Makine Öğrenmesi Modellerini Kullanarak Akademik Veri Tabanlarına İlişkin Tahminler Oluşturma*

Generating Predictions for Academic Databases using Machine Learning Models

Ertuğrul Burak Eroğlu
Çankırı Karatekin Üniversitesi Bilgi ve Belge Yönetimi Bölümü Arş. Gör. Dr., ertugrulburakeroglu@gmail.com

Kasım Binici
Çankırı Karatekin Üniversitesi Bilgi ve Belge Yönetimi Bölümü Öğretim Üyesi, kbinici@karatekin.edu.tr

Bilgi Yönetimi Dergisi

Cilt: 7 Sayı: 2 Yıl: 2024

https://dergipark.org.tr/tr/pub/by

Hakemli Makaleler

Araştırma Makalesi

Makale Bilgisi

Gönderildiği tarih: 28.05.2024
Kabul tarihi: 24.07.2024
Yayınlanma tarihi: 31.12.2024

Article Info

Date submitted: 28.05.2024
Date accepted: 24.07.2024
Date published: 31.12.2024

Anahtar Sözcükler

Makine Öğrenmesi, Kütüphane Danışma Hizmetleri, Yapay Zekâ

Keywords

Machine Learning, Library Advisory Services, Artificial Intelligence

DOI numarası

10.33721/by.1491489

ORCID

0000-0002-6874-5725 (1)

0000-0002-8071-9693 (2)

Öz

Makine öğrenmesi, birçok uygulamayı güçlendirerek günümüzde hemen her alanda yaygın kullanılan bir teknoloji hâline gelmiştir. Temel anlamda türlü veri kümeleriyle eğitilen bir makinenin amaca uygun algoritmalar kullanarak programlanması neticesinde insan eliyle gerçekleştirilen işlerin bilgisayar sistemlerine devredilmesini hedefleyen makine öğrenmesi uygulamaları, otonom sistemlere güç vermektedir. Kütüphane danışma hizmetlerinin makine öğrenmesi teknikleriyle otonom biçimde yürütülmesinde kullanışlı makine öğrenmesi algoritmalarının belirlenmek istendiği bu çalışmada, akademik veri tabanlarıyla eğitilen bir makine öğrenmesi modelinin, herhangi bir konudaki bilgi kaynağı gereksinimini betimleyen doğal dil sorularına verdiği yanıtların başarım düzeyleri çeşitli makine öğrenmesi algoritmaları çerçevesinde incelenmektedir. Bunun için öncelikle T.C. Millet Kütüphanesinde listelenen akademik veri tabanları referans alınmış, öznitelikleri tanımlanarak bir eğitim veri seti oluşturulmuş ve çeşitli veri madenciliği teknikleri kullanılarak model eğitilmiştir. Ardından modeli sınamak amacıyla gereksinim duyulan test veri setinin ortaya çıkarılması amacıyla 7300 soruluk bir liste oluşturulmuştur. Bir konu hakkındaki bilgi kaynağı gereksinimini betimleyen ve kütüphane danışma birimlerine sorulma potansiyeli bulunan bu yapay sorular, doğal dil işleme ve metin madenciliği teknikleri kullanılarak işlenmiştir. Çalışmadaki veri matrislerine uygun olduğu tespit edilen yedi farklı makine öğrenmesi algoritmasının başarım düzeyleri hem varsayılan hem de optimize edilmiş hiper parametre ayarlarıyla test edilmiş ve en uygun algoritmanın %92,7 oranında doğru tahminde bulunan Destek Vektör Makinesi olduğu tespit edilmiştir. Buna alternatif algoritmalar ise Derin Öğrenme ve Olasılıksal Sinir Ağı olarak belirlenmiştir. Bu algoritmalar sırasıyla %72,3 ve %70,5 oranında doğru tahminde bulunmuştur.

Abstract

Machine learning has become a widely used technology in almost every field today, empowering many applications. Machine learning applications, which aim to transfer the tasks performed by human hands to computer systems as a result of programming a machine trained with various data sets using appropriate algorithms, power autonomous systems. In this study, in which it is aimed to determine useful machine learning algorithms for the autonomous execution of library reference services using machine learning techniques, the performance levels of the answers given by a machine learning model trained with academic databases to natural language questions describing the information source requirement on any subject are examined within the framework of various machine learning algorithms.

**Bu makalenin araştırma ve yayın süreci "Araştırma ve Yayın Etiğine" uygun şekilde yürütülmüştür.*
*** Bu makale Ertuğrul Burak Eroğlu'nun Çankırı Karatekin Üniversitesi'nde yaptığı doktora tezine dayanmaktadır.*

For this, first of all, T.R. Academic databases listed in the National Library were taken as reference, a training data set was created by defining their attributes, and the model was trained using various data mining techniques. Then, a list of 7300 questions was created to reveal the test data set needed to test the model. These artificial questions, which describe the need for information sources on a subject and have the potential to be asked to library information units, were processed using natural language processing and text mining techniques. The performance levels of seven different machine learning algorithms, which were found to be suitable for the data matrices in the study, were tested with both default and optimized hyperparameter settings, and the most suitable algorithm was determined to be Support Vector Machine. Alternative algorithms to this are determined as Deep Learning and Probabilistic Neural Network.

1. Giriş

Teknolojinin hayatımızdaki her alana nüfuz etmesi ve gelişmeye devam etmesi hem kurumların hem de kullanıcıların (müşterilerin) kütüphane hizmetlerinden daha yenilikçi ve faydalı bir şekilde yararlanma isteklerini artırmaktadır. Bu durum, geleneksel yöntemlerle sunulan kütüphane hizmetlerinin daha gelişmiş ve modern yaklaşımlarla sürdürülmesini teşvik etmektedir. Teknolojik gelişmelere ayak uyduran ve hem kurumlara hem de kullanıcılara büyük kolaylıklar sunan bilgisayar destekli araçlar, zamanla çeşitlenerek daha fazla uygulamaya alan açmaktadır. Bu çerçevede, kütüphane hizmetlerini teknolojik açıdan inceleyen birçok teorik ve uygulamalı çalışma yapılmaktadır. Bu çalışmalar, kütüphanelerin teknolojik yazılım ve donanımlardan faydalanarak hizmet kalitesini ve performansını artırması, maliyetleri düşürmesi ve kullanıcı memnuniyetini en üst düzeye çıkarması üzerinde durmaktadır. Son teknolojik gelişmeler ışığında bu çalışmaların, makine öğrenmesi ve onun güç verdiği yapay zekâ gibi yöntemlere dayalı uygulamalarla şekillenmeye başladığı görülmektedir.

Yapay öğrenmenin akıllı sistemler geliştirmeye öncülük etmesiyle birlikte, kurum ve kuruluşlarda yapay zekâ ve makine öğrenmesi gibi yeni nesil teknolojilere dayalı uygulamalar da hızla yaygınlaşmaktadır (Daniel, 2021). Günümüzde, küçükten büyüğe tüm işletmelerde olduğu gibi (Coşkun ve Gülleroğlu, 2021, s. 948), bilgi merkezleri de kalıcılığı ve başarıyı yakalamak aynı zamanda toplumun bilgi gereksinimlerini karşılamak için son teknolojilere dayalı hizmet sistemlerini geliştirmek ve uygulamak zorunda kalmıştır (Asemi, Ko ve Nowkarizi, 2020, ss. 413-414). Bu durum, yapay öğrenmenin kütüphaneler üzerindeki akıllı hizmet uygulamalarına yönelik baskısını artırmaktadır (Tektaş, Akbaş ve Topuz, 2002). Yapay öğrenme ve yapay zekâ gibi yeni nesil araçların bilginin yönetildiği kurumlara giderek daha fazla nüfuz etmesi bu dönüştürücü teknolojinin kütüphane hizmetlerinde köklü değişimleri beraberinde getirmektedir (Fernandez, 2016, s. 7). Kütüphaneler yeni teknolojilere uyum sağlamada gecikmeli olsalar da yapay öğrenmenin otonom sistemlerden bilgi erişim araçlarına kadar birçok alanda hizmet sunumunu kolaylaştırması nedeniyle, kütüphane ekosisteminin önemli bir parçası haline gelen bu yeni teknolojiden bilgi kurumlarının yararlanması gerekli hale gelmiştir (Wheatley ve Hervieux, 2019, s. 347).

Yapay öğrenmenin kütüphane faaliyetlerine ve kullanıcılarına sağladığı yararlar, yeni teknolojilerin gelişimi ve hizmetlere entegrasyonunun önünü açan güçlü bir etken olmuştur. Yapay öğrenme tekniklerinin gelişmesi ve akıllı sistemlerin yaygınlaşması, kütüphanelerde yeni dönüşümü tetiklemektedir. Son yıllarda kütüphaneler, bilgiye erişim ve hizmet sunma şeklini geliştirmek için çeşitli yapay zekâ araçlarına yönelmektedir. Bu araçlar, kütüphanelerin daha verimli, kullanıcı dostu ve erişilebilir hâle gelmesine yardımcı olan uygulamaların hayata geçirilmesine yardımcı olmaktadır. Uzman sistemler, doğal dil işleme, örüntü tanıma, makine öğrenmesi, robotik ve akıllı arayüz teknolojilerindeki gelişimin yansıması olan bu yeni nesil kütüphane araçları; içerik indeksleme, belge eşleştirme, kataloglama ve sınıflandırma, seçim ve sağlama, zeki alıntılar, içerik özetleme, diyaloğa dayalı bilgi erişim, sanal danışma hizmetleri ve sohbet robotları, yeni nesil etki faktörü araçları, veriye dayalı izleme, kullanıcı tanımlama, robot/drone yardımları, yapay zekâ alarmları, yapay zekâya dayalı kullanıcı eğitimleri ve operasyonel verimlilik olarak sıralanabilir (Daniel, 2021; Khanzode ve Sarode, 2020, s. 32-33; Nawaz, Gomes ve Saldeen, 2020).

Yeni nesil bilgi hizmeti araçlarıyla birlikte kütüphane ekosistemindeki personel davranışlarının bilgisayar sistemleri tarafından modellenmesi ve geleneksel sistemlerle entegrasyonu, operasyonel verimliliği artırma hedeflerine ulaşmada önemli bir rol oynamaktadır. Bu sayede, kütüphane personeli tarafından geleneksel yollarla yapılan pek çok iş, yapay zekâ teknolojik alt yapısıyla güçlendirilmiş

bilgisayar sistemlerine devredilerek çeşitli kazançlar elde edilmektedir. Öte yandan kütüphane kullanıcılarının profillerinde ve eğilimlerinde önemli değişiklikler yaşanmaktadır (Cox, 2022, ss. 372-373; Kaya, 2017, s. 83; Yılmaz, 2021, s. 56). Kütüphane kullanıcıları gereksinim duydukları bilimsel bilgilere hızlı ve kolay erişmek istemeleri; yığın bilgi artışı; araştırmacıların büyük verilerle değil, büyük literatürle başa çıkmak zorunda kalmaları yapay öğrenmeye dayalı sistemlere olan ilgiyi artırmaktadır. Bilgi ve kaynak keşfi sürecinde büyük fark yaratan makine öğrenmesinin güç verdiği bilgisayar sistemlerinin kütüphaneler üzerindeki etkisi de giderek artmaktadır (Cox, Pinfield ve Rutter, 2019, s. 420).

Makine öğrenmesi araştırmaları, kütüphane ve bilgi bilimi literatüründe uzun zamandır önemli bir yer tutmaktadır. Birçok araştırmacı, bu çalışmada olduğu gibi, makine öğrenmesi algoritmalarının performansını inceleyen çalışmalar yürüterek, model arayışına girmişlerdir. Bu çalışmaların önemli bir kısmı, sınıflandırma problemlerine ve uygun algoritma seçimine yöneliktir. Örneğin, Golub, Hagelback ve Ardö (2020), İsveç Akademik ve Araştırma Kütüphanelerinin ortak kataloğu olan İsveç Ulusal Birlik Kataloğu LIBRIS'ten elde edilen katalog kayıtları üzerinde Dewey Onlu Sınıflama Sistemi'ne göre makine öğrenmesi yaklaşımıyla bir sınıflandırma çalışması gerçekleştirmiştir ve makine öğrenmesi algoritmalarının başarımlarını ölçmüşlerdir. Benzer şekilde Wagstaff ve Liu (2018), kütüphane koleksiyonlarında ayıklama işleminin makine öğrenmesi ile otomatikleştirilmesini inceleyen bir araştırma gerçekleştirmişler ve algoritma seçimine odaklanmışlardır. Öte yandan, Kütüphane ve Bilgi Bilimi alanındaki denetimli makine öğrenmesi çalışmalarında, Destek Vektör Makinesi (DVM) algoritmasının sıklıkla tercih edildiği görülmektedir. Örneğin Binici (2019), Destek Vektör Makinesi kullanarak elektronik belgelere otomatik dosya plan numarası atamayı başarmıştır. Waqas, Anjum ve Afzal (2023) ise bu algoritmayı araştırma makalelerinden üst veri çıkarmak için kullanmış ve benzer şekilde başarılı sonuçlar almıştır.

Farklı disiplinler çeşitli algoritmaları beslemekte ve kendi projelerinde alana uygun algoritmaları tercih etmektedir. Örneğin psikoloji ve matematiksel optimizasyon Destek Vektör Makinesini; istatistik Bayes'i; felsefe ve mantık gibi alanlar ise ters tümdengelim algoritmalarını beslenmektedir (Domingos, 2017, s. 20). Dolayısıyla bu disiplinlerde sürdürülen makine öğrenmesi uygulamalarında, alandan beslenen makine öğrenmesi algoritmaları diğerlerine göre daha sık tercih edilmektedir. Algoritmalar, kendilerine verilen çeşitli görevleri gerçekleştirmek için diğer algoritmalara nazaran daha yüksek ya da daha düşük performans sergileyebilirler. Örneğin, hastalık teşhisinde genellikle Naive Bayes algoritmasına başvurulurken, kitap/film tavsiyesi gibi amaçlar doğrultusunda K- En Yakın Komşu algoritmasına başvurulmaktadır (Domingos, 2017, s. 53-54).

Makine öğrenmesi projelerinde, makine öğrenmesi algoritmalarının ve neticede sınıflandırma yöntemlerinin başarım düzeylerini etkileyen önemli bir etken de sınıflandırılacak verilerin özellikleridir (Boateng, Otoo ve Abaye, 2020, s. 343). Ayrıca algoritmaların hiper parametre ayarlarında yapılan iyileştirmelerin, sonuçlar üzerinde önemli etkileri bulunmaktadır. Her model için optimize edilmesi gereken farklı hiper parametre türleri vardır ve bu nedenle parametre ayarları makine öğrenmesi modellerine bağlı olarak önemli ölçüde değişmektedir (Talaei Khoei ve Kaabouch, 2023, s. 2).

Farklı disiplinlerden beslenen ve çeşitli problemlere çözüm sunan makine öğrenmesi algoritmalarının varlığı ve başarımlarının değişkenliği, bilgi ve belge yönetimi alanına özgü veri setlerinde bu algoritmaların başarısının değerlendirilmesini ve uygun hiper parametre ayarlarının belirlenmesini gerekli kılmaktadır.

Son yıllarda kütüphaneler için artırılmış verimlilik ve kullanıcılar için kolay bilgi erişimi sağlayarak en popüler yöntemlerden biri hâline gelen makine öğrenmesi teknikleri, bu çalışmanın ortaya çıkmasına zemin hazırlamıştır. Genel anlamda kütüphane danışma hizmetlerinin makine öğrenmesi yöntemiyle otonom biçimde yerine getirilmesi düşüncesinden hareketle bu çalışma, bilimsel bir konuyla ilgili bilgi kaynağı ihtiyacının betimlendiği doğal dil sorularına akademik veri tabanlarına ait özniteliklerle eğitilen bir makine öğrenmesi modelinin ilgili akademik veri tabanını adres göstermesi üzerine tasarlanmıştır.

Bu araştırmada, bilimsel bir konu hakkındaki bilgi gereksinimini tasvir eden doğal dil sorularına akademik veri tabanlarının öznitelikleriyle eğitilmiş bir makine öğrenmesi modelinin yanıt verebilme kabiliyeti üzerinde durulmuştur. Model üzerinde veri matrislerine uygun olduğu anlaşılan yedi makine

öğrenmesi algoritmasının[1] varsayılan ve optimize edilmiş hiper parametre ayarlarıyla başarım değerlerinin belirlenmesi, karşılaştırılması ve en iyi başarım sağlayan algoritmanın tespit edilmesi hedeflenen bu araştırmada amaç aşağıda belirtilmiştir.

- Kütüphanelerde danışma birimlerine yöneltilen soruları temsilen, herhangi bir konu hakkında erişilmek istenen bilimsel bilgi kaynağını tasvir eden doğal dil sorularına, bu fonksiyonu gerçekleştirebilmek üzere programlanmış bir makine öğrenmesi modelinin ilgili veri tabanını/veri tabanlarını adres gösterebilme becerisinin değerlendirilmesi,
- Algoritmalar için en iyi başarım sağlayan çekirdek fonksiyonlarının ve hiper parametre ayarlarının tespit edilmesi,
- Akademik veri tabanlarına yönelik danışma sorularını yanıtlamak için en iyi başarımı sağlayan makine öğrenmesi algoritmasının belirlenmesidir.

Çalışmanın amaçları doğrultusunda araştırmada kütüphane danışma hizmetleri kapsamında akademik veri tabanları için hazırlanmış veri setleri çerçevesinde yanıtı aranan sorular aşağıdaki şekilde oluşturulmuştur.

- Modelde veri matrislerine uygun makine öğrenmesi algoritmaları nelerdir?
- Model üzerinde sınanan algoritmalardan hangileri kabul edilebilir performans göstermektedir?
- Model için en kullanışlı makine öğrenmesi algoritması nedir?
- Makine öğrenmesi algoritmalarının çekirdek hesaplama yöntemlerinde ve hiper parametre ayarlarında yapılan iyileştirmelerin başarıma etkisi nedir?
- Modelin, ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler Rehberi'nde yer alan konular çerçevesinde yapılandırılmış doğal dil sorularına yanıt olarak akademik veri tabanlarına yönlendirme kabiliyeti nedir?

2. Yöntem

Bir konu hakkında erişilmek istenen bilgi kaynağını tasvir eden doğal dil sorularına akademik veri tabanları öznitelikleriyle eğitilen bir makine öğrenmesi modelinin çeşitli makine öğrenmesi algoritmalarına başvurularak test edildiği çalışmada öncelikle veri matrislerine uygun, farklı disiplinlerde üstün performans gösteren yedi algoritma tespit edilmiştir. Bunlar Destek Vektör Makinesi, Olasılıksal Sinir Ağı, K- En Yakın Komşu, Naive Bayes, Bulanık Mantık, Karar Ağacı ve Derin Öğrenme algoritmalarıdır. Çeşitli hiper parametre ayarları gerçekleştirilerek optimize edilmiş bu algoritmaların başarım değerleri, varsayılan hiper parametre ayarlarından elde edilen değerlerle karşılaştırılarak raporlaştırılmıştır.

2.1. Sınırlılıklar

Araştırma kapsamında eğitim veri seti oluşturmak amacıyla, 15 Ocak - 15 Mayıs 2022 tarihleri arasında T.C. Millet Kütüphanesi Veri Tabanları web sayfasında listelenen 133 akademik veri tabanından öznitelikler elde edilmiştir. Bu veri tabanlarının tercih edilme sebebi, araştırma için kapsamlı ve çeşitli bir veri kümesi sunmalarıdır. Bu listede yer alan akademik veri tabanlarının konu, tür, dil, format, bilim alanı, anahtar kelime gibi özellikleri belirlenerek makine öğrenmesi modelinin eğitimi için gereksinim duyulan eğitim veri seti ortaya çıkarılmıştır. Ardından model bir bilgi kaynağı gereksinimini tasvir edici yapıda, yapay olarak hazırlanan, doğal dil sorularından meydana gelen test veri setiyle sınanmıştır. Doğal dil sorularının hazırlanmasında ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler Rehberi'nde yer alan konular temel alınmıştır. Kütüphane danışma birimlerine yöneltilebilecek soruların çeşitliliği ve uygulama alanının genişliği göz önüne alındığında, bu çalışmanın akademik veri tabanları ile sınırlandırılması kararı alınmıştır. Bu doğrultuda örnek olay yöntemine başvurularak araştırmanın gerektirdiği zaman, emek ve maliyet önemli ölçüde azaltılmıştır.

[1] *Araştırma kapsamında başvurulan makine öğrenmesi algoritmaları; Destek Vektör Makinesi, Olasılıksal Sinir Ağı, K-En Yakın Komşu, Naive Bayes, Bulanık Mantık, Karar Ağacı ve Derin Öğrenmedir.*

2.2. Veri Toplama

Her makine öğrenmesi projesinde en temel unsur veridir. Bir makinenin eğitildiği verinin miktarı ve niteliği, yerine getirdiği görevdeki performansını önemli ölçüde etkilemektedir. Veri ne kadar fazla ve kaliteli olursa, makine o kadar iyi öğrenir ve karmaşık görevlerde o denli başarılı olur (Marr, 2023). Araştırma sürecinde, modelin ortaya konmasında kullanılacak verinin yeterli, nitelikli ve amaca yönelik olması gerektiği göz önünde bulundurularak, verinin elde edilmesi ve hazırlanması aşamasına yoğun çaba harcanmıştır.

Çalışmada veri hazırlama, düzenleme, modelleme ve değerlendirme aşamalarında KNIME[2] analitik platformu kullanılarak, çeşitli veri işleme operatörleri ve makine öğrenmesi algoritmaları ile bir doğal dil modelleme çalışması gerçekleştirilmiştir.

Çalışmada, danışma hizmetlerinin görev alanlarından biri olan kütüphane kullanıcılarının bilgi keşfi sürecine yardımcı olma bağlamında akademik veri tabanları üzerinde örnek bir uygulama gerçekleştirilmiştir. Bu çalışmada, kullanıcıların herhangi bir konuda, türde (makale, kitap, bildiri vb.), dilde ya da erişim türünde (açık erişim, abone erişim vb.) doğal dil ile yöneltmesi muhtemel sorulara, eğitilmiş bir makinenin otomatik olarak doğru veri tabanlarını önermesi üzerine bir sistem geliştirilmiştir. Sistem, ortaya çıkan model ve kullanılan algoritmalar değerlendirilerek test edilmiştir. Genel olarak, bir makine öğrenmesi modelinin performansının değerlendirilmesi amacıyla hem eğitilmesi hem de test edilmesi için iki ayrı veri setine gereksinim duyulmaktadır.

2.2.1. Eğitim Veri Seti

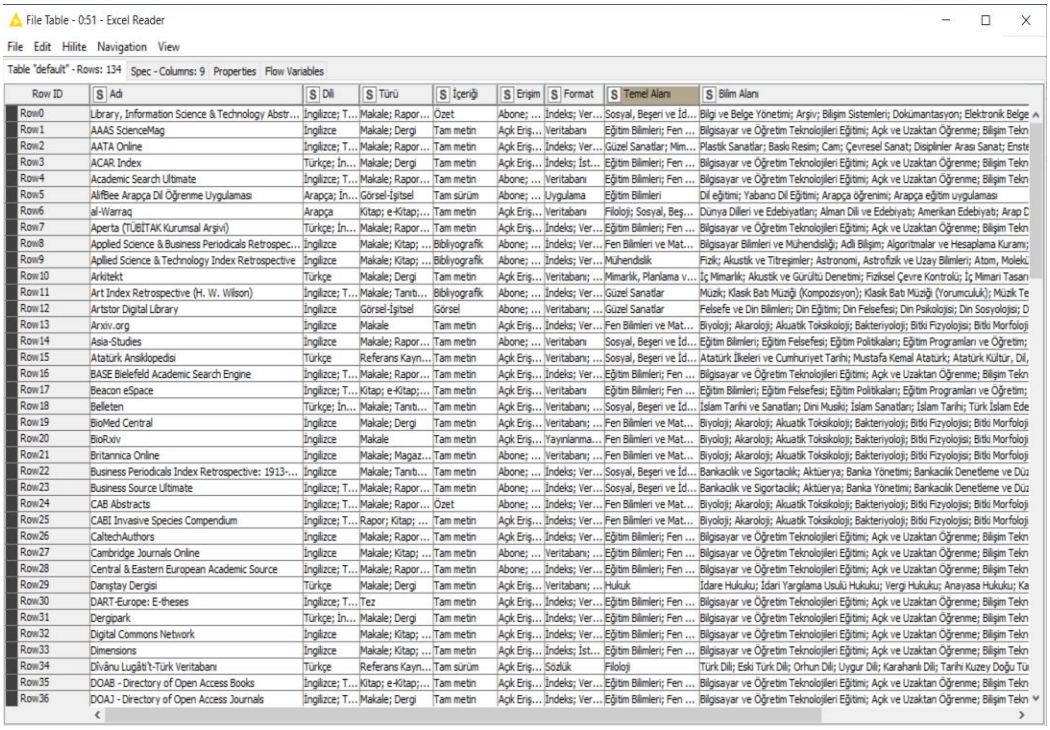
Çalışmanın kapsamı doğrultusunda, makinenin eğitilmesi için kullanılacak veri setinin oluşturulabilmesi amacıyla, T.C. Millet Kütüphanesinde 15 Ocak-15 Mayıs 2022 tarihleri arasında listelenen 133 adet veri tabanı, tür, dil, format, temel alan, bilim alanı, konular, içerik ve erişim türü bakımından kayıt altına alınarak tanımlanmıştır. T.C. Millet Kütüphanesinin veri tabanı listesinin bu çalışmada tercih edilmesinin en önemli sebepleri ülkemizdeki en fazla veri tabanına sahip olması ve araştırma kapsamında veri tabanlarında kısıtsız gezinme imkânı sunmasıdır. Veri tabanlarının öz niteliklerinin belirlenmesinde en önemli unsur, içerdikleri konuların ve anahtar kelimelerin kapsamlı ve standart bir şekilde saptanmasıdır. Bu amaçla, incelenen veri tabanlarının kapsadığı konular "ÜAK 2022 Mart Dönemi Doçentlik Başvurularına Ait Bilim Alanları ve Anahtar Kelimeler" Rehberi'ne dayandırılmıştır.

Veri tabanlarının nitelik özelliklerinin tek biçimli bir şekilde kayıt altına alınabilmesi için MS Access yazılımı kullanılmıştır. Bu amaçla, öncelikle veri tabanlarının niteliklerinin girilebileceği yapılandırılmış bir form oluşturulmuştur. İncelenen veri tabanlarından elde edilen özellikler, bu form aracılığıyla MS Access veri tabanına aktarılmıştır. Kayıtlar, veri tabanının adı, dili, türü, içeriği, erişim türü, formatı, temel alanı, bilim alanı ve konusu gibi nitelik unsurlarını içeren ilişkisel tablolarda tutulmuştur. Veri tabanları üzerindeki çalışmalar tamamlandıktan sonra, elde edilen veriler MS Access'ten MS Excel'e aktarılarak KNIME yazılımı için uygun veri seti elde edilmiştir. Denormalizasyon sağlanarak gerçekleştirilen bu aktarım, MS Access'in dışa aktarma işlevi kullanılarak gerçekleştirilmiştir.

[2] *https://www.knime.com/*

Şekil 1

Eğitim Veri Seti



File Table - 0:51 - Excel Reader

File Edit Hilite Navigation View

Table "default" - Rows: 134 Spec - Columns: 9 Properties Flow Variables

Row ID	Adı	Dili	Türü	İçeriği	Erişim	Format	Temel Alanı	Bilim Alanı
Row0	Library, Information Science & Technology Abstr…	İngilizce; T…	Makale; Rapor…	Özet	Abone; …	İndeks; Ver…	Sosyal, Beşeri ve İd…	Bilgi ve Belge Yönetimi; Arşiv; Bilişim Sistemleri; Dokümantasyon; Elektronik Belge
Row1	AAAS ScienceMag	İngilizce	Makale; Dergi	Tam metin	Açık Eriş…	Veritabanı	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row2	AATA Online	İngilizce; T…	Makale; Rapor…	Tam metin	Açık Eriş…	İndeks; Ver…	Güzel Sanatlar; Mim…	Plastik Sanatlar; Baskı Resim; Cam; Çevresel Sanat; Disiplinler Arası Sanat; Enste
Row3	ACAR İndex	Türkçe; İn…	Makale; Dergi	Tam metin	Açık Eriş…	İndeks; İst…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row4	Academic Search Ultimate	İngilizce; T…	Makale; Rapor…	Tam metin	Abone; …	Veritabanı	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row5	Alifbee Arapça Dil Öğrenme Uygulaması	Arapça; İn…	Görsel-İşitsel	Tam sürüm	Abone; …	Uygulama	Eğitim Bilimleri	Dil eğitimi; Yabancı Dil Eğitimi; Arapça öğrenimi; Arapça eğitim uygulaması
Row6	al-Warraq	Arapça	Kitap; e-Kitap;…	Tam metin	Açık Eriş…	Veritabanı	Filoloji; Sosyal, Beş…	Dünya Dilleri ve Edebiyatları; Alman Dili ve Edebiyatı; Amerikan Edebiyatı; Arap D
Row7	Aperta (TÜBİTAK Kurumsal Arşivi)	Türkçe; İn…	Makale; Rapor…	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row8	Applied Science & Business Periodicals Retrospec…	İngilizce	Makale; Kitap; …	Bibliyografik	Abone; …	İndeks; Ver…	Fen Bilimleri ve Mat…	Bilgisayar Bilimleri ve Mühendisliği; Adli Bilişim; Algoritmalar ve Hesaplama Kuramı;
Row9	Applied Science & Technology İndex Retrospective	İngilizce	Makale; Kitap; …	Bibliyografik	Abone; …	İndeks; Ver…	Mühendislik	Fizik; Akustik ve Titreşimler; Astronomi, Astrofizik ve Uzay Bilimleri; Atom, Molekü
Row10	Arkitekt	Türkçe	Makale; Dergi	Tam metin	Açık Eriş…	Veritabanı; …	Mimarlık, Planlama v…	İç Mimarlık; Akustik ve Gürültü Denetimi; Fiziksel Çevre Kontrolü; İç Mimari Tasarı
Row11	Art İndex Retrospective (H. W. Wilson)	İngilizce; T…	Makale; Tanıtı…	Bibliyografik	Abone; …	İndeks; Ver…	Güzel Sanatlar	Müzik; Klasik Batı Müziği (Kompozisyon); Klasik Batı Müziği (Yorumculuk); Müzik Te
Row12	Artstor Digital Library	İngilizce	Görsel-İşitsel	Görsel	Abone; …	Veritabanı; …	Güzel Sanatlar	Felsefe ve Din Bilimleri; Din Eğitimi; Din Felsefesi; Din Psikolojisi; Din Sosyolojisi; D
Row13	Arxiv.org	İngilizce	Makale	Tam metin	Açık Eriş…	İndeks; Ver…	Fen Bilimleri ve Mat…	Biyoloji; Akaroloji; Akuatik Toksikoloji; Bakteriyoloji; Bitki Fizyolojisi; Bitki Morfoloji
Row14	Asia-Studies	İngilizce	Makale; Rapor…	Tam metin	Abone; …	Veritabanı	Sosyal, Beşeri ve İd…	Eğitim Bilimleri; Eğitim Felsefesi; Eğitim Politikaları; Eğitim Programları ve Öğretim;
Row15	Atatürk Ansiklopedisi	Türkçe	Referans Kayn…	Tam metin	Açık Eriş…	Veritabanı; …	Sosyal, Beşeri ve İd…	Atatürk İlkeleri ve Cumhuriyet Tarihi; Mustafa Kemal Atatürk; Atatürk Kültür, Dil,
Row16	BASE Bielefeld Academic Search Engine	İngilizce; T…	Makale; Rapor…	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row17	Beacon eSpace	İngilizce; T…	Kitap; e-Kitap;…	Tam metin	Açık Eriş…	Veritabanı	Eğitim Bilimleri; Fen …	Eğitim Bilimleri; Eğitim Felsefesi; Eğitim Politikaları; Eğitim Programları ve Öğretim;
Row18	Belleten	Türkçe; İn…	Makale; Tanıtı…	Tam metin	Açık Eriş…	Veritabanı; …	Sosyal, Beşeri ve İd…	İslam Tarihi ve Sanatları; Dini Musiki; İslam Sanatları; İslam Tarihi; Türk İslam Ede
Row19	BioMed Central	İngilizce	Makale; Dergi	Tam metin	Açık Eriş…	Veritabanı; …	Fen Bilimleri ve Mat…	Biyoloji; Akaroloji; Akuatik Toksikoloji; Bakteriyoloji; Bitki Fizyolojisi; Bitki Morfoloji
Row20	BioRxiv	İngilizce	Makale	Tam metin	Açık Eriş…	Yayınlanma…	Fen Bilimleri ve Mat…	Biyoloji; Akaroloji; Akuatik Toksikoloji; Bakteriyoloji; Bitki Fizyolojisi; Bitki Morfoloji
Row21	Britannica Online	İngilizce	Makale; Magaz…	Tam metin	Abone; …	Veritabanı; …	Fen Bilimleri ve Mat…	Biyoloji; Akaroloji; Akuatik Toksikoloji; Bakteriyoloji; Bitki Fizyolojisi; Bitki Morfoloji
Row22	Business Periodicals İndex Retrospective: 1913-…	İngilizce	Makale; Tanıtı…	Tam metin	Abone; …	İndeks; Ver…	Sosyal, Beşeri ve İd…	Bankacılık ve Sigortacılık; Aktüerya; Banka Yönetimi; Bankacılık Denetleme ve Düz
Row23	Business Source Ultimate	İngilizce; T…	Makale; Rapor…	Tam metin	Abone; …	İndeks; Ver…	Sosyal, Beşeri ve İd…	Bankacılık ve Sigortacılık; Aktüerya; Banka Yönetimi; Bankacılık Denetleme ve Düz
Row24	CAB Abstracts	İngilizce; T…	Makale; Rapor…	Özet	Abone; …	İndeks; Ver…	Fen Bilimleri ve Mat…	Biyoloji; Akaroloji; Akuatik Toksikoloji; Bakteriyoloji; Bitki Fizyolojisi; Bitki Morfoloji
Row25	CABI Invasive Species Compendium	İngilizce; T…	Rapor; Kitap; …	Tam metin	Açık Eriş…	İndeks; Ver…	Fen Bilimleri ve Mat…	Biyoloji; Akaroloji; Akuatik Toksikoloji; Bakteriyoloji; Bitki Fizyolojisi; Bitki Morfoloji
Row26	CaltechAuthors	İngilizce	Makale; Rapor…	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row27	Cambridge Journals Online	İngilizce	Makale; Kitap; …	Tam metin	Abone; …	Veritabanı; …	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row28	Central & Eastern European Academic Source	İngilizce; T…	Makale; Rapor…	Tam metin	Abone; …	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row29	Danıştay Dergisi	Türkçe	Makale; Dergi	Tam metin	Açık Eriş…	Veritabanı; …	Hukuk	İdare Hukuku; İdari Yargılama Usulü Hukuku; Vergi Hukuku; Anayasa Hukuku; Ka
Row30	DART-Europe: E-theses	İngilizce; T…	Tez	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row31	Dergipark	Türkçe; İn…	Makale; Dergi	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row32	Digital Commons Network	İngilizce	Makale; Kitap; …	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row33	Dimensions	İngilizce	Makale; Kitap; …	Tam metin	Açık Eriş…	İndeks; İst…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row34	Dîvânu Lugâti't-Türk Veritabanı	Türkçe	Referans Kayn…	Tam sürüm	Açık Eriş…	Sözlük	Filoloji	Türk Dili; Eski Türk Dili; Orhun Dili; Uygur Dili; Karahanlı Dili; Tarihi Kuzey Doğu Tü
Row35	DOAB - Directory of Open Access Books	İngilizce; T…	Kitap; e-Kitap;…	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn
Row36	DOAJ - Directory of Open Access Journals	İngilizce; T…	Makale; Dergi	Tam metin	Açık Eriş…	İndeks; Ver…	Eğitim Bilimleri; Fen …	Bilgisayar ve Öğretim Teknolojileri Eğitimi; Açık ve Uzaktan Öğrenme; Bilişim Tekn

Şekil 1, akademik veri tabanlarının dil, tür, içerik, format, temel alan ve bilim alanlarının eğitim veri setinde nasıl tanımlandığını göstermektedir. Örneğin, ilk satırda yer alan "Library, Information Science & Technology Abstract" veri tabanı, bilgi ve belge yönetimi alanıyla ilgili bir bilgi kaynağıdır. Bu nedenle, veri tabanının konuları ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler Rehberi'ne dayalı olarak "Bilgi ve Belge Yönetimi, Arşiv, Bilişim Sistemleri, Dokümantasyon, Elektronik Belge Yönetim Sistemleri, İnternet ve Kütüphanecilik" terimleriyle tanımlanmıştır. Veri tabanının içerdiği kaynakların dili, türü, içeriği, formatı ve temel alanı da veri tabanı özelliklerinin tanımlanması amacıyla kaydedilmiştir.

2.2.2. Test Veri Seti

Çalışmanın amacı doğrultusunda makinenin sınanabilmesi için ihtiyaç duyulan test veri setinin oluşturulması aşamasında, ilk olarak bu çalışmanın yürütücüsü tarafından doğal dilin esas alındığı, çeşitli ve rastgele kombinasyonlardan oluşan 50 farklı soru kalıbı ortaya çıkarılmıştır. Soru kalıplarının ortaya çıkmasının ardından, "ÜAK 2022 Mart Dönemi Doçentlik Başvurularına ait Bilim Alanları ve Anahtar Kelimeler" isimli Rehber'deki anahtar kelimeler, bilim alanlarına göre hiyerarşik sırayla MS Excel'e aktarılmış ve bir konu dizini elde edilmiştir.

Şekil 2

Konu Dizini

	A	
1	Konu	Alt Konular
1396	Beşeri ve İktisadi coğrafya	Bölge Planlama
1397	Beşeri ve İktisadi coğrafya	Şehir ve Bölge Planlama Eğitimi
1398	Beşeri ve İktisadi coğrafya	Tarihi Coğrafya
1399	Beşeri ve İktisadi coğrafya	Türkiye Beşeri Coğrafyası
1400	Beşeri ve İktisadi coğrafya	Türkiye Ekonomik Coğrafyası
1401	Beşeri ve İktisadi coğrafya	Türkiye Fiziki Coğrafyası
1402	Beşeri ve İktisadi coğrafya	Ulaşım Planlaması
1403	Beşeri ve İktisadi coğrafya	Ülkeler Coğrafyası
1404	Beşeri ve İktisadi coğrafya	Yerleşme Coğrafyası
1405	Bilgi ve Belge Yönetimi	Bilgi ve Belge Yönetimi
1406	Bilgi ve Belge Yönetimi	Belge Yönetimi
1407	Bilgi ve Belge Yönetimi	Bilgi Yönetimi
1408	Bilgi ve Belge Yönetimi	BBY
1409	Bilgi ve Belge Yönetimi	Bilgi bilimi
1410	Bilgi ve Belge Yönetimi	Kütüphanecilik ve Bilgi Bilimi
1411	Bilgi ve Belge Yönetimi	Arşiv
1412	Bilgi ve Belge Yönetimi	Bilişim Sistemleri
1413	Bilgi ve Belge Yönetimi	Dokümantasyon
1414	Bilgi ve Belge Yönetimi	Elektronik Belge Yönetimi
1415	Bilgi ve Belge Yönetimi	Elektronik Belge Yönetim Sistemleri
1416	Bilgi ve Belge Yönetimi	İnternet
1417	Bilgi ve Belge Yönetimi	Kütüphane
1418	Bilgi ve Belge Yönetimi	Kütüphanecilik
1419	Bölgesel Çalışmalar	Bölgesel Çalışmalar
1420	Bölgesel Çalışmalar	Bölge Planlama
1421	Bölgesel Çalışmalar	Bölge-Bölüm-Yöre Analizleri
1422	Bölgesel Çalışmalar	Bölgesel Gelişme ve Küreselleşme
1423	Bölgesel Çalışmalar	Bölgesel İktisat
1424	Bölgesel Çalışmalar	Ekonomik Coğrafya
1425	Bölgesel Çalışmalar	Kültürlerarası İletişim
1426	Bölgesel Çalışmalar	Bölgesel Analizler

MS Excel üzerinde oluşturulan konu dizini, önceden hazırlanmış olan soru kalıplarına uygulanarak, tüm konularla ilgili farklı kombinasyonlarda, doğal dil ile oluşturulmuş 7300 sorudan oluşan bir test seti (Şekil 3) ortaya çıkarılmıştır.

Şekil 3

Test Veri Seti

Table "default" - Rows: 7300 | Spec - Column: 1 | Properties | Flow Variables

Row ID	S Sorular
Row3550	Ben bilgi ve belge yönetimi alanında görevli bir hocayım. Alanımla ilgili makalelere nereden erişebilirim?
Row3551	bilgi ve belge yönetimi alanında dergi makalelerine nereden erişebilirim?
Row3552	Türkçe dilinde bilgi ve belge yönetimi konusunu ilgilendiren dergilere nereden göz atabilirim?
Row3553	Bilimsel çalışmaları bulabileceğim bir platform var mı? bilgi ve belge yönetimi konusunda olursa iyi olur.
Row3554	bilgi ve belge yönetimi ile ilgili yayınlanmış makale istiyorum. Ne yapmalıyım?
Row3555	bilgi ve belge yönetimi konusunda Türkçe tez var mı?
Row3556	Tam metin okuyabileceğim e-kitap ihtiyacım var. Özellikle bilgi ve belge yönetimi konusunda e-kitaplara nereden ulaşabilirim?
Row3557	bilgi ve belge yönetimi disiplinindeki raporlara nereden ulaşabilirim?
Row3558	Abone olduğumuz veri tabanlarından bilgi ve belge yönetimi konusundaki akademik çalışmaları nereden tarayabilirim?
Row3559	bilgi ve belge yönetimi alanında yabancı tezler için nereye bakmalıyım?
Row3560	bilgi ve belge yönetimi konusunda kitap bulmam gerekli. Ne yapmam lazım?
Row3561	Magazin ihtiyacım var. bilgi ve belge yönetimi hakkında magazinler için hangi veri tabanına bakmam lazım?
Row3562	bilgi ve belge yönetimi alanında kitap ya da kitap bölümü arıyorum. Nasıl ulaşabilirim?
Row3563	bilgi ve belge yönetimi alanında dergi makalesi istiyorum. Nerede bulabilirim?
Row3564	bilgi ve belge yönetimi ile ilgili akademik çalışmalara nereden erişebilirim?
Row3565	bilgi ve belge yönetimi konusunda Türkçe dergi arıyorum. Hangi veritabanını önerirsiniz?
Row3566	bilgi ve belge yönetimi disiplininde bilimsel dergi arıyorum. Ne yapmam gerekir?
Row3567	Üniversitenizin akademik personeliyim. bilgi ve belge yönetimi anabilim dalında görev yapıyorum. Alanımla ilgili bilimsel yayın ihtiyacım var. Nereden bulabilirim?
Row3568	Ücretsiz makaleleri nereden bulabilirim? Çalışma konum bilgi ve belge yönetimi .Özellikle bu konuda olursa iyi olur.
Row3569	bilgi ve belge yönetimi ile ilgili ücretsiz dergi makalelerine ulaşmak için hangi veri tabanını kullanmalıyım?
Row3570	bilgi ve belge yönetimi alanında yayınlanmış açık erişimli dergi istiyorum. Hangi veri tabanına bakmam gerekir?
Row3571	bilgi ve belge yönetimi disipliniyle ilgili bilimsel dergilere ihtiyacım var. Ne yapmamı tavsiye edersiniz?
Row3572	bilgi ve belge yönetimi alanında ücretsiz kitaplara nereden erişebilirim?
Row3573	Alanımla ilgili dergilere ihtiyacım var. bilgi ve belge yönetimi konulu olanları nereden bulabilirim?
Row3574	bilgi ve belge yönetimi ile ilgili Türkçe makale istiyorum.
Row3575	bilgi ve belge yönetimi konusundaki Türkçe tezlere nasıl erişebilirim?
Row3576	bilgi ve belge yönetimi disiplinindeki bilimsel çalışmalara nereden ulaşabilirim?
Row3577	bilgi ve belge yönetimi alanında tez lazım. Nereye bakmalıyım?
Row3578	bilgi ve belge yönetimi alanında Türkçe tezlere nereden erişebilirim?
Row3579	Bu kurumda çalışıyorum. bilgi ve belge yönetimi hakkında kitap arıyorum. Nereden bulabilirim.
Row3580	Yüksek lisans öğrencisiyim. bilgi ve belge yönetimi konusunda yabancı tezlere nereden erişebilirim.
Row3581	bilgi ve belge yönetimi ile ilgili ekitap okumak istiyorum. Nereden erişebilirim?
Row3582	Okulumuzda doktora yapıyorum. bilgi ve belge yönetimi konusunda çalışıyorum. Konumla ilgili yabancı doktora tezlerine nereden erişebilirim?
Row3583	bilgi ve belge yönetimi konusunda makale ihtiyacım var.
Row3584	bilgi ve belge yönetimi alanında dergi makalesi arıyorum. Ne yapmalıyım?
Row3585	bilgi ve belge yönetimi konusunda atıfları bulabileceğim bir indeks arıyorum.
Row3586	Makalelerin kaç atıf aldığına nereden bakabiliriz? bilgi ve belge yönetimi konusunda yüksek atıf alan makaleleri bulmak istiyorum.
Row3587	Benim bazı yayınlara ihtiyacım var. bilgi ve belge yönetimi konulu kitaplara nereden erişebilirim?
Row3588	bilgi ve belge yönetimi konusunda Türkçe tez arıyorum. Yardımcı olabilir misiniz?
Row3589	Hoca bilgi ve belge yönetimi konusunda literatür taraması yapmamızı istedi. Bu konudaki makaleleri bulabileceğim bir yer var mı?
Row3590	bilgi ve belge yönetimi ile ilgili bilimsel dergi arıyorum. Nereye bakmalıyım?
Row3591	bilgi ve belge yönetimi konusunda e-kitap ihtiyacım var.
Row3592	Ücretsiz erişebileceğim kitaplara ihtiyacım var. İlgilendiğim konu bilgi ve belge yönetimi . Bu konudaki kaynakları nereden temin edebilirim?
Row3593	Türkçe yazılmış tezleri nereden bulabilirim? Özellikle bilgi ve belge yönetimi konusundakileri arıyorum.
Row3594	bilgi ve belge yönetimi alanında açık erişimli tam metin kitap istiyorum. Ne yapmalıyım?
Row3595	Merhaba. bilgi ve belge yönetimi konusunda Türkçe ücretsiz makale arıyorum. Yardımcı olabilir misiniz?
Row3596	bilgi ve belge yönetimi disiplininde türkçe tam metin dergileri nereden bulabilirim?
Row3597	bilgi ve belge yönetimi alanında tam metin ücretsiz makale arıyorum. Nereyi taramalıyım?
Row3598	Türkçe dilinde bilgi ve belge yönetimi konusunda erişebileceğim açık erişimli tezler var mı? Nereden bulabilirim?
Row3599	bilgi ve belge yönetimi disiplininde dergi makalelerine ihtiyacım var. Açık erişimli olursa iyi olur. Ne tavsiye edersiniz?

Şekil 3, her bilim alanı için farklı kombinasyonlarda 50 soru hazırlandığını ve toplamda 7300 sorunun yer aldığını göstermektedir. "Bilgi ve belge yönetimi" konusunu örnek teşkil etmesi amacıyla, test veri setindeki doğal dil sorularına yer verilmiştir. Bu sorular, bir araştırmacının herhangi bir konudaki bilgi kaynağı gereksinimini Türkçe doğal dille ifade etmesi üzerine kurgulanmıştır.

Örnek soru: “Bilgi ve belge yönetimi ile ilgili akademik çalışmalara nereden erişebilirim?”

2.3. Veri Hazırlama

Araştırma amaçları doğrultusunda elde edilen modelden en iyi performansı alabilmek için veri hazırlama aşaması oldukça önemlidir. Bu çalışmada, veri setleri içerisindeki eksik, hatalı, gürültülü ve kirli verilerin ayıklanması için yoğun çaba sarf edilmiştir. Bu işlem aşağıdaki başlıklar altında detaylı olarak anlatıldığı üzere model performansı izlenerek ve sık sık veri düzenleme aşamasına geri dönülerek gerçekleştirilmiştir. Bu sayede model üzerinde en verimli sonuçlar elde edilmiştir.

2.3.1. Eğitim Veri Seti Hazırlama

Makinenin eğitilmesi amacıyla metinsel formatta derlenmiş olan eğitim veri setinin makinece anlaşılabilmesi için öncelikle ön işlemlerden geçirilmesi gerekmektedir. Akademik veri tabanlarından derlenmiş olan niteleme unsurlarının eğitim setinde kullanılabilmesi için öncelikle tür, dil, içerik, temel alan, bilim alanı, konu, erişim türü ve formata ilişkin özelliklerin karşılık geldiği veri tabanını niteleyen

binominal bir matrisin oluşturulması gerekmektedir. Bu matrisin elde edilmesinde izlenen süreç şu şekilde açıklanabilir:

1. Kütüphane danışma hizmetlerinde veri tabanı keşfi sürecinde doğal dille yöneltilen soru kalıpları göz önünde bulundurularak, yaygın kullanılan özyapı çerçevesinde, akademik veri tabanlarına ait öznitelikler listesi oluşturulmuştur.
2. Veri tabanlarının tanımlandığı eğitim veri setinde veri tabanına ait ad, dil, tür, konu, temel alan, bilim alanı, içerik, format ve erişim türü sütunları yer almaktadır. Ancak araştırmanın gerektirdiği yoğun çalışma zamanı ve makine performansı gözetilmiş; makinenin eğitimi için akademik veri tabanının adı, dili, türü, bilim alanı ve anahtar kelimelerine ait özniteliklerin temel amaçlara ulaşabilmek için yeterli olduğuna karar verilmiştir.
3. Veri tabanları tasarlanırken, ilişkili tablolara kaydedilen ad, tür ve bilim alanı gibi özellikler bütünleştirilmiştir. Normalizasyon sağlanarak oluşturulan bu veri seti ile veri kaybı ve tekrarı önlenerek veri bütünlüğü sağlanmıştır.
4. Veri setinden istenmeyen ve yanlış bilgileri ayıklamak için veri temizleme işlemi gerçekleştirilmiştir. Bu işlem kapsamında, boş ve bilinmeyen karakterlere sahip hücreler filtrelenerek veri setinden kaldırılmıştır.
5. Metin verilerinin makine tarafından daha kolay işlenebilmesi için bazı ön işlemleme adımları gerçekleştirilmiştir. Bu işlemler büyük-küçük harf dönüştürme ve noktalama işaretlerinin kaldırılmasıdır.
6. Elde edilen veri seti içerisinde yer alan eksik/kayıp, gürültülü ve kirli veriler kontrol edilerek gerekli düzenlemeler yapılmıştır. Bu sayede, veri setinin kalitesi ve güvenilirliği artırılmıştır.

Gerçekleştirilen bu işlemlerin ardından eğitim veri seti için ortaya çıkarılan binominal matris Şekil 4'te gösterilmektedir.

Şekil 4

Eğitim Veri Seti Matrisi

Table "default" - Rows: 134 | Spec - Columns: 205 | Properties | Flow Variables

Row ID	S Adi	I atmosfer-bil...	I bahce-bitki...	I bilgisayar-bilimleri-muhendisligi
Row49	Humanities & So...	0	0	0
Row50	Huthi Trust: Digi...	0	0	0
Row51	IEEE	0	0	1
Row52	IRCICA Farabi S...	0	0	0
Row53	Internet Archive	0	0	0
Row54	Islamic Heritage...	0	0	0
Row55	JSTOR	0	0	0
Row56	JSTOR Open Co...	0	0	0
Row57	Karakaş psikoloji...	0	0	0
Row58	Kelime.com	0	0	0
Row59	Konuşan Kitaplık	0	0	0
Row60	KoreaScience	1	1	1

Şekil 4, 134 satır ve 205 sütundan oluşan eğitim veri seti matrisinin bir kesitini sunmaktadır. Bu matriste, her bir satır bir bilim alanını, her bir sütun ise bir akademik veri tabanını temsil etmektedir. Matristeki değerler, ilgili bilim alanının o veri tabanında yer alıp almadığını gösterir: "1" (var), "0" (yok). Örneğin, "Bilgisayar Bilimleri ve Mühendisliği" bilim alanı için IEEE ve KoreaScience Veri Tabanlarında "1" değeri görülmektedir. Bu, her iki veri tabanında Bilgisayar Bilimleri ve Mühendisliği ile ilgili yayınlar bulunduğu anlamına gelmektedir.

2.3.2. Test Veri Seti Hazırlama

Çalışmada amaçlanan değerlendirmeleri gerçekleştirmek için, eğitilmiş olan makinenin test edilmesi gerekmektedir. Bu amaçla, "Test Veri Setinin Elde Edilmesi" başlığı altında test veri setinin nasıl elde edildiği ayrıntılı olarak açıklanmıştır. Test veri seti elde edildikten sonra, insanlar tarafından konuşulan

dilin bilgisayarlar tarafından anlaşılabilir hâle getirilebilmesi amacıyla, veri seti içerisinde yer alan metinler (sorular) üzerinde ön işleme ve doğal dil işleme adımları gerçekleştirilmiştir. Bu adımlar şu şekilde açıklanabilir:

1. Alışılmış makine öğrenmesi projelerinde başvurulan veri setlerinden farklı olarak modelin eğitimi için, akademik veri tabanlarından elde edilen yapılandırılmış öznitelikler kullanılırken, test aşamasında ise doğal dil ile oluşturulmuş salt metinlerden oluşan bir veri kümesi üzerinde çalışılmıştır. Doğal dil işleme ve metin madenciliği teknikleri ile bu sorulardan (salt metinlerden) öznitelikler elde edilmiştir. Bir konu hakkında bilgi kaynağı ihtiyacını betimleyen sorulardan oluşturulan test verisine ait öznitelikler sınıflandırılmış ve modelin tahmin başarımının ölçümünde referans olarak kullanılmıştır. Sınıflandırma işlemi sadece değerlendirme ve sonuçların başarısını takip etmek için yapılmış, modelin eğitiminde ise sadece eğitim veri seti kullanılmıştır.
2. Test veri kümesi, bilgisayar işlem yükünü ve test süresini optimize etmek için veri havuzundaki soruların %10'u (730 soru) rastgele seçilerek oluşturulmuştur. Bunun nedeni yoğun testlerin bilgisayar performansını olumsuz etkileyip süreyi uzatmasıdır. Fakat algoritmaların en yüksek başarı seviyelerini belirlemek için tekrar tekrar test edilmeleri şarttır. Bu nedenle, 730 sorudan oluşan ve farklı soru kalıplarını ve konuları kapsayan bir test veri seti, modelin genel performansını farklı soru türlerine karşı değerlendirmek için yeterli görülmüştür.
3. Test veri seti, doğal dil ile oluşturulmuş metin formatında sorulardan oluşmaktadır. Bu soruları değerlendirmek için öncelikle metinlerin istenen özelliklere göre parçalara ayrılması ve bu parçalardaki ögelerin (kelimelerin) belirlenmesi gerekmektedir. Bu amaçla, metinlere "işaretleme" adı verilen bir işlem uygulanmıştır.
4. Metin içerisindeki özellikler ayrıştırıldıktan sonra, noktalama işaretleri, semboller ve özel karakterler temizlenmiş, bağlaçlar ve önemsiz kelimeler çıkarılarak sadeleştirilmiştir.
5. Anlamlı ögelerde büyük-küçük harf duyarlılığının model performansını olumsuz etkileyebileceği göz önünde bulundurularak, eğitim veri setinde olduğu gibi tüm ifadeler küçük harflere dönüştürülmüştür. Bu ön işlem, modelin farklı yazım şekillerini daha iyi eşleştirebilmesine ve daha tutarlı sonuçlar üretmesine yardımcı olmuştur.
6. Metin ön işleme işleminin son aşamasında, anlamlı ögeler üzerinde yazım hataları, boşluklar ve kısaltmalar düzeltilerek metin temizleme işlemleri gerçekleştirilmiştir.
7. Test veri setinden elde edilecek özniteliklerin eğitim veri setiyle uyumlu olması, makine öğrenmesi projesinin başarısı için oldukça önemlidir. Bu uyumun sağlanması için, her iki veri setinde de benzer karakterlere sahip özniteliklerin tutulması gerekmektedir. Bu amaçla, tüm süreç benzer karakterli özniteliklerin oluşturulmasına odaklanmıştır. Ayrı ayrı oluşturulan eğitim ve test veri setlerinin uyumlu hâle getirilmesi için "ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler" Rehberi'ndeki bilim alanları esas alınarak bir sözlük hazırlanmıştır. Kütüphanecilik alanında yaygın kullanılan "denetimli kavramlar dizini"nden esinlenerek oluşturulan bu sözlük sayesinde test veri setinden çıkarılması istenen özniteliklerin eğitim veri setiyle uyumlu olması sağlanmıştır.
8. Sözlük kullanarak dar terimler hiyerarşik olarak geniş terimlere bağlanarak etiketlemeler gerçekleştirilmiş ve bu sayede matris boyutu küçültülmüştür. Elde edilen öznitelikler binominal matris formatına dönüştürülerek 723 doğal dil sorusunu temsil eden bir test veri seti oluşturulmuştur.

Şekil 5

Test Veri Seti Matrisi

Table "default" - Rows: 723 Spec - Columns: 145 Properties Flow Variables

Row ID	Sorular	bahce-bit...	biyoloji	bolgesel...	enerji-sistemleri...	psikoloji...	iletisim-calismalari
Row550	sosyal psikoloji ile ilgili yayınlanmış makale istiyorum. Ne yapmalıyım?	0	0	0	0	1	0
Row549	sosyal psikoloji alanında dergi makalelerine nereden erişebilirim?	0	0	0	0	1	0
Row548	sosyal politika ile ilgili akademik çalışmalara nereden erişebilirim?	0	0	0	0	0	0
Row547	sosyal politika alanında kitap ya da kitap bölümü arıyorum. Nasıl ulaşabilirim?	0	0	0	0	0	0
Row546	sosyal hizmet konusunda Türkçe tez arıyorum. Yardımcı olabilir misiniz?	0	0	0	0	0	0
Row545	sosyal hizmet disiplininde türkçe tam metin dergileri nereden bulabilirim?	0	0	0	0	0	0
Row544	sosyal hizmet alanında açık erişimli tam metin kitap istiyorum. Ne yapmalıyım?	0	0	0	0	0	0
Row543	sosyal bilimler eğitimi konusundaki Türkçe tezlere nasıl erişebilirim?	0	0	0	0	0	0
Row542	siyasi tarih disiplinindeki bilimsel çalışmalara nereden ulaşabilirim?	0	0	0	0	0	0
Row541	siyasi tarih alanında tam metin ücretsiz makale arıyorum. Nereyi taramalıyım?	0	0	0	0	0	0

Şekil 5, 723 satır ve 145 sütundan oluşan test veri setine ilişkin ortaya çıkarılan matrisin bir kesitini göstermektedir. Matriste her satır, bir doğal dil sorusunu temsil ederken, her sütun ise bir terimi temsil etmektedir. Etiketlenen terimler "1" (var) olarak işaretlenmiş, etiketlenmeyen terimler ise "0" (yok) olarak gösterilmiştir. Örneğin, "Sosyal psikoloji alanında dergi makalelerine nereden erişebilirim?" doğal dil sorusuna ait satıra bakıldığında, "sosyal psikoloji" konusunun etiketlenmiş ve "1" olarak işaretlenmiş olduğu görülmektedir. Bu durum, doğal dil sorusunda "sosyal psikoloji" konusunun ele alındığını göstermektedir.

2.4 Modelleme

Veri kümelerinin modellenmesinde en kullanışlı araçlar makine öğrenmesi algoritmalarıdır. Modelleme süreci öğrenme hedefinin belirlenmesi ile başlar. Ardından uygun algoritmanın seçilmesi ve model performansının test edilmesi gerekir. Test işlemi projenin amaçları doğrultusunda belirlenen test verileri ile gerçekleştirilir. Modelin performansı yetersizse, modelde düzenlemeler, verilerde değişiklikler veya algoritma değişikliği gibi çözümler değerlendirilebilir (Gökalp, 2022, s.2; Şeyranlıoğlu, 2022, s. 58-59).

Geliştirilen araştırma modeli, denetimli makine öğrenmesi türünde ve tahmin edici bir yaklaşım kullanılarak oluşturulmuştur. Modelin performansı, yukarıda sıralanan parametreler çerçevesinde değerlendirilmiştir. Değerlendirme sonucunda modelin bazı aksayan yönleri tespit edilmiştir. Bu aksayan yönler, veri düzenleme aşamasına geri dönülerek gerekli düzenlemeler ve geliştirmeler yapılarak; algoritmaların hiper parametre ve çekirdek hesaplama yöntemlerinde iyileştirmeler yapılarak çözüme kavuşturulmuştur.

2.4.1. Değerlendirme ve Skor

Bu çalışma kapsamında oluşturulan makine öğrenmesi modeli değerlendirilmiş, modelin çalışabilirliği ve test edilebilirliği görülünceye kadar defalarca veri düzenleme aşamasına dönülmüş, eğitim ve test veri setleri üzerinde düzenleme işlemleri gerçekleştirilmiştir. Model için uygun veri matrisleri elde edilinceye kadar bu adımlar tekrarlamıştır.

Türkçe metinlerin işlenmesinde işletim sistemi, veri tabanı, yazılım ve algoritmaların dil ve karakter uyuşmazlığı büyük bir sorun oluşturmaktadır. Türkçe karakterlerin hem eğitim hem de test veri setlerinde sorunlara yol açması nedeniyle makine öğrenmesi algoritması tarafından doğru şekilde yorumlanamadığından, modelin performansı olumsuz etkilenmiştir. Proje kapsamında, veri setlerindeki terimlerin makine tarafından doğru şekilde anlaşılması ve etiketlemelerin hatasız bir şekilde yapılabilmesi amacıyla, araştırmacı tarafından kapsamlı bir sözlük hazırlanmıştır. Bu sözlük, projenin temel bileşenlerinden biri olarak veri setlerinden bilim alanları ve anahtar kelimelerin otomatik ve hassas bir şekilde çıkarılmasını sağlamış, projenin başarısına önemli katkılar sunmuştur. Sözlük sayesinde veri setleri içerisindeki terimlerin makine tarafından anlaşılabilmesi ve otomatik olarak işlenebilmesi, etiketleme işlemenin hatasız ve tutarlı bir şekilde yapılarak veri setlerinin kalitesinin ve güvenilirliğinin artırılması sağlanmıştır. Böylelikle model, verilerden istenilen sonuçları üretebilmiştir.

Sözlük kullanımıyla çözülen bir diğer sorun da alt alanlarıyla birlikte tüm konu alanlarından elde edilen büyük bir matrisin boyutunun azaltılmasıdır. Özellikle bilim dallarına ait alt alanların (diğer bir deyişle anahtar kelimelerin) enlemesine büyük bir matris oluşturması önemli performans sorunlarına yol açmaktadır. Veri madenciliği alanında boyut laneti olarak da adlandırılan bu sorun sözlük yardımıyla ana konulara dönüştürülmüş, böylelikle öznitelik sayısında azaltma sağlanarak sorun aşılmıştır.

Şekil 6

Sözlük Kullanımı

Table "default" - Rows: 1664 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Bul	degistir
Row287	arnavut dili ve edebiyatı	dunya-dilleri-edebiyatlari
Row1622	arı ve ipek böceği yetiştiriciliği ve ıslahı	zootekni
Row210	arıtma tesisi tasarımı	cevre-bilimleri-muhendisligi
Row67	arşivcilik	bilgivebelgeyonetimi
Row1127	askeri coğrafya	siyasi-tarih
Row1376	askeri psikiyatri	tip-bilimi
Row1155	askeri psikoloji	psikoloji
Row1377	askeri sağlık hizmetleri	tip-bilimi
Row1182	askeri sosyoloji	sosyoloji
Row1257	askeri tarih	tarih

Etiketleme aşamasında Türkçe karakterlerin ve boşlukların neden olduğu problemleri ortadan kaldırmak ve makine performansını düşüren aşırı büyük bir matris oluşumunu engellemek için kullanılan sözlük sayesinde, örneğin "arşivcilik" gibi Türkçe karakter içeren bir terim, bağlı olduğu Bilgi ve Belge Yönetimi bilim dalına "bilgivebelgeyonetimi" şeklinde dönüştürülmüştür.

Dünya problemlerinin çözümüne yönelik eğitim verileri üzerinde ideal tek bir öğrenme algoritması olmadığından, algoritma seçimi deneysel yöntemlerle yapılmaktadır. Sınıflandırıcılar, eğitim verisine göre değişen modeller oluşturur ve "en iyi" algoritma diye bir şey yoktur. Bu nedenle, eldeki verilere uygun algoritmalar deneysel metotlarla belirlenmelidir (Aydın ve Aslan, 2017). Dolayısıyla makine öğrenmesine dayalı bir projenin başarısı için hedeflere uygun veriler, doğru işlemler ve deneysel algoritma seçimi kritik öneme sahiptir. Bu çalışmada da kullanılacak algoritma/algoritmalar ve ortaya konacak model, deneysel seçimlerle belirlenerek en yüksek başarıya ulaşılması hedeflenmiştir.

Modelin değerlendirilmesinde önemli bir diğer unsur ise proje için en faydalı algoritmaların en yüksek başarı değerlerini elde edebilmeleri için doğru çekirdek fonksiyonlarının ve hiper parametre değerlerinin belirlenmesidir. Bu amaçla, her algoritma için hesaplama yöntemi ve parametre kombinasyonları üzerinde kapsamlı bir çalışma yapılmış ve sınıflandırıcıdan en yüksek performansın elde edilebileceği ayarlar belirlenmiştir.

Makine öğrenmesi algoritmalarının performansını ölçmek için eğitim ve test veri setlerine ek olarak bir doğrulama veri setine ihtiyaç duyulmuştur. Bu set, eğitim veri setindeki akademik veri tabanlarının özelliklerine göre oluşturulmuştur. Test veri setindeki her sorgu, araştırmacılar tarafından akademik veri tabanları çerçevesinde yanıtlanmış ve bu sayede doğrulama veri seti oluşturulmuştur. Böylelikle tüm makine öğrenmesi algoritmalarının sınıflandırma performansı aynı doğrulama veri seti kullanılarak ölçülmüş ve skorlar otomatik olarak hesaplanmıştır.

Tablo 1

Doğal Dil Sorularına Algoritmalar Tarafından Verilen Yanıtlar

Doğal Dil Sorusu (Test)	Destek Vektör Makinesi	Derin Öğrenme	Olasılıksal Sinir Ağı	K-En Yakın Komşu	Naive Bayes	Bulanık Mantık	Karar Ağacı
Abone olduğumuz veri tabanlarından matematik konusundaki akademik çalışmaları nereden tarayabilirim?	Zentralblatt Math Database	Zentralblatt Math Database	Zentralblatt Math Database	DOAJ - Directory of Open Access Journals	DOAB - Directory of Open Access Books	İSAM Risaleler Veri Tabanı	Zentralblatt Math Database

Örneğin test veri setinden gelen "Abone olduğumuz veri tabanlarından matematik konusundaki akademik çalışmaları nereden tarayabilirim?" sorusuna çalışmada kullanılan makine öğrenmesi

algoritmalarının verdikleri yanıtlar Tablo 1'de gösterilmektedir. Buna göre, Destek Vektör Makinesi, Derin Öğrenme, Olasılıksal Sinir Ağı ve Karar Ağacı algoritmaları "Zentralblatt Math Database" Veri Tabanını adres göstermiştir. Doğrulama veri seti içerisinde, test veri setinden gelen bu sorgunun yanıtı "Zentralblatt Math Database" olarak tanımlandığından, bu veri tabanını adres gösteren algoritmaların verdiği yanıtlar makine tarafından doğru olarak kabul edilmiştir. K-En Yakın Komşu, Naive Bayes ve Bulanık Mantık algoritmaları sırasıyla "DOAJ - Directory of Open Access Journals", "DOAB - Directory of Open Access Books" ve "İSAM Risaleler Veri Tabanı" yanıtları vermiştir. Bu yanıtlar doğrulama veri setinde yer alan doğru yanıtla eşleşmediğinden, yanlış olarak kabul edilmiştir.

Gerçekleştirilen makine öğrenmesi projesinde dikkat edilen önemli bir diğer unsur da "yetersiz uyum" ve "aşırı uyum" durumlarıdır. Bu durumlar, öğrenicinin genelleme yeteneğini ve sınıflandırma performansını olumsuz etkileyebilir. Yetersiz uyumda, model eğitim verilerini yeterince öğrenemediği için genelleme yeteneği zayıf olur. Aşırı uyumda ise, model eğitim verisini ezberler ve yeni veriler üzerinde doğru tahminler yapamaz (Demirhan, 2015, s. 32). Çalışmada, yetersiz uyum ve aşırı uyum problemlerinden kaçınmak için çeşitli stratejiler uygulanmıştır:

- Yetersiz Uyumdan Kaçınma:
  - Kapsamlı Özellik Tanımlama: Her akademik veri tabanına ait yeterli öznitelik tanımlanarak, modelin veriyi daha iyi öğrenmesi ve genelleme yeteneğini geliştirmesi sağlanmıştır.
- Aşırı Uyumdan Kaçınma:
  - Test Seti Kullanımı: Model, eğitim veri setinden bağımsız bir test setiyle sınanarak, aşırı uyum probleminin önüne geçilmiştir.
  - Veri Etiketleme: Sınıflandırıcının öğrendiklerini genelleyebilmesi için, akademik veri tabanı içerisinde yer alan bilim alanları (konu kategorileri), tür, dil ve erişim biçimleri etiketlenerek eğitim veri setine dâhil edilmiştir.
  - En Uygun Sınıf Özellikleri: En uygun sınıf özelliklerinin belirlenmesine özen gösterilerek, modelin karmaşıklığı kontrol altına alınmıştır.

2.5.Başarım Ölçümü

Bir konu hakkında kaynağa erişim sürecinde doğal dil sorularına makinece verilen yanıtların örnekleri Şekil 7'de gösterilmektedir. Buna göre, örneğin, “Hoca bilgi ve belge yönetimi konusunda literatür taraması yapmamızı istedi. Bu konudaki makaleleri bulabileceğim bir yer var mı?” sorusuna makine tarafından verilen yanıtın “Library, Information Science & Technology Abstract” olduğu görülmektedir. Çalışmada, test veri seti içerisinde yer alan her soru için, yukarıdaki örnekte olduğu gibi çalışmada kullanılan makine öğrenmesi algoritmalarının çıkarım (tahmin) yapması sağlanmıştır.

Şekil 7

Proje Çıktıları

Table "default" - Rows: 723 | Spec - Columns: 206 | Properties | Flow Variables

Row ID	Sorular	Prediction (Adi)
Row79_?	Bu kurumda çalışıyorum. uçak-havacılık-uzay mühendisliği hakkında kitap arıyorum. Nereden bulabilirim.	ProQuest E-Book Central
Row80_?	Bu kurumda çalışıyorum. veteriner hekimlik hakkında kitap arıyorum. Nereden bulabilirim.	PubMed
Row81_?	Endüstri Mühendisliği konusunda atıfları bulabileceğim bir indeks arıyorum.	Web of Science
Row82_?	Hoca batı sanatı ve çağdaş sanat konusunda literatür taraması yapmamızı istedi. Bu konudaki makaleleri bulabileceğim bir yer var mı?	Humanities & Social Sciences Index Retrospective
Row83_?	Hoca bilgi ve belge yönetimi konusunda literatür taraması yapmamızı istedi. Bu konudaki makaleleri bulabileceğim bir yer var mı?	Library, Information Science & Technology Abstracts

Model üzerinde çalıştırılan makine öğrenmesi algoritmalarından elde edilen çıktıların başarım ölçümleri bu çalışmanın bulgular başlığı altında sunulmuştur. Elde edilen bulgular, her bir algoritma için ayrı başlıklar ve tablolar kullanılarak sunulmuştur. Tablolarda, modelin doğru yaptığı tahminlerin sayısı için "Doğru Tahmin (DT)", modelin yanlış yaptığı tahminlerin sayısı için "Yanlış Tahmin (YT)", doğru yapılan tahminlerin toplam tahminlere oranı için "Doğru Tahmin Yüzdesi (DTY)", yanlış yapılan tahminlerin toplam tahminlere oranı için "Yanlış Tahmin Yüzdesi (YTY)” ve modelin tesadüften daha iyi performans gösterip göstermediğini ölçmek için "Cohen's Kappa (K)" değerleri yer almaktadır. Cohen's Kappa (K) skoru için; < 0: Zayıf; 0.0-0.20 arası hafif; 0.21-040 arası makul; 0.41-0.60 arası

orta; 0.61-0.80 arası önemli ve 0.81-1.00 arası mükemmel uyuşmaya işaret etmektedir (Jin, 2019; Landis ve Koch, 1977, s. 165; Özhan, 2020, s. 56; Widmann, 2020).

Bu araştırma için açıklanması gereken bir diğer unsur da kullanılan algoritmalardaki hiper parametre seçimleridir. Kullanılan algoritmaların çekirdek ve parametre seçimlerinde titiz bir çalışma gerçekleştirilmiş, projenin hedefleri doğrultusunda, veri setine ve modele en uygun çekirdek ve hiper parametreler belirlenmiştir. Görmez'in (2021) de belirttiği üzere, modelin başarısı büyük ölçüde parametre seçimine bağlıdır. Bu nedenle, her bir algoritma için en iyi performansı sağlayacak çekirdeğin tespit edilmesi ve parametrelerin ayarlanması amacıyla, uzun süreli deneysel çalışmalara girilmiştir. Bu çalışmaların sonuçları, elde edilen bulguların yer aldığı alt başlıklarda detaylı bir şekilde sunulmuştur.

3. Bulgular ve Tartışma

3.1. Bulgular

Akademik veri tabanlarının özniteliklerinden oluşan veri seti ile eğitilmiş bir makine öğrenmesi modelinin Türkçe doğal dil sorularına doğru veri tabanını adres gösterebilme yeteneğinin değerlendirildiği çalışmaya ilişkin bulgular sunulmuştur. Modelin performansı, farklı algoritmalar ve hiper parametre ayarları kullanılarak test edilmiştir.

Çalışma kapsamında elde edilen veri matrislerine uygun algoritmalar çalıştırılarak 7300 sorudan oluşan bir havuzdan rastgele seçilen 723 soruya model tarafından yanıt vermesi sağlanmıştır. Algoritmalar tarafından üretilen tahminlerin başarımları hem varsayılan hem de iyileştirilmiş hiper parametrelerle çalıştırılan her bir algoritma için tablolar üzerinde *N: Tahmin Sayısı* ve *%: Tahmin Yüzdesi* olmak üzere gösterilmiş ve yorumlanmıştır. Algoritmaların hiper parametre değerlerinin açıklanmasında Nodepit (2023) platformundan yararlanılmıştır.

3.1.1. Destek Vektör Makinesi (DVM) Algoritmasına İlişkin Bulgular

Destek Vektör Makinesi, bu çalışmada olduğu gibi, sınıflandırma odaklı makine öğrenmesi projelerinde sıklıkla tercih edilen bir algoritmadır. KNIME yazılımı üzerinde gerçekleştirilen çalışmada, Destek Vektör Algoritmasına ait Polinom (Polynomial), Hiper Tanjant (HyperTangent) ve Radyal Tabanlı Fonksiyon (RBF) olmak üzere üç farklı çekirdek hesaplama yöntemi ve bunlara ait hiper parametreler bulunmaktadır. Polinom çekirdeği Güç (Power), Sapma (Bias) ve Gama (Gamma); Radyal Tabanlı Fonksiyon çekirdeği Sigma; Hiper Tanjant çekirdeği ise Kappa ve Delta parametrelerine sahiptir.

KNIME yazılımında Destek Vektör Makinesi (DVM) varsayılan olarak "Polinom" çekirdeği ile çalışmaktadır. Bu çekirdekte 'Örtüşen Ceza' 1.0, 'Güç' 1.0, 'Sapma' 1.0 ve 'Gama' 1.0 değerlerine sahiptir. Çekirdek ve hiper parametreler üzerinde manuel olarak yapılan testler sonucunda, en iyi performans 'Hiper Tanjant' çekirdeği ile elde edilmiştir. Bu çekirdekte 'Kappa' 9, 'Delta' 3 ve 'Örtüşen Ceza' 10 değerleri kullanılmıştır. 'Örtüşen Ceza' değeri, yanlış sınıflandırılan her noktaya ne kadar ceza verileceğini belirlemekte ve modelin performansını önemli ölçüde etkilemektedir.

Destek Vektör Makinesi (DVM) algoritması için varsayılan ve iyileştirilmiş hiper parametre değerlerinin başarımları aşağıdaki tabloda gösterilmektedir:

Tablo 2

Destek Vektör Makinesi Algoritması'nın Başarım Değerleri

Parametreler	Başarım	*N*	*%*	Metrikler	
Varsayılan	DT	547	75,7	K	0,742
	YT	176	24,3		
	Toplam	723	100		
İyileştirilmiş	DT	670	92,7	K	0,922
	YT	53	7,3		
	Toplam	723	100		

Not: DT: Doğru Tahmin YT: Yanlış Tahmin K: Cohen's kappa

Tablo 2'de Destek Vektör Makinesi algoritmasının test veri seti içerisindeki sorulara yanıt verme kabiliyeti gösterilmektedir. Algoritmanın hem varsayılan değerlerinden hem de çekirdek hesaplama yönteminin değiştirilerek hiper parametrelerde yapılan ayarlamalarla elde edilen iyileştirilmiş değerlerinden sağlanan tahmin performansına ait verilerin karşılaştırıldığı tablo, başarım düzeylerinde önemli bir fark olduğunu göstermektedir. Destek Vektör Makinesi algoritmasının performansında gözlemlenen önemli farklılıklar, büyük ölçüde modele en uygun çekirdek hesaplama yönteminin seçilmiş olmasından kaynaklanmaktadır.

Tablonun 'Varsayılan Değerler' satırında; sınıflandırıcının varsayılan çekirdek hesaplama yöntemi olarak kullandığı, genellikle görüntü işleme için tercih edilen ve Destek Vektör Makinesinin diğer çekirdek işlevlerine göre daha az verim alınan (Chaturvedi, 2023), polinom hesaplama yöntemine göre elde edilen değerler gösterilmektedir.

Tablonun 'İyileştirilmiş Değerler' satırında ise, model için en uygun Destek Vektör Makinesi çekirdek fonksiyonu sağlayan, kökeni sinir ağı teorisine dayanan ve özellikle lineer olmayan sınıflandırma problemlerinde başvurulan Hiper Tanjant hesaplamasına ilişkin veriler yer almaktadır (Fadel vd., 2016).

Buna göre varsayılan hesaplama çekirdeğinde çalıştırılan algoritmanın %75,7 (n=547) oranında doğru, %24,3 (n=176) yanlış tahminde bulunduğu görülmüştür. Ayrıca elde edilen sonuçların güvenirliğinin teyidi için ölçülen ve gözlenen ile beklenen değerler arasında önemli bir uyuşma olduğuna işaret eden Cohen's Kappa (K) değerinin K=0,742 olduğu gözlemlenmiştir.

Algoritma üzerinde farklı hesaplama çekirdekleri ve hiper parametreler sınanmış ve yapılan testler sonucunda en verimli işlemi yerine getiren Hiper Tanjant çekirdeğinin %92,7 (n=670) oranında doğru tahminde bulunduğu anlaşılmıştır. Yanlış tahmin oranı ise %7,3 (n=53) olarak belirlenmiştir. Algoritmanın optimize edilmiş hâliyle, iki değerleyici arasında "mükemmel" uyuşmaya işaret eden Cohen's Kappa değerinin K=0,922 olduğu görülmüştür.

Varsayılan ve iyileştirilmiş değerlerle test edilen algoritmanın başarım değerleri arasındaki fark incelendiğinde, doğru tahmin yüzdesinde %17 (n=123) artış yaşandığı tespit edilmiştir. Ayrıca Cohen's Kappa değerinde 0,180 artış olduğu tespit edilmiştir.

3.1.2. Olasılıksal Sinir Ağı (PNN) Algoritmasına İlişkin Bulgular

Araştırma kapsamında kullanılan makine öğrenmesi algoritmalardan biri olan PNN, çakışan kurallar için aktivasyonun üst sınırını tanımlamak için Teta Eksi (Theta Minus), çakışmayan kurallar için etkinleştirmenin alt sınırını tanımlamak için Teta Artı (Theta Plus) hiper parametrelerine sahiptir. Varsayılan hiper parametre değerleri olan 'Teta Eksi: 0,2' ve 'Teta Artı: 0,4' ile çalıştırılan Olasılıksal Sinir Ağı algoritmasına ait başarım değerleri Tablo 3'te yer almaktadır. Her ne kadar farklı hiper parametre değerleriyle algoritma tahmin becerisinin artırılması istense de hiper parametrelerde yapılan ayarlardan elde edilen sonuçlar, varsayılan değerlerde sağlanan başarım değerlerinin üzerine çıkamamıştır.

Tablo 3

Olasılıksal Sinir Ağı Algoritması'nın Başarım Değerleri

Parametreler	Başarım	*N*	%	Metrikler	
Varsayılan	DT	510	70,5	K	0,688
	YT	213	29,5		
	Toplam	723	100		
İyileştirilmiş	DT	510	70,5	K	0,688
	YT	213	29,5		
	Toplam	723	100		

Not: DT: Doğru Tahmin YT: Yanlış Tahmin K: Cohen's kappa

Olasılıksal Sinir Ağının başarım ölçümüne ilişkin verilerin gösterildiği Tablo 3 incelendiğinde algoritmadan elde edilebilecek maksimum başarımın, varsayılan hiper parametre değerleriyle sağlandığı görülmektedir. Algoritma, test veri setinde yer alan sorulara %70,5 (n=510) oranında doğru yanıt vermiştir. Yanlış cevap oranı ise %29,5 (n=213) olarak ölçülmüştür. Olasılıksal Sinir Ağı algoritması, Destek Vektör Makinesi (DVM) kadar yüksek bir performans sergilemese de %70'in üzerindeki sınıflandırma oranı ile model için alternatif bir algoritma olduğu anlaşılmaktadır.

Tabloda önemli görülen bir diğer önemli veri ise Cohen's kappa değeridir. Algoritmanın başarımıyla uyumlu olması beklenen bu değer (K=0,688), "önemli" uyuşmaya işaret etmektedir.

3.1.3. K-En Yakın Komşu (KNN) Algoritmasına İlişkin Bulgular

Sınıflandırma projelerinde kolay uygulanabilirliği nedeniyle yaygın olarak başvurulan K-En Yakın Komşu (KNN) Algoritması, çalışmada test edilen algoritmalardan biridir. KNIME yazılımı üzerinde, Algoritma'ya ait üç hiper parametre bulunmaktadır. Bunlardan biri, Algoritma'nın yeni bir örneği sınıflandırmak için kullandığı en yakın komşuların sayısını belirlemek için başvurulan *dikkate alınacak komşu sayısı (number of neighbours to consider)* hiper parametresidir. Diğeri, komşuları veri noktasına olan mesafelerine göre ağırlıklandırmak için kullanılan *komşuların mesafeye göre ağırlığı (weight neighbours by distance)* parametresidir. Sonuncusu ise sınıflandırma sonuçlarına ilişkin olasılık değerlerini belirlemek için kullanılan *çıkış sınıf olasılıkları (output class probabilities)* parametresidir.

Araştırmada, Algoritma'nın performansını optimize etmek için iki aşamalı bir hiper parametre optimizasyon süreci uygulanmıştır. İlk aşamada, "dikkate alınacak komşu sayısı" hiper parametresi manuel olarak test edilmiş ve varsayılan 3 değerinin 1 ile değiştirilmesi sonucu algoritmadan elde edilebilecek maksimum başarım belirlenmiştir. İkinci aşamada ise diğer hiper parametrelerin performans üzerindeki etkisi araştırılmış ve bu parametrelerin sınıflandırıcının performansına etki etmediği gözlemlenmiştir.

Tablo 4

K-En Yakın Komşu Algoritması'nın Başarım Değerleri

Parametreler	Başarım	*N*	*%*	Metrikler	
Varsayılan	DT	458	63,3	K	0,612
	YT	265	36,7		
	Toplam	723	100		
İyileştirilmiş	DT	468	64,7	K	0,62
	YT	255	35,3		
	Toplam	723	100		

Not: DT: Doğru Tahmin YT: Yanlış Tahmin K: Cohen's kappa

Tablo 4'te, K-En Yakın Komşu Algoritması'nın sınıflandırma başarısına ilişkin veriler yer almaktadır. Varsayılan hiper parametre değerleriyle çalıştırılan sınıflandırıcı, test veri seti içerisindeki sorulara %63,3 (n=458) oranında doğru yanıt vermiştir. Sınıflandırıcının hatalı yanıt oranı ise %36,7 (n=265)'dir. Hiper parametre değerlerinde yapılan optimizasyonlar sonrasında ise algoritma, %64,7 (n=468) oranında doğru, %35,3 (n=255) oranında yanlış tahminde bulunmuştur.

Araştırmada, varsayılan ve iyileştirilmiş hiper parametre değerleriyle çalıştırılan algoritmanın doğru yanıt oranında %1,4 oranında bir artış gözlemlenmiştir (n=10). Bu artış, algoritmalara ilişkin en iyi performans değerlerinin saptanması açısından önemsiz olarak değerlendirilmiştir.

Tabloda ayrıca K-En Yakın Komşu Algoritması'na ilişkin metrikler yer almaktadır. Varsayılan ve iyileştirilmiş hiper parametrelerle çalıştırılan algoritmanın metrikleri incelendiğinde, Cohen's Kappa'nın varsayılan değerlerde K=0,612, iyileştirilmiş değerlerde ise K=0,62 olduğu tespit edilmiştir.

Kappa için "önemli" uyuşma olduğunu gösteren bu değer, algoritma başarımı için güvenirliği teyit etmektedir.

3.1.4. Naive Bayes (NB) Algoritmasına İlişkin Bulgular

Olasılık hesaplamalarına dayalı sınıflandırma gerçekleştiren ve Naive Bayes Teoremi'ni temel alan Naive Bayes Algoritması, projede kullanılan algoritmalardan biridir. Naive Bayes, *varsayılan olasılık (default probability), standart sapma eşiği (threhold standard deviation), minimum standart sapma (minimum standard deviation)* ve *özellik başına maksimum benzersiz nominal değer sayısı (maximum number of unique nominal values per attribute)* hiper parametrelerine sahiptir.

Tablo 5'te, varsayılan değerler satırında, belirli bir nitelik/sınıf değer çifti için olasılık parametresi olan 'varsayılan olasılık: 0,0001' ve yeterli (çeşitli) veri bulunmayan gözlemler için kullanılacak minimum standart sapmayı belirlemek amacıyla kullanılan 'minimum standart sapma: 0,0001' parametre değerleriyle çalıştırılan Naive Bayes Algoritması'nın varsayılan özelliklerine ilişkin başarım değerleri gösterilmektedir. Algoritma için en yüksek performansı sağlayan 'varsayılan olasılık: 0,1', 'minimum standart sapma: 0,1' hiper parametre değerleriyle çalıştırılan algoritmanın optimize edilmiş özelliklerine ait başarım değerlerine ise tablonun iyileştirilmiş değerler satırında yer almaktadır.

Model üzerinde Naive Bayes sınıflandırıcısını çalıştırmak için, her özellik için izin verilen maksimum benzersiz nominal değer sayısı, eğitim veri kümesindeki sınıf sayısından fazla olmalıdır. Bu nedenle, bu değer 1000 olarak belirlenmiştir. Bu sayede, model üzerinde çalıştırılabilen algoritmanın performans sonuçları Tablo 5'te gösterilmektedir.

Tablo 5

Naive Bayes Algoritması'nın Başarım Değerleri

Parametreler	Başarım	*N*	*%*	Metrikler	
Varsayılan	DT	461	63,8	K	0,612
	YT	262	36,2		
	Toplam	723	100		
İyileştirilmiş	DT	466	64,5	K	0,622
	YT	257	35,5		
	Toplam	723	100		

Not: DT: Doğru Tahmin YT: Yanlış Tahmin K: Cohen's kappa

Tablo 5'e göre varsayılan hiper parametreler ile çalıştırılan Algoritma'nın %63,8 (n=461) oranında doğru, %36,2 (n=262) oranında ise yanlış sınıflandırma gerçekleştirdiği görülmektedir. Optimize edilmiş hiper parametre değerleriyle bu oranların doğru sınıflandırma için %64,5 (n=466), yanlış sınıflandırma için ise 35,5 (n=257) olduğu tespit edilmiştir. Varsayılan ve optimize edilmiş versiyonlar arasındaki doğru sınıflandırma yüzdesinde %0,7 (n=5) artış olduğu belirlenmiştir.

Algoritmaya ilişkin bulgularda ayrıca sınıflandırıcı performansını ölçen Cohen's Kappa (K) ve Seçicilik (S) metrikleri yer almaktadır. Buna göre, test veri setindeki doğal dil sorularına, akademik veri tabanlarına ait özniteliklerle eğitilen modelin performansına ilişkin sunulan verilere olan güveni doğrulayan Cohen's Kappa değeri göze çarpmaktadır. Naive Bayes, varsayılan özellikleriyle K=0,612; iyileştirilmiş özellikleriyle K=0,622 Cohen's Kappa değerine sahiptir. Bu değerler "önemli" uyuşmaya işaret etmektedir.

3.1.5. Bulanık Mantık (Fuzzy) Algoritmasına İlişkin Bulgular

Geleneksel sınıflandırma algoritmalarının aksine, Bulanık Mantık Algoritması, bir elemanın bir kümeye aidiyet derecesini belirleyerek esnek bir sınıflandırma sunar. Bu sayede, günlük yaşamda karşılaşılan belirsiz ve karmaşık problemlerin çözümünde oldukça faydalıdır. Araştırmada, her bulanık aralığın üyelik değerlerini bir kural için birleştirerek, tüm kurallar üzerinden nihai bir çıktı hesaplayan 'Bulanık norm (Fuzzy norm): Min/Max Norm' ve farklı sınıfların kuralları arasındaki çatışmaları önlemek için kuralları indirgeyen bir küçültme yöntemi olan 'Küçültme Fonksiyonu (Shrink Function):

VolumeBorderBased' hiper parametre değerleriyle, Bulanık Mantık Algoritması'nın varsayılan özelliklerine ait başarım sonuçları Tablo 6'da gösterilmektedir. Algoritmanın performansını artırmak için parametreler üzerinde kapsamlı optimizasyon çalışmaları gerçekleştirilmiştir. Fakat, istenilen seviyede bir iyileştirme elde edilememiştir. Gerçekleştirilen iyileştirme ise yeni bir kural oluşturulduktan sonra çakışan kuralları azaltmak ve farklı sınıflardaki kurallarla çakışmayı önlemek için "Kaydettikten sonra küçült" (Shrink after commit) seçeneğinin kaldırılmasıyla sağlanmıştır. Bulanık mantık algoritmasına ilişkin bulgular Tablo 6'da özetlenmiştir.

Tablo 6

Bulanık Mantık Algoritması'nın Başarım Değerleri

Parametreler	Başarım	*N*	*%*	Metrikler	
Varsayılan	DT	432	59,8	K	0,577
	YT	291	40,2		
	Toplam	723	100		
İyileştirilmiş	DT	433	59,9	K	0,579
	YT	290	40,1		
	Toplam	723	100		

Not: DT: Doğru Tahmin YT: Yanlış Tahmin K: Cohen's kappa

Bulanık Mantık Algoritması'nın varsayılan ve iyileştirilmiş hiper parametre değerleriyle başarım düzeyleri Tablo 6'da yer almaktadır. Buna göre Algoritma'nın varsayılan hiper parametre değerleriyle %59,8 (n=432), iyileştirilmiş hiper parametre değerleriyle ise %59,9 (n=433) oranında doğru tahmin gerçekleştirdiği görülmektedir. Yanlış tahmin oranına bakıldığında ise, varsayılan hiper parametre değerlerinde %40,2 (n=291), iyileştirilmiş hiper parametre değerlerinde ise %40,1 (n=290) oranında hatalı sınıflandırma yapmıştır.

Algoritma'nın başarımına ilişkin metriklere bakıldığında ise, iyileştirilmiş hiper parametrelerle K=0,577 Cohen's Kappa değerini sağlayan Algoritma'da, gözlenen ve beklenen değerler arasında "orta" dereceli bir uyuşma olduğu belirlenmiştir.

3.1.6. Karar Ağacı (KA) Algoritmasına İlişkin Bulgular

Sınıflandırma problemleri için yaygın olarak kullanılan Karar Ağacı Algoritması başlangıçta, bölünmenin hesaplanacağı kalite ölçüsünü seçmek için 'kalite ölçüsü (quality measure): gini index', her düğüm için minimum kayıt sayısını belirlemek için 'düğüm başına minimum kayıt sayısı (min number records per node): 2', genelleme performansını artırmak için 'budama yöntemi (pruning method): budama yok (no pruning)' ve ağaçta saklanan kayıt sayısını seçmek için 'görünüm için saklanacak kayıt sayısı (number records to store for view): 10.000' varsayılan hiper parametre değerlerini kullanmaktadır. Algoritmadan elde edilebilecek maksimum başarımın belirlenebilmesi amacıyla, parametreler çok çeşitli varyasyonlarla sınanmış ve 'görünüm için saklanacak kayıt sayısı: 1', 'düğüm başına minimum kayıt sayısı: 1' değerleriyle algoritmanın başarı sınırı tespit edilmiştir. Varsayılan ve iyileştirilmiş hiper parametre değerleriyle test edilmiş olan Karar Ağacı Algoritması'na ilişkin veriler Tablo 7'de yer almaktadır.

Tablo 7

Karar Ağacı Algoritması'nın Başarım Değerleri

Parametreler	Başarım	*N*	*%*	Metrikler	
Varsayılan	DT	11	1,5	K	0,008
	YT	712	98,5		
	Toplam	723	100		
İyileştirilmiş	DT	416	57,5	K	0,533
	YT	307	42,5		
	Toplam	723	100		

Not: DT: Doğru Tahmin YT: Yanlış Tahmin K: Cohen's Kappa

Tablo 7'de, öncelikli olarak Algoritma'nın varsayılan ve iyileştirilmiş değerleri arasındaki önemli farklar göze çarpmaktadır. Varsayılan hiper parametre değerleriyle %1,5 (n=11) oranında doğru sınıflandırma yüzdesine sahip olan karar ağacı, hiper parametreler üzerinde gerçekleştirilen optimizasyonların ardından %57,5 (n=416) oranında doğru sınıflandırma gerçekleştirmiştir. Öte yandan sınıflandırıcı, varsayılan hiper parametre değerleriyle %98,5 (n=712) oranında yanlış sınıflandırma gerçekleştirmişken, iyileştirilmiş hiper parametre değerlerinde bu oran %42,5'e (n=307) düşmüştür. Sınıflandırıcıda sağlanan iyileştirmenin ve doğru tahmin yüzdesinin uyumunu kontrol için başvurulan Cohen's Kappa, varsayılan hiper parametrelerde K=0,008, iyileştirilmiş hiper parametrelerde K=0,533 değeri ile "orta" uyuşmaya işaret etmiştir. Bu durum, karar ağacına ilişkin performans ölçümünü teyit etmektedir. Gerçekleştirilen optimizasyonlar sonucunda elde edilen skorlar, test veri setindeki sorulara algoritmanın büyük oranda doğru yanıt verdiğini göstermektedir. Fakat bu sonuç, istenilen başarı seviyesinin altında kalmaktadır.

3.1.7. Derin Öğrenme (DL4J) Algoritmasına İlişkin Bulgular

Çalışmada, performansı değerlendirilen makine öğrenmesi yöntemlerinden biri de derin öğrenmedir. Proje hedeflerine uygunluğu ve kullanışlılığı göz önünde bulundurularak, etiketlenmiş verilerden öğrenen bir teknik olan denetimli öğrenme için derin ağlar (deep networks for supervised learning) mimarisi ve teknikleri tercih edilmiştir.

Derin öğrenme kütüphanesinde yer alan DL4J Algoritması, varsayılan olarak 'ağırlık başlatma stratejisi (weight initialisation strategy): XAVIER', 'kayıp fonksiyonu (loss function): mean squared error', 'öğrenme oranı (learning rate): 0,1' çıktı katmanı parametrelerini ve 'döngü (epochs): 1', 'grup boyutu (batch size): 1' veri parametreleriyle 'rastlantısal dereceli azalma (stochastic gradient descent)' öğrenme yöntemini kullanmaktadır.

Algoritma'nın hiper parametrelerinden 'ağırlık başlatma stratejisi', katman için başlangıç ağırlıklarını ayarlamak amacıyla kullanılacak stratejiyi belirlemek; 'kayıp fonksiyonu', katman için kullanılması gereken kayıp fonksiyonunun türünü seçmek; 'öğrenme oranı', katman için kullanılması gereken öğrenme oranını ayarlamak; 'döngü', tüm veri seti üzerinde yürütülen eğitimlerin sayısını girmek; 'grup boyutu', küçük gruplar için örnek sayısını vermek ve 'rastlantısal dereceli azalma', öğrenme yönetimini tayin etmek için kullanılmaktadır.

Algoritma üzerinde gerçekleştirilen uzun uğraşların sonucunda, model üzerinde en iyi başarım yakalayan hiper parametreler, çıktı katmanı için 'ağırlık başlatma stratejisi: XAVIER', 'kayıp fonksiyonu: Cosine proximity', 'öğrenme oranı: 0,5' ve veri parametreleri için 'döngü: 2000', 'grup boyutu: 1000' şeklinde tespit edilmiştir. Ayrıca algoritma için tek bir 'yoğun katman' (dense layer) ileri besleme katmanından yararlanılmış ve bu katman için aktivasyon fonksiyonunun türü 'aktivasyon fonksiyonu (activition function): ReLU' olarak atanmıştır. Buna göre varsayılan ve iyileştirilmiş hiper parametre değerleriyle çalıştırılan DL4J Algoritması'nın başarım durumuna ait veriler Tablo 8'de sunulmaktadır.

Tablo 8

Derin Öğrenme Algoritması'nın Başarım Değerleri

Parametreler	Başarım	*N*	*%*	Metrikler	
Varsayılan	DT	12	1,7	K	-0,001
	YT	711	98,3		
	Toplam	723	100		
İyileştirilmiş	DT	523	72,3	K	0,704
	YT	200	27,7		
	Toplam	723	100		

Not: DT: Doğru Tahmin YT: Yanlış Tahmin K: Cohen's kappa

Model üzerinde çalıştırılan, DL4J Algoritması'nın başarım değerlerine Tablo 8'de yer verilmektedir. Buna göre varsayılan ve iyileştirilmiş hiper parametre değerleriyle çalıştırılan Algoritma'nın iki versiyonu arasında önemli performans farkları olduğu göze çarpmaktadır. Varsayılan parametrelerle çalıştırılan Algoritma'nın doğru sınıflandırma oranı %1,7 (n=12), yanlış sınıflandırma oranı ise %98,3 (n=711)'tür. Hiper parametrelerde gerçekleştirilen optimizasyonlar neticesinde doğru sınıflandırma oranının %70,6 (n=511) artarak %72,3'e (n=523) çıktığı, yanlış sınıflandırma oranının ise aynı oranda azalarak %27,7'ye (n=200) düştüğü tespit edilmiştir.

DL4J Algoritması'na ilişkin önemli bir diğer bulgu da Cohen's Kappa metriğine aittir. Optimize edilmiş hiper parametrelerle çalıştırılan Algoritma'nın bu değeri K=0,704'tür. Bu değer, "önemli" uyuşmaya işaret etmektedir. Sonuç olarak derin öğrenmenin model için alternatif bir algoritma olduğu anlaşılmıştır.

3.1.8. Makine Öğrenmesi Algoritmalarının Başarım Değerlerinin Karşılaştırılması

Çalışmada, model için uygun olduğu belirlenen makine öğrenmesi algoritmaları, varsayılan ve optimize edilmiş parametre değerleriyle çalıştırılmış ve elde edilen performans değerleri ayrı ayrı sunulmuştur. Fakat model için en uygun algoritmanın seçilebilmesi için bu değerlerin karşılaştırılması büyük önem taşımaktadır. Bu sayede hem model için kullanışlı algoritmalar hem de en iyi performansı sağlayan algoritma belirlenebilecektir.

Tablo 9

Makine Öğrenmesi Algoritmalarının Başarım Değerleri

Başarım Değerleri	*F*			*%*		
Algoritmalar	DT	YT	T	DTY	YTY	T
Destek Vektör Makinesi	670	53	723	92,7	7,3	100
Olasılıksal Sinir Ağı	510	213	723	70,5	29,5	100
K-En Yakın Komşu	463	260	723	64,0	36,0	100
Naive Bayes	461	262	723	63,8	36,2	100
Bulanık Mantık	432	291	723	59,8	40,2	100
Karar Ağacı	416	307	723	57,5	42,5	100

Not: DT: Doğru tahmin; YT: Yanlış Tahmin; DTY: Doğru Tahmin Yüzdesi; YTY: Yanlış Tahmin Yüzdesi; T: Toplam

Çalışmada kullanılan makine öğrenmesi algoritmalarının doğru ve yanlış tahmin skorlarının gösterildiği Tablo 9'da en yüksek başarının %92,7 (n=670) oranla destek vektör makinesi tarafından sağlandığı anlaşılmaktadır. Bunu %70,5 (n=510) başarı oranı ile olasılıksal sinir ağı izlemektedir. K-En Yakın Komşu, Naive Bayes, Bulanık Mantık ve Karar Ağacı Algoritmalarından ise sırasıyla %64,0 (n=463), %63,8 (n=461), %59,8 (432) ve %57,5 (n=416) oranında doğru tahmin skorları elde edilmiştir. Dolayısıyla, kütüphane danışma hizmetleri çerçevesinde, bir araştırma konusuyla ilgili bilgi kaynağı gereksiniminin doğal dille tasvir edildiği soruların, akademik veri tabanlarına ait özniteliklerle eğitilen bir makine öğrenmesi modeli tarafından yanıtlanmasında en kullanışlı algoritmanın benzerlik muhakemelerinden dış değerlendirme yaparak öğrenen destek vektör makinesi olduğu söylenebilir. Ayrıca Olasılıksal Sinir Ağı Algoritması da benzer çalışmalar için alternatif bir algoritma olarak ön plana çıkmaktadır. Bununla birlikte araştırmada sınanan diğer sınıflandırıcılardan K-En Yakın Komşu ile Naive Bayes; Bulanık Mantık ile Karar Ağacı Algoritmaları'nın birbirlerine oldukça yakın performans sağladığı tespit edilmiştir.

Çalışmada, Derin Öğrenme Algoritması'nın diğer makine öğrenmesi algoritmalarına kıyasla nasıl bir performans sergilediği de merak konusu olmuştur. Bu kapsamda, model üzerinde test edilen makine öğrenmesi algoritmalarından %70 doğru sınıflandırma eşiğini aşan Destek Vektör Makinesi ve Olasılıksal Sinir Ağı Algoritmaları ile derin öğrenme için kullanılan DL4J Algoritması'nın başarım değerlerine ilişkin karşılaştırma Tablo 10'da yer almaktadır.

Tablo 10

Derin Öğrenme Algoritması'nın (DL4J) Diğer Makine Öğrenmesi Algoritmalarına (DVM Ve PNN) Göre Başarım Durumu

Başarım Değerleri	*F*			*%*		
Algoritmalar	DT	YT	T	DTY	YTY	T
Destek Vektör Makinesi	670	53	723	92,7	7,3	100
Derin Öğrenme (DL4J)	523	200	723	72,3	27,7	100
Olasılıksal Sini Ağı	510	213	723	70,5	29,5	100

Not: DT: Doğru tahmin; YT: Yanlış Tahmin; DTY: Doğru Tahmin Yüzdesi; YTY: Yanlış Tahmin Yüzdesi; T: Toplam

Çalışmada iyileştirilmiş parametre değerleriyle çalıştırılan ve %70'in üzerinde doğru tahmin başarısı elde edilen Destek Vektör Makinesi (DVM) ve Olasılıksal Sinir Ağı (PNN) Algoritmaları ile modele uygunluğu nedeniyle tercih edilen DL4J Algoritması'nın karşılaştırılması önemli görülmüştür. Tablo 10'da DL4J Algoritması'nın, %72,3 (n=523) oranında doğru tahmin yeteneğine sahip olduğu görülmektedir. Bu oran, derin öğrenme algoritmasını makine öğrenmesi algoritmaları arasında en yüksek performansı gösteren (DTY: %92,7) Destek Vektör Makinesi'nin hemen ardından ikinci sıraya yerleştirmektedir. Dolayısıyla DL4J Algoritması'nın, Olasılıksal Sinir Ağı Algoritmasıyla birlikte Destek Vektör Makinesinin alternatifi olduğu tespit edilmiştir.

3.2.Tartışma

Makine öğrenmesinde, her problem için tek bir algoritma çözüm sunmaz. Farklı bilim dalları farklı algoritmalar geliştirir ve çalışmalarında öncelikle bu algoritmalara başvurulur (Domingos, 2017). Bu araştırma, kütüphane ve bilgi bilimi alanında kütüphane danışma hizmetleri çerçevesinde ele alınmıştır. Akademik veri tabanlarının özellikleriyle eğitilmiş bir makine öğrenmesi modeli üzerinde çalıştırılan yedi farklı makine öğrenmesi algoritmasının bilgi kaynağı (kitap, dergi vb.) gereksinimini betimleyen Türkçe doğal dil sorgularına uygun veri tabanını önerebilme başarısı değerlendirilmiştir. Veri matrisine uygun olan Destek Vektör Makinesi, Olasılıksal Sinir Ağı, K-En Yakın Komşu, Naive Bayes, Bulanık Mantık, Karar Ağacı ve Derin Öğrenme Algoritmaları test edilmiştir. Araştırma sonuçlarına göre, modele en uygun algoritmanın Destek Vektör Makinesi olduğu belirlenmiştir.

Destek Vektör Makinesi Algoritması, örüntü tanıma ve sınıflandırma gibi birçok alanda yaygın olarak kullanılmakta ve bu görevlerde yüksek başarı göstermektedir (Bray ve Han, 2004, s. 265). Algoritma'nın genelleme yeteneği, optimal çözüm bulma becerisi ve ayırt edici gücü sayesinde diğer makine öğrenmesi algoritmalarına kıyasla pek çok araştırmada daha iyi performans gösterdiği gözlemlenmiştir (Cervantes vd., 2020, s. 189). Kütüphane ve bilgi bilimi disiplininde daha önce yapılan makine öğrenmesi çalışmaları değerlendirildiğinde, Destek Vektör Makinesi Algoritması'nı merkeze alan birçok çalışma yapıldığı anlaşılmıştır. Örneğin, İsveç Akademik ve Araştırma Kütüphanelerinin ortak kataloğu olan İsveç Ulusal Birlik Kataloğu LIBRIS'ten temin edilen katalog kayıtları üzerinde Dewey Onlu Sınıflama Sistemi'ne göre makine öğrenmesi teknik ve yöntemleri kullanarak bir sınıflandırma çalışması gerçekleştiren Golub, Hagelback ve Ardö (2020), farklı makine öğrenmesi algoritmalarını sınamışlar, en yüksek başarım sağlayan algoritmanın Destek Vektör Makinesi olduğunu belirlemişlerdir. Diğer bir araştırmada ise, Wagstaff ve Liu (2018), kütüphane koleksiyonunda ayıklama işlemini makine öğrenmesi ile inceleyen bir araştırma gerçekleştirmiştir. Wesleyan Üniversitesi Kütüphanesi koleksiyonu için makine öğrenmesi sınıflandırıcılarının tahminleri ile kütüphanecilerin ayıklama kararları arasında istatistiksel bir uyum olduğunu bulan araştırmacılar, Destek Vektör Makinesi sınıflandırıcısından oldukça yüksek başarım elde etmişlerdir. Öte yandan, Kütüphane ve bilgi bilimi alanındaki denetimli makine öğrenmesi çalışmalarında, Destek Vektör Makinesi Algoritması'nın yoğun bir şekilde kullanıldığı görülmektedir. Örneğin, Binici (2019) elektronik belgelere otomatik dosya plan numarası atamak için Destek Vektör Makinesi kullanmış ve yüksek başarı elde etmiştir. Waqas, Anjum ve Afzal (2023) ise bu algoritmayı araştırma makalelerinden üst veri çıkarmak için kullanmış ve benzer şekilde başarılı sonuçlar almıştır.

Bu çalışmada, akademik veri tabanlarının özellikleriyle eğitilmiş bir makine öğrenmesi modeli kullanılarak, test veri seti içerisindeki doğal dil sorgularına ilgili akademik veri tabanlarını önerme yeteneği değerlendirilmiştir. Farklı algoritmalar test edilmiş ve en yüksek başarı oranı Destek Vektör Makinesi tarafından elde edilmiştir. Sonuç olarak, kütüphane ve bilgi bilimi alanındaki denetimli makine öğrenmesi çalışmaları için Destek Vektör Makinesi Algoritması'nın öncelikli olarak tercih edilebileceği söylenebilir.

Çalışmada, Derin Öğrenme Algoritması (DL4J) da yüksek başarı gösteren sınıflandırıcılardan biri olarak öne çıkmıştır. Doğal dil işleme, sanal asistanlar, metinden resme çeviri, sahte haber tespiti ve otomatik dil çevirileri gibi birçok alanda kullanılan derin öğrenme (Mathew, Amudha ve Sivakumari, 2021, s. 607), otomatik özellik çıkarma yeteneği sayesinde diğer makine öğrenmesi algoritmalarına kıyasla makine öğrenmesi projelerinde sıklıkla tercih edilen bir teknik haline gelmiştir (Patterson ve Gibson, 2017, s. 6). Artan bilgi işleme gücüyle birlikte, derin öğrenme modelleri son yıllarda duygu analizi, haber sınıflandırması, soru yanıtlama ve doğal dil çıkarımı gibi çeşitli sınıflandırma görevlerinde klasik makine öğrenimi tabanlı yaklaşımları geride bırakmaya başlamıştır (Minaee vd., 2021). Bu çalışmada, makine öğrenmesi yaklaşımıyla bir sınıflandırma modeli test edilmiş ve DL4J adlı Derin Öğrenme Algoritması kullanılmıştır. Yüksek doğruluk ve karmaşık problemleri çözme becerisi nedeniyle tercih edilen bu Derin Öğrenme Algoritması, diğer makine öğrenmesi algoritmalarına kıyasla daha yavaş işlem yapmaktadır. Bunun nedeni hesaplama karmaşıklığının yüksek olması ve çok sayıda parametre kullanmasıdır. Ancak, Derin Öğrenme Algoritması'nın, yüksek bir doğruluk oranı sunması ve diğer algoritmaların çözemediği karmaşık problemleri çözebilmesi gibi önemli avantajları bulunmaktadır. Bu durumda Derin Öğrenme Algoritması'nın benzer çalışmalar için bir alternatif olabileceği ifade edilebilir.

Yapay sinir ağlarından biri olan ve diğer sinir ağı yöntemlerine göre daha etkili bir sınıflandırma potansiyeli bulunan Olasılıksal Sinir Ağı (Alweshah vd., 2022, s. 1810), çalışmada kullanılan ve yüksek performans sağladığı anlaşılan bir diğer sınıflandırıcıdır. Etkili bir sınıflandırma algoritması olan Olasılıksal Sinir Ağı, finansal risk tahmini, biyomedikal mühendisliği, sibernetik gibi farklı araştırma alanlarında yaygın olarak kullanılmaktadır (Chaki, Routray ve Mohanty, 2022, s. 2). Algoritma'nın bu çalışmada sağladığı yüksek sınıflandırma başarımı, sınıflandırıcının kütüphane ve bilgi bilimi disiplininde gerçekleştirilecek benzer çalışmalar için alternatif bir tercih olabileceğine işaret etmektedir.

Çalışmada K-En Yakın Komşu, Naive Bayes, Bulanık Mantık ve Karar Ağacı Algoritmaları da test edilmiştir. Bu Algoritmalar, daha önceki birçok sınıflandırma çalışmasında yüksek doğruluk elde etmesine rağmen (Awad ve ELseuofi, 2011; Mohamed, 2017; Shah vd., 2020; Osisanwo vd., 2017; Uddin vd., 2019), bu çalışmada istenilen başarıyı yakalayamamıştır. Bunun sebebi, sınıflandırma yöntemlerinin başarısının büyük ölçüde sınıflandırılan verilerin özelliklerine bağlı olmasıdır. Bu nedenle, en iyi sınıflandırma yöntemini belirlemek için deneme yanılma yöntemi kullanılmıştır (Boateng, Otoo ve Abaye, 2020, s. 343). Yapılan testler sonucunda, en yüksek başarıyı Destek Vektör Makinesi Algoritması elde etmiştir. Buna alternatif olarak Derin Öğrenme ve Olasılıksal Sinir Ağı Algoritmaları da yüksek başarı göstermiştir.

Makine öğrenmesi algoritmalarının hiper parametrelerinin optimize edilmesinin sınıflandırıcıların performansı üzerinde önemli bir etkiye sahip olduğu çalışmada gözlemlenen bir diğer önemli konudur. Hiper parametre optimizasyonu, bir makine öğrenmesi modeli için en iyi hiper parametre kombinasyonunu bulma sürecidir (Emeç ve Özcanhan, 2023). Talaei Khoei ve Kaabouch (2023), hiper parametre ayarlarının makine öğrenmesi modellerinde önemli sonuç farklılıkları yaratabileceğini ve her modelin kendine özgü optimize edilmesi gereken hiper parametrelere sahip olduğunu vurgulamaktadır. Bu optimizasyon, manuel olarak veya otomatik optimizasyon teknikleri kullanılarak gerçekleştirilebilir. Manuel optimizasyon, özellikle çok sayıda parametre, karmaşık modeller ve doğrusal olmayan hiper parametre etkileşimleri içeren durumlarda zorlayıcı ve zaman alıcı olabilir. Bu araştırmada sınıflandırıcılardan en yüksek performansı elde etmek için, kullanılan algoritmaların hiper parametreleri üzerinde yoğun bir şekilde çalışma yapılmıştır. Farklı hiper parametre kombinasyonlarının test edilmesi sonucunda, birçok algoritmanın performansında önemli artışlar elde edilmiştir. Örneğin, Destek Vektör Makinesi Algoritması'nda yapılan optimizasyon ile %17'lik bir doğru tahmin artışı sağlanmıştır. Benzer şekilde, diğer algoritmalarda da önemli performans artışları gözlemlenmiştir. Bu sonuçlar, makine

öğrenmesi projelerinde algoritma seçiminin yanında hiper parametre optimizasyonunun da başarının anahtarı olduğunu göstermektedir.

4. Sonuç ve Öneriler

Bir makine öğrenmesi modelinin, bir konu hakkında bilgi kaynağına erişmek için kütüphane danışma birimlerine yöneltilen sorulara yanıt verebilme kabiliyetinin sınandığı bu çalışmada, akademik veri tabanlarına ait konu, tür, erişim formatı gibi özniteliklerle eğitilen bir makinenin ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler Rehberi'nde yer alan konular çerçevesinde yapılandırılmış doğal dil sorularına ilgili akademik veri tabanını önerme kabiliyeti çeşitli makine öğrenmesi algoritmaları karşılaştırılarak test edilmiştir.

2022 yılı Ocak-Mayıs ayları arasında, çalışmada kullanılan makine öğrenme modeli için eğitim veri seti oluşturulmuştur. Bu veri seti, 133 akademik veri tabanından toplanan konu, tür ve erişim formatı gibi öznitelikleri içeren verilerden oluşmaktadır. Veri madenciliği ve makine öğrenmesi teknikleri kullanılarak bu veriler bilgisayarca analiz edilmiş ve modele öğretilmiştir. Daha sonra, makinenin performansını değerlendirilebilmek için, "Yöntem" bölümünde açıklandığı gibi doğal dil soruları içeren bir test veri seti hazırlanmıştır. Bu test veri seti, bilgi kaynağı ihtiyacının doğal dille ifade edildiği sorulardan oluşmaktadır. Metin madenciliği ve doğal dil işleme teknikleri kullanılarak bu sorular işlenmiş ve makine öğrenmesi algoritmalarının test edilebileceği bir model oluşturulmuştur. Bu çerçevede çalışmada, denetimli makine öğrenmesi yöntemlerinden yedi farklı algoritma kullanılarak tahmin edici bir model geliştirilmiş ve modelin performansı değerlendirilmiştir. Elde edilen sonuçlar ışığında araştırma sorularına aşağıdaki cevapları verebilmek mümkündür:

- Çalışma kapsamında yanıt aranan ilk araştırma sorusuna (Modelde veri matrislerine uygun makine öğrenmesi algoritmaları nelerdir?) yanıt veren yedi algoritma tespit edilmiştir. Veri matrisleriyle uyumlu olan bu algoritmalar Destek Vektör Makinesi, Olasılıksal Sinir Ağı, K- En Yakın Komşu, Naive Bayes, Bulanık Mantık, Karar Ağacı ve Derin Öğrenmedir.
- "Model üzerinde sınanan algoritmalardan hangileri kabul edilebilir performans göstermiştir?" sorusu araştırmada yanıtı aranan ikinci sorudur. Veri matrisleriyle uyumlu algoritmalar arasından üçünün başarım değerlerinin kabul edilebilir olduğu tespit edilmiştir. Bu bağlamda, Destek Vektör Makinesi (DVM), Derin Öğrenme (DL4J) ve Olasılıksal Sinir Ağı (PNN) Algoritmaları'nın %70'in üzerinde doğru sınıflandırma gerçekleştirdiği belirlenmiştir. Model üzerinde iyileştirilmiş hiper parametre ayarlarıyla çalıştırılan algoritmalardan DVM'nin %92,7, DL4J'nin %72,3 ve PNN'nin %70,5 oranında başarılı sınıflandırma yaptığı tespit edilmiştir.
- Üçüncü araştırma sorusu olan "Model için en kullanışlı makine öğrenmesi algoritması nedir?" sorusuna yanıt olarak, Destek Vektör Makinesi (DVM) Algoritması'nın en uygun algoritma olduğu tespit edilmiştir. Hiper Tanjant çekirdek fonksiyonunda "kappa: 9" ve "delta: 3" parametre değerleri ve yanlış sınıflandırma için her noktaya 10 ceza puanı (örtüşen ceza: 10) ile optimize edilen algoritma, %92,7'lik performansıyla bu makine öğrenmesi projesi için en uygun sınıflandırıcı olarak belirlenmiştir.
- Dördüncü araştırma sorusu olan "Makine öğrenmesi algoritmalarının çekirdek hesaplama yöntemlerinde ve hiper parametre ayarlarında yapılan iyileştirmelerin başarıya etkisi nedir?" sorusunu yanıtlamak için, model üzerinde çalıştırılan algoritmaların çekirdek hesaplama yöntemleri ve hiper parametreleri optimize edilmiştir. Bu optimizasyon sonucunda, birçok algoritmanın sınıflandırma performansında önemli bir artış olduğu gözlemlenmiştir. Örneğin, model için en yüksek başarıyı gösteren Destek Vektör Makinesi Algoritması üzerinde yapılan iyileştirmeler sonucunda, başarıda %15'ten fazla bir artış sağlanmıştır. Karar Ağacı ve Derin Öğrenme Algoritmaları'nda ise bu artış %50'nin üzerindedir. Bu durum, algoritmaların çekirdek hesaplama yöntemleri ve hiper parametreleri üzerinde yapılan optimizasyonların ne kadar önemli olduğunu göstermektedir.
- Beşinci araştırma sorusunun (Modelin, ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler Rehberi'nde yer alan konular çerçevesinde yapılandırılmış doğal dil sorularına yanıt olarak akademik veri tabanlarına yönlendirme kabiliyeti nedir?) yanıtlanması amacıyla, akademik veri

tabanlarının özniteliklerinden elde edilen eğitim veri seti ile bir konudaki bilgi kaynağına erişim gereksinimini betimleyen doğal dil sorularından oluşan test veri seti oluşturulmuştur. Eğitim veri seti içerisindeki akademik veri tabanlarından elde edilen öznitelikler, veri madenciliği teknikleri kullanılarak makineye öğretilmiştir. 'ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler Rehberi'nde yer alan konular çerçevesinde bir konu hakkında bilimsel bilgi kaynağına erişimi betimleyen, yapay olarak oluşturulmuş doğal dil sorularından oluşan test veri seti üzerinde ise metin madenciliği ve doğal dil işleme tekniklerine başvurulmuştur. Eğitim veri setindeki akademik veri tabanlarından elde edilen özelliklerden oluşturulan bir sözlük kullanılarak, test veri setindeki doğal dil soruları etiketlenmiştir. Böylece, eğitim ve test veri setlerinden elde edilen matrislerin uyumluluğu kontrol altına alınmıştır. Model üzerinde çalıştırılan makine öğrenmesi algoritmalarına ilişkin testler neticesinde, başta Destek Vektör Makinesi olmak üzere Derin Öğrenme Algoritması (DL4J) dahil birçok sınıflandırıcının başarılı sonuçlar verdiği görülmüştür.

Bu çalışma, kütüphane danışma hizmetlerinin otomasyonu için makine öğrenmesi teknik ve yöntemlerinin etkinliğini göstermektedir. Araştırma bulguları, danışma hizmetlerinin güncel teknolojilerle sürdürülmesine ve geliştirilmesine katkıda bulunmaktadır. Elde edilen makine öğrenmesi algoritmaları ve hiper parametre değerleri, kütüphane danışma hizmetleri ile ilgili gelecekteki makine öğrenmesi çalışmalarına rehberlik edecektir.

Bu araştırmanın kapsamı ve elde edilen bulgular ışığında, aşağıdaki öneriler sunulmuştur:

- Elde edilen bulgular, kütüphane danışma hizmetleri bağlamında makine öğrenmesi temelli sınıflandırma çalışmaları için Destek Vektör Makinesi Algoritması'nın en uygun seçenek olduğunu göstermektedir. Bu nedenle, kütüphane ve bilgi bilimi disiplininde benzer sınıflandırma problemleri için Destek Vektör Makinesi Algoritması'nın kullanılması önerilmektedir.
- Bu araştırmada akademik veri tabanlarıyla eğitilen bir makine öğrenmesi modeli değerlendirilmiş ve model başarımı ÜAK Doçentlik Bilim Alanları ve Anahtar Kelimeler Rehberi'nde yer alan konulara dayalı olarak oluşturulan, bir bilgi kaynağı ihtiyacını betimleyen doğal dil sorularıyla sınanmıştır. Kütüphanelerde danışma hizmeti taleplerindeki çeşitlilik ve format göz önünde bulundurularak tasarlanacak farklı veri kümeleriyle oluşturulacak bir makine öğrenmesi projesinde, bu çalışmada uygulanan modelleme sürecine benzer bir yaklaşım benimsenmesi önerilmektedir.
- Farklı hiper parametre ayarları ile çalıştırılan çok sayıda makine öğrenmesi algoritmasının performans verileri arasında önemli farklılıklar gözlemlenmiştir. Parametre değerlerinin seçimi, makine öğrenmesi projesinin yapısına ve kullanılan algoritmaya göre değişkenlik göstermektedir. Bu nedenle, gelecekteki makine öğrenmesi çalışmalarında her algoritma için en uygun çekirdek hesaplama ve hiper parametre ayarlarının belirlenmesi ve raporlanması literatüre önemli katkılar sağlayabilir.
- Elde edilen bulgular, Derin Öğrenme Algoritması'nın bu görevde önemli bir başarı elde ettiğini göstermektedir. Diğer makine öğrenmesi algoritmalarına kıyasla doğal olarak daha yavaş bir öğrenme sürecine sahip olsa da benzer projelerde Derin Öğrenme Algoritması da değerlendirilmelidir.
- Bu çalışmada, modelin oluşturulması, özellikle veri hazırlama aşamasında büyük özen gerektirmiştir. Türkçe karakter ve boşluk içeren kelimelerin etiketlenmesinde karşılaşılan zorluklar, doğal dil sorularının yer aldığı test veri kümesinden eğitim veri kümesiyle uyumlu etiketleme sağlamak için kullanılan sözlük sayesinde çözülmüştür. Benzer projelerde Türkçe ve boşluk içeren kelimelerin etiketlenmesinde sorun yaşanmaması için alanda çalışanlar benzer yaklaşım sergileyebilirler.

Bu çalışmada hem sınıflandırma performansını hem de modelin genel başarısını yükseltmek için, bir bilim alanına ait alt konular ve anahtar kelimeler, sözlük yardımıyla en üst hiyerarşideki bilim alanına dönüştürülmüştür. Bir çeşit denetimli kavram dizini olan bu sözlük sayesinde eğitim ve test aşamalarında çok boyutlu matrislerin oluşması engellenmiştir. Benzer makine öğrenmesi projelerinde yüksek performans elde etmek için, geniş boyutlu matrislerin oluşmasını önlemek amacıyla bu yöntem kullanılabilir. Bu çalışmada, akademik veri tabanlarından elde edilen özniteliklerle eğitilmiş bir makine

öğrenmesi modelinin, kütüphane danışma hizmetleri kapsamında, bilgi ihtiyacını ifade eden doğal dil sorularına cevap verme yeteneği test edilmiştir. Gelecekteki çalışmalarda, kütüphane danışma birimlerine yöneltilen Türkçe sıkça sorulan sorular üzerinde benzer bir makine öğrenmesi çalışması yapılması önerilmektedir.

Bu çalışma, kütüphane danışma hizmetleri çerçevesinde akademik veri tabanlarına odaklanmaktadır. Ancak kütüphanelerde, danışma hizmetlerinin boyutu oldukça büyüktür ve kullanıcıların bilgi ihtiyaçlarını karşılamak için çok çeşitli hizmetler sunulmaktadır. Bu hizmetler, kullanıcı davranışları ve bilgi arama eğilimleri hakkında zengin veri kümelerini ortaya çıkarmaktadır. Bu veriler, büyük dil modelleri gibi yeni teknolojilerin kütüphanelerde daha geniş bir şekilde kullanılmasına olanak sağlayabilir. Bu modeller, kullanıcıların sorularını daha doğru bir şekilde yorumlayabilir, alakalı sonuçlar önerebilir ve bilgiye erişimlerini kolaylaştırarak kütüphanelerde kullanıcı deneyimini önemli ölçüde geliştirme potansiyeline sahiptir.

Bu çalışmada sunulan makine öğrenmesi modeli, akademik veri tabanlarından metin verilerini işlemek için metin madenciliği tekniklerini ve değişkenleri tanımlayan denetimli kavramlar dizinini kullanmaktadır. Modelin metin madenciliği bileşeni, özellikle veri setleri bakımından, büyük dil modelleriyle entegrasyon için oldukça uygundur. Bu entegrasyonun, modelin doğal dili daha iyi anlamasını ve kullanıcılara daha kapsamlı, alakalı ve faydalı yanıtlar vermesini sağlayacağı öngörülmektedir. İleriki çalışmalarda, modelin uygulamaya konması durumunda, büyük dil modellerinin uygulama programlama arayüzüne (API) entegrasyonu olağan bir durum olup başarım düzeyinin ölçülmesine yönelik değerlendirmelerin yapılması önerilmektedir. Değerlendirmenin, kullanıcı memnuniyeti, bilgi erişimi, verimlilik ve başarı oranı gibi kriterlere göre gerçekleştirilmesi önemli görülmektedir.

Bu çalışmada, akademik veri tabanlarından elde edilen bilgilerle eğitilmiş bir makine öğrenmesi modeli, kütüphanelerde danışma hizmeti sunma potansiyeli açısından test edilmiştir. Farklı makine öğrenmesi algoritmaları kullanılarak doğal dil sorularına cevap verme yeteneği değerlendirilmiş ve başta Destek Vektör Makinesi Algoritması olmak üzere birçok algoritmanın bu görevi başarıyla yerine getirebildiği anlaşılmıştır. Bu bulgular, kütüphanelerde danışma hizmetlerini otomatikleştirmek için makine öğrenmesi tabanlı sohbet robotlarının uygulanabilirliğini desteklemektedir. Kütüphane sohbet robotlarının maliyet tasarrufu, zaman optimizasyonu ve kullanıcı memnuniyetini artırma gibi birçok fayda sağlayabileceği düşünülmektedir. Bu nedenle, kütüphane danışma hizmetlerini otonom hâle getirecek sohbet robotu formatında makine öğrenmesi uygulamalarının geliştirilmesi önerilmektedir.

Etik Standartlar ile Uyumluluk

Çıkar Çatışması: Yazarlar herhangi bir çıkar çatışmasının olmadığını beyan eder.

Etik Kurul İzni: Bu çalışma için etik kurul iznine gerek yoktur.

Yazar Katkı Beyanı: Yazarlar çalışmaya eşit derecede katkı vermiştir.

Finansal Destek: Finansal destek yoktur.

Kaynakça

Alweshah, M., Rababa, L., Ryalat, M. H., Al Momani, A. ve Ababneh, M. F. (2022). African Buffalo Algorithm: Training the Probabilistic Neural Network to Solve Classification Problems. *Journal of King Saud University-Computer and Information Sciences, 34*(5), 1808-1818.

Asemi, A., Ko, A. ve Nowkarizi, M. (2020). Intelligent Libraries: A Review on Expert Systems, Artificial Intelligence, and Robot. *Library Hi Tech, 39*(2), 412-434. https://doi.org/10.1108/LHT-02-2020-0038

Awad, W. A., Ve Elseuofi, S. M. (2011). Machine Learning Methods For Spam E-mail Classification. *International Journal of Computer Science & Information Technology (IJCSIT), 3*(1), 173-184. https://doi.org/10.5121/ijcsit.2011.3112

Aydın, F. ve Aslan, Z. (2017). Yapay Öğrenme Yöntemleri ve Dalgacık Dönüşümü Kullanılarak Nöro Dejeneratif Hastalıkların Teşhisi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 32*(3), 749-766. https://doi.org/10.17341/gazimmfd.337621

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

Boateng, E. Y., Otoo, J. ve Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, *8*(4), 341-357. https://doi.org/10.4236/jdaip.2020.84020

Bray, M. ve Han, D. (2004). Identification of Support Vector Machines for Run off Modelling. *Journal of Hydroinformatics, 6*(4): 265–280. https://doi.org/10.2166/hydro.2004.0020

Cervantes, J., Garica-Lamont, F., Rodriguez-Mazahua, L. ve Lopez, A. (2020). A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing*, *408*, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

Chaki, S., Routray, A. ve Mohanty, W. K. (2022). A Probabilistic Neural Network (PNN) Based Framework for Lithology Classification Using Seismic Attributes. *Journal of Applied Geophysics*, *199*, 104578. https://doi.org/10.1016/j.jappgeo.2022.104578

Coşkun, F. ve Gülleroğlu, H. D. (2021). Yapay Zekânın Tarih İçindeki Gelişimi ve Eğitimde Kullanılması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *54*(3), 947-966. https://doi.org/10.30964/auebfd.916220

Cox, A. M., Pinfield, S. ve Rutter, S. (2019). The Intelligent Library: Thought leaders' Views on the Likely Impact of Artificial Intelligence on Academic Libraries. *Library Hi Tech*, *37*(3), 418-435. https://doi.org/10.1108/LHT-08-2018-0105

Cox, A. (2022). How Artificial Intelligence Might Change Academic Library Work: Applying the Competencies Literature and the Theory of the Professions. *Journal of the Association for Information Science and Technology*, 1-14. https://doi.org/10.1002/asi.24635

Daniel. (2021, 11 Ocak). 7 *ways Artificial Intelligence is Changing Libraries.* IRIS AI. https://iris.ai/academics/7-ways-ai-changes-libraries/

Demirhan, T. (2015). Makine Öğrenmesi Algoritmalarının Karmaşıklık ve Doygunluk Analizinin Bir Veri Kümesi Üzerinde Gerçekleştirilmesi (Doktora tezi). Trakya Üniversitesi, Fen Bilimleri Enstitüsü.

Domingos, P. (2017). *Master Algoritma: Yapay Öğrenme Hayatımızı Nasıl Değiştirecek?* (Çev. Tufan Göbekçin). İstanbul: Paloma Yayınevi.

Emeç, M. ve Özcanhan, M. H. (2023). Makine Öğrenmesi Algoritmalarında Hiper Parametre Belirleme. *Mühendislikte Öncü ve Çağdaş Çalışmalar* içinde, 71-98. https://as-books.com/index.php/mocc/article/download/39/33

Fernandez, P. (2016). "Through the Looking Glass: Envisioning New Library Technologies" How Artificial Intelligence will Impact Libraries. *Library Hi Tech News, 33*(5), 5-8. https://doi.org/10.1108/LHTN-05-2016-0024

Golub, K., Hagelback, J. ve Ardö, A. (2020). Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches. *Journal of Data and Information Science*, *5*(1), 2020, 18–38. https://doi.org/10.2478/jdis-2020-0003

Gökalp, Ö. M. (2022). *Makine Öğrenmesi.* https://www.academia.edu/68874574/Machine_Learning

Görmez, B. (2021). Adli Bilişimde Makine Öğrenmesi: Makine Öğrenmesi Algoritmaları ile Terör Olaylarının Tahmin Edilmesi Çalışması (Yayımlanmamış yüksek lisans tezi). Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Jin. (2019, 25 Aralık). *7 Methods to Evaluate Your Classification Models*. (Medium). https://medium.com/analytics-vidhya/everything-you-need-about-evaluating-classification-models-dfb89c60e643

Kaya, E. (2017). Değişen Kullanıcı Alışkanlıkları Doğrultusunda Bir Web Keşif Aracı Model Önerisi (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Khanzode, K. C. A. ve Sarode, R. D. (2020). Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literature Review. *International Journal of Library and Information Science (IJLIS), 9*(1), 30-36. https://sdbindex.com/Documents/index/00000018/00000-09008

Landis, J. R. ve Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33* (1), 159-174. https://dionysus.psych.wisc.edu/iaml/pdfs/landis_1977_kappa.pdf

Marr, B. (2023, 28 Şubat). *Beyond ChatGPT: 14 Mind-blowing AI Tools Everyone Should be Trying Out Now*. https://www.forbes.com/sites/bernardmarr/2023/02/28/beyond-chatgpt-14-mind-blowing-ai-tools-everyone-should-be-trying-out-now/

Mathew, A., Amudha, P., ve Sivakumari, S. (2021). Deep Learning Techniques: An Overview. *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020 içinde* (599-608). https://link.springer.com/chapter/10.1007/978-981-15-3383-9_54#Sec3

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., ve Gao, J. (2021). Deep Learning-based Text Classification: A comprehensive Review. *ACM Computing Surveys (CSUR), 54*(3), 1-40.

Mohamed, A. E. (2017). Comparative Study of Four Supervised Machine Learning Techniques for Classification. *International Journal of Applied, 7*(2), 1-15. https://www.academia.edu/download/54482697/2.pdf

Nawaz, N., Gomes, A. M., ve Saldeen, M. A. (2020). Artificial Intelligence (AI) Applications for Library Services and Resources in COVID-19 Pandemic. *Artificial intelligence (AI), 7*(18), 1951-1955.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O. ve Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT), 48*(3), 128-138. https://doi.org/10.14445/22312803/IJCTT-V48P126

Özhan, E. (2020). Makine Öğrenmesi Yöntemleri ile Web'den Bilgi Çıkarımı Sürecinin İyileştirilmesi. *Afyon Kocatepe Üniversitesi Uluslararası Mühendislik Teknolojileri ve Uygulamalı Bilimler Dergisi, 3*(2), 52-59. https://dergipark.org.tr/en/download/article-file/1252531

Patterson, J. ve Gibson, A. (2017). *Deep Learning: A practitioner's Approach.* O'Reilly Books.

Shah, K., Patel, H., Sanghvi, D. ve Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research, 5*(12), 1-16. https://doi.org/10.1007/s41133-020-00032-0

Şeyranlıoğlu, O. (2022). Şirket Değerlemesinde Makine Öğrenmesi Algoritmalarının Kullanımı: Holding Şirketleri Üzerine Bir Araştırma (Yayımlanmamış doktora tezi). Giresun Üniversitesi Sosyal Bilimler Enstitüsü, Giresun.

Talaei Khoei, T. ve Kaabouch, N. (2023). Machine Learning: Models, Challenges, and Research Directions. *Future Internet, 15*, 1-29. https://doi.org/10.3390/fi15100332

Tektaş, M., Akbaş, A., ve Topuz, V. (2002). Yapay Zekâ Tekniklerinin Trafik Kontrolünde Kullanılması Üzerine Bir İnceleme. *Uluslararası Trafik ve Yol Güvenliği Kongresi, Gazi Üniversitesi, Ankara*.

Uddin, S., Khan, A., Hossain, M. E., Moni, M. A. (2019). Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. *BMC Medical Informatics and Decision Making, 19*(281), 1-16. https://doi.org/10.1186/s12911-019-1004-8

Wagstaff, K. L. ve Liu, G. Z. (2018). Automated Classification to Improve the Efficiency of Weeding Library Collections. *The Journal of Academic Librarianship, 44*(2), 238-247. https://doi.org/10.1016/j.acalib.2018.02.001

Waqas, M., Anjum, N. ve Afzal, M. T. (2023). A Hybrid Strategy to Extract Metadata from Scholarly Articles by Utilizing Support Vector Machine and Heuristics. *Scientometrics, 128*(8), 4349-4382. https://econpapers.repec.org/article/sprscient/v_3a128_3ay_3a2023_3ai_3a8_3ad_3a10.1007_5fs11192-023-04774-7.htm

Wheatley, A. ve Hervieux, S. (2019). Artificial Intelligence in Academic Libraries: An Environmental Scan. *Information Services & Use, 39*(4), 347-356. https://doi.org/10.3233/ISU-190065

Widmann, M. (2020, 21 Eylül). *Cohen's Kappa: Learn It, Use It, Judge It.* (KNIME). https://www.knime.com/blog/cohens-kappa-an-overview

Yılmaz, E. (2021). *Bilgi Merkezlerinin Varlık Sebebi ve Müşterisi Olarak Kullanıcı.* Bilgi Merkezlerinde Yönetim-1 *içinde.* İstanbul: Hiper Yayın.