T.C.

MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞ ANALİTİĞİ ANABİLİM DALI
İŞ ANALİTİĞİ BİLİM DALI

# VERİ BİLİMİ POZİSYONLARINA YÖNELİK İŞ İLANLARINDA ARANAN YETKİNLİKLERİN İNCELENMESİ

Yüksek Lisans Tezi

ZEHRA ÖZSÜRÜNÇ

İstanbul, 2022

T.C.

MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞ ANALİTİĞİ ANABİLİM DALI
İŞ ANALİTİĞİ BİLİM DALI

# VERİ BİLİMİ POZİSYONLARINA YÖNELİK İŞ İLANLARINDA ARANAN YETKİNLİKLERİN İNCELENMESİ

Yüksek Lisans Tezi

ZEHRA ÖZSÜRÜNÇ

Danışman: Doç. Dr. Özgür ÇAKIR

İstanbul, 2022

# ÖZET

Büyük verinin yükselişe geçmesi ile şirketlerde veriyi işleyen, anlamlandıran, veriden iç görüler elde etmesi beklenen pozisyonlar hızla ortaya çıkmıştır. Veriden fayda sağlama iştahı ve pazardaki rekabet, şirketleri bu pozisyonlar için yetkin çalışan istihdamına yöneltmektedir. Şirketler, büyük veri ile çalışacak pozisyonlar oluştururken şirket ihtiyaç ve beklentilerini göz önünde bulundurmaktadırlar. Çalışan istihdamı da her şirket özelinde belirlenen kriterlere göre gerçekleşebilmektedir. Bu durum, iş ilanlarında nesnel bir yaklaşımın ortaya konmasına engel olmakta ve ilgili iş pozisyonu için aranan yetenek ve becerilerin bazen ütopik sayılabilecek düzeyde yüksek bazen de iş tanımını karşılamayacak düzeyde düşük kalmasına neden olmaktadır.

Bu çalışmada, iş ilanlarında veri yönetiminden yapay öğrenme uygulamalarına kadar geniş bir alanı kapsayan sekiz farklı pozisyon tanımı bir üst kavram olarak veri bilimi alanında kabul edilmiştir. Bu iş pozisyonları için Türkiye genelinde açılan 1,465 iş ilanı toplanarak metin madenciliği teknikleri ile incelenmiş ve bu pozisyonlar için aranan yetkinliklerin çerçevesi ortaya konulmuştur. Çalışma sonucunda, işletmelerin veri bilimi pozisyonlarını ve bu pozisyonlar için gerekli yetkinlikleri aynı şekilde tanımlamadıklarını gösteren anlamlı farklılıklara rastlanmıştır.

**Anahtar Kelimeler:** *Veri bilimi, Çalışan yetkinlikleri, Metin madenciliği*

# ABSTRACT

With the rise of big data, positions that process and make sense of data and that are expected to gain insights from data have emerged rapidly in companies. The appetite for benefiting from data and competition in the market are driving companies to recruit competent employees for these positions. Companies consider company needs and expectations when creating positions to work with big data. Employee employment can also be realized according to the criteria determined for each company. This situation prevents an objective approach in job postings and causes the skills and abilities sought for the relevant job position to be high at a level that can be considered utopian and sometimes low at a level that cannot meet the job description.

In this study, eight different position definitions covering a wide area from data management to artificial learning applications in job postings have been accepted as a top concept in the field of data science. For these job positions, 1,465 job postings opened throughout Turkey were collected and analyzed using text mining techniques, and the framework of competencies sought for these positions was revealed. As a result of the study, significant differences were found showing that businesses do not define data science positions and the competencies required for these positions in the same way.



**Keywords:** *Data science, Employee competencies, Text mining*

# ÖNSÖZ

Tez sürecim boyunca her daim desteklerini hissettiğim canım ailem ve arkadaşlarıma, tez çalışmam süresince bilgi ve tecrübesi ile ilham veren kıymetli danışman hocam Doç. Dr. Özgür Çakır'a, gösterdiği inanç ve destekleri için çok teşekkür ederim.

Zehra ÖZSÜRÜNÇ

İSTANBUL, 2022

# İÇİNDEKİLER

ÖZET ..... i
ABSTRACT ..... ii
ÖNSÖZ ..... iii
İÇİNDEKİLER ..... iv
KISALTMALAR ..... vi
TABLOLAR LİSTESİ ..... vii
ŞEKİLLER LİSTESİ ..... viii
GİRİŞ ..... 1
BÖLÜM 1. VERİ BİLİMİ ..... 2
1.1. VERİ ANALİZİNİN KÖKENİ: İSTATİSTİK ..... 2
1.2. VERİ TABANI YÖNETİM SİSTEMİ ..... 4
1.2.1. Operasyonel Veri Deposu (ODS) ..... 5
1.2.2. NoSQL ..... 6
1.2.3. Veri Ambarı ..... 7
1.2.4. Data Mart ..... 8
1.3. VERİ TABANLARINDA BİLGİ KEŞFİ ..... 9
1.4. VERİ MADENCİLİĞİ ..... 10
1.4.1. Yapay Zekâ ..... 12
1.4.2. Makine Öğrenmesi ..... 13
1.4.3. Derin Öğrenme ..... 15
1.5. İŞ ZEKÂSI VE BÜYÜK VERİ ..... 15
1.6. VERİ BİLİMİ ..... 18
1.7. VERİ BİLİMİ YETKİNLİK GEREKSİNİMLERİ ..... 18
BÖLÜM 2. METİN MADENCİLİĞİ ..... 23
2.1. METİN MADENCİLİĞİ SÜRECİ ..... 25
2.1.1. Metin Ön İşleme ..... 26
2.1.2. Normalizasyon İşlemleri ..... 26
2.1.3. Durak Kelimeler (Stopwords) ..... 27
2.1.4. Bölütleme (Tokenization) ..... 27
2.1.5. Gövdeleme (Lemmatization) ..... 27
2.1.6. Vektör Oluşturma/Kelime Torbası (Bag of Words) ..... 28
2.2. METİN MADENCİLİĞİ YÖNTEMLERİ ..... 29
2.2.1. Doğal Dil İşleme (NLP) ..... 29
2.2.2. Bilgi Çıkarma (Information Extraction) ..... 30

2.2.3. Duygu Analizi (Sentimental Analysis) ..... 31
2.2.4. Sınıflandırma (Classification / Categorization) ..... 32
2.2.5. Kümeleme (Clustering) ..... 35
BÖLÜM 3. VERİ BİLİMİ ALANINDAKİ İŞ İLANLARININ ANALİZİ ..... 39
3.1. AMAÇ, KAPSAM VE YÖNTEM ..... 39
3.2. VERİ KÜMESİ VE VERİ HAZIRLAMA ..... 41
3.3. VERİ ANALİZCİSİ İŞ İLANLARININ ANALİZİ ..... 47
3.4. VERİ BİLİMCİSİ İŞ İLANLARININ ANALİZİ ..... 53
3.5. İŞ ZEKASI İŞ İLANLARININ ANALİZİ ..... 59
3.6. İŞ ANALİSTİ İŞ İLANLARININ ANALİZİ ..... 65
3.7. VERİ MÜHENDİSİ İŞ İLANLARININ ANALİZİ ..... 70
3.8. VERİ AMBARI UZMANI İŞ İLANLARININ ANALİZİ ..... 76
3.9. VERİ TABANI UZMANI İŞ İLANLARININ ANALİZİ ..... 81
3.10. MAKİNE ÖĞRENMESİ MÜHENDİSİ İŞ İLANLARININ ANALİZİ ..... 86
3.11. ANLAMLI FARKLILIKLARIN SINANMASI ..... 92
3.11.1. Tüm İlanlar Bazında Anlamlı Farklılıkların Sınamaları ..... 93
3.11.2. Pozisyonlar Bazında Anlamlı Farklılıkların Sınanması ..... 98
3.12. BULGULAR VE TARTIŞMA ..... 105
SONUÇ ..... 114
KAYNAKÇA ..... 116

# KISALTMALAR

| | |
|---|---|
| **CRISP – DM** | Cross Industry Standard Processing – Data Mining |
| **ODS** | Operational Data Store |
| **IMS** | Information Management System |
| **CICS** | Customer Information Control System |
| **OLAP** | Online Analytical Processing |
| **OLTP** | Online Transaction Processing |
| **NLP** | Natural Language Processing |
| **IE** | Information Extraction |
| **IR** | Information Retrieval |
| **SQL** | Structured Query Language |
| **SAP** | System, Application & Products |
| **ODI** | Oracle Data Integrator |
| **AWS** | Amazon Web Servis |

# TABLOLAR LİSTESİ

**Sayfa No.**

**Tablo 1.** Pozisyon bazlı ilan sayısı & ilan bazlı yetkinlik adeti ..... 94
**Tablo 2**. Pozisyon bazlı ilan oranları & ilan bazlı yetkinlik oranları ..... 95
**Tablo 3.** Toplam ilan sayısındaki yetkinlik oranları ve sıralamaları ..... 97
**Tablo 4.** Yetkinlik bazlı anlamlılık test sonuçları matrisi ..... 97
**Tablo 5.** İş analisti pozisyonu-anlamlılık test sonuçları ..... 98
**Tablo 6.** İş zekâsı pozisyonu anlamlılık test sonuçları ..... 99
**Tablo 7.** Veri analizcisi-anlamlılık test sonuçları ..... 100
**Tablo 8.** Veri mühendisi-anlamlılık test sonuçları ..... 101
**Tablo 9.** Veri bilimcisi-anlamlılık test sonuçları ..... 102
**Tablo 10.** Veri ambarı uzmanı-anlamlılık test sonuçları ..... 102
**Tablo 11.** Veri tabanı uzmanı-anlamlılık test sonuçları ..... 104
**Tablo 12.** Makine öğrenmesi mühendisi-anlamlılık test sonuçları ..... 104

# ŞEKİLLER LİSTESİ

**Sayfa No.**

**Şekil 1.**Operasyonel Veri Deposu Mimarisi ..... 6
**Şekil 2.** OLTP Sistemi (Veri Tabanı) ile Veri Ambarı Sisteminin Karşılaştırılması ..... 8
**Şekil 3.**Veri tabanlarından bilgi keşfi ..... 9
**Şekil 4.** Veri Madenciliği Süreci ..... 11
**Şekil 5.** Yapay Zekâ Şeması ..... 13
**Şekil 6.** Makine öğrenmesi algoritmalarının sınıflandırılması ..... 14
**Şekil 7.** Büyük verinin özellikleri ..... 16
**Şekil 8.** Büyük veri tipleri ..... 17
**Şekil 9.** Bir metin madenciliği çerçevesi ..... 24
**Şekil 10.** Metin madenciliği için 7 pratik alan ..... 25
**Şekil 11.** İngilizce dilindeki bazı durak kelimeler örneği ..... 27
**Şekil 12.** Oyun kartı örneği ..... 36
**Şekil 13.** Python kütüphanelerinin yüklenmesi ..... 41
**Şekil 14.** Veri dosyasının Python üzerinden görüntülenmesi ..... 42
**Şekil 15.** Büyük-küçük harf dönüşümü ..... 42
**Şekil 16.** Noktalama işaretlerinin silinmesi ..... 42
**Şekil 17.** Sayıların silinmesi ..... 43
**Şekil 18.** İngilizcedeki yararsız kelimeler (stopwords) listesi ..... 43
**Şekil 19.** Özel olarak yararsız kelimeler (stopwords) listesinin oluşturulması ..... 44
**Şekil 20.** Yararsız kelimelerin metinden çıkarılması ..... 44
**Şekil 21.** Kelime saydırma ..... 45
**Şekil 22.** Az geçen kelimelerin tespit edilmesi ..... 45
**Şekil 23.** Az geçen kelimelerin silinmesi ..... 46
**Şekil 24.** Kelime bölütleme (tokenization) işlemi ..... 47
**Şekil 25.** Kelimenin köklere ayrıştırılması (Lemmatization) işlemi ..... 47
**Şekil 26.** Veri analizcisi aranan nitelikler-terim frekansı ..... 48
**Şekil 27.** Veri analizcisi aranan nitelikler-kelime bulutu ..... 48
**Şekil 28.** SQL-İletişim-Python en yakın kelimeler analizi ..... 49
**Şekil 29.** Veri analizcisi aranan nitelikler bigram analizi ..... 50
**Şekil 30.** Veri analizcisi aranan nitelikler trigram analizi ..... 50
**Şekil 31.** Veri analizcisi iş tanımı bigram analizi ..... 51
**Şekil 32.** Veri analizcisi iş tanımı trigram analizi ..... 51

**Şekil 33.** Veri analizcisi üniversite bölümü bigram analizi ..... 52
**Şekil 34.** Veri analizcisi-üniversite bölümü-trigram analizi ..... 52
**Şekil 35.** Veri analizcisi-üniversite bölümü-kelime bulutu ..... 53
**Şekil 36.** Veri bilimcisi-aranan nitelikler-terim frekansı ..... 54
**Şekil 37.** Veri bilimcisi-aranan nitelikler-kelime bulutu ..... 54
**Şekil 38.** Python-SQL-İletişim(communication) en yakın kelime analizi ..... 55
**Şekil 39.** Veri bilimcisi-aranan nitelikler-bi-gram analizi ..... 56
**Şekil 40.** Veri bilimcisi-aranan nitelikler-tri-gram analizi ..... 56
**Şekil 41.** Veri bilimcisi-iş tanımı-bigram analizi ..... 57
**Şekil 42.** Veri bilimcisi-iş tanımı-trigram analizi ..... 57
**Şekil 43.** Veri bilimcisi-üniversite bölümü-bigram analizi ..... 58
**Şekil 44.** Veri bilimcisi-üniversite bölümü-trigram analizi ..... 58
**Şekil 45.** Veri bilimcisi-üniversite bölümü-kelime bulutu ..... 59
**Şekil 46.** İş zekâsı-aranan nitelikler-terim sıklığı ..... 59
**Şekil 47.** İş zekâsı-aranan nitelikler-kelime bulutu ..... 60
**Şekil 48.** SQL-ETL-SAP için en yakın kelime analizi ..... 61
**Şekil 49.** İş zekâsı-aranan nitelikler-bigram analizi ..... 61
**Şekil 50.** İş zekâsı-aranan nitelikler-trigram analizi ..... 62
**Şekil 51.** İş zekâsı-aranan nitelikler-kelime bulutu ..... 62
**Şekil 52.** İş zekâsı-iş tanımı-bigram analizi ..... 63
**Şekil 53.** İş zekâsı-iş tanımı-trigram analizi ..... 63
**Şekil 54.** İş zekâsı-üniversite bölümü-bigram analizi ..... 64
**Şekil 55.** İş zekâsı-üniversite bölümü-trigram analizi ..... 64
**Şekil 56.** İş zekâsı-üniversite bölümü-kelime bulutu ..... 65
**Şekil 57.** İş analisti-aranan nitelikler-terim sıklığı ..... 66
**Şekil 58.** İletişim(communication)-SQL-Çevik(agile) kelimelerinin en yakın kelime analizi ..... 66
**Şekil 59.** İş analisti-aranan nitelikler-kelime bulutu ..... 67
**Şekil 60.** İş analisti-aranan nitelikler-bigram analizi ..... 68
**Şekil 61.** İş analisti-aranan nitelikler-trigram analizi ..... 68
**Şekil 62.** İş analisti-iş tanımı-bigram analizi ..... 69
**Şekil 63.** İş analisti-iş tanımı-trigram analizi ..... 69
**Şekil 64.** İş analisti-üniversite bölümü-bigram analizi ..... 70
**Şekil 65.** İş analisti-üniversite bölümü-trigram analizi ..... 70
**Şekil 66.** Veri mühendisi-aranan nitelikler-terim sıklığı ..... 71
**Şekil 67.** SQL-Spark-Python kelimelerinin en yakın kelime analizi ..... 72

**Şekil 68.** Veri mühendisi-aranan nitelikler-kelime bulutu ..... 72
**Şekil 69.** Veri mühendisi-aranan nitelikler-bigram analizi ..... 73
**Şekil 70.** Veri mühendisi-aranan nitelikler-trigram analizi ..... 73
**Şekil 71.** Veri mühendisi-iş tanımı-bigram analizi ..... 74
**Şekil 72.** Veri mühendisi-iş tanımı-trigram analizi ..... 74
**Şekil 73.** Veri mühendisi-üniversite bölümü-bigram analizi ..... 75
**Şekil 74.** Veri mühendisi-üniversite bölümü-trigram analizi ..... 75
**Şekil 75.** Veri mühendisi-üniversite bölümü-kelime bulutu ..... 76
**Şekil 76.** Veri ambarı-aranan nitelikler-terim frekansı ..... 76
**Şekil 77.** SQL-Oracle-ETL kelimelerinin en yakın kelime analizi ..... 77
**Şekil 78.** Veri ambarı uzmanı-aranan nitelikler-kelime bulutu ..... 78
**Şekil 79.** Veri ambarı uzmanı-aranan nitelikler-bigram analizi ..... 78
**Şekil 80.** Veri ambarı uzmanı-aranan nitelikler-trigram analizi ..... 79
**Şekil 81.** Veri ambarı uzmanı-iş tanımı-bigram analizi ..... 79
**Şekil 82.** Veri ambarı uzmanı-iş tanımı-trigram analizi ..... 80
**Şekil 83.** Veri ambarı uzmanı-üniversite bölümü-bigram analizi ..... 80
**Şekil 84.** Veri ambarı uzmanı-üniversite bölümü-trigram analizi ..... 81
**Şekil 85.** Veri ambarı uzmanı-üniversite bölümü-kelime bulutu ..... 81
**Şekil 86.** Veri tabanı uzmanı-aranan nitelikler-terim sıklığı ..... 82
**Şekil 87.** SQL-Oracle-İletişim kelimelerinin en yakın kelime analizi ..... 82
**Şekil 88.** Veri tabanı uzmanı-aranan nitelikler-kelime bulutu ..... 83
**Şekil 89.** Veri tabanı uzmanı-aranan nitelikler-bigram analizi ..... 83
**Şekil 90.** Veri tabanı uzmanı-aranan nitelikler-bigram analizi ..... 84
**Şekil 91.** Veri tabanı uzmanı-iş tanımı-bigram analizi ..... 84
**Şekil 92.** Veri tabanı uzmanı-iş tanımı-trigram analizi ..... 85
**Şekil 93.** Veri tabanı uzmanı-üniversite bölümü-bigram analizi ..... 85
**Şekil 94.** Veri tabanı uzmanı-üniversite bölümü-trigram analizi ..... 86
**Şekil 95.** Veri tabanı uzmanı-üniversite bölümü-kelime bulutu ..... 86
**Şekil 96.** Makine öğrenmesi mühendisi-aranan nitelikler-terim frekansı ..... 87
**Şekil 97.** Python-Tensorflow-Spark kelimelerinin en yakın kelime analizi ..... 87
**Şekil 98.** Makine öğrenmesi-aranan nitelikler-kelime bulutu ..... 88
**Şekil 99.** Makine öğrenmesi mühendisi-aranan nitelikler-bigram analizi ..... 89
**Şekil 100.** Makine öğrenmesi mühendisi-aranan nitelikler-trigram analizi ..... 89
**Şekil 101.** Makine öğrenmesi mühendisi-iş tanımı-bigram analizi ..... 90
**Şekil 102.** Makine öğrenmesi mühendisi-iş tanımı-trigram analizi ..... 90

**Şekil 103.** Makine öğrenmesi mühendisi-üniversite bölümü-bigram analizi..... 91
**Şekil 104.** Makine öğrenmesi mühendisi-üniversite bölümü-trigram analizi ..... 91
**Şekil 105.** Makine öğrenmesi mühendisi-üniversite bölümü-kelime bulutu..... 92
**Şekil 106.** Pozisyonlar bazında öne çıkan yetkinlikler..... 113

# GİRİŞ

Veri tabanlarının ortaya çıkması, biriken veriden yararlı bilgiler elde edilmeye çalışılması istatistik biliminin gelişimi verinin ve veri analiz çalışmalarının popüler olmasını sağlamıştır. Bununla beraber 2000'li yıllardan sonra internetin yaygınlaşması, mobil teknolojilerin gelişimi, sosyal medya platformlarının popülerlik kazanması veri miktarını her geçen gün artırmıştır. Şirketler için ciddi bir rekabet avantajı sağlayan veri ve veri analiz çalışmaları da bu gelişmelerle beraber her geçen gün daha dikkate değer noktaya ulaşmıştır. Şirketler için verilerden öngörüler elde etmek ve bu öngörüler ışığında gelecek planlaması yapmak ve stratejik kararlar almak her zamankinden daha önemlidir. Bu kapsamda veri analizi şirketler için ajandanın ilk sırasında yer alırken, bu alanda istihdam edilecek çalışanlar şirketler için büyük önem taşımaktadır.

Şirketler verilerden iç görü elde etmek istemesine karşın, veriden iç görü elde etmek, veri analizi sürecinin son adımı sayılabilir. Veri analizi birden çok basamaktan oluştuğundan, veri bilimi alanında veri ile uğraşan farklı pozisyonlar ortaya çıkmıştır. Ancak bu pozisyonların yetkinlik setleri ve görev tanımları tam anlamıyla ayrıştırılabilmiş, bir çerçeveye oturtulabilmiş değildir.

Çoğu kullanımda birbirinin yerine geçen Veri Bilimi pozisyonları ve bu pozisyonlarda çalışanların hangi beceri ve yetkinliğe sahip olması gerektiği, kafa karıştırıcı olmaktadır. Bu tez çalışması, büyük veri ile doğan ve henüz tam anlamıyla yetkinlik setlerinde ayrımın net bir şekilde yapılmadığı ve kafa karışıklığı olduğu düşünülen Veri Bilimi pozisyonları için gerekli yetkinlik ve beceri setlerinin çerçevesini, bu alanda yayınlanmış iş ilanları üzerinden incelemeyi ve Türkiye'deki şirketlerin veri bilimi yetkinlik setlerine yaklaşımlarını ortaya koymayı amaçlamaktadır. Bu çalışma ile Veri Bilimi pozisyonlarında istihdam edilecek adayların teknik ve sosyal becerilerinin neler olması gerektiği, bu pozisyonların kendi arasında ne kadar ayrıştığı ve benzeştiği, ortak yetkinlik setlerinin neler olduğu ortaya konulmaya çalışılmıştır.

Bu tez çalışması, üç bölümden oluşmaktadır. İlk bölümde veri bilimi tarihçesi istatistik biliminin doğuşu, veri tabanlarının ortaya çıkışı, veri madenciliği, büyük veri ile doğan veri bilimi kavramları üzerinde durulmuş ve veri bilimi alanındaki yetkinlik ihtiyaçları aktarılmıştır. İkinci bölümde tez çalışması kapsamında kullanılan metin madenciliği metin ön işleme adımları aktarılmıştır. Bununla beraber metin madenciliği yöntemlerine de genel hatları ile değinilmiştir. Üçüncü bölümde, tez çalışmasının amaç, kapsam ve yöntemi aktarılarak veri kümesi tanıtılmıştır. Sonrasında analiz sonuçları her başlıkta ayrı ayrı paylaşılmıştır. Metin madenciliği kapsamında elde edilen analiz sonuçları, anlamlılık testleri yapılarak bölüm içerisinde aktarılmıştır. Analiz ve anlamlılık test sonuçları bulgular kısmında ifade edilmiştir. Sonuç ve tartışma kısmında uygulama sonucu elde edilen bulgular değerlendirilerek yorumlanmıştır.

# BÖLÜM 1. VERİ BİLİMİ

Bu bölümde, istatistik disiplini ve veri tabanlarının ortaya çıkışı aktarılacaktır. Veri madenciliği kapsamında, Yapay Zekâ, Makine Öğrenmesi ve Derin Öğrenme kavramları tanımlanacaktır. Büyük veri ve veri bilimi hakkında bilgi verilerek veri biliminde yetkinlik gereksinimlerine genel bir bakış ortaya konacaktır.

## 1.1. VERİ ANALİZİNİN KÖKENİ: İSTATİSTİK

İstatistikten bahsetmeden önce bir adım geriye giderek analitik yaklaşımdan kısaca bahsetmek gerekir. Analitik, elde bulunan verilerde anlamlı örüntüler bulabilmek için, matematik, istatistik, tahminlere dayalı modelleme ve makine öğrenimi tekniklerinin kullanılması olarak tanımlanabilir (Knight, 2018). 1960'ların sonlarında bilgisayarların karar destek sistemleri haline gelmesiyle analitiğin yükselişi başlamış oldu (Foote, 2021). Bu bölümde 1960'lardan günümüze veri biliminin ortaya çıkış süreci, istatistik bilimi milat alınarak aktarılacaktır.

İngiltere'de 1663'lerde süren veba salgını sırasında Londralı bilim insanı John Graunt (1620–74), ölüm kayıtlarını inceleyerek ölüm olaylarının düzenli olarak ortaya çıktığını fark eder ve ölüm olaylarını önleyebilmek amacıyla bir erken uyarı sistemi tasarlar. Graunt'un sunduğu bu analiz ilk resmi istatistiksel veri analizi olarak kayıtlara geçer. Yine o dönemde yaşayan, kuyruklu yıldızın isim babası olarak bilinen, kuyruklu yıldızın görülme olasılığını hesaplamaya çalışan Edmund Halley (1656–1742) Graunt'un veri analiz çalışmalarını ileri bir noktaya taşır. Halley de Graunt gibi ölüm kayıtlarını inceleyerek insanların hayatta kalma olasılıklarını hesaplamaya çalışır (Tabak, 2014).

John Graunt ve Edmund Halley, ölüm kayıtları üzerinden yaptıkları analizlerden bazı aydınlatıcı sonuçlar elde ederler ancak sistematik bir şekilde dikkate almadıkları bir sorun vardır, bu da kesinlik sorunudur. Graunt ve Halley kesinlik sorunundan bahsetmişlerdir ancak; çalıştıkları veriler kesinlik hakkında matematiksel bir tartışmaya uygun değildir. Sayısal verilerden kesin sonuçlar çıkarma sorunu ilk olarak 19. yüzyılın başlarında Fransız matematikçi Adrien-Marie Legendre (veya Le Gendre; 1752-1833) tarafından ele alınmıştır. Legendre, bir kuyruklu yıldızın yörüngesini ölçümlerden tahmin etme problemi üzerine çalışarak bu noktada istatistiğe katkıda bulunur. Bir kuyruklu yıldızın yörüngesini hesaplamak basit bir işlem gibi gözükebilir. Gerekli sayıda ölçüm yaptıktan sonra hesaplamaların yapılabileceği düşünülebilir ancak her ölçüm belli miktarda hata oranı içerir ve ölçümler arttıkça hata oranları da artar. Legendre kesinlik problemini bugün "en küçük kareler yöntemi (method of least squares)" denilen metodla çözmüştür (Tabak, 2014). Bu durum göstermektedir ki her yeni keşif, buluş beraberinde yeni bir soru doğurmuştur. Birbiri ardına sorulan sorular ve verilen cevaplar ile veri analizi bugünkü gelişmiş formuna dönüşmüştür.

İstatistiğin ortaya çıkış serüvenine bakıldığında, istatistik kabaca, gözlemlenebilir ve ölçülebilir olaylardan veri toplanması ve bu verilerin anlamlı bilgiye dönüştürülmesi olarak tanımlanabilir. İstatistik biliminde araştırmacı belli bir soruya cevap aramaktadır.

İstatistik, veri analizinde bir hipotez ile yola çıkar. Bu hipotezi test etmek için veriler toplanır. Veriler betimsel (descriptive) veya çıkarımsal (inferential) istatistik yöntemleri ile analiz edilir. Betimsel istatistik analizinin temel hedefi, elde edilen verileri tablo veya grafik olarak göstermektir. Bu kapsamda ortalama, mod, medyan, standart sapma, varyans ve korelasyon katsayısı gibi metriklerden yararlanılır. Çıkarımsal analiz yönteminde ise elde edilen sonuçlar üzerinden yorumlamalar yapılır.

Tümevarım yöntemiyle ana kütle hakkında çıkarımlar yapılır. Bu yaklaşımda momentler yöntemi, en büyük olabilirlik, en büyük artçıl, bayes-tipi kestirimci, minimum uzaklık, maksimum aralık verme, kestirim, güven aralığı ve inanılır aralık gibi ölçüler kullanılır. Bunlar dışında, gruplar arası farklılıkları ölçmek amacıyla fark analizi yöntemi; araştırma yapılan veriler arasındaki ilişkileri tespit etmek üzere ilişki analizi yöntemi; geleceğe yönelik tahminlerde bulunmak üzere, tahmin analiz yöntemi kullanılır. Tahmin yönteminde regresyon analiz tekniklerinden faydalanılır. Analiz edilen verilerin anlamlılık düzeyleri ve güven aralıkları gibi metriklerle araştırmacının hipotezi doğrulanır veya reddedilir. Bu anlamda, istatistiki yaklaşımda ele alınan konunun bir kapsamı ve amacı vardır. Çerçevesi önceden çizilmiş bir yapıda, veri analizi gerçekleştirilir. Sonuçlar yine belli bir çerçevede değerlendirilir.

İstatistik bilimi belli bir olgunluğa ulaştığında bu alanda da yeni sorgulamalar ortaya çıkmıştır. İstatistik biliminin veri analiz yaklaşımı, Tukey'in "Veri Analizinin Geleceği (The Future of Data Analysis)" isimli makalesiyle bir tartışma zeminine oturtulmuştur. Tukey 1962'de veri analizi tartışmasını başlatarak veri analizinin yöntemlerini ve geleceğini irdelemiştir. Tukey, (1962) makalesinde veri analizinde kullanılan analiz araçları ve yaklaşımlarını ele almıştır. O dönem için hali hazırda kullanılan analiz araçlarına ek olarak yeni arayışların da mevcut olduğunu söylemiştir. Tukey, bir anlamda bugünkü veri madenciliği çalışmalarına ve çeşitlenen analiz araçlarına işaret etmiştir. Veri analizinde temel yaklaşımın *"Veri bize ne anlatmak istemektedir?"* sorusu üzerine kurulması gerektiğini ifade etmiştir. Tukey'in bu yaklaşımı ise bilgi keşif sürecine (KDD) işaret etmektedir.

İstatistik yaklaşımına baktığımızda, araştırmacı sorusuna cevap bulmak, hipotezini test etmek için yola çıkar. Dolayısıyla veride çok sesliliğe izin vermez. Ancak veri, çoğu zaman sorulan sorunun dışında da yanıtlar içerir. Bugün iç görü olarak tanımlanan bu yanıtlar, veri analiz yaklaşımıyla yakından ilgilidir.

Tukey, (1962) de makalesinde iç görü meselesine değinmektedir. Veri analizinde elde edilen sonuçlardan ziyade, bu sonuçlardan bir öngörü çıkarmanın önemli olacağını, grafiklerin bilgisayarlar tarafından kolayca elde edilebileceğini ve araştırmacıların analizlerde makinelerin yaptığı işlemleri denetleyen konumda olacağını ileri sürmüştür. Bu anlamda veri analizinde grafiklerin ötesine geçmek ve veriyi derinden incelemenin önemine vurgu yapmıştır. Bugün gelişmiş bilgisayar ve analiz araçlarıyla

Tukey'in belirttiği gibi analiz makinalarca yapılmakta ve görselleştirilmektedir. Veri analizinin geleceğine ışık tutan ve yön veren Tukey'in yaklaşımları ve değerlendirmeleri, veri analizinde bugün gelinen noktanın habercisi durumundadır. Veri analizi tartışmalarının erken dönemlerinde Tukey'in tespit ve önermeleri yol gösterici olmuştur ve bugün de geçerliliğini koruyan tespit ve önermelerdir.

Veri analizinin temellerinin istatistiğe dayanması (Foote, 2021), internetin ortaya çıkması ve beraberindeki teknolojik gelişmeler daha gelişmiş analiz yöntemlerini beraberinde getirmiştir (Marr, 2015). Büyük verinin ortaya çıkması gelişmiş yöntemlere olan ihtiyacı artırmıştır (Bryson et al., 1999). İstatistiksel veri analizi bu anlamda, büyük veri için de milat olarak kabul edilir (Marr, 2015).

İstatistikten veri bilimine doğru giden yolda, bazı temel istatistiksel yaklaşımlardan kısaca bahsettik. Bunun yanında, istatistiğin bir adım ötesine gidildiğinde ele alınacak kavramları tek başına aktarmak zorlu olmaktadır. Bu nedenle istatistikten bahsederken, aynı zamanda veri madenciliğinden bahsedecek; veri madenciliğinden bahsederken de veri tabanlarına, veri ambarlarına da değinilecektir.

Veri tabanları bilgisayarların ortaya çıkmasından bağımsız olarak değerlendirilemezken, internetin yaygınlaşması ile artan büyük veri miktarı ve veri üzerinden hem raporlama hem de analiz yapma ihtiyacı veri ambarlarından bahsi zorunlu kılmaktadır. Bu bahiste, büyük veriye de zaten değinilmiş olmaktadır. Özetle, bu aşamadan sonra anlatılacak kavramların iç içe geçmesi ve birbiri ile bağlantılı olması bazen geriye dönüş ihtiyacını ortaya çıkaracaktır. Böylece veri analiz sürecinin gelişim süreci daha doğru anlaşılacaktır.

## 1.2. VERİ TABANI YÖNETİM SİSTEMİ

Veri tabanı yönetim sistemi, veri tabanlarının oluşturulması, düzenlenmesi, güncellenmesi, veri tabanına yeni veri eklenmesi veya veri tabanından veri silinmesi, verilerin işlenmesi için kullanılan sistemlerdir. Veri tabanı yönetim sistemi, veri tabanları kullanılarak yazılım geliştirme, performans takibi, güvenliğin sağlanmasını kolaylaştırmayı amaçlamaktadır. Her şirket, kurum ihtiyacına göre, farklı veri tabanı sistemleri kullanabilmektedir. Kullanılacak veri tabanı sistemi, kullanıcı sayısı ve maddi olanaklara göre belirlenir ve buna göre bir seçim yapılır. Örneğin Access Veri Tabanı Microsoft Office Access uygulaması kurulup yönetilebilmektedir (Akcay, 2019).

Veri tabanları ilişkisel ve ilişkisel olmayan veri tabanları olarak ikiye ayrılır. MSSQL, PostgreSQL, MySQL, Oracle, dBase, Informix, Ingres ilişkisel veri tabanlarına örnek olarak verilebilir (Berg et al., 2013). İnternetin yükselişe geçmesi ve ortaya çıkan veri hacmi ve bu verinin etkin bir şekilde işlenememesi ilişkisel olmayan veri tabanlarını (NoSQL) da ortaya çıkarmıştır. Google ve Amazon NoSQL sistemleri olan Big Table ve DynamoDB'yi kullanmaktadır. Google devasa verilerini ilişkisel veri tabanında değil, Big Table üzerinde tutmaktadır.

Veri tabanı yönetim sistemi hem operasyonel veri tabanını hem de veri ambarını kapsadığından genel bir ön bilgilendirme yapma ihtiyacı duyulmuştur.

### 1.2.1. Operasyonel Veri Deposu (ODS)

İngiltere, 1943 yılında İkinci Dünya Savaşı sırasında düşmanın gizli haberleşme kodlarını çözmek için teorik bir bilgisayar ve ilk veri işleme makinelerinden birini icat etti. İlk büyük ölçekli elektronik bilgisayar olan Colossus, büyük hacimli verileri analiz etmek için kullanıldı (Innova, 2021). 1965'te ABD'de vergi beyannamesi ve parmak izlerinin manyetik bantta saklanması için ilk veri merkezleri inşa edildi. 1996 yılında dijital veri depolama, kâğıda bilgi depolamaktan daha az maliyetli hale geldi. Veri depolama maliyetinin düşmesi ve bilgisayarların popüler hale gelmesiyle veri tabanı alanında yeni bir dönem başlamış oldu (DuCharme, 2005).

Veri tabanı bilgilerin bilgisayar sisteminde elektronik olarak saklandığı düzenli bir koleksiyondur. Veri miktarı arttıkça, bilgisayarlarda veri depolama için kullanılan geleneksel yöntemler yetersiz kalmaya başlamıştır. Böylece verilerin kolay erişilebilir, değiştirilebilir, güncellenebilir özetle yönetilebilir şekilde saklandığı veri tabanları ortaya çıkmıştır.

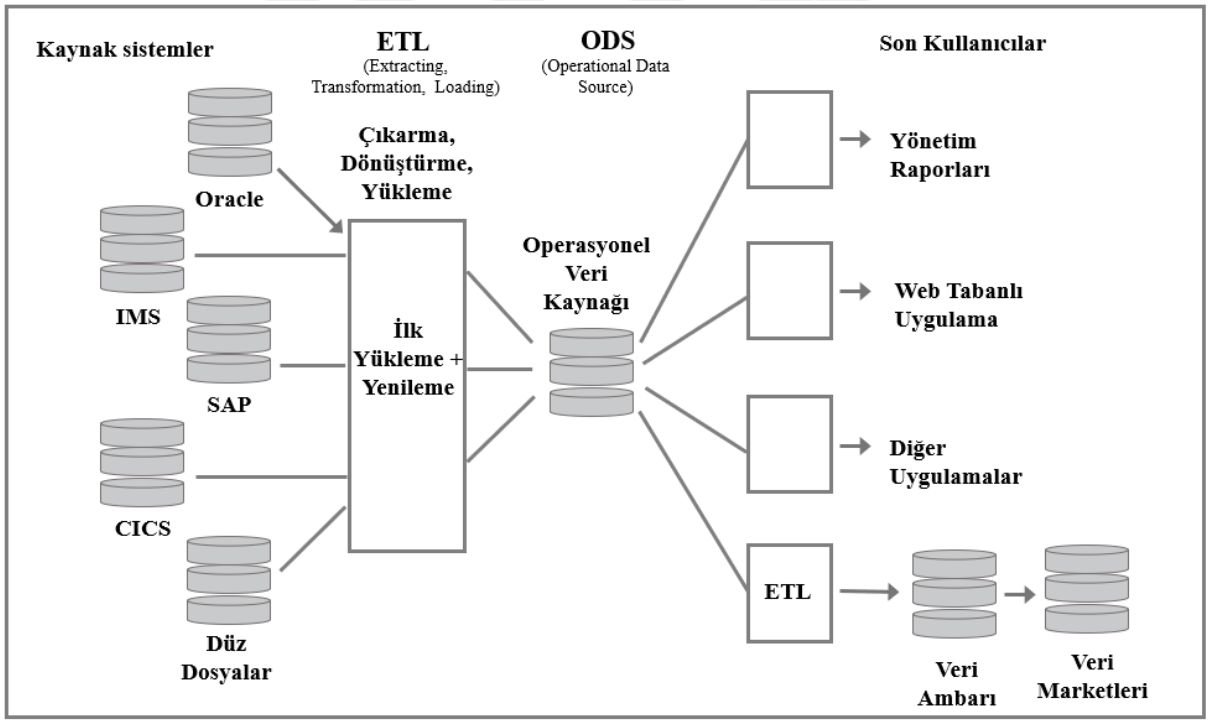
Veri tabanlarının ortaya çıkışı 1960'lı yıllara dayanmaktadır. Veri tabanları ortaya çıkışından bu yana, değişen ihtiyaçlar doğrultusunda radikal değişimler geçirmiştir. Hiyerarşik veri tabanı gibi navigasyonel veri tabanları (ağaç benzeri bir modeli temel alır ve yalnızca bir kaynaktan çoklu alıcıya doğru ilişkiye olanak sağlar) ve ağ veri tabanı (çoklu ilişkilere olanak sağlayan daha esnek bir model), verileri depolamak ve değiştirmek üzere kullanılan orijinal sistemlerdi (Oracle, 2022b).

Veri tabanlarının ilişkisel ve ilişkisel olmayan veri tabanları olarak ikiye ayrıldığından bahsetmiştik. İlişkisel veri tabanı modeli, 1970 yılında E.F. Codd tarafından tasarlanmıştır (Berg et al., 2013). 1980'li yıllarda ise popülarite kazanmıştır. İlişkisel veri tabanındaki veriler, sütunlar ve satırlardan oluşan bir tablo kümesi şeklinde tutulur. İlişkisel veri tabanları şirketlerin, yaptıkları işlemleri bağlantılı olarak yorumlamasına ve takip edebilmesine olanak tanıyan mimarilerdendir. Böylece bir şirket, sattığı ürünü, miktarını, kime sattığını, fiyatını ve buna benzer birçok metriği tek bir yerden ilişkili bir şekilde takip etme imkanına sahip olmaktadır. İlişkisel veri tabanlarında verileri sorgulamak, değiştirmek ve tanımlamak ve aynı zamanda erişim kontrolü sağlamak üzere SQL (Structured Query Language) programlama dili geliştirilmiştir. SQL, yapılandırılmış verilerden veri çekme, düzenleme, işlemi gerçekleştirir. Kullanıcının SQL komutuyla yaptığı her işleme *veri tabanı sorgusu* (Database Query) denir.

1985'li yıllarda ise nesneye dayalı veri tabanları ortaya çıkmıştır (Berg et al., 2013). Nesne odaklı veri tabanındaki bilgiler, nesneler halinde temsil edilir. Yüksek performanslı işleme ihtiyacının olduğu alanlarda, iş alanlarında bu veri tabanı daha işlevsel olmaktadır. Finansal hizmetler, telekomünikasyon, mühendislik gibi alanlarda talep görmektedir. Bununla birlikte, Python, Java, C#, Visual Basic gibi nesne yönelimli programlama dilleri nesne yönelimli veri tabanının popülaritesini artırmıştır (Netinbag, 2022).

Inmon and Imhoff (1996) Operasyonel Veri Deposunu (Operational Data Store-ODS), nesne-odaklı, entegre, uçucu, güncel değerli verileri içeren bir veri deposu olarak tanımlamıştır. Bu veri deposunun nesne odaklı olması, veri tabanının organizasyonun ana veri varlığı üzerine inşa edilmesi anlamına gelmektedir. Örneğin bir üniversitenin veri nesnesi, öğrenciler, öğretmenler, dersler iken; bir şirketin nesnesi, çalışanlar, hizmetler olabilir. ODS birden çok sistemden gelen verileri birleştirmektedir. Bu yüzden entegredir. ODS eski verileri içermediğinden günceldir. Kısa süreli bir kurumsal bellek olmasından dolayı uçucu olarak tanımlanmaktadır. ODS veri ambarı oluşturmak için kullanılabilecek ara bir katman olabilir (Bhatia, 2019).

ODS'deki veriler, birden çok veri kaynağından çıkarma, dönüştürme ve yükleme (Extraction, Transformation, and Loading (ETL)) yoluyla alınır. ETL işlemi farklı şekillerde yapılabilir. Kaynaktan veri alındıktan sonra, T-SQL, PL/SQL gibi sistemler ile veri dönüştürme işlemi yapılarak veri hedef sisteme yüklenir. Buna karşın, ETL işlemi SSIS, Data Integrator, Power Center gibi bir ETL araçları yoluyla yapılabilir. ETL aracı kullanmak daha maliyetlidir ancak bir kodlama işlemi gerektirmediğinden bu araçla çalışılması daha kolay olabilmektedir. Veri dönüştürme işlemi verinin belli bir standarda tutulmasını sağlar. Böylece veri çekme işlemi daha hızlı ve etkin olmaktadır (Bhatia, 2019).




**Şekil 1.**Operasyonel Veri Deposu Mimarisi

**Kaynak:** (Bhatia, 2019), syf 391

### 1.2.2. NoSQL

Yapılandırılmış veri türlerinin yanında, web sayfaları, sosyal medya kaynakları, mailler gibi kaynaklar yapılandırılmamış veri tipine örnek olarak verilebilir. Bu veri türünün standart bir formu

bulunmamaktadır. İnternetin ivme kazanması ve yapılandırılmamış verilerin daha hızlı bir şekilde işlenmesine duyulan ihtiyaç NoSQL veri tabanlarının ortaya çıkmasına neden olmuştur

1998 yılında "NoSQL" sözcüğü ilk defa Carlo Strozzi tarafından kullanılmıştır. Geliştirdiği veri tabanının sorgulama dili olarak SQL'i kullanmadığını belirtmek isteyen Strozzi, açık kaynak kodlu veri tabanı için NoSQL DB (Database) ismini kullanmıştır. NoSQL'in genel kabul görmüş bir tanımı bulunmamakla beraber, bazı özellikleri ile açıklamak mümkündür (Bhatia, 2019).

NoSQL, ilişkisel olmayan bir modeldir ve genellikle açık kaynaklıdır. 21 yy.'ın başında oluşan veri web ihtiyaçlarını karşılamak için kullanılmaktadır. NoSQL'de bir şema olmadan çalışma imkânı bulunmaktadır. Bu durum da aynı türde olmayan verilerle çalışma kolaylığı sağlamaktadır (Oracle, 2022a).

### 1.2.3. Veri Ambarı

Amerikalı bilgisayar bilimcisi Bill Inmon veri ambarı kavramının isim babası olarak bilinmektedir. Inmon veri ambarını nesne odaklı, entegre, kalıcı ve yöneticiler için karar verme sürecini destekleyen bir veri koleksiyonu olarak tanımlamaktadır (Bhatia, 2019).

Merkezi bir veri havuzu olan veri ambarı, özel olarak hızlı sorgulama ve analiz amaçlarıyla tasarlanmış bir veri tabanı türüdür (Oracle, 2022b). Şirketler veri tabanlarından günlük işlemler gerçekleştirirken bir yandan raporlamalar da yapmak istemektedir. Bu durum sistemlerde aksamalara neden olmaktadır. Bu durum *veri ambarları* ihtiyacının ortaya çıkmasına neden olmuştur.

Veri ambarı geçmiş verilerin tutulduğu tarihsel bilgilerin yer aldığı bir veri tabanıdır. Kurumların uzun süreli bellekleri olarak düşünülebilir. Veri ambarlarından yönetimsel sorgulamalar yapabilmek için Çevrimiçi Analitik İşleme (Online Analytical Processing-OLAP) sistemi kullanılır. Böylece yöneticilerin cevap aradığı sorulara yanıtlar bulunabilmektedir. OLAP sistemi sayesinde, "En etkin kanallar hangileridir?", "En yüksek marjlı müşteriler kimlerdir?" gibi sorgulamaların cevapları elde edilebilmektedir.

Veri ambarı veri erişimi ve veri temizlemeye olanak tanır. Çevrimiçi analiz ve karar desteği için kullanılır (Fayyad et al., 1996). Şirketler veri depolama ve temizleme sorunu çözdükten sonra veri ile ne yapabilecekleri sorusunu sormaktadırlar. Bu noktada Bilgi Keşfi (Knowledge Discovery on Databases- KDD) süreci ortaya çıkar. Veri ambarlarının analizi için Codd et al., (1993) tarafından çevrimiçi analitik işleme uygulaması olan OLAP (Online Analytical Processing) uygulaması önerilmiştir. Codd, (1993) OLAP'ı "büyük hacimli çok boyutlu verilerin dinamik sentezi, analizi ve konsolidasyonu" olarak tanımlamıştır. OLAP, SQL'den daha üstün olanaklar tanıyarak çok boyutlu veri analizi yapılmasına olanak sağlamaktadır. (Bhatia, 2019)

OLAP fiziksel veri tabanından faydalı bilgiler çıkarmak için tasarlanmış bir sistemdir. Birden çok kaynaktan veri toplar ve yöneticilere özetlenmiş bilgiler sunar. OLAP'a ek olarak veri madenciliği

de yöneticilere bilgi sunulmasına hizmet eder ancak aralarında bazı farklılıklar vardır. OLAP verileri özetleyip tahminler sunar. Örneğin, "Bölgeye ve yıla göre ortalama satışlar nedir?" soruna OLAP cevap verir. Veri madenciliği ise, gizli örüntüleri bulmaya çalışır. Örneğin, telekomünikasyon şirketlerinde müşteri kaybı önemli bir metriktir. Veri madenciliği "Hangi müşteriler aldığı bu hizmeti değiştirecek ve nedeni nedir?" sorusuna cevap arar. Özetlemek gerekirse, OLAP "Bu doğru mu?" sorusuna cevap ararken, veri madenciliği "Bu olay neden gerçekleşti?" sorusuna cevap aramaktadır (Bhatia, 2019). Böylece OLAP ve veri madenciliği birbirini tamamlar ve yönetimin daha iyi karar almasına yardımcı olur.

Veri tabanları ile veri ambarı arasındaki en temel fark kullanım amaçlarıdır. Veri tabanı günlük operasyonel verileri tutarken, veri ambarı analiz ve raporlama için işletmelerin geçmişe dönük verisini saklamaktadır. Veri tabanlarından da analiz yapmak mümkündür ancak veri ambarına göre daha karmaşık sorgulara ihtiyaç duyulmaktadır. Veri tabanlarında tekrarlı verilerin oluşmaması için tablolar birbirleriyle ilişkilendirilir. Veri ambarlarında böyle bir durum söz konusu değildir. Veri tabanlarının 7/24 erişebilir olması büyük önem taşımaktadır. Veri ambarları yüksek hacimli analitik işlemler için tasarlanmıştır (OLAP vb.). Veri tabanlarında veriler işlem odaklı tutulurken, veri ambarlarında işlemler nesne odaklı tutulmaktadır. Veri tabanlarından Online Transaction Processing (OLTP) sistemi kullanılırken, veri ambarlarında OLAP sistemi kullanılmaktadır.

| **OLTP Sistemi (Veri Tabanı)** | **Veri Ambarı Sistemi** |
|---|---|
| Güncel veriler tutulur | Sadece tarihsel veriler tutulur |
| Ayrıntılı veriler saklanır | Ayrıntılı veya derinlemesine özetlenmiş verileri depolamak için yeteri kadar esnektir |
| Verinin doğası dinamiktir | Verinin doğası durağandır |
| Veri işleme doğası gereği özyinelemelidir | Veri işleme yapılandırılmamış ve geçicidir. |
| İşlem çıktısı yüksek seviyededir | İşlem çıktısı, çoğunlukla düşük veya orta seviyededir |
| Kullanım şekli tahmin edilebilir | Kullanım şekli tahmin edilemez |
| İşlem odaklıdır | Analiz odaklıdır |
| OLTP sistemleri uygulamaya yöneliktir | Nesneye daha çok odaklanırlar. |
| Günlük kararlar desteklenir | Uzun vadeli veya stratejik kararlar alınır |
| Çok sayıda operasyonel kullanıcıya hizmet vermektedir. | Çoğunlukla yöneticilere hizmet vermektedir |

**Şekil 2.** OLTP Sistemi (Veri Tabanı) ile Veri Ambarı Sisteminin Karşılaştırılması

**Kaynak:** (Bhatia, 2019) , syf 402

### 1.2.4. Data Mart

Veri ambarlarının yanında data mart denilen veri tabanlarından da bahsedilebilir. Data Mart şirketlerdeki farklı departmanlar için oluşturulmuş özel veri tabanları olarak düşünülebilir. Veri tabanlarının bir alt formu olarak kullanılmaktadırlar. Örneğin bir data mart sadece finans veya sadece

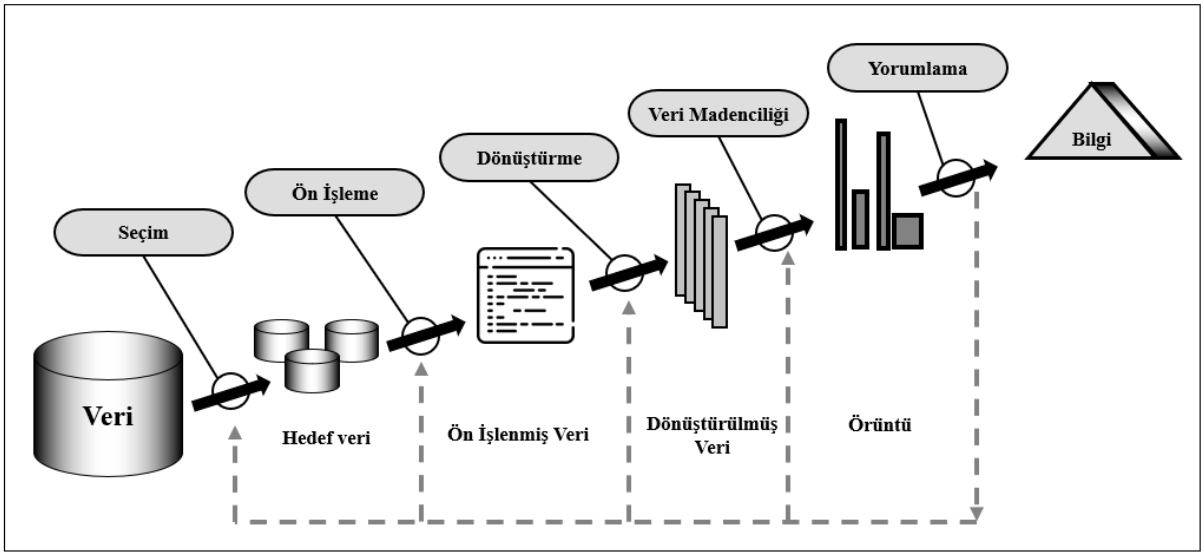
satış biriminin verilerini içerebilmektedir. Data mart sayesinde veri erişimi kolaylaşmaktadır (Kezay, 2020).

## 1.3. VERİ TABANLARINDA BİLGİ KEŞFİ

Bilgi keşfi, veri tabanlarından örtük, gizli, önceden bilinmeyen ve potansiyel olarak yararlı bilgilerin çıkarılmasıdır (Frawley et al., 1992).

Geçmişten bugüne verilerden yararlı bilgiler elde etme birçok isimle adlandırılmıştır. Bilgi çıkarımı, bilgi keşfi, bilgi toplama, veri arkeolojisi, veri deseni işleme, veri madenciliği bu isimlerdendir. İstatistikçiler, veri analizciler ve yönetim bilişim sistemleri en çok veri madenciliği terimini kullanmıştır. Veri keşfi veriden yararlı bilgiler elde etme sürecinin tamamını kapsarken, veri madenciliği tüm bu süreçteki belirli bir adımı ifade eder. Veri madenciliğinde veriden örüntüler çıkarmak için özel algoritmalar kullanılır (Fayyad et al., 1996).

Fayyad vd., (1996) veri tabanlarında veri madenciliği ve bilgi keşfinin hem birbiriyle olan hem de makine öğrenmesi, istatistik ve veri tabanları gibi ilgili olduğu diğer alanlarla olan ilişkisini ele almıştır. Fayyad et al., (1996) dijital çağın ya da başka bir deyişle bilgi çağının aşırı veri yüklemesi sorununu ortaya çıkardığını ifade etmiştir. Bu noktada veri tabanlarından bilgi keşfi (KDD)'nin bu sorunu çözme girişimi olduğunu dile getirmektedir. Fayyad'ın aşırı veri yüklemesi (data overload) olarak tanımladığı, bugün büyük veri dediğimiz verinin boyutunun artması veri tabanlarından bilgi keşfi ihtiyacını ortaya çıkarmıştır.




**Şekil 3.**Veri tabanlarından bilgi keşfi

Kaynak: (Fayyad et al., 1996)

İşletmelerde ana veri madenciliği ve bilgi keşfi uygulama alanları pazarlama, finans, sahtekarlık tespiti, üretim, telekomünikasyon alanlarında gerçekleşmektedir. İnternetin ortaya çıkması ve gelişimi

başarılı KDD uygulamalarının ortaya çıkmasını sağlamıştır. 1996 yılında Fayyad makalesinde başarılı bir KDD uygulamasına FIREFLY isimli müzik öneri aracını örnek vermektedir. FIREFLY kullanıcıya çeşitli müzik parçaları ile ilgili fikrini sorarak, sonrasında beğenebileceği müzik önerilerinde bulunmaktadır (Fayyad et al., 1996).

Veri tabanlarından bilgi keşfinde sürecin bütününe odaklanılır. Adım adım verinin nasıl saklandığı/depolandığı, veriye nasıl erişim sağlandığı, büyük veri setlerine algoritmaların nasıl uygulanacağı ve nasıl verimli çalışacağı, sonuçların nasıl görselleştirilmesi ve yorumlanması gerektiği gibi adımların tamamı bilgi keşfi sürecinde yer alır. Böylece insan-makine etkileşimli olan bu sürecin ne şekilde modellenebileceği üzerine çalışılır.

Veri tabanlarından bilgi keşfi sürecinin ilk adımı, uygulama yapılacak alan ile ilgili ön bilgi edinmek ve bu sürecin amacını belirlemektir. İkinci olarak hedef bir veri seti seçilmelidir. Üçüncü adımda veri temizleme ve ön işleme aşaması gelmektedir. Bir sonraki adımda verilerin azaltılması ve bir projeksiyon çizilmesi gerekmektedir. Böylece veriyi temsil etmek için faydalı özellikler elde edilecektir. Beşinci adımda veri keşfi sürecinin amacı belirli bir veri madenciliği yöntemiyle eşleştirilir. Sınıflandırma, regresyon, kümeleme gibi veri madenciliği yöntemlerinden birisi seçilebilir. Altıncı adımda keşifsel bir analiz yapılarak model ve hipotez seçilir. Bu süreçte hangi modellerin ve parametrelerin uygun olabileceğine karar verilir. Yedinci adımda veri madenciliği yapılır. Sonraki adımda elde edilen örüntüler görselleştirilir ve yorumlanır. Eğer elde edilen sonuçlar istenilen düzeyde değilse, ilk adıma yeniden gidilir ve süreç tekrarlanır. Sonuçlar yeterli görülünce bir sonraki adıma geçilir ve keşfedilen bilgi kullanılmak üzere dokümante edilir ve ilgili taraflara raporlanır.

## 1.4. VERİ MADENCİLİĞİ

Veri madenciliği, akıllı matematiksel algoritmalar kullanarak gelecekteki olayların gerçekleşme olasılığını tahmin etmeye çalışan, bilgisayar bilimleri alanında sıklıkla kullanılan popüler kavramlardan birisidir. Veriden bilgi keşfi, veri tabanlarından bilgi keşfi olarak da bilinmektedir.

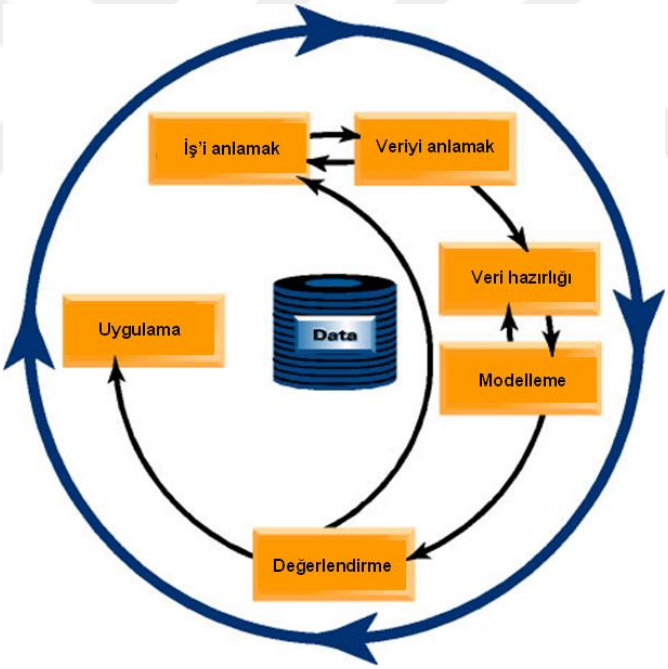
Veri madenciliğinin yaklaşık 50 yıllık bir geçmişi vardır ancak internetin ortaya çıkmasıyla önemli bir teknoloji haline gelmiştir. Veri tabanlarının 1960'larda kullanılmaya başlanması ve 1990'larda internetin hayatımıza girmesiyle üretilen veri miktarı artmaya başlamış ve veri madenciliği de yükselişe geçmiştir. Çağımızda veri miktarı çok yüksek olmasına rağmen bu veri değerli bir bilgiye dönüştürülmediği sürece, veri içinde yüzer ama bilgi açlığı çekeriz. Bu noktada veri madenciliği değerli bilgi yokluğu probleminin çözülmesine yardımcı olmaktadır (Bhatia, 2019).

Veri madenciliğinde küçük veri boyutları ile çalışılabildiği gibi, büyük veri setleri ile çalışmak daha kesin sonuçlar elde edilmesine olanak tanımaktadır. Veri madenciliği sayesinde büyük hacimli verilerden, daha önce keşfetmediğimiz varlığını bilmediğimiz bilgiler, başka bir deyimle örüntüler elde ederiz. Bu büyük veri hacmi, organizasyonların veri ambarlarındaki tarihsel verileridir (Bhatia, 2019).

Ganesh, (2002) verilerin büyük boyutlu olmasından bahsederken veri boyutunun terabayta kadar çıkabildiğini aktarmıştır.

Modern veri tabanları milyonlarca kayıt içerdiğinden standart istatistiksel teknikler yetersiz kalmaktadır. İstatistiksel tekniklerin yetersiz kalması elbette istatistik bilimini saf dışı bırakmamaktadır. Veri madenciliği ve istatistik arasında yakın bir ilişki bulunmaktadır. Veri madenciliğinde devasa kaynaklardan verinin çekilmesi ve analiz edilerek sunulması sürecinde istatistik ve veri madenciliği ortak bir çalışma içerisindedir. Veri madenciliğinde istatistiksel yöntemler olan kümeleme analizi, yapay sinir ağları, diskriminant analizi, karar ağaçları, regresyon analizi, box-jenkins metodları, korelasyon analizi, birliktelik kuralları kullanılmaktadır (Tüzüntürk, 2010). Veri madenciliği aynı zamanda veri tabanları disiplini ile de yol almıştır. Veri madenciliği, veri tabanı, veri ambarı, makine öğrenimi ve istatistik gibi alanların ortak sınırlarında yer almaktadır (Ganesh, 2002; Shmueli et al., 2020).

Veri madenciliği, sağlık, telekomünikasyon, finans (bankacılık, borsa), dolandırıcılık tespiti, pazarlama, sigortacılık, biyoloji, tıp, pazar segmentasyonu, trend analizi, müşteri kaybı ve web sitesi tasarımı gibi alanlarda uygulanmaktadır (Gtech, 2022).




**Şekil 4.** Veri Madenciliği Süreci

Kaynak: (Chapman, 1999)

Veri madenciliği süreci, problemin tanımlanması ile başlar. Problem tanımından sonra verinin tanınması ve anlaşılması gerekmektedir. Bu noktada verinin problemin çözümüne nasıl katkı sağlayacağını anlamak önem arz etmektedir. Bir sonraki adımda veri hazırlama süreci başlar. Sonrasında veri modelleme aşamasına geçilir ve model sonucu çıkan sonuçlar değerlendirilir. Son aşama elde edilen sonuçların hayata geçirilmesi aşamasıdır.

Veri madenciliğinin kökenini veri tabanları ve istatistik oluşturmaktadır. Bir veri madenciliği tekniği uygulayabilmek için veri tabanına ve makine öğrenmesine gereksinim vardır. Veri madenciliğinde Şekil 4'te görüldüğü gibi veriyi anlama, modelleme, yorumlama süreci vardır ve adım adım bu süreç takip edilir. Makine öğrenmesinde ise veriler girdi olarak kullanılır ve tahminleme için algoritmalar kullanılır. Özetle, veri madenciliğinde veriyi temizleyen, işleyen bir insan varlığı vardır ve veriden elde edilen örüntüleri anlamaya çalışır. Makine öğrenmesinde ise insan varlığı, algoritmanın tanımlanması, seçilmesi sürecinde vardır. Sonrasında algoritma süreci devralmaktadır.

### 1.4.1. Yapay Zekâ

Makine öğrenmesi kavramına geçmeden önce, makine öğrenmesinin bir üst başlığı olarak adlandırılabilecek olan yapay zekadan kısaca bahsetmek gerekir. McCarthy, (2004) yapay zekayı şöyle tanımlamaktadır: *"Özellikle akıllı bilgisayar programları olmak üzere akıllı makineler yapma bilimi ve mühendisliğidir. İnsan zekasını anlamak için bilgisayar kullanma göreviyle alakalıdır; ancak yapay zekâ, kendini biyolojik olarak gözlemlenebilir yöntemlerle sınırlamaz."* Bu anlamda yapay zekâ insan zekâsı ile rekabet edecek güçte bir teknolojik sistem olarak düşünülebilir. Çalışma sistemini de insanın öğrenme yolculuğunu baz alarak gerçekleştirir ve deneme yanılma yoluyla öğrenir.

Turing, (1950) yayımlanan "Bilgisayar Makineleri ve Zekâ (Computing Machinery And Intelligence)" isimli çalışmasında yapay zekâ tartışmalarının zeminini hazırlamıştır. Turing yayımladığı çalışmasında "Makineler düşünebilir mi?" sorusunu sormaktadır. Turing Testi olarak bilinen çalışmada Turing bir sorgu memurunun bilgisayar yanıtı ile insan metni arasındaki farkı ayırt edip edemeyeceğini test ederek sorduğu soruya cevap bulmaya çalışmaktadır.

1956 yılında John McCarthy düzenlenen bir konferansta yapay zekâ terimini ilk kez kullanmıştır. 1980 yıllarına gelindiğinde, nöral ağlar yapay zekâ uygulamalarında yaygın olarak kullanılmaya başlanmıştır. Nöral ağlar, kendi kendini eğitmek için geri yayılım algoritması kullanmaktadır. 1997 yılında IBM'in Deep Blue isimli yapay zekâ uygulaması dünya satranç şampiyonunu mağlup etmiştir. 2016 yılında ise AlphaGo isimli yapay zeka programı, dünya şampiyonu Go oyuncusunu yenmiştir (IBM, 2020b). Bu gelişmeler yapay zekanın yükselişe geçişinde önemli yapı taşları olarak literatüre geçmiştir.

Bugün sıklıkla kullanılan konuşma tanıma (speech recognition), çevrim için konuşma robotları, bilgisayar görüşü (computer vision), öneri motorları, otomatik hisse senedi alım satımı gibi uygulamalar yapay zekaya örnek olarak verilebilir.




**Şekil 5.** Yapay Zekâ Şeması

Kaynak: (IBM, 2020b)

### 1.4.2. Makine Öğrenmesi

IBM programcısı ve yapay zekâ biliminin öncüsü Arthur Samuel 1952'de "Makine Öğrenimi (Machine Learning-ML)" terimini ortaya atmıştır. Makine öğrenimini, bilgisayarlara programlanmadan öğrenme yeteneği veren alan olarak tanımlamıştır. Makine öğrenmesinin çıkış tarihine bakıldığında yeni bir alan olmadığı görülmektedir fakat terimin ortaya çıktığı dönemde makine öğrenmesi tekniklerinin uygulanabilmesi için bilgisayarlar yeterince hızlı değildi. 1990'larda bilgisayarların büyük depolama alanlarına sahip olması ve işlem güçlerinin artması makine öğrenmesinin yükselişe geçmesine olanak tanımıştır. Bu kapsamda makine öğrenimi web sitelerinde arama yapan kullanıcıları sınıflandırmak için en yaygın kullanılan araç olarak ortaya çıkmıştır.

Gün içinde yapılan internet aramalarında makine öğrenmesi algoritmalarına sıklıkla maruz kalınır. Google ve Bing gibi arama motorları kullanıcıların yaptığı aramalarda makine öğrenmesi algoritmalarını kullanarak web sayfası sıralaması yaparlar ve hangi kullanıcıya hangi reklamın gösterileceğine karar verirler. Facebook ve instagram gibi sosyal ağlar da makine öğrenmesi algoritmalarından yararlanırlar. Bunun yanında bankalar yapılan çevrim içi işlemlerin gerçek mi yoksa sahte mi olduğunu yine makine öğrenmesi algoritmaları sayesinde tespit edebilmektedirler. Gmail, Yahoo, and Hotmail gibi mail ortamı sağlayıcıları da makine öğrenmesi yoluyla, spam olan yani güvenli olmayan mailleri tespit edebilmektedir. Makine öğrenmesinin nihai amacı insan zekasına benzer bir Yapay Zekâ (AI) platformu oluşturmaktır (Bhatia, 2019). Bu anlamda, makine öğrenmesi Yapay Zekâ uygulamasının bir koludur, bir yan uygulamasıdır denilebilir.

Makine öğrenimi, bilgisayar bilimi, yapay zekâ, veri tabanları ve istatistikleri kapsayan disiplinler arası bir alan haline gelmiştir. (Foster et al., 2017). Makine öğrenmesi gelişen ve yükselen veri bilimi alanında kritik bir bileşendir. Makine öğrenmesinde algoritmalar istatistiksel yöntemler kullanılarak

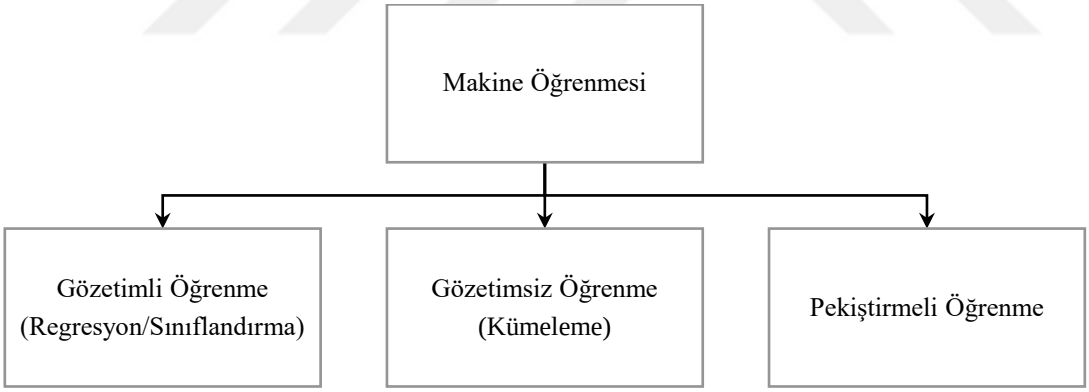
eğitilir ve algoritmaların sınıflandırma, tahminleme gibi işlemler yapabilmesi sağlanır. Böylece makine öğrenmesi kullanılarak veri madenciliği çalışmalarında iç görü elde edilebilmesi sağlanmaktadır.

Makine öğrenmesi algoritmaları üçe ayrılır: Gözetimli/Denetimli Öğrenme, Gözetimsiz/Denetimsiz Öğrenme ve Pekiştirmeli Öğrenme.

Gözetimli/Denetimli öğrenme problemleri, sınıflandırma veya regresyon problemleri olarak kategorize edilir. Veri seti etiketli bir veri setinden oluşur. Sinir ağları, destek vektör makineleri, karar ağaçları teknikleri kullanılır.

Gözetimsiz/Denetimsiz öğrenme, sonuçların nasıl görüneceği hakkında çok az veya hiçbir fikir olmadan problem ele alınmaktadır. Veriler etiketli değildir ve değişkenlerin kesin etkisi bilinmeden sonuçlar elde edilir. Genellikle K-ortalama tekniği kullanılır. Gözetimsiz öğrenmede insan müdahalesi olmadan gizli kalıpların veya grupların keşfedilmesi beklenir.

Pekiştirmeli öğrenme ise, makinelerin insanlara karşı oynadığı bilgisayar oyunları, sürücüsüz araçlar, robot navigasyonu gibi uygulamalarda karşımıza çıkmaktadır. Bu algoritma deneme yanılma yoluyla çalışır. Böylece pekiştirmeli öğrenmede, ideal aksiyon belirli bir bağlamda makinelerce öğrenilir ve en üst düzey performans elde edilmeye çalışılır. Google'ın Go oyununda dünya şampiyonunu mağlup eden AlphaGo programı, Tesla Motors'un kendi kendini süren arabası ve Amazon'un birinci sınıf hava teslimatı, pekiştirmeli öğrenmeye algoritmalarına dayanmaktadır.




**Şekil 6.** Makine öğrenmesi algoritmalarının sınıflandırılması

Kaynak: (Bhatia, 2019), syf 5

Makine öğrenmesinin uygulandığı alanlara bakıldığında, finans, sağlık, eğitim gibi sektörler karşımıza çıkmaktadır. Ses tanıma, otonom araçlar, sahtekarlık tespiti, kişiselleştirilmiş reklamlar, yüz tanıma makine öğrenmesinin kullanıldığı çalışmalara örnek olarak verilebilir. Aynı zamanda, Siri, Alexa, or Google Assistant gibi sanal kişisel asistanlar; trafik sıkışıklığını gerçek zamanlı olarak modelleyen ve gidilecek noktaya varış süresini tahminleyen Google haritalar; kameraların manuel izlenmesini ortadan kaldırmaya yardımcı olan ve insanların uzun süre hareketsiz durması gibi olağandışı durumları veya buna benzer oluşan şüpheli durumlarda görevlileri uyaran, makine öğrenimi tabanlı

video gözetim sistemleri; el yazılarını okumaya ve birbirinden ayırt etmeye yarayan makine öğrenimi tabanlı uygulamalar; sürücüsüz araçlar ve otonom helikopterler makine öğrenmesinin kullanıldığı çalışmalara örnek olarak verilebilir (Bhatia, 2019).

### 1.4.3. Derin Öğrenme

Makine öğrenmesi ve derin öğrenme kimi zaman birbiri yerine kullanılmaktadır. İkisi ortak birçok özellik taşısa da derin öğrenme makine öğrenmesinin bir alt dalı olarak tanımlanabilir. Derin öğrenmenin makine öğrenmesinden ayrıldığı temel nokta, algoritmaların öğrenme şeklidir. Makine öğrenmesinde algoritmalara insan müdahalesi daha fazladır. Derin öğrenmede ise manuel insan müdahaleleri ortadan kaldırılarak makinelerin daha büyük veri kümeleri yardımıyla öğrenmeleri sağlanır. Derin öğrenme *"ölçeklenebilir makine öğrenmesi"* olarak düşünülebilir (IBM, 2020a).

## 1.5. İŞ ZEKÂSI VE BÜYÜK VERİ

1865'te yazar Richard Miller Devens "Cyclopædia of Commercial and Business Anecdotes" adlı kitabında İş Zekâsı (Business Intelligence) kavramını ilk kez kullanır. Devens kitabında, bir bankacının çevresinden gelen bilgileri nasıl kullandığına yer verir.

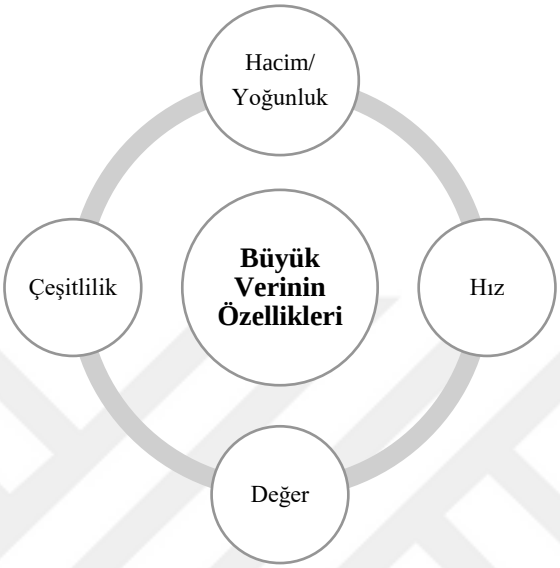
OLAP ve/veya veri madenciliği gibi araçları içeren bir veri ambarı/veri martlarından oluşan bir ortam topluca İş Zekası (BI) teknolojileri olarak tanımlanabilir (Bhatia, 2019).

1969 yılında Dağıtık kontrol ve TCI/IP protokollerini içeren ilk geniş alan ağı olan Advanced Research Projects Agency Network (ARPANET) kuruldu. Bu, günümüz internetinin temelinin atılması anlamına geliyordu. Tim Berners-Lee ve Robert Cailliau, World Wide Web, HTML, URL ve HTTP protokolünü geliştirdiler. Böylece verilere yaygın ve kolay erişim ile internet çağı başlamış oldu (Innova, 2021).

Büyük Veri kavramı, ilk defa 1997 yılında Nasa araştırmacıları Michael Cox ve David Ellsworth tarafından kullanılmıştır (Yıldız & Topal, 2015). 2000 yılı itibari ile büyük veri bilinirliğini artırarak yeni bir dönemin kapısını açmıştır. 2005 yılına gelindiğinde, büyük veri kümelerini depolamak ve işlemek için kullanılan açık kaynaklı yazılım araçları koleksiyonu Apache Hadoop geliştirildi.

Davos'ta 2012 yılında gerçekleştirilen Dünya Ekonomik Forumu'nda veri, bir para birimi veya altın gibi yeni bir ekonomik varlık sınıfı olarak ilan edildi (Peters, 2012). Verinin değeri her geçen gün artarken çağımız büyük veri çağı olarak adlandırılmaktadır (Singh & Reddy, 2015). Büyük verinin yüksek etkinliği olduğundan, büyük veri ile ilgili farklı açılardan çeşitli görüşler vardır. Bir yandan, büyük verinin insanlığın gösterge paneli olacağı, yoksullukla mücadele ve suçlar konusunda önem taşıdığı savunulurken, bir yandan veriye "büyük birader" göndermesi yapılarak daha karamsar bir görüş ortaya konmaktadır (Peters, 2012).

Chris Eaton & Paul Zikopoulos, (2011) büyük veriyi geleneksel veri tabanı yönetimi araçları kullanılarak işlenemeyecek kadar büyük olan veri setleri olarak tanımlar. Bazı araştırmacılar ise, büyük veriyi, özelliklerini göz önünde bulundurarak tanımlar. Gantz et al., (2012)' a göre büyük veri, çok çeşitli ve büyük hacimli verilerden yüksek hızda keşif ve analiz imkânı veren, ekonomik olarak değer elde etmek için tasarlanmış yeni nesil teknoloji ve mimarilerdir. Gantz vd., (2012) veriye dört temel özellik atfetmişlerdir (şekil 6) Hacim (Volume), Çeşitlilik (Variety), Hız (Velocity) ve Değer (Value).




**Şekil 7.** Büyük verinin özellikleri

Kaynak: (Gandomi & Haider, 2015)

Yönetim Bilimi Profesörü Russell Lincoln Ackoff , bilgi piramidi olarak adlandırdığı şemasında verinin geçtiği adımlardaki dönüşümünü ve çıktılarını sırasıyla şöyle sınıflandırmıştır: Veri (Data), Enformasyon (Information), Bilgi (Knowledge) ve Bilgelik (Wisdom) (Bellinger et al., 2004). Bugün bilgi piramidi sıklıkla kullanılmaktadır ve veri analizinde bu pramidin hangi aşamasında olunduğu takip edilebilmektedir.

Farklı kaynaklardan ve çeşitli biçimlerde gelen ve hızla biriken veri, büyük veri analitiğine olan ilgiyi artırmaktadır (Chiang et al., 2018). Büyük verinin popüler olması ve yükselişe geçmesini sağlayan şey verilerin depolanmasını, paylaşılmasını, işlenmesini kolaylaştıran teknolojik gelişme ve ilerlemelerdir (De Mauro et al., 2016). Hadoop gibi popüler teknolojiler (Davenport T.H., Dyché, 2013) ve kullandığın kadar öde imkanı ile maliyet noktasında avantaj sağlayan bulut tabanlı teknolojiler, büyük verinin dolaşıma girmesini kolaylaştırmıştır (De Mauro et al., 2016).

Veri odaklı çalışan şirketler hem finansal hem de operasyonel ölçekte, daha iyi performans ortaya koymaktadır (Mcafee & Brynjolfsson, 2012). Şirketlerde maliyet düşürme, karar alma süreçlerine katkı sağlama, ürün ve hizmetlerde iyileştirme anlamında yol gösteren büyük veri (T. Davenport, 2014) ve büyük veriden analitik yoluyla elde edilen kazanımlar şirketler için sürdürülebilir rekabet avantajı sağlamaktadır (T. H. Davenport, 2006). Örneğin; Marriot büyük veriden yararlanarak optimum oda

fiyatını dinamik olarak hesaplarken, Amazon, tedarik zinciri akışlarını simüle ederek envanter ve stok dışı kalma oranlarını azaltmaktadır (Phillips-Wren et al., 2015). Büyük veri ile çevrim içi işletmeler, müşteri işlemlerinin ötesinde, müşterinin sitede ne kadar kaldıklarını neye tıkladıklarının bilgisine erişebilmektedir. Dinamik promosyonlara verilen yanıtlar, sosyal ağ etkileşimlerinden, duygu ve düşüncelerin paylaşıldığı bloglardan, müşteri hakkında birçok bilgiyi edinilebilmektedir (Phillips-Wren et al., 2015).

Büyük veri çağı olarak adlandırılan günümüzdeki bilginin işlevi ve formu değişmektedir. Akıllı makinalar çağında veri tarihsel veri olarak elde edilirken, bugün gerçek zamanlı verilerle çalışma imkânı bulunmaktadır. Geçmişte veri kaynakları kendi kendine oluşturulmuş, yüksek kalitedeki veriler iken, bugün güvenilir olmayan harici veri kümeleri de dahil büyük miktardaki veriler kaynak olarak kullanılmaktadır. Yine geçmiş dönemde veriye ulaşım sadece stratejik düzeyde kalırken, bugün operasyonel seviyede de veriye erişim imkanı vardır (Berner et al., 2014). Böylece şirketlerde sadece stratejik düzeyde değil, operasyonel düzeyde de kararlar veri ile alınabilmektedir. Öyle ki veri, stratejik bir iş varlığı haline gelmiştir (Miller, 2014).

Büyük veri miktarı artarken bir yandan veri tipleri de değişmektedir. Geçmişte yapılandırılmış büyük veri saklanırken, bugün farklı tipte ve yapılandırılmamış büyük veri, farklı kaynaklarda saklanabilmektedir. Telefondaki bir fotoğraf, çağrı merkezindeki bir ses kaydı, mailler, tweet'ler, mesajlar farklı tipteki büyük veriye örnek olarak verilebilir (Galbraith, 2014).

Büyük veri 3'e ayrılmaktadır. Yapılandırılmış Büyük Veri, Yarı-yapılandırılmış büyük veri ve yapılandırılmamış veri. Oracle, DB2, Teradata, MySQL, PostgreSQL, OLTP sistemleri yapılandırılmış veriye; CSV dosyaları, XML and JSON belgeleri yarı-yapılandırılmış veriye; sosyal medya, telefon mesajları, website içerikleri, fotoğraf ve videolar atmosfer ve uydu verileri ise yapılandırılmamış veriye örnek olarak verilebilir.




**Şekil 8.** Büyük veri tipleri

Kaynak: Gandomi & Haider, (2015)

Büyük veri her geçen gün artarken ve farklı şekillerde karşımıza çıkarken önemi ve değeri de günden güne artmaktadır. 2020 yılında 198,08 milyar dolar olan küresel büyük veri analitiği pazar büyüklüğünün 2030 yılına kadar 684,12 milyar dolara ulaşacağı tahmin edilmektedir (Allied Market Research, 2021). Bu durum, büyük veri analitiğinin önemini açıkta ortaya koymaktadır.

## 1.6. VERİ BİLİMİ

Veri bilimi kavramı 2001 yılında ilk kez William Cleveland tarafından kullanılmıştır (Çelik, 2018). Veri bilimi, yapılandırılmış veya yapılandırılmamış verilerden bilgi ve iç görü elde etmek için bilimsel yöntemler, süreçler, algoritmalar ve sistemler kullanan bir alandır (Vikipedi, 2022).

Büyük verinin ortaya çıkması ile yapılandırılmış veri tipinin yanında yapılandırılmamış veri tipinin de ortaya çıktığından bahsetmiştik. Şirketler Twitter, Facebook gibi sosyal ağlardan, forumlardan, maillerden ve bunlara benzer kaynaklardan gelen yapılandırılmamış verileri de kullanarak iç görü elde etmek istemektedir. Veri alanında çalışanlar, farklı veri kaynaklarından beslenerek büyük veriyi analiz etme, hipotez kurma ve bu hipotezlerin doğruluğunu test etme işlerini yürütürler. Bu kapsamda veri alanında çalışan kişiler, çok yönlü veri setleri ile çalıştıklarından, disiplinler arası yetkinliklere ihtiyaç duymaktadırlar (Vikipedi, 2022).

Veri biliminde dikkat kesilecek nokta veri değil, bilim olmalıdır. Veri bilimi, eldeki verilerle bir sorular yanıtlanabildiğinde gerçekten faydalıdır. Bu, veri bilimindeki bilim kelimesinin tam karşılığıdır denebilir. Veri biliminde temel soru şu olmalıdır: "Elimdeki veri ile mevcut problemlere çözüm bulabiliyor muyum?" Eğer veri problemin çözülmesine imkân tanıyorsa, veri biliminin gerçekleştiği söylenebilir. (Leek, 2013)

İstatistikten veri bilimi yolculuğuna bakıldığında, süreç şöyle özetlenebilir. Bilgisayarların veri depolama amacıyla kullanılması veri tabanlarının ortaya çıkmasını sağlamıştır. Veri miktarının artması ile veri tabanlarının donanımsal olarak gelişmesi ihtiyacı ortaya çıkmıştır. Verilerin tutulacağı ortamların genişlemesi ile veri ambarları geliştirilmiştir. Verilerin depolanması sorunu aşılınca veriye hızlıca erişmek ve bu veriden yararlı bilgiler elde edilmek istenmiştir. Bu durum veri tabanlarından bilgi keşfetme sürecini ortaya çıkarmıştır. Bilgi keşfi süreci genel bir kavramdır ve derinlemesine analiz ihtiyacı olduğu durumlarda yeni kavramların gelişimine ihtiyaç duyulmuştur. Böylece, veriden yararlı bilgilerin elde edildiği ve gizli örüntülerin ortaya çıkarıldığı veri madenciliği ortaya çıkmıştır. 1990'lı yıllarda internet devrimi ile yeni bir dönemin kapısı açılmıştır. İnternetin gelişmesi ve yaygınlaşması ile 2000'li yıllardan sonra büyük veri çağı başlamıştır. Büyük verinin ortaya çıkması her geçen gün artması ile veri bilimi zemini oluşmuştur. Her geçen gün bu zemin yeni gelişmelerle genişlemektedir.

## 1.7. VERİ BİLİMİ YETKİNLİK GEREKSİNİMLERİ

Veri bilimi, mevcut disiplinlerdeki sorunlara yeni veri analitiği olanaklarını göz önünde bulundurarak bakmak için yeni bir paradigmadır (Dumontier & Kuhn, 2017). Büyük veri analitiği, şirketlerin yeni ürünler ortaya çıkarmasına ve firma değerini artırmasına imkân sağlar (Balar et al., 2013). Birçok kuruluş analitiği uygulamaya çalışmasına rağmen, az sayıda kuruluş bu yeterliliğe erişebilmiş durumdadır.

Organizasyonlarda başarı sadece teknoloji kullanımı ile sağlanamaz. İşyerindeki yetenekler de şirket başarısında önemlidir (Gupta & George, 2016). Analitik yeteneklerin bir şirkette nasıl konumlandırılacağını ele alan Grossman & Siegel, (2014) sundukları modelde kültür, çalışanlar, süreçler, yönetişimden oluşan 4 başlık ortaya koymuşlardır. Grossman & Siegel, (2014) 'e göre veri bilimciler, veri ve analitikten anlayan, iş bilgisi olan yani ürün, hizmet ve operasyonu bilen, aynı zamanda bilgi teknolojileri alanına hâkim olup bu alandaki araçları etkin kullanabilen kişilerdir.

Büyük veri ve analitik yetenek tartışması büyük oranda tek bir role -veri bilimcisine- odaklanmış durumdadır (Miller, 2014). Veri Bilimcisi son yılların popüler iş unvanlarından biridir. Şirketlerde birçok çalışan veri bilimcisi olarak nitelendirilebilmektedir. Ancak Veri bilimcisi profilini yeterince tanımlamak (Costa & Santos, 2017) ve bu profili veri alanında çalışan diğer unvanlarla farkını ve benzerliklerini ortaya koymak önemlidir.

Büyük veri ve analitik alanındaki yetenek ihtiyacı sadece veri bilimcisi ile sınırlı değildir. Bu alandaki ihtiyaç veri bilimcilerden ötedir. Çünkü; veri analiz süreçleri iç görü elde etme adımına gelene kadar, birçok farklı aşama barındırmaktadır. Bu aşamaların her biri kendi içerisinde özelleşmekte ve bir sonraki adıma yardımcı olmaktadır. Bu kapsamda veri analizi, verinin toplanması, saklanması, temizlenmesi, işlenmesi, analiz edilmesi adımlarının her birinde ayrı yetenekler gerektirmektedir. Tüm bu süreçleri veri bilimcisi rolüne sığdırmak mümkün görünmemektedir. Öyle ki, veri bilimi alanındaki iş pozisyonlarının her biri temel teknik yetenekler gerektirmektedir. Bu yeteneklerdeki uzmanlık, yeterlilik ölçüleri rolden role, yani iş ihtiyacına göre farklılık gösterebilmektedir. Örneğin veri çekme işi ile ilgilenen bir çalışanın ileri düzey SQL bilgisine sahip olması gerekirken, veri görselleştirme araçlarına ileri düzeyde hâkim olması beklenmeyebilir.

Büyük veri çağında kullanılan analiz araçları, organizasyonların web sitesi, e-mail, mobil uygulamalar, sosyal medya gibi çeşitli kanallardan bilgi almasını kolaylaştırmaktadır (Chong & Shi, 2015). Organizasyonların bilgiye erişim kolaylığının yanında, elde edilen bilginin, bu alandaki yetkin çalışanlarca işlenerek iç görüye dönüştürebilmesi ihtiyacı ortaya çıkmaktadır. Teknolojik gelişmeler ve internetin yükselişe geçmesiyle ortaya çıkan büyük veri, şirketlerin analiz ihtiyaçlarını ve bu alanda yetkin çalışan ihtiyacını derinleştirmiştir. Büyük verinin etkin analiz edilebilmesi bu alanda geliştirilen araçların etkin kullanımı ile mümkün olabilmektedir. Bu kapsamda veri biliminin halen gelişiyor olması ve her geçen gün yeni araç ve teknolojilerin ortaya çıkması, bu alandaki istihdam edilecek çalışanların sahip olması gereken yetkinliklerin tanımlanabilmesini zorlaştırmaktadır. Çünkü çalışanların sürekli yetkinlik gelişimine ihtiyacı bulunmaktadır. Öyle ki, Amazon "2025 Yetkinlik Geliştirme" programı kapsamında çalışanların daha üst yetkinlik gerektiren işler için altı yıl boyunca eğitilmesine yönelik 700 milyon ABD Doları yatırım yapacağını duyurmuştur (McKinsey, 2020).

Küresel ölçekte mevcut otomasyon, yapay zekâ ve dijital teknolojilerin benimsenmesi, dünya ekonomisinin %50'sini etkileyebileceği düşünülmektedir. Bu oran ise 1.2 milyar çalışana denk gelmektedir (McKinsey, 2020).

Türkiye'de 2030 yılına kadar otomasyon, yapay zekâ ve dijitalleşme nedeniyle 7,6 milyon işin kaybolması öngörülürken, 8,9 milyon yeni işin ortaya çıkması beklenmektedir. Türkiye'nin yetenek dönüşümü sırasında 21,1 milyon kişinin, mevcut mesleğine devam ederken yeni yetkinlikler geliştirmesi gerekmektedir. İşgücüne katılacak 7,7 milyon çalışanın ise gerekli yetkinliklerle donatılması önem taşımaktadır. 2030 yılında, öngörülen yetenek dönüşümü gerçekleştiği takdirde, en büyük yetkinlik değişiminin % 63 oranı ile teknolojik yetkinliklerinde gerçekleşmesi beklenmektedir (McKinsey, 2020).

Büyük veri yöneticilerin ajandasında ilk sıralarda ancak hala büyük veri alanında çalışacak kişilerin ilgili pozisyondaki iş tanımı ve yetenek setleri netleştirilebilmiş değildir. Her organizasyon kendi ihtiyaçlarını göz önünde bulundurarak büyük veri alanında pozisyonlar açarak kendi perspektiflerinden iş tanımları yapmaktadırlar. Bu öznel iş ve yetenek tanımları bu alanda çalışan kişiler için ve yine bu alanda ilerlemek isteyen şirketler için kafa karıştırıcı olabilmektedir. Bu durum işletme yöneticilerinin de veri bilimi konusunda net bir bilgiye sahip olmadığını göstermektedir. Zaman zaman, organizasyonlarda veri alanında çalışan ve veri bilimcisi olarak tanımlanan rollere, neredeyse mitolojik bir rol atfedilmektedir. Bu yaklaşım, veri ile çalışma sırasında iç görü çıkarma veya karar alma aşamasına gelene kadar, verinin geçtiği adımları -veri toplama/çekme, veri temizleme, sınıflandırma, organize etme, görselleştirme- neredeyse görmezden gelerek, bu alanda çalışan kişinin şirketin tüm analitik ihtiyaçlarını çözmesi beklenmektedir (De Mauro vd., 2016).

Veri bilimi diğer geleneksel bilimlerden farklı olarak akademik bir perspektiften tam olarak açıklanmamıştır (De Mauro et al., 2016). Provost & Fawcett, (2013) bunun iki sebebi olduğunu ileri sürmektedir. Birincisi veri bilimi, büyük veri ve veriye dayalı karar verme kavramlarıyla güçlü bir şekilde bağdaştırılır ve karıştırılır. İkincisi ise bu veri biliminin daha çok pratik ve deneysel aşamada kalması ve teorik, metodolojik düzeyde ele alınmamış olmasıdır. Bu açıdan bakıldığında veri bilimi alan tanımı, bu alanda pratik çalışma yapanların tanımları ile karıştırılmaktadır (De Mauro et al., 2016). Derin analitik becerilere sahip çalışanlara ve veri okuryazarlığı olan yöneticilere her geçen gün ihtiyaç artmaktadır (Peters, 2012). Veri bilimcisi olarak tanımlananlar için ilk adım bu kişilerin şirkette ne iş yaptıklarıdır. Sonrasında ise bu kişilerin hangi yeteneklere sahip olması gerektiği ve bu yeteneklerin hangi alanda en kolay bulunabildiği sorusu sorulmalıdır (T. H. Davenport & Pat, 2012). Büyük veriyi çevreleyen terimlerin anlamını standartlaştırmak için endüstri ve akademi iş birliği ve hem akademinin hem de endüstrinin ortak hareket etmesi önem arz etmektedir (Carillo, 2017).

Çalışanların hangi yeteneklere sahip olması gerektiği literatürde tartışılagelen bir konudur. Bilgi sistemleri çalışanlarının hangi yeteneklere sahip olması gerektiği inceleme konusu olmuştur. Todd, McKeen ve Gallupe, (1995) sistem analisti, programcı ve bilgi sistemleri yöneticilerinin hangi becerilere

sahip olmaları gerektiğini gazete ilanları üzerinden 20 yıllık periyottaki iş ilanları inceleyerek araştırmışlardır.

Provos ve Fawcett, (2013) veri bilimcisinin daha az veri işleme daha çok veri analitiği tasarımı ve yorumlaması gerektiğini ifade etmektedir. Bunun yanından veri biliminin genel ilkelerinin çerçevesinin geçmişte de çok farklı olmadığını ve ilerde de bunun çok farklılaşmayacağının altını çizmektedir. Bu nedenle veri biliminin genel ilkelerine odaklanmanın önemli olduğunu vurgulamaktadır.

Veri Analitiği veya Veri Bilimcisi rolleri belli bir alanda uzmanlaşmanın yanında, T şeklinde beceri dağılımı denilen geniş bir yelpazede, belli bir düzeyde farklı alanlarda da bilgi ve beceri gerektiren yetkinliklerin bir karışımından oluşmaktadır denebilir. Markow vd. (2017) veri bilimcisi profilini iş analisti gibi benzer profillerden ayırt etmek için kavramsal bir model önermiştir. Veri bilimcisi profilinin multi-disipliner bir profil çizdiğini ve bilgisayar bilimleri, istatistik, matematik gibi farklı alanlardan yetenek ve beceriler barındırdığını ve bu yüzden veri bilimcisini bilgi sistemleri şemsiyesi altında konumlandırmanın uygun olacağını ve aynı zamanda veri bilimcilerin iletişim, iş bilgisi ve merak gibi güçlü sosyal ve kişisel yeteneklere sahip olmaları gerektiğini ifade etmişlerdir.

Costa ve Santos, (2017)'a göre veri bilimcisi, veri ile ilgili işlerde, zorluklara rağmen veriden model ve yaklaşım çıkarabilen kişidir. Veri ile uğraşan pozisyonlarda çalışan kişiler için analitik-teknik becerilerin yanında, sosyal beceriler de önem arz etmektedir.

Bowne-Anderson, (2018) veri ile uğraşan rollerde, iyi bir sunum yapma becerisinin, karmaşık bir makine öğrenmesi becerisinden daha önemli olduğunu dile getirmiştir.

De Mauro vd., (2016) büyük veri ile ilişkili en önemli iş ailelerinin neler olabileceğine netlik getirmeye ve şirketler tarafından aranan ortak yetenekleri tanımlamaya çalışmışlardır. Çalışmalarında büyük veri kelimelerini içeren tüm iş başlıklarını veya iş içeriklerini toplayarak bu verileri metin madenciliği yoluyla analiz etmişlerdir. Veri bilimcisi rolünün, net bir şekilde olmasa da inceledikleri pozisyonlarda üst bir başlık olarak, konumlanabileceğini ifade etmişlerdir. De Mauro vd.,(2016) veri bilimcilerin temel becerilerinin analitik yetkinlik setlerinden oluştuğunu ve organizasyonda diğer birimlerle etkin bir iletişim kurabilmeleri için çalıştıkları işin bağlamını anlamaları ve proje yönetimi tekniklerini bilmeleri, kullanmaları gerektiğini ifade etmişlerdir. Ayrıca bu kişilerin veri tabanlarından veri çekebilmek için sorgu yazabilme becerisine sahip olması gerektiğini belirtmişlerdir.

Philips-Wren vd., (2015) büyük veri analitiği çerçevesini kaynak bulma, hazırlık, depolama, analiz, erişim ve kullanım bileşenlerinden oluştuğunu ifade ederken bu çerçevede 3 tür kullanıcıya işaret etmiştir: İş Kullanıcıları, İş Analistleri ve Veri Bilimcileri.

Markow, W. vd., (2017) Veri Analitiği şemsiyesinde Veri Analisti, Veri Madencisi, İş Zekâsı Analisti, Veri Ambarı Uzmanı pozisyonlarına yer verirken; Veri Bilimcisi şemsiyesinde, Veri Bilimcisi, Veri Mühendisi, İstatistikçi rollerine yer vermişlerdir.

Miller, (2014) veri uzmanlarından bahsetmiş ve bu çalışanların nasıl geliştirileceğine dair birtakım önerileriler dile getirmiştir. Veri bilimcisi konusunda henüz bir standart olmadığını, herhangi bir kişinin kendisini veri bilimci olarak tanıtabileceğini dile getiren Miller, (2014) bugün cevabı henüz tam olarak netleşmemiş sorular sormaktadır. *Veri bilimcisi uygulamalı bir rol müdür, yoksa doktora mı yapması gerekir? Veri bilimcisi dediğimiz kişiler tüm yeteneklere sahip olmalı mıdır, yoksa bu iş bir takım işi midir? Bütün veri bilimciler makine öğreniminde uzmanlaşmış olmalı mıdır? Bir veri bilimcinin ihtiyaç duyduğu temel beceriler nelerdir ve bunlar istatistik veya iş analitiği programlarında öğretilenlerden farklı mıdır, nasıl farklılaşmaktadır?* Miller, (2014) üniversite ve sektörün iş birliği yaparak bu sorulara cevaplar bulması gerektiğini ifade etmektedir.

Bu çalışma kapsamında Miller'in de yukarıda bahsettiği ve diğer araştırmacıların da sorguladığı sorulara, analiz edilecek iş ilanları üzerinden cevaplar aranacaktır. Elde edilen sonuçların literatüre Türkiye örneği üzerinden katkı sağlaması arzu edilmektedir.

# BÖLÜM 2. METİN MADENCİLİĞİ

Metin madenciliği veya metinden bilgi keşfi ilk olarak Feldman & Dagan, (1995) tarafından tanıtılmıştır. Metin madenciliği yapılandırılmamış metinden ilginç bilgilerin çıkarılması sürecini ifade eder (Hotho et al., 2005). Metin madenciliği, metnin veri olarak kullanıldığı (Şeker, 2015) birtakım analiz araçları yardımıyla analizlerin gerçekleştirildiği, bilgi yoğun bir süreç olarak kabul edilen bir veri madenciliği uygulamasıdır (Feldman & Sanger, 2006, p 1).

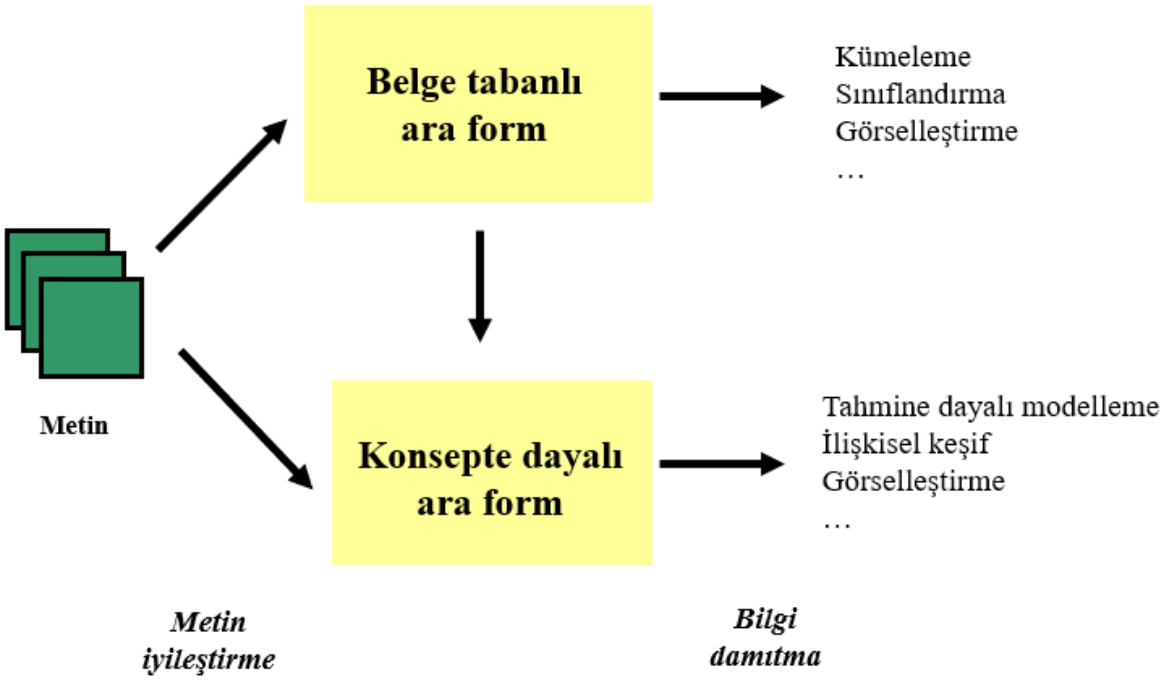
Veri madenciliğinde, verilerin önceden yapılandırılmış bir şekilde depolandığı varsayıldığından, veri ön işleme sırasında iki önemli adım vardır: Verilerin temizlenmesi ile normalleştirilmesi ve çok sayıda birleştirilmiş tabloların oluşturulması. Buna karşılık, metin madenciliğinde, veri ön işleme operasyonları, doğal dil belgeleri için temsili özelliklerin tanımlanması ve çıkarılmasına odaklanır. Özetle, yapılandırılmamış verilerle çalışılan metin madenciliğinde veri ön işleme adımları ile metinler yapılandırılmış ara bir formata dönüştürülür (Feldman & Sanger, 2006, p. 1). Nasıl ki, veri madenciliği verilerden model elde etme olarak tanımlanıyorken, metin madenciliği de metinden model elde etme olarak tanımlanabilir (Witten, 2004a).

Bilgi depolamanın en doğal biçimi metin olduğu için, metin madenciliğinin veri madenciliğine göre daha yüksek bir ticari potansiyele sahip olduğu düşünülmektedir. Yapılan çalışmalar, bir şirkete ait bilgilerin %80'inin metin belgelerinde bulunduğunu göstermektedir. Yapılan araştırmalar iş bilgilerinin %85'inin metin biçiminde tutulduğunu tahmin etmektedir (Hotho et al., 2005). Bununla birlikte, metin madenciliği, doğası gereği yapılandırılmamış ve bulanık olan metin verileriyle uğraşmayı gerektirdiğinden veri madenciliğinden çok daha karmaşık süreçler içermektedir (Tan, 1999).

Metin madenciliği algoritmaları, bir insanın metin dokümanını okumasına gerek kalmadan metni anlamasına yardımcı olmaktadır. Metin madenciliğinde iki süreç metni anlamlandırmada yardımcı olur. Metinde dilin yapısı, cümleler, paragraflar metnin nasıl oluştuğu metnin sözdizimini (syntax) göstermektedir. Bunun yanında tek tek kelimelerin anlamları (semantics) metindeki bağlamları ortaya koymaktadır. Metinden pratik sonuçlar çıkarabilmek için metindeki kelimelerin anlamından bağımsız olarak, sözdiziminden yararlanılabilmektedir. Belge sınıflandırma veya bilgi alma gibi metin madenciliği süreçlerinin arkasındaki algoritmalar, sözdizimsel benzerlikten yararlanarak çalışmaktadır. Bu algoritmalarda sözdizimsel benzerliğin, anlamsal benzerlik anlamına geldiği varsayılmaktadır. (Miner et al., 2012, p. 44)

Metin Madenciliği, verilerdeki trendleri belirlemek için (Lent et al., 1997) çeşitli veri madenciliği yaklaşımları üzerine inşa edilmiştir. Metin madenciliğinde trend analizi, bir koleksiyon içindeki belgelerin tarih ve saat etiketlerine dayanır. Böylece bir döneme ilişkin bir belge alt kümesi ile başka bir döneme ilişkin belge alt kümesi arasında karşılaştırmalar yapma imkânı ortaya çıkar. Belgeler genelinde trend analizi, *iki dönem arasındaki haber konularının genel eğilimi nedir?, Haber konuları*

*hemen hemen aynı mı yoksa iki dönem arasında büyük ölçüde farklı mı?* sorularını cevaplamaya çalışır. (Feldman & Sanger, 2016, p. 9).



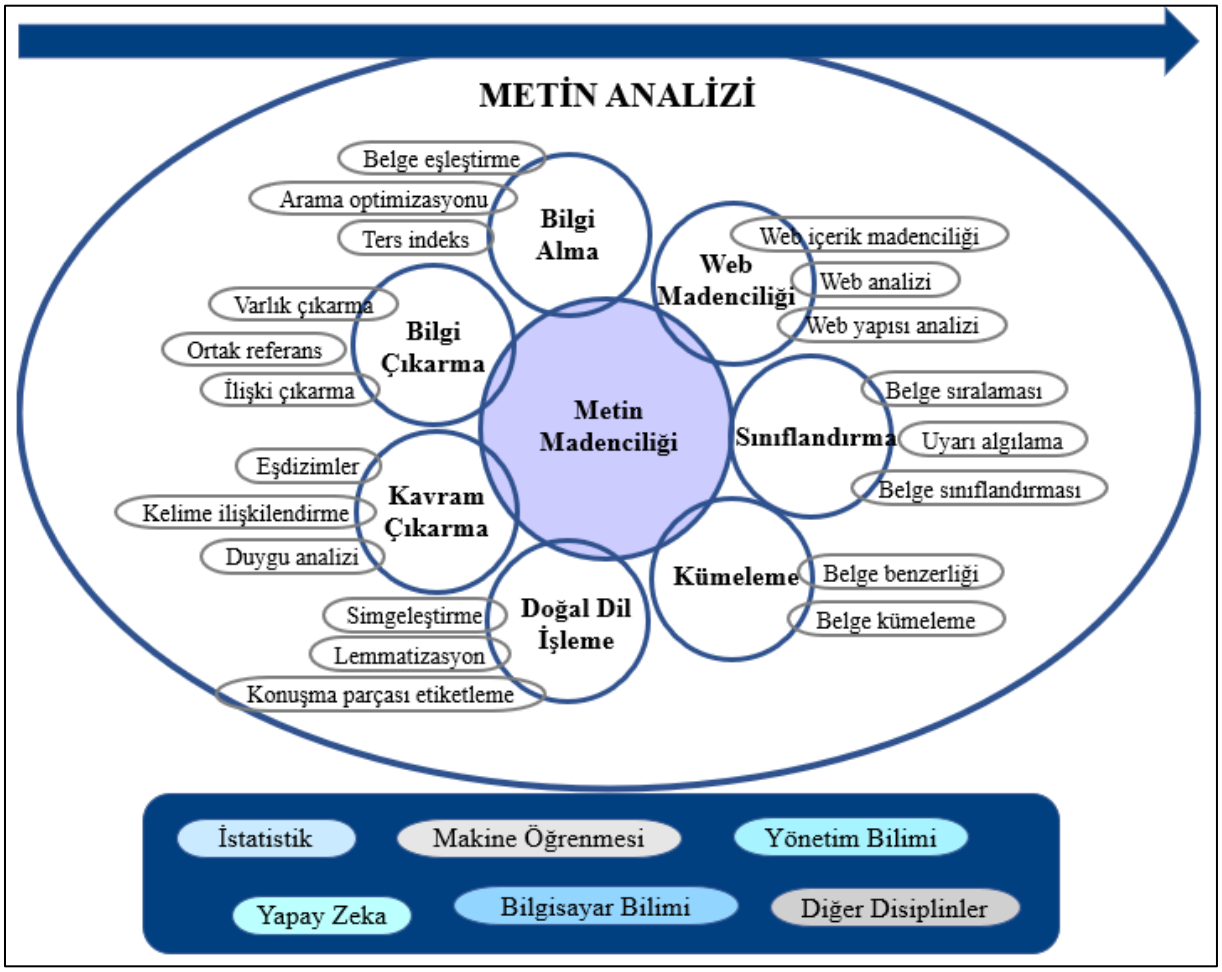

**Şekil 9.** Bir metin madenciliği çerçevesi

Kaynak: (Tan, 1999)

(Miner et al., 2012) metin madenciliği çalışmaları için yedi uygulama alanından bahsetmektedir. Buna göre;

- Bilgi alımı (information retrieval), anahtar kelime aramaları ile metin belgelerine ulaşılmasını ifade etmektedir. Doküman bulma, arama optimizasyonu bu kapsamda gerçekleştirilmektedir.
- Belge kümeleme (document clustering) işleminde, veri madenciliği kümeleme yöntemleri kullanılarak terim, paragraf veya belgelerin gruplama işlemi yapılmaktadır. Kümeleme işlemi sayesinde belgelerin (paragrafların) benzerlikleri tespit edilebilmektedir.
- Belge sınıflandırmasında (document classification), paragraf veya belgelerin kategorilere ayrılması işlemi gerçekleştirilir. Sınıflandırma işlemlerinde mevcut yapıya yeni katılan dokümanların/belgelerin ait olduğu sınıfa atama yapılabilmektedir.
- Web madenciliği (web mining) sayesinde, internet üzerinden çeşitli araçlar yardımıyla veri çekilerek ve metin madenciliği işlemleri yapılabilmektedir.
- Bilgi çıkarma (Information extraction (IE)), yapılandırılmamış ve yarı yapılandırılmış metinden, yapılandırılmış metnin elde edilme sürecini ifade etmektedir. Yapılandırılmamış metinden ilişkiler keşfedilmeye çalışılır.
- Doğal Dil İşleme (NLP), düşük seviyeli dil işleme ve anlam çıkarma işlemlerine imkân tanımaktadır.

- Kavram çıkarma (concept extraction), sözcüklerin ve deyimlerin anlamsal olarak benzer gruplara ayrılmasını ifade etmektedir. Duygu analizi bu kapsamda gerçekleştirilmektedir.




**Şekil 10.** Metin madenciliği için 7 pratik alan

Kaynak: (Miner et al., 2012, p. 40)

Veri tabanlarından bilgi keşfi sürecinde, veri kümesine uygulanması gereken adımlar vardır. Bu adımlar yinemeli olarak gerçekleştirilir ve bu yolla veriden yararlı, işe yarar bilgiler elde edilir. Veri Madenciliği için uygulanan Sektörler Arası Standart Süreç (Crisp DM) olarak adlandırılan Crisp DM Modeli altı adımdan oluşmaktadır (Chapman, 1999). Bu adımlar sırasıyla işi anlama, veriyi anlama, veri hazırlama, modelleme, değerlendirme ve uygulama'dır.

## 2.1. METİN MADENCİLİĞİ SÜRECİ

Etkili metin madenciliği işlemleri, karmaşık veri ön işleme metodolojilerine dayanmaktadır. Aslında metin madenciliği, ham ve yapılandırılmamış veri kaynaklarından temsiller çıkaran veri ön işleme tekniklerine o kadar bağımlıdır ki, metin madenciliğinin bir dereceye kadar bu ayrıntılı hazırlık teknikleri tarafından tanımlandığı bile söylenebilir (Feldman & Sanger, 2016, p. 57). Hazırlık işlemleri veriyi üzerinde işlem yapılacak uygun bir forma dönüştürerek ham verinin üzerinde daha fazla dilsel

işleme yapmaya imkân tanır. Örneğin, işlenecek metin girdisi bir pdf dosyası, taranmış bir sayfa hatta bir ses kaydı bile olabilir. Hazırlık işlemi ile ham girdideki metin etiketlenerek metin paragraf, sütun veya tablolara dönüştürülür.

İnsan dilini işlemenin kesin algoritmaları bilinmemekle birlikte, bazı sistemsel aşamalarla dil işleme tek ve tutarlı bir işlemde birleştirilmeye çalışılmaktadır ancak tatmin edici bir çözüm henüz bulunamamıştır. Bu nedenle metin anlama sistemlerinin çoğunluğu tüm sorunu birkaç alt parçaya bölerek ve bunları bağımsız olarak çözerek geleneksel böl ve yönet stratejisini kullanmaktadır. Alan bilgisi olmadan sadece dilbilimi kullanarak metin analizinde ilerlemek mümkündür. Bu yolla geliştirilen Doğal Dil İşleme (NLP) bileşenleri, genel bir kullanım sunması ve çözüm üretmesi açısından değerlidir (Feldman & Sanger, 2016, p. 60).

### 2.1.1. Metin Ön İşleme

Veri madenciliği süreçlerinde en zorlayıcı adım, veri hazırlığı sürecidir. Metin madenciliğinde, veri kaynağı metin olduğundan dolayı, veri ön hazırlığı adımı özel bir süreç gerektirmektedir (Hotho et al., 2005).

Uysal & Gunal, (2014) metin ön işlemenin metin sınıflandırma üzerindeki etkisini incelemişlerdir. Metin ön işlemenin sınıflandırma, öznitelik çıkarma kadar önemli olduğu sonucunu ortaya koymuşlardır.

Metin madenciliğinde kaynak veri metin olduğu için ve metin verisinin standart bir formatı olmadığından, analiz öncesi metin ön işleme adımları gerekmektedir. Bu adımlar analiz kapsamında kullanılacak metin kaynağının belli bir standarda oturtularak aynı forma getirilmesini sağlamaktadır. Işik & Dağ, (2020) metin ön işleme tekniklerinin etkilerini analiz ettikleri çalışmasında, metnin küçük harflere dönüştürülmesi, stopwords'lerin çıkarılması ve N-gram kombinasyonları sınıflandırma yönteminde diğer ön işleme yöntemlerinden daha iyi sonuçlar verdiğini ortaya koymuştur.

### 2.1.2. Normalizasyon İşlemleri

Yazım kuralları, metinlerin temel yapısını oluşturmaktadır. Bu kapsamda, cümleler büyük harfle başlar. Özel isimler büyük harfle yazılır. Metnin bir veri kaynağı olduğu durumda, büyük harfle yazılan bir kelime ile küçük harfle yazılan aynı kelime farklı veriler olarak algılanmaktadır. Metin ön işleme adımlarında harf dönüşümü ile normalizasyon sağlanmaktadır (Miner et al., 2012, pp. 49–50).

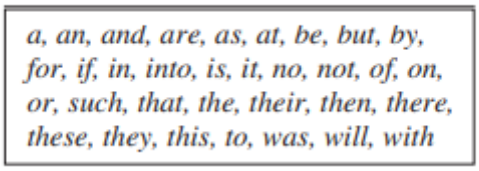
Metinlerde kullanılan noktalama işaretleri, boşluklar veya özel karakterler, analiz sonuçlarına direkt etki etmezler ancak sistem kaynaklarını tükettiklerinden dolayı bunların silinmesi gerekmektedir (Silva & Ribeiro, 2003). Aynı şekilde metinde yer alan sayılar da analiz sonuçlarını etkilemeyeceği düşünülüyorsa bunların da metinden kaldırılması gerekmektedir.

### 2.1.3. Durak Kelimeler (Stopwords)

Metinlerde kullanılan, edat, bağlaç, zarf gibi sözcükler cümle içinde akışı sağlamalarına karşın, metnin analizinde ciddi bir yük oluştururlar. Bu nedende metin analizinde verimlilik nedeniyle durak kelimeler atılmaktadır (Witten, 2004b).

İngilizcede "stopwords" olarak adlandırılan ve Türkçe diline "durak kelimeler" şeklinde çevrilen bu kelime grupları (a, an, are, as, at, be, but..) analizin nihai sonuçları üzerinde çok az etkiye sahip olduklarından bilgi kaybı olmadan bunların kaldırılması mümkündür (Miner et al., 2012, p. 47).

Bir metin veri seti üzerinde çalışıldığında veri setinin indirgenmesi iki önemli avantaj sağlar. Birincisi algoritma performansını iyileştirir. Diğeri ise veri boyutunun azalması, kullanılan depolama alanı ve hesaplamalar için harcanan zamanda tasarruf sağlamaya imkan verir (Silva & Ribeiro, 2003).




**Şekil 11.** İngilizce dilindeki bazı durak kelimeler örneği

Kaynak: Syf 47. Miner 2012

Durak kelimeler için kesin bir çerçeve çizmek mümkün değildir. Genellikle metinde sık geçen kelimelerin anlamlı bir etkisi olup olmadığına bakılır. Buna göre durak kelimeler listesine metinde sık geçen ve analize katkı sağlamayacağı düşünülen kelimeler de eklenir (Cady, 2017, p. 236).

### 2.1.4. Bölütleme (Tokenization)

Bir metinde kullanılan tüm sözcüklerin işlenebilir bir forma getirilebilmesi için, metin madenciliği sistemlerinde, metin kendisini oluşturan parçalara yani kelimelere ayrıştırılarak bölütleme (tokenization) işlemi gerçekleştirilmiş olur (Cady, 2017, p. 60; Miner et al., 2012, p. 47).

Metnin tek tek kelimelere ayrılmasıyla metin bir sözcük akışına bölünmüş olur. Bu süreç noktalama işaretleri, belirteçler ve kısaltmalar nedeniyle karmaşık bir süreçtir (Cady, 2017, p. 233). Bölütleme (tokenization) işlemden sonra, metin daha ileri analizler için uygun hale getirilmiş olur.

Webster & Kit, (1992) çalışmasında doğal dil işleme adımının başlangıcı sayılan bölütleme işleminin önemini ve yapısını ele almıştır.

### 2.1.5. Gövdeleme (Lemmatization)

Aynı köke sahip kelimelerin ortak bir forma dönüştürülmesi işlemi olan gövdeleme algoritması ilk olarak (Lovins, 1968) tarafından çalışılmıştır. Böylece ortak dil sorunları ile başa çıkma yöntemleri teorik ve pratik zeminde tartışılmıştır.

Kök çıkarma işlemi olarak tanımlanan gövdeleme işleminde ön ekler, son ekler kelimelerden çıkarılarak tek bir formda normalleştirme işlemi gerçekleştirilir. Örneğin, İngilizce dilindeki walking, walks, walked, walker kelimeleri eklerden kurtarılarak "walk" şekline dönüştürülür. Kelime boyutlarındaki bu azalmalar, sınıflandırma, kümeleme gibi metin madenciliği görevlerinde algoritmaların çalışmalarında iyileştirmeler sağlar (Miner et al., 2012, pp. 47–48).

Kelimelerin köklerine ayrıştırılması sürecinde Porter kök ayırıcı (Porter stemmer) kullanılan popüler algoritmalardandır. Porter stemmer kelimelerin ortak sonlarını çıkararak, ortak kök elde edilmeye çalışmaktadır (Porter, 1980). Porter stemmer'ın daha iyileştirilmiş hali ise Snowball stemmer'dır (Miner et al., 2012, p. 48).

### 2.1.6. Vektör Oluşturma/Kelime Torbası (Bag of Words)

Doğal dil işlemedeki en temel kavram frekans dağılımı olarak da adlandırılabilecek "kelime torbası" kavramıdır. Serbest bir metnin makine öğrenmesi algoritmalarında kullanılabilmesi için sayısal bir vektöre dönüştürülme işlemidir (Cady, 2017, p. 233).

Metin madenciliğinde, metnin sözdizimsel yapısı ve semantiği için çeşitli yöntemler kullanılsa da çoğu metin madenciliği yaklaşımı, bir metin belgesinin bir dizi kelime ile temsil edilebileceği fikrine dayanır. Metin belgesi içerdiği kelime kümesine göre tanımlanır. Belgelerdeki kelimelerin öneminin tanımlanabilmesi için her kelimeye sayısal bir değer atfedilerek bir vektör gösterimi kullanılır. Buna dayalı olarak, uzay vektör modeli (Salton et al., 1975) yaklaşımı ortaya çıkmıştır.

Metin ön işleme adımları tamamlandıktan sonra, metin madenciliği uygulamaları için kelimelerin bir vektör temsiline dönüştürülmesi gerekmektedir. Bu vektör gösterimi, ikili gösterim veya tamsayı şeklinde gösterilebilir. Aşağıda basit bir örnek ile vektör gösterimi açıklanmıştır (Miner et al., 2012, p. 50).

Doc 1: My dog ate my homework.

Doc 2: My cat ate the sandwich.

Doc 3: A dolphin ate the homework.

Tüm dokümanlardaki kelime sayısı sayılarak aşağıdaki şekilde gösterilir.

a (1), ate (3), cat (1), dolphin (1), dog (1), homework (2), my (3), sandwich (1), the (2)

Tüm kelimeler her bir dokümanda aratılarak, ilgili kelime var ise 1, yoksa 0 olarak işaretlenmektedir. Buna göre ikili (binary) vektör gösterimi aşağıdaki gibi olmaktadır:

Doc 1: 0,1,0,0,1,1,1,0,0

Doc 2: 0,1,1,0,0,0,1,1,1

Doc 3: 1,1,0,1,0,1,0,0,1

Doküman 1'de "my" kelimesi iki kez geçmesine rağmen, 1 olarak kodlanmıştır. İkili gösterimde kelimenin kaç kez geçtiğinden ziyade, var olup olmamasına bakılmaktadır. İlgili kelime var ise 1, yoksa 0 olarak işaretlenmektedir.

Tamsayı gösterimi ise aşağıdaki gibi olmaktadır.

Doc 1: 0,1,0,0,1,1,2,0,0

Doc 2: 0,1,1,0,0,0,1,1,1

Doc 3: 1,1,0,1,0,1,0,0,1

Doküman 1'de "my" kelimesi iki kez geçtiği için, kelime sayısı kadar, yani iki kelime olarak gösterim yapılmıştır.

İkili gösterimde belgede geçen her terim 1 olarak tutulurken, tamsayı gösteriminde o kelimenin sayısı kadar belgede saklanır (Miner et al., 2012, p. 50).

Metin ağırlıklı vektörlere dönüştürüldükten sonra bu vektörler, en sık kullanılan TF-IDF şeması ile saklanabilmektedir. TF-IDF, frekans-ters belge frekans anlamına gelmektedir. Terim sıklığı, terimin belgede görünme sayısını ifade eder. Belge 1'deki "my" kelimesinin terim sıklığı iki olarak ifade edilmektedir.

TF-IDF aşağıdaki şekilde formüle edilmektedir.

TF= (Bir belgedeki t terimin sıklığı) / (Belgedeki toplam terim sayısı)

IDF=log_e (Toplam belge sayısı / t terimini içeren belge sayısı)

$$Idf(d,t) = log\frac{n}{1+df\ (t)}$$

$$df - idf(d,t) = tf(d,t).idf(t)$$

## 2.2. METİN MADENCİLİĞİ YÖNTEMLERİ

Bu bölümde metin madenciliği yöntemlerinden Doğal Dil İşleme (NLP), Bilgi Çıkarma (Information Extraction), Duygu Analizi (Sentimental Analysis), Sınıflandırma (Classification/Categorization), Kümeleme (Clustering) yöntemleri genel hatları ile aktarılacaktır.

### 2.2.1. Doğal Dil İşleme (NLP)

NLP'nin genel amacı, bilgisayarları kullanarak doğal dilin daha iyi anlaşılmasını sağlamaktır (Kodratoff, 1999). Metnin işlenmesi için diğer şeylerin yanı sıra dilbilimsel analiz teknikleri kullanılır. Metin madenciliği otomatik doğal dil işlemenin tamamını ve hatta daha fazlasını kapsıyor gibi görünmektedir. Ancak çoğu metin madenciliği çalışması, doğal dil işlemenin derinlemesine ve bilişsel olan yönlerinden kaçınarak daha basit teknikler kullanır (Witten, 2004a).

Doğal dil işlemenin tarihsel gelişimine bakıldığında, 1940'ların sonları ile 1950'lerin başlarındaki otomatik çeviri projeleri, metin madenciliğinde neden basit teknikler kullanıldığını daha iyi açıklamaktadır. Öyle ki, bahsedilen dönemde otomatik çeviri projelerinde kelime kelime çeviriler ile temel söz dizimsel temelli tekniklerle düzgün ve kullanışlı çevirilerin yapılabileceği varsayılmıştır. Ne var ki bu projelerin sonucunda, doğal dil işlemenin okuma yazma bilmeyen bir çocuğun konuşma seviyesinde bile çözümlenemeyecek kadar zorlu, karmaşık bir yapı olduğu ve algoritmik bir şekilde kodlanmasının ve kullanılmasının oldukça zor olduğu anlaşılmıştır (Lenat, 1995). Araştırmacılar pratikteki zorluklardan dolayı çalışma sistemlerini konudan ve dilden bağımsız olarak kurgulamaktadır. Bu yüzden girdi, dil değil de veri imiş gibi ele alınmaya çalışılmaktadır (Witten, 2004a).

Metin madenciliği olgunlaşma sürecinde iki temel noktaya varılmıştır. Buna göre geniş kapsamlı metinlerde daha sığ/yüzeysel bir metin madenciliği işlemi uygulanması önerilirken, buna karşın herhangi bir alana özgü metinlerde daha derinlemesine analizlerin gerçekleştirilmesi önerilmiştir (Witten, 2004a).

### 2.2.2. Bilgi Çıkarma (Information Extraction)

Bilgi çıkarma (IE), metinde bilgi keşfi sürecinin ilk adımı olarak adlandırılabilir (Soderland, 1999). Bilgi çıkarma yöntemlerinin amacı, metin belgelerinden kişi adı, yer adı, kurum adı gibi belirli bilgilerin çıkarılmasıdır (Allahyari et al., 2017).

Bilgi çıkarma tablo bilgiler içeren yapılandırılmış metinlerden, haber metinleri gibi yapılandırılmamış metinlere kadar değişik türdeki metinlerin analizinde kullanılır. Yapılandırılmış metinlerde bilgi çıkarımı belirlenen etiket seçimlerinin ayıklanması ile gerçekleştirilir. Yapılandırılmamış metinlerde ise metin çıkarma adımına ek olarak sözdizimsel analiz, anlamsal etiketleme, kişi, şirket adlarının tespiti gibi işlemler yapılmaktadır (Soderland, 1999).

Doğal dildeki metinler, bir bilgisayarın doğrudan otomatik analiz yapamayacağı kadar çok bilgi içerir. Bilgisayarlar, büyük miktarda metni elemek ve tek kelimelerden, cümlelerden veya pasajlardan faydalı bilgiler çıkarmak için kullanılabilir. Bu yüzden bilgi çıkarma, ne tür bir bilgi aradığımızı önceden bildiğimiz, tam doğal dil anlayışının sınırlı bir biçimi olarak kabul edilebilir (Witten, 2004b).

Bilgi çıkarma süreci bilgi yoğun bir süreçtir ve bu süreçte her metin için farklı kural setleri oluşturmak gerekmektedir. Bu nedenle makine öğrenimi, bilgi çıkarma sürecinde manuel oluşturulan kural setlerinin oluşturduğu zorluktan dolayı, bilgi çıkarma sürecinde iyi bir alternatif olmaktadır (Soderland, 1999).

Fiyatları ve kısa açıklamaları olan ürün listeleri, iş veya kiralama ile ilgili ilanlar, kurumsal satın alma ile ilgilenen şirketler hakkında haberler bilgi çıkarma sürecine örnek olarak verilebilir (Soderland, 1999).

### 2.2.3. Duygu Analizi (Sentimental Analysis)

Duygu analizi Wilks & Bien, (1983)'nın çalışmalarına kadar uzanmaktadır. Ancak 2001 yılından sonra duygu analizi büyük bir popülerlik kazanmıştır (Miner et al., 2012, p. 21).

Duygu analizi metin içeriğinin, duygusal kutupsallıklar üzerinden sınıflandırılarak olumlu ve olumsuz duygu çıkarımı yapılmasıdır. Basit bir duygu analizi manuel anahtar kelime tanımlamasıyla yapılır (Cady, 2017, p. 239). Ürün, hizmet, film değerlendirmeleri duygu analizi konusu olabilmektedir.

Kaynar et al., (2016) yaptıkları çalışmada film yorum içeriklerini duygu analizi yöntemi ile incelemişlerdir. Naive Bayes, Merkez Tabanlı Sınıflayıcı, Çok Katmanlı Yapay Sinir Ağları (MLP) ve Destek Vektör Makineleri (SVM) sınıflandırma algoritmaları kullandıkları çalışmalarında, yapay sinir ağları ve destek vektör makinalarının diğer algoritmalara göre daha iyi sonuçlar verdiği görülmüştür.

Akgül et al., (2016) twitter verileri üzerinden gerçekleştirdikleri duygu analizinde, tweet'leri olumlu, olumsuz ve nötr olarak 3 kategoride sınıflandırmışlardır. Yaptıkları çalışmada hem sözlük hem de N-gram modeli geliştirmişlerdir. Sözlük modelinin göreceli olarak daha iyi sonuçlar verdiğini ortaya koymuşlardır.

Eliaçik & Erdogan, (2015) finans topluluklarına ait duygu polarite ölçümü için, mikro blog servisleri üzerinden duygu analizi gerçekleştirmişlerdir. Bu analiz ile finansal alandaki kullanıcı verileri kullanılarak finans alanınki insan topluluklarının duygu durumları ile borsa fiyatları arasındaki bağdaşıklık oranı hesaplanmıştır.

Akın et al., (2018) sosyal medya üzerinden bir televizyon kanalına ait seyirci görüşlerini analiz etmişlerdir. Twitter üzerinden alınan verilerin her biri pozitif, negatif veya nötr olacak şekilde etiketlenerek kanalın reyting değerleri incelenmiştir. Analiz sonucunda, seyirciler tarafından atılan pozitif veya nötr duygulu tweet'lerin kanalın reytinginde artış sağladığı, negatif duygulu tweet'lerin ise kanalın reyting değerlerini düşürdüğü ortaya konmuştur.

Alpkoçak et al., (2019) TREMO isimli bir veri seti üzerinden, bir metin sınıflandırma problemi olarak duygu analizi gerçekleştirmişlerdir. Analiz kapsamında Yapay Sinir Ağları (YSA), Destek Vektör Makineleri (DVM), Random Forest (RF) ve K-En Yakın Komşu (k-NN) algoritmaları kullanılmıştır. Mutluluk, korku, öfke, üzüntü, tiksinme ve şaşırma başlıklarındaki her duygu için en önemli ilk 500 kelime Karşılıklı Bilgi (Mutual Information-MI) yöntemi ile çıkarılmıştır. Sınıflandırma sonuçlarında yapay sinir ağlarının en iyi sonucu verdiği gözlemlenmiştir.

Tuzcu, (2020) bir kitap satış sitesinin kullanıcı yorumları üzerinden duygu analizi gerçekleştirmiştir. Python programlama dili üzerinden Çok Katmanlı Algılayıcı (MLP) algoritması uygulamış, daha sonra da rapid miner uygulaması üzerinden aynı veri setini kullanarak Naive Bayes (NB), Destek Vektör Makineleri (DVM) ve Lojistik Regresyon (LR) algoritmaları uygulayarak duygu

analizini gerçekleştirmiştir. Algoritmaların sınıflandırma başarıları karşılaştırıldığı çalışmada Çok Katmanlı Algılayıcı (MLP)'nın en iyi sonuçları veren algoritma olduğu görülmüştür.

İlhan & Sağaltıcı, (2020) twitter üzerinden elde ettikleri veri setini kullanarak makine öğrenmesi yöntemi ile duygu analizi gerçekleştirmişlerdir. Naive Bayes ve Destek Vektör Makineleri algoritmalarından faydalanarak yapılan analizde, Destek Vektör Makineleri'nin daha iyi sınıflandırma sonucu verdiği ortaya çıkmıştır.

### 2.2.4. Sınıflandırma (Classification / Categorization)

Metin sınıflandırma doğal dil belgelerinin önceden tanımlanmış kategorilere atanması işlemidir (Sebastiani, 2002). Metin sınıflandırma meta verilerin otomatik olarak ortaya çıkarılmasını sağlama, bir belgenin kapsadığı konuların tespit edilmesi, web kaynaklarının büyük kataloglarının düzenlenmesi gibi pratik uygulamalar sağlar (Witten, 2004a).

1990'lara kadar metin sınıflandırma işlemleri için uzmanlar tarafından sınıflandırma kuralları oluşturulmaya çalışılmıştır. Bu kurallar yeni belgelere uygulanarak oluşturulacak sistemde bilgi mühendisliği teknikleri kullanılmaya çalışılmıştır (Witten, 2004b).

Metin sınıflandırma işlemleri 1960'lara kadar uzanmaktadır. 1990'lara kadar daha çok bilimsel literatürün indekslenmesi amacıyla kullanılmıştır. 1990 sonrası, dijital metin belgelerinin artması ve belgelerin kullanım kolaylığı için bu belgelerin düzenlenmesi ihtiyacı ortaya çıkmıştır. Günümüzde metin sınıflandırma işlemleri, posta filtreleme, web sayfası sınıflandırma, metin türü tespiti gibi değişik bağlamlarda kullanılmaktadır (Feldman & Sanger, 2006, p. 64).

Hayes et al., (1990) metin sınıflandırma işlemlerini, "true" ve "false" şeklinde iki olası değere sahip ikili değişken kombinasyonlarına dayalı olan bilgi getirimi yöntemi ve bilgiye dayalı sınıflandırma olmak üzere iki sınıfta değerlendirmiştir.

Metin sınıflandırma, metin belgelerini önceden tanımlanmış sınıflar atamayı amaçlar. Bu atama Boolen veri tipi şeklinde gerçekleştirilir. Yani sınıflandırma mantığı 0 ve 1 şeklinde kurgulanmıştır. İlgili metin tanımlı sınıfa ait ise 1, değilse 0 değerini alır (Apté et al., 1994).

Metin sınıflandırmada bilgi mühendisliği (knowledge engineering) ve makine öğrenmesi olmak üzere iki temel yaklaşım kullanılmıştır. Bilgi mühendisliğinde, uzman sistemler (expert systems) yaklaşımıyla kural tabanlı bilgisayar sistemleri tasarlanmaktaydı. Bu yaklaşımdaki temel zorluk, çok fazla nitelikli işgücü ihtiyacı ortaya çıkarmasıydı (Feldman & Sanger, 2006, p. 64). Bu nedenle sınıflandırma işlemleri için daha çok makine öğrenmesine odaklanılmıştır. Günümüzde güncel sistemlerin büyük çoğunluğu sınıflandırma problemlerinde makine öğrenmesi yaklaşımını kullanmaktadır. Makine öğrenmesinde, daha önce sınıflandırılmış örnekler üzerinden makine eğitilerek yeni sonuçları tahmin etmeye çalışmaktadır. Daha az maliyetli olan bu yaklaşımda, manuel olarak sınıflandırılmış eğitim örneklerine ihtiyaç bulunmaktadır.

Metin sınıflandırmanın ilk dönemlerinde CONSTRUE isimli uzman bir sistem kullanılmıştır. Hayes et al., (1990) bu sistemi kullanarak Reuters haber servisinin içeriklerini sınıflandırmaya çalışmışlardır. Bu uzman sistem, haber içeriklerine, konu kategorilerini atamak için, manuel kurallarla yönetilen, kural tabanlı ve %90 üzerinde kesinlik sonucu veren bir sistemdir (Hayes & Weinstein, 1991).

Metin sınıflandırma için makine öğrenmesi uygulamalarında farklı algoritmalar kullanılmaktadır. Naive Bayes, Karar Ağaçları (Decision Tree), Yapay Sinir Ağları (Artificial Neural Network), Örnek Tabanlı Sınıflandırıcılar (Example Based Classifier), Destek Vektör Makineleri (Support Vector Machine) ve İstatistiksel Dil Modeli (Statistical Language Model) tabanlı sınıflandırıcılar yaygın olarak tercih edilen algoritmalardandır.

Metin sınıflandırmanın popüler uygulamalarından birisi e-posta spam filtrelemesidir. Buna göre gelen bir e-posta, ikili (binary) bir sınıflandırma etiketi kullanılarak "spam" veya "spam değil" şeklinde sınıflandırılır (Miner et al., 2012, p. 882).

- Bilgiye Erişim-Enformasyon Getirimi (Information Retrieval (IR))

Bilgi erişimi otomatik indeksleme için ilk girişimler 1975'te yapılmıştır (Salton et al., 1975). IR eski bir araştırma alanı olmasına rağmen, World Wide Web'in yükselişi ve gelişmiş arama motorlarına ihtiyaç duyulması ile artan bir ilgi görmüştür (Hotho et al., 2005). Bilgi alma, direk cevapları sunmaz, sorulan soruların cevaplarını içeren belgelerin bulunmasını sağlar (Hearst, 1999).

Bir metin sınıflandırma yöntemi olarak indeks oluşturulması için anahtar sözcükler üzerinden arama gerçekleştirilir (Tantuğ, 2012). Böylece gereksinim duyulan belgelere erişim belgelerdeki anahtar kelimelerin varlığı üzerinden gerçekleştirir. Bu yöntem kolay uygulanabilir ve etkili bir yöntem olmasına karşın, kavramsal içerikle ilgili bilgi edinilmesi zor bir yöntemdir. Kelime bazlı arama yapılarak sınıflandırma işlemi yapıldığı için bu zorluk ortaya çıkmaktadır. Hayes et al., (1990) bu duruma İngilizce dilindeki "takeover" ve "acquisition" kelimelerini örnek vermektedir. Şirket satın alma anlamında "takeover" kelimesi ile arama yapıldığında, ilgili metnin yazarı bu durumu anlatmak için sadece "acquisition" kelimesini kullanmış olabilir. Problem yaratan bu duruma, metindeki ağırlıklı kelime koleksiyonları kullanımıyla, başka bir deyişle, gelişmiş istatistik tabanlı yöntemlerle (gizli anlamsal indeksleme-(latent semantic indexing) çözüm bulunabilmektedir (Hayes et al., 1990).

Metin sınıflandırma için en yaygın kullanılan algoritmalar, yapay sinir ağları, karar ağaçları, destek vektör makineleridir (Miner et al., 2012, p. 14). Metin sınıflandırma makine öğrenmesi algoritmalarının ilk örnekleri, sinir ağları alanında Li et al., (1991) tarafından; karar ağaçları alanında Apté et al., (1994) tarafından; destek vektör makinaları alanında Cortes & Vapnik, (1995) tarafından gerçekleştirilmiştir.

- Naive Bayes Sınıflandırıcı

Naive Bayes sınıflandırıcı Bayes teoremine dayanan basit olasılıksal sınıflandırıcılardır. Naive Bayes, sınıflandırılacak tüm nesnelerin, özelliklerinin birbirinden tamamen bağımsız olduğu varsayımına dayanır. Bu basit varsayım, değişken sayısı yüksek olduğunda verimli ve doğru sonuçlar üretebilmektedir (Miner et al., 2012)

Olasılığa dayalı sınıflandırma yöntemi olan naive bayes'te sınıflandırma durum değeri olan CSV (d,c)' yi, d belgesinin c kategorisine ait olma olasılığı bayes teoremi ile hesaplanır.

$$P(c\backslash d) = \frac{P(d\backslash c)P(c)}{P(d)}$$

Naive Bayes sınıflandırıcı algoritması, sınıflandıracağı bir öge, atanacak sınıfa ait birkaç öge içerdiğinde, sınıflandırmasını bu yönde yapacaktır. Örneğin, bir kişinin kadın mı yoksa erkek mi olduğunu tespit etmek için, saçlarının uzun olduğu saç tokası kullandığı bilgileri bir kişinin kadın olması lehine orta bir ağırlığa sahiptir. Bayes sınıflandırıcı ise aşırı bir tahminleme ile kişinin kadın olduğunu söyleyecektir. Oysa bu tanıma uyan erkek de vardır. Bayes varsayımları çarpıcı derecede basit bir mantıkta kurgulanmasına rağmen, pratik uygulamalarda etkili sonuçlar vermektedir (Cady, 2017, p. 293).

(Liu et al., 2013) yaptıkları çalışmada film incelemelerini naive bayes sınıflandırıcı kullanarak sınıflandırmışlardır.

- Karar Ağaçları

Metin sınıflandırmanın ilk dönemlerinde kullanılan sınıflandırma teknikleri kural tanımlama ve karar ağaçlarıdır. Apté et al., (1994) çalışmasında WHEAT isimli bir kategoriye belgelerin atanabilmesi için bir kural seti oluşturmuştur. Buna benzer kurallar makine öğrenmesi teknikleri kullanılarak da oluşturulabilmektedir (Witten, 2004b).

Olasılıksal metotlar doğası gereği sayısal sonuçlar verdiğinden insanlar bu sonuçları yorumlamada zorlanmaktadır. Bu durumdan etkilenmeyen algoritma sınıfı karar ağaçlarıdır (Sebastiani, 2002).

Mitchell, (1997, p. 53) karar ağaçları ile metin sınıflandırma örneğini göstermektedir. Çizdiği karar ağacında bir cumartesi sabahının tenis oynamaya elverişli olup olmamasını analiz etmektedir.

- Örnek Sınıflandırma Çalışmaları

Çalış et al., (2012) reklam içerikli e-postaların metin madenciliği yöntemleri kullanılarak sınıflandırılması için bir çalışma gerçekleştirmiştir. 400 normal ve 400 reklam içerikli e-posta Destek Vektör Makinesi, k En Yakın Komşu ve Naive Bayes sınıflandırma algoritmaları yardımıyla sınıflandırılmaya çalışılmıştır. Çalışma sonucunda k En Yakın Komşu (k-NN) algoritması kullanıldığında %96,5 doğruluk oranı ile en iyi sonuç elde edilmiştir.

Yapılan bir diğer metin madenciliği çalışmasında, kullanıcıların web site içerikleri, sınıflandırma algoritmaları kullanılarak analiz edilerek ilgili sayfanın e-ticaret sitesi olup olmadığına karar verilmiştir. K-NN En-Yakın-Komşu ve Naive Bayes sınıflandırma algoritmalarının kullanıldığı çalışmada Naive Bayes algoritmasının daha iyi sonuçlar verdiği görülmüştür (Kaşikçi & Gökçen, 2013).

Zhou et al., (2016) Yazılım süreçlerindeki hata raporlarının sınıflandırılması için metin madenciliği yöntemi kullanmışlardır.

Goh & Ubeynarayana, (2017) Amerika'daki OSHA isimli web sitesinden elde ettikleri 1000 inşaat kazasına ait bilgiyi metin madenciliği sınıflandırma tekniklerinden faydalanarak sınıflandırmışlardır. Çalışmada, destek vektör makinesi (SVM), doğrusal regresyon (LR), rastgele orman (RF), k-en yakın komşu (KNN), karar ağacı (DT) ve Naive Bayes (NB) olmak üzere altı makine öğrenme algoritması kullanılmıştır. Sınıflandırma çalışmasında destek vektör makinesi (SVM) algoritmasının en iyi sonuçları verdiği gözlemlenmiştir.

Göker & Tekedere, (2017) Türkiye'de yürütülen FATİH projesine ait görüşleri içeren metin dosyalarını Naive Bayes, K-En Yakın Komşu (k-NN), Karar Ağaçları, Ardışık Minimal Optimizasyon (SMO- Sequential Minimal Optimization), RBF Network sınıflandırma algoritmaları kullanılarak sınıflandırmışlardır. Ardışık Minimal Optimizasyon Algoritmasının %88.73 ile en yüksek başarıyı gösterdiğini gözlemlemişlerdir.

Hu et al., (2009) yaptıkları çalışmada, şarkıların ilgili olabilecekleri ruh hali kategorisine atanması için, şarkı sözleri metin madenciliği yöntemi ile analiz edilerek, ilgili ruh hali kategorisinin belirlenmesi sağlanmıştır. 5585 şarkının kullanıldığı çalışma sonucunda, ses özelliklerinin ruh hali sınıflandırması için şarkı sözlerinden her zaman daha iyi performans göstermediğini, şarkı sözleri ve ses içeriklerinin birleşmesi ruh hali kategorisinin belirlenmesinde daha efektif sonuçlar verdiğini ortaya koymuşlardır.

Başkaya & Aydın, (2017) ekonomi, politika, spor ve sağlık olmak üzere 4 kategoriye ait haber metinlerini, metin madenciliği tekniklerinden faydalanarak sınıflandırmışlardır. Çalışma sonucunda 80 haber metni Kelime Kökü Öznitelik Yöntemi, Terim Frekansı Terim Ağırlıklandırma Yöntemi ve Rastgele Orman Sınıflandırma algoritması kullanılarak %100 başarı oranı ile sınıflandırılmıştır.

### 2.2.5. Kümeleme (Clustering)

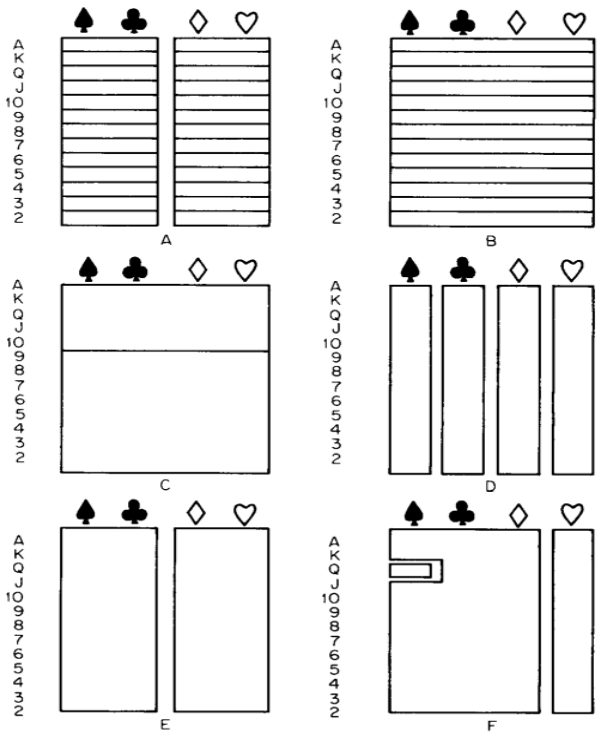
Kümeleme analizi uygulaması erken dönemlerde Bartels vd. (1970) tarafından tanımlanmıştır (Bartels vd., 1970, akt. Anderberg, 1973:syf.5)

Kümeleme algoritmasında temel odak noktası birbirinden anlamlı şekilde farklılaşmış kümelerin ortaya çıkmasıdır (Anderberg, 1973).

Kümeleme analiz sonuçları, verilerle ilgili akıl yürütmeye yardımcı olur. Aslında, analiz sonucunda elde edilen bir küme seti, verinin nasıl organize edileceği ile ilgili bir önerme sunmaktadır.

Bu önerme tamamen yanlış ve saçma unsurlar içerebilir. Daha önce fark edilmeyen ilişkileri açıklamaya yardımcı olabilir  veya veriler arasında bilinen ilişkiler konusunda yeni fikirlerin ortaya atılmasına yardımcı olabilir (Anderberg, 1973)

Anderberg, (1973) kümeleme analizini iskambil kağıtları üzerinden açıklamaya çalışmıştır. Bir iskambil destesinde jokerleri bir tarafa koyup 52 kartın kümelere ayrılmasının istenmesi durumunda oluşabilecek kümeleri şekildeki gibi göstermiştir. Bu altı sınıflandırma bir problemin çözümünde karşılaşılabilecek olası sonuçların kümesini temsil etmektedir. Bu sonuç, aynı zamanda bir problem çözümünün ele alış şeklinin sonuçları nasıl farklılaştırdığı veya anlamlı sonuçlar elde edebilmek için, probleme nasıl yaklaşmak gerektiğine dair mesajlar içermektedir.




**Şekil 12.** Oyun kartı örneği

Kaynak: (Anderberg, 1973, p. 17)

Kümeleme analizinde ortaya çıkan ilk soru, analiz sonucunda iyi bir kümeleme analizinin yapılıp yapılmadığının nasıl anlaşılacağıdır. Bir kümeleme analizinde sonuçların iyi veya kötü olarak nitelendirilmesinden ziyade sorgulanması gereken şey, prosedürlerin veriye doğru bir şekilde uygulanıp uygulanmadığıdır. Prosedürler geçerli ve uygun kabul edildiği sürece kümelerin kendileri sorgulanmaz (Anderberg, 1973).

Kümeleme analizi, bir grup içindeki nesnelerin, kendilerini tanımlayan bilgilerden yola çıkılarak gruplandırılmasıdır. Kümeleme analizinde verinin içerisindeki benzeşen veya farklılaşan grupların ayrıştırılması amaçlanmaktadır (Steinbach et al., 2004).

Kümeleme bir belge koleksiyonunda benzer grupların bulunması işlemidir. Kümeleme belgeler arasında yapılabileceği gibi, paragraflar, cümleler, terimler düzeyinde de yapılabilmektedir (Allahyari et al., 2017).

Gözetimli öğrenme tekniği kullanılarak uygulanan metin sınıflandırma tekniğinde (Kodratoff, 1999) kategoriler önceden bilinmektedir. Eğitim verisi ise buna göre eğitilmektedir. Gözetimsiz öğrenme tekniğinin kullanıldığı kümeleme yönteminde ise bir sınıf veya küme önceden tanımlanmamıştır (Witten, 2004b).

Kümeleme analizinde önceden tanımlanmış sınıflar yoktur. Dokümanda yer alan verinin benzerlik ve farklılıklarına göre kendi içerisinde gruplanması beklenmektedir (Steinbach et al., 2004).

Denetimsiz öğrenme yöntemleri temel olarak üç başlıkta ele alınabilir (Steinbach et al., 2004).

- K-Ortalama Kümeleme Yöntemi
- Hiyerarşik Yöntemler (Agglomerative, Divisive)
- Yoğunluk tabanlı yöntemler

Hiyerarşik kümelemede metin belgeleri arasındaki yakınlığı ölçmek için benzerlik tabanlı kümeleme algoritması kullanılmaktadır (Allahyari et al., 2017). Hiyerarşik kümeleme yaklaşımında temel mantık benzer özniteliklerin bir araya getirilmesi veya bölünmesidir (Steinbach et al., 2004).

Anderberg (1973) hiyerarşik kümeleme yöntemlerinin, araştırmacının incelediği veri setinde kaç grup bulunduğunu başlangıçta bilmediği durumlarda çok uygun bir yöntem olduğunu belirtmektedir.

Hiyerarşik kümeleme algoritması benzerlik matrisi üzerinden çalışır. Ağaç dallarına ayrılmış görüntüsüyle, dalların her biri bir varlığı temsil ederken, kök tüm varlık koleksiyonunu temsil eder. Dallardan köklere doğru oluşan hiyerarşik yönteme birleştirici yöntem (agglomerative hierarchical methods) denir. Köklerden başlayıp dallara ayrılan yönteme de bölücü yöntem (divisive hierarchical methods) denir (Anderberg, 1973).

Ward & Hook, (1963) hiyerarşik kümeleme yönteminde, araştırmacının probleminin amaç fonksiyonunu maksimize edecek şekilde çok genel bir hiyerarşik kümeleme yöntemi tanımlamışladır. Bu yöntemde amaç fonksiyonunun hata toplamları gösterilmiştir. Bu formülasyon bir çok hiyerarşik kümeleme yöntemini kapsamaktadır (Anderberg, 1973). Bunun yanında Wishart, (1969) küme merkezleri arasındaki Öklid uzaklıklarının karesi alınmış bir matrisinin güncellenmesi yoluyla Ward algoritmasının nasıl uygulanabileceğini göstermiştir.

K-ortalama kümeleme yönteminde, rastgele bir orta nokta seçimi yapılmaktadır. Bu durum boş kümelerin ortaya çıkmasına neden olmaktadır. Etkili sonuçlar elde edilebilmesi için başlangıç noktası seçimi etkin bir şekilde yapılmalıdır (Bradley & Fayyad, 1998).

MacQueen, (1967) k-ortalamalar terimini, her bir veri biriminin en yakın ağırlık merkezine atanması olarak tanımlamaktadır. MacQueen, (1967) bu süreçteki m veri biriminin k kümeye ayrılması algoritmasını aşağıdaki şekilde aktarmaktadır:

I. Veri kümesindeki ilk k veri birimi, her biri bir üyeden oluşan kümeler olarak alınır.
II. Kalan (m-k) kadar veri birimlerinin her biri kendisine en yakın ağırlık merkezine sahip kümeye atanır. Her atama işleminden sonra, kümenin ağırlık merkezi yeniden hesaplanır.
III. İkinci adımda tüm veriler kümelere atandıktan sonra, mevcut küme merkezleri sabit çekirdek noktalar olarak kabul edilir. Her veri birimi en yakın çekirdek noktaya atanarak kümeler arasında bir kez daha geçiş sağlanır.

K-ortamalar yöntemindeki önemli noktalardan birisi, optimal bir kümeleri bulma noktasında bir garantinin olmamasıdır. Bu nedenle yaygın olarak, k-ortalamalar işleminde farklı, rastgele başlangıç küme merkezleri ile birkaç kez yeniden başlatılarak işlem gerçekleştirilir. Scikit-learn bunu varsayılan olarak gerçekleştirir (Cady, 2017, p. 146)

Kümeleme algoritmaları çeşitli alanlarda uygulanabilmektedir. ABD Hava Kuvvetleri, birçok Hava Kuvvetleri uzmanlık alanındaki iş envanterlerini analiz etmek için CODAP isimli bir iş kümeleme programı kullanmıştır (Anderberg, 1973).

Godfrey et al., (2014) kümeleme analizi kullanarak, dünya kupası öncesi twitter üzerinden atılan 30 bin tweeti incelemiştir. Çalışmada kümeleme algoritması olarak k-means ve Negatif Olmayan Matris Çarpanlara ayırma (NMF) kullanarak kümeleme işlemi gerçekleştirilerek sonuçlar karşılaştırılmıştır. Her İki algoritma da benzer sonuçlar verdiği ancak NMF'nin daha hızlı olduğunu ve kolay yorumlanabildiği görülmüştür.

# BÖLÜM 3. VERİ BİLİMİ ALANINDAKİ İŞ İLANLARININ ANALİZİ

Bu bölümde, yapılan tez çalışmasının amaç, kapsam ve yöntemi detaylı olarak aktarılacaktır. Veri kümesi hakkında bilgi verilerek veri hazırlama sürecinden bahsedilecektir. Her pozisyon bazında, metin madenciliği yöntemi kullanılarak yapılan analizler ve analiz sonuçları detaylı şekilde aktarılacaktır. Bunun yanında, metin verisinden elde edilen sayısal sonuçların anlamlılık test sonuçları paylaşılacaktır.

## 3.1. AMAÇ, KAPSAM VE YÖNTEM

Şirketlerin teknolojik yeteneklerinin artması, bulut teknolojiler, Hadoop gibi platformların veri saklamaya ve iletmeye imkân tanıması, "kullandığın kadar öde" uygulamalarıyla da verinin daha düşük maliyetlerle depolanabilmesi, ardından gelen başarılı büyük veri uygulamaları, büyük verinin önemini her geçen gün artırmaktadır. Buna ek olarak, anlamlandırılabilen ve iç görü elde edilebilen verinin müşteri deneyimi, pazarlama gibi alanlarda rekabet avantajı ve operasyonel alanlarda verimlilik sağlaması şirketlerin büyük veriye olan iştahını artırmaktadır.

Büyük veri ile şirketlerde veriyi işleyen anlamlandıran veriden iç görüler elde etmesi beklenen pozisyonlar hızla ortaya çıkmıştır. Veriden fayda sağlama iştahı ve pazardaki rekabet, şirketleri bu pozisyonlar için yetkin çalışan istihdamına yöneltmektedir. Şirketler, büyük veri ile çalışacak pozisyonlar oluştururken şirket ihtiyaç ve beklentilerini göz önünde bulundurmaktadırlar. Çalışan istihdamı da her şirket özelinde belirlenen kriterlere göre gerçekleşebilmektedir. Bu durum büyük veri ile uğraşan Veri Bilimi alanında aranan pozisyonlar için öznel tanımlamaların ortaya çıkmasına neden olmaktadır. Bu tanımlar zamanla, yazılı olmayan, her şirket ihtiyacına paralel gelişen ve farklılıklar barındıran uygulamalara dönüşebilmektedir.

Veri biliminin dijitalleşme ile şirketlerde uygulama alanı bulması, ancak bu konunun literatürde tartışılarak henüz belirli bir olgunluğa ulaşmamış olması, şirketlerin istihdam için belirlediği yetenek setleri ile teorik çerçevenin benzer olduğu gibi bir algı oluşturmaktadır. Bu durum, yayınlanan iş ilanlarında nesnel bir yaklaşımın ortaya konmasına engel olmakta ve ilgili iş pozisyonu için istenen yetenek ve beceriler bazen ütopik düzeylere çıkarak istihdam edilecek kişiye mitolojik ögeler yüklenirken, bazen de aday için istenilen yetenekler zayıf kalmaktadır.

Bu çalışma, büyük veri ile doğan ve henüz yetkinlik setlerinde ayrımın net bir şekilde yapılmadığı ve kafa karışıklığı olduğu düşünülen Veri Bilimi pozisyonlarına yönelik platform ve metodolojilere ilişkin gerekli bilgi ve yetkinlik setlerinin çerçevesini incelemeyi ve literatürdeki mevcut çalışmalarla karşılaştırarak Türkiye'deki şirketlerin bu pozisyonlara yönelik yaklaşımlarını ortaya koymayı amaçlamaktadır. Çoğu kullanımda birbirinin yerine geçen Veri Bilimi pozisyonları ve bu pozisyonlarda çalışanların hangi bilgi ve yetkinliğe sahip olması gerektiği, kafa karıştırıcı olmaktadır. Yapılan literatür incelemesi sonrası görüldüğü kadarı ile bu çalışma, Türkiye'de bu alanda yapılmış olan

ilk çalışma olarak Veri Bilimi alanındaki bilgi ve yetenek tartışmasını ele alarak literatüre Türkiye örneği üzerinden katkı sağlamaya gayret etmiştir.

Bu tez çalışması ile Türkiye'deki platformlar üzerinden yayınlanan veri bilimi pozisyonlarına yönelik iş ilanlarının içerikleri teknik ve sosyal yetkinlikler açısından ele alınarak incelenmiştir. Analiz süreci öncesinde analiz edilecek verinin çerçevesi çizilmiştir. Bu çerçeve çizilirken literatürde yapılmış çalışmalar ışığında ve piyasadaki pratik uygulamalardan ilham alınarak ilerlenmiştir. Böylece "veri", "data", "analiz", "analysis", "analist" anahtar kelimeleri baz alınarak ilanlar taranmıştır. Bu doğrultuda, veri bilimi ile ilişkili olduğu düşünülen "Veri Bilimcisi (Data Scientist)", "Veri Analizcisi (Data Analyst)", "İş Zekâsı Uzmanı (Business Intelligence Speacialist)", "Veri Tabanı Uzmanı (Database Specialist)", "Veri Ambarı Uzmanı (Data Warehouse (DWH) Specialist)", "Makine Öğrenmesi Mühendisi (Machine Learning (ML) Engineer)", "İş Analisti (Business Analyst)" olmak üzere toplamda 8 pozisyona ait ilanlar analiz kapsamında kullanılmıştır.

Veri toplama aşamasından sonra veri analiz aşamasına geçilmiştir. Veri analizinde veriler metin kaynaklarından oluştuğu, dolayısıyla sayısal bir veri seti olmadığı için metin madenciliği yöntemi kullanılmıştır. Metin madenciliği yöntemi aynı zamanda çalışmanın amaçları ile uyumlu olduğundan ve örtüştüğünden tercih edilmiştir.

Metin madenciliğinde kullanılan birçok araç bulunmaktadır. Python, R, Knime, Rapid Miner gibi uygulamalar metin madenciliğine imkân tanımaktadır. Bu tez çalışmasında veri analizi için metin madenciliğine olanak tanıyan Python 3.9 uygulaması kullanılmıştır.

Python uygulama dili Guido van Rossum tarafından 1990'lı yılların başlarında geliştirilmeye başlanmıştır. Python, programlama yapısı içirişinde birçok yardımcı kütüphane barındırdığından daha az kodlama ile analizlerin gerçekleştirilmesine imkân tanımaktadır. Python uygulamasının daha kolay kullanımı için çeşitli arayüzler yani IDE (Integrated Development Environment)'ler mevcuttur. Spyder, PyCharm, Jupyter / Anaconda Python IDE uygulamalarındandır. Bu çalışmada Anaconda üzerinden JupyterLab 2.1.5 versiyonu kullanılmıştır. JupyterLab kod sayfalarının web üzerinden görüntülenerek işlem yapılmasına olanak tanımaktadır. Python, web tabanlı uygulamalar, makine öğrenmesi, veri madenciliği, metin madenciliği gibi alanlarda kullanılabilmektedir.

Bu çalışma kapsamında Python uygulaması metin madenciliği yönteminde kullanılmıştır. Python üzerinden kod yazılarak ve metin madenciliği sürecine yardımcı kütüphaneler (nltk, numpy, pandas) kullanılarak metin analiz süreci gerçekleştirilmiştir. Aşağıda metin madenciliği süreci, veri dosyası okuma adımından, metin görselleştirme ve diğer tüm veri işleme süreçlerine kadar adım adım paylaşılmıştır.

## 3.2. VERİ KÜMESİ VE VERİ HAZIRLAMA

İş ilanlarının elde edileceği platformlar belirlenirken, Türkiye'de şirketlerin en çok kullandığı ve iş arayanların da en sık ziyaret ettiği platformlar tercih edilmeye çalışılmıştır. Bu kapsamda kariyer.net ve Linkedin platformları veri kaynağı için uygun görülmüştür. Bu platformlardan elde edilen içerikler yapılan analizler için nitelikli veri sağlamaya yardımcı olmuştur.

Kariyer.net ve Linkedin üzerinden yayınlanan ilanlar "Aranan Nitelikler" ve "İş Tanımı" olmak üzere 2 temel başlıktan oluşmaktadır. Her şirket iş ilanında aranan nitelikler başlığına yer vermesine karşın bazı şirketlerin iş ilanlarında iş tanımı yer almamaktadır. Bu nedenle veri toplama sürecinde bazı ilanlarda iş tanımı satırları boş kalmıştır. Buna ek olarak üniversite bölüm mezuniyeti de çoğunlukla kriterler arasında yer almaktadır.

Belirlenen platformlardan elde edilen veriler, excel dosyası üzerinde tutulmuştur. Excel dosyasındaki veri tablosu, *ilan adı, aranan nitelikler, iş tanımı ve üniversite bölümü* olmak üzere dört sütundan oluşmaktadır. Buna ek olarak her ilan için ilan veren şirketin isim bilgisi de veri listesinde tutulmuştur.

Veri toplama süreci Mart 2021-Mart 2022 tarih aralığını kapsamaktadır. Toplamda 1,465 ilan elde edilmiştir. *Veri Analizcisi (264 ilan), Veri Bilimcisi (232 ilan), İş Zekâsı Uzmanı (277 ilan), İş Analisti (253 ilan)* ilanlarında 200 üzerinde ilan sayısına ulaşılmışken; *Veri Ambarı Uzmanı (105 ilan), Veri Mühendisi (144 ilan), Veri Tabanı Uzmanı (110 ilan)* pozisyonlarında 100 üzerinde ilan sayısına ulaşılabilmiştir. *Makine Öğrenmesi Mühendisi (80 ilan)* pozisyonunda ise 100 sınırının altında ilan elde edilmiştir.

Veri analizine, kullanılacak Python kütüphanelerinin yüklenmesi ile başlanmıştır. Böylece, veri dosyası Python üzerinden okutularak içeriği görüntülenmiştir. Verinin sorunsuz bir şekilde yüklenmesinden sonra veri işleme adımlarına geçilmiştir.

```
import numpy as np
import pandas as pd
import nltk
```

**Şekil 13.** Python kütüphanelerinin yüklenmesi

```
data = pd.read_excel('data_source/business_int_277.xlsx')
data.head()
```

| | job_title | qualifications | job_description | university_department | company |
|---|---|---|---|---|---|
| 0 | Business Intelligence Senior Expert | Have minimum 5 years experience in Data Wareho... | Work closely with stakeholders to understand t... | Computer Science, Industrial,Computer or Data ... | Aras Kargo |
| 1 | Business Intelligence Specialist | Intermediate level of PL/SQL or SQL & experien... | Ability to work in multicultural environment\n... | Computer Engineering, Information Systems Engi... | Watsons |
| 2 | Business Intelligence Associate | Good knowledge and experience in Reporting Arc... | Designing accurate analysis models to answer v... | NaN | Mavi Giyim Sanayi ve Ticaret A.Ş. |
| 3 | Business Intelligence and Reporting Specialist | Graduated from a university with a Bachelors d... | Work closely with stakeholders to understand t... | Computer Science, Industrial, Electric-Elec... | Marm Assistance Destek Hizmetleri A.Ş. |
| 4 | Business Intelligence and Reporting Specialist | Bachelor's degree in Computer Science / Indust... | Working closely with the customer, BU, BU IT, ... | Computer Science / Industrial Engineering | DHL Supply Chain |

**Şekil 14.** Veri dosyasının Python üzerinden görüntülenmesi

Veriler excel dosyasından okutulduktan sonra, "aranan nitelikler (qualifications)" sütunundan başlanarak büyük-küçük harf dönüşüm işlemi gerçekleştirilmiştir. Bir üst satırda yer alan büyük harflerin aşağıda yer alan görselde küçük harflere dönüştüğü gözlemlenmiştir.

```
new_data["qualifications"].apply(lambda x:" ".join(x.lower() for x in x.split()))
```

```
0      have minimum 5 years experience in data wareho...
1      intermediate level of pl/sql or sql & experien...
2      good knowledge and experience in reporting arc...
3      graduated from a university with a bachelors d...
4      bachelor's degree in computer science / indust...
                             ...
272    problem solving and analytical thinking skills...
273    at least 3 years of experience in business int...
274    in similar positions min. 5 years of experienc...
275    microsoft business intelligence (ssis, ssrs) e...
276    minimum +2 years of experience impeccable atte...
Name: qualifications, Length: 277, dtype: object
```

**Şekil 15.** Büyük-küçük harf dönüşümü

Harf dönüşüm işlemleri yapıldıktan sonra, veri dosyasındaki noktalama işaretleri silinmiştir.

```
data_1.str.replace("[^\w\s]"," ").head(5)
```

```
<ipython-input-12-f7888e2844ef>:1: FutureWarning: The default value of regex will change f
  data_1.str.replace("[^\w\s]"," ").head(5)
0    have minimum 5 years experience in data wareho...
1    intermediate level of pl sql or sql   experien...
2    good knowledge and experience in reporting arc...
3    graduated from a university with a bachelors d...
4    bachelor s degree in computer science   indust...
Name: qualifications, dtype: object
```

**Şekil 16.** Noktalama işaretlerinin silinmesi

Noktalama işaretleri silindikten sonra, satırlarda yer alan sayılar silinerek ön işleme adımlarına devam edilmiştir.

```
data_2.str.replace("\d","").head()
```

```
<ipython-input-14-2a24a63435e4>:1: FutureWarning: The default value of regex will
  data_2.str.replace("\d","").head()
0    have minimum  years experience in data warehou...
1    intermediate level of plsql or sql  experience...
2    good knowledge and experience in reporting arc...
3    graduated from a university with a bachelors d...
4    bachelors degree in computer science  industri...
Name: qualifications, dtype: object
```

**Şekil 17.** Sayıların silinmesi

Sayıların silinmesinin ardından metinde büyük oranda ön işleme adımları tamamlanmıştır. Artık metin verisi üzerinden daha anlamlı sonuçlar elde edebilmek adına metinde yer alan kelimeler incelenmiştir. Bu adımda her metinde yer alan ve atıldığında metinde ciddi anlam bozukluklarına sebep olmayan bağlaçlar, özel isimler, kısaltmalar atılmıştır. İngilizcede "stopwords" olarak adlandırılan, durak kelimeler/yararsız kelimeler metinden çıkartılacak daha yararlı bilgiler elde edilmesi amaçlanmıştır.

Yararsız kelimelerin elimizdeki veri dosyasından silinebilmesi için, nltk kütüphanesi yüklenerek İngilizce dilinde önceden tanımlı olan "stopwords" listesi indirilmiştir. Bu kelimeler sw isimli bir değişkene atanmıştır. İngilizcede tanımlı olan stopwords listesi analiz için temel olarak yararsız olan kelime listesini içermektedir (am, is, are, to, with, etc…). Ancak analiz boyunca, kendi veri kümemizde, "skill","preferably","years" gibi sık tekrarlanan ve aynı zamanda analizde anlamlı sonuçlara katkı sağlamayacağı düşünülen kelimeler de metin içerisinden silinmiştir. Çalışmamıza özel bir yararsız kelimeler listesi hazırlanarak bu işlem tamamlanmıştır.

```
!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\zehra\anaconda3\lib\site-packages (3.5)
Requirement already satisfied: joblib in c:\users\zehra\anaconda3\lib\site-packages (from nltk) (0.16.0)
Requirement already satisfied: tqdm in c:\users\zehra\anaconda3\lib\site-packages (from nltk) (4.47.0)
Requirement already satisfied: regex in c:\users\zehra\anaconda3\lib\site-packages (from nltk) (2020.6.8)
Requirement already satisfied: click in c:\users\zehra\anaconda3\lib\site-packages (from nltk) (7.1.2)
```

```
nltk.download("stopwords")
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\zehra\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
```

```
from nltk.corpus import stopwords
sw = stopwords.words("english")
```

**Şekil 18.** İngilizcedeki yararsız kelimeler (stopwords) listesi

Kütüphanede halihazırda var olan yararsız kelimeler listesine, aşağıda yer alan kelimeler eklenmiştir. Bu kelimeler analiz için katkı sağlamayacağı düşünülen kelimelerden oluşturulmuştur. Kelime seçiminden önce, analiz öncelikle ön tanımlı olan yararsız kelimeler (stopwords) listesi

üzerinden yapılmıştır. Katkı sağlamayan kelimeler analiz boyunca yapılan denemelerle bulunmuş, metin anlamlı bir analiz imkânı sunacak hale gelinceye kadar kelimeler analizden çıkartılmak üzere yararsız kelimeler listesine atılmıştır. Custom_sw listesine atılan bu kelimeler analiz boyunca metin kümesi dışında tutulmuştur.

```
customlist_sw = ["skill","preferably","years","year","etc", "good","skills","minimum","least","min","work","result","integration",
                 "business","strong","plus","tools","able","nan","abilities","use","ability","data","similar","responsible","worked",
                 "experienced","experience","using","working","analysis","knowledge","reporting","excellent","architecture","attention",
                 "written","advanced","language","related","high","basic","report","new","fluent","user","test","object","processes",
                 "report","creating","big","complex","oriented","spoken","power","development","bi","intelligent","projects","technologies",
                 "intelligence","learning","learn","project","thinking","one","solving","prone","service","proficient","applications",
                 "office","open","command","management","problem","application","analytical","english","microsoft","platforms","objects",
                 "process","team","solution","developed","system","analytics","time","level","detail","design","platform","reports","systems",
                 "environment","write","understanding","services","solutions","technical","databases"]

sw.extend(customlist_sw)
```

**Şekil 19.** Özel olarak yararsız kelimeler (stopwords) listesinin oluşturulması

Metin dosyası içinde sık geçen ve yararsız olarak tespit edilen kelimelerin yanında, metinde az geçen kelimeler de analiz sonuçlarının kalitesine etki etmektedir. Bu sebeple metinde az geçen kelimeler de analiz için herhangi bir katkı sağlamadığından bu kelimeler de metinden silinerek işlemlere devam edilmiştir.

```
data_3["qualifications"].apply(lambda x:" ".join(x for x in x.split() if x not in sw))

0      warehousing sql mssql queries issues validatio...
1      intermediate plsql sql oracle rdbms previous u...
2      deliverables sql server tsql queries demonstra...
3      graduated university bachelors degreee compute...
4      bachelors degree computer science industrial e...
                             ...
272    teamwork acting sense discipline responsibilit...
273    qlikview involved warehouse etl etl odi ssis s...
274    positions previous qlikview tableau equivalent...
275    ssis ssrs warehouse methodologies sql server a...
276    impeccable computer eg excel access sqltableau...
Name: qualifications, Length: 277, dtype: object
```

**Şekil 20.** Yararsız kelimelerin metinden çıkarılması

Aşağıda kelime saydırma işlemi yapılarak kelimelerin geçme sıklıkları listelenmiştir. Kelimeler en çok tekrar eden kelimeden sadece bir kere tekrar eden kelimelere doğru sıralanmıştır.

```
data_4 = pd.DataFrame(data_4, columns = ["qualifications"])
```

```
pd.Series(" ".join(data_4["qualifications"]).split()).value_counts()
```

```
sql              233
etl              149
warehouse        137
sap              127
communication    111
                ...
kpis               1
driving            1
roadmap            1
determination      1
industries         1
Length: 1564, dtype: int64
```

**Şekil 21.** Kelime saydırma

Kelimelerin en çok geçen kelimeden en az geçen kelimeye doğru sıralanmasının ardından, kelimeler sondan saydırılarak üç kereden az geçen kelimeleri tespit edilmiştir. Toplamda 1,564 kelimeden metin içinde bir ve iki kere geçen 872 kelime olduğu görülmüştür. Bu kelimeler tanımlanan sil isimli bir değişkene atanmıştır.

```
sil = pd.Series(" ".join(data_4["qualifications"]).split()).value_counts()[-872:]
sil
```

```
scenario         2
sqlplsql         2
installation     2
synapse          2
teach            2
                ..
kpis             1
driving          1
roadmap          1
determination    1
industries       1
Length: 872, dtype: int64
```

**Şekil 22.** Az geçen kelimelerin tespit edilmesi

Metinde bir ve iki kez geçen kelimeler sil isimli değişken yardımı ile silinmiştir.

```
data_4["qualifications"].apply(lambda x: " ".join(i for i in x.split() if i not in sil))

0      warehousing sql mssql queries issues familiari...
1      intermediate plsql sql oracle rdbms previous u...
2      deliverables sql server tsql queries demonstra...
3      university bachelors computer science computer...
4      bachelors degree computer science engineering ...
                           ...
272    teamwork acting sense discipline responsibilit...
273    qlikview involved warehouse etl etl odi ssis s...
274    positions previous qlikview tableau equivalent...
275    ssis ssrs warehouse methodologies sql server a...
276    computer eg excel access willingness develop e...
Name: qualifications, Length: 277, dtype: object
```

**Şekil 23.** Az geçen kelimelerin silinmesi

Az geçen kelimelerin silinmesinin ardından, tokenization yani bölütleme işlemi yapılmıştır. Bu adımda kelimeler parçalara ayrıştırılmaktadır. Yani her kelime bir vektör olmaktadır. Bu işlem adımında TextBlob kütüphanesi yüklenerek cümlelerin kelimelere ayrıştırılması sağlanmıştır.

Metindeki cümleler kelimelere ayrıştırıldıktan sonra, her kelimenin en yalın formu olan kök hali elde edilmeye çalışılmaktadır. Bu süreçte Lemmatization işlemi uygulanmaktadır. Böylece yalın halde aynı anlamda olan ancak ek alarak analizde farklı kelimeler imiş gibi görünen kelimelerin tek forma dönüştürülerek analize dahil edilmesi sağlanmaktadır.

Kelimelerin köklerine ayrıştırılması işleminden sonra, metinde geçen kelimelerin terim frekanslarına bakılarak geçme sıklıkları tespit edilmiştir. Sonrasında sort_values komutuyla en çok geçen kelimeden en az geçen kelimeye doğru bir sıralama yapılmıştır.

```
nltk.download("punkt")
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\zehra\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

```
!pip install textblob
```

```
Requirement already satisfied: textblob in c:\users\zehra\anaconda3\lib\site-packages (0.15.3)
Requirement already satisfied: nltk>=3.1 in c:\users\zehra\anaconda3\lib\site-packages (from textblob) (3.5)
Requirement already satisfied: tqdm in c:\users\zehra\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (4.47.0)
Requirement already satisfied: click in c:\users\zehra\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (7.1.2)
Requirement already satisfied: regex in c:\users\zehra\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (2020.6.8)
Requirement already satisfied: joblib in c:\users\zehra\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (0.16.0)
```

```
import textblob
from textblob import TextBlob
```

```
data_4["qualifications"].apply(lambda x: TextBlob(x).words)
```

```
0      [warehousing, sql, mssql, queries, issues, val...
1      [intermediate, plsql, sql, oracle, rdbms, prev...
2      [deliverables, sql, server, tsql, queries, dem...
3      [graduated, university, bachelors, degreee, co...
4      [bachelors, degree, computer, science, industr...
                             ...
272    [teamwork, acting, sense, discipline, responsi...
273    [qlikview, involved, warehouse, etl, etl, odi,...
274    [positions, previous, qlikview, tableau, equiv...
275    [ssis, ssrs, warehouse, methodologies, sql, se...
276    [impeccable, computer, eg, excel, access, sqlt...
Name: qualifications, Length: 277, dtype: object
```

**Şekil 24.** Kelime bölütleme (tokenization) işlemi

```
from textblob import Word
nltk.download('wordnet')
data_4["qualifications"]= data_4["qualifications"].apply(lambda x: " ".join([Word(word).lemmatize() for word in x.split()]))
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\zehra\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

```
data_4["qualifications"]
```

```
0      warehousing sql mssql query issue validation c...
1      intermediate plsql sql oracle rdbms previous u...
2      deliverable sql server tsql query demonstrate ...
3      graduated university bachelor degreee computer...
4      bachelor degree computer science industrial en...
                             ...
272    teamwork acting sense discipline responsibilit...
273    qlikview involved warehouse etl etl odi ssis s...
274    position previous qlikview tableau equivalent ...
275    ssis ssrs warehouse methodology sql server adm...
276    impeccable computer eg excel access sqltableau...
Name: qualifications, Length: 277, dtype: object
```

**Şekil 25.** Kelimenin köklere ayrıştırılması (Lemmatization) işlemi

## 3.3. VERİ ANALİZCİSİ İŞ İLANLARININ ANALİZİ

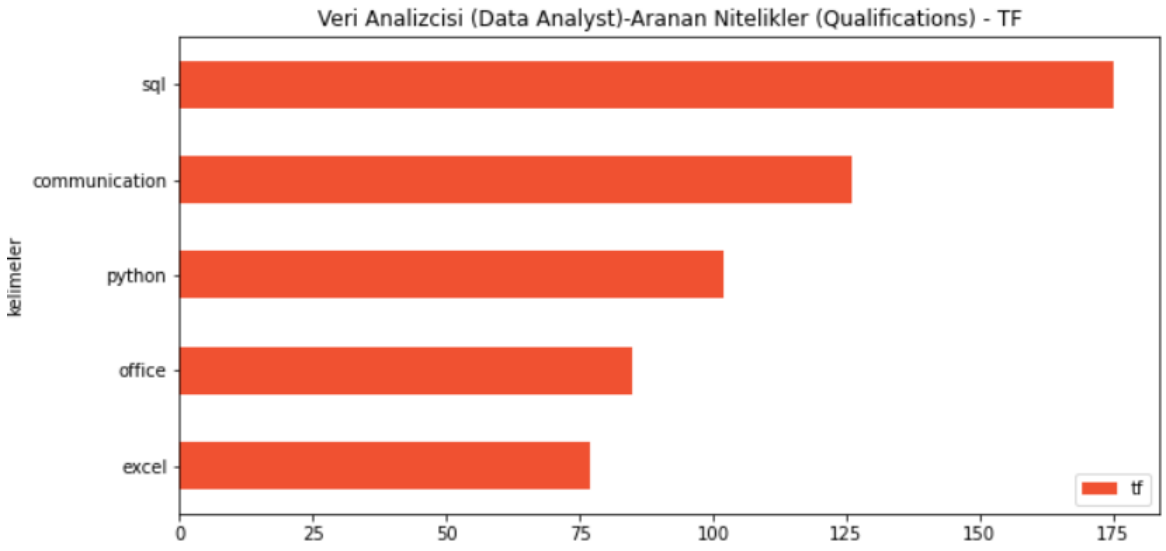
Veri Analizcisi (Data Analyst) pozisyonunda toplamda 264 ilan elde edilmiştir.

- **Aranan Nitelikler (Qualifications)**

İncelenen 264 ilan üzerinden elde edilen kelime sıklık grafiği, veri analizcisi istihdam etmek isteyen şirketlerin SQL, iletişim ve Python yetkinliklerini ilk sırada tercih ettikleri görülmüştür. Buna göre ilk üç sırada iki teknik yetkinlik aranırken, bir tane sosyal yetkinlik arandığını görülmektedir.

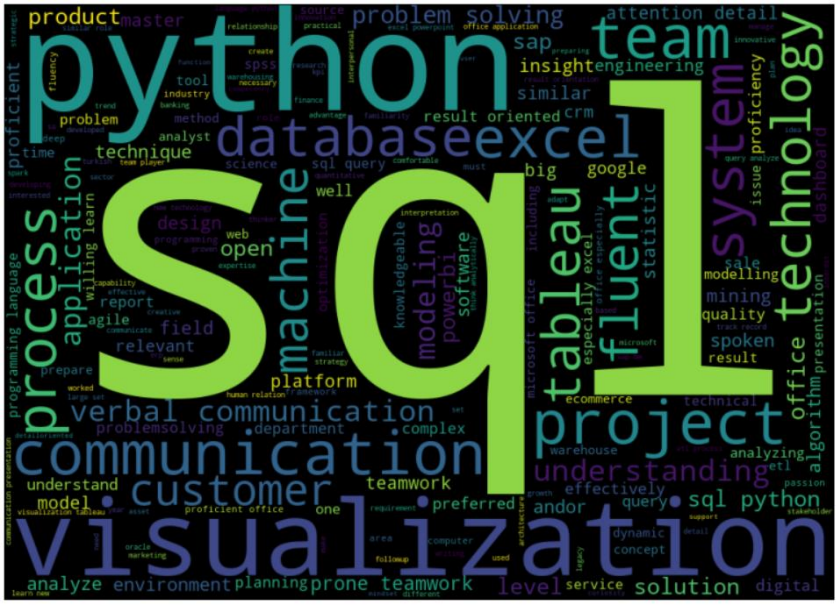
SQL (Structured Query Language) yapılandırılmış bir sorgulama dilidir. Veri tabanlarından istenilen bilgilerin alınmasına imkân tanımaktadır. SQL yardımı ile veri tabanından veri alınabilir, veri tabanına kayıt eklenebilir, veri tabanındaki kayıtlar güncellenebilir (Turhost, 2021).

Python bir veri analizi programlama dili olarak tanımlanabilir. Python'da kullanılan NumPy, Pandas ve Matplotlib, gibi kütüphaneler matematik ve istatistiksel işlemleri hızlandırır ve veri görselleştirilmesine olanak sağlar (Bilginç, 2022).




**Şekil 26.** Veri analizcisi aranan nitelikler-terim frekansı

Veri analizcisi pozisyonu için veri kümesi üzerinden elde edilen kelime bulutunda, SQL, Python, görselleştirme, iletişim kelimeleri ön plana çıkmaktadır. Kelime bulutu veri analizcisi iş tanımı için sunduğu çerçeve şu şekilde yorumlanabilir. SQL yetkinliği olan veri çeken, Python programlama ile veri analizi gerçekleştiren ve veri görselleştirmesi yapan bir pozisyon tanımı görülmektedir.




**Şekil 27.** Veri analizcisi aranan nitelikler-kelime bulutu

İletişim sosyal yetkinliği, veri analizcisi pozisyonu için ve aşağıdaki bölümlerde analiz sonuçları paylaşılacak diğer pozisyonlar için de önemli bir sosyal yetkinlik olarak karşımıza çıkmaktadır.

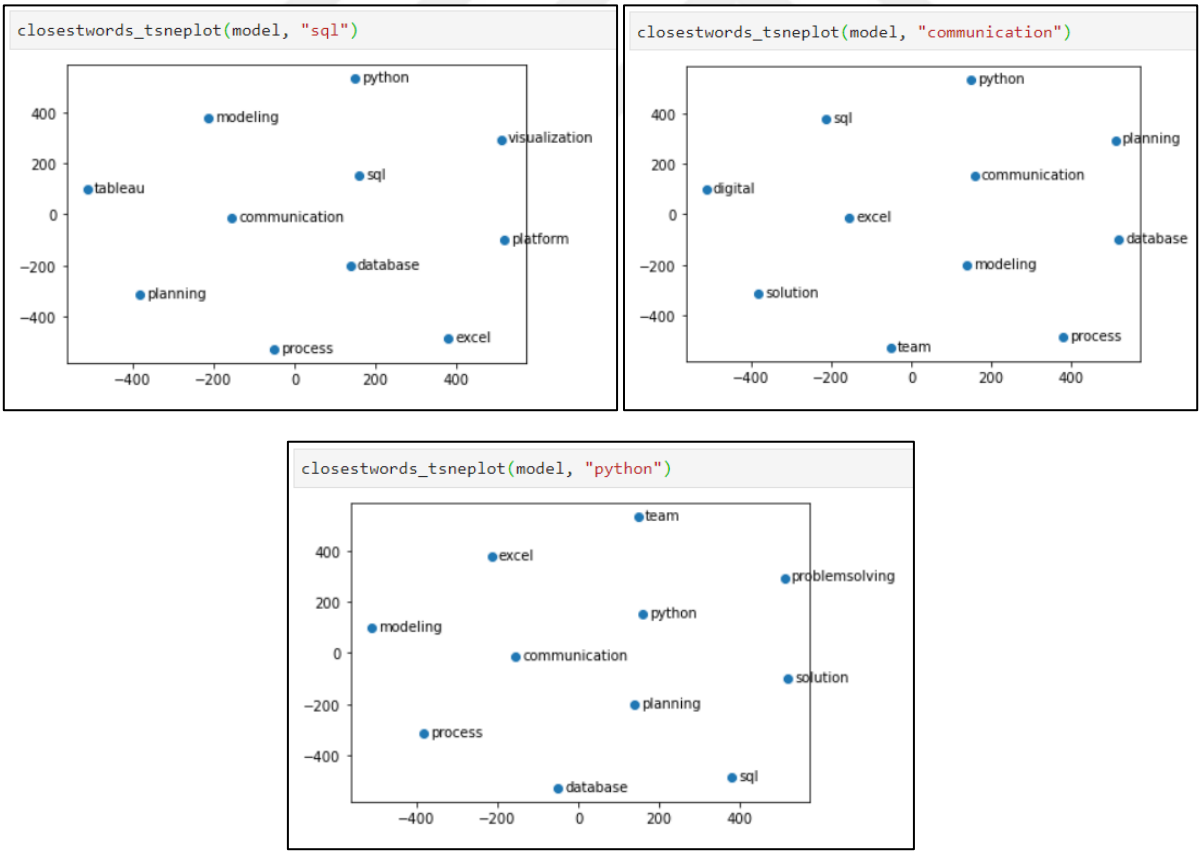
- **En Yakın Kelime Analizi**

Sık geçen kelimeler ve kelime bulutu, veri analizcisi pozisyonu için şirketlerin vurgu yaptığı temel platform, metodoloji, bilgi ve yetkinliklere işaret etmektedir. Bu kapsamda işaret edilen bu alanların beraberinde ilişkili olduğu platform ve yetkinlikleri tespit etmek çalışan yetkinlik setlerini tanımlamada yararlı sonuçlar ortaya koyacaktır. Bu kapsamda, SQL, İletişim ve Python yetkinliklerinin en yakın kelimeler analizi yapılmıştır.

SQL kelimesinin, en yakın kelimeler analizine bakıldığında, SQL kelimesi etrafında iletişim, modelleme, Python, görselleştirme, veri tabanı kelimelerinin kümelendiği görülmektedir. Bu sonuç, SQL yetkinliği ile beraber bahsi geçen platform ve yetkinliklerin de öne çıktığını göstermektedir.

İletişim kelimesine en yakın kelime kümesine bakıldığında, SQL, Excel, Python, modelleme, veri tabanı ve planlama kelimelerinin yer aldığı görülmektedir.

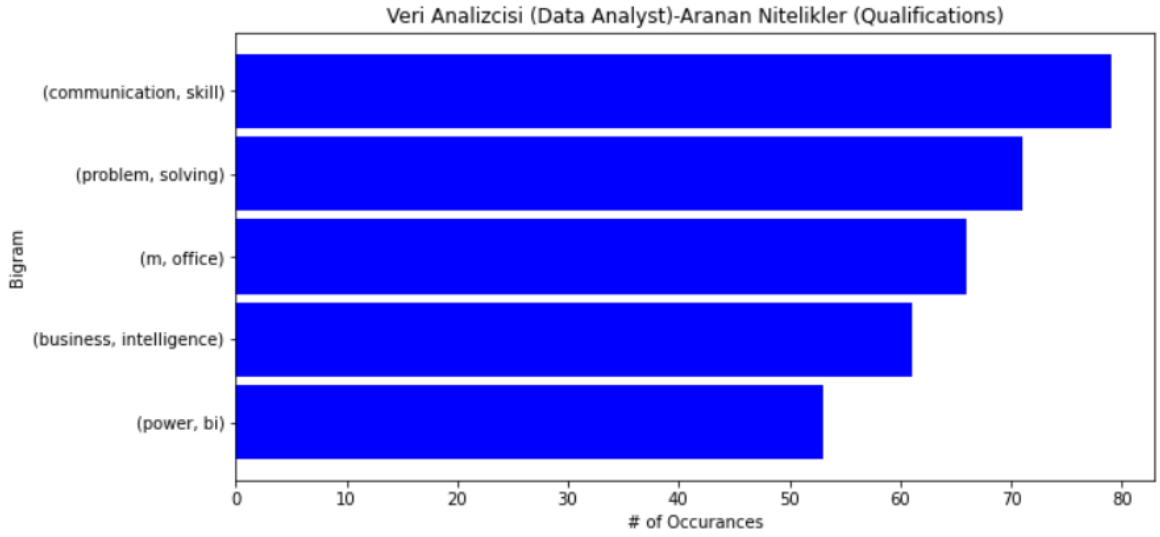
Python kelimesinin en yakın kelime kümesine bakıldığında, takım, problem çözme, çözüm, planlama, iletişim ve Excel kelimelerinin olduğu görülmektedir.




**Şekil 28.** SQL-İletişim-Python en yakın kelimeler analizi

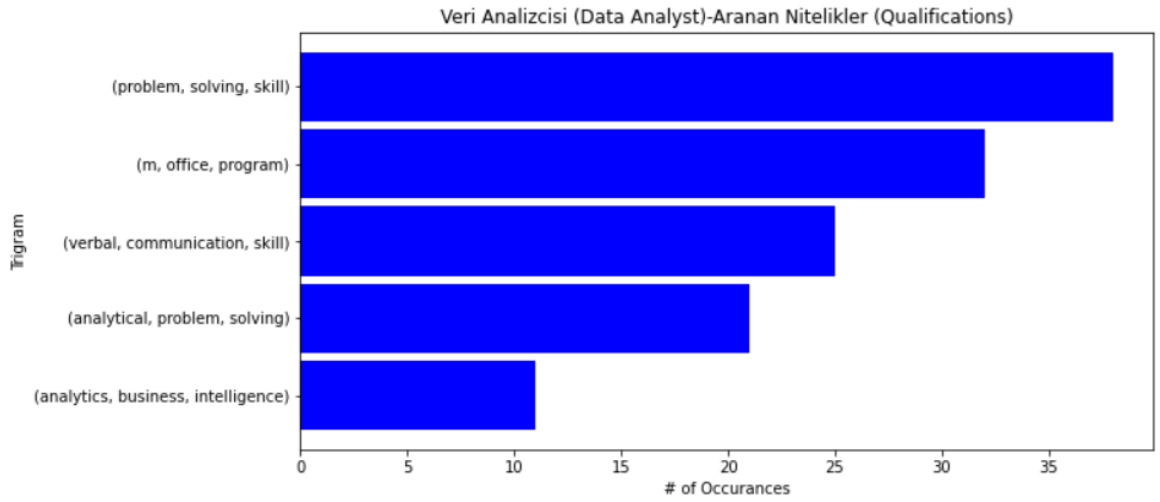
- **Aranan Nitelikler-N-Gram Analizi**

Veri analizcisi pozisyonu veri kümesindeki, aranan nitelikler sütünü N-gram analizi ile de incelenmiştir. N gram analizi ile kelime öbekleri üzerinden yetkinlik tanımlarının daha net bir şekilde ortaya konması amaçlanmıştır. Bu kapsamda, veri analizcisi için ilk beş sırada aranan nitelikler arasında, iletişim becerisi, problem çözme, office programı, iş zekâsı ve power bi gelmektedir. Bu kelimelerin iki tanesi sosyal alandaki yetkinliklere işaret ederken, üç tanesi teknik alanındaki yetkinliklerdendir.




**Şekil 29.** Veri analizcisi aranan nitelikler bigram analizi

N-gram analizinde, ikili kelime gruplarının yanı sıra üçlü kelime grupları da analiz edilebilmektedir. Bu kapsamda veri analizcisi pozisyonu için ikili kelime grubundan farklı olarak üçlü kelime gruplamada yeni bir başlığın ortaya çıkmadığını, ikili kelime gruplarını tamamlayıcı eklemelerin olduğunu görmekteyiz.

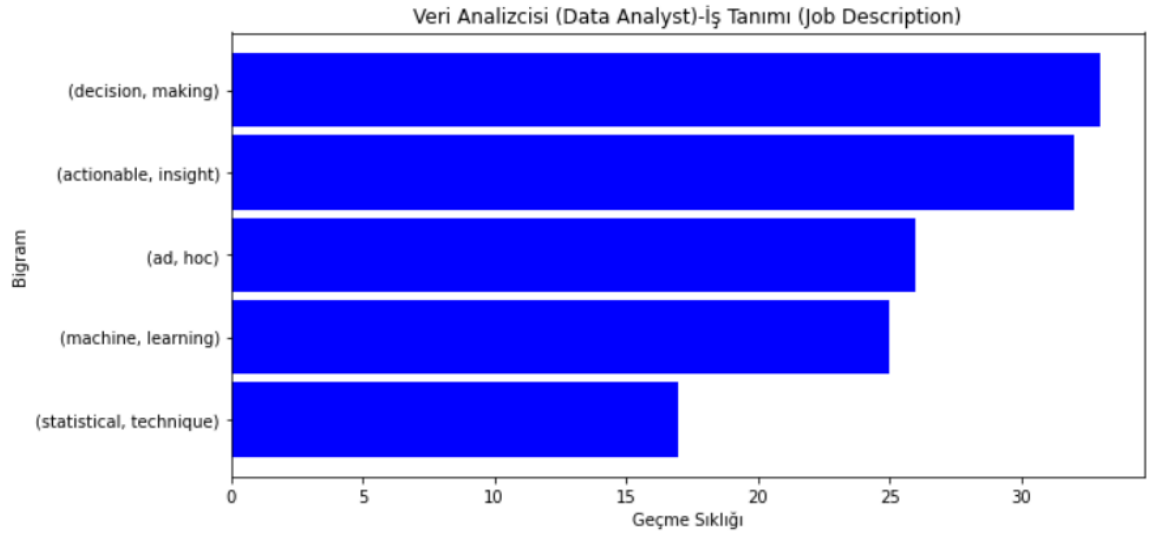



**Şekil 30.** Veri analizcisi aranan nitelikler trigram analizi

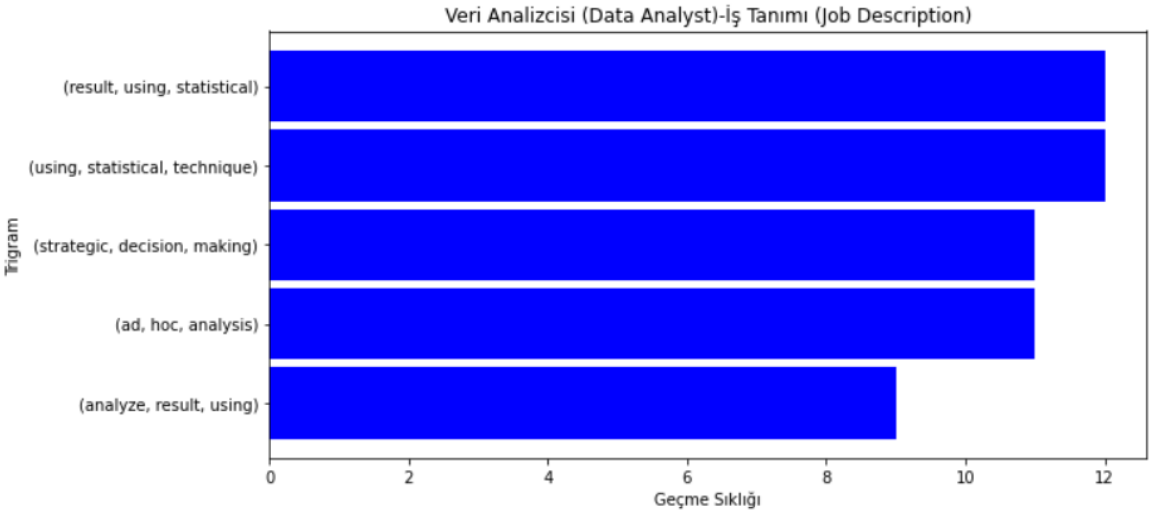
- **İş Tanımı (Job Description)**

Veri analizcisi pozisyonu için veri setinde aranan niteliklere ek olarak, iş tanımı sütununun olduğundan bahsetmiştik. Aranan nitelikler sütunu yukarıda analiz edilmiştir. Aynı adımlar iş tanımı sütunu için de aşağıda tekrar edilerek yorumlamalar yapılmıştır.

Veri analizcisi iş tanımında ikili kelime gruplarının daha etkin ve anlamlı sonuçlar ortaya koyduğu gözlemlendiğinden, yukarıdaki kelime sıklık analizi iş tanımı için gerçekleştirilmemiştir. Veri Analizcisi pozisyon tanımında ikili kelime gruplarına bakıldığında, karar verme, uygulanabilir içgörü, ad hoc, makine öğrenmesi ve istatistiksel teknikler iş tanımı içerisinde en sık geçen ikili kelime gruplarıdır. Makine öğrenmesi literatürde daha çok veri bilimcisi pozisyonu ile ilişkilendirilen ve veri analizinden üst bir katman olan bir kavram olduğundan veri analizcisi iş tanımında yer alması ilgi çekici bir bulgu olarak karşımıza çıkmaktadır.



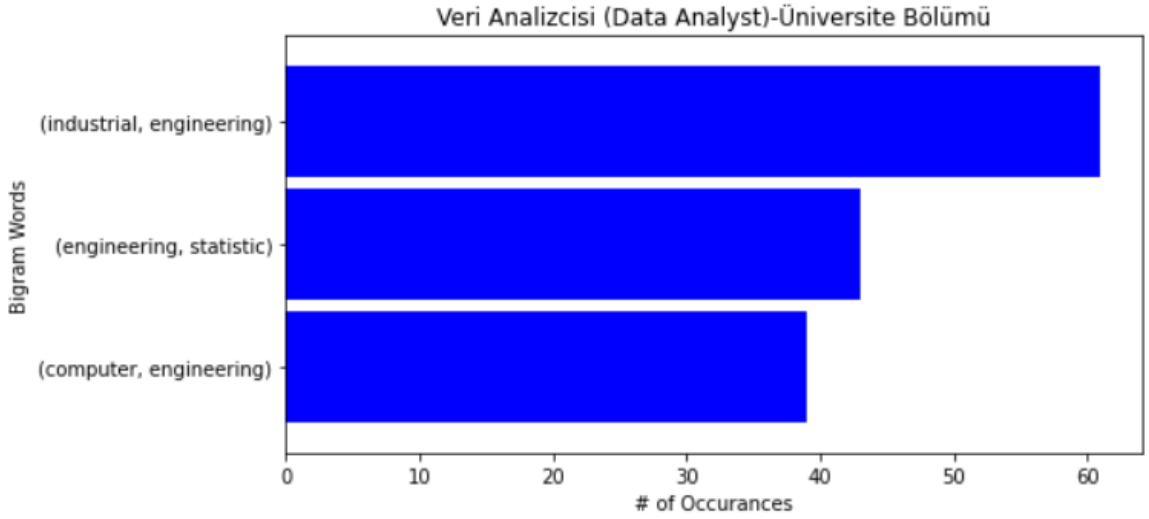

**Şekil 31.** Veri analizcisi iş tanımı bigram analizi



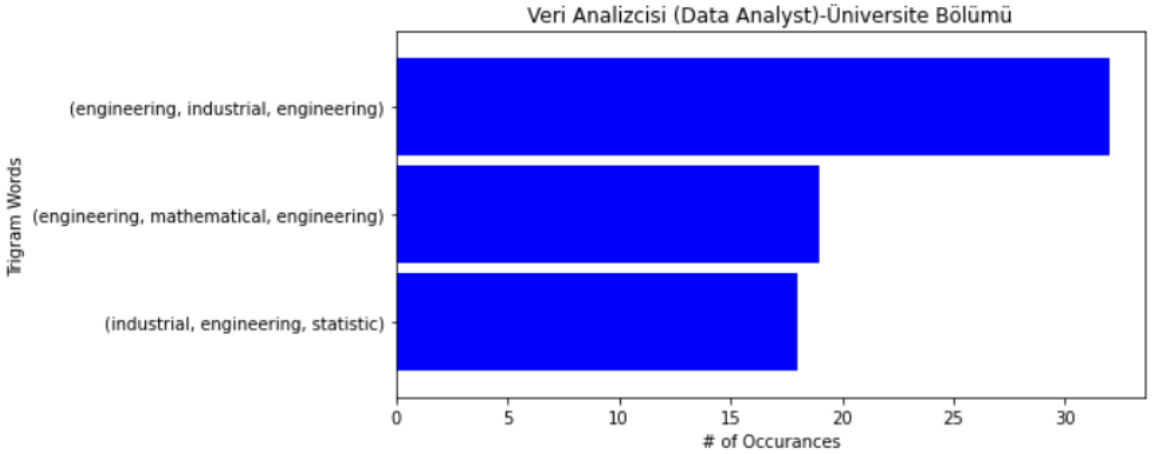

**Şekil 32.** Veri analizcisi iş tanımı trigram analizi

- **Üniversite Bölümü (University Department)**

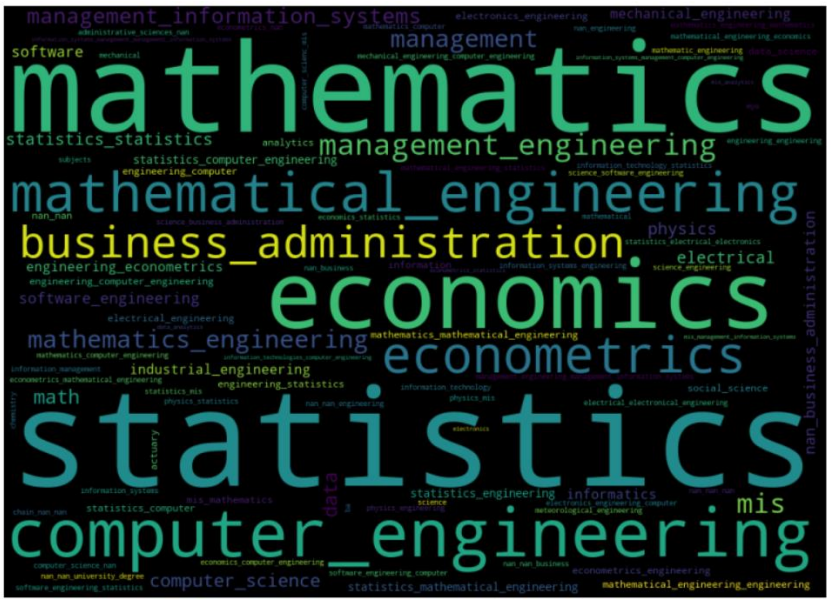
Veri analizcisi pozisyonunda istihdam edilmek istenen profiller için aranan mezuniyet bölümlerine bakıldığında endüstri mühendisliği, istatistik ve bilgisayar mühendisliğinin ilk üç sırada geldiği görülmektedir. Burada da aşağıdaki sonuçlar N-gram analizi kullanılarak elde edilmiştir.




**Şekil 33.** Veri analizcisi üniversite bölümü bigram analizi




**Şekil 34.** Veri analizcisi-üniversite bölümü-trigram analizi




**Şekil 35.** Veri analizcisi-üniversite bölümü-kelime bulutu

Veri kümesinde yer alan üniversite bölümü sütunu, kelime bulutu üzerinden analiz edildiğinde matematik, istatistik, ekonomi, işletme, bilgisayar mühendisliği, matematik mühendisliği gibi bölümlerin göze çarptığını görmekteyiz. Şirketlerin veri analizcisi pozisyonu için mezuniyet şartlarında geniş bir çerçeve çizdiği ve net bir odak belirlemediği söylenebilir. İşletme, ekonomi gibi bölümlerin veri analizi için tercih ediliyor olması ilgi çekici bir sonuç olarak ortaya çıkmaktadır.

## 3.4. VERİ BİLİMCİSİ İŞ İLANLARININ ANALİZİ

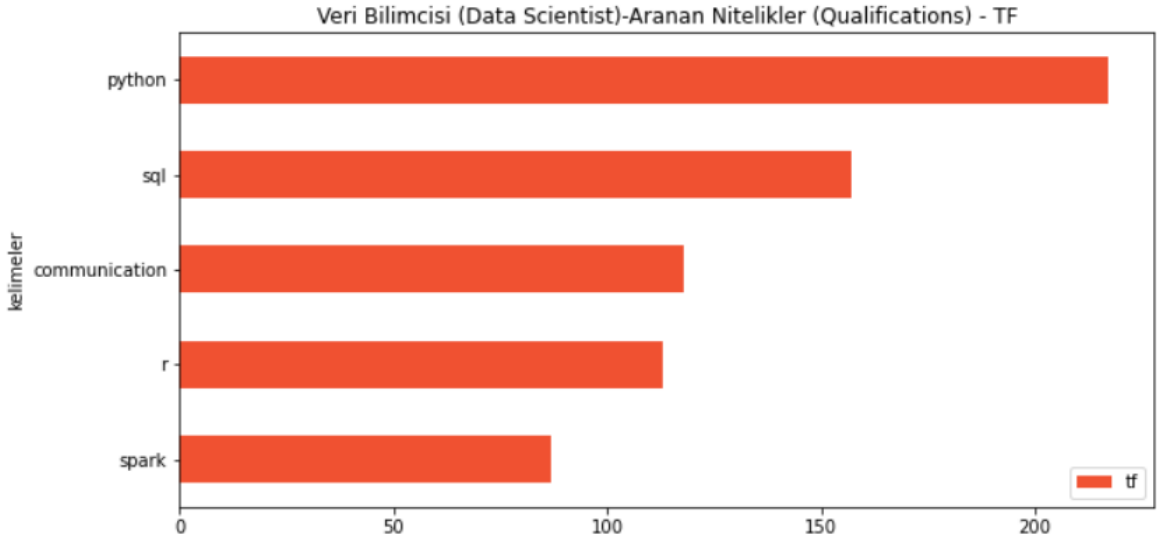
Veri bilimcisi (Data Scientist) pozisyonu için toplamda 232 ilan elde edilmiştir.

- **Aranan Nitelikler (Qualifications)**

Veri bilimcisi pozisyonu için elde edilen iş ilanları üzerinden aranan nitelikler sütunu analiz edildiğinde, Şekil 36'da terim frekansı grafiğinde Python, SQL ve İletişim yetkinliklerinin ilk üç sırada geldiği görülmektedir. Sonrasında ise R ve Spark bilgi ve yetkinlikleri gelmektedir.

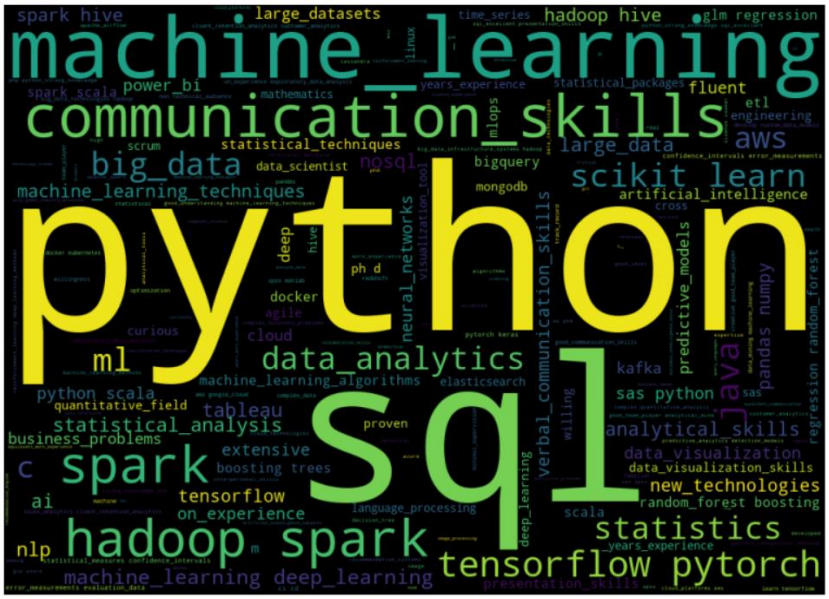
Veri bilimcisi pozisyonu için Python'a ek olarak veri analiz aracı olarak R yetkinliğinin de sayıldığını görmekteyiz.

Spark büyük veri işlemeye yarayan araçlardan birisidir. Hızı, kullanım kolaylığı, dinamik veri (streaming) ile çalışma imkânı, makine öğrenmesi amacıyla kullanılması Spark'ın veri mühendisleri tarafından büyük talep görmesini sağlasa da veri bilimcileri de veri analizi için Spark'tan faydalanmaktadır (Yeşilyurt, 2019). Bu kapsamda veri bilimcisi rolünde Spark konusunda bilgi ve yetkinliğin arandığı görülmektedir.




**Şekil 36.** Veri bilimcisi-aranan nitelikler-terim frekansı

Veri bilimi pozisyonu için aranan nitelikler için oluşan kelime bulutunda ön plana çıkan kelimeler, Python, SQL, makine öğrenmesi ve iletişim olmuştur. Bunun yanında ağırlığı küçük olsa da Spark, Hadoop da göze çarpan bilgi ve yetkinlikler arasındadır.




**Şekil 37.** Veri bilimcisi-aranan nitelikler-kelime bulutu

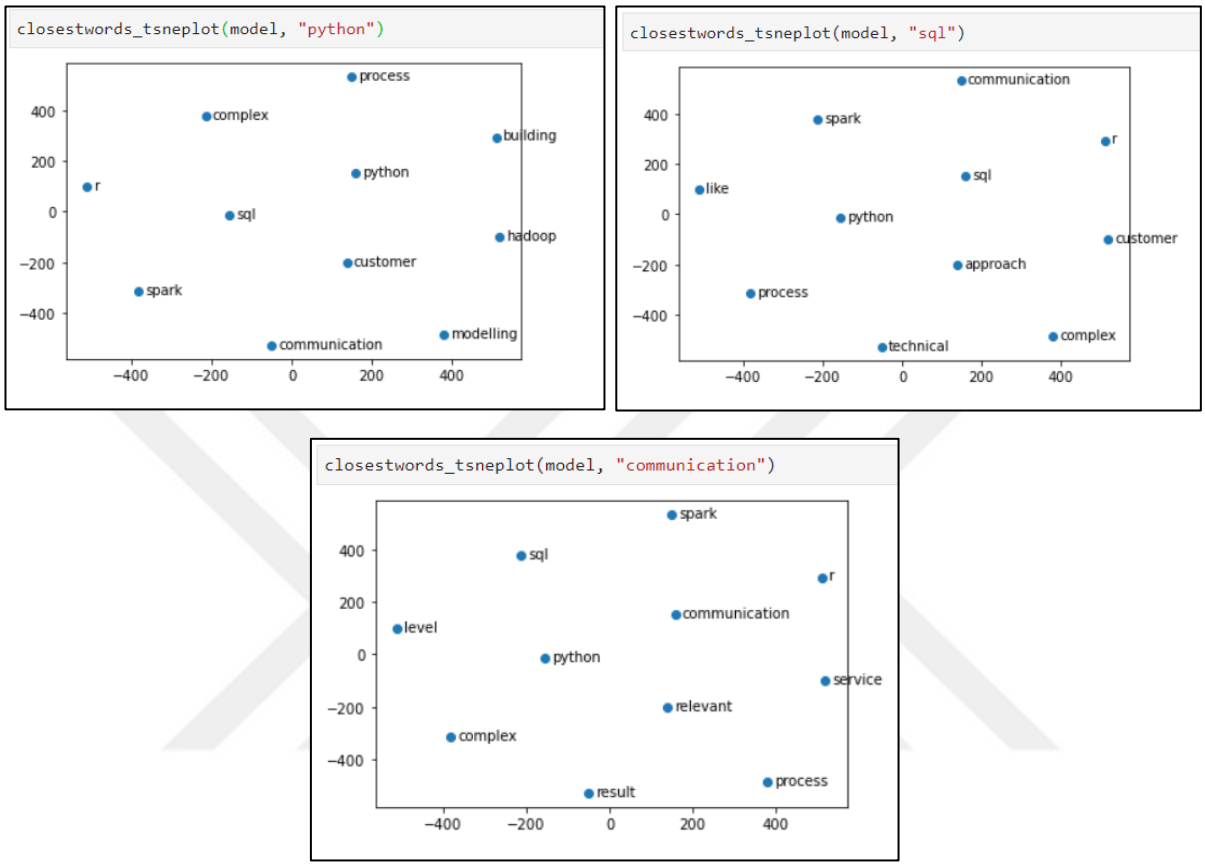
- **En Yakın Kelime Analizi**

Benzer kelimeler analizi Python, SQL, iletişim kelimeleri etrafındaki kelime kümesi gözlemlenmiştir.

Python yetkinliği etrafındaki en yakın kelime kümesine bakıldığında SQL ve Hadoop bilgi ve yetkinliklerinden anlamlı olarak bahsedilebileceği görülmektedir.

SQL yetkinliğine ait en yakın kelime kümesine bakıldığında, Python, Spark, iletişim, R bilgi ve yetkinlikleri olduğu görülmektedir.

İletişim yetkinliğine en yakın kelime kümesine bakıldığında, Python, SQL, Spark, R bilgi ve yetkinlikleri olduğu görülmektedir.

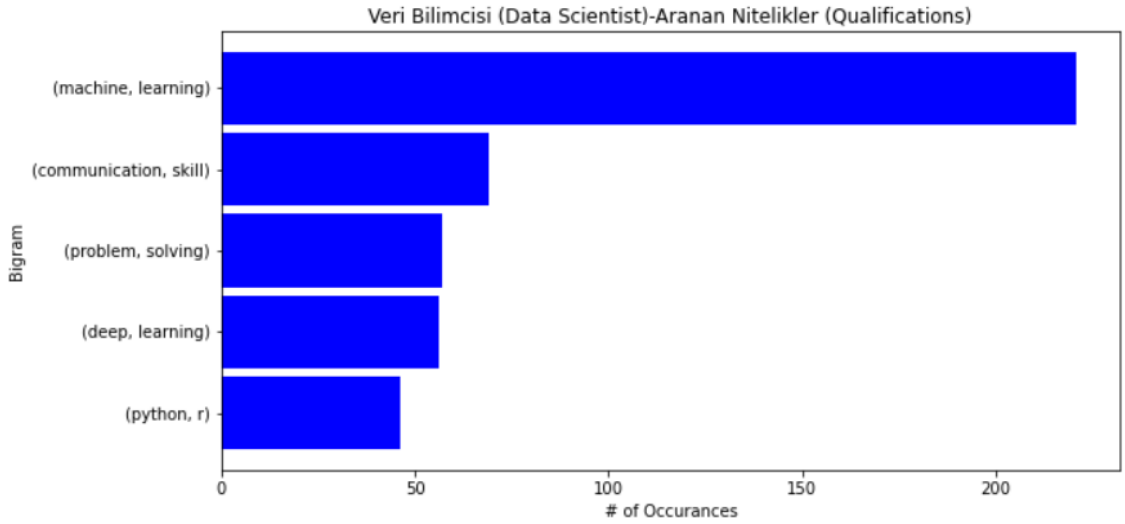



**Şekil 38.** Python-SQL-İletişim(communication) en yakın kelime analizi

- **Aranan Nitelikler-N-Gram Analizi**

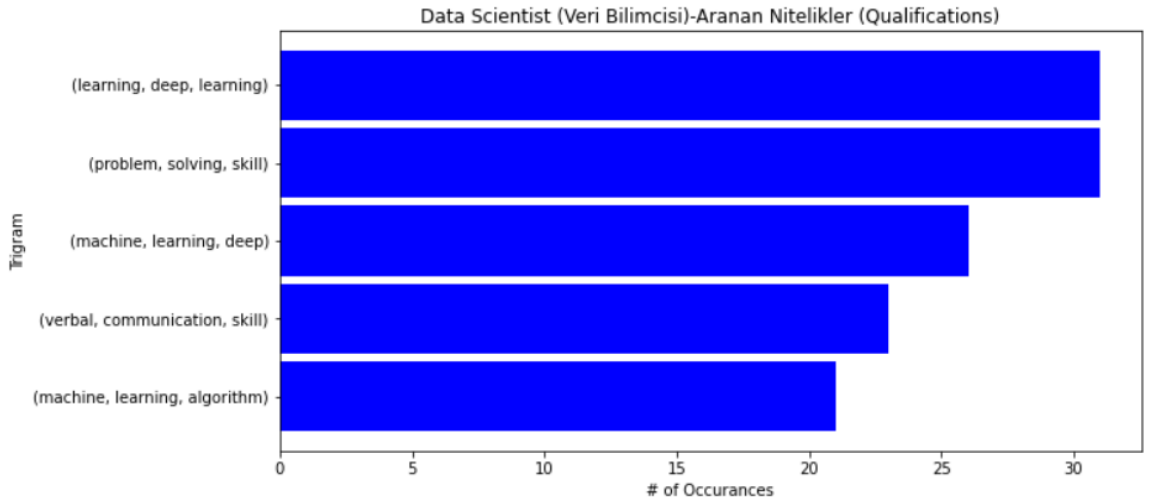
Veri bilimcisi pozisyonu için aranan nitelikler veri kümesindeki kelime grupları da tespit edilmek istenmiştir. Bu kapsamda ikili ve üçlü kelime gruplarına bakıldığında aşağıdaki tablolar elde edilmiştir.

İkili kelime grubunda, yoğun bir şekilde makine öğrenmesi kelime grubunun geçtiği görülmüştür. Bu durum Veri bilimcisi pozisyonu için ileri analiz teknikleri ve yetkinliği gerektiren makine öğrenmesinin önemli bir yetkinlik olarak öne çıktığı görülmektedir. Bununla beraber, iletişim becerileri, problem çözme, derin öğrenme, Python, R kelime gruplarının da makine öğrenmesine görece daha az sıklıkla geçen kelime grupları olduğu görülmektedir.




**Şekil 39.** Veri bilimcisi-aranan nitelikler-bi-gram analizi

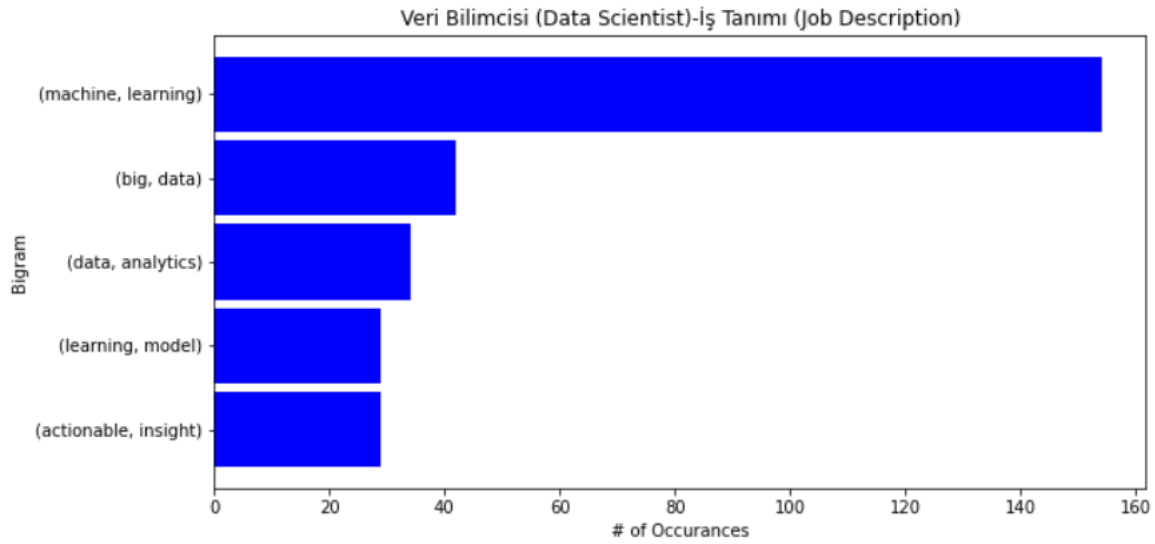
Veri Bilimcisi pozisyonu için aranan nitelikler veri kümesindeki üçlü kelime grupları analiz edildiğinde, ikili kelime gruplarını tamamlayıcı şekilde tablonun oluştuğu görülmektedir.




**Şekil 40.** Veri bilimcisi-aranan nitelikler-tri-gram analizi

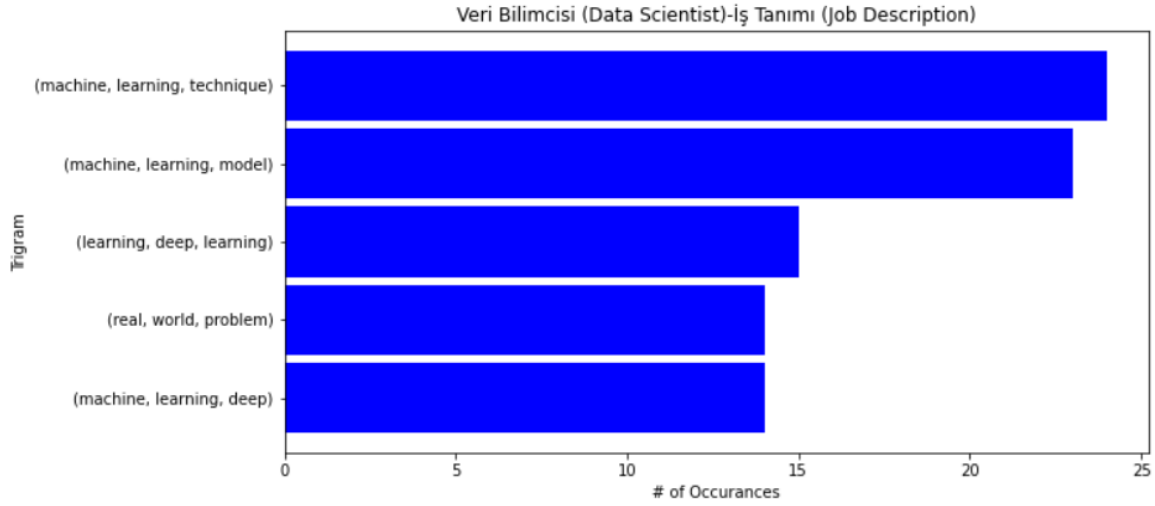
- **İş Tanımı (Job Description)**

Veri bilimcisi iş ilanları veri kümesinde iş tanımı sütunu N-gram analizine tabi tutulmuştur. İş tanımı veri kümesinde baskın olarak “makine öğrenmesi” kelime grubunun öne çıktığı görülmektedir. “Büyük veri”, “veri analitiği”, “eyleme geçirilebilir içgörü (actionable insight)” kelime grupları makine öğrenmesine görece sıklık derecesi az olmakla beraber geçen kelimeler olduğu görülmüştür. Aranan nitelikler veri kümesinden elde edilen sonuçlarda olduğu gibi, iş tanımı veri kümesinden elde edilen sonuçlarda da de makine öğrenmesi yetkinliğine işaret edilmektedir.




**Şekil 41.** Veri bilimcisi-iş tanımı-bigram analizi

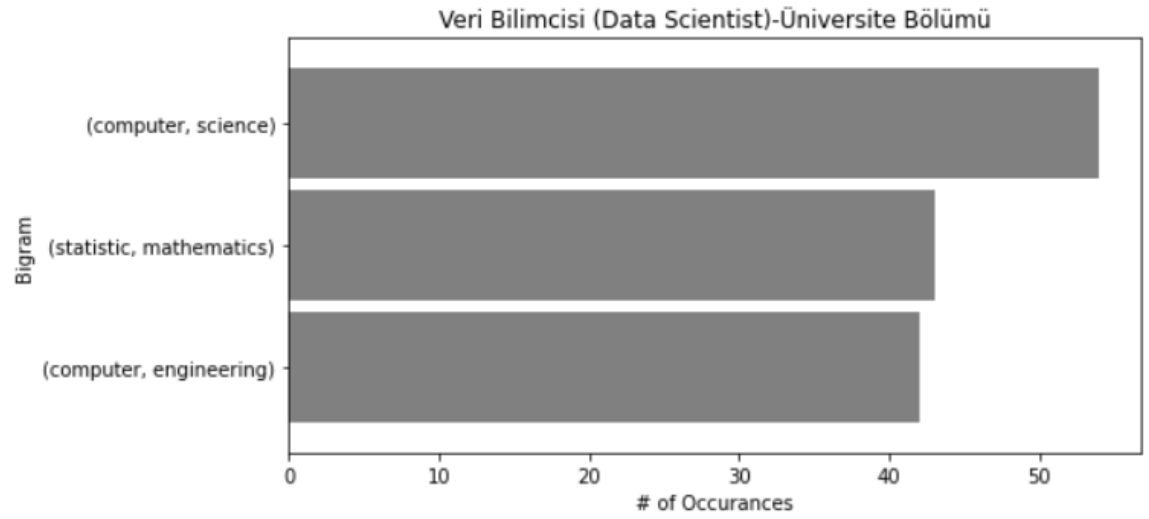
Üçlü kelime gruplarına bakıldığında, ikili kelime gruplarını tamamlayıcı şekilde kelimelerin oluştuğu görülmektedir. Bununla beraber, "gerçek dünya problemleri" kelime grubu veri biliminin problem çözme kapsamının gerçek problemlerden oluştuğu şeklinde yorumlanabilir.




**Şekil 42.** Veri bilimcisi-iş tanımı-trigram analizi

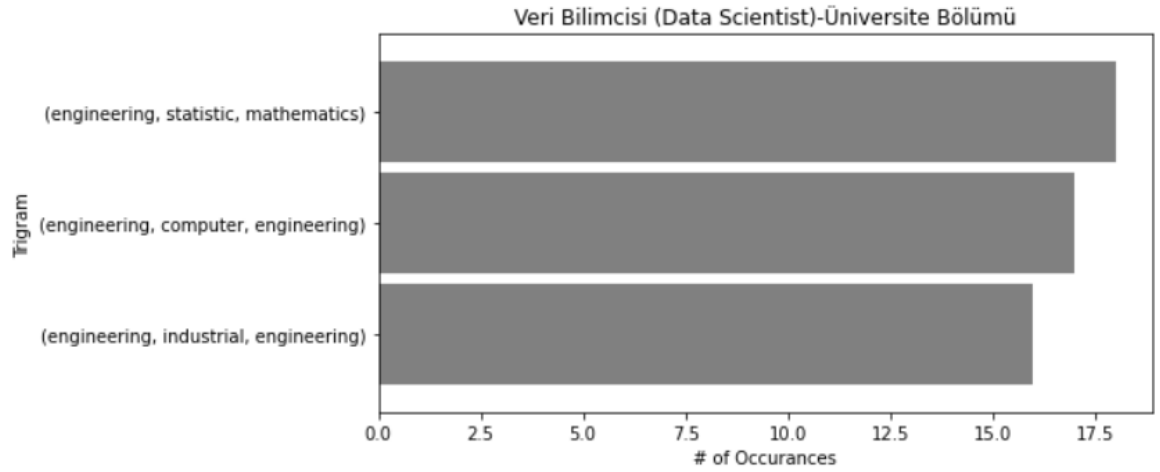
- **Üniversite Bölümü (University Department)**

Veri bilimcisi pozisyonu için mezuniyet şartlarına bakıldığında, *"bilgisayar bilimleri", "istatistik matematik", "bilgisayar mühendisliği"* bölümlerinin tercih edilen ilk üç bölüm arasında olduğu görülmektedir.



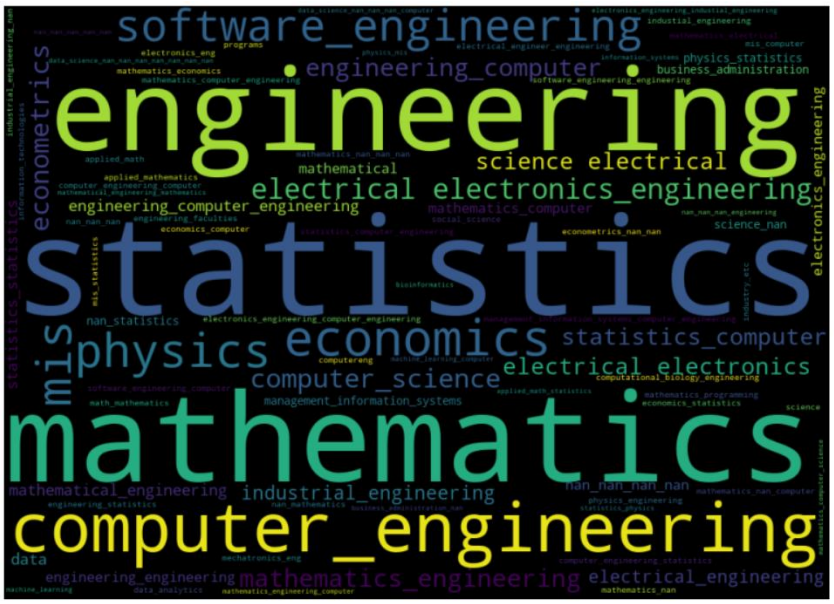

**Şekil 43.** Veri bilimcisi-üniversite bölümü-bigram analizi

Üçlü kelime grupları analizi bölüm konusunda yeni bir bilgi sunmamıştır.




**Şekil 44.** Veri bilimcisi-üniversite bölümü-trigram analizi

Bölüm mezuniyeti için kelime bulutuna bakıldığında, istatistik ve matematik bölümlerinin aynı oranda ön plana çıktığını bununla beraber, genel bir çerçeve olarak mühendislik bölümünün ön plana çıktığı görülmektedir. Bu durum, herhangi bir mühendislik bölümü şeklinde yorumlanabilir.




**Şekil 45.** Veri bilimcisi-üniversite bölümü-kelime bulutu

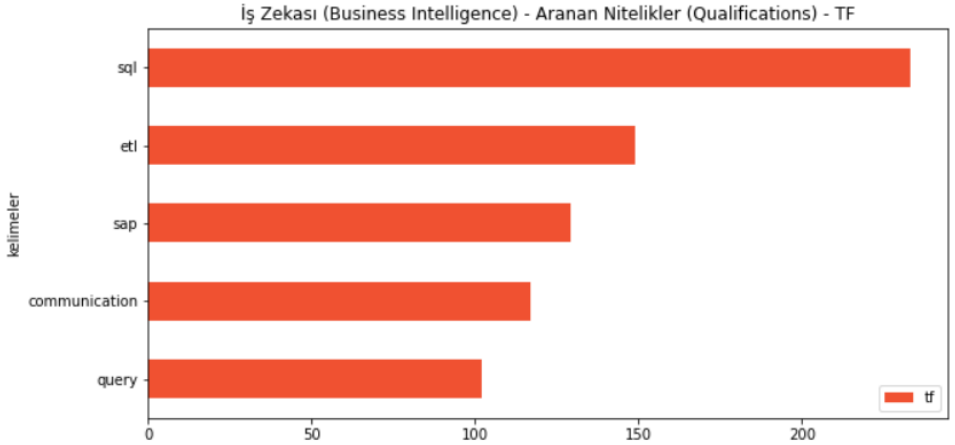
## 3.5. İŞ ZEKASI İŞ İLANLARININ ANALİZİ

İş Zekâsı (Business Intelligence) pozisyonu kapsamında 277 ilan elde edilmiştir.

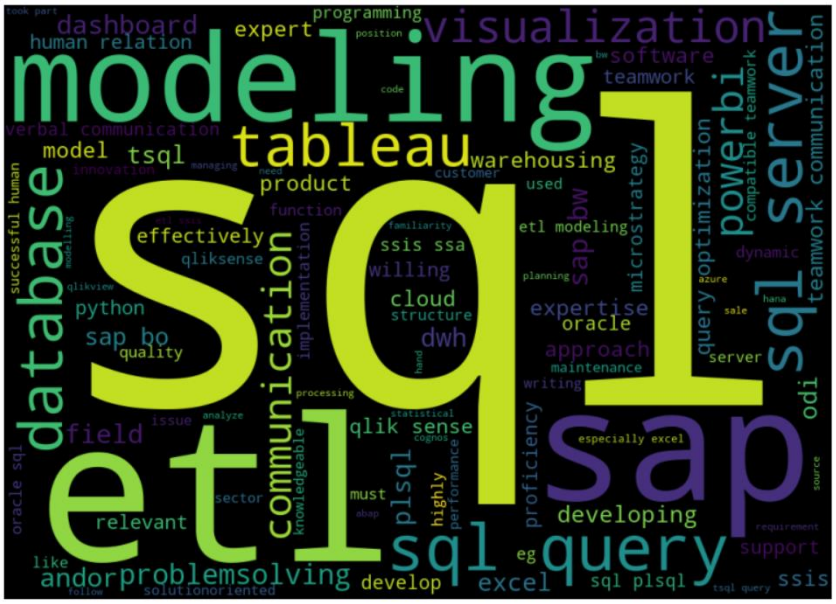
- **Aranan Nitelikler (Qualifications)**

İş Zekâsı pozisyonu kapsamında elde edilen iş ilanları veri kümesinin aranan nitelikler sütunu analiz edilmesi sonucu Şekil 46'da yer alan terim frekansı grafiği elde edilmiştir. Buna göre, şirketler tarafından ilk üç sırada aranan bilgi yetkinlik ve platformlar arasında SQL, ETL ve SAP yer aldığı görülmektedir.

ETL (Extract Transform Load) verinin farklı kaynak sistemlerden alınıp hedef sisteme yüklenmesini sağlayan bir veri entegrasyon platformudur. İlk üç sırada sayılan teknik yeterliliklerden sonra, iletişim yetkinliğinin iş zekâsı pozisyonu kapsamında da öne çıktığı görülmektedir.




**Şekil 46.** İş zekâsı-aranan nitelikler-terim sıklığı




**Şekil 47.** İş zekâsı-aranan nitelikler-kelime bulutu

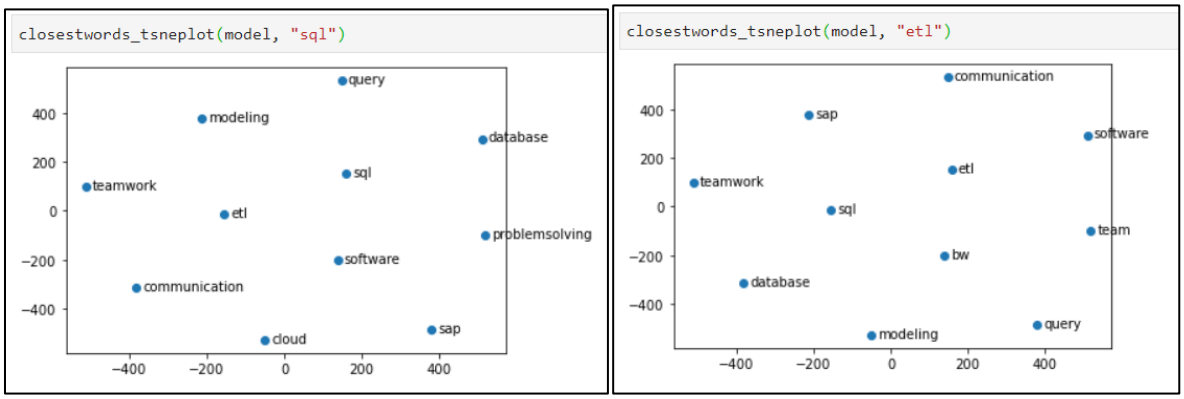
İş zekâsı pozisyonu için kelime bulutuna bakıldığında, SQL kelimesinin baskın bir şekilde ön planda olduğu görülmektedir. Bununla beraber kelime bulutunda, ETL, modelleme, SAP yetkinlik ve platformları görece daha az olsa da ön plana çıkan kelimeler arasında yer almıştır.

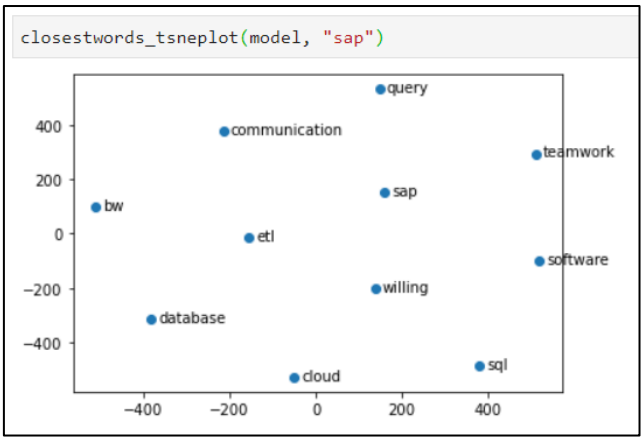
- **En Yakın Kelime Analizi**

SQL kelimesinin en yakın kelime kümesine bakıldığında, ETL, Modelleme, Sorgulama, Veri tabanı, problem çözme kelimelerinin olduğu görülmektedir.

ETL kelimesinin en yakın kelime kümesine bakıldığında, SQL, SAP, İletişim, yazılım, takım kelimelerinin olduğu görülmektedir. ETL kelimesinin en yakın kelime kümesinde, iletişim ve takım kelimeleri sosyal yetkinliklere vurgu yapması açısından dikkat çekicidir.

SAP, kelimesinin en yakın kelime kümesine bakıldığında ETL, İletişim, Sorgulama, Takım çalışması, yazılım kelimelerinin olduğu görülmektedir.






**Şekil 48.** SQL-ETL-SAP için en yakın kelime analizi

- **N-Gram Analizi**

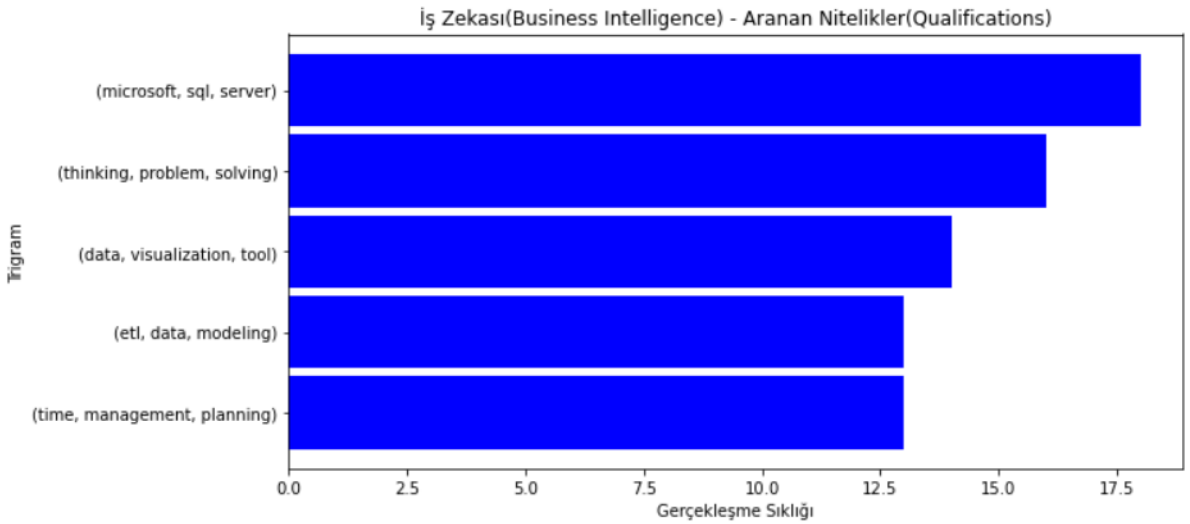
İş Zekası pozisyonu aranan nitelikler sütunu N-gram analizinde veri ambarı, power bi, analitik düşünme kelime grupları ilk üç sırada geldiği görülmektedir. Power bi aracından bahsedilmesi, iş zekası pozisyonunda şirketlerin veri görselleştirme ve raporlama yetkinliğine önem verdiği şeklinde değerlendirilebilir. Analitik düşünme sosyal yetkinliği de bu pozisyon kapsamında aranan yetkinlikler arasında sayılmıştır.

Problem çözme ve SQL Server, kelime grupları ilk beş sırada aranan yetkinlikler arasında sayılmıştır.




**Şekil 49.** İş zekâsı-aranan nitelikler-bigram analizi

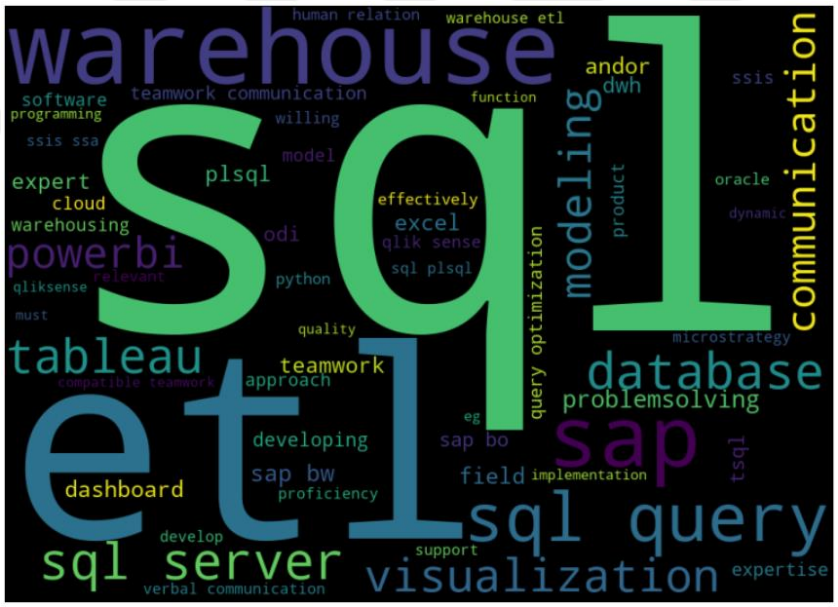
Üçlü kelime gruplarına bakıldığında, bazı kelimelerin ikili kelime gruplarının tamamlayıcıları olan kelimeler olduğu görülmektedir. Microsoft SQL Server bunlardandır. Bununla beraber, Veri görselleştirme araçları kelimesi, ikili kelime grubunda geçen power bi aracına ait yetkinlik vurgusunu destekleyici niteliktedir. ETL Veri Modelleme kelime grubu ETL platformu ve bu alandaki yetkinliğe işaret etmektedir.




**Şekil 50.** İş zekâsı-aranan nitelikler-trigram analizi

- **Kelime Bulutu-Aranan Nitelikler**

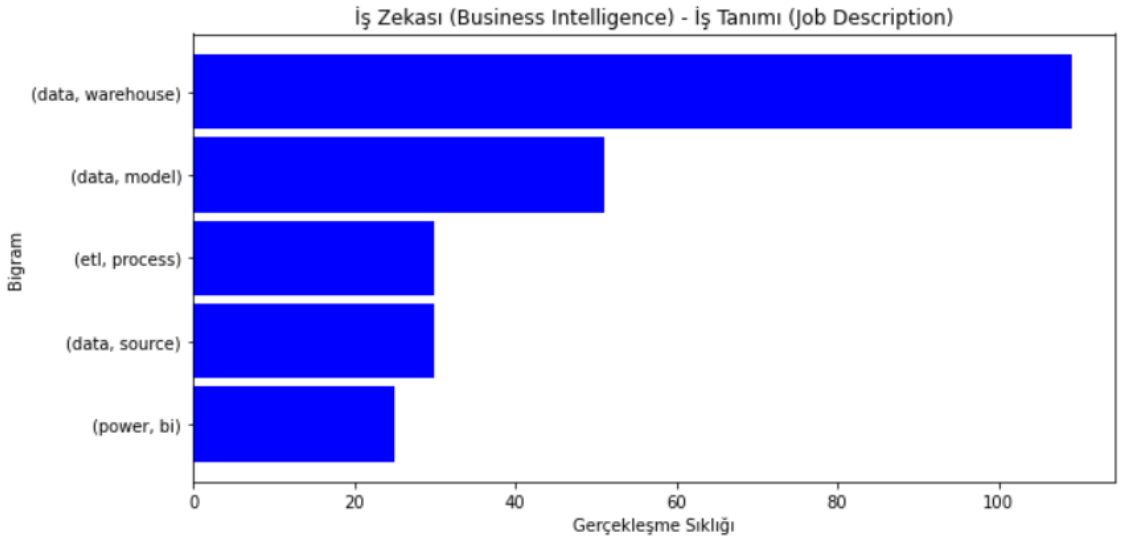
Kelime bulutunda iş zekâsı pozisyonunda aranan nitelikler arasında SQL yetkinliğinin ön plana çıktığı görülmektedir. Sonrasında ETL ve Warehouse kelimeleri göze çarpmaktadır. Yine SQL Server, SQL Query, Tableau, SAP teknik yetkinliklerinin ön plana çıktığı görülmektedir.




**Şekil 51.** İş zekâsı-aranan nitelikler-kelime bulutu

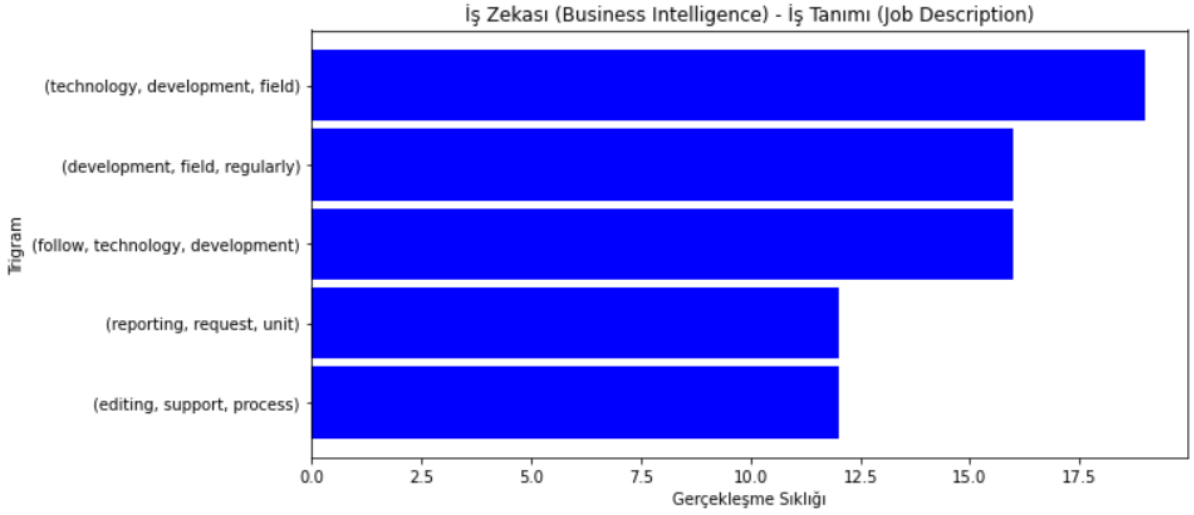
- **İş Tanımı (Job Description)**

İş zekâsı pozisyonunda veri ambarı kelimesi ikili kelime grubunda ilk sırada gelmektedir. Sonrasında daha düşük sıklıkta veri modelleme, ETL süreçleri, veri kaynağı ve Power BI kelimelerinin yer aldığı görülmüştür.




**Şekil 52.** İş zekâsı-iş tanımı-bigram analizi

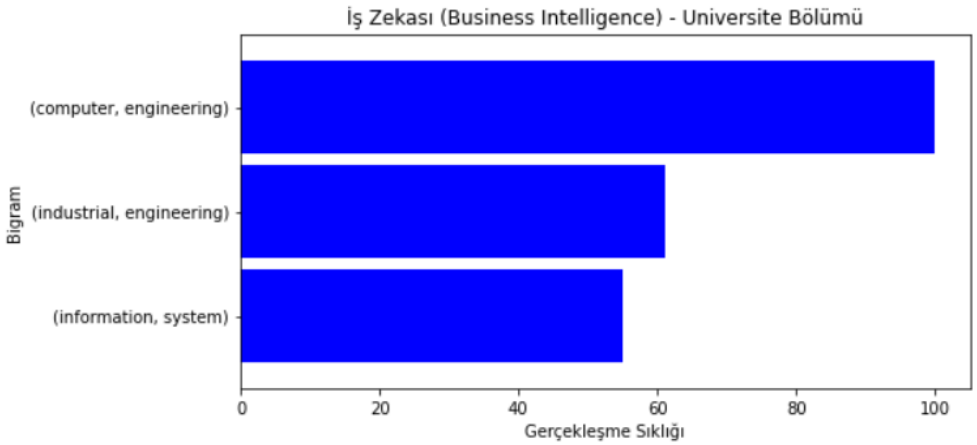
Üçlü kelime gruplarına baktığımızda teknolojik gelişmeleri takip etmeye yönelik, "teknoloji gelişim alanı" ve "teknoloji gelişimini takip" kelime gruplarının olduğunu görmekteyiz.




**Şekil 53.** İş zekâsı-iş tanımı-trigram analizi

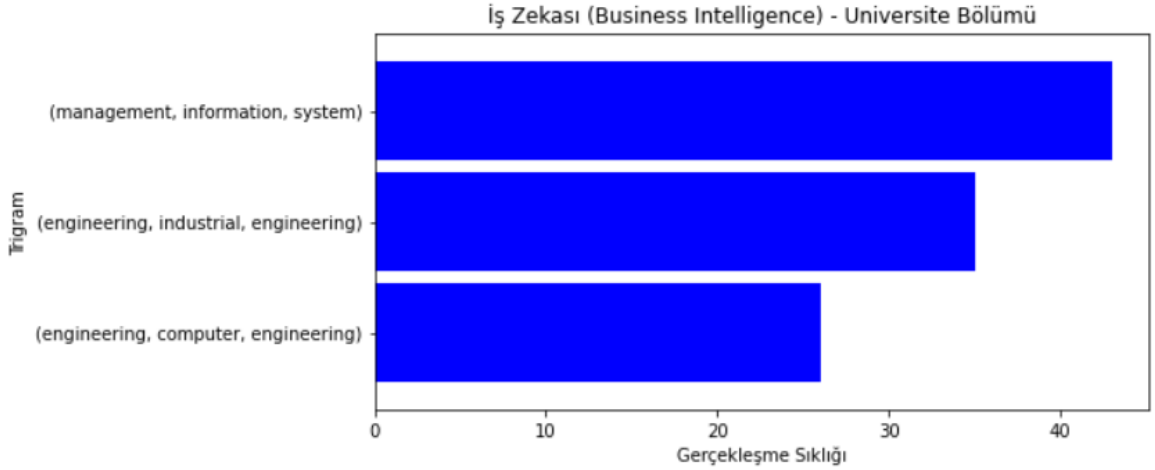
- **Üniversite Bölümü**

İş zekâsı pozisyonunda bölüm mezuniyet kriterleri olarak bilgisayar mühendisliği, endüstri mühendisliği ve bilişim sistemleri şeklinde ilk üç sıralamada yer alan kelimeler görülmektedir.




**Şekil 54.** İş zekâsı-üniversite bölümü-bigram analizi

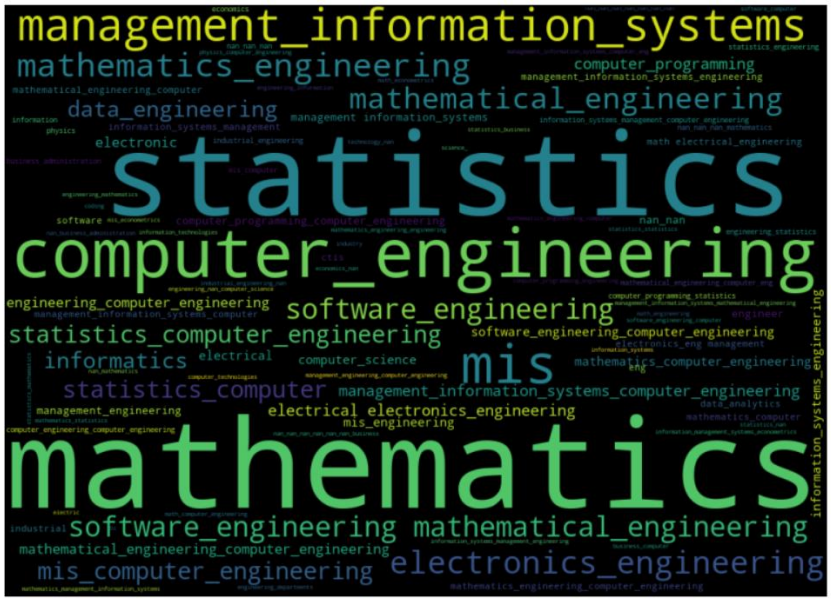
İkili kelime grubunda bilişim sistemleri olarak geçen kelime grubunun yönetişim bilişim sistemleri olarak tamamlandığını görmekteyiz. Diğer kelime grupları ise ikili kelime gruplarından bir farklılık göstermemiş veya bir bilgi tamamlama işlevi göstermemiştir.




**Şekil 55.** İş zekâsı-üniversite bölümü-trigram analizi

- **Kelime Bulutu-Üniversite Bölümü**

Kelime bulutu üzerinden bölüm kriterlerine bakıldığında matematik ve istatistik bölümlerinin ön plana çıktığı görülmektedir. Bilgisayar mühendisliği ve yönetim bilişim sistemleri de ön plana çıkan bölümler arasında yer almaktadır.




**Şekil 56.** İş zekâsı-üniversite bölümü-kelime bulutu

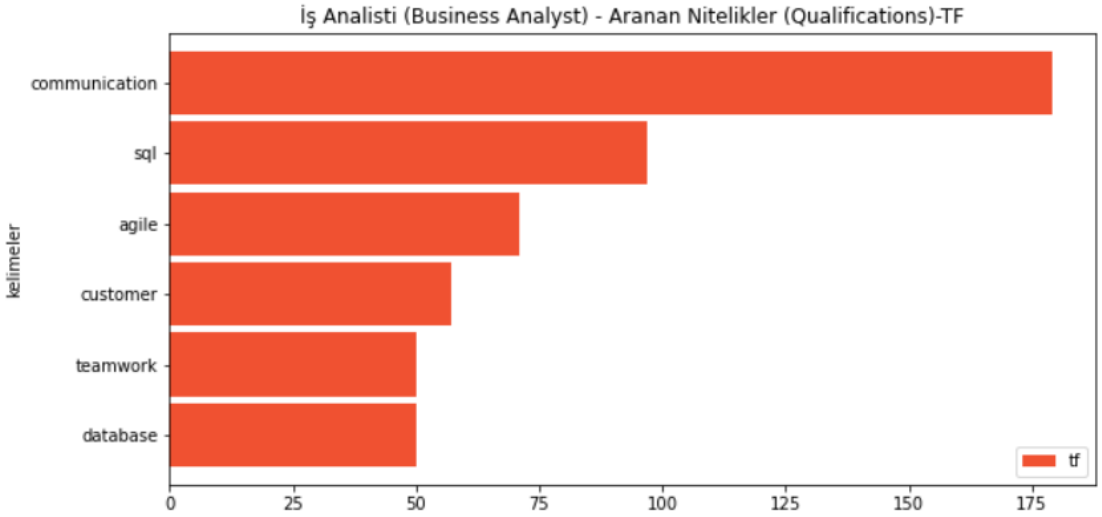
## 3.6. İŞ ANALİSTİ İŞ İLANLARININ ANALİZİ

İş Analisti (Business Analyst) pozisyonu için toplamda 253 ilan toplanmıştır.

- **Aranan Nitelikler (Qualifications)**

İş Analisti pozisyonunda aranan nitelikleri sütunu terim frekansı yardımıyla analiz edildiğinde Şekil 57'de görüldüğü gibi iletişim, SQL, "çevik (agile)" kelimeleri ilk üç sırada yer almaktadır. İletişim kelimesi baskın olarak cön plana çıkmıştır. Teknik bir yetkinlikten ziyade sosyal bir becerinin ilk sırada gelmesi dikkat çekici bir sonuç olarak karşımıza çıkmaktadır. En çok tekrar eden ikinci kelime, SQL yetkinliği olmuştur. Çevik (agile) kelimesi proje yönetimi süreçlerinde son dönemde sıkça karşımıza çıkan bir yetkinliktir. İş analisti pozisyonu için çevik yönteme vurgu yapan agile kelimesi, İş analisti pozisyonunda bu yöntemle uyumlu şekilde çalışma disiplinine işaret edildiği şeklinde değerlendirilebilir.

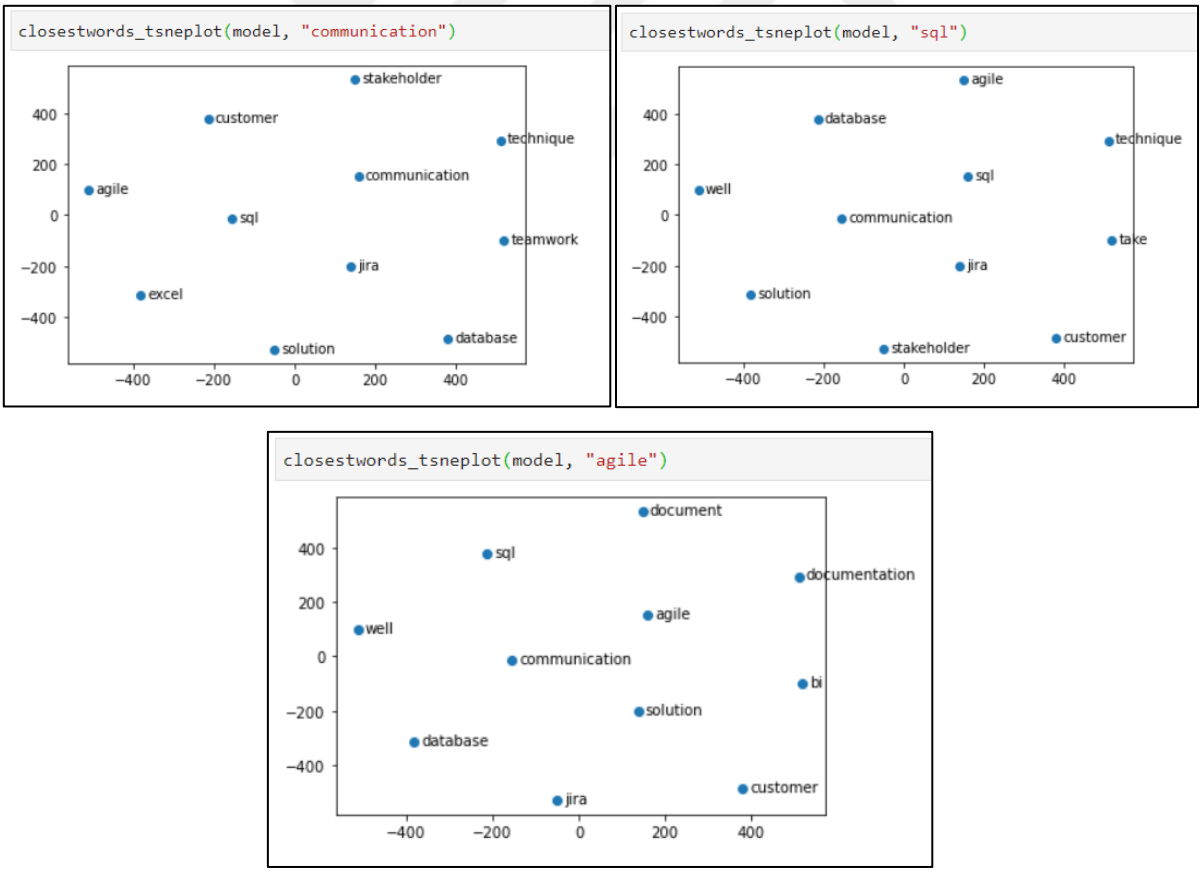
Müşteri (customer), takım çalışması (teamwork), veri tabanı (database) kelimeleri de ilk üç kelimeden sonra en çok tekrar eden kelimeler olmuştur.




**Şekil 57.** İş analisti-aranan nitelikler-terim sıklığı

- **En Yakın Kelime Analizi**

İletişim kelimesinin en yakın kelime kümesine bakıldığında, SQL, Müşteri, takım çalışması, jira kelimeleri gelmektedir.




**Şekil 58.** İletişim(communication)-SQL-Çevik(agile) kelimelerinin en yakın kelime analizi

SQL, kelimesinin en yakın kelime kümesine bakıldığında, İletişim, veri tabanı, çevik, jira kelimeleri olduğu görülmektedir.

Çevik (agile) kelimesinin en yakın kelime kümesine bakıldığında, SQL, doküman, dokümantasyon, çözüm ve iletişim kelimelerinin olduğu görülmektedir.



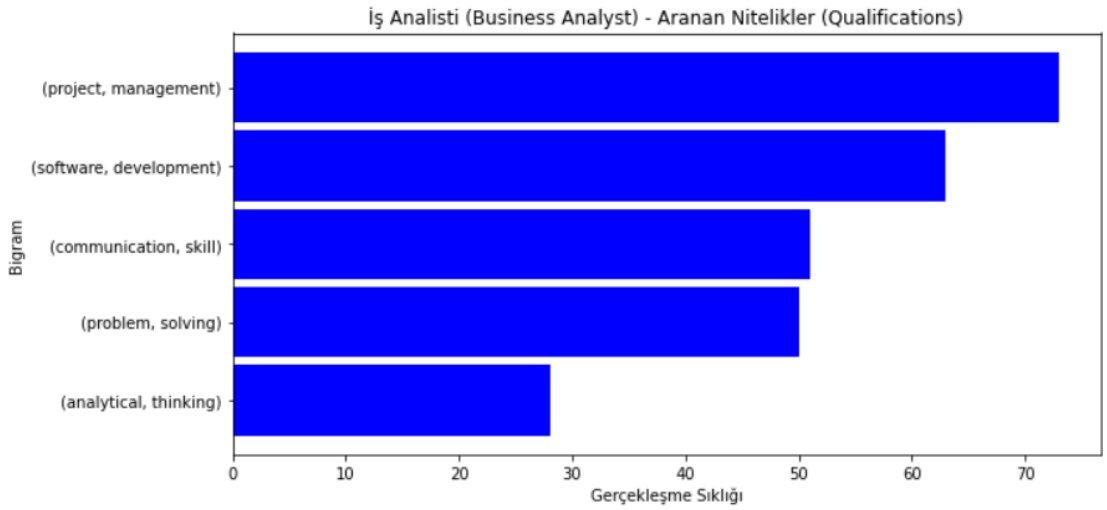

**Şekil 59.** İş analisti-aranan nitelikler-kelime bulutu

İş Analisti pozisyonu için kelime bulutuna bakıldığında, aranan yetkinlikler arasında SQL yetkinliğinin ön plana çıktığı görülmektedir. İletişim becerisi ise, ikinci sırada ön plana çıkan bir yetkinlik olarak karşımıza çıkmaktadır.

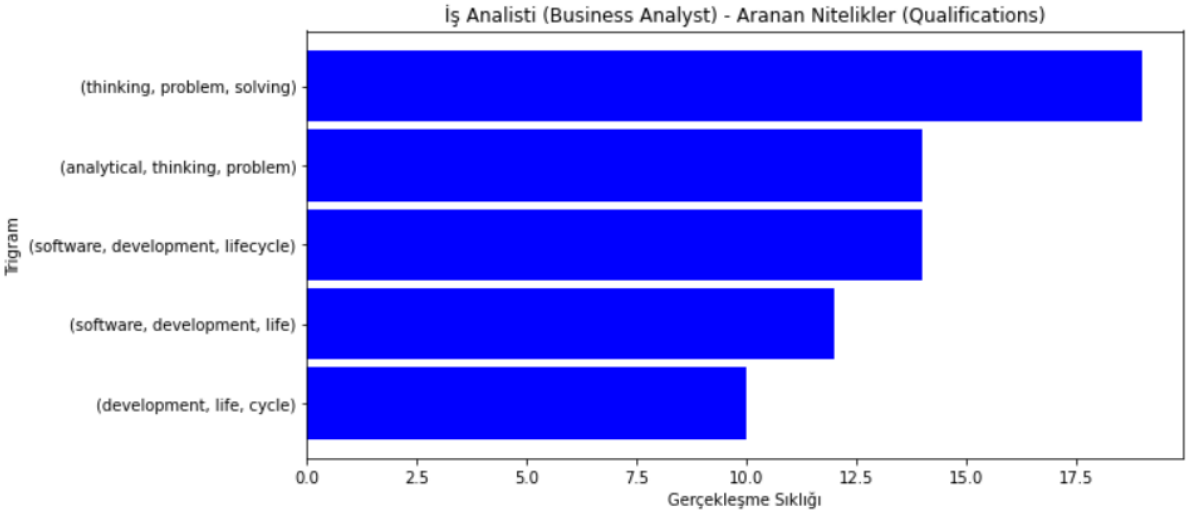
- **N-Gram Analizi-Aranan Nitelikler**

İş Analisti pozisyonunda tekrar eden ikili kelimeler kümesi, kelime sıklıkları grafiğinde eksik kalan alanları netleştirerek, genel çerçeveyi daha belirgin kılmaktadır. En çok geçen kelime ikilisi "Proje Yönetimi (Project management)" kelimesi olmuştur. Bir önceki grafikte geçen "iletişim, çevik, müşteri, takım çalışması, dokümantasyon, Jira, sunum, scrum" kelimeleri proje yönetiminde en çok ihtiyaç duyulan yetkinler olarak sayılabilir.

İkinci sırada gelen yazılım geliştirme (software development) ise, bir önceki grafikte, SAP, UML kelimeleri ile ilişkilidir. Sonrasında, problem çözme, iletişim becerileri, analitik düşünme, yüz yüze iletişim (verbal communication), analitik beceriler, zaman yönetimi (time management) kelime grupları karşımıza çıkmaktadır.



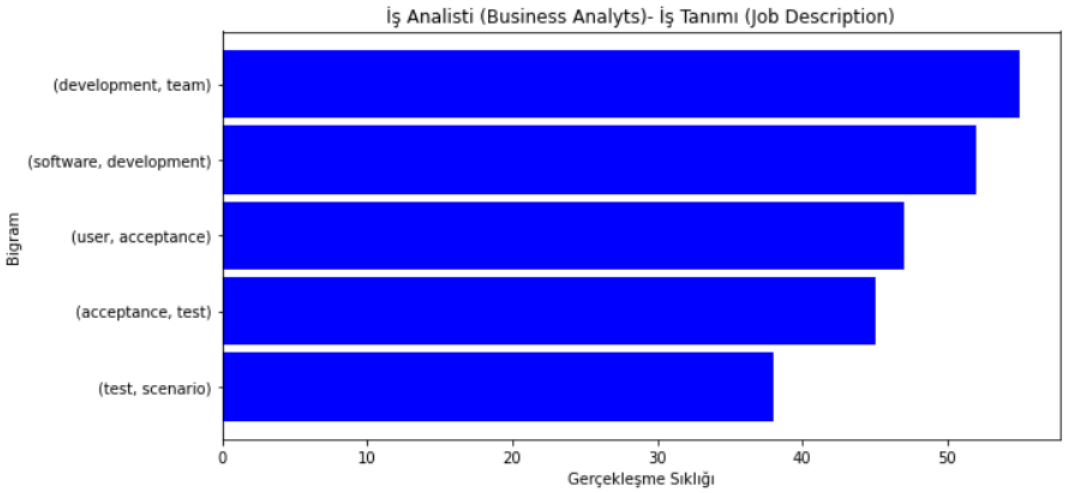

**Şekil 60.** İş analisti-aranan nitelikler-bigram analizi




**Şekil 61.** İş analisti-aranan nitelikler-trigram analizi

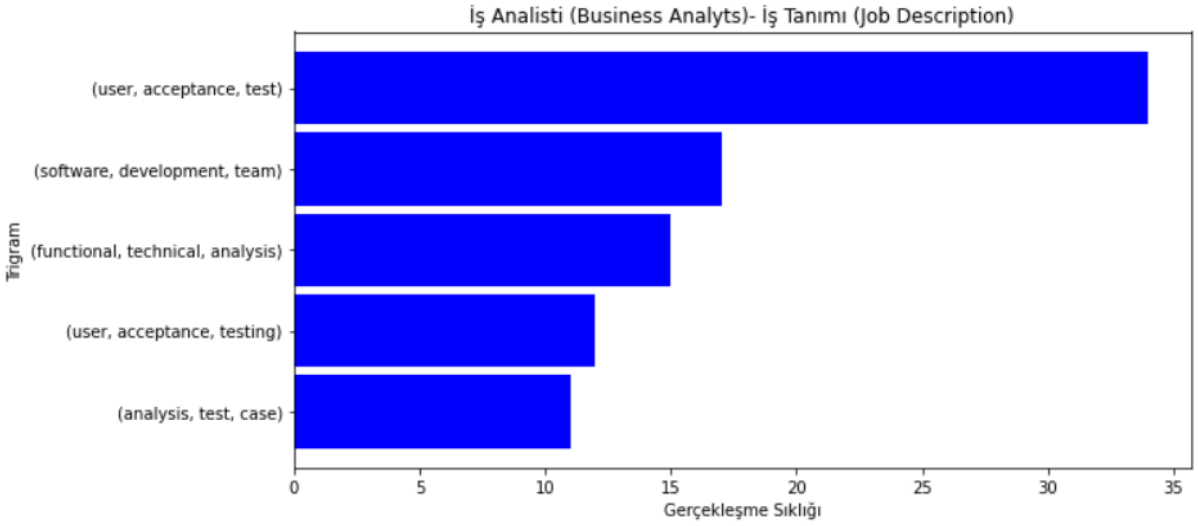
- **İş Tanımı (Job Description)**

İş analisti pozisyonu için iş tanımı veri kümese N-gram analizi uygulandığında, ikili kelime grubunda yeterince anlamlı sonuçlar elde edilemediği görülmüştür.




**Şekil 62.** İş analisti-iş tanımı-bigram analizi

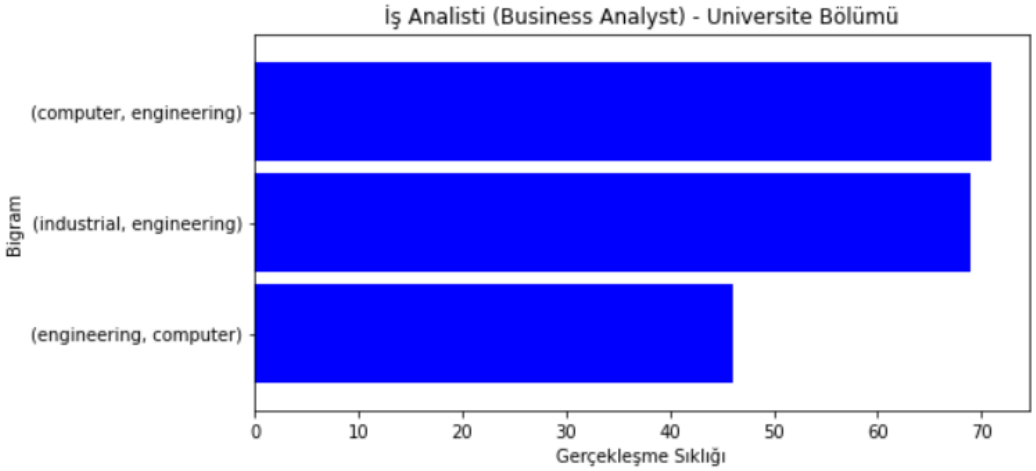
Üçlü kelime gruplarına bakıldığında, kullanıcı kabul testi, yazılım geliştirme takımı, fonksiyonel ve teknik analiz, kullanıcı kabul testleri kelime gruplarının olduğu görülmektedir.




**Şekil 63.** İş analisti-iş tanımı-trigram analizi

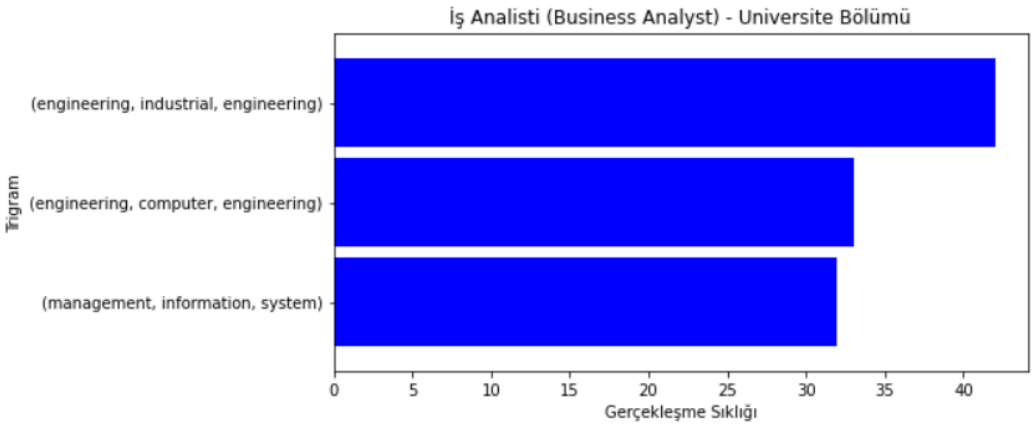
- **Üniversite Bölümü (University Department)**

İş analisti pozisyonu için beklenen mezuniyet bölümlerinin bilgisayar mühendisliği, endüstri mühendisliği olduğu görülmektedir.




**Şekil 64.** İş analisti-üniversite bölümü-bigram analizi

Üçlü kelime gruplarında yönetim bilişim sistemleri bölümünün de talep edilen bölümler arasında sayıldığı görülmektedir.



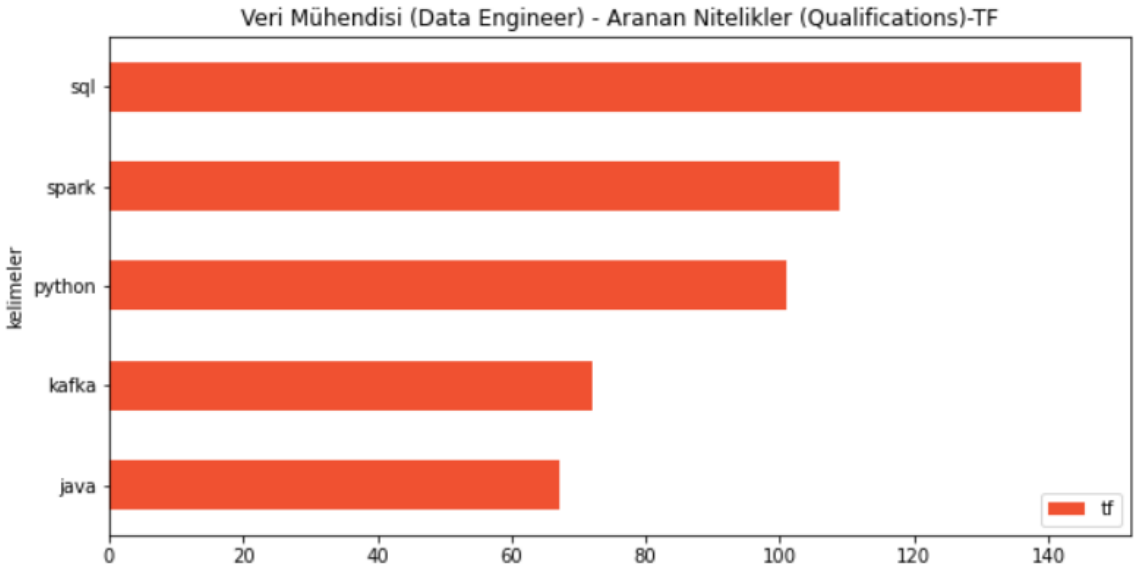

**Şekil 65.** İş analisti-üniversite bölümü-trigram analizi

## 3.7. VERİ MÜHENDİSİ İŞ İLANLARININ ANALİZİ

Veri mühendisi (Data Engineer) pozisyonu alanında toplamda 144 ilan elde edilmiştir.

- **Aranan Nitelikler (Qualifications)**

Veri mühendisi pozisyonu için aranan nitelikler sütunu üzerinden kelime sıklık analizine bakıldığında SQL, Spark ve Python kelimelerin ilk üç sırada yer aldığı görülmektedir. Sonrasında ise Kafka ve Java kelimelerinin geldiği görülmektedir.

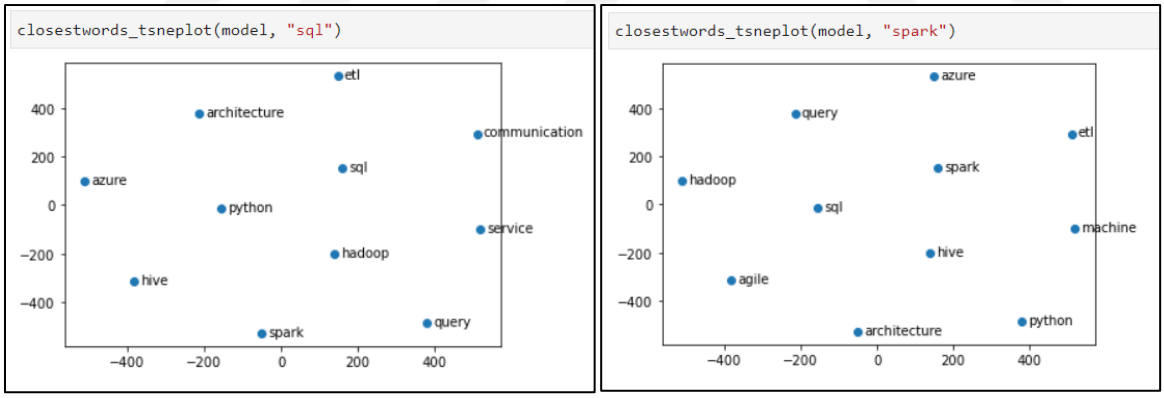


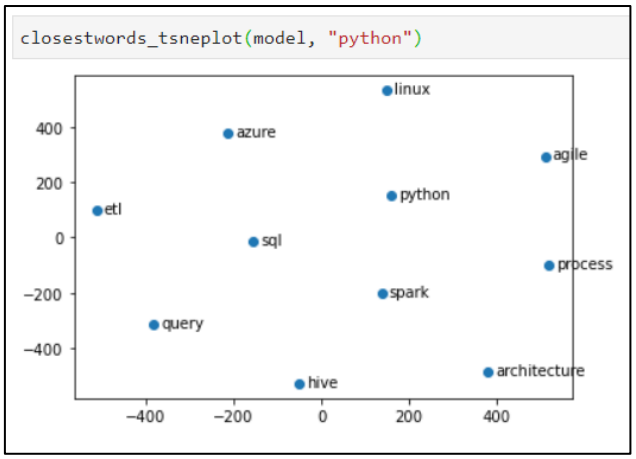

**Şekil 66.** Veri mühendisi-aranan nitelikler-terim sıklığı

- **En Yakın Kelime Analizi**

Terim frekansı grafiğinden elde edilen bilgiler doğrultusunda, en sık geçen kelimelere en yakın kelime kümesinde neler olduğu tespit edilmeye çalışılmıştır.

SQL kelimesinin en yakın kelime kümesine bakıldığında, Python, mimari, ETL, iletişim, Hadoop teknik ve sosyal yetkinliklerinin yer aldığı görülmektedir.





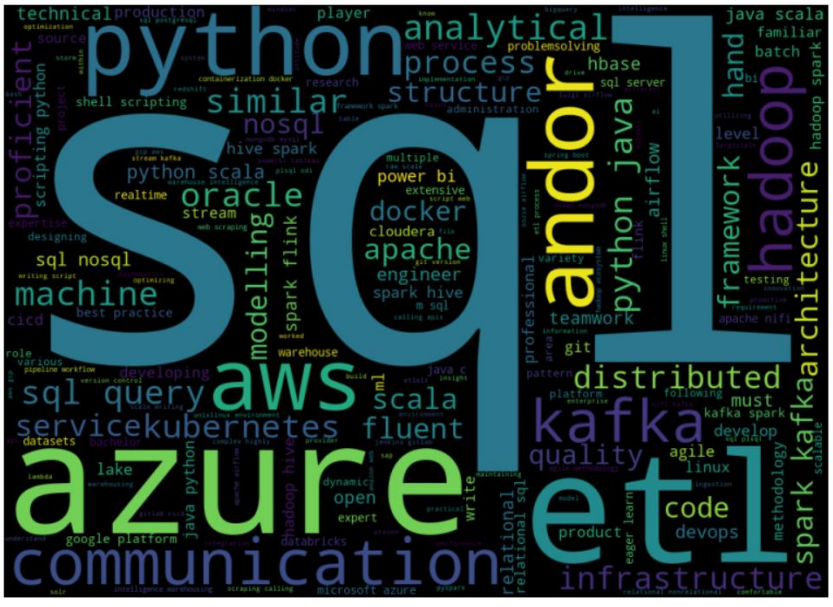

**Şekil 67.** SQL-Spark-Python kelimelerinin en yakın kelime analizi

Spark kelimesinin en yakın kelime kümesine bakıldığında SQL, Azure, sorgu, ETL, Hive kelimelerinin olduğu görülmektedir.

Python kelimesinin en yakın kelime kümesine bakıldığında, SQL, Azure, Linux, çevik, süreç, Spark kelimelerinin olduğu tespit edilmiştir.

- **Kelime Bulutu-Aranan Nitelikler**

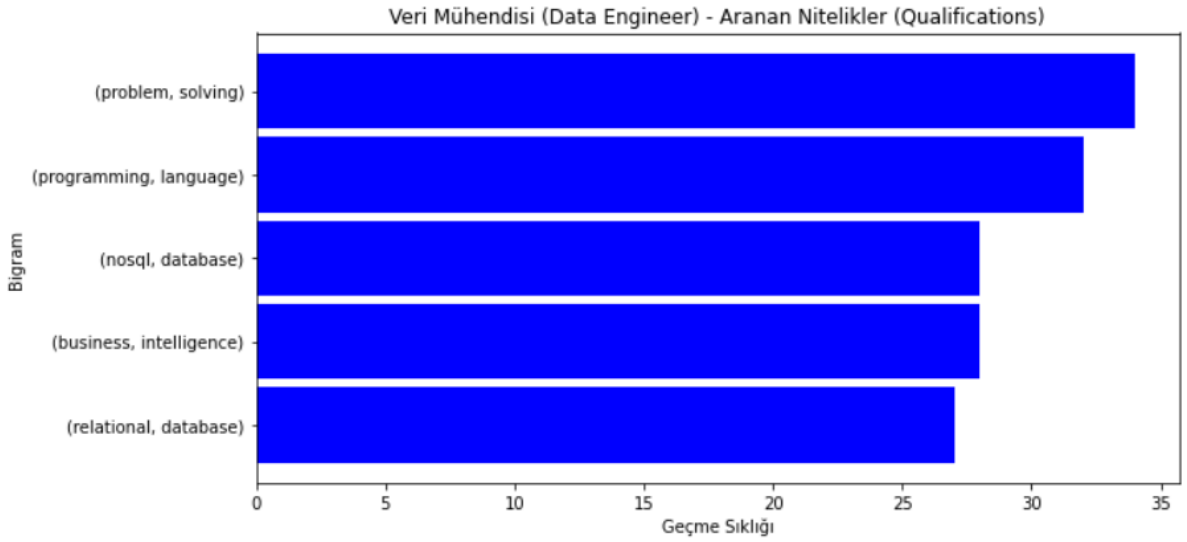
Veri mühendisi pozisyonunun aranan nitelikler veri kümesinin kelime bulutunda SQL kelimesinin ilk ön planda olduğu görülmektedir. Sonrasında, Azure, ETL ve Python kelimeleri kısmen öne çıkan kelimeler olmuştur.




**Şekil 68.** Veri mühendisi-aranan nitelikler-kelime bulutu

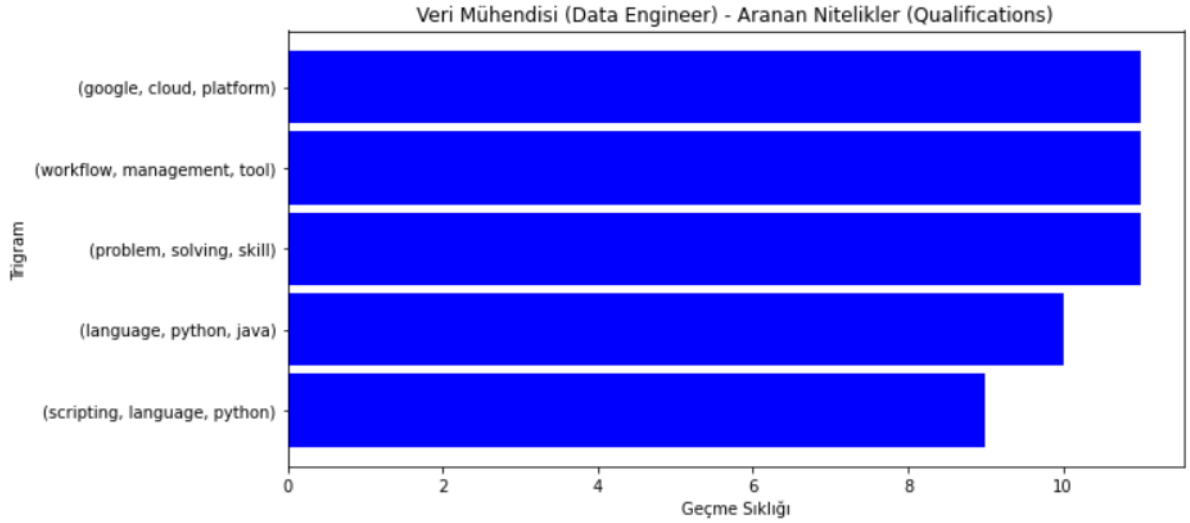
- **N-Gram Analizi-Aranan Nitelikler**

Veri mühendisi pozisyonu aranan nitelikler N-gram analizine bakıldığında, problem çözme, programlama dilleri, NoSQL veri tabanı kelimelerinin ilk üç sırada yer aldığı görülmektedir. Sonrasında, iş zekâsı ve ilişkisel veri tabanı kelimeleri gelmektedir.




**Şekil 69.** Veri mühendisi-aranan nitelikler-bigram analizi

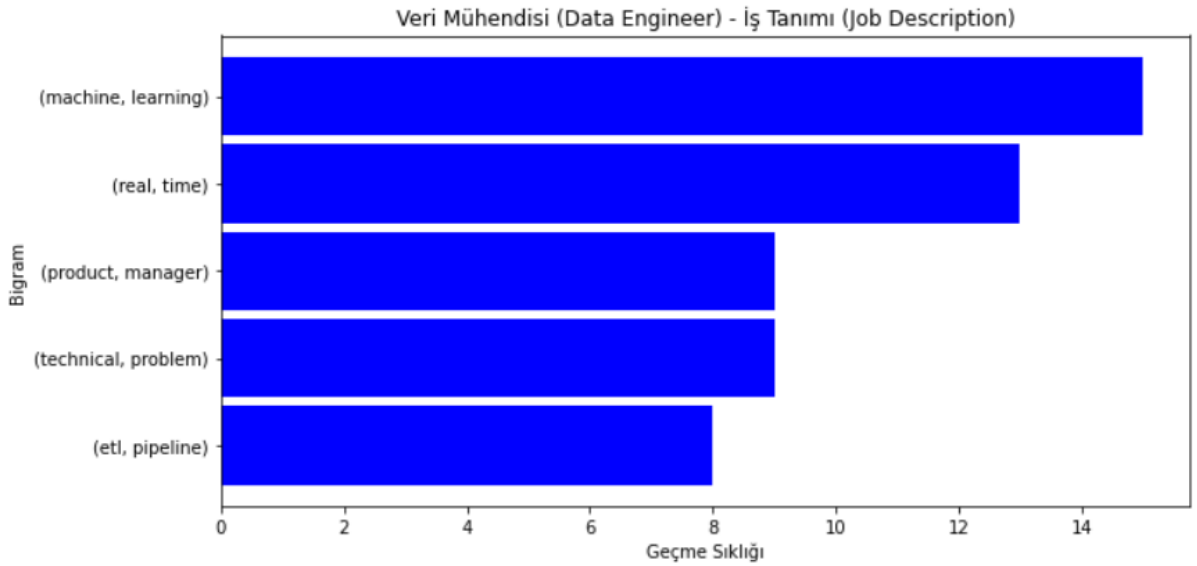
Üçlü kelime gruplarına bakıldığında, sırasıyla Google bulut platformları, iş akışı yönetim araçları, problem çözme becerileri ilk üç sırada gelmektedir. Sonrasında, yukarıda programlama dillerinden Python ve Java'nın tercih edildiğini dördüncü sıradaki "Python, java dili" kelime grubundan anlaşılmaktadır.




**Şekil 70.** Veri mühendisi-aranan nitelikler-trigram analizi

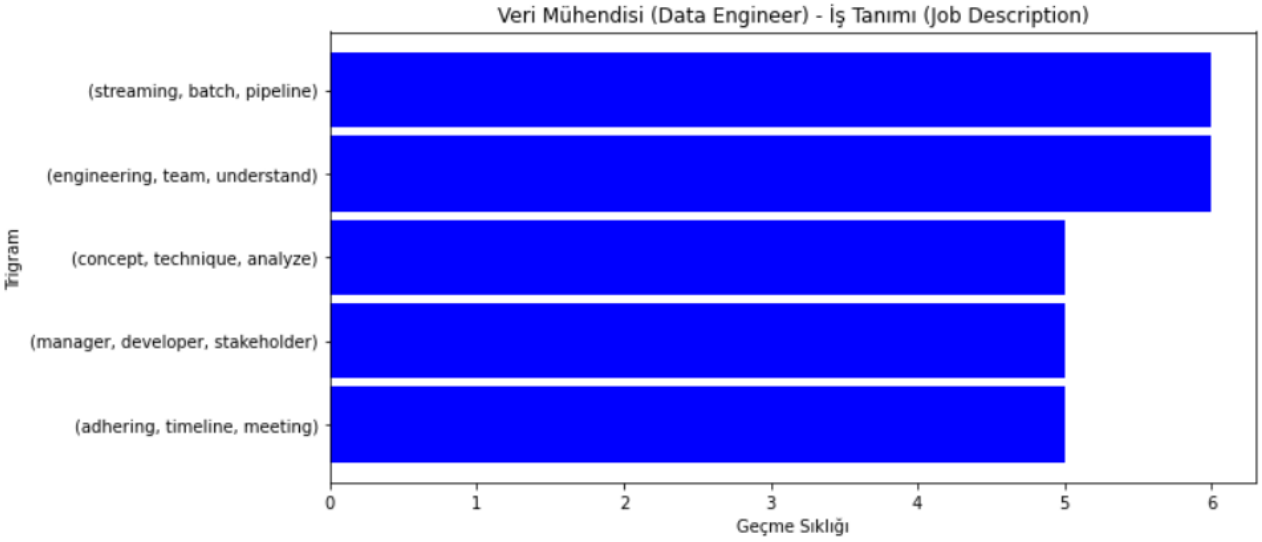
- **İş Tanımı (Job Description)**

Veri mühendisi iş tanımının n-gram analizine bakıldığında makine öğrenmesi, gerçek zamanlı, ürün yöneticisi, kelime gruplarının ilk üç sırada geldiği görülmektedir.




**Şekil 71.** Veri mühendisi-iş tanımı-bigram analizi

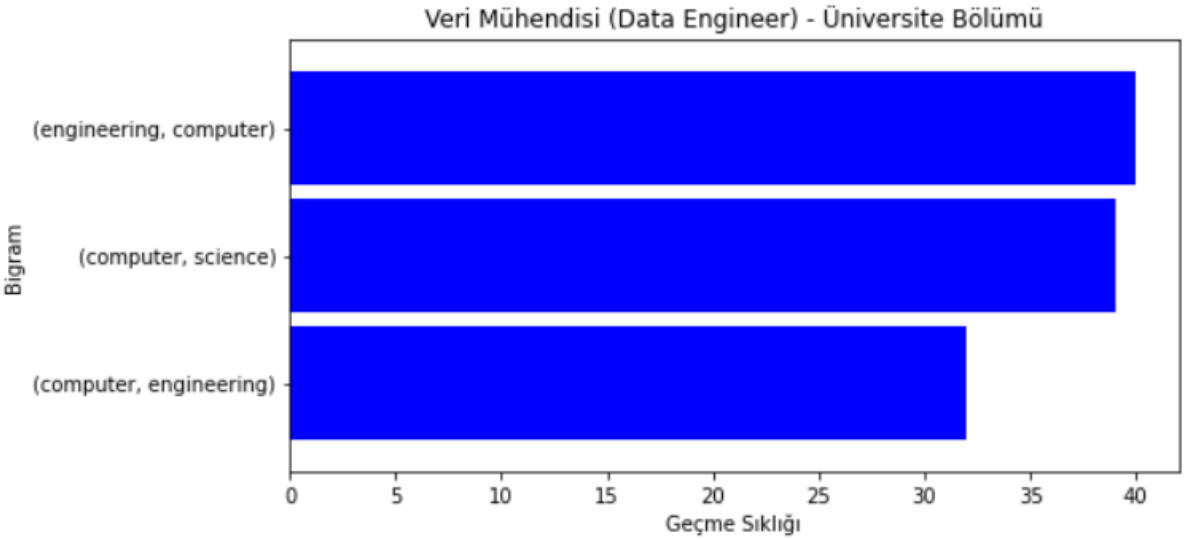
Üçlü kelime gruplarına bakıldığında yeterince anlamlı sonuçlar elde edilemediği görülmektedir.




**Şekil 72.** Veri mühendisi-iş tanımı-trigram analizi

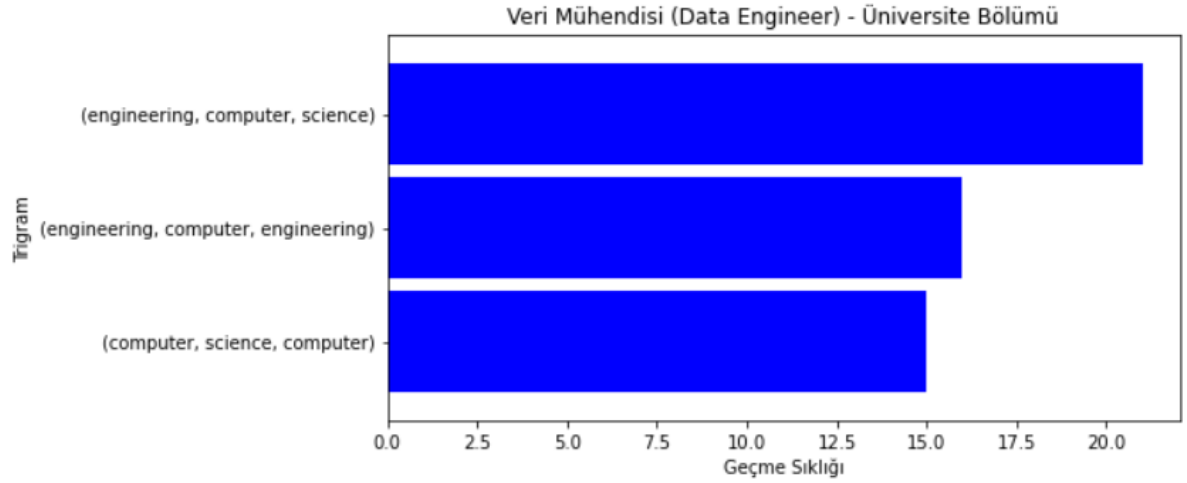
- **Üniversite Bölümü**

Veri mühendisi pozisyonunda istihdam edilecek adayların bilgisayar mühendisliği bölümünden mezun olmasının tercih edildiği görülmektedir.




**Şekil 73.** Veri mühendisi-üniversite bölümü-bigram analizi

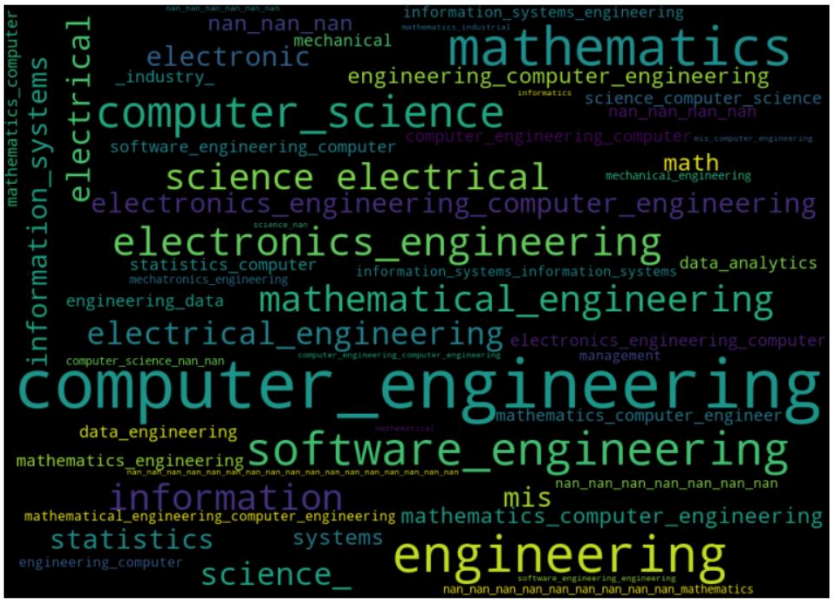
Üçlü kelime grubunda bölüm mezuniyeti ile ilgili, ikili kelime grubundan farklı olarak yeni bir bilgi bulunmamaktadır.




**Şekil 74.** Veri mühendisi-üniversite bölümü-trigram analizi

- **Kelime Bulutu-Üniversite Bölümü**

Bölüm mezuniyet şartları arasında kelime bulutunda baskın şekilde öne çıkan bir bölüm görünmemektedir. Bilgisayar mühendisliği bölümü diğer bölümlere görece ön plana çıkmıştır.



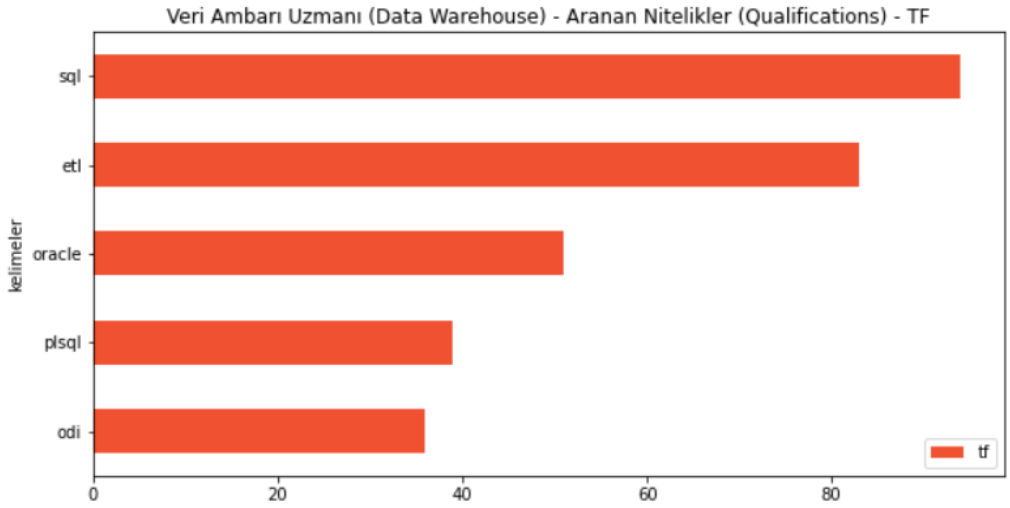

**Şekil 75.** Veri mühendisi-üniversite bölümü-kelime bulutu

## 3.8. VERİ AMBARI UZMANI İŞ İLANLARININ ANALİZİ

Veri Ambarı uzmanı (DWH Specialist) pozisyonu için toplamda 105 ilan elde edilmiştir.

- **Aranan Nitelikler (Qualifications)**

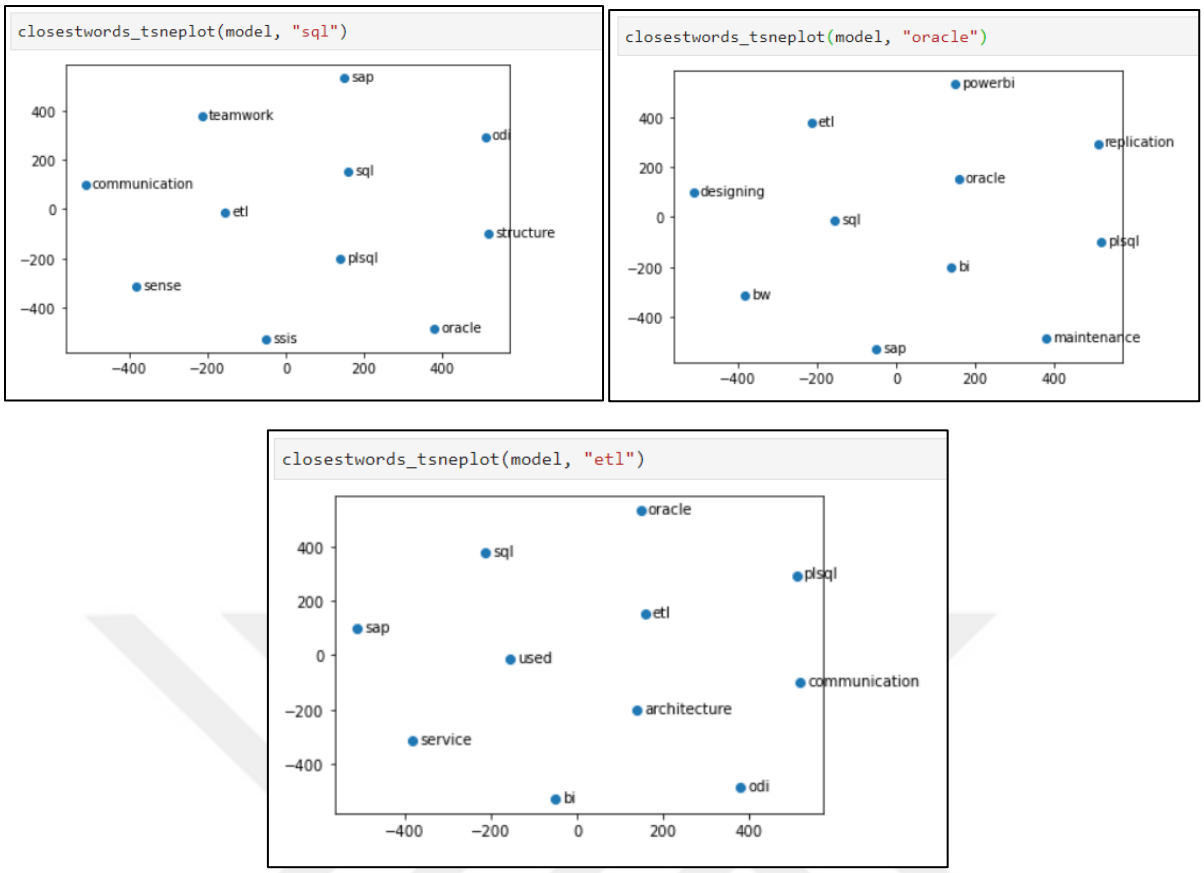
Veri ambarı uzmanı pozisyonu için aranan nitelikler analiz edildiğinde, kelime sıklık grafiğinde SQL, ETL ve Oracle kelimelerinin ilk üç sırada yer aldığı görülmektedir. Sonrasında PLSQL ve Odi kelimelerinin olduğu görülmektedir.




**Şekil 76.** Veri ambarı-aranan nitelikler-terim frekansı

- **En Yakın Kelime Analizi**

Terim frekansı grafiğinde ilk üç sırada yer alan kelimelerin en yakın kelime kümesine bakılmıştır. SQL kelimesinin en yakın kelime kümesine bakıldığında, ETL, takım çalışması, SAP, Odi, PLSQL kelimelerinin olduğu görülmektedir.



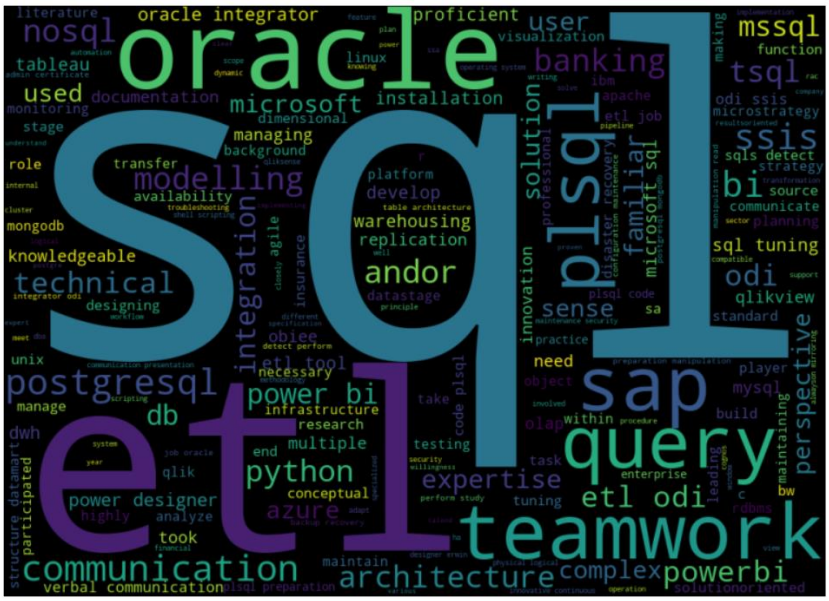

**Şekil 77.** SQL-Oracle-ETL kelimelerinin en yakın kelime analizi

Oracle kelimesinin en yakın kelime kümesine bakıldığında, SQL, ETL, Power BI, PLSQL kelimelerinin olduğu görülmektedir.

ETL kelimesinin en yakın kelime kümesine bakıldığında SQL, Oracle, PLSQL, iletişim, mimari kelimelerinin olduğu görülmektedir.

- **Kelime Bulutu-Aranan Nitelikler**

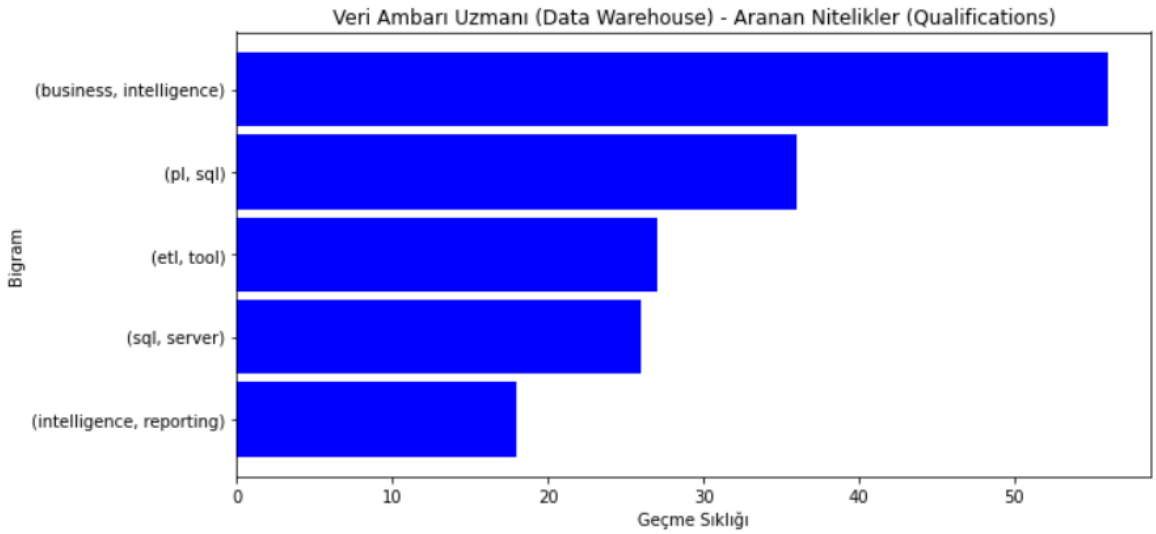
Veri ambarı uzmanı aranan nitelikler veri kümesinin kelime bulutuna bakıldığında SQL kelimesinin ilk sırada yer aldığı görülmektedir. Bununla beraber ETL, Oracle kelimeleri de öne çıkan kelimeler arasındadır.




**Şekil 78.** Veri ambarı uzmanı-aranan nitelikler-kelime bulutu

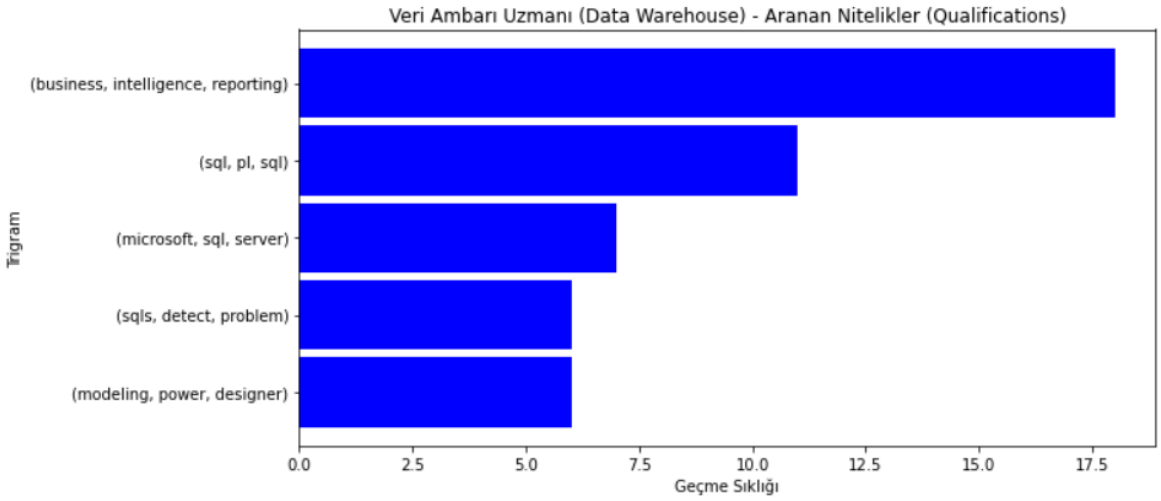
- **N-Gram Analizi-Aranan Nitelikler**

Veri ambarı aranan nitelikler n-gram analizine bakıldığında, iş zekâsı, PLSQL, ETL araçları kelimelerinin ilk üç sırada yer aldığı görülmektedir. Sonrasında ise, SQL Server kelimesinin olduğu görülmektedir.




**Şekil 79.** Veri ambarı uzmanı-aranan nitelikler-bigram analizi

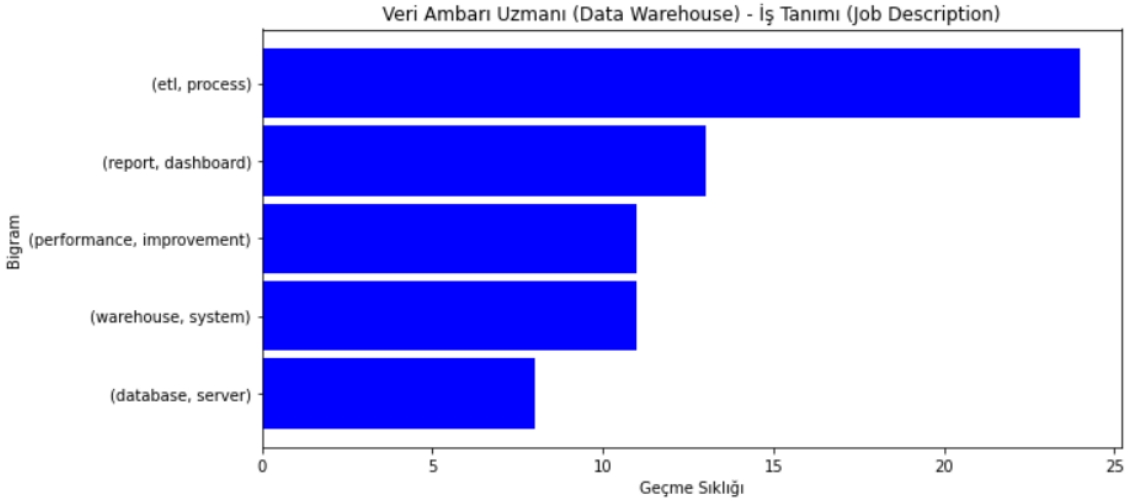
Üçlü kelime gruplarına bakıldığında, İş zekâsı raporlama, SQL PLSQL ve Microsoft SQL Server kelimelerinin ilk üç sırada geldiği görülmektedir. İkili kelime grubunda yer alan zekâ raporlama kelime grubunun iş zekâsı raporlama şeklinde tamamlandığı görülmektedir.




**Şekil 80.** Veri ambarı uzmanı-aranan nitelikler-trigram analizi

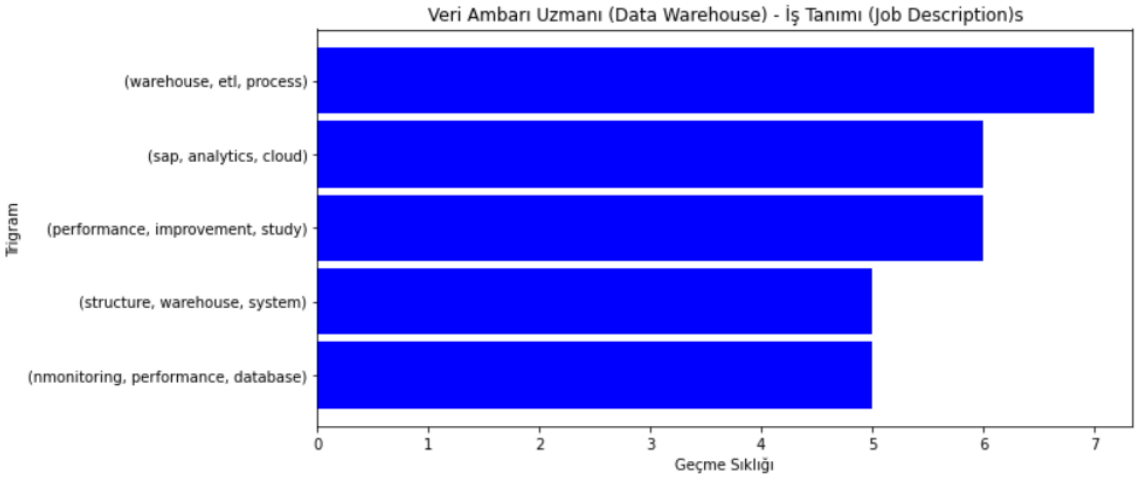
- **İş Tanımı (Job Description)**

Veri ambarı iş tanımı n-gram analizinde ETL süreçleri, gösterge tablosu raporlama, performans gelişim/iyileşme kelimelerinin ilk üç sırada yer aldığı görülmektedir.




**Şekil 81.** Veri ambarı uzmanı-iş tanımı-bigram analizi

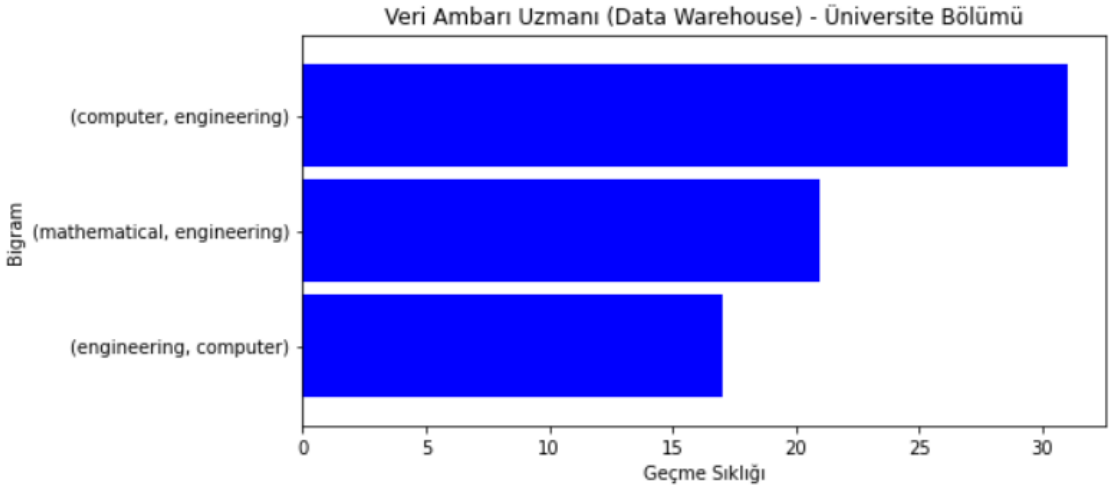
Üçlü kelime grubunda ikili kelime grubuna ek olarak yeni bir bilgi içermediği görülmektedir.




**Şekil 82.** Veri ambarı uzmanı-iş tanımı-trigram analizi

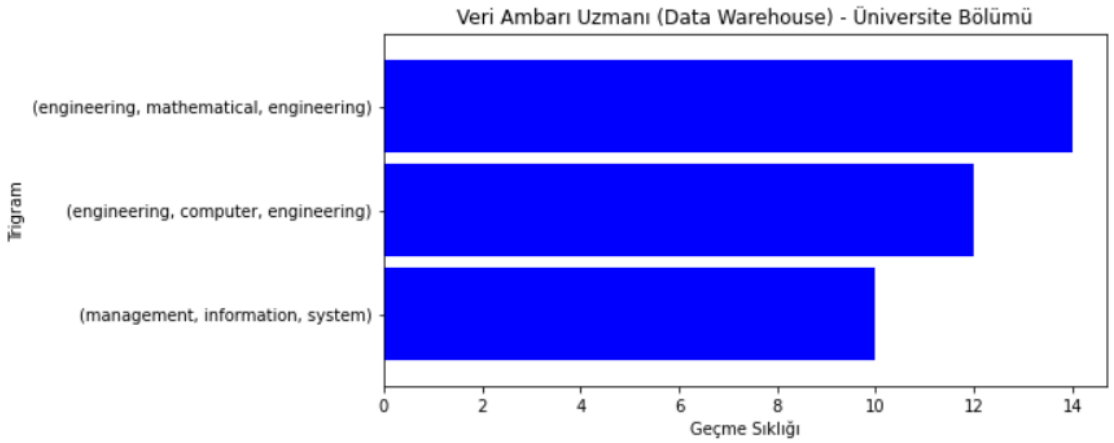
- **Üniversite Bölümü**

Veri ambarı uzmanı için mezuniyet bölüm tercihleri arasında bilgisayar mühendisliği, matematik mühendisliği olduğu görülmektedir.




**Şekil 83.** Veri ambarı uzmanı-üniversite bölümü-bigram analizi

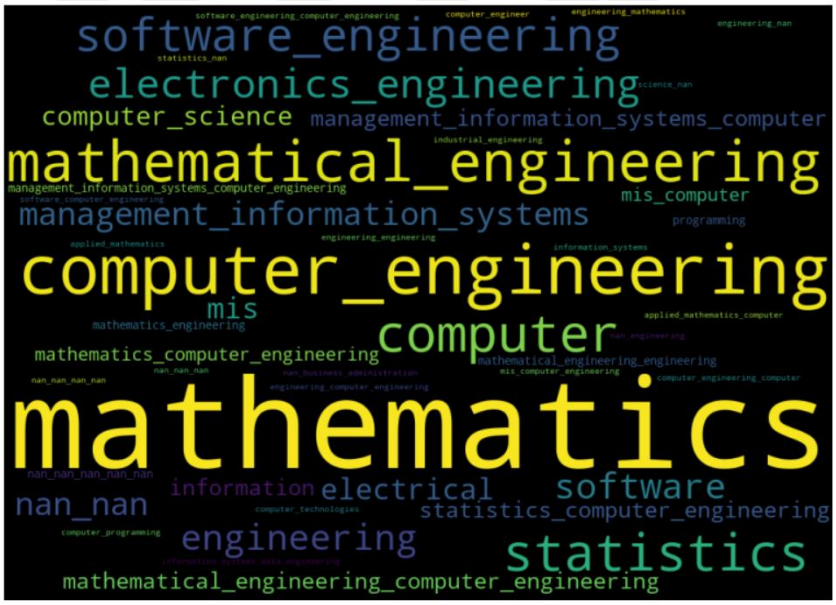
Üçlü kelime grupları arasında, yönetim bilişim sistemleri bölümü yeni bir bilgi olarak görünmektedir.




**Şekil 84.** Veri ambarı uzmanı-üniversite bölümü-trigram analizi

- **Kelime Bulutu-Üniversite Bölümü**

Veri ambarı uzmanı için mezuniyet tercihleri arasında matematik bölümünün ön plana çıktığı görülmektedir. Sonrasında bilgisayar ve matematik mühendisliği bölümlerinin göze çarptığı görülmektedir.



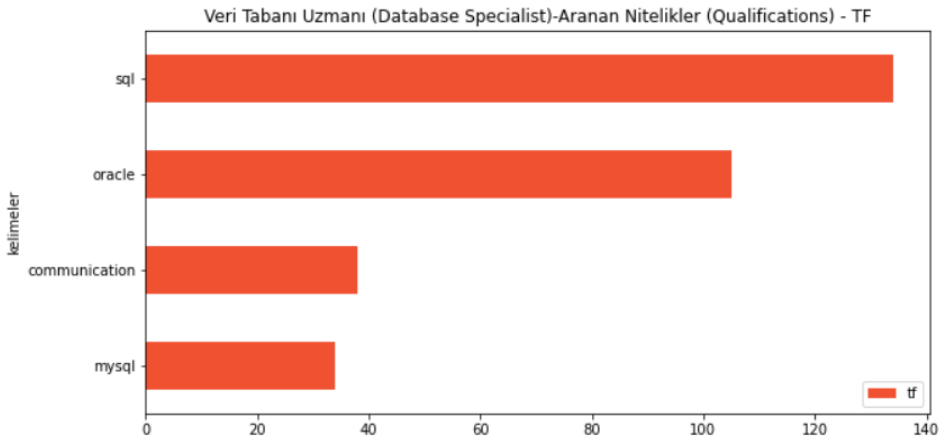

**Şekil 85.** Veri ambarı uzmanı-üniversite bölümü-kelime bulutu

## 3.9. VERİ TABANI UZMANI İŞ İLANLARININ ANALİZİ

Veri tabanı uzmanı (Database Specialist) pozisyonunda toplamda 110 ilan elde edilmiştir.

- **Aranan Nitelikler (Qualifications)**

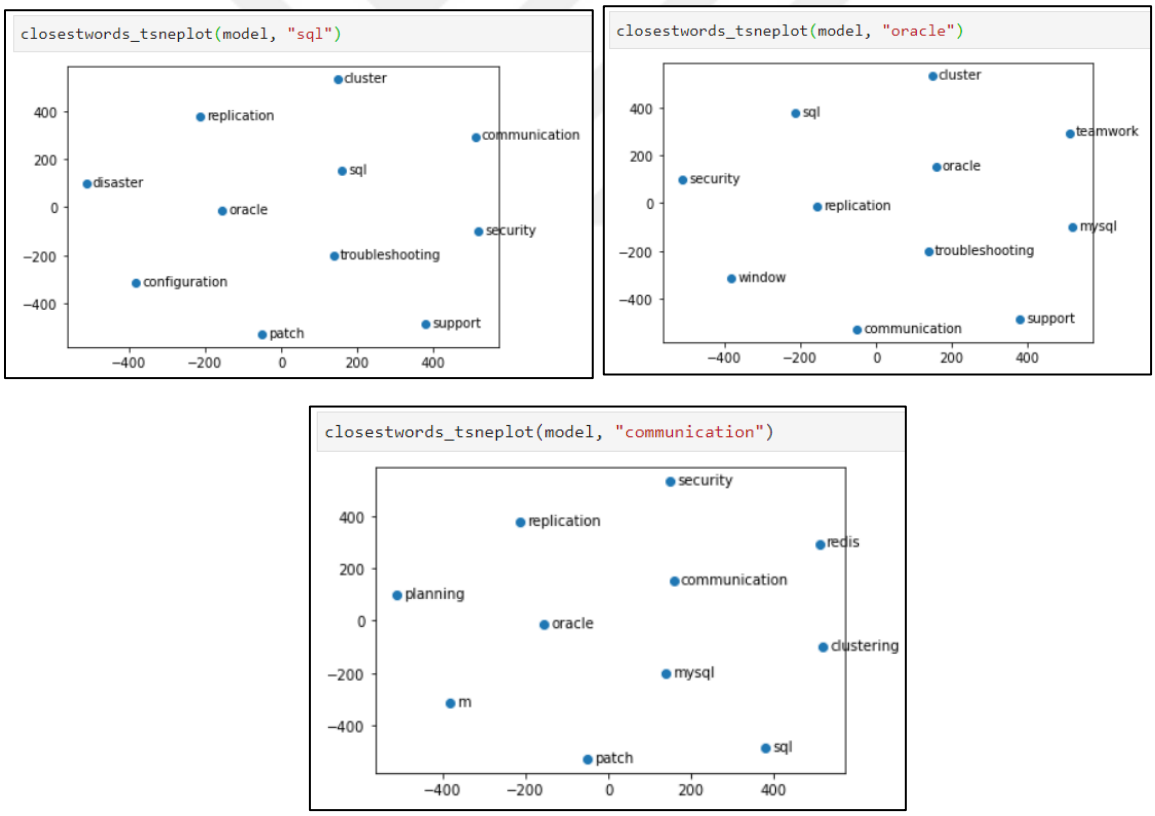
Veri tabanı uzmanı için aranan nitelikler terim frekansı analizine bakıldığında SQL, Oracle ve iletişim kelimelerinin ilk üç sırada geldiği görülmektedir. Sonrasında ise MySQL kelimesi gelmektedir.




**Şekil 86.** Veri tabanı uzmanı-aranan nitelikler-terim sıklığı

- **En Yakın Kelime Analizi**

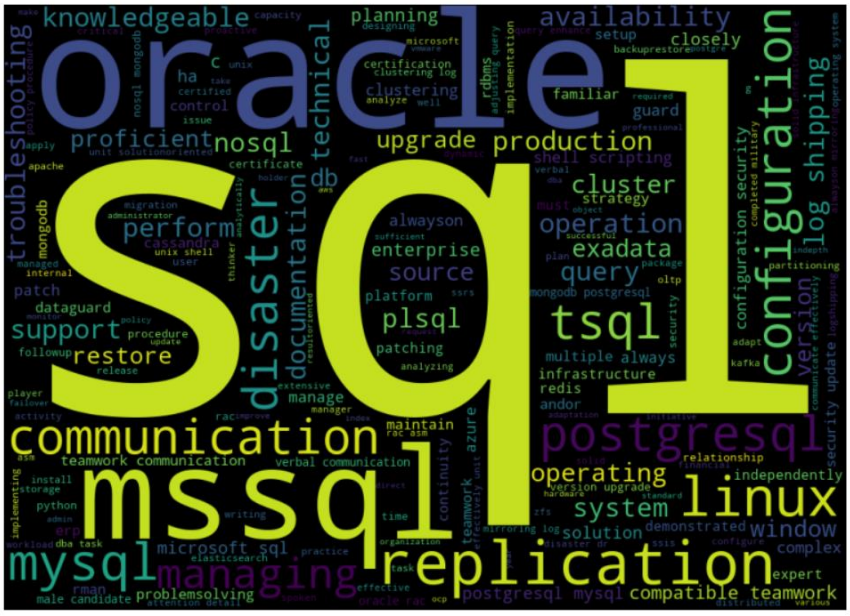
Terim frekansı grafiğinde ilk sırada yer alan SQL kelimesinin en yakın kelime kümesine bakıldığında, Oracle, iletişim, sorun giderme (troubleshooting), güvenlik kelimelerinin olduğu görülmektedir.




**Şekil 87.** SQL-Oracle-İletişim kelimelerinin en yakın kelime analizi

Oracle kelimesinin en yakın kelime kümesine bakıldığında SQL, iletişim, takım çalışması, sorun giderme kelimelerinin olduğu görülmektedir.

İletişim kelimesinin en yakın kelime kümesine bakıldığında, Oracle, güvenlik, Redis, sorun giderme, MySQL kelimelerinin olduğu görülmektedir.



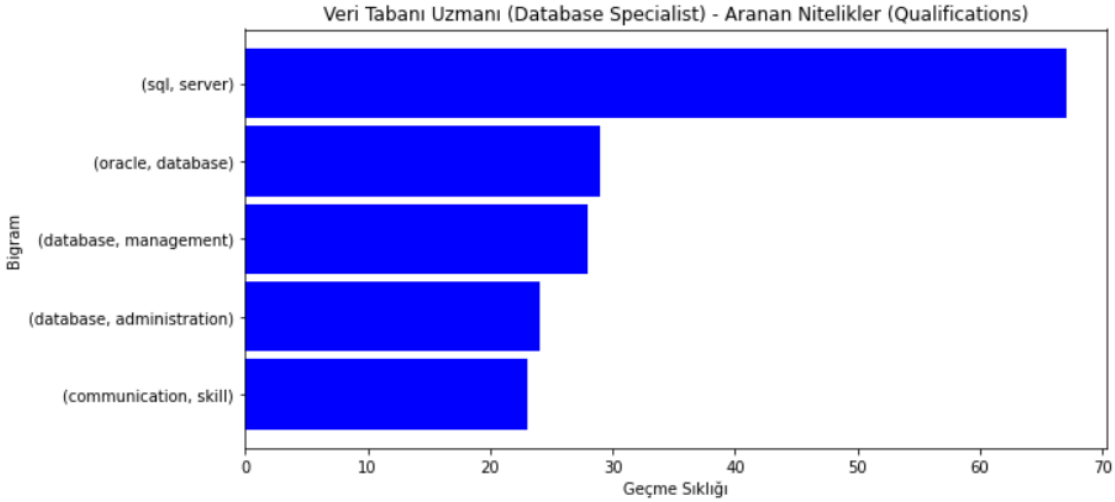

**Şekil 88.** Veri tabanı uzmanı-aranan nitelikler-kelime bulutu

- **Kelime Bulutu-Aranan Nitelikler**

Kelime bulutunda aranan nitelikler için SQL yetkinliğinin ön plana çıktığı görülmektedir. Sonrasında Oracle, MsSQL kelimelerinin ön plana çıktığı söylenebilir.

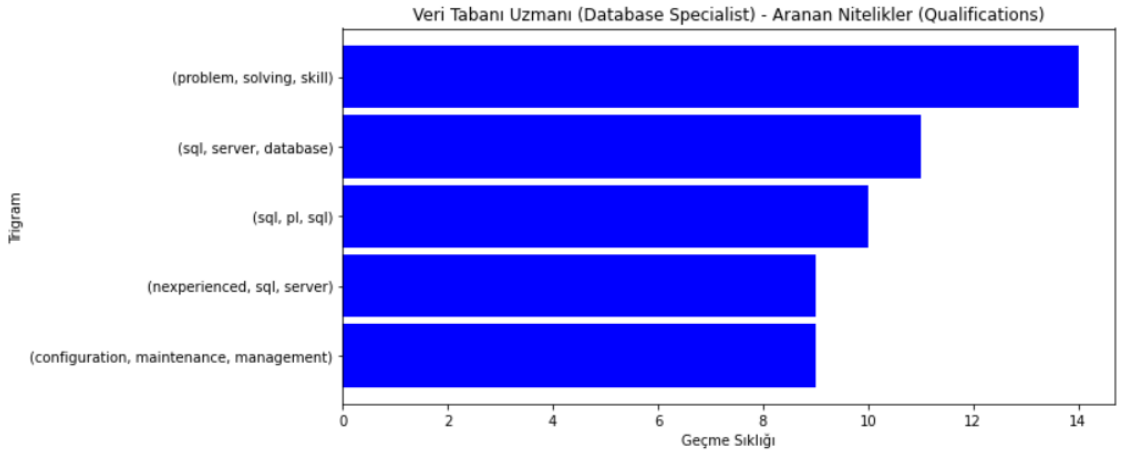
- **N-Gram Analizi-Aranan Nitelikler**

Veri tabanı pozisyonu aranan nitelikler N-gram analizine bakıldığında, SQL Server kelimesinin ilk sırada diğer kelimelere oranla yüksek bir farkla ortaya çıktığı görülmektedir. Sonrasında Oracle veri tabanı, veri tabanı yönetimi ve iletişim becerileri kelime gruplarının yakın sıklıklarla sıralandığı görülmektedir.




**Şekil 89.** Veri tabanı uzmanı-aranan nitelikler-bigram analizi

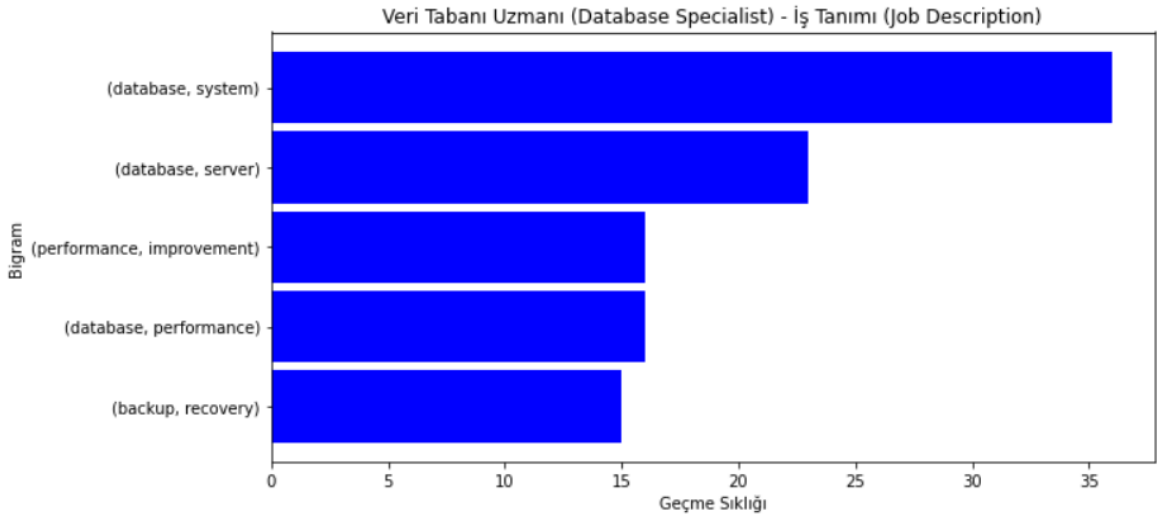
Üçlü kelime grubunda problem çözme becerileri ikili kelime grubundan farklı olarak ilk sırada yer almaktadır.




**Şekil 90.** Veri tabanı uzmanı-aranan nitelikler-bigram analizi

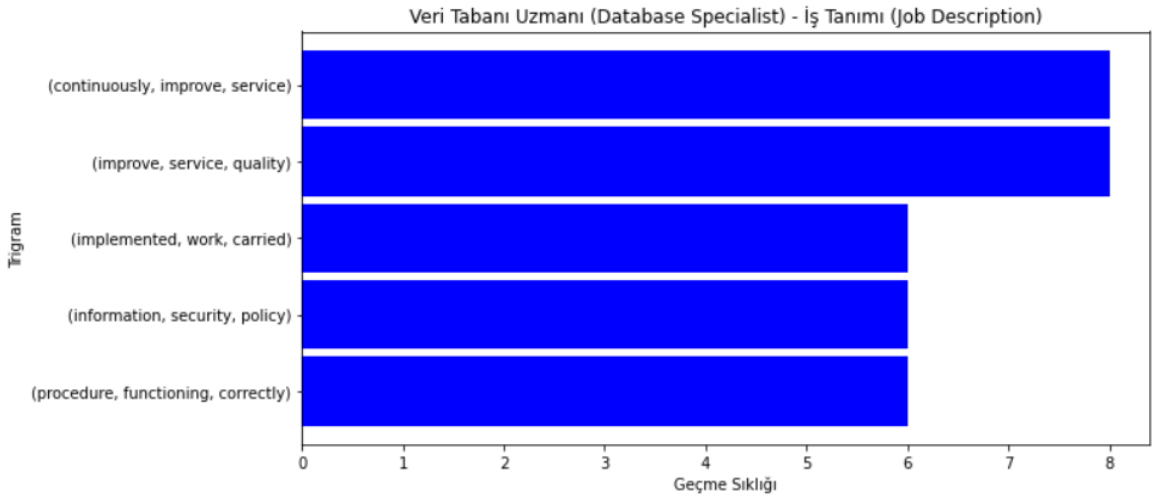
- **İş Tanımı (Job Description)**

Veri tabanı iş tanımı N-gram analizinde veri tabanı sistemi, veri tabanı hizmeti, performans iyileştirme kelimeleri ilk üç sırada yer almaktadır.




**Şekil 91.** Veri tabanı uzmanı-iş tanımı-bigram analizi

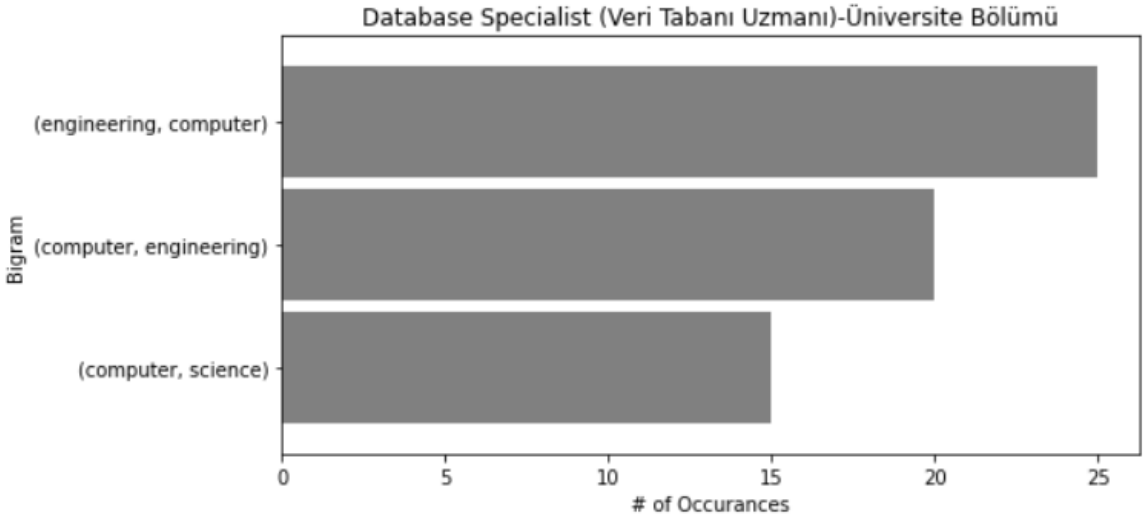
Üçlü kelime gruplarına bakıldığında, servis kalitesinin iyileştirilmesi, bilgi güvenliği politikaları kelime gruplarının yer aldığı görülmektedir.




**Şekil 92.** Veri tabanı uzmanı-iş tanımı-trigram analizi

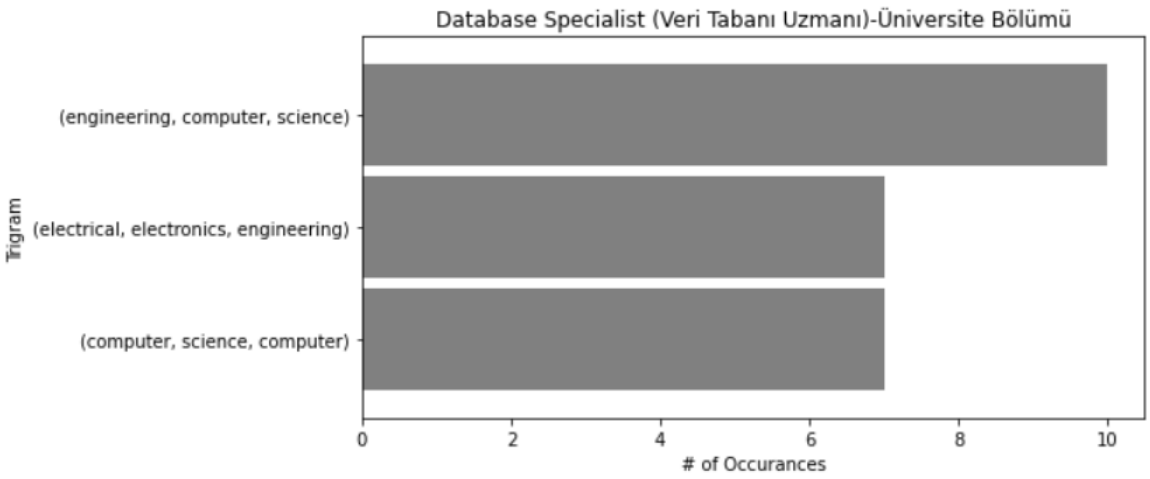
- **Üniversite Bölümü**

Veri tabanı uzmanı pozisyonu için aranan mezuniyet bölümleri arasında bilgisayar mühendisliği ve bilgisayar bilimlerinin yer aldığı görülmektedir.




**Şekil 93.** Veri tabanı uzmanı-üniversite bölümü-bigram analizi

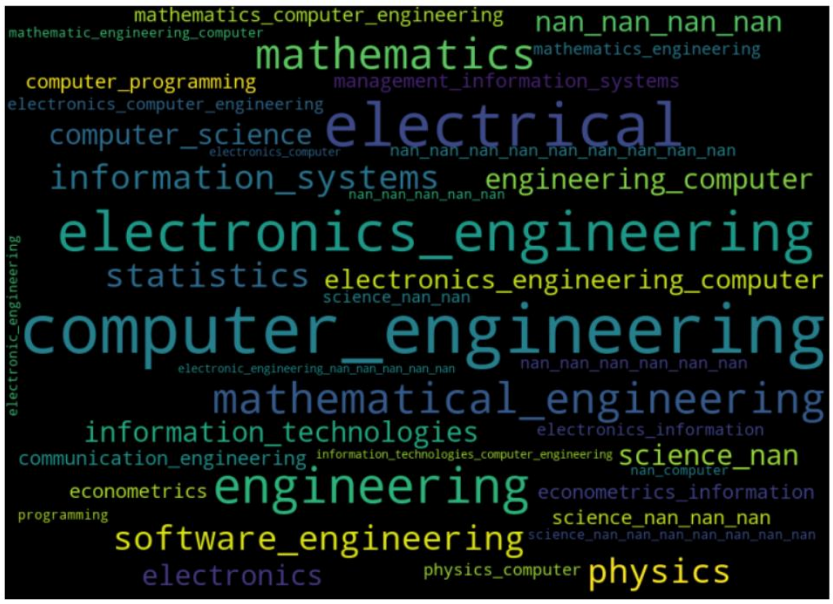
Üçlü kelime gruplarında bilgisayar mühendisliğine ek olarak, elektrik elektronik mühendisliğinin yer aldığı görülmektedir.




**Şekil 94.** Veri tabanı uzmanı-üniversite bölümü-trigram analizi

- **Kelime Bulutu-Üniversite Bölümü**

Veri tabanı uzmanı pozisyonu için mezuniyet bölümleri kelime bulutuna bakıldığında anlamlı şekilde ön plana çıkan kelime veya kelime grubunun olmadığı buna rağmen bilgisayar mühendisliği bölümünün öne çıkan bölüm arasında yer aldığı söylenebilir.



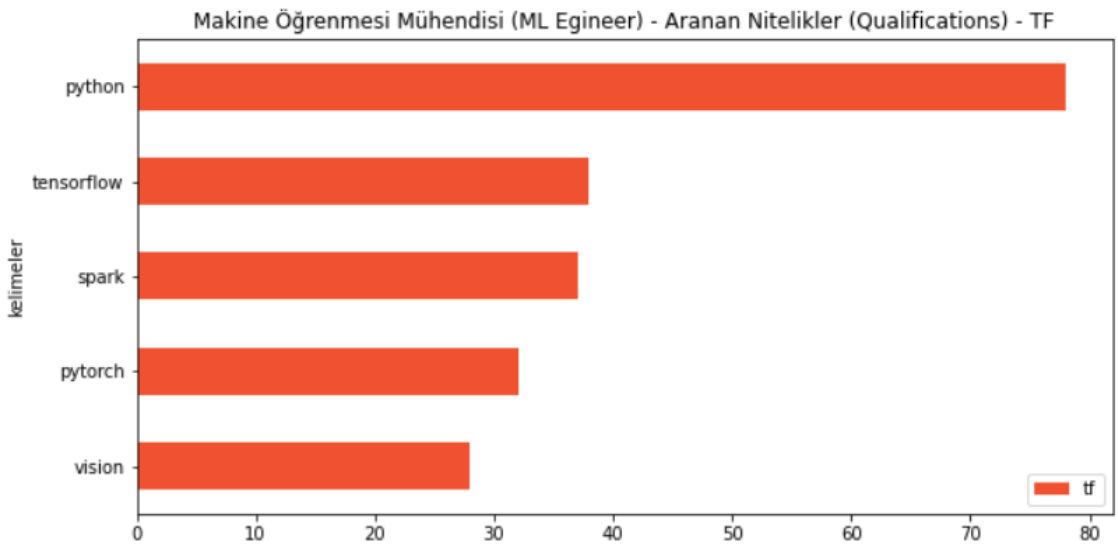

**Şekil 95.** Veri tabanı uzmanı-üniversite bölümü-kelime bulutu

## 3.10. MAKİNE ÖĞRENMESİ MÜHENDİSİ İŞ İLANLARININ ANALİZİ

Makine Öğrenmesi Mühendisi (ML Engineer) pozisyonu için 80 ilan elde edilmiştir.

- **Aranan Nitelikler (Qualifications)**

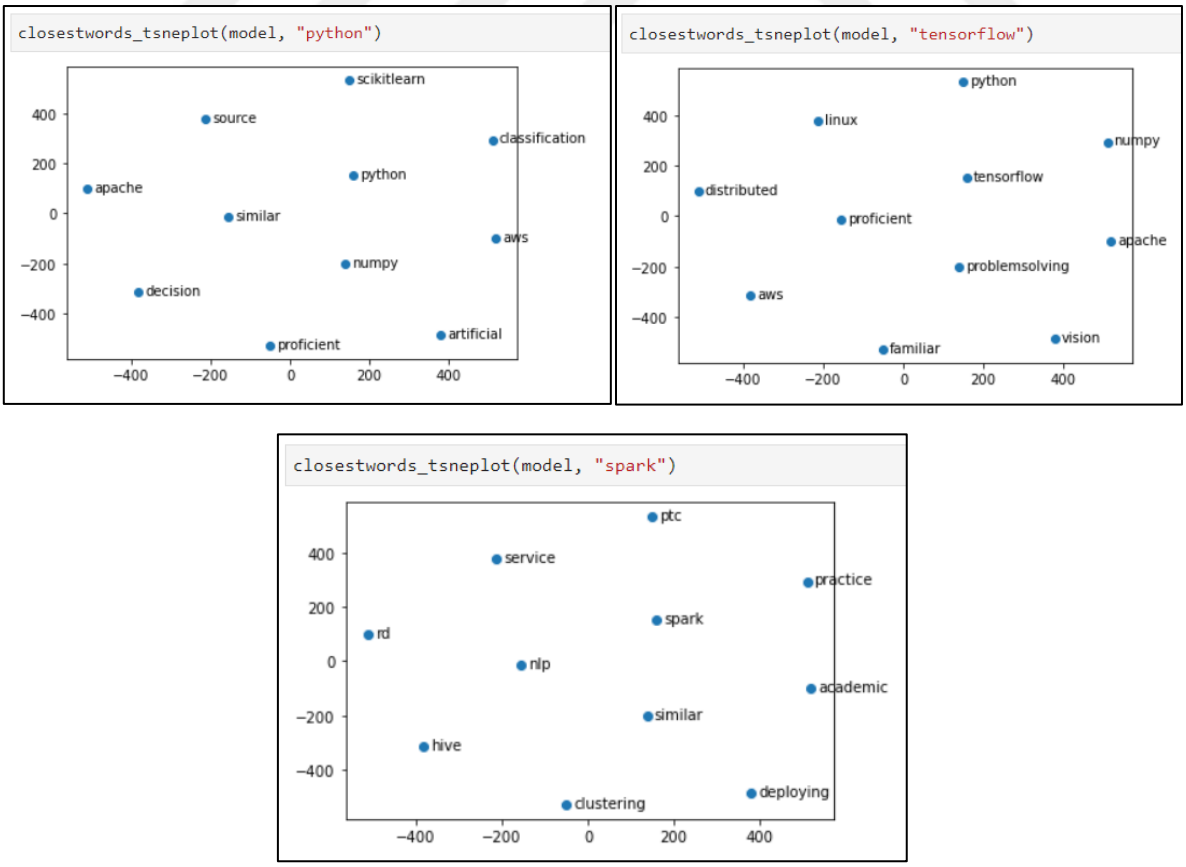
Makine öğrenmesi mühendisi aranan nitelikler terim frekansı analizine bakıldığında, Python, Tensorflow, Spark kelimelerinin ilk üç sırada yer aldığı görülmektedir. Sonrasında ise Pytorch ve görüş kelimelerinin geldiği görülmektedir.




**Şekil 96.** Makine öğrenmesi mühendisi-aranan nitelikler-terim frekansı

- **En Yakın Kelimeler Analizi**

Python kelimesinin en yakın kelime kümesine bakıldığında, benzer, kaynak, scikitlearn, sınıflandırma, aws, Numpy kelimelerinin yer aldığı görülmektedir. Makine öğrenmesi alt dallarından birisi sınıflandırmadır. Numpy ve scikitlearn ise Python programlama analizlerinde kullanılan kütüphanelerdir.



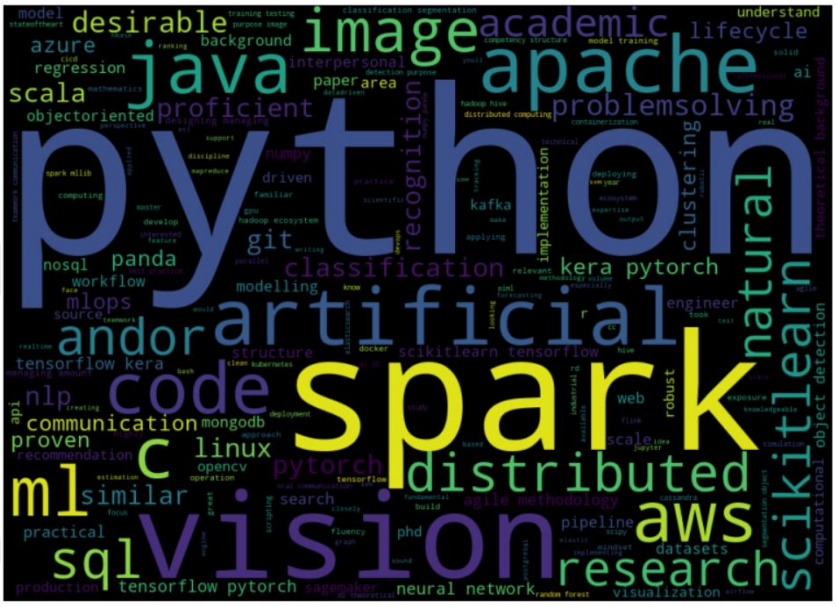

**Şekil 97.** Python-Tensorflow-Spark kelimelerinin en yakın kelime analizi

Tensorflow kelimesinin en yakın kelime kümesine bakıldığında Linux, Python, Numpy, apache, problem çözme kelimelerinin yer aldığı görülmektedir.

Spark kelimesinin en yakın kelime kümesine bakıldığında, NLP, hizmet, akademik, benzer kelimelerinin yer aldığı görülmektedir.

- **Kelime Bulutu-Aranan Nitelikler**

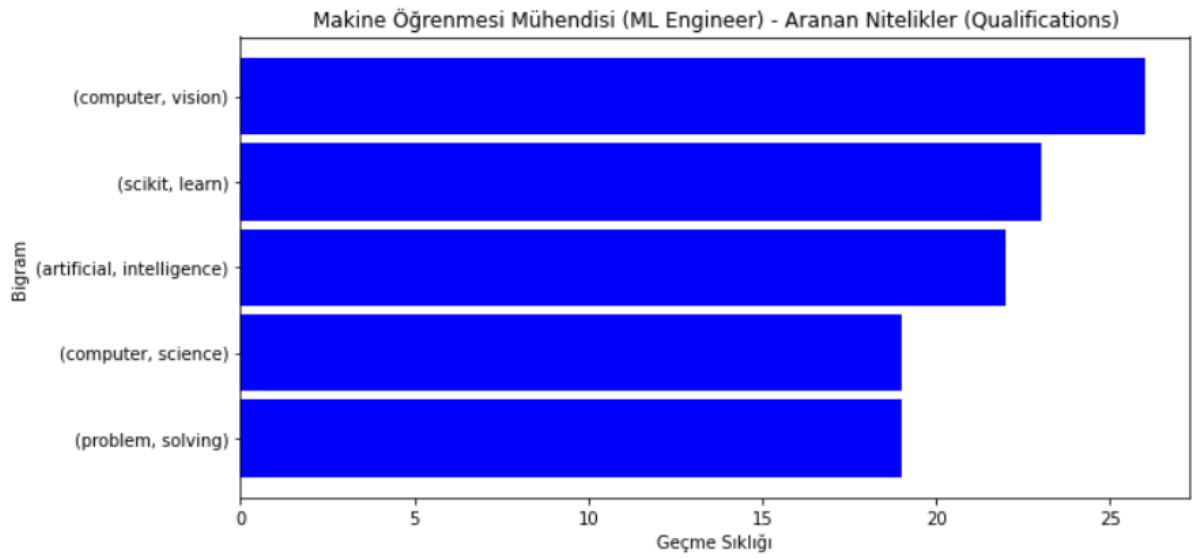
Makine öğrenmesi aranan nitelikler kelime bulutu incelendiğinde, Python kelimesinin ön plana çıktığı, Spark kelimesinin Python kelimesini takip ettiği görülmektedir.




**Şekil 98.** Makine öğrenmesi-aranan nitelikler-kelime bulutu

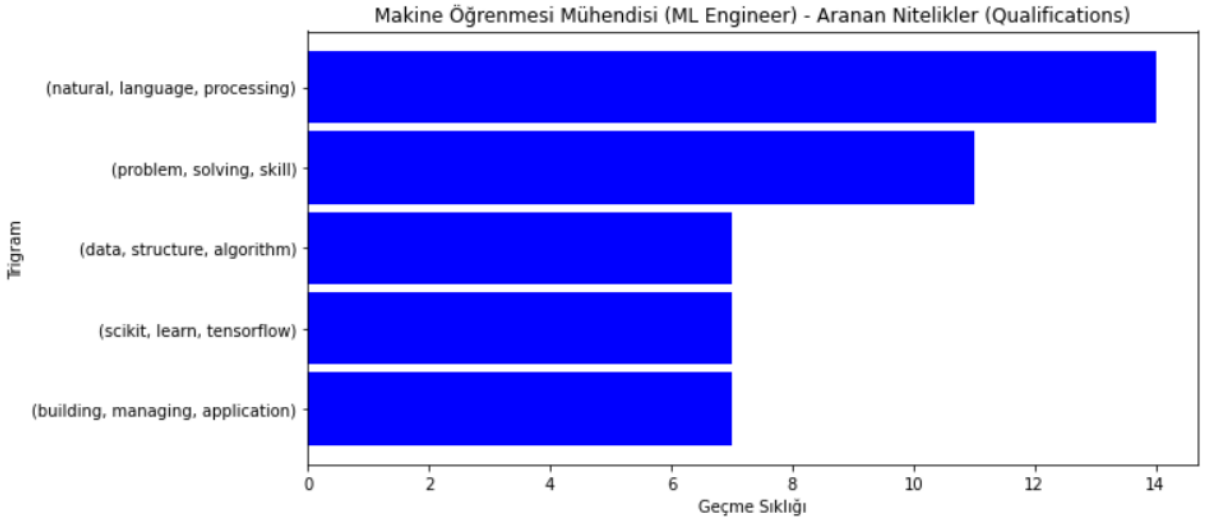
- **N-Gram Analizi -Aranan Nitelikler**

Makine öğrenmesi aranan nitelikler N-gram analiz sonuçlarına bakıldığında bilgisayar görüşü, scikit learn, yapay zekâ kelimelerinin ilk üç sırada yer aldığı görülmektedir. Sonrasında ise, bilgisayar bilimi ve problem çözme kelimelerinin sıralandığı görülmektedir.




**Şekil 99.** Makine öğrenmesi mühendisi-aranan nitelikler-bigram analizi

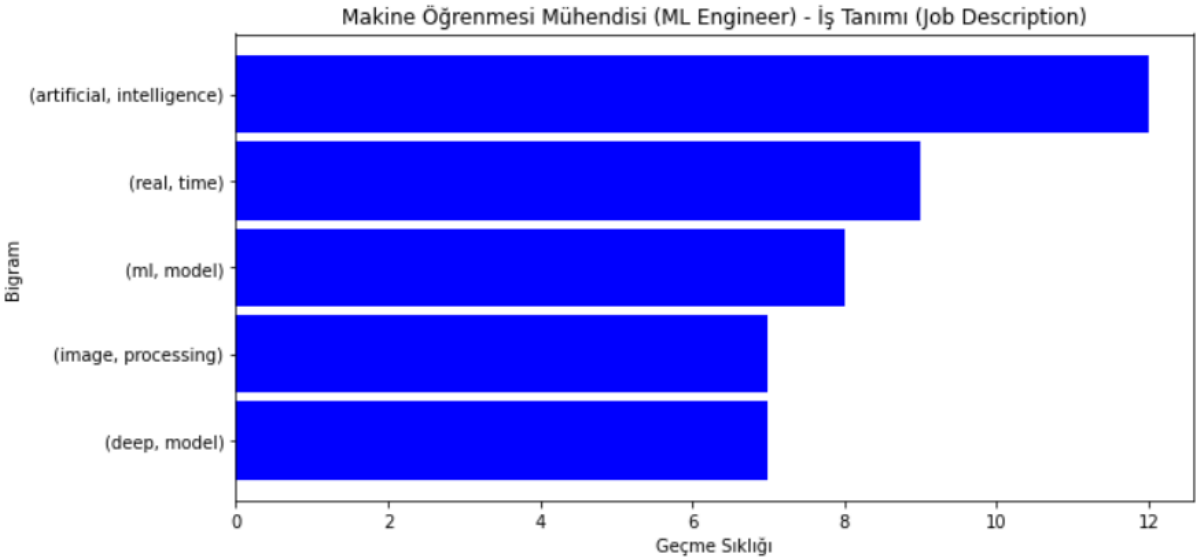
Makine öğrenmesi mühendisi pozisyonu için aranan nitelikler N-gram analizine bakıldığında ilk üç sırada doğal dil işleme, problem çözme becerileri ve veri algoritma yapıları kelimelerinin yer aldığı görülmektedir.




**Şekil 100.** Makine öğrenmesi mühendisi-aranan nitelikler-trigram analizi

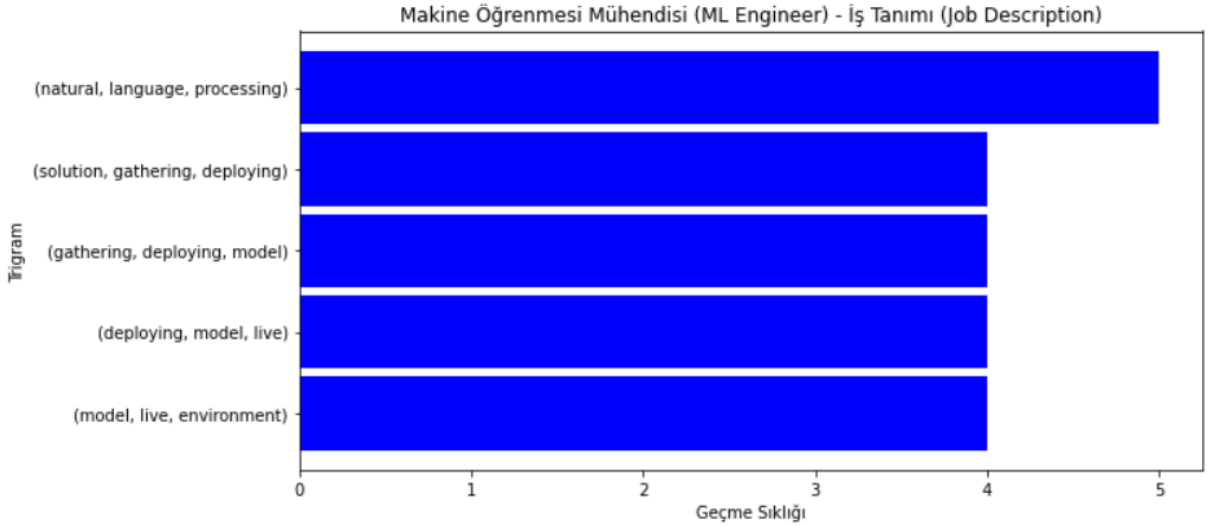
- **İş Tanımı (Job Description)**

Makine öğrenmesi iş tanımına bakıldığında ise, yapay zekâ, gerçek zamanlı, ML model kelimelerinin ilk üç sırada geldiği görülmektedir. Sonrasında resim işleme ve derin model kelimeleri gelmektedir.




**Şekil 101.** Makine öğrenmesi mühendisi-iş tanımı-bigram analizi

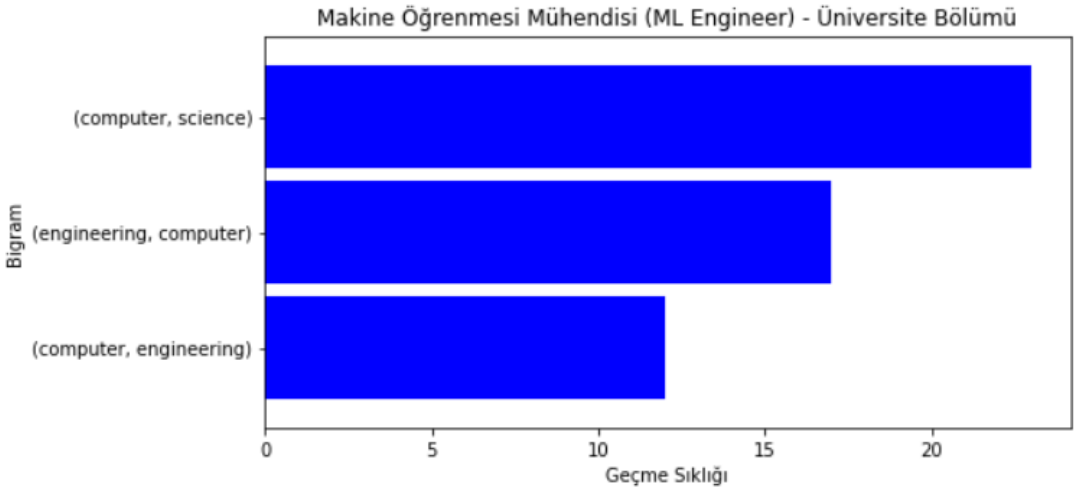
Üçlü kelime gruplarına bakıldığında, doğal dil işleme kelime grubunun ilk sırada yer aldığı görülmektedir. Diğer kelime grupları aynı sıklıktadır ve anlamsal olarak dikkat çekici bir sonuç sunmamaktadır.



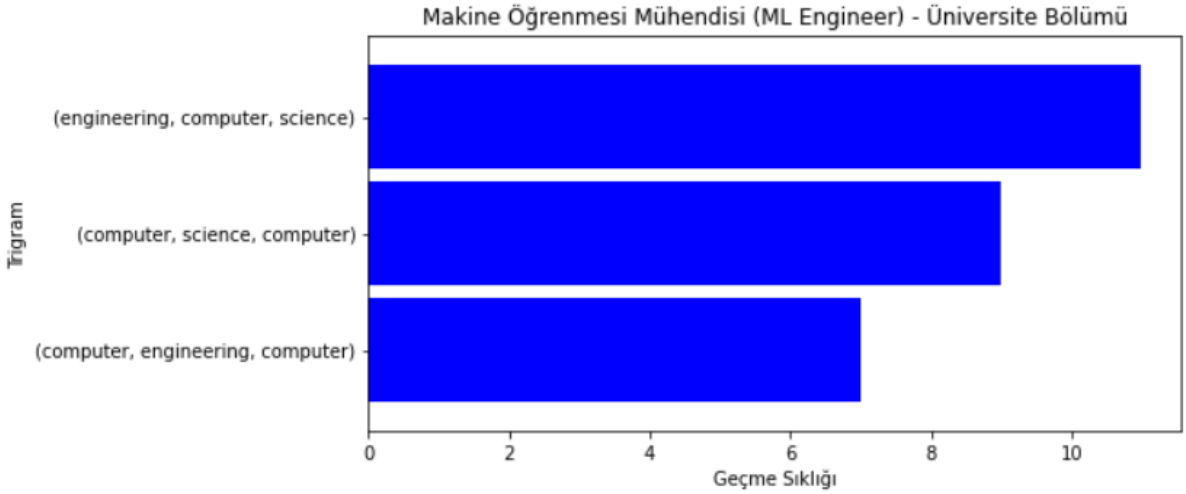

**Şekil 102.** Makine öğrenmesi mühendisi-iş tanımı-trigram analizi

- **N-gram Analizi-Üniversite Bölümü**

Makine Öğrenmesi Mühendisi için mezuniyet bölümlerine bakıldığında, bilgisayar mühendisliği bölümünün ilk sırada geldiği görülmektedir.



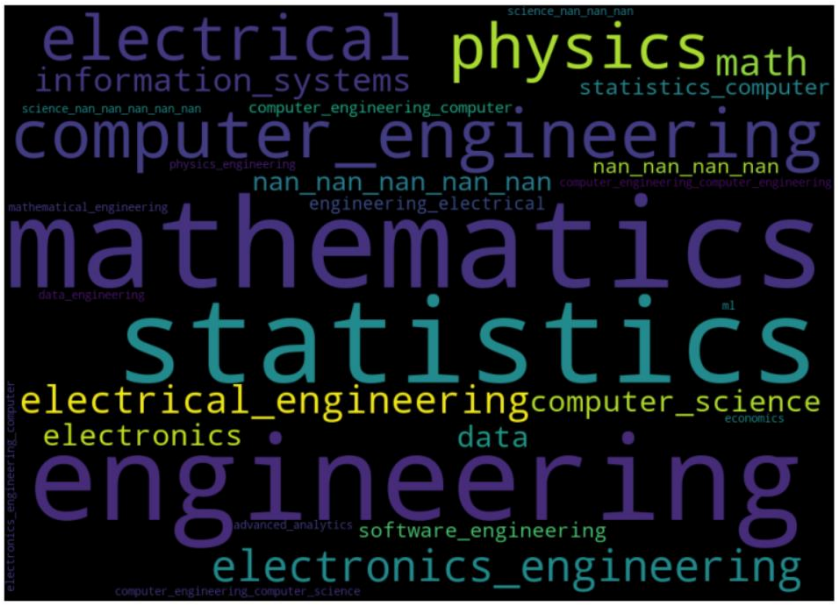

**Şekil 103.** Makine öğrenmesi mühendisi-üniversite bölümü-bigram analizi




**Şekil 104.** Makine öğrenmesi mühendisi-üniversite bölümü-trigram analizi

- **Kelime Bulutu-Üniversite Bölümü**

Kelime bulutunda mezuniyet bölümleri matematik, istatistik, mühendislik aynı ölçüde ön plana çıkmaktadır.




**Şekil 105.** Makine öğrenmesi mühendisi-üniversite bölümü-kelime bulutu

### 3.11. ANLAMLI FARKLILIKLARIN SINANMASI

Bu tez çalışması kapsamında sekiz pozisyon bazında 1,465 ilan elde edilmiştir. Her pozisyon bazında ilanlarda en çok geçen kelimelerin kelime sıklık grafikleri çıkarılmıştır. Kelime sıklık grafikleri (terim frekansı) baz alınarak bütün pozisyonlarda en çok tekrar eden ilk üç yetkinlik ortaya konmuştur. Her pozisyonda aranan ilk üç yetkinliğin diğer pozisyonlardaki tercih edilme/aranma durumlarına bakılarak, pozisyonlar arası yetkinlik beklentilerinin nasıl farklılaştığı ortaya konmaya çalışılmıştır. Tüm pozisyonlarda aranan ilk üç yetkinliğe bakıldığında 10 adet farklı yetkinliğin olduğu görülmüştür.

Tablo 1'de yetkinlikler tablonun solundaki satırlara yerleştirilmiş, sütunlarda ise pozisyonlara yer verilmiştir. Tablo 1'de söz konusu yetkinliklerin her pozisyon bazındaki toplam ilan sayısında kaç kere geçtiğini göstermektedir. Kelime sıklık grafiğindeki yaklaşımdan farklı olarak, bu tablodaki değerler, söz konusu yetkinlik her bir ilanda birden çok kez geçse bile 1 olarak, eğer ilgili yetkinlik geçmiyorsa 0 olarak sayılmıştır. Örneğin, Veri analizcisi pozisyonundaki 264 iş ilanının 92 tanesinde Python yetkinliği geçmektedir. Kelime sıklık grafiğinde tekrar eden kelimeler de sayıldığından bu sayı 102 adet idi. Özetle; her yetkinliğin iş ilanlarında geçip geçmeme durumuna bakılarak yukarıda yer alan tablo oluşturulmuştur. Kelime sıklık grafiğinde, pozisyon bazında ilk üç sırada yer alan yetkinliklerin, Tablo 1'de de aynı şekilde ilk üç sırada yer aldığı görülmüştür. Tablo 1'den yola çıkarak, adetler ve oranlar üzerinden her pozisyonu tanımlayan ilk üç yetkinlik sıralanabilmektedir. Ancak, sayısal gösterimlere ek olarak, anlamlılık testleri ile sonuçlar daha güvenli bir zeminde sunulmuştur.

Anlamlılık testlerinin yapılabilmesi için, "ilan sayısı" x "yetkinlikler" matrisi oluşturulmuştur. Bu matriste ilan satırında ilgili yetkinlik geçiyorsa 1 geçmiyorsa 0 olarak kodlanmıştır. Tüm ilanlar bazında 1465 x 10'luk bir matris oluşturulmuşken, pozisyon bazında ise, ilan sayısı x 10 şeklinde matrisler oluşturulmuştur. Matris oluşturma işlemleri Python yardımı ile gerçekleştirilmiştir.

Aşağıda anlamlılık test sonuçları, tüm ilanlar ve sekiz farklı pozisyon bazında gösterilmiştir.

### 3.11.1. Tüm İlanlar Bazında Anlamlı Farklılıkların Sınamaları

Tüm ilanların yer aldığı matris 1465 satır ve 10 sütundan oluşmaktadır.

1,465 ilanın 840 tanesinde SQL yetkinliğinin geçtiği görülmektedir. %57 oranıyla SQL yetkinliği birinci sırada yer almaktadır. SQL yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimi alanındaki pozisyonlarda en çok aranan birinci yetkinliğin, SQL yetkinliği olduğunu ortaya koymaktadır. Diğer yandan, veri bilimi alanında çalışmak isteyen adaylar SQL yetkinliğine odaklanmalıdırlar.

1,465 ilanın 621 tanesinde iletişim yetkinliğinin geçtiği görülmektedir. %42 oranıyla iletişim yetkinliği 10 yetkinlik arasında ikinci sırada yer almaktadır. İletişim yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimi alanındaki pozisyonlarda en çok aranan ikinci yetkinliğin, iletişim yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar iletişim yetkinliğine odaklanması önem arz etmektedir. Veri bilimi alanında ikinci sırada aranan yetkinliğin sosyal bir yetkinlik olması hem şirketler hem de bu alanda çalışmak isteyen adaylar için dikkat çekici bir durumdur.

**Tablo 1.** Pozisyon bazlı ilan sayısı & ilan bazlı yetkinlik adeti

| | Pozisyon Adı | Veri Analizcisi | Veri Bilimcisi | İş Zekâsı Uzmanı | İş Analisti | Veri Mühendisi | Veri Ambarı Uzmanı | Veri Tabanı Uzmanı | Makine Öğrenmesi Mühendisi | Toplam |
|---|---|---|---|---|---|---|---|---|---|---|
| | | İlan Bazlı Yetkinlik Sıklığı (Adet) | | | | | | | | |
| Yetkinlikler | Python | **92** | **197** | 30 | 7 | **98** | 9 | 7 | **69** | **509** |
| | SQL | **161** | **144** | **179** | **93** | **105** | **68** | **71** | 19 | **840** |
| | R | 52 | **109** | 19 | 2 | 8 | 5 | 3 | 7 | 205 |
| | İletişim (Communication) | **114** | **113** | **105** | **160** | 42 | 29 | **35** | 23 | **621** |
| | Spark | 10 | 78 | 3 | 0 | **86** | 3 | 1 | **27** | 208 |
| | Oracle | 15 | 1 | 50 | **14** | 29 | **36** | **52** | 2 | 199 |
| | ETL | 20 | 8 | **122** | 3 | 56 | **56** | 1 | 3 | 269 |
| | SAP | 27 | 1 | 68 | **24** | 5 | 14 | 1 | 0 | 140 |
| | Çevik (Agile) | 19 | 19 | 11 | **60** | 32 | 5 | 2 | 14 | 162 |
| | Tensorflow | 1 | 46 | 0 | 0 | 3 | 0 | 0 | **38** | 88 |
| | Pozisyon Bazlı İlan Sayısı | **264** | **232** | **277** | **253** | **144** | **105** | **110** | **80** | |
| | Toplam İlan Sayısı | **1,465** | | | | | | | | |

**Tablo 2**. Pozisyon bazlı ilan oranları & ilan bazlı yetkinlik oranları

| | Pozisyon Adı | Veri Analizcisi | Veri Bilimcisi | İş Zekâsı Uzmanı | İş Analisti | Veri Mühendisi | Veri Ambarı Uzmanı | Veri Tabanı Uzmanı | Makine Öğrenmesi Mühendisi | Toplam |
|---|---|---|---|---|---|---|---|---|---|---|
| | | İlan Bazlı Yetkinlik Oranı (%) | | | | | | | | |
| Yetkinlikler | Python | **35%** | **85%** | 11% | 3% | **68%** | 9% | 6% | **86%** | **35%** |
| | SQL | **61%** | **62%** | **65%** | **37%** | **73%** | **65%** | **65%** | 24% | **57%** |
| | R | 20% | **47%** | 7% | 1% | 6% | 5% | 3% | 9% | 14% |
| | İletişim (Communication) | **43%** | **49%** | **38%** | **63%** | 29% | 28% | **32%** | 29% | **42%** |
| | Spark | 4% | 34% | 1% | 0% | **60%** | 3% | 1% | **34%** | 14% |
| | Oracle | 6% | 0% | 18% | 6% | 20% | 34% | **47%** | 3% | 14% |
| | ETL | 8% | 3% | **44%** | 1% | 39% | **53%** | 1% | 4% | 18% |
| | SAP | 10% | 0% | 25% | 9% | 3% | 13% | 1% | 0% | 10% |
| | Çevik (Agile) | 7% | 8% | 4% | **24%** | 22% | 5% | 2% | 18% | 11% |
| | Tensorflow | 0% | 20% | 0% | 0% | 2% | 0% | 0% | **48%** | 6% |
| | Pozisyon Bazlı İlan Oranı | **18%** | **16%** | **19%** | **17%** | **10%** | **7%** | **8%** | **5%** | |
| | Toplam İlan | **100%** | | | | | | | | |

1,465 iş ilanının 509 tanesinde Python yetkinliğinin geçtiği görülmektedir. %35 oranıyla Python yetkinliği 10 yetkinlik arasında üçüncü sırada yer almaktadır. Python yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimi alanındaki pozisyonlarda en çok aranan üçüncü yetkinliğin, Python yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar Python yetkinliğine odaklanmalıdırlar.

1,465 iş ilanının 269 tanesinde ETL platformu konusundaki yetkinliğin geçtiği görülmektedir. %19 oranıyla ETL platformunda çalışabilme yetkinliği 10 yetkinlik arasında dördüncü sırada yer almaktadır. ETL yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimi alanındaki pozisyonlarda en çok aranan dördüncü yetkinliğin, ETL yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar ETL yetkinliğine odaklanmalıdırlar.

1,465 ilanın içerisinde, R, Spark ve Oracle yetkinliklerinin sırasıyla, 205, 208, 199 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde R, Spark ve Oracle yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya konmuştur. %14 oranıyla R, Spark ve Oracle yetkinlikleri beşinci sırada yer almaktadır. Bu durum, Türkiye'de veri bilimi alanındaki pozisyonlarda R, Spark ve Oracle yetkinliğinin önem derecesinin yakın/benzer olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar R, Spark ve Oracle yetkinliklerine eşit oranda odaklanmalı veya bunlardan bir tanesini seçerek bu alanda gelişmeye çalışmalıdırlar.

1,465 ilanın içerisinde, SAP platformu ve çevik (agile) metodolojisi ile uyumlu çalışabilme yetkinliğinin sırasıyla, 140, 162 ilanda geçtiği görülmektedir. %5 anlamlılık seviyesinde SAP platformu, çevik metodolojisi ile çalışabilme yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya konmuştur. %10 oranıyla SAP, çevik yetkinlikleri altıncı sırada yer almaktadır. Bu durum, Türkiye'de veri bilimi alanındaki pozisyonlarda SAP platformu, çevik metodolojisi ile çalışabilme yetkinliğinin önem derecesinin yakın/benzer olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar SAP platformu, Agile metodolojisi ile çalışabilme yetkinliklerine eşit oranda odaklanmalı veya bunlardan bir tanesini seçerek bu alanda gelişmeye çalışmalıdırlar. Elde edilen analiz sonuçları, eşit derece aranan yetkinliklerde, şirketlerin bu yetkinlikleri birbirinin alternatifi olarak aramadığı, her ikisinin de gerekli ve eşit derece önemli yetkinlikler olduğu şeklinde yorumlanabilir.

1,465 ilanın 88 tanesinde Tensorflow yetkinliğinin geçtiği görülmektedir. %6 oranıyla Tensorflow konusundaki bilgi ve yetkinlik, 10 yetkinlik arasında yedinci sırada yer almaktadır. Tensorflow yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimi alanındaki pozisyonlarda en çok aranan yedinci yetkinliğin, Tensorflow yetkinliği olduğunu ortaya koymaktadır. Veri bilimi alanında çalışmak isteyen adaylar Tensorflow konusunda bilgi ve yetkinliğini geliştirebilirler.

**Tablo 3.** Toplam ilan sayısındaki yetkinlik oranları ve sıralamaları

| Total | Python | SQL | R | Communication | Spark | Oracle | ETL | SAP | Agile | Tensorflow |
|---|---|---|---|---|---|---|---|---|---|---|
| **1465** | 509 | 840 | 205 | 621 | 208 | 199 | 269 | 140 | 162 | 88 |
| **Proportion** | 0,35 | 0,57 | 0,14 | 0,42 | 0,14 | 0,14 | 0,18 | 0,10 | 0,11 | 0,06 |
| **Rank** | 3 | 1 | 5 | 2 | 5 | 5 | 4 | 6 | 6 | 7 |

**Tablo 4.** Yetkinlik bazlı anlamlılık test sonuçları matrisi

| | Python | SQL | R | Communication | Spark | Oracle | ETL | SAP | Agile | Tensorflow |
|---|---|---|---|---|---|---|---|---|---|---|
| **Python** | | -12,60** | 13,48** | -4,26** | 13,32** | 13,81** | 10,22** | 17,23** | 15,90** | 20,67** |
| **SQL** | | | 27,46** | 8,18** | 27,28** | 27,83** | 23,75** | 31,79** | 30,25** | 35,81** |
| **R** | | | | -18,00** | -0,16 | 0,32 | -3,22** | 3,73** | 2,40** | 7,27** |
| **Communication** | | | | | 17,84** | 18,34** | 14,65** | 21,85** | 20,49** | 25,40** |
| **Spark** | | | | | | 0,48 | -3,06** | 3,89** | 2,56** | 7,43** |
| **Oracle** | | | | | | | -3,54** | 3,41** | 2,08** | 6,96** |
| **ETL** | | | | | | | | 6,93** | 5,61** | 10,41** |
| **SAP** | | | | | | | | | -1,34 | 3,59** |
| **Agile** | | | | | | | | | | 4,91** |
| **Tensorflow** | | | | | | | | | | |

### 3.11.2. Pozisyonlar Bazında Anlamlı Farklılıkların Sınanması

Bir önceki bölümde veri bilimi pozisyonlarına yönelik elde edilen toplam 1,465 ilan üzerinden anlamlılık test sonuçları Tablo 3 ve Tablo 4'te paylaşılmıştır. Bu bölümde ise, her pozisyon bazında elde edilen anlamlılık test sonuçları tablolarda gösterilerek yorumlanmaya çalışılmıştır.

- **İş Analisti Anlamlılık Test Sonuçları**

253 iş ilanının 160 tanesinde iletişim yetkinliğinin geçtiği görülmektedir. %63 oranıyla iletişim yetkinliği iş analisti pozisyonu için beş yetkinlik arasında birinci sırada yer almaktadır. İletişim yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de iş analisti pozisyonu için en çok aranan birinci yetkinliğin, iletişim yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar iletişim yetkinliğine odaklanmalıdırlar.

253 iş ilanının 93 tanesinde SQL yetkinliğinin geçtiği görülmektedir. %37 oranıyla SQL yetkinliği beş yetkinlik arasında ikinci sırada yer almaktadır. SQL yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de iş analisti pozisyonu için en çok aranan ikinci yetkinliğin, SQL yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar SQL yetkinliğine odaklanmalıdırlar.

253 iş ilanının 60 tanesinde çevik (agile) metodolojisine ilişkin yetkinliğin geçtiği görülmektedir. %24 oranıyla çevik (agile) metodolojisine ilişkin yetkinlik beş yetkinlik arasında üçüncü sırada yer almaktadır. Çevik (agile) metodolojisine ilişkin yetkinlik, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de iş analisti pozisyonu için en çok aranan üçüncü yetkinliğin, çevik (agile) metodolojisi ile uyumlu çalışabilme yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adayların proje yönetimi süreçlerinde kullanılan bir metodoloji olan agile metodolojisi ile çalışabilecek yetkinliğe sahip olmalarının şirketler için önem arz ettiğini göstermektedir.

253 iş ilanının içerisinde, Oracle ve SAP platformlarına ilişkin yetkinliklerin sırasıyla, 14, 24 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde Oracle ve SAP platformlarına ilişkin yetkinlikler arasında anlamlı bir farklılık olmadığı ortaya konmuştur. Oracle ve SAP yetkinliği iş analisti pozisyonu dördüncü sırada ve eşit derecede aranan yetkinliklerdir. Bu durum her iki yetkinliğin şirketlerce arandığı ve eşit derece önem taşıdığı şeklinde yorumlanabilir.

**Tablo 5.** İş analisti pozisyonu-anlamlılık test sonuçları

| İş Analisti | SQL | Communication | Oracle | SAP | Agile |
|---|---|---|---|---|---|
| **253** | 93 | 160 | 14 | 24 | 60 |
| **Proportion** | 0,37 | 0,63 | 0,06 | 0,09 | 0,24 |
| **Rank** | 2 | 1 | 4 | 4 | 3 |

| | SQL | Communication | Oracle | SAP | Agile |
|---|---|---|---|---|---|
| **SQL** | | -6,18** | 9,31** | 7,69** | 3,23** |
| **Communication** | | | 17,20** | 15,15** | 9,78** |
| **Oracle** | | | | -1,69** | -5,99** |
| **SAP** | | | | | -4,38** |
| **Agile** | | | | | |

Anlamlılık test sonuçları, İş analisti pozisyonunu tanımlayan ilk üç yetkinliğin sırasıyla, iletişim, SQL ve çevik metodolojisine ilişkin çalışabilme yetkinliği olduğunu ortaya koymuştur.

- **İş Zekâsı Uzmanı Anlamlılık Test Sonuçları**

277 iş ilanının 179 tanesinde SQL yetkinliğinin geçtiği görülmektedir. %65 oranıyla SQL yetkinliği beş yetkinlik arasında birinci sırada yer almaktadır. SQL yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de iş zekâsı pozisyonu için en çok aranan birinci yetkinliğin, SQL yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar SQL yetkinliğine odaklanmalıdırlar.

277 iş ilanının içerisinde, iletişim yetkinliğinin ve ETL platformuna ilişkin yetkinliğin sırasıyla, 105 ve 122 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde iletişim yetkinliği ve ETL platformuna ilişkin yetkinlik arasında anlamlı bir farklılık olmadığı ortaya konmuştur. İletişim yetkinliği ve ETL platformuna ilişkin yetkinlik iş zekâsı pozisyonu için ikinci sırada ve eşit derecede aranan yetkinliklerdir.

277 iş ilanının içerisinde, Oracle ve SAP platformuna ilişkin yetkinliklerin sırasıyla, 50 ve 68 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde Oracle ve SAP yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. Oracle ve SAP yetkinliği iş zekâsı pozisyonu için üçüncü sırada ve eşit derecede aranan yetkinliklerdir.

Anlamlılık test sonuçları, İş Zekâsı pozisyonunu tanımlayan ilk yetkinliğin SQL olduğunu, ikinci sırada, iletişim ve ETL yetkinliğinin aynı önem derecesiyle ikinci sırada yer aldığını, üçüncü sırada ise Oracle ve SAP yetkinliklerinin aynı önem derecesinde olduğu görülmüştür.

**Tablo 6.** İş zekâsı pozisyonu anlamlılık test sonuçları

| İş Zekası | SQL | Communication | Oracle | ETL | SAP |
|---|---|---|---|---|---|
| **277** | 179 | 105 | 50 | 122 | 68 |
| **Proportion** | 0,65 | 0,38 | 0,18 | 0,44 | 0,25 |
| **Rank** | 1 | 2 | 3 | 2 | 3 |

| | SQL | Communication | Oracle | ETL | SAP |
|---|---|---|---|---|---|
| **SQL** | | 6,53** | 12,63** | 4,97** | 10,37** |
| **Communication** | | | 5,34** | -1,47 | 3,43** |
| **Oracle** | | | | -6,89** | -1,87 |

| | |
|---|---|
| **ETL** | 4,94** |
| **SAP** | |

- **Veri Analizcisi Anlamlılık Test Sonuçları**

264 ilanın 161 tanesinde SQL yetkinliğinin geçtiği görülmektedir. %61 oranıyla SQL yetkinliği beş yetkinlik arasında birinci sırada yer almaktadır. SQL yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri analisti pozisyonu için en çok aranan birinci yetkinliğin, SQL yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar SQL yetkinliğine odaklanmalıdırlar.

264 ilanın 114 tanesinde iletişim yetkinliğinin geçtiği görülmektedir. %43 oranıyla iletişim yetkinliği beş yetkinlik arasında birinci sırada yer almaktadır. İletişim yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri analizcisi pozisyonu için en çok aranan ikinci yetkinliğin, iletişim yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar iletişim yetkinliğine odaklanmalıdırlar.

264 ilanın 92 tanesinde Python yetkinliğinin geçtiği görülmektedir. %35 oranıyla Python yetkinliği beş yetkinlik arasında üçüncü sırada yer almaktadır. Python yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri analizcisi pozisyonu için en çok aranan üçüncü yetkinliğin, Python yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar Python yetkinliğine odaklanmalıdırlar.

264 ilanın 52 tanesinde R yetkinliğinin geçtiği görülmektedir. %20 oranıyla R yetkinliği beş yetkinlik arasında dördüncü sırada yer almaktadır. R yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri analizcisi pozisyonu için en çok aranan dördüncü yetkinliğin, R yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar R yetkinliğine odaklanmalıdırlar.

264 ilanın 27 tanesinde SAP platformuna ilişkin yetkinliğin geçtiği görülmektedir. %10 oranıyla SAP yetkinliği beş yetkinlik arasında beşinci sırada yer almaktadır. SAP yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri analizcisi pozisyonu için en çok aranan beşinci yetkinliğin, SAP platformuna ilişkin bilgi ve yetkinliğin olduğunu ortaya koymaktadır.

**Tablo 7.** Veri analizcisi-anlamlılık test sonuçları

| **Veri Analizcisi** | **Python** | **SQL** | **R** | **Communication** | **SAP** |
|---|---|---|---|---|---|
| **264** | 92 | 161 | 52 | 114 | 27 |
| **Proportion** | 0,35 | 0,61 | 0,20 | 0,43 | 0,10 |
| **Rank** | 3 | 1 | 4 | 2 | 5 |

| | **Python** | **SQL** | **R** | **Communication** | **SAP** |
|---|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **Python** | -6,23** | 3,97** | -1,97** | 7,08** |
| **SQL** | | 10,66** | 4,16** | 14,36** |
| **R** | | | -6,01** | 3,08** |
| **Communication** | | | | 9,22** |
| **SAP** | | | | |

Anlamlılık test sonuçları, veri analizcisi pozisyonunu tanımlayan ilk üç yetkinliğin sırasıyla SQL, iletişim, Python olduğunu ortaya koymaktadır. Bununla beraber, veri analizcisi pozisyonu için sıralanan beş yetkinlik %5 seviyesinde anlamlı derecede farklılaşmıştır. Veri Analizcisi pozisyonu, bu beş yetkinlik bazında farklı oranlarda temsil edilerek tanımlanmaktadır.

- **Veri Mühendisi Anlamlılık Test Sonuçları**

144 iş ilanının içerisinde, Python, SQL, Spark yetkinliklerinin sırasıyla, 98, 105 ve 86 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde Python, SQL, Spark yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. Python, SQL, Spark yetkinliği veri mühendisi pozisyonunda birinci sırada ve eşit derecede aranan yetkinliklerdir.

144 ilanın içerisinde, İletişim yetkinliği ve ETL platformuna ilişkin çalışabilme yetkinliğinin sırasıyla, 42 ve 56 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde İletişim ve ETL yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. Python, SQL, Spark yetkinliği veri mühendisi pozisyonunda birinci sırada ve eşit derecede aranan yetkinliklerdir.

Anlamlılık test sonuçları, Veri Mühendisi pozisyonunu tanımlayan ilk iki yetkinliğin ilk setinde Python, SQL ve Spark olduğunu, ikinci yetkinlik setinde ise iletişim ve ETL olduğunu ortaya koymuştur. Veri mühendisi pozisyonunda beş yetkinliğin her birinin %5 seviyesinde anlamlı bir şekilde farklılaşmadığı gruplaşma olduğu görülmüştür.

**Tablo 8.** Veri mühendisi-anlamlılık test sonuçları

| **Veri Mühendisi** | **Python** | **SQL** | **Communication** | **Spark** | **ETL** |
|---|---|---|---|---|---|
| **144** | 98 | 105 | 42 | 86 | 56 |
| **Proportion** | 0,68 | 0,73 | 0,29 | 0,60 | 0,39 |
| **Rank** | 1 | 1 | 2 | 1 | 2 |

| | **Python** | **SQL** | **Communication** | **Spark** | **ETL** |
|---|---|---|---|---|---|
| **Python** | | -0,91 | 7,17** | 1,48 | 5,19** |
| **SQL** | | | 8,26** | 2,39** | 6,19** |
| **Communication** | | | | -5,48** | -1,75 |
| **Spark** | | | | | 3,62** |
| **ETL** | | | | | |

- **Veri Bilimcisi Anlamlılık Test Sonuçları**

232 iş ilanının 197 tanesinde Python yetkinliğinin geçtiği görülmektedir. %85 oranıyla Python yetkinliği beş yetkinlik arasında birinci sırada yer almaktadır. Python yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimcisi pozisyonu için en çok aranan birinci yetkinliğin, Python yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar Python yetkinliğine odaklanmalıdırlar.

232 iş ilanının 144 tanesinde SQL yetkinliğinin geçtiği görülmektedir. %62 oranıyla SQL yetkinliği beş yetkinlik arasında ikinci sırada yer almaktadır. SQL yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimcisi pozisyonu için en çok aranan ikinci yetkinliğin, SQL yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar SQL yetkinliğine odaklanmalıdırlar.

232 iş ilanının içerisinde, R ve iletişim yetkinliklerinin sırasıyla, 109 ve 113 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde R ve İletişim yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. R ve iletişim yetkinliği veri bilimcisi pozisyonunda üçüncü sırada ve eşit derecede aranan yetkinliklerdir.

232 iş ilanının 78 tanesinde Spark yetkinliğinin geçtiği görülmektedir. %34 oranıyla Spark yetkinliği beş yetkinlik arasında dördüncü sırada yer almaktadır. Spark yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri bilimcisi pozisyonu için en çok aranan dördüncü yetkinliğin, Spark yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar Spark yetkinliğine odaklanmalıdırlar.

**Tablo 9.** Veri bilimcisi-anlamlılık test sonuçları

| **Veri Bilimcisi** | **Python** | **SQL** | **R** | **Communication** | **Spark** |
|---|---|---|---|---|---|
| **232** | 197 | 144 | 109 | 113 | 78 |
| **Proportion** | 0,85 | 0,62 | 0,47 | 0,49 | 0,34 |
| **Rank** | 1 | 2 | 3 | 3 | 4 |

| | **Python** | **SQL** | **R** | **Communication** | **Spark** |
|---|---|---|---|---|---|
| **Python** | | 5,77** | 9,41** | 8,97** | 13,18** |
| **SQL** | | | 3,30** | 2,92** | 6,40** |
| **R** | | | | -0,37 | 2,96** |
| **Communication** | | | | | 3,34** |
| **Spark** | | | | | |

- **Veri Ambarı Uzmanı Anlamlılık Test Sonuçları**

105 iş ilanının içerisinde, SQL ve ETL yetkinliklerinin sırasıyla, 68 ve 56 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde SQL ve ETL yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. SQL ve ETL yetkinliği veri ambarı uzmanı pozisyonunda birinci sırada ve eşit derecede aranan yetkinliklerdir.

105 iş ilanının içerisinde, İletişim yetkinliği ve Oracle platformu ile çalışabilme yetkinliğinin sırasıyla, 29 ve 36 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde İletişim ve Oracle yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. İletişim ve Oracle yetkinliği veri ambarı uzmanı pozisyonunda ikinci sırada ve eşit derecede aranan yetkinliklerdir.

105 iş ilanının 14 tanesinde SAP platformu ile çalışabilme yetkinliğinin geçtiği görülmektedir. %13 oranıyla SAP yetkinliği beş yetkinlik arasında üçüncü sırada yer almaktadır. SAP yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri ambarı uzmanı pozisyonu için en çok aranan üçüncü sıradaki yetkinliğin, SAP platformu ile çalışabilme olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar SAP platformu ile konusunda bilgi ve yetkinlik edinmeye odaklanmalıdırlar.

**Tablo 10.** Veri ambarı uzmanı-anlamlılık test sonuçları

| **Veri Ambarı Uzm.** | **SQL** | **Communication** | **Oracle** | **ETL** | **SAP** |
|---|---|---|---|---|---|
| **105** | 68 | 29 | 36 | 56 | 14 |
| **Proportion** | 0,65 | 0,28 | 0,34 | 0,53 | 0,13 |
| **Rank** | 1 | 2 | 2 | 1 | 3 |

| | **SQL** | **Communication** | **Oracle** | **ETL** | **SAP** |
|---|---|---|---|---|---|
| **SQL** | | 5,82** | 4,64** | 1,70 | 8,99** |
| **Communication** | | | -1,05 | -3,93** | 2,61** |
| **Oracle** | | | | -2,83** | 3,68** |
| **ETL** | | | | | 6,79** |
| **SAP** | | | | | |

- **Veri Tabanı Uzmanı-Anlamlılık Test Sonuçları**

110 iş ilanının 71 tanesinde SQL yetkinliğinin geçtiği görülmektedir. %65 oranıyla SQL yetkinliği beş yetkinlik arasında birinci sırada yer almaktadır. SQL yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri tabanı uzmanı pozisyonu için en çok aranan birinci yetkinliğin, SQL yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar SQL yetkinliğine odaklanmalıdırlar

110 iş ilanının 52 tanesinde Oracle platformu ile çalışabilme yetkinliğinin geçtiği görülmektedir. %47 oranıyla Oracle platformu ile çalışabilme yetkinliği beş yetkinlik arasında ikinci sırada yer almaktadır. Oracle yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri tabanı uzmanı pozisyonu için en çok aranan ikinci yetkinliğin, Oracle yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar Oracle platformu konusunda bilgi ve yetkinlik edinmeye odaklanmalıdırlar.

110 iş ilanının 35 tanesinde iletişim yetkinliğinin geçtiği görülmektedir. %32 oranıyla iletişim yetkinliği beş yetkinlik arasında üçüncü sırada yer almaktadır. İletişim yetkinliği, %5 anlamlılık

seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de veri tabanı uzmanı pozisyonu için en çok aranan üçüncü yetkinliğin, iletişim yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar iletişim yetkinliğine odaklanmalıdırlar.

105 iş ilanının içerisinde, Python ve R yetkinliklerinin sırasıyla, yedi ve üç kere geçtiği görülmektedir. %5 anlamlılık seviyesinde Python ve R yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. Python ve R yetkinliği veri tabanı uzmanı pozisyonunda dördüncü sırada ve eşit derecede aranan yetkinliklerdir.

**Tablo 11.** Veri tabanı uzmanı-anlamlılık test sonuçları

| Veri Tabanı Uzm. | Python | SQL | R | Communication | Oracle |
|---|---|---|---|---|---|
| **110** | 7 | 71 | 3 | 35 | 52 |
| **Proportion** | 0,06 | 0,65 | 0,03 | 0,32 | 0,47 |
| **Rank** | 4 | 1 | 4 | 3 | 2 |

| | Python | SQL | R | Communication | Oracle |
|---|---|---|---|---|---|
| **Python** | | -11,36** | 1,30 | -5,08** | -7,72** |
| **SQL** | | | 12,83** | 5,14** | 2,62** |
| **R** | | | | -6,18** | -8,90** |
| **Communication** | | | | | -2,37** |
| **Oracle** | | | | | |

- **Makine Öğrenmesi Mühendisi-Anlamlılık Test Sonuçları**

80 iş ilanının 69 tanesinde Python yetkinliğinin geçtiği görülmektedir. %86 oranıyla Python yetkinliği beş yetkinlik arasında birinci sırada yer almaktadır. Python yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de makine öğrenmesi mühendisi pozisyonu için en çok aranan birinci yetkinliğin, Python yetkinliği olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar Python yetkinliğine odaklanmalıdırlar.

80 iş ilanının 38 tanesinde Tensorflow konusundaki yetkinliğin geçtiği görülmektedir. %48 oranıyla Tensorflow yetkinliği beş yetkinlik arasında ikinci sırada yer almaktadır. Tensorflow yetkinliği, %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaşmaktadır. Bu durum, Türkiye'de makine öğrenmesi mühendisi pozisyonu için en çok aranan ikinci yetkinliğin, Tensorflow konusundaki bilgi ve yetkinliğin olduğunu ortaya koymaktadır. Bu alanda çalışmak isteyen adaylar Tensorflow yetkinliğine odaklanmalıdırlar.

**Tablo 12.** Makine öğrenmesi mühendisi-anlamlılık test sonuçları

| Makine Öğrenmesi Müh. | Python | Communication | Spark | Agile | Tensorflow |
|---|---|---|---|---|---|
| **80** | 69 | 23 | 27 | 14 | 38 |
| **Proportion** | 0,86 | 0,29 | 0,34 | 0,18 | 0,48 |

| | | | | | |
|---|---|---|---|---|---|
| **Rank** | 1 | 3 | 3 | 3 | 2 |

| | **Python** | **Communication** | **Spark** | **Agile** | **Tensorflow** |
|---|---|---|---|---|---|
| **Python** | | 9,04** | 8,03** | 11,99** | 5,71** |
| **Communication** | | | -0,68 | 1,70 | -2,49** |
| **Spark** | | | | 2,40** | -1,79 |
| **Agile** | | | | | -4,28** |
| **Tensorflow** | | | | | |

80 iş ilanının içerisinde, iletişim, Spark ve Agile yetkinliklerinin sırasıyla, 23, 27 ve 14 kere geçtiği görülmektedir. %5 anlamlılık seviyesinde iletişim, Spark ve Agile yetkinlikleri arasında anlamlı bir farklılık olmadığı ortaya çıkmıştır. İletişim, Spark ve Agile yetkinliği makine öğrenmesi mühendisi pozisyonunda üçüncü sırada ve eşit derecede aranan yetkinliklerdir.

## 3.12. BULGULAR VE TARTIŞMA

Dijital teknolojilerin hızlı gelişimi ve yaygınlaşması ve beraberinde ortaya çıkan büyük veri şirketlerin ajandalarında ilk sırada yer almaktadır. Dijitalleşmenin etkisi ile müşteri deneyimi ön plana çıkmış, şirketler hızlı ve hatasız hizmet vermeye odaklanmıştır. Müşteri deneyimi, stratejik kararlar, operasyonel süreçlerin iyileştirilmesi, artan müşteri verisinin sorunsuz ve güvenli bir şekilde saklanması gibi birçok başlık veri bilimi başlığında toplanmaktadır.

Şirketler için büyük rekabet avantajı sağlayan veri, veri bilimi alanına ilgiyi her geçen gün artırmaktadır. Dolaşıma giren verinin artması ise, veri bilimi alanında istihdam edilen çalışan sayısının artmasına neden olmaktadır. Veri bilimi alanındaki pozisyonların birçoğunun günümüz ihtiyaçları ile yakın geçmişte doğmuş olması, yeni ortaya çıkan bu pozisyonların çalışan yetkinliklerini ve iş tanımını belirlemede zorluk oluşturmaktadır. İleride de yeni rollerin ortaya çıkmasının beklenmesi, mevcut rollerin iş tanımı ve yetkinlik setlerinin henüz tam anlamıyla netleştirilmemiş olması, aynı zamanda bu alandaki akademik çalışmaların az olması veri bilimi yetkinlik setlerinin incelenmesi ihtiyacını doğurmuştur.

Bu çalışmada, iş ilanlarında veri yönetiminden yapay öğrenme uygulamalarına kadar geniş bir alanı kapsayan sekiz farklı pozisyon tanımı bir üst kavram olarak veri bilimi alanında kabul edilmiştir. Bu iş pozisyonları için Türkiye genelinde açılan 1,467 iş ilanı toplanarak metin madenciliği teknikleri ile incelenmiş ve bu pozisyonlar için aranan yetkinliklerin çerçevesi ortaya konulmuştur. Çalışma sonucunda, işletmelerin veri bilimi pozisyonlarını ve bu pozisyonlar için gerekli yetkinlikleri aynı şekilde tanımlamadıklarını gösteren anlamlı farklılıklara rastlanmıştır.

**Veri analizcisi** pozisyonu için 264 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda istihdam edilecek kişiler için en çok geçen ilk üç yetkinliğin SQL, İletişim ve Python olmuştur.

SQL, İletişim ve Python kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- SQL kelimesine en yakın kelimelerin, iletişim, veri tabanı, modelleme, Python, görselleştirme kelimeleri olduğu;
- İletişim kelimesine en yakın kelimelerin, excel, SQL, modelleme, Python, planlama, veri tabanı olduğu;
- Python kelimesine en yakın kelimelerin ise, iletişim, planlama, excel, takım, *problem çözme* kelimelerinin olduğu görülmüştür.

Aranan nitelikler metinlerinde en çok geçen kelimeler ve bu kelimelerin etrafındaki kelime kümeleri, veri analizcisi pozisyonu, SQL ve Python kullanabilen, iletişim yeteneği güçlü kişilerin bu alanda istihdam edilirken tercih edildiğini göstermektedir. Buna ek olarak, modelleme ve görselleştirme yapabilen ve problem çözme yeteneğine sahip kişiler yine bu pozisyon için tercih edilecek adaylar arasında yer alabilecektir.

Aranan nitelikler metninde N-gram analizi ile ikili kelime gruplarına bakıldığında, yukarıda tek bir kelime olarak ifade edilen kelimeleri tamamlayıcı bir bilgi elde edilmiştir. Buna göre, sıklık sırasıyla iletişim becerisi, problem çözme, Ms Office, iş zekâsı ve power bi kelime gruplarının olduğu görülmüştür. Görsel raporlamaya imkân tanıyan Power bi aracı yukarıda geçen görselleştirme yetkinliğini tamamlayıcı niteliktedir. Bu kapsamda bu alanda istihdam edilecek kişi, görselleştirme aracı olarak power bi kullanılması beklenmektedir şeklinde yorumlanabilir.

Veri analizcisi iş tanımı veri kümesi N-gram analizi ile analiz edilmiştir. İş tanımı başlığında kelime sıklık grafikleri veya kelime bulutu ile anlamlı sonuçlar elde edilememiştir. Bu nedenle iş tanımı veri kümesine sadece N-gram analizi uygulanmıştır. Bu başlık için ikili ve üçlü kelime grupları incelendiğinde, iş tanımında, eyleme geçirilebilir iç görüler, özel analizler, makine öğrenmesi, istatistik tekniklerin kullanılması, stratejik karar verme kelime gruplarının olduğu görülmüştür.

Veri analizcisi üniversite bölümleri için kelime bulutuna bakıldığında, matematik ve istatistik bölümlerinin yanı sıra, matematik mühendisliği, işletme, ekonomi gibi bölümlerin yer aldığı görülmüştür. Veri analizcisi üniversite bölümü n-gram analizi ile de analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde endüstri mühendisliği, bilgisayar mühendisliği bölümlerinin ön plana çıktığı görülmüştür.

**Veri bilimcisi** pozisyonu için 232 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda istihdam edilecek kişiler için en çok geçen ilk üç yetkinliğin, Python, SQL ve İletişim olduğu görülmüştür. Bu noktada dikkat çekici durum, veri analizcisi pozisyonu için de aranan ilk üç yetkinliğin aynı olmasıdır. Veri bilimcisi pozisyonu için, Python yetkinliği ilk sırada gelirken, veri analizcisi pozisyonu için, SQL ilk sırada gelmekteydi.

Python, SQL ve İletişim kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- Python kelimesine en yakın kelimelerin, SQL, karmaşık, süreç, Hadoop, müşteri kelimeleri olduğu;
- SQL kelimesine en yakın kelimelerin, Python, Spark, İletişim, R, müşteri, kelimeleri olduğu;
- İletişim kelimesine en yakın kelimelerin ise, Python, SQL, Spark, R, hizmet kelimelerinin olduğu görülmüştür.

Aranan nitelikler veri kümesinde, en çok geçen kelimeler ve bu kelimelerin etrafındaki kelime kümeleri incelendiğinde, veri bilimcisinin Python ve SQL kullanabilen, iletişim yeteneği güçlü kişiler olması beklendiği görülmüştür. Buna ek olarak, büyük veri analiz araçlarından olan Hadoop kelimesinin geçmesi, büyük veriye, Spark kelimesinin geçmesi gerçek zamanlı veri işleme ve veri entegrasyon süreçlerine işaret etmektedir.

Veri bilimcisi aranan nitelikler veri kümesi ayrıca N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, sırasıyla, makine öğrenmesi, iletişim becerileri, problem çözme, derin öğrenme, Python ve R kelime gruplarının olduğu, üçlü kelime gruplarında da benzer şekilde bu kelime gruplarının yer aldığı görülmüştür. Bu bağlamda, veri bilimcisi, Python ve R kullanabilen, SQL ile sorgu yapabilen, makine öğrenmesi ve derin öğrenme konusunda bilgisi olan, iletişimi güçlü bir rol olarak tanımlanabilir.

Veri bilimcisi iş tanımı başlığındaki veri kümesi N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, iş tanımında, makine öğrenmesi modelleri, makine öğrenmesi teknikleri, büyük veri, veri analitiği, gerçek dünya problemleri kelime gruplarının geçtiği görülmüştür. Yukarıda, büyük veriye işaret eden Hadoop aracından ve gerçek zamanlı veri işleme ve veri entegrasyon süreçlerine işaret eden Spark platformundan bahsetmiştik.

Veri bilimcisi üniversite bölümleri için kelime bulutu incelendiğinde, *matematik ve istatistik* bölümlerinin ön plana çıktığını, bununla beraber, bilgisayar mühendisliği ve genel mühendislik başlığının ön plana çıktığı görülmüştür. Baskın olmamakla beraber, ekonomi bölümünün de mezuniyet tercihleri arasında olduğu görülmektedir.

**İş zekâsı** pozisyonu için 277 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda en çok geçen ilk üç yetkinliğin SQL, ETL ve SAP platformu konusundaki yetkinliğin olduğu görülmüştür. SQL yetkinliği veri analizcisi ve veri bilimcisi pozisyonu ile aynı iken, iş zekâsı pozisyonunda ETL SAP kelimeleri ile ilk üç yetkinlik sıralamasında farklılaşmaktadır. ETL extract (ayıklama), transform (dönüştürme) ve load (yükleme) işlemlerinden oluşan bir veri entegrasyon sürecidir. SAP ise şirketlerde departmanların ortak verilere ulaşmasını sağlayan merkezi bir sistem oluşturulmasını sağlar. SAP yaygın olarak kullanılan kurumsal kaynak planlama yazılımlarındandır. Bu bağlamda, iş zekâsı pozisyonunda istihdam edilecek kişilerin veri

analizinden ziyade, veri analizine giden süreçte, verinin ayıklanması, yüklenmesi, veriye erişim sağlanması gibi işlemleri gerçekleştirdiği görülmektedir.

SQL, ETL ve SAP kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- SQL kelimesine en yakın kelimelerin, *ETL, modelleme, sorgu, veri tabanı, problem çözme, yazılım* kelimeleri olduğu;
- ETL kelimesine en yakın kelimelerin, *SQL, SAP, İletişim, yazılım, takım* kelimeleri olduğu;
- SAP kelimesine en yakın kelimelerin ise, *ETL, iletişim, sorgu, takım çalışması, yazılım* kelimelerinin olduğu görülmüştür.

Aranan nitelikler veri kümesinde, en çok geçen kelimeler ve bu kelimelerin etrafındaki kelime kümeleri incelendiğinde, iş zekâsı uzmanının SQL, ETL ve SAP konusunda bilgili, aynı zamanda iletişim yeteneği güçlü, veri tabanlarından sorgu yapabilen, takım çalışmasına yatkın kişiler olması beklendiği görülmüştür.

İş zekâsı uzmanı aranan nitelikler veri kümesi ayrıca N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, sırasıyla veri ambarı, power bi, analitik düşünme, problem çözme, veri görselleştirme araçları, ETL veri modelleme, zaman yönetimi ve planlama kelime gruplarının geçtiği görülmektedir. Bu bağlamda, power bi aracı veri görselleştirme ve raporlama araçlarından birisidir. İş zekâsı uzmanı pozisyonundaki kişilerin bu aracı kullanabilmesinin beklenen bir durum olduğu görülmektedir.

İş zekâsı uzmanı iş tanımı başlığındaki veri kümesi N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, iş tanımında veri ambarı, veri model, ETL süreçleri, power bi, teknoloji gelişimi takibi kelime gruplarının ön plana çıktığı görülmüştür.

İş zekâsı uzmanı olarak istihdam edilecek kişilerin mezuniyet kriterlerine N-gram analizi ve kelime bulutuna bakıldığında, matematik, istatistik, bilgisayar mühendisliği, endüstri mühendisliği, yönetim bilişim sistemleri bölüm mezunlarının bu alanda istihdam için tercih edildiği görülmektedir. İş zekâsı pozisyonunda, veri bilimcisi ve veri analizcisi pozisyonundan farklı olarak yönetim bilişim sistemleri bölümünün mezuniyet bölümleri arasında sayıldığı görülmüştür.

**İş analisti** pozisyonu için 253 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda en çok geçen ilk üç yetkinliğin iletişim, SQL ve çevik (agile) kelimelerinin olduğu görülmüştür. Diğer pozisyonlarda geçen kelimelerden farklı olarak iş analisti pozisyonunda çevik (agile) kelimesi geçmektedir. Çevik yöntem, yazılım geliştirmede kullanılan proje yönetimine özel bir yaklaşımdır. Geleneksel proje yaklaşımlarına bir tepki olarak doğmuştur.

İletişim, SQL ve Çevik (agile) kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- İletişim kelimesine en yakın kelimelerin, SQL, müşteri, takım çalışması, jira kelimeleri olduğu;
- SQL kelimesine en yakın kelimelerin, iletişim, veri tabanı, çevik, jira kelimeleri olduğu;
- Çevik (agile) kelimesine en yakın kelimelerin ise, iletişim, SQL, dokümantasyon kelimelerinin olduğu görülmüştür.

Bu bağlamda jira kelimesi çevik proje yönetiminde kullanılan proje takip uygulamalarından birisidir. İş analisti pozisyonunda istihdam edilecek kişilerin jira uygulaması konusunda yetkinliğinin olması beklenmektedir.

İş analisti aranan nitelikler veri kümesi ayrıca N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, sırasıyla proje yönetimi, yazılım geliştirme, iletişim becerisi, problem çözme, analitik düşünme kelime gruplarının ön plana çıktığı görülmüştür.

İş analisti iş tanımı başlığındaki veri kümesi N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, iş tanımında kullanıcı kabul testleri, yazılım geliştirme takımı, fonksiyonel teknik analiz kelime gruplarının olduğu görülmüştür.

İş analisti uzmanı olarak istihdam edilecek kişilerin mezuniyet kriterlerine N-gram analizi ve kelime bulutu yardımı ile bakıldığında, bilgisayar mühendisliği, endüstri mühendisliği ve yönetim bilişim sistemleri bölümleri mezunlarının bu alanda istihdam için tercih edildiği görülmektedir.

**Veri mühendisi** pozisyonu için 144 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda en çok geçen ilk üç yetkinliğin SQL, Spark, Python kelimelerinin olduğu görülmüştür. Spark yetkinliği veri mühendisi pozisyonunu diğer pozisyonlardan ayrıştırıcı bir yetkinlik olarak karşımıza çıkmaktadır. Spark büyük veri kümeleri üzerinde gerçek zamanlı ve hızlı veri analizine imkân tanıyan açık kaynak kodlu bir kütüphanedir.

SQL, Spark ve Python kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- SQL kelimesine en yakın kelimelerin, Python, mimari, ETL, iletişim, Hadoop kelimeleri olduğu;
- Spark kelimesine en yakın kelimelerin, SQL, sorgu, azure, ETL, hive, kelimeleri olduğu;
- Python kelimesine en yakın kelimelerin ise, SQL, azure, linux, çevik, süreç, Spark kelimelerinin olduğu görülmüştür.

Veri mühendisi pozisyonu için daha önce incelenen ilanlarda rastlanmayan kelimelerle ilgili kısaca bilgi vermek gerekirse;

- Azure, Microsoft'un bulut uygulamasının altyapı geliştirme platformudur. Verilerin depolanması ve yönetilmesi konusunda farklı yöntemler sunar. SQL veri tabanı, Storage Blobs, Storage Tables bu yöntemlerdendir.

- Hive SQL benzeri bir sorgulama aracıdır. Hadoop üzerinden veri sorgulama ve analiz işlemleri yapmak amacıyla kullanılmaktadır.
- Linux ise açık kaynak kodlu bir işletim sistemidir.

Bu kapsamda, veri mühendisi pozisyonunda istihdam edilecek kişilerin, veri sorgulamada SQL kullanabilme yetkinliğinin yanı sıra, büyük veri ile çalışabilme kabiliyeti gerektiren Hive aracını kullanabilmesi de beklenmektedir.

Veri mühendisi aranan nitelikler veri kümesi ayrıca N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, sırasıyla problem çözme, Python, java programlama dilleri, NoSQL veri tabanı, iş zekâsı, Google bulut platformları, iş akışı yönetim araçları kelime gruplarının geçtiği görülmüştür.

Veri mühendisi iş tanımı başlığındaki veri kümesi n-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, makine öğrenmesi, ürün yöneticisi kelime gruplarının ön plana çıktığı görülmektedir.

Veri mühendisi olarak istihdam edilecek kişilerin mezuniyet kriterlerine N-gram analizi ve kelime bulutu yardımı ile bakıldığında, bilgisayar mühendisi ve matematik bölümleri mezunlarının bu alanda istihdam için tercih edildiği görülmektedir.

**Veri ambarı uzmanı** pozisyonu için 105 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda en çok geçen ilk üç yetkinliğin SQL, ETL, Oracle kelimelerinin olduğu görülmüştür. Oracle platformu konusundaki yetkinliğin veri ambarı uzmanı pozisyonunu diğer pozisyonlardan ayrıştırıcı bir yetkinlik olarak karşımıza çıkmaktadır. Oracle ilişkisel veri tabanı yönetim sistemidir.

SQL, ETL ve Oracle kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- SQL kelimesine en yakın kelimelerin, ETL, takım çalışması, SAP, odi, PLSQL kelimeleri olduğu;
- *ETL* kelimesine en yakın kelimelerin, SQL, Oracle, PLSQL, iletişim, mimari kelimeleri olduğu;
- *Oracle* kelimesine en yakın kelimelerin ise, SQL, ETL, Powerbi kelimelerinin olduğu görülmüştür.

Veri ambarı uzmanı pozisyonu için daha önce incelenen ilanlarda rastlanmayan kelimelerle ilgili kısaca bilgi vermek gerekirse;

- Odi, Oracle tarafından geliştirilmiş bir ETL aracıdır.

Veri ambarı uzmanı aranan nitelikler veri kümesi ayrıca n-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, sırasıyla iş zekâsı, ETL araçları, kelime gruplarının geçtiği görülmüştür.

Veri ambarı uzmanı iş tanımı başlığındaki veri kümesi n-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, ETL süreçleri, performans geliştirme çalışması, veri tabanı performansı izleme kelime gruplarının ön plana çıktığı görülmektedir.

Veri ambarı uzmanı olarak istihdam edilecek kişilerin mezuniyet kriterlerine N-gram analizi ve kelime bulutu yardımı ile bakıldığında, bilgisayar mühendisi, matematik yönetim bilişim sistemleri bölümleri mezunlarının bu alanda istihdam için tercih edildiği görülmektedir.

**Veri tabanı uzmanı** pozisyonu için 110 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda en çok geçen ilk üç yetkinliğin SQL, Oracle, iletişim kelimelerinin olduğu görülmüştür.

SQL, Oracle, iletişim kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- SQL kelimesine en yakın kelimelerin, Oracle, sorun giderme, iletişim kelimeleri olduğu;
- Oracle kelimesine en yakın kelimelerin, SQL, takım çalışması, sorun giderme kelimeleri olduğu;
- İletişim kelimesine en yakın kelimelerin ise, Oracle, güvenlik, Redis, MySQL kelimelerinin olduğu görülmüştür.

Veri tabanı uzmanı pozisyonu için daha önce incelenen ilanlarda rastlanmayan kelimelerle ilgili kısaca bilgi vermek gerekirse;

- Redis, açık kaynak kodlu, ilişkisel olmayan (NoSQL) veri tabanıdır.

Veri tabanı uzmanı aranan nitelikler veri kümesi ayrıca N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, sırasıyla Oracle veri tabanı, veri tabanı yönetimi, iletişim becerisi, problem çözme becerisi kelime gruplarının geçtiği görülmüştür.

Veri tabanı uzmanı iş tanımı başlığındaki veri kümesi N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, performans geliştirme, yedekleme-kurtarma, servis kalitesini iyileştirme, bilgi güvenlik politikaları kelime gruplarının ön plana çıktığı görülmektedir.

Veri tabanı uzmanı olarak istihdam edilecek kişilerin mezuniyet kriterlerine N-gram analizi ve kelime bulutu yardımı ile bakıldığında, bilgisayar mühendisi, matematik bölümleri mezunlarının bu alanda istihdam için tercih edildiği görülmektedir.

**Makine öğrenmesi mühendisi** pozisyonu için 80 ilan incelenmiştir. İlanlar kapsamında aranan nitelikler başlığı analiz edildiğinde, bu pozisyonunda en çok geçen ilk üç yetkinliğin Python, Tensorflow, Spark kelimelerinin olduğu görülmüştür.

Python, Tensorflow, Spark kelimelerine en yakın kelime grupları analiz edilmiştir. Buna göre:

- Python kelimesine en yakın kelimelerin, scikitlearn, sınıflandırma, aws, numpy, kelimeleri olduğu;
- Tensorflow kelimesine en yakın kelimelerin, Linux, Python, numpy, apache, problem çözme kelimeleri olduğu;
- Spark kelimesine en yakın kelimelerin ise, NLP, hizmet, uygulama kelimelerinin olduğu görülmüştür.

Makine Öğrenmesi Mühendisi pozisyonu için daha önce incelenen pozisyonlarda rastlanmayan kelimelerle ilgili kısaca bilgi vermek gerekirse;

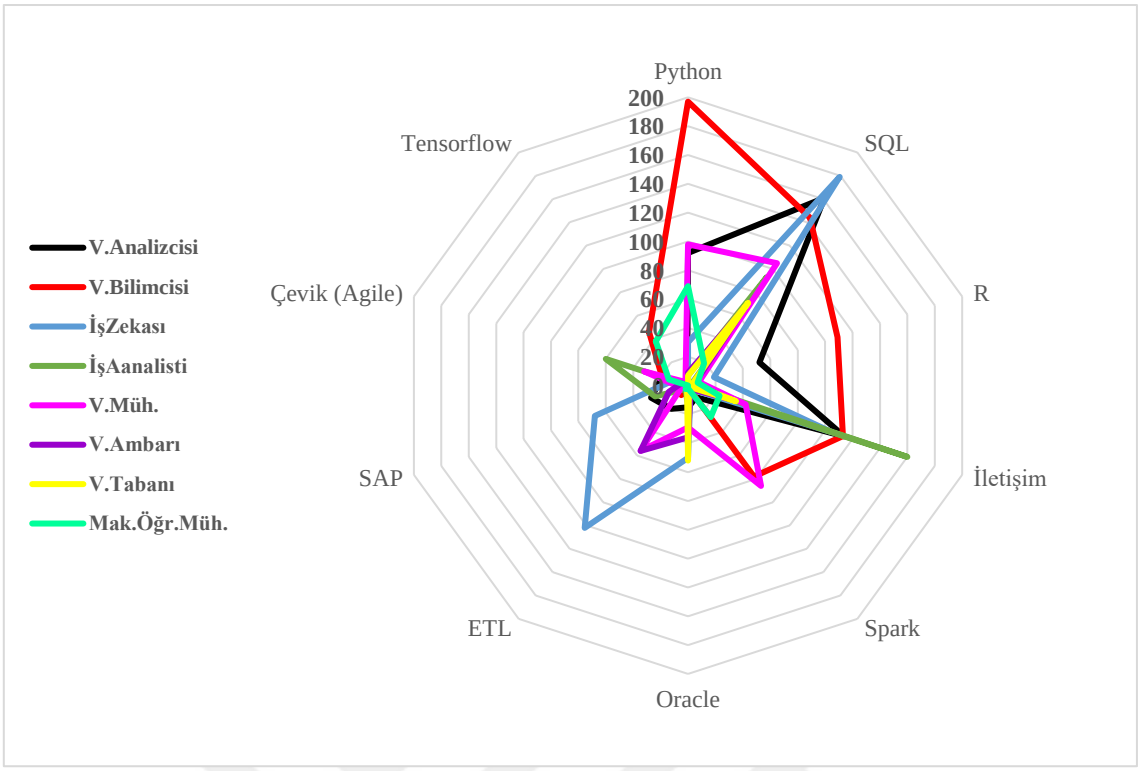
- Scikitlearn veri bilimi ve makine öğrenmesi için kullanılan Python paketlerden birisidir.
- Amazon Web Servis (AWS) bulut platformudur. Bulut tabanlı ürünler sunmaktadır. Depolama, veri tabanları, kurumsal uygulamalar bunlardan bazılarıdır.
- Numpy, veri bilimi çalışmalarında kullanılan bir Python kütüphanesidir.

Makine Öğrenmesi Mühendisi aranan nitelikler veri kümesi ayrıca N-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, sırasıyla bilgisayar görüşü, yapay zekâ, problem çözme becerileri, doğal dil işleme (NLP) kelime gruplarının geçtiği görülmüştür.

Makine Öğrenmesi Mühendisi iş tanımı başlığındaki veri kümesi n-gram analizi ile analiz edilmiştir. İkili ve üçlü kelime grupları incelendiğinde, yapay zekâ, makine öğrenmesi modeli, görüntü işleme, doğal dil işleme kelime gruplarının ön plana çıktığı görülmektedir.

Makine Öğrenmesi Mühendisi olarak istihdam edilecek kişilerin mezuniyet kriterlerine n-gram analizi ve kelime bulutu yardımı ile bakıldığında, *bilgisayar mühendisi, matematik, istatistik* bölümleri mezunlarının bu alanda istihdam için tercih edildiği görülmektedir.

Sekiz iş pozisyonunu 10 yetkinlik bazında görsel olarak aşağıdaki şekilde göstermek mümkündür. Python yetkinliği *veri bilimcisi, makine öğrenmesi mühendisi* pozisyonları için ilk sırada yer alırken; SQL yetkinliği *iş zekâsı, veri analizcisi, veri mühendisi, veri ambarı uzmanı, veri tabanı uzmanı* pozisyonları için ilk sırada aranan nitelikler arasındadır. İletişim yetkinliği ise *iş analisti* pozisyonu için ilk sırada aranan nitelik olmuştur.




**Şekil 106.** Pozisyonlar bazında öne çıkan yetkinlikler

# SONUÇ

Büyük veri ile şirketlerde veriyi işleyen anlamlandıran veriden iç görüler elde etmesi beklenen pozisyonlar hızla ortaya çıkmıştır. Veriden fayda sağlama iştahı ve pazardaki rekabet, şirketleri bu pozisyonlar için yetkin çalışan istihdamına yöneltmektedir. Şirketler, büyük veri ile çalışacak pozisyonlar oluştururken şirket ihtiyaç ve beklentilerini göz önünde bulundurmaktadırlar. Çalışan istihdamı da her şirket özelinde belirlenen kriterlere göre gerçekleşebilmektedir. Bu durum büyük veri ile uğraşan Veri Bilimi alanında aranan pozisyonlar için öznel tanımlamaların ortaya çıkmasına neden olmaktadır. Bu tanımlar zamanla, yazılı olmayan, her şirket ihtiyacına paralel gelişen ve farklılıklar barındıran uygulamalara dönüşebilmektedir.

Bu tez çalışması ile Türkiye'de yayınlanan veri bilimi pozisyonlarına yönelik iş ilanları teknik ve sosyal yetkinlikler açısından incelenmiştir. Analiz edilecek verinin çerçevesi, literatürde yapılmış çalışmalar ışığında ve piyasadaki pratik uygulamalardan ilham alınarak çizilmiştir. Bu doğrultuda "veri", "data", "analiz", "analysis", "analist" anahtar kelimeleri kullanılarak ilanlar taranmıştır. Bu doğrultuda, veri bilimi ile ilişkili olduğu düşünülen "Veri Bilimcisi (Data Scientist)", "Veri Analizcisi (Data Analyst)", "İş Zekâsı Uzmanı (Business Intelligence Speacialist)", "Veri Tabanı Uzmanı (Database Specialist)", "Veri Ambarı Uzmanı (Data Warehouse (DWH) Specialist)", "Makine Öğrenmesi Mühendisi (Machine Learning (ML) Engineer)", "İş Analisti (Business Analyst)" olmak üzere toplamda sekiz pozisyona ait ilanlar veri bilimi üst başlığı altında ele alınarak analiz kapsamında kullanılmıştır. Sekiz pozisyon bazında toplamda 1,465 ilan elde edilmiştir.

Her bir pozisyon için elde edilen metin veri kümesinin kelime sıklık analizine bakılmıştır. Buna göre her pozisyonda ilk sırada gelen ilk üç yetkinlik sıralanmıştır. Pozisyonlar bazında sıralanan ilk üç yetkinlikte ortak yetkinlikler olduğu görülmüştür. Ortak yetkinlikler tek bir kez sayılarak, Python, SQL, R, İletişim, Spark, Oracle, ETL, SAP, Çevik (Agile), Tensorflow olmak üzere 10 adet yetkinlik elde edilmiştir. Bu yetkinlikler pozisyon bazlı bakılarak, her bir ilanda geçip geçmeme durumu 0 veya 1 olarak kodlanarak bir matris oluşturulmuştur. Bu matris üzerinden anlamlılık testleri yapılmıştır. SQL, İletişim, Python, ETL, Tensorflow yetkinliklerinin %5 anlamlılık seviyesinde diğer yetkinliklerden anlamlı bir şekilde farklılaştığı görülmüştür. R, Spark ve Oracle yetkinliklerinin arasında %5 anlamlılık seviyesinde anlamlı bir farklılık olmadığı eşit derece aranan yetkinlikler olduğu görülmüştür. Aynı şekilde SAP, Çevik (Agile) yetkinlikleri arasında %5 anlamlılık seviyesinde anlamlı bir farklılık olmadığı ve eşit derece aranan yetkinlikler olduğu görülmüştür.

Python yetkinliği veri bilimcisi, makine öğrenmesi mühendisi pozisyonları için ilk sırada yer alırken; SQL yetkinliği iş zekâsı, veri analizcisi, veri mühendisi, veri ambarı uzmanı, veri tabanı uzmanı pozisyonları için ilk sırada aranan nitelikler arasında olduğu görülmüştür. İletişim yetkinliği ise iş analisti pozisyonu için ilk sırada aranan nitelik olmuştur.

Özetle; bu tez çalışmasında, iş ilanlarında veri yönetiminden yapay öğrenme uygulamalarına kadar geniş bir alanı kapsayan sekiz farklı pozisyon tanımı bir üst kavram olarak veri bilimi alanında kabul edilmiştir. Bu iş pozisyonları için Türkiye genelinde açılan 1,467 iş ilanı toplanarak metin madenciliği teknikleri ile incelenmiş ve bu pozisyonlar için aranan yetkinliklerin çerçevesi ortaya konulmuştur. Çalışma sonucunda, işletmelerin veri bilimi pozisyonlarını ve bu pozisyonlar için gerekli yetkinlikleri aynı şekilde tanımlamadıklarını gösteren anlamlı farklılıklara rastlanmıştır.

Bu çalışmanın, veri bilimi pozisyonlarına yönelik yayınlanan iş ilanları üzerinden bu alandaki bilgi ve yetkinlik setlerini Türkiye örneği üzerinden inceleyerek analiz eden ilk çalışma olması nedeniyle özgün ve veri bilimi konusundaki yetkinlik ihtiyaçlarını Türkiye örneği üzerinden tartışması açısından da öncü bir çalışma olduğu düşünülmektedir.

Bu çalışmanın bulgu ve sonuçlarının pratik zeminde işletmeler için yol gösterici olması, teorik zeminde ise akademik alanda literatüre katkı sağlaması umulmaktadır.

Bu çalışma ile ortaya konan bulgu ve sonuçlar önümüzdeki dönemde iş ilanları üzerinden tekrar analiz edilerek, yetkinlik çerçevesi yeniden sunulabilir. Buna ek olarak veri kümesindeki sayı artırılarak, farklı ülkelerden de elde edilecek iş ilanları dahil edilerek çalışma sonuçları daha genellenebilir düzeye ulaştırılabilir. Sektör bazında, yetkinlik setlerinin ne denli benzerlik veya farklılık gösterdiği ilerideki araştırma konularından birisi olabilir. Yüksek öğretim kurumlarında bu alandaki müfredat ihtiyaçları mevcut durum ve beklenen durum üzerinden araştırılabilir.

# KAYNAKÇA


Akcay, A. (2019). *Veri Tabanı*. https://aycaakcay.medium.com/veri-tabani-95a062cfdd8c

Akgül, E. S., Ertano, C., & Diri, B. (2016). Sentiment analysis with Twitter. *Pamukkale University Journal of Engineering Sciences*, *22*(2), 106–110. https://doi.org/10.5505/pajes.2015.37268

Akın, B., Şimşek, G., & Tuğba, U. (2018). Sosyal medya analitiği ile değer yaratma: Duygu analizi ile geleceğe yönelim. *Mehmet Akif Ersoy Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *5*(3), 797–811. https://doi.org/10.30798/makuiibf.435804

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. http://arxiv.org/abs/1707.02919

Allied Market Research. (2021). *Big Data and Business Analytics Market Statistics - 2030*. https://www.alliedmarketresearch.com/big-data-and-business-analytics-market

Alpkoçak, A., Tocoglu, M. A., Çelikten, A., & Aygün, İ. (2019). Türkçe metinlerde duygu analizi için farklı makine öğrenmesi yöntemlerinin karşılaştırılması. *Deu Muhendislik Fakultesi Fen ve Muhendislik*, *21*(63), 719–725. https://doi.org/10.21205/deufmd.2019216303

Anderberg, M. R. (1973). Cluster Analysis for Application. In *Academic Press Inc* (Vol. 1).

Apté, C., Damerau, F., & Weiss, S. M. (1994). Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems (TOIS)*, *12*(3), 233–251. https://doi.org/10.1145/183422.183423

Balar, A., Malviya, N., Prasad, S., & Gangurde, A. (2013). Forecasting consumer behavior with innovative value proposition for organizations using big data analytics. *2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2013*. https://doi.org/10.1109/ICCIC.2013.6724280

Başkaya, F., & Aydın, İ. (2017). Haber metinlerinin farklı metin madenciliği yöntemleriyle sınıflandırılması. *International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–5.

Bellinger, G., Castro, D., & Mills, A. (2004). *Data , information , knowledge , and wisdom*.

Berg, K. L., Seymour, T., & Goel, R. (2013). History of databases. *International Journal of Management Information Systems*, *17*(1), 29–36. https://doi.org/10.1002/(SICI)1097-4571(198707)38:4<309::AID-ASI16>3.0.CO;2-#

Berner, M., Graupner, E., & Maedche, A. (2014). The Information Panopticon in the Big Data Era. *Journal of Organization Design*, *3*(1).

Bhatia, P. (2019). Data mining and data warehousing:Principles and Practical Techniques. In *Cambridge University Press*. Cambridge University Press. https://doi.org/10.1007/978-3-540-48399-1_10

Bilginç. (2022). *Python nedir? Python hakkında her şey*. https://bilginc.com/tr/blog/158/python-nedir-python-hakkinda-hersey

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for k-means clustering. *ICML*, *98*, 91–99. https://doi.org/10.7567/JJAP.54.061701

Bryson, S., Kenwright, D., Cox, M., Ellsworth, D., & Haimes, R. (1999). Gigabyte Data Sets in Real Time. *Communications of the ACM*, *42*(8), 82–90.

Cady, F. (2017). *The data science handbook*. John Wiley & Sons, Inc.

Çalış, K., Gazdağı, O., & Yıldız, O. (2012). Reklam içerikli epostaların metin madenciliği yöntemleri ile otomatik tespiti. *International Journal of Informatics Technologies*, *6*(1), 1-7–7.

Carillo, K. D. A. (2017). Let's stop trying to be "sexy" – preparing managers for the (big) data-driven business era. *Business Process Management Journal*, *23*(3), 598–622. https://doi.org/10.1108/BPMJ-09-2016-0188

Çelik, S. (2018). *Büyük Veri ve İstatistikteki Uygulamaları*. Doktora Tezi, Bursa, Uludağ Üniversitesi,Sosyal Bilimler Enstitüsü.

Chapman, P. (1999). The CRISP-DM User Guide. *The CRISP-DM User Guide*, 14.

Chiang, R. H. L., Grover, V., Liang, T.-P., Dongsong, &, Guest, Z., & Zhang, D. (2018). Special Issue: Strategic Value of Big Data and Business Analytics. *Journal of Management Information Systems*, *35*(2), 383–387. https://doi.org/10.1080/07421222.2018.1451950

Chong, D., & Shi, H. (2015). Big data analytics: a literature review. *Journal of Management Analytics*, *2*(3), 175–201. https://doi.org/10.1080/23270012.2015.1082449

Chris Eaton, & Paul Zikopoulos. (2011). *Understanding Big Data | Guide books*. https://dl.acm.org/doi/abs/10.5555/2132803

Codd, E. F., Codd, S. B., & Salley, C. T. (1993). Providing OLAP to User-Analysts - An IT Mandate. *Technical Report*.

Cortes, C., & Vapnik, V. (1995). “Support-vector networks.” *IEEE Expert-Intelligent Systems and Their Applications*, *20*(3), 273–297. https://doi.org/10.1109/64.163674

Davenport, T. (2014). Big data at work: dispelling the myths, uncovering the opportunities. *Choice Reviews Online*, *51*(11), 51-6260-51–6260. https://doi.org/10.5860/choice.51-6260

Davenport, T. H. (2006). Competing on Analytics. *Harvard Business Review*.

Davenport, T. H., & Pat, D. J. (2012). Data Scientist : The Sexiest Job of the 21st Century. *Harvard Business Review*.

Davenport T.H., Dyché, J. (2013). Big data in big companies. *Baylor Business Review*, *32*(1), 20–21. http://search.proquest.com/docview/1467720121?accountid=10067%5Cnhttp://sfx.lib.nccu.edu.tw/sfxlcl41?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ:abiglobal&atitle=VIEW/REVIEW:+BIG+DATA+IN+BIG+COMPANIES&title=Bay

De Mauro, A., Greco, M., Grimaldi, M., & Nobili, G. (2016). Beyond Data Scientists: a Review of Big Data Skills and Job Families. *International Forum on Knowledge Asset Dynamics 2016*, *June*, 1844–1857. http://www.ifkad.org/Proceedings/2016/papers/119_IFKAD2016.pdf

DuCharme, B. (2005). *25 years of database history (starting in 1955)*. https://www.bobdc.com/blog/25-years-of-database-history-s-1/

Dumontier, M., & Kuhn, T. (2017). Data Science – Methods, infrastructure, and applications. *Data Science*, *1*(1–2), 1–5. https://doi.org/10.3233/ds-170013

Eliaçik, A. B., & Erdogan, N. (2015). Mikro bloglardaki finans toplulukları için kullanıcı ağırlıklandırılmış duygu analizi yöntemi. *Ulusal Yazılım Mühendisliği Sempozyumu*, 782–793.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, *17*(3), 37–54. https://doi.org/10.1007/978-3-319-18032-8_50

Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). *International Conference on Knowledge Discovery and Data Mining (KDD)*, *95*, 112–117. http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf

Feldman, R., & Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. In *Cambridge University Press*.

Feldman, R., & Sanger, J. (2016). The Text Minning Handbook. In *Journal of Chemical Information*

*and Modeling* (Vol. 53, Issue 9).

Foote, K. D. (2021). *A Brief History of Analytics*. https://www.dataversity.net/brief-history-analytics/#:~:text=Data analytics is based on,of planning activities%2C including taxation

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2017). Big Data and Social Science - A Practical Guide to Methods and Tools. In *Journal of Statistical Software* (Vol. 78, Issue Book Review 2). https://doi.org/10.18637/jss.v078.b02

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, *13*(3), 407–426. https://doi.org/10.1007/3-540-57253-8_65

Galbraith, J. R. (2014). Organization Design Challenges Resulting From Big Data. *Journal of Organization Design*, *3*(1).

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Ganesh, S. (2002). Data Mining: Should It Be Included In The "Statistics" Curriculum? In *The 6th international conference on teaching statistics (ICOTS 6)*. https://doi.org/10.1002/aris.1440360107

Gantz, B. J., Reinsel, D., & Shadows, B. D. (2012). Big Data , Bigger Digital Shadows , and Biggest Grow th in the Far East Executive Summary: A Universe of Opportunities and Challenges. *Idc*, *2007*(December 2012), 1–16.

Godfrey, D., Johns, C., Meyer, C., Race, S., & Sadek, C. (2014). A case study in text mining: Interpreting twitter data from world cup tweets. *ArXiv Preprint ArXiv:1408.5427*. http://arxiv.org/abs/1408.5427

Goh, Y. M., & Ubeynarayana, C. U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis and Prevention*, *108*(August), 122–130. https://doi.org/10.1016/j.aap.2017.08.026

Göker, H., & Tekedere, H. (2017). FATİH projesine yönelik görüşlerin metin madenciliği yöntemleri ile otomatik değerlendirilmesi. *Bilişim Teknolojileri Dergisi*, *10*(3), 291–299. https://doi.org/10.17671/gazibtd.331041

Grossman, R. L., & Siegel, K. P. (2014). Organizational Models for Big Data and Analytics. *Journal of Organization Design*, *3*(1). https://doi.org/10.7146/jod.3.1.9799

Gtech. (2022). *Veri madenciliği nedir ve nasıl yapılır?* https://www.gtech.com.tr/veri-madenciligi-nedir-ve-nasil-yapilir/

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information and Management*, *53*(8), 1049–1064. https://doi.org/10.1016/j.im.2016.07.004

Hayes, P. J., Andersen, P. M., Nirenburg, I. B., & Schmandt, L. M. (1990). TCS: A Shell for Content-Based Text Categorization. In *In Sixth conference on artificial intelligence for applications*. IEEE.

Hayes, P. J., & Weinstein, S. P. (1991). Adding value to financial news by computer. *In Proceedings of the 1st International Conference on Artlificial Intelligence Applications on Wall Street*, 2–8. https://doi.org/10.1109/aiaws.1991.236581

Hearst, M. A. (1999). *Untangling text data mining*. 3–10. https://doi.org/10.3115/1034678.1034679

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. Computational Linguistics and Language Technology. *Journal of Food Science*, *43*(1), 19–72.

Hu, X., Stephen Downie, J., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Ismir*, 411–416.

IBM. (2020a). *Makine Öğrenmesi*. https://www.ibm.com/tr-tr/cloud/learn/machine-learning

IBM. (2020b). *Yapay zeka*. https://www.ibm.com/tr-tr/cloud/learn/what-is-artificial-intelligence

İlhan, N., & Sağaltıcı, D. (2020). Twitter'da duygu analizi. *"Twitter'da Duygu Analizi", Harran Üniversitesi Mühendislik Dergisi, 5(2): 146-156, (2020).*, *5*(2), 146–156.

Innova. (2021). *Büyük verinin kısa tarihçesi*. https://www.innova.com.tr/tr/blog/buyuk-veri-blog/buyuk-verinin-kisa-tarihcesi

Işik, M., & Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(3), 1405–1421. https://doi.org/10.3906/elk-1907-46

Kaşikçi, T., & Gökçen, H. (2013). Metin madenciliği ile e-ticaret sitelerinin belirlenmesi. *Bilişim Teknolojileri Dergisi*, *7*(1), 0. https://doi.org/10.12973/bid.2014

Kaynar, O., Yildiz, M., Görmez, Y., & Albayrak, A. (2016). *Makine öğrenmesi yöntemleri ile duygu analizi*. *September*, 234–241.

Kezay, A. (2020). *Data mart nedir? Data mart vs data warehouse*. https://www.datasciencearth.com/veri-ambari-egitimi-bolum-8-data-mart-nedir-data-mart-vs-data-warehouse/

Knight, M. (2018). *What Is Analytics?* https://www.dataversity.net/what-is-analytics/

Kodratoff, Y. (1999). Knowledge discovery in texts: A definition, and applications. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1609*, 16–29. https://doi.org/10.1007/BFb0095087

Leek, J. (2013). *The key word in "Data Science" is not Data, it is Science*. https://web.archive.org/web/20180821112933/https://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

Lenat, D. B. (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, *38*(11), 33–38. https://doi.org/10.1145/219717.219745

Lent, B., Agrawal, R., & Srikant, R. (1997). Discovering Trends in Text Databases. *Proc 3rd Int Conf Knowledge Discovery and Data Mining KDD*, 227–230. citeseer.ist.psu.edu/lent97discovering.html

Li, W., Lee, B., Krausz, F., & Sahin, K. (1991). Text Classification by a Neural Network. *23rd Annual Summer Computer Simulation Conference*, 313–318.

Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013). Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, 99–104. https://doi.org/10.1109/BigData.2013.6691740

Lovins, J. B. (1968). Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, *11*(1–2), 22–31. https://doi.org/10.1111/j.1440-1681.1996.tb02836.x

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281–297. https://doi.org/10.1007/s11665-016-2173-6

Marr, B. (2015). *A brief history of big data everyone should read*. https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/

Mcafee, A., & Brynjolfsson, E. (2012). Spotlight on Big Data Big Data: The Management Revolution, 2012. Acedido em 15-03-2017. *Harvard Business Review*, *October*, 1–9. http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf

McCarthy, J. (2004). What is Artificial Intelligence? *Journal of Food Quality*, 1–14. https://doi.org/10.1155/2021/4535567

McKinsey. (2020). *İşimizin Geleceği*.

Miller, S. (2014). Collaborative Approaches Needed to Close the Big Data Skills Gap. *Journal of Organization Design*, *3*(1), 26. https://doi.org/10.7146/jod.9823

Miner, G. D., Elder, J., & Nisbet, R. A. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. In *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. https://doi.org/10.1016/C2010-0-66188-8

Mitchell, T. M. (1997). Machine Learning. New York: McGraw-hill.

Netinbag. (2022). *Nesneye Dayalı Veri Tabanı Yönetim Sistemi Nedir?* https://www.netinbag.com/tr/internet/what-is-an-object-oriented-database-management-system.html

Oracle. (2022a). *NoSQL nedir?* https://www.oracle.com/tr/database/nosql/what-is-nosql/

Oracle. (2022b). *Veritabanı Nedir?* https://www.oracle.com/tr/database/what-is-database/

Peters, B. (2012). The Age of Big Data - Forbes. *Forbes*, 4–9. http://www.forbes.com/sites/bradpeters/2012/07/12/the-age-of-big-data/

Phillips-Wren, G., Iyer, L. S., Kulkarni, U., & Ariyachandra, T. (2015). Business analytics in the context of big data: A roadmap for research. *Communications of the Association for Information Systems*, *37*, 448–472. https://doi.org/10.17705/1cais.03723

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Şeker, S. E. (2015). Metin Madenciliği (Text Mining). *YBS Ansiklopedi*, *2*(3), 2–4.

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). Data Mining For Business Analytics: Concepts, Tehniques, and Applications in Python. In *Mining of Massive Datasets* (Vol. 2, Issue

January 2013). Wiley. https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

Silva, C., & Ribeiro, B. (2003). The Importance of Stop Word Removal on Recall Values in Text Categorization. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 1661–1666. https://doi.org/10.1109/ijcnn.2003.1223656

Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of Big Data*, *2*(1), 1–20. https://doi.org/10.1186/s40537-014-0008-6

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning*, *34*(1), 233–272. https://doi.org/10.1023/a:1007562322031

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. *New Directions in Statistical Physics*, 273–309. https://doi.org/10.1007/978-3-662-08968-2_16

Tabak, J. (2014). Probability and statistics:The Science of Uncertainity. In *Infobase Publishing*. https://doi.org/10.1007/978-1-4939-2565-0_73

Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, *8*, 65–70. https://doi.org/10.1.1.38.7672

Tantuğ, A. C. (2012). Metin Sınıflandırma(Text Classification). *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *5*(2). http://www.bmbb.info/dergi/index.php/dergi/article/view/63

Turhost. (2021). *SQL nedir? Ne işe yarar?* https://blog.turhost.com/sql-nedir-ne-ise-yarar/#serp

Turing, B. A. M. (1950). Computing Machinery and Intelligence. *Boston Studies in the Philosophy and History of Science*, 433–460. https://doi.org/10.1007/978-3-319-53280-6_11

Tuzcu, S. (2020). Çevrimiçi kullanıcı yorumlarının duygu analizi ile sınıflandırılması. *ESTUDAM Bilişim Dergisi*, *1*(2), 1–5.

Tüzüntürk, S. (2010). Veri Madenciliği ve İstatistik. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *1*(2001), 65–90.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Vikipedi. (2022). *Veri bilimi*. https://tr.wikipedia.org/wiki/Veri_bilimi#cite_note-1

Ward, J. H., & Hook, M. E. (1963). Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*, *23*(1), 69–81.

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. *Japanese Society of Biofeedback Research*, *19*, 709–715. https://doi.org/10.20595/jjbf.19.0_3

Wilks, Y., & Bien, J. (1983). Beliefs, points of view, and multiple environments. *Cognitive Science*, *7*(2), 95–119. https://doi.org/10.1016/S0364-0213(83)80007-X

Wishart, D. (1969). 256. Note: An algorithm for hierarchical classifications. *Biometrics*, 165–170.

Witten, I. H. (2004a). Text mining. *The Practical Handbook of Internet Computing*, 14-1-14–22. https://doi.org/10.1201/9780203507223

Witten, I. H. (2004b). Text mining. *The Practical Handbook of Internet Computing*, 1–23. https://doi.org/10.1201/9780203507223

Yeşilyurt, A. (2019). *Apache Spark nedir? Spark'a giriş*. https://amineyesilyurt.medium.com/apache-spark-nedir-sparka-giriş-582d2e0059af

Yıldız, U. A., & Topal, S. (2015). *Büyük Veri Kahramanı Veri Bilimci* (Issue April).

Zhou, Y., Tong, Y., Gu, R., & Gall, H. (2016). Combining text mining and data mining for bug report classification. *Journal of Software: Evolution and Process*, *28*(3), 150–176. https://doi.org/10.1002/smr