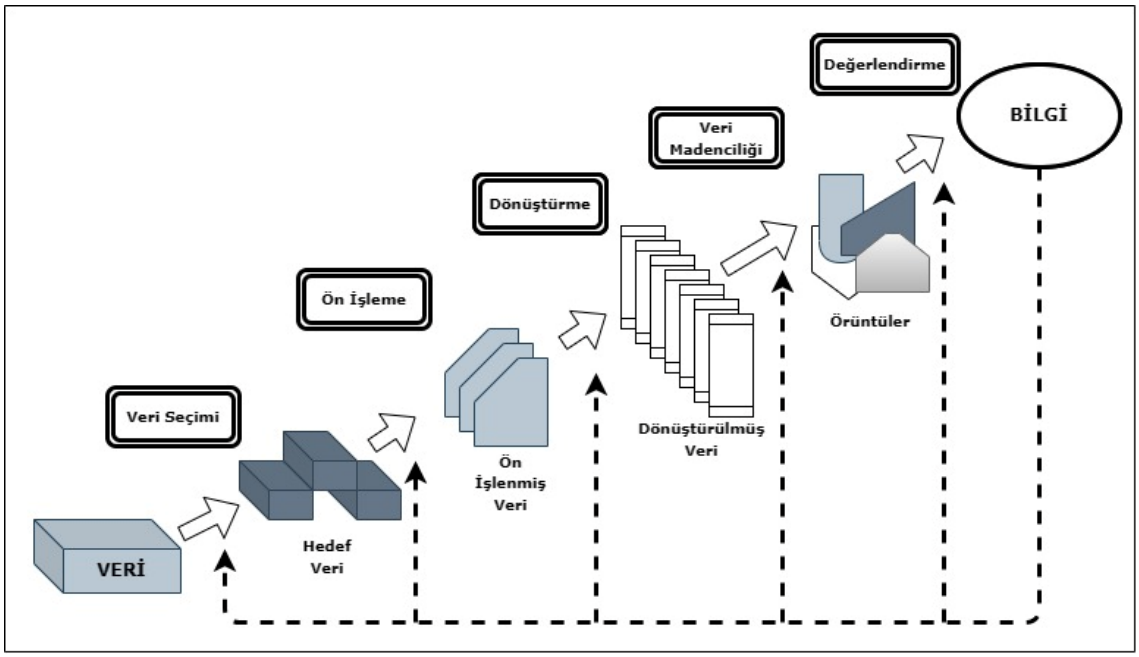
Şekil 1' de görüldüğü gibi, hiyerarşik yapının her bir kademesi sırasıyla örnekleme, keşfetme, düzenleme, modelleme ve değerlendirme aşamalarını temsil etmektedir.

1.3.2. KDD Modeli

KDD (Knowledge and Data Discovery) modeli “Veri Tabanlarında Bilgi Keşfi (VTBK)” şeklinde çevrilmiştir. Araştırma amaçlı konularda kullanımı daha fazladır. VTBK kavramı ilk olarak 1995 yılında Montreal’ de veriden bilgi edilmesine ilişkin tüm süreci tanımlamak amacıyla kullanılmıştır. Bu yöntem genelde araştırma amaçlı konularda kullanılır (Gürsoy, 2009: 26).

Veri madenciliği, verilerden örüntüleri çıkarmak için özel algoritmaların uygulanması sürecidir. Öte yandan veriler üzerinden belirli bir örüntü veya model üreten veri analizi ve keşif algoritmalarının uygulamasından oluşan veri tabanlarından bilgi keşfi sürecindeki bir adımdır. Veri hazırlama, veri seçimi, veri temizleme ve madencilik sonuçlarının doğru yorumlanması gibi aşamalar verilerden yararlı bilgilerin elde edilmesi için son derece önemlidir (Fayyad vd., 1996: 39).

Matheus vd. (1993)’ e göre Veri Tabanlarında Bilgi Keşfi, çok sayıda ticari ve bilimsel alanda yüksek getiri vaat eden aktif bir araştırma alanıdır. Kurumsal, kamusal ve bilimsel topluluklar çevrimiçi veri tabanlarında rutin olarak tutulan yoğun bir veri akışına maruz kalmaktadırlar. Bu verileri zamanında analiz etmek ve anlamlı kalıplar oluşturmak, bilgisayar yardımı ve güçlü analitik araçlar olmadan çok zor bir iş olabilir. Bununla birlikte, tek başına standart bilgisayar paket istatistiksel ve analitik programlar alanında uzmanlaşmış istatistikçilerin katkısı olmadan sınırlı fayda sağlar. Veri tabanlarında bilgi keşfinin en büyük zorluğu, büyük miktarda ham veriyi otomatik olarak işlemek, bunlar içinden en önemli ve anlamlı kalıpları belirlemek ve de bunları kullanıcının hedeflerine ulaşmak için kullanabileceği uygun bilgi halinde onlara sunmaktır. Şekil 2’ de VTBK adımları verilmektedir.



Şekil 2. Veri Tabanlarında Bilgi Keşfi Adımları (Fayyad, 1996: 1590; Baldaniya vd., 2014: 442).

VTBK Modelinin aşamaları aşağıda gösterilmiştir (Terzi vd., 2011: 31; Albayrak ve Koltan Yılmaz, 2009: 36; Savaş vd., 2011: 8):

a. Problemin tanımlanması

Uygulamanın hangi amaçla yapılacağının açık seçik bir biçimde tanımlanması veri madenciliği sürecinin ilk ve en önemli, adımıdır. Amaç, kesinlikle sorun üzerine odaklanmalı ve elde edilecek sonuçların başarı durumlarının ne şekilde ölçüleceğinin tanımlanması olmalıdır. Problem ile örtüşmeyen model sadece problemi çözememekten ziyade yeni problemlerin de ortaya çıkmasına sebep olur. Bununla birlikte yanlış kararlar ile katlanılacak ayni ve maddi zararlar ile doğru kararların getirileri de bu aşamada belirlenmelidir.

b. Verilerin hazırlanması

Veri hazırlama, ham veriden, son veriye kadar yapılan bütün düzenlemeleri içeren aşamadır. Çeşitli veri madenciliği yöntemleri için veri hazırlama süreci, tablo, kayıt, veri dönüşümü gibi uygulamaları içermektedir.

Bu aşama modelleme esnasında yaşanacak çeşitli sorunları ortadan kaldırmak için oldukça önemlidir. Sağlıklı sonuçlar elde etmek için verilerin hazırlanması, vakit ve enerji açısından veri keşfi sürecinin %50- %85' ini kaplamaktadır. Verilerin hazırlanması süreci genel olarak "verilerin toplanması", "verilerin temizlenmesi" ve "indirgeme" işlemlerini içeren alt süreçlerden oluşmaktadır.

c. Modelin kurulması

Çok model denenmek suretiyle uygun modelin bulunması bu aşamanın yinelemeli bir yapıda olduğunun göstergesidir. Çözüme ulaştırılması beklenen problem için hem veri hazırlama hem de model kurma çok sayıda yineleme gerektirebilir.

d. Modelin kullanılması

Güvenilirlik ve geçerliliği kabul edilen model kendi başına uygulanabilir ya da başka bir modelin alt parçası olarak da kullanılabilir.

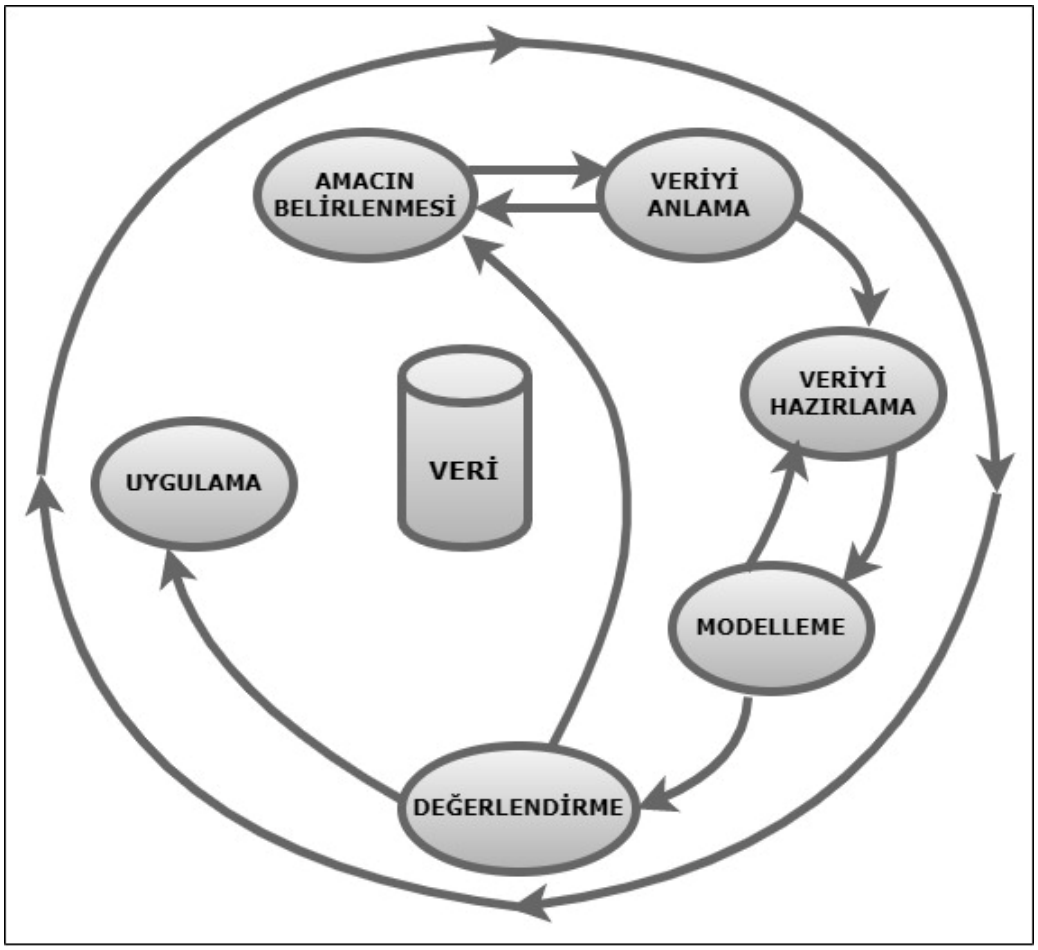
e. Modelin izlenmesi

Kurulan modelin performansının izlenmesi her disiplinde gerekli bir uygulama olmakla beraber zaman zaman revizyon gerekebilir. Sürecin önceki aşamalarına dönmek suretiyle gerekli iyileştirmeler yapılmalıdır.

1.3.3. CRISP-DM Modeli

CRISP-DM (CRoss Industry Standard Process for Data Mining) Modeli, "işi anlamak, veriyi anlamak, veri hazırlama, modelleme, değerlendirme, sahaya sürme" aşamalarını içeren 6 aşamalı yinelemeli ve etkileşim içeren bir süreçtir. CRISP-DM Daimler Chryrler, SPSS ve NCR şirketlerinin oluşturduğu bir konsorsiyumun çabaları sonucunda geliştirilmiştir. Bu metodoloji, bir veri madenciliği projesinin yürütülmesinde yapısal bir yaklaşım ortaya koymaktadır (Peker ve Kırbaş, 2016: 86).

Veri madenciliği projeleri, yürütüldüğü alanla ilgili çözüm arayışı içerisindeyken doğru verinin elde edilmesi, veri dönüşümü, uygun veri madenciliği modelinin seçilmesi, sonuçların değerlendirilmesi ve raporlamasında standart bir yapı çerçevesinde yürütülmelidir. CRISP-DM, kullanılan teknolojiden ve sektörden bağımsız olarak veri madenciliği projelerinin az maliyetle, daha güvenilir ve daha hızlı bir şekilde yürütülmesi için gerekli sistematik yapıyı sağlayan bir süreç olarak tanımlanabilir (Cihan vd., 2018: 86). Şekil 3' te CRISP-DM Modelinin aşamaları gösterilmektedir.



Şekil 3. CRISP-DM Süreç Modeli Aşamaları (Cihan vd., 2018: 86; Çınar ve Arslan, 2008: 306)

CRISP-DM Modelinin aşamaları aşağıda ayrıntılı bir şekilde açıklanmıştır (Çınar ve Arslan, 2008: 306).

a. Amaç ve Hedefin Belirlenmesi (Business Understanding): Amaç ve hedeflerin belirlenmesiyle birlikte bu aşamadan elde edilen veriler doğrultusunda problem tanımlanır ve taslak plan oluşturulur.

b. Veriyi Anlama (Data Understanding): Veri yapısı, kalitesi, veri hakkındaki ilk izlenimler ortaya konulur. Keşifçi veri analiz yöntemleri ile veri özellikleri, hatalı ve eksik verilerin de durumu tespit edilir.

c. Veriyi Hazırlama (Data Preparation): Bu aşamada veri, analizler için hazır hale getirilmesi için gerekli işlemlerden geçer. Veri kalitesinin yükseltilmesi için yapılan çalışmalar analizlerin de daha sağlıklı yapılmasını sağlar. Birden çok tablo veya kayıt için bütünleme (entegreasyon) işlemi yapılıp yeni kayıtlar oluşturulabilir, gerektiğinde dönüştürme veya normalizasyon işlemi yapılabilir.

d. Modelleme (Modeling): Veri türüne ve proje amaçları doğrultusunda ne şekilde modelleme yapılacağına karar verilir. Modellerle ilgili çeşitli ayarlamalar yapılır. Örneğin parametrelerin, eğitim ve test gruplarının belirlenmesi bu aşamada yapılır. Veri yapısı ve kalitesine göre işlemlerin türü ve tekrarlanma sayısı değişebilir.

e. Değerlendirme (Evaluation): Modellerin güvenilirlik ve geçerliliği ile ilgili bir değerlendirme yapılır. Ayrıca nihai uygulamaya geçmeden hedef ve amaçlara uygunluk açısından bir değerlendirme yapılır.

f. Sonuçları Kullanma/Uygulama (Deployment): Bu aşamada elde edilen bulguların ne şekilde uygulanacağı belirlenir. Projenin durumuna göre sadece raporlama yapılabileceği gibi sürece tekrardan dönülebilir. Genel olarak bu aşamada sonuçlar ilgili birim ya da kuruma bağlıdır.

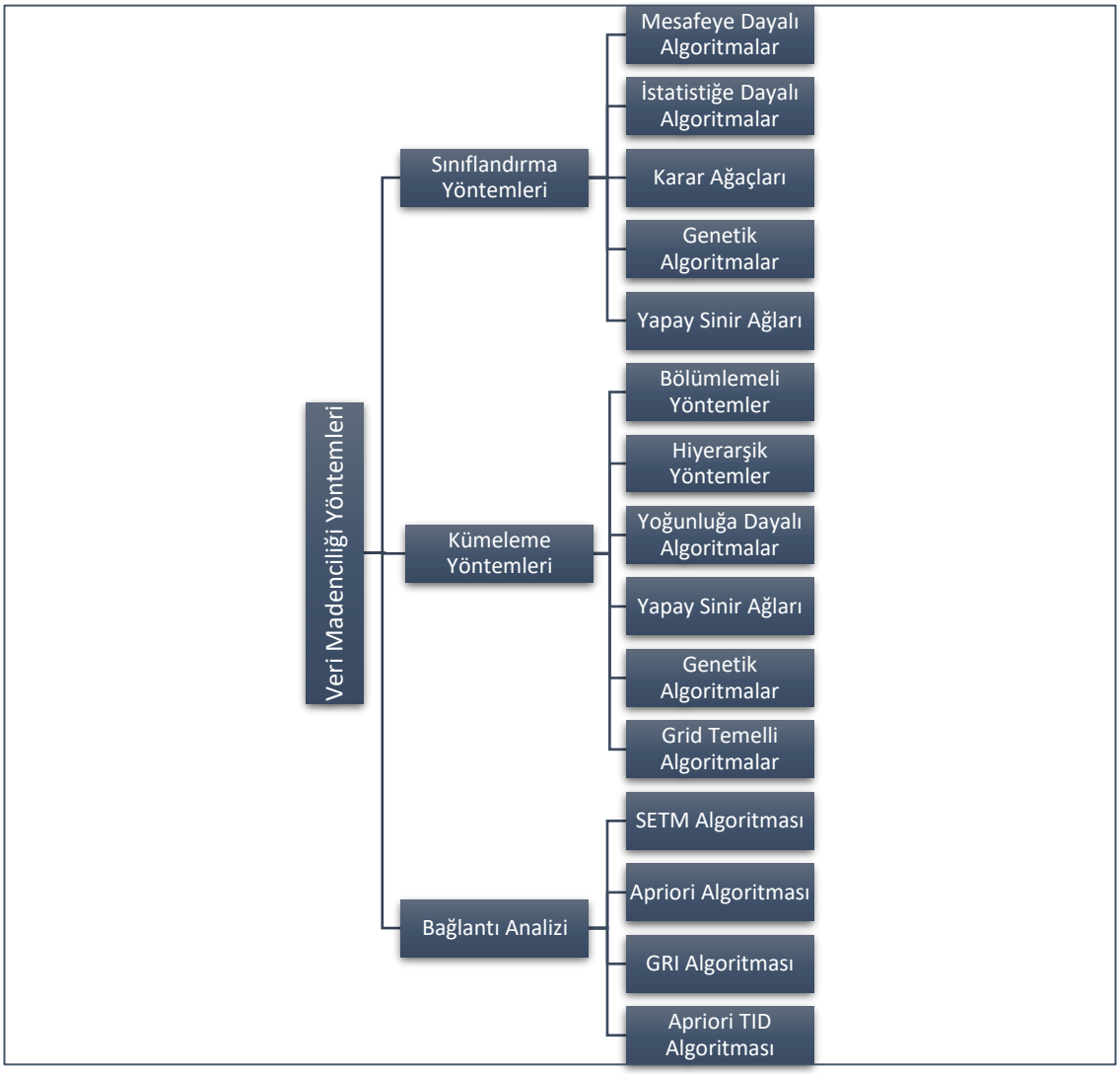
1.4. Veri Madenciliği Yöntemleri

Veri madenciliği konusunda birçoğu istatistiksel tabanlı olmak üzere çok sayıda yöntem geliştirilmiştir (Özkan, 2020: 32). İş dünyası günümüzde önemli ekonomik dalgalanmalar, yeni trendler ve sürekli değişen dinamik bir pazar yapısı içinde yüksek risk taşıyan kararlar almak durumundadır. Bu tip kararların sadece tecrübe ve önseziye dayalı verilmesi karar vericileri büyük sıkıntılara sokabilir. Veri madenciliği yöntemleri sağlıklı kararlar verilmesi adına oluşturulacak karar destek sistemlerine önemli katkılar sunabilir. Bunun için de bilgi teknolojilerinden yararlanılmalıdır (Savaş vd., 2012: 3).

Veri madenciliği uygulamalarında kullanılan veri tabanları içeriğinde binlerce kayıt veya değişken barındırabilir. Bu kadar fazla değişkenin aralarında bir korelasyon olmaksızın tamamının bağımsız olması olası bir durum değildir. Veri tabanlarında bilgi keşfi sürecinde bahsedildiği gibi veri bilimi ile uğraşanların bazı tahmin edici değişkenler arasında görülen çoklu bağlantı durumuna karşı dikkatli olmaları gerekir. Çoklu bağlantı durumu çözüm uzayında birtakım kararsızlıklara yol açarak tutarsız sonuçlar elde edilmesine sebebiyet verebilir. Bu tür durumlardan kaçınmak için tekrar eden değişkenlerden kaçınmak gerekir (Larose, 2006: 1).

Veri madenciliği, çok geniş uygulama alanı olan ve çok sayıda yönteme ve algoritmaya sahip bir disiplindir. Veri yapısına uygun yöntemlerin kullanılması sağlıklı sonuçlar elde etmek için son derece önemlidir. Veri madenciliği yöntemleri genel olarak sınıflandırma, kümeleme ve birliktelik kuralları şeklinde üç başlık altında incelenebilir

(Özsürünç, 2022: 17; Kör, 2017: 34). Veri madenciliği yöntemleri Şekil 4' te gösterilmiştir.



Şekil 4. Veri Madenciliği yöntemleri (www.veribilimiokulu.com, 2023)

Şekil 4' te görüldüğü gibi veri madenciliği yöntemleri sınıflandırma yöntemleri, kümeleme yöntemleri ve bağlantı analizi şeklinde üç ana başlık altında gruplandırılmıştır. Bu gruplara dahil olan çeşitli algoritmalar çalışmanın devamında açıklanmıştır.

1.4.1. Sınıflandırma Yöntemleri

Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi sınıflandırma olarak adlandırılır (Özkan, 2020: 39). Makine öğrenmesinde sıklıkla başvurulan yöntemlerden birisi olan sınıflandırma, denetimli öğrenme yöntemlerinden biridir. Yöntem öncelikle veri setinin eğitim ve test şeklinde ikiye ayrılması ile başlar. Eğitim verisi kullanılarak model hazırlanıp, test verisi ile de bu modelin test edilmesi amaçlanır. Oluşturulan model

yardımıyla bir sonraki gelecek verinin hangi sınıfa ait olduğu tahmin edilmeye çalışılır. Sınıflandırmadaki temel amaç benzer özelliklere sahip verinin hangi gruba ait olduğunun tahminidir (Alan ve Karabatak, 2020: 533).

Sınıflandırma, veri madenciliğinde en çok bilinen uygulamasıdır. Veri setini oluşturan nesnelerin sınıflayıcı bir model tarafından ilgili sınıflara atanması sürecidir. Başka bir deyişle eldeki nesnelerin hangi sınıfa atanması gerektiğinin ya da atanmaması durumu varsa bunun belirlenmesidir. Modelde girdiler her biri bir sınıf etiketi ile etiketlenecek gözlem veya örneklerden oluşan bir eğitim kümesinden, çıktılar ise modelin her bir gözleme niteliklere dayalı olarak atadığı sınıf etiketlerinden oluşur (Emel, 2005: 224).

1.4.1.1. İstatistiğe Dayalı Algoritmalar

İstatistiğe dayalı algoritmalar Bayes Sınıflandırması, Regresyon Analizi ve CHAID Karar Ağacı Algoritmasıdır.

a. Bayes Sınıflandırması

Bayes sınıflandırması, koşullu olasılıklara dayalı bir sınıflandırma algoritması olup veri seti içinden rastgele seçilen bir verinin hangi olasılıkla hangi sınıfa ait olacağını tespit etmeye çalışır. Bu yöntem, koşullu olasılık ilkeleri doğrultusunda verilerin sınıfını tespit etme amacı güden çeşitli hesaplamalar ile çalışmaktadır (Kır Savaş vd., 2014: 347). Bayes sınıflandırma metodu aşağıdaki formül ile ifade edilebilir.

$$\left(A_i \cap A_j\right) = \emptyset, i \neq j \quad (1)$$

$$P\left(\frac{A_i}{B}\right) = \frac{P\left(\frac{B}{A_i}\right) * P(A_i)}{\sum_{i=1}^{k} P\left(\frac{B}{A_i}\right) * P(A_i)} \quad (2)$$

b. Regresyon Analizi

Regresyon denklemi, herhangi bir bağımlı değişkenin bir veya daha fazla bağımsız deşişken ile arasındaki ilişkinin matematiksel ifadesidir. Regresyon denklemi ile bağımlı değişkeninin, bağımsız değişkenler karşılığında alacağı değerler tahmin edilmeye çalışılır. Tabi burada aradaki ilişkinin gücünü derece olarak ifade eden korelasyon katsayıda ayrı bir durumu ifade eder. Regresyon analizi bağımsız değişken sayısı doğrultusunda "basit regresyon analizi" ve "çoklu regresyon analizi" şeklinde ikiye ayrılır (Şimşek Gürsoy, 2009: 91).

Regresyon analizi, deneysel olmayan çalışmalardan elde edilen verilerden nedensel durumların belirlenmesinde önemli bir araç olduğu için sosyal araştırmalarda sıklıkla kullanılır. Regresyon analizi, bir bağımlı değişken ile bir veya birden fazla bağımsız değişken arasındaki olası ilişkilerin tespit edilmesi modellenmesi için kullanılan bir tekniktir. Sağlıklı kurulan bir regresyon modeli değişkenler arasındaki ilişkiyi doğru yansıtıyorsa, bu model bağımlı değişkenin tahmin edilmesi, önemli bağımsız değişkenlerin belirlenmesi ve bu değişkenler arsındaki ilişkinin nedenselliğinin ortaya çıkarılması için başarılı olabilir (Budak, 2021: 13).

c. CHAID Algoritması

Karar ağacı algoritmaları istatistikçiler tarafından 1970'lerde kullanılmaya başlanmıştır. CHAID algoritması Kaas (1980) tarafından karar ağacı oluşturmak için geliştirilmiştir (Albayrak ve Koltan Yılmaz, 2009: 33). Başlangıçta tasarlanan Ki-kare otomatik etkileşimli algoritma yöntemi daha sonra CHAID (Ki-kare otomatik etkileşim dedektörü) olarak öz nitelik için ilişki düzeyine göre anlamlı farklılıkları olan değer çiftlerini tespit edebilmektedir. Burada anlamlı farklılıklar istatistiksel analizlerden elde edilen oransal değer ile temsil edilmektedir. Bu yöntemde kullanılan hedeflenen öznitelik değeri sürekli ise F, nominal ise Pearson Ki-kare, sıralı ise olasılık oran sıralaması kullanılmaktadır. CHAID ile diğer algoritmaların en önemli farkı CHAID' ın çoklu karar ağaçları türetebilmesidir. CHAID yöntemi veri setinin daha homojen bölümlemesi için güçlü bir tekniktir (Büyükkıran, 2020: 4). CHAID algoritması, bağımlı değişkenin sürekli veya kesikli olması ile ilgili bir kısıta sahip olmadığı için pratik ve dolayısıyla çok tercih edilen bir modeldir (Kılıç ve Turhan, 2022: 27).

1.4.1.2. Mesafeye Dayalı Algoritmalar

Mesafeye dayalı algoritmalar; mevcut veri setinde her bir öğe arasındaki mesafelere göre çalışan algoritmalardır.

a. KNN Algoritması

Mesafeye temelli yöntemlerden birisi de en yakın k-komşu (K-nearest neighbors - KNN) algoritmasıdır. Uygulama alanı geniş olan bu yöntem sınıfları belli olan örnek kümesindeki gözlem değerlerinden yaralanarak yeni gelen değerin hangi sınıfa dahil edileceğinin belirlenmesi amaçlar. Bu yöntem veri kümesinde yer alan gözlem değerlerinin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının

hesaplanıp, en küçük uzaklığa sahip **k** sayıda gözlemin veri setinden seçilmesi temelinde çalışır. Bu yöntemde uzaklık hesaplanmasında Öklid uzaklık formülü kullanılabilir (Özkan, 2020: 141).

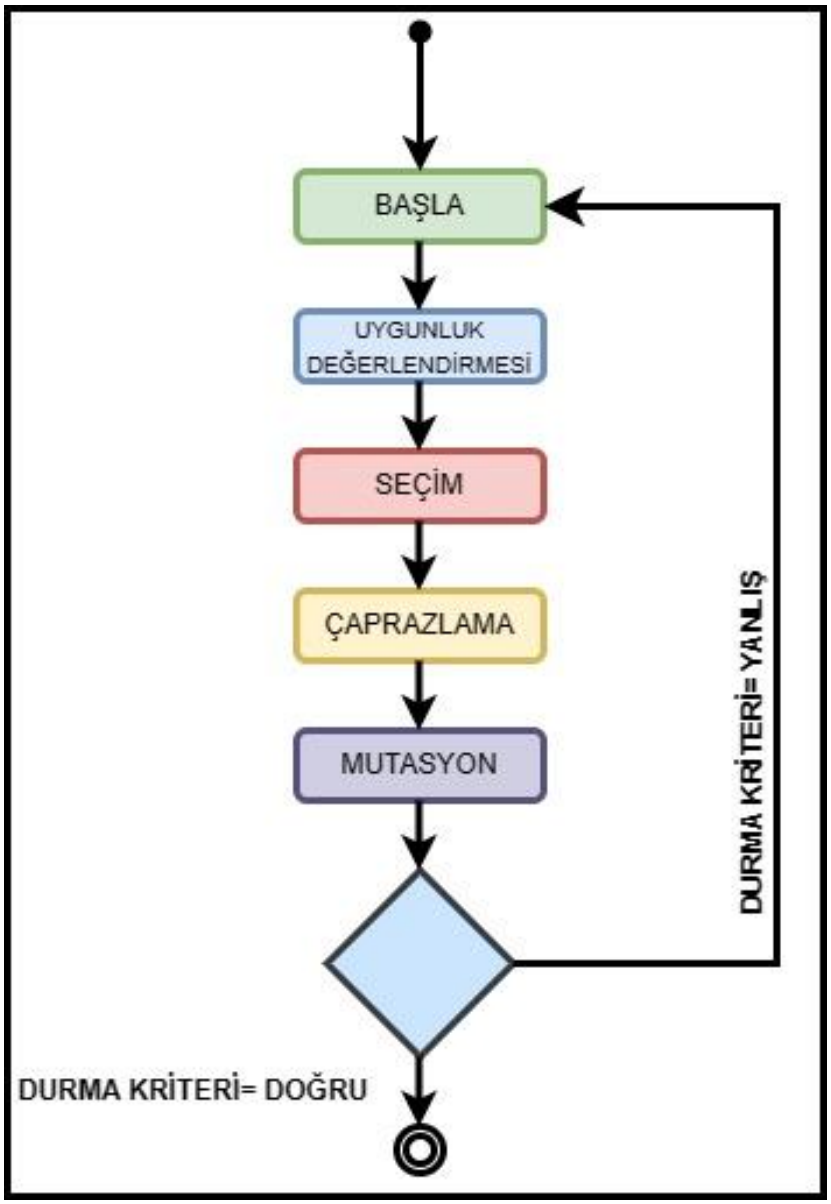
$$d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ij} - x_{jk})^2} \quad (3)$$

Veri setindeki kayıtların uzaklık ölçüsüne bağlı olarak sınıflanmasını sağlayan yöntemlerden biridir. Burada önemli olan eğitim veri setinin kategorisinin bilinmesidir. Yeni bir kaydın hangi sınıfa ait olacağının tespiti için bu noktanın eğitim veri setindeki tüm noktalara olan uzaklığı hesaplanır ve elde edilen sonuçlar küçükten büyüğe sıralanır. Önceden belirlenmiş komşu sayısı parametresi dikkate alınmak suretiyle en yakın uzaklıklar seçilir. Seçilen K tane farklı sınıf eğitim veri noktaları içerisinde yeni kaydın sınıfını belirlemek için genellikle oylama veya ağırlıklı ortalama yöntemi kullanılır. Çoğunluk oylamasına göre, yeni örnek için sınıf tespiti yapılırken K adet komşu arasında en çok tekrarlanan sınıf seçilir. Ağırlıklı oylama yönteminde ise yeni örnek noktasına daha yakın olan komşuların sınıflama yaparken söz hakkının yüksek olması amaçlanır (Kemalbay ve Alkış, 2021: 558).

1.4.1.3. Genetik algoritmalar

Deneysel çalışmalardan elde edilen verilerin ve veriler arasındaki bağıntıların modellenmesi zaman zaman doğrusal olmayabilir ya da anlaşılması zor olabilir. Özellikle kalabalık ve karmaşık veri yapıları arasındaki ilişkilerin ortaya konmasında sezgisel metotların veya yapay zekâ tekniklerinin kullanılması kaçınılmazdır. Bulanık mantık, yapay sinir ağları ve genetik algoritma sezgisel metotların en çok kullanılan araçlarıdır (Özel ve Topsakal, 2013: 44).

Genetik algoritma, çok fazla kısıt içeren optimizasyon problemlerinin çözümü için gerekli evrimsel hesaplama yöntemidir. Problemler için evrimsel bir süreç sonunda optimum sonuca yinelemeli bir şekilde çözüm aranır (Candan vd., 2019: 31). Genetik algoritma akış şeması Şekil 5’ te verilmiştir.



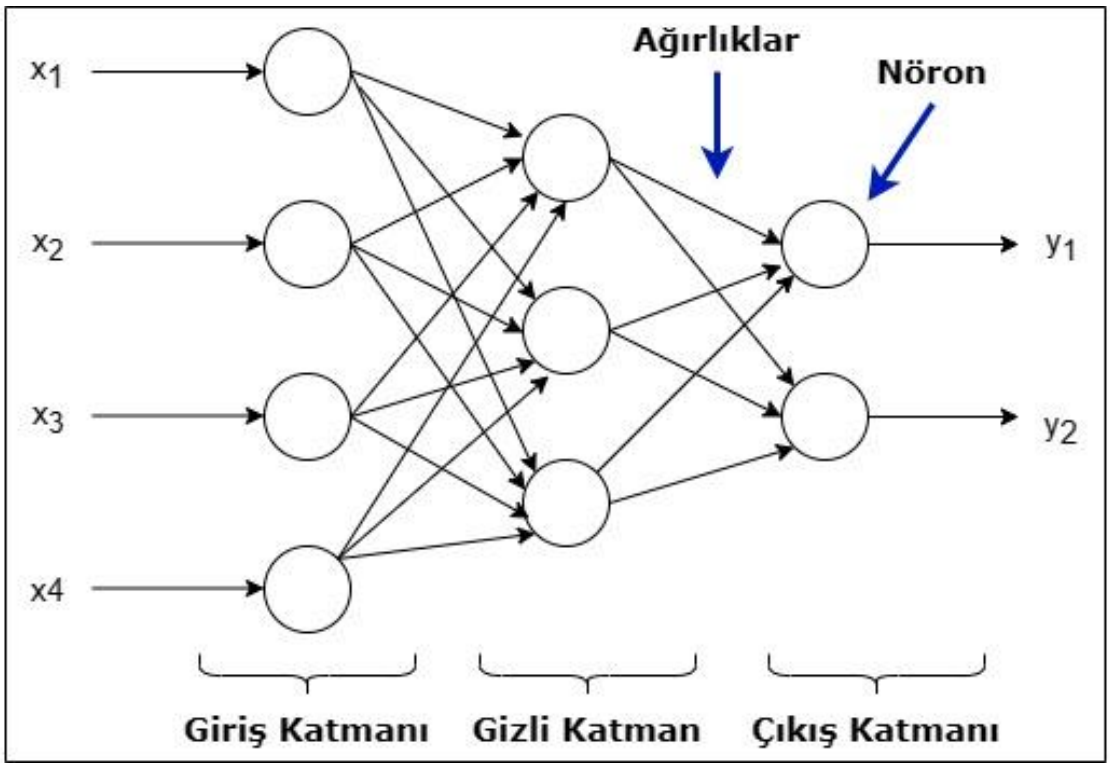
Şekil 5. Genetik algoritma akış diyagramı (Candan vd., 2019: 32)

Öncelikle rastgele bir başlangıç popülasyonu oluşturulur. Her popülasyon için uygunluk ölçütü olarak aranan değerleri için uygunluk değerlemesi yapılır. Sonrasında seçim, çaprazlama ve mutasyon gibi genetik algoritma uygulamaları çalıştırılır ve uygunluk kriterleri tekrardan gözden geçirilir. Bu işlemler belirlenen belirli bir uygunluk kriteri değerine varıncaya kadar devam eder. İstenen çözüm kümesine ulaşıldığında sistem en iyi çözümü çıktı olarak verir (Aktürk, 2020: 352).

1.4.1.4. Yapay Sinir Ağları

Yapay sinir ağları (YSA) insan beyninin bilgi işleme yapısından esinlenerek ve insan beyninin işlevsel özelliklerine benzer şekilde oluşturulmuş, örnekler üzerinden olayları öğrenebilen, sınıflandırma yapabilen, verileri optimize edebilen, bir sonucu

tahmin eden başarılı bir yöntemdir. Karmaşık konularda çıkarımlarda bulunmak için insan beyninin bir canlandırması gibi işleyerek insan beynindeki nöron ağına benzer şekilde birbirine bağlı düğümler şeklinde modellenir. Doğrusal olmayan, eksik ve gürültülü verinin olduğu durumlarda başvurulan bir yöntemdir. YSA bünyesinde yer alan yapay sinir hücreleri süreç elemanları olarak adlandırılır. Bunlar "girdiler", "ağırlıklar", "toplama fonksiyonu", "aktivasyon fonksiyonu" ve "çıkış" şeklindedir. Şekil 6' da YSA modeli gösterilmiştir (Aghalarova ve Bozkurt Keser, 2021: 22).



Şekil 6. YSA' nın Yapısı (Aghalarova ve Bozkurt Keser, 2021: 22)

YSA, bilgiyi depolamak ve analitik bir şekilde kullanıma hazır hale getirmek için basit işlem birimlerinden oluşan büyük ölçüde paralel dağıtılmış bir yapıdadır. YSA, sınıflandırma gibi çok değişkenli ve doğrusal olmayan modelleme problemlerini çözmedeki sağlamlığı nedeniyle genellikle bir vekil model veya yanıt yüzeyi yaklaşım modeli olarak kullanılır (Bre vd., 2018: 1430).

Şekil 6' da görüldüğü üzere YSA' yı çok katmanlı yapılar oluşturur. Çeşitli ortamlardan elde edilen veriler giriş katmanında işleme girer, buradan ortada bulunan katmana iletilir. Bu katmanlarda işleme alınan bilgi ileri, geri besleme ile değerlendirilir.

1.4.1.5. Karar ağaçları

Emel ve Taşkın (2005)' a göre karar ağaçları basit öğrenme algoritmasına sahip önemli sınıflandırma araçlarındandır. Açığa çıkarılan sonuçlar rahatlıkla anlaşılabilir. Karar ağaçlarında sadece karar değil aynı zamanda bu kararlarla ilgili açıklamaları da içerirler. Karar ağaçları tümevarım şeklinde bir eğitim yapısına sahiptir. Bir veri kümesi karar ağacı oluşturulurken kullanılan yöntem "ağaç tümevarımı (tree induction)" şeklinde adlandırılır. Bu yöntem sınıflama veya tahmin modelleri için kullanılan, çeşitli ağaç benzeri örüntüleri tespit etmek için kullanılmaktadır. Karar ağaçları algoritmaları, çok sayıda test yapmak suretiyle hedef tahmininde en iyi seçeneği bulmaya çalışırlar. Yapılan her test karar ağacının dallarını oluşturur ve diğer test bu dallar kullanılarak gerçekleştirilir. Bu uygulama, bir yaprak düğümünde (leaf node) sonlanır. Ağacın kökünde son yaprağa kadar her yol "kural" olarak adlandırılır. Kurallar "if-then (eğer-sonra)" şeklindedir.

CHAID (Chi- Squared Automatic Interaction Detector), Exhaustive CHAID, CART (Classification and Regression Trees), ID3, C4.5, MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree), C5.0, SLIQ (Supervised Learning in Quest), SPRINT (Scalable Paralleizable Induction of Decision Trees) gibi çok sayıda karar ağacı algoritması bulunmaktadır. Bunların bazıları aşağıda açıklanmıştır (Albayrak ve Koltan Yılmaz, 2009: 40).

a. CART Algoritması

CART algoritması, Breiman ve diğerleri (1984) tarafından önerilen hem sayısal hem de nominal veri türlerini girdi değişkeni olarak kabul edebilen bir algoritmadır. Genel olarak sınıflandırma problemlerinin çözümü için bir araç olarak kullanılabilir. CART karar ağacı, ikili ve tekrar eden bir biçimde bölünen bir yapıya sahiptir. Dallanmada kriter olarak Gini indeksinden yararlanır. Gini indeksi bağımlı değişken kategorik olduğu durumlarda kullanılan bir seçim kriteridir (Öztürk, 2022: 14). Yani, Gini Indeksi, karar ağacının oluşturulmasında bir saflık ölçütü olarak kullanılır. Nitelik seçiminde düşük Gini indeksine sahip olan nitelik tercih edilir (Efeoğlu, 2022: 4). CART ağacı, oluşturma aşamasında durma ile ilgili herhangi bir kural olmadan devamlı olarak bölünmek sureti ile büyümektedir. Artık herhangi bir bölünmenin gerçekleşemeyeceği durumda, uçtan köke doğru budama işlemi başlatılır. Mümkün olan en doğru karar ağacı,

her budama işlemi sonrası bir test verisi ile değerlendirme yaparak tespit edilir (Çalış vd., 2014: 6).

b. ID3 Algoritması

Quinlan tarafından sınıflandırma işlemlerini gerçekleştirmek amacıyla geliştirilen bu karar ağacı algoritması entropi tabanlı bir algoritmadır. Bu yöntemde hangi niteliğe göre dallanmanın yapılacağı bir sistemdeki belirsizliğin ölçüsü olan *entropi* kavramı sayesinde yapılır (Özkan, 2020: 42) Entropi genel olarak karar ağaçlarında dallanmanın yapılması sırasında faydalanılan ölçütlerden biridir. Kısaca bir sistemdeki belirsizliğin ölçüsü olarak da adlandırılabilir (Timor ve Yüzbaşı Künç, 2021: 5). Entropi, bir veri setinde yer alan belirsizlik ve rastgelelik ölçüsünü gösterir. Birbirine benzeyen veya benzemeyen öğelerden oluşan alanları belirlemek için kullanılır. Entropi veri setinin homojenliğini, olayın gerçekleşme olasılıklarına göre hesapladığı için 0 ve 1 değerlerini alır (Takma ve Hızlı, 2023: 3).

c. C4.5 Algoritması

J. Ross Quinlan 1993 yılında ID3 algoritmasının ilkelerini geliştirmek suretiyle C4.5 algoritmasını oluşturmuştur. C4.5 sınıflandırma öğrenme algoritması, en yaygın kullanılan verimli ve iyi bilinen algoritmalardan biridir. ID3'e göre daha çeşitli ve daha yeni öğrenme algoritmalarına sahip olan C4.5 algoritması, ID3 algoritmasının güncel versiyonu olarak ortaya çıkmıştır (Kaçmaz vd., 2020: 1758).

1.4.2. Kümeleme Yöntemleri

Kümeleme işlemi, veri setinde yer alan nesnelerin benzer özelliklerde olanlarının gruplara ayrılmasını şeklinde tanımlanabilir. Her bir grup küme olarak adlandırılır ve kendi içinde benzer diğer kümelerden ise farklı olan nesnelerden oluşur. Küme içerisinde yer alan nesneler arasındaki uzaklık küme dışındaki nesnelere nazaran daha kısadır. Kümeleme yönteminde veri noktalarından ya da nesnelerden oluşan bir koleksiyonun öznitelik değerleri doğrultusunda homojen kümelerin oluşturulup tanımlayıcı bir analiz yapmak temel amaçtır (Karaatlı ve Altıntaş, 2018: 873).

Kümeleme yöntemleri; mesafe, benzerlik ya da farklılık ölçülerinden faydalanarak veri kümesinde yer alan birimleri kendi içinde homojen ve birbirleri içinde heterojen kümelere ayırmak için kullanılır. Kümeleri oluştururken kullandıkları yaklaşımlara göre farklı biçimlerde sınıflandırılabilirler. Kümeleme yöntemleri bu

çalışma kapsamında hiyerarşik yöntemler, bölümlemeli yöntemler, yoğunluğa dayalı algoritmalar, grid temelli algoritmalar şeklinde 4 başlık altında açıklanmıştır.

1.4.2.1. Hiyerarşik Yöntemler

Hiyerarşik kümeleme yöntemleri, değişkenleri birbirleri ile farklı aşamalarda bir araya getirerek ardışık biçimde kümeler belirlemeyi ve bu kümelere girecek elemanların hangi uzaklık düzeyinde küme elemanı olduğunu belirlemeye yönelik yöntemlerdir.

a. SLINK Algoritması

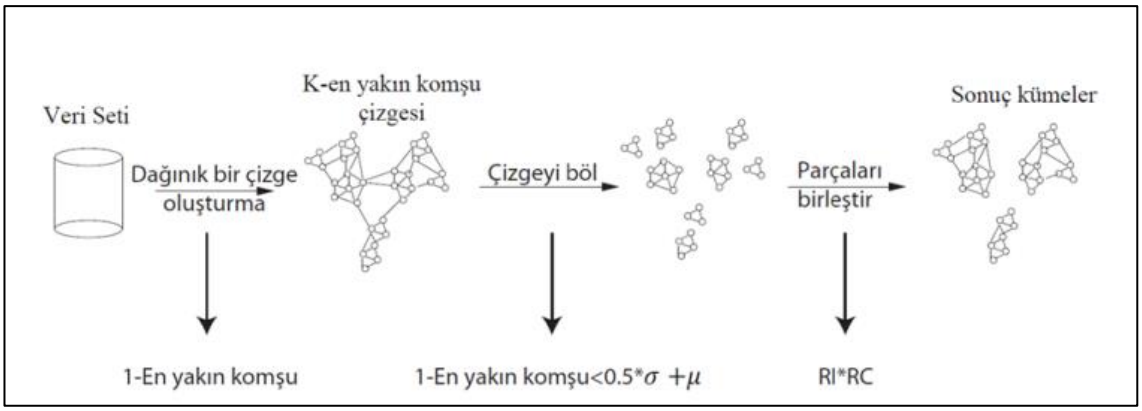
Bu algoritma, ayrık kümeler olarak ele alınan her bir kümenin aşamalı olarak birleştirildiği bir yapıda çalışır. Burada iki kümenin birbirine uzaklığı, bu kümelere dahil birbirine en yakın verilerin uzaklığı olarak alınır. Eğer uzaklık bir eşik değeri altındaysa kümeler birleştirilir (Tapkan vd., 2011: 250).

b. CURE Algoritması

Hiyerarşik bir kümeleme algoritması olan CURE (Clustering Using REpresentatives) ilk olarak 1998’ de SIGMOD konferansında Guha, Rastogi ve Shim tarafından sunulmuştur. Birleştirici bir metot olan CURE küresel olmayan ve farklı boyutlu kümeleri bulunmasında diğer hiyerarşik metotlardan ziyade daha hassas olmasından ötürü geliştirilmiştir (Demiralay ve Çamurcu, 2005: 2-3).

c. CHAMELEON Algoritması

Chameleaon algoritmasının, daha önce tartışılan bazı kısıtlardan kurtularak benzer küme çiftini belirlemede hem birbirine bağlılığı hem de yakınlığı hesaba katan bir yöntemdir. Bunun için her bir küme çifti arasındaki birbirine bağlılığın ve yakınlığın derecesini modellemek için yeni bir yaklaşım kullanır. Bu yaklaşım, kümelerin dahili özelliklerini dikkate alır. Bu nedenle kullanıcı tarafından sağlanan statik modele bağlı olmak yerine birleştirilmiş kümelerin dahili özelliklerine otomatik olarak uyum sağlayabilir (Karyips vd. 1999: 68). Şekil 7’ de Chamaleon algoritmasının iki aşamalı yapısı görülmektedir.



Şekil 7. CHAMELEON Algoritması (Karyips vd. 1999: 69)

Chameleon, önce veri öğelerini alt kümelere bölen ve daha sonra nihai kümeleri elde etmek için bu alt kümeleri tekrar tekrar birleştiren iki aşamalı bir algoritma kullanır.

Chameleon, düğümlerin veri öğelerini temsil ettiği ve ağırlıklı kenarların veri öğeleri arasındaki benzerlikleri temsil ettiği seyrek bir grafik üzerinde çalışır. Bu seyrek grafik gösterimi, Chameleon' un büyük veri kümelerine ölçekleme yapmasına ve metrik alanlarda değil, yalnızca benzerlik alanında bulunan veri kümelerini başarıyla kullanmasına olanak tanır. Bir metrik uzaydaki veri kümeleri, her veri öğesi için sabit sayıda özniteliğe sahipken, benzerlik alanındaki veri kümeleri yalnızca veri öğeleri arasındaki benzerlikleri sağlar (Karyips vd. 1999: 68).

Chameleon, veri kümesindeki kümeleri iki aşamalı bir algoritma kullanarak bulur. İlk aşamada Chameleon, veri öğelerini nispeten küçük birkaç alt kümede kümelemek için bir grafik bölümleme algoritması kullanır. İkinci aşamada, bu alt kümeleri tekrar tekrar birleştirerek gerçek kümeleri bulmak için bir algoritma kullanır.

Chameleon, *RI (Ci, Cj)* bağıl bağlanabilirlik ve *RC (Ci ,Cj)* bağıl yakınlıklarına bakarak *Ci* ve *Cj* küme çiftleri arasındaki yakınlığa karar verir (Çetinkaya, 2014: 50).

Göreceli bağlantı: *Ci* ve *Cj* kümeleri arasındaki göreceli bağlantı, *Ci* ve *Cj* kümeleri arasındaki mutlak bağlantının *Ci* ve C*j* kümelerinin dâhili bağlanabilirliğinin normalize edilmiş değeridir.

$$RI(C_i, C_j) = \frac{\left|EC_{(C_i,C_j}\right|}{\frac{1}{2}(\left|EC_{C_i}\right|+\left|EC_{C_j}\right|} \quad (4)$$

Burada; $EC_{(C_i,C_j)}$: C_i ve C_j kümelerinin her ikisini birlikte içeren kümenin kenar kesitidir (edge cut). $EC_{(C_i)}$ ve $EC_{(C_j)}$ ise çizgeyi iki eşit parçaya bölen kenarların ağırlıklı toplamıdır.

Göreceli yakınlık: iki kümenin birbirine olan yakınlığıdır. C_i ve C_j kümelerinin görece yakınlığı, C_i ve C_j kümelerinin mutlak yakınlıklarının yine bu kümelerin dahili yakınlık değerine bağlı olarak normalize edilmiş değeridir. Aşağıdaki gibi hesaplanır.

$$RC(C_i, C_j) = \frac{\bar{S}_{EC(C_i,C_j)}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_i}} + \frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_j}}} \quad (5)$$

Burada; $\bar{S}_{EC(C_i,C_j)}$, Ci ve Cj kümelerinin köşegen değerlerinin ağırlıklı ortalamalarıdır. $\bar{S}_{EC_{C_i}}$ve $\bar{S}_{EC_{C_j}}$ Ci ve Cj kümelerinin dar bölgelerinin kenarlarının ağırlık değerleridir (Çetinkaya, 2014: 50).

d. BIRCH Algoritması

Toplayıcı bir hiyerarşik kümeleme tekniği olan BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) algoritması Zhang, Ramakrishnan ve Livyn tarafından ilk olarak SIGMOD 96 konferansında sunulmuştur. Algoritma, kümelenme özelliği (Clustering Feature, CF) ve kümelenme özelliği ağacı (Clustering Feature Tree, CF-tree) şeklinde iki kavram üzerine kuruludur. Bu iki kavram kümelerin özetlenmesi büyük veri yapılarının ölçeklendirilmesine imkân vermektedir. BIRCH algoritmasının amacı mevcut veri seti ile iyi kümelerin oluşturulmasıdır. Veri setini bir kez tarayarak iyi bir küme oluştursa da opsiyonel olarak tekrardan yapılan taramalar da güvenilirliği artırır. Bununla birlikte algoritma aykırı değerlerin etkisini de azaltır (Koldere Akın, 2008: 94).

e. CLUCDUH Algoritması

CLUCDUH (Clustering Categorical Data Using Hierarchies) algoritması sadece nicel verilerin kümelenmesine olanak veren BIRCH algoritmasına alternatif olarak Silahtaroğlu tarafından (2009) sunulmuştur. Bu algoritma, kökten başlayarak sonuna kadar ağırlık dengesini korumaya özen gösterecek şekilde hesaplanan dallanmalardan

oluşan ağaç yapısıyla kümeleme yapmaktadır. CLUCDUH algoritması, eşit ayırma parametresi (EP) ile çalışmaktadır:

$$EP = \sum_{i=1}^{n} |CA - NV_i| \quad (6)$$

$$CA = \frac{N}{NY} \quad (7)$$

1- $N = bir\ değişkendeki\ kayıt\ sayısı$
2- $NV = bir\ değişkendeki\ kategori\ sayısı$
3- $NV_i = değişkendeki\ i.kategorinin\ sıklığı$

EP, 0 a yakın değerler aldığında optimuma yakındır. Bu durum öğelerin yarısı istenen değer yani "*true"* değeri aldığında mümkün olur. Ayrıca kategorilerin aldığı değerlerin de birbirine eşit olması ve tüm nesnelerin *"true"* olması durumunda *EP* sıfır olur. En küçük *EP* değerine sahip değişken kök düğüm olur ve ardından iteratif bir şekilde düğümler bulunarak ağaç oluşturulur. Bu ağaçta kök düğüm sonrası oluşan düğümler alt kümeleri temsil eder. Bu şekilde *NV* adet alt küme oluşur. Yaprak düğümden kök düğüme doğru birbiriyle aynı değeri alan benzer alt kümeler birleştirilir (Altınok, 2019: 135).

1.4.2.2. Bölümlemeli Yöntemler

Bu tür kümeleme, verileri birkaç bölüme ayırır, her bölüm bir küme olarak kabul edilir. Bu amaca ulaşmak için, başlangıç grupları oluşturulur ve son kümeleri elde etmek için belirli kriterlere göre birleştirilir (Idrıssı vd., 2015: 1).

a. K-Ortalama Algoritması

K-ortalama algoritması ilk olarak Mac Queen (1967) tarafından önerilmiştir. Bir veri kümesinde yer alan gözlem değerlerinden önceden belirlenmiş sayıda (k tane) küme oluşturan denetimsiz bir öğrenme algoritmasıdır. Bu yöntem esas olarak başlangıç noktası olarak seçilen ve küme merkezi olarak atanan değere bağlı kalarak kümeleme yapar (Özarı vd., 2019: 1118). Makine öğrenmesi, veri madenciliği, görüntü bölümleme gibi bilişim uygulamaları ile pazarlama araştırmaları, müşteri ilişkileri yönetimi mühendislik ve tıp çalışmaları gibi çoğu alanda en fazla kullanılan yöntemlerdendir. Bu algoritmanın en önemli avantajı basit olmasıdır. Bu algoritmada verilerin benzer özelliklerine

gruplandırılması için küme merkezine olan uzaklıklar esas alınmaktadır. K-ortalama algoritmasında Denklem 8 yer alan J amaç fonksiyonunun minimizasyonu hedeflenir (Temel ve Kahraman, 2021: 331).

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (8)$$

Burada j amaç fonksiyonu, k küme sayısını, n nesne sayısını ifade eder. Ayrıca $x_i^{(j)}$, j. Kümeye ait i. nesne ve c_j, j kümesinin merkezidir.

$$d_{ij} = \left\| x_i^{(j)} - c_j \right\|^2 \quad (9)$$

Denklem 9' deki d_{ij}, i. nesne' nin j. küme merkezine olan uzaklığını ifade eder.

$$c_j = \sum_{i=1}^{n_j} \frac{x_{ij}}{n_j} \; ; 1 \leq j \leq k \quad (10)$$

K-ortalama yönteminin adımları aşağıdaki gibidir (Temel ve Kahraman, 2021: 331):

1- Rastgele k adet küme merkezi seçilir.
2- Nesneler ile küme merkezleri arasındaki uzaklıklar hesaplanır.
3- Nesneler kendilerine en yakın merkezlerin ait olduğu kümelere atanır.
4- Küme merkezleri 3.denkleme göre güncellenir.
5- Küme değiştiren nesneleri yoksa birbirini izleyen 2 adımda hata karelerindeki artış tanımlanmış bir yaklaşma değerine eşit veya küçükse kümeleme işlemi sona erdirilir. Aksi durumda 2.adımdan itibaren tekrarlanır.

b. PAM Algoritması

PAM algoritması Kaufmann ve Rousseuw (1990) tarafından geliştirilmiştir. K-ortalama algoritmasının gürültülü ve aykırı değerler karşısındaki problemlerini azaltmak için nispeten daha küçük boyutlu veri yapılarında etkili sonuçlar vermektedir. Bununla birlikte daha büyük boyutlu veri setlerinde hesaplama karmaşıklığından ötürü problemler yaşanmaktadır. Bölümlemeli bir kümeleme yönetimi olan PAM algoritmasında küme merkezleri medoidlerden oluşmaktadır. Bir kümeye ait medoid, kümeyi en iyi temsil eden elemandır. Burada amaç, k adet kümede yer alan elemanların en yakın kümenin medoidlerine olan uzaklıklarının minimize edilmesidir (Akgül ve Başkır, 2013: 55).

PAM algoritmasının uygulama adımları aşağıdaki gibidir:

1) n sayıda birim (X' in gözlem değerleri) ve p sayıda değişkene ilişkin gözlemlerin oluşturduğu veri matrisi oluşturulur.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \quad (11)$$

2) k tane eleman, X matrisinin elemanlarından rastgele seçilir. Seçilen elemanlar, k-adet kümeyi temsil eden birimler olarak kabul edilir. Bu birimler medoid olarak adlandırılır.
3) Belirlenen k-adet medoid ile diğer gözlem değerlerini temsil eden elemanlar arasındaki uzaklıklar hesaplanır. Her bir eleman, medoidin temsil ettiği kümeye atanır. Bu şekilde k-adet küme oluşturulur.
4) Tekrarlamalı bir şekilde, k-adet kümede medoid olmayan elemanların tamamı medoidlerle yer değiştirilir. Bu işlem, k-adet kümenin elemanlarının en yakın küme medoidlerine uzaklıkları toplamı minimum oluncaya kadar devam ettirilir. Böylece, k-adet kümenin her biri için en uygun temsili değerin (medoidin) seçimi sağlanır (Akgül ve Başkır, 2013: 55).

c. CLARA Algoritması

CLARA (Clustering Large Application) algoritması, PAM algoritmasına göre daha büyük veri yapılarında uygulamaya uygundur. CLARA algoritması bütün veri setini tarayıp çeşitli kümeler için temsilciler seçmek yerine veri setinden rastgele temsilci bir küme alıp PAM algoritmasını bu temsilciye uygular. Bu işlem sonunda oluşan kümelerin temsilcileri belirlenir ve ana kümeyi oluşturan veri setinden bir örneklem daha seçilir. Bu aşamada ilk temsilcilerin rastgele seçilmesi yerine önceden belirlenmiş temsilciler kullanılır. Temsilci seçimi sürecinin azaltılması algoritmanın daha hızlı ve güvenilir sonuçlar vermesini sağlayacaktır. (Silahtaroğlu, 2016: 190).

d. CLARANS Algoritması

CLARANS (Clustering Large Applications based on Randomized Search) algoritması, CLARA algoritmasının farklı bir versiyonudur. CLARA algoritması veri kümesinin tamamı yerine sabit bir temsilci olmak kaydıyla PAM algoritmasını uygular. CLARANS ise her iterasyonda farklı bir temsilci küme kullanır. Kısacası CLARA

algoritmasında seçilen temsilci sabit CLARANS' ta ise farklıdır (Aslanyürek ve Mesut, 2021: 55).

1.4.2.3. Yoğunluğa Dayalı Algoritmalar

Yoğunluk tabanlı kümeleme algoritmalarında nesneler yoğunluk bölgelerine göre sınıflandırılır. Bu algoritmalar, rastgele şekil sınıflarını keşfedebilir (Idrıssı vd., 2015: 2).

a. DBSCAN Algoritması

Ester vd. (1996) tarafından ilk olarak öne sürülen bir model olan DBSCAN yoğunluğa dayalı bir kümeleme algoritmasıdır. Bu model, yalnızca bir girdi parametresi gerektirir ve kullanıcıyı buna uygun bir değer belirlemede destekler. Nesnelerin komşuları ile olan mesafelerini hesaplamak suretiyle önceden belirlenmiş eşik değerinden daha fazla nesne bulunan alanları gruplandırarak kümeleme yapar. DBSCAN, veri madenciliği uygulamalarına yaklaşım ve uygulamalar getiren bir algoritmadır (Bilgin ve Çamurcu, 2005: 139).

b. OPTICS Algoritması

DBSCAN algoritmasının geliştirilmiş bir versiyonu olarak da ifade edilebilen OPTICS algoritması Ankerst vd. (1999) tarafından önerildi. DBSCAN algoritmasıyla benzer yapıdadır. İki algoritma arasındaki farklılık ise OPTICS' te verilerdeki yoğunluk değişimlerinin anlamlı kümeler bulunması sürecinde sorun teşkil etmemesidir (Yousefi vd., 2020; 182).

c. DENCLUE Algoritması

DENCLUE (DENsity Based CLUstEring) algoritması ilk olarak Hinneburg ve Keim (1998) tarafından önerilmiştir. Bu algoritmada temel fikir, genel nokta yoğunluğunu, etki fonksiyonlarının toplamı olarak modellemektir. Daha sonra kümeler yoğunluk değer kümeleri belirlenerek tanımlanabilir ve rastgele kümeler, genel yoğunluk fonksiyonuna dayalı basit bir denklem yardımıyla tanımlanabilir. DBSCAN algoritmasıyla karşılaştırıldığında bu algoritmanın; sağlam bir matematiksel temele sahip olma, büyük miktarda gürültülü veri içeren veri kümelerinde iyi kümeleme özelliklerine sahip olma, yüksek boyutlu veri setlerinde rastgele şekillendirilmiş kümelerin matematiksel olarak modellenmesine olanak sunma ve mevcut algoritmalardan önemli ölçüde daha yüksek bir hıza sahip olma gibi üstünlükleri vardır.

1.4.2.4. Grid Temelli Algoritmalar

Grid tabanlı kümeleme algoritmalarında veriler gridlere bölünür. Algoritma doğrudan veri tabanı üzerinde uygulanmak yerine grid üzerinde taşımak suretiyle uygulanır (Idrıssı vd., 2015: 2). Aşağıda grid temelli algoritmalar açıklanmıştır.

a. STING Algoritması

Wang vd. (1997) tarafından geliştirilen STING (STatistical Information Grid-based Method) metodu, girdilerin gömülü uzaysal alanlarının dikdörtgen yapıda hücrelere ayrıldığı, grid tabanlı ve çoğul çözünürlüklü bir kümeleme tekniğidir. Burada kullanılan algoritmada katmanlardan yararlanılmak suretiyle hiyerarşik bir yapı oluşturulur ve üst katmanlarındaki hücreler alt katmanlara doğru daha küçük hücrelere bölünerek kümeleme analizi yapılır. Her hücre için yapılan sorgulama işlemlerinde, hücrede bulunan nokta sayısı, değerler ortalaması, tüm değerler standart sapması, en küçük ve en büyük değerler ve bu nesnelerin dağılımları gibi bilgiler kayıt altına alınmaktadır. Bu değerler en alttaki hücreler için otomatik, üst katmanlarda ise alt katmanların sonuçları doğrultusunda hesaplanır (Yıldız, 2014: 53).

b. Dalga Kümeleme

Dalga kümelemesi (Wave Cluster), wavelet dönüşümü kullanan bir çoklu çözüm kümeleme yöntemidir. Öncelikle veri uzayını çok boyutlu bir grid yapısına dönüştürür. Sonra wavelet dönüşümü kullanarak yoğun bölümleri tespit edip orijinal uzayda dönüşüm yapar. Farklı seviyelerdeki çözümlerde bağıl mesafe verimliliği wavelet dönüşümü sayesinde korunur. Bu durum, doğal kümeleri daha rahat seçilebilir hale getirir. İlgili alanlardaki yoğun bölgeler taranarak kümeler tanımlanabilir. Noktaların yoğun olduğu bölgeler vurgulanır, aynı zamanda zayıf bilgileri küme dışında bırakır. Bunun anlamı; veri kümelerinde yer alan veriler otomatik olarak belirlenir ve bölgelerde temizlik yapılır. Wavelet dönüşümü sınır dışındaki verileri otomatik olarak temizler. Wavelet dönüşümü, kümelerin farklı seviyelerdeki doğruluğunu keşfeder (Karaibrahimoğlu, 2014: 57).

c. CLIQUE Algoritması

Agrawal vd. (1998) tarafından önerilen CLIQUE, çok boyutlu veri yapılarında kümeleme yapmak amacıyla grid (ızgara) ve yoğunluk tabanlı yaklaşım benimseyen bir algoritmadır. Bununla birlikte her bir boyut ızgara yapısı gibi bölümlenir ve hücrelerin

sahip olduğu nokta sayısına göre yoğunluğu ölçer (Yousefi, 2020: 184). Yöntemin adımları aşağıdaki gibidir:

1- Her boyut aynı sayıda eşit mesafelere ayrılır.
2- N boyutlu bir veri alanı farklı dikdörtgen dörtgen birimlere bölünür.
3- Toplam gözlem değerlerinin oranı, başlangıç modeli parametre değerinden yüksek bir değer alırsa birim yoğundur.
4- Bir alt düzey içindeki bağlı yoğun birimler bir kümeyi oluşturur.

CLIQUE algoritmasının, veri giriş sırasına önem vermemesi, ölçeklenebilir olması, veri boyutu büyüdükçe performansının artması gibi avantajları vardır. Bunun yanında bazı dezavantajları mevcuttur. Izgara boyutunun ve yoğunluk eşiğinin belirlenmesi gerekir. Eşik değeri sabit olursa kümeler çok farklı yoğunluk değerleri alır ki bu da kümeleme işlemini olumsuz etkiler. Maliyet yüksek olabilir. Düşük ve yüksek boyutlar için aynı yoğunluk değerine sahiptir (Yousefi, 2020: 185).

1.4.3. Birliktelik Kuralları Analizi

Birliktelik kuralları analizi, veri seti içinde birlikte sık işlem gören kayıtların belirli bir güven eşik değeri yardımıyla tespit edilip analiz edilmesine dayanır. Bu şekilde veri seti içindeki ilişkilerin ortaya çıkarılması amaçlanır. Sıklıkla verilen birliktelik analizi örneği market sepet analizidir. Burada beraber satın alınan ürün gruplarının tespit edilmesi ve bu doğrultuda satış stratejilerinin oluşturulması amaçlanmaktadır (Sabah ve Bayraktar, 2020: 73).

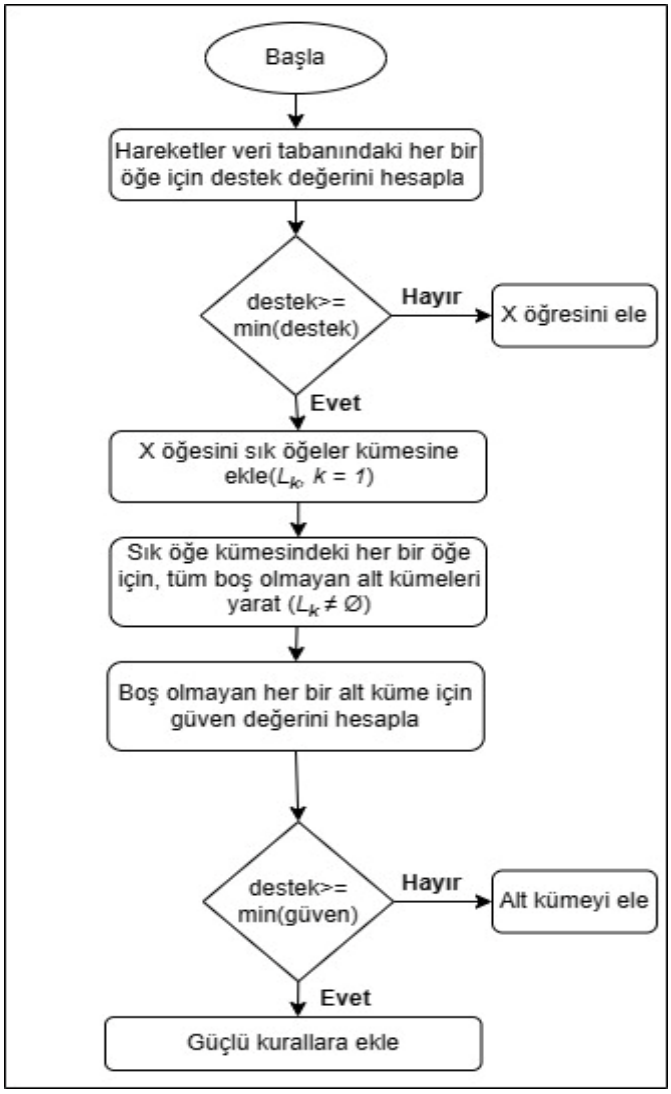
Birliktelik kuralları analizi ilk olarak Agrawal vd. (1993) tarafından ele alınmış olup kullanılan ilk veri madenciliği tekniklerden birisidir (Agrawal vd., 1993). Belirli olayların bir arada gerçekleşme durumlarını analiz veri madenciliği uygulamalarına Birliktelik Kuralları Analizi denir. Bu uygulamalar, bir arada olma veya birlikte gerçekleşme kurallarını belirli olasılıklarla ortaya koyar. Birliktelik kuralları, geçmiş verilerin analiz edilerek bu veriler içindeki birliktelik davranışlarının tespiti ile geleceğe yönelik çalışmalar yapılmasını destekleyen bir yaklaşımdır (Veri Bilimi Okulu, 2023).

Birliktelik kuralı analizi için Apriori, SETM, GRI, Apriori TID gibi çeşitli algoritmalar kullanılabilmektedir. Bunlar aşağıda açıklanmıştır.

1.4.3.1. Apriori Algoritması

Apriori algoritması, Agrawal ve Srikant (1994) tarafından geliştirilmiş olan ve çoğunlukla market sepet analizinde birlikte satın alınan ürünlerin tespiti ve aralarındaki ilişkilerin ortaya çıkarılması amacıyla kullanılan bir birliktelik kuralı algoritmasıdır. Elde edilen veriler sayesinde firmalar müşterilerinin satın alma eğilimlerini hakkında fikir sahibi olabilirler. Bu sayede satışların artması, optimum stok miktarı ve müşteri tatmini gibi konularda başarılı politikalar geliştirebilirler (Budak vd., 2018: 216).

Öncelikle veri seti içerisinde bir öğe içeren tüm öğe kümeleri ($(k = 1\ olmak\ üzere\ L_k)$ tespit edilir. Sonrasında L_k' dan aday $(k + 1)$ öğe kümeleri elde edilir (C_{k+1}). Sadece destek değeri, minimum destek değerine büyük ya da eşit olan aday kümeleriyle L_{k+1} oluşturulur. $L_k = \emptyset$ koşulu sağlanana kadar iterasyon devam eder. Algoritmanın akışını özetleyen şema, şekil 8' da verilmiştir (Budak vd., 2018: 217).



Şekil 8. Apriori Akış Diyagramı (Budak vd., 2018: 217)

Apriori algoritması aday küme oluşturmada form azaltılmasından dolayı avantajlıdır. İlk bağlantıda tüm büyük tekli öğe kümeleri oluşturulur. Daha sonraki tüm geçişler için, yalnızca önceki geçişte büyük olduğu tespit edilen öğe kümeleri aday öğe kümeleri olarak kabul edilir. Burada temel düşünce, büyük alt kümenin öz alt kümesinin kendisinden büyük olacağıdır. Böylece ikisi, k boyutunda büyük öğe kümeleri oluşturur, bunun için gereken, (k-1) boyutunda öğe kümelerini birleştirmek. Bu şekilde, önceki algoritmalarda olduğu gibi, aday nesne kümeleri oluşturmak için çok sayıda öğe kümesinin dikkate alınması gerekmez (Agrawal ve Srikant, 1994: 3).

1.4.3.2. SETM Algoritması

SETM algoritması küme yönelimli bir algoritma olup sıralama, birleştirme ve tarama gibi basit veri tabanı ilkeleri kullanmaktadır. İlk olarak Houtsma ve Swami (1995) tarafından oluşturulan bu algoritma parametrelerin aldığı değerler aralığında basit, hızlı ve kararlıdır. SETM algoritması SQL gibi genel sorgu dilleri kullanılarak geliştirilebilir ve küme yönelimli doğası sayesinde çeşitli eklentiler geliştirmeye olanak sağlar (Houtsma ve Swami, 1995: 25).

1.4.3.3. Apriori TID Algoritması

Apriori TID Algoritması, sık kullanılan öğe setlerini anlamak için zaman serileri için uygulanabilir bir birliktelik kuralı algoritmasıdır (Sarma ve Mishra, 2016: 160).

Apriori TID, Apriori ile aynı aday oluşturma işlevine sahiptir. İlginç olan özelliği ilk geçişten sonra sayım desteği için veri tabanı kullanmamasıdır. Önceki geçişte kullanılan aday öğe kümelerinin bir kodlaması kullanılır. Daha sonraki geçişlerde, kodlamanın boyutu veri tabanından çok daha küçük olabilir, böylece okuma çabasından tasarruf edilir (Agrawal ve Srikant, 1994: 3).

1.4.3.4. GRI Algoritması

GRI Algoritması Smith ve Goodman (1992) tarafından geliştirilmiştir. Bir hipotezin bilgi içeriğinin sınanması için kullanılan j-ölçeği ile hesaplama yapmaktadır. Bu algoritma Apriori algoritmasından farklı olarak girdi ve çıktı değerlerinin numerik değerler almasına imkân vermektedir (Yaşar, 2016: 96)

İKİNCİ BÖLÜM
METİN MADENCİLİĞİ

Günümüzde veriler çok farklı yapılarda karşımıza çıkmaktadır. Bu farklılıklar verilerin analizini zorlaştırdığı gibi analiz yöntemlerinin de çeşitlendirilmesini zorunlu kılmaktadır. Gerçek hayatta yapılandırılmış verilerin yanı sıra yapılandırılmamış veri içeren metinler de içeriğinde bazı gizli ve yararlı olabilecek verileri barındırırlar. Metin madenciliği yöntemleri bu gibi verilerin açığa çıkarılmasını sağlayacak yöntemler sunar. Bu bölümde metin madenciliği ile ilgili kavramsal çerçeve ele alınmıştır.

2.1. Metin Madenciliği Kavramı

Metin madenciliği doğal dil metninden anlamlı bilgiler ortaya çıkarmaya çalışan yeni bir alandır. Gelişmekte olan bu alan Belirli amaçlar için yararlı olan bilgileri çıkarmak için metinlerin analiz edilmesi süreci olarak basitçe tanımlanabilir. Veri tabanlarında saklanan çeşitli veri türleri ile karşılaştırıldığında, metinler, yapılandırılmamış, biçimsiz ve algoritmik olarak ele alınması zor yapıdadırlar. Bununla birlikte, modern kültürde metinler, resmi bilgi alışverişi için en yaygın araçtır. Metin madenciliği genellikle, işlevi olgusal bilgi veya fikirlerin iletilmesi olan metinlerle ilgilenir. Bu tür metinlerden otomatik olarak bilgi çıkarımı son derece zordur (Witten, 2004: 1).

Temelde metin madenciliği, yapısal olmayan metinlerden, bilgi içeren yapısal metinler elde etme işlemidir. Ön işleme ve özellik çıkarımı gibi işlemleri içeren adımlar, metinlerin işlenmesi yoluyla bilgi kazanımı için önemlidir. Bu adımlardan sonra yapısal olmayan metinler, metin madenciliği yöntemlerinin kullanılacağı ve bilgisayar tarafından işlenebilecek yapısal bir şekle dönüştürülmüş olur. Büyük veri içinden bilgi keşfi bu şekilde gerçekleşir. Üretilen bu anlamlı bilgilerden kurum ya da kuruluşlar faydalanabilirler. Metin madenciliği yöntemleri esasen matematik ve istatistik temelli yöntemlerdir (Kılınç vd., 2016: 90).

Veri madenciliği ve metin madenciliği birbirine benzemekle beraber aralarındaki temel fark, metin madenciliğinin yapılandırılmış veriler yerine metin verisi ile çalışmasıdır. Bu yüzden metin madenciliğinde ilk ve en önemli adım verilerin düzenlenmesi ve yapılandırılmasıdır. Bu işlem yapıldıktan sonra verileri analiz etmek daha sağlıklı sonuçlar verecektir (Çelik, 2020: 1345).

Metin madenciliği, doğal dil işleme ve veri madenciliği yöntemlerinin bir arada kullanılmasıyla sonuçlar elde etmeye çalışır. Doğal dil işleme insanların kullandığı dillerin işlenmesi ve kullanımı ile ilgili araştırma yapan bilim dalıdır. Veri madenciliği ise geçmişle ilgili analiz ya da gelecekle ilgili tahmin yapılabilmesine yardımcı olacak örüntülerin çeşitli bilgisayar programları kullanarak büyük veri yığınları içerisinden elde edilmesidir. Metin madenciliğinde, veri madenciliği yöntemleri kullanılarak eldeki verileri gruplandırma, sınıflandırma ya da daha farklı istatistiksel analiz araçları kullanılarak modeller oluşturulmaya çalışılır. Örnek olarak; oluşturulan model, mevcut veri kümesine dahil olmayan yeni bir veri elde edildiğinde bu yeni veri hakkında tahmin yapmaya imkân verir. Metin sınıflandırma da metin madenciliği için diğer bir uygulama alanıdır ve çeşitli dokümanları bilinen sınıflara atama işlemi yapmaya yarar (Kaşıkçı ve Gökçen, 2013: 25).

Metinler üzerinde yapılan çalışmalarda kullanılan yöntemlerden olan içerik analizinden farklı olarak metin madenciliği yöntemleri, verilerdeki örtülü kalıpları veya örüntüleri veriye dayalı bir şekilde meydana çıkarmaktır (Tsantis ve Castellani, 2001: 39).

2.2. Metin Madenciliği Süreci

Metin madenciliği sürecinde araştırmanın amacına ve metinlerin yapısına göre süreç değişiklik gösterse de genel olarak metin madenciliği süreci aşağıdaki gibidir.

2.2.1. Ön İşleme

Ön işleme metin madenciliği teknikleri ve uygulamalarında çok önemli bir rol oynamaktadır. Metin madenciliği sürecinin ilk adımıdır (Vijayarani vd., 2015: 9). Ön işleme sürecinde metin verisinin eğitim ve sınıflandırma uygulamalarına hazırlanması sağlanır. Ön işleme, veri setinde yer alan sorunları ortadan kaldırmak, verinin doğal yapısını öğrenerek kaliteli analiz yapabilmek ve veriden daha anlamlı bilgi üretebilmek gibi amaçlar için yapılır (Kaşıkçı ve Gökçen, 2013: 26).

2.2.1.1. Bölütleme

Bölütleme (tokenization) işlemi ile metinde yer alan cümlelere ait kelimelerin diğer kelimelerle ilişkilerine bakılmaktadır (Bektaş ve Bektaş, 2021: 32). Web dosyaları için ilk adım HTML etiketlerden arındırılarak metinlere dönüştürülmesidir. Bölütleme, genellikle bir tür doğal dil metninin ön işlemesi olarak anlaşılır. Bölütleme, benzersiz

tanımlama sembollerine sahip hassas verilerin değiştirilmesi işlemidir (Kadhim, 2018: 24).

2.2.1.2. Etkisiz sözcüklerin temizliği

Etkisiz sözcükler (stopwords), her metin belgesinde bulunan ve yaygın olarak tekrarlanan sözcüklerin listesidir. “ve, veya, ama, fakat, çünkü, üstelik, hatta” gibi bağlaçlar veya “ben, sen, o, şu” gibi zamirlerin tek başlarına anlamları olmadığı ve bu kelimelerin kategorizasyon sürecine çok az değer kattığı veya hiç değer katmadığı için çıkarılması gerekir. Aynı nedenle özellik özel bir karakter veya sayı ise o özellik kaldırılmalıdır. Etkisiz kelimelerini bulmak için, terim listemizi sıklığa göre düzenleyebilir ve sık kullanılanları semantik değer eksikliğine göre seçebiliriz (Kadhim vd., 2014: 70).

2.2.1.3. Sayıların ve noktalama işaretlerinin temizliği

Veri kümesi, sayıları kaldırmak ve metni küçük harfe dönüştürmek işlemlerine de ön işleme sırasında tabi tutulur. Genel olarak noktalama işaretleri de analiz için bir anlam ifade etmeyebilir. Bu yüzden temizlenmesi daha uygundur. Örneğin analize konu olan metinde yer alan dört haneli bir yıl ifade ile dört haneli parasal bir ifade karışıklığa neden olabilir (Oza ve Naik, 2016: 470).

2.2.1.4. Kelime köklerine inme

Kelime köklerini belirleme (stemming), bir kelimenin kökünü belirlemek için kullanılır. Bu yöntemin amacı, çeşitli ekleri kaldırmak, kelime sayısını azaltmak, doğru eşleşen köklere sahip olmak, zamandan ve hafıza alanından tasarruf etmektir. Kelime köklerini belirlemede, bir kelimenin morfolojik biçimlerinin köküne çevrilmesi, her birinin anlamsal olarak ilişkili olduğu varsayılarak yapılır. Bu işlerler sırasında dikkat edilmesi gereken iki husus vardır:

1- Aynı anlamı taşımayan kelimeler ayrı tutulmalıdır.
2- Bir kelimenin morfolojik biçimlerinin aynı temel anlama sahip olduğu varsayılır ve bu nedenle aynı köke eşlenmelidir.

Bu iki kural, metin madenciliği veya dil işleme uygulamalarında iyi ve yeterlidir (Vijayarani vd., 2015: 10).

2.2.1.5. N-Gramlar

N-gramlar, bölütlenmiş bir belge içinde bir diziden n öğenin oluşturduğu sürekli kelime gruplarıdır. Özellikle unigramlar n = 1 olan terimlerdir, Bigramlar, n = 2 olan bitişik terim çiftleridir, trigramlar ve quadgramlar sırasıyla üç ve dört sürekli terim içerir (Bharadwaj ve Shao, 2019: 19).

2.2.2. Metin Dönüştürme

Bu süreç, ön işlemden geçen metin verilerinin yapısal bir temsilinin oluşturulması sürecidir. Yani ön işlemden geçen metin verileri sayısal formada gösterilir. Bu dönüşüm süreci belge gösterimi olarak da bilinir ve çeşitli gösterim yöntemleri vardır. Vektör uzay modeli, olasılıksal konu modeli ve istatistiksel dil modeli gibi değişik belge gösterim yöntemleri bulunmaktadır. Bunların en sık kullanılanı vektör uzay modelidir. Bu modelde her belge bir vektör olarak temsi edilir. Vektör uzay modelinde belgeler n-boyutlu bir alanda vektörler şeklinde temsil edilirler. Burada n, kelime haznesindeki benzersiz terimlerin sayısını ifade eder (Singh vd., 2017: 1780; Ağdeniz, 2017: 83; Pekin, 2020: 28).

2.2.3. Özellik Seçimi

Literatürde “Feature Extraction” şeklinde de kullanılan özellik seçimi sürecinde konu ile ilgisi olmayan veriler silinmek suretiyle yüksek boyutu azaltmak amaçlanmaktadır. Bu şekilde sadece arama uzayını küçülmek ile işlemlerin kalitesi de arttırılabilir (Polat, 2021: 140). Başka bir deyişle özellik seçimi veri setinin tamamını en iyi şekilde temsil edebilecek alt kümenin tespiti olarak tanımlanabilir. Bunun için veri setinde yer alan *n* adet özellik arasından *k* tanesinin seçilmesi işlemidir. Bu yüzden veri setinde ilgili olmayan ya da etki düzeyi düşük veriler silinerek veri seti daha anlaşılır hale getirilir. Dolayısıyla gürültülü veri daha anlaşılır hale getirilir (Forman, 2003: 1291).

Özellik seçimi, herhangi bir veri yığınına dahil olacak yeni öğenin, bütün özellikleri ile değil de öğeyi oluşturan belirli özelliklerin öne çıkarılması ile sisteme dahil edilmesi durumudur. Ayrıca sistem de bu çıkarılan özellikler üzerine kurulur. Örneğin birkaç resim içerisinden içinde çimen bulunanlar tespit edilmek isteniyor olsun. Bilindiği üzere çimenlerin ayırt edici özelliklerinden birisi yeşil olmalarıdır ve bu resimlerden yeşil tonun ağırlıkta olanlarının çimen resmi olma ihtimali yüksektir. Öyleyse sisteme giren bir resmin bütün özelliklerinin işlenmesi yerine resmin renk kodlarının dağılımının

(histogram) işlenmesi buna bir örnek daha kolay olabilir. Resim, aslında histogramından çok daha karmaşık ve büyük bir veridir. Bu veri histogramı çıkarılmak suretiyle küçültülmüş ve istenen amaca yönelik olarak işlenmiştir. Özellik çıkarımına aslında bir boyut azaltma (dimension reduction, dimensionality reduction) işlemi de denilebilir. Buna göre karmaşık ya da düzensiz olan bir veri yığının boyutları azaltılarak daha basit bir problem haline dönüştürülebilir. Doğru uygulanmış bir özellik çıkarımı işlemi ve bu özelliklere uygun bir sistem tasarımı, başarılı sonuçlar elde edilmesi için önemlidir. Ayrıca özellik çıkarımı işlemi sonucunda elde edilen ve birden fazla özelliğin temsil eden veri yapısına özellik vektörü (feature vector) denilmektedir (https://bilgisayarkavramlari.com/, 2023).

2.2.4. Veri Madenciliği Yöntemleri Kullanımı

Veri madenciliği algoritmalarının kullanılarak veriden anlamlı bazı örüntüler çıkarmak ya da verinin mevcut yapısı üzerinden çıkarımlarda bulunmak bu aşamada gerçekleşir. Burada analiz için sınıflama, kümeleme, özetleme veya bilgi erişimi gibi yöntemlere başvurulabilir (Polat, 2021: 140). Veri madenciliği algoritmalarının sağlıklı çalışabilmesi için veri içerisinde gerçekten ilgisiz olduğu düşünülen verilerden kaçınılmalıdır. Özellikle düşük derece faydası olan verilerin, yüksek derecede faydası olan verilerle karıştığı durumlarda hangi verinin elverişli olduğuna karar veren algoritmalar mevcuttur. Yani eldeki verinin tümünün kullanılması yerine maksimum bilgiyi elde edeceğimiz şekilde veri yapısını minimuma indirgemek gerekmektedir (Abu Hamde, 2018: 34).

2.2.5. Görselleştirme

Gerek veri madenciliğinde gerekse metin madenciliğinde verilerin yorumlanması için görselleştirme oldukça önemlidir. Sağlıklı bir görselleştirme süreci analizlerin de sağlıklı yapılmasına ve herkesin rahatlıkla anlayabileceği bir yapının oluşmasını sağlar. Analizler için kullanılan paket programların görselleştirme için seçenekleri vardır ya da bu uygulamalar için açık kaynak yazılımların da kütüphaneleri mevcuttur. Aşağıda görselleştirme için kullanılan çeşitli araçlardan örnekler yer almaktadır.

2.2.5.1. Kelime Bulutu

Bir kelime bulutu (wordcloud), kelime sıklığının grafiksel bir temsilidir (Nasir vd., 2020: 8). Kelime bulutu, analize konu olan metin belgesinin içeriğini çağrıştıran

birçok kelimeden oluşan bir görüntüdür. Sözcükler, belgedeki geçiş sıklıklarına ve önemlerine göre çeşitli şekil ve renklerde sunulur. Metin formu ne kadar büyük olursa, oluşum sayısı da o kadar fazla olur. Kelime bulutu metin madenciliğinde baskın kelimelerin sıklığını bulmak ve ardından çalışmanın konusuna ve durumuna göre sonuçlar çıkarmak için kullanılır (Mustaqim vd., 2020: 3).

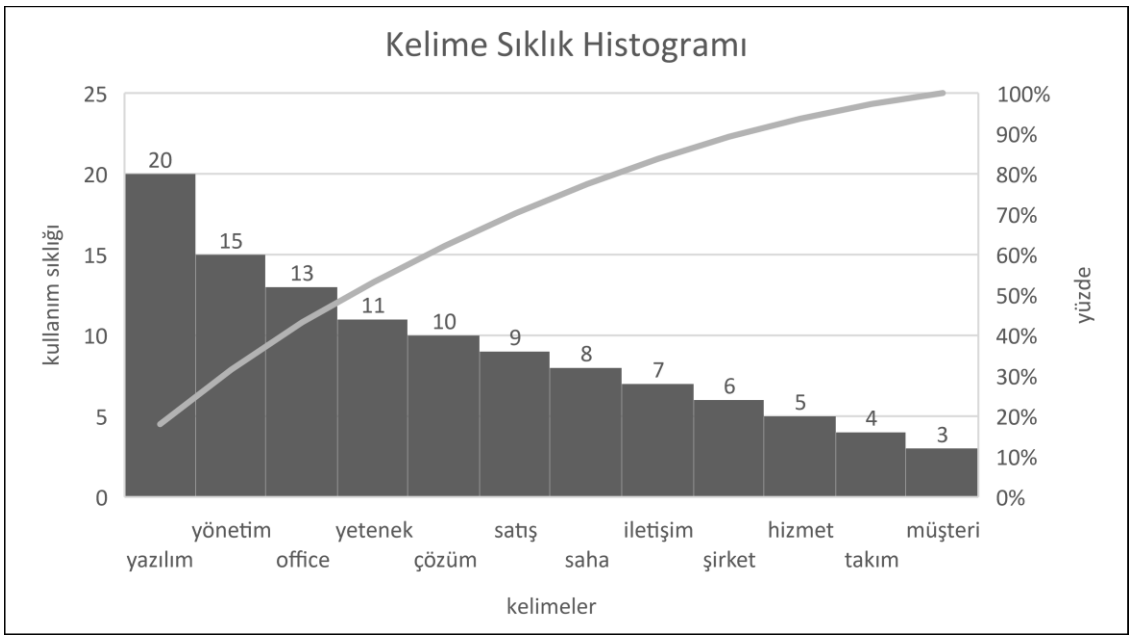
Şekil 9. Kelime bulutu örneği

Şekil 9' da görüldüğü gibi daha büyük puntoda yazılan kelimelerin kullanım sıklığı daha fazladır. Daha küçük yazılanlar ise daha az kullanılmıştır. Ayrıca farklı bir boyutu temsil etmek için renkler de kullanılabilir. Her bir renk farklı bir sınıfı temsil edecek şekilde bir uygulama da yapılabilir.

2.2.5.2. Histogram

Histogram, istatistikte ve sayısal verilerin görselleştirilmesinde kullanılan yararlı bir yöntemdir. Görsel olarak sütun grafiğine benzese de aralarında bazı farklılıklar mevcuttur. Sütun grafiği gösteriminde *X* eksenindeki her bir sütunun karşılığı bir *Y* değeri vardır. Histogramda ise *X* ekseninde yer alan değerler "bin" adı verilen kutucuklarla temsil edilirler ve her kutucuğa karşılık gelen bir Y değeri vardır. Bir başka deyişle, sütun

grafiğinde X ekseni ayrık değerlerden oluşur, histogramda ise Y ekseni süreklidir (www.medium.com, 2023). Şekil 10’ da örnek bir histogram grafiği yer almaktadır.

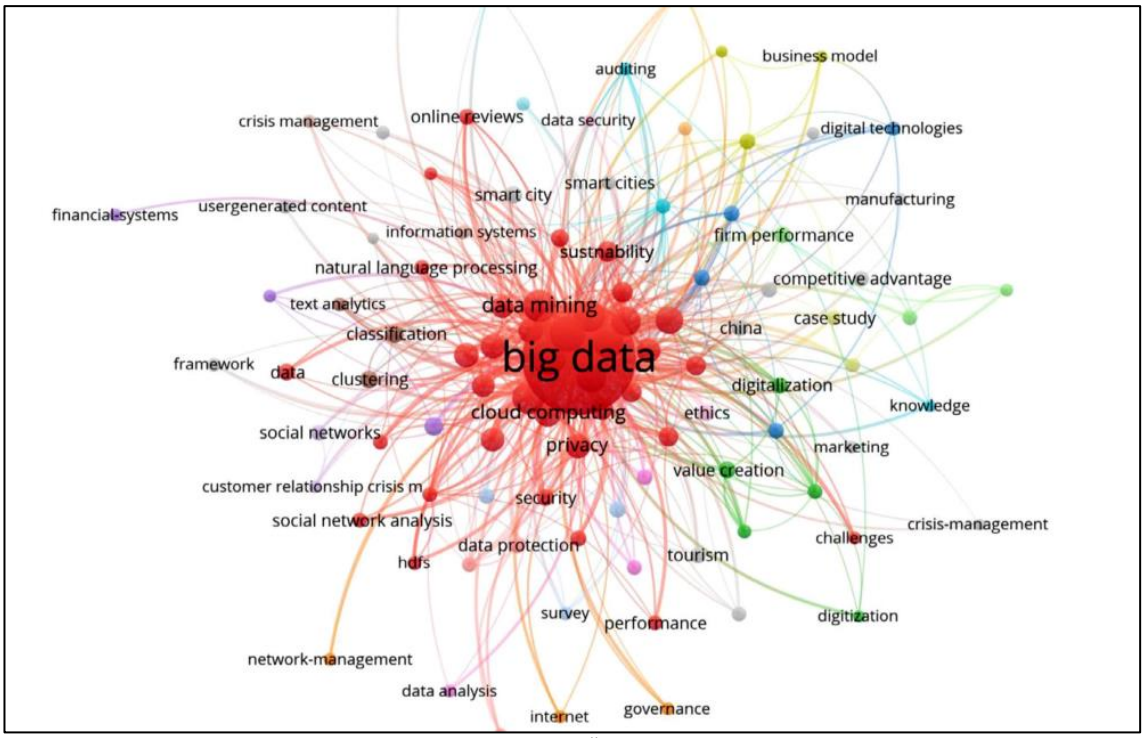


Şekil 10. Kelime Sıklık Diyagramı Histogramı

Şekil 10’ da yer alan grafikte görüldüğü üzere metinde yer alan kelimeler kullanım sıklıklarına göre histogram grafiğinde yer almaktadır. Metinde sıklıkla kullanılan kelimelerin kullanım miktarları ile bir arada görülmesini sağlayan kullanışlı bir araçtır.

2.2.5.3. Kelime Ağ Diyagramları

Ağ diyagramları, veri setinde yer alan değerlerin birbirleri ile olan bağlantılarını düğümler, köşeler ve bağlarla göstermeye yarayan bir araçtır. Bu bağlantılar, veri kümeleri arasındaki ilişkilerin tanımlanmasını sağlar. Uygulamada düğümler küçük noktalar veya dairelerle temsil edilir. Bunun yerine çeşitli şekillerden de yararlanılabilir. Bağlantılar basit çizgiler vasıtasıyla düğümleri birbirine bağlar. Bununla birlikte, çoğu zaman ağ diyagramlarında yer alan düğüm ve çizgiler eşit boyutlarda olmaz. Renk veya ebat değişkenlerin farklı kullanılmasıyla zengin bir görselleştirme sağlanabilir. Örneğin, görselleştirme esnasında düğümlerin büyüklüğü, renkleri veya çizgilerin uzunluğu ve kalınlığı değerlerin orantısına göre değişebilir (https://datavizcatalogue.com, 2023). Şekil 11’ de bir ağ diyagramı örneği görülmektedir.



Şekil 11. Ağ Diyagramı Örneği (Kushava vd., 2021: 4)

Şekil 11' de yer alan ağ diyagramı, Kushwaha vd. (2021) tarafından hazırlanan ve günümüzde büyük verileri ışığında yeni yönetim yaklaşımlarının incelendiği bir literatür çalışmasının sonuçlarının görselleştirilmesi kullanılmıştır.

2.2.6. Yorumlama ve Değerlendirme

Elde edilen sonuçların yorumlanması değerlendirilmesi bu aşamada gerçekleştirilmektedir. Burada önemli olan noktalardan bir tanesi elde edilen bulguların sağlıklı değerlendirilmesi için uzman görüşlerinden yararlanmak ve gerekirse daha farklı analiz yöntemleri ile kıyaslama yapmaktır.

2.3. Metin Madenciliği Yöntemleri

Metin madenciliği günümüzde pek çok alanda kullanımı olan, metinleri veri kaynağı olarak ele alıp gerektiğinde klasik veri madenciliği yöntemleri gerektiğinde ise kendine has yöntemlerin kullanıldığı bir disiplindir. Enformasyon getirimi, metinler arası benzerlik, konu özetleme, duygu analizi ve konu modelleme gibi uygulama alanları bulunmaktadır. Bu bölümde metin madenciliği ile ilgili çeşitli kullanım alanları açıklanacaktır.

2.3.1. Bilgi getirimi

Bilgi getiriminin (Information Retrieval) amacı, belge koleksiyonunda bulunan ilgili olabilecek belgelere erişmek, ilgisi olmayan belgeleri ise elemektir. Bilgi erişiminde, anahtar sözcükler vasıtasıyla metin verileri içerisinden istenilen özellikteki metinlere ulaşılması amaçlanmaktadır. Google gibi arama motorları bu tip uygulamaların en tipik örneklerindendir. Bilgi erişim sistemlerinin performansı duyarlılık veya hassasiyet olarak adlandırılan (precision), duyarlılık (recall) ve bu iki ölçüt kullanılarak oluşturulan F-skoru (F_1) ölçütleri kullanılarak değerlendirilmektedir (Yıldız ve Ağdeniz, 2018: 293; Oğuzlar, 2011: 16). Kısacası, metin madenciliği uygulaması eğer web üzerindeki veri kaynaklarında yapılacaksa; web sayfaları, adresleri veya dosya sistemi üzerinde dosyaların tarihleri, kullanıcı bilgileri, dosya isimleri, dizin bilgileri gibi bilgilerin toplanması ile ilgili işlemlerdir (https://ybsansiklopedi.com/, 2023).

2.3.2. Metinler Arası Benzerlik ve İntihal Tespiti

Yapılandırılmamış veri olarak nitelendirilen metin belgeleri arasındaki benzerliklerin tespiti çözümünde özel bazı tekniklerin kullanılmasını gerektiren bir problemdir. Bu benzerliklerin hesaplanması, intihal tespiti veya yazar tanıma gibi uygulamaların temelini oluşturur. Bunun için çeşitli metin eşleştirme algoritmalarından yararlanılmaktadır. Metin eşleştirme algoritmaları tam metin eşleştirme algoritmaları ve yaklaşık metin eşleştirme algoritmaları şeklinde iki grupta incelenmektedir. Tam metin eşleştirme algoritmaları iki metin verisinin birebir eşleşip eşleşmediğini belirleyen basit karşılaştırma yöntemleridir. Örnek olarak Boyer Moore, Karp-Rabin, Morris Pratt, Quick Search, Brute Force, Shift Or, Apostolico-Giancarlo, Turbo-BM, ve Skip Search gibi çeşitli algoritmalar verilebilir. Yaklaşık metin eşleştirme algoritmaları ise "belirli bir metin içinde yaklaşık alt metin eşleştirmelerini tespit etme" ve "yaklaşık olarak örüntüyle eşleşen sözlük metinlerini tespit etme" şeklinde iki temel işlem üzerinde yoğunlaşmıştır (Kaya Keleş ve Özel, 2017: 316).

2.3.3. Konu Özetleme ve Kelime Çıkarımı

Yazarların eserlerinin özetlenmesi ve anahtar kelime çıkarımı da metin madenciliği uygulamalarındandır. Bu yöntemler sayesinde yazara ait metni özetlemek suretiyle içerisindeki örüntüleri tespit edip çeşitli çıkarımlarda bulunmaya imkân sağlar (Abdalla ve Angın, 2022: 698).

2.3.4. Kavram Bağlanımı Tespiti

Farklı metinlerde kullanılan bazı kavramların analize tabi tuttuğumuz bir metin içerisinde tespit edilmesi kavram bağlanımı (concept of linkage) olarak adlandırılmaktadır. Bu yapıyla birden fazla belge arasında fark edebildiğimiz kavramsal ilişkilerin yansıra dışarıdan fark edemediğimiz potansiyel ilişkiler de tespit edilebilmektedir (Atan, 2018: 228).

2.3.5. Duygu Analizi

Doğal dil işlemenin araştırma alanında olan duygu analizi (sentiment analysis) insanların tutum ve değerlendirmelerini bir metin üzerinden analiz eden bir çalışma alanıdır. Veri madenciliği, web madenciliği ve metin madenciliği uygulamalarına sıklıkla konu olan duygu analizi işletmeler ve toplum açısından son derece önemlidir. Bu açıdan yönetim, pazarlama, sosyoloji gibi sosyal bilimlerin çoğu alanında uygulama alanı bulmaktadır. Duygu analizi, kullanıcı veya müşteri yorumları, bloglar, sosyal ağlardaki metinler, forum tartışmaları gibi mecralardan elde edilen metinlerin analizi için son derece kullanışlı bir yöntemdir. Duygu analizi sayesinde metin üzerinde olumlu düşüncelerle olumsuz düşünceleri birbirinden ayırma imkânımız vardır. Bu sayede fikir odaklı bilgi çıkarımı ve soru cevaplama sistemlerinin sağlıklı çalışması gibi avantajlar elde edilebilinmektedir (Kılıç ve diğ., 2020: 132).

2.3.6. Konu Modelleme

Konu modelleme, bilgisayar ve web teknolojilerindeki gelişmeler nedeniyle üretilen büyük hacimli verileri düşük boyutta sunmak ve uygulama içeriğine bağlı olarak çok büyük boyutta veriler içerisinde yer alan gizli kavramlarını, öne çıkan özelliklerini veya gizli değişkenlerini sunmak son derece kullanışlı bir metodoloji olarak düşünülebilir (Kherwa ve Bansal, 2019: 3).

Metin madenciliği yöntemleri kullanılarak metinlerin hangi konularla ilişkili olduğu belirlenebilmektedir. Bunun için kullanılan konu modelleme algoritmaları makine öğrenmesi de kullanarak yapılandırılmamış metinlerin otomatik olarak etiketlenmesini başarabilirler. Bunun için gözetimli veya gözetimsiz algoritmalar kullanılabilir. Gözetimli modellerde daha önceden etiketler belirlenip yeni gelen konunun hangi etiket altında olacağı tahmin edilmeye çalışılabilir. Gözetimsiz modellerde ise bir metin kümesi içinde olası konu başlıkları tespit edilmeye çalışılır. Çeşitli konu modelleme yöntemleri

olmakla birlikte en sık kullanılan yöntem Gizli Dirichlet Ayrımı (Latent Dirichlet Allocation)' dır (Atan, 2020: 227).

Deerwester vd. (1990), konu modelleme yöntemlerinin ilki olan Latent Semantic Analysis (LSA) modelini ilk olarak oluşturmuştur. Bu yöntem elimizdeki belge ya da terim matrisini, belge-konu/konu-terim matrislerine ayrıştırmak üzerine kuruludur. Akabinde Hoffman (1999) Probabilistic Latent Semantic Analysis (pLSA) şeklinde olasılıksal bir model önermiştir. Bu modelde amaç belge-terim matrisinde gözlemlenen verileri temsil eden gizli başlıkları olasılıksal bir yapı ile ortaya çıkarmaktır. Nihayetinde Blei vd. (2003), pLSA' nın daha geliştirilmiş bir versiyonu olan Latent Dirichlet Allocation (LDA) yöntemini önermişlerdir (Şahin, 2021: 53).

2.3.6.1. Gizli Dirichlet Ayrımı

Konu modelleme (topic modelling), büyük ölçekli belgelerin oluşturduğu yapılandırılmamış veri yapılarından anlamlı bilgiler çıkarmaya yarayan denetimsiz bir makine öğrenmesi tekniğidir. Literatürde çok sayıda konu modelleme tekniği bulunmakla beraber Gizli Dirichlet Ayrımı (GDA) en yaygın ve kullanışlı yöntemdir. Ayrık veri türü olarak nitelendirilen belgelerin modellenmesi için geliştirilmiş olan GDA belgeleri oluşturan örtük konuları ortaya çıkarabilir. GDA esas olarak olasılıklara dayalı bir yöntemdir. Buna göre konular sabit bir külliyat üzerinde bir olasılık dağılımına sahiptir ve belgelerin bu örtük konuların rastgele bir bileşiminden oluşmaktadır. Dolayısıyla GDA belge koleksiyonunda yer alan konuları, konuları oluşturan kelimelerin olasılıklarını, hangi kelimelerin hangi konulara atandığını ve her belge için konuların dağılımını ortaya çıkarmaktadır (Ekinci vd., 2019: 68; Blei vd., 2003: 993; Agrawal vd.,2018: 74).

GDA yöntemi son dönemde, doğal dil işleme, duygu analizi, bilgi çıkarımı, sosyal araştırma ve eğilimler analitiği gibi farklı kullanım alanları olan metin madenciliği disiplininde sıklıkla kullanılmaktadır. Yine iş ilanlarının analizinin yapıldığı çalışmalarda da uygulama örneklerini görmek mümkündür. Önceleri sadece metinsel verilerin analizi için kullanılsa da genetik veriler, resimler, videolar ve sosyal ağlar farklı türde verilerde de uygulanabilmektedir (Gürcan ve Çağıltay, 2019: 82543).

GDA yönteminde analize konu olan belgeler kümesi derlem (corpus), külliyat içindeki her bir öğe belge (document), belge içindeki her bir kelime ise terim (term) olarak adlandırılır. GDA yönteminde amaç, belgelerin oluşturduğu külliyatı kapsamlı bir şekilde

temsil edecek konular (topics) şeklinde ayrıştırmak ve bu şekilde örtük içeriklere ulaşmaktır. GDA yönteminde kelime-konu ve belge-konu şeklinde iki farklı matristen yararlanılmaktadır. Kelime-konu matrisinin boyutlar K (konu sayısı) ve V (külliyattaki kelimeler), belge-konu matrisinin boyutları ise K (konu sayısı) ve D (külliyattaki belgeler) şeklindedir. GDA yönteminde konular temsil eden kelimler olasılıklara göre değerlendirilir ve her bir kelime birden fazla konu içerisinde bulunabilir. Ayrıca bir konu içerisinde en yüksek olasılık değerine sahip kelimeler o konu hakkında fikir verebilirler (Çallı vd., 2021: 2362).

Bir belge koleksiyonundaki konu dağılımını modellemek için GDA yöntemi süreç aşağıdaki gibidir (Gürcan, 2017: 58):

1. Öncelikle konu sayısı belirlenir. Belirleme işlemi daha önceki çalışmalar örnek alınarak tahmin etme ya da deneme-yanılma yoluyla yapılabilir. Farklı tahminler için istatistiksel olarak en yüksek anlamlılık düzeyi ya da kesinlik vaat eden konu sayısı seçilebilir.
2. Yöntem her bir terimi geçici bir konuya atama yapar. Konu adımları 3.adımda tekrardan güncelleneceği için geçicidir. Geçici konular ve terimler rastgele şekilde atanır. Her bir terim farklı konulara atanabilir.
3. Bu adımda "tekrarlı yaklaşım" söz konusudur. GDA her belgedeki her bir kelime için konuları günceller. Güncelleme iki duruma göre yapılır; "konularda bu kelimeler ne kadar yaygın?" ve "Belgede bu konular ne kadar yaygın?".

Şekil 12' de GDA yönteminin aşamalarını anlatan grafik gösterimi yer almaktadır. Burada düğümler, rastgele değişkenleri temsil etmektedir. Düğümler arasındaki ilişikler de bağlantılar kullanılarak gösterilmektedir.

Şekilde 12' de verilen grafik gösteriminde;

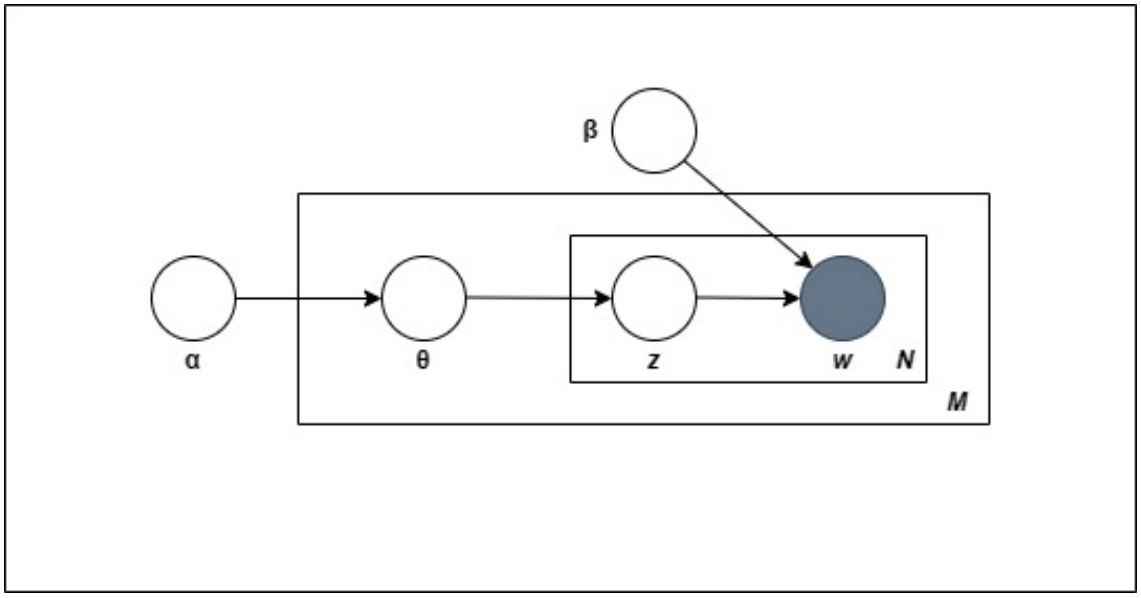
- α, her bir doküman için konu dağılımını gösterir.
- β, her bir konu başına kelime dağılımını gösterir.
- θ, belirli bir doküman için konu dağılımını gösterir.
- z, her bir kelime için atanan konuları belirtir.
- w, gözlemlenen kelimeleri belirtir.

GDA modeli için gözlemlenebilir veya gizli değişkenlerin dağılımı Denklem 12' de verilmiştir.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, W_{1:D}) =$$

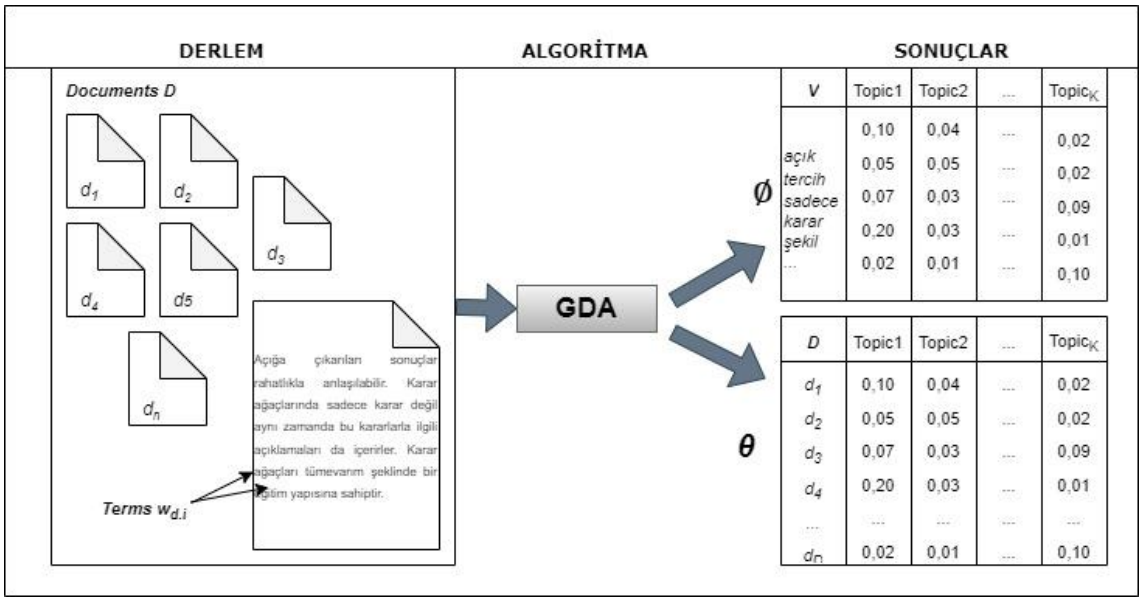
$$\prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d)(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:k}, z_{d,n})) \quad (12)$$

Denklemde, K, konu sayısını, D belge sayısını, N kelime sayısını, $\beta_{1:K}$ kelimenin konu içindeki oranı, $\theta_{1:D}$ belge içinde konuların oranı, $z_{1:D}$ konuların belgelere atanma oranını ve $w_{1:D}$ ise belgelerde gözlemlenen kelimeleri ifade etmektedir.



Şekil 12. GDA Modeli Grafik Gösterimi (Blei vd., 2003: 997)

Şekil 12' de yer grafik gösteriminde kutular kopyaları temsil eden "plakalardır". Dış plaka belgeleri temsil ederken, iç plaka bir belge içindeki konuların ve sözcüklerin tekrarlanan seçimini temsil eder. Şekil 12' de gösterilen yapıda α ve β değerleri işlemler sırasında bir kez örneklenmektedir. Belge seviyesinde bulunan konu değişkenleri ise işlemlerdeki her bir metin belgesi için örneklenirken, kelime seviyesindeki değişkenler ise belgelerle bulunan tüm kelimeler için tekrardan örneklenmektedir (Blei vd., 2003: 997). α değerinin düşük ya da yüksek olması belgelerin içerdiği konu sayısı ile ilgilidir. α değeri yüksek ise, belgeden çıkarılacak konu sayısı fazla; düşük ise konu sayısı az olmalıdır. α değerinin yüksek olması durumunda veri kümesinde çok sayıda kelime bulunarak işlemler gerçekleştirilir. α değerinin düşük olduğunda ise veri kümesinde az sayıda kelime dikkate alınarak işlem yapılır. Şekil 13' te örnek bir GDA uygulaması verilmiştir.



Şekil 13. GDA Uygulaması (Maier vd., 2018: 94)

Şekil 13' te verilen GDA uygulamasından görüldüğü üzere GDA algoritması çeşitli belgelerden oluşan bir derlemi, belge konu matrisi ve konu kelime matrisi oluşturmak suretiyle modeli oluşturmaktadır. Bu şekilde kelimeler ve ağırlıkları kullanılarak kullanışlı bilgiler ortaya çıkarılmakta ve bu kapsamda konular tespit edilebilmektedir.

ÜÇÜNCÜ BÖLÜM

BİLGİ VE İLETİŞİM SEKTÖRÜNDEKİ ÇEVRİMİÇİ İŞ İLANLARININ VERİ VE METİN MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

Günümüzde iş dünyasında kaynakların doğru kullanımı adına doğru kişilerin doğru işlere yerleşmesi işletmeler için son derece önemlidir. En yüksek maliyet kalemini oluşturan iş gücü maliyetlerinin karşılığı olarak yüksek verimlilik beklentisi işletmelerin başlıca gündemlerinden bir tanesidir. İnsan kaynakları planlamasının işletmeler adına doğru bir şekilde yapılması verimlilik konusunda işletmeleri rekabette avantajlı konuma getirebilir. Bu doğrultuda doğru sayıda ve uygun niteliklere sahip kişilerin işletmelere kazandırılması ve akabinde bu kişilerin kariyer ile ilgili beklentilerinin de karşılanarak uzun vadede işletmeye katkı sağlamaları önemli bir konudur.

İş görenler açısından bakıldığında ise durum benzer özellikler taşımaktadır. İş görenlerin kariyer planlarını yaparlarken sektörel beklentileri inceleyip becerilerini bu yönde geliştirmeleri doğru işe doğru kişilerin yerleşmesi için önemlidir. Ayrıca iş görenlerin çalıştıkları pozisyonlar için gerekli yetkinlikleri taşımaları, yaşadıkları iş stresini de minimum seviyeye indirecektir. Bu süreçte iş gören adayları kariyer planlarını yaparlarken öncelikle kendi kişilik özelliklerini doğru tespit edip yeteneklerini keşfetmeye çalışmalı, hedefledikleri sektörlerin ve iş alanlarının beklentileri doğrultusunda becerilerini geliştirmelidirler. İş gören adaylarının bunu mümkün kılmasının yolu, eğitim hayatı ve devamında profesyonel iş yaşamı boyunca iş dünyasının beklentilerinin takip etmekten geçmektedir. Dolayısıyla iş dünyası ve iş gören adaylarının birbirlerinin beklentilerini anlamak adına belirli alanlarda buluşması son derece önemlidir. Bunun için çeşitli organizasyonlar olmakla birlikte iş ilanları da beklentilerin takip edilmesi adına son derece önemlidir.

İşverenlerin boş pozisyonları doldurmak ve aday havuzu oluşturmak için iş gören adayları ile buluşabilecekleri çeşitli yollar vardır. İŞKUR vasıtasıyla, Yetenek Kapısı Platformu, kariyer sitelerinde yer alan çevrim içi iş ilanları, gazete veya dergiler gibi kitle iletişim araçları, sosyal ağlar ve işletmelerin kendi web siteleri iş ilanlarının yayınlandığı alanlardır.

Bu bölümde Kariyer.net web sitesinde yayınlanan iş ilanlarının Veri Madenciliği ve Metin Madenciliği yöntemleri ile analiz edilmesi ilgili süreçler, yöntemler ve nihayetinde bulgulara yer verilmiştir. Öncelikle verilerin elde edilmesi ön işleme ile ilgili süreçler ele alınmıştır. Ardından veri ve metin madenciliği ile verilerin analizi ve bulgular yer almaktadır.

Türkiye'de iş ve iş bulma süreçleri için kullanılan platformlardan bir tanesi de iş ilanlarının paylaşıldığı internet siteleridir. Bu sitelerden bazıları aşağıda incelenmiştir;

Kariyer.net: 1999' dan itibaren hizmet veren Kariyer.net, iş arama ve işe alım süreçlerinde yeni nesil teknolojilerle, iş arayanlarla işverenleri internet ortamında bir araya getiren bir insan kaynakları sitesidir. Kariyer.net' in aday veri tabanında 25 milyonu aşkın özgeçmiş bulunmaktadır. Kariyer.net üyesi 94 binin üzerinde şirket, ihtiyaç duyduğu iş gücünü Kariyer.net aracılığıyla aramaktadır. Kariyer.net birçok farklı ürün ile üyelerine hizmet vermektedir. Bunlar; "Kariyer.net İlan Paketi", "Pratik Aday Servisi", "Özgeçmiş Havuzu", "Markan Ana Sayfada", "İmaj Paketi", "Select Paket", "Mavi Yaka İşe Alım Çözümü", "Memnuniyet Analizi" ve "10. Kat" şeklindedir.

Yenibiris.com: Yenibiris.com, büyük ölçekli ve çokuluslu şirketlerden orta ve küçük ölçekli şirketlere kadar her boyutta, farklı sektörlerden 280 bini aşkın firmaya internet bazlı seçme ve değerlendirme hizmeti sunan bir internet çevrim içi iş ilanı sitesidir. Yenibiris.com' un özgeçmiş bankasında yaklaşık 19 milyon özgeçmiş bulunmaktadır. Yenibiris.com sunduğu içerik ile bireysel üyelerinin kariyer ve kişisel gelişimlerine rehberlik etmeyi misyon olarak benimsemiştir. Teknolojiyi yoğun bir şekilde kullanarak bireysel ve kurumsal üyelerinin iş ve aday arama süreçlerini hızlandırıp kolaylaştırmaya yönelik çözümler üretir.

Secretcv.com: Secretcv.com, 2000 yılında faaliyetlerine başlayan, iş arayan kimseler ile eleman arayan firmaların aynı ortamda buluşmalarını amaçlayan bir insan kaynakları sitesidir. Secretcv' nin verilerine göre 2017 yılı itibariyle 22.000.000 aday özgeçmişi ve yaklaşık 65.000 üye firma Secretcv veri tabanına kayıtlıdır. Diğer insan kaynakları sitelerinde olduğu gibi burada da iş arayanlar herhangi bir ücret ödemezler. Secretcv.com doğru işe doğru insan felsefesinden yola çıkarak firmalar ve iş arayanları en doğru şekliyle buluşturmayı hedeflemektedir. Özgeçmişlerin firmalar tarafından incelenmesi iki şekilde sağlanmaktadır. Birincisi, üye firmalar yayınladıkları ilanlara

bireysel kullanıcıların başvuru yapmaları ve başvuruların ilgili firma tarafından incelenmesi şeklindedir. İkincisi ise üye firmaların aday veri tabanı içerisinde belirledikleri kriterlerle arama yapmaları sonucu çıkan özgeçmişleri incelemeleridir (https://www.secretcv.com/, 04.06.2023).

Ayrıca verilerin ön işlenmesi sırasında sektör etiketleri oluşturulurken TUİK Sınıflama Sunucusu kullanılmış ve sektörler bu sunucuda yer alan NACE rev.2- Altılı Ekonomik Faaliyet Sınıflaması kodlarına göre etiketlenmiştir. NACE rev.2- Altılı Ekonomik Faaliyet Sınıflaması beş düzeyden oluşmaktadır. Yirmi iki faaliyet grubunun yer aldığı birinci sınıflama düzeyi bu çalışmada sektörler etiketlenirken kullanılmıştır.

3.1. Çevrimiçi İş İlanları ile İlgili Yapılan Çalışmalar

Bu bölümde çevrim içi iş ilanları ile ilgili yapılan literatür çalışmasından örnekler yer almaktadır.

Koong vd. (2002), bilgi işlem alanında talep edilen ve sürekli değişen yeteneklerin keşfi için Monster.com ve Hotjobs.com sitelerinde yer alan ilanları incelemek suretiyle veri tabanı, programlama dili, web tasarımı gibi alanlarda aranan nitelikleri tespit etmeye çalışmışlardır.

Webb (2006), yönetim bilişim sistemleri (YBS) müfredatlarını iş gücü piyasasındaki taleplere uyumlu hale getirme çabası doğrultusunda yaptığı çalışmada bir yıllık Monster.com verisinden yararlanılmıştır. Burada özellikle lisans derecesi gerektiren işlerin ön plana çıktığı görülmüştür. Ayrıca sözlü ve yazılı iletişim becerileri de aranmaktadır. YBS öğrencileri için yapılan beceri listeleri sonucunda “veri tabanı”, “ağ”, “sistem analizi” ve “programlama” başlıklarından oluşan dört farklı kariyer grubu oluşturuldu. Ayrıca ortaya çıkan veriler doğrultusunda; “analiz”, “ağ oluşturma” ve “programlama” gibi üç beceri kümesi de oluşturulmuştur.

Huang vd. (2009), Nisan 2008 ve Haziran 2008 tarihleri arasında Monster.com sitesinde yayınlanan ve bilgi işlem alanı ile ilgili 241 çevrimiçi iş ilanını incelemişlerdir. Ayrıca bilimsel makaleler, uygulayıcı literatürü ve çevrim içi iş ilanları doğrultusunda aranan nitelikleri “teknik, insani ve iş becerileri” şeklinde sınıflandırmışlardır. Bu şekilde her üç kategori için bilimsel makaleler, uygulayıcı literatürü ve çevrim içi iş ilanlarında en popüler yeteneklerin tespitine çalışmışlardır.

Liecky vd. (2009), geliştirdikleri yazılım ile Temmuz 2007 ve Nisan 2008 tarihleri arasında 3 farklı siteden yaklaşık 210000 iş ilanı elde ettiler. Daha sonra kümeleme analizi kullanarak % 91 başarı oranı ile yaptıkları sınıflandırma ile 20 farklı iş tanımı ortaya koydular. Bu verileri girdi olarak kullanarak her bir iş tanımında en sık bahsedilen yetenekleri belirlemeye çalıştılar.

Kanık vd. (2012), 2005 – 2012 yılları arasını kapsayan Kariyer.net ve İŞKUR verileri ile yaptıkları analizlerde iş bulma oranı ile emek piyasası arasında yoğun bir ilişki tespit etmişlerdir. Özellikle kriz öncesi ve sonrasında görülen farklılıklar ve işsiz havuzlarının niteliklerinin de farklılaşması çalışmanın diğer bulguları olarak görülebilir. Bu tarz çalışmalar sayesinde Türkiye özelinde istihdamla ilgili önemli ipuçları bulmak mümkündür.

Smith ve Ali (2014), web programlama kurslarında verilen programlama dillerinin pazardaki talebi ne şekilde karşıladıklarının tespit edilmesini amaçlamışlardır. Bunun için web programlama kurslarında verilen müfredatlar incelenmiş ve karşılaştırma için Pazar talebi ve bu talebi oluşturan faktörler değerlendirilmiştir. Elde edilen bulgular doğrultusunda müfredat oluşturma adına tavsiyeler sunulmuştur.

Motlogelwa vd. (2014) yaptıkları içerik analizinde Şubat 2008 ve Aralık 2008 tarihleri arasında Botswana' da günlük ve haftalık yayınlanan gazete ve dergilerde yer alan iş ilanlarını incelemişler ve sektör tarafından talep edilen iş yeteneklerini tespit etmeye çalışmışlardır. Araştırmanın bulguları doğrultusunda Botswana Üniversitesi Bilgisayar Bilimleri Bölümü müfredatının düzenlenmesine veri sağlanması amaçlanmıştır.

Koçer ve Öksüz (2015), elektronik işe alma süreçlerini ele alan çalışmalarında Kariyer.net ve Adecco' yu incelemişlerdir. Çalışmada özel istihdam büroları ile yapılan istihdam süreçlerinin detaylı olarak incelenmesi amaçlanmıştır. Bu süreçte Türkiye' de faaliyetlerini sürdüren Adecco Türkiye ve Kariyer.net yöneticileri ile görüşülmüştür. Çalışma nitel bir analize dayalı olup yüz yüze görüşmelerden elde edilen veriler analize tabi tutulup bulgular işe alım süreçlerinde özel istihdam bürolarının önemli bir rol oynadığı yönündedir. Bu sayede özellikle özel sektör firmaları neredeyse tüm iş gören ihtiyaçlarını özel istihdam büroları aracılığı ile karşılamaktadır.

Kureková vd. (2015), iş gücü piyasasını analiz etmek için açık pozisyon verilerinin ve web tabanlı gönüllülük esasına dayalı anketlerden elde edilen verilen kullanılmasını karşılaştırmışlardır. Bu şekilde metodolojik sorunları ortaya koymuş ve her iki tür veri kullanımı için avantaj ve dezavantajları belirlemişlerdir. Bununla birlikte ortaya çıkan metodolojik sorunlar için bazı stratejiler önerilmiştir. Sonuç olarak internet kullanımının yaygınlaşması ve bununla birlikte veri toplama yöntemlerinin de gelişmesi ile çevrim içi verilerin kullanımının anketlere göre daha güvenilir sonuçlar verdiği görülmektedir.

Okay (2016) çalışmasında; beş mühendislik bölümünü (bilgisayar, elektrik, endüstri, makine, inşaat) içeren 17347 iş ilanını veri madenciliği teknikleri incelemiştir. Türkiye' de bu bölümlerden mezun olacak öğrenci sayıları, üniversitelerin kontenjanları kullanılarak yaklaşık olarak hesaplanmış ve bu bölümlerin iş ilanlarındaki ilişkileri çeşitli analiz yöntemleri kullanılarak incelenmiştir. Çalışma ders müfredatlarının belirlenmesi ve ilgili mühendislik alanlarında ne gibi teknik becerilerin ön plana çıktığının tespiti için önemli bulgular sunmaktadır.

Gaediner vd. (2018), büyük veri ve veri analizi gibi işler için endüstri tarafından önemli beceri ve yetkinlikleri daha iyi anlamak için büyü veri içeren 1216 iş ilanının analizini yapmışlardır. Çalışmada, yeni ürün ve hizmetler geliştirme de dahil olmak üzere stratejik girişimleri desteklemek için büyük verinin potansiyelini keşfetmenin önemi bunun için veri bilimcilerin istihdam edilmesi ön plana çıkarılan konulardan biridir. Sonuç olarak, büyük veri becerileri kavramsal bir yapıda ortaya konulmuştur. Ayrıca birçok büyük veri iş ilanının gelişen analitik bilgi sistemlerini vurguladığını ve ortaya çıkan zor teknik becerilere verilen öneme ek olarak sosyal becerilerin de yüksek önem taşıdığını ortaya konulmuştur.

Ayaz vd. (2018), denizcilik sektörüne yönelik yaptıkları çalışmada Kariyer.net verilerinden yararlanmışlardır. Denizcilik sektörüne ait 160 ilan "pozisyon", "yabancı dil", "tecrübe" gibi çeşitli başlıklar özelinde içerik analizine tabi tutulmuştur. Sonuç olarak lojistik ve gümrük komisyon işletmelerinin en çok ilan veren işletmeler olduğu, bu ilanların büyük bir çoğunluğunun İstanbul ilinde yer aldığı, denizcilik fakültesi mezunlarının daha çok tercih edildiği gibi bulgulara ulaşılmıştır. Ayrıca departman bazında operasyon elemanı ve uzman pozisyonlarının daha fazla ilanda yer aldığı ve yabancı dil becerisinin de ilanlarının neredeyse tamamında istendiği sonuçlarına

ulaşılmıştır. Bununla birlikte adaylara yemek, ulaşım ve özel sağlık sigortası gibi olanaklarında çoğu firma tarafından adaylara sağlandığı görülmektedir. İşletmeler adaylarda kişisel beceri olarak analitik düşünme, takım çalışmasına yatkınlık ve iletişim yeteneklerini de öncelikli olarak aramaktadırlar.

Güler (2018), hazırladığı yüksek lisans tezinde aile işletmeleri ile kurumsal işletmelerin işe işe alım süreçlerinin farklılıklarını ortaya koymaya çalışmıştır. Bu amaçla Kariyer.net' te yayınlanan "yeni mezun" etiketli iş ilanlarını incelemiş ve bir nitel araştırmaya tabi tutmuştur. Araştırma sonuçlarına göre Office programları kullanımı, yabancı dil, ehliyet ve deneyim ön plana çıkan becerilerdir. İşletmeler bu beceriler yan sıra teşvikler, örgüt kültürüne uyum, öğrenmeye açıklık gibi kriterlere göre işe alımlarını gerçekleştirmektedirler. Kurumsal işletmeler eğitime daha çok önem vermektedirler ve çoğunun bünyesinde eğitim programları bulunmaktadırlar. Ayrıca hem aile işletmeleri hem kurumsal işletmeler staj imkanları sunup staj sonunda yeterli gördüğü yeni mezunları istihdam edebilmektedirler.

Olszak ve Lorek (2019), çağdaş iş dünyasında büyük veri kullanımı, büyük verinin altında yatan temel kavramlar ve büyük veri analizi için mevcut yöntem ve araçları inceledikleri çalışmalarında kariyer sitelerinden elde edilen verileri analiz etmişlerdir. Sonuç olarak, iş portallarını yönetmek için akademisyen ve insan kaynakları uzmanlarına yönelik uygun yöntemler ve araçlar önermişlerdir. Aynı zamanda bu çalışma, işgücü piyasasının gereklilikleri ve ihtiyaçları hakkında daha iyi bilgi sahibi olmak isteyen çalışanlar ve daha iyi yönlendirilmiş iş gören adaylarını işe almak isteyen işverenler için bir değerli bilgiler de içermektedir.

Belov vd. (2019), Rusya iş gücü piyasasının analizi için metin analizine yönelik bir yaklaşım önermişlerdir. Bu amaçla, işgücü piyasası için büyük veri teknolojilerine dayalı analitik bir sistemin yapısal şeması oluşturulmuştur. Özel metinlerden nesneler ve aralarındaki bağlantılar (örneğin iş ilanlarından) hakkında anlamsal bilgilerin çıkarılması amaçlanmıştır. Sonuç olarak Rusya işgücü piyasasının izlenmesine yönelik bir sistem oluşturulmuş olup, diğer ülkelerin de analize dahil edilmesi için çalışmalar devam etmektedir. Dikkate alınan yaklaşımlar ve yöntemler, büyük miktarda metinden bilgi çıkarmak için yaygın olarak kullanılabilir.

Matsuda vd. (2019)' da yaptıkları çalışmada Pakistan' ın en popüler iş portalı Rozee.pk' dan alınan verilerle Pakistan'daki işgücü piyasası koşulları ve beceri talebi-arzı hakkında yeni tanımlayıcılar elde etmeyi amaçlamışlardır. Çalışmada, Pakistan ekonomisinin sürekli artan genç nüfus karşısında oluşan işgücü için daha fazla iş yaratması gerektiği ve eğitim öğretim sisteminin özel sektördeki beklentileri karşılayamadığı için eğitimli genç iş gücünün işsizlik problemi yaşadığı endişe verici bir durum ele alınmıştır. Çalışma sonucunda, yüksek eğitimli iş gören arzında fazlalık olmasına rağmen, bilgi ve iletişim teknolojisi gibi bazı endüstrilerde özel becerilere ve deneyime sahip işçilerden yoksun olduğunu ortaya koyulmuştur. Çalışmada aynı zamanda niteliklerin ve becerilerin tam olarak eşleşmesinin işverenler için önemli olduğu tespit edilmiştir. Dolayısıyla iş ilanlarında belirtilen niteliklere sahip olmayan veya gereğinden fazla niteliklere sahip olan iş gören adaylarının, nitelikleri iş gerekleriyle tam olarak örtüşen kişilere göre istihdam edilme olasılığı daha düşük olduğu sonucuna ulaşılmıştır.

Dakhill (2019), müfredat geliştirmek adına sayısal bir metodoloji ortaya koymak amacıyla yazdığı tezde altı-sigma kalite yaklaşımından yararlanmıştır. Bu amaçla İmalat Mühendisliği Bölümü seçilmiş ve uygulama ile bu bölümün müfredatı geliştirilmeye çalışılmıştır. Seçilen farklı imalat mühendisliği bölümleri müfredatlarının içeriklerinden oluşturulan bilgi beceriler ve yetenekler seti üzerinde öğrenciler, öğretim üyeleri ve endüstri temsilcileri gibi farklı paydaşların görüşlerine başvuru yapılacak bir anket geliştirilmiştir. Bu anketin sonuçları analiz edilerek çeşitli önerilerde bulunulmuş ve bu öneriler Kariyer.net' te yer alan ilgili bölüme ait ilanlar incelenmek suretiyle geçerlilikleri tespit edilmiştir.

Vankevich ve Kalinouskaya (2021)' de işgücü piyasası hakkındaki bilgileri önemli ölçüde zenginleştiren ve ekleyen ve etkili karar vermeyi kolaylaştıran ayrıca beceri ve yeterlilikler bağlamında işgücü talebinin analizine ve tahminine odaklı bir çalışma yapmışlardır. Çalışmada, büyük veri kullanılarak işgücü piyasasının analizi, kariyer sitelerinde yayınlanan açık pozisyon açıklamasında listelenen beceri ve yeterlilikler bağlamında işgücü talebinin ve işgücü arzının değerlendirilmesi ve işverenlerin ve bireylerin daha iyi istihdam ve eğitim almalarına yardımcı olmak için beceriler ve yeterlilikler arasındaki eşleşmelerin belirlenmesi amaçlanmıştır. Çalışma sonucunda kariyer sitelerindeki bilgilerin işgücü piyasasındaki arz ve talep hakkında

bilgiler içerdiği ve bu bağlamda yetkinliklerin geliştirilmesi için eğitim alanlarının belirlenmesi için bir yola haritası oluşturulmuştur.

Üçler ve Büyükçelikok (2021), yaptıkları çalışmada iletişim fakültesi müfredatlarında yer alan derslerin sektördeki nitelikli iş gücü ihtiyacının karşılanması konusundaki yeterliliğin tespit edilmesi amaçlamıştır. Bu amaçla iletişim fakültesi müfredatları incelenmiş, mezunların sektörel donanım yeterliliği ve istihdam problemleri tespit edilmeye çalışılmıştır. Kariyer.net medya sektörün kategorisinde yer alan çevrimiçi ilanlar içerik analizine tabi tutulmuştur. Yapılan analiz çerçevesinde araştırma müfredatlardaki ders içeriklerinin üniversitelere göre farklılık göstermelerine rağmen teorik derslerin ağırlıkta olduğudur. Sektör beklentileri doğrultusunda bölümler "dijital pazarlama", "sosyal medya pazarlaması", "Seo", "Sem", "Google Adwords", "Photoshop" gibi teknikleri içeren derslerden müfredatlar oluşturmaları istihdam adına daha sağlıklı sonuçlar verecektir.

Lyu ve Liu (2021), yaptıkları çalışmada enerji sektöründeki istihdamda hassas beceriler (sosyal, bilişsel, insan yönetimi, proje yönetimi ve müşteri hizmetleri becerisi) ve teknik beceriler (ürünler ve pazarlama, mühendislik ve genel bilgisayar becerisi) arasındaki kalıpları göstermek adına Burning Glass Technologies (BGT)' ten elde ettiği verileri analiz etmişlerdir. BGT, şirket web siteleri ve çevrimiçi iş ilanlarından bilgi elde eden bir iş analitiği ve iş gücü piyasası bilgi firmasıdır. Elde edilen bulgulara göre hassas becerilerin iş gereksinimlerinde hızlı bir artış gösterdiği ve teknik becerilerin ise bu süreçte sabit kaldığı yönündedir.

Ergüt (2021), çalışmasında ekonometri mezunlarının istihdam olanaklarını ve onları istihdam etmek isteyen firmaların ihtiyaçlarını ortaya koymaya çalışmıştır. Bu doğrultuda iş arayan ve işverenleri bir araya getiren çevrimiçi iş ilanı sitesinde içerik analizi yapmış ve iş ilanlarındaki ihtiyaç duyulan beceri ve özelliklerin ortaya çıkarılması amacıyla metin madenciliği uygulamalarına başvurulmuştur. Yapılan incelemeler sonucunda bankacılık, enerji ve otomotiv sektörünün en fazla olanağı sağlayan sektörler olduğu görülmüştür.

Hutter (2021), işverenlerin ve iş gören adayanlarının iş ve işçi arama faaliyetlerini incelemek için Alman Federal İstihdam Bürosu' nun çevrim içi faaliyetlerden elde ettiği büyük veriyi analiz ederek bir çalışma yapmıştır. Veriler, iş ve işgücü piyasası döngüsü

boyunca arama ve yerleştirme faaliyetlerinin davranışını ve bunların mevsimsel değişimlerini modellemek için kullanılmıştır. Sonuçlar, firmaların ve istihdam bürolarının arama faaliyetlerinin döngüsel olduğunu göstermektedir. Buna karşılık, iş arayanların arama yoğunluğu, en azından COVID-19 krizinden önce, konjonktüre karşı bir model gösterdiği görülmektedir.

Decorte vd. (2021), yapay sinir ağlarından yararlanarak önceden eğitilmiş bir dil modeli oluşturmuş ve bununla iş unvanlarının boş pozisyonlardaki iş yeteneklerini ne şekilde temsil ettiklerini tahmin etmeye çalışmışlardır. Burada amaç hızlı ve hatasız bir biçimde iş unvanının belirlemesiyle birlikte yeteneklerin de ortaya konulmasıdır.

3.2. Çalışmanın Amacı

İşletmelerin başarısına katkı yapan en önemli unsur sahip olduğu insan kaynağıdır. Bu açıdan bakıldığında insan kaynakları yönetiminde en önemli konular iş gören bulma ve seçimi süreçleridir. İşletmenin ihtiyaçlarını karşılayacak niteliklerde kişilerin bulunması ve bunların işletmeye kazandırılması çözülmesi gereken önemli bir problemdir. İş gören bulma; işletmelerin, yeterli miktarda ve uygun niteliklere sahip kişilerin iş başvurularını, ihtiyaç duyulan zamanda ve işletme ihtiyaçları doğrultusunda kendine çekebilme sürecidir şeklinde tanımlanabilir. İş gören seçimi ise; başvuru yapan adaylar arasından, iş gereklerine uygun niteliklere sahip olanların seçilmesi anlamına gelmektedir. Doğru kişilerin işletmeye kazandırılması çalışanlardan üst düzey performans elde etme imkanını artırırken, çalışanlara da iş tatmini ve potansiyellerini ortaya koyma fırsatı sunabilir. İşletmeler işe alım süreçlerinde boş pozisyonları doldururlarken iç kaynaklardan veya dış kaynaklardan yararlanma gibi farklı yaklaşımlardan yararlanabilirler. boş pozisyonların iç kaynaklardan yararlanılarak doldurulması, terfi veya transfer gibi uygulamalar vasıtasıyla yapılabilir. Bu yöntemler iş tatmini açısında da son derece önemlidir. Çünkü çalışanların beklentileri arasında işletme de terfi gibi durumlar da yer almaktadır. İşletmeler eğer iç kaynaklardan boş pozisyonlarını dolduramıyorlarsa dış kaynaklara başvurabilirler. Dış kaynaklardan temin yöntemleri ise iş ilanları, duyurular, bireysel başvurular, İŞKUR ilanları, eğitim kurumları ya da internet üzerinden çalışan kariyer siteleri şeklindedir. İnternet siteleri vasıtasıyla yapılan aday bulma ve seçim faaliyetleri maliyet avantajı, yüksek verimlilik, geniş kitlelere ulaşma adına avantajlar sunmaktadır. Ayrıca iş gören adayları açısından da çok

fazla işletmeye ulaşma ve bunları karşılaştırma gibi avantajlar sağlayabilir. Bundan dolayı kariyer siteleri dünyada ve ülkemizde geniş bir kullanıcı kitlesine sahiptir.

2000'li yılların başlarından itibaren özellikle internet kullanımın artmaya başlamasıyla çevrim içi iş ilanları iş arayanların öncelikle müracaat ettikleri kurumların başında gelmektedir. Çevrim içi iş ilanlarını içeren kariyer siteleri sayesinde çok sayıda, farklı sektörlere ait ve farklı lokasyonlardaki ilanlara erişim mümkündür. İş arayan adaylar farklı eğitim seviyelerine göre bu siteler vasıtasıyla çeşitli firmalar ile etkileşim içerisine girebilmekte ve kariyer planlarına buna göre yön vermeye çalışmaktadırlar. İş görenlerin kariyer planlarını yaparken dikkat etmesi gereken hususlar, geliştirmesi gereken becerilerin tespit edilmesi amaçlanmıştır.

İşverenler açısından bakıldığında ise kariyer siteleri işletmelere çok geniş bir veri tabanı sunmaktadırlar. Çevrim içi iş ilanları işverenlerin iş gören adaylarının bulunma seçilme sürecine katkı yapan önemli araçlardandır. İşverenler aradıkları nitelikteki iş görenlere bu sayede ulaşabilmekte ve kendileri için geniş bir aday havuzu oluşturabilmektedirler.

Çalışmada, çevrimiçi iş ilanlarında yer alan bilgilerin hem işverenler hem de iş görenler için daha verimli kullanılması adına veri ve metin madenciliği yöntemleri ile analiz edilmiştir. Aynı zamanda elde edilen bulgular eğitim kurumları için iş gücü piyasasını doğru analiz etme adına önem arz etmektedir. Eğitim kurumları açısından bakıldığında ise elde edilen bulgular ışığında müfredatların düzenlenmesi ve derslerin planlanması için bazı yönlendirici bilgiler elde edilmesi amaçlanmıştır.

3.3. Çalışmanın Kapsamı

Çalışma kapsamında Türkiye' de faaliyet gösteren bir kariyer sitesinin 2022 Mart – 2023 Nisan tarih aralığında yayınladığı iş ilanları incelenmiştir. Metin madenciliği ile ilgili analizler için 18 sektörün arasından "Bilgi ve İletişim Sektörü" seçilmiştir. "Bilgi ve İletişim Sektörü", veri seti içerisinde "İmalat" ve "Toptan ve Perakende Ticaret" sektörlerinin ardından en çok ilan sayısına sahip üçüncü sektör durumundadır. Gelişen teknoloji ve sektördeki artan nitelikli eleman açığının artması ayrıca ilanlarda teknik becerilerin daha ön planda olması gibi nedenler bu sektörün analiz için seçilmesinde önemli etkenlerden bir kaçıdır.

Teknolojik değişimlerin tüm iş dünyasını etkisi altına aldığı günümüzde en çok etkilenen sektörlerin başında Bilgi ve İletişim sektörü gelmektedir. Bununla birlikte ARGE yatırımlarının ve teknolojik yeniliklerin en fazla görüldüğü sektör konumdadır. Yapay zekâ teknolojileri ile ilgili mühendislik alanları, bulut depolama teknolojileri, mobil yazılım uygulamaları, web tasarım faaliyetleri sektör içerisinde ön plana çıkmaktadır. Bütün bunlar sektörün insan kaynağı planlamasının da hızlı bir değişim göstermesine sebep olmaktadır. Bilgi ve iletişim sektörü istihdamı oransal olarak en çok artış gösteren sektördür. Sektörün, net istihdam değişim oranı yüzde 6,1 ile Türkiye Genelinin yaklaşık iki katına yakındır. Kişi sayısı olarak bakıldığında 15 bin 544 kişi ile bilgi ve iletişim sektörü alt sıralarda yer aldığı görülmektedir. Fakat nicelik olarak sektördeki net istihdam az olsa bile ilgili sektörün nitelik açısından Türkiye ekonomisi için önemi göz ardı edilemez bir gerçektir. (İŞKUR, 2021: 38).

Dolayısıyla bilgi ve iletişim sektörü, özellikle yüksek nitelikte personel ihtiyacı olan ve sektörel beklentilerin teknolojik değişimlerden yoğun olarak etkilenen bir yapıdadır. Bundan dolayı çalışmada bilgi ve iletişim sektörünün analiz edilmesi uygun görülmüştür.

3.4. Çalışmanın Kısıtları

Çalışma kullanılan veriler elde edilirken iş ilanlarının tamamına ulaşılmak istenmiş ve bu doğrulta bir yazılım kullanılmıştır. Fakat web sitesinin veri yapısından ötürü sitede beyan edilen kadar iş ilanına ulaşılamamıştır. Yine de bir yıllık sürede 292.277 iş ilanına erişilmiştir.

İş ilanları içerinde “üst düzey yönetici” pozisyonu için yayınlanan iş ilanları az sayıda yer almaktadır. Bunun sebebi hem pozisyon olarak daha az sayıda aranılması hem de bu seviyede için aranılan iş gören adaylarının genellikle dış transfer, iç transfer ya da terfi gibi uygulamalarla bulunmasıdır. Bu da araştırma için bir kısıt olarak görülmektedir.

Ayrıca mavi yaka olarak tabir edilen daha fazla beden gücü gerektiren işçi pozisyonları için de daha çok İŞKUR duyuruları kullanılmaktadır. Dolayısıyla elde edilen ilan sayıları bu pozisyonlar için talebi tam olarak temsil etmeyebilir.

3.5. Çalışmada Yararlanılan Araçlar

Çalışmada analize konu olan çevrim içi iş ilanlarının elde edilmesi için C++ programlama dili kullanılarak oluşturulan özel bir yazılımdan yararlanılmıştır. C++, Bjarne Stroustrup tarafından 1979 yılından itibaren geliştirilen orta seviyeli ve genel amaçlı bir programlama dilidir (www.yusufsezer.com.tr). Kullanılan yazılım sayesinde kariyer sitesi üzerinden web kazıma yapılarak veriler elde edilmiştir.

Elde edilen veriler çok sayıda ön işlemden geçmiştir. Bu amaçla Microsoft Excel ve Google e-tablolar gibi uygulamalardan yararlanılmıştır.

Ön işlemeye tabi tutulan verilerin analizi için ise Rapidminer Studio Version 10.1 programından yararlanılmıştır. RapidMiner makine öğrenmesi, veri madenciliği, metin madenciliği, tahmin edici analiz ve iş analizi amaçlarına yönelik olarak geliştirilmiş bir yazılım platformudur. Genel olarak iş ve ticaret dünyasında kullanıldığı gibi aynı zamanda araştırma, eğitim, hızlı prototipleme ve uygulama geliştirme gibi amaçlarla da kullanılabilir. Ayrıca veri madenciliği sürecinin tüm adımları RapidMiner tarafından desteklenmektedir, bu yüzden veri hazırlama, veri modelleme, doğrulama ve optimizasyon gibi amaçlarla da kullanılmaktadır (www.dqturkiye.com, 13.05.2023).

Veriler elde edilip ön işlemlerden geçirildikten sonra sınıflayıcı veri madenciliği yöntemlerinden olan karar ağacı analizi yapılmıştır. Karar ağacı analizleri yapılırken Random Forest algoritmasından yararlanılmıştır. Ardından metin madenciliği uygulamaları kullanılarak analizlere devam edilmiştir. Bunun için olasılıksal konu modelleme yöntemi olan Gizli Dirichlet Ayrımı (GDA) kullanılmıştır.

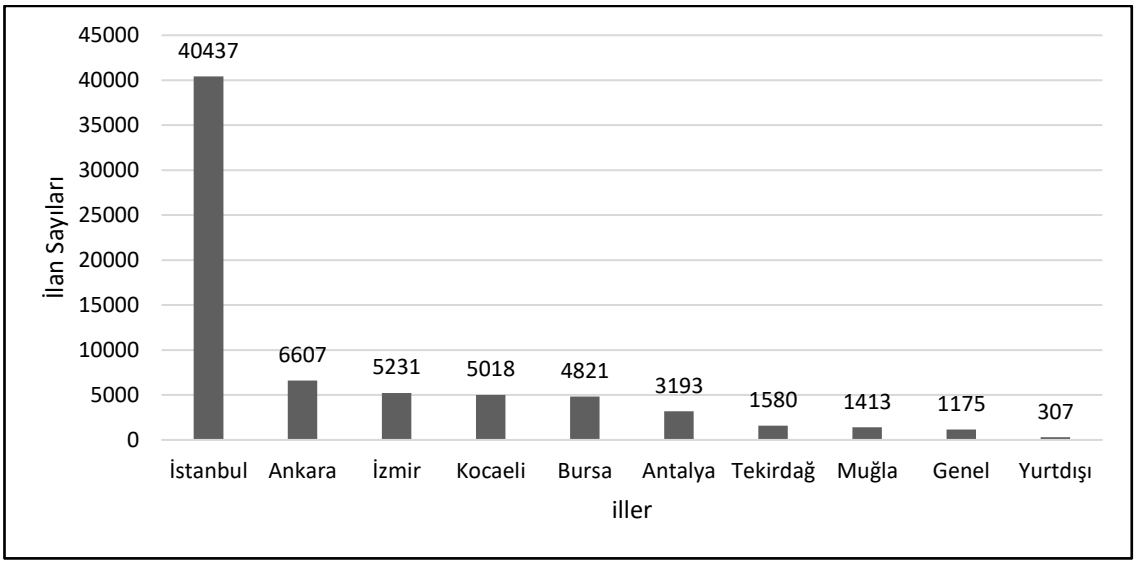
3.6. Veriler

Çalışmada kullanılan veriler Kariyer.net web sitesi üzerinden web kazıma kullanılarak elde edilmiştir. Web sayfası üzerinden veri kazıma; bir kullanıcı gibi davranarak ekranda görselleştirilen verilerin toplanmasına imkân veren otomatik taramaya dayalı bir yöntemdir. Bu yöntemleri kullanmak suretiyle internet kaynaklarından veri elde etme, veri analizi çalışmaları için son derece önemlidir (Akbulut, 2022: 450).

Veriler 07.04.2022 ve 18.03.2023 tarihleri arasında yaklaşık 1 yıllık süre içerisinde Kariyer.net web sitesinde yayınlanan ilanları içermektedir. İlgili web sitesinde

günlük ortalama yaklaşık 80 – 90 bin iş ilanı yer almaktadır. Verilerin elde edilmesi sürecinde toplam 292.277 ilana ulaşılmıştır. Ön işlemler sırasında aynı ilanların elenmesi ile toplam 81.890 eşsiz ilan veri setinde kalmıştır. Ayrıca veri seti 11 sütundan oluşmaktadır ve bu 11 sütunda yer değişkenlerin tamamı kategorik değişkenlerdir. Bu değişkenler; "firma adı", "il", "eğitim seviyesi", "tecrübe", "iş tanımı", "sektör", "çalışma şekli", pozisyon seviyesi", "departman", "lisan" ve "ilan linki" şeklindedir.

Şekil 14' te yer alan grafikte en yüksek sayıda iş ilanına sahip 8 il, genel ilanlar ve yurt dışı ilanlarının dağılımı görülmektedir.



Şekil 14. İş İlanlarının İllere Göre Dağılımı

İş ilanlarının illere göre dağılımına bakıldığında İstanbul toplam ilan sayısının yaklaşık %50' sine sahiptir. Ankara, İzmir, Kocaeli, Bursa gibi iller de yüksek miktarda iş ilanının verildiği illerdir. Ayrıca 307 adet de yurtdışı için verilen iş ilanları mevuttur.

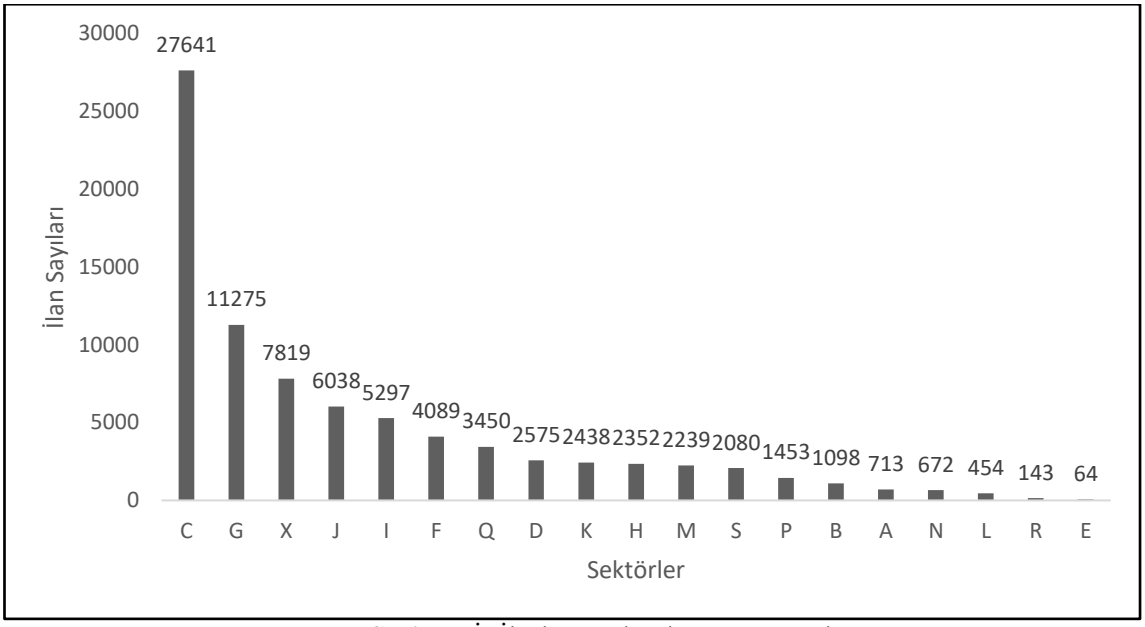
Veri setindeki iş ilanlarının NACE rev.2- Altılı Ekonomik Faaliyet Sınıflaması esas alınarak yapılan sektörel dağılımı Tablo 1' de görülmektedir.

Sektör	Sektör (Harf Kodu)	Frekans
İmalat	C	27641
Toptan ve perakende ticaret; motorlu kara taşıtlarının ve motosikletlerin onarımı	G	11275
Genel	X	7819
Bilgi ve iletişim	J	6038
Konaklama ve yiyecek hizmetleri faaliyetleri	I	5297
İnşaat	F	4089
İnsan sağlığı ve sosyal hizmet faaliyetleri	Q	3450
Elektrik, gaz, buhar ve iklimlendirme üretimi ve dağıtımı	D	2575
Finans ve sigortacılık faaliyetleri	K	2438
Ulaştırma ve depolama	H	2352
Mesleki, bilimsel ve teknik faaliyetler	M	2239
Diğer hizmet faaliyetleri	S	2080
Eğitim	P	1453
Madencilik ve taş ocakçılığı	B	1098
Tarım, ormancılık ve balıkçılık	A	713
İdari ve destek hizmet faaliyetleri	N	672
Gayrimenkul faaliyetleri	L	454
Kültür, sanat, eğlence, dinlence ve spor	R	143
Su temini; kanalizasyon, atık yönetimi ve iyileştirme faaliyetleri	E	64
Kamu yönetimi ve savunma; zorunlu sosyal güvenlik	O	0
Hane halklarının işverenler olarak faaliyetleri; hane halkları tarafından kendi kullanımlarına yönelik olarak ayrım yapılmamış mal ve hizmet üretim faaliyetleri	T	0
Uluslararası örgütler ve temsilciliklerinin faaliyetleri	U	0
TOPLAM		81.890

Tablo 1. NACE kodlarına göre sektörlere ait harf kodları ve ilan sayıları

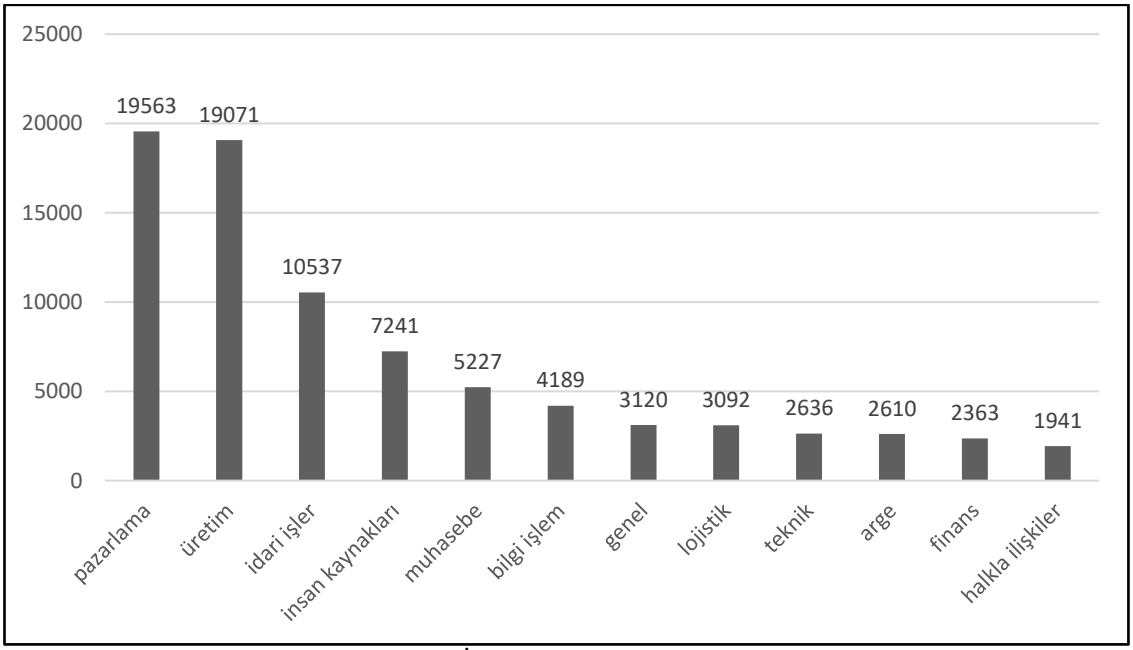
Sektörel olarak verilerin etiketlenmesi sürecinde TUİK Sınıflama Sunucusu NACE rev.2- Altılı Ekonomik Faaliyet Sınıflaması kodları esas alınmıştır. Bu şekilde birinci seviyede 22 başlık altında sektörler etiketlenmiştir. 22 başlık içerisinde "kamu yönetimi ve savunma; zorunlu sosyal güvenlik", "hane halklarının işverenler olarak faaliyetleri; hane halkları tarafından kendi kullanımlarına yönelik olarak ayrım yapılmamış mal ve hizmet üretim faaliyetleri" ve "uluslararası örgütler ve temsilciliklerinin faaliyetleri" başlıklarına ilişkin ilanlar ilgili web sitesinde yer almadığı için veri seti içerisinde de bu başlıklar bulunmamaktadır. Ayrıca herhangi bir sektör belirtilmeyen ilanlar "genel" etiketi ile veri setinde yer almaktadır. Bu şekilde toplam 19 sektör bulunmaktadır. Tablo 1' de sektörlere ait ilan sayıları ve bu sektörlerin harf kodları

yer almaktadır. Harf kodları NACE rev.2- Altılı Ekonomik Faaliyet Sınıflaması dahilinde çalışmaya dahil edilmiş olup çalışmanın ilerleyen bölümlerinde zaman zaman bu kodlardan yararlanılacaktır. Çalışma kapsamında "Bilgi ve İletişim Sektörü" için yayınlanan iş ilanlarından yararlanılmıştır. Şekil 15' te iş ilanlarının sektörel dağılımı sütun grafik yardımıyla gösterilmiştir.



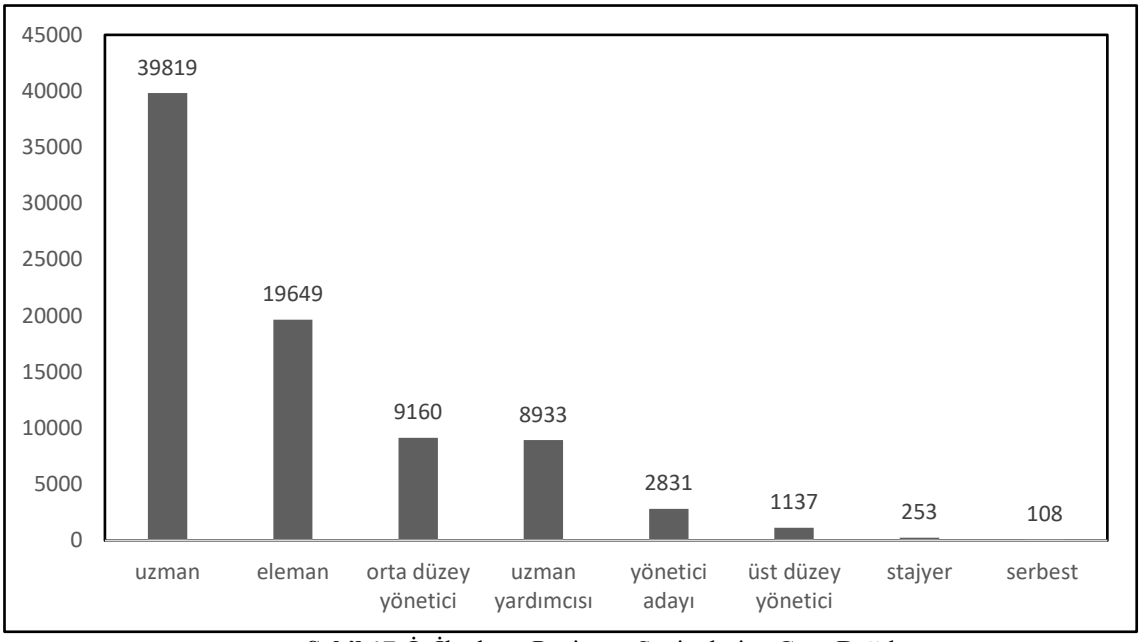
Şekil 15. İş İlanların Sektörlere Göre Dağılımı

Sektörlere göre dağılımın gösterildiği Şekil 15' te imalat sektörünün en yüksek paya sahip olduğu görülmektedir. Departmanlara göre iş ilanların dağılımı Şekil 16 yer verilen grafikte gösterilmektedir.



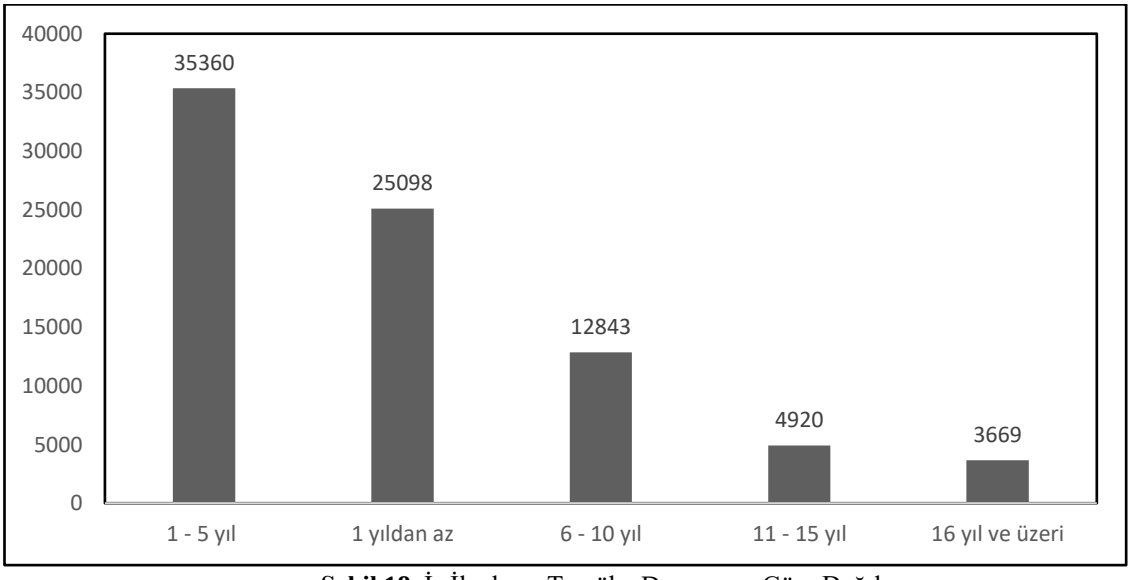
Şekil 16. İlanların Departmanlara Göre Dağılımı

Buna göre en yüksek ilan sayılar pazarlama, üretim ve idari işler departmanları için verilmiştir. İlanların pozisyon seviyesine göre dağılımlarında Şekil 17' de yer alan grafikte gösterilmektedir.



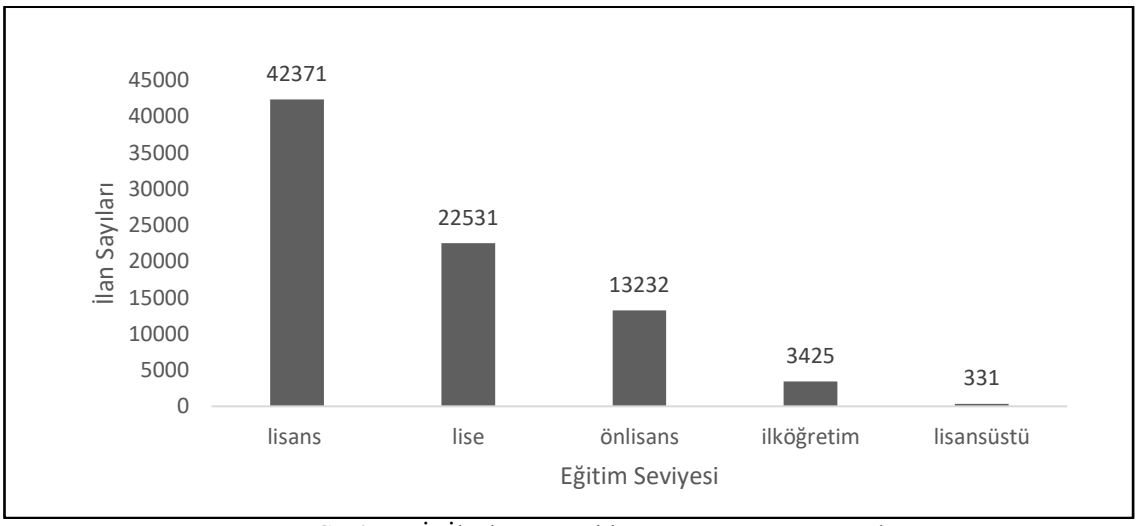
Şekil 17. İş İlanların Pozisyon Seviyelerine Göre Dağılımı

Uzman pozisyonu için verilen ilanlar toplam ilanlar içerisinde en yüksek sayıdadır. Bunun yanında işletmeler günümüzde önemi artan stajyer ve serbest çalışma (freelance) iş ilanlarını da bu platformlarda yayınlayabilmektedir. İş görenlerin tecrübelerine göre ilanların dağılımı da Şekil 18' de yer alan grafikteki gibidir.



Şekil 18. İş İlanların Tecrübe Durumuna Göre Dağılımı

Şekil 18’ de görüldüğü üzere özellikle 1-5 yıl arasına tecrübe aranan ilanlar en yüksek sayıya sahiptir. Bununla beraber tecrübesiz olarak nitelendirebileceğimiz bir yıldan az tecrübeye sahip ilanlar da %30 gibi azımsanmayacak bir orandadır. Bu durum özellikle yeni mezun iş gören adaylarının doğru kariyer planlaması ve becerilerini sektör beklentileri doğrultusunda geliştirmeleri ile istihdam edilme oranlarında artış olabileceğini göstermektedir. İş ilanlarının eğitim seviyelerine göre dağılımı ise Şekil 19’ da verilmiştir.



Şekil 19. İş İlanlarının Eğitim Durumuna Göre Dağılımı

İş ilanlarının eğitim seviyelerine göre dağılımına bakıldığında ise lisans mezuniyeti yaklaşık %50’ lik bir orana sahipken lise mezunları da %28 gibi azımsanmayacak orana sahiptir. Ayrıca ön lisans ve lise mezunu ilanlarının içerikleri incelendiğinde teknik personellerin yoğunlukla arandığı görülmektedir. Bu da mesleki eğitimin önemine işaret edebilir.

3.7. Veri Ön İşleme

Verilerin ön işlemeleri Microsoft Excel ve Google E-tablolar yardımıyla yapılmıştır. Bu süreçte yapılan işlemler aşağıda sıralanmıştır;

1- 292.277 satırdan oluşan ilanların aynı ilanı içeren satırları silinmiş ve 81.890 eşsiz ilan veri setinde kalmıştır.

2- Veri setinde yer alan bütün değerler küçük harfe dönüştürülmüştür.

3- “Eğitim seviyesi” değişkenine ait değerler “ilköğretim”, “lise”, “ön lisans”, “lisans” ve “lisansüstü” şeklinde etiketlenmiştir.

4- “Tecrübe” değişkenine ait değerler “1 yıldan az”, “1 – 5 yıl arası”, “5 – 10 yıl arası”, “11 – 15 yıl arası” ve “16 yıl ve üzeri” şeklinde etiketlenmiştir.

5- "Çalışma Şekli" değişkenine ait değerler "stajyer", "yarı zamanlı", "serbest", "tam zamanlı" ve "dönemsel" şeklinde etiketlenmiştir.

6- "Pozisyon Seviyesi" değişkenine ait değerler "eleman", "orta düzey yönetici", "serbest", "stajyer", "uzman", "uzman yardımcısı", "üst düzey yönetici" ve "yönetici adayı" şeklinde etiketlenmiştir.

7- "Departman" değişkenine ait değerler "pazarlama", "üretim", "idari işler", "insan kaynakları", "muhasebe", "bilgi işlem", "lojistik", "teknik", "ar-ge", "finans", "halkla ilişkiler" şeklinde etiketlenmiştir. Herhangi bir departman belirtmeyen ilanlar ise "genel" etiketi ile veri setinde yer almıştır.

8- "Lisan" değişkenine ait değerler "İngilizce", "Türkçe" ve "Diğer" şeklinde etiketlenmiştir.

9- "İl" değişkeni sütunu 81 iline uygun şekilde düzenlemiştir. Bu illere ek olarak "Tüm Türkiye" ve "Yurtdışı" değerleri de dahil edilmiştir.

10- "İş Tanımı" sütununda yer alan tüm noktalama işaretleri, sık tekrar eden bazı kelimeler silinmiştir. Ayrıca analizlerin daha sağlıklı yapılabilmesi adına bu sütunda yer alan metinler İngilizce' ye çevrilmiştir. Çeviriler yapıldıktan sonra ilanlar arasından rastgele seçim yapılarak bir kısım ilanın çevirisi kontrol edilmiştir.

11- "İş Tanımı" sütununda yer alan "QUALIFICATIONS", "JOB DESCRIPTION", "GENEL NİTELİKLER", "İŞ TANIMI", "tercihen", "summary" gibi bütün ilanlarda görülen kelimeler silinmiştir.

12- Çeviriler yapıldıktan sonra veri setinde yer alan Türkçe karakterlerin Excel programı formülleri yardımıyla İngilizce' ye uygun hale getirilmiştir.

13- "Sektör" değişkenine ait değerler NACE rev.2- Altılı Ekonomik Faaliyet Sınıflaması kodlarına göre etiketlenmiştir.

3.8. Verilerin Analizi ve Bulgular

Verilerin analizinde öncelikle iş ilanlarında yer alan değişkenlerin birbirleri ile olan ilişkilerini tespit etmek adına karar ağacı algoritmalarından yararlanılmıştır. Sonrasında "iş tanımı" değişkeni üzerinde olasılıksal bir yöntem olan Gizli Dirichlet Ayrımı (GDA) tekniği kullanılarak konu modelleme analizi yapılmıştır.

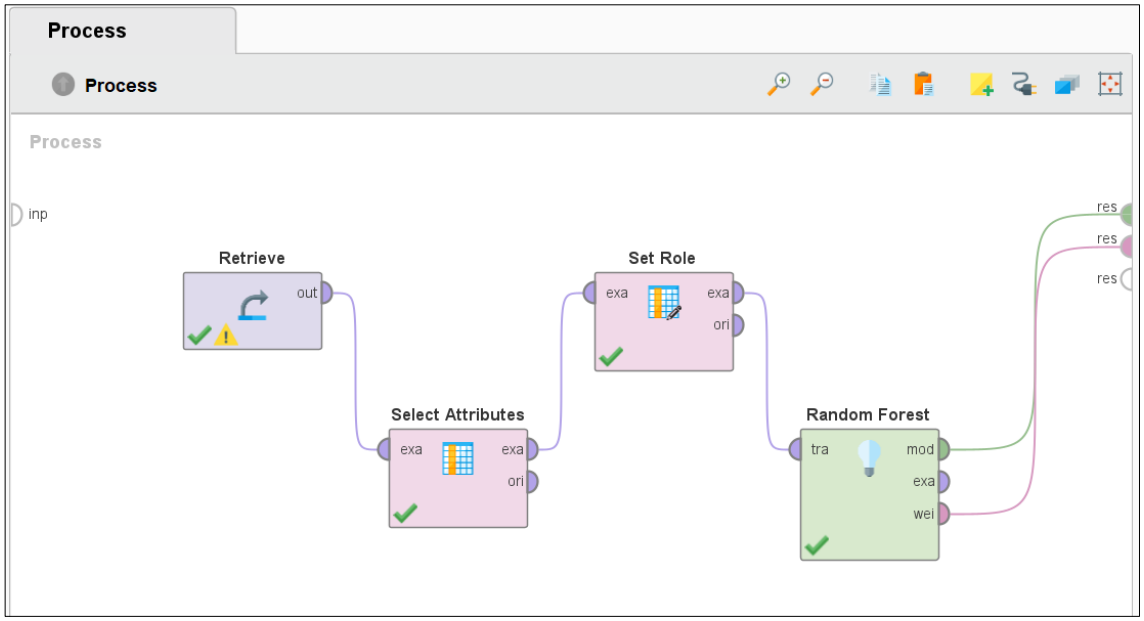
GDA analizi, “Bilgi ve İletişim Sektörü” ilanları kullanılarak yapılmıştır. Bunun sebebi öncelikle hızla gelişen teknoloji karşısında bu sektördeki beklentilerin nasıl şekillendiğini gizli kalmış konuların tespiti ile ortaya koymaktır.

Analizler yapılırken öncelikle veri setinde yer alan toplam on iki aylık veri üçer aylık periyotlara ayrılmış ve analiz her üç aylık periyoda uygulanmıştır. Bu şekilde bir yıllık süreç içerisinde iş ilanlarında sektörün beklentilerinde olabilecek değişimlerin tespit edilmesi amaçlanmıştır. Ardından ilanlar “lisan mezunu” ve “ön lisans” mezunu şeklinde filtrelenip analizler yapılmış ve aradaki farklar tespit edilmeye çalışılmıştır. Yine ilanlar, tecrübeli ve tecrübesiz adaylarda aranan niteliklerde farklar olup olmadığının tespiti için “bir yıldan az tecrübeli” ve “bir yılın üzerinde tecrübeye sahip” şeklinde filtrelenip tekrardan analiz edilmiştir.

3.8.1. Bilgi ve İletişim Sektörü İlanlarının Karar Ağacı Yöntemi ile Analizi

Çalışanlar profesyonel kariyer yaşamları boyunca farklı seviyelerde ve farklı departmanlarda çalışma fırsatları bulabilirler. Muhakkak herkes çalışmak istediği departmanda ve ulaşabileceği en yüksek pozisyonda çalışmak ister. Bu bölümde “pozisyon seviyesi” ve “departman” değişkenleri üzerinde etkisi olan değişkenler ve ağırlıkları tespit edilmiş ve belirli pozisyon seviyeleri ve departmanlar için mümkün olan kurallar bulunmaya çalışılmıştır. Söz konusu analizler için bilgi ve iletişim sektörü ilanları ele alınmış olup Random Forest karar ağacı algoritması kullanılmıştır.

Random Forest karar ağacı algoritması veri setinde yer alan en iyi özelliklerden seçilmiş düğümleri dallara ayırmadan her düğümde rastgele alınan özelliklerin en iyisini seçip bütün düğümler, dallara ayırır. Her bir veri kümesi, ana veri setinden yer değiştirmek suretiyle üretilir. Budama işlemi yoktur. Random Forest algoritmasının diğer algoritmalara göre daha hızlı ve doğru olmasının sebebi bu yöntemdir (Akçetin ve Çelik, 2014: 51). Analizler, RapidMiner Studio Version 10.1 kullanılarak yapılmıştır. Şekil 20’ de karar ağacı analizine ilişkin RapidMiner 10.1 ekran görüntüsü yer almaktadır.



Şekil 20. Random Forest Algoritması Rapidminer Ekran Görüntüsü

Şekil 20' de görüldüğü üzere random forest algoritması uygulanırken veri seti için "retrieve", analizde kullanılacak değişkenlerin seçilmesi için "select attributes", etiketleme için "set role" ve algoritmanın uygulanması için "random forest" operatörlerinden faydalanılmıştır.

Literatürde karar ağacı algoritmaları için çeşitli özellik çıkarım yöntemleri bulunmaktadır. Bilgi kazancı (information gain), kazanç oranı (gain ratio) ve gini katsayısıdır (gini index) bunlardan bazılarıdır.

Bilgi kazancı ve kazanç oranı entropi temelli özellik seçimi yöntemlerindendir. Entropi, bir sistemdeki olması muhtemel belirsizliğin ölçüsü şeklinde tanımlanabilir. Entropi, gelişigüzel bir örnek koleksiyonunun saflığını karakterize eden bilgi teorisi ölçüsünde yaygın olarak kullanılır. Bilgi kazancı ve kazanç oranı gibi öznitelik sıralama yöntemlerinin temelinde yer alır. Entropi ölçüsü, sistemin öngörülemezliğinin bir ölçüsü olarak kabul edilir. Entropi Denklem 13' teki gibi hesaplanır (Novakovic vd., 2011: 122):

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)) \qquad (13)$$

Denklem 13' te $p(y)$, Y rasgele değişkeni için marjinal olasılık yoğunluk fonksiyonudur. Eğitim veri setinde (S) gözlenen Y değerleri, ikinci bir X özelliğinin değerlerine ve Y değerinin entropisine göre bölümlendirilirse X tarafından uyarılmış

bölümler, *Y* değerinin bölümlemeden önceki entropisinden daha küçüktür, o zaman *Y* ve *X* özellikleri arasında bir ilişki vardır. O zaman *X* gözlemlendikten sonra *Y* ' nin entropisi Denklem 14' teki gibidir.

$$H(Y\backslash X) = -\sum_{x\in X} p(x) \sum_{y\in Y} p(y\backslash x) \log_2(p(y\backslash x)) \quad (14)$$

Denklem 14' te de *p(y|x),* verilen *x* için *y*' nin koşullu olasılığıdır (Novakovic vd., 2011: 122).

Buna bağlı olarak, bilgi kazancı Denklem 15' teki şekilde hesaplanmaktadır:

$$Bilgi\ Kazancı = H(Y) - H(Y\backslash X) = H(X) - H(X\backslash Y) \quad (15)$$

Ayrıca 0 ile 1 arasında değer alan kazanç oranı ise Denklem 16' deki şekilde hesaplanır:

$$Kazanç\ Oranı = \frac{Bilgi\ Kazancı}{H(X)} \quad (16)$$

Burada oranın bire eşit olması durumu X bilgisi ile Y bilgisini tamamen tanımlanabildiğini ifade eder. Oran eğer sıfıra değeri alırsa X ile Y arasında herhangi bir ilgi olmadığını ifade eder (Novakovic vd., 2011: 122).

Öncelikle "pozisyon seviyesi" etiketi ile yapılan analizlerde pozisyon seviyesi değişkenine etki eden değişkenlerin ağırlıkları belirlenmiştir. Ağırlıklar belirlenirken kazanç oranı kriteri esas alınmıştır. Sonuçlar Tablo 2' de görülmektedir.

Değişkenler	Tecrübe	Lisan	Çalışma Şekli	Eğitim	Departman	Toplam
Ağırlıklar	0,205	0,456	0,061	0,068	0,210	1,000

Tablo 2. Pozisyon Seviyesi değişkenine etki eden değişkenlerin ağırlıkları

Tablo 2' de yer alan sonuçlara bakıldığında k en önemli değişkenlerin tecrübe ve lisan olduğu görülmektedir. Çalışma şekli, eğitim seviyesi veya departman, pozisyon seviyesi için etkili değişkenler olduğu söylenemez. "Pozisyon seviyesi" etiketi ile yapılan analiz sonucunda elde edilen kurallardan bazıları Tablo 3' te verilmiştir.

Mezuniyet	Tecrübe	Yabancı dil	Pozisyon seviyesi	Çalışma şekli	Departman	Yüzde (%)
Lise mezunu	1-5 yıl deneyimli	NA	Eleman	Tam zamanlı	İdari işler	5,06
Lise mezunu	1 yıldan az deneyimli	NA	Eleman	Tam zamanlı	Pazarlama	5,18
Lisans mezunu	1-5 yıl deneyimli	İngilizce	Uzman	Tam zamanlı	Pazarlama	13,05
Lisans mezunu	1-5 yıl deneyimli	NA	Uzman	Tam Zamanlı	Üretim	13,19
Lisans mezunu	1 yıldan az deneyimli	İngilizce	Uzman	Freelance	Bilgi işlem	3,44

Tablo 3. Pozisyon seviyesi değişkenine etki eden faktörler

Tablo 3' te farklı pozisyon seviyeleri ortaya çıkan örüntülerden örnekler verilmiştir. Buna göre örneğin, "üretim departmanı için lisans mezunu, 1-5 yıl deneyim sahibi, İngilizce bilen tam zamanlı uzman olarak çalışma ihtimali % 13,19' dur" şeklinde bir örüntü tanımlanabilir. Departman değişkenine etkisi olan değişkenlerin ağırlıkları Tablo 4' te görülmektedir.

Değişkenler	Lisan	Tecrübe	Çalışma Şekli	Eğitim	Pozisyon Seviyesi	Toplam
Ağırlıklar	0,287	0,397	0,040	0,098	0,178	1,000

Tablo 4. Departman değişkenine etki eden değişkenlerin ağırlıkları

Tablo 4' te görüldüğü üzere lisan, tecrübe ve pozisyon seviyesi de departman üzerinde etkili olan değişkenlerdir. Eğitim seviyesi nispeten daha düşük düzeyde etkiye sahiptir. "Departman" etiketi ile yapılan analiz sonucunda elde edilen kurallardan bazıları Tablo 5' te verilmiştir.

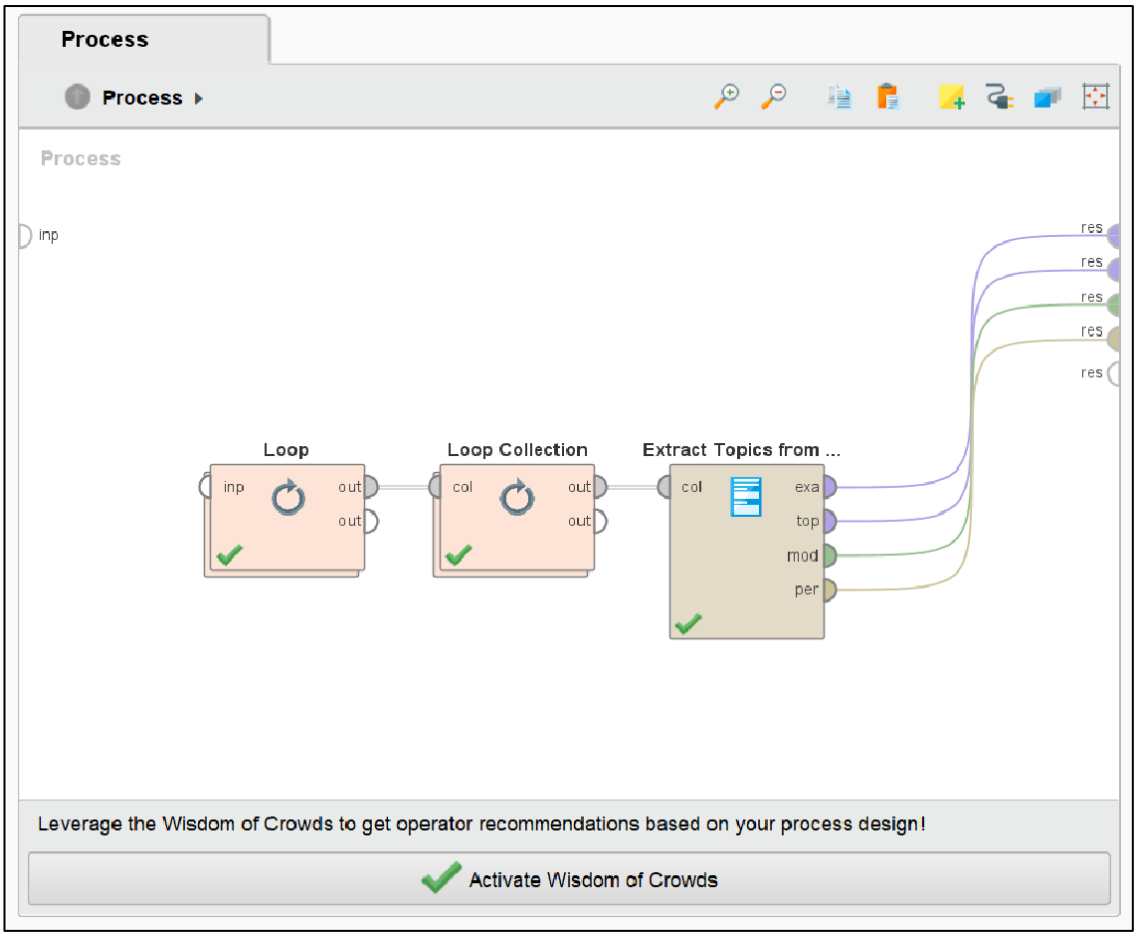
Mezuniyet	Tecrübe	Yabancı dil	Pozisyon seviyesi	Çalışma şekli	Departman	Yüzde (%)
Lisans mezunu	1-5 yıl deneyimli	NA	Uzman	Tam zamanlı	Pazarlama	3,56
Lise mezunu	1 yıldan az deneyimli	NA	Eleman	Tam zamanlı	Üretim	3,89
Lisansüstü mezun	1-5 yıl deneyimli	İngilizce	Üst düzey yönetici	Tam zamanlı	Üretim	0,01
Lisans mezunu	1 yıldan az deneyimli	İngilizce	Uzman yardımcısı	Tam zamanlı	Pazarlama	5,81

Tablo 5. Departman değişkenine etki eden faktörler

Tablo 5' te farklı departmanlar ortaya çıkan örüntülerden örnekler verilmiştir. Buna göre örneğin, "lisans mezunu, İngilizce bilen, 1 yıldan az deneyimli, uzman yardımcısının pazarlama departmanında çalışma ihtimali % 5,81' dir" şeklinde bir örüntü tanımlanabilir.

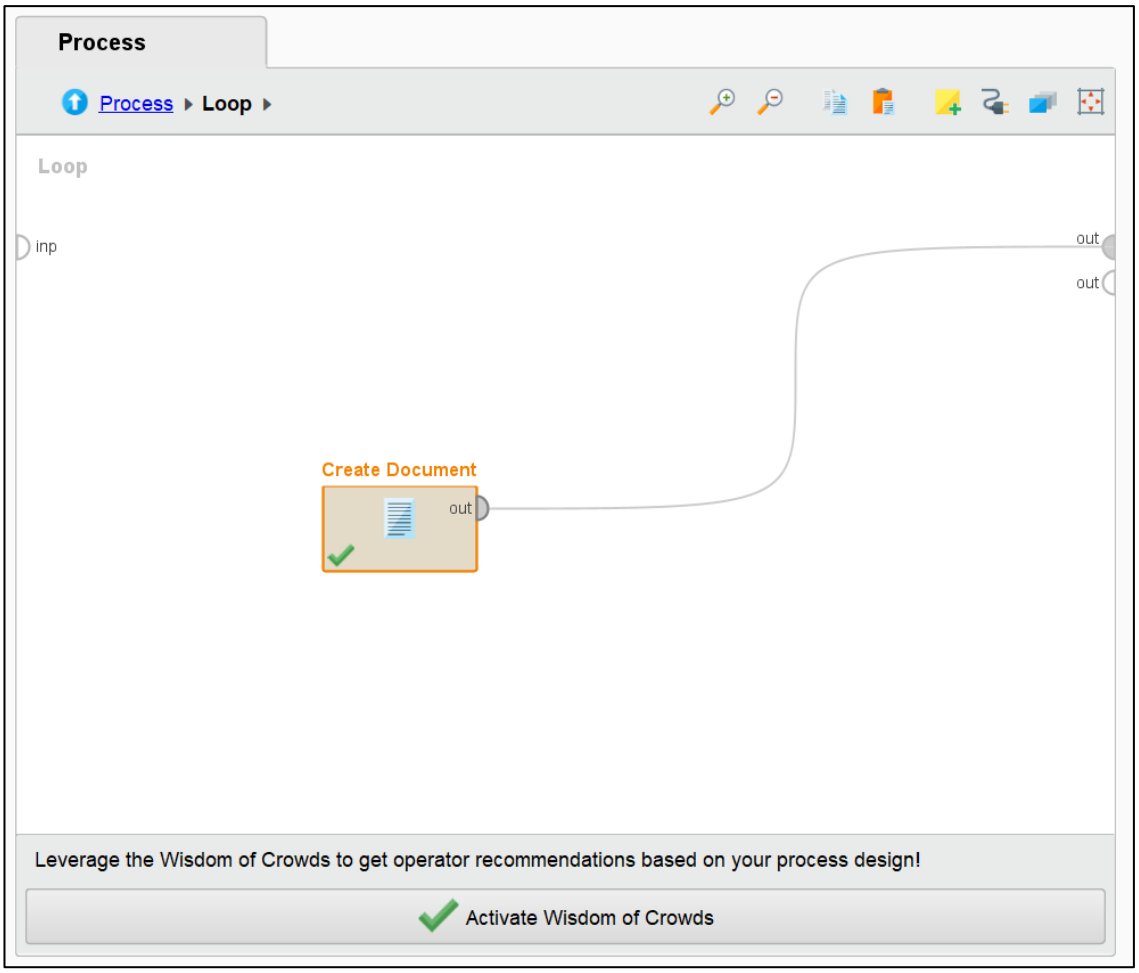
3.8.2. GDA Analizi İçin Rapiminer Studio 10.1 Kullanımı

Bu bölümde GDA analizleri için Rapidminer Studio 10.1 programının kullanımı ekran görüntüleri yardımıyla açıklanmıştır. Şekil 21' de GDA analizi için oluşturulan model görülmektedir.



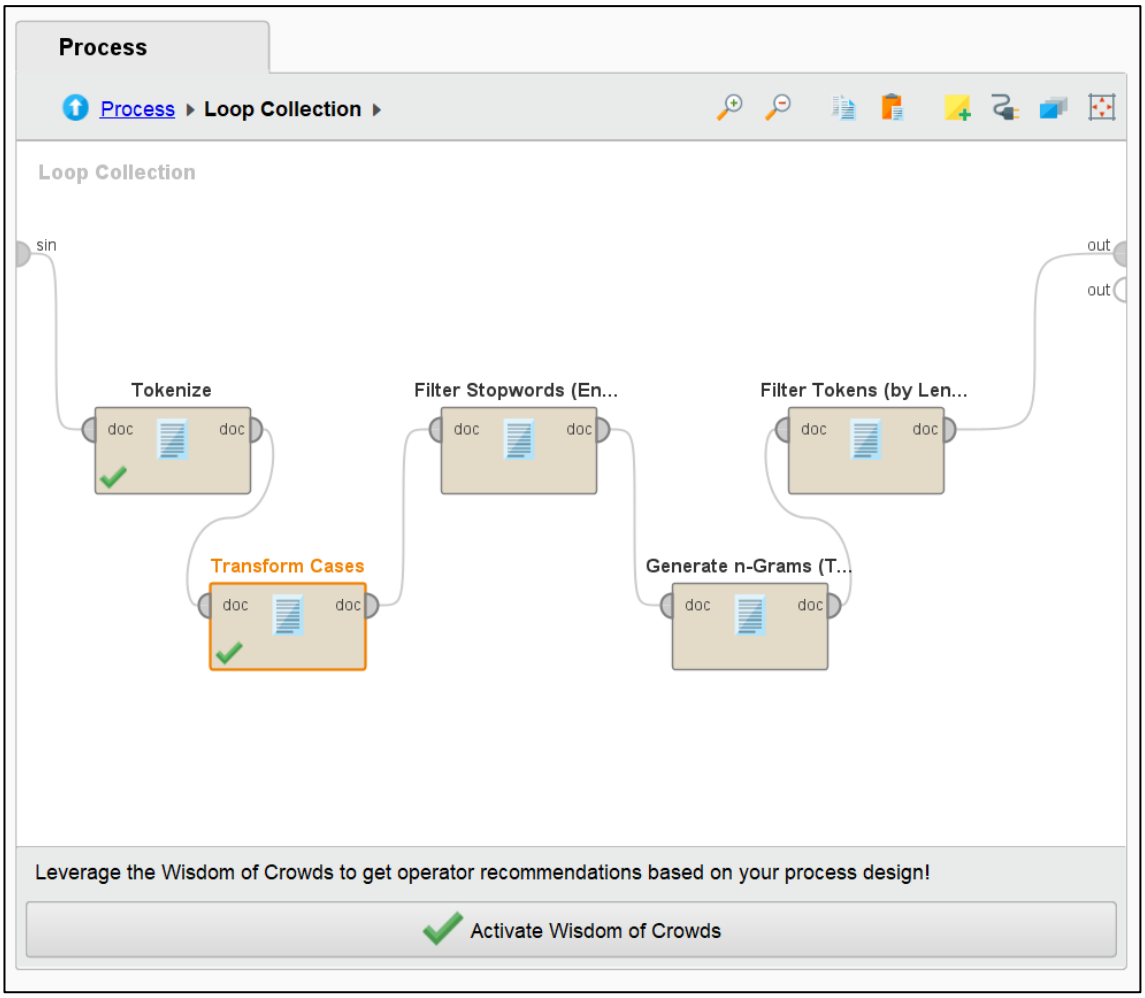
Şekil 21. GDA Analizi İçin Kullanılan Operatörler

Şekil 21' de görülen operatörler RapidMiner programı ana ekranda yer alan "operators" kısmından bulunup, seçilip ardından tutup-sürüklemek suretiyle ana ekrana bırakılmıştır. Ardından bu operatörler birbirlerine ve sonuç kısmına bağlanarak model oluşturulmuştur. Her bir operatör için bazı parametreler belirlemek gerekmektedir. Öncelikle Şekil 22' de "loop" operatörü için gerekli operatör ve parametrelerin ekran görüntüleri yer almaktadır.



Şekil 22. Loop Operatörü Ekran Görüntüsü

Şekil 22' de görülen "loop" operatörü çift tıklanmak suretiyle açılır ve boş bir "process" ekranı gelir. Ardından bu ekrana yine "operators" kısmından "create document" operatörü bulunup boş ekrana bırakılır. "create document" operatörü üzerine tıklanarak sağ tarafta yer alan "parameters" kısmından "edit text" kısmı seçilerek GDA analizi yapılacak metin bulunduğu yerden kopyalanıp buraya yapıştırmak suretiyle programa yüklenir. İşlem tamamlandıktan sonra ana ekrana dönülür. Şekil 23' te "loop collection" operatörü içeriğinde yer alan diğer operatörler için ekran görüntüsü yer almaktadır.



Şekil 23. Loop Collection Operatörü Ekran Görüntüsü

"Loop collection" operatörü çift tıklanmak suretiyle boş bir "process" sayfası açılır. Sırasıyla "tokenize" operatörü "non letters" parametresi seçilerek, "transform cases" operatörü "lower cases" parametresi seçilerek, "filter stopwords (english)" operatörü seçilerek, "create n-grams(terms)" operatörü "2 kelime" seçilerek ve "filter tokens(by lenght)" operatörü min=4 – max=25 şeklinde seçilerek Şekil 21' deki ekrana ulaşılır. Bu işlemler tamamlandıktan sonra ana ekrana dönülür. Şekil 21' de görülen ana ekrana dönüldükten sonra "Extract Topic from Documents (LDA)" operatörü üzerine tek tıklanarak ekranın sağında yer alan parametreler belirlenir. Burada konu sayısı, her konuyu temsil eden kelime sayısı, modelin tekrar etme sayısı gibi parametreler belirlendikten sonra üstte yer alan araç çubuğundaki "play" simgesi tıklanarak model çalıştırılır ve sonuç ekranına ulaşılır.

3.8.3. GDA Analizi İçin Optimum Konu Sayısının Belirlenmesi

GDA yönteminde sonuçların sağlıklı olması adına önemli parametrelerden birisi de konu sayısıdır. Konu sayısının belirlenmesi ile ilgili literatürde farklı uygulamalar

bulunmaktadır. Bazı çalışmalarda konu sayısı uygulayıcı tarafından yoruma dayalı olarak belirlenmiştir. Budak ve Sökmen (2022), Ekinci ve Omurca (2016), Altıntaş vd. (2021), Ankaralı ve Külcü (2020), GDA yöntemi kullanırken konu sayılarını yoruma dayalı olarak belirlemişlerdir.

Literatüre bakıldığında tutarlılık değeri de konu sayısının belirlenmesinde kullanılan göstergelerden birisidir. Aşağıda bununla ilgili yapılan çalışmalardan bazı örnekler bulunmaktadır.

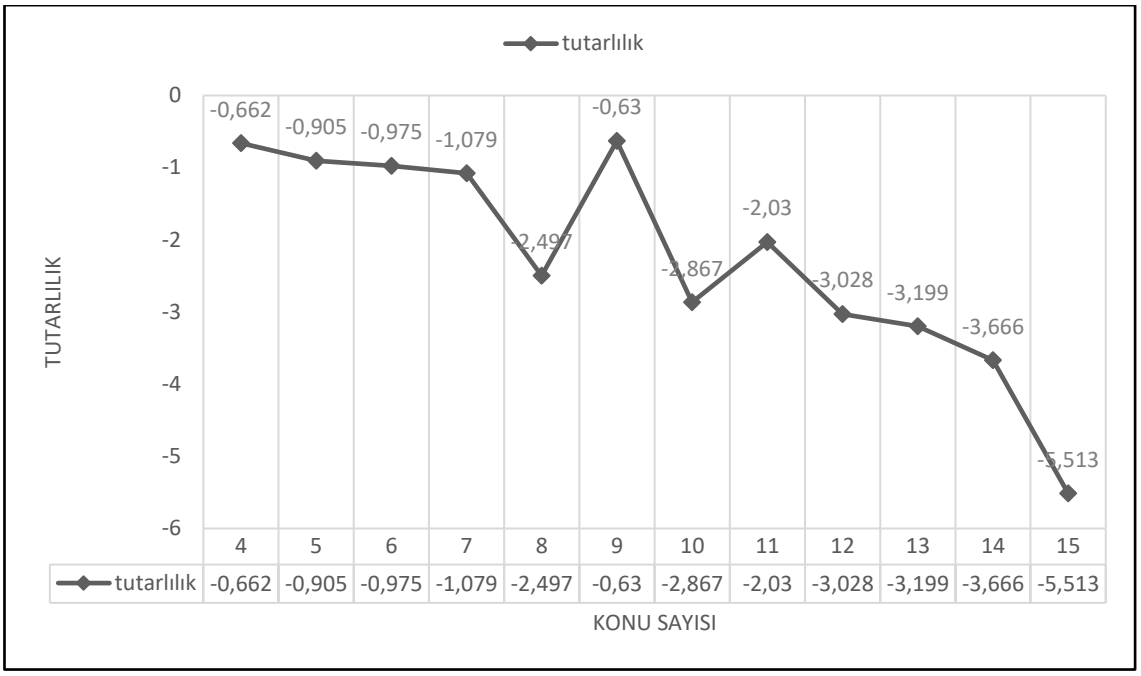
Gürcan ve Özyurt (2020) konu modelleme yöntemi kullanarak E-öğrenme araştırmalarındaki temel eğilim ve bilgi alanları ile ilgili boyutları tespit etmeye çalışmışlardır. Bu süreçte analizin önemli parametrelerinden biri olan konu sayısını 25 – 75 arasında değişen değerlerde yöntemi veri setine uygulayıp konu sayısını k=42 şeklinde tespit etmişlerdir. Bu yöntemde elde konu isimleri ortaya kendilerini temsil eden kelimeler doğrultusunda uygulayıcı ya da onun tarafından görevlendirilen bir uzman tarafından belirlenmektedir.

Ala ve Uğuz (2021), Türkiye' deki bölgesel kalkınmanın girişimcilik, inovasyon ve ar-ge çalışmalarıyla ilişkisini araştırdıkları çalışmada konu sayısını en yüksek tutarlılık değerine sahip k=9 şeklinde belirlemişlerdir.

Konu sayısının belirlenmesi için iterasyona dayalı optimizasyon yöntemlerinin haricinde yoruma dayalı olarak konu sayısının belirlenmesi de uygun sonuçlar vermektedir. Bu doğrultuda Çallı ve Alma (2021), Covid-19 Aşı tereddüdüne sahip hekimlerin Twitter paylaşımlarını GDA Yöntemi ile analiz ettikleri çalışmalarında konu sayısını k=6 şeklinde belirlemişlerdir.

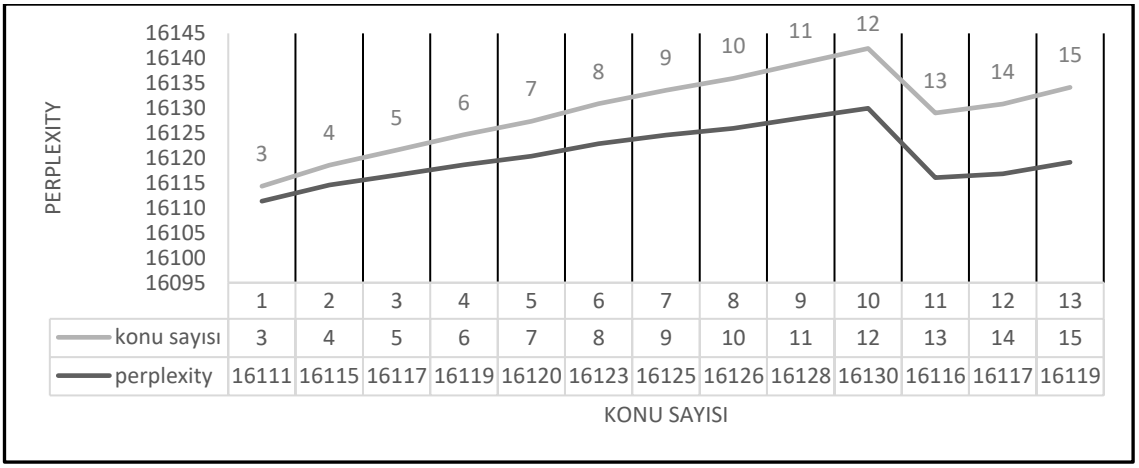
Nguyen ve Ho (2021) otel hizmetlerindeki müşteri deneyimlerini konu modelleme ile analiz ettikleri çalışmada konu sayısını en yüksek tutarlılık değerine bakarak k=14 şeklinde belirlemişlerdir.

Bu çalışmada konu sayısının belirlenmesi için model 3 konu sayısından 15 konu sayısına kadar tekrar tekrar çalıştırılmıştır. Yapılan analiz sonuçlarına göre farklı konu sayılarında yapılan analizlere dair tutarlılık ve perplexity değerleri de Şekil 24' te yer alan grafikte verilmiştir.



Şekil 24. Farklı Konu Sayılarına Göre Tutarlılık Değerleri

Tutarlılık değerinin sıfıra yakın değerler alması konuyu temsil eden değerlerin bir arada bulunma eğilimlerinin yüksek olduğunu ifade eder. Şekil 24' te yer alan değerlere göre 0' a en yakın tutarlılık değer k=9 konu sayısındadır. Ayrıca Şekil 25' te yer alan grafikte farklı konu sayılarına göre perplexity değerleri de gösterilmektedir.



Şekil 25. Farklı Konu Sayılarına Göre Perplexity Değerleri

Perplexity değerinin düşük olması durumu konu sayısının da daha anlamlı olduğunu ifade eder (Çallı vd., 2021: 2364). Şekil 25' te görüldüğü üzere en düşük perplexity değeri k=3' te görülmektedir. Fakat yapılan pilot çalışmalarla üç konu başlığının analizler için yeterli olmadığı ve bunun için perplexity değeri ile tutarlılık değerinin birlikte ele alınıp tutarlılık değerinin sıfıra en yakın değeri aldığı k=9, konu sayısı olarak belirlenmiştir.

3.8.4. Bilgi ve İletişim Sektörü İlanlarının Üçer Aylık Dönemlerinin GDA Analizleri

Bu bölümde mevcut veri seti elde edilmeye başlandığı aydan itibaren üçer aylık bölümlere ayrılmıştır. GDA analizi her üç aylık veriler için uygulanmıştır. Daha önce de bahsedildiği gibi analiz sonucunda ortaya çıkan konular uygulayıcı veya alan uzmanları tarafından isimlendirilebilir. Çalışmada sadece baskın konular için konu başlıkları belirlenmiştir.

Bilgi ve İletişim Sektörü Nisan, Mayıs ve Haziran 2022 İlanları

Tablo 6' te 2022 yılı nisan, mayıs ve haziran aylarına ait GDA analizi sonuçları bulunmaktadır.

Konular	Kelimeler	Confidence
Konu 1	Raporlama, lisans mezun, tedarikli zincir işlemler, işleme, veri tabanı, yaratıcı, Windows, mekanik, banka, organizasyon, veri analisti	0,071
Konu 2	Müşteri, deneyim, sistem yönetmek, askerlik, ürün ve hizmet, takım çalışması, sistem, yetenekler, Office programları, proje, lisans mezunu	0,177
Konu 3	Takım çalışması, iş çözümleri, iletişim, tecrübeli, veri, bilgi, analiz, tasarım, proje, hizmet, raporlama	0,194
Konu 4	Yönetici, değerleme, etkili iletişi0m, sorumluluk, müşteri sağlamak, yazılım çözümleri, kullanıcı desteği, muhasebe kayıtları, raporlama, deneyim	0,045
Konu 5	Değer zinciri, uyum sağlamak, yazılı iletişim, organizasyon, lisans mezunu, etkili iletişim, dijital pazarlama, bilgisayar, oryantasyon, yedek parça, etik	0,050
Konu 6	Koordinasyon, sonuç odaklı, programlar, pozitif, destek, raporlama, işlemler, ulusal firma, satış ve pazarlama, problemleri çözmek, aktif araç kullanımı	0,062
Konu 7	Teknolojik gelişmeler, yenilenmiş ürün, takım çalışması, mekatronik mühendisliği, proje paydaşları, işletme, ulaşım, grafik, yazılım ekibi, otomasyon	0,031
Konu 8	Deneyim, kalite, veri, lisans mezun, analitik düşünce, iletişim, takım çalışması, tasarım, hazırlık, MS Office	0,145
Konu 9	MS Office, yazılım, satış sonrası, yetenek, uzaktan çalışma, teknoloji, destek, mühendislik, işletme, programlar	0,224
	Toplam	1,000

Tablo 6. Bilgi ve İletişim Sektörü 2022 Yılı Nisan, Mayıs ve Haziran İlanları

Tablo 6' da görüldüğü gibi baskın konular Konu 9 (0,224), Konu 3 (0,194), Konu 2 (0,177) ve Konu 8 (0,145) şeklinde oluşmuştur. Burada yer alan confidence değeri, Bilgi ve İletişim Sektörü için yayınlanan bir sonraki ilanın bu konu ile alakalı olma

olasılığının yaklaşık % 22 olduğunu ifade etmektedir. Konu 9' da, "MS Office", "yazılım", "satış sonrası", "yetenek", "uzaktan çalışma", "teknoloji", "destek" gibi kelimeler ön plana çıkmıştır. Bir sonraki baskın konu olan Konu 3' te ise "takım çalışması", "iş çözümleri", "iletişim", "tecrübeli", "veri", "bilgi", "analiz", "tasarım", "proje", "hizmet", "raporlama" kelimeleri ön plana çıkmıştır. Konu 2' de "müşteri", "deneyim", "sistem yönetmek", "askerlik", "ürün ve hizmet", "takım çalışması" gibi kelimeler konuyu temsil etmektedirler.

Buradaki kelimelere bakıldığında MS Office programlarının kullanımı, yazılım bilgisi, veri analizi, tasarım, sistem yönetimi gibi teknik becerilerin yanı sıra iletişim becerileri, takım çalışması gibi hassas becerilerin de bilgi ve iletişim sektöründe aranan nitelikler olduğunu görülmektedir. Ayrıca ilanların genelinde erkek adayların askerliklerini yapmış olmaları göze çarpan iş veren taleplerindendir.

Bilgi ve İletişim Sektörü Temmuz, Ağustos ve Eylül 2022 İlanları

Bilgi ve İletişim Sektörü için 2022 yılı Temmuz, Ağustos ve Eylül aylarını GDA analizi sonuçları da Tablo 7' de verilmiştir.

Konular	Kelimeler	Confidence
Konu 1	Deneyim, bilgi, Yazılım platformu, Excel, bilgisayar mühendisliği, sunum becerileri, Kodlama İlkeleri, Child R Servis, sistem, sorun çözme, İngilizce	0,163
Konu 2	Routing switching, Teknoloji Danışmanlığı, Deneyimli HTML, Develops ALM, Programlama Deneyimi, UI UX, dijital müşteri, mühendislik, fiber, iletişim, üretim operasyonları, medya trendleri	0,035
Konu 3	Microsoft Windows, teknik servis, kariyer gelişimi, deneyimli adaylar, ürün müdürü, Neurocheck, Frontline, anahtar teslim proje, mobil arayüz, Raporlar Yönetimi, SDLC ürünü, SCRUM	0,096
Konu 4	Servis sağlayıcı, kullanıcı desteği, ürün portföyü, değer zincirleri, portal, analitik düşünce, yazılım mühendisleri, cursor SQL, Instagram, korsan yazılım	0,035
Konu 5	Takım çalışması, sorumluluk, endüstri, müşteri ziyareti, chip, Web Application, askerlik görevi, Canias ERP, D Technologies, sinyalizasyon	0,038
Konu 6	Kart yazıcıları, yönetim, dijital iletişim, robotic, node-red, Mühendislik Matematiği, raporlama, lisans mezunu, MS Office, veri, linux,	0,077
Konu 7	Veri tabanı, json, takım çalışması, medya teknolojisi, SAP, Oracle, ssi legislation, güvenlik duvarı, networks, medya dijital yayıncılık, sunum kabiliyeti	0,035
Konu 8	İletişim, yazılım, işletme, hizmet, takım, çözümler, destek, müşteriler, sektör, tasarım	0,317
Konu 9	Zaman yönetimi, yazılım geliştirici, iş, müşteri, süreçler, askerlik görevi, teknik, teknoloji, MS Office, analiz, sistemler, sorun	0,204
	Toplam	1,000

Tablo 7. Bilgi ve İletişim Sektörü 2022 Yılı Temmuz, Ağustos ve Eylül İlanları

Buna göre Konu 8 baskın konu olarak göze çarpmaktadır. Tablo 7' ye göre baskın konular sırasıyla Konu 8 (0,317), Konu 9 (0,204) ve Konu 1 (0,163) şeklinde oluşmuştur. Konu 8 altında çıkan kelimeler "iletişim", "yazılım", "işletme", "hizmet", "takım", "çözümler", "destek", "müşteriler", "sektör", "tasarım" şeklindedir. Konu 9' u temsil eden kelimelere bakıldığında ise "zaman yönetimi", "yazılım geliştirici", "müşteri", "süreçler", "askerlik görevi", "teknik, teknoloji", "MS Office", "analiz", "sistemler", "sorun" ön plana çıkmaktadır. Konu 1 altında da "deneyim", "bilgi", "yazılım platformu", "Excel", "bilgisayar mühendisliği", "sunum becerileri", "kodlama ilkeleri", "Child R Service", "sistem", "sorun çözme", "İngilizce" kelimeleri görülmektedir.

Bilgi ve İletişim Sektörü Ekim, Kasım ve Aralık 2022 İlanları

Tablo 8' de Bilgi ve İletişim Sektörü 2022 Ekim, Kasım ve Aralık aylarına ait ilanların GDA analizi sonuçları görülmektedir.

Konular	Kelimeler	Confidence
Konu 1	Yetenek, temsil, montaj süreci, deneyim, medical hizmetler, müşteri ihtiyaçları, stratejik ortaklık, servis takımları, uzaktan çalışmak	0,041
Konu 2	Kişisel gelişim, tasarım geliştirme, tedarik yönetimi, yazılım tecrübesi, MS Office, takım çalışması, teknoloji markaları, lisansüstü, operasyonel süreçler, pazar araştırması	0,041
Konu 3	Dinamik, müşteri çözümleri, ithalat, hızlı uyum, lojistik, web api, lisans çalışması, proje planlama, sunucular, yazılım	0,042
Konu 4	Yazılım, askerlik görevi, MS Office, üretken, parasal stratejiler, kalite, uygulama, fikir, kart, raporlama, muhasebe	0,113
Konu 5	Lisans mezun, bağlantı, deneyim, JIRA ZEPHYR, mühendislik, istatistik, denetim, iş odaklı, ERP uygulamaları, muhasebe	0,041
Konu 6	Planlama, e ticaret, yıllık rapor, Design application, geliştirme, tasarım, SQL, bilgi, yüksek lisans, inovasyon	0,043
Konu 7	Yenileme, yazılım uygulamaları, ürünler, sunum kabiliyeti, bilgisayar hizmetleri, yazılım yetenekleri, proje bazlı, portal, tasarım, Müşteri Ziyaretleri	0,042
Konu 8	İş, deneyim, müşteri, yönetmek, takım, gelişim, satış, iletişim, hizmet, alan, yıl, yetenekler, bilgi, işletme, işlem, ofis, takip etmek, Destek, süreçler, yetenek	0,592
Konu 9	Finans, ulaştırma, deneyim, yazılım geliştirme, süreç geliştirme, marka değeri, reaktif, MS Office, müşteri, planlama	0,045
	Toplam	1,000

Tablo 8. Bilgi ve İletişim Sektörü 2022 Yılı Ekim, Kasım ve Aralık İlanları

Burada Konu 8 oldukça yüksek bir skorla (0,592) baskın konu olarak ön plana çıkmaktadır. Konu 8 ile Konu 4 de diğer konulara göre daha baskın görülmektedir. Konu 8' i temsil eden kelimler "iş", "deneyim", "müşteri", "yönetmek", "takım", "gelişim", "satış", "iletişim", "hizmet", "alan", "yıl", "yetenekler", "bilgi", "işletme", "işlem", "ofis", "takip etmek", "destek", "süreçler", "yetenek" şeklinde oluşmuştur. Konu 4 ise "yazılım", "askerlik

görevi", "MS Office", "üretken", "parasal stratejiler", "kalite", "uygulama," "fikir", "kart", "raporlama", "muhasebe" kelimeleri ile temsil edilmektedir.

Bilgi ve İletişim Sektörü Ocak, Şubat ve Mart 2023 İlanları

Tablo 9' da son üç aylık dönem olan 2023 yılı Ocak, Şubat ve Mart aylarına ait iş ilanlarının GDA analizi sonuçları yer almaktadır.

Konular	Kelimeler	Confidence
Konu 1	Kalite, MS Office, bilişim teknolojileri, esnek çalışma, veri tabanı, program, fikir, deneyim, kişisel gelişim, dijital çözümler	0,077
Konu 2	Lisans mezunu, takım çalışması, yazılım teknolojileri, dijital dönüşüm fikirleri, pazarlama analitiği, dijital gelecek, projeler, bilgisayar kullanımı, değişim hedefleri, enerji	0,036
Konu 3	Fonksiyonel destek, sistem yöneticisi, java software, yazılım, donanım, uygulama projeleri, etkili iletişim, TLBIE, Stres Yönetimi, kapsamlı kütüphane	0,036
Konu 4	İş, deneyim, yönetmek, müşteri, satış, yazılım, iletişim, hizmet, takım, işletme, bilgi, süreçler, gelişim, tecrübeli, yetenekler, alan, müşteriler, destek, ürünler, çözümler	0,501
Konu 5	Gelişmiş, eğitim fırsatları, iletişim, sipariş yönetimi, yazılım mimari, uygulama, uluslararası pazarlar, deneyim, IOT çözümleri	0,034
Konu 6	Güçlü işletme, enerji santrali, MS Office, inovasyon, etik, analistler, süreç yönetimi, Kosgeb, TÜBİTAK, dijital yerli	0,035
Konu 7	Sorumluluk, marka, deneyimli, dinamik personel, e ticaret, dijital operasyon, RD Design, Yönetim Danışmanlığı, Teknopark	0,035
Konu 8	Yıl, teknik proje, sistem yöneticisi, inovatif teknolojiler, mühendislik, veri, tasarım, raporlama, takım çalışmasına yatkın	0,208
Konu 9	Yaratıcı, sosyal medya, teknoloji market, matematik ve istatistik, değişim yönetimi, iş odaklı, süreç yönetimi, kitle iletişim, veri güvenliği, analitik düşünme	0,038
	Toplam	1,000

Tablo 9. Bilgi ve İletişim Sektörü 2023 Yılı Ocak, Şubat ve Mart Ayları İlanları

Burada da Konu 4 baskın konu olarak görülmektedir. Konu 4 (0,501), "iş", "deneyim", "yönetmek", "müşteri", "satış", "yazılım", "iletişim", "hizmet", "takım", "işletme", "bilgi", "süreçler", "gelişim", "tecrübeli", "yetenekler", "alan", "müşteriler", "destek", "ürünler", "çözümler" kelimeleri ile temsil edilmektedir. Ayrıca Konu 8 de Konu 4' ten sonra en yüksek confidence değerine (0,208) sahiptir ve bu konu başlığı altında "yıl", "teknik proje", "sistem yöneticisi", "inovatif teknolojiler", "mühendislik", "veri", "tasarım", "raporlama", "takım çalışmasına yatkın" gibi kelimeler yer almaktadır.

Bilgi ve İletişim Sektörü İş İlanlarının Üçer Aylık Dönemlerdeki Baskın Konular

Tablo 10' da her üç aylık verinin analizi sonucu elde edilen baskın konular bir arada bulunmaktadır.

Dönem	Konu Adı	Kelimeler	Confidence
1	**Satış Sonrası Destek**	MS Office, yazılım, satış sonrası, yetenek, uzaktan çalışma, teknoloji, destek, mühendislik, işletme, programlar	0,224
2	**Müşteri Hizmetleri**	İletişim, yazılım, işletme, hizmet, takım, çözümler, destek, müşteriler, sektör, tasarım	0,317
3	**Müşteri İlişkileri Yönetimi**	İş, deneyim, müşteri, yönetmek, takım, gelişim, satış, iletişim, hizmet, alan, yıl, yetenekler, bilgi, işletme, işlem, ofis, takip etmek, Destek, süreçler, yetenek	0,592
4	**Süreç Yönetimi**	İş, deneyim, yönetmek, müşteri, satış, yazılım, iletişim, hizmet, takım, işletme, bilgi, süreçler, gelişim, tecrübeli, yetenekler, alan, müşteriler, destek, ürünler, çözümler	0,501
1.Dönem: Nisan-Mayıs-Haziran 2022 2.Dönem: Temmuz-Ağustos-Eylül 2022		3.Dönem: Ekim-Kasım-Aralık 2022 4.Dönem: Ocak-Şubat-Mart 2023	

Tablo 10. Bilgi ve İletişim Sektörü İş İlanlarının Üçer Aylık Dönemlerdeki Baskın Konular

Görüldüğü üzere özellikle üç ve dördüncü üçer aylık dönemlerde çok yoğun benzerlikler mevcuttur. Buradan hareketle, "deneyim", "müşteri", "iletişim", "çözüm", "destek", "takip" gibi kelimelere bakıldığında her iki grup için konu başlıkları "Müşteri İlişkileri Yönetimi" şeklinde belirlenmiştir. Birinci ve ikinci üç aylık sonuçlara bakıldığında ise bazı farklılıklar görülmektedir. Birinci üç aylık dönemde "satış sonrası destek" gibi bir konu başlığı uygun görülmüştür. Buradan hareketle, İlanda belirtilen kelimelerden yola çıkılarak, 2022 yılının mart, nisan mayıs dönemlerinde satış sonrası destek elemanına için yayınlanan ilanların öne çıktığını söyleyebiliriz. İkinci üç aylık dönemde ise "müşteri hizmetleri" gibi bir konu başlığı uygun olabilir.

Bu bölümde hızla değişim ve gelişim gösteren Bilgi ve İletişim Sektörü için yayınlanan iş ilanlarının bir yıllık süreç içerisinde değişiklik gösterip göstermediğine baskın konular çerçevesinde bakılmıştır. Bariz farklılıklar görülmemekle beraber genel olarak Satış Sonrası Destek, Müşteri Hizmetleri, Müşteri İlişkileri Yönetimi baskın konu başlıkları olarak ön plana çıkmıştır. Özellikle üç ve dördüncü dönemlerde yaklaşık % 70 gibi bir benzerlik oranı bulunmaktadır. Fakat bu ve buna benzer analizlerin daha uzun süreleri kapsayacak şekillerde yapılması hem dönemsel değişiklikleri hem de zaman içerisindeki gelişmelerin iş ilanlarına yansımalarını tespit etmek için kullanışlı bir araç olabilir.

3.8.5. Bilgi ve İletişim Sektörü Lisans Mezunu İlanlarının GDA Analizleri

Çalışmada Bilgi ve İletişim Sektörü ilanları lisans-ön lisans ve tecrübeli tecrübesiz şeklinde ayrılıp her bir grup ayrı analiz edilmiş ve aralarındaki farklılıklar tespit edilmeye

çalışılmıştır. Tablo 11' de Bilgi ve İletişim Sektörü lisans mezunu ilanlarının GDA sonuçları yer almaktadır.

Konular	Kelimeler	Confidence
Konu 1	Karar verme, dinamik çevre, Web programlama, problem çözme, finansal planlar, bankalar, satış müdürü, ar - ge, otomotiv	0,033
Konu 2	Kalite yönetimi, uygulama geliştirme, VMWARE hipervisor, uluslararası şirket, müşteriler hissedarlar, mobil, sektörel deneyim, yapay zekâ	0,033
Konu 3	Aylık izleme, planlama, ilişkiler geliştirmek, makine, ihracat, sunucu yönetimi, telekomünikasyon, code version, sipariş, teknolojik çözümler	0,032
Konu 4	Saha Teknolojileri, Uzman Yardımcısı, Takım çalışması, Sunucu Yönetimi, operasyonel süreçler, Deneyim, müşteri İlişkiler, Pazar araştırması, veri analisti	0,033
Konu 5	Bilgi teknoloji, entegrasyon hizmeti, Telekom, bilişim, depo yönetimi, Yardım Projeleri, ürün, Katalog	0,033
Konu 6	Güvenlik, operasyon, deneyim, mimari, Satış Deneyimi, iletişim becerileri, redux, organizasyon, MS Office, Uyumlu Tasarımlar	0,035
Konu 7	Web sayfa, prosedür, aktif araç, veri analiz, agile scrum metodolojileri, organizasyon, sosyal platformlar, teknik beceriler, mühendisler, İçerik yönetimi, iş akışları, kalite Desteği	0,033
Konu 8	Performans, Tedarik, Hizmet Denetimi, takım çalışması, Verileri Bağlamak, Sistem, Satın Alma, veri toplamak, Bildirimler, takip sistemi	0,033
Konu 9	İş, deneyim, yazılım, takım, gelişim, süreçler, şirket, müşteri, iletişim, yetenekler, hizmet, bilgi, işlem, satış, bilgi	0,735
	Toplam	1,000

Tablo 11. Bilgi ve İletişim Sektörü Lisans Mezunu İlanları

Tablo 11' de yer alan sonuçlara bakıldığında Konu 9 (0,735) çok yüksek bir confidence değeri ile baskın konu olarak görülmektedir. Konu 9' un altında yer alan kelimeler ise daha çok hassas becerilere işaret etmektedir. Hassas beceriler, geleneksel anlamda beceri veya yetenekler olarak görülmeyen fakat teknik becerilerin etkisini artıran kişilerarası iletişim becerilerinin ve kişisel niteliklerin bir kombinasyonudur (Hirudayaraj vd., 2021: 2). Fakat diğer konular altında daha teknik becerileri işaret eden kelimeler görülmektedir. Teknik detayların daha fazla olduğu ve daha fazla teknik becerilerin talep edildiği iş ilanlarının sayısının daha az olması ve analizlerde baskın konuların gerisinde kalması günümüzde özellikle belirli sektörlerde insan kaynağı ihtiyacının nicelik olarak daha az fakat nitelik olarak daha yüksek olmasına işaret etmektedir.

Burada da "WEB programlama", "finansal planlar", "satış müdürü", "ar-ge", "VMWARE HIPERVISOR", "uluslararası şirket", "yapay zekâ", "sunucu yönetimi", "code version", "uzman yardımcısı", "sunucu yönetimi", "veri analisti", "redux", "veri analiz", "agile scrum metodolojileri", "sosyal platformlar", "teknik beceriler",

“mühendisler”, “sistem”, “takip sistemi” gibi analiz sonucu ön plana çıkan kelime ve kelime grupları bu sektör için gerekli teknik beceri göstermektedir.

Tablo 11’ de yer alan analiz sonuçlarına göre ortaya çıkan konular için konu başlıkları oluşturmak yerine uzman görüşüne dayalı olarak iş tanımları yapılmıştır. İş tanımları öncelikle baskın konu için yapılmış ayrıca örnek olarak bir konu daha seçilerek iş tanımı yapılmıştır. Buna göre baskın konu olan Konu 9 “yazılım konusunda deneyimli, takım çalışmasına yatkın, gelişime açık, iletişim yetenekleri güçlü müşteri ilişkileri yönetimi personeli” şeklinde bir iş tanımını temsil edebilir.

Konu 3 için ise “Otomasyon yoğun üretim faaliyetlerinde makinalar arası veri transferi süreçlerini yönetebilecek, sunucu yönetiminde deneyimli, teknoloji çözüm uzmanı” şeklinde bir iş tanımı uygun olabilir.

3.8.6. Bilgi ve İletişim Sektörü Ön Lisans Mezunu İlanlarının GDA Analizleri

Bu bölümde de Bilgi ve İletişim Sektörü ön lisans mezunları için yayınlanan ilanlara dair analiz sonuçları yer almaktadır. Sonuçlar Tablo 12’ de görülmektedir.

Konular	Kelimeler	Confidence
Konu 1	Küresel, sorun çözme, uluslararası pazarlar, diksiyon, içerik üretimi, yayıncılık, Hollanda, tercüman yeminli, tasarım kataloğu	0,031
Konu 2	Süreç yönetimi, Nitelikli Üniversite, muhasebe departmanı, medya kanalları, Ek gelir, CSS arayüzü, M kodları, Powerpoint, disiplin	0,031
Konu 3	Hepsiburada, çizimler, askerlik görevi, Ethernet, SAP, etkili iletişim, Yönetim İş birliği, Takım çalışması, Jengal takımı, VPN bağlantısı, perakende zincirleri, Destek Uzmanları	0,031
Konu 4	Organizasyon yeteneği, web desteği, ilgili okullar, raporlama yeterlilikleri, teknik sertifika, MS Office, uzman, sigorta, doğrudan sektör, İstanbul	0,030
Konu 5	Dergi, bordro işlemleri, dünya çağında, ürün, Excel kullanımı, Memnuniyet Takip, Ürün İşlem, Sorun iletişimi, üretim süreçleri, Word	0,029
Konu 6	Yoğun tempo, müşteri ilişkileri yönetimi, sipariş almak, kurulum eğitimi, sorumluluk, hazırlık raporları, sistemler nesne, ERP müşteri, lojistik, kullanıcı sorunları	0,032
Konu 7	Saha çalışması, analitik düşünme, veri, Fransızca, bağımsız çözümler, Euromessage, Visilabs, internet altyapısı, İstanbul, Antalya	0,030
Konu 8	İş, müşteri, hizmet, şirket, satış, MS Office, deneyim, takım çalışması, veri, iletişim, Destek, askerlik görevi	0,697
Konu 9	Esnek saatler, öğrenme, rapor, problem çözme, görünüm, takım arkadaşı, network, outsourcing, ihracat, müşteri şikayetleri	0,090
	Toplam	1,000

Tablo 12. Bilgi ve İletişim Sektörü Önlisans Mezunu İlanları

Konu 8 (0,697) baskın olarak ön plana çıkmıştır. Lisans mezunları ilanlarındaki gibi benzer kelimeler burada da ön plandadır. Baskın konu olan Konu 8' in altında "müşteri", "hizmet", "şirket", "satış", "MS Office", "deneyim", "takım çalışması", "veri", "iletişim", "destek", "askerlik görevi" gibi kelimeler yer almaktadır. Lisans mezunu ilanlarında olduğu gibi benzer kelimeler burada da ön plana çıkmıştır. Dolayısıyla ister lisans ister ön lisans mezunu olsun Bilgi ve İletişim Sektörü için yayınlanacak ilanların burada yer alan hassas becerileri içermesi beklenebilir.

Ayrıca "içerik üretimi", "tasarım", "CSS arayüzü", "M kodları", "SAP", "etkili iletişim", "JENGAL takımı", "VPN bağlantısı", "Destek Uzmanları", "web desteği", "İstanbul", "ERP müşteri", "lojistik", "veri", "Euromessage", "Visilabs", "internet altyapısı", "network", "ihracat", "müşteri şikayetleri" gibi kelimeler de ön lisans mezunlarından beklenen teknik becerileri ifade etmektedir.

Burada baskın konu olan Konu 8 için "hizmet sektöründe faaliyet gösteren bir şirkette, Ofis programlarına hâkim ve veri girişi yapabilecek, deneyimli, iletişim yetenekleri güçlü, erkek adaylar için askerlik görevini tamamlamış, destek personeli" şeklinde bir iş tanımı yapılabilir.

Konu 5 için ise "Ürün satışı ve satış sonrası süreçlerin takibini yapabilecek, Word excel programlarını kullanabilen, ön muhasebe ve basit veri girişi bilgisi olan ofis ortamında çalışacak lise veya ön lisans mezunu" şeklinde bir iş tanımı özellikle önlisans veya lise mezunu iş ilanları için uygun olabilir.

Sonuç olarak lisans veya ön lisans mezunları için yayınlanan ilanlar genel olarak mezun olunan okul veya eğitim seviyesinden ziyade yetkinlik odaklı bir yapıya sahiptir. Problem çözme ve belirli teknik becerilere sahip olmak çoğu ilanda daha fazla önem arz etmektedir.

3.8.7. Bilgi ve İletişim Sektörü Tecrübeli Adayların İlanlarının GDA Analizleri

Bilgi ve İletişim sektörü için bir yıldan fazla tecrübe aranan adaylar için yayınlanan ilanların analizi Tablo 13' de yer almaktadır.

Konular	Kelimeler	Confidence
Konu 1	Görsel Tasarım, bilgi yazılım, medya platformları, para, deneyim, mobil cihazlar, lisansüstü, mağaza, Matematik Mühendisliği, sipariş, verimlilik	0,030
Konu 2	Ulusal Gizlilik, analiz, İşletme Mezun, Yapılan Projeler, Satış Geliştirme, bilgi ağı, SDLC işlemi, Cloud data, bakım geliştirme	0,030
Konu 3	Yazılım, iletişim, projeler, çözüm, tasarım, askerlik görevi, sistem, destek, müşteri odaklı	0,322
Konu 4	Fotoğraf, iletişim yeteneği, Raporlama, koordinasyon, Sistem, Strateji geliştirme Visilab, Üniversiteler, takım çalışması	0,031
Konu 5	Python programlama, veri analizi, Müşteriler Teknik, üretim işletmeleri, Raporlama Performansı, Güçlü İletişim	0,030
Konu 6	İş öğrenmek, Yazılım programları, Webpack NPM, Sürekli güncelleme, lisansüstü, destek, Üretim, VPN IPSEC, çalışma saatleri	0,031
Konu 7	Uygulamalar, Süreçler, konaklama, İkna yetenekleri, teknik çizim, danışmanlık, ücret, finansal araçlar, İşlemler	0,031
Konu 8	Deneyim, müşteri, hizmet, takım, bilgi, yönetim, süreçler, MS Office, satış, işlem	0,466
Konu 9	Takım, Projeler, yönetim, elektrik Mühendisliği, lisans, endüstri, Solidworks, MS Office, proaktif, iletişim	0,030
	Toplam	1,000

Tablo 13. Bilgi ve İletişim Sektörü Tecrübeli Adayların İlanları

Bu sonuçlara göre Konu 8 (0,466) baskın konu olarak görülmektedir. Konu 8, “deneyim”, “müşteri”, “hizmet”, “takım”, “bilgi”, “yönetim”, “süreçler”, “MS Office”, “satış”, “işlem” kelimeleri ile temsil edilmektedir. Buradan hareketle müşteri hizmetleri, satış operasyonları veya tecrübeye dayalı olarak yönetim pozisyonları Konu 8 ile ilişkilendirilebilir.

Ayrıca Konu 8, “hizmet süreçlerinde deneyimli, takım çalışmasına yatkın, ofis programlarına hâkim, müşteri ilişkileri yöneticisi veya satış yöneticisi” şeklinde bir iş tanımı da yapılabilir.

Konu 5 için ise “Üretim süreçlerindeki verileri ve müşterilerin talep veya şikayetlerini değerlendirmeye alıp analiz edebilecek ve elde edilen bulguları raporlayabilecek Python programlama dili kullanabilen veri analizi uzmanı” gibi bir iş tanımı uygundur.

İlanlarda her ne kadar 1 yıldan fazla tecrübe etiketleri olsa da sektörel olarak tecrübenin yanında teknik becerilerin de son derece önemli yer tuttuğunu söylemek mümkündür. Çünkü analiz sonuçlarında iki konuda “deneyim” kelimesi anahtar kelime olarak görünmektedir. Bu iki konu da baskın konu durumunda değildir.

3.8.8. Bilgi ve İletişim Sektörü Tecrübesiz Adayların İlanlarının GDA Analizleri

Bu bölümde ise Bilgi ve İletişim Sektörü için yayınlanan bir yıldan az tecrübeye sahip aday ilanlarının analiz sonuçları yer almaktadır. Sonuçlar Tablo 14' te yer almaktadır.

Konular	Kelimeler	Confidence
Konu 1	Yenilik, Deutsch, Sermaye QIA, finansal teknoloji, Call Center, Personel ile ilgili, COPC, saha müşterisi, eğitim bursu, Platform	0,037
Konu 2	Müşteri, takım, hizmet, şirket, iletişim, satış, işletme, yetenekler, yönetmek, MS Office, yazılım geliştirme, saha, çözüm	0,692
Konu 3	Gelecek Teknolojiler, Lisans Hizmetleri, Depo, Hizmet, programlar, marka, mevzuat, dil becerileri, Yeterlilik belgesi, şirket politikaları	0,040
Konu 4	Arama motoru, askerlik tamamlandı, EINES, deneyim, dinamik, Destek ekipleri, kredi kartı, Seyahat Engelleri	0,039
Konu 5	Planlama, organizasyon, Endüstri, kişisel bilgisayarlar, perakendecilik, akıllı şehir, mühendislik, Dağıtım, zaman yönetimi, Veri, tedarik, kariyer yönetimi	0,038
Konu 6	Dinamik enerjik, rehber, Küresel Şirketler, Üniversiteler, Yıllık Satış, Takım başarı, Deneyim, spor ile ilgili, oryantasyon Programı, dijital düşünce	0,038
Konu 7	Süreç yönetimi, Sunucu Çalışma, İyileştirme, altyapı hizmetleri, e export, dönemsel, dijitalleşme endüstrisi, Ulaşım, Teknik	0,038
Konu 8	Üniversiteler, endüstriyel, odak, kontrol, information network, gelecekteki müşteri, core c	0,039
Konu 9	Hizmet işletmeleri, üniversiteler, askeri görev, SANLAB, alanında deneyimli, mezuniyet derecesi, hizmetler alanı, teknik yönlendirme, ödeme hizmetleri	0,039
	Toplam	1,000

Tablo 14. Bilgi ve İletişim Sektörü Tecrübesiz Adayların İlanları

Burada yer alan tecrübesiz etiketi bir yıldan az tecrübeye sahip adaylar için kullanılmıştır. Analiz sonuçlarına göre Konu 2 (0,692) baskın konu olarak ortaya çıkmıştır. Konu 2 için anahtar kelimeler ise "müşteri", "takım", "hizmet", "şirket", "iletişim", "satış", "işletme", "yetenekler", "yönetmek", "MS Office", "yazılım geliştirme", "saha", "çözüm" şeklinde oluşmuştur. Bu kelimelerden hareketle yazılım ekibi, takım çalışması, saha çalışması gibi etiketler bu konu için uygun olabilir.

Ayrıca baskın konu olan Konu 2 için "saha satış tecrübesi olan, takım çalışmasına yatkın, ofis programlarına hâkim, müşteri hizmetleri yetkilisi" şeklinde de bir iş tanımı yapılabilir.

Diğer seçilen konu olan Konu 5 için ise "Teknoloji firması için ürün tedariği ve satış süreçlerinin planlanmasında ve organizasyonunda görev alacak, veri ve zaman yönetimi konusunda uzman endüstri mühendisi" uygun bir iş tanımı olabilir.

Diğer konulara bakıldığında ise confidence değeri düşük olmasına rağmen teknik becerileri temsil eden detaylar, bazı önemli uluslararası firmalar ve çeşitli hassas beceriler göze çarpmaktadır.

Bilgi ve İletişim Sektörü özelinde tecrübeli ve tecrübesiz adaylar için yayınlanan ilanların analiz sonuçlarına bakıldığında bariz farklılıklar görülmemekle beraber sektörün beklediği teknik ve hassas becerilerin ön plana çıktığı görülmektedir. Donanım, yazılım, web programlama gibi teknik unsurlar teknik beceriler olarak adlandırılırken, iletişim öğeleri, takım projelerinde işbirliği yapma yeteneği gibi adayların işlerine yönelik tutum ve yaklaşımlarını içeren bereciler ise hassas beceriler olarak ifade edilir (Patacsil ve Tablatin, 2017: 350).

3.8.9. Bilgi ve İletişim Sektörü İlanlarının GDA Analizleri

Bu bölümde 07.04.2022-18.03.2023 tarih aralığında yayınlanan Bilgi ve İletişim Sektörü ilanlarının tamamı için yapılan GDA analizi sonuçları bulunmaktadır. Tablo 15'te 9 konu ve her konu için on kelime içeren analiz sonuçlarını göstermektedir.

Konular	Kelimeler	Confidence
Konu 1	Mühendislik fakültesi, yönetim, devreye alma, MS Office, Redux Toolkit, Google, Web, Access, ürün geliştirme, planlama	0,029
Konu 2	Elektronik mühendisliği, Big data uygulaması, code reviews, Python, tercihen deneyimli, ürün hattı, tekstil endüstrisi, video içeriği, lisans mezunu, fabrika otomasyonu	0,029
Konu 3	Sistemler, json, planlama, ulusal gizlilik, web sitesi, verimlilik, hızlı düşünmek, reklam, stratejik danışmanlık, araştırma geliştirme	0,029
Konu 4	Medya Stratejileri, veri departmanları, girişim, yazılım, hazırlık tasarımı, sipariş, iletişim, akademik çalışmalar, satış operasyonları, uygulamalar, kalite sistemi	0,028
Konu 5	Yönetim danışmanlık, aktivite, genç profesyonel, muhasebe, dijital, bilgisayar, sorumluluk sahibi, veri görselleştirme, hedefler, kullanıcılar	0,032
Konu 6	Uyumlu, ürünler, kontrol otomasyonu, malzeme Scrum, CRM programı, Test uygulaması, İstanbul, WEVE, otomasyon tabanlı, kapasite kulanım, erkek	0,028
Konu 7	İş, deneyim, hizmet, yönetmek, işletme, satış, müşteri, iletişim, MS Office, işlem, çözüm	0,494
Konu 8	Takım çalışması, veri, destek, proje, yazılım, teknolojiler, teknik, kalite, askerlik tamamlanmış, parça, İstanbul	0,304
Konu 9	Ek iş, teknik taşıma, ödeme prosedürleri, program, Mobil Deneyimler, Teknoloji Uzmanlığı, Raporlama, plan düzenlemek, kendini geliştirmek	0,028
	Toplam	1,000

Tablo 15. Bilgi ve İletişim Sektörü İlanları

Tablo 15’ te görüldüğü üzere baskın konular Konu 7 (0,494) ve Konu 8 (0,304) şeklindedir. Konu 7 altında çıkan kelimeler “iş”, “deneyim”, “hizmet”, “yönetmek”, “işletme”, “satış”, “müşteri”, “iletişim”, “MS Office”, “işlem”, “çözüm” şeklindedir. Konu 8 altında ise “takım çalışması”, “veri”, “destek”, “proje”, “yazılım”, “teknolojiler”, “teknik”, “kalite”, “askerlik tamamlanmış”, “parça”, “İstanbul” kelimeleri yer almaktadır. Bu iki konu değerlendirildiğinde Konu 8’ de daha çok teknik becerilerin ön planda olduğunu Konu 7’ de ise hassas becerilerin daha ön planda olduğunu söylemek mümkündür. Bu bağlamda Bilgi ve İletişim Sektörü için de hassas becerilerin önemini vurgulamak gerekir.

Şekil 26’ da, Tablo 15’te yer alan analiz sonuçlarının kelime bulutu şeklinde gösterimi yer almaktadır.

Şekil 26. Bilgi ve İletişim Sektörü İlanlarının GDA Analizleri Kelime Bulutu Gösterimi

Şekilde görülen kelimeler büyüklüklerine göre daha büyük olanın kullanım sıklığı ve ağırlığı daha fazla daha küçük olanların ise daha az şeklinde yorumlanabilir. Kelime bulutu gösterimleri çeşitli analiz sonuçlarının görselleştirilmesi adına kullanışlı bir araç olabilir.

SONUÇ

Büyük veri yığınlarının barındırdığı kullanışlı bilgilerin ortaya çıkarılması, günümüz iş dünyasının problemlerinden biridir. Geleneksel yöntemlerle çözülmesi zor olan problemlerin veri veya metin madenciliği yöntemleri kullanılarak daha hızlı ve etkili bir şekilde çözüme ulaştırılması mümkündür. Bilgisayar destekli bir çözümleme süreci olarak bakıldığında veri ve metin madenciliği sağlık, eğitim, finans, çevre, pazarlama gibi birçok alanda kullanılmaktadır. Veri madenciliği, matematik, istatistik, bilgisayar bilimleri gibi farklı alanları kapsayan multidisipliner bir yapıya sahiptir. Aynı şekilde metin madenciliği de bahsi geçen alanlarla ile dil bilimi ve morfoloji ile de yakından ilgilidir. Veriye değer veren ve modern sistemlerle veri analizine kaynak ayıran işletmeler yoğun rekabet ortamında avantajlı konuma ulaşmaları kaçınılmaz olacaktır.

Veri ve metin madenciliğinin veri yığınlarını değerli hale getirme, büyük veri ile ilgili işlemleri daha hızlı bir şekilde yapma, esnek veri kullanımı, kaynakların daha verimli kullanımı gibi avantajları vardır. Bununla birlikte veri ve metin madenciliği yöntemlerinde üstesinden gelinmesi gereken bazı zorluklar da vardır. Alan bilgisi entegrasyonu, değişen kavramların ayrıntı düzeyi, birden fazla lisan içeren metin iyileştirme ve doğal dil işlemedeki belirsizlikler, metin madenciliği sürecinde ortaya çıkan başlıca sorunlar ve zorluklardır (Talib vd. 2016: 417). Veri madenciliğinde ise kirli verilerin çokluğu, sağlıksız veri depolama, kullanılacak algoritmanın seçimi ve süre kısıtları gibi zorluklar mevcuttur. Bu gibi zorluklar özellikle veri tabanı yönetimi, modern sistemlerin kullanılması ve detaylı ön işlemler ile aşılabilir. Bu çalışmada da doğru algoritma seçimi için detaylı literatür taraması yapılmış ve özellikle analizlerin sağlıklı yapılabilmesi için de detaylı bir veri ön işleme sürecinden geçilmiştir.

Günümüzde iş gören adayları iş ararlarken bazı zorluklarla karşılaşmaktadır. Eğitim hayatını tamamlamış iş gören adayları istedikleri imkanları sunan işler bulamazken işverenler de aradıkları niteliklerde iş görenler bulamamaktadır. İş gören adaylarının, kariyer planlarını yaparlarken dikkate almaları gereken konulardan birisi sektörel beklentilerin hangi yönde olduğunun belirlenmesi ve eğitim öğretim programlarını buna göre planlaması gerekmektedir. Aynı zamanda eğitim öğretim kurumlarının da teknolojiye bağlı olarak hızla değişen sektörel beklentileri karşılayacak programlar hazırlaması gerekmektedir. Bu şekilde sektörün beklentileri doğrultusunda işgücünün yetişmesi daha kolay olabilir. Ayrıca iş gören adayları da beceri ve

yetkinliklerini sektör beklentileri doğrultusunda geliştirip daha sağlıklı bir kariyer planı oluşturabilirler.

Son yıllarda internet kullanımın artması ile işletmeler hem mavi yaka hem de beyaz yaka personel ihtiyaçları için kariyer sitelerini daha sık kullanmaya başlamışlardır. Kariyer sitelerinde yayınlanan iş ilanları işveren ve iş gören adaylarının buluşması açısından önemli bir araçtır. Binlerce işletme ile milyonlarca iş gören adayı bu ortamda hızlı bir şekilde buluşmaktadırlar. Dolayısıyla bu alanlarda yayınlanan çevrim içi iş ilanları da büyük bir veri yığını oluşturmaktadır. Oluşan bu büyük veri içerisinde sektörel beklentilere ilişkin önemli bilgiler yer almaktadır. Bunun için veri ve metin madenciliği yöntemlerinin kullanılması anlamlı ve kullanışlı bilgilerin açığa çıkarılmasını sağlar.

Bu çalışmada, gelişen teknoloji karşısında Bilgi ve İletişim Sektörünün iş gören adaylarından beklentilerinin hangi yönde oluştuğunun tespit edilmesi amaçlanmıştır. Aynı zamanda yapılan çalışmalar iş gören adayları ve eğitim öğretim planlaması için de önemli bilgiler içermektedir. Çalışmada Kariyer.net web sitesinde yayınlanan iş ilanlardan yararlanılmıştır. İlanlar web kazıma yöntemi ile Kariyer.net sitesinden ".xlsx" uzantılı bir dosya şeklinde alınmış ve MS Excel programında ön işlemlere tabi tutulmuştur. İş ilanında yer alan bilgiler çerçevesinde oluşturulan veri seti "firma adı, il, eğitim durumu, tecrübe, iş tanımı, sektör, çalışma şekli, pozisyon seviyesi, departman, lisan, ilan linki" değişkenlerinden oluşmuştur. Sektörel olarak verilerin etiketlenmesi sürecinde TUİK Sınıflama Sunucusu NACE rev.2- Altılı Ekonomik Faaliyet Sınıflaması kodları esas alınmıştır. Veriler, aynı satırların silinmesi, değerlerin küçük harflere dönüştürülmesi, değişkenlerin etiketlenmesi, noktalama işaretlerinin silinmesi, sık tekrar eden ve tek başına anlam ifade etmeyen kelimelerin silinmesi gibi detaylı ön işlemeye tabi tutulmuştur.

Çalışmada, veri ve metin madenciliği yöntemleri kullanılmıştır. Analizler RapidMiner Studio 10.1 programı kullanılarak yapılmıştır. Bu aşamada Random Forest algoritmasından yararlanılarak Karar Ağacı Analizi yapılmıştır. Pozisyon seviyesi ve departman değişkenleri ele alınmıştır.

Pozisyon seviyesi değişkenine etki eden en önemli faktörler lisan ve tecrübedir. Bu da işletmelerde daha yüksek pozisyonlara yapılan atamalarda tecrübe ve bilinen yabancı dilin, eğitim ve diğer değişkenlere nazaran daha önemli olduğunu ortaya

koymaktadır. Ayrıca karar ağacı analizi sonuçlarına göre belirli kurallar elde edilmiştir. Buna göre, bilgi ve iletişim sektörü için "üretim departmanında, lisans mezunu, 1-5 yıl deneyimli, İngilizce bilen uzman çalışma ihtimali %13,19' dur", "pazarlama departmanında, lise mezunu, 1 yıldan az deneyimli, eleman çalışma olasılığı %5,18' dir" ve "idari işler departmanında, lise mezunu, 1-5 yıl deneyimli, tam zamanlı eleman çalışma olasılığı %5,06' dır" şeklinde kurallar elde edilmiştir. Departman seçimine ise en fazla etki eden değişken ise tecrübe, lisan ve pozisyon seviyesi şeklindedir. Buna göre bilgi ve iletişim sektörü için "lisans mezunu, 1 yıldan az deneyimli İngilizce bilen, tam zamanlı uzman yardımcısının pazarlama departmanında çalışma olasılığı %5,81' dir", "lise mezunu, 1-5 yıl deneyimli, tam zamanlı uzmanın üretim departmanında çalışma olasılığı %3,89' dur" şeklinde yüksek olasılıklı örüntülere ulaşılmıştır. Bununla birlikte "lisansüstü mezunu, 1-5 yıl deneyimli, İngilizce bilen, tam zamanlı üst düzey yöneticinin üretim departmanında çalışma olasılığı yüzde birin altındadır" şeklinde düşük olasılıklı örüntüler de elde edilmiştir. Bunun sebebi ise üst düzey yönetici için yayınlanan ilanların genel iş ilanları içerisinde çok az sayıda olmasıdır.

Çalışmada yer alan diğer analizler ise GDA yöntemi ile yapılan Konu Modelleme analizleridir. Bunun için öncelikle GDA analizi için önemli bir parametre olan konu sayısı belirlenmiştir. Konu sayısının belirlenmesi için kurulan model 3 konu sayısından 15 konu sayısına kadar iteratif bir şekilde çalıştırılmış ve 9 sayısında optimum tutarlılık ve çarpaşıklık değerlerine ulaşılmıştır. Konu sayısı belirlendikten sonra analizler için öncelikle bir yıllık süreyi kapsayan veri seti üçer aylık dönemlere ayrılmış ve her üç aylık dönem ayrı ayrı analiz edilmiştir. İlk üç aylık dönemde Konu 9, baskın konu olarak görülmektedir. Konu 9 altında çıkan kelimeler ise "MS Office, yazılım, satış sonrası, yetenek, uzaktan çalışma, teknoloji, destek, mühendislik, işletme, programlar" şeklindedir. Bu kelimelere göre Konu 9 için "satış sonrası destek" şeklinde bir konu başlığı uygun görülmüştür. İkinci üç aylık dönemde Konu 8 baskın konu olarak belirlenmiştir. Konu 8, "iletişim, yazılım, işletme, hizmet, takım, çözümler, destek, müşteriler, sektör, tasarım" kelimeleri ile temsil edilmektedir ve bu kelimelere göre "müşteri hizmetleri" şeklinde isimlendirilmiştir. Üçüncü üç aylık dönemde de "İş, deneyim, müşteri, yönetmek, takım, gelişim, satış, iletişim, hizmet, alan, yıl, yetenekler, bilgi, işletme, işlem, ofis, takip etmek, destek, süreçler, yetenek" kelimeleri ile temsil edilen Konu 8 baskın konudur. Buna göre üçüncü üç aylık dönemdeki baskın konu da "müşteri ilişkileri yönetimi" şeklinde isimlendirilmiştir. Son olarak dördüncü üç aylık

dönemde Konu 4 baskın konu olarak görülmektedir. Konu 4 için "iş, deneyim, yönetmek, müşteri, satış, yazılım, iletişim, hizmet, takım, işletme, bilgi, süreçler, gelişim, tecrübeli, yetenekler, alan, müşteriler, destek, ürünler, çözümler" kelimeleri analiz sonucunda çıkmış ve "süreç yönetimi" olarak adlandırılmıştır.

Analizlerin devamında, Bilgi ve İletişim Sektörü ilanları lisans-ön lisans ve tecrübeli tecrübesiz şeklinde ayrılıp her bir grup ayrı analiz edilmiştir. Buna göre her bir analiz sonucunda ortaya çıkan konular için uzman görüşüne başvurularak iş tanımları oluşturulmuştur. Lisans mezunları Bilgi ve İletişim Sektörü ilanları için baskın konuyu temsil eden kelimelerden "yazılım konusunda deneyimli, takım çalışmasına yatkın, gelişime açık, iletişim yetenekleri güçlü müşteri ilişkileri yönetimi personeli" şeklinde bir iş tanımı yapılmıştır. Bilgi ve iletişim sektörü için yayınlanacak ilanların bu tanıma benzer şekilde gelme olasılığı 0,735 olarak belirlenmiştir. Ön lisans mezunları için yayınlanan iş ilanlarının analiz sonuçlarına göre baskın konu "hizmet sektöründe faaliyet gösteren bir şirkette, ofis programlarına hâkim ve veri girişi yapabilecek, deneyimli, iletişim yetenekleri güçlü, erkek adaylar için askerlik görevini tamamlamış, destek personeli" şeklinde tanımlanmıştır. Buna göre ön lisans mezunları için yayınlanacak iş ilanlarının baskın konuda yer alan iş tanımı doğrultusunda yayınlanma olasılığı 0,697 olarak belirlenmiştir.

GDA analizi, veri setinde yer alan tecrübeli adaylara (bir yıldan daha uzun süre deneyime sahip adaylar) yönelik iş ilanları için yapılmıştır. Buna göre "deneyim, müşteri, hizmet, takım, bilgi, yönetim, süreçler, MS Office, satış, işlem" kelimeleri baskın konuyu temsil eden kelimeler olarak görülmektedir. Bu kelimelere istinaden "yazılım konusunda deneyimli, takım çalışmasına yatkın, gelişime açık, iletişim yetenekleri güçlü müşteri ilişkileri yönetimi personeli" şeklinde bir iş tanımı tecrübeli adaylar için yapılmıştır. Bu iş tanımına uygun iş ilanların yayınlanma olasılığı 0,466 olarak belirlenmiştir. Tecrübesiz adaylar (1 yıldan daha az deneyime sahip adaylar) için yapılan GDA analizi sonuçlarında ise baskın konu için anahtar kelimeler, "müşteri, takım, hizmet, şirket, iletişim, satış, işletme, yetenekler, yönetmek, MS Office, yazılım geliştirme, saha, çözüm" şeklinde oluşmuştur. Bu kelimelere istinaden yazılım ekibi, takım çalışması, saha çalışması gibi etiketler bu konu için uygun olabilir. Ayrıca baskın konu için "saha satış tecrübesi olan, takım çalışmasına yatkın, ofis programlarına hâkim, müşteri hizmetleri yetkilisi" şeklinde de bir iş tanımı yapılmıştır.

Bir yılı kapsayan Bilgi ve İletişim Sektörü iş ilanlarının tamamı için de GDA analizi yapılmıştır. Sonuçlara bakıldığında dokuz konu içerisinden iki konunun en yüksek confidence değerlerine sahip olduğu görülmüş ve her iki konu da baskın konu olarak değerlendirilmiştir. İlk baskın konu olan Konu 7 için anahtar kelimeler "iş, deneyim, hizmet, yönetmek, işletme, satış, müşteri, iletişim, MS Office, işlem, çözüm" şeklindedir. İkinci baskın konu olan Konu 8 için ise baskın kelimeler "takım çalışması, veri, destek, proje, yazılım, teknolojiler, teknik, kalite, askerlik tamamlanmış, parça, İstanbul" şeklinde oluşmuştur. Bu iki baskın konu değerlendirildiğinde Konu 8' de daha çok hassas becerilerin ön planda olduğunu, Konu 7' de ise teknik becerilerin daha ön planda olduğunu görmekteyiz. Bu bağlamda teknik becerilerin daha önemli düşünülen Bilgi ve İletişim Sektörü için de hassas becerilerin önemini vurgulamak gerekir. Bu sektörde iş arayan adayların teknik becerilerin yanında hassas becerilerini de geliştirmeleri gerekmektedir.

Bu çalışma hem farklı beklentileri olan işverenler hem de kariyerini planlayan iş gören adayları için önem taşımaktadır. İşverenler açısından büyük veri, veri ve metin madenciliği yöntemleri ile anlamlı hale getirilmiştir. Kariyerini planlayanlar için iş ilanlarında iş görenden beklentiler ve gençlerin kendilerini geliştirmesi gereken alanların ortaya konulması açısından önemli bir çalışma olduğu söylenebilir. İş ilanlarında sık olmayıp nadir görülen özelliklerden yola çıkılarak gelecekte sektörde aranacak özellikler çıkarılabilir. Eğitimciler açısından da sektörün beklentilerini karşılayacak programların (derslerin) ortaya konularak gençler geleceğe hazırlanması sağlanabilecektir.

Gelecekte yapılacak araştırmalarda, farklı veri ve metin madenciliği yöntemleri kullanılarak sonuçlar ortaya konulabilir. Bu yöntemlerle farklı sektörler ele alınıp iş ilanları incelenebilir. Ayrıca farklı kariyer sitelerinde yer alan iş ilanları da analiz edilip karşılaştırmalı bir araştırma yapılabilir. Hatta araştırmaya yurt dışında yayınlanan iş ilanları da dahil edilip ülkeler iş ilanları açısından karşılaştırılabilir. Kamu ve özel sektör iş ilanları analiz edilip karşılaştırılabilir. İş ilanları uzun dönemde incelenerek zaman boyutu da araştırmaya dahil edilip değişim trendleri gözlemlenebilir.

KAYNAKLAR

Abdalla, S. ve Angın, Z. (2022). "Angela Carter'ın Reflections Adlı Kısa Hikâyesinin Metin Madenciliği ile Analizi ve Simgesel Bağlamda Değerlendirilmesi.", Sosyal, Beşerî ve İdari Bilimler Dergisi, 5/6, 696-707.

Abu Hamde, M. (2018). "Kurumsal Belgelere (Metin Verilerine) Metin Madenciliği Tekniği ile Erişimin Değerlendirilmesi: Türk Özel Sektörüne Yönelik Bir İnceleme.", Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Abu Saa, A., Al-Emran, M. ve Shaalan, K. (2019). "Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques.", Technology, Knowledge and Learning, 24/4, 567-598.

Aghalarova, S. ve Keser, S. B. (2021). "Önerilen Yapay Sinir Ağı Algoritması ile Ortaokul Öğrencilerin Akademik Performansının Tahmini.", Veri Bilimi, 4/2, 19-32.

Agrawal, A., Fu, W., ve Menzies, T. (2018). "What is wrong with topic modeling? And how to fix it using search-based software engineering.", Information and Software Technology, 98/1, 74-88.

Agrawal, R., ve Srikant, R. (1994). "Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases.", VLDB, 1215/1, 487-499.

Agrawal, R., Imieliński, T., ve Swami, A. (1993). "Mining association rules between sets of items in large databases.", In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 207-216.

Agrawal, S. (2013). "Data mining: Data mining concepts and techniques.", In 2013 international conference on machine intelligence and research advancement, IEEE, 203-207.

Ağdeniz, Ş. (2017). "Finansal raporların analizinde metin madenciliğinin kullanımı: Borsa İstanbul şirketlerinin kurumsal yönetim niteliklerinin tahmini.", Doktora Tezi, Osmangazi Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.

Ağdeniz, Ş. ve Yıldız, B. (2018). "Muhasebede Analiz Yöntemi Olarak Metin Madenciliği.", Muhasebe Bilim Dünyası Dergisi, 20/2, 286-315.

Akbulut, M. (2022). “E-Perakende Firmalarının Web Kazıma Yöntemiyle Veri Analizi.”, 26.Pazarlama Kongresi Bildiri Kitabı, Kırşehir, 452- 467.

Akçetin, E., ve Çelik, U. (2014). “İstenmeyen elektronik posta (spam) tespitinde karar ağacı algoritmalarının performans kıyaslaması.”, Journal of Internet Applications and Management, 5/2, 43-56.

Aken, A., Litecky, C., Ahmad, A., ve Nelson, J. (2009). “Mining for computing jobs.”, IEEE software, 27/1, 78-85.

Akgül, F. G. ve Başkır, M. B. (2013). “Bankaların 2008-2012 Yılları Arasında Aktif Büyüklüklerini Etkileyen Kriterler Bakımından Hiyerarşik Kümeleme ve PAM Algoritması ile Sınıflandırılması.”, Bankacılık ve Sigortacılık Araştırmaları Dergisi, 1/5, 48-63.

Akın, Y. K. (2008). “Veri madenciliğinde kümeleme algoritmaları ve kümeleme analizi.”, Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Aktürk, C. (2020). “Pazarlama 4.0 için genetik algoritma tabanlı bir karar destek modeli önerisi.”, Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 9/1, 346-356.

Ala, T. ve Uğuz, S. (2021). “Türkiye’de Bölgesel Kalkınmanın Girişimcilik, İnovasyon ve Ar-Ge Çalışmalarıyla İlişkisinin Bibliyometrik Analizi ve LDA Mallet Uygulaması.”, Erzincan Üniversitesi Sosyal Bilimler Enstitüsü Dergisi Bölge Bilimi ve Planlama Kongresi Özel Sayısı, 14/20, 13-29.

Alan, A. ve Karabatak, M. (2020). “Veri Seti-Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi.”, Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 32/2, 531-540.

Albayrak, A. S., ve Yilmaz, S. K. (2009). “Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama.”, Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences, 14/1.

Altınok, Y. (2019). “Veri Madenciliğinde Hiyerarşik Kümeleme Algoritmalarının Uygulamalı Karşılaştırılması”, Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Altıntaş, V., Albayrak, M., ve Topal, K. (2021). “Kanser hastalığı ile ilgili paylaşımlar için Dirichlet ayrımı ile gizli konu modelleme.”, Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 36/4, 2183-2196.

Ankaralı, E., ve Külcü, Ö. (2020). "Rapidminer ile Twitter verilerinin konu modellemesi." Bilgi Yönetimi, 3/1, 1-10.

Ankerst, M., Breunig, M. M., Kriegel, H. P., ve Sander, J. (1999). "OPTICS: Ordering points to identify the clustering structure.", ACM Sigmod record, 28/2, 49-60.

Antons, D., Grünwald, E., Cichy, P., ve Salge, T. O. (2020). "The application of text mining methods in innovation research: current state, evolution patterns, and development priorities.", R&D Management, 50/3, 329-351.

Aslanyürek, M., ve Mesut, A. (2021). "Kümeleme Performansını Ölçmek için Yeni Bir Yöntem ve Metin Kümeleme için Değerlendirmesi.". Avrupa Bilim ve Teknoloji Dergisi, 27/1, 53-65.

Atan, S. (2020). "Metin Madenciliği: İmkânlar, Yöntemler ve Kısıtlar.", Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 31/1, 220-239.

Ayaz, İ. S., Baran, E., ve Açık, A. (2018). "Denizcilik Sektörüne Yönelik İş İlanlarının Analizi: Kariyer. Net Örneği.", In SETSCI Conference Indexing System, 3/1, 711-715.

Aydemir, E., ve Yavuz, M. (2019). "Mevsimlere Göre İlaç Satış Verilerinin Birliktelik Analizi ile İncelenmesi.", Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi, 3/1, 23-30.

Bağdatlı Kalkan, S., ve Yücel Bahar, Y. (2017). "Determination Of Factors Affecting Happiness Level By Classification Tree Technique.", European Journal of Business and Social Sciences, 6/2, 54 – 62.

Baldaniya, Rushabh H. (2014). "Overviewing Issues of Data Mining With Highlights of Data Warehousing.", International Journal of Scientific & Engineering Research, 5/4, 442.

Baykal, A. (2006). "Veri madenciliği uygulama alanları.", Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi, 7, 95-107.

Bektas, Y., ve Bektaş, J. (2021). "Avrupa Topluluğu Ekonomik Faaliyetlerin İstatistiki Sınıflandırılması Kullanılarak Dengesiz Veri Setlerinde Sınıflandırma Problemine Bakış." Icontech International Journal, 5/3, 31-37.

Belov, S., Javadzade, J., Kadochnikov, I., Korenkov, V., ve Zrelov, P. (2019). "Big data technologies for labour market analysis.", In Proceedings of the 27th International Symposium Nuclear Electronics and Computing (NEC'2019), Budva, Becici, Montenegro, 469-472.

Benoit, G. (2002). "Data Mining.", Annual Review of Information Science and Technology, 36, 265–310.

Bharadwaj, P., ve Shao, Z. (2019). "Fake news detection with semantic features and text mining.", International Journal on Natural Language Computing (IJNLC), 8.

Bilgin, T. T., ve Çamurcu, Y. (2005). "DBSCAN, OPTICS ve K-Means Kümeleme Algoritmalarının Uygulamalı Karşılaştırılması.", Politeknik Dergisi, 8/2, 139-145.

Blei, D. M., Ng, A. Y., ve Jordan, M. I. (2003). "Latent Dirichlet Allocation.", Journal Of Machine Learning Research, 3(Jan), 993-1022.

Bre, F., Gimenez, J. M., ve Fachinotti, V. D. (2018). "Prediction of wind pressure coefficients on building surfaces using artificial neural networks." Energy and Buildings, 158, 1429-1441.

Budak, İ. (2021). "Veri ve metin madenciliği ile hava yolu işletmelerinin sosyal medya yorum ve skorlarının değerlendirilmesi", Doktora Tezi, Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü, Denizli.

Budak, İ. ve Sökmen, A. (2022). "Otel Hizmetlerinin Değerlendirilmesinde Gizli Dirichlet Ayrımı ile Analiz: Kastamonu.", Journal of Tourism and Gastronomy Studies, 10/4, 2942-2954.

Budak, V. Ö., Kartal, E., ve Gülseçen, S. (2018). "Site-içi Aramalar ve Apriori Algoritması Kullanılarak Web Sitesi Ziyaretçilerinin İhtiyaç Tespitine Yönelik Bir Örnek Olay İncelemesi.", Bilişim Teknolojileri Dergisi, 11/2, 211-222.

Büyükarıkan, U. (2020). "Finansal Performansa Etki Eden Finansal Değişkenlerin CHAID Karar Ağacıyla Belirlenmesi: Tekstil Sektörü Örneği.", Aydın İktisat Fakültesi Dergisi, 5/1, 1-10.

Can, M. B., Eren, Ç., Koru, M., Özkan, Ö., ve Rzayeva, Z. (2012). "Veri kümelerinden bilgi keşfi: veri madenciliği.", Başkent Üniversitesi Tıp Fakültesi XIV. Öğrenci Sempozyumu, Ankara.

Candan, H., Durmuş, A., ve Harman, G. (2019). "Genetik Algoritma ve Sınıflandırıcı Yöntemler ile Kanser Tahmini.", Veri Bilimi, 2/1, 30-34.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., ve Blei, D. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models.", Advances In Neural İnformation Processing Systems, 22.

Cihan, Ş., Karabulut, B., Arslan, G., ve Cihan, G. (2018). "Koroner Arter Hastalığı Riskinin Veri Madenciliği Yöntemleri ile İncelenmesi.", International Journal of Engineering Research and Development, 10/1, 85-93.

Coşkun, M., ve Bülbül, H. İ. (2019). "Hanehalkı Bilişim Teknolojileri Kullanımının Veri Madenciliği Teknikleri ile Analizi.", TÜBAV Bilim Dergisi, 12/2, 1-17.

Çalış, A., Kayapınar, S., ve Çetinyokuş, T. (2014). "Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama." Endüstri Mühendisliği, 25/3, 2-19.

Çallı, L. ve Çallı, B. A. (2021). "Covid-19 Aşı Tereddütüne Sahip Hekimlerin Gizli Dirichlet Ayrımı (GDA) Algoritmasıyla Twitter Paylaşımlarının Konu Modellemesi.", 8. Uluslararası Yönetim Bilişim Sistemleri Konferansı. Marmara Üniversitesi, İstanbul Türkiye.

Çallı, L., Çallı, F., ve Çallı, B. A. (2021). "Yönetim Bilişim Sistemleri Disiplininde Hazırlanan Lisansüstü Tezlerin Gizli Dirichlet Ayrımı Algoritmasıyla Konu Modellemesi." MANAS Sosyal Araştırmalar Dergisi, 10/4, 2355-2372.

Çelik, S. (2020). "Metin madenciliği ile Shakespeare külliyatının incelenmesi.", Manas Sosyal Araştırmalar Dergisi, 9/3, 1343-1357.

Çelik, S., Bozkurt, Ö. Ç., ve Ekşili, N. (2022). "Çalışan Performansı Ölçeğindeki İfadelerin Karar Ağacı Algoritması ile Belirlenmesi.", Mehmet Akif Ersoy Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 9/1, 561-584.

Çetinkaya, S. (2014). "Kartografik genelleştirmede bina dizilimlerinin karakterizasyonu ve yorumlanmasına ilişkin yeni yaklaşımlar.", Doktora Tezi, Yıldız Teknik Üniversite Fen Bilimleri Enstitüsü, İstanbul.

Çınar, H. ve Arslan, G., (2008). "Veri madenciliği ve CRISP-DM yaklaşımı", XVII. İstatistik Araştırma Sempozyumu, 304-314, Ankara.

Çöllü, D. A., Akgün, L., ve Eyduran, E. (2020). "Karar ağacı algoritmalarıyla finansal başarısızlık tahmini: Dokuma, giyim eşyası ve deri sektörü uygulaması." Uluslararası Ekonomi ve Yenilik Dergisi, 6/2, 225-246.

Dakhil, A. (2019). "Development Of Effective Methodology For Improving Undergraduate Program Curriculum in Higher Education Utilizing Six Sigma Approach.", Doktora Tezi, Atılım Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., Pradhan, B., ve Kløve, B. (2019). "Urban Flood Risk Mapping Using The GARP and QUEST Models: A Comparative Study Of Machine Learning Techniques.", Journal of Hydrology, 569, 142-154.

Decorte, J. J., Van Hautte, J., Demeester, T., ve Develder, C. (2021). "JobBERT: Understanding Job Titles Through Skills.", arXiv preprint arXiv:2109.09605.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). "Indexing by Latent Semantic Analysis.", Journal Of The American Society For İnformation Science, 41/6, 391-407.

Demiralay, M. ve Çamurcu, A. Y. (2005). "CURE, AGNES ve K-MEANS Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması.", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 4/8, 1-18.

Efeoğlu, E. (2022). "Kablosuz Sinyal Gücünü Kullanarak İç Mekân Kullanıcı Lokalizasyonu için Karar Ağacı Algoritmalarının Karşılaştırılması.", Acta Infologica , 6/2, 163-173.

Ekinci, E. ve Omurca, S. İ. (2017). "Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkartılması.", Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 9/1, 51-58.

Ekinci, E., Omurca, S. İ., Kırık, E., ve Taşçı, Ş. (2020). "Tıp Veri Kümesi İçin Gizli Dirichlet Ayrımı.", Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 22/64, 67-80.

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

Ergüt, Ö. (2021). “Metin Madenciliği Yaklaşımıyla İşverenlerin Nitelik Taleplerinin İncelenmesi.”, İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi, 20/40, 138-157.

Ester, M., Kriegel, H. P., Sander, J., ve Xu, X. (1996). “A Density-Based Algorithm For Discovering Clusters in Large Spatial Databases With Noise.” In KDD, 96/34, 226-231.

Fayyad, U. M. (1996). “Data mining and Knowledge discovery in databases: Applications in Astronomy and Planetary Science.”, American Association for Artificial Intelligence, Menlo Park, CA, United States.

Fayyad, U., Piatetsky-Shapiro, G., ve Smyth, P. (1996). “From Data Mining to Knowledge Discovery in Databases.”, AI magazine, 17/3, 37-37.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., ve Romero, C. (2019). “Text mining in education.”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9/6, 1332.

Forman, G. (2003). “An extensive empirical study of feature selection metrics for text classification.”, J. Mach. Learn. Res., 3, 1289-1305.

Frawley, William J., Gregory Piatetsky-Shapiro, ve Christopher J. Matheus (1992). "Knowledge discovery in databases: An overview." AI magazine 13/3: 57-57.

Ganesh, S. (2002). “Data mining: Should it be included in the statistics curriculum?”, In The 6th international conference on teaching statistics (ICOTS 6), Cape Town, South Africa.

Gardiner, A., Aasheim, C., Rutner, P., ve Williams, S. (2018). “Skill requirements in big data: A content analysis of job advertisements.”, Journal of Computer Information Systems, 58/4, 374-384.

Gayathri, M., Shankar, R., ve Duraisamy, S. (2020). “Data Mining Technique Used For Air Pollution Prediction.”, International Research Journal of Modernization in Engineering Technology and Science, 02/10.

Gazioğlu, K., ve Şeker, Ş. E. (2017). “Veri Madenciliği Yöntemleri ile TWITTER Üzerinden Girişimcilik Analizi.”, YBS Ansiklopedi, 4/4.

Güler, M. (2018). "Aile İşletmelerinde ve Kurumsal İşletmelerdeki Yeni Mezun İşe Alım Süreçlerinin Karşılaştırılması, Yüksek Lisans Tezi", Sakarya Üniversitesi İşletme Enstitüsü, Sakarya.

Gürcan, F. (2017). "Yeni Nesil Yazılım Geliştirme Eğilimlerine Yönelik Uzman Bilgi ve Becerilerinin Olasılıksal Konu Modelleme Yordamıyla Belirlenmesi", Doktora Tezi, Karadeniz Teknik Üniversitesi/Fen Bilimleri Enstitüsü.

Gürcan, F., ve Cağıltay, N. E. (2019). "Big Data Software Engineering: Analysis Of Knowledge Domains And Skill Sets Using LDA-Based Topic Modeling.", IEEE access, 7, 82541-82552.

Gürcan, F., ve Özyurt, Ö. (2020). "Emerging Trends And Knowledge Domains in E-Learning Researches: Topic Modeling Analysis With The Articles Published Between 2008-2018.", Journal of Computer and Education Research, 8/16, 738-756.

Hand, D. J. (2007). "Principles Of Data Mining.", Drug safety, 30/7, 621-622.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., ve Yeganegi, M. R. (2020). "Text mining in big data analytics.", Big Data and Cognitive Computing, 4/1, 1.

Hassani, H., Huang, X., ve Silva, E. (2018). "Digitalisation and big data mining in banking.", Big Data and Cognitive Computing, 2/3, 18.

Hinneburg, A., ve Keim, D. A. (1998). "An efficient approach to clustering in large multimedia databases with noise.", Bibliothek der Universität Konstanz, 98, 58-65.

Hirudayaraj, M., Baker, R., Baker, F., ve Eastman, M. (2021). "Soft skills for entry-level engineers: What employers want.", Education Sciences, 11/10, 641.

Houtsma, M., ve Swami, A. (1995). "Set-oriented mining for association rules in relational databases.", In Proceedings of the eleventh international conference on data engineering, 25-33.

https://bilgisayarkavramlari.com/2008/12/01/ozellik-cikarimi-feature-extraction/, (Erişim Tarihi: 15.06.2023)

https://datavizcatalogue.com/TR/yontemleri/ag_diyagrami.html, (Erişim Tarihi: 29.05.2023).

https://medium.com/@sddkal/python-ile-g%C3%B6r%C3%BCnt%C3%BC-i%CC%87%C5%9Fleme-histogram-normalle%C5%9Ftirilmi%C5%9F-histogram-ve-histogram-e%C5%9Fitleme-3d0052174f1f, (Erişim Tarihi: 29.05.2023).

https://rapidminer.dqturkiye.com/, (Erişim Tarihi: 06.04.2023).

https://www.digitalvidya.com/blog/what-is-text-mining-guide/, (Erişim Tarihi: 09.11.2022)

https://www.secretcv.com/, (Erişim Tarihi: 04.06.2023)

https://www.yusufsezer.com.tr/cpp-nedir/, (Erişim Tarihi: 06.04.2023).

https://ybsansiklopedi.com/wp-content/uploads/2015/08/MetinMadenciligi30_32.pdf, (Erişim Tarihi: 29.05.2023).

Huang, H., Kvasny, L., Joshi, K. D., Trauth, E. M., ve Mahar, J. (2009). "Synthesizing IT job skills identified in academic studies, practitioner publications and job ads.", In Proceedings of the special interest group on management information system's 47th annual conference on Computer personnel research, 121-128.

Huber, S., Wiemer, H., Schneider, D., ve Ihlenfeldt, S. (2019). "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model.", Procedia Cirp, 79, 403-408.

Hutter, C. (2021). "Cyclicality of labour market search: a new big data approach.", Journal for labour market research, 55/1, 1.

Idrissi, A., Rehioui, H., Laghrissi, A., ve Retal, S. (2015). "An improvement of DENCLUE algorithm for the data clustering.", In 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA), 1-6.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., ve Zhao, L. (2019). "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey.", Multimedia Tools and Applications, 78, 15169-15211.

Kaçmaz, A., Yıldız, K., ve Buldu, A. (2020). "An application on technology addiction with c4.5 classification algorithm." Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 9/4, 1756-1765.

Kadhim, A. I. (2018). “An evaluation of preprocessing techniques for text classification.”, International Journal of Computer Science and Information Security (IJCSIS), 16(6), 22-32.

Kadhim, A. I., Cheah, Y. N., ve Ahamed, N. H. (2014). “Text document preprocessing and dimension reduction techniques for text document clustering.” In 2014 4th international conference on artificial intelligence with applications in engineering and technology, IEEE, 69-73.

Kahraman, H., ve Temel, S. (2021). “Yapay Sinir Ağları ve K-Ortalamalar Tabanlı Büyük Veri Azaltma Algoritmasının Tasarımı ve Uygulaması.”, Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 9/6, 329-342.

Kalemci, Ö. (2018). “Veri Madenciliği Yöntemi ile Prostat Kanseri İçin Erken Uyarı Protokollerinin Geliştirilmesi”, Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Kanık, B., Sunel, E., ve Taskin, T. (2012). “Beveridge Egrisi ve Eslesme Fonksiyonu: Türkiye Ornegi.”, Research and Monetary Policy Department, Central Bank of the Republic of Turkey.

Karaatlı, M., ve Altıntaş, E. (2018). “Clustering The Companies Listed On Stock Exchange Istanbul By Data Mining.”, Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10/26, 871-886.

Karaibrahimoğlu, A. (2014). “Veri madenciliğinden birliktelik kuralı ile onkoloji verilerinin analiz edilmesi: Meram tıp fakültesi onkoloji örneği”, Doktora Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya.

Karamaşa, Ç., ve Erdoğan, N. K. (2018). “Bayramlarda Gerçekleşen Trafik Kazalarının Birliktelik Kuralları ile Analiz Edilmesi.”, Karadeniz Uluslararası Bilimsel Dergi, 40, 386-411.

Karypis, G., Han, E. H., ve Kumar, V. (1999). “Chameleon: Hierarchical clustering using dynamic modeling.”, Computer, 32/8, 68-75.

Kaşıkçı, T., ve Gökçen, H. (2013). “Metin madenciliği ile e-ticaret sitelerinin belirlenmesi.”, Bilişim Teknolojileri Dergisi, 7/1, 25-32.

Keleş, M. K., ve Özel, S. A. (2017). "Similarity detection between Turkish text documents with distance metrics." In 2017 International Conference on Computer Science and Engineering (UBMK), IEEE, 316-321.

Kherwa, P., ve Bansal, P. (2019). "Topic modeling: a comprehensive review.", EAI Endorsed transactions on scalable information systems, 7/24, 1-16.

Kılıç, G., Budak, İ., ve Kılıç, B. S. (2020). "Kara cuma etiketlerinin Tweet istatistikleri ve duygu analizi ile sıralanması." Selçuk Üniversitesi Sosyal Bilimler Meslek Yüksekokulu Dergisi, 23/1, 131-140.

Kılıç, T. M., ve Turhan, Ş. (2022). "Organik gıda tüketim davranışlarına etki eden faktörlerin CHAID algoritması ile incelenmesi.", Gıda ve Yem Bilimi Teknolojisi Dergisi, 27, 26-35.

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M., ve Özçift, A. (2016). "KNN algoritması ve R dili ile metin madenciliği kullanılarak bilimsel makale tasnifi.", Marmara Fen Bilimleri Dergisi, 28/3, 89-94.

Koçer, S., ve Öksüz, G. (2015). "Elektronik İşe Alma Sürecinde Özel İstihdam Bürolarının Rolü: Adecco Türkiye ve Kariyer. Net İncelemesi.", Uluslararası Yönetim İktisat ve İşletme Dergisi, 11/24, 181-203.

Koong, K. S., Liu, L. C., ve Liu, X. (2002). "A study of the demand for information technology professionals in selected internet job portals.", Journal of Information Systems Education, 13/1, 21-28.

Kör, H. (2017). "Bulut Tabanlı Çevrimiçi Öğrenme Ortamında Etkinlik Öneri Sistemi Tasarımı: Eğitimsel Veri Madenciliği Uygulaması", Doktora Tezi, Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü, Kırıkkale.

Kureková, L. M., Beblavý, M., ve Thum-Thysen, A. (2015). "Using online vacancies and web surveys to analyse the labour market: a methodological inquiry.", IZA Journal of Labor Economics, 4, 1-20.

Kushwaha, A. K., Kar, A. K., ve Dwivedi, Y. K. (2021). "Applications of big data in emerging management disciplines: A literature review using text mining." International Journal of Information Management Data Insights, 1/2, 100017.

Larose, D. T. (2006), "Data Mining Methods and Models", Wiley , New Delhi, India.

Lyu, W., ve Liu, J. (2021). "Soft skills, hard skills: What matters most? Evidence from job postings.", Applied Energy, 300, 117307.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ve Adam, S. (2018). "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures.", 12/2-3, 93-118.

Matheus, C. J., Chan, P. K., ve Piatetsky-Shapiro, G. (1993). "Systems for knowledge discovery in databases." IEEE Transactions on knowledge and data engineering, 5/6, 903-913.

Matsuda, N., Ahmed, T., ve Nomura, S. (2019). "Labor market analysis using big data: The case of a Pakistani online job portal.", World Bank Policy Research Working Paper, 9063.

Mikut, R., ve Reischl, M. (2011). "Data mining tools.", Wiley interdisciplinary reviews: data mining and knowledge discovery, 1/5, 431-443.

Motlogelwa, P., Mbero, Z. A., Ayalew, Y., Nkgau, T. Z., ve Masizana-Katongo, A. (2011). "Computing knowledge and skills demand: A content analysis of job adverts in Botswana.", International Journal of Advanced Computer Science and Applications, 2/1, 1-10.

Mustaqim, T., Umam, K., ve Muslim, M. A. (2020). "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm.", In Journal of Physics: Conference Series, 1567/3, 032024.

Nguyen, V. H., ve Ho, T. (2021). "Analyzing customer experience in hotel services using topic modeling.", Journal of Information Processing Systems, 17/3, 586-598.

Novaković, J., Strbac, P., ve Bulatović, D. (2011). "Toward optimal feature selection using ranking methods and classification algorithms.", Yugoslav Journal of operations research, 21/1, 119-135.

Okat, N. (2019). "Predicting relative job placement potentials and mining skill sets by analyzing online job ads", Yüksek Lisans Tezi, Melikşah Üniversitesi, Kayseri.

Olszak, C. M., ve Lorek, P. (2019). "Big Data Approach to Analyzing Job Portals for the ICT Market." In Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology–ISAT 2018: Part II, 276-285.

Oza, K. S., ve Naik, P. G. (2016). "Prediction of online lectures popularity: a text mining approach.", Procedia Computer Science, 92, 468-474.

Özarı, Ç., Eren, Ö. ve Alıcı, A. (2019). "K-Ortalamalar Yönteminin Başlangıç Merkez Seçim Sorunsalı Üzerine Bir Çalışma." Business & Management Studies: An International Journal, 7/2, 1117-1135.

Özel, C. ve Topsakal, A. (2014). "Veri madenciliği kullanarak beton basınç dayanımının belirlenmesi.", Cumhuriyet Üniversitesi Fen Edebiyat Fakültesi Fen Bilimleri Dergisi, 35/1, 1-11.

Özkan, Y. (2020). "Veri Madenciliği Yöntemleri", Papatya Bilim, İSTANBUL.

Özsürünç, R. (2022). "Veri Madenciliğinde Lojistik Regresyon Modellerinin İncelenmesi", Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Öztürk, A. (2015). "Açık ve uzaktan öğrenme sistemlerinde kümeleme analizi yöntemiyle öğrenen gruplarının belirlenmesi", Yüksek Lisans Tezi, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.

Öztürk, H. (2022). "Dengesiz veri setlerinde farklı dengeleme algoritmalarının optimum denge oranlarının sınıflandırma ve regresyon ağaçları yöntemi ile incelenmesi: simülasyon çalışması", Doktora Tezi, Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü, Aydın.

Patacsil, F. F., ve Tablatin, C. L. S. (2017). "Exploring the importance of soft and hard skills as perceived by IT internship students and industry: A gap analysis.", Journal of Technology and Science education, 7/3, 347-368.

Pejić Bach, M., Krstić, Ž., Seljan, S., ve Turulja, L. (2019). "Text mining for big data analysis in financial sector: A literature review.", Sustainability, 11/5, 1277.

Peker, M., ve Kırbaş, İ. (2016). "Veri madenciliği süreç modeli ile el hareketlerinin myoelektrik kontrolü.", Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 7/1, 84-93.

Pekin, S. (2021). "Finansal Performans Tahmininde Metin Madenciliğinin Kullanımı: BIST İmalat Sanayi İşletmelerinde Bir Araştırma", Doktora Tezi, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir.

Polat, H. (2021). "Halkla İlişkiler 2.0 Kapsamında Hedef Belirleme ve Ölçme Sorunsalına Bakış: Alternatif Bir Yöntem Olarak Veri Madenciliğinin Kullanılmasına Yönelik Örnek Bir Uygulama", Doktora Tezi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü, Erzurum.

Pradipta, A., Hartama, D., Wanto, A., Saifullah, S., ve Jalaluddin, J. (2019). "The Application of Data Mining in Determining Timely Graduation Using the C45 Algorithm.", International Journal of Information System & Technology, 3/1, 31-36.

Rygielski, C., Wang, J. C., ve Yen, D. C. (2002). "Data mining techniques for customer relationship management." Technology in society, 24/4, 483-502.

Sabah, L., ve Bayraktar, H. (2020). "Veri madenciliği birliktelik kuralları yöntemi kullanarak binaların risk durumlarının belirlenmesi.", Gazi Mühendislik Bilimleri Dergisi, 6/1, 70-78.

Salman Saeed, M., Mustafa, M. W., Sheikh, U. U., Jumani, T. A., Khan, I., Atawneh, S., ve Hamadneh, N. N. (2020). "An efficient boosted C5. 0 decision-tree-based classification approach for detecting non-technical losses in power utilities.", Energies, 13/12, 3242.

Saritas, M. M., ve Yasar, A. (2019). "Performance analysis of ANN and Naive Bayes classification algorithm for data classification.", International journal of intelligent systems and applications in engineering, 7/2, 88-91.

Saura, J. R. (2021). "Using data sciences in digital marketing: Framework, methods, and performance metrics.", Journal of Innovation & Knowledge, 6/2, 92-102.

Savaş, B. K., İlkin, S., Hangisi, S., ve Şahin, S. (2017). "Gölge Tespitinde Kullanılan Bayes Sınıflandırma, Otsu Bölütleme ve Histogram Dağılımı Yöntemlerinin Karşılaştırılması.", Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 5/2, 345-355.

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.

Seyrek, İ. H., ve Ata, H. A. (2010). “"Veri zarflama analizi ve veri madenciliği ile mevduat bankalarında etkinlik ölçümü.”, BDDK Bankacılık ve Finansal Piyasalar Dergisi, 4/2, 67-84.

Shafique, U., ve Qaiser, H. (2014). “A comparative study of data mining process models: KDD, CRISP-DM and SEMMA.”, International Journal of Innovation and Scientific Research, 12/1, 217-222.

Silahtaroğlu, G. (2009). “Clustering Categorical Data Using Hierarchies: CLUCDUH.”, International Journal of Computer and Information Engineering, 3/8, 2006-2011.

Silahtaroğlu, G. (2016). “Veri madenciliği kavram ve algoritmaları.”, İstanbul: Papatya Yayıncılık Eğitim.

Singh, K., Devi, H. M., ve Mahanta, A. K. (2017). “Document representation techniques and their effect on the document Clustering and Classification: A Review.”, International Journal of Advanced Research in Computer Science, 8/5, 1780-1784.

Smith, D., ve Ali, A. (2014). “Assessıng market demand for WEB programming languages/technologies.”, Issues in Information Systems, 15/2, 411-420 .

Şahin, C. (2021). “Veri madenciliği yaklaşımının yeni ürün geliştirme sürecinde kullanımı: Akıllı telefonlar üzerine bir uygulama”, Doktora Tezi, Karadeniz Teknik Üniversitesi Sosyal Bilimler Enstitüsü, Trabzon.

Şeker, Ş. E. (2018). “CRISP-DM”, YBS Ansiklopedi, 5/2, 10-16.

Şentürk, A. (2006). “Veri Madenciliği Kavram ve Teknikler”, Ekin Yayınevi, BURSA.

Şimşek Gürsoy, U. T. (2009). “Veri Madenciliği ve Bilgi Keşfi.”, PEGEM Akademi, ANKARA.

Talib, R., Hanif, M. K., Ayesha, S., ve Fatima, F. (2016). “Text mining: techniques, applications and issues.”, International Journal of Advanced Computer Science and Applications, 7/11, 414-418.

Tapkan, P., Özbakır, L., ve Baykasoğlu, A. (2011). “Weka ile veri madenciliği süreci ve örnek uygulama.”, Endüstri Mühendisliği Yazılımları ve Uygulamaları Kongresi, 30, 247-262.

Telciler, C. (2013). “Veri Tabanı Kavramı ve MS-SQL Uygulamaları”, Pusula Yayıncılık, İSTANBUL.

Terzi, Ö., Küçüksille, E. U., Ergin, G., ve İlker, A. (2011). “Veri Madenciliği Süreci Kullanılarak Güneş Işınımının Tahmini.”, Uluslararası Teknolojik Bilimler Dergisi, 3/2, 29-37.

Timor, M., ve Yüzbaşı Künç, G. (2021). “Ekonomik Gelişmişliği Etkileyen Bilgi Ekonomisi Değişkenlerinin Veri Madenciliği ile Belirlenmesi.”, Optimum Ekonomi ve Yönetim Bilimleri Dergisi, 8/1, 1-18.

Tsantis, L., ve Castellani, J. (2001). “Enhancing learning environments through solution-based knowledge discovery tools: Forecasting for self-perpetuating systemic reform.”, Journal of Special Education Technology, 16/4, 39-52.

Türkiye İş Kurumu, (2021). “2021 Yılı Bilgi ve İletişim Sektörü İşgücü Piyasası Raporu”, İŞKUR İş Gücü Piyasası Raporu, 1 – 38.

Uluyardımcı, M. M., ve Zontul, M. (2019). “Veri madenciliği yöntemleri ile uçuş biletleme analizi.” Aurum Mühendislik Sistemleri ve Mimarlık Dergisi, 3/2, 153-168.

Uysal, İ., Bilen, M., ve Ulukuş, S. (2014). “TWOING Algoritması ile Sınıflandırma, Kalp Hastalığı Uygulaması.”, XVI Akademik Bilişim Konferansları, Mersin Üniversitesi, Şubat, 443-452.

Üçler, N., ve Büyükçelikok, T. Ö. (2021). “İletişim Fakültesi Müfredatlarının Medya Sektörü İş İlanları Üzerinden Sektörel Beklentileri Karşılama Yeterliğinin İncelenmesi.” Gaziantep University Journal of Social Sciences, 20/3, 1245-1269.

Vankevich, A., ve Kalinouskaya, I. (2021). “Better understanding of the labour market using Big Data.” Ekonomia i prawo. Economics and law, 20/3, 677-692.

Vijayarani, S., Ilamathi, M. J., ve Nithya, M. (2015). “Preprocessing techniques for text mining-an overview.” International Journal of Computer Science & Communication Networks, 5/1, 7-16.

Webb, G. K. (2006). “The market for IS and MIS skills and knowledge: analysis of on-line job postings.”, Communications, 51, 711.

Witten, I. H. (2004). “Text mining. The Practical Handbook of Internet Computing”, 1/23.146.

Yang, S., Guo, J. Z., ve Jin, J. W. (2018). “An improved Id3 algorithm for medical data classification.”, Computers & Electrical Engineering, 65, 474-487.

Yaşar, A. (2016). “Olumlu Görüş Dışındaki Denetim Görüşlerinin Veri Madenciliği Yöntemleriyle Tahminine İlişkin Karar ve Birliktelik Kuralları.”, Mali Çözüm Dergisi/Financial Analysis, 26/133, 81-109.

Yousefi, T., Odabas, M. S., ve Oktaş, R. (2020). “Kümeleme Algoritmalarında Kullanılan Farklı Yöntemlere Genel Bakış.” Black Sea Journal of Engineering and Science, 3/4, 173-189.

Zhou, Z. H. (2003). “Three perspectives of data mining.” Artificial Intelligence,143/1, 139-146