T.C.
MARMARA ÜNİVERSİTESİ
TÜRKİYAT ARAŞTIRMALARI ENSTİTÜSÜ
BİLGİ VE BELGE YÖNETİMİ ANABİLİM DALI

ELEKTRONİK BELGE YÖNETİM SİSTEMLERİNDE VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KULLANIMI

DOKTORA TEZİ

MUHAMMET EMİN GEDİKLİ

İSTANBUL, 2023

T.C.
MARMARA ÜNİVERSİTESİ
TÜRKİYAT ARAŞTIRMALARI ENSTİTÜSÜ
BİLGİ VE BELGE YÖNETİMİ ANABİLİM DALI

ELEKTRONİK BELGE YÖNETİM SİSTEMLERİNDE VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KULLANIMI

DOKTORA TEZİ

MUHAMMET EMİN GEDİKLİ

TEZ DANIŞMANI : PROF. DR. BERAT BIRFIN BİR

İSTANBUL, 2023

İÇİNDEKİLER

ÖNSÖZ

Uzmanlar içerisinde bulunduğumuz çağı artık "Bilgi Çağı"nın ötesinde bir kavram olan, dinamiklerinin yaygın ve ucuz internet, kişisel bilgisayar ve akıllı cihazlar olduğu 'Katılım (Bağlantılılık) Çağı' olarak nitelendirmektedir. Bu çağın en önemli özelliği akıl almaz ve takibi mümkün olmayan bir hızla gelişen, değişen teknolojik süreçlerdir. Bu noktada ortaya çıkan veri miktarı her geçen saniye muazzam bir biçimde artmaktadır. Bu artış, aynı zamanda bu büyüklükteki veri yığınlarının analiz edilmesi ve mevcut verilerden anlamlı, değerli bilgilerin çıkarılması problemini de beraberinde getirmektedir.

Günümüzde kullanımı yoğun biçimde devam eden ilişkisel veri tabanları büyük miktardaki verilerin saklanmasını sağlamakla beraber, bu veriler üzerinde yapılan analizler ve anlamlı bilgiler çıkarma işlemi de oldukça zorlaşmaktadır. Bu çalışma ile veri madenciliği tekniklerinin yeterince uygulanmadığı elektronik belge yönetim sistemlerinin gerekli verimlilik ve kalite standartlarına uygun biçimde çalışamayacağı ve bağlı olduğu birçok sürecin performansı ve sonuçlarını da olumsuz etkileyeceği hipotezi ortaya konulmuş ve çalışma bu doğrultuda ilerlemiştir.

Bu bağlamda, elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi algoritmalarına başvurularak bahsi geçen problemlere yönelik çözümler üreten bir model ortaya koyulmuştur. Standart dosya planı kodu doğru şekilde belirlenmiş ve ilgili üst veriler kullanılarak oluşturulmuş bu belirli model, aslında tüm elektronik belge yönetim sistemleri için de genel bir çerçeve ve öneri şeması sunmaktadır. Sınıflandırma yöntemleri aracılığıyla gerçekleştirilen işlemler sonucunda, söz konusu yöntemlerin başarı ölçütlerine göre F1 skoru 0.96 elde edilmiştir.

Uzun ve meşakkatli bu süreçte benden desteklerini esirgemeyen tez danışmanım Prof. Dr. Berat Bir'e ve yardımlarına hiç ara vermemiş önceki tez danışmanım Prof. Dr. Hamza Kandur'a teşekkürü borç bilirim. Bu süreçte fikirlerinden sürekli faydalandığım tez izleme komitemde bulunan Prof. Dr. Tuba Karatepe ve Dr. Öğr. Üyesi Hüseyin Yüce'ye, desteğini esirgemeyen M. Oytun Cibaroğlu'na, tezin son, zor ve sıkıntılı süreçlerinde maddi ve manevi destekleriyle yardımcı olan ve elinden geleni fazlasıyla yapan Halid Metin Yolcu'ya, mesai arkadaşım Doç. Dr. Bahattin Yalçınkaya'ya ve onun nezdinde tüm bölüm hocalarıma ve arkadaşlarıma teşekkürlerimi sunarım.

M. Emin GEDİKLİ
İstanbul, 2023

ÖZET

GEDİKLİ, Muhammet Emin. *Elektronik Belge Yönetim Sistemlerinde Veri Madenciliği ve Makine Öğrenmesi Yöntemlerinin Kullanımı*. Doktora Tezi, İstanbul.

Bu tez çalışmasının odağındaki üç temel unsur elektronik belge yönetim sistemleri, veri madenciliği ve makine öğrenmesi teknikleridir. Bu bağlamda çalışmada bilgi ve belge yönetimi alanında, özellikle de elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi tekniklerinin kullanımına yönelik bir çalışma ortaya koymak ve bu alanda söz konusu tekniklerin daha fazla konuşulması ve uygulamaya geçirilmesi için bir katkı sağlamak hedeflenmektedir.

Çalışmanın hipotezi şu şekilde oluşturulmuştur: Kamu kurum ve kuruluşlarında kullanılan elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi teknikleri kullanılmamakta ve bu da üstveri yönetimini, e-arşiv sistemine geçişi olumsuz etkilemektedir. Bu bağlamda çalışma kapsamında gerçekleştirilen incelemeler ile elektronik belge yönetim sistemlerinde kullanıcı kaynaklı hatalar, saklı örüntüler tespit edilmeye ve performansı arttırmaya yönelik tavsiyelerde bulunulmaya çalışılmış ve bir model önerisinde bulunulmuştur.

Anahtar Sözcükler:

Elektronik Belge Yönetim Sistemi, Üstveri, Veri Madenciliği, Makine Öğrenmesi

ABSTRACT

GEDİKLİ, Muhammet Emin. *The Use of Data Mining and Machine Learning Methods in Electronic Records Management Systems*. PhD Thesis, Istanbul, 2023.

The three main elements in the focus of this thesis study are electronic records management systems, data mining and machine learning techniques. In this context, it is aimed to present a study on the use of data mining and machine learning techniques in the field of information and records management, especially in electronic records management systems, and to contribute to the further discussion and implementation of these techniques in this field.

The hypothesis of the study is formed as follows: Data mining and machine techniques are not used in electronic records management systems used in public institutions and organizations, which negatively affects metadata management and the transition to the e-archive system. In this context, with the examinations carried out within the scope of the study, it is tried to identify user-caused errors, hidden patterns in electronic records management systems and to make recommendations to improve performance and a model was proposed.

Keywords:

Electronic Records Management System, Metadata, Data Mining, Machine Learning

KISALTMALAR

CRISP-DM	Cross-Industry Standard Process for Data Mining
EBYS	Elektronik Belge Yönetim Sistemi
ELAS/RM	Elektronik Arşivleme Sistemi Referans Modeli
MÖ	Makine Öğrenmesi
TS	Türk Standardı
TSE	Türk Standartları Enstitüsü
VM	Veri Madenciliği
YZ	Yapay Zekâ

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

BİRİNCİ BÖLÜM

1. GİRİŞ

Bilgi teknolojileri alanında yaşanan ve özellikle son 20 yılda internetin yaygınlaşması nedeniyle geçmişle karşılaştırılamayacak boyutlara varan gelişmelerin kaçınılmaz bir sonucu olarak veri miktarı her geçen saniye muazzam bir biçimde artmaktadır. Bu artış, aynı zamanda bu büyüklükteki veri yığınlarının analiz edilmesi ve mevcut verilerden anlamlı, değerli bilgilerin çıkarılması problemini de beraberinde getirmektedir. Günümüzde kullanımı yoğun biçimde devam eden ilişkisel veri tabanları büyük miktardaki verilerin saklanmasını sağlamakla beraber, bu veriler üzerinde yapılan analizler ve anlamlı bilgi çıkarma işlemleri de oldukça zorlaşmaktadır. Bu zorlukları aşmak amacıyla ortaya çıkan yeni veritabanı kavramları ve veri analiz teknikleri ilişkisel veri tabanlarına nispetle çok daha verimli ve işlevsel bir fonksiyon icra etmektedir. Bu bağlamda, büyük miktarlardaki verileri yönetmek için "veri ambarı" ve bu verileri analiz etmek, onlardan hareketle anlamlı ve değerli bilgiler elde etmek için ise "veri madenciliği" kavramları öne çıkmaktadır.

Veri madenciliği kavramı, hemen hemen tüm kaynaklarda, veri tabanlarında bulunan verilerden hareketle bir bilgi keşfetmek veya çıkarmak şeklinde tanımlanmaktadır (Han ve diğerleri, 2011, s.1; Larose ve Larose, 2015, s.1; Nisbet ve diğerleri, 2009, s.34). Veri madenciliği tekniklerinin temel amacı ise; büyük miktardaki veri yığınlarından, bulundukları haldeyken fark edilmesi oldukça güç olan değerli ve anlamlı bilgiler elde edebilmektir (Maimon ve Rokach, 2010, s.1; Larose, 2006, s.xi).

Veri madenciliği teknikleri kullanmanın temelde iki amacı olduğunu dile getirmemiz mümkündür; geçmişi anlamak ve gelecek hakkında tahminlerde bulunmak. Bu tekniklerle büyük miktarlarda veri bulunduran veri tabanlarındaki ve/veya veri ambarlarındaki ilişkileri ve olayların düzenli bir biçimde birbirini takip ederek gelişmesi anlamına gelen saklı örüntüleri bulmak ve bu örüntülerden hareketle gelecekte ortaya çıkabilecek çeşitli durumları tahmin etmek için bir iş kuralları dizisi çıkarmak amaçlanmaktadır (Han ve diğerleri, 2022, s.5).

Günümüzde veri madenciliği bankacılık, pazarlama, sigortacılık, sağlık, haberleşme, endüstri, mühendislik, istihbarat, müşteri ilişkileri, insan kaynakları ve benzeri sayısız alanda kullanılmakta ve bu alanlarda ortaya çıkan problemlerin çözümünde kritik bir rol oynamaktadır (Han ve diğerleri, 2011, s.27).

Geleneksel yöntemlerle hesaplanamayacak büyüklükteki bir verinin her an toplandığı ve biriktiği bir çağda yaşamaktayız. Web'de, iş hayatında, bilimde, sağlıkta, mühendislikte, kısacası neredeyse günlük hayatımızın her alanında gerçekleşen dijital dönüşümün ve bilgisayar-teknoloji dünyası ile veri toplama-depolama araçlarının hızla gelişmesinin kaçınılmaz bir sonucu olarak her gün terabaytlarca ve hatta petabaytlarca veri üretmekte ve depolamaktayız (Kotu ve Despande, 2018, s.8). Söz konusu veri birikiminin işlevsiz yığınlara dönüşmemesi ve sadece depolanmak suretiyle fiziksel ya da sanal alanları işgal etmekten öte bir anlam ifade etmesi için bu verilerin etkili, ölçeklenebilir ve esnek bir şekilde analizi önemli bir ihtiyaç haline gelmeye başlamıştır.

Sonuç olarak insanların eylem, işlem ve bilgi birikimlerinin dijital girdi ve çıktıları olarak görülebilecek bu veri yığınları bir bakıma dijitalleşen bütün bir dünyanın ya da yaşamın sanal bir izdüşümü şeklinde görülebilir. Bu durumda söz konusu yığının içinde, bahsi geçen eylem, işlem ve bilgi birikimlerine dair anlamlı sonuçların bulunduğu da gayet açıktır. İşte bu çok büyük miktardaki verilerden orada olduğunu bildiğimiz fakat bir yığın halinde bulunduklarından erişmemizin olanaklı olmadığı anlamlı ve değerli bilgileri ortaya çıkarmak, bu tür verileri organize bilgi haline dönüştürmek için güçlü ve çok yönlü araçlara ihtiyaç duyulması sonucunda veri madenciliği doğmuştur (Han ve diğerleri, 2011, s.2).

Veri madenciliği, büyük miktardaki veri yığınları içerisinde bulunan, daha önceden fark edilmemiş anlamlı bilgileri keşfetme ve saklı örüntüleri bulma işlemi olarak tanımlanmaktadır. Söz konusu işlemler literatürde ortaya çıktığı ilk zamanlarda "veriden bilgi keşfetme işlemi" olarak da ifade edilmiş (Frawley ve diğerleri, 1992, s.57), sonraları "veri madenciliği" kavramı, bu işlemleri karşılamak için kullanılan ortak terim haline gelmiştir. Herhangi bir veri koleksiyonunu işleyerek bilgiye dönüştüren veri madenciliği bu bakımdan disiplinler arası bir kavram olması sebebiyle de birçok farklı tanıma sahip olmuştur.

Veri madenciliği sürecini oluşturan temel adımları sıralamak gerekirse; veri temizleme, veri entegrasyonu, veri seçimi, veri dönüşümü, örüntülerin bulunması,

modelin değerlendirilmesi ve bilgi sunumu (görselleştirme) aşamalarından söz etmek mümkün olabilir (Kantardzic, 2011). İşlenmemiş halde bulunan veriler üzerinde veri madenciliği işlemlerini uygulayabilmek için söz konusu verilerin hazır hale getirilmesini sağlayan bu süreçlerle ilgili çalışmalar da önemli bir yere sahiptir.

Veri madenciliği, bilgi teknolojisinin doğal gelişiminin bir sonucu olarak da görülebilir. Veri tabanı ve veri yönetimi endüstrisi; veri toplama, veri tabanı oluşturma, veri yönetimi ve gelişmiş veri analizi gibi çeşitli kritik işlevlerin geliştirilmesi şeklinde ilerleme kaydetmiştir. Veri toplama ve veri tabanı oluşturma mekanizmaları için yapılan ilk çalışmalar, veri depolama, sorgulama ve işlem yönetimi gibi etkili mekanizmaların geliştirilmesi için bir temel oluşturmuştur. Günümüzde çok sayıda veri tabanı sistemi, sorgulama ve işlem yönetimini varsayılan yetenekleri arasında sunmakta ve sonraki adım olarak gelişmiş veri analizini sistemlerine dahil etmektedir (Han ve diğerleri, 2011, s.2).

1960'lı yıllardan itibaren veri tabanı ve bilgi teknolojisi, ilkel dosyalama sistemlerinden başlayarak sofistike ve çok güçlü veri tabanı sistemlerine kadar düzenli bir gelişme göstermiştir. 1970'lerden 1980'lerin başına kadar veri tabanı yönetim sistemlerindeki araştırma ve geliştirmeler, hiyerarşik ve ağ veri tabanı sistemlerinden ilişkisel veri tabanı sistemlerine, veri modelleme araçlarına, indeksleme ve erişim yöntemlerine kadar ilerlemiştir. Bunlara ilave olarak sorgulama dilleri, kullanıcı arayüzleri, sorgu optimizasyonu ve işlem yönetimi aracılığıyla daha rahat ve esnek bir veri erişiminin elde edilmesi mümkün olmuştur (Han ve diğerleri, 2011, s.3).

1980'lerin ortalarından itibaren veri tabanı yönetim sistemlerinin yaygınlaşmasıyla, gelişmiş veri tabanı sistemleri, veri ambarları, web üzerinden erişilen veri tabanları ve veri analizi kavramları ön plana çıkmaya başlamıştır. Gelişmiş veri analizi kavramı ise 1980'li yılların sonlarından itibaren telaffuz edilmeye başlanmıştır. Son 30 yılda donanım teknolojisinin hızlı ilerleyişi, güçlü ve uygun fiyatlı bilgisayarların, veri toplama ve depolama ortamlarının üretilmesine yol açmış ve bu kavramın popülerliğini arttırmıştır. Veriler, artık birçok farklı türde veri tabanında ve bilgi deposunda saklanabilmektedir (Kelleher ve Tierney, 2018, s.8).

Ön plana çıkan bir veri tabanı yapısı da veri ambarları olmuştur. Veri ambarları, birden fazla heterojen veri kaynağındaki verileri tek bir veri tabanı altında toparlayan ve düzenleyen veri tabanları olarak tanımlanmaktadır. Veri ambarı teknolojisi veri temizleme, veri entegrasyon, özetleme ve toplama (aggregation) gibi eldeki veriyi farklı

açılardan görüntüleme yeteneğine sahip çevrimiçi analitik işlemeyi (online analytical processing-OLAP) bünyesinde barındırmaktadır. Çevrimiçi analitik işlem araçları çok boyutlu veri analizini ve karar vermeyi desteklemesine rağmen, daha detaylı veri analizi için sınıflandırma, kümeleme, aykırı değer ve anormallik tespiti gibi özellikleri içeren veri madenciliği tekniklerine ihtiyaç vardır.

Özet olarak, veri miktarının artması ve güçlü veri analizi araçlarına olan ihtiyaç, veri açısından bolluğun yaşandığı fakat anlamlı bilginin çok az olduğu bir durumu ortaya çıkarmıştır. Büyük ve çok sayıda veri tabanında toplanan ve hızlı bir şekilde büyüyen muazzam miktardaki veri, güçlü veri analizi araçları olmadan insanlar tarafından işlenemeyecek boyutlara ulaşmıştır. Bu durum da veri madenciliğinin gelişmesine sebep olmuştur.

Bu çalışmada veri madenciliği tekniklerinin uygulanacağı alan olan elektronik belge yönetim sistemlerinin tarihsel arka planına kısaca değinmek yerinde olacaktır. Aslında belge yönetiminin tarihini 1980'lerdeki dijital dönüşümün çok daha öncesine götürmek olanaklıdır. "Belge" fikrinin var olduğu günden beri, özellikle onun otantikliğini temin etmek ve süreçleri takip etmek amacıyla bir "belge yönetimi" de hep söz konusu olmuştur. Örneğin, önceki dönemlerde belgeler gelen ve giden defterlerine kaydedilirdi. Belgelerin hangi birimden hangi birime gönderildiği, konusu ve tarihi gibi belge hakkında sadece temel birkaç üst veri tutulurdu.

Günümüzdeki belge yönetim sistemlerinde ise, dijital dönüşümün getirdiği imkanlar aracılığıyla belge hareketlerinin bütün detaylarını saklamak olanaklı hale gelmiştir. Saklanan bu belgeler, belge hareketleri ve üst veriler sayesinde bir birimin zaman içerisindeki tüm belge hareketlerine ulaşmak, belge yöneticilerinin işlem hareketlerini incelemek, sistem kullanıcılarının davranışlarını analiz etmek mümkün hale gelmiştir.

Bu çalışmada sunduğumuz modelde, veri analizi marifetiyle belge üst verileri ve içeriği incelenerek belli bir standart dosya planı dahilinde gruplanabilir; yeni bir belge için potansiyel standart dosya planı numarası tahmin edilebilir/belirlenebilir; sistem kullanıcılarının ve birimlerin zaman içindeki belge hareketleri incelenerek onların davranışları ile ilgili tahminler yapılabilir hale gelmektedir. Modelin kullanılmasını teklif ettiğimiz sistemlerde binlerce belge ve kullanıcının olabileceği düşünüldüğünde bu analizin gözle ve elle yapılamayacağı, otomatik olarak yapılmasının gerektiği açıktır.

1.1. Araştırmanın Konusu

Bu tez çalışmasında, elektronik belge yönetim sistemleri alanında veri madenciliği tekniklerinin ve makine öğrenmesi algoritmalarının kullanımı konu edinilmektedir. Çalışmada; belgelerde bulunan belge sayısı, belge tarihi, referans numarası, başlık, konu, kategori, gizlilik derecesi, üretildiği yer, üretim tarihi, belge sahibi, üretici gibi üst verileri ve belgenin metni veri madenciliği ve makine öğrenmesi algoritmaları kullanılarak incelenecektir. İncelemeler sonucunda da elektronik belge yönetim sistemlerinde kullanıcı kaynaklı hatalar ile saklı örüntüler tespit etmeye ve performansı arttırmaya yönelik tavsiyelerde bulunmaya çalışılacaktır.

Kurum ve kuruluşların tüm belgelerini ve iş süreçlerini barındıran bu sistemlerin ideal şekillerde kullanılmasını temin etmek hem kurum içindeki işlerin daha düzgün bir şekilde yapılmasını hem de belgelerin kullanılması ve elektronik arşive transferinde daha düzgün bir yapı oluşmasını sağlayacaktır. Elektronik belge yönetim sistemlerinde üst veri standardının da yaygınlaşmasıyla veri madenciliği ve makine öğrenmesi teknikleri kullanılarak daha başarılı sonuçlar alınacağı değerlendirilmektedir.

Kurumların gündelik faaliyetlerini kayıt altına alması ve bu bilgileri paydaşları ile paylaşması kurumsal faaliyetlerin ayrılmaz bir parçasıdır. Herkesin, her zaman, her yerden kolaylıkla ulaşabileceği şeffaf, verimli ve sade bir kurum yapısı günümüzde modern ve demokratik kurumların temel hedefi haline gelmiştir (Kandur, 2006, s.15).

Kurumların gündelik işlerini yerine getirirken oluşturdukları her türlü dokümantasyonun içerisinden kurum faaliyetlerinin delili olabilecek belgelerin ayıklanarak bunların içerik, format ve ilişkisel özelliklerini koruyan ve bu belgeleri üretimden nihai tasfiyeye kadar olan süreç içerisinde yöneten sistemlere elektronik belge yönetim sistemleri adı verilmektedir (TS 13298, 2015).

Elektronik belge yönetimi, standartlarla belirlenmiş olsa da kullanılan elektronik belge yönetim sistemi programları kurum ve kuruluşlar arasında farklılık gösterebilmekte, sistemlerin kullanımı esnasında kullanıcı kaynaklı hatalar meydana gelmekte ve EBYS ideal şekilde kullanılamamaktadır. Belgeler standartlara uygun olarak oluşturulmuş EBYS'lerde saklansa da amaçlandığı şekilde kullanılmayan sistemler hem kendi bütünlüklerini koruyamamakta hem de bulunduğu kurum ve kuruluşlardaki diğer yönetim bilgi sistemlerine karar verme noktasında istenilen katkıyı sağlayamamaktadır. EBYS'lerde yapılan hataların tespiti ve performans geliştirmesi gibi iyileştirmeler bu

eksiklikleri en aza indirmeye çalışacak ve EBYS'lerin amaçlandığı doğrultuda kullanılmasını sağlayacaktır.

Diğer yandan, son yıllarda çokça konuşulan ve popüler bir konu olan veri madenciliği ve makine öğrenmesi, içerisinde veri barındıran her sistemde kendisine yer edinmeye ve uygulama konusu olmaya başlamıştır. Veri madenciliği ve makine öğrenmesinin, belge yönetimi alanında da kullanılması söz ettiğimiz bağlamlarda oldukça çeşitli faydalar sağlayabilecektir.

1.2. Araştırmanın Amacı ve Hipotezleri

Bu tez çalışmasının amacı, bilgi ve belge yönetimi alanında özellikle de elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi tekniklerinin kullanımına yönelik bir çalışma ortaya koymak ve bu alanda veri madenciliği ve makine öğrenmesinin konuşulmaya ve uygulamaya geçirilmesini sağlamaktır.

Kamu kurum ve kuruluşlarında elektronik belge yönetim sistemleri doğru bir şekilde kullanılmamakta ve bu da üst veri yönetimini, e-arşiv sistemine geçişi olumsuz etkilemektedir. Veri madenciliği ve makine öğrenmesi tekniklerinin kullanılmadığı elektronik belge yönetim sistemlerinde oluşan kullanıcı hataları ve bunların meydana getirdiği sorunları tespit etmek, ilgili teknikler kullanılarak sistemlerin gözden geçirilmesi sonucu hatalara yönelik düzeltmeler yapmak ve önerilerde bulunmak da bu tezin amaçlarındandır.

Çalışmanın hipotezleri şu şekilde belirlenmiştir:

H1: Elektronik belge yönetim sistemlerinin başarılı bir şekilde çalışabilmesi için veri madenciliği ve makine öğrenmesi teknikleri kullanılabilir.

H2: Kurum ve kuruluşlarda kullanılan elektronik belge yönetim sistemlerinde veri madenciliği ve/veya makine öğrenmesi teknikleri yeteri kadar kullanılmamaktadır.

H3: Elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi teknikleri kullanılması arşiv yönetim sistemlerinin başarısını olumlu etkileyecektir.

Literatürde veri madenciliği ve makine öğrenmesi ile ilgili çok fazla kaynak ve çalışma bulmak mümkündür. Fakat söz konusu yöntemlerin hem bilgi ve belge yönetimi alanında hem de elektronik belge yönetim sistemlerinde kullanımıyla ilgili az sayıda kaynak olması, bu çalışmanın yapılmasının bir diğer amacıdır.

Tez çalışması kapsamında araştırmanın amacına dair temel maddeler aşağıdaki gibi listelenebilir:

- Elektronik ortamda oluşturulan belgelerin iş süreçlerine uygun bir şekilde oluşmasını sağlamak
- Belge oluşturulurken yapılan hataları tespit etmek ve önüne geçmek
- Belgelere yetkisiz erişimleri tespit etmek
- Belge oluşturulurken kullanıcıların dosya planında kaybolması problemine çözüm üretmek
- Seri veya dosya bazında saklama planı oluşturmak yerine belge bazında saklama planı oluşturmayı engellemek
- Gelen belgelerdeki dosyalanma hatalarını tespit etmek
- Tanımlanabilirlik problemi olan belgeleri tespit etmek (Yetersiz üstveri, hatalı dosya planı, vb).

1.3. Araştırma Problemi

E-dönüşüm ve e-devlet çalışmaları neticesinde ortaya çıkan TS 13298 standardı ve bu standarda uyma zorunluluğu getiren 2008/16 sayılı Başbakanlık genelgesi ile kamu kurum ve kuruluşlarında, elektronik belge yönetim sistemi uygulamalarının belirli bir standarda ve çerçeveye uygun olması sağlanmaya çalışılmıştır. Ancak, elektronik belge yönetim sistemlerinde kullanıcı kaynaklı birçok hatanın yapıldığı da görülmektedir.

Çalışmanın problemini şu şekilde ifade etmek mümkündür: "Türkiye'de kamu ve özel kurumlarda kullanılan elektronik belge yönetim sistemleri, TS 13298 standardına uygun şekilde hazırlanmış olsa da günlük kullanımlarında kullanıcı kaynaklı birçok hata yapılmakta ve verimli bir şekilde kullanılmamaktadır."

1.4. Araştırmanın Kapsamı ve Sınırlılıkları

Kurum ve kuruluşların yönetimini ilgilendiren diğer alanlarda olduğu gibi, elektronik belge yönetimi alanında da verilerin analizi ve çıkan sonuçların değerlendirilmesi hem kurum ve kuruluşların yönetimi hem de EBYS'nin standartlar dahilinde kullanılıp kullanılmadığını kontrol etmek açısından oldukça önemlidir. Şüphesiz, tez kapsamında bütün kamu kurum ve kuruluşlarındaki EBYS'lerini incelemek ve veri toplamak mümkün değildir. Bundan dolayı tez kapsamında, gerekli izinlerin alındığı bir devlet üniversitesinin sistemindeki belgeler ele alınmış ve değerlendirilmiştir.

Elbette çıkan sonuçlar sadece incelenen kamu kurum ve kuruluşları için değil, standart dahilindeki tüm EBYS'ler için geçerli olacaktır.

Tez çalışması kapsamında kullanılan bütün veriler anonimleştirilmiş, 6698 sayılı Kişisel Verilerin Korunması Kanunu'nun belirlediği çerçevede kişisel verilerin tamamı kaldırılmış ve bu konuda azami ölçüde hassasiyet gösterilerek hareket edilmiştir.

1.5. Özgün Değer

Ülkemizde e-devlet çalışmaları gün geçtikçe artmakta ve bu doğrultuda birçok alanda çalışmalar yapılmaktadır. Bu çalışma alanlarından bir tanesi de kurum ve kuruluşların gündelik faaliyetleri sonucunda oluşan belgelerin elektronik ortamda üretilmesi, korunması ve kurum ve kuruluşlar arasında paylaşılmasının daha kolay bir şekilde gerçekleşmesini sağlayan Elektronik Belge Yönetim Sistemleri hakkındaki çalışmalardır.

Bilgi kaynaklarının çeşitlenmesi ve yeni teknolojiler sayesinde zamana ve mekâna olan bağımlılığın ortadan kalkması devlet tarafından verilen hizmetlerde yeniliklere yol açmıştır. E-devlet başlığı altında toplanan bu yeni hizmet ve çalışmaların hız kazanmasıyla beraber ülkemizde elektronik belge yönetimi alanında 2007 yılında "TSE 13298 Elektronik Belge Yönetim Sistemi" standardı oluşturulmuş ve bu standart, 2008/16 sayılı Başbakanlık Genelgesiyle kamu kurumlarında elektronik belge yönetim sistemlerinin uygulanmasını zorunlu hale getirmiştir. Bu sayede tüm kamu kurum ve kuruluşlarında ve özel sektörde elektronik belge yönetim sistemleri yaygınlaşmaya başlamıştır.

Büyük miktardaki verilerden anlamlı, faydalı bilgiler keşfetmek için kullanılan veri madenciliği ve makine öğrenmesi teknikleri, günümüzde tüm dünyada olduğu gibi Türkiye'de de yaygınlaşmaya başlamıştır. Hâlihazırda sağlıktan ekonomiye, bankacılıktan pazarlamaya, mühendislikten istihbarata kadar birçok alanda kullanılan veri madenciliği ve makine öğrenmesinin, bu alanla çok yakın bağları olan bilgi ve belge yönetimi alanında konuşulmaya, araştırılmaya ve kullanılmaya başlanması önemlidir. Ortaya koyulan modelin sadece elektronik belge yönetimi alanında değil, e-devlet sürecindeki bilgi yönetim sistemi olarak değerlendirilebilecek diğer projelerde de kullanılmasının özgün fayda sağlayacağı düşünülmektedir.

1.6. Literatür Analizi

Çalışmanın bu kısmında, tez yazım sürecindeki araştırmalara dayanarak literatür ile ilgili genel bir değerlendirmede bulunmak ve bilhassa Türkçede öne çıkan kaynaklardan kısaca söz etmek amaçlanmaktadır. Öncelikle belirtmek gerekir ki veri biliminin bir alt dalı olarak görülebilecek olan veri madenciliği alanı birçok farklı disiplinle hem iç içe geçmiş hem de sürekli etkileşim halinde olduğundan, bizzat bu alanla ilgili literatür kuşatılamayacak ölçüde geniş ve kapsamlıdır. Bu bağlamda bizim çalışmamız elektronik belge yönetim sistemleri ile ilişkisi içinde veri madenciliği ve makine öğrenmesi alanları ile ilgilendiği için literatür araştırması ve analizini de bu özel ilişkiyle sınırlı tutarak hareket edeceğiz.

İlk olarak söylenmesi gereken bir husus da bu alanda yapılan çalışmaların tamamına yakınının yabancı dillerde ve özellikle İngilizcede gerçekleştirilmiş olduğu olgusudur. Türkçede alanla ilgili yapılan çalışmalar son yıllardaki popülaritenin etkisiyle hızla artsa da halen yabancı literatür ile karşılaştırılamayacak bir seviyededir. Burada Türkçe literatürde belge yönetimi ile veri madenciliği ilişkisi bağlamında yazılmış tez ve makaleleri sıralayıp kısaca içeriklerinden bahsetmekle yetineceğiz. Bu çalışmanın konusu olan elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi uygulamalarının ele alındığı doğrudan çalışmalar dilimizde yok denecek kadar azdır.

Türkçe literatürde veri madenciliği alanında karşımıza çıkan ilk kaynak Takçı ve Soğukpınar'ın (2002) kaleme aldığı "Kütüphane Kullanıcılarının Erişim Örüntülerinin Keşfi" makalesidir. Bu çalışmada veri madenciliğinin web verileri üzerinde kullanılmasını konu alan web madenciliği teknikleri sayesinde kütüphane kullanıcılarının kütüphane web sitesi üzerindeki davranışları incelenerek erişim örüntüleri tespit edilmeye çalışılmıştır. Araştırma kapsamında geliştirilen uygulama ve istatistiksel yöntemlerle kullanıcılar hakkında bazı örüntüler bulunmuş ve benzer kullanıcıların kümelenmesi gerçekleştirilmiştir.

Aynı yıl yayınlanan Sever ve Oğuz (2002) "Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorguları ve Algoritmalar" isimli çalışmada veri tabanlarında bilgi keşfi kavramı ve bu süreci oluşturan temel adımlar anlatılmaktadır. Daha sonrasında veri madenciliğinde karşılaşılan problemlere değinilmiş ve veri madenciliğinde kullanılan algoritmalar örneklerle açıklanmaya çalışılmıştır. Ayrıca eşleştirme sorguları kavramı teorik olarak tanımlanmış ve eşleştirme algoritmalarından bahsedilmiştir.

Takip eden yılda yine aynı yazarlar tarafından kaleme alınan Sever ve Oğuz (2003) "Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım: Kısım II- Eşleştirme Sorgularının Biçimsel Kavram Analizi ile Modellenmesi", yukarıda bahsedilen makalenin bir devamı niteliğinde görülebilir. Bu çalışmada ise pazar sepeti analizi örneği üzerinde eşleştirme sorgularının detayları anlatılmış ve eşleştirme kuralı çıkarımı ile biçimsel kavram analizi arasında bağ kurabilecek bir model önerisinde bulunulmuştur.

Arslantekin (2003) "Veri Madenciliği ve Bilgi Merkezleri" adlı makalesinde veri madenciliğinin genişleyen kullanımını göz önünde bulundurarak bilgi merkezlerindeki süreçlerde nasıl uygulanabileceği hakkında değerlendirmelerde ve kullanım örnekleri vererek önerilerde bulunmuştur.

Türkçe literatürde veri madenciliği hakkında yapılmış yukarıda bahsettiğimiz bu ilk çalışmalardan sonra yakın dönemde Türk Kütüphaneciliği dergisinde yayınlanmış bir makale olan Çalışkan ve Can (2018) "Türkçe Metinler Üzerine Yapılan Sayısal Üslup Araştırmalarını İnceleyen ve Benim Adım Kırmızı Çevirilerinin Aslına Olan Sadakatini Ölçen Bir Çalışma"da sayısal üslup analizi yönteminin tanıtılması amacıyla çeviri eserlerin aslına ne kadar sadık olduğunu ölçmeye çalışan bir uygulama ortaya konulmuştur. Uygulama kapsamında Orhan Pamuk'un "Benim Adım Kırmızı" adlı romanı ve çevirilerindeki üslup uyumu araştırılmış ve çevirilerde istatiksel olarak yazar üslubuna sadık kalındığı gözlemlenmiştir. Ayrıca Türkçe metinlerde üslup analizinin nasıl uygulanabileceği tartışılmış ve Türkçe eserleri konu alan çalışmaları derleyen kapsamlı bir kaynak taraması yapılmıştır.

Bilgi Yönetimi dergisinde yayınlanan Aktan (2018) "Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu" isimli çalışmada dünyadaki veri hacminin ve çeşitliliğinin hızla artmasıyla ortaya çıkan "Büyük Veri" kavramı ele alınmış, uygulama alanları ve sağladığı avantajlar hakkında literatür incelemesi yapılmıştır. Ayrıca sunduğu avantajlara karşın kişisel bilgilerin mahremiyeti konusunda sorun teşkil edebilecek güvenlik açıklarına neden olabileceğine de dikkat çekilmiştir.

Yine aynı isimli dergide bir sonraki yıl yayınlanan iki makale dikkat çekmektedir. Bu makalelerden ilki olan Yıldırım ve Özdemirci (2019) "Kurumlarda Örtük Bilginin Yapay Zekâ Destekli Tavsiye Sistemleri Aracılığıyla Ortaya Çıkarılması"nda günümüzde kurumların rekabetçi bir ortamda var olabilmeleri için bilgiye olan ihtiyaçlarının arttığına dikkat çekilmiş ve kurumların başarısının örtük bilgilerin tespiti, ortaya çıkarılması ve bu bilgilerden yararlanılmasıyla bağlantılı olduğu belirtilmiştir. Yapay zekâ destekli tavsiye sistemlerinin kurum içindeki örtük bilgileri ortaya çıkarma noktasında yardımcı

olabileceği, bu sayede kurum performansı ve kalitesinin artabileceği ifade edilmiştir. Ayrıca çalışma kapsamında oluşturulan "Profil Öğrenme Modeli"nin kurumlardaki örtük bilginin keşfedilmesindeki önemine değinilmiştir.

Söz ettiğimiz iki makaleden diğeri ise Cibaroğlu ve Yalçınkaya (2019) "Belge ve Arşiv Yönetimi Süreçlerinde Büyük Veri Analitiği ve Yapay Zekâ Uygulamaları"dır. Bu çalışmada, yapay zekâ ve büyük veri kavramlarının birçok alanda kullanılmaya başlandığı ve etkili olduğu, bu alanlardan birinin de belge ve arşiv yönetimi alanı olduğu ifade edilmektedir. Bu teknolojilerin kullanılmasıyla belge ve arşiv süreçlerinin otomatikleştirildiği ve hızlandığından bahsedilmiştir. Yapay zekâ ve büyük veri teknolojilerinin belge ve arşiv yönetimi alanında uygulanabilirliği konusunda birçok uluslararası çalışmanın yapıldığı ancak Türkiye'de bu alanda yapılan çalışmaların azlığı vurgulanmıştır. Ayrıca bahsedilen teknolojilerin belge ve arşiv yönetimi süreçlerinde nasıl uygulanabileceğine dair değerlendirmelerde bulunulmuştur.

Türk Kütüphaneciliği dergisinde Çakmak ve Eroğlu (2020) tarafından kaleme alınan "Sosyal Medyada Kullanıcı Etkileşimi ve İçerik Kategorizasyonu: Ankara'daki Halk Kütüphanelerinin Facebook Gönderilerinin Analizi" adlı makalede, Ankara'daki halk kütüphanelerinin Facebook platformunda yayınladıkları içerikler ve kullanıcı etkileşimlerinden oluşturulan bir veri seti üzerinde metin madenciliği algoritmalarıyla veri analizi yapılmıştır. Yapılan analizler sonucunda halk kütüphanelerinin Facebook sayfalarını aktif olarak kullandığı ve genellikle çocuklar ile ilk/ortaöğretim düzeyindeki kullanıcılara yönelik etkinliklerin ön planda olduğu belirlenmiştir. Ancak koleksiyona yönelik gönderi sayısının daha az olduğu tespit edilmiştir. Sonuçta kütüphanelerin etkileşim sayılarını artırmak ve hedef kitlelere yönelik içerikler üretmek için sosyal medya stratejileri geliştirmeleri gerektiği önerilmektedir.

Çakmak ve Tırnavalı (2020) "Danışma Hizmetlerinin Değerlendirilmesi: Türkiye Büyük Millet Meclisi Kütüphanesi Danışma Hizmetlerine Yönelik Bir Araştırma" isimli makale Bilgi Dünyası'nda yayınlanmıştır. Çalışmada Türkiye Büyük Millet Meclisi (TBMM) Kütüphanesi'nde sunulan danışma hizmetlerinin değerlendirilmesi ve iyileştirilmesi amacıyla kullanıcı talepleri ile bu taleplere verilen cevaplar kullanılarak bir veri analizi çalışması gerçekleştirilmiştir. Yapılan analiz sonucunda, kullanıcı taleplerinin genellikle kitap türü bilgi kaynaklarına yönelik olduğu, taleplerin çoğunun e-posta ile yanıtlandığı ve büyük bir kısmının bir saat içinde yanıtlandığı tespit edilmiştir. Çalışmanın sonunda TBMM Kütüphanesi danışma hizmetlerinin etkinliğini artırmak için iyileştirme önerileri sunulmuştur.

Ankaralı ve Külcü (2020) "RapidMiner ile Twitter Verilerinin Konu Modellemesi" isimli çalışmada Twitter'da kullanıcılar tarafından belirlenen kelimeleri içeren tweet verileri elde edilerek, bu veriler üzerinde metin ve veri madenciliği teknikleriyle bir analiz çalışması yapılmıştır. Öncelikle tweetlerde sık kullanılan kelimeler belirlenip kullanım sıklıklarına göre kelime gruplarının oluşturulduğu çalışmada daha sonra Latent Dirichlet Allocation (LDA) algoritması kullanılarak tweetlerin konu modellemesi yapılmıştır.

Öztürk (2021) tarafından kaleme alınan "Arşivler ve Yapay Zekâ" adlı makalede, bilgi teknolojilerindeki gelişmelerin ve artan belge hacminin arşivlerin yönetimini zorlaştırdığı vurgulanıp yapay zekâ uygulamalarıyla bu yönetim sorunlarının üstesinden gelinebileceği ifade edilmiştir. Yapay zekânın arşivcilik alanında örnek kullanımları sunulmuş ve mesleki değişim ve gelişmeler açısından önemli bir rol oynayacağı belirtilmiştir.

Öztürk ve Özel (2021)'in Bilgi Dünyası'nda yazdığı "Yapay Zekâ ve Kütüphaneler" isimli makalede, kütüphanelerin hizmet anlayışını ve kurum yapısını değiştirme potansiyeline sahip olduğu düşünülen yapay zekânın kütüphanelerdeki kullanımı hakkında literatür incelemesi gerçekleştirilmiştir. Ayrıca Ankara'da bulunan üniversite kütüphanelerinde çalışan kütüphanecilerle bir anket çalışması yapılmış ve kütüphanecilerin yapay zekâ teknolojisi ve uygulamalarına yönelik algıları, farkındalıkları ve beklentileri incelenmiştir. Anket sonuçları yorumlanarak farkındalığın oluşması için bazı önerilerde bulunulmuştur.

Bilgi ve Belge Araştırmaları dergisinde Kaya (2021) tarafından "Bir Araştırma Kaynağı Olarak Arşivlenen Sosyal Medya Verilerinin Kullanımı" isimli bir makale yayınlanmış ve makalede yeni bir veri kaynağı türü olarak ifade edilmeye başlanan sosyal medya verileri ve kullanım alanları üzerine odaklanılmıştır. Bilgi ve belge yönetimi alanında da fark edilmeye başlanan bu yeni veri kaynağı türüyle ilgili literatür araştırması gerçekleştirilmiş ve kullanım alanları hakkında değerlendirmeler yapılmıştır.

Bilgi Yönetimi dergisinde Efe (2022) tarafından kaleme alınan "Yargısal ve Hukuki Süreçlerde Yapay Zekâ Kullanan Araçlar Üzerine Bir Araştırma" adlı makalede, hukuk alanında yapay zekâ temelli uygulamaların kullanılmaya ve süreçleri değiştirmeye başladığı ifade edilerek kullanım örnekleri verilmiştir. Ayrıca alanda yapay zekâ tabanlı yönetim bilişim sistemlerinin kullanımı ve gelecekteki beklentiler hakkında değerlendirmelerde bulunulmuştur.

Aynı dergide sosyal medya platformlarının sahip olduğu büyük veriyi nasıl elde ettikleri ve hangi amaçlarla kullandıkları konusunda yapılmış çalışmaların incelendiği Eker (2022) "Sosyal Ağlar Büyük Veriden Nasıl Yararlanır: Facebook ve Twitter"da bahsedilen platformların büyük veri sayesinde kullanım amaçları sıralanmıştır.

Bilgi ve Belge Yönetimi alanında yayınlanmış makalelerin ardından, Türkiye'de Bilgi ve Belge Yönetimi bölümlerinde veri madenciliği ve makine öğrenmesi konularında yapılmış 8 adet lisansüstü çalışmayı ele alacağız. Bunlardan 2 tanesi yüksek lisans, 6 tanesi doktora seviyesinde kaleme alınmış tez çalışmalarıdır.

İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde "Kütüphane ve Bilgi Bilimi Çalışmalarında Dönemsel Konu Analizi" isimli doktora tezi 2016 yılında Binici tarafından kaleme alınmıştır. Çalışmada kütüphane ve bilgi bilimi alanında kaleme alınmış bilimsel yayınların başlık, öz ve anahtar kelimelerinden oluşturulan veri seti üzerinde metin madenciliği, kelime birlikteliği ve ağ analizi tekniklerinin kullanıldığı bir analiz çalışması yapılmıştır. Çalışmanın amacı bahsedilen alanlardaki yayınların temalarının belirlenmesi ve dönemsel olarak ön plana çıkan konuların tespit edilmesidir. Analiz sonucunun daha anlaşılabilir olması amacıyla yayın temaları ve dönemlik konu yönelimleri metin görselleştirme araçları kullanılarak gösterilmiştir. Çalışmanın sonucunda, kütüphane ve bilgi bilimindeki tematik yapının 12 ana konu ve bunların alt konularını kapsayacak içerikte olduğu gözlemlenmiştir.

Ele alacağımız ikinci doktora çalışması 2017 yılında Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde Taşkın tarafından hazırlanan "İçerik Tabanlı Atıf Analizi Modeli Tasarımı: Türkçe Atıflar İçin Metin Kategorizasyonuna Dayalı Bir Uygulama" isimli tezdir. Atıf analizini konu alan çalışmada, Türkçe atıflara yönelik içerik tabanlı bir analiz modeli tasarlanmıştır. Araştırmada kütüphanecilik ve bilgi bilim alanında yayınlanmış makalelerin üst veri, kaynakça ve tam metinleri toplanarak bir veri seti meydana getirilmiştir. Makalelere verilen atıflar dört taksonomik kategoride gruplanmış ve her kategori de kendi içerisinde alt kategorilere ayrılmıştır. Daha sonra her bir atıf kategorisi için sınıflandırma algoritmalarıyla gerçekleştirilen sınıflandırmanın başarıları değerlendirilmiş ve çalışmanın sonunda atıfların nasıl değerlendirilmesi gerektiğini ortaya koyan içerik tabanlı bir atıf analizi modeli oluşturulmuştur.

"Kurumsal Belgelere (Metin Verilerine) Metin Madenciliği Tekniği ile Erişimin Değerlendirilmesi: Türk Özel Sektörüne Yönelik Bir İnceleme" isimli Hamde (2018) tarafından İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde

kaleme alınan doktora tezinde, kurumların bilgi yönetim sistemlerinde gerçekleştirilen metin madenciliği uygulamaları hakkında genel bir çerçeve çizilmiş ve kurumlarda bu uygulamaların kullanıldığı sistemler belirlenmiştir. Metin madenciliği uygulamaları sayesinde şirketlerin rakipleri hakkındaki bilgileri otomatik bir şekilde elde etmesinin karar verme ve rekabet etme konusunda yardımcı olacağı ifade edilmiştir. Ayrıca kurumsal belge yönetim sistemlerinde metin madenciliği uygulamaları hakkında bilgi edinmek amacıyla Türkiye'de özel sektörde çalışan bir grup üzerinde anket çalışması gerçekleştirilmiş ve sonuçlar değerlendirilmiştir. Sonuçta, metin madenciliği sayesinde özel sektördeki şirketlerin performanslarını nasıl artıracağına ve belgelere erişimin nasıl geliştirilebileceğine yönelik tespit ve değerlendirmelerde bulunulmuştur.

Çelik (2021)'in Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde hazırladığı "Türkçe Akademik Yayınlar İçin Yapısal Öz Çıkarım Sistemi" isimli doktora tezinde, makale özetlerinin, makale metinlerinin yapısal içeriğini kapsamlı biçimde içermediği tespitinden hareketle, araştırmacı/okurun aradığı makaleleri bulmasına yardımcı olabilmenin yolları araştırılmıştır. Bu kapsamda, bilimsel makale özlerinin metin madenciliğinin alt alanlarından biri olan metin özetleme teknikleriyle yapısal özlerini çıkarabilecek bir özetleme sistemi geliştirilmiştir. Türkçe kütüphanecilik ve bilgi bilim literatüründe yayınlanan makalelerden oluşan veri seti üzerinde yapılan denemeler sonucunda otomatik özetleme sistemiyle meydana getirilmiş özetlerin klasik özetlere göre daha başarılı olduğu gözlemlenmiştir. Geliştirilen özetleme sisteminin sadece yukarıda bahsedilen veri seti üzerinde değil, diğer makale araştırmalarında da kullanılabileceği ifade edilmiştir.

Ankara Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde Doğan (2022) tarafından kaleme alınan "Bilgi Merkezi ve Hizmetlerinde Veri Madenciliğinin Kullanılabilirliği: Üniversite Kütüphaneleri Örneği" isimli doktora tezi, üniversite kütüphane sistemlerinde tutulan kullanıcı verilerinin veri madenciliği teknikleri aracılığıyla analiz edilmesi ve sonuçların kütüphane hizmetleri ve işlemlerinin iyileştirilmesinde kullanımının incelenmesini amaçlamaktadır. Veri toplama aşamasıyla ilgili yapılan ilk çalışmalar neticesinde, üniversite kütüphane sistemlerinde tutulan verilerin ileride yapılabilecek veri madenciliği çalışmaları için uygun olmadığı tespit edilmiştir. Bunun üzerine hipotetik veri üretme işlemleriyle, üniversite kütüphane kullanıcılarının davranışlarını simüle edecek bir yönteme başvurulmuş ve kullanıcı verileri rastgele olacak şekilde üretilmiştir. Üretilen bu veri seti üzerinde uygulanan veri

madenciliği teknikleriyle kütüphane işlem ve hizmetlerinin iyileştirilmesi için karar destek süreçlerinde kullanılabilecek sonuçlar ve bir model elde edilmiştir.

Ele alacağımız son doktora bu yıl Ankara Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde Kurt (2023) tarafından hazırlanan "Bilgi Yönetiminde Veri ve Metin Madenciliği: Bir Dijital İçerik Analizi Uygulaması" adlı çalışmadır. Kripto paralar ve blokzincir hakkında dijital içerikler üreten bir platformun sitesinden elde edilen verilerin işlenmesini konu alan çalışmada veri ve metin madenciliği teknikleri kullanılarak iki farklı analiz gerçekleştirilmiştir. Birinci analizde, platformun ürettiği içeriklerin okunma sayılarını etkileyen hususlar, ikinci analizde ise okunma sayıları ve dijital içerik platformunun gelirleri arasında nasıl bir ilişki olduğu tespit edilmeye çalışılmıştır. Her iki analizin sonucunda da uygulanan veri ve metin madenciliği tekniklerinin başarıları istatiksel olarak değerlendirilmiş ve sonuçların istatistik bakımından geçerli olduğu ifade edilmiştir. Ayrıca çalışmada, bilgi ve belge yönetimi alanında veri madenciliği sürecinin başından sonuna kadar uygulandığı bir örnek uygulama sunulduğu belirtilmiştir.

Türkiye'de Bilgi ve Belge Yönetimi Bölümü altında veri madenciliği ve makine öğrenmesi alanında yüksek lisans düzeyinde iki çalışmayı ele alacağız. Bunlardan ilki İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi'nde Akçay (2014) tarafından hazırlanmış "Bilgi ve Belge Yönetiminde Veri Madenciliği" isimli yüksek lisans tezidir. Çalışmada öncelikle veri madenciliği kavram, yöntem ve teknikleri hakkında temel teorik bilgi verildikten sonra tüm alanlarda kullanılmaya başlayan veri madenciliğinin bilgi ve belge yönetimi alanında edineceği yer, önemi ve kullanım alanlarından bahsedilmiştir.

Ele alacağımız ikinci yüksek lisans tezi ise Doğan (2015) tarafından Ankara Üniversitesi Sosyal Bilimler Enstitüsü Bilgi ve Belge Yönetimi Bölümü'nde hazırlanan "Büyük Verinin Üniversite Kütüphaneleri Açısından Önemi ve Kütüphane Hizmetlerinde Kullanımı: Ankara Üniversitesi Kütüphaneleri Örneği" adlı çalışmadır. Tezde, birçok alanda dönüşüme neden olan büyük verinin kütüphanelerde sağlayacağı dönüşüm, dermelerinin içinde büyük miktarda verinin bulunması ile kütüphane işlem ve hizmetlerinin geliştirilmesi için kullanılabilirliği olmak üzere iki başlıkta incelenmiştir. Araştırma kapsamında büyük veriye yönelik farkındalığı tespit etmek amacıyla bilgi ve belge yönetimi mezunu olan kütüphane çalışanları ile iki anketten oluşan bir çalışma yapılmıştır. İlk olarak yapılan anketle büyük veri kavramı hakkındaki farkındalık tespit edilmiş, sonrasında kütüphanecilere büyük veri hakkında bir sunum yapılmıştır.

Sunumdan sonra katılımcılara tekrar anket soruları yöneltilerek bu kavram hakkında farkındalığın oluşup oluşmadığı bulunmaya çalışılmıştır. Anket sonucunda katılımcıların kavram hakkındaki farkındalığında anlamlı bir artış gözlenmiştir.

10-11 Ekim 2019 tarihleri arasında *"Endüstri 4.0 Sürecinde Bilgi Yönetimi ve Bilgi Güvenliği: eBelge-eArşiv-eDevlet-Bulut Bilişim-Büyük Veri-Yapay Zekâ"* temalı gerçekleştirilen 4. e-BEYAS sempozyumunda daha önceki sempozyumlara oranla çok daha fazla sayıda "Büyük Veri" ve "Yapay Zekâ" konulu bildiri sunulmuştur. Bu durum da alanımızda yeni teknolojilerin kullanılmasının kaçınılmaz bir gereksinim olmaya başladığını göstermektedir.

EBYS ve yapay zekâ algoritmalarının kullanımından bahseden bildiriler arasından iki tanesi ön plana çıkmaktadır. Bildirilerden ilki (Şimşek ve diğerleri, 2019), Cumhurbaşkanlığı Bilgi ve Belge Yönetimi Daire Başkanlığında, EBYS üzerindeki verilerden faydalanarak ve kural tabanlı algoritmaları kullanılarak yapılan bir çalışma hakkındadır. Kurumsal uygulamalar içerisinde en çok kullanılanlardan birisi olan EBYS uygulaması içinde oluşturulan bir belgenin üst verisi ve içeriği, yapay zekâ algoritmalarıyla eğitilmiş ve bu eğitim sonucuna göre standart dosya planı (SDP) tahmini yapılmış veya diğer bazı alanlar tahmin edilmeye çalışılmıştır. Böylece hem kullanıcı kaynaklı hataların en aza indirgenmesi hem de bilgiye erişim safhasında standart dosya planlarından maksimum seviyede yararlanılması sağlanmış olmaktadır. Bu sayede, bilginin yönetilebilmesiyle kurumsal faaliyetlerde geleceğe dair öngörüler elde edilebileceği, kurum ve kuruluşların bu öngörüleri kullanarak tespit ve değerlendirme imkânına sahip olacağı için kurumsal hedeflerine (Yıllık ve Stratejik Planlar ve Kalkınma Planları vb.) dair başarı oranının yükseleceği belirtilmiştir.

Bir diğer bildiride (Binici, 2019) ise makine öğrenmesi algoritmalarıyla belgelere standart dosya planlarının otomatik olarak atanması üzerine bir çalışma yapılmıştır. Bu çalışmada, öncelikle belli sayıda belgeyi içeren bir veri seti oluşturulmuş ve bu veri setinden her konudan orantılı olmak koşuluyla rastgele yöntemlerle belgelerin üçte biri (1/3) sınıflandırmak için seçilmiştir. Sınıflandırılmak üzere seçilen belgeler üzerinde, son zamanlarda popüler olan ve yoğun biçimde kullanılmaya başlanan "Destek Vektör Makinesi (Support Vector Machine-SVM)" algoritması çalıştırılmıştır. Çalışma sonucunda manuel olarak yapılan sınıflama ile otomatik olarak yapılan çıkarımın isabet oranı %87.72 olarak bulunmuştur. Bir diğer ifadeyle, belgelerin %87.72'si makine öğrenmesi algoritmasıyla doğru olarak sınıflandırılmıştır.

1.7. Kaynaklar

Araştırma ve tez yazım aşamalarında matbu ve elektronik birçok farklı kaynaktan çalışma boyunca yararlanılmıştır. Tezin teorik altyapısını teşkil eden başlıca kaynakların daha çok açık veri tabanları, alanda öne çıkmış İngilizce temel kitap ve makaleler olduğu dile getirilebilir. Çalışmamızın temel ilgisi bağlamında ülkemizde ve kendi dilimizde yazılmış çok fazla kaynak bulunmamaktadır. Bu bağlamda, kullanılan elektronik kaynaklar aşağıda sıralanmıştır.

DergiPark
Directory of Open Access Journals (DOAJ)
Ebrary
EBSCOHost
Emerald
Google Scholar
IEEE Xplore
Journal Storage (JStor)
Library, Information Science & Technology Abstracts (LISTA)
ProQuest
Sage Journals Online
Science Direct
Scopus
Springer Link
Taylor & Francis
ULAKBİM - Mühendislik ve Temel Bilimler Veri Tabanı
ULAKBİM - Sosyal Bilimler Veri Tabanı
ULAKBİM - TR Dizin
Web of Science
Wiley Online Library
YÖK Tez Kataloğu

Çalışmada arama motorlarında yaygın olarak kullanılan anahtar kelimeler aşağıda verilmiştir.

Data mining	Veri madenciliği
Text mining	Metin madenciliği

Machine learning	Makine öğrenmesi
Records management	Belge yönetimi
Electronic records management	Elektronik belge yönetimi
Electronic records management systems	Elektronik belge yönetim sistemleri
Metadata management	Üst veri yönetimi
Crisp-dm	Crisp-dm
Classification algorithms	Sınıflandırma algoritmaları
Clustering algorithms	Kümeleme algoritmaları
Association rule analysis	Birliktelik kuralı analizi
Evaluation metrics	Değerlendirme metrikleri/ölçütleri
Data visualization	Veri görselleştirme
Artificial intelligence	Yapay zeka
Open data	Açık veri

Tez yazımında Marmara Üniversitesi Türkiyat Araştırmaları Enstitüsünün "Tez Yazım Kılavuzu"na müracaat edilmiştir.[1] Bununla birlikte tez yazımında EndNote referans yönetimi programı kullanılmıştır.

[1] Marmara Üniversitesi Türkiyat Araştırmaları Enstitüsü Tez Yazım Kılavuzu'na https://turkiyat.marmara.edu.tr/dosya/tae/ogrenci/2022%20Tez%20Yaz%C4%B1m%20K%C4%B1lavuzu.pdf adresinden erişilmiştir.

İKİNCİ BÖLÜM

2. VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ

Veri bilimi alanında çalışanların pek çok kez duyduğu klişelerden biri de Russell L. Ackoff'un meşhur bilgi piramidinden (The Knowledge Pyramide) ilhamla oluşturulmuş "From Data to Wisdom" yani "Veriden Bilgeliğe" sloganıdır (Ackoff, 1989). Her ne kadar "anlayış (understanding)" ile birlikte Ackoff'un makalesindeki versiyonuyla piramit beş dilimden oluşsa da ardından gelenler "anlayış"ı göz ardı etmiş ve aşağıdan yukarıya sırasıyla şu kavramları içerecek şekilde piramidi yorumlamıştır: veri, enformasyon, bilgi ve bilgelik. Piramidin en altında yer alan veriden en üstünde yer alan bilgeliğe yolculuk aslında veri biliminin temel amacı olarak değerlendirilebilir. Çalışmamızın da veri bilimi başlığı altında yer aldığını düşündüğümüzde söz konusu piramitteki kavramları ele almak ve bunların birbiriyle ilişkisini göstermek, sunacağımız modelin altında yatan teorik zemini ve temel dinamikleri göstermesi bakımından yerinde olacaktır.

Veri

Piramidin en alt basamağında yer alan "veri" (data) kavramının karşıladığı nesneler düşünüldüğünde, söz konusu kavramı tanımlamanın ne kadar güç olduğu ortadadır. Veri bilimi özelinde de birçok farklı şekilde tanımlanan terimle ilgili taramamızı bu konuda oldukça kapsamlı bir makale ortaya koymuş olan Rowley (2007) üzerinden gerçekleştireceğiz. Genel itibarıyla veriler nesnelerin, olayların ve çevrelerinin özelliklerini temsil eden sembollerdir. Ackoff'un benzetmesine başvuracak olursak bu veriler metal cevherleri gibi, işe yarar biçimde işlenmedikçe hiçbir değer taşımazlar. Bu yüzden veri ile enformasyon arasındaki ilişki yapısal değil, işlevseldir. Yani tıpkı bir maden gibi veri de işlendiğinde ve enformasyona dönüştüğünde miktar olarak azalsa da değer bakımından artmaktadır (Ackoff, 1989). Bu durumun piramidin zemininden zirvesine çıkarken söz konusu olan her aşama için geçerli olduğu söylenebilir.

Rowley sözünü ettiğimiz makalesinde veriyi tanımlamaya çalışan birçok örnek sunar. "*Bir bağlamı olmadığı ve yorumlanmadığı müddetçe veri bir anlama sahip değildir.*", "*Veriler, organize edilmemiş ve işlenmemiş olan ve belirli bir anlam ifade etmeyen ayrık, nesnel gerçekler veya gözlemlerdir.*" ya da "*Veri öğelerin, şeylerin,*

olayların, faaliyetlerin ve işlemlerin temel ve kayıtlı bir açıklamasıdır." gibi veriyi tanımlama girişimlerine yönelik Rowley'nin tespiti kayda değerdir. Tüm bu açıklamalar şaşırtıcı bir biçimde verinin yoksun olduğu şeylerle ilgilidir; örneğin, veriler anlam veya değerden yoksundur, düzenlenmemiş ve işlenmemiştir. Dolayısıyla bizzat veriyi tanımlamak yerine enformasyonda olup veride olmayan özellikleri sıralayarak bir tanım yapmaya çalışırlar (Rowley, 2007).

Enformasyon

Verinin tanımlanmasında enformasyona bağlı olarak açıklamada bulunma girişiminin kaçınılmaz bir sonucu, enformasyonun da daima veriyle ilişki olarak tanımlanmasıdır. Bu tanımlara örnek olarak Rowley (2007) şunları sıralar: "*Enformasyon biçimlendirilmiş veridir ve gerçekliğin temsili olarak tanımlanabilir.*", "*Enformasyon, bir konunun anlaşılmasına değer katan veridir*.", "*Enformasyon, alıcı için anlam ve değere sahip olacak şekilde düzenlenmiş verilerdir.*" ve "*Enformasyon, bir amaç doğrultusunda işlenen veridir*."

Dikkat edilecek olursa tüm bu açıklamalar enformasyonu bir veri olarak tanımlar ve ona çeşitli özellikler atfeder. "Anlamlı", "bir amacı olan", "bir bağlamı olan" veriler, enformasyon özelliği kazanmış olur. Sonuç olarak hem bilgi sistemleri ders kitaplarında hem de bilgi yönetimi literatüründe enformasyon, veri olarak tanımlanır ve "organize" veya "yapılandırılmış" veri olarak değerlendirilir. Enformasyon varlığını belirli bir amaç ya da bağlamla verinin ilişkilendirilmesine borçludur ve böylece anlamlı, değerli, faydalı ve alakalı hale gelir (Rowley, 2007).

Bilgi

Veri ve enformasyonun tanımlanmalarında ve ayrıca piramit hiyerarşisindeki ilişki biçiminde gördüğümüz unsurlar bilginin tanımlanmasında da söz konusudur. Tarih boyunca birçok farklı disiplinden çok sayıda ismin üzerinde tartıştığı "bilgi" kavramının tanımlanması ve anlaşılması oldukça zor bir terim olduğu konusunda bir fikir birliğinden söz edilebilir. Bu bağlamda Rowley'nin sunduğu örneklerden konumuzla doğrudan ilgili olan birkaçını inceleyebiliriz. Örneğin ilk olarak "*Bilgi, karar almaya yardımcı olmak için kullanılabilecek değerli bir varlık elde etmek için uzman görüşü, becerileri ve deneyiminin eklendiği veri ve enformasyonun bir birleşimidir.*" tanımına bakacak olursak bilginin, veri ve enformasyona "eklenen" başka nitelikler ile açıklanmaya çalışıldığını görürüz.

Fakat "*Bilgi, verilerden çıkarılan enformasyon üzerine inşa edilir. Veri nesnelerin bir özelliğiyken, bilgi, onları belirli bir şekilde iş görmeye uygun hale getiren insanın bir özelliğidir.*" tanımında enformasyonun "dönüştürülmesine" odaklanılmaktadır. Benzer şekilde "*Bilgi, mevcut bir soruna veya faaliyete uygulandıkları şekliyle anlayış, deneyim, birikmiş öğrenme ve uzmanlığı aktarmak için düzenlenmiş ve işlenmiş veri ve/veya enformasyonlardır.*" tanımında da veri ve enformasyonun "düzenlenmesi" ve "işlenmesi"ne dikkat çekilmektedir. Sonuç olarak bu tanımları özetlersek her ne kadar iki farklı yaklaşım biçimi öne çıksa da bilgi; enformasyon, anlayış, yetenek, deneyim, beceri ve değerlerin bir karışımı olarak ifade edilebilir.

Bilgelik

İnsanlık tarihi kadar eski olan "bilgelik" kavramı ve ona yönelik arayışa ilişkin açıklamaları bir kenara bırakıp bilgi yönetimi bağlamında bu terimi ele alanlara göz atacak olursak şu ana kadar değerlendirdiğimiz piramidin diğer kavramlarına nispetle üzerine daha az düşünülmüş olduğunu görebiliriz. Burada bilgi yönetimi literatüründe öne çıkan tanımlardan biri olarak Awad ve Ghaziri'nin "*Bilgelik, vizyon öngörüsü ve ufkun ötesini görme yeteneği ile soyutlamanın en yüksek seviyesidir*" ifadesini anabiliriz. Bununla birlikte piramitteki diğer unsurlarla ilişkisi bakımından "entegre bilgi", "değer katılmış bilgi" olarak değerlendirilen bilgelik esasen "neden"in bilgisidir, denebilir.

Veri Madenciliği

Disiplinler arası bir çalışma alanı olan "veri madenciliği" kavramı birçok farklı şekilde tanımlanmaktadır:

Veriden bilgi keşfi (**K**nowledge **d**iscovery from **d**ata (**KDD**)) olarak da isimlendirilen veri madenciliği, büyük miktardaki veri yığınları içerisinden saklı, ilginç örüntüleri bulma ve bilgi çıkarma işlemi olarak tanımlanabilir (Han ve diğerleri, 2011, s. 8). Kelleher ve Tierney (2019, s.11-15)'a göre veri madenciliği, hacimli veri kümelerinden kolayca saptanamayacak faydalı örüntüler çıkarma sürecidir ve genel olarak yapılandırılmış (yapısal) verinin analiziyle ilgilenmektedir. Kantardzic (2019, s.2)'e göre veri madenciliği, büyük hacimli verilerde yeni, değerli ve önemli bilgilerin aranmasını sağlayan döngüsel (recursive) bir süreçtir. Bir başka tanımda ise verilerden daha önceden bilinmeyen, potansiyel olarak anlamlı ve yararlı bilgilerin otomatik veya genellikle yarı otomatik işlemlerle çıkarılması şeklinde ifade edilmektedir (Witten, Frank ve Hall, 2011).

Tanımları bu şekilde çoğaltmak mümkündür. Tanımlar incelediğinde şu ortak noktalar göze çarpmaktadır; büyük miktarda veri olması, veri içerisinde daha önceden fark edilmemiş örüntüler ile davranış kalıpları olması ve potansiyel olarak anlamlı ve fayda sağlayabilecek bilgi(ler) çıkarma süreci. Bu ortak noktalar göz önüne alındığında veri madenciliği, büyük hacimli veri yığınlarından daha önceden keşfedilmemiş saklı örüntüleri bulma ve faydalı bilgi(ler) çıkarma işlemi olarak tanımlanabilir.

2.1. VERİ MADENCİLİĞİ

2.1.1. Veri Madenciliği Kavramı ve Tarihsel Gelişimi

Veri biliminin önemli alt dallarından biri olan veri madenciliğinin mevcut duruma gelene dek geçirdiği gelişim ve dönüşümlerden kısaca söz etmek yerinde olacaktır. Veri madenciliğinin tarihsel gelişim sürecindeki kayda değer aşamaları ve dönüm noktalarını tespit ederek başlayabiliriz. Söz konusu bağlamda bu gelişimi dört ayrı dönem ve başlık altında incelemek mümkündür.

İlk olarak 20. yüzyılın başlarından itibaren istatistik biliminin müstakil bir hale gelerek geliştirdiği atılımı, veri madenciliğinin kökenleri olarak belirleyebiliriz. Bu yıllardan itibaren istatistikçiler daha sonra veri madenciliğinde tahmine dayalı modelleme halini alacak olan uygulamaların ilk örneklerini vermiştir. Söz konusu çalışmalar esnasında toplanan verilerin, anlamlı bilgiler çıkarmak için çok önemli bir kaynak olduğunun keşfedilmesiyle birlikte bu verilere dayalı tahminler yapmak için tekniklerin gelişiminin önü açılmıştır. Bununla birlikte, yüzyılın ortalarında bilgisayarların tarih sahnesine çıkmasıyla büyük boyutlardaki veriyi işlevsel bir şekilde depolamak ve kullanmak için gerekli altyapıyı sağlayan veritabanı yönetim sistemlerinin ortaya çıkması da bir başka önemli tarihsel dönüm noktası olarak ifade edilebilir (Han ve diğerleri, 2011).

Sözünü etmiş olduğumuz bu başlangıç noktalarından sonra “veri madenciliği” terimi ilk olarak 1960’larda büyük veri yığınlarındaki temel örüntüleri keşfetme sürecini tanımlamak için kullanıldı. İlgili tarihsel aralıkta araştırmacılar, veri madenciliğinin asli unsurlarından biri olan makine öğrenimi için algoritmalar geliştirmeye başladılar. Örnek vermek gerekirse 1986 yılında Ross Quinlan tarafından geliştirilen ID3 algoritması, veri madenciliğinde halen popüler bir yöntem olan karar ağaçları için ilk ve en etkili algoritmalardan biriydi (Quinlan, 1986). Bununla birlikte 1990’lara gelene dek veri madenciliğinin gelişim sürecini etkileyen temel birkaç gelişmeyi sıralayacak olursak ilişkisel veritabanı modeli ve sistemlerinin ortaya çıkması, bu gelişmeye paralel olarak

veritabanı sorgulama dillerinin üretilmesi ve çevrimiçi hareket işleme (OLTP) teknolojisinin kullanılmaya başlamasını anabiliriz (Han ve diğerleri, 2011).

Yukarıda kısaca dile getirdiğimiz veri madenciliğinin ilk temelleri ve belirlenişinden sonra alandaki esas gelişmelerin yaşandığı 1990'lardan sonraki süreç, hızlı bir büyüme ve olgunlaşma dönemi olarak görülebilir. Dijital verilerin kullanılabilirliğinin artması ile depolama ve işleme hususundaki gelişmeler, daha karmaşık ve ileri düzeyde veri madenciliği tekniklerinin geliştirilmesini bir zorunluluk haline getirmiştir. Örneğin alanda bir devrim niteliği taşıyan keşiflerden biri olarak 1994 yılında Rakesh Agrawal ve Ramakrishnan Srikant tarafından geliştirilen Apriori algoritmasını söyleyebiliriz (Agrawal ve Srikant, 1994). Veri madenciliği uygulamalarının pek çok alanında temel bir görev olan birliktelik kural çıkarımı probleminin çözümü için verimli bir yöntem geliştirmişlerdir. Birçok gelişmenin yaşandığı bu atılım döneminden önemli birkaç temel keşfi sıralayacak olursak büyük veri (big data) kavramının ortaya çıkması ve konuşulup üzerine düşünülen temel bir ilgi odağına dönüşmeye başlaması ve buna paralel olarak ilişkisel veri tabanlarının büyük veriyi ele almadaki kısıtlılıkları sebebiyle ilişkisel olmayan (NoSQL) veri tabanlarının ortaya çıkması dile getirilebilir (Kelleher ve Tierney, 2019).

2010'lardan günümüze gelene kadarki tarihsel süreçte teknolojinin her alanında daha önce benzeri görülmemiş bir gelişime tanıklık edilmiştir. Veri madenciliği alanı da söz konusu dönemi önemli ölçüde belirleyen büyük veri ve derin öğrenmenin yükselişinden etkilenmiştir (LeCun ve diğerleri, 2015). Bu aralıkta yaşanan gelişmeleri kuşatıcı bir şekilde sıralayabilmek olası değildir. Fakat birkaç madde halinde öne çıkarılacak noktaları anacak olursak Hadoop ve Spark gibi dağıtılmış bilgi işlem çerçevelerinin geliştirilmesinin büyük ölçekli veri kümelerinin işlenmesini sağlaması, sinir ağlarındaki gelişmeler, bu veri kümelerindeki karmaşık kalıpları keşfedebilen derin öğrenme tekniklerinin geliştirilmesinden söz etmek mümkündür (Dean ve Ghemawat, 2008). Bunlarla birlikte dikkat çekilmesi gereken belki de en önemli husus yapay zekâ ve doğal dil işleme alanında yaşanan çığır açıcı gelişmelerin veri bilimi alanındaki bütün alt dalları radikal bir şekilde etkileyip dönüştürmüş olmasıdır. Bu bağlamda sadece belirli ve özel bir veri kümesi üzerinde öğrenme yapılmayıp açık erişimdeki tüm verileri kendine veri kaynağı olarak alabilen yapay zekâ uygulamalarının ve önceden eğitilmiş farklı dil modellerinin ortaya çıkması, veri madenciliğinin geleneksel çalışma modelleri ve yaklaşım tekniklerini değiştirmeye ve dönüştürmeye başlamıştır.

Sonuç olarak veri madenciliği alanı, istatistik ve veri tabanı yönetimindeki tarihsel köklerinden veri biliminin vazgeçilmez bir parçası olarak bugünkü durumuna dek geçen zaman aralığında kayda değer biçimde değişmiş ve gelişmiştir. Burada kısaca sunmaya çalıştığımız tarihsel perspektif, ele aldığımız alanı şekillendiren etmenler hakkında bize bir bakış açısı sağlayacaktır.

2.1.1.1. Veri Ambarları ve Çevrimiçi Analitik İşleme

Veri ambarları, karar destek sistemlerine yardımcı olması amacıyla oluşturulmuş konu odaklı, entegre, zamana bağlı ve geçici olmayan veri koleksiyonu (deposu) olarak tanımlanmaktadır (Han ve diğerleri, 2011). Veri ambarları ve operasyonel veri tabanları, veri madenciliği tekniklerinin kullanımı açısından bazı farklılıklar göstermektedir. Bu sistemler arasında farklı işlevler ve farklı veri türleri bulunduğundan, veri ambarlarını operasyonel veri tabanlarından ayrı olarak ele almak gerekmektedir.

Veri ambarları, bünyesinde bulundurduğu verileri birleştirmekte ve genelleştirmektedir. Veri ambarları oluşturulurken veri temizleme, veri entegrasyonu ve veri dönüşümü işlemleri yapılmaktadır. Bu işlemler aynı zamanda veri madenciliği teknikleri için önemli ön işlem adımları olarak görülebilir. Ayrıca veri ambarları, veri madenciliği ve veri genellemesine fayda sağlayan çok boyutlu verilerin analizi için çevrimiçi analitik işleme araçlarını da bünyesinde barındırmaktadır. Bu nedenle veri ambarları, veri analizi ve çevrimiçi analitik işleme için daha önemli bir duruma gelmekte ve veri madenciliği için de birçok fayda sağlamaktadır. Sonuç olarak veri ambarları, elde bulunan veri içerisinden bilgi keşfi sürecinde önemli bir adım oluşturmaktadır.

2.1.1.2. Veri Küpü Teknolojisi

Veri ambarı sistemleri, yukarıda da belirtildiği gibi çok boyutlu verilerin analizi için çevrimiçi analitik işleme (OLAP) araçları sunmaktadır. Çevrimiçi analitik araçları, özet veriye hızlı ve etkili erişim sağlamak için veri küpü konsepti ve çok boyutlu bir veri modeli getirmektedir. Veri küpü kavramı başlangıçta çevrimiçi analitik işleme (OLAP) için tasarlanmış olmasına rağmen veri madenciliği için de fayda sağlamakta ve aktif olarak kullanılmaktadır.

Çok boyutlu veri madenciliği, bilgi keşfetme teknikleri ile çevrimiçi hareket işleme tabanlı veri analizini birleştiren bir veri madenciliği yaklaşımıdır. Bu kavram aynı zamanda çevrimiçi analitik madencilik (OLAM) olarak da bilinmektedir. Çevrimiçi analitik madencilik (OLAM) çok boyutlu uzayda verileri inceleyerek ilginç örüntüleri

keşfetmeye çalışmaktadır. Bu da herhangi bir boyutta verinin daha alt küçük kümelerine odaklanma imkânı vermektedir.

2.1.1.3. Sık Görülen Örüntülerin Tespiti, Birliktelik Analizi ve Korelasyonlar

Çok büyük miktarda veri arasından sık görülen örüntülerin, ortak noktaların ve korelasyon ilişkilerinin keşfedilmesi pazarlama, karar analizi, işletme yönetimi gibi birçok alanda fayda sağlamaktadır. Sık sık birlikte ve/veya sırayla satın alınan ürünleri inceleyerek müşterilerin satın alma alışkanlıklarını bulmaya çalışan pazar sepeti analizi uygulaması bu alanda yapılan popüler uygulamalardan biridir. Sık görülen kalıpların bilgisi firmaların daha verimli satış yapmasına ve raf alanlarını planlamalarına yardım ederek satışların artmasına yardımcı olmaktadır.

Sık görülen kalıpların bulunması birlikteliklerin, korelasyonların ve veriler arasındaki diğer ilginç kalıpların keşfedilmesinde önemli bir rol oynamaktadır. Ayrıca veri sınıflandırmaya, kümelemeye ve diğer veri madenciliği uygulamalarına katkı sağlamaktadır. Bundan dolayı, sık görülen kalıpların bulunması önemli bir veri madenciliği görevi ve aynı zamanda veri madenciliği araştırmalarında önemli bir başlık haline gelmektedir.

2.1.2. Veri Madenciliği İşlevleri

2.1.2.1. Sınıflandırma

Sınıflandırma, önemli veri sınıflarını tanımlayan modelleri çıkaran veri analizi biçimlerinden bir tanesidir. Sınıflandırıcılar olarak adlandırılan veri analiz biçimi kategorik olarak sınıf etiketlerini tahmin etmektedir. Veri madenciliği araştırmalarındaki önemli alanlardan bir tanesi de büyük miktarda verileri işleyebilen sınıflandırma ve tahmin etme tekniklerini geliştirmek olmuştur. Sınıflandırma teknikleri dolandırıcılık tespiti, pazarlama, performans tahmini, imalat ve tıbbi teşhis gibi birçok alanda uygulanmaktadır.

Veri sınıflandırma, bir sınıflandırma modelinin oluşturulduğu öğrenme aşaması ve bu model ile yeni verilerin sınıf etiketlerinin tahmin edildiği sınıflandırma aşaması olmak üzere iki aşamadan oluşmaktadır. İlk aşamada, önceden sınıf etiketleri belirlenmiş veriler kullanılarak bir sınıflandırma modeli oluşturulmakta ve ikinci aşamada ise oluşturulan bu sınıflandırma modeli kullanılarak yeni veriler sınıflandırılmaktadır.

2.1.2.2. Kümeleme Analizi

Kümeleme işlemi, bir veri kümesini birden çok grup veya küme halinde gruplama işlemidir. Bu kümeler veya gruplar oluşturulurken her bir küme içerisindeki birbirleriyle yüksek oranda benzerlik gösteren verilerin diğer gruplardaki verilerden de yüksek oranda farklı olmasına önem gösterilmektedir. Farklılıklar ve benzerlikler, verileri tanımlayan nitelik değerleri baz alınarak değerlendirilir ve çoğu zaman mesafe oranları içermektedir.

Bir veri madenciliği tekniği olarak kullanılan kümeleme biyoloji, güvenlik, iş zekâsı ve web aramaları gibi birçok alandaki uygulamalarda kullanılmaktadır. İş zekâsı uygulamalarında kümelemeden, çok sayıda müşteriyi benzer özellikleri olan müşteri grupları halinde organize etmek için faydalanılmaktadır. Bu sayede, gelişmiş müşteri ilişkileri yönetimi için iş stratejileri geliştirmek kolaylaşmaktadır.

Kümeleme, web aramalarında da birçok şekilde kullanılmaktadır. Örneğin; bir anahtar kelime arandığında milyarlarca web sayfası bulunmasından dolayı çok fazla sayıda arama sonucu dönmektedir. Kümeleme teknikleri, arama sonuçlarını gruplar halinde organize etmeyi ve sonuçların hızlı ve kolay erişilebilir bir şekilde sunulmasını sağlamaktadır. Ayrıca, bilgiye erişim uygulamalarında yaygın olarak kullanılan, belgeleri veya dokümanları belirli başlıklar altında toplama işlemi için de kümeleme teknikleri geliştirilmiştir.

Kümeleme, veri dağılımı hakkında fikir edinmek için tek başına bir veri madenciliği tekniği olarak kullanılabileceği gibi daha önceden tespit edilen veri kümeleri üzerinde çalışacak diğer veri madenciliği algoritmaları için bir ön işlem adımı olarak da kullanılabilmektedir.

2.1.2.3. Aykırı Durumların Tespiti

Aykırı durumların tespiti işlemi, veriler içerisinde beklenenden çok farklı niteliklere sahip olan verileri tespit etme işlemidir. Literatürde anomali tespiti olarak da adlandırılan bu işlemde farklı davranışlara sahip olan verilere aykırı değerler veya anomaliler denmektedir.

Aykırı durum tespiti tekniğinin dolandırıcılık tespiti, tıbbi bakım, kamu emniyeti ve güvenliği, sanayide hata tespiti, görüntü işleme, sensör / video ağların takip edilmesi ve saldırı tespiti gibi birçok alanda uygulamaları vardır.

Aykırı durumların tespit edilmesi ve kümeleme işlemleri birbiriyle alakalı işlemlerdir. Kümeleme, bir veri kümesinin içerisindeki çoğunluğun oluşturduğu kalıpları bulup veriyi organize etmeye çalışırken, aykırı durumların tespitinde çoğunluk kalıplarından büyük ölçüde sapmış olağandışı durumlar yakalanmaya çalışılmaktadır.

Aykırı durumlar ile "gürültülü veriler"i birbirine karıştırmamak gerekir. Gürültülü veri, ölçülen bir değişkende meydana gelmiş rastgele bir hatadır. Aykırı durum tespiti de dahil olmak üzere veri analizi konusunda genellikle gürültülü verilerle ilgilenilmez. Diğer birçok veri analizi ve veri madenciliği uygulamalarında olduğu gibi aykırı durumların tespit edilmesi işleminden önce gürültülü veriler veri kümesinden çıkarılmalıdır.

2.1.3. Veri Madenciliği Süreci

Veriden bilgi keşfetme süreci aşağıdaki adımlardan oluşmakta ve bu adımlar daha başarılı sonuçlar elde etmek için bir döngü biçiminde kullanılabilmektedir: veri temizleme, veri entegrasyonu, veri seçimi, veri dönüşümü, veri madenciliği, modelin değerlendirilmesi ve veri görselleştirilmesi. İlk dört adım birçok kaynakta veri önişleme adımları olarak kabul edilmektedir.

Bir uygulamada veri madenciliği tekniklerini kullanmadan önce, uygulama içerisinde bulunan verileri hazırlamak gerekmektedir. Bunun için de nitelikler ve veri değerleri daha yakından incelenmelidir. Teorik olarak bu işlemler kolay gibi gözükse de pratikte uygulama verilerini hazırlama işlemi zordur ve zaman almaktadır. Bunun temel sebepleri; veri madenciliği tekniklerinin uygulanacağı verilerin gürültülü ve hacim bakımından çok büyük olması ile birbirinden farklı kaynaklardan elde edilmesi olarak sayılabilir. Uygulama verisi hakkında bilgi sahibi olmak, veri madenciliği süreçleri ve başarılı sonuçlar elde etmek için faydalıdır.

Uygulama verilerini incelemeye başladıkça çeşitli nitelik türleri ve temel istatistik kavramları karşımıza çıkmaktadır. Veriler hakkında yapılan bu ön çalışmalarda elde edilen nitelikler, niteliklerin değerleri ve temel istatistikler hem veri ön işleme adımında hem de veri entegrasyonu adımında eksik değerlerin doldurulmasına, aykırı değerlerin tespit edilmesine ve veri tutarsızlıklarının giderilmesine yardımcı olmaktadır.

Kurumlarda kullanılan veri tabanları, büyük boyutlarda olmaları ve çok sayıda heterojen kaynaktan beslenmeleri sebebiyle gürültülü, eksik ve tutarsız verilere sahip olabilmektedir. Bu durum veri kalitesinin düşmesine ve veri madenciliği sonuçlarının istenen başarıyı elde edememesine neden olmaktadır. Eldeki verinin kaliteli olup

olmadığı ise doğruluk, tutarlılık, eksiksizlik, inandırıcılık ve yorumlanabilirlik kavramları ile tespit edilmektedir.

Kaliteli ve başarılı sonuçlar elde etmek için verinin ön işlemlerden geçirilmesi gerekmektedir. Bu ön işlemler; veri temizleme, veri entegrasyonu (bütünleştirme), veri seçimi ve veri dönüşümü gibi adımlardan oluşmaktadır.

Verilerdeki gürültüyü gidermek ve tutarsızlıkları düzeltmek için *veri temizleme* işlemi uygulanmaktadır. Birden fazla heterojen kaynaktan gelen verileri, veri ambarı gibi tutarlı bir veri deposunda birleştirmek için *veri entegrasyonu* işleminden geçirmek gerekmektedir. *Veri seçimi* işleminde toplama, gereksiz özellikleri ortadan kaldırma ve kümeleme yapılarak veri boyutu azaltılmaktadır. *Veri dönüşümü* işleminde ise veriler belli bir aralığa ölçeklendirilmektedir.

Veri madenciliği tekniklerinden önce, veri ön işlemlerinin uygulanması ortaya çıkacak örüntülerin kalitesini arttırırken aynı zamanda sonuçların elde edilmesi için geçen süreyi de oldukça azaltmaktadır.

2.1.3.1. Veri Temizleme

Bu adımda temel olarak, eksik verilerin giderilmesi (doldurulması veya çıkarılması), gürültülü verilerin elenmesi ve verideki tutarsızlıkların giderilmesi işlemleri yapılmaktadır.

Eksik verilerin giderilmesi için tercih edilen yöntemlerden bazıları aşağıda listelenmiştir:

- *Eksik veri içeren kayıtları göz ardı etme*: Kayıtlardaki eksik veri sayısının çok düşük olduğu durumlarda, o verileri içeren kayıt veri setinden çıkarılarak göz ardı edilebilir. Oldukça kolay bir çözüm olmasına rağmen, kayıt içerisindeki birden çok veri alanı eksik olmadığı müddetçe bu yöntem çok etkili olmayacaktır.
- *Eksik verileri elle doldurma*: Her ne kadar uygun bir yöntem olarak görünse de birçok eksik değer içeren büyük bir veri seti göz önüne alındığında zaman alıcı ve etkin olmayan bir yönteme dönüşebilir.
- *Eksik veriyi doldurmak için sabit bir değer kullanma*: Eksik verilerin yerine sabit olarak belirlenen bir değer (“Bilinmeyen”, “Eksik”, -1 vs) kullanılabilir. Böyle bir durumda eksik veri yerine seçilen sabit değerin, veri madenciliği algoritmaları tarafından yanlışlıkla ilgilenilecek bir kavram olarak tespit edilebileceği de

unutulmamalıdır. Bu yüzden, basit bir yöntem olmasına rağmen çok verimli bir çözüm değildir.

- *Merkezî eğilim ölçüsü kullanarak eksik verileri doldurma*: Eksik olan veriler, veri setinde aynı sütundaki diğer verilerin ortalaması, modu veya medyanı hesaplanarak doldurulabilir. Bir başka yöntem ise, makine öğrenmesi algoritmaları kullanarak tahmini değerler elde etme ve bu değerleri eksik verilerin yerine kullanmaktır.
- *Eksik veriyle aynı kategorideki/sınıftaki kayıtların ortalamasını veya medyan değerini kullanma*: Eksik verinin olduğu veri setindeki tüm kayıtların ortalama veya medyan değerini kullanmanın yerine aynı kategorideki/sınıftaki kayıtların ortalaması veya medyan değeri alınarak eksik veriler doldurulabilir.
- *En muhtemel değeri kullanarak eksik verileri doldurma*: Bu yöntemde öncelikle regresyon, Bayesian formülü veya karar ağaçları kullanılarak en olası değer belirlenir ve bu değer, eksik verilerin yerine kullanılabilir.

Veri içindeki rastgele hataya veya sapmaya "gürültü" denilmektedir. Veri temizleme aşamasındaki önemli adımlardan biri de verideki rastgele hataları veya sapmaları tespit edip çıkarmaktır. Gürültülü verilerin çıkarılması için tercih edilen yöntemlerden bazıları aşağıda listelenmiştir:

- *Binning*: Veri setindeki kayıtlar, eksik verinin bulunduğu sütuna göre sıralanır ve belirlenen kayıt sayısı kadar küçük parçalara ayrılır. Eksik veri değerleri "komşu"larına, yani çevresindeki verilere göre değerlendirilerek düzeltilir.
- *Regresyon*: Veri değerlerini uygun hale getiren bir teknik olan regresyon kullanılarak eksik veriler düzeltilebilir.
- *Aykırı değerler analizi*: Benzer verilerin bir küme içerisinde toplanması ve bu kümelerin dışında kalan verilerin aykırı veri olarak değerlendirilmesi sonucunda veri setinden çıkarılmasıyla düzeltilebilir.

2.1.3.2. Veri Entegrasyonu

Veri madenciliği projelerinin çoğunda, birden çok veri kaynağından gelen verilerin birleştirilmesi olarak tanımlanan veri entegrasyonu gerekli olabilir. Dikkatli ve düzgün bir şekilde yapılan entegrasyon, veri setindeki fazlalıkları ve tutarsızlıkları azaltmaya ve önlemeye yardımcı olur. Bu durum da veri madenciliği sürecinin sonraki adımlarında doğru sonuçlar elde etmeyi ve işlemlerin hızlanmasını sağlar.

2.1.3.3. Veri Seçimi

Veri seti içerisindeki veri madenciliğinin çözmeye çalıştığı problemi etkileyecek ya da ilgilendirecek değişkenlerden, orijinal veri setinin bütünlüğünü bozmadan daha küçük bir veri seti oluşturma işlemidir. Seçilecek veri seti, istatistiksel hesaplamada veya model oluşturmada bir anlam ifade edecek kadar veriyi içermelidir. Yani, seçilen veri setindeki madencilik işlemleri daha verimli olmalı, ancak neredeyse aynı sonuçları üretmelidir.

Veri seçimi için tercih edilen yöntemlerden bazıları aşağıda listelenmiştir:

- *Boyut azaltma*: Veri setindeki özniteliklerin sayısının azaltılması işlemidir.
- *Sayısallığı azaltma*: Orijinal veri setinin hacmini alternatif, daha küçük veri temsil formlarıyla değiştirme işlemidir.
- *Veri sıkıştırma*: Orijinal veri setinin azaltılmış veya sıkıştırılmış bir temsilini elde etmek için dönüşüm uygulama işlemidir.

2.1.3.4. Veri Dönüşümü

Veri madenciliği işleminin sonuçlarının daha verimli olması ve tespit edilen örüntülerin daha kolay anlaşılabilmesi için veri setinin dönüştürülmesi veya birleştirilmesi işlemidir. Veri dönüşümü için tercih edilen yöntemlerden bazıları aşağıda listelenmiştir:

- *Düzeltme (smoothing)*: Veri setindeki gürültüleri giderme işlemidir. Binning, regresyon ve kümeleme teknikleri kullanılarak düzeltme yapılabilir.
- *Birleştirme-toplama (aggregation)*: Verilere, birleştirme veya toplama işlemlerinin uygulandığı bir yöntemdir.
- *Normalleştirme*: Verileri, belirlenen bir ölçeğe (0-1 aralığı veya -1 ile 1 aralığı vs.) indirgeme işlemidir. Bu sayede, farklı ölçekteki verilerin birlikte ele alınması sağlanabilir.
- *Ayrıştırma (discretization)*: Veri setindeki sayısal bir niteliğin ham değerlerini, aralıklar veya kavramsal etiketlerle değiştirme işlemidir.
- *Nitelik(ler) ekleme*: Madencilik işlemine yardımcı olmak için, veri setindeki niteliklerden yeni niteliklerin oluşturulması ve eklenmesi işlemidir.

2.1.3.5. Veri Madenciliği

Burada söz konusu olan işlem veri madenciliği algoritmaları kullanılarak bir model oluşturulmasıdır. Çalışmamızın esas konusunu oluşturan bu kısım ilerleyen bölümlerde detaylı olarak anlatılacaktır.

2.1.3.6. Modelin Değerlendirilmesi

Veri madenciliği sürecinde, oluşturulan modellerin başarı oranlarının tespit edilmesi ve sonuçların ne denli isabetli olduğunun bilinmesi önemlidir. Bu noktada, seçilen algoritma türlerine göre modelin değerlendirilmesi için kullanılan birçok metrik türü bulunmaktadır. En çok kullanılan metrik türleri 4 ana başlık altında incelenecektir:

Sınıflandırma modellerinde değerlendirme metrikleri:

True Positive (TP - Doğru Pozitif): Sınıflandırma algoritması tarafından doğru bir şekilde pozitif belirlenmiş pozitif kayıtları ifade eder.

True Negative (TN - Doğru Negatif): Sınıflandırma algoritması tarafından doğru bir şekilde negatif belirlenmiş negatif kayıtları ifade eder.

False Positive (FP - Yanlış Pozitif): Sınıflandırma algoritması tarafından yanlış bir şekilde pozitif olarak belirlenmiş negatif kayıtları ifade eder.

False Negative (FN - Yanlış Negatif): Sınıflandırma algoritması tarafından yanlış bir şekilde negatif olarak belirlenmiş pozitif kayıtları ifade eder.

		Gerçek Değer	
		Pozitif	**Negatif**
Tahmin Edilen Değer	**Pozitif**	TP	FP
	Negatif	FN	TN

Tablo 1: Karışıklık Matrisi

Karışıklık Matrisi (Confusion Matrix): Sınıflandırma algoritmasının farklı sınıflardaki kayıtları ne kadar iyi belirleyebildiğini analiz etmek için kullanılan faydalı bir araçtır. TP ve TN değerleri sınıflandırma algoritmasının ne kadar doğru tahmin ettiğini gösterirken, FP ve FN değerleri de ne kadar yanlış tahmin ettiğini göstermektedir. Yüksek oranda başarı gösteren sınıflandırma algoritmalarında, FP ve FN değerleri sıfır veya sıfıra çok yakındır. Öte yandan TP ve FP değerlerinin toplamı da tüm kayıtların toplam sayısına eşit veya çok yakındır.

Doğruluk (Accuracy): Bir özelliğin doğru tahmin edilen değerlerinin tüm değerlere oranına "doğruluk" denilmektedir. Bir başka ifadeyle, bir test veri setinde, sınıflandırma algoritması tarafından doğru şekilde sınıflandırılan kayıt sayısının tüm kayıt sayısına oranı "doğruluk" olarak tanımlanır.

Doğruluk = TP+TN / P + N

Hata / Yanlış Sınıflandırma Oranı (Error / Misclassification Rate): Bir test veri setinde, sınıflandırma algoritması tarafından yanlış şekilde sınıflandırılan kayıt sayısının tüm kayıt sayısına oranına "hata oranı" denilmektedir.

Hata Oranı = FP + FN / P + N = 1 - Doğruluk Oranı

Hassaslık (Sensitivity) / Yakalama (Recall): "Doğru pozitif oranı" olarak da adlandırılan bu ölçüt, gerçekte pozitif olan kayıtların, sınıflandırma algoritması tarafından ne kadarının pozitif tahmin edildiğini ifade eder.

Hassaslık = TP / TP + FN = TP / P

Kesinlik (Precision): Sınıflandırma algoritması tarafından pozitif olarak tahmin edilen kayıtların, gerçekte hangi oranda pozitif olduğunu ifade eder.

Kesinlik = TP / TP + FP

Belirlilik (Specifity): "Doğru negatif oranı" olarak da adlandırılan bu ölçüt, gerçekte negatif olan kayıtların, sınıflandırma algoritması tarafından ne kadarının negatif tahmin edildiğini ifade eder.

Belirlilik = TN / TN + FP = TN / N

Yanlış Pozitif Oranı (False Positive Rate): Bu ölçüt, gerçekte negatif olan kayıtların, sınıflandırma algoritması tarafından ne kadarının pozitif tahmin edildiğini ifade eder.

Yanlış Pozitif Oranı = FP / FP + TN = 1 - Belirlilik

F1 Skoru (F1 Score): Kesinlik ve hassaslık değerlerini beraber kullanarak elde edilen bir ölçüttür. Kesinlik ve hassaslık ölçütlerinin harmonik ortalaması olarak hesaplanan F1 skorunun, doğruluk ölçütünden daha kullanışlı olduğu söylenebilir.

F1 Skor = 2 x Kesinlik x Hassaslık / (Kesinlik + Hassaslık)

ROC Eğrisi (ROC Curve): Sınıflandırma algoritmalarının başarı performanslarını karşılaştırmak için kullanılan görsel bir değerlendirme ölçütüdür.

Performans karşılaştırması bir grafik üzerinde gösterilir. Grafik, yanlış pozitif oranının x ekseninde, doğru pozitif oranının y ekseninde gösterilmesiyle çizilir. Bu grafikteki her nokta, belirli bir karar eşiğine denk gelen "hassaslık / 1 - belirlilik" oranını temsil etmektedir.

Eğrinin Altında Kalan Alan (AUC-Area Under Curve): ROC eğrisi altında kalan alan kullanılarak sınıflandırma algoritmalarının başarı performansları karşılaştırılabilir. Sınıflandırma algoritmasının AUC değeri 1'e ne kadar yakın ise o kadar doğru sınıflandırıyor demektir. Aslında F1 skoru ne kadar yüksekse, ROC eğrisi altında kalan alan da o kadar yüksektir ve AUC değeri de 1'e o kadar yakındır.

Regresyon modellerinde değerlendirme metrikleri:

Maksimum Hata (Max Error): Regresyon algoritmasında, tahmin edilen değer ile gerçek değer arasındaki en kötü durum hatasını hesaplayan bir ölçüttür.

Max Error = max(|yi - yî|)

Ortalama Kare Hata (MSE-Mean Squared Error): Regresyon eğrisinin bir dizi noktaya ne kadar yakın olduğunu hesaplayan ölçüttür. Kare (karesel) hatanın veya kaybın, beklenen değerine karşılık gelen bir ölçüt olarak da tanımlanabilir. MSE değeri sıfıra ne kadar yakınsa, regresyon algoritmasının başarısı o kadar yüksektir.

Ortalama Kare Hatanın Kökü (RMSE -Root Mean Square Error): Bir regresyon algoritmasının tahmin ettiği değerler ile gerçek değerleri arasındaki uzaklığın bulunmasında kullanılan ve hatanın büyüklüğünü ölçen bir ölçüttür. RMSE değerinin 0 olması, algoritmanın hiç hata yapmadığı anlamına gelmektedir.

Ortalama Mutlak Hata (MAE - Mean Absolute Error): Ortalama mutlak hata, iki sürekli değişken arasındaki farkı bulmak için kullanılan ölçüttür. MAE, gerçek değer ile veriye en iyi uyan çizgi arasındaki ortalama dikey mesafedir.

2.1.3.7. Veri görselleştirilmesi

Sıraladığımız bu adımlardan sonra keşfedilen bilginin kullanıcıların anlayabileceği formatta aktarılması için görselleştirilmesi ve sunumunun gerçekleştirilmesi adımıdır.

Veri görselleştirme, eldeki karmaşık verilerin daha kolay anlaşılabilmesi ve yorumlanabilmesi için görsellere dönüştürme işlemidir. Veri madenciliği algoritmaların gelişmesi ve teknolojik gelişmeler sayesinde büyük ve karmaşık veri setleri kolaylıkla

analiz edilebilmekte ve başarılı sonuçlara ulaşılmaktadır. Bu sayede karşılaştırma, yorumlama ve analiz etme gibi işlemler çok daha verimli şekilde gerçekleştirilebilmektedir. Burada da verinin bir bütün halinde ele alınarak görselleştirilmesinin kıymeti ön plana çıkmaktadır.

Veri görselleştirme teknikleri kuralların, kavramların daha iyi anlaşılması, yeni yapıların keşfedilmesi veya bu yapıları düzenlemek gibi çeşitli amaçlar için kullanılabilmektedir. Bunlardan kuralların ve kavramların daha iyi anlaşılması için kullanılan görselleştirmeye "bilgi görselleştirilmesi", grafikler ve tablolarla yeni yapılar keşfetmek veya bu yapıları düzenlemek için kullanılan görselleştirmeye "görsel bilgi keşfi" denmektedir.

2.1.4. Kullanım Alanları

Her alan ve disipline yayılabilecek oldukça geniş bir uygulama sahasına sahip olan veri biliminin kritik bir parçası olan veri madenciliğinin kullanımı, son on yılda çeşitli sektörlerin değer ve fayda üreten birçok uygulamasında ciddi bir artış göstermiş ve giderek daha görülür hale gelmiştir. Elbette söz konusu teknik ve uygulamaların yer aldığı tüm alanlardan söz etmek olanaklı olmamakla birlikte bu kısımda, veri madenciliğinin birkaç temel kullanım alanındaki örneklerine başvurarak genel bir resim vermeye çalışacağız. Bu bağlamda günümüz yaşamının belki de en hayati parçalarını oluşturan dört alandan, yani sağlık ve biyomedikal araştırmalar, finans ve bankacılık, (e-) ticaret ve telekomünikasyondan örnekleri ele alabiliriz.

Veri madenciliğinin yoğun olarak kullanıldığı ve etkili sonuçlar alınan alanlardan biri olarak ilk önce sağlık ve biyomedikal araştırmalardan söz etmemiz mümkündür. Örneğin, bir veri madenciliği tekniği olan öngörücü modelleme aracılığıyla hasta verilerinin değerlendirilmesi, hastalığın ilerlemesi hakkında tahminlerde bulunulabilmesi ve böylelikle önleyici tedavilerin gerçekleştirilmesi mümkün olmaktadır. Bir başka örnek ise yaş, kilo ve genetik yatkınlık gibi faktörlerin makine öğrenmesi algoritmaları kullanılarak değerlendirilmesi sayesinde diyabet başlangıcının tahmin edilebilmesidir (Kavakiotis ve diğerleri, 2017).

Bunun yanı sıra gen bilimi çalışmalarında da veri madenciliği, hastalıklarla ilişkili genlerin tanımlanmasını sağlayacak biçimde de kullanılmaktadır. Örnek olarak hastalıklarla arasında bağlantı olan genetik varyantları tespit etmek için oluşturulmuş

Genome-Wide Association Study (GWAS) isimli çalışma, söz konusu alanda dikkat çeken bir uygulama verilebilir (Moore ve Williams, 2009).

Veri madenciliğinin kullanıldığı başka uygulama sahalarından biri olan finans ve bankacılık sektöründe veri madenciliğinin çok sayıdaki kullanım alanlarından birkaçını sıralamak gerekirse kredi puanlama, dolandırıcılık tespiti ve müşteri segmentasyonu dile getirilebilir. Örneğin, en temel veri madenciliği algoritmalarından biri olan karar ağaçları, uzun bir süredir müşterilerin finansal kayıtlarına dayalı olarak kredi skorlarını belirlemek için kullanılmaktadır (Thomas, 2000).

Bunun yanı sıra dolandırıcılık tespiti de veri madenciliğinin önemli bir role sahip olduğu başka bir alandır. Anomali tespitinde kullanılan teknikler sayesinde, hileli işlemleri belirleyebilecek olağandışı örüntülerin tespit edilmesi olanaklı hale gelmektedir (Bolton ve Hand, 2002).

Perakende ve e-ticaret sektöründe sıkça kullanılan pazar sepeti analizi, müşteri segmentasyonu ve öneri sistemleri için de veri madenciliğine başvurulmaktadır. Örneğin, pazar sepeti analizi, genellikle birlikte satın alınan ürünleri belirlemek için birliktelik kuralı analizini/madenciliğini kullanır. Bu sayede ürün yerleştirme ve çapraz satış stratejileri hakkında bu ürünleri tedarik eden kimselere nitelikli bilgi sağlanmaktadır (Tan, Steinbach ve Kumar, 2005). Böylece genellikle birlikte satın alınan ürünler sözgelimi birbirlerine yakın raflara dizilerek ya da online alışverişlerde bir arada sunularak bir ürünü alan müşterinin diğer ürüne de kolaylıkla erişebilmesi sağlanmaktadır.

Veri madenciliğinin uygulandığı bir başka alan olan tavsiye sistemleri, alışveriş sitelerinden film sitelerine ya da sosyal medya platformlarından online eğitim sitelerine kadar birçok sektörde kullanılmaktadır. Müşterilerin/kullanıcıların geçmişteki davranışları ile benzer müşterilerin/kullanıcıların davranışlarını değerlendirerek müşterilere/kullanıcılara ürün (film, müzik, online eğitim, arkadaş, vb.) önermek için işbirlikçi filtreleme teknikleri kullanılır (Linden, Smith ve York, 2003). Böylelikle müşteri/kullanıcının deneyimi daha zengin ve isabetli bir hal almaktadır.

Son olarak veri madenciliği tekniklerinin kullanıldığı telekomünikasyon endüstrisinden kısaca söz edebiliriz. Kayıp tahmini, ağ optimizasyonu ve dolandırıcılık tespitinde veri madenciliği oldukça faydalı sonuçlar vermektedir. Örneğin müşteri kayıp tahmin modelleri, kullanım kalıplarına ve geri bildirimlerine göre müşteriler arasından

aboneliklerini iptal etme olasılığı yüksek olanları bulmak için veri madenciliği tekniklerini kullanır (Mozer ve diğerleri, 2000). Bu sayede söz konusu müşterilere avantajlı teklifler sunularak kaybın gerçekleşmesi önlenebilmektedir.

2.1.5. Veri Madenciliğiyle İlişkili Alanlar ve Diğer Hususlar

2.1.5.1. Yapay Zekâ

Son zamanlarda adından çokça söz ettiren ve her geçen gün önemi artan bir teknoloji olan yapay zekâ; makinelerin öğrenme, akıl yürütme, sorun çözme, algılama ve dili anlama gibi insan davranışlarını sergileyen makinelerin geliştirilmesi olarak tanımlanabilir. Veri madenciliği bağlamında yapay zekâ, büyük veri setlerindeki örüntüler, korelasyonlar ve kuralları keşfetme sürecini otomatikleştirmekte ve makine öğrenmesi, sinir ağları ve derin öğrenme gibi çeşitli yapay zekâ teknikleri sayesinde, geleneksel programlamaya ihtiyaç duymadan tahmine dayalı modeller oluşturmak ve veriye dayalı kararlar almayı olanaklı kılmaktadır. Yapay zekanın bu yetenekleri yalnızca veri madenciliği süreçlerinin verimliliğini ve doğruluğunu artırmakla kalmaz, ayrıca büyük boyutlardaki farklı türde verilerden öngörüler çıkarılmasını sağlayarak bu süreçlere önemli ölçüde katkıda bulunur.

2.1.5.2. Metin Madenciliği

Metin madenciliği, veri kaynağı olarak metinleri ele alan veri madenciliği türüdür. Burada yapılan işlem, belirlenen amaçlar doğrultusunda metinden değerli bilgileri çıkarmak için metnin analiz edilmesidir. Metin madenciliğinde başlıca amaçlar; metinlerin sınıflandırılması, kümelenmesi, metinlerden konu çıkarılması, metinlerin özetlenmesi olarak sıralanabilir. Ayrıca metin madenciliği ile yüksek miktardaki metin verisinden gözle görülmeyecek içeriklerin, ilişkilerin ve örüntülerin çıkartılması ve bunların kurum ve kuruluşlara çeşitli yönlerden fayda getirmesi amaçlanmaktadır.

Metin madenciliği, yapılandırılmamış veya yarı yapılandırılmış metin verilerini veri kaynağı olarak ele alıp işleme ve analiz etme işlemidir. Veri madenciliğinde olduğu gibi metin madenciliğinde de birçok teknik ve yaklaşım söz konusudur. Bu tekniklerin ve yaklaşımların temel amacı metin verisini algoritmaların anlayabileceği şekilde sayılara dönüştürmek ve sayılar üzerinde işlem yaparak metni analiz etmektir.

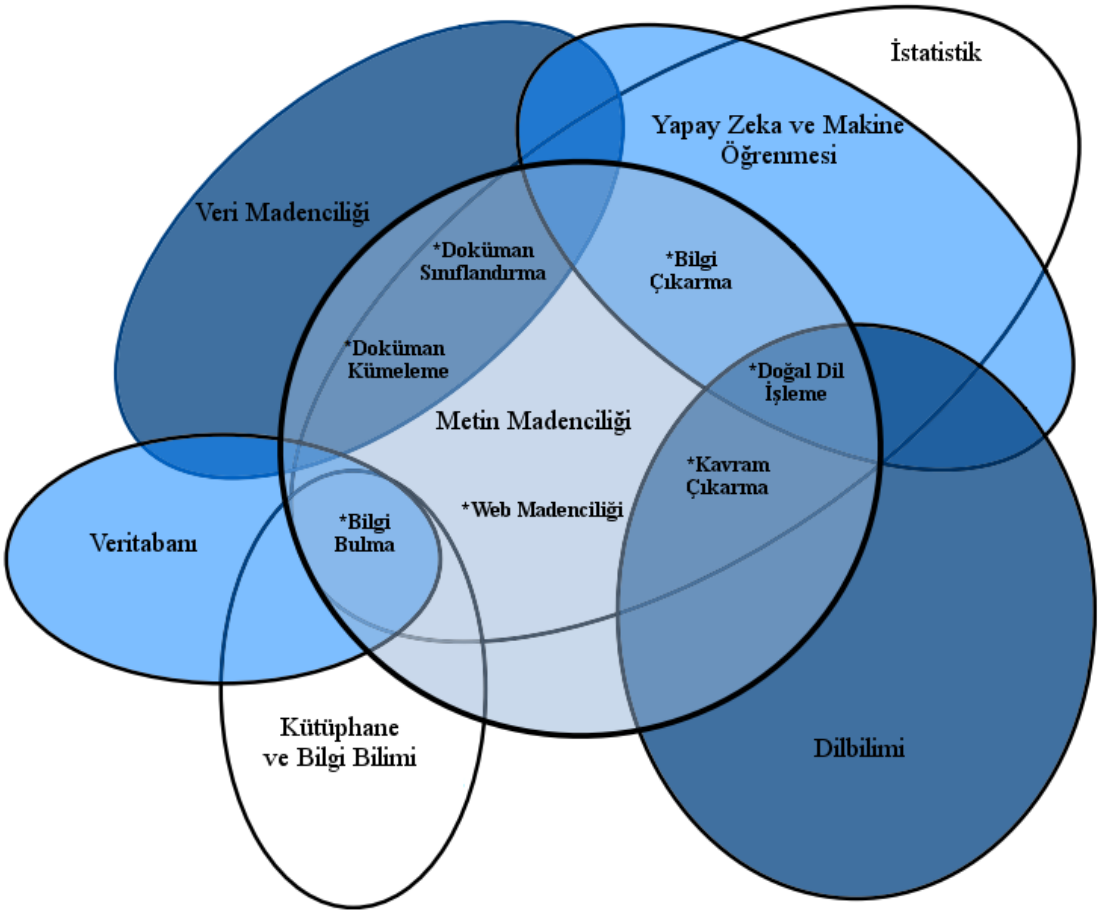
Metin madenciliği kavramı, altı temel alanla ilişki içerisindedir. Bu alanları şu şekilde sıralayabiliriz:

1. Veri Madenciliği
2. Veri tabanı
3. Kütüphane ve Bilgi Bilimi
4. Dilbilim
5. Yapay Zekâ ve Makine Öğrenmesi
6. İstatistik

Metin madenciliği yukarıda belirtilen alanlarda yedi farklı amaçla kullanılmaktadır. Bu kullanım alanlarını da şu şekilde sıralayabiliriz:

- **Arama ve bilgi bulma:** Anahtar kelime(ler) aranması ve arama motorları kullanılarak metin dokümanlarının bulunması ve saklanması.
- **Doküman kümeleme:** Veri madenciliğindeki kümeleme teknikleri kullanılarak terimlerin, paragrafların, kalıp metinlerin ve dokümanların gruplanması ve kategoriler halinde ayrılması.
- **Doküman sınıflandırma:** Veri madenciliğindeki sınıflandırma teknikleri kullanılarak kalıp metinlerin, paragrafların veya dokümanların gruplanması ve kategoriler halinde ayrılması.
- **Web madenciliği:** Belli bir alana odaklanarak internet üzerinde veri ve metin madenciliğinin kullanılması.
- **Bilgi çıkarma:** Yapılandırılmamış veya yarı yapılandırılmış metinlerin belli bir düzene göre yapılandırılması ve anlamlı sonuçların çıkarılması.
- **Doğal dil işleme:** Düşük seviyeli dil işlemesi ve kelimelerin anlaşılması.
- **Kavram çıkarma:** Kelimelerin ve cümlelerin semantik olarak benzer gruplara ayrılması.

Metin madenciliğinin ilişkili olduğu altı temel alan ve bu alanlardaki kullanımları aşağıdaki şekilde bir arada gösterilmiştir.



Şekil 1: Metin Madenciliği ve İlişkili Olduğu Alanlar

1. Metin Sınıflandırma ve İşleme

Verilen bir metnin sınıflandırılması veya hangi kategoriye ait olduğunun belirlenmesi veri bilimi / veri madenciliği alanındaki önemli konulardan birisidir. Metin sınıflandırma olarak isimlendirilen yöntem, farklı alanlarda farklı amaçlar için kullanılmaktadır. Örneğin; sosyal medyada paylaşılan yorumların duygu analizinin yapılması, ürün bilgilerinin değerlendirilerek ürünün doğru kategoriye yerleştirilmesi veya bir metnin yazarının otomatik olarak tespit edilmesi önemli metin sınıflandırma problemleri arasında yer almaktadır. Sınıflandırma işleminin yapılabilmesi için metinlerin özniteliklerinin (üst verilerinin) çıkartılmış olması, yani metnin yapısal veriye dönüştürülmesi gerekmektedir.

Kurumsal uygulamalar içerisinde en çok kullanılanlardan birisi olan elektronik belge yönetim sistemlerinde belgelerin üst verileri ve içeriği ele alınıp metin sınıflandırma yöntemleri aracılığıyla "Standart Dosya Planı (SDP)" kodunun tahmin edilmesi de bu alandaki en temel kullanımlarından birini oluşturmaktadır. Bu sayede EBYS'lerde en çok yapılan hatalardan biri olan *yanlış standart dosya planı kodu*

*seçimi*nin önüne geçilerek hem kullanıcı kaynaklı hatalar minimum seviyede tutulacak hem de belgelere erişim sırasında standart dosya planından maksimum seviyede yararlanılmış olacaktır. Böylelikle bilgi ve belgelerin düzgün bir şekilde yönetilmesiyle kurumsal faaliyetlerde geçmiş işlemlerin kontrolü sağlanabileceği gibi geleceğe dair öngörüler de elde edilebilecektir.

Belge metninden standart dosya planının tahmin edilmesine ilave olarak belgedeki anahtar kelimelerin tespit edilmesi, sdp kodu-anahtar kelime(ler), sdp kodu-konu ve sdp kodu-birim/kullanıcı eşleştirmeleri de yapılabilmektedir. Bu sayede, birçok online platformda sıklıkla kullanılan önerme sistemlerine benzer şekilde, elektronik belge oluşturulurken kullanıcılara sdp kodu ve konu başlığı gibi önermeler yapılabilecektir.

Kurum dışından gelen elektronik belgelerin EBYS'lere dahil edilmesi aşamasında, gönderici kurumda belgeye verilen sdp kodunun aynısının alıcı kurumda da aynı şekilde kullanılması gerektiği gibi bir yanlış algı bulunmaktadır. Gelen belgeye, alıcı kurumda başka bir sdp kodu ataması da yapılabileceğinden, yukarıda bahsedilen yaklaşım her zaman doğru sonuç vermemektedir. Bu noktada gelen belge metni değerlendirilip sistemde bulunan belgelere benzerliğine bakılarak bir sdp kodu tahmini yapılabilmektedir.

Metin Sınıflandırma Yöntemleri / Modelleri

Kelime Çantası (Bag of Words, BoW) Modeli

Belgedeki kelimelerin belgeyi temsil etmesi ve belgenin öznitelikleri olmasına dayanarak tasarlanmış bir modeldir. Bu modelin kullanılabilmesi için öncelikle ön işleme adımlarının yapılması ve mümkün olduğu kadar öznitelik sayısının azaltılması gerekmektedir.

Metin ön işleme adımlarından bazıları aşağıda listelenmiştir:

- Belgedeki bütün kelimelerin küçük harfe dönüştürülmesi
- Tüm noktalama işaretlerinin ve yazdırılamayan karakterlerin kaldırılması ve boşluk karakteri ile yer değiştirmesi
- Tek başına anlam ifade etmeyen ve sık kullanılan kelimelerin (stop words) metinden çıkarılması
- Kelimelerdeki son eklerin çıkarılmasıyla morfolojik köklerin elde edilmesini amaçlayan kök bulma işleminin yapılması

Sayma Vektörü Yöntemi (Count Vector)

Belgeler içerisinde geçen kelimelerin (terimlerin) görünürlüğünü tanımlamada kullanılan yöntemlerden biridir. Bu yöntemde, kelimenin belge içerisinde geçip geçmediği veya kaç kez geçtiği hesaplanmaktadır.

Terim Sıklığı-Ters Doküman Sıklığı (TF-IDF) Yöntemi

Belge içerisindeki kelimelerin ağırlık değerinin hesaplanması için kullanılan yöntemlerden biri olan "Terim Sıklığı-Ters Doküman Sıklığı" (Term Frequency-Inverse Document Frequency), metin sınıflandırma alanında yaygın olarak kullanılan yöntemlerden biridir. TF-IDF kapsamında TF ve IDF olmak üzere iki sayısal gösterim vardır. TF değeri, bir kelimenin bir belgede görünme sayısını, DF değeri ise bir kelimenin göründüğü belge sayısını ifade etmektedir.

Aynı zamanda TF-IDF yöntemi, belgelerden anahtar kelime(ler) çıkarma yöntemi olarak da kullanılmaktadır. Konunun genişliğine veya derinliğine bağlı olarak kelimenin belgeler içerisindeki değerinin, belgeye verdiği katkı sayesinde, anahtar kelime veya belge indeks terimi olarak tanımlanmasını sağlamaktadır.

Word2Vec Modeli

Word2Vec az katmanlı yapay sinir ağı yapısını kullanarak gözetimsiz öğrenme yapan ve kelimelerin vektör olarak temsil edilmesini sağlayan bir modeldir. Diğer modellerden en büyük farkı, kelimeler arasındaki anlam benzerliklerinin kelime vektörleri arasındaki uzaklığın hesaplanmasıyla bulunmasıdır. Ayrıca kelimelerin anlamları da kelime vektörlerinin yönleri ile temsil edilmektedir. Başka bir ifadeyle, Word2Vec anlam olarak birbirine yakın kelimeleri birbirine yakın koordinatlarda kümelemektedir.

Word2Vec benzeri kelime vektörlerinden bazıları aşağıda listelenmiştir:

- GloVe (Global Vectors for Word Representation)
- FastText
- BERT
- RoBERT

Önceden Eğitilmiş Dil Modeli ve Önemi

Metin işleme ve doğal dil işlemedeki zorlukların başında eğitim verisinin yetersizliği ve bu veriye erişimdeki sorunlar gelmektedir. Bu noktada doğal dil işlemedeki

çalışmalar yön değiştirerek, dil görevlerine ayrı ayrı modeller geliştirmek yerine daha büyük bir çerçeveden bakılarak dilin modellenmesi üzerine genel bir yaklaşım tercih edilmeye başlanmıştır. Özellikle büyük bir modeli eğitmek için büyük miktarda veri gerekmektedir. Bu da hem zaman hem de hesaplama (donanım) kaynakları açısından çok maliyetli olması sebebiyle tercih edilmemektedir.

"Önceden eğitilmiş dil modelleri (pre-trained language models)" bu amaçla, büyük miktarda metin verisi ile eğitilen genelleştirilmiş dil modeli sunarak ilgili dildeki tüm dilbilimsel yapıyı bir çerçevede ele almış ve ihtiyaç duyulduğunda diğer dil işleme görevleri için transfer edilmesine olanak sağlamıştır. Bu sayede farklı farklı dil görevleri için model eğitimi minimum düzeye indirilmiştir.

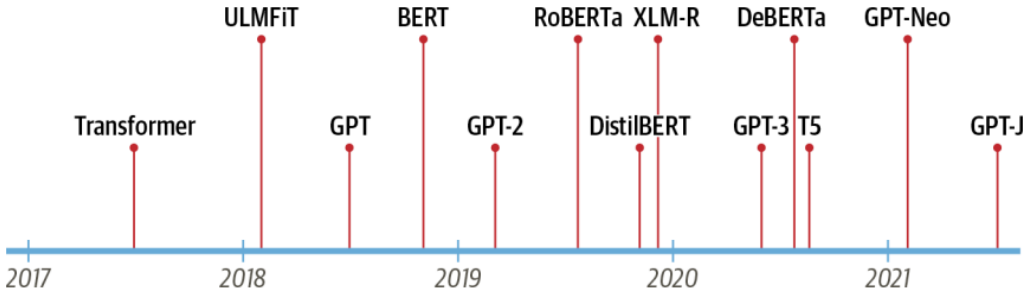
Doğal dil işleme alanında 2018 yılında peş peşe tanıtılan ELMo, ULMFIT, OpenAI Transform ve son olarak BERT gibi önceden eğitilmiş dil modelleri çok farklı dil görevlerinde büyük başarı sağlamıştır. Örneğin; 2018 Ekim ayında Google tarafından tanıtılan **BERT (Bidirectional Encoder Representations for Transformers)** modeli, doğal dil işleme için geliştirilen etkin transfer öğrenme metotlarından biri olmuştur. Bu model BooksCorpus (800 milyon kelime) ve İngilizce Wikipedia (2.5 milyar kelime) dahil, toplam 3.3 milyar kelimelik bir korpus üzerinden eğitilmiştir.

Önceden eğitilmiş dil modelleri ilk oluşturulmaya başladığında genellikle İngilizce metinler üzerinde eğitimler gerçekleştiriliyordu. Son yıllarda yapılan çalışmalarla birlikte Türkçe'nin de dahil olduğu çok sayıda dildeki metinler üzerinde eğitimler gerçekleştirilmekte ve dillerin modelleri oluşturulmaktadır. Bu sayede farklı alanlardaki doğal dil işleme çalışmaları için fayda sağlayacak hazır dil modelleri elde edilmiş olmaktadır. Bu dil modelleri, doğal dil işleme alanında bakış açısının değişmesine ve birçok görevde elde edilen başarıların oldukça artmasına ve en önemlisi de her görev için büyük eğitim kümesine olan ihtiyacın azalmasına yol açmıştır.

EBYS'lerdeki yapay zekâ ve veri madenciliği uygulamaları için büyük önem arz ettiği düşünülen önceden eğitilmiş dil modelleri sayesinde, tez kapsamında yapılan birçok metin işleme görevinin başarı oranlarının arttığı ilk denemelerden itibaren açıkça gözlemlenmiştir. Çalışmamızın gerçekleştiği süreçte ortaya çıkan ve gelişen bu dil modellerinin EBYS alanında yapay zekâ, makine öğrenmesi ve veri madenciliği ile yürütülecek çalışmalara çok büyük katkı sağlayacağı açıktır ve bu durumdan sonuç bölümünde ayrıca söz edilecektir. Önceden eğitilmiş dil modelleri EBYS verileri

üzerinde özelleştirilerek birim, kullanıcı, standart dosya planı kodu ve konu bazlı metinlerin oluşturulmasıyla EBYS'lerin daha verimli kullanılması mümkündür.

İncelenen dil modellerine Şekil-2'den bakılabilir.



Şekil 2: Dil Modellerinin Gelişim Süreci

2.1.5.3. Veri Madenciliğinde Yaşanan Sorunlar

Veri madenciliği uygulamalarının sonuçlarının kaliteli ve başarılı olması için verinin yukarıda söz edilmiş olan ön işlemlerden geçirilmesi gerekmektedir. Bu ön işlem adımları veri temizleme, veri entegrasyonu (bütünleştirme), veri seçimi ve veri dönüşümü şeklinde sıralanabilir. Veri madenciliğinin önemli bir aşamasını teşkil eden veri ön işleme, veri madenciliği sürecinin yaklaşık %60 - %80'ini kapsamaktadır. Bu orandan da anlaşılacağı üzere hem ilk aşama olması hem de zaman bakımından büyük bir yüzdeyi teşkil etmesi, ön işlem adımlarının ne kadar kritik olduğunu göstermektedir. Burada yapılacak yanlışlar, sonuçların kalitesiz ve başarısız olmasına sebep olacaktır. Bu yüzden veri ön işleme adımlarının her birini özenle yerine getirmek gerekmektedir.

Veri temizleme adımında verideki gürültü giderilmekte, eksik veriler birtakım yöntemlerle doldurulmakta ve tutarsızlıklar düzeltilmektedir. Bu işlemleri yaparken seçilecek yöntemler de sonuçları doğrudan etkilemektedir. Bu bağlamda karşımıza çıkabilecek sorunlardan biri, eksik veri içeren kayıtları silme durumunda elimizde bulunan veriden değerli kısımları kaybetme riskidir. Bunun yanında, eksik verileri elle doldurma kısmında yanlış veriler işlenmesi söz konusu olabilir ve bu durum da sonuçları olumsuz bir biçimde etkilemektedir. Eksik veriyi doldurmak için sabit bir değer kullandığımızda da eğer yanlış bir değer seçersek model sonucu yine olumsuz yönde etkilenecek ve sonuçlar başarısız ya da kalitesiz olacaktır.

Farklı veri kaynaklarından veri alma işlemlerinin gerçekleştirildiği veri entegrasyonu sürecinde karşımıza çıkabilecek sorunlardan biri, söz konusu kaynaklardan

bir ya da birkaçında alan adları aynı olsa da alanın altındaki verilerin farklı olabilmesidir. Bu durumda, ilgili alanları tek tek kontrol etmek gerekebilir. Örneğin, belirli bir kaynaktan aldığımız bir belgenin üst verisi ile bir başka kaynaktan aldığımız belgenin aynı isimli üst verisini eşleştirmemiz gerektiği durumlarda, söz konusu üst verilerin altındaki diğer veriler birbirinden farklı olabilir. Veri entegrasyonu adımına da tam olarak bu yüzden ihtiyaç duyulmaktadır. Bir başka örnek olarak birleştireceğimiz farklı kaynaklardaki verilerde alanların birbirleriyle nasıl eşleşmesi gerektiği de sorunlarla karşılaşma ihtimalimiz olan bir diğer alandır. Yalnızca alan adına bakarak bir kaynaktan gelen verideki alanı, bir diğerindeki alanla eşleştirmek yerine, bu alanların altında yer alan verileri de kontrol ederek sürecin ilerletilmesi gerekmektedir. Çünkü bir kaynakta daha fazla sayıda veri niteliği bulunurken diğer kaynakta daha az veri niteliği olabilir.

Veri seçimi adımında gereksiz özellikleri ortadan kaldırma ve kümeleme yapılarak veri boyutu azaltılmaktadır. Boyut azaltma, veri madenciliği için önemli bir yöntemdir. Önemli olmasının sebepleri aşağıdaki gibi sıralanabilir;

- Veriler çok fazla boyuta (özniteliğe) sahip olmakta ve boyut büyüdükçe veri temizlemeden model kurmaya bütün süreçlerde harcamamız gereken zaman ve kaynaklar artmaktadır.
- Verinin çok fazla sayıda boyuta sahip olması, veri görselleştirmenin de oldukça zor bir şekilde yapılmasına neden olmaktadır. 3 boyuttan sonrasını düşünmek zordur, fakat yapılan çalışmanın da bir şekilde görselleştirilmesi gerekmektedir.
- Hemen her veri setinde bazı öznitelikler arasında yüksek korelasyon bulunmaktadır. Bu da tekrar eden bilgiler ortaya çıkarmakta ve oluşturulan modelde aşırı uyum (overfitting) problemine sebep olmaktadır.

Veri seçiminde yapılan iş, veriyi en iyi tanımlayan öznitelikleri bulup diğerlerini atmaktır. Yani, öznitelikler arasında veriyi en iyi temsil ettiği düşünülen öznitelikleri seçmektir. Örnek olarak, 50 boyutlu bir veri setinden 20 tanesinin önemli olduğunu belirleyip kalan 30 özniteliği atmak burada tercih edilecek çözümdür. Fakat yukarıda da belirttiğimiz üzere bu çözümün de veri kaybına sebep olabileceği unutulmamalıdır.

Veri dönüşümü aşamasında karşılaşılabilecek sorunlardan biri, normalleştirmede alınan belirli bir aralığın beklenen sonucu vermemesi şeklinde gerçekleşebilir. Örneğin farklı alanlarda oluşturulan normalleştirme aralıklarının birbirleriyle uyumlu olmaması

hesap hatalarına sebep olabilmektedir. Bunun yanı sıra birleştirme sürecinde, verileri kayıt bazında değil, toplam olarak ele almak da söz konusu aşamada veri kayıplarına ve bazı önemli ve değerli bilgileri kaçırma ihtimaline sebep olabilmektedir.

Sonuç olarak veri madenciliği uygulamalarında karşılaşılabilecek bu türden sorunların, arzu edilen kaliteli ve başarılı sonuçların elde edilmesine engel olduğu ortadadır. Dolayısıyla ön işlem süreçlerinin titiz ve tutarlı bir biçimde gerçekleştirilmesi birçok sorunun meydana gelmeden çözülebilmesini sağlayacaktır.

2.1.5.4. Veri Tabanı Tasarımında Dikkat Edilmesi Gereken Hususlar

Verileri tuttuğumuz veri tabanlarının tasarımı da sonuçlar üzerinde en az veri madenciliği ön işlemleri kadar etkilidir. Veri tabanı tasarımı ne kadar iyi ve tasarım kurallarına bağlı kalınarak hazırlanmışsa, o veri tabanlarından alınan verilerin kullanıldığı veri madenciliğinden elde edilen sonuçlar da o kadar kaliteli ve başarılı olacaktır. Ayrıca bu tasarım veri üzerinde yapılacak ön işlemlerde harcanacak zamanın azalmasını sağlayacaktır.

Veri tabanı tasarımında dikkat edilmesi gereken hususlardan bazıları aşağıda listelenmiştir:

- Tekrar eden verilerin önüne geçilmelidir. Bu veriler hem veri tabanı alanını boşa harcamakta hem de hataları ve veri tutarsızlıklarını arttırmaktadır.
- Verilerin doğru bir biçimde girilmesi sağlanmalı ve eksiksiz olmasına önem gösterilmelidir. Doğru olmayan veriler, bu verilerin kullanıldığı her yerde yanlış sonuçlar alınmasına neden olacaktır. Eksik veriler ise tüm raporlarda ve sorgularda doğru olmayan ve verinin bütünlüğünü bozacak problemlere neden olacaktır. Bu verilere dayanarak alınacak kararlar veya yapılacak işlemler de eksik ve yanlış olacaktır.

Verinin tutulduğu tablolarda bulunan her bir sütunla ilgili boş bırakılabilme/bırakılamama, tekil olma, belli kurallar dahilinde veri girişinin sağlanması ve varsayılan değer ataması gibi kısıtların belirlenmesi gerekmektedir. Örneğin, bu kısıtlar sayesinde veri girişinin zorunlu olduğu bir alanın boş bırakılmasının önüne geçilecektir. Bu sayede veri ön işlem aşamalarından biri olan eksik verilerin tamamlanması sürecine ihtiyaç kalmayacaktır.

Sonuç olarak aslında bütün sürecin zemininde yer alan veri tabanı tasarımının, modelin üreteceği sonuçlar üzerinde oldukça etkili bir unsur olduğu açıktır. Bu yüzden

tıpkı veri ön işleme süreçlerinde olduğu gibi bu aşamanın da kural tabanlı, tutarlı ve titiz bir şekilde ortaya konması kalite ve başarı oranını doğrudan etkileyecek ve yukarıda söz ettiğimiz sorunlar meydana gelmeden verileri elde edebilmemizi olanaklı hale getirecektir.

2.2. MAKİNE ÖĞRENMESİ

2.2.1. Makine Öğrenmesi Kavramı ve Tarihsel Gelişimi

Son yıllarda bilgi ve iletişim teknolojileri alanındaki radikal değişim ve gelişimler, gündelik yaşamımızın bütün alanlarında dönüştürücü etkilerini ortaya koymaktadır. Kesintisiz bir yenilik ve ilerleme ile her yana yayılan bu dönüşüm yapay zekâyla ilişkili kavramların daha görünür olması ve sürekli kullanılmasına sebep olmuştur.

Günümüzde veri üretimi eski dönemlerle kıyaslanamayacak boyutlara ulaşmıştır. Medikal teknolojilerden finansa, ulaşımdan spora kadar birbirinden bütünüyle ayrı sahalarda üretilen veri büyük miktarda ve çok çeşitlidir. Söz konusu koşullar, bu verileri kullanarak faydalı bilgiler elde etmek ve bunları stratejik anlamda kullanmak için büyük bir fırsat yaratmıştır. Bu bağlamda bilgiden söz ettiğimizde, belirli hedeflere varmak veya kayda değer bir anlayış ortaya koymak adına verinin, bir dönüşüm ve analiz süreci sonucunda faydalı bir biçime sokulması söz konusudur (Gökçen, 2011). Ayrıca, donanım teknolojilerindeki yenilikler, bu geniş veri setlerini hızla işleyerek değerli içgörüler elde etme kapasitesini de artırmıştır.

Belirttiğimiz koşullar altında günümüzün en kıymetli cevheri olarak görülebilecek bilgiyi ele geçirebilmek için farklı alanlarda atılımlar da yoğunluk kazanmıştır. Bu bağlamda ilk olarak veri bilimi alanı anılabilir. Karmaşık problemleri çözmek için tüm veri türlerini faydalı, anlamlı ve değerli bilgiye dönüştürmeye yarayan veri biliminin yanı sıra anmamız gereken bir diğer alan da makine öğrenmesidir. Makine öğrenmesi, kısaca bilgisayarların örnek veri ya da geçmiş deneyimleri kullanarak belirli bir ölçüte göre programlanması şeklinde tanımlanabilir (Alpaydın, 2009).

Alan Turing, yaşamı boyunca hak ettiği ilgiyi görmemiş bir İngiliz matematikçi, mantıkçı ve kriptoloji uzmanıdır; ancak günümüzde bilgisayar biliminin ve yapay zekanın öncüsü olarak tanınmaktadır (BBC, 2019). 1936'da yayımladığı bir makalede, insanların belirli bir işi ya da görevi nasıl tamamladığına dair izlediği yöntem ve süreçleri incelemiştir. Bu çalışma, her türlü komut dizisini çözümleyebilen ve belirtilen görevi gerçekleştirebilen bir "Evrensel Makine" konseptini doğurmuştur (Turing, 1937). Bu

bahsi geçen makale, günümüzde bilgisayar biliminin üzerinde inşa olunduğu zemin olarak görülmektedir.

1950 yılında Alan Turing, "Computing Machinery and Intelligence" başlıklı makalesinde, yapay zekanın nasıl tanımlanabileceği ve değerlendirilebileceği üzerine kapsamlı bir analiz sundu. Bu makalede ortaya koyduğu "Taklit Oyunu" kavramı, daha sonra "Turing Testi" olarak bilinir hale geldi ve bir yapay zekanın gerçekten zeki olup olmadığını belirlemek için bir ölçüt olarak kabul edildi. Turing'in bu öncü yaklaşımı, yapay zekâ araştırmalarının temelini oluşturmuş ve bu alandaki ilerlemelerin yönünü belirleyici bir şekilde etkilemiştir.

Dünya satranç şampiyonu Garry Kasparov'un 1997 yılında IBM'nin Deep Blue isimli bilgisayarına yenilmesiyle yapay zekâ kavramına yönelik farkındalık ve ilgi de artmaya başlamıştır (Google Cloud, 2017). 2014 yılına gelindiğindeyse ekrandaki piksel hareketlerini analiz edip video oyunları oynamayı öğrenebilecek bir sinir ağı geliştirerek ön plana çıkan DeepMind, Google tarafından satın alındı (Wiki, 2014). DeepMind'ın araştırmacı ekibi tarafından tasarlanan AlphaGo, sinir ağı teknolojisi kullanılarak hem insanlarla hem de bilgisayarlarla oynayarak eğitildi. Bu yapay zekâ tabanlı program, 2015'te profesyonel Go oyuncularına karşı ilk zaferini elde etti. Bu başarısını 2016'da dünya sıralamasında ikinci olan Lee Sedol'u mağlup ederek ve 2017'de bir numaralı oyuncu Ke Jie'yi yenerek sürdürdü (BBC, 2018).

Tarihsel seyrini özellikle popüler kültürdeki karşılıkları ile böylelikle kısaca özetlediğimiz gelişmelere paralel biçimde makine öğrenmesi, derin öğrenme ve transfer öğrenme gibi alanlarda yapılan çalışmalar da gün geçtikçe çoğalmakta ve çeşitli ihtiyaçları karşılamaktadır.

Makine öğrenmesi, belirli bir problemi çözmekte kullanılan algoritmaların performanslarını bizzat iyileştirerek daha başarılı çözümler üretmesiyle ilgilenen bir bilimsel çalışma olarak da ifade edilebilir. Bu bağlamda, algoritmaların farklı koşullara uyum yeteneği kazanması sağlanarak daha gelişmiş ya da başka bir ifadeyle çalıştığı ortamı öğrenen algoritmalar oluşturmak istenmektedir. Aslında söz konusu algoritmalar, eğitim verileri olarak adlandırılan veri setleri üzerinde çalışarak tahminde bulunmak ve/veya karar vermek için söz konusu verilerin matematiksel bir modelini üretmeye çalışmaktadır. Matematik model oluşturulduktan sonraki adım, test verilerinin üzerinde algoritmaların çalıştırılması ve sonuçların değerlendirilmesidir.

Veriden bilgi elde edilebilmesini sağlayan şey kural öğrenmedir. Kuralı, veriyi açıklayan basit bir model olarak tanımlayabiliriz ve bu modele bakarak verinin altında yatan süreç hakkında bir açıklama elde ederiz. Verinin altında yatan bu süreci tanımlamak imkân dahilinde olmasa da veride önemli örüntü ve düzenlilikleri tespit edebileceğimizi varsayarız. Makine öğrenmesinin iş gördüğü alan da tam olarak burasıdır.

Veriyi değerli ve faydalı kılan şey onda mevcut olan düzendir. Bilhassa, belli standartlar dahilinde oluşturulmuş sistemlerde, örneğin EBYS'de, bu türden pek çok düzenin var olduğunu rahatlıkla dile getirebiliriz. Burada keşfedilecek bir düzen, süreci anlamamıza yardımcı olacak ve onun aracılığıyla öngörüde bulunmak olası hale gelecektir: Böylece, eğer yakın geleceğin, verinin elde edildiği geçmişten çok farklı olmadığını varsayacak olursak bu veriden oluşturulacak kural setiyle gerçekleştireceğimiz tahmin ve öngörüler de doğru olacaktır.

Makine öğrenmesi ile oluşturulan ve mevcut bazı parametrelere bağlı olarak tanımlanmış bir model ile veri ya da geçmiş deneyim üzerinde modelin başarımını ölçmek için tanımlı bir ölçüte ihtiyaç vardır. Burada amaç, modelin parametrelerini bu başarım ölçütüne göre en iyi şekilde işleten parametre değerlerini elde etmektir. Model, gelecekle ilgili öngörüler yapmak için kullanılabilen öngörücü veya veriden bilgi çıkarmaya yönelik açıklayıcı bir model olabilir.

Makine öğrenmesinin kullanıldığı bir başka işlev de kurala uymayan ve ender rastlanan aykırılıkların tespitidir. Bu türden bir uygulamada veriden bir kural öğrendikten sonra kuralla değil, kurala uymayan aykırı örneklerle ilgileniriz. Sahtekârlık, işlem anomalileri gibi olağandışı durumların tespit edilmesi bu başlık altında ele alınabilir.

Günümüz teknolojisinin vardığı noktada verinin çok büyük oluşu, gerekli bilgiye sahip insanların nadirliği ve elle işlemenin oldukça masraflı yapısı yüzünden veri işleme ve bilgi çıkarma işlemlerinim insanlar tarafından yürütülmeye devam etmesi artık olanaklı değildir. Tam olarak bu sebeple veriyi inceleyip bilgi çıkarabilecek, yani "öğrenecek" bilgisayar yazılımlarına duyulan gereksinim her zamankinden çok daha fazladır.

Makine öğrenmesinin son yıllarda çeşitli uygulamalardaki başarıları ve bu uygulamalara yönelik yapılan değerlendirmeler, aslında bize ihtiyaç duyduğumuz şeyin yeni algoritmalardan ziyade büyük miktarda örnek veri ve yeterli hesaplama gücü olduğunu göstermektedir. Örneğin, bugün kullanmakta olduğumuz destek vektör

makinelerinin kökenleri, 1950 ve 1960'larda önerilmiş potansiyel işlevler, doğrusal modeller ve en yakın komşu tabanlı yöntemlere dayanmasına rağmen, o günlerde bu algoritmaların potansiyellerini bütünüyle ortaya koyabilmelerine imkân verecek hızlı bilgisayarlar ve yeterli veri saklama yeteneği olmadığı için bugün gördükleri işlevi görmüyorlardı (Alpaydın, 2009).

Böylece, makine öğrenmesinin tarihsel gelişimi ve temel bir girişten sonra, makine öğrenmesi işlevlerini ele alabiliriz.

2.2.2. Makine Öğrenmesi İşlevleri

Bilgisayar teknolojisinin ve donanım altyapısının ilerlemesi sayesinde, büyük veri kümeleri kolaylıkla depolanabilir, işlenebilir ve makine öğrenmesi algoritmalarıyla etkili sonuçlar alınabilir hale gelmiştir. Bilgisayarlarda bir görevin yerine getirilmesi veya bir sorunun çözülmesi, sıralı komutlar bütünü olan "algoritma"lar aracılığıyla gerçekleşir. Bazı işlemler için elimizde belirli algoritmalar bulunsa da bazı işlemler için henüz tanımlanmış bir algoritma olmayabilir. Bu tür durumlarda, mevcut veriyi analiz ederek sorunun nasıl ele alınacağını "öğrenmek" esastır. Diğer bir deyişle, bilgisayardan belirli bir işlem için otomatik olarak bir algoritma oluşturmasını talep etmek olanaklıdır.

Problemin çözümüne yönelik öğrenme sürecinde, çözüm yolunun nasıl izlendiği belirlenir. Her zaman bu süreci kesin bir biçimde tanımlamak mümkün olmasa da yaklaşık bir açıklama sağlanabilir. Bu yaklaşık açıklama, tüm veri seti için kapsamlı olmasa da verinin büyük ve kritik bir bölümü için genellikle yeterlidir. Burada ana amaç, veri içerisindeki kritik örüntüleri ve düzenlilikleri keşfetmektir. Dolayısıyla, makine öğrenmesinin iş gördüğü alanı bu şekilde ifade etmek de mümkündür (Alpaydın, 2009).

Tüm akıl sahibi canlıların doğal bir biçimde edinmiş olduğu önceki tecrübelerden öğrenme yeteneğinin makineler için veri aracılığıyla gerçekleştirildiği bir eğitim metodu olarak da tanımlanabilecek makine öğrenmesi, temelinde algoritmaların, matematiğin ve istatistiğin yer aldığı bir veri analizi yöntemi ve yapay zekânın bir alt dalıdır.

Makine öğrenimi algoritmalarının temel amacı, bilgisayarlara, sunulan girdi verilerindeki yapıları "öğrenme" ve bu öğrenilen bilgiyi genelleştirme yeteneği kazandırmaktır. Öğrenme kavramı, deneyimin bilgi veya uzmanlık haline getirilmesi süreci olarak tanımlanabilir. Bu perspektiften bakıldığında, bir makine öğrenme algoritmasında girdi, deneyimi yansıtan veri setleridir; çıktı ise, belirli görevleri yerine

getirebilen, bu deneyimden türetilen bilgi veya uzmanlıkla şekillenen bir bilgisayar programıdır (Shalev-Shwartz ve Ben-David, 2014).

Makine öğrenmesi algoritmaları, öğrenme sürecinin nasıl gerçekleştirildiğine göre çeşitli kategorilere ayrılır (Alpaydın, 2009). Literatürde bazen dört veya beş başlık altında ele alınsa da çoğunlukla aşağıdaki tasnif takip edilmektedir.

1. Denetimli (Gözetimli) Öğrenme

2. Denetimsiz (Gözetimsiz) Öğrenme

3. Takviyeli (Pekiştirmeli) Öğrenme

Bu bağlamda denetimli öğrenme ile denetimsiz öğrenme arasındaki temel farkı kısaca dile getirebiliriz. Denetimli öğrenmede amaç, girdi ile bir gözetmen tarafından doğru değeri verilmiş çıktı arasında bir eşleşme öğrenmektir. Denetimsiz öğrenmede ise böyle bir gözetmen bulunmaz ve elimizde sadece girdi verisi mevcuttur; amaç, girdideki düzenlilikleri tespit etmektir. Şimdi bu alt kategorilerden ilk ikisini, yani denetimli ve denetimsiz öğrenmeyi detaylı olarak ele alabiliriz.

Denetimli (Gözetimli) Öğrenme

Denetimli öğrenme, belirli girdi değerlerinin hangi çıktı değerleriyle eşleştiğinin bilindiği durumlar için uygulanan bir öğrenme yaklaşımıdır. Bu yöntem, eğitim verisindeki girdi-çıktı çiftlerini kullanarak bir fonksiyon oluşturur ve bu fonksiyonu, daha önce görülmemiş test verisi üzerinde değerlendirir. Bu tür öğrenme, etiketli veri kümeleri üzerinde çalışır, yani her girdi değeri için bir çıktı değeri (etiket) bilinir. Denetimli öğrenme algoritmalarının temel amacı, girdi ve çıktı arasındaki ilişkiyi en iyi şekilde yakalayan matematiksel bir modeli (fonksiyonu) belirlemektir. Bu model, yeni girdi verileri geldiğinde doğru çıktıları tahmin etmek için kullanılır (Mohri, Rostamizadeh ve Talwalkar, 2018).

Denetimli öğrenmeyi somut bir temsil ile açıklamak gerekirse, şu örneği kullanabiliriz: Bir sürücü kursunda eğitim alan bir öğrenciyi ve onun eğitmenini düşünelim. Bu durum, denetimli öğrenmenin bir yansıması olarak kabul edilebilir. Eğitmen, eğitim veri kümesini temsil ederken; öğrenci, bu veri kümesi üzerinde çalışan makine öğrenme algoritmasını temsil eder. Eğitmenin, öğrenciye trafik kurallarını ve sürüş tekniklerini öğretmesiyle öğrenci, güvenli bir biçimde araç kullanma yeteneğine

sahip olur. Benzer şekilde, algoritma da eğitim veri kümesi sayesinde belirli bir örüntüyü öğrenmektedir.

Denetimli öğrenme algoritmaları uygulamalarına bir başka örnek olarak nesnelerin görüntülerini sınıflandırma problemi verilebilir. Bu senaryoda, denetimli öğrenme algoritması, her bir görüntünün hangi nesneyi temsil ettiğini belirten etiketlerle birlikte sunulan eğitim setini kullanarak görüntü ve nesne türleri arasındaki ilişkiyi öğrenir. Eğitim sürecinin ardından algoritma, daha önce rastlamadığı etiketsiz yeni nesne görüntülerini doğru bir şekilde sınıflandırma yeteneğini kazanmaktadır.

Bu bağlamda, denetimli öğrenme algoritmalarını ele alarak başlayabiliriz. Bu algoritmalar kendi içinde iki ana kategoriye ayrılmaktadır; Sınıflandırma ve Regresyon. İlk olarak sınıflandırma algoritmaları altında yer alan başlıca birkaç örneği kısaca değerlendirebiliriz.

Sınıflandırma

Sınıflandırma algoritmaları, belirli etiketlere sahip eğitim veri setini kullanarak modelin eğitilmesini sağlayan ve yeni, etiketsiz verinin hangi kategoriye ait olduğunu tahmin eden algoritmalardır. Eğer veri setinde sadece iki kategori (örn. erkek/kadın, hasta/sağlıklı, pozitif/negatif vb.) bulunuyorsa bu ikili sınıflandırma olarak tanımlanır. Ancak, kategorilerin sayısı ikiden fazla olduğunda bu durum çoklu sınıflandırma olarak adlandırılır (Aggarwal, 2014).

Sınıflandırma algoritmalarının uygulama alanlarından biri; bir bankanın müşterilere kredi verip vermemesi kararının belirlenmesi problemidir. Banka, müşterinin kredi geçmişi, gelir seviyesi, mevcut borçları ve diğer finansal göstergeler gibi parametreleri dikkate alarak bir değerlendirme yapar ve bu değerlendirmeye göre müşteriye kredi "verilir" ya da "verilmez". Bu durumda, "kredi verilecek" ve "kredi verilmeyecek" kategorileri, problemdeki sınıf etiketlerini temsil eder ve bu ikili sınıflandırmaya bir örnektir.

Çoklu sınıflandırma için bir örnek verecek olursak; bir e-ticaret platformunda, otomatik şekilde duygu analizi yaparak ürün incelemelerini "olumlu", "olumsuz" ve "nötr" kategorilerine ayırmak istendiğini varsayabiliriz. Bu durumda kullanıcılar tarafından yazılan yorumlar analiz edilir ve bu yorumlar genel duygusal tonuna göre bir kategoriye atanır ve "olumlu", "olumsuz" ve nötr" kategorileri, problemdeki sınıf etiketlerini temsil eder.

En çok kullanılan sınıflandırma algoritmaları şu şekilde sıralanabilir:

- Karar Ağaçları
- Lojistik Regresyon
- Naive Bayes Algoritması
- K-En Yakın Komşular Algoritması

Karar Ağaçları

Makine öğrenmesi algoritmaları içerisinde, insanın düşünce süreçlerini en yakından yansıtanın Karar Ağaçları (Decision Trees) olduğu söylenebilir. Günlük yaşantımızda basit gibi görünen birçok eylem, aslında ardışık karar alma süreçlerinden meydana gelir. Bu kararlar bazen basit ve doğrudan olabilirken, bazen de birçok bağlantılı kararı beraberinde getiren karmaşık bir yapıda olabilir (Şimşek, 2018).

Karar ağacı, bir kararın ve potansiyel sonuçlarının ağaç benzeri bir grafik veya model aracılığıyla gösterildiği bir sınıflandırma algoritmasıdır. Bu, algoritmanın görsel bir biçimde sunulmasına olanak tanır. Karar ağaçlarının temel amacı, son karara ulaşmak için izlenmesi gereken tüm potansiyel yolların bir haritasını oluşturmaktır.

Karar ağacı, büyük bir veri setini, belirli karar kurallarını takip ederek daha yönetilebilir alt kümelerine bölmek için kullanılır. Bu yapı, hiyerarşik bir yaklaşım sunarak en üstten başlayıp en alta doğru ilerler (Kantardzic, 2011).

Lojistik Regresyon

Lojistik regresyon, temel sınıflandırma algoritmalarından biridir. Bu algoritma, çıktı değişkeninin kesin değerini tahmin etmekten ziyade, belirli bir değerin ne olasılıkla gerçekleşebileceğini belirlemeye odaklanır. Genelde, tahmin edilen değişkenin iki olası değeri olduğunda, yani daha önce bahsettiğimiz ikili sınıflandırma problemlerinde tercih edilir (Kantardzic, 2011).

Örneğin, bir hastanın belirli bir tedaviye olumlu yanıt verip vermeyeceğini tahmin etmek yerine, lojistik regresyon yöntemi kullanılarak tedavi süresine bağlı olarak tedavinin başarılı olma olasılığı hesaplanmaya çalışılmaktadır.

Regresyon

Regresyon, temelde istatistiksel bir analiz yöntemi olup, makine öğrenmesi uygulamalarında da sıkça kullanılmaktadır. Bu yöntem, bir bağımlı değişkenin (çıktı) bir ya da birden fazla bağımsız değişken (girdi) ile nasıl ilişkilendiğini modellemeyi amaçlar.

Regresyon analizinde, bağımsız değişkenlerin bağımlı değişken üzerindeki etkileri için ağırlık katsayıları belirlenir. Bu teknik, çıktının sayısal bir değer olduğu durumlarda, yani bir sonucun niceliksel olarak tahmin edilmesi gerektiğinde tercih edilir (Harrington, 2012).

Regresyon yöntemi, birçok uygulama senaryosunda kullanılmaktadır. Örneğin, bir şirketin reklam harcamaları ve web sitesi ziyaretçi sayıları dikkate alınarak gelecek ayın ürün satış miktarını tahmin etme problemi, regresyon analizinin kullanılabileceği durumlardan biridir. Bu tür bir analiz, bağımsız değişkenler (reklam harcamaları ve web sitesi ziyaretçi sayıları) ile bağımlı değişken (gelecek ayın satış miktarı) arasındaki ilişkiyi belirlemek için kullanılır.

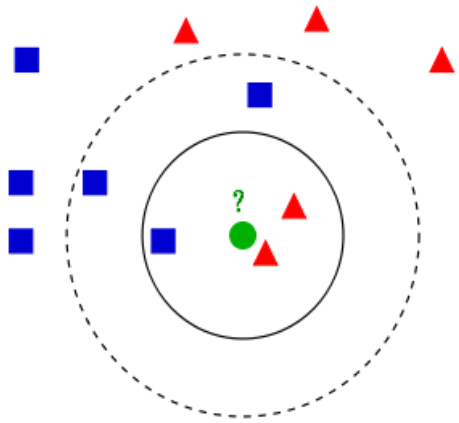
En çok kullanılan regresyon algoritmaları aşağıda verilmiştir:

- K-En Yakın Komşular Algoritması
- Karar Ağaçları
- Lineer Regresyon
- Karar Destek Makinesi

K-En Yakın Komşular Algoritması

K-En Yakın Komşular (k-Nearest Neighbours - kNN) algoritması, T.M. Cover ve P.E. Hart tarafından 1967'de sunulmuştur. Algoritmanın temel çalışma mantığı şu şekildedir: Sınıfı belirlenmemiş bir veri noktası alındığında, bu nokta veri kümesindeki diğer tüm noktalarla karşılaştırılır. Ardından, bu yeni veriye en benzer olan "k" adet veri seçilir ve bu verilerin sınıf etiketlerine göre bir çoğunluk oyu ile yeni verinin sınıfı belirlenir.

K-En Yakın Komşular algoritması hem sınıflandırma hem de regresyon görevleri için uygundur; ancak, literatürde bu algoritmanın sınıflandırma problemleri için daha sık tercih edildiği görülmektedir (Ray, 2017).



Şekil 3: k-En Yakın Komşular Örneği (Wiki)

Şekilde, örnek bir veri kümesindeki kayıtların hangi sınıfa ait olduğu verilmiştir. Mavi renkli kareler bir sınıfı, kırmızı renkli üçgenler ise başka bir sınıfı temsil etmektedir. Problem; yeni bir veri olan yeşil dairenin hangi sınıfa dahil edileceğine karar vermektir. Yeşil dairenin hangi sınıfa dahil olacağını tespit etmeyi sağlayan etmen, algoritmanın adında da geçen, "k" ifadesidir. "k"; kaç tane komşu veriye bakılacağına karar vermek için kullanılır.

Eğer k = 3 olarak seçilirse, siyah düz çizgiyle çizilmiş dairenin içinde 2 tane kırmızı 1 tane de mavi sınıf vardır. Algoritma, kırmızı sayısı fazla olduğu için yeşil daireyi de kırmızı üçgen sınıfına dahil edecektir.

Eğer k = 5 olarak seçilirse, siyah kesikli çizgiyle çizilmiş dairenin içinde 3 tane mavi 2 tane de kırmızı sınıf vardır. Algoritma, mavi sayısı fazla olduğu için yeşil daireyi de mavi kare sınıfına dahil edecektir.

Lineer Regresyon

Lineer regresyon, bir çıktı değişkeninin veri setindeki girdi değişkenlerine bağlı olarak nasıl değiştiğini belirlemek amacıyla kullanılan temel bir regresyon algoritmasıdır. Bu algoritma, nicel bir çıktı fonksiyonunun belirlenmesini hedefler. Bu hedefe ulaşmak için, girdi değişkenleri temel alarak bir çıktı denkleminin oluşturulması gerekmektedir (Harrington, 2012).

Evin konumu, oda ve banyo sayısı, yaşı, merkezî yerlere mesafesi gibi özellikler dikkate alınarak ev fiyatlarının tahmin edildiği bir problem burada örnek olarak verilebilir. Çıktı değeri olan evin fiyatını hesaplayabilecek formülün bulunmaya çalışıldığı bu örnekteki temel amaç, bir değişkendeki artış veya azalışın ev fiyatına nasıl

etkide bulunduğunu doğru biçimde tespit edecek katsayılar elde etmek ve yeni bir ev fiyatını tahmin etmeye çalışmaktır.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Formülde, bir evi niteleyen bazı özellikleri aşağıda verilmiştir:

- y : Ev Fiyatı
- x_1 : Oda sayısı
- x_2 : Evin yaşı
- …
- x_n : Merkeze olan mesafe

Lineer regresyon algoritmalarında tespit edilmek istenen katsayılar, formülde evin niteliklerinin fiyata olan etkisini (artış veya azalış) hesaplayacak olan β_0, β_1 gibi katsayılardır. Oda sayısı arttıkça evin fiyatının artması ya da yaşı arttıkça evin fiyatının azalmasını belirten katsayıların hesaplanmaya çalışılması örnek olarak anılabilir.

Denetimsiz (Gözetimsiz) Öğrenme

Etiketlenmemiş veri örneklerini içeren veri setlerini işlemek için başvurulan öğrenme yaklaşımı denetimsiz öğrenme olarak tanımlanır. Bu tür eğitim verilerinde sadece girdi değerleri bulunurken, bu girdilere karşılık gelen çıktı değerleri mevcut değildir. Denetimsiz öğrenme algoritmalarının ana hedefi, benzer özelliklere sahip girdileri gruplandırmaktır (Alpaydın, 2009).

Denetimsiz öğrenme yaklaşımını, eğitmen rehberliği olmadan sürücünün kendi başına araba sürmeye çalıştığı bir senaryo ile açıklayabiliriz. Bu durumda eğitmen etiketsiz veri kümesi problemini, sürücü ise bu veri kümesi üzerinde çalışan makine öğrenmesi algoritmasını temsil etmektedir.

Denetimsiz öğrenme algoritmalarının uygulama alanlarına bir örnek olarak, e-ticaret platformlarındaki müşteri segmentasyonu gösterilebilir. Örneğin, bir e-ticaret platformu, kullanıcıların site içi davranışlarına (ürün aramaları, gezinme süreleri, inceledikleri ürünler, sepete eklemeleri, satın alma eğilimleri) ve demografik bilgilerine (cinsiyet, yaş, coğrafi konum) dayanarak, benzer davranış özelliklerine sahip müşteri grupları oluşturabilir. Bu gruplamaların sonucunda platform, farklı müşteri segmentlerine özelleştirilmiş pazarlama stratejileri ve teklifler sunabilir. Bu örnek denetimsiz öğrenme algoritmalarının potansiyel uygulama alanlarından sadece biridir.

Denetimsiz öğrenme algoritmaları, genel olarak "kümeleme" ve "birliktelik kuralı çıkarımı" başlıkları altında incelenmektedir.

Kümeleme

Kümeleme algoritmaları, veri seti içerisindeki örneklerin niteliklerine dayanarak benzerliklerini belirlemeyi ve bu benzerliklere göre de gruplandırmayı amaçlamaktadır. Bu algoritmalarda, veri örneklerinin niteliklerindeki yakınlık veya uzaklık bilgileri temel alınarak benzerlik hesaplamaları yapılır. Bu sayede, birbirine benzeyen veri örnekleri aynı kümeye atanırken, birbirinden farklı olan örnekler ise farklı kümelerde sınıflandırılır. Bu yaklaşım, veri setlerindeki doğal yapıları veya örüntüleri ortaya çıkarmak için oldukça etkilidir (Xu ve Wunsch, 2005).

Optimal bir kümeleme, aynı kümeye dahil edilen örnekler arasında yüksek benzerliklerin ve farklı kümelerdeki örneklerle ise düşük benzerliklerin bulunduğu durumda elde edilir. Diğer bir deyişle bir küme içerisindeki örnekler birbirine oldukça benzerken, farklı kümelerdeki örneklere benzerliği minimum seviyede olmalıdır (Kılınç ve Başeğmez, 2018).

Kümeleme, temelde bir veri madenciliği yöntemi olarak bilinse de, makine öğrenmesi, görüntü işleme, bilgi çıkarımı, web araştırması ve biyoinformatik gibi geniş bir yelpazede farklı uygulama alanlarında başarıyla kullanılabilmektedir (Han ve diğerleri, 2011).

Kümeleme yöntemine sağlık sektöründen bir örnek vermek gerekirse hastanelerin, hastaların tıbbi geçmişi, yaş, cinsiyet, yaşam tarzı ve genetik bilgileri gibi verilere dayanarak "yüksek riskli", "orta riskli" ve "düşük riskli" şeklinde gruplara ayırılması gösterilebilir. Bu gruplandırma, hastalara özel tedavi planları oluşturmak veya önleyici sağlık programlarına yönlendirmek için kullanılabilir.

K-Means Kümeleme

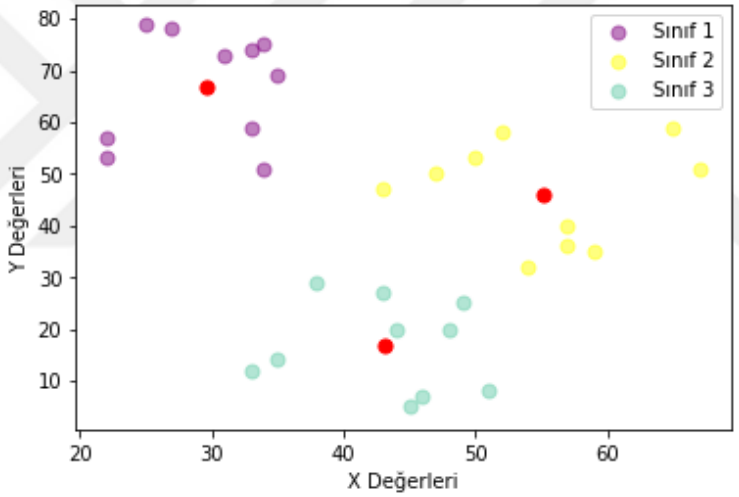
K-Means algoritması, belirli bir veri setindeki etiketsiz verileri "k" sayıda kümeye ayırma yöntemidir. Bu "k" değeri, kümelerin sayısını belirtir ve kullanıcının belirlediği bir parametredir. Her bir küme, "centroid" adı verilen bir merkez nokta etrafında oluşur. Centroid'ler, kümeye ait olan tüm veri noktalarının ortalamasını yansıtan bir konumu temsil etmektedir (Harrington, 2012).

K-Means algoritmasının işleyişi aşağıda sıralanan adımlarla gerçekleşmektedir: Başlangıçta, “k” miktarında merkez nokta (centroid) rastgele seçilir. Sonrasında, veri setindeki her bir veri noktası, kendisine en yakın olan merkez noktaya göre bir kümeye atanır. Bu atama, veri noktasının hangi merkez noktaya daha yakın olduğunun belirlenmesiyle yapılır. Atama işleminden sonra, her bir kümeye ait veri noktalarının ortalaması alınarak merkez noktaları (centroidler) yeniden hesaplanır ve güncellenir. Bu süreç, belirlenen bir kriter veya iterasyon sayısına ulaşana kadar devam eder.

K-Means algoritmasına şöyle bir örnek verilebilir. X ve Y niteliklerine sahip örneklerin olduğu etiketsiz veri setindeki içeriğin kümelenmesi ve belirlenen kümelerin merkez noktalarının (centroidlerin) yerleri aşağıda gösterilmektedir.

X	15	24	12	17	22	22	…	11
Y	69	41	43	68	48	63	…	46

Tablo 2: K-Means Kümeleme İçin Örnek Veri Seti



Şekil 4: K-Means Kümeleme Örneği

Birliktelik Kural Çıkarımı

Birliktelik kuralı çıkarımı, veri kümesi içerisinde sıkça bir arada görülen öğelerin kombinasyonlarını belirlemek amacıyla kullanılan bir yöntemdir. Bu yöntem, genellikle büyük veri kümelerinde yer alan öğeler arasındaki ilişkileri keşfetmek için kullanılır (Zhang ve Zhang, 2002).

Market sepeti analizi, birliktelik kuralı çıkarımının en bilinen uygulamalarından biridir. Bu analiz, müşterilerin bir alışveriş sırasında hangi ürünleri birlikte satın aldıklarını belirlemek için kullanılır. Örneğin, bir müşterinin ekmek aldığında sıklıkla süt de aldığı tespit edilebilir. Bu tür bilgiler, ürün yerleşimi, promosyonlar ve indirimler gibi

pazarlama stratejilerinin geliştirilmesinde kullanılabilir. Ayrıca, müşterilere özelleştirilmiş öneriler sunmak için de kullanılabilir. Bu sayede, satışların ve müşteri memnuniyetinin artırılması hedeflenir.

Apriori Algoritması

Apriori algoritması, birliktelik kuralı çıkarımı için geliştirilen ve büyük veri kümelerinde sıkça bir arada görülen öğe kümesini belirlemeyi amaçlayan popüler algoritmalardan biridir. Algoritmanın temel çalışma prensibi, sıkça bir arada görülen öğe kümesini belirlerken, daha az sık görülen öğe kombinasyonlarını erken aşamada elemektir. Özellikle perakende sektöründe, müşterilerin birlikte satın aldığı ürünleri belirlemek için kullanılır. Bu bilgi, ürün yerleşimi, promosyonlar, indirimler ve öneri sistemleri gibi pazarlama stratejilerinin geliştirilmesinde kullanılabilir (Han ve diğerleri, 2011).

Alışveriş Numarası	Alınan Ürünler
1	Ekmek, Süt, Su
2	Ekmek, Tereyağı, Yumurta
3	Ekmek, Süt, Yumurta, Su
4	Süt, Su, Kola
5	Ekmek, Süt, Kola

Tablo 3: Ürün Veri Seti

Yukarıdaki tabloda müşterilerin alışverişlerinde satın aldığı ürünlerin bir kısmı örnek olarak verilmiştir. Bu veriler baz alınarak Apriori algoritması kullanıldığında, en sık bir arada satın alınan ürünlerin "Ekmek, Süt" ve "Süt, Su" olduğu ortaya çıkmaktadır.

Takviyeli (Pekiştirmeli) Öğrenme

Takviyeli öğrenme, belirli bir ortam içerisinde etkileşimde bulunarak otonom kararlar alabilen bir sistemin, belirlenen bir hedefe erişmek için ardışık eylemler gerçekleştirmesini kapsar. Bu öğrenme yaklaşımı, sistemin çoklu karar seçenekleri arasından hedefe ulaşmasını maksimize edecek en uygun eylem dizisini belirlemeyi ve öğrenmeyi hedeflemektedir (Sutton ve Barto, 2011).

Takviyeli öğrenme yaklaşımında sistem, doğru eylemler için pozitif geri bildirimlerle ödüllendirilirken, yanlış eylemler için negatif geri bildirimlerle cezalandırılır veya daha düşük ödüller alır. Esas hedef, sistemin maksimum toplam ödülü elde edebilmek için hangi eylem dizisinin en optimal olduğunu öğrenmesidir. Bu öğrenme süreci, sistemin çok sayıda eylem gerçekleştirmesi ve ardışık deneme-yanılma döngülerinden geçmesini gerektirir (Kılınç ve Başeğmez, 2018).

Takviyeli öğrenme yaklaşımı, özellikle oyun tasarımı ve robotik uygulamalarında yaygın olarak kullanılmaktadır. Örneğin, labirent içerisinde hareket eden ve çıkışa ulaşmayı hedefleyen bir robot, takviyeli öğrenme yaklaşımıyla bu amacına ulaşabilir. Robot, dört temel hareket yönüne sahip olup, çıkışa ulaşana kadar hareketlerini sürdürür. Ardışık denemeler sonucunda robot, labirentin başlangıcından çıkışına en kısa sürede ve herhangi bir engelle karşılaşmadan nasıl ulaşacağını öğrenir. Bu süreçte, robotun doğru hareketleri için ödüllendirilmesi ve yanlış hareketler için cezalandırılması, onun en etkili hareket dizisini öğrenmesine yardımcı olur.

Q-Learning Algoritması

Q-Learning algoritması, takviyeli öğrenme yaklaşımları arasında öne çıkan ve yaygın olarak kullanılan bir yöntemdir. Bu algoritmanın merkezinde, bir sistemin (örneğin bir robot) belirli bir durumda alabileceği eylemlerin beklenen ödüllerini önceden tahmin ederek en yüksek ödülü elde edecek eylemi seçme fikri bulunmaktadır. Özellikle belirli bir durumda gerçekleştirilebilecek hareketlerin değerlerini öğrenmek ve bu değerlendirmelere dayanarak en uygun hareketi seçmek algoritmanın temelini oluşturur. Bu yaklaşım, sistemin karşılaştığı durumlar ve alabileceği eylemler arasındaki ödül beklentilerini optimize ederek, uzun vadede en yüksek toplam ödülü elde etmesini amaçlar.

Q-Learning algoritmasında, R (Reward = Ödül) ve Q (State = Durum-Değer) olmak üzere iki temel matris bulunmaktadır. Ödül matrisi, sistemin hareket edebileceği durumlar arasındaki geçişler için tanımlanan ödülleri içerir. Bu matriste, ulaşılamayan durumlar için genellikle negatif bir değer veya sıfır, ulaşılabilen durumlar için ise pozitif bir değer atanır. Öte yandan durum matrisi, sistemin belirli bir durumda alabileceği eylemlerin beklenen uzun vadeli ödüllerini temsil eder. Algoritmanın başlangıcında durum matrisinin tüm değerleri genellikle sıfır olarak başlatılır. Algoritma ilerledikçe durum matrisi güncellenir ve optimal sonuca yaklaşır. Bu güncellemeler sistemin deneyimlerine ve aldığı ödüllere dayanarak gerçekleştirilir. Bu iki matris, sistemin öğrenme sürecinde optimal eylemleri seçerken kullandığı temel yapıları oluşturmaktadır.

Algoritma, başlangıçta rastgele bir konumdan harekete geçerek hedefe ulaşabileceği potansiyel yolları keşfetmeye çalışır. Bu keşif süreci sırasında, algoritma ödül matrisine dayanarak tecrübe kazanır. Her adımda, algoritma maksimum ödülü elde etmeyi hedefler ve bu süreçte edindiği deneyimleri kullanarak hareket eder. Edindiği bu

deneyimler durum matrisine kaydedilir ve algoritma her iterasyonda bu matristeki değerleri güncelleyerek optimal hareket adımlarına yaklaşmaya çalışır. Bu sürekli güncelleme mekanizması algoritmanın zamanla daha bilinçli ve etkili kararlar almasını sağlar (Russell ve Norvig, 2009).

2.2.3. Makine Öğrenmesi Süreci

Makine Öğrenmesi (MÖ) süreci, MÖ modellerinin geliştirilmesi ve dağıtılmasındaki belli başlı aşamaları gösteren döngüsel bir süreçtir. MÖ süreci, projelerle ilgili karmaşıklıkların nasıl ele alınması ve yönetilmesi gerektiğini gösteren yapısal bir yaklaşım sunduğundan, MÖ projelerinin başarılı bir şekilde uygulanmasını sağlaması bakımından oldukça büyük bir önem taşımaktadır. Bu süreç genel olarak altı aşamada ele alınabilir; problem tanımlama, veri toplama ve hazırlama, model seçimi ve eğitimi, model değerlendirme ve ayarlama, model dağıtımı ve model izleme-bakımı.

2.2.3.1. Problem Tanımı

MÖ sürecindeki ilk aşama olan "problem tanımı", MÖ modelinin çözmeyi amaçladığı sorunun tespitinin gerçekleştirilmesidir. Bu aşamanın iş hedeflerini anlama, proje paydaşlarını belirleme ve başarı ölçütlerini tanımlama bölümlerinden oluştuğu dile getirilebilir. Örnek vermek gerekirse, daha önce kullanım alanlarında söz ettiğimiz uygulama sahalarından biri olan perakende şirketin müşteri kayıplarını tahmin etmek için MÖ'ne müracaat ettiği durumda paydaşların pazarlama ile satış ekipleri ve başarı ölçütünün de tahminlerin doğruluk oranı olacağı söylenebilir (Provost ve Fawcett, 2013).

2.2.3.2. Veri Toplama ve Hazırlama

Problem tanımı aşaması geçildikten sonraki adım verilerin toplanması ve hazırlanması aşamasıdır. Burada çeşitli kaynaklardan veri toplamak, veri hatalarını veya tutarsızlıkları gidermek için veriyi temizlemek ve bu verileri MÖ algoritmaları tarafından verimli ve doğru bir şekilde kullanılabilecek bir formata dönüştürmek söz konusudur. Bir kez daha müşteri kaybı tahmini örneğine başvuracak olursak veriler şirketin müşteri veri tabanından ve/veya şirketin müşteriler hakkında veri barındıran/sağlayan diğer veri kaynaklarından toplanacaktır. Ardından yinelenen kayıtların elenmesi, verilerdeki eksik değerlerin doldurulması ve kategorik değişkenlerin kodlanması aracılığıyla veri hazırlama aşaması gerçekleştirilecektir (Hastie, Tibshirani ve Friedman, 2009).

2.2.3.3. Model Seçimi ve Eğitimi

MÖ sürecindeki üçüncü aşama, ilk adımda tespit edilen probleme uygun bir MÖ algoritmasının seçilmesi ve bu doğrultuda modelin eğitilmesini içerir. Algoritma seçimi, problemin doğasına ve veri tipine bağlı olarak değişiklik gösterebilir. Yine müşteri kaybı tahmin problemini ele alacak olursak lojistik regresyon veya karar ağacı gibi bir ikili sınıflandırma algoritması buradaki sorunun çözümünde uygun olacaktır. Model, eldeki verilerden oluşturulacak bir alt küme olan eğitim veri seti kullanılarak eğitilir (Géron, 2019). Farklı bir problem hakkında başka bir örnek olarak da emlakçılık sektöründe yer alan ev/konut fiyatlarının tahmin edilmesinden söz edebiliriz. Burada, konutun yatak odası ve banyo-tuvalet sayısı, büyüklüğü, konumu gibi kategorik ve sayısal nitelikler bir arada ele alınarak oluşturulan konut verileriyle modelin eğitilmesi gerekir ve bu örnek bağlamında nitelikler ve hedef değişken arasında daha karmaşık ve doğrusal olmayan ilişkiler bulunduğu için buna uygun yeteneklere sahip algoritmalardan olan rastgele orman seçmek daha yerinde bir tercih olacaktır (Breiman, 2001).

2.2.3.4. Model Değerlendirmesi ve Ayarlama (Model Evaluation and Tuning)

Üçüncü aşamada sözünü ettiğimiz prosedürler ile model eğitildikten sonra, performans tespiti için bir değerlendirmeye ihtiyaç duyulmaktadır. Bu adımda model, eldeki tüm verilerden oluşturulacak bir başka alt küme olan test verisi setinde sınanmakta ve tahminlerin gerçek değerlerle karşılaştırılması gerçekleştirilmektedir. Modelin performansı, problem tanımlama aşamasında tanımlanan başarı metriklerine başvurularak ölçülüp ortaya konacaktır. Modelin performansı istenen düzeyde değilse, kullanılan algoritmanın parametreleri ayarlanabilir veya farklı bir algoritma kullanılma çözümüne yönelinebilir (Brownlee, 2019).

2.2.3.5. Model Dağıtımı (Model Deployment)

Model ilgili ölçütler kullanılarak değerlendirildikten ve gerekirse parametre ayarlamaları gerçekleştirildikten sonra bir ürün/uygulama haline getirilir. Bu adımda, eğitildiği veri setleri dışındaki yeni veriler üzerinde tahminlerde bulunmak amacıyla kullanılabilmesi için eğitilmiş modelin, mevcut sistemin altyapısına entegrasyonunun sağlanması, girdi-çıktı için gerekli veri kanallarının kurulması ve modelin beklenen iş yükünün üstesinden gelebilmesini temin etmek gerekmektedir. Örneğimize dönecek olursak müşteri kaybı tahmini için modelin şirketin müşteri ilişkileri yönetimi (CRM) sistemine entegre edilmesi ve modele müşterilerle ilgili en güncel verilere ulaşmasını

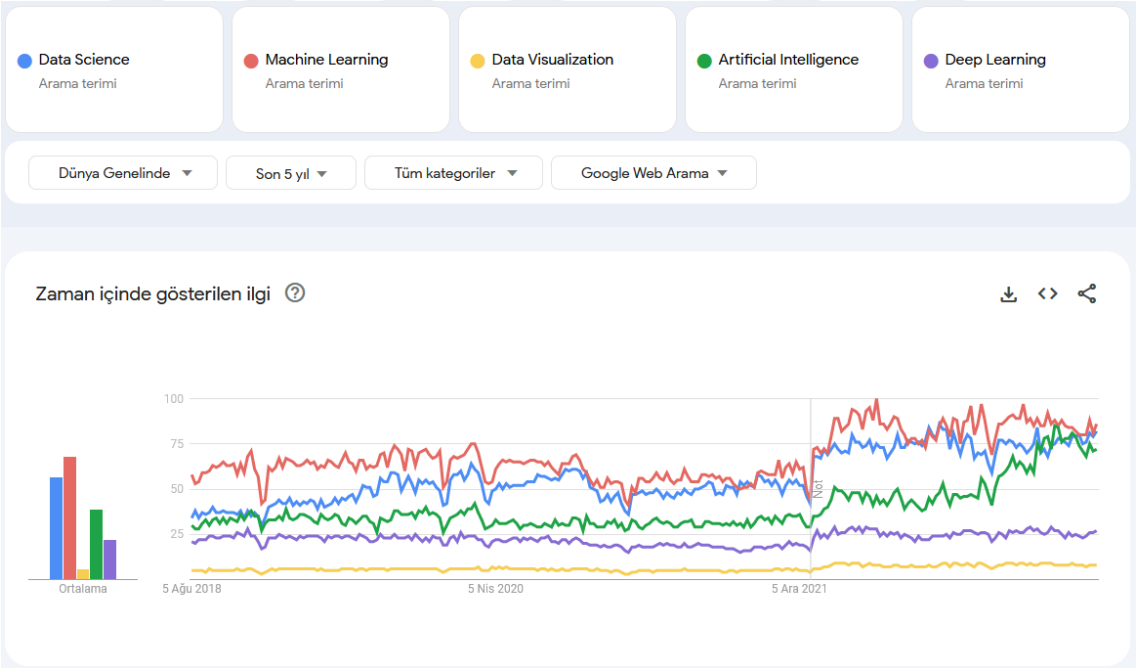
sağlayacak kanalların oluşturulması sıradaki adımı oluşturacaktır. (Bostrom ve Bostrom, 2017).

2.2.3.6. Model İzleme ve Bakım (Model Monitoring and Maintenance)

MÖ sürecinin son aşaması, modelin performansının zaman içindeki hareketlerinin izlenmesi ve gerektiğinde modelin güncellenmesi işlemlerinden oluşmaktadır. Bunun nedeni, modelin eğitim ve test aşamasında kullanılan verilerde zamanla ortaya çıkabilecek farklar ve bununla ilişkili olarak MÖ modellerinin performansında yaşanan düşmelerdir. Dolayısıyla, modelin düzenli olarak yeni veri setleriyle tekrar eğitilmesi ve gerekli durumlarda güncellenmesi büyük önem arz etmektedir. Müşteri kaybı tahmini örneğinde, müşterilerden elde edilecek en güncel verilerle modeli yeniden eğitme ve performansını değerlendirmek bu aşamayı teşkil eder (Provost ve Fawcett, 2013).

Sonuç olarak, MÖ süreci, MÖ modellerini geliştirmek ve dağıtmak, yani sistemlere dahil edilecek bir ürün haline getirmek için yapılandırılmış bir yaklaşım sağlamaktadır. Sözü edilen bu süreci takip ederek MÖ projelerinin başarılı olması ve istenen sonuçları vermesi sağlanabilir.

2.2.4. Kullanım Alanları



Şekil 5: Yapay Zekâ ve Yenilikçi Teknolojiler Hakkında Arama Trendleri

2018-2023 yılları arasında Google'da gerçekleştirilen aramalara dair elde edilen verilere göre veri bilimi, yapay zekâ ve makine öğrenmesi gibi yenilikçi teknolojiler

hakkında ciddi bir ilgi artışı gözlemlenmektedir. Bu ilgi artışı ulaşımdan insan kaynakları yönetimine, finans sektöründen bilgi ve iletişim teknolojilerine kadar birçok farklı sektörde ve iş modelinde dönüştürücü etkilere yol açmıştır. Bu trend, teknolojik gelişmelerin ve inovasyonun modern dünyada nasıl bir etki yaptığının somut bir göstergesi olarak değerlendirilebilir.

Çalışmamızın merkezinde yer alan belge yönetimi sistemleri de teknolojik ilerlemelerle yapay zekâ ve makine öğrenmesi algoritmalarının etkisi altında önemli bir dönüşüm geçirmeye başlamıştır. Bu alandaki yenilikler sayesinde belgelerin standart dosya planı kodlarının tespit edilmesi, saklama planlarının belirlenmesi, doğru biçimde sınıflandırılması ve yaşam döngülerinin etkin bir şekilde yönetilmesi gibi hedefler (McDonald ve Léveillé, 2014) mümkün hale gelmeye başlamıştır. Yapay zekâ ve makine öğrenmesi teknolojilerinin entegrasyonu belge yönetimi süreçlerini daha verimli, hızlı ve hatasız hale getirerek belge yönetimi uygulamalarının modern iş dünyasındaki gereksinimlere daha uygun hale gelmesini sağlamaya başlamıştır.

Arşiv sistemlerinde makine öğrenmesi algoritmalarının kullanımı, belge yönetimi süreçlerini otomatikleştirme ve verimliliği artırma noktasında büyük bir potansiyele sahiptir. Özellikle büyük veri kümelerinin sınıflandırılması ve yönetilmesi konusunda bu algoritmaların katkısı yadsınamaz derecede önemlidir. New South Wales Devlet Arşivlerinde gerçekleştirilen çalışma, sınıflandırma algoritmalarının, yapılandırılmamış ve sınıflandırılmamış verilerin sınıflandırılmasında ve imhasında ne kadar etkili olduğunu göstermektedir (Rolan ve diğerleri, 2018). Aynı şekilde, Avustralya Ulusal Arşivlerinde yapılan çalışma ise makine öğrenmesi algoritmalarının belge yönetim süreçlerini nasıl dönüştürebileceğini ortaya koymaktadır. Bu çalışmada üst veri çıkarımı, semantik analiz, taksonomi ve ontoloji oluşturma gibi konular üzerinde durulmuş ve bu süreçlerin otomatikleştirilmesi için algoritmaların nasıl kullanılabileceği tespit edilmeye çalışılmıştır (Rolan ve diğerleri, 2018).

Geniş bir yelpazede kullanım alanı olan yapay zekâ ve makine öğrenme algoritmalarının çeşitli sektörlerdeki kullanım örnekleri aşağıda verilmiştir (Kızrak, 2019).

Ulaşım ve Lojistik • Trafik tahmini • Araç bakımı • Otonom araçlar	Gayrimenkul ve İnşaat • Fiyat tahmini • Yatırım getirisi analizi
Eğitim	Finans ve Bankacılık

• Öğrenci performansı tahmini • Özelleştirilmiş öğretim uygulamaları • Otomatik değerlendirme	• Dolandırıcılık tespiti • Kredi verme ve puanlama
Sağlık • Hastalık teşhisi • Kişiselleştirilmiş tedavi • İlaç keşfi • Gen analizi	Perakende ve E-Ticaret • Tavsiye sistemleri • Talep tahmini • Müşteri segmentasyonu
İnsan Kaynakları • İşe alım • Performans yönetimi • İK analizi • Dijital asistan	Enerji ve Çevre • Enerji tüketimi tahmini • Sürdürülebilir kaynak optimizasyonu • Çevresel değişiklik izleme

Tablo 4: Makine Öğrenmesi Kullanım Alanları

2.3. CRISP-DM Modeli

Veri bilimi alanında proje yönetimi, kurum ve kuruluşları başarılı sonuçlara ulaştırmada önemli bir rol oynamaktadır. Veri bilimi projeleri, çok sayıda değişkeni, veri kaynağını ve analitik teknikleri içermesinden dolayı karmaşık bir yapıya sahiptir. Aşamaları iyi tanımlanmış bir proje yönetimi metodolojisi olmadan proje hedeflerine ulaşmak oldukça zordur. Etkili proje yönetimi, projelerin yapılandırılmış ve sistematik bir şekilde yürütülmesini, verimliliğin üst düzeye çıkarılmasını, risklerin en aza indirilmesini ve proje paydaşlarına somut katkılar sunulmasını sağlamaktadır.

Veri madenciliği aşamalarını sistematik bir hale getirecek (Olson ve Delen, 2008, s.9) ve daha genel olarak veri bilimi projelerinin etkili bir şekilde yönetilebilmesini sağlayacak bir metodolojiye ihtiyaç duyulmuştur. Böyle bir metodolojiyi oluşturmak amacıyla aralarında otomobil sektöründen Daimler Chrysler (daha sonra Daimler-Benz), veri madenciliği çözümleri sunan, istatistik yazılımı üreticisi SPSS (daha sonra ISL) ve veri ambarı, donanım ve yazılım üreticisi NCR şirketi 1996 yılı sonunda bir araya gelerek çalışmaya başlamışlardır (Chapman ve diğerleri, 2000, s.3). Birkaç yıl içerisinde sigorta sektöründen OHRA, önde gelen veri bilimi şirketleri, son kullanıcılar, danışmanlık şirketleri ve araştırmacılar da bu çalışma grubuna dahil olarak bir konsorsiyum oluşturmuştur (Kelleher ve Tierney, 2019, s.57). Bu çalışma sonucunda "Veri Madenciliğinde Sektörler Arası Standart Süreç (**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)" isminde bir proje yönetimi metodolojisi oluşturulmuş ve **CRISP-DM** kısaltmasıyla anılmaya başlanmıştır. CRISP-DM projesine Avrupa Komisyonu, ESPRIT

programı kapsamında sponsor olmuş ve CRISP-DM metodolojisi 1999'da bir çalıştayda ilan edilmiştir.

CRISP-DM, veri bilimcilere veriye dayalı projelerin karmaşıklıklarının üstesinden gelmelerini sağlayacak yapılandırılmış bir yol haritası sunmakta ve tüm kritik yönleriyle sürecin ele alınmasını sağlamaktadır. Veri bilimcileri, CRISP-DM ilkelerine bağlı kalarak proje karmaşıklıklarını etkili bir şekilde yönetebilmekte, riskleri azaltabilmekte ve anlamlı sonuçlar elde edebilmektedir. CRISP-DM, veri bilimcileri ve proje paydaşları arasındaki iş birliği ve iletişimi teşvik etmektedir. Paydaşları erken aşamalardan itibaren projeye dahil ederek, proje hedeflerinin iyi tanımlanmış olmasını ve kurumun hedefleriyle uyumlu olmasını sağlamaktadır. Buna ilave olarak, CRISP-DM'in yinelemeli ve esnek bir metodolojiye sahip olduğu dile getirilmelidir. Yeni içgörüler elde edildikçe veya proje gereksinimleri gelişip değiştikçe, veri bilimi projelerinin genellikle birden çok yinelemeye ihtiyacı olduğu ve bazı ayarlamalar yapılması gerektiği kabul edilmektedir. Bu yinelemeli yapısı, veri bilimcilerin yaklaşımlarını ve modellerini iyileştirmelerine olanak tanıyarak proje sonuçlarının genel kalitesi ve başarısını arttırmaktadır.

Bir veri bilimi projesinin yaşam döngüsünü tanımlayan CRISP-DM, *işin anlaşılması (problemin tanımlanması), verinin anlaşılması, verinin hazırlanması, modelleme, değerlendirme* ve *dağıtım (modelin kullanılması, kullanıcılarla paylaşım)* olmak üzere altı aşamadan oluşmaktadır. Veri, tüm aşamaların ana konusu olduğundan diyagramın merkezinde yer almaktadır. Yukarıda söz ettiğimiz esneklik doğrultusunda bu aşamaların sırası da katı sınırlarla belirlenmiş değildir. Gerektiğinde, farklı aşamalar arasında ileri-geri hareket etmek ve geçişler yapmak mümkündür (Chapman ve diğerleri, 2000, s.10). Yani, aşamalar arasında daima doğrusal bir şekilde hareket etme zorunluluğu yoktur. Belirli bir aşamanın neticesine bağlı olarak, önceki aşamalardan birine dönülebilir, mevcut aşama yenilenebilir veya sıradaki aşamaya geçilebilir (Kelleher ve Tierney, 2019, s.58). Her aşamanın sonucu, bir sonraki aşamanın girdisidir. Oklar, aşamalar arasındaki en önemli ve sık bağımlılıkları göstermektedir (Chapman ve diğerleri, 2000, s.10).

CRISP-DM sürecindeki dış daire, veri bilimi projesinin yinelemeli yapısını simgelemektedir. Veri bilimi projesi, bir çözüm üretildikten sonra sona ermez. Süreç sırasında öğrenilen bilgiler ve oluşturulan model genellikle yeni problemleri tetikleyebilmektedir. Sonraki veri bilimi süreçleri, daha önceki süreçlerin sonuçlarından

ve deneyimlerinden faydalanır (Chapman ve diğerleri, 2000, s.10). Veri bilimi projesi geliştirilip model oluşturulduktan sonra, düzenli aralıklarla, modelin ihtiyaçları karşılayıp karşılamadığı ve eskiyip eskimediği bakımından kontrol edilmesi gerekmektedir. Bu kontrolün sıklığı, modelin kullandığı verinin ne kadar hızlı değiştiğine bağlıdır (Kelleher ve Tierney, 2019, s.62).

2.3.1. İşin Anlaşılması (Business Understanding)

CRISP-DM'in ilk aşaması, veri bilimcilerin projenin hedefleri, gereksinimleri ve istenen sonuçları hakkında kapsamlı bilgi edinmek için paydaşlarla yakın bir şekilde çalıştığı "İşin Anlaşılması" aşamasıdır. Bu aşama, cevaplanması gereken temel iş sorularının belirlenmesini ve bunların veri analizi yoluyla ulaşılabilecek ölçülebilir hedeflere dönüştürülmesini içermektedir.

İşin anlaşılması aşaması, dört adımdan oluşmaktadır. Bunlar;

- İş hedeflerinin belirlenmesi,
- Mevcut durumun değerlendirilmesi,
- Veri madenciliği hedeflerinin belirlenmesi ve
- Proje planının geliştirilmesidir (Chapman ve diğerleri, 2000, s.12).

Sürecin ilk aşamasında, iş perspektifinden hareketle projenin amaçları ve ihtiyaçları detaylı bir şekilde incelenir ve analiz edilir. Bu analiz sonucunda elde edilen bilgiler ışığında, spesifik bir veri madenciliği problemi tanımı yapılır. Bu tanımlama sayesinde projenin hedeflerine erişebilmesi için bir ön plan oluşturulmaya çalışılır (Chapman ve diğerleri, 2000, s.10).

CRISP-DM'in İşin Anlaşılması aşamasında, veri bilimcileri, projenin hedeflerini ve gereksinimlerini tanımlamak için iş paydaşlarıyla yakın bir şekilde çalışmaktadır. Kurumun veri bilimi projesi aracılığıyla neyi/neleri başarmayı amaçladığına dair net bir anlayışa sahip olması önemlidir.

Veri bilimcileri, proje hedeflerini açıkça tanımlayarak, projenin kurumun stratejik öncelikleriyle uyumlu olmasını sağlamaktadır. Ayrıca gerçekçi beklentilerin belirlenmesine yardımcı olmakta ve projenin sonraki aşamaları için bir yol haritası oluşturmaktadır. Veri bilimcilerin, proje hedeflerini tanımlamasına ek olarak, projenin değerlendirmesini ve başarısını ölçmek için başarı kriterlerini belirlemesi gerekmektedir. Net iş hedefleri ve başarı kriterleri tanımlamak, kurum üzerinde somut etkisi olan

sonuçlar sağlamaya odaklanılmasını sağlar. Ayrıca, görevlerin önceliklendirilmesine ve proje boyunca kaynakların etkili bir şekilde tahsis edilmesine yardımcı olur.

Veri bilimi projelerinde başarılı sonuçlar elde etmek, projenin ne için yapılacağının net bir şekilde tanımlanmasına bağlıdır (Olson ve Delen, 2008, s.11). Projede ele alınacak sorunların yanlış belirlenmesi, yanlış sonuçlara ulaşılmasına neden olur. Bu sebepten sorunların doğru bir şekilde belirlenmesi oldukça önemlidir.

2.3.2. Verinin Anlaşılması (Data Understanding)

Proje hedefleri belirlendikten sonra, bir sonraki aşama, veri bilimcilerin mevcut veri kaynaklarını araştırdığı ve değerlendirdiği "Verinin Anlaşılması" aşamasıdır. Bu, gerekli veri kümelerinin elde edilmesini, yapılarının ve kalitelerinin anlaşılmasını ve verilerdeki potansiyel sınırlamaların veya yanlılıkların belirlenmesini içermektedir.

Verinin anlaşılması aşaması, dört adımdan oluşmaktadır. Bunlar;

- Verilerin toplanması / bir araya getirilmesi,
- Verinin tanımlaması,
- Verilere ilişkin ilk öngörülerin keşfedilmesi ve
- Veri kalitesinin doğrulanmasıdır (Chapman ve diğerleri, 2000, s.12).

Sürecin ikinci aşaması, verilerin toplanması ile başlar ve verilere aşina olunmasını, veri kalitesi sorunlarının belirlenmesini, verilere ilişkin ilk öngörülerin gerçekleştirilmesini ve/veya saklı bilgilerle ilgili hipotezlerin oluşturulması için ilginç alt veri kümelerinin tespit edilmesini sağlayan bir dizi işlemle devam eder (Chapman ve diğerleri, 2000, s.10).

CRISP-DM'in Verinin Anlaşılması aşamasında, veri bilimcileri, yapısı, kalitesi ve projeyle alaka düzeyi hakkında fikir edinmek için mevcut verileri toplamaya ve keşfetmeye odaklanırlar. Keşif aşamasında, özelliklerini anlamak için veriler incelenmekte ve bu verilerin değişkenleri, dağılımları ve potansiyel korelasyonları analiz edilmektedir. Özet istatistikler, görselleştirme ve veri profili çıkarma gibi keşfedici veri analizi teknikleri verilerdeki kalıpları, aykırı değerleri ve herhangi bir veri kalitesi sorununu ortaya çıkarmak için bu aşamada kullanılmaktadır.

Veri bilimciler, sonraki aşamalara geçmeden önce toplanan verilerin kalitesini ve alaka düzeyini değerlendirmektedir. Bu değerlendirme, proje için verilerin eksiksizliği, doğruluğu, tutarlılığı ve uygunluğu gibi faktörlerin değerlendirilmesini içermektedir.

Veri bilimciler, mevcut verileri kapsamlı bir şekilde anlayarak ve bunların kalitesini ve uygunluğunu değerlendirerek, CRISP-DM'in sonraki aşamaları için verilerin nasıl ele alınacağı ve hazırlanacağı konusunda bilinçli kararlar vermektedir. Bu, modelleme için kullanılan verilerin güvenilir, temsilî ve projenin hedeflerine ulaşmak için uygun olmasını sağlamaktadır.

Verilerin dikkatli bir şekilde seçilmesi, veri madenciliği algoritmalarının yararlı bilgi örüntülerini hızlı bir şekilde keşfetmesini kolaylaştırabilmektedir (Olson ve Delen, 2008, s.12). Dolayısıyla, proje hedeflerine ulaştıracak verilerin seçilmesi oldukça önemlidir.

2.3.3. Verinin Hazırlanması (Data Preparation)

"Verinin Hazırlaması" aşaması, verileri analiz ve modellemeye uygun hale getirmek için temizleme, bütünleştirme ve dönüştürmeye odaklanılan aşamadır. Bunların yanı sıra verilerdeki eksik ve aykırı değerlerle ilgilenmek, özellik mühendisliğini gerçekleştirmek ve farklı veri kümelerini birleştirmek gibi görevleri de içermektedir.

Verinin hazırlanması aşaması, beş adımdan oluşmaktadır. Bunlar;

- Verilerin seçilmesi,
- Verilerin temizlenmesi,
- Yeni verilerin ve/veya niteliklerin oluşturulması,
- Verinin birleştirilmesi ve/veya bütünleştirilmesi,
- Uygun formata dönüştürülmesidir (Chapman ve diğerleri, 2000, s.12).

Sürecin üçüncü aşaması olan verinin hazırlanması, ilk ham verilerden modelleme aşamasına gönderilecek nihai veri setini oluşturmak için gereken tüm işlemleri kapsar.

Verilerin temizlenmesi adımı, Verinin Anlaşılması aşamasında tanımlanan eksik, aykırı değerlerin ve tutarsızlıkların ele alınmasını içermektedir. Veri bilimciler, eksik değerlere sahip satırları kaldırmayı, uygun teknikler kullanarak eksik değerleri tamamlamayı ve aykırı değerleri tespit ederek bunları düzenleyecek yöntemleri tercih etmektedir. Bu, verilerin eksiksiz olmasını ve modelleme sürecini olumsuz etkileyebilecek anormalliklerden kaçınabilmeyi sağlamaktadır.

Birden fazla veri kaynağı söz konusu olduğunda veri entegrasyonu gerçekleştirilmektedir. Veri bilimcilerin farklı veri kümelerini birleştirmeleri ve bunları ortak değişkenlere veya anahtar alanlara göre düzenlemeleri gerekmektedir. Bu adım,

ilgili tüm verilerin tek bir veri kümesinde birleştirilmesini sağlayarak kapsamlı analiz ve modellemeyi kolaylaştırır.

Veri dönüştürme, verilerin uygulanacak modelleme tekniklerine uygun formatta hazırlanmasını kapsamaktadır. Bu, mevcut değişkenleri birleştirerek veya dönüştürerek yeni özelliklerin oluşturulduğu özellik mühendisliğini içermektedir.

Eksik ve aykırı değerlerle uğraşmak, verinin hazırlanması aşamasında kritik bir adımdır. Eksik değerler, veri toplama hataları veya eksik kayıtlar gibi çeşitli nedenlerle ortaya çıkabilmektedir. Veri bilimciler, verilerin özelliklerine ve projenin amaçlarına dayalı olarak eksik değerleri ele almak için en uygun yaklaşıma karar vermelidir.

Verinin hazırlanmasıyla ilgili görevler, birden çok kez gerçekleştirilebilir ve bahsedilen görevlerin önceden belirlenmiş bir sırası yoktur. Bu görevler; tablo, kayıt ve öznitelik seçimi, veri dönüşümü ve temizleme işlemlerini kapsamaktadır (Chapman ve diğerleri, 2000, s.11). Bu aşamada, daha derin bir veri keşfi araştırması uygulanabilir ve saklı örüntüleri görme fırsatı sağlayan ek modeller kullanılabilir (Olson ve Delen, 2008, s.9). Verilerin kalitesi ve bütünlüğü sonraki modelleme aşamalarının başarısını doğrudan etkilediğinden, verinin hazırlanması aşaması oldukça önemlidir.

2.3.4. Modelleme (Modeling)

Verinin hazırlanmasından sonraki aşama, uygun modelleme tekniklerinin seçildiği ve hazırlanan veriler kullanılarak modellerin eğitildiği "Modelleme" aşamasıdır. Bu aşama, proje hedeflerine, hazırlanan verinin özelliklerine ve aranan içgörü türlerine en uygun modelleme tekniklerinin seçilmesini içermektedir.

Modelleme aşaması, dört adımdan oluşmaktadır. Bunlar;

- Modelleme tekniklerinin seçilmesi,
- Test tasarımının oluşturulması,
- Modelin inşa edilmesi ve
- Modelin değerlendirilmesidir (Chapman ve diğerleri, 2000, s.12).

Sürecin bu dördüncü aşamasında, çeşitli modelleme teknikleri seçilir ve bir önceki aşamada hazırlanan veriye uygulanır. Sonuçlar göz önünde bulundurularak modelleme tekniklerinin parametreleri optimum, yani en uygun değerlerle ayarlanmaya çalışılır. Genellikle, aynı veri madenciliği problem türü için kullanılabilecek birkaç modelleme tekniği bulunur. Bu tekniklerden bir kısmının veri formuyla ilgili özel gereksinimleri

olabilir. Bu sebeple, çoğu zaman verinin hazırlanması aşamasına geri dönmek gerekebilir (Chapman ve diğerleri, 2000, s.11).

CRISP-DM'in Modelleme aşamasında, veri bilimciler, ele alınan problemin türüne ve mevcut verilerin yapısına bağlı olarak regresyon, sınıflandırma, kümeleme veya derin öğrenme gibi bir dizi algoritma kullanabilir. Projenin hedeflerine ve eldeki verilerin yapısına uygun olan modelleme tekniğini seçmek önemlidir.

Modelleme tekniği seçildikten sonra, veri bilimciler hazırlanan verileri eğitim ve doğrulama setlerine ayırır. Eğitim seti, modelin eğitilmesi amacıyla kullanılırken, doğrulama seti modelin performansının ölçülmesi ve gerektiğinde ayarlamaların yapılması için kullanılır. Model performansını belirlemek için çapraz doğrulama gibi teknikler kullanılabilir. Model eğitimi sırasında veri bilimciler, mümkün olan en iyi performansı elde etmek için modelin parametrelerini optimize eder. En doğru ve etkili modeli bulmak için farklı algoritmalar, özellik seçimleri veya hiperparametre yapılandırmaları ile birçok deney gerçekleştirilebilir.

Modeller eğitildikten sonra, uygun değerlendirme ölçütleri ve doğrulama teknikleri kullanılarak performansları değerlendirilir. Değerlendirme ölçütlerinin seçimi, ele alınan problemin türüne bağlıdır. Örneğin, bir sınıflandırma probleminde yukarıda söz ettiğimiz doğruluk, kesinlik, hatırlama veya F1 skoru gibi ölçütler kullanılabilir. Regresyon problemlerinde ise ortalama karesel hata (MSE) veya kök ortalama karesel hata (RMSE) gibi ölçütler de yaygın olarak kullanılır.

Veri bilimciler farklı modellerin performansını karşılaştırır ve projenin hedeflerini ve İşin Anlaşılması aşamasında belirlenen başarı kriterlerini en iyi karşılayan model(ler)i seçer. Model değerlendirmesi, her bir modelin güçlü ve zayıf yönlerini belirlemeye yardımcı olur ve hangi modelle devam edileceği konusunda karar vermeye rehberlik eder.

2.3.5. Değerlendirme (Evaluation)

CRISP-DM'in Değerlendirme aşamasında, Modelleme aşamasında geliştirilen modellerin performansı ve geçerliliği değerlendirilmektedir. Veri bilimciler, uygun değerlendirme ölçütlerini kullanarak modellerin projenin hedeflerini ne kadar iyi karşıladığını ölçmektedir. Bu aşama, modellerin performanslarını iyileştirmek için gereken parametre ayarlamalarını ve bunların optimize edilmesini içermektedir.

Değerlendirme aşaması, üç adımdan oluşmaktadır. Bunlar;

- Model sonuçlarının değerlendirilmesi,
- Sürecin yeniden incelenmesi (review process),
- Bundan sonraki adımların belirlenmesidir (Chapman ve diğerleri, 2000, s.12).

Sürecin bu beşinci aşamasında, bir önceki aşamada oluşturulan model(ler)in veri analizi açısından sonuçları değerlendirilir. Model dağıtımına, yani son aşamaya geçmeden önce, modelin birinci aşamada belirlenen iş hedeflerine uygun şekilde oluşturulup oluşturulmadığından emin olmak için kapsamlı bir değerlendirme ve yapılan işlemleri gözden geçirilmesi önemlidir. Yeterince dikkate alınmamış bazı iş sorunlarının olup olmadığını belirlemek de bu aşamadaki bir diğer önemli husustur. Bu aşamanın sonunda veri madenciliği sonuçlarının kullanımına ilişkin bir karar verilmelidir (Chapman ve diğerleri, 2000, s.11). Eğer model, hedeflenen başarıyı elde edememişse sürecin önceki aşamalarına geri dönülerek daha uygun veya başarılı model(ler) oluşturmak için CRISP-DM sürecinin tekrarlanması gerekebilir.

Veri bilimciler, model performansının ötesinde modelin, kurumun iş hedefleri üzerindeki etkisini de değerlendirir. Modelin tahminlerinin veya tavsiyelerinin karar alma süreçlerine ve operasyonel iş akışlarına nasıl entegre edilebileceğini anlamak için alan uzmanları ve paydaşlarla iş birliği yaparlar. Bu değerlendirme, modelin beklenen faydaları sağlayıp sağlamadığını ölçmeye ve eyleme dönüştürülebilir içgörüler üretmeye yardımcı olur.

2.3.6. Dağıtım (Deployment)

CRISP-DM'in son aşaması, geliştirilen modellerin operasyonel sistemlere veya karar verme süreçlerine entegre edildiği "Dağıtım" aşamasıdır. Bu aşama, modellerin bir üretim ortamında uygulanmasını, erişilebilirlik için kullanıcı arayüzleri veya API'ler ve bunların performanslarının devam etmesini sağlamak için izleme mekanizmaları oluşturulmasını içermektedir.

Dağıtım aşaması, dört adımdan oluşmaktadır. Bunlar;

- Dağıtımın planlanması,
- Modelin izlenmesi ve bakımı,
- Nihai raporun oluşturulması,
- Projenin gözden geçirilmesidir (Chapman ve diğerleri, 2000, s.12).

Başarılı bir modelin oluşturulması, genellikle veri madenciliği sürecinin bittiği anlamına gelmemektedir. Modelin amacı, veri(ler)den faydalı bilgi(ler) çıkarmak olsa bile, elde edilen bilginin kullanılabilecek şekilde düzenlenmesi ve anlaşılır şekilde sunulması gerekmektedir. Bu da modelin ve elde edilen bilgilerin veri bilimi projesini kullanacak kuruluşun karar verme süreçlerine dahil edilmesi anlamına gelmektedir (Chapman ve diğerleri, 2000, s.11).

Veri bilimciler, modelleri operasyonel sistemlere entegre etmek için bilişim teknolojileri ekipleriyle birlikte çalışarak ölçeklenebilirlik, güvenilirlik ve güvenlik sağlarlar. Ayrıca, modelin zaman içinde etkili bir şekilde çalışmaya devam etmesi için model izleme, performans izleme ve sürekli bakım gibi hususları da dikkate alırlar. Düzenli izleme, performanstaki tüm sapmaları veya bozulmaları belirlemeye yardımcı olarak zamanında güncellemeler veya iyileştirmeler yapılmasına imkân verir.

Dağıtım aşamasının sonunda, veri bilimciler tüm CRISP-DM sürecini yansıtmak, öğrenilen dersleri belirlemek ve gelecekteki projeler için en iyi uygulamaları belgelemek amacıyla bir proje incelemesi yürütürler. Projenin tanımlanmış hedeflere ulaşmadaki başarısını ve CRISP-DM metodolojisinin projenin yürütülmesine rehberlik etmedeki etkinliğini değerlendirirler. Veri bilimciler, veri kalitesi sorunları, modelleme teknikleri veya iletişim stratejileri gibi iyileştirme alanlarını da göz önünde bulundurarak bu içgörülerden gelecekteki çalışmalarda faydalanırlar. Proje incelemesi ve projeden öğrenilen dersler, veri bilimi ekibi ve bir bütün olarak kurum için değerli kaynaklar olan sürekli iyileştirmeyi ve veriye dayalı bir kültürün oluşturulmasını teşvik eder.

Sonuç olarak veri bilimi alanında başarılı sonuçlara ulaşması istenen bir proje yönetimi sürecini belirli bir standarda bağlayan CRISP-DM döngüsünün önemi oldukça açıktır. Büyük veri yığınları ve karmaşık teknikler içeren bu zorlu süreci yönetebilmek için ihtiyaç duyulan metodolojiyi ortaya koyan CRISP-DM veri bilimcilerinin takip edeceği bir harita işlevi görmektedir. Söz ettiğimiz altı aşama ile birlikte andığımız kural dizilerine sadık kalınması durumunda proje yönetimi kolaylaşmakta ve istenen anlamlı sonuçlara ulaşılabilmektedir.

ÜÇÜNCÜ BÖLÜM

3. ELEKTRONİK BELGE YÖNETİM SİSTEMLERİNDE VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ KULLANIMI

Bu bölümde belge yönetimi, elektronik belge yönetimi, elektronik belge yönetim sistemleri ve EBYS'lerde üst veri kullanımı, yapay zekâ, makine öğrenmesi ve veri madenciliği hakkında bu sistemlere entegre edilmesi için neler yapılması gerektiğinden bahsedilecektir. Bu bağlamda şimdi sırasıyla belge ve doküman kavramlarına, elektronik belge yönetimi disiplinine, bu disiplin altında oluşturulan standartlar, sistemler ve bunlara ilişkin bazı hususlara değinerek ilerleyebiliriz.

Belge ve Doküman

Gündelik yaşamdaki kullanımlarında her ne kadar bu iki sözcük birbiri yerine geçebilecek şekilde kullanılsa da teknik olarak oldukça farklı anlamlara sahiptirler. Belge "herhangi bir bireysel veya kurumsal fonksiyonun yerine getirilmesi için alınmış, ya da fonksiyonun sonucunda üretilmiş, içerik, ilişki ve formatı ile ait olduğu fonksiyon için delil teşkil eden kayıtlı bilgi" anlamına gelirken doküman ise "çok daha geniş anlamlı bir kavramdır. Resmî belge niteliği taşımayan ancak kurumsal aktivitelerin gerçekleştirilmesinde kullanıcıların bilgi amaçlı olarak kullanabilecekleri kaynaklardan oluşur. Her doküman bir belge değildir ama her belge bir doküman görevi görebilir" şeklinde tanımlanabilir (Kandur, 2006; TS 13298, 2015).

Belge Yönetimi

Belge yönetimi ve arşiv sistemi; bir kurum içinde yürütülen tamamlayıcı bir süreç yönetimidir ve kurumsal hafızayı oluşturan bilgi ve belgeleri güvence altına alır. Kurumun belge ve arşiv ile ilgili işlemlerinin etkin ve verimli bir biçimde yürütülmesini sağlar (Özdemirci ve diğerleri, 2009). Temelde bir süreç yönetimi olarak görülebilecek belge yönetimi ve arşiv işlemlerinin yürütüldüğü yapıyı ortaya koymakla birlikte, belge yönetimi arşiv sisteminin de temelini oluşturmaktadır. Bu bakımdan "belge ve arşiv işleri, belgenin üretimi/oluşumu ile başlayan, kaydedilmesi, dosyalanması, düzenlenmesi, kullanımı, tasfiyesi vb. aşamalarıyla devam eden bir süreci kapsamaktadır." (Özdemirci ve diğerleri, 2009) olarak da ifade edilmektedir.

Elektronik Belge Yönetimi

2015'te yayınlanan standartta elektronik belge yönetimi şu şekilde tanımlanmıştır: "Kurumların gündelik işlerini yerine getirirken oluşturdukları her türlü dokümantasyonun içerisinden kurum faaliyetlerinin delili olabilecek belgelerin ayıklanarak bunların içerik, format ve ilişkisel özelliklerini korumak ve bu belgeleri üretimden nihai tasfiyeye kadar olan süreç içerisinde yönetmek" (TS 13298, 2015).

Elektronik Belge Yönetim Sistemi

Son derece geniş ve karmaşık bir alan olan elektronik belge yönetimi, sistematik olarak değerlendirilmeli ve bu sistemdeki öğelerin uyumlu bir şekilde bir arada çalışmasına yönelik gerekli önlemler alınmalıdır (TS 13298, 2015).

Sistemi oluşturan öğelerin başında Elektronik Belge Yönetim Sistemi (EBYS) yazılımı gelmektedir. EBYS'ye geçmek isteyen kurumların bu konuda sertifikasyona sahip bir yazılıma ihtiyaçları olacaktır. TS 13298 standardı bir EBYS yazılımında bulunması gereken asgari fonksiyonel özellikleri tanımlamaktadır (TS 13298, 2015).

3.1 Üst Veri Yönetimi

Literatürde genellikle *"veri hakkında veri"* olarak tanımlanan üst veri kavramı, birçok farklı alan ve disiplinde farklı şekillerde ele alınmaktadır. Kullanım alanları ve ihtiyaçlara sunduğu çözümler itibarıyla oldukça önem kazanmış olan üst veriyi kaynağın biçim ve içerik olarak özeti şeklinde tanımlamak da mümkündür (Yalçınkaya, 2014). Üst veriler bir belgenin içeriği hakkında yazarı, oluşturulma tarihi, biçimi ve konusuyla ilgili anahtar kelimeler gibi değerlendirilebilecek bilgiler sağlamaktadır.

Belgelerin üretimi ve ardından gelen yaşam döngüsü süresince üst verilerin oluşturulduğu ve uygulandığı birçok farklı alan bulunmaktadır. Belgelerin alınma, kaydedilme ve sınıflandırma süreçleri boyunca gerçekleştirilen üst veri tanımları belgelerin adeta kimlikleri olarak değerlendirilebilir. Bu bakımdan belgelerden anlamlı ve yararlı bilgiler elde etmeye yönelik bir süreçte başvurulacak esas nitelikler de ilk planda üst veri tanımları olmaktadır.

İyi kurgulanmış bir üst veri yönetimi sayesinde, kullanıcılar binlerce belge arasında kolaylıkla arama yapabilmekte ve ihtiyaç duyduğu belgeyi ve belge hakkındaki bilgiyi hızlı bir şekilde elde edebilmektedir. Büyük bir belge kümesini, etkili bir şekilde taranabilen organize bir yapıya dönüştüren üst veri yönetimi, elektronik belge sayılarının çok büyük boyutlara ulaştığı günümüz dünyasında kaynaklara erişim ve kullanım

imkanlarını muhafaza etmekte oldukça önemli bir rol oynamaktadır. Bunun yanı sıra, verilerin ne zaman ve kim tarafından değiştirildiğini takip edebilme imkânı sunduğu için güvenlik açısından belgelerin bütünlüğünün korunması da üst veri tanımları ile sağlanmaktadır.

EBYS'lerde kullanılan üst veriler belgenin nasıl bir hiyerarşik yapıda tutulacağı, ne kadar süre muhafaza edileceğine ve bu süre sona erdiğinde belgelere hangi işlemin uygulanacağı da üst veriden hareket edilerek belirlenmektedir.

Ülkemizde TS 13298 standardının oluşturulmasıyla EBYS'lerde kullanılan üst verilerin standart hale gelmesi sayesinde, üst veriler de düzenli bir yapı kazanmış ve farklı bilgi yönetim sistemlerinin birbiriyle etkileşebilir hale gelmiştir. Verilerin birlikte çalışabilirliğini temin eden üst veri yönetimi, bütüncül bir dijital dönüşüme hizmet etmesi yanında, belgelerin elektronik ortamda depolanmasına yönelik ihtiyaçlara da karşılık vermektedir.

Dolayısıyla, EBYS'lerde gerçekleştirilecek veri madenciliği ve makine öğrenmesi uygulamalarında oluşturulacak model içinde kullanılan üst veri alanlarında bulunan değerlerin oldukça önemli olduğu açıktır.

TS 13298 - Elektronik Belge ve Arşiv Yönetim Sistemi Üstveri Elemanları

"TS 13298 - Elektronik Belge ve Arşiv Yönetim Sistemi" standardına tabi olan EBYS yazılımlarının, sistemlerinde olması gereken üst veri elemanları standardın "Üst Veri Yönetimi" bölümünde aşağıdaki başlıklar altında listelenmiştir:

- Belge tanımları üstveri elemanları
- Dosya tasnif planı üstveri elemanları
- Saklama planı üstveri elemanları
- Birim/Alt Birim tanımları üstveri elemanları
- Seri tanımları üstveri elemanları
- Alt Seri tanımları üstveri elemanları
- Klasör/Dosya tanımları üstveri elemanları
- Kullanıcı profil tanımları üstveri elemanları
- Kullanıcı rol tanımları üstveri elemanları
- Kullanıcı grup tanımları üstveri elemanları
- Güvenlik seviye tanımları üstveri elemanları

- Tasfiye işlem tanımları üstveri elemanları
- Sistem kullanımı üstveri elemanları

Yukarıda belirtilen başlıklar altında zorunlu üst veri elemanları olduğu gibi seçmeli üstveri elemanları da bulunmaktadır. Yapılan tez çalışmasında öncelikle zorunlu üstveri elemanları dikkate alınmakta ve veri hazırlama aşamasında bu üstveriler kullanılmaktadır. Ayrıca seçmeli üstveri elemanlarının da kullanılabileceği esnek bir veri modeli tasarlanmaktadır.

TS 13298 standardına dayanan ortak bir veri modeli tasarlayarak kurum bazlı EBYS yazılımlarının değerlendirilmesi yerine, bu standardın kriterlerini sağlamış ve ürün sertifikasyonunu almış EBYS yazılımlarında kullanılabilecek bir yapının oluşturulması amaçlanmaktadır. Bu veri modeli sayesinde, EBYS yazılımları kendi sistemlerindeki belgeler ve diğer veriler üzerinde kolay bir şekilde veri madenciliği ve makine öğrenmesi algoritmalarını çalıştırabilecektir.

3.2. Veri Madenciliği ve Makine Öğrenmesi Uygulamalarının Kullanılması İçin Gerekli Yapılar

Yapay zekanın alt alan uygulamalarının elektronik belge yönetim sistemlerinde kullanılmasını sağlamak için yapılması gereken çalışmalara aşağıda kısaca değinilmiştir.

3.2.1. TS 13298'deki Üstverilerin Yapay Zekâ Çalışmalarında Kullanımı

TS 13298 ile EBYS'lerde kullanılan üstverilerin standartlaştırılması belgeler için kullanılan üstverileri düzenli bir yapıya dönüştürmüş ve bilgi yönetim sistemlerinin birbiriyle haberleşmesine olanak sağlamıştır. Bu da üstveri ve üstveri yönetiminde merkezî bir veritabanı yapısının oluşturulması için önemli adımlardan biri haline gelmiştir. Ayrıca belgeleri elektronik ortamda muhafaza etmek için gerekli olan en önemli etkenlerden biri belgenin üstverisidir.

Bilgi yönetimi, işlevsel bir öneme sahiptir. İnternet sayfaları, elektronik belgeler, dijital dokümanlar, fiziksel belgeler ve veritabanları gibi çeşitli bilgi kaynaklarının tanımlanması ve bu kaynaklara erişimde üstveri kritik bir rol oynamaktadır. Kontrollü terim listeleri, kodlama şemaları ve diğer standart tanımlayıcılar gibi araçlar, belgelerin bulunmasında ve gerekli yardımın alınmasında üstveri, en temel yardımcı unsurlar arasında yer alır.

EBYS'lerde kullanılan üstveriler belgenin yerleştirileceği klasöre, klasörün konacağı seriye, klasörün veya serinin saklama süresine, saklama süresi dolduğunda akıbetlerinin ne olacağına karar vermede belirleyici etken olarak karşımıza yine üstveri alanları çıkmaktadır.

Yukarıda kısaca belirtilen hususlar göz önüne alındığında, EBYS'lerde yapılacak yapay zekâ uygulamaları için sistemde kullanılan üst veri alanları çok önemli bir yere sahiptir.

3.2.2. EBYS'ler İçin Kontrollü Sözlüklerin Oluşturulması

Kontrollü sözlükler, benzer öğeleri gruplandırmak için oluşturulan terimlerden meydana gelmektedir. Konu şema veya başlıklarında, taksonomilerde ve diğer bilgi yönetim sistemlerinde kullanılan bu sözlükler, daha sonra erişim için bilgiyi organize etme ve yönetme yolu sunmaktadır. Bir kavram için doğru terim belirlendikten sonra, bu terimi kullansın ya da kullanmasın tüm benzer öğelerin bir araya geldiği yapılar olan kontrollü sözlükler, bu çalışma kapsamında belge oluştururken kullanıcıların belli listelerden veri girişini sağlamak ve ardından verileri, yani belgeleri belli bir düzen dahilinde ve hiyerarşik bir yapıda veri tabanında tutmaya yaramaktadır.

Kontrollü sözlüklerin amacı, bilgileri organize etmek, kataloglamak ve istenildiği şekilde erişmek için bir terminoloji oluşturmaktır. Değişken terimlerin çeşitliliğini kapsayan kontrollü sözlükler aynı zamanda tercih edilen terimlerde bir tutarlılığa ve aynı terimlerin benzer içeriğe atanmasına da imkân tanımaktadır. Kontrollü bir sözlüğün sunduğu en önemli işlevler olarak kavramların farklı terimlerini ve eşanlamlılarını bir araya getirmek ve söz konusu kavramları mantıksal bir yapı içinde birbiriyle ilişkilendirmek ya da gruplara ayırmak sayılabilir. Kontrollü bir sözlükteki bağlantılar ve ilişkiler hem belgelerin oluşturulması hem de erişim için tanımlanması ve sürdürülmesine hizmet etmektedir. Kısacası, kontrollü kelime dağarcığı, aynı kavrama farklı isimlerin atanabildiği normal konuşma dilimizden farklı olarak belirsizliği azaltıp tutarlılık sağlamaktadır (Harpring, 2010, s.12).

Kontrollü sözlükler başlığı altında birçok alt kategori bulunmakla birlikte, bu çalışma kapsamında ele alacaklarımız yalnızca kontrollü listeler, ontolojiler ve taksonomiler olacaktır. Kontrollü sözlükler birçok farklı alanı ilgilendirdiği için farklı sorunlara farklı çözümler üreten birçok alt türe sahip olsa da belge yönetimi alanında bizi esas olarak ilgilendiren bu üç sözlük tipidir.

Kontrollü Listeler

Kontrollü liste, terminolojiyi kontrol etmek için kullanılan basit bir terim listesi olarak tanımlanabilir. İyi yapılandırılmış bir kontrollü listenin özellikleri şöyle sıralanacaktır:

- Her terim benzersizdir.
- Terimler birbirleriyle aynı anlama gelmez.
- Terimlerin tümü aynı sınıfa üyedir (yani, bir sınıflandırma sisteminde aynı düzeydedirler).
- Terimler nitelik bakımından birbirine denktir.
- Terimler alfabetik olarak veya başka bir mantıksal düzen içinde sıralanır (Harpring, 2010, s.19).

Kontrollü listeler genellikle özel bir veri tabanı veya durum için tasarlanır ve özel olarak üretildiği bu bağlam dışında iş görmeyebilir. Kısa bir değer listesinin yeterli ve uygun olduğu, terimlerin eşanlamlı veya yardımcı bilgilere sahip olma ihtimalinin bulunmadığı bir veri tabanında oldukça verimlidirler. Bu tür listelerin avantajı, seçilebilecek yalnızca kısa bir terim listesine sahip olması ve böylece daha fazla tutarlılık sağlayarak hata olasılığını minimize etmesidir. Kullanıcı bakımından düşünüldüğünde, bu tür kısa listelerde gezinmek, özellikle uzman olmayanlar için karmaşık listelerden daha kolay olacaktır (Harpring, 2010, s.20).

Belge oluştururken belge ile ilgili girmemiz gereken zorunlu veya seçmeli üst verilerin tutarlılığını sağlayıp hata olasılığını en aza indirmek için belge yönetiminde bu listelerin kullanılması tercih edilmektedir. Böylelikle EBYS'den makine öğrenmesi için ihtiyaç duyulan verilerin daha önce bahsettiğimiz düzenlenme aşamasında söz konusu olan işlemlerin kolaylaşmasını sağlayacaktır.

Taksonomiler

Taksonomi, tanımlanmış bir alan için düzenli bir sınıflandırma biçimi olarak ifade edilebilir. Hiyerarşik bir yapı halinde organize edilmiş kontrollü sözlük terimlerini (genellikle yalnızca tercih edilen terimleri) içerir. Bir taksonomideki her terim, diğer terimlerle -ağaç yapısı benzeri- bir veya daha fazla "ebeveyn-çocuk" ilişkisine sahiptir. Bütün-parça, cins-tür veya örnek ilişkileri gibi farklı ana-alt ilişki türleri bulunması da mümkündür. Bununla birlikte, iyi tasarlanmış bir taksonomide belirli bir ana türün tüm alt türleri aynı ilişkiyi paylaşmaktadır. Bir taksonomi, genellikle daha yalın hiyerarşilere

ve daha az karmaşık bir yapıya sahip olduğu için bir eş anlamlılar sözlüğünden (thesauri) farklılık gösterebilir. Çünkü, taksonomilerde genellikle eş anlamlı veya değişken terimler ya da birbiriyle ilişkili terimler ilişkiler bulunmaz (Harpring, 2010, s.22).

Standart dosya planı haricinde hiyerarşik bir yapı oluşturmaya ihtiyaç duyduğumuz durumlarda taksonomilere başvurabiliriz. Yalnızca belgelerde değil, bir bütün olarak EBYS içerisinde daha anlamlı ve düzenli bir yapı sağlayabilmek için taksonomilerin kullanılması gerekmektedir. Belgelerin birbiriyle olan ilişkisi dışındaki hiyerarşiler söz konusu olduğunda, örneğin bir kurumdaki çalışanların kendi kurumsal hiyerarşisi de sisteme dahil edilmek durumunda olduğu için burada taksonomileri kullanarak hareket edilebilir.

Ontolojiler

Bilgisayar biliminde, bir ontoloji genellikle kavramlar, nitelikler, bağıntılar ve işlevlerin tümünün açıkça tanımlandığı kavramsal bir modelin büyük resmi, makine tarafından okunabilen bir özelliği olarak görülmektedir. Bu bakımdan bir ontoloji tam anlamıyla bir kontrollü sözlük değildir, fakat belirli bir model içinde bir veya daha fazla kontrollü sözlüğü (örneğin yukarıda saydığımız kontrollü listeler, taksonomiler vb.) kullanır ve bunları, anlamlı bir sonuç ifade edecekleri biçimde terimleri kullanarak temsilî bir dilde ifade etmeye yarar (Harpring, 2010, s.24).

Ontolojiler, üzerinde çalıştıkları verileri; bireyler, sınıflar, nitelikler, ilişkiler ve olaylar şeklinde çeşitli gruplara ayırmaktadır. Ontolojinin sahip olduğu dilbilgisi, bu grupları, sözlük terimlerinin bir arada kullanılma biçimleri ile ilgili kısıtlamalar oluşturarak birbiriyle ilişkilendirir. Çeşitli gramer veya dillere sahip olan ontoloji standartları, sorgulama oluşturmak için de kullanılabilmektedir (Harpring, 2010, s.25).

Çok yönlü taksonomiler ve eş anlamlılar dizileri ile bazı ortak özellikleri bulunan ontolojiler, bunlardan farklı olarak makine tarafından okunabilir biçimde (machine-readable form) bilgiyi temsil edebilecek terimler ve nitelikler arasında anlamsal ilişkiler kurmaktadır. Semantik Web'de, yapay zekada, yazılım mühendisliğinde ve bilgi mimarisinde, elektronik formda bir bilgi temsili biçimi olarak da ontolojiler kullanılmaktadır (Harpring, 2010, s.25).

Sonuç olarak veriyi düzenli ve hiyerarşik bir yapıda tutmamızı sağlayan bu sözlük türleri ile algoritmaların bu yapıyı anlayabilmesi olanaklı hale gelmektedir. EBYS'de de bu yapıların (ontoloji, taksonomi, kontrollü sözlükler, vb.) bulunması, bir modelin daha

başarılı sonuçlar elde etmesini sağlayacaktır. Çünkü söz konusu yapılar mevcut değilse, istenen başarı seviyesine ulaşmak oldukça zorlaşacak ve süreç için harcanan zaman ve efor artacaktır. Dolayısıyla daha önce belirtildiği üzere model oluşturulmadan önce bu yapıların kurulması başarı için bir önkoşul olarak değerlendirilebilir.

3.2.3. EBYS'lerde Yapay Zekâ Uygulamalarıyla Yapılacak İşlemlerin Belirlenmesi

Tezimizde ortaya koymak istediğimiz çalışma için izlenecek en uygun yöntem, kamu kurum ve kuruluşlarının EBYS'lerindeki belgeler üzerinde bir veri analizi süreci yürütmek ve bu kapsamda bir uygulama çalışması yapmaktır. Yukarıda detaylı bir şekilde ele aldığımız CRISP-DM modelinin sunduğu aşamalar üzerinden kendi proje yönetim sürecimizi aktarabiliriz. Bu bağlamda, bilindiği üzere ilk olarak işin anlaşılmasına ihtiyaç duyulmaktadır. Bu aşamanın altında yer alan adımların iş hedefleri ve mevcut durumun tespit edilmesi, veri madenciliği ve makine öğrenmesi hedeflerinin belirlenmesi ve proje planının oluşturulması olduğu hatırlanacaktır. Burada ortaya koyacağımız uygulamanın iş hedefi, belirlenen kamu kurum ve kuruluşlarda veri toplama işleminin gerçekleştirilmesi ve elde edilen verilerin değerlendirmeye tabi tutulması şeklinde özetlenebilir.

İkinci alt adım olan mevcut durumun tespitine gelince, ilk olarak belge yönetimi alanında olmasa da literatürde ve diğer alanlarda yapılmış olan veri ve metin madenciliği ile makine öğrenmesi uygulamalarını incelemek gerekmektedir. Çalışmanın birçok yerinde bu incelemelerden aktarımlarda bulunduğumuz göz önüne alındığında, burada halihazırdaki durumun tespit edilmesinin gerekli olduğu açık hale gelmektedir. Bu bağlamda elektronik belge yönetim sistemlerinin genel işleyişi incelenmiş, yapılan hatalar ana ve alt başlıklar halinde belirlenmiştir. Bu başlıklar arasında belge metniyle alakalı uygun standart dosya planı kodunun seçilmemesi, standart dosya planı kodları altında bulunan belgeler içerisinde yanlışlıkla yerleştirilmiş belgelerin olması ve dışarıdan sisteme dahil edilen fiziksel belgelerin olması sayılabilir.

Ardından üçüncü alt adım, yani veri madenciliği ve makine öğrenmesi hedeflerinin belirlenmesi doğrultusunda, mevcut durumda tespit ettiğimiz bu hataların nasıl düzeltileceği ile ilgili varsayımsal çözümler ortaya atılmıştır. Ayrıca EBYS'lerdeki ilgili hataların hangi yapay zekâ algoritmaları kullanılarak düzeltilebileceği belirlenmiştir.

Böylelikle işin anlaşılması süreci tamamlandıktan sonra verinin anlaşılması aşamasına geçilmiştir. Burada modelimiz içinde kullanılacak veriler bir araya getirilmiş, tanımlanmış ve işin anlaşılması sürecindeki varsayımların yerinde olup olmadığı tespit edilmeye çalışılmıştır. Bu bağlamda üzerinde çalışacağımız veri kümesinin özelliklerinin keşfedildiği, veri kalitesine yönelik bir fikir edinildiği ve daha sağlıklı öngörüler üretme imkanının doğduğu ifade edilebilir.

Modeli oluşturma aşamasına geçmeden önce yukarıda önemini birçok kez ifade ettiğimiz verinin hazırlanması süreci başlatılmıştır. Burada çalışma konumuzu teşkil eden veriler seçilmiş, temizlenmiş, yeni veriler ve nitelikler oluşturularak veri daha bütüncül bir hale getirilmiştir. Bu bağlamda özellikle belge metinleri üzerinde işlem yapıldığı için temel metin ön işleme teknikleri kullanılarak belge metinlerine algoritmaların anlayacağı bir düzen kazandırılmıştır. Veri temizleme başlığı altında, kullanılan veri setinde bulunan HTML diline özel etiketler (tag) başta olmak üzere metin haricindeki tüm özel karakterler kaldırılmış ve tüm kelimeler küçük harfe çevrilmiştir. Farklı tablolarda bulunan belgeyle ilgili süreçleri detaylandıracak tüm veriler yeni nitelikler oluşturularak tek bir dosya altında birleştirilmiştir.

Bütün detaylarını modeli tanıttığımız bölümde ele alacağımız bu hazırlık sürecinin ardından modelin oluşturulması aşamasına geçilmiştir. Burada kullanılacak yapı ve teknikler tespit edilerek başlanmıştır. Örneğin, çalışma kapsamında ilişkisel veritabanı yönetim sistemi olarak SQL Server, veri madenciliği ve makine öğrenmesi algoritmaları için Python programlama dili ve ilgili kod kütüphaneleri kullanılmıştır. Belgelerde bulunan belge sayısı, belge tarihi, referans numarası, başlık, konu, kategori, gizlilik derecesi, üretildiği yer, üretim tarihi, belge sahibi, belge üreticisi gibi üst veriler veri madenciliği ile makine öğrenmesi ve belgenin metin kısmı da metin madenciliği teknikleri kullanılarak incelenmiştir.

Değerlendirme aşamasında model test edilmiş ve elde edilen sonuçlarla ilgili çıkarımlarda bulunulmuştur. Bu bağlamda anlamlı sonuçlar elde edilmeye, kullanıcı kaynaklı hatalar, saklı örüntüler tespit edilmeye çalışılmış ve sistem üzerinde iyileştirmeler yapılması için önerilerde bulunulmuştur. Ayrıca, yapılan uygulama çalışması neticesinde, EBYS'lerin nasıl kullanıldığı ve kullanılması gerektiğine ilişkin tespitlerde bulunulmuştur.

Sonuçlar; grafikler, tablolar ve şekiller kullanılarak görselleştirilmiş ve bu sayede sonuçların yorumlanması daha da kolay hale getirilmiştir. Amacımıza/problemimize göre EBYS üzerindeki belgelerin belirtilen üst verileri kullanılarak makine öğrenmesi algoritmaları aracılığıyla aşağıda belirtilen başlıklar altında faydalı sonuçlar elde edilebilmektedir. Bu başlıklar;

- Belge metninden standart dosya planı kodunun tahmin edilmesi,
- Standart dosya planı kodlarına anahtar kelimelerin atanması,
- Dışarıdan sisteme dahil edilen elektronik belgenin standart dosya planı koduyla eşleştirilmesi,
- Standart dosya planı kodları altında bulunan belgeler içerisinde yanlışlıkla konmuş/yerleştirilmiş belgelerin tespit edilmesi,
- Birim bazlı en çok kullanılan standart dosya planı kodlarının belirlenmesi şeklinde sıralanabilir.

3.2.4. Elektronik Belge Yönetim Sistemi Alanında Yapılan Örnek Çalışmaların Analizi

Yapay zekâ uygulamaları elektronik belge yönetim sistemlerinde, arşivleme süreçlerinde ve uzun süreli saklama dönemlerinde kullanılmaya başlanmıştır. Bununla birlikte belge süreçlerinin otomatikleştirilmesi ve iyileştirilmesi sağlanmış ve elde edilen veriler anlamlandırılarak bilgi keşfi olanaklı hale gelmiştir. Çoğunluğu yurtdışında yapılan çalışmalar, genellikle büyük veri setlerindeki belgelerin seçimi, değerlendirilmesi, sınıflandırılması, ayıklama ve imhası gibi süreçlere odaklanırken, bazı çalışmalar tarihsel değeri olan belgelerin içerik analizini gerçekleştirip bu belgeleri dijital ortamda geniş kitlelere sunmayı hedeflemektedir.

Kısaca bu şekilde özetlenebilecek olan örnekler dışında akademik çalışmaları literatür analizi kısmında ele almıştık. Çalışmamız bağlamında özellikle önemli gördüğümüz Cumhurbaşkanlığı Bilgi ve Belge Yönetimi Toplantıları, Ulusal Veri Sözlüğü, Açık Veri ve Açık Veri Projesi kapsamında yürütülen çalışmalardan ayrıca söz etmek yerinde olacaktır.

3.2.4.1. Bilgi ve Belge Yönetimi Toplantıları

Kamu kurumlarının elektronik belgeleri yeteri kadar sağlıklı yönetemediği bilinmektedir. Bu sorunu tartışmak ve çeşitli çözüm önerileri geliştirmek amacıyla T.C. Cumhurbaşkanlığı Bilgi ve Belge Yönetimi Daire Başkanlığı'nın ev sahipliğinde bir dizi

toplantı düzenlenmiştir. Bu toplantılardan bir oturumunda büyük veri, yapay zekâ ve derin öğrenme konularının EBYS'lerde nasıl kullanılabileceği tartışılmış ve bununla birlikte karşılaşılan sorunlara yönelik tedbirler irdelenmiştir.

Büyük veri ve yapay zekâ kavramlarının EBYS'lerde kullanılmaya başlanması için yapılması gereken çalışmalar hakkında aşağıdaki öneriler dile getirilmiştir:

- "Büyük veri" kavramının tam olarak netleştirilmesi ve EBYS bünyesinde nasıl ve ne amaçla kullanılacağının belirlenmesi gerekmektedir.
- Her bir kurum için kurumsal taksonomilerin ve veri sözlüklerinin oluşturulması, yapay zekâ kavramının EBYS'lerde uygulanabilmesi için kritik öneme sahiptir.
- Yapay zekâ uygulamalarının çalışabilmesi için olmazsa olmaz unsur veri'dir. Yapay zekâ alanında çalışan akademisyenler ve araştırmacıların önündeki en büyük engel veriye erişimdir. Fakat EBYS'lerdeki verilerin kullanılması hem yasal açıdan hem de veri güvenliği açısından sorun oluşturabileceği için verilerin anonimleştirilerek kullanılması önerilmiştir.
- Büyük veri paylaşımının yasal açıdan nasıl yapılacağı netleştirilmelidir.
- EBYS'ye sadece yazılı değil sözlü metinlerin de dahil edilebilmesinin yararlı olacağı dile getirilmiştir.
- Yapay zekânın EBYS'lerde düzgün bir şekilde çalışması ve sağlıklı sonuçlar elde edilebilmesi için veri'ye ihtiyaç olduğu kadar, bu verilerin "doğru veri" olması da gerekmektedir. Bunun için de belge yöneticileri tarafından sağlıklı bir şekilde dosyalandığına ve yönetildiğine onay verilen belgelerin kullanılması son derece önemlidir.
- Yukarıda bahsedilen sorunlar çözüldükten sonra EBYS'lerde yapay zekâ uygulamalarının geliştirilmesiyle içerik analizi yapılması, belgelerin özetlenmesi ve sevk süreçlerinin otomatikleştirilmesi gibi iş süreçlerini her anlamda iyileştirecek yapılar ortaya konacaktır. Böylelikle 7/24 çalışan bir sistemin yapay zekâ tarafından yönetilmesi büyük ölçüde sağlanacaktır.

Bilgi ve Belge Yönetimi Daire Başkanlığının çeşitli kamu kurumlarıyla yaptığı önceki toplantılarda dile getirilen hususlar da değerlendirilmiş ve katılımcıların bu konularla ilgili düşünceleri alınmıştır.

- Dosyalamanın son derece önemli olduğu ve kurumların dosyalama rehberlerinin olması gerektiği belirtilmiştir.

- Elektronik belgelerin arşive aktarılmasında yaşanacak sorunlardan bahsedilmiş ve kamuda ortak veri tabanlarının oluşturulmasının önemine bir kez daha dikkat çekilmiştir.
- EYP'nin ilerleyen zamanlarda UYP'ye (Ulusal Yazışma Paketi) dönüştürülmesinin faydalı olacağı dile getirilmiştir.
- E-imza süreçlerinin geliştirilmesi, kullanımının daha kolay hale getirilmesi için yapılan çalışmalardan (mobil imza, biyometrik imza ve temassız imzalama) bahsedilmiştir.
- SDP üzerinde incelemeler yapılarak kullanım kolaylığı açısından SDP'lerin sadeleştirilmesi gerektiği söylenmiştir.
- Yapay zekâ kullanımı için EBYS'lerde öncelikli olarak belge üreten kişilere doğru dosya planını seçtirmenin önemi vurgulanmıştır.

3.2.4.2. Ulusal Veri Sözlüğü Çalışmaları

Kamu, kurum ve kuruluşlarının bilgi sistemlerindeki entegrasyon zorlukları, mükerrer ve çelişen verilerin olması, bilişim sistemlerinde dil birliğinin sağlanamaması ve veri sahipliğinin belli olmaması gibi birçok sorunu çözmek amacıyla "Ulusal Veri Sözlüğü" çalışmaları başlatılmıştır. Veri sözlüğü; yönetilen verilerin üst verilerinden oluşan bir sözlük olarak ifade edilebilir. Veri sözlüğü oluşturmadaki amaçlar;

- Tekrarlı ve çelişen verilerin önüne geçmek,
- Kişilere ya da yazılım firmalarına bağımlılığı azaltmak,
- Entegrasyon zorluklarını azaltmak,
- Bilgi çıkarımına hazır olmama durumunun önüne geçmek,
- Bilgiye erişim güçlüğünü azaltmak ve
- Veri envanterini çıkarmak olarak sayılabilir.

Ulusal Veri Sözlüğü ile verinin standartlaştırılması çalışması başlatılmış olup, Cumhurbaşkanlığı Dijital Dönüşüm Ofisi sorumluluğunda "Veri Sözlüğü Portalı" oluşturulmuş ve ilgili kullanıcılara eğitimler verilmeye başlanmıştır.

Ulusal Veri Sözlüğü kapsamında;

- Ulusal veri envanterinin çıkarılması,
- Veri sahipliğinin belirlenmesi,
- Ulusal veri entegrasyon mimarisi ile yönetim ve izleme süreçlerinin yapılandırılması,

- Terminoloji birliğine varılması ve kurumsal hafızanın oluşturulması,
- Ulusal veri modellerinin belirlenmesi,
- Yeni yazılımlar için süreç iyileştirmeleri ve
- Merkezî servis tasarım platformunun oluşturulması hedeflenmektedir.

Bu çalışmalar neticesinde, kurumlar belli kurallar dahilinde kendi veri sözlüklerini oluşturacaktır.

"Ulusal Veri Sözlüğü" çalışmaları, dijital dönüşüm içerisinde var olan bütün sistemleri ilgilendirdiği gibi EBYS'leri de ilgilendirmekte ve bu açıdan önem arz etmektedir.

3.2.4.3. Açık Veri Alanındaki Çalışmalar

Yapay zekâ ve yenilikçi teknolojilerin geliştirilebilmesi için büyük miktarda ve çeşitlilikte veriye ihtiyaç duyulmaktadır. Ancak bu verilere erişimdeki kısıtlamalar, teknolojik ilerlemenin önündeki en büyük engellerden biri haline gelmiştir. Kamu kurumları iş süreçleri ve hizmet sunumları sırasında önemli miktarda veri üretmektedir. Bu verilerin kişisel veri, ulusal güvenlik ve ticari sırların mahremiyeti gibi kritik konular gözetilerek paylaşılması hem katma değerli yeni hizmetlerin üretimini ve teknolojik gelişmeleri destekleyecek hem de kamu hizmetlerinin şeffaflığını ve hesap verilebilirliğini artıracaktır. Bu bağlamda, T.C. Cumhurbaşkanlığı Dijital Dönüşüm Ofisi'nin başlattığı "AçıkVeri Projesi", kamu kurumlarının ürettiği verilere erişimi kolaylaştırmayı ve bu verilerin yenilikçi hizmetlerin üretiminde kullanılmasını teşvik etmeyi amaçlamaktadır.

Proje Kapsamında On Birinci Kalkınma Planı[2] kapsamında öngörülen;

- "815. Kamu verisi şeffaflık, hesap verebilirlik ve katılımcılığı artırmak ve katma değerli yeni hizmetlerin üretimine imkân sağlamak üzere ve mahremiyet ilkeleri çerçevesinde açık veri olarak kullanıma sunulacaktır."
- "815.1. Kamu verisinin paylaşımına yönelik düzenlemeler yapılacaktır."
- "815.2. Kamu verisinin paylaşılacağı Ulusal Açık Veri Portalı hayata geçirilecek ve veri anonimleştirmeye ilişkin ilkeler belirlenecektir."

politika ve tedbirleri çerçevesinde çalışmalar gerçekleştirilmektedir.

[2] https://www.sbb.gov.tr/wp-content/uploads/2022/07/On_Birinci_Kalkinma_Plani-2019-2023.pdf (Erişim Tarihi: 29.07.2022)

Proje kapsamında kamu kurumlarının iş süreçlerinde ürettikleri verilerin anonimleştirilip mahremiyetleri sağlanarak paylaşılması, açık devlet ilkeleri doğrultusunda şeffaflığın ve hesap verilebilirliğin teşvik edilmesi hedeflenmektedir. Bu yaklaşım sadece kamu yönetiminde değil, aynı zamanda yapay zekâ ve yenilikçi teknolojilerin geliştirilmesi sürecinde de kritik bir rol oynamaktadır. Anonimleştirilmiş verilerin paylaşılması, özel sektör ve akademik çevreler için önemli bir kaynak oluşturarak, bu veriler üzerinde yapılan analizlerle sosyal ve ekonomik değer üretme potansiyelini artırmaktadır. Bu durum hem teknolojik yeniliklerin teşvik edilmesine hem de ülkemizin dijital dönüşüm sürecinin hızlandırılmasına katkıda bulunacaktır.

Açık devlet verisi, kamu kurumlarının iş süreçlerinde üretilen ve genel erişime sunulan verileri ifade etmektedir. Bu veriler herhangi bir maliyet olmaksızın, yeniden kullanılabilir ve paylaşılabilir bir biçimde halkın erişimine sunulmaktadır. Ancak bu verilerin paylaşımı sırasında bazı kritik sınırlamalar bulunmaktadır. Özellikle kişisel veri, ulusal güvenlikle ilgili sırlar ve ticari sırlar gibi hassas bilgilerin korunması gerekmektedir. Bu nedenle bu tür verilerin açık devlet verisi olarak tanımlanması ve paylaşılması söz konusu değildir. Ulusal Açık Veri Portalı, bu verilerin uygun bir şekilde anonimleştirilerek ve yasal sınırlamalar gözetilerek kamuyla paylaşılmasını sağlamak amacıyla oluşturulmuş bir platformdur. Bu portal, açık veri ilkelerine uygun olarak verilerin standartlaştırılması, düzenlenmesi ve erişilebilir hale getirilmesi için çalışmalar yürütmektedir.

Ulusal Açık Veri Portalı'nın uluslararası standartlara uygun şekilde yapılandırılması amacıyla yürütülen çalışmalar, bu verilerin hem teknik hem de yasal açıdan uygun bir şekilde paylaşılmasını sağlamayı hedeflemektedir. Bu bağlamda, verilerin uygun formatlarda, erişilebilir ve yeniden kullanılabilir olmasını sağlamak için teknik altyapı çalışmaları yürütülmektedir. Ayrıca verilerin paylaşılması sırasında karşılaşılabilecek yasal ve idari engellerin aşılması amacıyla yasal ve idari düzenleme altyapıları da oluşturulmaktadır. Kamu kurumlarına ve diğer ilgili taraflara rehberlik etmek amacıyla açık devlet verisi konusunda veri yönetimi prosedürleri, rehberlik dokümanları ve eğitim materyalleri de hazırlanmaktadır.

"AçıkVeri Projesi"nde yapılan çalışmalar neticesinde araştırmacılara ve bilim insanlarına sunulacak anonimleştirilmiş veri, dijital dönüşüm içerisinde var olan bütün sistemleri ilgilendirdiği gibi EBYS'leri ve bu alanda çalışan kişileri de yakından

ilgilendirmektedir. Bu açıdan "AçıkVeri Projesi" önem arz etmekte ve yakından takip edilmektedir.

AçıkVeri Projesi hakkında kısaca şunlar söylenebilir:

- AçıkVeri işlenebilir, kullanılabilir, üçüncü kişiler tarafından üzerinde değişiklikler yapılabilir veri olarak tanımlanmaktadır. AçıkVeri herhangi bir kontrol mekanizmasına ve kısıtlamaya tabi değildir.
- AçıkVeri kişilere/kurumlara birlikte çalışabilirlik imkânı sağlamaktadır. Müşterek çalışma ile bilimsel araştırmalardan daha fazla verim elde edilebilecek ve etkileşimli bir ekosistem oluşturulabilecektir.
- Veriyi açık olarak sunan yayıncılar ile kamu kurumları arasında etkileşim sağlanacak, etkin, şeffaf ve hesap verebilir kamu hizmetleri ile ekonomik büyüme odaklı bir yönetişim oluşturulabilecektir.
- AçıkVeri, üniversitelerde çeşitli yayınlar, kitaplar vb. akademik çıktılarda kullanılabileceği gibi planlama, politika geliştirme, yeni teşvik mekanizmaları oluşturulması, olası yeni iş birliklerinin kurulması ve bürokratik süreçlerin azaltılmasında da kullanılabilir.
- AçıkVeri için verinin anonimleştirilmesi elzemdir. Anonimleştirme, kimlik ve hassas bilgiler içeren verilerin ifşasının önlenmesi amacıyla mahremiyet modelleri tarafından yarı tanımlayıcı öznitelikler üzerinde yapılan dönüşüm işlemleridir.
- Anonimleştirmeyle, verinin tipi ve biçimi korunarak paylaşılmış büyük veri kümelerinde yer alan veri sahiplerinin kimlik bilgileri ve hassas verilerinin ifşa edilmesi zorlaştırılır ya da engellenir.

Ulusal veri sözlüğü ve AçıkVeri projesi çalışmaları bir arada düşünüldüğünde birçok alanda yapılacak bilimsel çalışmaya hız kazandıracağı gibi EBYS alanındaki çalışmalara da fayda sağlayacağı açık bir şekilde görülmektedir.

3.2.4.4. Açık Veri Zirvesi

10 Haziran 2021'de Türkiye'nin kamu ve özel sektör paydaşlı ilk toplumsal açık veri konferansı olan "Açık Veri Zirvesi" gerçekleştirilmiştir. Dijital dönüşümün ayrılmaz ve çok önemli bir parçası olan Açık Veri'nin ele alındığı zirvede açık devlet verilerinin; kamu yönetimlerinde koordinasyonun sağlanmasında hız kazandıracağı, yönetimsel süreçlerin optimizasyonuna katkı sağlayacağı, iş dünyası açısından bakıldığında;

yenilikçi hizmet üretiminin artacağı, inovatif iş alanlarının oluşturulması ile iş gücünün artırılmasına ve ekonomiye katkı sağlayacağı ifade edilmiştir. Ayrıca açık devlet verilerinin akademi dünyasına katkıları konusunda da araştırma kalitesini artırması ve yenilikçilik süreçleri hızlandırmasının beklendiği ve vatandaşlara katkılarının ise hizmetlerin daha kaliteli bir şekilde sağlanması şeklinde gerçekleşeceği dile getirilmiştir.

Bunlara ilave olarak, kamu kurumlarının tüm uygulama, platform ve altyapı katmanlarında kullandıkları verilere ilişkin standart ve tanımlamaların yer alması için yapılan Ulusal Veri Sözlüğü (UVS) çalışmalarının başlatıldığı dile getirilip yapılan bu çalışmaların bir sonucu olarak kamu kurumlarında ortak bir veri dili sayesinde ulusal veri envanterinin oluşturulmasının amaçlandığı vurgulanmıştır. Zirvede "Açık Veri Nedir? Ne değildir?", "Açık Veri Daha İyi Bir Gelecek İçin Hangi Sorunlarımızı Çözer? (Örnek Uygulamalar)", "Türkiye'de Açık Veri: Fırsatlar ve Sorunlar", "Dünyada Açık Veri: Fırsatlar ve Sorunlar", "Açık Veri Hukuku" ve "Açık Veri Ekonomisi" başlıklarında toplam altı oturum gerçekleştirilmiştir.

Son olarak, çok yakın zamanda Türkiye'nin açık veri portalı veri.gov.tr'nin erişime açılacağı ve ulusal güvenlik, kişisel verilerin korunması ilkesi ve ticari sırlar konusunda azami dikkat gösterileceği ifade edilmiştir.

Ulusal veri sözlüğü ve Açık Veri projesi çalışmaları bir arada düşünüldüğünde kamu, özel sektör ve akademi gibi birçok alanda yapılacak bilimsel çalışmaya hız kazandıracağı, ayrıca EBYS alanındaki çalışmalara da fayda sağlayacağı açık bir şekilde görülmektedir.

DÖRDÜNCÜ BÖLÜM

4. ELEKTRONİK BELGE YÖNETİM SİSTEMLERİNDE VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ UYGULAMASI VE MODEL ÖNERİSİ

Bu bölümde öncelikli olarak uygulamada kullanılan kod kütüphaneleri hakkında bilgi verilmektedir. Sonrasında ise elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi süreçlerinin etkisinin ortaya konulması amacıyla bir devlet üniversitesinin elektronik belge yönetim sistemi üzerinde gerçekleştirilen uygulama ve analizler yer almaktadır. Bu bölümde son olarak TS 13298 standardına dayalı tüm elektronik belge yönetim sistemleri için kullanılabilecek bir model önerisi sunulmaktadır.

4.1.Veri Madenciliği ve Makine Öğrenmesinde Sık Kullanılan Kod Kütüphaneleri

Python Programlama Dili

Veri madenciliği ve makine öğrenmesi alanlarındaki uygulamaların geliştirilmesinde Python programlama dili birçok araştırmacı ve uygulama geliştiricisi tarafından öncelikli olarak kullanılmaktadır.

Python nesne yönelimli, modüler, etkileşimli ve yüksek seviyeli bir programlama dilidir. Modüler yapısı sayesinde, sınıf yapısını ve her türlü veri alanı girişini desteklemektedir. Python Windows, Linux / Unix ve Mac-OS üzerinde çalışmasının yanı sıra Java ve .NET sanal makinelerinde de çalışmaktadır. Açık kaynak kodlu ve ücretsiz bir yazılımdır.

Bilimsel araştırmalar da dahil olmak üzere birçok alanda yaygın olarak kullanılan Python'un yoğun bir şekilde tercih edilme nedenlerinden kısaca bahsetmek faydalı olacaktır. Öncelikle Python'un tutarlı, modüler ve anlaşılır bir sözdizimine sahip olması, kodun semantik açıdan okunabilirliğini artırarak, araştırmacıların ve uygulama geliştiricilerin karmaşık algoritmaları daha etkin bir şekilde gerçekleştirebilmesine olanak tanır. Ayrıca bu nitelikler Python'un diğer birçok programlama diline göre daha kolay öğrenilmesini ve bu alanda yeni başlayanlar için bile kolayca adapte olunabilir bir dil olmasını sağlar.

İkinci olarak Python, veri analizi, veri madenciliği, veri görselleştirmesi ve makine öğrenmesi gibi alanlar için hazırlanmış kütüphaneler ve araçlar konusunda oldukça zengin bir dildir. Python diline özgü kütüphaneler, bahsedilen alanlarda nitelikli çalışmalar yapmayı ve uygulamalar geliştirmeyi kolaylaştırır. Özellikle pandas, scikit-learn, TensorFlow ve PyTorch gibi kütüphaneler, bu disiplindeki temel ihtiyaçlara yanıt vermek üzere tasarlanmıştır. Bu kütüphanelerin sağladığı hazır fonksiyonlar ve araçlar, araştırmacıların ve uygulama geliştiricilerin karmaşık algoritmaları sıfırdan yazma zorunluluğunu ortadan kaldırarak, araştırmalarını ve projelerini daha hızlı ilerletmelerini sağlamaktadır.

Son olarak, Python'un geniş ve aktif bir kullanıcı topluluğuna sahip olması, bu dilin öne çıkmasını destekleyen bir diğer kritik faktördür. Bu aktif topluluk sayesinde, karşılaşılan problemler için çözümler hızla bulunabilmekte, yeni geliştirilen kütüphaneler ve teknikler hakkında bilgi alışverişinde bulunulabilir.

4.1.1. Temel Kütüphaneler

4.1.1.1 NumPy

NumPy, Python ile bilimsel hesaplama yapmak için kullanılan bir programlama kütüphanesidir. Numerical Python'un kısaltması olan NumPy kütüphanesi çok boyutlu diziler ve matrisler oluşturmaya yarar ve bu diziler üzerinde çalışmak için bir matematiksel fonksiyon kütüphanesine sahiptir. Temel doğrusal cebir, Fourier dönüşümü ve rastgele sayı işlemlerini destekleyen NumPy, çeşitli bilimsel hesaplamalarda matematiksel ve sayısal görevlerde önemli bir işlev görmektedir.

4.1.1.2.SciPy

Çok sayıda bilimsel hesaplamayı içeren bir Python kütüphanesi ve Scientific Python'un kısaltması olan SciPy, NumPy'nin üzerine inşa edilmiştir. Sayısal entegrasyon, interpolasyon, optimizasyon, lineer cebir gibi görevler için oldukça verimli fonksiyonlara sahiptir. Birçok farklı alanda bilimsel hesaplamalar yapılabilen bu kütüphanede ayrıca çok sayıda farklı alt modül de mevcuttur.

4.1.1.3. Pandas

Pandas veri analitiği alanında, özellikle veri analizi ve manipülasyonu için sıklıkla kullanılan bir Python kütüphanesidir. Bu kütüphane, heterojen veri yapılarıyla çalışma yeteneğine sahiptir. Çok boyutlu veri setleri üzerinde esneklikle çalışabilme olanağı sağlayan Pandas, DataFrame ve Series gibi temel veri yapıları sayesinde hem sıralı hem

de isimlendirilmiş indekslere sahip veri koleksiyonları ile etkili bir şekilde çalışmayı mümkün kılar.

Veri temizleme, dönüştürme ve önişleme gibi veri analizinin önemli adımlarında ihtiyaç duyulan fonksiyonları ve metotları bünyesinde barındırır. Eksik veri işleme, veri birleştirme, yeniden şekillendirme ve gruplama gibi işlevsellikleri sayesinde veri analizi süreçlerini optimize eder.

Pandas ile veri analizi süreçlerinde performans optimizasyonu elde edilebilir. Kütüphane, büyük veri setleri üzerinde hızlı ve etkili işlemler yapabilen algoritmalar içerdiğinden veri setlerinin hızla işlenmesi ve analiz edilmesi sağlanır.

Sonuç olarak pandas kütüphanesi esnek veri yapıları, kapsamlı veri manipülasyon araçları ve performans avantajları sunmaktadır.

4.1.2. Veri Görselleştirme Kütüphaneleri

4.1.2.1. Matplotlib

Matplotlib, grafik işlemleri ve iki boyutlu görselleştirme işlemleri için en çok kullanılan Python kütüphanelerinden biridir. Statik, animasyonlu ve etkileşimli görselleştirmeler oluşturmak için kullanılan Matplotlib çizimleri uygulamalara yerleştirmek için nesne yönelimli bir API ve özelleştirmeye izin veren fonksiyonlara sahiptir.

4.1.2.2. Seaborn

Seaborn, Python'da ilgi çekici ve bilgilendirici istatistiksel grafikler oluşturmak için kullanılan bir kütüphanedir. Bu kütüphane, veri görselleştirmeyi verileri araştırmanın ve keşfetmenin merkezi haline getirmeyi amaçlamaktadır. Özellikle birden fazla değişken içeren karmaşık veri kümelerini görselleştirmek için kullanılabilen Seaborn'un bir başka kullanışlı yanı da Pandas veri yapılarıyla sorunsuz bir şekilde entegre olmasıdır. Ayrıca, Matplotlib grafiklerini şekillendirmek için tek veya iki değişkenli dağılımları çizmek ya da veri alt kümeleri arasında karşılaştırmalar yapmak için geniş bir araç setine sahiptir.

4.1.2.3. Plotly

İnteraktif görseller oluşturmaya yarayan bir grafik kütüphanesi olan Plotly'da, kullanıcıların Jupyter platformunda kullanabildiği ya da web uygulamalarına gömülebilen etkileşimli grafikler ve gösterge tabloları oluşturulabilmektedir. Plotly'nin

grafikleri görsel olarak sunma biçimi ve verilerden daha fazla bilgi elde etmek için sunulan araçlar oldukça kullanıcı dostu bir yapıya sahiptir.

4.1.3. Makine Öğrenmesi Kütüphaneleri

4.1.3.1. Scikit Learn

Scikit-learn makine öğrenmesi alanında en yaygın olarak kullanılan kütüphanelerden biridir. Scikit-learn kütüphanesinin bu kadar popüler olmasının birkaç sebebi vardır. Öncelikle scikit-learn kütüphanesi, sınıflandırma, regresyon, kümeleme ve boyut indirgeme gibi birçok temel makine öğrenmesi algoritmasını kapsamlı bir şekilde sunar. Bu algoritmaların standartlaşmış ve optimize edilmiş uygulamaları sayesinde, uygulama geliştirmek isteyenler kendi algoritmalarını sıfırdan yazma yükümlülüğünden kurtulmaktadır.

Scikit-learn veri önişleme, model seçimi ve değerlendirme gibi makine öğrenmesi süreçlerinin kritik aşamaları için hazırlanmış araçlar ve fonksiyonlar sunar. Bu durum, model eğitimi ve testi sırasında karşılaşılan zorlukların üstesinden gelmeyi kolaylaştırır, böylece araştırmacılar modelin doğruluğunu ve genelleştirme kapasitesini optimize edebilir.

Scikit-learn'ın modüler ve uyumlu yapısı sayesinde NumPy ve Pandas gibi veri işleme kütüphaneleriyle uyumlu bir şekilde çalışabilme yeteneğine sahiptir. Bu özelliğiyle veri hazırlama ve modelleme süreçlerinin birbirine sorunsuz bir şekilde bağlanmasını sağlar.

Netice itibariyle scikit-learn kütüphanesi, makine öğrenmesi alanında geniş algoritma yelpazesi, kapsamlı yardımcı araçlar ve modüler bir yapı sunmaktadır.

4.2. Elektronik Belge Yönetim Sisteminde Veri Madenciliği ve Makine Öğrenmesi Uygulaması: Üniversite Örneği

Bu bölümde uygulama modelimizi oluşturmadan önceki veri hazırlama ve temizleme ile algoritmaların çalıştırılması ve sonuçların elde edilmesi aşamalarında gerçekleştirilen işlemler sıra düzeni içinde sunulacaktır. Bu noktada söz konusu süreçteki tüm detayları aktarmak yerine temsil kabiliyeti yüksek ve ana çerçeveyi ortaya koyan unsurlar sıralanacaktır.

4.2.1. Veri Hazırlama İşlemleri

Uygulamada yapılan işlemler ve konu başlıkları

- Standart dosya planı koduyla ilgili işlemler,
- Belge ve belgeyle ilgili işlemler,
- Birimlerle ilgili işlemler,
- Kullanıcılarla ilgili işlemler olarak sıralanabilir.

Standart Dosya Planı Kodlarıyla İlgili İşlemler

TS 13298 standardının "*4.1.1. EBYS ait olduğu kurumun yapısını ve fonksiyonlarını yansıtacak bir dosya tasnif planını içinde barındırmalı ve / veya kurum dosya tasnif planı ile uyumlu olmalıdır.*" ve "*4.1.2. EBYS içerisinde temsil edilecek olan dosya tasnif planı hiyerarşik bir yapıda olmalı ve asgari üç seviyeden oluşmalıdır. Asgari seviye tercih edildiğinde birim, seri ve dosya seviyeleri tercih edilmelidir.*" maddeleri göz önünde bulundurularak ilgili kontroller yapılmıştır.

Ana konular, "Üç karakter sayısal kodla tanımlanan ana konular, konu grubundaki temel işlevleri/dosyaları tanımlar. Yalın olarak verilen ana konular, yazışmalarda ve dosyalamada kullanılır. Ana konuların alt konulara bölünmesi halinde ise sadece dosya tanımlaması özelliğini taşır." olarak tanımlanabilir (Özdemirci ve diğerleri, 2009, s.182).

Alt konular ise "İki sayısal rakamla ifade edilen alt konular, ana konuya bağlı bölünmeleri gösterir. Planda, ana konu üç alt bölünmeyle sınırlandırılmıştır. Alt bölünmeler, ihtiyaç halinde kendi seviyelerinde "99" a kadar bölünebilir. Yalın olarak verilen alt konular, yazışmalarda ve dosyalamada kullanılır. Alt konuların kendine bağlı daha alt konulara bölünmesi halinde ise sadece dosya tanımlaması özelliğini taşır. Ana konu ve alt bölünmelerine ilişkin örnek tablo aşağıda verilmiştir." olarak ifade edilmiştir (Özdemirci ve diğerleri, 2009, s.183).

Standart dosya planı kodları değerlendirilirken alt konuların bölünme sayıları birbirinden farklı seviyelerde olabileceği göz önünde bulundurularak aşağıda gösterildiği şekilde bir düzenleme yapılması gerekmiştir.

Ana Dosya	**1.Alt Konu**	**2.Alt Konu**	**3.Alt Konu**	
000				***Genel***
010				***Mevzuat İşleri***
	01			Kanunlar
	02			Tüzükler
	03			Yönetmelikler
	04			Yönergeler

	05			Tebliğler
	06			Genelgeler
		01		*İç Genelgeler*
		02		*Dış Genelgeler*
…	**…**	**…**	**…**	**…**
100				**Eğitim - Öğretim İşleri (Genel)**
101				**Akademik Birimlerin Kurulması**
	01			Üniversite
		01		*Devlet Üniversiteleri*
		02		*Vakıf Üniversiteleri*
			01	Mütevelli Heyeti
			02	Kuruluş ve Eğitim-Öğretime Açılması
			03	Devlet Yardımları
			04	Öğrenci Bilgileri
			05	Akademik Personel Bilgileri
	02			Fakülte, Enstitü, Konservatuar, Yüksekokul, MYO
…	**…**	**…**	**…**	**…**

Tabloda "010. Mevzuat İşleri" altında bulunan 010.06. kodlu "Genelgeler";

- 010.06.01. İç Genelgeler ve
- 010.06.01. Dış Genelgeler olmak üzere ikiye ayrılmıştır.

Veri setinin standart dosya planını tutan ilgili tablosunda "010.06.01. İç Genelgeler" altında aşağıda gösterildiği gibi farklı yıllarda açılmış klasörler/bölümler bulunabilir.

Burada şu hususun ifade edilmesi önem arz etmektedir. Elektronik belge üzerindeki sayı kısmında SDP kodları "010.06.01" şeklinde gösterilse de EBYS veri tabanında bu kod "010.06.01.05.01" olarak kaydedilir.

010.06.01	İç Genelgeler	Genelgeler-İç Genelgeler
010.06.01.01	Klasör 01-2019	Genelgeler-İç Genelgeler-Klasör 01
010.06.01.01.01	Bölüm 01	Genelgeler-İç Genelgeler-Klasör 01-Bölüm 01
010.06.01.02	Klasör 02-2020	Genelgeler-İç Genelgeler-Klasör 02
010.06.01.02.01	Bölüm 01	Genelgeler-İç Genelgeler-Klasör 02-Bölüm 01
010.06.01.03	Klasör 03-2021	Genelgeler-İç Genelgeler-Klasör 03
010.06.01.03.01	Bölüm 01	Genelgeler-İç Genelgeler-Klasör 03-Bölüm 01
010.06.01.04	Klasör 04-2022	Genelgeler-İç Genelgeler-Klasör 04
010.06.01.04.01	Bölüm 01	Genelgeler-İç Genelgeler-Klasör 04-Bölüm 01
010.06.01.05	Klasör 05-2023	Genelgeler-İç Genelgeler-Klasör 05
010.06.01.05.01	Bölüm 01	Genelgeler-İç Genelgeler-Klasör 05-Bölüm 01

Farklı yıllarda da oluşturulmuş olsa "İç Genelgeler" konusunun altında bulunan bütün belgelerin metin içeriklerinin iç genelgelerle ilgili olacağı düşünüldüğünde, yukarıda gösterilen tablodaki tüm SDP kodlarını "010.06.01" olarak ele almamız mümkündür. Bu noktada istisnai bir durum söz konusu olabilir: Eğer makine öğrenmesi algoritmalarının eğitimi aşamasında belgeler yıllara göre ayrı ayrı değerlendirilmek isteniyorsa SDP kodlarının "010.06.01" olarak kullanılmaması ve bir alt kırılıma göre ele alınması gerekmektedir. Örnek olarak algoritmanın, "010.06.01.05" kodunun altındaki tüm belgeler bir üst konu kodu olan "010.06.01" kodu altında değil, kendi içinde eğitilmesi daha doğru olacaktır.

Veri setinde bulunan tüm belgelere, SDP kodlarını bahsedildiği şekilde de gösterecek ilave bir sütun eklenerek makine öğrenmesi algoritmalarının farklı yıllarda oluşturulmuş aynı konudaki belgeleri bir arada değerlendirmesi için gerekli düzenleme yapılmıştır. Bahsedilen durumun daha iyi anlaşılması için başka bir örnek daha vermek gerekirse, yukarıdaki örnekte "010. Mevzuat İşleri" ana konusu iki alt bölünmeye sahip olsa da birçok ana konuda üç alt bölünme olabilir. Bu noktada da yukarıdaki mantıkla hareket edildiğinde en alt konunun altında oluşturulan klasör/bölümleri, dahil olduğu alt konu altında değerlendirmek gerekeceği açıktır.

Örneğin, "101. Akademik Birimlerin Kurulması" altındaki "101.01. Üniversiteler" konusu kendi içinde "101.01.01. Devlet Üniversiteleri" ve "101.01.02. Vakıf Üniversiteleri" olmak üzere iki alt konuya ayrılmaktadır. "101.01.01." ve "101.01.02." kodları da bir kırılım daha yaparak "Mütevelli Heyeti", "Kuruluş ve Eğitim-Öğretime Açılması", "Devlet Yardımları", "Öğrenci Bilgileri", "Akademik Personel Bilgileri" şeklinde beş alt konuya ayrılmaktadır.

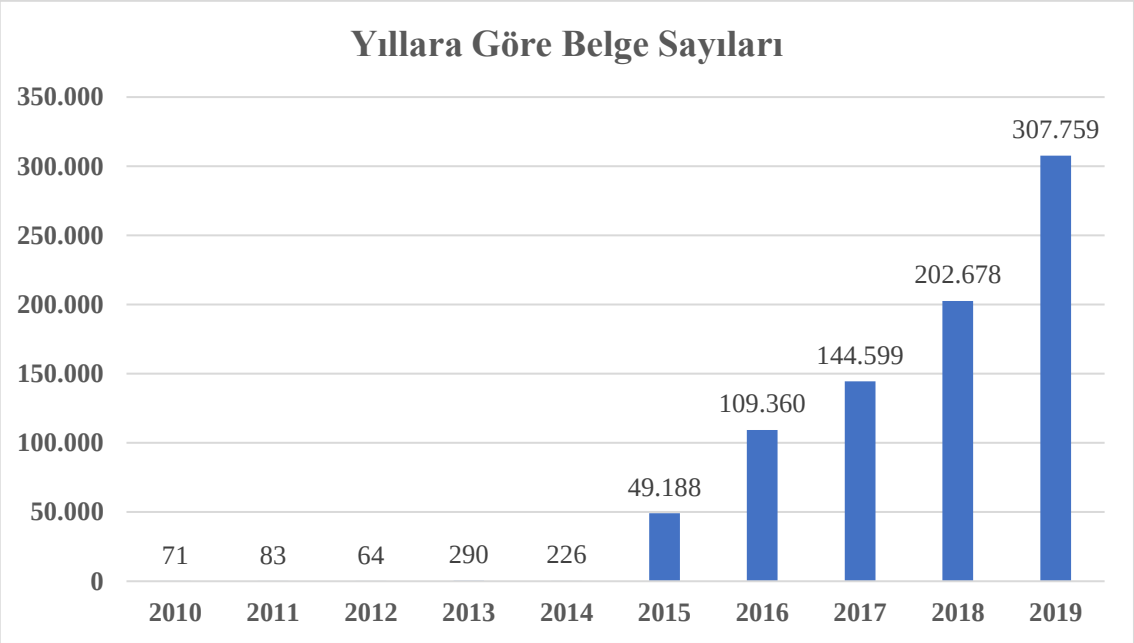
"010. Mevzuat İşleri" iki alt konuya ayrılırken "101. Akademik Birimlerin Kurulması" üç alt konuya ayrılmıştır. Tüm SDP ağacında bu tür farklı seviyede alt konu ayrımları olabileceği bilinmeli ve bu ayrımlar göz önünde bulundurularak veri setinde gerekli düzenlemelerin yapılmasına dikkat edilmelidir.

Belgelerle İlgili İşlemler

TS 13298 standardında ifade edildiği gibi, "11.1.1. Elektronik ortamda üretilen dokümanlardan belge statüsü kazananlar EBYS içerisinde tanımlanabilir olmalıdır. Tanımlanabilirlik, herhangi bir elektronik belge üreticisi, yazarı, alıcısı ve belgeye ait tarih bilgilerinin kayıt altına alınması ile sağlanır."

Yıllara Göre Belge Sayıları

Veri setindeki belgelerin yoğunluklarının tespit edilmesi amacıyla yıllara göre belge sayıları bulunmuş ve grafikte gösterilmiştir.



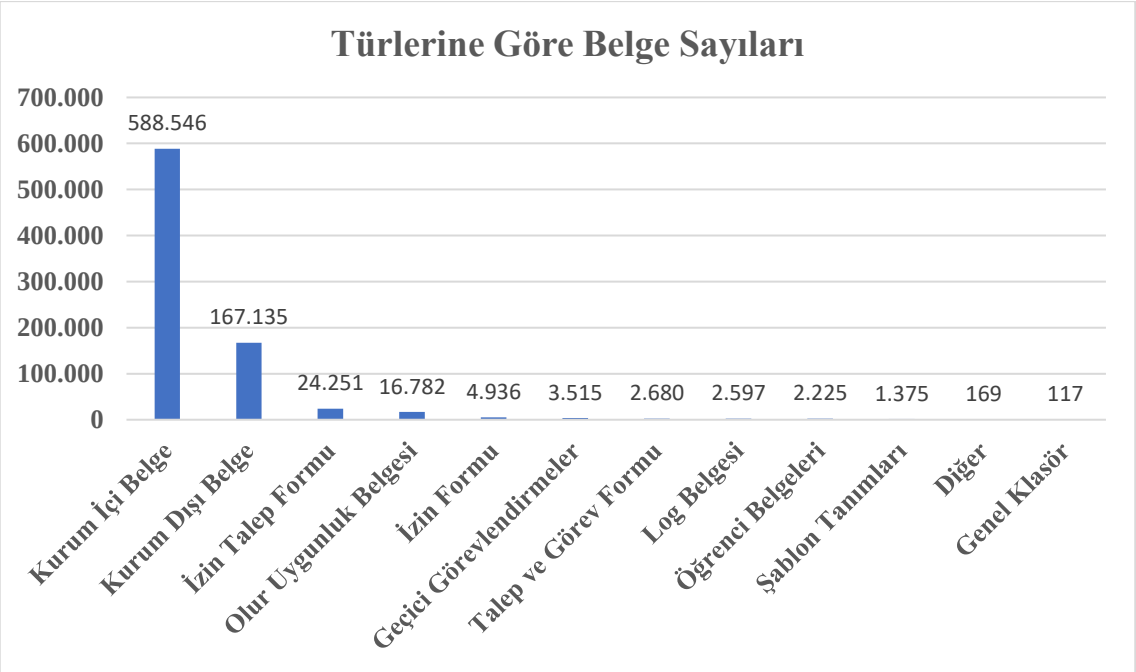
Şekil 6: Yıllara Göre Belge Sayıları

Tablodan da anlaşılacağı üzere 2015 yılından itibaren EBYS'de üretilen belgelerin sayısında artış gözükmektedir. 2010-2014 yılları arasında üretilen belgelerin metinlerine bakıldığında büyük çoğunluğunun sistemi test etmek amacıyla üretildiği tespit edilmiş ve bu durum yapılan görüşmeler sonucunda teyit edilmiştir. Bundan dolayı 2015 yılından itibaren oluşturulan belgeler üzerinde veri hazırlama sürecine devam edilmesine karar verilmiştir.

2015-2019 yıllarında toplam 813.584 adet belge bulunmaktadır. Algoritmaların eğitiminde kullanılacak nihai veri setinde ise 64.103 adet belgeye düşürülmüştür. Veri hazırlama sürecinin ilerleyen aşamalarında veri setinde kullanılacak belge sayılarındaki bu azalmanın sebepleri de ayrıca ele alınacaktır.

Türlerine Göre Belge Sayıları

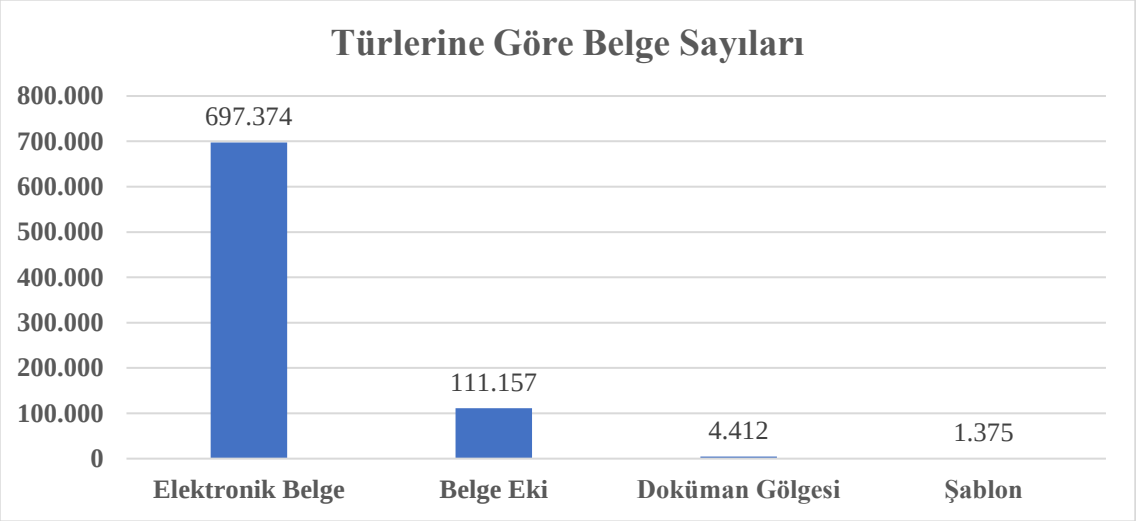
Veri setinde EBYS'nin işleyişinde kullanılan farklı belge türleri bulunmaktadır ve belge türlerine göre belge sayıları aşağıdaki grafikte gösterilmiştir.



Şekil 7: Türlerine Göre Belge Sayıları

Tablodan da anlaşılacağı üzere veri setinde bulunan belgelerin %72.27 oranıyla büyük çoğunluğu kurum içinde üretilen belgeler olarak gözükmektedir. "Diğer" türünde bulunan belge türlerinden bazıları ise Senato Kararları, Yönetim Kurulu Kararları, Fakülte Kurulu Kararları ve Test belgeleri olarak sayılabilir. Türlerine göre belge sayıları toplamı da yukarıda ifade edildiği gibi 813.584'tür.

Ayrıca belge türlerini ifade etmek için veri setinde başka bir tür tanımı daha yapılmış ve bunlarla ilgili grafik de aşağıda gösterilmiştir.



Şekil 8: Türlerine Göre Belge Sayıları

Belge ve Bağlantılı Tabloların Eşleşmelerinin Yapılması

Veri seti üzerinde yapılan bir diğer önemli işlem de belgelerle ilişkili verileri tutan diğer tabloların eşleşmelerinin yapılmasıdır. İlk yapılan eşleştirme belge ve belgenin SDP kodunun eşleştirmesi olmuştur. Bu işlem gerçekleştirildikten sonra her bir SDP kodunda var olan belge sayıları tespit edilmiştir. Çok fazla SDP kodu olduğundan tüm SDP kodlarındaki belge sayılarını bir grafik halinde göstermenin zor olacağı aşikardır. Bunun yerine SDP ile ilgili tablonun sadece küçük bir kısmını göstermenin yeterli olacağı düşünülmüştür.

SDP Kodu	**Standart Dosya Planı Adı**	**Belge Sayıları**
010.01.02.01	Genel İşler-Mevzuat İşleri-Kanunlar-Klasör 01-Bölüm 01	2
010.01.04.01	Genel İşler-Mevzuat İşleri-Kanunlar-Klasör 02-Bölüm 01	6
010.01.05.01	Genel İşler-Mevzuat İşleri-Kanunlar-Klasör 03-Bölüm 01	465
010.03.01.01	Genel İşler-Mevzuat İşleri-Yönetmelikler-Klasör 01-Bölüm 01	7
010.03.02.01	Genel İşler-Mevzuat İşleri-Yönetmelikler-Klasör 02-Bölüm 01	3
010.03.03.01	Genel İşler-Mevzuat İşleri-Yönetmelikler-Klasör 03-Bölüm 01	4
010.03.04.01	Genel İşler-Mevzuat İşleri-Yönetmelikler-Klasör 04-Bölüm 01	25
010.03.05.01	Genel İşler-Mevzuat İşleri-Yönetmelikler-Klasör 05-Bölüm 01	182
010.04.01.01	Genel İşler-Mevzuat İşleri-Yönergeler-Klasör 01-Bölüm 01	1
010.04.02.01	Genel İşler-Mevzuat İşleri-Yönergeler-Klasör 02-Bölüm 01	7
010.04.03.01	Genel İşler-Mevzuat İşleri-Yönergeler-Klasör 03-Bölüm 01	7
010.04.04.01	Genel İşler-Mevzuat İşleri-Yönergeler-Klasör 04-Bölüm 01	29
010.04.05.01	Genel İşler-Mevzuat İşleri-Yönergeler-Klasör 05-Bölüm 01	445
…	…	…

Belge-SDP kodu eşleşmesi ve SDP kodlarındaki belge sayıları bulunduktan sonra yukarıda bahsettiğimiz, algoritmaların tüm belgelerle ve her bir belgeyi de kendi SDP kodu içinde eğitilmesini sağlayacak SDP kodlarının en alt seviye konu kodunda birleştirilmesi işlemi yapılmış ve sonuç aşağıdaki tabloda gösterildiği gibi olmuştur.

SDP Kodu	**Standart Dosya Planı Adı**	**Belge Sayıları**
010.01.02.01	Genel İşler-Mevzuat İşleri-Kanunlar	473
010.03.01.01	Genel İşler-Mevzuat İşleri-Yönetmelikler	221
010.04.01.01	Genel İşler-Mevzuat İşleri-Yönergeler	489
…	…	…

Belge metinleri farklı bir tabloda bulunduğu için belge ve belge metninin eşleştirilmesi yapılan bir diğer işlemdir. Bu işlem esnasında metni aynı olan belgelerin

varlığı tespit edilmiş ve aynı metne sahip olan belge sayıları bulunmuştur. Aşağıdaki tabloda metni aynı olan bazı belgeler ve belge sayıları gösterilmektedir.

Belge Metni	**Sayı**
Üniversitemizde görev yapan ekli listede adı, soyadı, unvanı ve kadro durumu belirtilmiş olan akademik personelin 2914 sayılı Kanunun 8'inci maddesi, idari personelin ise 657 sayılı Kanunun 64'üncü maddesi uyarınca hizalarında belirtilen tarihlerden geçerli olmak üzere görev aylığı, kazanılmış hak aylığı ve emekli keseneğine esas aylıklarında kademe ilerlemesi yapılmasını ve kıdem yıllarının artırılmasını olurlarınıza arz ederim.	759
Üniversitemiz programlarına yurt dışından öğrenci kabulünde 2019-2020 Eğitim-Öğretim yılından itibaren kabul edilecek Yabancı Uyruklu Öğrenci Sınavları (YÖS) ekte sunulmuştur. Bilgilerinize arz ederim.	282
Yükseköğretim Kurulu Başkanlığından alınan "Kamu Çalışanı Anketi" konulu ilgi yazı ve eki ilişikte gönderilmiştir. Cumhurbaşkanlığı İnsan Kaynakları Ofisi tarafından kamu çalışanlarının mevcut İnsan Kaynakları süreçleriyle ilgili uygulamalar hususundaki görüş ve düşünceleri ile gelişime açık yönlerin belirlenmesi amacıyla oluşturulan söz konusu anketin, statüsüne bakılmaksızın tüm personele tebliğ edilmesi ve 26 Temmuz 2019 tarihine kadar anket.cbiko.gov.tr adresinden doldurulmasının sağlanması hususunda; Bilgilerinizi ve gereğini rica ederim.	156
Öğretim Üyeliğine Yükseltilme ve Atanma Yönetmeliği 12 Haziran 2018 tarih ve 30449 sayılı Resmi Gazete'de yayımlanarak yürürlüğe girmiş olup ilgili yönetmelik metni ekte gönderilmiştir. Bilgilerinizi ve gereğini rica ederim.	124

Makine öğrenmesi algoritmaları belge metni üzerinde eğitilirken önemli olan husus, aynı belge metninin veri setinde birçok defa bulunması değildir. Söz konusu belge metninin bir kez bulunması yeterli olduğu için tekrarlanan metinlerin bir işlevi bulunmamaktadır. Bu yüzden metinleri aynı olan belgeler tekilleştirilmiş ve diğer belgeler veri setinden silinmiştir. Ayrıca yukarıdaki belge metni örneklerinde metin haricinde HTML kodları da bulunmaktadır. Veri seti hazırlandıktan sonra veri temizleme aşamasında HTML kodları gibi metin dışı olan tüm ifadeler silinecek ve düzenli bir hale

getirilecektir. Veri temizleme işlemlerinin detaylarına ilerleyen aşamalarda değinilecektir.

Aynı metne sahip olan belgelerle ilgili detaylar incelendiğinde bu durumun dağıtımlı belgeler oluşturulurken çoğaltıldığı ve kaç tane birime ve/veya personele gönderilecekse o kadar sayıda oluşturulup veri tabanında tutulduğu tespit edilmiştir.

Belge metni tekilleştirilmesinden önce 813.584 olan belge sayısı bu işlemden sonra 64.103 adete düşürülmüştür.

Kullanıcılarla İlgili İşlemler

TS 13298 standardının kullanıcılarla ilgili bölümleri göz önünde bulundurularak her bir kullanıcı için bir fonksiyon (kullanıcı, yönetici gibi), rol ve grup tanımları kontrol edilmiş ve kullanıcılarla ilgili sadece kullanıcıID, durum (aktif / pasif), pozisyon ve birim verileri kullanılmıştır.

Birimlerle İlgili İşlemler

Standart dosya planı, belge ve belgelerle ilişkili tablolara göre daha düzenli ve az kayıt sayısı olan birim tablosunda herhangi bir sütun ekleme ve/veya düzenleme işlemine gerek duyulmamıştır. Birimleri gösteren tablonun bir bölümü aşağıda verilmiştir.

Birim Kodu	**Üst Birim Kodu**	**Seviye**	**Birim Adı**
6	-1	1	Rektörlük
6.2	6	2	Genel Sekreterlik
6.6	6	2	Enstitüler
6.7	6	2	Yüksekokullar
6.8	6	2	Fakülteler
6.10	6	2	Araştırma Merkezleri/Koordinatörlükler/Laboratuvar/Ofisler
6.11	6	2	Beden Eğitimi ve Spor Bölüm Başkanlığı
6.12	6	2	Türk Dili Koordinatörlüğü
6.13	6	2	Ortak Dersler Bölüm Başkanlığı
6.15	6	2	Etik Kurulu
6.16	6	2	Fikri Sınai Mülkiyet Hakları Kurulu
6.17	6	2	Fen, Mühendislik ve Sosyal Bilimleri Araştırmaları Etik Kurulu
6.18	6	2	Teknoloji Transfer Ofisi
6.19	6	2	Özel Kalem Birimi
6.20	6	2	Kurumsal Gelişim Ofisi
…	…	…	…

4.2.2. Modelin Oluşturulması ve Değerlendirilmesi

```
# Kullanılacak kod kütüphanelerinin dahil edilmesi
import pandas as pd
import re
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

```
# NLTK kütüphanesinin stopword listesini indirme
import nltk
nltk.download('punkt')
nltk.download('stopwords')
```

Modeli oluşturmadan önce veri seti üzerinde yapılacak veri analizi ve temizleme işlemleri için gerekli kod kütüphaneleri dahil edilmiş ve bir doğal dil işleme kütüphanesi olan NLTK (Natural Language Toolkit)'den noktalama işaretleri ve sıklıkla geçen kelimeleri ifade eden stopwords (durak kelimeleri) listesi indirilmiştir.

Stopwords, metin madenciliği ve doğal dil işleme (NLP) uygulamalarında genellikle önemli olmayan ve veri setinden çıkarılması gerektiği düşünülen kelimeler listesi olarak ifade edilmektedir. Bu kelimeler, metin içerisinde genellikle çok sık tekrar etmesine rağmen genellikle metnin anlamını belirlemede az rol oynamaktadır.

```
stop_words = stopwords.words('turkish')
```

```
print(stop_words)
```

```
['acaba', 'ama', 'aslında', 'az', 'bazı', 'belki', 'biri', 'birkaç', 'birşey', 'biz', 'bu', 'çok', 'çünkü', 'da', 'daha',
'de', 'defa', 'diye', 'eğer', 'en', 'gibi', 'hem', 'hep', 'hepsi', 'her', 'hiç', 'için', 'ile', 'ise', 'kez', 'ki', 'kim',
'mı', 'mu', 'mü', 'nasıl', 'ne', 'neden', 'nerde', 'nerede', 'nereye', 'niçin', 'niye', 'o', 'sanki', 'şey', 'siz', 'şu',
'tüm', 've', 'veya', 'ya', 'yani']
```

Sonraki aşamada stopwords listesinin içeriği yazdırılmıştır. Çıktıdan da görüleceği üzere metinler içerisinde sıklıkla geçen 'acaba', 'ama', 'için', 've', 'veya' gibi varsayılan bir kelime listesi bulunmaktadır. Bu listeye, uygulamanın geliştirildiği alana özel kelimeler ilave etmek mümkündür. Bu sebepten EBYS belge metinlerinin sonunda kullanılan 'bilgilerinize', 'bilgilerinizi', 'gereğini', 'arz', 'rica', 'ederim' ve 'ederiz' kelimeleri de bu listeye dahil edilmiştir.

```
## Stopwords listesine EBYS belge metinlerinde sık geçen kelimeleri ekleme
stop_words.extend(['bilgilerinize','bilgilerinizi', 'gereğini', 'arz', 'rica', 'ederim', 'ederiz'])
print(stop_words)
```

```
['acaba', 'ama', 'aslında', 'az', 'bazı', 'belki', 'biri', 'birkaç', 'birşey', 'biz', 'bu', 'çok', 'çünkü', 'da', 'daha',
'de', 'defa', 'diye', 'eğer', 'en', 'gibi', 'hem', 'hep', 'hepsi', 'her', 'hiç', 'için', 'ile', 'ise', 'kez', 'ki', 'kim',
'mı', 'mu', 'mü', 'nasıl', 'ne', 'neden', 'nerde', 'nerede', 'nereye', 'niçin', 'niye', 'o', 'sanki', 'şey', 'siz', 'şu',
'tüm', 've', 'veya', 'ya', 'yani', 'bilgilerinize', 'bilgilerinizi', 'gereğini', 'arz', 'rica', 'ederim', 'ederiz']
```

Stopwords listesinin son hali yukarıda gösterilmiştir.

```
def preprocess_text(text):
    # HTML tag'lerini kaldırma
    soup = BeautifulSoup(text, "html.parser")
    text = soup.get_text(separator=" ")

    # Metni küçük harfe çevirme
    text = text.lower()

    # Noktalama işaretlerini kaldırma
    text = re.sub(r'[^\w\s]', '', text)

    # Stopword'leri çıkar
    tokenized_text = word_tokenize(text)
    text = [word for word in tokenized_text if word not in stop_words]

    return " ".join(text)
```

Daha önceden bahsettiğimiz üzere veri setindeki belge metinlerinde bulunan metin harici HTML etiketlerinin çıkarılması, metnin küçük harfe dönüştürülmesi, noktalama işaretlerinin kaldırılması ve stopwords kelimelerinin çıkarılması için ileride kullanılacak bir fonksiyon tanımlanmıştır.

```
veri = pd.read_csv('siniflandirma_v01.csv', sep = ';', encoding='utf-8')
veri.head()
```

	belgeID	konu	metin	SDPKodu	SDPKodGrubu	SDPAdı	SDPTamAdı
0	71867	İzleme ve Değerlendirme	<p>Üniversitemizin 2014 yılı yatırım programı ...	602.03.02.01.01	602.03.02	İzleme ve Değerlendirme	Danışma-Denetimle İlgili Faaliyetler|Plan ve P...
1	71889	İzleme ve Değerlendirme	<p style="text-align: justify;">Üniversitemizi...	602.03.02.01.01	602.03.02	İzleme ve Değerlendirme	Danışma-Denetimle İlgili Faaliyetler|Plan ve P...
2	71906	İzleme ve Değerlendirme	<p style="text-align: justify;">Üniversitemizi...	602.03.02.01.01	602.03.02	İzleme ve Değerlendirme	Danışma-Denetimle İlgili Faaliyetler|Plan ve P...
3	71996	Danışman Atama İşleri	<p>Üniversitemiz Öğretim Üyelerinden Doç...	302.13.01.01	302.13	Danışman Atama İşleri	Ana Hizmet Faaliyetleri|Öğrenci Özlük İşleri*|...
4	72006	İlan Yayınlama	<p style="margin-bottom: 0.0001pt; text-align:...	934.02.07.01.01	934.02.07	İlan Metni	Yardımcı Hizmetlerle İlgili Faaliyetler|Satına...

Gerekli hazırlıklar yapıldıktan sonra oluşturulan veri seti programa dahil edilmiş ve ilk beş kaydı gösterilmiştir.

```
veri['temizMetin'] = veri.apply(lambda row: preprocess_text(row['metin']), axis=1)
veri.head()
```

	belgeID	konu	metin	SDPKodu	SDPKodGrubu	SDPAdı	SDPTamAdı	temizMetin
0	71867	İzleme ve Değerlendirme	<p>Üniversitemizin 2014 yılı yatırım programı ...	602.03.02.01.01	602.03.02	İzleme ve Değerlendirme	Danışma-Denetimle İlgili Faaliyetler|Plan ve P...	üniversitemizin 2014 yılı yatırım programı izl...
1	71889	İzleme ve Değerlendirme	<p style="text-align: justify;">Üniversitemizi...	602.03.02.01.01	602.03.02	İzleme ve Değerlendirme	Danışma-Denetimle İlgili Faaliyetler|Plan ve P...	üniversitemizin 2014 yılı yatırım programı izl...
2	71906	İzleme ve Değerlendirme	<p style="text-align: justify;">Üniversitemizi...	602.03.02.01.01	602.03.02	İzleme ve Değerlendirme	Danışma-Denetimle İlgili Faaliyetler|Plan ve P...	üniversitemizin 2014 yılı yatırım programı izl...
3	71996	Danışman Atama İşleri	<p>Üniversitemiz Öğretim Üyelerinden Doç...	302.13.01.01	302.13	Danışman Atama İşleri	Ana Hizmet Faaliyetleri|Öğrenci Özlük İşleri*|...	üniversitemiz öğretim üyelerinden doç dr deniz...
4	72006	İlan Yayınlama	<p style="margin-bottom: 0.0001pt; text-align:...	934.02.07.01.01	934.02.07	İlan Metni	Yardımcı Hizmetlerle İlgili Faaliyetler|Satına...	bursa teknik üniversitesi kablosuz ağ sistemi ...

Belge metinlerindeki HTML etiketlerinin çıkaracak ve yukarıda bahsedilen metin ön işleme adımlarını gerçekleştirecek fonksiyon kullanılarak veri setine 'temizMetin' isminde bir sütun eklenip belge metni buraya temizlenmiş şekilde kopyalanmıştır.

```
# Modeli oluşturacak makine öğrenmesi kütüphanelerinin dahil edilmesi
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
```

Kullanılacak makine öğrenmesi algoritmalarını içeren kod kütüphanesi ve ilgili algoritmalar programa dahil edilmiştir.

```
X = veri['temizMetin']
X.tail()
```

Algoritmanın belge metnine göre SDP kodunu tahmin etmek için kullanacağı sütun olan 'temizMetin' sütunu veri setinden ayrı bir değişkene aktarılmıştır.

```
y = veri['SDPAdı']
```

Algoritmanın tahmin edeceği SDP kodu da ayrı bir değişkene aktarılarak, model çalıştırılmaya hazır hale getirilmiştir.

```
# TF-IDF
vectorizer = TfidfVectorizer()
vectors = vectorizer.fit_transform(X)

# Eğitim ve test verilerini ayırma
X_train, X_test, y_train, y_test = train_test_split(vectors, y, test_size=0.2, random_state=58)

# Algoritmaları oluşturma
models = [
    MultinomialNB(),
    LinearSVC(),
    LogisticRegression(),
    RandomForestClassifier()
]

# Her bir modeli eğitme, tahmin yapma ve performansı değerlendirme
for model in models:
    model_name = model.__class__.__name__
    model.fit(X_train, y_train)
    predictions = model.predict(X_test)

    print(f"\n{model_name} Performance:")
    print(classification_report(y_test, predictions))
```

Yukarıda metinleri algoritmaların anlayabileceği hale dönüştürmek için kullanılan TF-IDF işlemi yapıldıktan sonra veri seti eğitim ve test veri setleri olmak üzere ikiye

ayrılmıştır. Son olarak metin sınıflandırma işlemlerini yapacak algoritmalar çalıştırılarak sonuçlar elde edilmiştir.

Naive Bayes algoritmasının çıktıları aşağıdaki gibi elde edilmiştir.

	precision	**recall**	**f1-score**	**support**
Kararlar	0.95	0.86	0.90	264
Yatay Geçiş	0.97	0.92	0.94	147
Raporlama	0.99	0.94	0.97	175
Yönetim Kurulu	0.97	0.96	0.96	389
Devir, Giriş-Çıkış İşlemleri	0.97	0.93	0.95	203
Mali İşler (Genel)	0.92	0.90	0.91	373
Asaleten	0.92	0.97	0.95	478
Kurum dışı	0.93	0.97	0.95	883
Sigorta Primleri Bildirgeleri	0.96	0.97	0.96	212
Faturalar	0.97	0.94	0.96	158
accuracy			**0.94**	3282
macro avg	0.96	0.94	0.95	3282
weighted avg	0.94	0.94	0.94	3282

Tablo 5: Naive Bayes algoritmasının sonuçları

Support vector machines algoritmasının çıktıları aşağıdaki gibi elde edilmiştir.

	precision	**recall**	**f1-score**	**support**
Kararlar	0.95	0.95	0.95	264
Yatay Geçiş	0.98	0.96	0.97	147
Raporlama	0.99	0.98	0.99	175
Yönetim Kurulu	0.98	0.98	0.98	389
Devir, Giriş-Çıkış İşlemleri	0.99	0.98	0.98	203
Mali İşler (Genel)	0.94	0.96	0.95	373
Asaleten	0.97	0.97	0.97	478
Kurum dışı	0.97	0.98	0.98	883
Sigorta Primleri Bildirgeleri	1.00	0.95	0.97	212
Faturalar	0.99	0.97	0.98	158
accuracy			**0.97**	3282
macro avg	0.98	0.97	0.97	3282
weighted avg	0.97	0.97	0.97	3282

Tablo 6: Support Vector Machines algoritmasının sonuçları

Logistic regression algoritmasının çıktıları aşağıdaki gibi elde edilmiştir.

	precision	**recall**	**f1-score**	**support**
Kararlar	0.96	0.91	0.93	264
Yatay Geçiş	1.00	0.93	0.96	147
Raporlama	1.00	0.98	0.99	175
Yönetim Kurulu	0.97	0.97	0.97	389
Devir, Giriş-Çıkış İşlemleri	0.99	0.96	0.98	203
Mali İşler (Genel)	0.93	0.96	0.94	373
Asaleten	0.97	0.98	0.98	478
Kurum dışı	0.94	0.99	0.96	883
Sigorta Primleri Bildirgeleri	0.99	0.93	0.96	212
Faturalar	0.99	0.96	0.97	158
accuracy			**0.96**	3282
macro avg	0.98	0.96	0.97	3282
weighted avg	0.97	0.96	0.96	3282

Tablo 7: Logistic Regression algoritmasının sonuçları

Random forest algoritmasının çıktıları aşağıdaki gibi elde edilmiştir.

	precision	**recall**	**f1-score**	**support**
Kararlar	0.97	0.95	0.96	264
Yatay Geçiş	0.96	0.95	0.96	147
Raporlama	1.00	0.98	0.99	175
Yönetim Kurulu	0.99	0.97	0.98	389
Devir, Giriş-Çıkış İşlemleri	0.99	0.97	0.98	203
Mali İşler (Genel)	0.95	0.92	0.93	373
Asaleten	0.96	0.97	0.97	478
Kurum dışı	0.94	0.98	0.96	883
Sigorta Primleri Bildirgeleri	0.99	0.94	0.97	212
Faturalar	0.99	0.97	0.98	158
accuracy			**0.96**	3282
macro avg	0.97	0.96	0.97	3282
weighted avg	0.97	0.97	0.97	3282

Tablo 8: Random Forest algoritmasının sonuçları

Algoritma	**F1 Skor**
Lineer Support Vector Machine (SVM)	0,97
LogisticRegression	0,96
RandomForest	0,96
Naive Bayes	0,94

Tablo 9: Sınıflandırma Algoritmalarının Karşılaştırılması

Sonuçlardan görüleceği üzere belge metninden SDP kodunu bulma işleminde Destek Vektör Makineleri (Support Vector Machine (SVM)) algoritmasının F1 skoru 0,97 olarak elde edilmiştir. Diğer sınıflandırma algoritmalarının da iyi sonuçlar elde ettiğini söylemek mümkündür.

SONUÇ VE ÖNERİLER

Veri madenciliği, değerli bilgiyi arar. Bu değerli bilgileri bulabilmek için gereken her şeyden önce büyük miktarlardaki veri yığınlarıdır. Bu noktada başarılı sonuçlar elde etmek için öncelikli olan, yeterli miktarlarda kaliteli, yani hata ve eksiklerin olmadığı veridir. Çünkü sonuçların, yani çıkarılan bilginin kalitesi, verinin kalitesine bağlıdır. Söz konusu veri yığınları istenen halde olmadıkça veri madenciliğinin amacı gerçekleşmez ve aradığı değerli bilgi ortaya çıkmaz.

Yaşamın her alanında, insanların oluşturduğu tüm kurumlarda belki de yazı var olduğu günden beri belgeler vazgeçilmez bir yere sahip olmuştur. Tarih boyunca belgenin doğası sürekli değişip yeni bir hal kazansa da bu yeri hiç kaybetmemiştir. Yakın tarihte teknoloji alanında yaşanan benzeri görülmemiş değişim ve dönüşümler içinde belgeler de elektronik varlıklar haline gelmiştir. Bu durum da ister istemez hem belgelere bakışımızı hem de onları ele alma ve yönetme biçimlerimizi bütünüyle dönüştürmüştür.

Belgelerin böylelikle elektronikleşmesi tahmin edileceği üzere miktar bakımından da bu varlıkların adeta sınırsız bir şekilde çoğalmasıyla sonuçlanmıştır. Başlangıçta bir sorun olarak görülebilecek bu durum, esasen yukarıda veri madenciliğinin her şeyden önce ihtiyaç duyduğunu dile getirdiğimiz büyük miktarda veri yığınları olarak görüldüğünde önemli bir fırsat olarak da değerlendirilebilir. Çünkü elektronik belgelerin yer aldığı bu yığınlar aslında içsel olarak bir düzene sahiptir ve veri madenciliği ya da makine öğrenmesi teknikleriyle bu yığınları hem kolaylıkla yönetebilmemiz hem de onlarda mevcut olan düzeni keşfederek buradan ilginç ve değerli bilgiler çıkarabilmemiz mümkündür.

Bu tez çalışmasında, tüm kurum ve kuruluşların iş ve işlemlerini yürütebilmesine olanak sağlayan elektronik belge yönetim sistemlerinin veri madenciliği teknikleri ve makine öğrenmesi algoritmaları aracılığıyla nasıl tasarlanabileceğine odaklanılmıştır. Bu bağlamda EBYS'lerde elektronik belge yönetiminin yapısal işleyişini bozan kullanıcı kaynaklı hataları ve standart dosya planında yanlış yere yerleştirilmiş aykırı belgeleri; sınıflandırma, kümeleme ve aykırı değerlerin analizi yöntemleriyle tespit edip oluşturduğumuz standart dosya planı kodu öneri sistemiyle bu hatalı işlemlerin önüne geçmeyi hedefleyen bir model ortaya koyulmuştur.

Çalışma kapsamında bir devlet üniversitesinin elektronik belge yönetim sisteminde 2015-2019 yılları arasındaki veriler kullanılmıştır. Üzerinde doğrudan bir model oluşturmaya elverişsiz halde bulunan bu veri seti, uzun bir veri hazırlama ve temizleme sürecinden sonra uygun hale getirilmiştir. Bu süreçte ayrıca ilgili elektronik belge yönetim sisteminin sorumluları ve kullanıcılarıyla görüşmeler gerçekleştirilmiş ve üzerinde çalışılan veri setinin yapısı daha iyi anlaşılmaya çalışılmıştır.

Bununla birlikte, veri hazırlama ve temizleme sürecinde; dağıtımlı belgeler söz konusu olduğunda belgelerin gereksiz biçimde çoğaltılarak veri tabanına kaydedildiği, standart dosya planı kodu atanmamış belgelerin var olduğu, belge metinlerinin içerisinde bir fazlalık olarak görülebilecek html etiketlerinin bulunduğu ve esasen elektronik belge yönetimi standartlarına aykırı ve çeşitli düzensizlikler yaratan bir yapının söz konusu olduğu tespit edilmiştir. Bu sorunlar; verilerin seçilmesi, temizlenmesi, ihtiyaç duyulan yeni niteliklerin oluşturulması, birleştirilmesi ve uygun formata dönüştürülmesine yönelik çeşitli işlemler aracılığıyla ortadan kaldırılmıştır.

Çalıştığımız veri setinin bu işlemlerden sonra uygulamaya elverişli hale getirilmesinin ardından sınıflandırma işlemleri kapsamında dört farklı algoritmayla bir model oluşturulmuştur. Söz konusu algoritmalar metin sınıflandırma işlemlerinde hem akademik literatürde hem de ilgili sektörlerde sıklıkla müracaat edilen ve başarı oranları hakkında bir uzlaşının bulunduğu Naive Bayes, SVM, lojistik regresyon ve Random Forest algoritmalarıdır. Bu algoritmaların uygulanması sonucunda arzu edilen sonuçlar yüksek F1 skorlarıyla elde edilmiştir.

Bunlar arasında en iyi performansı sağlayan algoritma, metin sınıflandırma algoritmalarının doğruluk ölçütü olarak görülebilecek F1 skoru 0,97 çıkan SVM, yani destek vektör makineleridir. Diğer üç algoritmanın F1 skorları da sırasıyla, Logistic Regression 0,96, Random Forest 0,96 ve Naive Bayes 0,94 olarak görülmüştür. Aktardığımız sonuçlara bakılacak olursa kullanılan tüm algoritmaların isabetli çıkarımlarda bulunduğu oldukça açıktır.

Elektronik belge yönetim sistemlerinde veri madenciliği ve makine öğrenmesi algoritmaları kullanılarak kullanıcı temelli hataların ve bunların meydana getirdiği sorunların tespit edilmesi ve ilgili teknikler kullanılarak sistemin gözden geçirilmesi sonucunda yukarıda sıraladığımız bulgular da göz önünde bulundurulacak olursa sistemlerin verimli ve başarılı bir şekilde çalışabilmesi için veri madenciliği ve makine

öğrenmesinin tekniklerinin oldukça gerekli olduğuna yönelik hipotezimiz doğrulanmış olmaktadır.

Bu sonuçların ardından elektronik belge yönetim sistemlerinin süreç ve yönetimlerine yönelik veri madenciliği ve makine öğrenmesi teknikleri başta olmak üzere yapay zekâ ve yenilikçi teknolojiler perspektifinden birtakım önerilerde bulunmak mümkündür. İlk olarak söylenmesi gereken söz konusu teknolojilerin elektronik belge yönetim sistemlerinde kullanılmasının bugün artık bir zorunluluk haline gelmiş olmasıdır. Dolayısıyla ilgili sistemlerde hem veri madenciliği ve makine öğrenmesi teknikleri hem de yapay zekâ teknolojisi altında geliştirilen diğer farklı araçların kullanımı temel bir unsur olarak görülmelidir.

Elektronik belge yönetim sistemlerinde yapay zekâ ve yenilikçi teknolojilerin arzu edildiği şekilde çalışabilmesi için doğru veriye ihtiyacımız olması sebebiyle ilgili kurumda belge yöneticileri ile doğru standart dosya planı kodunun sistem tarafından eşleştirilmesi gerekliliği bir öneri olarak dile getirilebilir. Belirli aralıklarla kontrol sağlanarak belgelerin doğru standart dosya planı kodu seçiminin temin edilmesi birçok sorunu ortaya çıkmadan çözecektir.

Bu noktada, veri madenciliği ve makine öğrenmesi tekniklerinin uygulanacağı eğitim veri seti oluşturulurken dikkat edilmesi gereken en önemli husus da tamamıyla standartlara (EBYS standartı/TS 13298, vs.) uygun biçimde oluşturulmuş belgelerin seçilmesidir. Seçilen belgelerdeki her bir hata, modelin yanlış kurulmasına ve yanlış sonuçlar çıkmasına neden olacak, bu durum da EBYS'deki veri madenciliği ve makine öğrenmesi algoritmalarının başarısını düşürecektir.

Açık Veri projesi hayata geçtiğinde öncelikli olarak EBYS'lerle ilgili veri kümesi artmış olacağı için EBYS alanında yapılan ve yapılacak olan YZ, MÖ ve derin öğrenme projelerinde yaşanan kamu verisine erişim problemi ortadan kalkacaktır. Ayrıca algoritmaların daha büyük bir veri kümesiyle eğitilebilmesi mümkün hale geleceğinden bu tarz uygulamaların başarılarının artmasını da sağlayacaktır. Bununla birlikte, Ulusal Veri Sözlüğü çalışmasıyla da bu ekosistem daha da genişleyecek ve daha yapısal verileri içeren veri kümeleri elde edilmiş olacaktır. Bu da yapay zekâ ve yenilikçi teknolojilerin tüm sistemlerde aktif ve verimli bir şekilde kullanılmasını ve bilgi sistemlerinin birbirleriyle iletişimlerini sağlamış olacaktır.

Hâlihazırda kullanılmakta olan bilgi yönetim sistemlerine, bu çalışmada sunduğumuz veri madenciliği ve makine öğrenmesi uygulamaları dışındaki yapay zekâ ve yenilikçi teknolojilerin uygulamaları başlamış ve başarılı bir şekilde devam etmektedir. Örneğin bu modelin ortaya koyduğu ölçüde başarılı sonuçlar elde edilebilmesi için "derin öğrenme" algoritmaları kullanılarak aynı işlemler yapılabilir. Doğal dil işleme alanındaki yeni gelişmeler ve bunlarla beraber ortaya çıkan "önceden eğitilmiş dil modelleri" olarak adlandırılan, örneğin OpenAI şirketinin geliştirdiği GPT4 tarzı dil modelleri kullanılarak yukarıda bahsedilen tüm işlemler oldukça kolay ve başarılı sonuçlar elde edecek şekilde yapılabilir.

Normalde, sıfırdan bir dil modeli geliştirmek hem çok uzun zaman almakta hem de çok fazla bilgisayar kaynağına ihtiyaç duymaktadır. Bahsedilen dil modellerinin ikinci defa eğitilmesi mümkün olduğu için alana özel eğitilmiş modeller oluşturulabilmektedir. Örneğin, önceden eğitilmiş dil modelleriyle EBYS verileri üzerinde ikinci bir defa eğitim yapılarak EBYS içindeki yapıyı ve belgeleri daha iyi anlayabilmemiz sağlanabilir. Yani elimizdeki veriyle de dil modelini eğiterek daha başarılı ve alana özel işlemler yapabilir. Sözgelimi, üniversiteler veya belirli bir kurumdaki EBYS verileriyle bu modelleri özelleştirerek daha verimli bir şekilde çalışmalarını sağlayabiliriz.

Değinilebilecek bir başka husus, literatürde kuantum teknolojisinin belge ve arşiv yönetiminde kullanılmasına dair teorik spekülasyonlardır. Fakat bu henüz tam olarak yaygınlaşmamış ve son halini almamış teknolojinin hangi biçimde alanda kullanılacağına ilişkin çok fazla detay bulunmamaktadır. Ama yakın gelecekte, belki birkaç sene içerisinde yukarıda söz ettiğimiz dil modellerinin EBYS içerisine dahil edilmesi ve kullanılmasına şahit olmamız kuvvetle muhtemeldir.

Geleceğin belge yöneticileri olacak Bilgi ve Belge Yönetimi bölümü öğrencilerinin yukarıda söz ettiğimiz yapay zekâ ve yenilikçi teknolojilerin sunduğu ve her gün bir yenisi eklenen araçlara yönelik ilgisinin artması oldukça önemlidir. Bu bağlamda ilgili bölümlere düşen vazife, hem bu teknolojileri güncel bir biçimde takip etmeleri hem de bu takibin verimlerini öğrencilerine özellikle eğitim müfredatında yapılacak değişiklik ve yeniliklerle aktarmalarıdır.

Sonuç olarak günümüzde teknolojinin belirleyici ve dönüştürücü gücünün her geçen gün arttığı göz önünde bulundurulduğunda her alanda olduğu gibi hem bilgi ve belge yönetimi alanında hem de elektronik belge yönetim sistemlerinde de ortaya koyulan

yeni teknik ve araçları göz ardı etmek olanaklı değildir. Çalışmada sunmayı denediğimiz teorik arka plan ve pratik modelin ortaya koyduğu bulgular da bu tespiti açıkça doğrulamaktadır.

KAYNAKÇA

Abbasoğlu, B. (2020). Ortaokul Öğrencilerinin Akademik Başarılarının Eğitsel Veri Madenciliği Yöntemleri İle Tahmini. *Veri Bilimi Dergisi, 3*(1), 1-10.

Ackoff, R. L. (1989). From data to wisdom. Journal of applied systems analysis, 16(1), 3-9.

Aggarwal, C. C. (2014). Data Classification: Algorithms and Applications: CRC press.

Aggarwal, C. C., & Han, J. (2014). *Frequent pattern mining*: Springer.

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

Agrawal, R., Imieliński, T. ve Swami, A. (1993). *Mining Association Rules Between Sets of Items In Large Databases.* Paper presented at the Proceedings of the 1993 ACM SIGMOD international conference on Management of data.

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

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İstanbul Üniversitesi İşletme Fakültesi Dergisi, 29*(1), 1-22.

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

Alpaydın, E. (2000). Zeki veri madenciliği: ham veriden altın bilgiye ulaşma yöntemleri. *Bilişim 2000 eğitim semineri*.

Alpaydın, E. (2009). Introduction to Machine Learning: MIT press.

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

Arıcı, G., & Kandur, H. (2016). Elektronik Belge Yönetim Sistemleri (EBYS) Yazılımlarının Geliştirilmesinin Kurumsal Karar Destek Sistemleri (KDS) İçin Önemi. *KURUMSAL BELLEKLERİN GELECEĞİ*, 65.

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

Aruğaslan, E. ve Çivril, H. (2021). Türkiye'de Eğitim Alanında Yapılan Veri Madenciliği ve Yapay Zeka Çalışmaları. *Uluslararası Teknolojik Bilimler Dergisi*, *13*(2), 81-89.

AWS. (2019). Amazon Web Services. Erişim Adresi: https://aws.amazon.com

Baker, R. S. ve Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61-75). New York, NY: Springer New York.

Baker, R. S. ve Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1*(1), 3-17. doi:10.5281/zenodo.3554657

BBC. (2018). The History of Machine Learning. Erişim Adresi: https://www.bbc.com/timelines/zypd97h

BBC. (2019). Alan Turing: Creator of Modern Computing. Erişim Adresi: https://www.bbc.com/timelines/z8bgr82

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

Binici, K. (2019). Makine Öğrenmesi Yaklaşımıyla e-Belgelere Standart Dosya Plan Numaralarının Otomatik Olarak Atanması Üzerine Bir Çalışma. Bilgi Yönetimi, 2 (2), 116-126. DOI: 10.33721/by.654464

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 235-249.

Bostrom, H., & Bostrom, R. (2017). A Lifecycle Model for Data Science Projects. Proceedings of the 50th Hawaii International Conference on System Sciences.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brownlee, J. (2019). Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End. Machine Learning Mastery.

Calders, T. ve Pechenizkiy, M. (2012). Introduction To The Special Section On Educational Data Mining. *SIGKDD Explor. Newsl., 13*(2), 3–6. doi:10.1145/2207243.2207245

Can, E. (2017). *Temel Eğitimden Ortaöğretime Geçiş Sınavı Kazanımlarının Veri Madenciliği Yöntemleri ile Değerlendirilmesi.* (Yüksek Lisans), Afyon Kocatepe Üniversitesi.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. & Shearer, C., "CRISP-DM 1.0 Step-by-step data mining guide", 2000.

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering, 8*(6), 866-883.

Chisholm, A. (2013). *Exploring Data with RapidMiner*: Packt Publishing Ltd.

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

Cichosz, P. (2014). *Data Mining Algorithms: Explained Using R*: John Wiley & Sons.

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

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

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

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

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

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

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

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

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

Frawley, W. J., Piatetsky-Shapiro, G. ve Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine, 13*(3), 57. doi:10.1609/aimag.v13i3.1011

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

Google Cloud. (2017). A History of Machine Learning. Erişim Adresi: https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/

Google Trends. (2023). Arama Trendleri. Erişim Adresi: https://trends.google.co.in/trends/explore?date=2019-01-01%202023-01-09&q=Data%20Science,Machine%20Learning,Data%20Visualization,Artificial%20Intelligence,Deep%20Learning

Gökçen, H. (2011). Yönetim Bilgi / Bilişim Sistemleri: Analiz ve Tasarım (2. Baskı ed.). Ankara: Afşar Matbaacılık.

Harpring, P. (2010). *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*: Getty Research Institute.

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

Han, J., Kamber, M. ve Pei, J. (2011). *Data Mining: Concepts and Techniques-3rd ed.*: Elsevier.

Han, J., Pei, J. ve Tong, H. (2022). Data Mining: Concepts and Techniques (4 ed.): Morgan Kaufmann.

Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques: Elsevier.

Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery, 8*(1), 53-87.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*: MIT press.

Harpring, P. (2010). What Are Controlled Vocabularies. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works.* Los Angeles, CA: Getty Publications.

Harrington, P. (2012). Machine Learning In Action: Manning Publications Co.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data mining use cases and business analytics applications*: CRC Press.

ISO 16175-1:2010, *Information and documentation -- Principles and functional requirements for records in electronic office environments -- Part 1: Overview and statement of principles*

ISO 16175-2:2011, *Information and documentation -- Principles and functional requirements for records in electronic office environments -- Part 2: Guidelines and functional requirements for digital records management systems*

ISO 16175-3:2010, *Information and documentation -- Principles and functional requirements for records in electronic office environments -- Part 3: Guidelines and functional requirements for records in business systems*

ISO/TSE 15489-1:2001, *Information and documentation -- Records management -- Part 1: General*

ISO/TSE 15489-2:2001, *Information and documentation -- Records management -- Part 2: Guidelines*

Kandur, H. (2006). Elektronik Belge Yönetimi: Sistem Kriterleri Referans Modeli (v. 2.0). *Gözd. Geç, 2*, 53.

Kandur, H. (2011). Türkiye'de kamu kurumlarında elektronik belge yönetimi: Mevcut durum analizi ve farkındalığın artırılması çalışmaları. *Bilgi Dünyası, 12*(1), 2-12.

Kantardzic, M. (2019). *Data Mining: Concepts, Models, Methods, and Algorithms-2nd ed.*: John Wiley & Sons.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104-116.

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

Kelleher, J. D. ve Tierney, B. (2018). *Data Science*: MIT Press.

Kılıçer, S. ve Şamlı, R. (2021). Veri Madenciliği İle Türkiye'deki Ve Avrupa Birliği Ülkelerindeki Bilgisayar Mühendisliği Programlarının Karşılaştırılması. *European Journal of Science and Technology*. doi:10.31590/ejosat.873157

Kılınç, D., & Başeğmez, N. (2018). Uygulamalarla Veri Bilimi: Abaküs Yayıncılık.

Kızrak, A. (2019). Yapay Zeka Kullanım Alanları ve Uygulamalarına Derinlemesine Bir Bakış. Erişim Adresi: https://medium.com/@ayyucekizrak/yapay-zeka-kullanım-alanları-ve-uygulamalarına-derinlemesine-bir-bakış-d0fecaf7f61b

Kotu, V. ve Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*: Morgan Kaufmann.

Kotu, V. ve Deshpande, B. (2018). *Data Science: Concepts and Practice*: Elsevier Science.

Kurt, L. (2023). *Bilgi Yönetiminde Veri ve Metin Madenciliği: Bir Dijital İçerik Analizi Uygulaması*. Yayımlanmamış Doktora Tezi, Ankara Üniversitesi, Ankara. Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

Larose, D. T. (2006). *Data Mining Methods and Models*: Wiley.

Larose, D. T. ve Larose, C. D. (2015). *Data Mining and Predictive Analytics* (2 ed.): John Wiley & Sons

Larose, D. T., & Larose, C. D. (2015). *Data mining and predictive analytics*: John Wiley & Sons.

Laudon, K. C., & Laudon, J. P. (2011). Yönetim Bilişim Sistemleri. *Çeviri Editörü: Prof. Dr. Uğur Yozgat) Ankara: Nobel Akademik Yayıncılık*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76-80.

Maimon, O. ve Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*: Springer US.

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

McDonald, J., & Léveillé, V. (2014). Whither The Retention Schedule In The Era Of Big Data and Open Data? Records Management Journal, 24(2), 99-121.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning.

Moore, J. H., & Williams, S. M. (2009). Epistasis and its implications for personal genetics. The American Journal of Human Genetics, 85(3), 309-320.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Transactions on Neural Networks, 11(3), 690-696.

Nisbet, R., Elder, J. F. ve Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*: Academic Press/Elsevier.

Nisbet, R., Miner, G., & Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*: Academic Press.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Özarslan, S. (2014). *Öğrenci Performansının Veri Madenciliği İle Belirlenmesi.* (Yüksek Lisans), Kırıkkale Üniversitesi

Özdemirci, F., Torunlar, M. ve Saraç, S. Üniversiteler için belge yönetimi ve arşiv sistemi işlemleri (BEYAS) el kitabı. Ankara. 2009.

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

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

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining.* Paper presented at the Library of congress.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

Ray, S. (2017). Commonly Used Machine Learning Algorithms (with Python and R Codes). Erişim Adresi:

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., & Stuart, K. (2018). More human than human? Artificial intelligence in the archive. Archives and Manuscripts, 47(2), 179-203.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. Journal of information science, 33(2), 163-180.

Russell, S., & Norvig, P. (2009). Artificial Intelligence: A Modern Approach: Prentice Hall Press.

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

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

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms: Cambridge University Press.

Sutton, R. S., & Barto, A. G. (2011). Reinforcement Learning: An introduction.

Şimşek, H. K. (2018). Makine Öğrenmesi Dersleri 3b: Karar Ağaçları (Regresyon). Erişim Adresi: https://medium.com/data-science-tr/makine-öğrenmesi-dersleri-3a-karar-ağaçları-regresyon-9cdfa9218a68

Şimşek, M.U., Kayalı, N. ve Cavar Akpınar, S. (2019). *Kamu Bilgi Yönetim Süreçlerinde Teknolojik Gelişmelerin ve Yapay Zekâ Yaklaşımlarının Kullanılması ve Değerlendirilmesi*. B. Yalçınkaya ve diğerleri (Yay. haz.). Bilgi Yönetimi ve Bilgi Güvenliği : eBelge-eArşiv-eDevlet-Bulut Bilişim-Büyük Veri-Yapay Zeka, e-BEYAS 2019 Sempozyumu, 10-11 Ekim 2019, Ankara, Türkiye, Bildiriler içinde (s.397-408), Ankara: Ankara Üniversitesi, BEYAS Koordinatörlüğü.

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

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Pearson Addison Wesley.

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

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149-172.

TS 13298. (2015). *Elektronik Belge ve Arşiv Yönetim Sistemi.* Ankara: Türk Standartları Enstitüsü.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London mathematical society, 2(1), 230-265.

Ünal, M. A., & Özdemirci, F. (2017). EBYS (e-Beyas) ve e-arşiv sistemlerinde/uygulamalarında yapay zekâ yaklaşımı. Bilgi Sistemleri ve Bilişim Yönetimi: Beklentiler.

Wiki. (2014). Deep Mind. Erişim Adresi: https://en.wikipedia.org/wiki/DeepMind

Wiki. (2018). Data Science. Erişim Adresi: https://en.wikipedia.org/wiki/Data_science

Wiki. k-Nearest Neighbors Algorithm. Erişim Adresi: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Witten, I. H., Frank, E. ve Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.). Amsterdam: Morgan Kaufmann.

Xu, R., & Wunsch, D. C. (2005). Survey of Clustering Algorithms.

Yalçınkaya, B. (2014). *E-devlet Üstveri Standardının Oluşturulması ve Türkiye için Modellenmesi*, Yayınlanmamış Doktora Tezi, Marmara Üniversitesi, İstanbul, Erişim adresi: https://tez.yok.gov.tr/ulusaltezmerkezi/

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

Zhang, C., & Zhang, S. (2002). Association Rule Mining: Models and Algorithms: Springer-Verlag.

Zhao, Y. (2012). *R and data mining: Examples and case studies*: Academic Press.

Zhao, Y., & Cen, Y. (2013). *Data mining applications with R*: Academic Press.