T.C.

İSTANBUL AYDIN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM PROGRAMI

ULUSLARARASI HABER RAPORLARININ RAPOR İÇERİKLERİNDE KULLANILAN İFADELERE GÖRE MAKİNE ÖĞRENMESİ YÖNTEMİYLE SINIFLANDIRILMASI VE DENETLENMESİ

YÜKSEK LİSANS TEZİ

Firdevs DURNAGÖL

Bilgisayar Mühendisliği Ana Bilim Dalı

Bilgisayar Mühendisliği Programı

Şubat,2020

T.C.

İSTANBUL AYDIN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM PROGRAMI

ULUSLARARASI HABER RAPORLARININ RAPOR İÇERİKLERİNDE KULLANILAN İFADELERE GÖRE MAKİNE ÖĞRENMESİ YÖNTEMİYLE SINIFLANDIRILMASI VE DENETLENMESİ

YÜKSEK LİSANS TEZİ

Firdevs DURNAGÖL

(Y1713.010070)

Bilgisayar Mühendisliği Ana Bilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Muttalip Kutluk ÖZGÜVEN

Şubat,2020

T.C.
İSTANBUL AYDIN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ

YÜKSEK LİSANS TEZ ONAY FORMU

Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı Y1713.010070 numaralı öğrencisi **Firdevs DURNAGÖL**'ün "**Uluslararası Haber Raporlarının Rapor İçeriklerinde Kullanılan İfadelere Göre Makine Öğrenmesi Yöntemiyle Sınıflandırılması ve Denetlenmesi**" adlı tez çalışması Enstitümüz Yönetim Kurulunun 31.01.2020 tarihli ve 2020/02 sayılı kararıyla oluşturulan jüri tarafından oybirliği/~~oyçokluğu~~ ile Tezli Yüksek Lisans tezi 18.02.2020 tarihinde kabul edilmiştir.

	Unvan	Adı Soyadı	Üniversite	İmza
		ASIL ÜYELER		
Danışman	Prof. Dr.	Muttalip Kutluk ÖZGÜVEN	İstanbul Aydın Üniversitesi	
1. Üye	Dr. Öğr. Üyesi	Ahmet GÜRHANLI	İstanbul Aydın Üniversitesi	
2. Üye	Dr. Öğr. Üyesi	Farzad KIANI	İstanbul Arel Üniversitesi	
		YEDEK ÜYELER		
1. Üye	Prof. Dr.	Ali GÜNEŞ	İstanbul Aydın Üniversitesi	———
2. Üye	Doç. Dr.	Metin ZONTUL	İstanbul Arel Üniversitesi	———

ONAY

Prof. Dr. Ragıp Kutay KARACA
Enstitü Müdürü

YEMİN METNİ

Yüksek Lisans tezi olarak sunduğum "Uluslararası Haber Raporlarının Rapor İçeriklerinde Kullanılan İfadelere Göre Makine Öğrenmesi Yöntemiyle Sınıflandırılması ve Denetlenmesi" adlı çalışmanın, tezin proje safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin Bibliyografya'da gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim. (18/02/2020)

Firdevs DURNAGÖL

İÇİNDEKİLER

Sayfa

KISALTMALAR

Σ : Toplam Sembolü

∏ : Çarpım Sembolü

Ac : Doğruluğun Kontrolü

Arff : Weka Dosya Formatı

Doc : Word Dosya Formatı

Fn : Olumsuz Yanlış

Fp : Olumlu Yanlış

IN : International News

K-NN : K-en Yakın Komşu

MN : Magazine News

MÖ : Makine Öğrenmesi

P : Kesinlik

Pdf : Taşınabilir Belge Biçimi

PGT : Preimplantasyon Genetik Tanı

RO : Rastgele Orman

SN : Sports News

Stddev : Standart Sapma

SVM : Destek Vektör Makinesi

Tn : Olumsuz Doğru

Tp : Olumlu Doğru

S : Kaynak

ŞEKİL LİSTESİ

Sayfa

TABLO LİSTESİ

Sayfa

ULUSLARARASI HABER RAPORLARININ RAPOR İÇERİKLERİNDE KULLANILAN İFADELERE GÖRE MAKİNE ÖĞRENMESİ YÖNTEMİYLE SINIFLANDIRILMASI VE DENETLENMESİ

ÖZET

Rapor verisinin miktarının çok olması durumunda giderek artan veri yoğunluğu içinde tasnifi ve arşivlenmesine yönelik işlemlerin yapılması zordur. Bu zorluğun aşılması, raporların denetlenmesi, düzenlenmesi ve düzeltilmesi, Karar Destek Sistemleri yollarından biri olan Makine Öğrenme ile aşılabilir. Raporların analiz edilmesi, anlamsız veriler arasından anlamlı verilerin çıkarılması, verinin kullanımı açısından büyük kolaylık sağlamaktadır. Bu yapılan araştırma, uluslararası yayın yapan büyük bir medya organının çevrimiçi olarak dünya çapında yayınladığı haber ve bilgi raporlarının makine öğrenme algoritmaları kullanılarak sınıflandırılmasına dayanmaktadır. Uygulamanın analiz aşamasında Rastgele Orman Karar Ağacı, ZeroR, Naif Bayes yöntemleri kullanılmıştır. Bu yöntemlerin sınıflandırma başarıları birbirleri ile karşılaştırılmıştır. Bunlar arasında en iyi sonuçları veren algoritma Rastgele Orman Karar Ağacı yönteminin dayandığı algoritmada parametrik değişiklikler ve düzenlemeler yapılması sonucu rapor sınıflandırmada sonuçlarda yüksek iyileştirmeler elde edilmiştir. Başarı oranı %91'e ve performans süresi 0.47s'e çıkmıştır. Araştırmadaki veri seti içerisinde her birinden 600 rapor olacak şekilde üç adet sınıf, uluslararası konularda raporlar, spor raporları, dergi (magazin) raporlarıdır. Veri setinin bir kısmı eğitim ve bir kısmı test kümesi olarak kullanılmış, 10-katlı çapraz doğrulama yöntemi ile algoritmik doğruluklar denetlenmiştir. Bu sayede, veri seti, hem test hem de eğitim kümesi olarak kullanılmıştır. Derleme ortamı olarak Weka veri madenciliği yazılımı kullanılmıştır.

Anahtar kelimeler: *Sınıflandırma, Metin Madenciliği, Makine Öğrenmesi, Gazetecilik, Rastgele Orman Algoritması*

CLASSIFICATION AND CONTROL OF INTERNATIONAL NEWS REPORTS BY TO EXPRESSIONS USED WITHIN REPORT CONTENTS THROUGH MACHINE LEARNING

ABSTRACT

Due to the large amount of data and the increasing density of data, it is difficult to process data. This hardship can be overcome by Data Mining. Analyzing the data, extracting meaningful data from meaningless data provides great convenience in terms of data usage. This study is a classification of the news that published on the website of an international channel by using artificial intelligence algorithm. Random Forest Decision Tree Algorithm, ZeroR Algorithm and Naif Bayesian Algorithm have been used in the analysis phase of the application The results of classification algorithms have been compared with each other. The algorithm that has given the best result among them is the Random Forest Decision Tree Algorithm. The success rate has been found as 91% and the duration of work has been found as 0.47 seconds. There are three classes in the dataset. These are International News (600), Sports News (600), Magazine News (600). Some of the dataset has been used as training and some has been used as test dataset. Algorithm accuracy has been checked by 10-fold cross validation method. Thus, the entire dataset has been used as both test and training dataset. Weka has been used as the compilation tool.

Key words: *Classification, Text Mining, Machine Learning, Journalism, Random Forest Algorithm*

I. GİRİŞ

Günümüzde verinin fazla olması ve bu durumun giderek artması sorunlara yol açmaktadır. Fazla olan veri arasında anlamlı verilerin kaybolması firmalar için, büyük sıkıntılar oluşturmaktadır. Teknolojinin gelişmesiyle bu sorun ortadan kalkmaktadır. Metin madenciliği yardımıyla anlamsız veriler arasından anlamlı verileri ayırt edilebilir. Metin madenciliği hayatın her alanında karşımıza çıkabilmektedir. Sağlık, bilim, askeriye gibi sektörler bu duruma örnek olarak verilebilmektedir. Metin madenciliği ile sınıflandırma yardımıyla anlamsız verilerden kurtulmak mümkündür. Metin madenciliği içerisinde sınıflandırma ve regresyon işlemleri yapılabilmektedir. Sınıflandırma işlemi, elde edilen verilerin hangi sınıfa ait olduğunu bulmayı amaçlamaktadır. Regresyon işlemi, mevcut bilgilere dayanarak sonucu tahmin etmeyi amaçlamaktadır. Metin madenciliğinde kullanılan çok sayıda yapay zeka algoritması mevcuttur. Mevcut probleme göre ilgili algoritmadan yararlanılarak analizler yapılmaktadır. Algoritma seçimi veri setinin doğru analiz edilmesi için önem taşımaktadır. Veri setini iyi tanıyarak, veri seti için hangi algoritmanın daha uygun olduğuna karar verilmesi ve algoritmanın ona göre seçilmesi gerekmektedir. Bazı algoritmalar sadece regresyon amaçlı kullanılırken bazı algoritmarda sınıflandırma amaçlı kullanılmaktadır. Literatürde bu konu ile ilgili bir çok çalışma bulunmaktadır. Sınıflandırma amaçlı kullanılan algoritmalardan bazıları, Naif Bayes, Destek Vektörler, İstatiksel Tabanlı Algoritmalar, Örnek Tabanlı Algoritmalar, Karar Ağaçları, Yapay Sinir Ağlarıdır.

Yapılan çalışmada, Uluslararası yayın yapan bir kanalın İnternet sitesi üzerinden yayınlanan metinler üzerinde sınıflandırma işlemi yapılmıştır. Naif Bayes, ZeroR ve Rastgele Orman Algoritmaları çalışma içerisinde kullanılmıştır. Çalışmanın amaç, belirli sınıflara dahil olan haber verileri üzerinde üç faklı tip ve üç farklı kategori yapay zeka algoritmaları kullanılarak makine eğitilmesi yöntemiyle yapılan sınıflandırmada, algoritmaların sınıflandırma analizlerinin karşılaştırılması ve haber metinlerinin yönetimlerinin saptanmasıdır. Weka derleyici ortamı kullanılarak analizler yapılmıştır.

II. YAPAY ZEKA VE HABER

A. Alan Araştırması

Yapay zeka algoritması kullanılarak veri madenciliği yapılması, birden fazla işten kazanç sağlamak için önemlidir. Anlamlı verilerin anlamsız verilerden ayrılması karar vermeyi kolaylaştırır, zamandan tasarruf sağlar. Zamandan tasarruf ise maddi kazancı beraberinde getirmektedir. En önemli şey zaman tasarrufudur. Yapılacak işlerin daha kısa sürede yapılması, firmaların kar etmesine de olanak sağlamaktadır. Yapay zeka algoritmaları ile makine öğrenmesi yapıldığında, birden fazla alanda işlevsel olarak bu durum kullanılmaktadır. Makinenin kullanılan algoritmaya göre sonuca ulaşması daha kolaydır. Günümüzde ve daha öncesinde bu konu ile ilgili makaleler, tezler ve bildiriler yayınlanmıştır. Literatür incelendiğinde konunun önemi daha da açık bir şekilde görülmektedir. Aşağıda literatürde bulunan bazı çalışmalar ve çalışmaların sonucu verilmiştir.

İşler & Narin (2012) 'WEKA Yazılımında k-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi' adlı yapmış olduğu çalışmasında amaç, Konjektif Kalp Yetmezliği Hastalığının k-Ortlama kümeleme kullanarak teşhis edilmesi ve algoritmanın başarı ölçümüdür. Hasta olan kayıtlar '0', hasta olmayan kayıtlar '1' sınıfında gösterilmiştir. '0' sınıfında 'cluster0've 'cluster1' kümeleri, '1' sınıfında 'cluster2've 'cluster3' kümeleri mevcuttur. Çalışmada çeşitli k değerleri denenerek testler yapılmıştır ve en başarılı ölçüm k=4 olduğunda bulunmuştur. K=4 alındığında başarı oranı %98,72 olarak bulmuşlardır. Veri setinde 83 adet kayıt bulunmaktadır. Bunlardan 29 adedi hasta, 54 adedi hasta olmayan kişilere ait kayıtlardır. Sonuç olarak, 83 adet kayıttan 82 adet kaydı doğru olarak tespit etmişlerdir.

Bilgin (2018) 'Metin Madenciliği Yöntemleri İle Yazar Tanıma: Divan Edebiyatı Örneği' adlı yapmış olduğu çalışmasında amaç, yazarı bilinmeyen eserlerin yazarlarını tespit etmektir. Çalışmada 25 adet divan edebiyatı şairine ait eserler kullanılmıştır. K-En Yakın Komşu algoritması, Destek Vektör Makinesi, Karar

Ağacı, Naive Bayes algoritmaları kullanılarak analizler yapılmıştır. 20 farklı model oluşturularak %91,45' lik doğruluk ve %90,23'lük f-değerine ulaşılmıştır.

Çınar & Atan (2019) 'Borsa İstanbul'da Finansal Haberler İle Piyasa Değeri İlişkisinin Metin Madenciliği Ve Duygu (Sentiment) Analizi İle İncelenmesi' adlı yapmış olduğu çalışmasında amaç, şirket için yapılan yorumların olumlu ya da olumsuz olduğunun anlaşılmasıdır. Çalışmada olumlu ve olumsuz toplam 14.108 adet şirket yayınları, medyada çıkan haberler ve sosyal medyadaki metinler incelenmiştir. Çalışmada duygu analizi yapılarak olumlu ve olumsuz yorumlar tespit edilmiştir. Duygu skoru yöntemiyle incelemeler yapılmıştır. 30 adet şirket için veriler incelenmiştir Sonuç olarak iki şirket haricinde 28 adet şirketin olumlu sonuç yaptığı tespit edilmiştir.

Yücel & Keskin Köylü (2018) 'Spam İçerikli E-Postaların Tespiti İçin Bir Metin Madenciliği Uygulaması: Terimlerin Gama İlişki Katsayısına Dayalı Polarizasyonu' adlı yapmış olduğu çalışmasında amaç, metin madenciliği ile gama ilişki katsayısı oluşturularak e-maillerin sınıflandırılmasıdır. Spam ve Notspam olmak üzere iki adet sınıf bulunmaktadır. 1480 adet e-mail veri olarak kullanılmıştır. Sınıflandırılma başarı oranı %81,2' dir.

Aydın (2018) 'Makine Öğrenmesi Algoritmaları Kullanılarak İtfaiye İstasyonu İhtiyacının Sınıflandırılması' adlı yapmış olduğu çalışmasında amaç, mevcuttaki ihtiyaçlar baz alınarak, bölgelere göre ihtiyaçların belirlenmesidir. Sınıflandırma çalışmasındaki bazı özellikler; aracın varış süresi, nüfus yoğunluğu, oraya yönlendirilen ana ve yan araçlar, bölgede bulunan itfaiyelerin sayısı. Veri seti İzmir Büyük Şehir Belediyesinden elde edilmiştir. 808 adet bölgeye ait veriler kullanılmıştır. En Yakın K komşu Algoritması ve Rastgele Orman Karar Ağacı Algoritması analizler için kullanılmıştır. Analiz sonucun da Rastgele Orman algoritması ile %93.84 başarı oranı elde edilmiştir.

Namous vd. (2018) 'Online News Popularity Prediction' adlı yapmış olduğu çalışmasında amaç, çevrim içi haberlerin popüleritesinin belirlenmesidir. Mashable web sitesinden haber verileri kullanılmıştır. Çalışmada, Çok Katlı Algı, Badding, Lojistik Regresyon, Naive Bayes Algoritması, Rasgtgele Orman Karar Ağacı Alagoritması, K- En Yakım Komşu Algoritması, Destek Vektör Makineleri

kullanılmıştır. Rastgele Orman Karar Ağacı ve Yapay Sinir Ağlarında en iyi sonuç elde edilmiştir. Yapılan çalışmada %65 başarı elde edilmiştir.

Patsis & Verhelst (2008) 'A Speech/Music/ Silence /Garbage/ Classifier for Searching and İndexing Broadcast News Material' adlı yapmış olduğu çalışmasında amaç, haber yayın materyallerinin sınıflandırılmasıdır. Veri seti Vtr dir. Vtr içerisinde konuşma, müzik, sessizlik ayırt edilmiştir. K-En Yakın Komşu Algoritması, B-Net, Karar Ağaçları, Destek Vektör Makineleri analiz esnasında kullanılmıştır. En iyi sonuç Destek Vektör Makineleri ile elde edilmiştir. Yapılan çalışma sonucunda %94 lük bir başarı elde edilmiştir.

Topaloğlu & Sur (2014) 'Sarılık Semptomlarında Yanlış Teşhisi Azaltmak için Karar Ağacı Uygulaması' adlı yapmış olduğu çalışmasında amaç, hastalık teşhisi konulurken yanlış konulan teşhisin azaltılmasıdır. Bu çalışmada sarılık tanısı konmuş 300 hastanın verileri kullanılmıştır. Analizler C5.0 ve j48 Karar Ağacı Algoritması kullanılarak yapılmıştır. Clemantine ortamı ve Weka ortamı kullanılmıştır. Weka içerisinde j48 Karar Ağacı Algoritması çalıştırılmıştır. Clemantine ortamında ise C5.0 Karar Ağacı Algoritması kullanılmıştır. Analizler Web tabanlı bir yazılım ile birleştirilmiştir. Çalışmanın doktorların karar vermesinde kolaylık sağlaması beklenmektedir. Sonuç olarak bir yazılım tasarlanıp, kullanıma sunulmuştur.

Şengür & Tekin (2013) 'Öğrencilerin Mezuniyet Notlarının Veri Madenciliği Metotları İle Tahmini' adlı yapmış olduğu çalışmasında amaç, öğretmenlerin yaptığı yanlış değerlendirmelerin önüne geçmektedir. Çalışma kapsamında Fırat Üniversitesi öğrencilerinin notları tahmin edilmeye çalışılmıştır. Çalışma analizleri Yapay Sinir Ağları ve Karar Ağaçları kullanılarak yapılmıştır. Çalışmanın dağlayacağı bir diğer yarar ise öğrencinin eğer not ortalaması belirli bir seviyenin altında ise öğrenci notunu yükseltmesi konusunda, öğrenciyi uyaran bir sistem olmasıdır. Çalışma iki farklı şekilde yapılmıştır. İlk çalışmada; öğrencilerin 1. ve 2. sınıf notları incelenerek yılsonu notu tahmin edilmiştir. İkinci çalışmada; 1., 2. ve 3. sınıf notları incelenerek yıl sonu notu tahmin edilmiştir. İki çalışma incelendiğinde Yapay Sinir Ağlarının, Karar Ağaçlarına göre daha iyi performansta sonuçlar ürettiği görülmüştür.

Namlı & Özcan (2017) 'Makine Öğrenmesi Algoritmaları Kullanarak Gişe Hasılatının Tahmini' adlı yapmış olduğu çalışmasında amaç, gişeden elden edilen gelirin tahmin edilmesidir. Veri seti 11 adet özellik içermektedir. Bu özelliklerden

yararlanılarak analizler yapılmıştır. Analizde Yapay Sinir Ağları, Destek Vektör Makineleri, Rastgele Orman Karar Ağacı Algoritması, C4.5 Karar Ağacı Algoritması kullanılmıştır. Yapay Sinir Ağları Algoritması en iyi sonuç verdiği görülmüştür. Yapılan çalışma sonucunda 91.6166% başarı oranı elde edilmiştir.

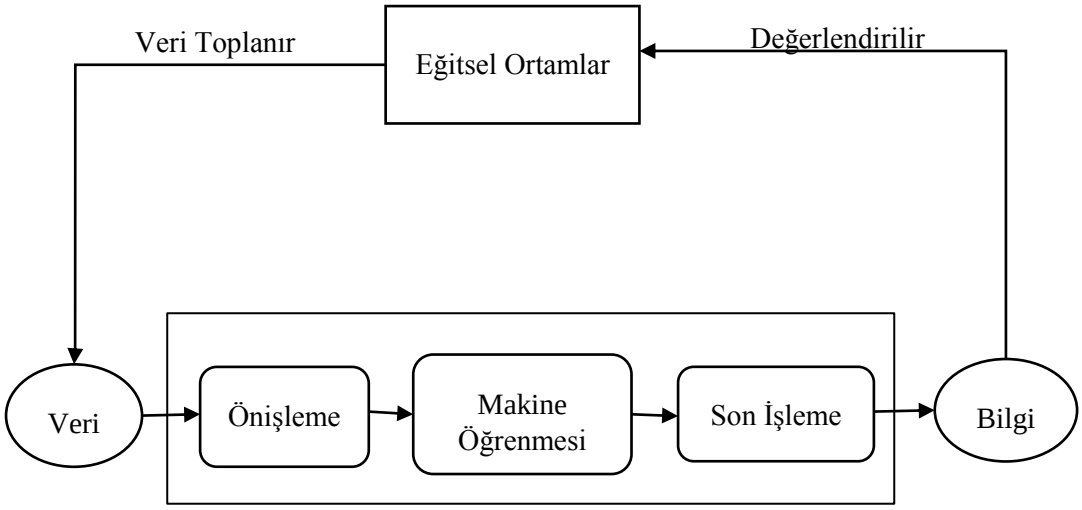
Göker & Tekedere (2017) 'FATİH Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi' adlı yapmış olduğu çalışmasında amaç, Fatih projesi için yapılan yorumların sınıflandırılmasıdır. 444 adet görüşün bulunduğu metinler incelenmiştir. Tr-idf ağırlıklandırma yöntemi ve vektörel olarak metinler gösterilmiştir. En iyi başarıyı %88.73 olarak Ardışık Minimal Optimizasyon Algoritması göstermiştir.

Gök (2017) 'Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi' adlı yapmış olduğu çalışmasında amaç, ortaokul öğrencilerinin başarılarının tahmin edilmesidir. Çalışma kapsamında 24 soruluk bir anket hazırlanmıştır ve bu anketi 6., 7. Ve 8. Sınıfa giden öğrenciler doldurmuştur. Matematik ve Türkçe derslerine ait dönem sonu notları tahmin edilmiştir. Veri setinde 1492 adet örnek bulunmaktadır. Analiz esnasında K- En yakın Komşu Algoritması, Naive Bayes Algoritması, Destek Vektör Makineleri- doğrusal ve radyal tabanlı fonksiyon, Rastgele Orman Karar Ağacı Algoritması kullanılmıştır. Analizler Weka ortamında yapılmıştır. Rastgele Orman Karar Ağacı Algoritması Türkçe ve genel başarı ortalaması tahmininde en iyi sonucu vermiştir. Matematik dersinde ise Doğrusal Regresyon en iyi sonucu vermiştir.

B. Makine Öğrenmesi

Verinin fazla olması, bilgiye ulaşımı zorlaştırmaktadır. Veri fazlalığı nedeniyle kullanılabilir veriye erişmek problem oluşturmaktadır. Problemin giderilmesi makine öğrenmesi ile olmaktadır. Makine öğrenmesi, anlamsız veriler içerisinden anlamlı verilere ulaşmamızı sağlamaktadır. Buda zamanda tasarruf demektir. Bir saatte yapılacak bir işlem, alakasız verilerin fazlalığı, anlamlı verilerin anlamsız veriler içerisinde kaybolmasından dolayı aylar ya da yıllar sürebilmektedir. Zamandan tasarruf bütün projeler için önemlidir. Zamandan tasarruf etmek maliyetten de tasarruf etmek anlamına gelmektedir. Bilgi her yerde mevcuttur. Bu sebeple makine öğrenmesi alanı geniştir. Tıp, ekonomi, bankacılık, teknoloji vs. birden fazla alanda bu konu hakkında söz edilebilir (Ertuğrul vd., 2012). Şekil 1'de makine öğrenmesi iş

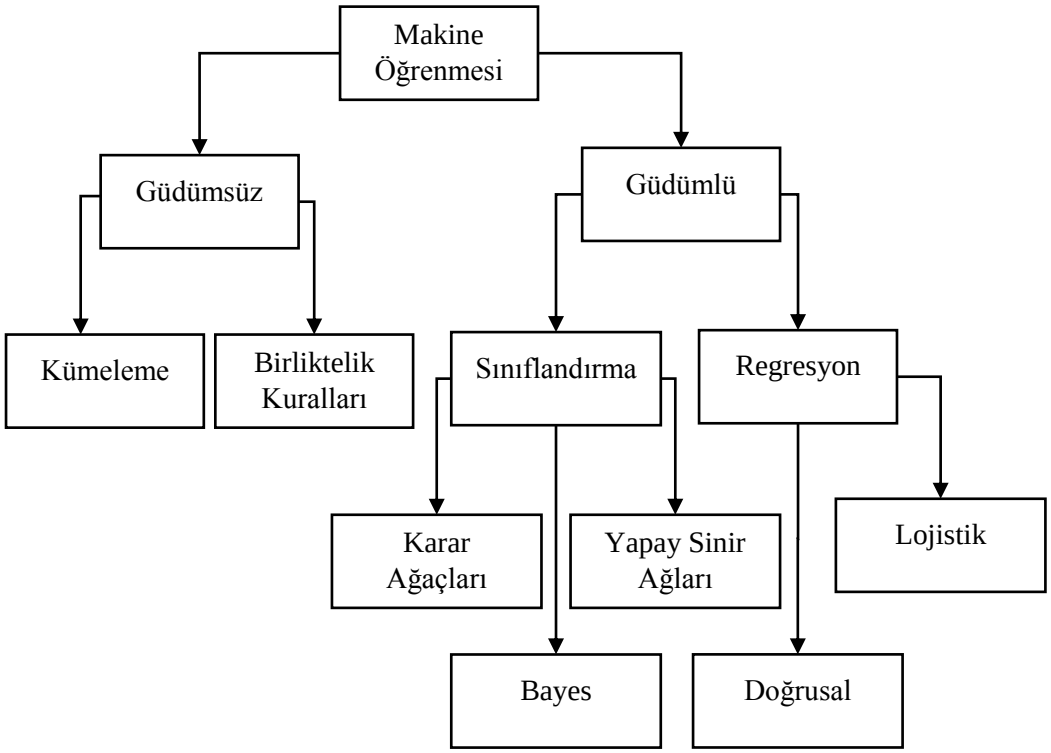
akışı adımları verilmiştir. Makine öğrenmesi iş akış adımlarında veri üzerindeki işlemler gözükmektedir. Veri ilk olarak toplanır ve derleyici ortamına yüklenir. Daha sonra veri üzerinde ön işleme yapılır. Veri setine göre uygun bir Yapay Zeka Algoritması seçilerek, veriye uygulanır. Algoritma uygulandıktan sonra sonuçlar değerlendirilir. Bu işlemler bir döngü şeklinde yapılmaktadır.



Şekil 1 Makine Öğrenmesi İş Akış Adımları

Kaynak: Akçapınar, 2014.

Makine Öğrenmesi yapılırken veri setine göre modeller oluşturulur. Veriyi iyi anlayıp, verinin akışına göre model oluşturulması gerekmektedir. Bu modellerin içerisinde Yapay Zeka Algoritmaları kullanılmaktadır (Yıldız & Şeker, 2016). Makine öğrenenmesi modelleri Şekil 2'de verilmiştir.



Şekil 2 Makine Öğrenmesi Modelleri

Kaynak: Narlı vd., 2014.

Makine öğrenmesi güdümsüz ve güdümlü modeller olarak ikiye ayrılmaktadır. Güdümsüz modeller, kümeleme ve birliktelik kuralları analizleridir. Güdümlü model ise sınıflandırma ve regresyon analizleri olarak ikiye ayrılmaktadır. Sınıflandırma analizlerinde karar ağaçları, bayes, yapay sinir ağları yöntemleri bulunmaktadır. Regresyon analizinde, lojistik ve doğrusal yöntemleri kullanılmaktadır. Bu kategorilerin altında çeşitli Yapay Zeka Algritmaları kullanılmaktadır.

Güdümüsüz modelerde, kümeleme ve birlikteliz analizleri olmak üzere iki çeşit analiz yapılmaktadır. Kümeleme analizi mevcut veri setinin alt kümelere ayrılması şeklinde olmaktadır. Sınıflandırmadan farkı, sınıflandırma analizinde belirli bir hedef sınıf bulunmaktadır. Kümeleme analizinde hedef söz konusu değildir. Örnğin: Bir mağaza için yeni müşteri potansiyeli ölçüleceği zaman, mevcut bilgiler yaş, kilo, boy vs şeklinde kümelere ayrılarak yapılmaktadır (Aydın, 2007). Birliktelik analizi, özelliklerin birlikte olduğunda verdiği sonucun durumuna göre yapılmaktadır. Literatürde sepet analizi olarak ta geçmektedir. Örneğin: Bir satış yapan mağazada, cipslerin yanındaki raflara hangi ürünlerin koyulacağı analizi. Bu durumun analizi

yapılırken daha önceki müşterilerin cips ile beraber hangi ürünleri aldığı incelenmektedir ve buna göre analiz gerçekleştirilmektedir (Aydın, 2007).

Güdümsüz modellerde, sınıflandırma ve regresyon olmak üzere iki çeşit analiz yapılmaktadır. Sınıflandırma analizi, daha önceden elde edilen gruplar üzerinden yapılmaktadır. Eğitim verisinde gruplar bellidir. Test verisi kullanılarak bunların hangi sınıfa dahil olduğu bulunmaktadır (Aydın, 2007). Regresyon analizinde tahmin yapılmaktadır. Mevcut veri setinin özelliklerine göre tahminler yapılmaktadır. Örneğin: Bir hastalığın teşhisinde doktora yardımcı olabilecek bir uygulama geliştirirken, mevcut eldeki bilgilerle hastanın yüzde kaç hasta olma durumu tahmin edilmektedir (Özcan, 2014).

C. Yapay Zeka'nın Makine Öğrenmesinde Kullanımı

Makine Öğrenmesi ve Yapay Zeka birlikte kullanılmaktadır. Yapay Zeka Algoritmaları kullanılarak makine öğrenmesi gerçekleşmektedir. Yapay Zeka Algoritmalarının kullanımı ve işlevselliği makine öğrenmesi için önemlidir. Veriyi doğru tanıyıp, veriyi doğru analiz yapacak Yapay Zeka Algoritmasıyla birleştirilmesi sonucun doğruluğu için hayati önem taşımaktadır. Burada ilk aşama veriyi anlamaktır. Doğru şekilde anlaşılmış bir veri, doğru Yapay Zeka Algoritması ile analiz edildiğinde, makine öğrenmesi analiz sonucu maksimum seviyede olacaktır (Atalay & Çelik, 2017).

1. Yapay Zeka'nın kullanım alanları

Sağlık sektöründe Yapay Zeka'nın yeri önemlidir. Geleceğin teknolojisi olarak Yapay Zeka görülmektedir. Sağlık sektörü içerisinde hemen hemen her alanda Yapay Zeka'ya yer vardır.

Hasta verileri üzerinde analizlerin yapılması, makinelerin kişi hakkındaki bilgisinin öğrenilmesi, kişinin genetik geçmiş bilgisi ve kişi üzerinde yapılan testler, tahliller makine öğrenmesi aracılığıyla makineye öğretilmektedir. Makine veri tabanındaki verilerle kişinin verilerini karşılaştırarak analiz yapılmaktadır. Yapay Zeka Algoritmaları kullanılarak, makine öğrenmesi aracılığıyla hastalıklarda erken teşhis yapılması, hastalıkların doğru teşhis edilmesi mümkündür. Erken teşhis hastalıklar için önemlidir ve hastanın hayatını kurtarmaktadır. Hastalıkların doğru teşhis edilmesinde, doktorun doğru kararlar vermesinde de Yapay Zeka Algoritmaları

kullanılmaktadır. Doğru kararın verilmesinde yardımcı olmaktadır (Demirhan vd., 2010).

Reçeteler, makine teşhisi koyduktan sonra, teşhise uygun ilaç tedavisi için reçete yazabilir. Hastalık teşhisi konulmuş birkişi için, kişinin genetik hastalıkları da göz önünde bulundurularak, kullanacağı ilaçlara karar verilir ve tedaviye başlatılır. İleride Yapay Zeka Robot Teknolojisi sayesinde, ameliyatların %100'e yakın bir başarı oranı ile gerçekleşmesi beklenmektedir (Koyuncugil & Özgülbaş, 2009).

Kişiye özel tedaviler, kişinin gen haritasının analiz edilmesiyle, kişiye özel tedaviler gerçekleştirilmektedir. Yapay Zeka ile bu en doğru şekilde ve daha hızlı gerçekleşmektedir. Hamilelik tedavileri, tüp bebek tedavilerinde, doğru sperm ve yumurtanın birleşmesine karar vermede Yapay Zeka Algoritmaları ve Veri Madenciliği kullanılmaktadır. Daha önceden denenmiş ve bu konuda başarı elde edilmiştir. Bu teknolojinin adı tıpta Preimplantasyon Genetik Tanı (PGT) teknolojisi olarak geçmektedir (Koyuncugil & Özgülbaş, 2009).

Gen analizleri, kişinin gen haritasının çıkarılması olası hastalıklara yakalanma riskinin tespit edilmesinde yardımcı olmaktadır. Erken teşhis ya da genlerin değiştirilmesi birden fazla hastalığın önlenmesinde, hastalığın hiç yaşanmamasında önemli rol oynamaktadır. Örneğin; Otizm hastalığının ana karnında tespit edildikten sonra genlerin değişmesi, Yapay Zeka ile yapılması hedeflenmektedir. Otizm hastalığına sahip bireylerin tedavisinde de Yapay Zeka'dan yararlanılması hedeflenmektedir (Demirhan vd., 2010).

Otomotiv sektöründe akıllı araç sisteminde Yapay Zeka' dan yararlanılmaktadır. Akıllı araç sistemi ile araçlarda Yapay Zeka Teknolojisi hayata geçmiştir. Birden fazla araç firmaları entegre olarak bu sistemi kullanmakta ve gelecekte araçsız sürücüler, zekaya sahip araçlar öngörülmektedir. Şuan kullanılan teknolojide araçlar kendi kendilerini park etme özelliğine sahiptir. Gelecekte bunu araçların kendi kendini kullanma, insanları araçların yönlendirmesi, en popüler mekanları bulma, en doğru tercihi verme gibi özelliklerde takip edeceği ön görülmektedir. Örneğin; Tesla firması araç çağırma sistemini şuan kullanmaktadır. Mercedes firması sürücüsüz araç sistemini şuan sadece 1 dakika olmak üzere kullanmaktadır (Yetim, 2015).

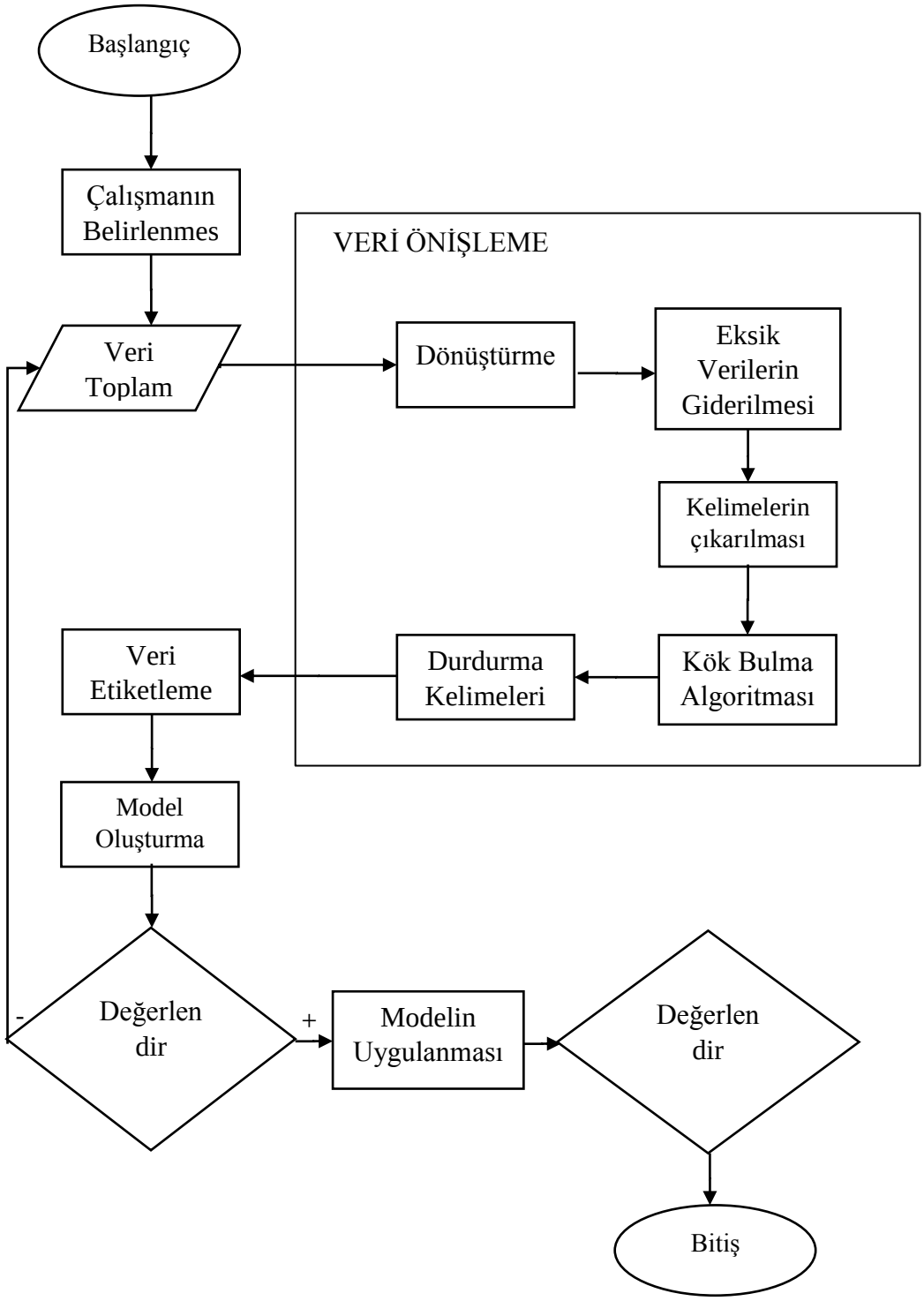
Eğitim alanında, meslek seçimi için öğrencilerin kişilik ve beceri analizlerini yaparak hangi meslek gurubunda başarılı olacağını ön görebilir. Bu şekilde kişilerin meslek

seçiminde hata yapmasının önüne geçer. Bununla birlikte işini mükemmele yakın yapan bireyler ortaya çıkar. Öğrencilerin bilgi ve becerilerinin tespitinden sonra becerilerini daha da geliştirecek testler hazırlamak için Yapay Zeka kullanılmaktadır. Yapay Zeka ile eğitimciler için testler yapılarak, öğrenciler için en iyi eğitmen seçimi yapılabilir. Yapay Zeka tabanlı uygulamalardan, bilgiye direkt erişim sağlanabilmektedir. Örneğin; Google, Siri vs. Yenilenme: Eğitmen ve öğrenciyi geliştiren Yapay Zeka her zaman kendi kendini de geliştirmektedir. Her tecrübesini kaydedip, yeni analizleri daha tutarlı gerçekleştirmektedir (Kazu & Özdemir, 2009).

Yapay Zeka'ya sahip robotların gelişmesiyle, askeri alanda büyük gelişim sağlanmaktadır. Özellikle savaşlarda, yer tespiti yapımı, sıfır insan kaybı, savunma da Yapay Zeka büyük önem taşımaktadır. Hemen hemen her ülke Yapay Zeka Teknolojisini kendi askeri bünyesinde kullanmak için çalışmalar yapmaktadır. Geleceğin kazananlarını Yapay Zeka'ya sahiplik belirleyeceği düşünülmektedir (Uçar, 2012). Günümüzde danseden robotlar, insan hareketlerini birebir takliteden robotlar prototip aşamasındadır. İleride ordunun tamamının oluşturulması öngörülmektedir.

D. Metin Madenciliği

Metin madenciliği metinlerden oluşan veri setlerinin analiz edilmesidir. Metin madenciliğide sınıflandırma işlemi yapılırken, karşılaşılan en büyük sorun metinlerin hangi sınıfa dahil olması gerektiğidir. Bu sorun veri madenciliğinden yararlanılarak giderilmiştir (Tantuğ, 2012). Veri madenciliği ve metin madenciliği arasındaki bağ şu şekilde tanımlanmaktadır; veri madenciliği yapısal verileri analiz etmek için kullanılırken, metin madenciliği yapısal olmayan veri analizinde kullanılmaktadır. Yapısal olmayan veriler, yapısal veri haline dönüştürülerek, makine öğrenmesinde yapay zeka algoritmaları aracılığıyla kullanılmaktadır (Yıldız & Ağdeniz, 2018). Metin madenciliği sınıflandırma adımları şekil 3'te verilmiştir. İlk olarak veri toplanır, veri önişleme adımları uygulandıktan sonra model oluşturulur, sonuç değerlendirilir ve oluşturulan model uygulanır. Eğer uygulama sonucunda uygulanan modelin sonucu iyi ise sonuçlar değerlendirilir, değil ise model baştan kurulur.



Şekil 3 Metin Madenciliği Sınıflandırma Adımları

Metin madenciliğinde veri ön işleme aşamasında, mevcut verilerin derleyici ortamına alınması için dönüştürme yapılması gerekmektedir. Dönüştürme işlemi kullanılan derleyicinin desteklediği formatlara göre değişiklik göstermektedir. Örneğin: Weka arff ve csv formatında dosyaları desteklemektedir. Bu sebeple bu formatta oluşturulan verilerin dönüştürme işlemini yapmasına gerek yoktur. Veri seti içerisinde eksik veriler bulunabilir. Analiz yapmadan önce eksik verilerin giderilmesi gerekmektedir. Eksik verilerin giderilmesi için çeşitli yaklaşımlar bulunmaktadır. Veri setine en uygun yaklaşım seçilerek uygulanır. Derleyiciler veri seti içerisinde eksik veriler mevcutken analiz yapmamaktadır. Veri setlerinde analizi etkilemeyecek değere sahip veriler bulunmaktadır. Bu verilerin analiz yapılmadan önce temizlenmesi gerekmektedir. Temizle işleminde özelliklerin analize etkisi ölçülür ve analizi etkilemeyecek kadar düşük değerlere sahip veriler, kolonlar veri seti içerisinden çıkartılarak analiz yapılır. Kullanılan yazım içerisinde bulunan çeşitli özellikler yardımıyla yapılabilir. Kök bulma algoritması, N-Gram yöntemi, Durdurma kelimeleri veri setinin temizlenmesi aşamasında yardım olmaktadır (Vijayarani vd., 2014; Jivani, 2011; Srividhya & Anitha, 2010). Kök bulma algoritmalarında, kelimelerin tekrarlı olarak aynı metin içerisinde analiz yapılmasını engellemektir. Bu ayrım yapıldıktan sonra analiz sonuçları daha doğru olacaktır. Eğer aynı köke sahip kelimeler bir arada analiz içerisinde kullanılırsa, iki kelimeyi farklı kelimeler olarak makine anlamaktadır. İki kelimenin aynı olmadığının makineye gösterilmesi gerekmektedir. Bunlarda Kök bulma algoritmaları ve durdurma kelimeleri aracılığı ile gerçekleşmektedir. N-gram yönteminde, N adet N' li parçalara kelimeleri ayırmaktadır. Bu şekilde kelimelerin köklerine inmektedir. Durdurma kelimeleri ise edat, bağlaç, zamir, sayılar, tarihler ve filleri cümleden çıkartmak için kullanılmaktadır. Bu kelimeler sınıflandırmayı etkilemeyecektir. Bu kelimelerin metinlerde çıkartılması analizin daha doğru yapılmasını sağlamaktadır.

Veri seti test ve eğitim kümesi olarak birlikte kullanılabilmektedir. Eğitim kümesi ile makine eğitilmesi gerçekleştirilmektedir. Test kümesi ile eğitilen makine test edilmektedir. Bununla ilgili birden fazla yöntem bulunmaktadır. Veri seti iki parçaya bölünerek ayrı ayrı eğitim ve test kümeleri de olarak kullanılabilmektedir. Bu işlemler analiz sonucunu değiştirmemektedir.

Yapay Zeka Algoritmaları kullanılarak verilerin makineye öğretilmesi gerçekleştirilmektedir. Veri seti makineye öğretilirken, öğrenme aşamasında

istenilmeyen durumlarla karşılaşılabilmektedir. Bu durumların önlenmesi gerekmektedir. Veri setinde ezberleme istenilmeyen bir durumdur. Bu sebeple makineye öğretirken, sınıflandırmayı etkilemeyecek veriler temizlenir. Makinenin ezberlemesine yol açacak veriler ise değerlendirmeye alınmaz isim, cinsiyet gibi değerler makinede ezberlemeye yol açmaktadır.

E. Sınıflandırma Amaçlı Yapay Zeka Algoritmaları

Sınıflandırma yapabilmek için birden fazla Yapay Zeka Algoritması bulunmaktadır. Makine öğrenmesi Yapay Zeka Algoritmaları kullanılarak gerçekleştirilmektedir. Bu algoritmalar içerisinde metin madenciliği için daha fazla tercih edilen algoritmalar kullanılmıştır. Bunlar, ZeroR, Naif Bayes, Rastgele Orman Algoritmalarıdır. Önemli olan mevcut veri seti için en uygun algoritmanın seçilmesidir.

1. ZeroR algoritması

Diğer sınıflandırma algoritmaları arasında daha ilkel olan bir algoritmadır. Çalışma mantığı basittir. Eğitim setinde frekansı en yüksek olanı seçer ve test verilerinin hepsini o sınıfa ait kabul eder (Nasa & Suman, 2012).

Örneğin: Bir sınıfta 40 erkek, 30 kız öğrenci vardır. Yeni gelen öğrencinin cinsiyetinin tahmin edilmesi gibi bir problemde, ZeroR algoritması erkek sayısı fazla olduğu için yeni gelen bütün öğrencileri erkek kabul edecektir.

2. Naif Bayes algoritması

Bayes teoremini baz alarak çalışan bir algoritmadır. Olasılık hesaplamalarına göre çalışır. Her özelliğin sonuca etkisi üzerine olasılık değerlerinin hesaplanması ile gerçekleşir. Verinin mevcut sınıflardan hangisine ait olma olasılığını hesaplar. Özelliklerin önem derecesini hepsinde eşit almaktadır. Bu şekilde daha doğru sonuçlara ulaşmaktadır. Bütün özellikler birbirinden bağımsız olarak kabul edilmektedir. Özet olarak olasılık değeri en yüksek olan kararın değerlendirilip sonuçlandırılmasıdır. Bayes teoreminde olduğu gibi koşullu olasılıktan faydalanır ve test veri setinde bulunan üyelerin hangi sınıfa ait olduğunu bulmaya çalışır. Metin madenciliğinde diğer algoritmalara göre daha iyi sonuçlar vermektedir. Başarı oranı daha yüksektir (Kalaycı, 2018; Karakoyun & Hacıbeyoğlu, 2014).

Bayes Teoremi Formülü (Çalış vd., 2013).

$$P(A|B) = (P(B|A) * P(A))/P(B) \quad \textbf{(1)}$$

P(A): A Olayının Bağımsız Olasılığı

P(B): B Olayının Bağımsız Olasılığı

P(B|A): A Olayı Olduğunda B Olayının Olma Olasılığı

P(A|B): B Olayı Olduğunda A Olayının Olma Olasılığı

Örneğin

Çizelge 1'den yararlanılarak, 30 yaşında 5 yıl tecrübeye sahip ve 3000tl maaş alan bir kişi hangi birimde çalışmaktadır.

Çizelge 1 Naif Bayes Algoritması Örneği Veri Çizelgesi

Departman	**Maaş(TL)**	**Yaş**	**Tecrübe(yıl)**
Yazılım	3000	26	4
Muhasebe	1500	22	2
Yazılım	5000	30	9
Muhasebe	2000	30	7
Muhasebe	500	18	3
Yazılım	2000	20	2
Yazılım	7000	29	5
Muhasebe	6000	45	15

Çizelge 2 Naif Bayes Algoritması Örneği Ortalama Değerleri

Departman	Maaş(TL)	Yaş	Tecrübe(yıl)
Muhasebe	2500	28,75	6,75
Yazılım	4250	26,25	5

Çizelge 3 Naif Bayes Algoritması Örneği Varyans Değerleri

Departman	Maaş(TL)	Yaş	Tecrübe(yıl)
Muhasebe	58	142,25	34,91
Yazılım	50	20,25	8,66

Olasılık formüllerini sınfılar için yazarsak;

$$\text{BD (yazılım)} = \frac{\text{p(yazılım)} * \text{p}\left(\frac{\text{maaş}}{\text{yazılım}}\right) * \text{p}\left(\frac{\text{yaş}}{\text{yazılım}}\right) * \text{p}\left(\frac{\text{tecrübe}}{\text{yazılım}}\right)}{\text{Normalleştirme}} \qquad (\mathbf{2})$$

Aynı formül muhasebe için yazılırsa;

$$\text{BD (Mh)} = \frac{\text{p(Mh)} * \text{p}\left(\frac{\text{maaş}}{\text{Mh}}\right) * \text{p}\left(\frac{\text{yaş}}{\text{Mh}}\right) * \text{p}\left(\frac{\text{tecrübe}}{\text{Mh}}\right)}{\text{Normalleştirme}} \qquad (\mathbf{3})$$

Normalleştirme değeri;

$$\text{N} = \text{P(Mh) p(maaş | Mh) p (Yaş | Mh) p (iş tecrübesi | Mh)} + \text{P(Yazılım) p(maaş | yazılım) p (Yaş | yazılım) p}\left(\text{iş tecrübesi} \middle| \text{yazılım}\right) \qquad (\mathbf{4})$$

BD: Beklenen Değer

N: Normalleştirme

Mh: Muhasebe

Not: Eğer iki sınıf varsa, saçim sadece bu ikisi arasında olduğunda normalleştirme göz ardı edilebilir.

$$\mathrm{P}(\text{yazılım}) = \frac{4}{8} = 0.5 \quad (\mathbf{5})$$

$$\mathrm{P}(\text{muhasebe}) = \frac{4}{8} = 0.5 \quad (\mathbf{6})$$

Koşullu olasılık değerlerinin hesaplanması(Gauss Dağılımı);

$$\mathrm{p}\left(\frac{\text{maaş}}{?}\right) = \frac{1}{\sqrt{2 * \pi * \sigma^2}} * \exp\left(\frac{-(\mathrm{x} - \mu)^2}{2 * \sigma^2}\right) \quad (\mathbf{7})$$

Hesaplama yapıldığında;

$$\mathrm{p}\left(\frac{\text{maaş}}{\text{yazılım}}\right) = \frac{1}{\sqrt{2 * \pi * 4.91\mathrm{E}6^2}} * \exp\left(\frac{-(3000 - 4250)^2}{24.91\mathrm{E}6^2}\right) = 6.84\mathrm{E} - 8 \quad (\mathbf{8})$$

$$\mathrm{p}\left(\frac{\text{maaş}}{\text{muhasebe}}\right) = \frac{1}{\sqrt{2 * \pi * 5.83\mathrm{E}6^2}} * \exp\left(\frac{-(3000 - 2500)^2}{25.83\mathrm{E}6^2}\right) = 8{,}11\mathrm{E} - 8 \quad (\mathbf{9})$$

Diğer kriterler için koşullu olasılıkların hepsi hesaplandığında çizelge 4 elde edilir.

Çizelge 4 Naif Bayes Algoritması Örneği Koşullu Olasılık Değerleri

	Maaş	**Yaş**	**Tecrübe**
Muhasebe	6,84074E-08	0,002805118	0,011414151
Yazılım	8,11614E-08	0,019705849	0,046043474

Naif Bayes' e göre hesaplama yapıldığında;

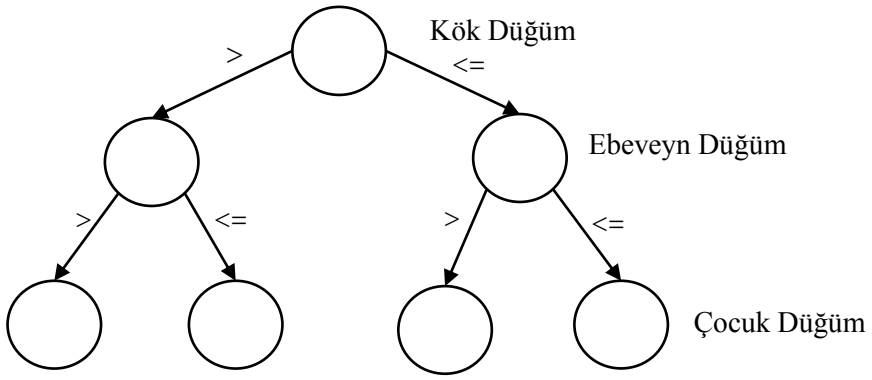
$$\mathrm{BD}(\text{yazılım}) = \frac{0.5 * 6{,}84\mathrm{E} - 8 * 0.0028 * 0.0114}{\text{Normalleştirme}} = 1.9\mathrm{E} - 12 \quad (\mathbf{10})$$

$$\mathrm{BD(muhasebe)} = \frac{0.5 * 8{,}11\mathrm{E} - 8 * 0.0019 * 0.046}{\mathrm{Normalleştirme}} = 3{,}68\mathrm{E} - 11 \qquad (\mathbf{11})$$

Sonuç olarak, 30 yaşında 5 yıl tecrübeye sahip ve 3000tl maaş alan bir kişi muhasebede çalışıcaktır (Şeker, 2013).

3. Karar Ağacı algoritması

Makine öğrenmesinde, regresyon ve sınıflandırma amaçlı kullanılmaktadır. Amaç, karar vermektir. Günlük yaşantımızda birden fazla işi art arda sıralarken dahi karar ağaçlarından yararlanırız. Günlük yaşantımızdaki iş planının ağaç seklinde ifade edilmesine Karar Ağacı denir (Ulusoy, 2013). Karar verme tabanlı bir algoritmadır. Veri içerisindeki özellikler üzerinden karar ağacını oluşturur. Karar ağaçları ağaç şeklinde oluşturulmaktadır. Ağacın yapısı kök düğüm, ebeveyn düğüm ve çocuk düğümden oluşmaktadır. Kök düğüm ilk düğüme karşılık gelmektedir. Kök düğüm aynı zamanda bütün düğümlerin ebeveyni olmaktadır. Kök düğümün alt düğümleri çocuk düğüm olarak adlandırılır. Her ebeveyn düğüm aynı zamanda bir üst düğüme göre çocuk düğüm olmaktadır. Yapı yukarıdan aşağı şeklinde çalışmaktadır. Kök düğümden dallanma başlar ve çocuk düğüme kadar devam eder (Tuncer, 2018). Şekil 4' te örnek bir karar ağacı verilmiştir.



Şekil 4 Karar Ağacı

Belli bir kurala göre ağaç oluşturulur. Bir duruma eşit olması '=', küçük olması '<', büyük olması '>', büyük eşit olması '>=', küçük eşit olması '<=' durumları kural olarak karar ağacı oluşturulurken koyulmaktadır (Uysal, 2014). Veri seti içerisinde eğitim ve test kümeleri bulunmaktadır. Eğitim kümesi kullanılarak karar ağacı oluşturulur ve test kümesi kullanılarak test edilir. Test kümesindeki elemanlar karar

ağacı üzerindeki dallanmayı devam ettirir. İlk eleman kök düğüme gider, daha sonra ilgili kurala göre sağ ya da sol düğümden gideceğine karar verir. Bu şekilde ağaç üzerinde çocuk düğümün alt kısmına yerleşene kadar bu adımlar devam ettirilir. En son adımda ise test kümesindeki eleman ağaca çocuk düğüm olarak yerleşir ve ağaç büyümüş olur (Pala, 2013).

Özellik seçimi, karar ağacının hangi kretere göre dallanması gerektiğini belirtmektedir. Karar ağacı özellik seçimi birden fazla yöntemle yapılmaktadır. En fazla kullanılan yöntemler aşağıda verilmiştir.

Bilgi kazancı, 'Shannon Bilgi Teorisi' temel almaktadır. Entropi tabanlıdır ve hesaplanırken entropi değerlerini kullanır. Entropi, sistem içerisindeki düzensizliğin tanımıdır. Örneğin: Yazı tura atılmasında, hileli bir para kullanıldığını varsayın. Sürekli para yazı geliyorsa ortamın entropisi 0 dır. Yani hiç düzensizlik yoktur. Başka bir ihtimal yoktur. Düzeni bozan hiç bir olasılık olmadığı için entropi 0 olmuştur (Çetinkaya, 1981). Aşağıdaki formülle hesaplanmaktadır.

$$H(S) = -\sum_{k=1}^{n} p_k \log_2(p_k) \qquad (\mathbf{12})$$

S: Kaynak

m_1, $m_{2,\dots}$ m_n olmak üzere n adet veri

m_k:p_k (m_k verisi için p_k olasılık değerini ifade eder)

p_1,p_2,..... p_k: olasılık değerleri

H(S): Kaynağın sahip olduğu entropi değeri

Bilgi kazancı düzensizliğin olmadığı durumların toplamıdır. 0 ile 1 arasında değer alır. Sınıflandırma içerisinde, bilgi kazancı sonuçların değerini ölçer. Sınıflandırmaya karar verilen özellikler sınıftan ne kadar bağımsız ise, bilgi kazancı değeri o kadar düşük çıkar. Her bir özelliğin bütün very seti üzerindeki kazanımı ölçülmektedir. Aşağıdaki formülle hesaplanmaktadır.

$$\text{Bilgi}_x(p) = \sum_{k=1}^{n} \left((|p_k|/p) * \text{Bilgi}(p_k) \right) \qquad (\mathbf{13})$$

Yukarıda her k özelliği için bilgi değeri hesaplanmaktadır.

$$\text{Kazanç}(\text{Özellik Y}) = \text{Bilgi}(\text{p}) - \text{Bilgi}_{\text{x}}(\text{p}) \qquad (\mathbf{14})$$

Yukarıda herhangi k değerinin kazancı hesaplanmaktadır (Odabaş 2017, 21).

Gini indensi, karar ağacı oluşturmak ve sınıflandırma yapabilmek için kullanılan indekstir. Diğer adı CRT Algoritmasıdır. İkili dallanma yaparak ağacı oluşturur. Bu algoritmada kök düğüm seçimi önemlidir. Eğer kök düğüm doğru seçilmez ise ağaç yanlış oluşturulur.

Kök düğüm seçimi: Bütün özelliklerin gini değerlerine göre kök düğüm seçimi yapılır. Aşağıdaki formülle gini indeks hesaplanmaktadır.

Gini-sol değer formülü:

$$\text{Gini}_{\text{sl}} = 1 - \sum_{\text{k=1}}^{\text{b}} \left[\frac{\text{Sl}_{\text{k}}}{|\text{A}_{\text{sl}}|}\right]^2 \qquad (\mathbf{15})$$

Gini-sağ değer formülü:

$$\text{Gini}_{\text{sg}} = 1 - \sum_{\text{k=1}}^{\text{b}} \left[\frac{\text{Sg}_{\text{k}}}{|\text{A}_{\text{sg}}|}\right]^2 \qquad (\mathbf{16})$$

Gini formülü:

$$\text{Gini}_{\text{j}} = \frac{1}{\text{n}}\left(|\text{A}_{\text{sl}}|\text{Gini}_{\text{sl}} + |\text{A}_{\text{sg}}|\text{Gini}_{\text{sg}}\right) \qquad (\mathbf{17})$$

Sl: Sol taraftaki dallanma sayısı (k kategorisinde)

Sg: Sağ taraftaki dallanma sayısı (k kategorisinde)

A_{sl}: Sol taraftaki dallanma sayısı

A_{sg}: Sağ taraftaki dallanma sayısı

b: Sınıf sayısı

A: Düğümdeki özelliklerin sayısı

Gini değerleri hesaplandıktan sonra düğüm seçilir. Düğüm seçilirken gini değeri en küçük olan seçilmektedir. Bu işlemden sonra geriye kalan diğer özellikler için tekrar gini değeri hesaplanır ve dallanma için seçim yapılır (Adak & Yurtay, 2013).

Kazanç oranı, bilgi kazancının normalize edilmiş halidir. Aşağıdaki formülle hesaplanır:

$$\text{Kazaç Oranı}(a_i, S) = \frac{\text{Bilgi Kazancı}(a_i, S)}{\text{Entropi}(a_i, S)} \qquad (\mathbf{18})$$

Tanıma göre entropinin sıfır olduğu yerlerde kazanç oranı tanımsızdır. Bilgi kazancı ne kadar fazla ise sonuç o kadar olumlu olacaktır. Bilgi kazancı hesaplandıktan sonar en iyi performansa sahip özellikler seçilir ve onlar kullanılır.Buda karmaşıklığı ve doğruluğu olumlu yönde etkileyecektir (Kuzey, 2012; Çimenli, 2015).

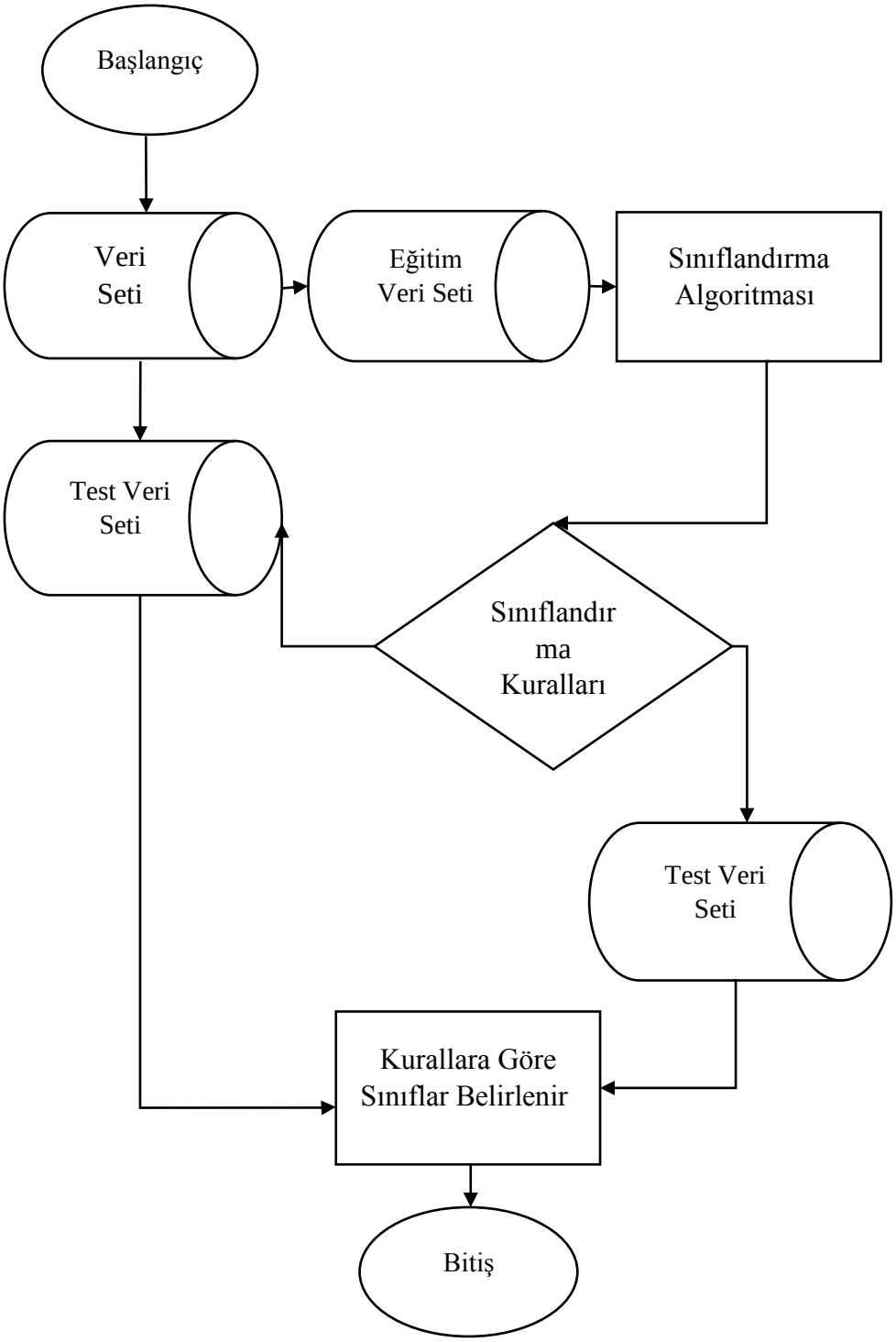
Karar Ağacında budama işlemi, ağacın derinliğine karar vermeye denir. Ağaç oluşturulduktan sonra sınıflandırma aşamasında budama işlemi yapılmaktadır. Budama işleminin amacı, sınıflandırma yaparken oluşabilecek hataları minimum seviyeye indirmektir. Eğer işlem sınıflandırmayı olumlu yönde etkileyecek ise bir dal budanır. Birden fazla budama yöntemi vardır. Bunlardan bazıları aşağıdan-yukarı, yukarıdan-aşağı, durdurma kelimeleri kullanılarak yapılan budamadır (Özdemir, 2014). Karar Ağaçlarının avantajları ve dezavantajları çizelge 5'te verilmiştir.

Çizelge 5 Karar Ağaçları Avantaj Ve Deazavantajları

Avantajlar	Dezavantajlar
Maliyeti azdır.	Ağaç boyutu büyüdükçe, takip etmek zorlaşır.
Kullanım alanı fazladır. Tıp, turizm, askeriye, banka vb.	Budama aşaması iyi yapılmazsa, yanlış sonuçlar doğuracaktır.
Hem sayısal hemde kategorik veriler üzerinde işlem yapabilir.	Öğrenme aşamasında yaşanan küçük sapmalar dahi, bütün ağacı etkileyebilir. Bu sebeple ağaç en iyi özelliği seçemeyebilir.
Sınıflandırma ve regresyon işlemleri yapabilir.	
Gösterimi ve anlaşılması kolaydır.	
Eksik veriler üzerinde iyi performans ile çalışır	

Kaynak: Hesarı, 2018; Sayıcı, 2013.

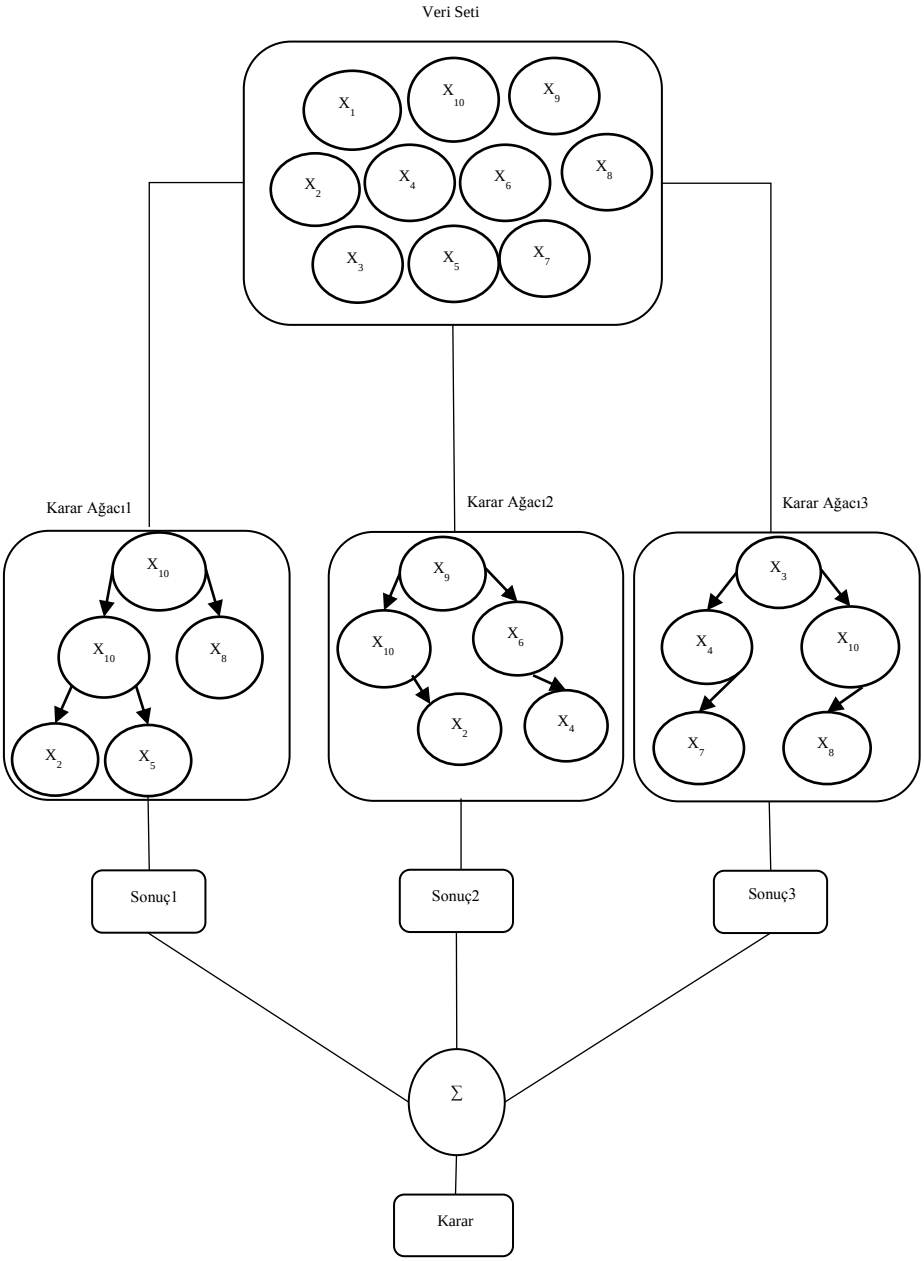
Karar Ağacı algoritmaları ile sınıflandırma işlemi kolay ve anlaşılır bir şekilde yapılmaktadır. Karar ağaçlarının eksiklerini girebilmek için, karar ağaçlarının daha gelişmiş şekilleridir. Karar ağaçları algoritmaları farklı özellik seçimlerini baz alarak ağaçları oluştururlar. Literatürde birden fazla karar ağacı algoritması mevcuttur. Bunlardan bazıları: Rastgele Orman, ID3, C4.5, C&RT, RepTree, CHAID dır (Ay, 2009). Karar ağacı algoritmasının çalışma adımları şekil 5'te verilmiştir.



Şekil 5 Karar Ağacı Algoritması Çalışma Adımları

Veri seti eğitim ve test verisi olmak üzereye ikiye ayrılır. Eğitim verisi üzerinde sınıflandırma algoritmaları çalıştırılarak makine öğrenmesi gerçekleştirilir. Bunun sonucunda sınıflandırma kuralları oluşur. Test veri seti bu kurallara göre test edilir ve uygun sınıfa yerleşir.

Rastgele Orman Karar Ağacı Algoritması ağaçlar topluluğu olarakta tanımlanmaktadır. Regresyon ve sınıflandırma için kullanılabilir. Rastgele Orman (RO) Karar Ağacı Algoritmasında bir tane karar ağacı yoktur. Karar ağacı n adettir ve n adet karar ağacı rastgele oluşturulur. N değeri değiştirilebirlir ve n değerini kullanıcı belirleyebilir. Gini indeksini baz alarak ağaçları oluşturur. Ağaçlar oluşturulurken sadece eğitim setleri kullanılır. Test veri seti test aşamasında kullanılmaktadır. Karar ağaçlarının sonuçları tek tek ele alınır ve hangi sonuçtan daha fazla ise o sınıfa dahil edilir (Hazım, 2018). Şekil 6'da karar ağacının uygulama adımları verilmiştir.



Şekil 6 Rastgele Orman Karar Ağacı Algoritması Uygulama Adımları

Şekil 6 ‘daki uygulama adımlarına göre, veri seti topluluğu içerisinden üç adet karar ağacı oluşturulmuştur. Karar ağaçları rastgele oluşmuştur ve içerisindeki elemanlar sadece bir defa karar ağacına yerleşmemiştir. Rastgele sayıda ve rastgele içerikte karar ağaçları oluşturulmuştur. Oluşan alt karar ağaçları sonuçunda, sonuç toplama işlemi uygulanmıştır. Algoritmanın matığına göre en çok oyu alan sonuç en doğru karardır. Sonuçların eşit sayıda çıkması sonucunda, hangi sonucun alındığı önemli değildir. RO, veri setinin tek bir ağaç yerine birden fazla ağaç oluşturması, algoritmanın daha iyi çalışmasına sebep olmaktadır. Bir olay yada duruma bir ağacın değilde birden fazla ağacın karar verip ortak kararın uygulanmasını sağlamaktadır. Parçalar bir araya gelerek daha iyi sonuç üretmektedir. Sonuç parçaların ortak kararıdır. Parçalardan bazıları yanlış olsada, diğer parçalar doğru olacaktır. Bütüne bakıldığında karar doğru olacağından, yanlış kararlar sonucu daha az etkileyecektir (Pervan, 2019). Buradaki en büyük sorun varyansı dengelemektir. Eğitim veri setlerindeki ufak değişiklikler bile sonucu ciddi şekilde etkileyecektir. Eğitim seti alt ağaçlara bölünürken çantalama algoritmasını kullanmaktadır.

Çantalama Algoritmasını RO’da kullanılmasının amacı, varyansı azaltarak ağaçların oluşturulmasıdır. Mevcut veri setini kullanarak n adet veri setleri üretir. Ürettiği yeni veri setleri ile n adet ağacı eğitir. Geri beslemelidir, bu sebeple bir eleman birden fazla yerde kullanılabilir. Veri setleri seçimi rastgele yapılır (Aslan, 2016).

Rastgele Orman Karar Ağacı Algoritmasının avantajları ve dezavantajları çizelge 6’da verilmiştir.

Çizelge 6 Rastgele Orman Karar Ağacı Avantaj Ve Deazavantajları

Avantajları	Dezavantajları
Regresyon ve sınıflandırma işlemlerinde kullanılmaktadır.	Maliyeti karar ağacına göre yüksektir.
Topluluk olarak karar verdiği için, doğru karar verme olasılığı diğer ağaçlardan daha iyidir. Özellikle sınıflandırma işleminde diğer ağaçlara göre daha iyi çalışmaktadır.	Aykırı değerler olmayan verilerde daha iyi çalışır.
Topluluktaki ağaçların karar sonucunun yanlış çıkması için %50 den fazla ağacın yanlış karar vermesi gerekmektedir.	Ağaç topluluklarını kısa sürede eğitir fakat sonucu oluştururken zaman aşımı yaşamaktadır.
Ağaç topluluklarını kısa sürede eğitir.	
Aşırı veri yüklemesini engeller	

Kaynak: Ekelik, 2019; Bilgen, 2014.

F. Uluslararası Habercilik

Haber, belirli olayların belirli bir sırada meydana gelmesinden oluşmaktadır. Merak unsuru haber için, olay için önemlidir. Olayları bilme isteği haberin oluşmasını sağlamaktadır. Olay, ilgi çekebilecek nitelikte olan durum, hadisedir. Bir olayın haber değeri taşıması için, o olayın herkes tarafından öğrenilme isteğinin olması gerekir. Olayların haber olabilmesi için, gerçek olması gerekmemektedir. Gerçek olmayan durumlarda haber yapılmaktadır. Etik olarak yapılan bir haberin gerçek olması gerekmektedir. Bir olay haber olduktan sonra gerçeklik kazanmaz. Olaylar haberleri oluşturmasına ragmen, her haber bir olay değildir. Haber olayları iletmek amacı ile yapılmaktadır. Bu iki kavram birbirleriyle ilişkilidir. Haber doğrudan anlatılmaz olaylara bölünür, hikaye, öykü şeklinde anlatılmaktadır (İlhan, 2019).

Uluslararası haber, haber niteliği taşıyan bir olayın daha büyük kitlelere hitap etmesidir. Bunun için bazı özellikleri haberin taşıması gerekmektedir. Bu özellikler:

haberin bir ulusu ilgilendiren öğeler içermesi, ulus çapında tanınan birkişi hakkında olması, savaş, teror, dünya ile ilgili haberler (Akın, 2010). Ulusal haber ajanslarından bazıları Anadolu ajansı, magnum photos, reuters, suriye arap haber ajansı, uriminzokkiri dir (Girgin, 2002).

1. Haber alanları ve türleri

Birden fazla alanda haber yazılabilmektedir. Haberler alanları dört bölümden oluşmaktadır. Niteliklerine göre haberler, genel, karmaşık ve basit haberler olmak üzere 3 başlıkta incelenmektedir. Genel haberlerde gündelik olaylar anlatılmaktadır. Mülakat, yarışmalar, yıl dönümleri, toplantılar vb. gibi konular bu kategori içerisinde incelenmektedir. Basit haberlerde yorum içermeyen konular ele alınmaktadır. Ölüm, cinayet, intihar, yangın, doğa olayları, hava durumu vb. gibi konular bu kategori içerisinde incelenmektedir. Karmaşık haberler ise uzmanlık gerektiren alanlarda yapılan haberlerdir. Eğitim, bilim, adalet vb. gibi konular bu kategoride incelenmektedir.

Özel konulu haberler, Spor, dedikodu, sanat, edebiyat vb. gibi konular bu kategoride incelenmektedir.

İçeriklerine göre haberler, İçeriklerin siyasi, ekonomi vb. gibi konulardan, spor, dedikodu vb. gibi konuların ayrılmasını sağlamaktadır. İçeriğe göre haberler bu kategoride incelenmektedir.

Yapılarına göre haberler, ilan, röportaj, tanıtım vb. gibi konular bu kategori altında incelenmektedir (Taylan & Ünal, 2017; Çifçi, 2011).

2. Dilbilimsel

Dilbilim ve yapay zeka doğal dil işlemenin üst dalıdır. Doğal dil işleme analizi, dilde kullanılan ses ve metinlerin bilgisayar ortamına taşınmasıyla ilgilenen bir bilimdir. Doğal dil işleme analizinin amacı, doğal dildeki bütün yapıyı anlayan ve işleyen bir yapı oluşturmak ve bu yapıyı bilgisayara öğretmektir. Bu sayede bilgisayar sorulan sorulara yanıt verebilmekte ve insan ile iletişime geçebilmektedir. Anlama, yazma ve okuma gibi insana özgü işlevlerin sistemsel bir şekilde modellenmesi ulaşılmak istenen hedeflerden birisidir (Erduran, 2017). Örneğin: İos uygulamasın da kullanılan siri sistemidir. Bu sistem kullanıcının sorduğu sorulara cevap vermek ve kullanıcıya tavsiyelerde bulunmak üzerine tasarlanmıştır. Siri bir kişisel asistan yazılımıdır.

Siri'nin kaynak olarak birden fazla veri tabanını aynı anda kullanabilir (Ertemel & Gürdal, 2016).

Bilgisayar ortamında metin madenciliğinde, verilerin analiz edilip bilgisayarın anlayacağı bir yapıya getirilmesi gerekmektedir. Doğal dil işlenerek sayısallaştırılır ve bu sayede bilgisayarda işlemler yapılabilir hale getirilir. Doğal dil ne kadar sağdeleştirilirse, bilgisayar o kelimeleri daha iyi anlar ve daha iyi işleyebilir. Doğal dilin yapısal hale dönüştürülmesi gerekmektedir. Metin içerisinde mevcutta bulunan bilgilerin çıkarılması gerekmektedir. Bilgisayar bu sayede dili daha iyi anlayabilir. Bunun içinde metin birden fazla aşamadan geçmektedir.

Söz dizim analizi, kullanılan dili yapısına göre, sözcük olarak değil de cümle olarak nasıl sıralanması gerektiğini inceler. Sözcükleri kullanarak metin oluşturacak şekilde sıralamaktadır. Bir diğer değişle cümle içerisindeki kelimelerin birbiri ile olan bağlantısını incelemektedir. Kullanılan her dilin kendisine ait bir söz dizim kuralı bulunmaktadır. Cümleler kelime öbeklerine ayrılır. Her kelime öbeğinin kullanılan dilin kurallarına göre edat, fiil, sıfat vs. gibi sınıfları bulunmaktadır. Kelimeler, kelime öbeklerinden oluşur. Kelime öbekleri ise dil kurallarına göre bir yada birden fazla olabilmektedir. Kelime öbekleri tek başına kullanıldığındaki anlamı ile cümle içerisinde kullanıldığı anlamı farklı olabilir. Metin madenciliği içerisinde, kelimeler öbeklere ayrılarak, kökler, ekler vs. ayrılır. Ayrım sonucunda bir kelimenin birden fazla kullanılmasının önüne geçilmiş olunur (Aravi, 2014).

Morofoloji, kelimelerin en küçük parçalara ayrılmasını ifade eder. Kelimeyi çekim ve yapım eklerine kadar parçalar. Bu sayede kelimenin köküne inerek kelimeyi inceler. Aynı kökten gelen kelimeleri tespit edebilmektedir (Şeker, 2015).

Anlambilim analizi, dilin yapısını matematiksel yapıya çevirerek, bilgisayarın anlamasını sağlamaktadır (Delibaş, 2008). Kullanılan cümlenin anlamını anlayarak olumlu yada olumsuz şekilde sınıflandırmayı amaçlamaktadır. Anlambilimi, bununla birlikte karmaşık duygu durumlarını, metinde geçen ruh halini anlamaya çalışır. Aslında kelimeleri etiketler ve bu etiketlere göre cümlenin hangi anlam taşıdığını anlar. Bazı kelimelerin anlamları olumlu bazı kelimelerin ise olumsuzdur. Metin madenciliği bu kelimeleri baz alarak etiketleme yapmaktadır. Örneğin, iki gruba ayrılmış haber metinleri içerisinde, olumlu ve olumsuz sınıflar mevcutta var ise, metin madenciliği bu sınıfları baz alarak kendi olumlu olumsuz cümle tablosunu

oluşturur. Bu tabloya göre sınıflandırma yapar. Olumlu yada olumsuz cümleler geçmiyor ise üçüncü ihtimal düşünülerek nötr sınıfı oluşturur. Eğer makine öğrenmesi ile yapay zeka algoritmalarından yararlanılıyor ise, bu şekilde ayrım yapılmasına gerek yoktur. Eğitim seti içerisinde daha önceden kelimeleri makine öğreneceği için, bir cümle tablosu oluşturmaz. Eğitim seti makineyi eğitirken, yapay zeka algoritmaları yardımıyla bu kelimeleri makineye öğretmiş olmaktadır (Atan & Çınar, 2019).

Söylem analizi, kelimelerin cümle içerisinde kullanıldığı anlamın iyi bir şekilde analiz edilip, doğru anlamının anlaşılması gerekmektedir. Kelime hangi alanda yazılıyor ise o alanla ilgili anlamının, kurulan cümle içerisinden çıkarılması gerekmktedir. Örneğin; 'run' kelimesi ingilizcede koşmak anlamına gelmektedir. Fakat 'run computer' cümlesi içersinde 'run' kelimesi çalışmak anlamında kullanılmaktadır. Meslek içerisindeki anlamı kelime için önemlidir. Bazı kelimeler kişilerin kültürleri bilgi birikiminede göre değişiklik göstermektedir. Bu sebeple kelimeler farklı anlamlarıylada kullanılmaktadır. Buda metin madenciliğinin problemlerinden bir tanesidir (Şeker, 2015). Kelimelerin pratikte hangi anlama geldikleri de önemlidir. Bunun analiz edilip makineye tanıtılması gerekmektedir.

III. METODOLOJİ

A. Hipotezler

Bu çalışmada alan araştırması ardından geldiğimiz noktada, karar ağacı metin madenciliği yöntemlerinin uluslararası haber yazılarının içerik ve söyleme dayalı analizinde kullanabileceğimi ve bu haber yazılarının bu algoritmalarla tahmin edilebileceğini öneriyoruz.

İkinci önerimiz de farklı karar ağacı algoritmaları arasından en iyi tahminin Rastgele Orman Algoritmasıyle elde edilebileceği önerisidir. Bu iki hipotezi kanıtlama amacıyla araştırma yöntemlerimizi ele alıyoruz.

Proje içerisinde ZeroR, Naif Bayes, Rastgele Orman Algoritmaları kullanılacaktır. Bu algoritmaların kullanılmasındaki amaç, literatür taraması yapıldığında metin madenciliği için en iyi sonuçları veren algoritmalar olduğu görülmesidir.

Veri seti Uluslararası yayın yapan kanalın İnternet sitesi üzerinden alınmıştır. Veri seti Weka programı içerisinde analiz yapılacağı için, Weka programının desteklediği 'arff' formatına çevirilmiştir.

Algoritmalar incelendiğinde en iyi başarı oranı ve en iyi çalışma süresini verecek olan algoritmanın Rastgele Orman Algoritması olması beklenmektedir.

ZeroR algoritmasının RO ile Naif Bayes Algotimasından daha az bir başarı oranı vermesi beklenmektedir. Bunun sebebi veri seti içerisinde üç adet sınfıın bulunmasıdır. Eğer sınıf sayısı iki olsaydı. Başarı oranının daha fazla olması beklenecektir.

Naif Bayes Algoritmasının ZeroR'dan daha fazla, RO'dan daha az bir başarı ile çalışması beklenmektedir. Sıklıkla metin madenciliğinde bu algoritmanın seçildiği yapılan literatür taramasında görülmüştür.

En kötü sonucu ZeroR Algoritmasının vermesi beklenmektedir. Rastgele Orman ve Naif Bayes Algoritmaları daha iyi sonuç verecektir. Bunun sebebi ise algoritmalar

incelendiğinde Rastgele Orman ve Naif Bayes algoritmalarının, ZeroR algoritmasına göre daha gelişmiş olmasıdır.

En iyi sonuç veren algoritmanın para metreleri algoritmanın yapısı ve veri seti göz önünde bulundurularak değiştirilecektir. Parametre değişikliği sonucunda Algoritmanın mevcut veri seti için optimum seviyeye ulaşması beklenmektedir.

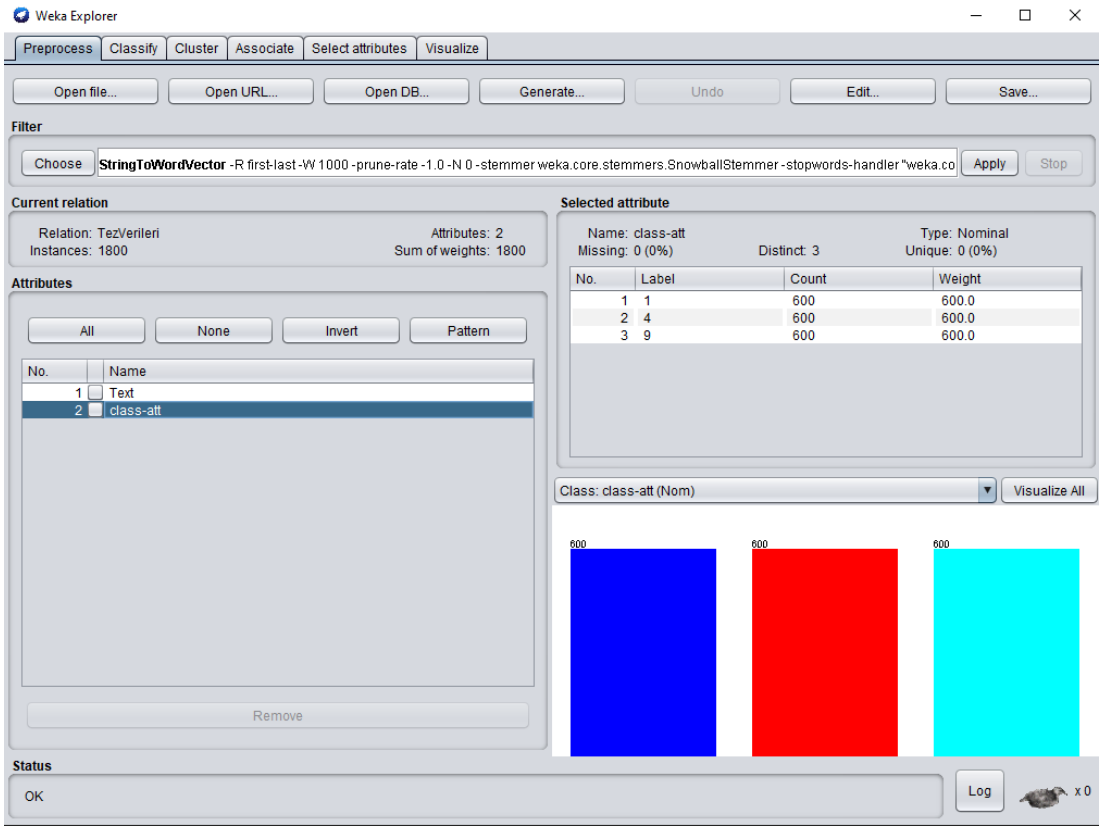
B. Veriyi Temin Etme ve Veri Ön İşleme Aşaması

Uluslararası bir kanaldan alınan veriler sınıflandırma yapmak amacı ile kullanılmıştır. Uluslararası kanalın web sitesinde yayınlanan haberler alınmıştır. Kanal ile görüşülüp, ilgili birimden haberler ‘csv’ formatında alınmıştır. Haberler 2019-2020 tarihleri arasında İnternet sitesi üzerinden yayınlanan haberlerdir. Veri seti içerisinde fazlaca gereksiz bilgiler bulunmaktadır. İlk olarak veri bu bilgilerden ayrıştırılmıştır. Veri seti içerisinde kategori ve metin kısımları kullanılacağından, bu kısımlar diğer gereksiz bilgilerden ayrıştırılmıştır. Bu ayrıştırma işlemi ‘Sublimetext’ adında bir metin editörü yardımıyla yapılmıştır. Verinin analizi Weka programı içerisinde yapılmıştır. Bu sebeple Weka programınında desteklediği arff dosya formatına veri dönüştürülmüştür. Veri seti içerisinde 1800 adet haber metni vardır. Çizelge 7’de metin kategorileri ve haber metin adetleri verilmiştir.

Çizelge 7 Veri Seti Sınıf Kategorileri

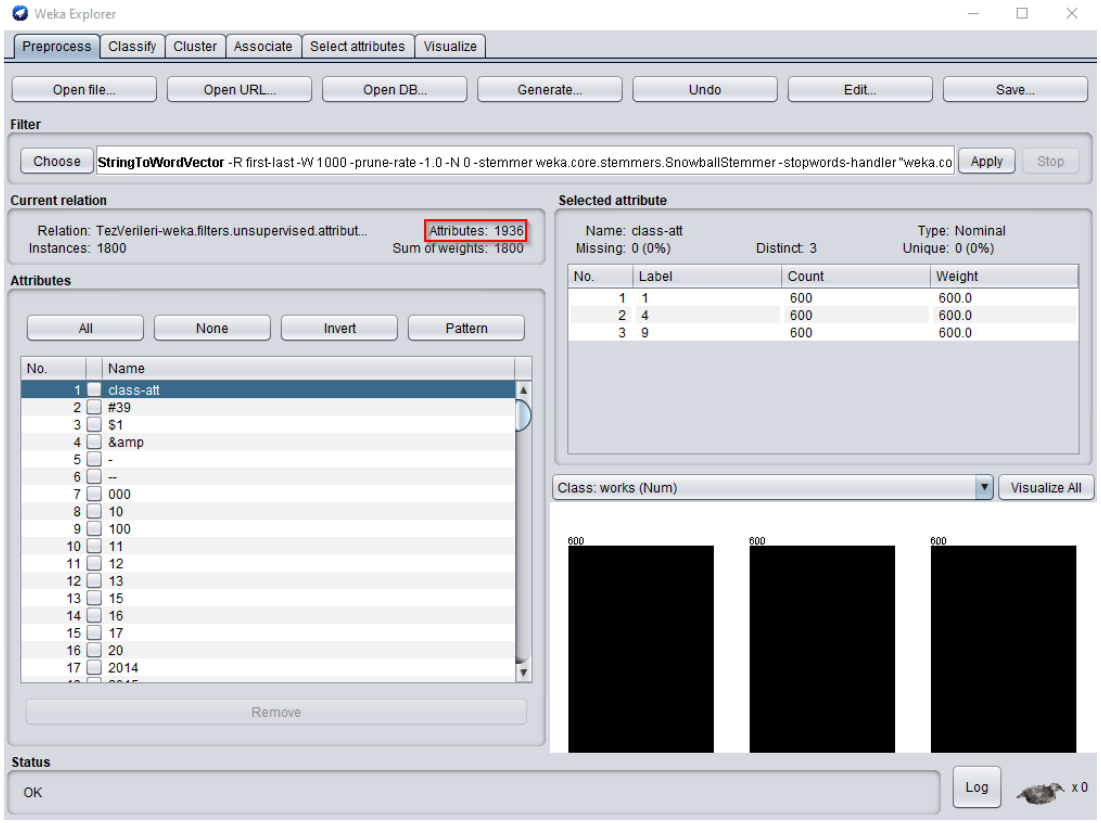
SINIFLANDIRMA BAŞLIĞI	HABER ADEDİ
International News	600
Sports News	600
Magazine News	600

Weka içerisindeki veri seti şekil 7'de verilmiştir.



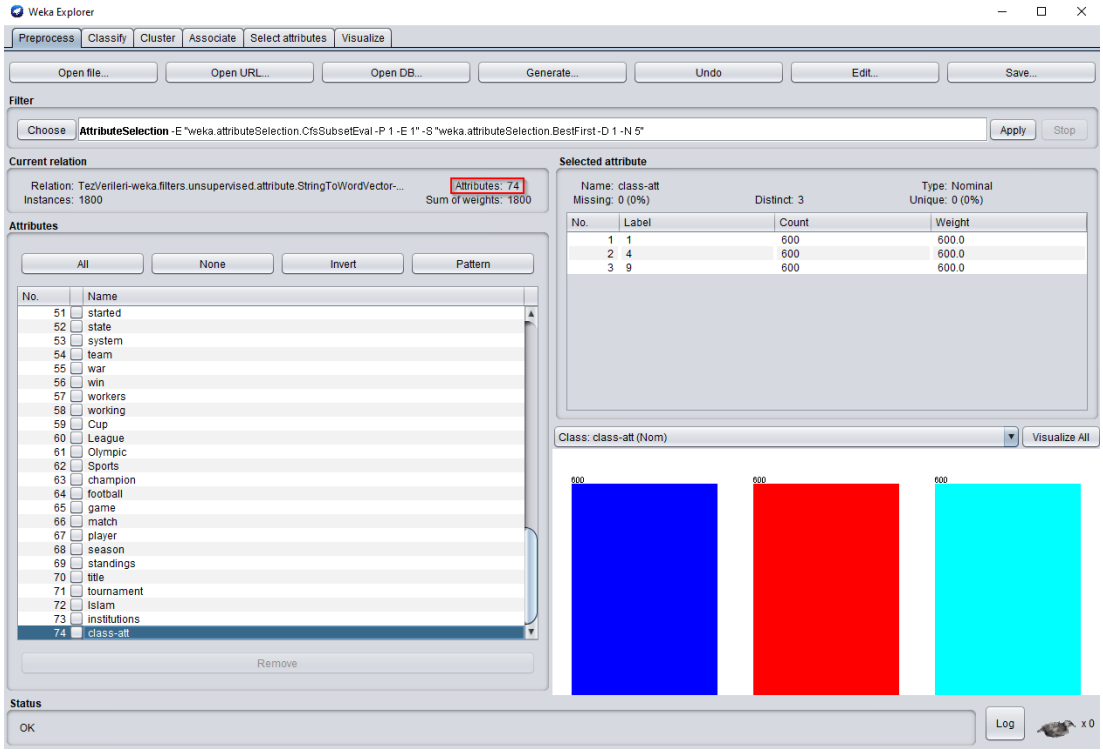
Şekil 7 Weka İçerisinde Veri Seti Görünümü

Başlangıçta 1935 adet özellik varken, ön işleme aşaması bittikten sonra 73 adet özellik kalmıştır. Şekil 8' de Weka içerisinde 1936 adet özelliğin gözükmesinin sebebi sınıfıda özellik sayısı olarak Weka'nın görmesidir.



Şekil 8 Ön İşleme Öncesi Özellik Sayısı

Ön işleme sonrası veri seti ön işleme aşaması sonrası özellikler şekil 9'da gösterilmektedir.



Şekil 9 Ön İşleme Sonrası Veri Seti

Veri dönüşümü, veriler toplandıktan sonra kullanılacak olan derleyicide çalışabilmesi için, derleyicinin desteklediği formata dönüştürülmesi gerekmektedir. Veriler toplanıp, Weka derleyicisi içerisine alınabilmesi için 'arff' dosya formatına çevrilmiştir.

Özellik çıkarımı, metin içerisinde kelime ayrımı işlemi yapılmıştır. Bütün kelimeler n'li gruplara ayrılarak işlemler yapılabilir. N=1 seçilerek, tekli kelime grupları şeklinde özellik çıkarımı yapılmıştır.

Gereksiz imleçleri metin içerisinden çıkartma, nokta, virgül, soru işaretler gibi noktalama işaretleri, boşluk karakteri, özel karakterler metin içerisinden çıkartılmıştır.

Kelime uzunluğu hesaplama, veri seti içerisinde kelime uzunlukları hesaplanmıştır. Minimum kelime uzunluğu üç'ten küçük olan kelimeler veri seti içerisinden çıkartılmıştır.

Büyük-küçük harf dönüşümleri, metin içerisinde aynı kelimelerin tekrar etmemesi için, bu dönüşümün yapılması gerekmektedir. Metindeki tüm veriler küçük harfe çevrilmiştir. Metin içerisinde aynı kelimelerin geçmesi ve aynı kelimelerin farklı kelimeler gibi algılanması önlenmiştir.

Durdurma kelimeleri, edat, bağlaç, zamir, sayılar, tarihler gibi sınıflandırmayı etkilemeyecek olan verilerin metin içerisinden çıkarılması gerekmektedir. Durdurma kelimesi olarak 630 adet İngilizce kelime hazırlanmıştır ve metine uygulanmıştır.

Kök Bulma Algoritması, kök bulma kullanılmasındaki amaç, aynı köke sahip, farklı ek alan kelimeler bir arada analiz içerisinde kullanılırsa makine, iki kelimeyi farklı kelimeler olarak algılamaktadır. İki kelimenin farklı kelimeler olduğunu makineye gösterilmesi gerekmektedir. Metin içerisinde aynı kökten gelen kelimeler bulunmuştur ve metin içerisinden çıkartılmıştır. Kartopu kök bulma algoritması kullanılmıştır. Bu aşamadan sonar veri seti analiz için uygun hale gelmiştir.

C. Yöntem Doğrulanması ve Yorumlanması

Metin madenciliğinde, mevcut veri seti üzerinde en iyi sonucu veren algoritma Rastgele Orman (RO) algoritmasıdır. Literatürde' de metin madenciliği ile ilgili yapılan sınıflandırma çalışmaları incelendiğinde RO' nın diğer algoritmalara göre daha iyi çalıştığı görülmektedir. Aşağıda bazı örnek çalışmalar verilmiştir.

Kılınç & Yazarlı (2018)' nın yapmış olduğu 'İstatistik Kitaplarının Metin Madenciliği Yöntemleri Kullanılarak Yazarlarının Eğitimine Göre Sınıflandırılması' adlı çalışmada, sınıflandırma algoritması olarak k-en yakın komşuluk (K-NN), destek vektör makinesi (SVM) ve rasgele orman (RF) kullanılmıştır. En iyi algoritma başarı sonucunu Rastgele Orman olarak bulmuşlardır.

Ünal & Şeker (2018)'in yapmış olduğu 'Metin Madenciliğinde Yazar Tanıma' adlı çalışmada, sınıflandırma algoritması olarak karar ağacı, k-en yakın komşuluk, naif bayes, rastgele orman kullanılmıştır. En iyi algoritma başarı sonucu rastgele orman algoritması olarak bulunmuştur.

Karasoy & Ballı (2016)'nın yapmış olduğu 'İçerik Tabanlı İstenmeyen SMS Filtreleme için Mobil Uygulama Geliştirilmesi ve Sınıflandırma Algoritmalarının Karşılaştırılması' adlı çalışmada, Bagging, Rastgele Orman, Rastgele Alt Uzay sınıflandırma algoritmaları kullanılmıştır. En iyi başarı sonucunu veren algoritma Rastgele Ormandır.

Tekin (2018)'nin yapmış olduğu 'Yazılım Geliştirme Taleplerinin Metin Madenciliği İle Sınıflandırılması Ve Önceliklendirilmesi' adlı yüksek lisans tezinde SMO, Ratation Forest, Random Forest, Naive Bayes, Naive Bayes Multional sınıflandırma algoritmasını kullanmıştır. En iyi başarı sonucunu veren algoritma Rastgele Ormandır.

Abidin vd., (2017)'nin yapmış olduğu 'Klasik Türk Müziğinde Makam Tanıma İçin Veri Madenciliği Kullanımı' adlı çalışmada Rastgele Orman Algoritmasının iyi başarı ile sınıflandırma işlemini yaptığını ortaya koymuştur.

Buna rağmen, Rastgele Orman Algoritmasının parametreleri değiştirilerek daha iyi sonuçlar da elde edilebilir. Rastgele Orman Algoritması için ormandaki ağaç sayısının değişmesi ve seçilecek olan ağaç elemanların seçiminin değiştirilmesi, RO başarı sonucunu ve çalışma süresini olumlu yönde etkileyecektir (Özdarıcı vd., 2011).

Doğrulama yöntemleri verinin doğru analiz edilmesi için önemlidir. Veriyi doğru analiz edebilmek için kullanılır. Veride test ve eğitim kümelerinin doğru seçilmesi analizin doğruluğunu arttırır. Yanlış seçilmiş bir eğitim kümesi, sonucu da yanlış çıkaracaktır. Yapılan projede k-katlı çapraz doğrulama yöntemi kullanılmıştır.

Klasik çapraz doğrulama yönteminde işlem 4 adımda gerçekleşmektedir. İlk olarak veri rastgele veri iki eşit parçaya bölünmektedir. Eğitim ve test kümeleri rastgele oluşturulur.

Veri iki parçaya bölündükten sonra ilk parça ile model oluşturularak analiz yapılır. İkinci parça ile model oluşturularak analiz yapılır. Daha sonra oluşturulan iki modelin sonuçlarının ortalaması alınır. Veri setinin tamamı kullanılarak model oluşturulur ve analiz yapılır. Elde edilen son iki sonuç karşılaştırılarak doğruluk oranı tespit edilmektedir (Muslu, 2009)

K-katlı Çapraz doğrulama yönteminde, veri eğitim ve test olarak k parçaya bölünmektedir. Bu şekilde test kümesi ve eğitim kümesi belirlenmiş olur. İlk olarak makine çalışır, daha sonra test kümesi sırayla eğitim kümesi ile yer değiştirir. Sonucun ortalaması alınarak algoritma doğruluğu belirlenmektedir. K kere işlem yapıldığı için, bütün veri seti hem test hem de eğitim kümesi olarak kullanılmaktadır (Rodriguez vd., 2010).

Örnek: K=4 olarak alınırsa;
Veri Seti 4 eşit parçaya bölünmektedir. 4 parçadan bir tanesi test, 3 tanesi eğitim olarak alınmaktadır. Her seferinde test verisi değiştirilerek işlem yapılmaktadır. Böylece bütün veri test ve eğitim olarak kullanılmaktadır. Böylelikle veri 4 defa işleme girmiş olmaktadır ve doğruluğu sağlanmış olmaktadır. Çizelge 8'de test ve eğitim parçaları gösterilmektedir.

Çizelge 8 K-Katlı Çapraz Doğrulama Örneği

1.Parça	2.Parça	3.Parça	4.Parça
Test	Eğitim	Eğitim	Eğitim
Eğitim	**Test**	Eğitim	Eğitim
Eğitim	Eğitim	**Test**	Eğitim
Eğitim	Eğitim	Eğitim	**Test**

Kaynak: Kırlıoğlu & Ceyhan, 2014.

Makine öğrenmesi gerçekleştikten sonra çıkan sonuçlarının yorumlanması gerekmektedir. Sonuçları değerlendirirken tablolar esas alınır. Sınıflandırma tablolarının değerlendirilmesinde karışıklık matrisi sonuçları ve buna bağlı değişkenler yorumlanır. Karışıklık matrisinin ve diğer değerlenir hangi işlemlerden geçtiği aşağıda açıklanmıştır.

Çizelge 9 Karışıklık Matrisi

	Tahmin Edilen		
	Negatif	**Pozitif**	
Gerçek	x	y	**Negatif**
	z	k	**Pozitif**

Kaynak: Kılınç vd., 2016.

x: Değerin olumsuz tahminlerinin gerçekte doğru olduğunu belirtmektedir.

y: Değerin olumlu tahminlerinin gerçekte yanlış olduğunu belirtmektedir.

z: Değerin olumsuz tahminlerinin gerçekte yanlış olduğunu belirtmektedir.

k: Değerin olumlu tahminlerinin gerçekte doğru olduğunu belirtmektedir.

x+y+z+k: veri setindeki eleman sayısını belirtmektedir.

x ve k değerleri doğru sınıflandırılmış verileri temsil emketedir. Z ve y ise yanlış sınıflandırılmış verileri temsil etmektedir.

x+k: Doğru sınıflandırılan veri sayısının toplamını belirtmektedir.

z+k: Yanlış sınıflandırılan veri sayısının toplamını belirtmektedir.

Hatırlatma: Doğru olan veri sayılarının toplamının, toplam veri sayısına bölünmesi ile bulunur. Çizelge 9'da ifade edilmiştir (Talan, 2016; Coşkun & Baysal, 2011).

$$\mathrm{AC} = \frac{\mathrm{x} + \mathrm{k}}{\mathrm{x} + \mathrm{y} + \mathrm{z} + \mathrm{k}} \qquad (\mathbf{19})$$

Olumlu doğru: Olumlu pozitif değerlerin, pozitif değerlere oranı ile bulunur. Aşağıdaki formülde ifade edilmiştir.

$$\mathrm{TP} = \frac{\mathrm{k}}{\mathrm{z} + \mathrm{k}} \qquad (\mathbf{20})$$

Olumlu yanlış: Olumlu yanlıs değerlerin, yanlış değerlere oranı ile bulunur. Aşağıdaki formülde ifade edilmiştir.

$$\mathrm{TN} = \frac{\mathrm{z}}{\mathrm{x} + \mathrm{y}} \qquad (\mathbf{21})$$

Olumsuz doğru: Olumsuz doğru değerlerin, doğru değerlere oranı ile bulunur. Aşağıdaki formülde ifade edilmiştir.

$$\mathrm{FP} = \frac{\mathrm{y}}{\mathrm{z} + \mathrm{k}} \qquad (\mathbf{22})$$

Olumsuz yanlış: Olumsuz yanlış değerlerin, yanlış değerlere oranı ile bulunur. Aşağıdaki formülde ifade edilmiştir.

$$\mathrm{FN} = \frac{\mathrm{x}}{\mathrm{x} + \mathrm{y}} \qquad (\mathbf{23})$$

Kesinlik: Pozitif tahmin edilen durumların, pozitif durumlara oranı ile bulurnur. Aşağıdaki formülde ifade edilmiştir.

$$\mathrm{P} = \frac{\mathrm{k}}{\mathrm{y}+\mathrm{k}} \qquad (\mathbf{24})$$

F-Ölçütü: Kesinik ve doğruluğun beraber kullanıldığı değerlerdir. Sonucun değerlendirilmesi ikisi bir arada kullanıldığında daha doğru yapılmaktadır. Aşağıdaki formülde ifade edilmiştir.

$$\mathrm{F} - \text{Ölçütü} = \frac{2 * \text{kesinlik} * \text{hassasiyet}}{\text{kesinlik} + \text{hassasiyet}} \qquad (\mathbf{25})$$

Örneğin: Çizelge 10'da karışıklık matrisi Verilen 76 adet veri için yorumlamayı etki edecek değeri hesaplayınız.

Çizelge 10 Karışıklık Martisi Örneği

	Tahmin Edilen		
Negatif	**Pozitif**		
12	24	**Negatif**	**Gerçek**
22	18	**Pozitif**	

12+22+24+18= 76 adet veri

22+24= 46 yanlış sınıflandırılan veri

12+18=30 doğru sınıflandırılan veri

$$\mathrm{AC} = \frac{12 + 18}{12 + 24 + 22 + 18} = 0.394$$

$$\mathrm{TP} = \frac{18}{22 + 18} = 0.450$$

$$\mathrm{TN} = \frac{22}{12+24} = 0.611$$

$$\mathrm{FP} = \frac{22}{22+18} = 0.55$$

$$\mathrm{FN} = \frac{12}{12+24} = 0.333$$

$$\mathrm{P} = \frac{18}{24+18} = 0.428$$

$$\mathrm{F-Ölçütü} = \frac{2 * 0.324 * 0.428}{0.324 + 0.428} = 0.368$$

D. İnceleme Ortamı

Yapay zeka algoritmalarının derlenebileceği birden fazla derleyici ortamı bulunmaktadır. Bunlardan bazıları çizelge 11'de verilmiştir.

Çizelge 11 Bazı Makine Öğrenmesi Ortamları

Derleyici Adı	Yayın Tarihi	Dil	Web Sitesi
Yale-Rapid Miner	2001-2006	Dilden Bağımsız	www.rapidminer.com
Knime	2004	Java	www.knime.org
Weka	1993	Java	www.cs.waikato.ac.nz/~ml/weka
R	1997	C, Fortran, R	www.r-project.org

Kaynak: Ranga & Bansal, 2014.

Çizelge 11'de 4 adet veri madenciliği ortamı ve bu ortamlara ait bazı özellikler verilmiştir. Aşağıda veri madenciliği ortamlarının tarihsel süreci, kapsam ve özellikleri açıklanmıştır.

Rapid Miner, 2001 yılında geliştirilmeye başlanmıştır. Ralf Klinkenberg, Simon Fischer ve Ingo Mierswa tarafından geliştirilmiştir. 2007 yılında ismi Yale'den Rapid

Miner olarak değiştirilmiştir. Açık kaynak kodlu ve geliştirilebilirdir. Bunun yanı sıra ücretli ve ücretsiz versiyonlarıda bulunmaktadır. Java tabanlı bir yazılımdır. Arayüz yönlendirmesi ile biden fazla işlem yapılmaktadır. Bu yüzden fazla kod yazmaya ihtiyaç duyulmamaktadır. Buda yapılacak olan hataları minimum indirilmesinde yardımcı olur. Kullanıcı dostudur. Her seviyeden kullanıcıya hitap edecek şekilde tasarlanmıştır. Veriyle alakalı bütün işlemler Rapid Miner üzerinden yapılabilmektedir. Veri ön işlemeden başlayan süreç, sonuç görselleştirme aşamasına kadar işlemlerin yapılması konusunda imkan sağlamaktadır. R ve Python dillerine de desteği vardır. Buda Rapid Miner'ı daha güçlü kılan yapıya sahip olmasını sağlar. Dosyaları png, svg, jpeg, eps veya pdf formatlarında olduğu gibi dışa aktarılabilir (Kaya & Özel, 2014).

Knime, Michael Berthold'ın başında bulunduğu ekip 2004 yılında Knime'ı geliştirmeye başlamışlardır. Merkezi Zürihtedir ve Konstanz, Austin ve Berlinde de ofisleri mevcuttur. Açık kaynak kodlu, ücretsiz bir platformdur. Knime kullanımı sürekle bırak mantığıyla çalışır ve kullanıcı dosturdur. Arka planda kod yazmaya gerek duymadan node'lar ile işlemler yapılmaktadır. Knime, Java tabanlı olmasına rağmen Python, Perl, JavaScript dillerinide desteklemektedir. Veri analizi ön işleme adımından, görselleştirmeye kadar bütün işlemleri içerisinde yapmaya imkan sağlamaktadır. Doc, pdf, xls, ppt ve diğerleri gibi belge formatlarına rapor şablonları oluşturabilmektedir (Doğan, 2017).

R hem programlama dili hemde yazılan programını çalıştırmaya yarayan bir ortamdır. John Chambers ve arkadaşları tarafından yazılan GNU projesidir. S dili ile benzerlik taşımaktadır. R istatistiksel ve algoritmik olarak verilerin incelenmesine olanak sağlamaktadır. Açık kaynak kodludur. Aynı zamanda bir programlama dili olduğu için diğer veri madenciliği analizi derleyicilerine göre kişilerin hata yapma olasılığı vardır. Grafik arayüzleri oldukça gelişmiş bir şekilde tasarlanmıştır. UNIX ortamlarında, Windows'ta ve MacOs'ta derlenebilir ve çalıştırılabilir. Ücretsiz R Studio ortamı mevcuttur. R yazıldıktan sonra geliştirilme aşamasında, diğer dillerin yanı sıra R da kullanılmıştır. Yani R dilini geliştirirken yazılan kodların bir kısmı R dili ile yazılmıştır (Dener vd., 2009).

Weka, Waikato Üniversitesinde (1993) geliştirilmeye başlanmıştır. Java dili ile yazılan bir yazılımdır. Ücretli ve üzretsiz sürümleri mevcuttur. Weka'da ekstra kod yazarak değil, arayüz kullanımı ile bütün işlemler yapılabilmektedir. Buda

kullanıcının hata yapma olasılığını düşürmektedir. Kullanım kolaylığı sebebiyle ve desteklediği algoritmalar ile sıklıkla tercih edilmektedir. Kullanıcı dostu arayüze sahiptir. Açık kaynak kodludur ve geliştirilebilir (Alan, 2012). Kullanıcı arayüzü şekil 10' da verilmiştir.

Şekil 10 Weka Kullanıcı Arayüzü

Kaynak: Narin & İşler 2012.

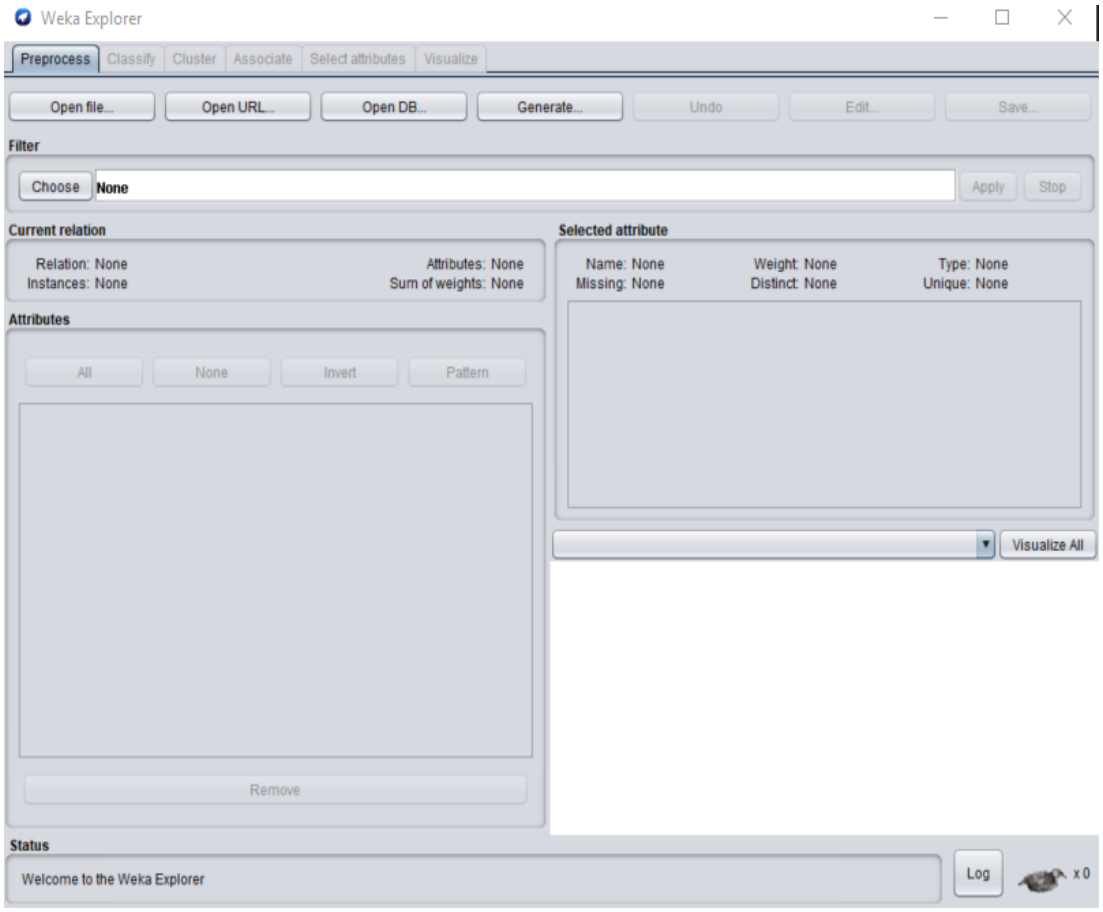
Explorer Arayüzü: Veri analizleri ve veri ön işleme yapılmaktadır.

Experimenter Arayüzü: Performans analizleri yapılmaktadır.

KnowledgeFlow: Grafik arayüzleri yapılmaktadır.

Simple CLI: Komut satırı kullanıllanımı yapılmaktadır

Weka içerisinde derleme işleminin yapılabilmesi için birden fazla arayüz bulunmaktadır.Yapılmak istenilen işleme göre ilgili arayüz seçilerek, analizler yapılabilir.



Şekil 11 Weka Explorer Sekmesi

Kaynak: Narin & İşler 2012

Explorer Arayüzü, veri analizleri ve veri ön işleme yapılmaktadır. Experimenter Arayüzü, performans analizleri yapılmaktadır. KnowledgeFlow, grafik arayüzleri yapılmaktadır. Simple CLI, komut satırı kullanıllanımı yapılmaktadır.

Weka ile sınıflandırma, bölütleme ve regresyon işlemleri yapılabilmektedir. Uçtan uca veri ile ilgili bütün işlemler yapılabilir. Veri ön işleme, algoritma analizi ve veri görselleştirme işlemleri buna dahildir. Ayrıca analiz yapılan algoritmaların performans değerlendirmeleri de yapılabilmektedir. İçerisinde geniş kütüphane ve filitreleme seçeneklerine sahiptir. Birden fazla dosya formatını desteklemektedir. Bunlar arff, names, data, csv, json, libsvm, m, dat, bsi, xrff dir. Weka içerisinde kullanılan algoritmalardan bazıları, j48 Karar Ağacı Algoritması, ZeroR Algoritması, OneR Algoritması, K-En Yakın Komşu Algoritması, Naive Bayes Algoritması vs. algoritmalar Weka içerisinde kullanılmaktadır (Takaoğlu, 2016).

Kullanıcı dostu olan Weka arayüzü sayesinde kullanıcılara kod yazmadan analiz yapma imkanı sağlamaktadır. Veri madenciliği ortamları karşılaştırıldığında

Weka'nın kullanım kolaylığı ve diğer avantajları göz önüne alındığında daha fazla tercih edildiği görülmektedir. Literatür'de Weka ve diğer veri madenciliği yazılımlarının karşılaştırılması incelenmiştir.

Veri madenciliği ortamları karşılaştırıldığında,

Kaya & Özal (2014) ve Rangra & Bansal (2014) çalışmalarında 6 adet veri madenciliği yazılımlarını incelemişler ve karşılaştırmışlardır. Bu yazılımlar Keel, Knime, R, Orange, RapidMiner (Yale), Weka'dır. Kaya ve Özal'a göre kullanıcının kolay örenmesi, kullanım kolaylığı, desteklediği dosya formatları ve içerdiği makine öğrenmesi algoritmaları açısından en başarılı yazılım Weka'dır. Ranga ve Baysal'a göre program bilgisi gerektirmeyen uygulamalar arasında ilk olarak Knime ile başlanmasını ve sonra daha güçlü bir yapıya sahip Weka ile devam edilmesi gerektiğini bildirmişlerdir.

Tekerek (2011) ve Doğan (2017) çalışmalarında 5 adet veri madenciliği yazılımlarını incelemişlerdir ve karşılaştırmıştırlar. Bunlar RapidMiner, Weka, Knime, Orange, R'dır. Tekerak'e göre RapidMiner Weka ve diğer veri madenciliği yazılımlarına göre daha iyidir. Fakat RapidMiner kod yazmaya imkan verirken Weka buna imkan vermez. Weka içerisinde kod yazmadan toollar aracılığıyla işlemler yapılmakta, buda kullanıcıyı yazılım alanine karıştırmadığı için dahaaz hataya yol açmaktadır. Doğan yapmış olduğu çalışma sonucu araştırdığı örneklem içerisindeki kişilerin veri madenciliği yazılımlarını tercih etme olasılıkları, Weka (%65.08), R (%58.73), RapidMiner (%46.03), Knime (%26.98), Orange (%23.81) olarak bulmuştur.

Çizelge 3'te veri madenciliği yazılımları karşılaştırılmıştır. Kullanıcının ihtiyacına ve bilgi birikimine göre, gerekli yazılım kullanılabilir.

Çizelge 12 Veri Madenciliği Yazılımları

	Weka	**RapidMiner**	**Knime**	**R**
PMML	Var	Var	Var	Var
Veri Görselleştirme	Var	Var	Var	Var
Veri Önişleme	Var	Var	Var	Var
Veri Analizi	Var	Var	Var	Var
Excel Desteği	Evet	Evet	Hayır	Evet
Yazıldığı Dil	Java	Java	Java	C, Fortran, R

Kaynak: Yıldız & Şeker, 2016.

Gerçekleştirilen projede Weka içerisinde sınıflandırma işlemi yapılmıştır. Weka'da sınıflandırma işlemi 'explorer' sekmesinde yapılmaktadır. Explorer içerisinde 'Preprocess' sekmesinde analiz yapılmadan önce ön işleme adımları gerçekleştirilmektedir. Ayrıca veri bu sekmede programın içerisinde dahil edilmektedir. 'Classify' sekmesinde sınıflandırma işlemi yapılmaktadır. Sınıflandırma işlemi yapılırken algoritmalar ve kısıtlamalar bu sekmede seçilmektedir. Sınıflandırma sonucu çıkan döküman bu sekmede görülmektedir.

IV. UYGULAMA

Veri ön işleme aşamasından sonra uygulama aşamasına geçilmiştir. Çizelge 14'te Veri ön işlem sonrası sınıflandırmaya etkisi olan kelime çizelgesi ve bu kelimelere ait ortalama ve standart sapma değerleri verilmiştir.

Çizelge 13 Ön İşlem Sonrası Veri Seti

Kelime	**Mean**	**StdDev**	**Kelime**	**Mean**	**StdDev**
Companies	0.054	0.227	Party	0.102	0.302
Congress	0.086	0.281	President	0.358	0.479
Department	0.095	0.293	States	0.191	0.393
Donald	0.228	0.42	Texas	0.043	0.202
House	0.141	0.348	Trump	0.253	0.435
Mexico	0.047	0.212	University	0.102	0.303
National	0.132	0.339	Washington	0.144	0.351
Areas	0.103	0.304	Population	0.093	0.291
World	0.359	0.48	Officials	0.195	0.396
Boeing	0.006	0.074	Response	0.09	0.286
Children	0.133	0.339	Protect	0.073	0.26
Citizens	0.076	0.264	Recent	0.211	0.408
Community	0.123	0.329	Recently	0.131	0.338
Country	0.359	0.48	Small	0.104	0.306

Kelime	Mean	StdDev	Kelime	Mean	StdDev
Elections	0.122	0.328	State	0.261	0.439
Federal	0.076	0.264	System	0.122	0.328
Life	0.167	0.373	Sports	0.02	0.14
Final	0.193	0.395	Team	0.208	0.406
Current	0.128	0.334	Started	0.17	0.376
Government	0.329	0.47	Win	0.219	0.414
Groups	0.153	0.36	Workers	0.048	0.215
Human	0.111	0.314	Working	0.138	0.345
Issue	0.132	0.338	Cup	0.116	0.32
Large	0.109	0.312	League	0.106	0.307
Led	0.165	0.371	Olympic	0.028	0.164
Forces	0.129	0.335	War	0.176	0.381
Living	0.09	0.286	Champion	0.083	0.276
Main	0.115	0.319	Football	0.067	0.25
Members	0.179	0.384	Game	0.154	0.361
Military	0.184	0.388	Match	0.138	0.345
Murder	0.036	0.185	Player	0.081	0.272
Office	0.123	0.329	Season	0.117	0.321
Order	0.119	0.324	Standings	0.011	0.105
Part	0.263	0.441	Title	0.129	0.335

Kelime	Mean	StdDev	Kelime	Mean	StdDev
People	0.467	0.499	Tournment	0.077	0.266
Political	0.271	0.445	Islam	0.047	0.212
Institutions	0.051	0.219			

Eldeki verilerin bir kısmı test, bir kısmı eğitim kümesi olarak kullanılmıştır. Veriler k-katlı çapraz doğrulama yöntemi kullanılarak test ve eğitim olarak ayrılmıştır. Literatürde genelde k değeri 10 kabul edilmektedir. Bu sebeple k değeri 10 alınmıştır.

A. ZeroR Algoritması

1800 adet veriden 600 adet veri doğru, 1200 adet veri yanlış sınıflandırılmıştır. Başarı oranı %33.3 tür. Karışıklık matrisi çizelge 15'te verilmiştir. Doğru sınıflandırılan alanlar koyu olarak gösterilmiştir. Koyu olmayan alanlar yanlış sınıflandırılan veriyi temsil etmektedir. Algoritmanın sınıflandırma mantığı gereği tek kategoride bütün veriler toplanmıştır.

Çizelge 14 ZeroR Algoritması Karışıklık Matrisi

Karışıklık Matrisi			
Sınıf	International News	Sports News	Magazine News
International News	**600**	0	0
Sports News	600	**0**	0
Magazine News	600	0	**0**

Çizelge 16 incelendiğinde International News kategorisinde F-Ölçütünün değeri diğer değerlerden daha yüksektir. F- ölçütü değeri diğer değerleride içerisinde barındırdığı için daha iyi sonuç vermektedir.

Çizelge 15 ZeroR Algoritması Sonuç Değerleri

Sınıf	TP Oranı	FP Oranı	Kesinlik	Hatırlatma	F-Ölçütü
International News	**1.00**	**1.00**	**0.33**	**1.00**	**0.50**
Sports News	0.00	0.00	?	0.00	?
Magazine News	0.00	0.00	?	0.00	?
Ağırlıklı Ortalama	0.33	0.33	?	0.33	?

B. Naif Bayes Algoritması

1800 adet veriden 1575 adet veri doğru, 225 adet veri yanlış sınıflandırılmıştır. Başarı oranı %87.5 dir. Karışıklık matrisi çizelge 17'de verilmiştir. Doğru sınıflandırılan alanlar koyu olarak gösterilmiştir. Koyu olmayan alanlar yanlış sınıflandırılan veriyi temsil etmektedir. En fazla veri Sport News kategorisinde toplandığı görülmektedir.

Çizelge 16 Naif Bayes Algoritması Karışıklık Matrisi

Karışıklık Matrisi			
Sınıf	International News	Sports News	Magazine News
International News	**514**	24	62
Sports News	23	**566**	11
Magazine News	82	23	**495**

Çizelge 18 incelendiğinde Sport News kategorisinde F-Ölçütünün değerinin daha yüksek olduğu gözlemlenmektedir.

Çizelge 17 Naif Bayes Algoritması Sonuç Değerleri

Sınıf	TP Oranı	FP Oranı	Kesinlik	Hatırlatma	F-Ölçütü
International News	0.85	0.08	0.83	0.85	0.84
Sports News	**0.94**	**0.03**	**0.92**	**0.94**	**0.93**
Magazine News	0.82	0.06	0.87	0.82	0.84
Ağırlıklı Ortalama	0.87	0.06	0.87	0.87	0.87

C. Rastgele Orman Karar Ağacı Algoritması

1800 adet veriden 1613 adet veri doğru, 187 adet veri yanlış sınıflandırılmıştır. Başarı oranı %89 dur. Ağaç sayısı:100, özellik seçimi:1 alınmıştır. Çalışma süresi 0.78 sn'dir. Karışıklık matrisi çizelge 19'da verilmiştir. Doğru sınıflandırılan alanlar koyu olarak gösterilmiştir. Koyu olmayan alanlar yanlış sınıflandırılan veriyi temsil etmektedir. En fazla veri Sport News kategorisinde toplandığı görülmektedir.

Çizelge 18 Rastgele Orman Algoritması Karışıklık Matrisi

Karışıklık Matrisi			
Sınıf	International News	Sports News	Magazine News
International News	**516**	20	64
Sports News	23	**561**	16
Magazine News	51	13	**536**

Çizelge 20 incelendiğinde Sports News kategorisinde F-Ölçütünün değerinin daha yüksek olduğu gözlemlenmektedir.

Çizelge 19 Rastgele Orman Algoritması Sonuç Değerleri

Sınıf	**TP Oranı**	**FP Oranı**	**Kesinlik**	**Hatırlatma**	**F-Ölçütü**
International News	0.86	0.06	0.87	0.86	0.86
Sports News	**0.93**	**0.02**	**0.94**	**0.93**	**0.94**
Magazine News	0.89	0.06	0.87	0.89	0.88
Ağırlıklı Ortalama	0.89	0.05	0.89	0.89	0.89

Algoritmaların sonuçları birbirleri ile karşılaştırılarak değerlendirmeler yapılmıştır. Çizelge 21'de algoritma sonuçları listelenmeketdir.

Çizelge 20 Algoritmaların Sonuçlarının Karşılaştırılması

Algoritma	Naif Bayes	ZeroR	Rastgele Orman
Başarı Oranları (%)	87	33	**89**
Doğru Sınıflandırılan Veri Sayısı	1571	600	**1613**
Yanlış Sınıflandırılan Veri Sayısı	229	1200	**187**
Çalışma Zamanı (sn)	0.06	0.02	**0.78**
F-Ölçütü	0.93	0.50	**0.94**

Çizelge 21 incelendiğinde, Rastgele Orman Algoritmasının ZeroR ve Naif Bayes Algoritmasına göre daha iyi sonuç verdiği gözlemlenmiştir. Zaten Algoritmaların yapısı incelenirse bunun beklenilen bir durum olduğu görülebilir. Hipotez kısmında da tahmin edildiği gibi RO diğer algoritmalara göre daha iyi başarı göstererek sınıflandırma işlemini yapmaktadır.

ZeroR algoritmasının başarı oranının %33 olmasının sebebi mevcutta üç adet sınıf bulunmasıdır. ZeroR Algoritması en çok hangi sınıfta veri varsa, daha sonar gelecek olan verileri o sınıfa dahil etme eğiliminde olduğu için bu durum tahmin edilmektedir. Mevcutta veri seti içerisinde üç adet eşit verilere sahip sınıf bulunduğundan sonuç %34 olmuştur. Sınıflar içerisinde veri sayısı eğer eşit olmasaydı, ZeroR Algoritması daha iyi başarı ile sınıflandırma yapabilirdi. Ancak

alınan sonuç verimli olmayacaktı. İlkel bir algoritma olduğu için fazla tercih edilmemektedir.

Naif Bayes Algoritması olasılık tabanlı bir algoritmadır ve metin madenciliğinde de fazlaca tercih edilmektedir. RO'dan daha az oranda başarı elde edilmesinin sebebi veri setidir. Aralarında çok fazla fark yoktur. Yaklaşık %2 oranında bir başarı farklı mevcuttur. Naif Bayes Algoritması mevcut veri seti içinde bazı durumlarda tercih edilebilir. Çalışma süresinin inemli olduğu bir durumda mevcut veri seti için tercih edilebilir. RO'ya göre daha iyi bir çalışma süresi ile sınıflandırma işlemi yapmıştır.

Algoritmalar karşılaştırıldığında, Rastgele Orman Algoritmasının en iyi başarı oranı ile sınıflandırma yaptığı görülmektedir. Bu algoritmanın yapısı incelendiğinde algoritmanın başarı oranını arttırmak için, parameter değişiklikleri yapılabilir. RO alt ağaçlardan oluşmaktadır. Bu sebeple ağaç sayısının ve özellik seçiminin değikliği çalışma zamanı ve başarı oranını olumlu yönde etkileyecektir. Ağaç sayısı, RO içerisinde sınıflandırma işlemi yaparken kullanılan alt ağaçların adedidir. Özellik seçimi ise alt ağaçlar için bir kerede seçtiği özellik sayısıdır. Çizelge 22'de RO ait parameter değişimi sonucu başarı oranı ve çalışma süresi değişiklikleri verilmiştir.

Çizelge 21 Rastgele Orman Parametre Seçimi

Ağaç Sayısı	Özellik Seçimi	Başarı Oranı(%)	Çalışma Süresi(sn)
100	1	89.0	0.78
100	2	90.7	0.80
100	3	90.5	0.90
50	1	90.4	0.38
50	**2**	**90.7**	**0.47**
50	3	90.2	0.55
25	1	89.6	0.23
25	2	89.4	0.27
25	3	89.8	0.30

İlk aşamada ağaç sayısı=100 ve özellik seçimi=1 iken başarı oranı %89, çalışma süresi 0.78 sn dir. Ağaç sayısı=50 ve özellik seçimi=2 olarak seçildiğinde Rastgele Orman Karar Ağacı Algoritmasının en iyi başarı oranını verdiği görülmüştür. Başarı oranı %90.7 dur. Çalışma süresi= 0.47 sn'dir. Yaklaşık 0.02 başarı oranı ve 0.31 sn daha iyi zaman ile sınıflandırma yapmıştır. Ağaç sayısı 100 ile 25 arasında alınmıştır. Özellik seçimi 3 adettir. Bunların sebebi veri seti için optimum seviyeye geldiğinde özellik değiştirmenin durdurulmasından kaynaklanmaktadır. Belirli değişiklikten sonar mevcut veri seti için özellikler optimum seviyeye ulaşır ve değişiklik devam ettikçe olumsuz yönde etkilenmeye başlar. Metin madenciliğinde Rastgele Orman

algoritması kullanırken parametreler ağaç sayısı ve özellik seçimi değişkenleri değiştirilirse daha iyi sonuç elde edilmektedir.

V. SONUÇ

Yapılan çalışma, belirli sınıflara dahil olan haber verileri üzerinde üç faklı tip ve üç farklı kategori yapay zeka algoritmaları kullanılarak makine eğitilmesi yöntemiyle yapılan sınıflandırmada, algoritmaların sınıflandırma analizlerinin karşılaştırılması ve haber metinlerinin yönetimlerinin saptanmasıdır. 1800 adet İngilizce haber metini üzerinden sınıflandırma işlemi yapılmıştır. Çalışma içerisinde üç adet algoritma mevcuttur. Bu algoritmalar: ZeroR, Naif Bayes, Rastgele Orman'dır. Algoritmaların başarı oranları karşılaştırıldığında yaklaşık %89 oranında ve çalışma zamanı 0.78 sn olarak Rastgele Orman Algoritmasının diğer algoritmalara göre daha iyi çalıştığı tespit edilmiştir. Fakat elde edilen bu başarı oranı ve çalışma süresi Rastgele Orman Algoritması içerisindeki parametreleri değiştirerek arttırılabilir. Bu sebeple 'ağaç sayısı' ve 'özellik seçimi' parametreleri değiştirilmiştir. Bu değişim sonucunda başarı oranı yaklaşık %91 ve çalışma zamanı 0.47 sn olarak bulunmuştur. Metin madenciliğinde Rastgele Orman Algoritması diğer algoritmalara göre daha iyi sonuç vermektedir. Eğer ağaç sayısı:50 ve özellik seçimi:2 olarak seçilirse, elde edilen başarı oranı ve çalışma süreside iyileştirilmektedir.

Gelecekte Rastgele Orman (RO) Algoritmasını geliştirmek için iki adet proje düşünülmüştür. RO içerisindeki hesaplamalar yapılırken kullanılan gini indeksi yerine verilerin ortalaması alınarak denemeler yapılması planlanmaktadır. Diğeri ise RO oluşturulurken alt ağaçları rastgele oluşturuyor, bu rastgeleliği değiştirerek bir sıralama algoritması yazmak ve bu sıralama algoritmasına göre alt ağaçları oluşturmasını sağlamaktır. Yeni oluşturulan bu rastgele orman artı adını vereceğim algoritma üzerinde denemeler yapılması hedeflenmektedir.

KAYNAKLAR

Abidin, S, Öztürk, Ö & Öztürk, TÖ (2017), 'Klasik Türk Müziğinde Makam Tanıma İçin Veri Madenciliği Kullanımı', Gazi Üniversitesi Mühendislik, Mimarlık Fakültesi Dergisi, c.32, s.4, syf.1221-1232.

Adak, MF & Yurtay, N (2013), 'Gini Algoritmasını Kullanarak Karar Ağacı Oluşturmayı Sağlayan Bir Yazılımın Geliştirilmesi', Bilişim Teknolojileri Dergisi, s.3, c.6, syf.1-6.

Akçapınar, G (2014), 'Çevrimiçi Öğrenme Ortamındaki Etkileşim Verilerine Göre Öğrencilerin Akademik Performanslarının Veri Madenciliği Yaklaşımı İle Modellenmesi', Doktora Tezi, Hacettepe Eğitim Bilimler Enstitüsü, Ankara, Türkiye.

Akın, ZO (2010), 'Uluslararası Haber Ajanslarının Türkiye Haberlerinde Eşik Bekçiliği Uygulamaları: Reuters ve Ap Örneği', Yüksek Lisans Tezi, Gazi Üniversitesi Sosyal Bilimler Enstitüsü, Ankara, Türkiye.

Alan, MA (2012), 'Veri Madenciliği ve Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama', Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, s.33, syf.165-174.

Aravi, G (2014), 'Metin Madenciliği İle Sosyal Medya Analizi', Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi, İstanbul, Türkiye.

Aslan, M (2016), 'Derinlik Kamerası İle Yaşlılarda Düşme Tespiti', Doktora Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ, Türkiye.

Atalay, M & Çelik, E (2017), 'Büyük Veri Analizinde Yapay Zekâ Ve Makine Öğrenmesi Uygulamaları', Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, c.9, s.22, syf.155-172.

Atan, S & Çınar, Y (2018), 'Borsa İstanbul'da Finansal Haberler İle Piyasa Değeri İlişkisinin Metin Madenciliği Ve Duygu (Sentiment) Analizi İle İncelenmesi', Ankara Üniversitesi SBF Dergisi, c.74, s.1, syf.1-34.

Ay, D (2009), 'Veri Madenciliği Ve Apriori Algoritması ile Süpermarket Analizi', Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya, Türkiye.

Aydın, S (2007), 'Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama', Doktora Tezi, Anadolu Üniversitesi Sosyal Bilimler Enstitüsü, Eskişehir, Türkiye.

Aydın, C (2018), 'Makine Öğrenmesi Algoritmaları Kullanılarak İtfaiye İstasyonu İhtiyacının Sınıflandırılması', Avrupa Bilim ve Teknoloji Dergisi, s.14, syf.169-175.

Bilgen, İ (2014), 'İnsan ve Hiv-1 Proteinleri Arasindaki Etkileşimlerin Rastgele Orman Yöntemi ve Birlikte Öğrenme Yaklaşımı ile Tahmin Edilmesi, Yüksek Lisans Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.

Bilgin, AO (2018), 'Metin Madenciliği Yöntemleri İle Yazar Tanıma: Divan Edebiyatı Örneği', Yüksek Lisans Tezi, Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü, Trabzon, Türkiye.

Coşkun, C & Baysal, A (2011), 'Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması', Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, İnönü Üniversitesi Fen Bilimleri Fakültesi, Malatya, Türkiye, 2 - 4 Şubat.

Çalış, K, Gazdağı, O & Yıldız, O (2013), 'Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti', Bilişim Teknolojileri Dergisi, c.6, s.1, syf.1-7.

Çetinkaya, OZ (1981), 'Belirsizliğin Ölçülmesi ve Entropi', Doğa ve Bilim Dergisi, s.5, syf.323-335, erişim tarihi: 12.09.2019, < https://dergipark.org.tr/en/download/article-file/8387>

Çifçi, S (2011), 'Yazılı Basında Haber Cümlelerinin Analizi', International Periodical for The Languages, c.6, s.1, syf.925-938, erişim tarihi: 18.10.2019, < http://www.acarindex.com/dosyalar/makale/acarindex-1423934334.pdf>

Çimenli, S (2015), 'Churn Analysis and Prediction with Decision Tree and Artificial Neural Network', Yüksek Lisans Tezi, Kadir Has Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.

Delibaş, A (2008), 'Doğal Dil İşleme ile Türkçe Yazım Hatalarının Denetlenmesi', Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.

Demirhan, A, Kılıç, YA & Güler, İ (2010), 'Tıpta Yapay Zeka Uygulamaları', Yoğun Bakım Dergisi, c.1, s.9, syf.31-41, erişim tarihi: 23.08.2019, < http://www.yogunbakimdergisi.org/managete/fu_folder/2010-01/2010-9-1-031-041.pdf>

Dener, M, Dörtler, M & Orman, A (2009), '', Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, Harran Üniversitesi, Şanlıurfa, 11-13 Şubat.

Doğan, O (2017), 'Ücretsiz Veri Madenciliği Araçları ve Türkiye'de Bilinirlikleri Üzerine Bir Araştırma, Ege Stratejik Araştırmalar Dergisi, c.8, s.1, syf.77-93.

Ekelik, H (2019), 'Dijital Reklam Verilerinden Yararlanarak Potansiyel Konut Alıcılarının Rastgele Orman Yöntemiyle Sınıflandırılması', Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, Türkiye.

Erduran, GY (2017), 'Online Müşteri Şikayetlerinin Veri Madenciliği İle İncelenmesi', Yüksek Lisans Tezi, Trakya Üniversitesi Sosyal Bilimler Enstitüsü, Edirne, Türkiye

Ertemel, V.A. & Gürdal, S. (2016), Crm'in Geleceği: Yaygin Bilişim Ve Ortam Duyarli Mobil Pazarlama Kavramlarinin İncelenmesi', Kafkas Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, c.7, s.13, syf.169-187.

Ertuğrul, İ, Organ, A & Şavlı, A (2012), 'Veri Madenciliği Uygulamasına İlişkin PAÜ Hastanesinde Hasta Profilinin Belirlenmesi', Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, c.19, s.2, syf.97-103.

Girgin, A (2002), 'Uluslar Arası Haber Ajansları', İstanbul Üniversitesi İletişim Fakültesi Dergisi, c.1, s.12, syf.169-185.

Gök, M (2017), 'Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi', Gazi Üniversitesi Fen Bilimleri Dergisi, c.3, s.5, syf.139-148.

Göker, H & Tekedere, H (2017), 'FATİH Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi', Bilişim Teknolojileri Dergisi, c.10, s.3, syf. 291-299.

Hesarı, S (2018), 'Finansal Başarısızlık Tahmini: Yapay Sinir Ağı ve Karar Ağacı Yöntemleri Üzerine Bir İnceleme', Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü, İzmir, Türkiye.

Hazım, LR (2018), 'Four Classification Methods Naïve Bayesian, Support Vector Machine, K-Nearest Neighbors and Random Forest Are Tested For Credit Card Fraud Detection', Yükseklisans Tezi, Altınbaş Üniversitesi, İstanbul, Türkiye.

İlhan, ÇK (2019), 'Televizyon Haberciliğinde Yeni Medyanın Kullanımı: Whatsapp İhbar Hattı', Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Gazetecilik Anabilim Dalı, İstanbul, Türkiye.

İşler, Y & Narin, A (2012), 'WEKA Yazılımında k-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi', Süleyman Demirel Üniversitesi Teknik Bilimler Dergisi, c.2, s.4, syf.21-29.

Jivani, AG (2011), 'A Comparative Study of Stemming Algorithms', International Journal of Computer Science, c.2, s.6, syf.1930-1938, erişim tarihi: 08.08.2019,<https://kenbenoit.net/assets/courses/tcd2014qta/readings/Jivani_ijcta2011020632.pdf>

Kalaycı, TE (2018), 'Kimlik hırsızı web sitelerinin sınıflandırılması için makine öğrenmesi yöntemlerinin karşılaştırılması', Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, c.5, s.24, syf.870-878.

Karakoyun, M & Hacıbeyoğlu, M (2014), 'Biyomedikal Veri Kümeleri İle Makine Öğrenmesi Sınıflandırma Algoritmalarının İstatistiksel Olarak Karşılaştırılması', Dokuz Eylül Üniversitesi Mühendislik Fakültesi Mühendislik Bilimleri Dergisi, c.16, s.48, syf.30-41.

Karasoy, O & Ballı, S (2016), 'İçerik Tabanlı İstenmeyen SMS Filtreleme için Mobil Uygulama Geliştirilmesi ve Sınıflandırma Algoritmalarının Karşılaştırılması', konferans bildirisi, International Artificial Intelligence and Data Processing Symposium (IDAP'16).

Kaya, M & Özal, SA (2014), 'Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması', Akademik Bilişim'14 - XVI. Akademik Bilişim Konferansı Bildirileri, Mersin Üniversitesi Mühendislik Fakültesi, Mersin, Türkiye, 5 - 7 Şubat.

Kaya, M & Özel, S (2009), 'Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması', Akademik Bilişim'14 - XVI. Akademik Bilişim Konferansı Bildirileri, Mersin Üniversitesi, Mersin, 5-7 Şubat.

Kazu, İY & Özdemir, O (2009), 'Öğrencilerin Bireysel Özelliklerinin Yapay Zeka ile Belirlenmesi (Bulanık Mantık Örneği)', Akademik Bilişim'09 - XI. Akademik Bilişim Konferansı Bildirileri, Harran Üniversitesi, Şanlıurfa, Türkiye, 11-13 Şubat.

Kılınç, D, Borandağ, E, Yücalar, F, Tunalı, V, Şimşek, M & Özçift, A (2016), 'KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi', Marmara Fen Bilimleri Dergisi, s.3, syf.89-94.

Kılınç Kan, B & Yazarlı, Y (2018). 'İstatistik Kitaplarının Metin Madenciliği Yöntemleri Kullanılarak Yazarlarının Eğitimine Göre Sınıflandırılması', Turkiye Klinikleri J Biostat, c.3, s.10, syf. 215-223.

Kırlıoğlu, H & Ceyhan, İF (2014), 'Mali Tablo Denetiminde Ön Analitik İnceleme Tekniği Olarak Veri Madenciliğinin Kullanımı: Borsa İstanbul Uygulaması, Akademik Yaklaşımlar Dergisi, c.5, s.1, syf. 13-36.

Koyuncugil, AS & Özgülbaş, N (2009), 'Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları', Bilişim Teknolojileri Dergisi, c.2, s.2, syf.21-31.

Kuzey, C (2012), 'Veri Madenciliğinde Destek Vektör Makinaları ve Karar Ağaçları Yöntemlerini Kullanarak Bilgi Çalışanlarının Kurum Performansı Üzerine Etkisinin Ölçülmesi ve Bir Uygulama', Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, Türkiye.

Muslu, D (2009). 'Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması', Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.

Namlı, ÖH & Özcan, T (2017), 'Forecasting of Box Office Revenue Using Machine Learning Algorithms', Yönetim Bilişim Sistemleri Dergisi, c.3, s.2, syf.130-143.

Namous, F, Rodan, A & Javed, Y (2018), 'Online News Popularity Prediction', The Fifth Hct Information Technology Trends, syf.28-29.

Narlı, S, Aksoy, E & Ercire, YE (2014), 'Investigation of Prospective Elementary Mathematics Teachers' Learning Styles and Relationships between Them Using Data Mining', International Journal of Educational Studies in Mathematics, c.1, s.1, syf.37-57.

Nasa, C & Suman (2012), 'Evaluation of Different Classification Techniques for WEB Data', International Journal of Computer Applications, c.52, s.9, syf.35-40, erişim tarihi: 15.07.2019, < https://pdfs.semanticscholar.org/aa28/256df9a08f2a1707147f85c1d26fb283b453.pdf?_ga=2.235086402.2121824527.1583049733-1619710388.1583049733>

Odabaş, Ö (2017), 'Veri Madenciliği Teknikleri İle Telekom Sektöründe Ayrilan Müşteri Analizi', Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.

Özcan, C (2014), 'Veri Madenciliğinin Güvenlik Uygulama Alanları ve Veri Madenciliği ile Sahtekârlık Analizi', Yüksek Lisans Tezi, İstanbul Bilgi Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, Türkiye.

Özdemir, S (2011), 'A Decision Tree Based Intrusion Detection System with Bootstrap Aggregating, Discretization, and Feature Selection' Yüksek Lisans Tezi, Boğaziçi Üniversitesi, İstanbul Türkiye.

Özdarıcı Ok, A, Akar, Ö & Güngör, O (2011), 'Rastgele Orman Sınıflandırma Yöntemi Yardımıyla Tarım Alanlarındaki ürün Çeşitliliğinin Sınıflandırılması', konferans bildirisi. ODTÜ, Jeodezi ve Coğrafi Bilgi Teknolojileri EABD, Ankara, Türkiye, Ocak.

Pala, T (2013), 'Tıbbi Karar Destek Sisteminin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi', Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.

Patsis, Y & Verhelsy, W (2008). 'A Speech/Music/ Silence /Garbage/ Classifier for Searching end Indexing Broadcast News Material', 19th International Conference on Database and Expert Systems Application, syf.585-589.

Pervan, N (2019). 'Derin Öğrenme Yaklaşımları Kullanarak Türkçe Metinlerden Anlamsal Çıkarım Yapma', Yüksek Lisans Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Türkiye.

Rangra, K & Bansal KL (2014), 'Comparative Study of Data Mining Tools', International Journal of Advanced Research in Computer Science and Software Engineering, c.4, s.6, syf.216-223.

Rodriguez, JD, Perez, A & Lozano, JA (2010), 'Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation, Ieee Transactıons On Pattern Analysıs And Machıne Intellıgence, c.32, s.3, syf.569-575.

Sayıcı, G (2013), 'Karar Ağaçları, Bayes Ağları Ve Etki Diyagramları Aracılığı ile Bilgi Keşfi ve Karar Verme', Yüksek Lisans Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, Türkiye.

Srividhya, V & Anitha, R (2010). 'Evaluating Preprocessing Techniques in Text Categor ization', International Journal of Computer Science, syf.49-51, erişim tarihi: 12.07.2019, < http://sinhgad.edu/ijcsa-2012/pdfpapers/1_11.pdf>

Şeker, SE (2015), 'Metin Madenciliği', Ybs Ansiklopedi, c.2, s.3, syf.30-32, erişim tarihi:10.2.2019,<https://www.researchgate.net/publication/281612863_YBS_Ansiklopedi_Cilt_2_Sayi_3_Eylul_2015>

Şengür, D (2013), 'Öğrencilerin Mezuniyet Notlarının Veri Madenciliği Metotları İle Tahmini Uygulaması', Yüksek Lisans Tezi, Fırat Üniversitesi Eğitim Bilimleri Enstitüsü, Elazığ, Türkiye.

Takaoğlu, M (2016), 'Birkaç Veri Kümesi ile WEKA ve MATLAB Üzerinde Kümeleme Algoritmalarının Karşılaştırılarak İncelenmesi', Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.

Talan, Mİ (2016), 'Veri Madenciliği İle Karpal Tünel Sendromuna Yönelik Ön Tanı Destek ve Hasta Takip Sisteminin Geliştirilmesi', Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara, Türkiye.

Tantuğ, AC (2012), 'Metin Sınıflandırma', Türkiye Bilişim Vakfı Bilgisayar Bilimleri Dergisi, c.6, s.6, syf.1-12, erişim tarihi: 08.06.2019, < https://dergipark.org.tr/tr/download/article-file/207208>

Taylan, A & Ünal, R (2017), 'Ana Akım Medyada Sansasyonel Habercilik: Sağlık İletişimi Örneği', III. Uluslararası Sağlık İletişimi Sempozyumu, Atatürk Üniversitesi Sağlık Bilimleri Fakültesi, Erzurum, Türkiye, 2-3 Kasım.

Tekerek, A (2011), 'Veri Madenciliği Süreçleri ve Açık Kaynak Kodlu Veri Madenciliği Araçları', Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, İnönü Üniversitesi Eğitim Bilimleri Fakültesi, Malatya, Türkiye, 2-4 Şubat.

Tekin, MC (2018). 'Yazılım Geliştirme Taleplerinin Metin Madenciliği İle Sınıflandırılması ve Önceliklendirilmesi', Yüksek Lisans Tezi, Maltepe Üniversitesi Fen Bilimleri Enstitüsü, Maltepe, İstanbul.

Topaloğlu, M & Sur, H (2014), 'Sarılık Semptomlarında Yanlış Teşhisi Azaltmak için Karar Ağacı Uygulaması', Research Article, c.11, s.3, syf.64-73, erişim tarihi: 06.05.2019, < https://www.nobelmedicus.com/Content/1/33/64-73.pdf>

Tuncer, D (2018), 'Veri Madenciliği Yöntemi İle Aylık Kullanılabilir Gelir Tahmini ve Göstergeleri', Yüksek Lisans Tezi, Trakya Üniversitesi Sosyal Bilimler Enstitüsü, Edirne, Türkiye.

Uçar, E (2012), 'Kablosuz Algilayici Ağlarin Uygulama Alanları ve Bir Algilayici Düğüm Tasarımı', Yüksek Lisans Tezi, Trakya Üniversitesi Fen Bilimleri Enstitüsü, Ed,rne, Türkiye.

Ulusoy, G (2013), 'Karar Ağacı Analizi ile Ab Genişleme Kriterlerinin Değerlendirilmesi', Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Ana Bilimdalı, İstanbul, Türkiye.

Uysal, M (2014), 'Veri Analizi için Genişleyebilir Bir Karar Ağacının Oluşturulması, Web ve Mobil Uygulamalarının Geliştirilmesi', Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara, Türkiye.

Ünal, Dİ & Şeker, ŞE (2018), 'Metin Madenciliğinde Yazar Tanıma', YBS Ansiklopedisi, c.5, s.1, syf.1-6,erişim tarihi: 1.04.2019, <https://ybsansiklopedi.com/wp-content/uploads/2018/06/yazar.pdf>

Vijayarani, S, Ilamathi, J & Nitya (2014), 'Preprocessing Techniques for Text Mining - An Overview', International Journal of Computer Science, c.5, s.1, syf.7-16.

Yetim, S (2015), 'Sürücüsüz Araçlar ve Getirdiği/Getireceği Hukuki Sorunlar', Ankara Barosu Dergisi. c.1, s.1, syf.127-183, erişim tarihi: 20.02.2019 < https://dergipark.org.tr/tr/download/article-file/398494>

Yıldız, B & Ağdeniz, Ş (2018), 'Muhasebe Analiz Yöntemi Olarak Metin Madenciliği', Muhasebe Bilim Dünya Dergisi, c.2, s.20, sfy.286-315, erişim tarihi: 01.04.2019, < https://dergipark.org.tr/en/download/article-file/491789>

Yıldız, M & Şeker, ŞE (2016), 'Veri Madenciliği Araçları (Data Mining Tools)', YBS Ansiklopedi, c.3, s.4, syf.10-19, erişim tarihi: 28.03.2019, <http://ybsansiklopedi.com/wp-content/uploads/2017/04/veri-madencili%C4%9Fi_araclari.pdf>

Yücel, A & Keskin Köylü, M (2018), 'Spam içerikli E-Postaların Tespiti için Bir Metin Madenciliği Uygulaması: Terimlerin Gama İlişki Katsayısına Dayalı Polarizasyonu', Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi, c.2, s.2, syf.95-104.

EKLER

Ek.1 Durdurma Kelimeleri

Durdurma Kelimeleri

a	b	d	from	ie
a's	be	definitely	further	if
able	became	described	furthermore	ignored
about	because	despite	g	immediate
above	become	did	get	in
according	becomes	didn't	gets	inasmuch
accordingly	becoming	different	getting	inc
across	been	do	given	indeed
actually	before	does	gives	indicate
after	beforehand	doesn't	go	indicated
afterwards	behind	doing	goes	indicates
again	being	don't	going	inner
against	believe	done	gone	insofar
ain't	below	down	got	instead
all	beside	downwards	gotten	into
allow	besides	during	greetings	inward
allows	best	e	h	is
almost	better	each	had	isn't
alone	between	edu	hadn't	it
along	beyond	eg	happens	it'd

already	both	eight	hardly	it'll
also	brief	either	has	it's
although	but	else	hasn't	itself
always	by	elsewhere	have	j
am	c	enough	haven't	just
among	c'mon	entirely	having	k
amongst	c's	especially	he	keep
an	came	et	he's	keeps
and	can	etc	hello	kept
another	can't	even	help	know
any	cannot	ever	hence	knows
anybody	cant	every	her	known
anyhow	cause	everybody	here	l
anyone	causes	everyone	here's	last
anything	certain	everything	hereafter	lately
anyway	certainly	everywhere	hereby	later
anyways	changes	ex	herein	latter
anywhere	clearly	exactly	hereupon	latterly
apart	co	example	hers	least
appear	com	except	herself	less
appreciate	come	f	hi	lest

appropriate	comes	far	him	let
are	concerning	few	himself	let's
aren't	consequently	fifth	his	like
around	consider	first	hither	liked
as	considering	five	hopefully	likely
aside	contain	followed	how	little
ask	containing	following	howbeit	look
asking	contains	follows	however	looking
associated	corresponding	for	i	looks
at	could	former	i'd	ltd
available	couldn't	formerly	i'll	m
away	course	forth	i'm	mainly
awfully	currently	four	i've	less
seeing	soon	thence	took	various
seem	sorry	there	toward	very
seemed	specified	there's	towards	via
seeming	specify	thereafter	tried	viz
seems	specifying	thereby	tries	vs
seen	still	therefore	truly	w
self	sub	therein	try	want
selves	such	theres	trying	wants

sensible	sup	thereupon	twice	was
sent	sure	these	two	wasn't
serious	t	they	u	way
seriously	t's	they'd	un	we
seven	take	they'll	under	we'd
several	taken	they're	unfortunately	we'll
shall	tell	they've	unless	we're
she	tends	think	unlikely	we've
should	th	third	until	welcome
shouldn't	than	this	unto	well
since	thank	thorough	up	went
six	thanks	thoroughly	upon	were
so	thanx	those	us	weren't
some	that	though	use	what
somebody	that's	three	used	what's
somehow	thats	through	useful	whatever
someone	the	throughout	uses	when
something	their	thru	using	whence
sometime	theirs	thus	usually	whenever
sometimes	them	to	uucp	where
somewhat	themselves	together	v	where's

somewhere	then	too	value	whereafter
introduction	methods	interest	document	approv
abstract	discussion	declar	article	work
result	competing	contribut	research	grant
results	compet	invest	read	paper
method	author	scanned	write	manuscript
3	4	5	6	7
one	rather	http	much	of
ones	rd	ftp	must	off
only	re	url	my	often
onto	really	email	myself	oh
or	reasonably	supplementary	n	ok
other	regarding	material	name	okay
others	regardless	abbreviation	namely	old
otherwise	regards	abbreviations	nd	on
ought	relatively	acknowledgments	near	once
our	respectively	acknowledgements	nearly	whereas
ours	s	footnote	necessary	whereby
ourselves	said	footnotes	need	wherein
out	same	selected	needs	whereupon

outside	saw	references	neither	wherever
over	say	conclusion	never	whether
overall	saying	conclusions	nevertheless	which
own	says	tables	new	while
p	second	pubmed	next	whither
particular	secondly	top	nine	who
particularly	see	1	no	who's
per	you	2	nobody	whoever
perhaps	you'd	9	non	whole
placed	you'll	many	none	whom
please	you're	may	noone	whose
plus	you've	maybe	nor	why
possible	your	me	normally	will
presumably	yours	mean	not	willing
probably	yourself	meanwhile	nothing	wish
provides	yourselves	merely	novel	with
q	z	might	now	within
que	zero	more	nowhere	without
quite	www	moreover	o	won't
qv	pdf	most	obviously	wonder
would	wouldn't	would	x	y

yes	yet	text	page	fig
figure	table	8	11	12
r	site	mostly		

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı: Firdevs DURNAGÖL

Doğum Yeri ve Tarihi: Fatih, 08.08.1991

E-mail: eser.firdevs@gmail.com

Eğitim Durumu

Lisans (2010 – 2015)

Karabük Üniversitesi Bilgisayar Mühendisliği

İş Deneyimi

TrtArabi 2015- 2017

Bilgi İşlem Uzmanı

TrtWorld 2017-

Bilgi İşlem Uzmanı