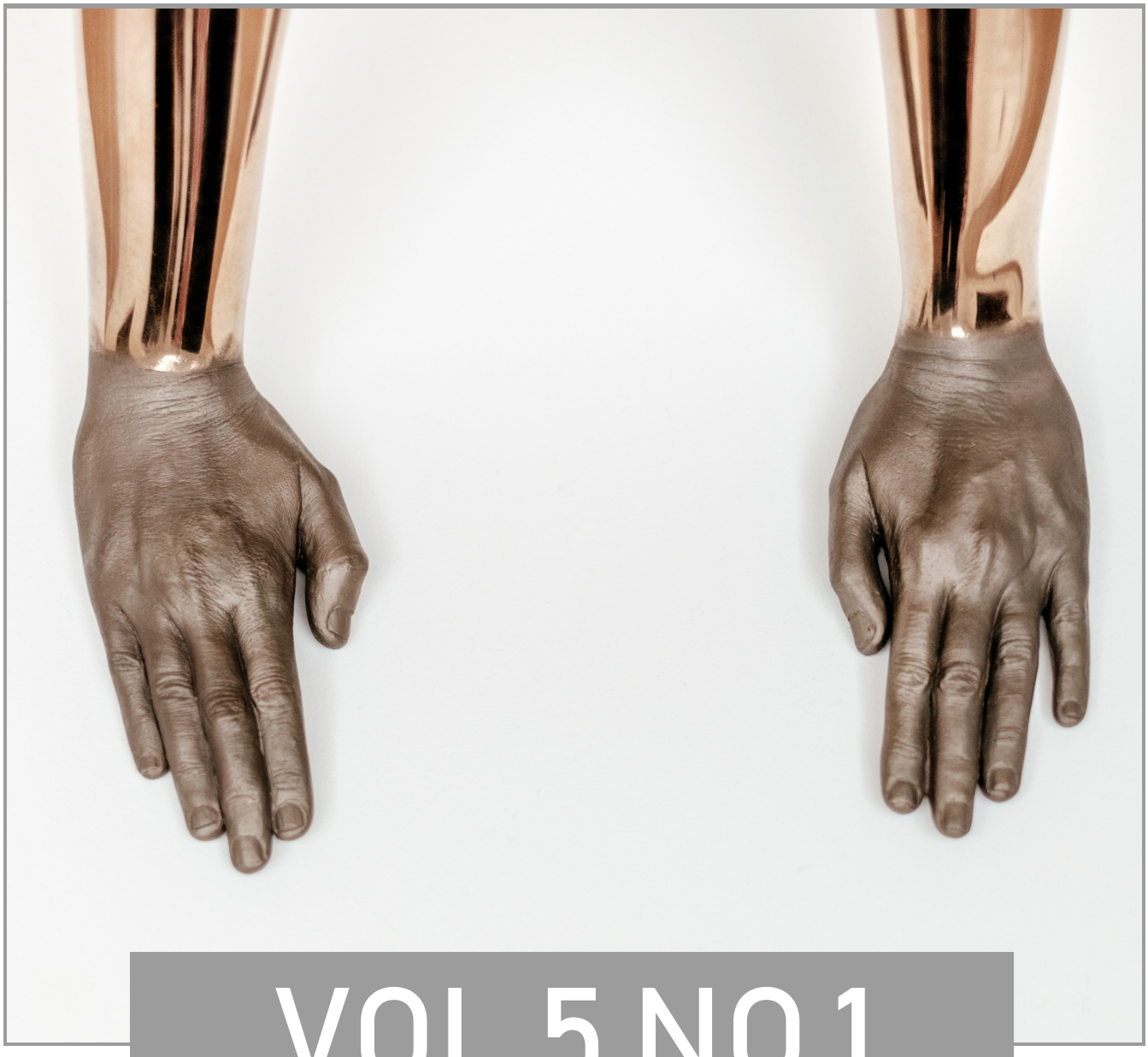
JISTA

Journal of Intelligent Systems: Theory and Applications

MARCH 2022

ISSN: 2651-3927



VOL 5 NO 1

ARTIFICIAL INTELLIGENT > MACHINE LEARNING > DEEP LEARNING

https://dergipark.org.tr/en/pub/jista

https://dergipark.org.tr/en/pub/jista

Journal of Intelligent Systems: Theory and Applications
2022 Volume: 5 Issue: 1

Editorial Boards

Editorial Assistants

https://dergipark.org.tr/en/pub/jista

Journal of Intelligent Systems: Theory and Applications

2022 Volume: 5 Issue: 1

Contents

Research Articles

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(1) (2022) 1-8
DOI: 10.38016/jista.950713

Yapay Zekâ Tabanlı Doğal Dil İşleme Yaklaşımını Kullanarak İnternet Ortamında Yayınlanmış Sahte Haberlerin Tespiti

Mesut Toğaçar[1]*, Kamil Abdullah Eşidir[2], Burhan Ergen[3]

[1] Fırat Üniversitesi, Bilgisayar Teknolojileri Bölümü, Teknik Bilimler Meslek Yüksekokulu, Elazığ, Türkiye

[2] Fırat Kalkınma Ajansı, Elazığ Yatırım Destek Ofisi, Elazığ, Türkiye

[3] Fırat Üniversitesi, Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Elazığ, Türkiye

mtogacar@firat.edu.tr, abdullahesidir@yahoo.com, bergen@firat.edu.tr

Öz

Sahte haber, bilinçli veya bilinçsiz bir şekilde çeşitli iletişim kanallarını kullanarak yayılan ve hiç bir gerçeklik payı olmayan uydurma haberlerdir. Günümüzde kitleler çoğu haberleri dijital ve sosyal medya üzerinden alıyorlar. Haberlerin hızlı bir şekilde kitlelere aktarabildiği bu tür iletişim ortamlarında çoğu zaman bu haberlerin doğruluğu suiistimal edinilebiliyor. Kökeni bilinmeyen haberler dezenformasyon veya yanlış bilgilendirme yapılarak toplumlarda ciddi sorunlar oluşturabilmektedir. Özellikle internet ortamında bilgi kirliliğine maruz kalan sahte haberler çok hızlı bir şekilde topluma etkisini gösterebilmektedir. Dijital ortamlarda bu tür problemlerin önüne geçilebilmesi için haberlerin doğruluğunu kavrayabilen ve hızlı bir şekilde teyit eden yapay zekâ tabanlı bir yaklaşım bu çalışmada önerilmektedir. Ayrıca, yapay zekânın bir alt dalı olan Doğal Dil İşleme (DDİ) yöntemi ile erişime açık veri setini kullanarak haberlerin gerçek veya sahte olduğunu tespit eden sınıflandırma analizi gerçekleştirildi. Veri seti, 6335 haber başlığı ve içerikten oluşmaktadır. Bu haberlerin 3171'i gerçek haber niteliği taşırken; 3164'ü ise sahte haber niteliği taşımaktadır. Çalışmanın analizinde DDİ yöntemi ile birlikte Uzun Kısa Süreli Bellek (UKSB) modeli kullanıldı ve veri setinin eğitimi bu model sayesinde gerçekleştirildi. Sonuç olarak, bu çalışmada eğitim verilerinden elde edilen genel doğruluk başarısı % 99,83 idi ve test verilerinden elde edilen genel doğruluk başarısı % 91,48 idi. Bu sonuçlar bize gösteriyor ki gelecekte düşünmeyi planladığımız benzeri çalışmalara umut verici olmuştur.

Anahtar kelimeler: Doğal Dil İşleme, Yapay Zekâ, Sahte Haber, Derin Öğrenme, Haber Sınıflandırma.

Detection of Fake News Published on the Internet Using Artificial Intelligence Based Natural Language Processing Approach

Abstract

Fake news is fabricated news that spread consciously or unconsciously through various communication channels and has no real share. Today, the masses receive most news on digital and social media. In such communication environments, where news can be transferred to the masses quickly, the accuracy of this news can often be abused. News of unknown origin can cause serious problems in societies by making disinformation or misinformation. Especially, fake news exposed to information pollution in the internet environment can show its effect on society very quickly. To prevent such problems in digital environments, an artificial intelligence-based approach that can grasp the accuracy of the news and confirm it quickly is proposed in this study. In addition, a classification analysis was performed using the Natural Language Processing (NLP) method, a sub-branch of artificial intelligence, to determine whether the news was real or false using the dataset that was accessible. The dataset consisted of 6335 news headlines and content. While 3171 of this news is real news; 3164 is fake news. In the analysis of the study, the Long Short Term Memory (LSTM) model was used together with the NLP method and the training of the dataset was carried out with this model. As a result, the overall accuracy success from the training data was 99.83%, and the overall accuracy success from the test data was 91.48%. These results show us that similar studies that we plan to think about in the future have been promising.

Keywords: Natural Language Processing, Artificial Intelligence, Fake News, Deep Learning, News Classification.

* Sorumlu yazar.
E-posta adresi: mtogacar@firat.edu.tr

Alındı : 10 Haziran 2021
Revizyon : 31 Ağustos 2021
Kabul : 22 Eylül 2021

1. Giriş (Introduction)

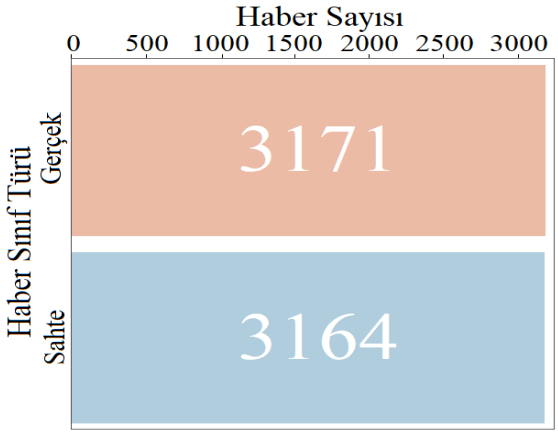
Geçmiş zamandan günümüze yanlış bilgiler ve sahte haberler çeşitli amaçları gerçekleştirme doğrultusunda kullanılmış ve süreç içerisinde tartışmalara yol açmıştır. Günümüzde ise bu durum internet ortamında daha hızlı bir şekilde hedef kitlelere ulaşmaktadır ve etkisi de kısa bir sürede görülmektedir. Zaman içerisinde anlık gelişen bilgilerin kontrolü zorlaşmaktadır ve internet ortamlarında paylaşılan bilgi ve haberler çoğu zaman kaynağı belirtilmeden yayınlanmaktadır. Bu durum sosyal medya üzerinden (Whatshapp, Facebook, Twitter vb.) kitlelere hızlı bir şekilde ulaşmaktadır ve kitleleri hızlı bir şekilde etkileşime geçirebilmektedir (Ünal ve Taylan, 2017). Haber kaynaklarının doğruluğunun teyit edilmesi ve yayınlanmadan önce belirli bir olgunluk seviyesine gelmesi gerektiği habercilik anlamında etik ve ahlaki bir kuraldır. Bu kural bazen kişiler tarafından bir an önce haberleştirme duygusu, hırs, çıkar ilişkisi, vb. sebepler ön plana çıktığı zaman bozabilmektedir (Figdor, 2017).

Bilgilerin ve haber içeriklerinin hızlı bir şekilde dezenformasyon olduğu bir internet ortamında bu durumu kontrol etmek zorunluluk haline gelmiştir. Küresel dünyada bu sürecin insanlar tarafından kontrolü zor ve imkânsızdır (Jayaseelan vd., 2020). Bu durumda insanların kontrol mekanizmasını makinelere aktaran bir sisteme ihtiyaç duyulmuştur ve adını yapay zekâ denilen kavram gelişmiştir. Yapay zekâ teknolojisi, insandaki idrak edebilme, düşünebilme, karar verme gibi mantıksal ve duyusal yetenekleri taklit ederek makineler tarafından bu eylemlerin gerçekleştirilmesini sağlar (Sun vd., 2020). Bu süreçte donanımları kontrol edebilen çeşitli yazılım dilleri (MATLAB, Python, R Studio vb.) geliştirilmiştir ve bu yazılımlar üzerinde yapay zekânın alt bir dalı olan derin öğrenme modelleri (AlexNet, GoogLeNet, ResNet, LSTM vb.) tasarlanmıştır (Doğan ve Türkoğlu, 2019). Milyarlarca kullanıcının anlık bulunduğu internet ortamında sahte haberlerin önüne geçebilmek için yapay zekâ tabanlı sistemlerin kullanılması zaruri bir ihtiyaç haline gelmiştir. Bu çalışmada hedefimiz, yapay zekâ teknolojisinin alt dalı olan Doğal Dil İşleme (DDİ) yöntemini kullanarak sahte haberlerin tespitini gerçekleştirmektir. Son zamanlarda yapay zekâ tabanlı sahte haber tespitini gerçekleştirmede birçok makaleler yayınlanmıştır. Bu makalelerden bazıları incelenirse; (Kaliyar vd., 2020) çalışmasında, sahte haber türlerinin sınıflandırılmasında FNDNet adını verdikleri bir derin öğrenme modeli geliştirdiler. Onlar, FNDNet modeli ile birlikte Hiper Parametre optimizasyon yöntemini de kullandılar ve elde ettikleri sınıflandırma başarısı % 98,36'di. (Altunbey Özbay ve Alataş, 2020), sahte haber verilerinin tespiti için çeşitli yapay zekâ modellerini kullandılar. Onlar, çalışmasında kullandıkları veri setini farklı oranlarda ayırarak modeller tarafından eğitimi gerçekleştirmişlerdir ve elde ettikleri sınıflandırma başarısı % 97,4'di. (S ve Chitturi, 2020) çalışmasında, sahte haber verilerinin tespiti için DDİ yaklaşımı ile birlikte UKSB modelini kullandılar. Onların bu çalışmadan elde ettikleri sınıflandırma başarısı % 91, 32'di. (Horne ve Adali, 2017) çalışmasında, iki sınıflı haber verisini kullanmışlardır. Onlar çalışmasında Kelime Hesap yöntemini kullanarak analizleri gerçekleştirdiler ve elde ettikleri sınıflandırma başarısı %71'di. (Onan, 2021) çalışmasında, kitlesel çevrimiçi kurslardan bireylerin duygu sınıflandırılmasını topluluk öğrenme ve derin öğrenme yaklaşımlarını kullanarak gerçekleştirmiştir. En iyi başarıyı UKSB ağları ve GloVe tekniğini birlikte kullanarak elde etmiştir. Duygu sınıflandırma sürecinde %95,80 oranında genel doğruluk başarısı sağlamıştır. (Yüksel ve Tan, 2018) çalışmasında, sosyal ağ veri tabanını kullanarak insanların görüşlerini gerçek zamanlı olarak yorumlayabilen akıllı karar destek sistemi önermişlerdir. Onlar çalışmasında özellik seçimi, gereksiz kelime çıkarımı gibi ön işlemleri el ile gerçekleştirmişlerdir. Duygu analizi için Google ara yüz uygulamasını kullanmışlar ve önerdikleri yaklaşımı İngilizce ve Türkçe kelimeler için ayrı ayrı uygulamışlar. Onlar çalışmasında Türkçe için %84,49 ve İngilizce için %95 sınıflandırma doğruluğu elde etmişler.

Bu makalenin bölüm özeti şu şekildedir; veri seti ile ilgili bilgiler ikinci bölümde yer almıştır. Derin öğrenme modelleri ve kullanılan yöntemler ile ilgili açıklamalar üçüncü bölümde yer almıştır. Deneysel analizler ve sonuçları dördüncü bölümde; tartışma ve sonuç bölümleri ise sırasıyla son iki bölümde yer almıştır.

2. Veri Seti (Dataset)

Veri seti dört sütun özellikten oluşmuş erişime açık "csv" uzantılı bir dosyadır. Sütun özelliklerinde; haber kimlik numarası, haber başlığı, haber içeriği ve haber durumu (sahte, gerçek) yer almaktadır. Haber durumu sütunu iki sınıflı bir haber içeriği olduğunu bize vermektedir. Haberler 2016-2018 yılları arasında uluslararası medyada gündem bulmuş internet verilerini içermektedir ve yazım dili İngilizce olarak yer almaktadır. Veri seti, Romanya Akademisi Biyokimya Enstitüsü'nde görev yapan araştırmacı Raluca Chitic tarafından Kaggle web sitesinde erişime açılmıştır (Chitic, 2021). 3171 adet gerçek haber içeriğinden ve 3164 adet sahte haber içeriğinden oluşmuştur. Toplamda 6335 adet haber başlığı ve içerikten oluşmuştur. Bu durum ile ilgili çubuk grafiği Şekil 1'de gösterildi. Veri setinin sütun özelliklerini ve örnek içeriklerini gösteren tasarım Şekil 2'de gösterildi. Bu çalışmanın deneysel analizinde veri setinin %25'i test verisi ve %75'i eğitim verisi olarak ayrıldı.



Şekil 1. Veri setinin istatistik bilgilerini gösteren çubuk grafiği (Bar graph showing statistics information of the dataset)

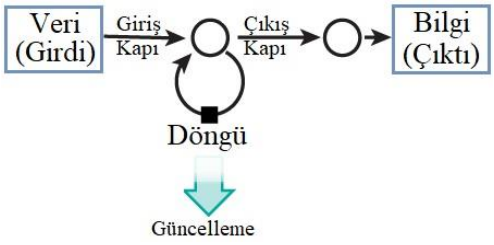
Haber No	Haber Başlık	Haber İçerik	Sınıf	
3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL	Gerçek
875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL	
95	'Britain's Schindler' Dies at 106	A Czech stockbroker who saved more than 650 Je...	REAL	
4869	Fact check: Trump and Clinton at the 'commande...	Hillary Clinton and Donald Trump made some ina...	REAL	
2909	Iran reportedly makes new push for uranium con...	Iranian negotiators reportedly have made a las...	REAL	
8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello..	FAKE	Sahte
10294	Watch The Exact Moment Paul Ryan Committed Pol	Google Pinterest Digg Linkedin Reddit Stumbleu..	FAKE	
10142	Bernie supporters on Twitter erupt in anger ag...	Kaydee King (@KaydeeKing) November 9, 2016 T..	FAKE	
6903	Tehran, USA	\nI'm not an immigrant, but my grandparents ..	FAKE	
7341	Girl Horrified At What She Watches Boyfriend D...	Share This Baylee Luciani (left), Screenshot o..	FAKE	

Şekil 2. Veri setinin haber sınıflarını gösteren örnek içerik şeması (Sample content diagram showing the news classes of the dataset)

3. Derin Öğrenme Modeli, Yöntemler (Deep Learning Model, Methods)

3.1. Uzun Kısa Süreli Bellek Modeli (Long Short Term Memory Model)

UKSB modeli tekrarlayan sinir ağı (TKA) modelinden türetilmiş bir modeldir. TKA modelinin yapısında döngüler vardır ve bu döngüler sayesinde geri bildirimler önceki ve sonraki katmanlar arasında olunabilir. Böylece kısa bilgiler TKA modeli tarafından öğrenilebilir. Ancak uzun süreli bilgiler TKA modeli için bir sorundur ve çözüm için UKSB modeli tasarlanmıştır. UKSB modeli TKA modelinin aksine tutulması gereken bilgiler tek bir sinir ağı kullanmak yerine özel ve çok katmanlı sinir ağları kullanarak doğrusal etkileşimle çalışır. Girdi verileri doğrusal zincir boyunca kolayca katmanlar arasında geçiş sağlar. UKSB modeli katmanlar arasında bilgi geçişlerini özel kapılar ile gerçekleştirir. UKSB modelinde eğer bir bilgi ağ yapısında kullanılacaksa "giriş kapıları" ile bilgilerin aktarımı sağlanır. Eğer bilgiler gereksizce veya hafızada yeterli bir alan oluşturulacak ise "unutma kapıları" kullanılır. UKSB tarafından "unut ve giriş" kapıları birleştirilecekse bu durumda "güncelleme kapıları" kullanılır. Neticede, UKSB modeli TKA modeli gibi sadece görüntüleri kullanarak bilgi aktarımı gerçekleştirmez; video, ses gibi verilerde de bu işlemi başarılı bir şekilde gerçekleştirir ve işlenen verilerde "çıkış kapılarına" aktararak tahmin, sınıflandırma adımlarını gerçekleştirir (Le vd., 2019). UKSB modelinin işlem adımlarını gösteren tasarım Şekil 3'te gösterildi.



Şekil 3. UKSB modelinin işlem adımlarını gösteren tasarım (Design showing the process steps of the LSTM model)

Bu çalışmada kullanılan UKSB modelinin katmanları ile ilgili temel bilgiler Tablo 1'de verildi.

Veri analizi için kullanılan metinsel ifadeler UKSB modelinin Girdi katmanında işlenir. LSTM katmanları ile işlenen veriler yoğun bir şekilde elde edilen özellikler Yoğunluk katmanına girdi olarak aktarılır. Yoğunluk katmanı ise kendi içerisinde aktivasyon fonksiyonlarını kullanarak (Softmax, ReLU, Sigmoid) (Wang vd., 2020) çıkış parametresini düşürtür ve haber verisinin sınıflandırma olasılığını çıktı olarak aktarır. Biz bu çalışmada üç aktivasyon fomksiyonunu sırasıyla Yoğunluk katmanlarında kullandık. Modelin eğitiminde optimizasyon algoritması olarak Adam (Bock ve Weiß, 2019) yöntemi tercih edildi ve öğrenme oranı ise 10^{-5} seçildi.

Tablo 1. UKSB modelinin bu çalışmada tercih edilen katman ve parametre bilgileri (Preferred layer and parameter information of the LSTM model in this study)

Katman	Çıkış Şekli	Parametre Sayısı
Embedding (Girdi)	(300, 100)	1,000,000
LSTM (UKSB)	(300, 128)	117,248
LSTM (UKSB)	64	49,408
Dense (Yoğunluk)	32	2,080
Dense (Yoğunluk)	16	528
Dense (Yoğunluk)	1	17

3.2. Doğal Dil İşleme Yöntemi (Natural Language Processing)

DDİ, sosyal medya internet ortamındaki verilerin anlamlı ve istenen bilgilerin dijital cihazlar tarafından öğrenilmesini sağlayan yapay zekâ destekli bir yöntemdir. DDİ yöntemi akıllı uygulamalarda (sohbet botları, dil çeviri ortamları, bilgi özetleme, spam tespiti, intihal yazılımları vb.) kullanılmaktadır (Ong vd., 2020). DDİ yöntemi ön işlem adımları olarak; kelime normalizasyonu, gürültü azaltma ve nesne standartı oluşturma tekniklerini kullanır. Kelime normalizasyon tekniğinde, aynı kelime kökünden gelen (örneğin; gidiyorum, gidecek, gittim vb.) sözcükleri normalizasyonunu gerçekleştirir. Gürültü azaltma tekniğinde; bağlaç sözcüklerini (ve, veya, ile vb.) cümleler içerisinde tespitini edebilmek ve anlam karmaşasını ortadan kaldırmaktadır. Nesne standart oluşturma tekniğinde ise; toplumsal veya dijital ortamlarda kullanılan kısaltmaların (örneğin; RT: retweet) anlamlarını tespit eden bir tekniktir. DDİ yöntemi ön işlem tekniklerinin ardından, kelimeleri oluşturan cümlelerin özne-nesne-yüklem ilişkisini tespit eden varlık çıkarma sürecini gerçekleştirir. Bunun dışında cümlelerin içerisinde geçen kelimelerin frekans bilgileri, sayısı gibi bilgiler de DDİ yöntemi ile tespit edilebilmektedir (Zhou vd., 2020). DDİ yönteminin işlem adımları Şekil 4'te gösterildi.

Şekil 4. Doğal dil işleme yönteminin adımları (Steps of the natural language processing method)

DDİ yönteminin adımları arasında kelime normalizasyonu, ilk adımı oluşturmaktadır. Bu adım "okudum, okuyorum, okuyacağım" gibi aynı kökten türeyen kelimelerin normalize oluşumu üzerinde işlemler gerçekleştirir. Gürültü giderme adımı genel olarak cümlede kullanılan bağlaçlar (ve, veya, ya da, ama, vb.) üzerine işlemlerin yürütülmesi ile alakalıdır. Bu tür bağlaçlar cümle içerisinde gereksiz kullanılmaktadır. Bunun tespitinde cümlelerin genlik değeri ile bağlaçların genlik değerleri arasında bir oran belirlenir. Bu oran sayesinde gürültü genliğinin artması engellenir (Adalı E, 2016). Nesne standart adımı herkes tarafından kabul edilmiş kısaltmaların (dm: direct message, rt: retweet, vs.) üzerinde yapılması öngörülen bir tekniktir. Bu üç adımın verilere uygulanması ile "ön işleme" süreci tamamlanmış olur ve "Varlık Çıkarma" adımına geçiş sağlanır. "Varlık Çıkarma" adımında ilk uygulanacak işlem; cümlenin özne, nesne ve yüklem kelimelerinin belirlenmesi ve ayırt edilmesidir. Varlık Çıkarma tekniğini gerçekleştiren, Python dilinde tasarlanmış açık kaynak kodlu kütüphaneler mevcuttur. Bu çalışma için "nltk" kütüphanesinin kod parametreleri kullanıldı. Son adımda "Glove" tekniği kullanılarak verilerin istatistiksel bilgileri, kelime sayısı, yoğunluğu, kelime bulut gösterimi gibi bilgilerin elde edilmesi sağlandı. Kısacası "Glove" tekniği olasılık istatistiklerinden yararlanarak veri setinin istatistiksel bilgilerini vermeye yarar. Bu işlemin gerçekleşmesinde Eşitlik (1)'de ifade edilen matematiksel formül kullanır. Bu eşitlikte, kelimelerin birlikte bulunma sayımlarının matris değeri X ile temsil edilmektedir. X_{ij} değişkeni, i. kelime bağlamında j. kelimenin kaç defa geçtiğini tablo halinde gösterimini sağlar. V değişkeni, kelime dağarcığının boyutunu temsil eder. Son olarak denklemde kullanılan b değişkeni, ön ayar değerini temsil ederken; W değişkeni ise kelime vektörlerini temsil eder (Pennington, Socher, ve Manning, 2015).

$$J = \sum_{i,j=1}^{V} f(X_{ij})(W_i^T W_J + b_i + b_j - \log X_{ij})^2 \quad (1)$$

Bu çalışmada DDİ yöntemi olarak Python yazılım dili kullanıldı ve "nltk" kütüphanesi kullanıldı ve varlık çıkarma işlemlerin gerçekleştirilmesinde "Glove" tekniği kullanıldı. Glove tekniği, kelimelerin vektörel

olarak gruplanmasını sağlayan ve vektör temsillerini elde edebilmek için denetimsiz bir öğrenme gerçekleştirmektedir (Pennington, Socher, ve Manning, 2015). Ayrıca, bu çalışmada kullanılmış DDİ yöntemi ile ilgili açık kaynak kodlar Python dilinde tasarlandı (Madz, 2021).

4. Deneysel Analiz (Experimental Analysis)

Çalışmanın deneysel analizleri Google Colab üzerinden gerçekleştirilmiştir ve kullanılan yazılım dili Python 3.6'dı. Python kodlarının tasarlanması ve derlenmesi Jupyter Notebook ara yüzü kullanılarak gerçekleştirildi. Deneysel analizler için karmaşıklık matrisi kullanıldı. Karmaşıklık matrisinin hesaplanmasında kullanılan metrikler ise; Duyarlılık (Duy), Özgüllük (Özg), Hassasiyet (Has), F-skoru (F-skr) ve Doğruluk (Dğr)'dir. Bu metriklerin hesaplanmasında Eşitlik (2) ile Eşitlik (6) arasındaki formüller kullanıldı. Eşitliklerde kullanılan değişkenler ise; Doğru Pozitif (DP), Doğru Negatif (DN), Yanlış Pozitif (YP), Yanlış Negatif (YN)'dir (Demir 2021; Sertkaya, Ergen, ve Togacar, 2019).

$$\text{Duy} = \frac{\text{DP}}{\text{DP+YN}} \quad (2)$$

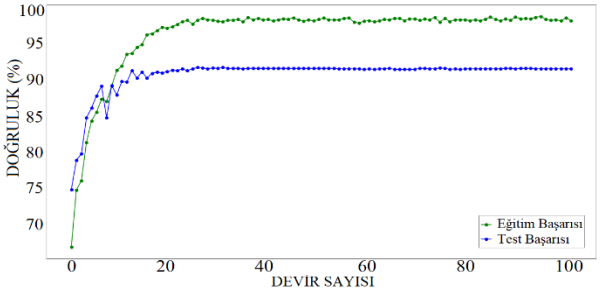
$$\text{Özg} = \frac{\text{DN}}{\text{DN+YP}} \quad (3)$$

$$\text{Has} = \frac{\text{DP}}{\text{DP+YP}} \quad (4)$$

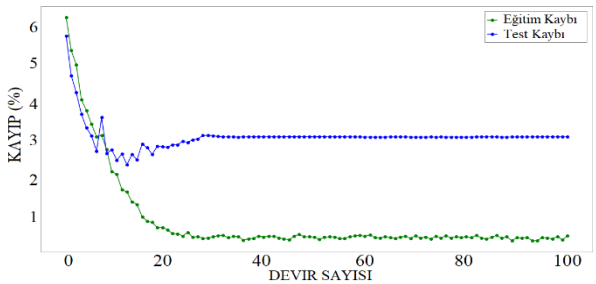
$$\text{F-skr} = \frac{2\text{xDP}}{2\text{xDP+YP+YN}} \quad (5)$$

$$\text{Dğr} = \frac{\text{DP+DN}}{\text{DP+DN+YP+YN}} \quad (6)$$

Deneysel analizde kullanılan veri setinin %25'i test verisi olarak ayrıldı. Derin öğrenme modeli için tercih edilen mini-batch parametresinin değeri 128 seçildi. Mini-batch, eğitimin gerçekleştirildiği donanımsal özellikler ile doğrudan bağlantılıdır ve bu parametre ile modelin eğitim esnasında aynı anda ne kadar veri işleneceğini bize verir (Yang vd., 2019). Deneyin eğitiminde tercih edilen devir sayısı 100 seçildi. Eğitim verilerinin analizinde elde edilen genel doğruluk başarısı %99,83'di ve test verilerinden elde edilen genel doğruluk oranı ise %91,48'di. Eğitim-test başarısını gösteren grafik Şekil 5'te gösterildi. Eğitim-test kaybını gösteren grafik ise Şekil 6'da gösterildi. Bu çalışmanın test verilerinin diğer analiz sonuçları ise; Duyarlılık başarısı %92,76'di, Özgüllük başarısı %90,25'di, Hassasiyet başarısı %90,09 ve F-skor başarısı %91,41'di. Analiz hesaplamaların gerçekleştirildiği karmaşıklık matrisi Şekil 7'de gösterildi ve analiz sonuçları ise Tablo 2'de verildi.



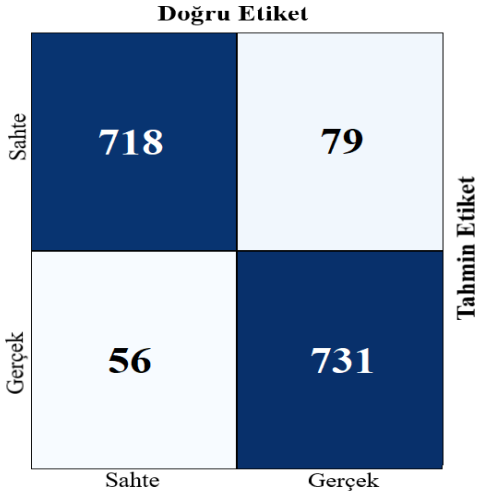
Şekil 5. Bu çalışmanın eğitim-test verilerine ait başarı grafikleri (The success graphs of the training-test data of this study)



Şekil 6. Bu çalışmanın eğitim-test verilerine ait kayıp grafikleri (The loss graphs of the training-test data of this study)

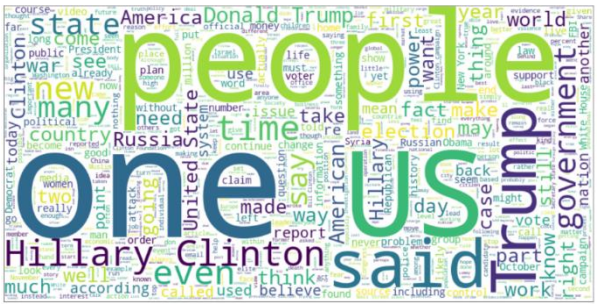
Tablo 2. Deneyin karmaşıklık matrisinden elde edilen analiz sonuçları (Analysis results obtained from the confusion matrix of the experiment)

Model ve Yöntem	Sınıf	F-Skr. (%)	Duy. (%)	Özg. (%)	Has. (%)	Dğr. (%)	Genel Dğr. (%)
UKSB & DDİ & Glove	Sahte	91,41	90,09	92,88	92,76	91,48	91,48
	Gerçek	91,55	92,88	90,09	90,25	91,48	



Şekil 7. Deneysel analiz sonucu elde edilen karmaşıklık matris grafiği (Confusion matrix graph obtained as a result of experimental analysis)

Deneyin son aşamasında sahte ve gerçek haberlerin kelime başlıklarının yer aldığı kelime bulut görseli oluşturuldu. Buradaki amacımız sahte haberlerin etiket görselleri ile gerçek haberin etiket görselini karşılaştırmaktı. Kelime bulutlarının oluşturulmasında python yazılımında yer alan "Word Cloud" parametresi kullanıldı. Kelime bulutlarının sağladığı yarar, istenilen anahtar kelimelerin okuyucuya görselleştirilerek sunulmasıdır. Böylece daha çok okur tarafından ilgi çekilmesi amaçlanmaktadır (Atenstaedt, 2012). Sahte haberlerin kelime bulut şeması Şekil 8'de gösterildi ve gerçek haberlerin kelime bulut şeması Şekil 9'da gösterildi. Şekil 8 ve Şekil 9 karşılaştırıldığında okuyucu gözüne doğrudan etkileyen büyük etiketler birbirinden farklıdır. Bu durum haberlerin sahte ve gerçek olma durumları ile doğrudan ilgilidir.



Şekil 8. Sahte haberlerin kelime bulut gösterimi (Word cloud representation of fake news)

Şekil 9. Gerçek haberlerin kelime bulut gösterimi (Word cloud representation of real news)

5. Tartışma (Discussion)

Sahte web haberlerinin bilinçli veya bilinçsiz bir şekilde sayısının son zamanlarda artması toplumlarda kaosa kadar sürüklenebilecek bir problem haline gelmiştir. Bu tür durumlarda insanlar tarafından önlem alınması milyarlarca kitlesi olan sosyal ve dijital medyada önüne geçilemez bir vaziyettir. Çünkü saniyesinde kontrol edilmeyen haber hedefindeki kitlelere hızlı bir şekilde yayılabilmektedir. Peki, bu durumun önüne nasıl geçilir veya nasıl minimize edilir? Aslında problemin çözümü problemin oluştuğu kaynağın kendisindedir. Yani yine çözüm sosyal ve dijital medyayı kontrol edebilen milyarlarca bilgiyi anlık bir zamanda süzebilen yazılımlar ile çözülebilir. Bizde bu yaklaşıma uyarak çalışmamızda yapay zekâ tabanlı UKSB ve DDİ yöntemlerini modelleyerek sahte haberleri algılayabilen bir yaklaşım sunduk. Çalışmamıza benzer model ile katkı sunan bir çalışmada (S ve Chitturi, 2020) tarafından 2018 yılında gerçekleştirmiştir. Bu çalışma ile ilgili analiz sonucu Tablo 3'te verildi.

Tablo 3. Benzer yaklaşım ile gerçekleştirilmiş çalışmaların karşılaştırılması (Comparison of studies carried out with a similar approach)

Makale	Yıl	Test Verisi (%)	Model / Yöntem	Dğr. (%)
(Sreekumar ve Chitturi)	2020	20	UKSB, DDİ	91,32
Bu çalışma	2021	25	UKSB, DDİ	91,48

(Sreekumar ve Chitturi, 2020) çalışmasında "George McIntires Fake News" açık erişimli veri setini (McIntire, 2017) kullandılar. Onlar, UKSB ve DDİ yöntemlerini modelleyerek sahte haberlerin tespitini %91,32 oranında bir başarı ile gerçekleştirdiler. Onların kelimeler üzerinde özellikler çıkartarak bu başarıyı elde ettiler. Ayrıca çalışmalarında test verisi oranı onların %20'di. Bizim çalışmamızda ise bu oran %25'di. Bizim eğitim veri oranımız onların eğitim verisi oranına göre daha düşüktür; fakat daha iyi performans elde edilmiştir. Bu sınıflandırma farklılığı modeller için seçilmiş parametreler ve değerleri ile alakalı olduğunu düşünüyorum. Ayrıca onların kullandığı veri seti sayısı

bizim kullandığımız veri setinden düşüktü. Sonuçta iki çalışmada benzer modeller ile farklı veri seti türünden yakın doğrulama sonucu almışlardır.

Çalışmamızın test başarısını artırmak için farklı optimizasyon yöntemlerini de (SGD, Adadelta vb.) analiz ettik; fakat Adam yöntemi ile elde edilen başarının üzerinde bir sonuç elde edemediğimiz için diğer optimizasyon yöntemlerinin sonuçlarını analizlere eklemedik. Bunun dışında UKSB modelinin yoğunluk katman sayısını artırmayı denedik; fakat bu denememizde başarıyı artırmadığı için en ideal model katmanları ile sınıflandırmayı gerçekleştirdik.

6. Sonuçlar (Conclusions)

Yakın zamana kadar toplumlar görsel ve basılı medya aracılığı ile haberleri takip edebilmekteydiler. Fakat son zamanlarda bu durum sosyal ve dijital medya üzerinden yürütülmektedir. Haberlerin aktarıldığı bu tip iletişim ortamlarında hızlı bir şekilde bilgiler yayıldığı için etkisi de büyük olabilmektedir. Bu makale ile sosyal ve dijital medya ortamlarında sahte haberlerin önüne geçebilmek için yapay zekâ tabanlı bir analiz gerçekleştirilmiştir. Çalışmanın amacı sahte haberlerin tespitinde ileri teknolojik yaklaşımları bir arada kullanarak başarılı bir şekilde gerçekleştirilmesini sağlamaktı. Bizde, UKSB ile beraber DDİ yöntemini modelleyerek haber içerikli veri setinin analizini gerçekleştirdik. Eğitim verilerinde elde edilen doğruluk başarısı %99,83'di ve test verilerinden elde edilen doğruluk başarısı %91,48'di. Bu sonuçlar bize sosyal platformlar üzerinde gerçekleştireceğimiz çalışmalara umut verici oldu.

Gelecek çalışmada, Türkçe haber web sitelerinden oluşturulmuş veri setleri üzerinden sahte haberlerin ve yanlış bilgilerin tespitini farklı yaklaşımlar ile birlikte analizleri gerçekleştirilecektir.

Kaynaklar (References)

Adalı, E., 2016, "Doğal Dil İşleme". Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi , 5 (2). https://dergipark.org.tr/tr/pub/tbbmd/issue/22245/238797

Altunbey Özbay, F., ve Alataş B, 2020, "Çevrimiçi Sosyal Medyada Sahte Haber Tespiti.", DÜMF Mühendislik Dergisi, 11 (1): 91–103. https://doi.org/10.24012/dumf.629368.

Atenstaedt, R., 2012, "Word Cloud Analysis of the BJGP.", British Journal of General Practice, 62 (596): 148 LP – 148, https://doi.org/10.3399/bjgp12X630142.

Bock, S., ve Weiß M., 2019, "A Proof of Local Convergence for the Adam Optimizer.", In 2019 International Joint Conference on Neural Networks (IJCNN), 1–8, https://doi.org/10.1109/ijcnn.2019.8852239.

Chitic, R., 2021, "REAL ve FAKE News Dataset." Kaggle, 2021, https://www.kaggle.com/rchitic17/real-or-fake.

Demir, F., 2021, "DeepCoroNet: A Deep LSTM Approach for Automated Detection of COVID-19 Cases from Chest X-Ray Images.", Applied Soft Computing, 103: 107160. https://doi.org/https://doi.org/10.1016/j.asoc.2021.107160.

Doğan, F., ve Türkoğlu İ., 2019, "Derin Öğrenme Modelleri ve Uygulama Alanlarına İlişkin Bir Derleme", DÜMF Mühendislik Dergisi 10 (2): 409–45, https://doi.org/10.24012/dumf.411130.

Figdor, C., 2017. "(When) Is Science Reporting Ethical? The Case for Recognizing Shared Epistemic Responsibility in Science Journalism.", Frontiers in Communication, 2: 3, https://doi.org/10.3389/fcomm.2017.00003.

Horne, Benjamin D., ve Adali S., 2017, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News.", http://arxiv.org/abs/1703.09398.

Kaliyar, Rohit K., Anurag G., Pratik N., ve Soumendu S., 2020, "FNDNet – A Deep Convolutional Neural Network for Fake News Detection.", Cognitive Systems Research, 61: 32–44. https://doi.org/https://doi.org/10.1016/j.cogsys.2019.12.005.

Le, Xuan H., Hung Viet H., Giha L., ve Sungho J., 2019, "Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting.", Water (Switzerlve), 11 (7). https://doi.org/10.3390/w11071387.

Madz, 2021, "NLP Using GloVe Embeddings." Kaggle, 2021, https://www.kaggle.com/madz2000/nlp-using-glove-embeddings-99-8-accuracy.

McIntire, G., 2017, "Machine Learning Finds 'Fake News' with 88% Accuracy." KD Nuggets, 2017, https://www.kdnuggets.com/2017/04/machine-learning-fake-news-accuracy.html.

Ong, Charlene J., Agni O., Rebecca Z., Francois Pierre M. C., Meghan H., Liang M., Darian F., vd., 2020, "Machine Learning ve Natural Language Processing Methods to Identify Ischemic Stroke, Acuity ve Location from Radiology Reports.", PLOS ONE, 15 (6): e0234908. https://doi.org/10.1371/journal.pone.0234908.

Onan A., 2021, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach.", Comput. Appl. Eng. Educ, 29: 572–589. doi:10.1002/cae.22253.

Pennington, J., Socher R., ve Manning C.D., 2015, "GloVe: Global Vectors for Word Representation." Stanford University, 2015, https://nlp.stanford.edu/projects/glove/.

Jayaseelan R., Brindha D., ve Kades W, 2020, "Social Media Reigned by Information or Misinformation About COVID-19: A Phenomenological Study.", SSRN Electronic Journal, https://doi.org/10.2139/ssrn.3596058.

Sreekumar D., ve Chitturi B., 2020, "Deep Neural Approach to Fake-News Identification.", Procedia Computer Science, 167: 2236–43, https://doi.org/https://doi.org/10.1016/j.procs.2020.03.276.

Sertkaya, M. E., Ergen B., ve Togacar M., 2019, "Diagnosis of Eye Retinal Diseases Based on Convolutional Neural Networks Using Optical Coherence Images.", In 2019 23rd International Conference Electronics, 1–5. https://doi.org/10.1109/electronics.2019.8765579.

Sun, Shaojing, Yujia Zhai, Bin Shen, ve Yibei Chen. 2020. "Newspaper Coverage of Artificial Intelligence: A Perspective of Emerging Technologies." Telematics ve

Informatics, 101433. https://doi.org/https://doi.org/10.1016/j.tele.2020.101433.

Ünal, R., ve Taylan A., 2017, "Sağlık İletişiminde Yalan Haber - Yanlış Enformasyon Sorunu ve Doğrulama Platformları.", Atatürk İletişim Dergisi / Dergi Park. https://dergipark.org.tr/tr/pub/atauniiletisim/issue/34005/360148.

Wang, Y., Li Y., Yong S., ve Rong X., 2020, "The Influence of the Activation Function in a Convolution Neural Network Model of Facial Expression Recognition.", Applied Sciences (Switzerlve), 10 (5). https://doi.org/10.3390/app10051897.

Yang, Z., Wang C., Zhang Z., ve Li J., 2019, "Mini-Batch Algorithms with Online Step Size." Knowledge-Based Systems, 165: 228–40. https://doi.org/10.1016/j.knosys.2018.11.031.

Yüksel A.S., Tan F.G., 2018, "A real-time social network-based knowledge discovery system for decision making", Automatika., 59: 261–273. https://doi.org/10.1080/00051144.2018.1531214.

Zhou, M., Nan D., Shujie L., ve Heung-Yeung S., 2020, "Progress in Neural NLP: Modeling, Learning, ve Reasoning.", Engineering, 6 (3): 275–90. https://doi.org/10.1016/j.eng.2019.12.014.

Research Article.
Journal of Intelligent Systems: Theory and Applications 5(1) 2022: 9-15
DOI: 10.38016/jista.922663

Estimation of High School Entrance Examination Success Rates Using Machine Learning and Beta Regression Models

Tuba Koç[1], Pelin Akın[2*]

[1,2] Çankırı Karatekin University, Faculty of Science, Department of Statistic, Çankırı/Turkey

tubakoc@karatekin.edu.tr , pelinakin@karatekin.edu.tr

Abstract
Education is the foundation of economic, social, and cultural development for every individual and society as a whole. Students are accepted to secondary education institutions with the high school entrance examination made by the Ministry of National Education in Turkey. In this study, the success rates of the students who took the high school entrance examination in Turkey's 81 provinces in 2019 were handled with the machine learning regression and beta regression model. The present paper aimed to model, predict, and explain students' success rates using variables such as divorce rate, gross domestic product, illiteracy, and higher education populations. Support vector regression, random forest, decision tree, and beta regression model were applied to estimate success rates. Two models with the highest R^2 value were found to be beta regression and random forest models. When the prediction errors of beta regression and random forest model were examined, it seemed to be that the random forest model is relatively superior to the beta regression model in predicting the success rates. While the beta regression model was the best predictor of the success rates of Çanakkale province, the random forest model predicted the success rates of Ankara well. Also, it was seen that the variables found to be significant in the beta regression model for success rates were also crucial in the random forest model. It is recommended to use both the beta and random forest models to estimate the students' success rates.

Keywords: Success rate of exam, Beta regression, Random forest, Classification and regression tree, Support vector regression

Makine Öğrenimi ve Beta Regresyon Modelleri Kullanılarak Lise Giriş Sınavı Başarı Oranlarının Tahmini

Öz
Eğitim, her birey ve bir bütün olarak toplum için ekonomik, sosyal ve kültürel gelişimin temelidir. Ülkemizde orta öğretim kurumlarına Milli Eğitim Bakanlığı tarafından yapılan lise giriş sınavı ile öğrenci kabul edilmektedir. Bu çalışmada, 2019 yılında Türkiye'nin 81 ilinde lise giriş sınavına giren öğrencilerin başarı oranları makine öğrenimi regresyon ve beta regresyon modeli ile ele alınmıştır. Bu makale, boşanma oranı, gayri safi yurtiçi hasıla, okuma yazma bilmeyenlerin sayısı ve yüksek öğretim nüfusu sayısı gibi değişkenleri kullanarak öğrencilerin başarı oranlarını modellemeyi, tahmin etmeyi ve açıklamayı amaçlamaktadır. Başarı oranlarını tahmin etmek için destek vektörü regresyon, rastgele orman, karar ağacı ve beta regresyon modeli uygulanmıştır. En yüksek R^2 değerine sahip iki modelin beta regresyon ve rastgele orman modelleri olduğu bulunmuştur. Beta regresyon ve rastgele orman modelinin tahmin hataları incelendiğinde, başarı oranlarını tahmin etmede rastgele orman modelinin beta regresyon modeline göre nispeten üstün olduğu görülmektedir. Beta regresyon modeli Çanakkale ilinin başarı oranlarının en iyi yordayıcısı iken, rastgele orman modeli Ankara'nın başarı oranlarını iyi tahmin etmiştir. Ayrıca beta regresyon modelinde başarı oranları için anlamlı bulunan değişkenlerin rastgele orman modelinde de önemli olduğu görülmüştür. Öğrencilerin başarı oranlarını tahmin etmek için hem beta hem de rastgele orman modellerinin kullanılması önerilir.

Anahtar kelimeler: Sınav başarı oranı, Beta regresyon, Rastgele orman, Sınıflandırma ve regresyon ağacı, Destek vektör regresyonu.

* Corresponding author.
E-mail: pelinakin@karatekin.edu.tr

Recieved : 20 April 2021
Revision : 30 June 2021
Accepted : 22 October 2021

1. Introduction

Education is one of the influential factors that provide social, cultural, and economic development. The economic development of a country and the progress of society take place with qualified workforce. Realizing this is one of the basic functions of education. Increasing the level of education is achieved by making that country more knowledgeable, skilled, and equipped. The education systems of the countries bring some political implications and obligations. Developments in science and technology cause changes in the needs of individuals. It is possible to train a qualified workforce that can adapt to the speed of developing technology as behaviour by innovating countries' education systems. Through innovations in the educational systems, countries can raise qualified human power that can adapt their behaviour to technological developments (MEB, 2018). For this purpose, education programs are prepared to raise qualified individuals who can solve problems, think critically, entrepreneurs, and contribute to society (MEB, 2018). For that purpose, curriculums are developed to raise individuals who can solve problems, think critically, and actively contribute to society. Factors affecting success in education have always been among the issues that are emphasized. At the stage of achieving success, the factors affecting this should be at a level that will create success at the maximum level to be good.

Measurement and evaluation are crucial in education systems as in other systems. Different countries use different variables in student selection for transition to high schools. It is seen that these variables are school graduation exams, central selection exams, school-based selection exams, school grades, and teachers' opinions (Gür et al., 2013). In Turkey, central selection exams are held in student selection for transition to high schools.

In recent years, using machine learning algorithms in the field of education is a general approach. Abbasoğlu (2020) analysed the effects of middle school students' demographic characteristics and socioeconomic status on their year-end general achievement averages using data mining methods. Gök (2017) estimated the end-of-term achievement averages of secondary school students using logistic regression and multi-class machine learning models. According to the results, both logistic regression and classification methods successfully estimate the average success score. Uskov et al. (2019) examined the machine learning predictions of student academic performance in STEM (Science, Technology, Engineering, and Mathematics) education. Abidi et al. (2019) investigated models for predicting confused students who try to do homework using ITS (Intelligent tutoring systems). In their studies, they used naïve Bayes (NB), generalized linear model (GLM), logistic regression (LR), deep learning (DL), decision tree (DT), random forest (RF), and gradient boosted trees (XGBoost) machine learning models. As a result, they showed that the RF, GLM, XGBoost and DL models achieved a high accuracy in predicting students' confusion in the algebra mastery skills in ITS. Using machine learning algorithms, Al Mayahi and Al-Bahri (2020) predicted whether university students would pass a particular course based on previous academic achievement data. In the study, the accuracy rate of the model was found to be 87%. Sethi et al. (2019) used three different machine learning algorithms in the subject/stream selection of middle school students. Rebai et al. (2020) investigated the success of secondary school education in Tunisia with a two-level algorithm. The study used decision trees and random forest algorithms to provide input for the data envelopment analysis (DEA) method. Rajak et al. (2020) applied different classification machine learning algorithms to data sets with characteristics such as family education, father's job, school attendance and calculated the model's performance. Yousafzai et al. (2020) used machine learning and data mining methods to predict students' performance at the secondary education level.

This study aims to show the use of machine learning in the education area. Machine learning algorithms and beta regression models were applied and compared to calculate the success rates of students in the high school entrance examination. The article is divided as follows: In section 2, the beta regression and machine learning algorithms are defined. Section 3 explains the application of the beta regression model and machine learning algorithms with success rate data. Finally, a brief discussion is given in Section 4.

2. Material and Methods

In this section, beta regression and machine learning methods used to estimate the success rates of high school entrance examinations are presented

2.1. Beta Regression Model

The beta regression model is widely used to model variables in the range (0, 1). This model is very flexible and can be used for random variables such as ratio and percentage (Ferrari and Cribari-Neto, 2004). It is commonly used in fields such as education, finance, and social sciences. (Cepeda-Cuervo, 2015).

The beta density is given by

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1 \quad (1)$$

Where Γ (.) is the gamma function p, q>0.

Ferrari and Cribari-Neto (2004) proposed a different parameterization with $\mu = \frac{p}{p+q}$ and $\phi = p + q$:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\varphi)\Gamma\big((1-\mu)\varphi\big)} y^{\mu\phi-1}(1-y)^{(1-\mu)\varphi-1}$$

$$,0<y<1 \quad , 0 < \mu < 1 \text{ and } \phi > 0 \quad (2)$$

$$y \sim B(\mu, \phi) \quad \text{and} \quad E(y) = \mu, \mathrm{Var}(y) = \frac{\mu(1-\mu)}{1+\phi}$$

μ is the mean of the response variable, and ϕ can be interpreted as a dispersion parameter for fixed μ.

Let $y_1, y_2, \dots, y_n$ be independent random variables each $y_t \sim B(\mu_t, \phi_t),\ t = 1, \dots, n$

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i \tag{3}$$

Where $\beta = (\beta_1, \beta_2, \dots\dots\dots, \beta_k)^T$ represent unknown regression parameters. x= (x_{t1},x_{t2},…,x_{tk}) denotes fixed covariates g(.) shows the link function, which strictly monotonic and twice differentiable. The Beta regression model can use different link functions such as log, logit, etc. (Dünder and Cengiz, 2020).

2.2. Machine Learning Regression Models

This study used support vector regression, decision tree, and random forest regression from machine learning regression models.

a) Support Vector Regression

Support vector machines, which have been suggested to resolve classification and regression problems, are supervised learning techniques based on statistical learning theory and the principle of structural risk minimization (Vapnik, 1992). Consider the problem of the set of training data

D= $\{(x^1, y^1), \dots, (x^l, y^l)\}$ with a linear function,

$$y = \langle w, x \rangle + b \tag{4}$$

The optimum regression function is provided from the minimum of the function,

Minimize

$$\frac{1}{2}\|w\|^2 + c\sum(\xi_i + \xi_i^*) \tag{5}$$

Constraints

$$\begin{aligned} &y_i - (wx_i) - b \le \varepsilon + \xi_i \\ &wx_i + b - y_i \le \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \ge 0 \end{aligned} \tag{6}$$

where (*) symbolizes both the vector with and without asterisks. ξ_i, ξ_i^* slack variable and c > 0 is a penalty parameter (Gunn, 1998). The constrained optimization problem is then reworded as a dual problem using Lagrange multipliers a_i, a_i^* for each constraint. Lagrange multipliers are determined by solving the issue with quadratic programming. After a_i, a_i^* are determined, the optimal weights w and the base b can be calculated, and the final predictor is given in the equation (Shokry et al., 2015).

$$y = \sum_i^m (a_i^* - a_i)(x_i - x) + b \tag{7}$$

b) Decision Tree

Decision tree algorithms are among the most preferred machine learning techniques because they are easy to interpret, detect errors, and apply easily (Kotsiantis, 2011). Breiman et al. (1998) proposed classification and regression tree (CART). The R function recursive partitioning (RPART) is an application of the CART. The RPART programs construct classification or regression models of a very general structure using a two-step process; the resulting models may represent binary trees (Therneau and Atkinson, 1997). The mean squared error is used for the split data in the RPART algorithm. MSE for a specific node is defined as;

$$\mathrm{MSE}_{node} = \frac{1}{m_{node}}\sum(y_i - \bar{y}_{node})^2 \tag{8}$$

If it is assumed that a binary split on each node on the tree will be divided into left and right. For each division,

$$\mathrm{MSE}_{left} = \frac{1}{m_{left}}\sum(y_i - \bar{y}_{left})^2 \tag{9}$$

$$\mathrm{MSE}_{right} = \frac{1}{m_{right}}\sum(y_i - \bar{y}_{right})^2 \tag{10}$$

For each attribute j, the following formula is calculated,

$$\min(\mathrm{MSE}_{left} + \mathrm{MSE}_{right}) \tag{11}$$

The smallest of the values is chosen. The splits the dataset recursively, which means that the subsets that meet a partition are partitioned until they reach a predetermined expiration criterion (Therneau and Atkinson, 1997).

c) Random Forest Regression

The Random Forests algorithm is one of the ensemble learning algorithms. The ensemble learning algorithms produce a prediction model by combining the strong points of a simpler group of fundamental models (Friedman and Sandow, 2011). The most widely used ensemble learning algorithms are bagging and random forest algorithms. Breiman's random forest classification is an improved version of the bagging technique by adding the randomness feature. The following steps are taken for the random forest algorithm:

i) Draw n bootstrap samples from the original dataset.

ii) For each of the bootstrap samples, grow an unpruned classification or regression tree (CART) is created.

iii) In random classification, two parameters are used, namely the number of variables used in each node (m) and the number of trees to be developed (N) to determine how best to split. A new estimate is made by combining the estimates made by the N number of trees separately. While the class with the majority votes in classification trees is chosen as the final estimate, for regression trees, estimation is made by taking the average of the average votes (Liaw and Wiener, 2002).

2.3. Evaluation Metrics for Regression models

Commonly used metrics to evaluate forecast accuracy are the mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE), and the coefficient of determination (R^2) (Uğuz, 2019). R^2 is used to measure the wellness of the fit by the trained models. A high R^2 value indicates that the prediction relationship is good. MAE, MSE, RMSE are the average error measure, so low values indicate good performance. The error measures are defined as follows

$$\text{MAE} = \frac{1}{N}(\sum|y_i - \hat{y}|)$$
$$\text{MSE}=\frac{1}{N}\sum(y_i - \hat{y}))$$
$$\text{RMSE}=\sqrt{\frac{1}{N}\sum(y_i - \hat{y})}$$
$$R^2 = 1 - \frac{\sum(y_i-\hat{y})^2}{\sum(y_i-\bar{y})^2} \quad (12)$$

3.Results

In this study, the high school entrance examination data's success rate was used for Turkey's 81 provinces for (in) 2019. The data obtained are available in URL1-2-3. The features of the variables are given in Table 1.

Table 1. Description of the variables

Variable	Description
y (response variable)	Success rate
x_1	Divorce rate
x_2	Gross domestic product (GDP-per city)
x_3	The number of illiteracy
x_4	Number of higher education population
x_5	Households Internet Access rate
x_6	Number of a theatre child audience
x_7	Book reading rate

Table 1 shows the description of the response variable and explanatory variables. The success rate was obtained as the number of questions answered correctly to the total number of questions by students who took the high school entrance examination in 81 provinces. The explanatory variables were chosen among several indicators that may have a potential influence on the success rate.

Table 2. Coefficients for the Beta regression model with Cauchit link

Coefficient	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0364	0.0884	0.4120	0.6803
x1	0.2084	0.0593	3.5140	0.0004
x2	0.6004	0.2981	2.0140	0.0440
x3	-0.0769	0.3412	-0.2250	0.8217
x4	0.5315	0.3999	1.3290	0.1839
x5	-0.5964	0.1118	-5.3370	9.44e-08
x6	0.7310	0.2877	-2.5410	0.0110
x7	0.2666	0.1394	1.9120	0.0500

Beta regression model was applied to success rates. Choosing the appropriate link function in the beta regression model can significantly improve the model (Koç, 2019). In the beta regression model, the information criteria of different link functions were examined, and the most suitable link function was the Cauchit link. When the beta regression model is applied to the data by selecting the Cauchit link function, the parameter estimates are given in Table 2. According to the results in Table 2, it is seen that the variables x1, x2, x6, and x7 affect the success rate positively and the variable x5 negatively. The results are consistent with the literature (Oral and McGivney, 2014; Yavuz, 2020). Beta regression estimated model is given by

$$\hat{y} = 0.03641 + 0.20837.x1 + 0.60036.x2 - 0.59641.x5 + 0.73103.x6 + 0.26662.x7$$

One of the methods used to prepare data for analysis is normalization. The purpose of normalization is to change the values of numeric columns in the data set to use a standard scale without breaking the differences in value ranges or losing information (Han et al., 2005). After normalization, we separated data to 90% of the data going to training and the remaining 10% to test. Shortly, randomly selected data for 73 cities in the data set are train, and eight cities are test data.

Machine learning algorithms and beta regression model was applied to train data to calculate success rates of test data. The analyses were performed using R software's 3.5.2 version.

Firstly, the support vector was applied. When applying support vector machine algorithm, it is essential to determine cost and error parameters. For this reason, the model with the best performance was selected after trying 1100 models in the range of error (0,1) and cost (1,100). The best model was obtained at error 0.2 and cost 2. Then CART, random forest, and beta regression model were performed to train data.

We compared the prediction capabilities of the machine learning algorithm and beta regression model on the test data.

Model validation for the machine learning was performed on the test data and the cities randomly

selected for the test data are extracted out too for the beta model. Performance measurements are given in Table 3.

Table 3. Performance measurement for models

Performance measurements	Beta Regression	Support Regression	Random Forest Regression	CART
MAE	0.0053	0.0603	0.0481	0.0435
MSE	0.0040	0.0055	0.0038	0.0043
RMSE	0.0636	0.0741	0.0617	0.0657
R^2	0.5909	0.5400	0.6638	0.5613

Table 3 shows that the random forest regression algorithm is the best model in machine learning algorithms. Table 4 shows a comparison between forecast success rates and actual success rates.

Table 4. Success rate prediction with models

Cities	Success rate	Beta_predicted	Beta_error	RF_predicted	RF_error
Afyonkarahisar	0.6244	0.6776	0.0532	0.6898	0.0654
Adıyaman	0.6211	0.5143	0.1068	0.5307	0.0904
Ankara	0.7289	0.7799	0.0510	0.7238	0.0051
Bayburt	0.6633	0.6121	0.0512	0.5620	0.1013
Çanakkale	0.7533	0.7462	0.0071	0.7341	0.0192
Kahramanmaraş	0.6078	0.6244	0.0166	0.6192	0.0114
Van	0.4744	0.5088	0.0344	0.4804	0.0060
Hakkari	0.3800	0.5012	0.1212	0.4405	0.0605

When the estimation errors of beta regression and random forest models are examined in Table 4, the random forest model gives less error for five cities. The RF model seems to be relatively superior to the beta regression model in predicting success rates.

In the beta regression model, x1, x2, x5, x6, and x7 variables were found to have a significant contribution to the model. When a random forest algorithm is applied to data, the order of importance of variables is given in Figure 1.

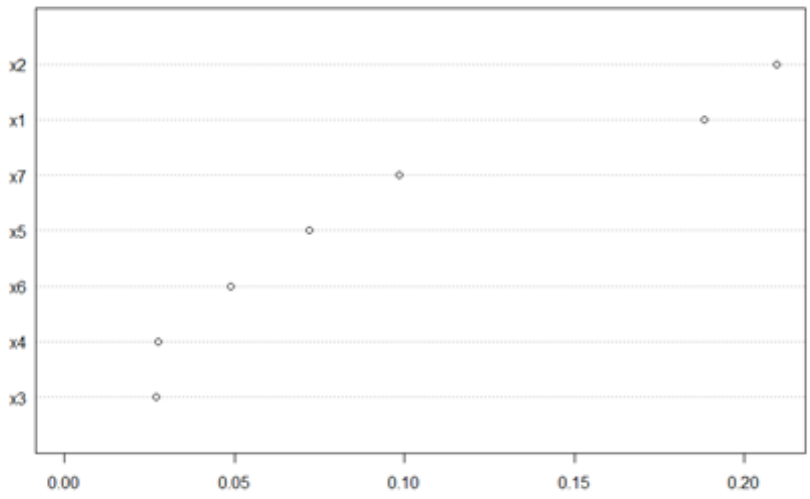


Figure 1. Random forest features importance

When we look at Figure 1, it is seen that the coefficients that are significant for the beta regression model are also important in the RF model. According to the RF model, the three most important variables were x2, x1, and x7.

4. Conclusion and Discussion

Education is one of the most significant elements that shape the future of society. Nowadays, the application of machine learning, which has successful applications in many fields, is a very favoured approach in education.

In this study, firstly, the beta regression model was used in 81 provinces of Turkey to determine the factors affecting the students who took the exam in 2019. According to the beta regression model, it is seen that the variables of divorce rate, GDP, household's internet access rate, number of theatre child audience, and book reading rate affect success rates. These results are very consistent with the literature (Özeren et al. 2020; Çömlekcioğulları, 2020). Then the data set was separated into train (73 cities) and test (8 cities). SVR, RF, CART, and beta regression models were applied to train data to calculate the success rates of test data. The two models with the highest R^2 values were R^2=0.5909 for beta regression and R^2= 0.6638 for the RF model. Also, it is seen that the best model is RF with the smallest MSE = 0.0038 among machine learning algorithms. When the prediction errors of beta regression and random forest model are examined, it is seen that the RF model is relatively superior to the beta regression model in predicting the success rates (Kikawa et al., 2020). While the beta regression model predicted the success rates of the best Çanakkale province, the RF model predicted the success rates of Ankara. Besides, variables significant in the beta regression model appear to be important in the RF model. According to the RF model, the three most important variables were found as "GDP," "divorce rate," and "book reading rate," respectively. The limitation of this study is that the data of LGS exam results for 2020 are still not disclosed in Turkey due to Covid-19. These two models are likely to provide a scientific basis for predicting students 'high school entrance examination success rates for all provinces in the following years.

References

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

Abidi, S. M. R., Hussain, M., Xu, Y. L., & Zhang, W., 2019. Prediction of Confusion Attempting Algebra Homework in an Intelligent Tutoring System through Machine Learning Techniques for Educational Sustainable Development. Sustainability, 11(1), 105. doi:ARTN 105 10.3390/su11010105

Al Mayahi, K. and Al-Bahri, M., 2020. Machine Learning Based Predicting Student Academic Success. Paper presented at the 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT).

Breiman, L., Friedman, J., Olshen, R., & Stone, C., 1998. CART. In: Chapman and Hall/CRC.

Cepeda-Cuervo, E., 2015. Beta regression models: Joint mean and variance modeling. Journal of Statistical Theory and Practice, 9(1), 134-145.

Çömlekcioğulları, A. (2020). Öğrenci başarısı ile ailelerin sosyo-ekonomik düzeyleri arasındaki ilişki (Denizli ili örneği).

Dünder, E., & Cengiz, M. A., 2020. Model selection in beta regression analysis using several information criteria and heuristic optimization. Journal of New Theory(33), 76-84.

Ferrari, S. L. P., & Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. Journal of applied statistics, 31(7), 799-815. doi:10.1080/0266476042000214501

Friedman, C., & Sandow, S., 2011. Utility-based learning from data. Boca Raton: Chapman & Hall/CRC.

Gök, M., 2017. Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 5(3), 139-148.

Gunn, S. R., 1998. Support vector machines for Classification and regression. ISIS technical report, 14(1), 5-16.

Gür, B., Çelik, Z., & Coşkun, İ., 2013. Türkiye'de ortaöğretimin geleceği: Hiyerarşi mi eşitlik mi. Seta analiz, 69, 1-26.

Han, H., Wang, W.-Y., & Mao, B.-H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Paper presented at the International conference on intelligent computing.

Kikawa, C. R., Ngungu, M. N., Ntirampeba, D., & Ssematimba, A. (2020). Support vector regression and beta distribution for modeling incumbent party for presidential elections. Appl. Math, 14(4), 721-727.

Koç, T., 2019. Türkiye'de boşanma oranlarini etkileyen faktörlerin beta regresyon modeli ile belirlenmesi. Avrasya Uluslararası Araştırmalar Dergisi, 7(16), 1111-1117.

Kotsiantis, S. B., 2011. Decision trees: a recent overview. Artificial Intelligence Review, 39(4), 261-283. doi:10.1007/s10462-011-9272-4

Liaw, A., & Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), 18-22.

MEB. ,2018). 2018 yılı performans programı Ankara Retrieved from sgb.meb.gov.tr.

Oral, I., & McGivney, E. J. ,2014. Türkiye eğitim sisteminde eşitlik ve akademik başarı, araştırma raporu ve analiz. İstanbul: Sabancı Üniversitesi Yayınları.

Özeren, E., Çiloğlu, T., Yilmaz, R., & Özeren, A. (2020). Öğrencilerin akademik kariyer hedefi seçiminde etkili olan faktörlerin veri madenciliği yöntemi ile belirlenmesi: Bartın başarı takip araştırması sonuçları üzerine bir inceleme. Bilgi ve İletişim Teknolojileri Dergisi, 2(2), 182-210.

Rajak, A., Shrivastava, A. K., & Vidushi. ,2020. Applying and comparing machine learning classification algorithms for predicting the results of students. Journal of Discrete Mathematical Sciences & Cryptography, 23(2, 419-427. doi:10.1080/09720529.2020.1728895

Rebai, S., Ben Yahia, F., & Essid, H. ,2020. A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socio-Economic Planning Sciences, 70, 100724. doi:ARTN 100724 10.1016/j.seps.2019.06.009

Sethi, K., Jaiswal, V., & Ansari, M. D. ,2020. Machine learning based support system for students to select stream ,subject). Recent Advances in Computer Science and Communications ,Formerly: Recent Patents on Computer Science), 13(3), 336-344.

Shokry, A., Audino, F., Vicente, P., Escudero, G., Moya, M. P., Graells, M., & Espuña, A. ,2015. Modeling and simulation of complex nonlinear dynamic processes using data-based models: Application to photo-Fenton process. In Computer Aided Chemical Engineering (Vol. 37, pp. 191-196): Elsevier.

Therneau, T. M., & Atkinson, E. J. ,1997. An introduction to recursive partitioning using the RPART routines. Retrieved from

Uğuz, S. ,2019. Makine Öğrenmesi Teorik Yönleri ve Python Uygulamaları ile Bir Yapay Zeka Ekolü. Ankara: Nobel.

Uskov, V. L., Bakken, J. P., Byerly, A., & Shah, A. ,2019. Machine learning-based predictive analytics of student academic performance in STEM education. Paper presented at the 2019 IEEE Global Engineering Education Conference (EDUCON).

Vapnik, V. ,1992. Principles of risk minimization for learning theory. Paper presented at the Advances in Neural Information Processing Systems.

Yavuz, A., 2020. Ortaöğretime geçiş sınavında öğrenci başarısını etkileyen etmenler.

Yousafzai, B. K., Hayat, M., & Afzal, S. , 2020. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. Education and Information Technologies, 25(6), 4677-4697. doi:10.1007/s10639-020-10189-1

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(1) (2022) 16-26
DOI: 10.38016/jista.979285

Sınıflama Algoritmalarının Yağışın Varlığını Kestirme Konusundaki Performanslarının Karşılaştırması

Hakan Koçak[1]*

[1] Meteoroloji Genel Müdürlüğü, Ankara

hkocak@gmail.com

Öz

Yağış tahmini başta hava tahmincileri, tarım ve ziraatla uğraşanlar olmak üzere tüm herkesi ilgilendiren önemli bir konudur. Son yıllarda büyük ivme yakalayan yapay zeka ve makine öğrenmesi uygulamaları diğer birçok alanda olduğu gibi yağış tahmininde de tatbik edilmekle beraber yüksek doğruluklu yağış kestirimi yapmak hala zorlu bir görev olarak karşımızda durmaktadır. Son yıllarda etiklerini daha fazla hissettiğimiz iklim değişikliği nedeniyle oluşan yağış rejimindeki değişiklikler bu zorlu görevi daha da zorlu hale getirmektedir.

Bu çalışmada veri seti üzerinde 5 farklı kategoriden 10 adet sınıflayıcı algoritma uygulanarak elde edilen performanslar karşılaştırılmıştır. Araştırmada ayrıca ana veri setinden bazı parametreler çıkarılarak farklı senaryolar oluşturulmuş, her bir senaryo için sınıflama algoritmaları uygulanarak performanslarındaki değişimler gözlemlenmiştir. Araştırma sonucunda tüm senaryolar göz önüne alındığında Fonksiyonlar kategorisi dört senaryodan üçünde en başarılı kategori olmuş ve en iyi performansa sahip sınıflayıcının da bu kategoriden MLP (Çok Katmanlı Yapay Sinir Ağı) sınıflayıcısı olduğu ortaya çıkmıştır. Araştırmada ayrıca oluşturulan senaryolar için en yüksek ortalama doğruluk oranlarının %83,4 ila %84,8 arasında değiştiği görülmüştür. Bu durum, veri setinden bazı parametrelerin çıkarılmasının sonuca büyük oranda etki etmediğini göstermektedir. Elde edilen bu sonuçlar, makine öğrenmesi yöntemlerinin yağışın varlığının kestirimi konusunda iyi derecede performans sağladığını ve bu amaçla kullanılabileceğini göstermiştir.

Anahtar kelimeler: Yağış Kestirimi, Sınıflandırma Algoritmaları, Weka, Yapay Zeka, Makine Öğrenmesi, Performans Karşılaştırması

Performance Comparison of Classification Algorithms in Rainfall Prediction

Abstract

Predicting rainfall is important issue that concerns everyone especially weather forecasters, farmers and those who work in agriculture sector. Although artificial intelligence and machine learning applications, which have gained great momentum in recent years, are applied in precipitation forecasting, as in many other areas, it is still a challenging task to make high-accuracy rainfall prediction. Changes in the precipitation regime due to climate change, the effects of which we have felt more and more in recent years, make this difficult task even more challenging.

In this study, the performances obtained by applying 10 classifier algorithms from 5 different categories on the data set were compared. In addition to that, different scenarios were created by removing some of the parameters from the original data set and the performance differences of the classification algorithms for each of the scenarios were noted. The results have shown that the Functions category was the most successful category in 3 of the 4 scenarios and MLP (Multi Layer Perceptron) algorithm which belongs to that category was the most successful classifier with the rate of 84.4%. Also, highest accuracy rates were between 83.4% and 84.8% considering all four scenarios. This shows that removing some of the parameters from the original parameter set does not have a significant impact on the classification accuracy. The study results have shown that machine learning techniques achieved good performance in predicting rainfall and could be used for that purpose.

Keywords: Rainfall Prediction, Classification Algorithms, Weka, Artificial Intelligence, Machine learning, Performance Comparison

* Sorumlu yazar.
E-posta adresi: hkocak@gmail.com

Alındı : 06 Ağustos 2021
Revizyon : 01 Ekim 2021
Kabul : 27 Ekim 2021

1. Giriş (Introduction)

Su, tüm canlıların hayatında vazgeçilmez bir öneme sahiptir. Su aynı zamanda tarım ve zirai faaliyetler için de şüphesiz çok önemli bir yere sahiptir. Bu nedenle su yeryüzünde en önemli doğal kaynaklardan biridir. Son yıllarda etkilerini daha fazla gördüğümüz küresel iklim değişikliği nedeniyle yağışlarda düzensizlik meydana gelmiştir. Bunun neticesinde ise yağışlarda genel olarak azalma, kuraklık ve bazen de kısa süreli aşırı yağış nedeni ile seller ve su taşkınları meydana gelmektedir. Tüm bunlar, değişen yağış rejimin örüntüsünü daha yakından analiz etme ve beklenmeyen aşırı yağış nedeniyle meydana gelebilecek doğal afetleri önleme noktasında yağış tahmini konusunu önemli bir hale getirmiştir. Bu nedenle günümüzün bilgi teknolojilerindeki gelişmelere paralel olarak yağış tahmini konusu özellikle makine öğrenmesi ve yapay zeka disiplinlerinin ve alanlarda çalışan araştırmacıların da ilgisini çeken bir araştırma konusu haline gelmiştir.

Geçmiş yüzyıllardan günümüze doğal ve teknik bilimlerin değişim ve gelişimi deneysel çalışmalar ve doğal olayların modellenmesi ile yakından ilişkilidir. Bu değişim ve gelişim neticesinde ise doğaya dair bilginin temel kaynağının sadece ölçümlerin olabileceği ve aynı zamanda doğal olayların özelliklerin de ölçülen gözlemler arasındaki ilişkiler ile açıklanabildiği temellerine dayanan fiziksel keşif kavramının temelleri ortaya çıkmıştır (Soucek, 1992). Doğa olaylarının fiziksel tasvirinin en önemli özelliği; eğer iki değişken arasındaki ilişki biliniyorsa bilinmeyen bir özelliğin veya değişkenin ilişkili olduğu diğer özellik veya değişen aracılığı ile nicel olarak kestirilebilmesidir. Değişkenler arasındaki ilişkilerin matematiksel soyut modeller olarak veya fizik kanunları şeklinde ifade edilebiliyor olması çözümlemeli bir araştırma için en uygun olanıdır, fakat uygulama için mekanik veya elektriksel gibi diğer bazı modeller bu amaç için daha uygun olabilir (Grabec, 1990). Bilgi işleme ve makine öğrenmesi konusundaki son yıllardaki gelişmeler, bilimsel gelişim ve evrimin geleceğinin, doğal olayların modellerini kendileri geliştirip değiştiren zeki sistemlerin geliştirilmesi yönünde olacağı görüşünü ve beklentisini artırmıştır (Dibike ve Solomatine, 2001).

Yağış tahmini en zorlu görevlerden biridir. Bu konuda önerilen birçok algoritma olmasına rağmen yağışı yüksek doğrulukla tahmin etmek hala oldukça zordur. Ziraatın önemli önemli olduğu ülkelerde her yıl üretilen tahıl miktarı ve su kıtlığı her zaman büyük bir ilgi ile takip edilmektedir. Yağış miktarındaki mevsimsel ufak dalgalanmalar bile ziraat sektörü üzerinde büyük ve yıkıcı etkilere sebep olabilmektedir. Yağışın doğru tahmininin ayrıca doğal afetlerin neden olduğu can kayıplarını ve maddi zararları önleme konusunda potansiyel faydaları vardır. Taşkın ve kuraklık gibi bazı özel durumlarda yüksek doğruluklu yağış tahmini zirai yönetim ve doğal felaketleri önleme açısından oldukça önemlidir (Shoba, 2014).

Yağış tahmini aynı zamanda sel baskınlarının önlenmesi ve ayrıca su kaynaklarının yönetimi konusunda yardımcı olur. Yağışın zamanlaması ve miktarı zirai hasat konusunda belirleyici unsur olabilmektedir. Bu nedenle yağışa dair önceden bilgi sahibi olunması hem ziraatla uğraşanlara hem de karar verici konumundakilere zirai hasarın azaltılması konusunda yardımcı olabilir.

Bu çalışmada tüm canlılar için büyük öneme sahip yağışın varlığı farklı makine öğrenmesi metotları kullanılarak tahmin edilmeye çalışılmıştır. Ayrıca araştırmada kullanılan makine öğrenmesi metotlarının yağış tahminindeki başarı oranlarının çeşitli metrikler kullanılarak karşılaştırılması da araştırmanın bir diğer odak noktasını oluşturmaktadır.

2. Makine Öğrenmesi Teknikleri ile Yağış Tahmine Yönelik Önceki Çalışmalar (Previous Study)

Makine öğrenmesi teknikleri kullanarak yağış tahmini yapmaya yönelik olarak farklı yerli ve yabancı araştırmacılar tarafından yapılmış birçok araştırma bulunmakla birlikte bu araştırmaların farklı makine öğrenmesi metotlarının karşılaştırılmasına yönelik olmaktan ziyade az ve sabit sayıda girdi parametre seti üzerinde belirli bir veya birkaç makine öğrenmesi tekniğinin (Ör. yapay sinir ağları veya ağaç yapılı sınıflayıcılar) uygulandığı araştırmalar olduğu görülmektedir. Yapılan bu araştırmada ise çok sayıda (10) makine öğrenmesi metodu yağış tahminin etki edeceği düşünülen yeterli sayıda meteorolojik parametre setine uygulanarak tahmin performansları karşılaştırılmış ve ayrıca parametre setinden bazı parametreler çıkarılarak oluşturulan farklı senaryolar için algoritmaların başarımlarındaki farklar gözlemlenmiştir. Aşağıda makine öğrenmesi metotları kullanılarak yağışın tahmin edilmesine yönelik bazı araştırmalar ve sonuçları hakkında özet olarak değinilmiştir.

Uzunali tarafından 2019 yılında yapılan bir araştırmada Kandilli Rasathanesi ve Deprem Araştırma Enstitüsü Müdürlüğünün, Kandilli Bölgesine ait Ocak 1918 ile Aralık 2018 yılları arasındaki 100 yıllık veri arşivi kullanılarak, önceki yıllara dayalı olarak, son yılların ortalama yağış değerlerinin tahmini yapılmıştır. Araştırmada yöntem olarak Yapay Sinir Ağı yöntemlerinden ANFIS (Adaptive Neuro-Fuzzy Interface System, Uyarlamalı Bulanık Ağ Çıkarım Sistemi) modeli kullanılmıştır. Araştırma sonucunda zaman serisi biçimindeki geçmiş yıllara ait yağış verilerini tek başına anlamlandırabilmek ne kadar zor olsa da veriyi yapay zekâ yöntemlerini uygulayarak kullandığımızda, bu verinin anlamlandırılmasında başarılı sonuçlar alınabileceği gözlemlenmiştir.

Xiao ve Chandrasekar tarafından 1997 yılında yapılan çalışmada ise YSA kullanılarak radar gözlemlerinden yağış tahmini yapılmıştır. Bu amaçla geliştirilen YSA modelinde girdi verisi olarak radar

gözlemleri; çıktı olarak ise yağış ölçer (rain gage) ölçümleri kullanılmıştır. Araştırmada yağış tahmini için YSA tekniğinin yanı sıra reflektivite-yağış miktarı (Z-R ilişkisi) ve reflektivite, diferansiyel reflektivite ve spesifik yayılım (propagation) gibi çoklu parametreli ölçüm denklemleri teknikleri ile de aynı verilerle yağış tahmini yapılmıştır. Sonuçlar karşılaştırıldığında YSA kullanılarak elde edilen yağış tahminlerinin diğer tekniklerle elde edilen yağış tahminlerinden daha iyi olduğu görülmüştür.

Rani ve Govardhan, çok Katmanlı Algılayıcılı Yapay Sinir Ağı (MLPNN) kullanarak yağış tahmini çalışması yapmışlar ve elde ettikleri sonuçları ARIMA tekniği kullanılarak bulunan sonuçlar ile karşılaştırmışlardır. Araştırmacılar sonuç olarak geri beslemeli algoritmasının en iyi sonucu verdiğini belirtmişlerdir.

Kannan, Prabhakaran ve Ramachandran 5 yıllık yağış verisi için Pearson katsayısı hesaplamışlar ve regresyon yaklaşımı ile yağış tahmini kestirimi yapmışlardır. Yağış tahmini kestirimi için çoklu doğrusal regresyon yönetmini kullanmışlardır. Araştırma sonucunda gerçek yağış değerleri hesaplanan yağış değerlerinden daha düşük çıkmıştır. Sonuçlara göre elde edilen modelin doğruluğu çok yüksek olmamakla birlikte kestirimi yapılan değerlere yakın değerler elde edilmiştir.

Soo-yeon, Sharad, Byunggu ve Dong (2012) yağış tahmini için CART ve C4.5 karar ağacı tekniklerini önermişlerdir. Yaptıkları çalışmada yağışı tahmin etmek için öncelikle yağış olasılığı belirlemişler; eğer yağış olasılığı var ise yağış tahmini yapmışlardır. Bu amaçla çalışmalarında rüzgar hızı, rüzgar yönü, ani rüzgar, nem, sıcaklık, buharlaşma, güneş ışınımı, rüzgar soğutma endeksi, çiy noktası, basınç irtifası, bulut tabanı, hava yoğunluğu, doymuş buhar basıncından oluşan 13 değişken kullanmışlardır. Önerilen model herhangi bir bölgedeki yağış olasılığını ve saatlik yağış miktarını zaman açısından etkili bir şekilde tahmin edebilen yararlı bir model olmuştur. Yağış olasılığını tahminde CART algoritmasının %99,2 ve C4.5 algoritmasının da %99,3 oranında başarılı olduğu görülmüştür. Saatlik yağış miktarı tahmininde ise CART ve C4.5 algoritmalarının doğruluk oranları sırasıyla %92,8 ve %93,4 olmuştur.

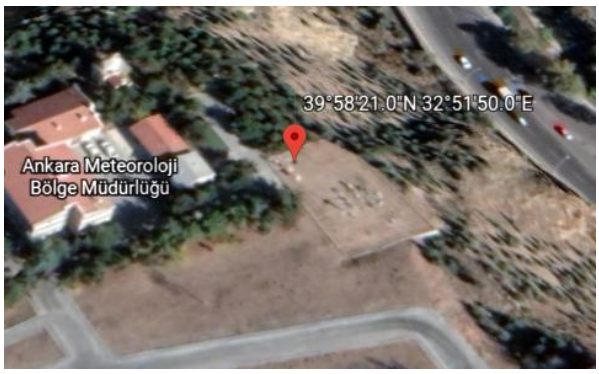
Sangari ve Balamurgan yağış tahmini konusunda K-En Yakın Komşu (KNN), Naive Bayes, Karar Ağaçları, Yapay Sinir Ağları ve Bulanık Mantık gibi farklı veri madenciliği tekniklerini karşılaştırmışlardır. Araştırmacıların kullandıkları farklı makine öğrenmesi tekniklerinin yağış tahmini konusundaki performanslarını da karşılaştırdıkları çalışma sonucunda yapay sinir ağlarının %85,77 oranı ile en iyi tahmin doğruluğuna sahip olduğu ortaya konmuştur.

Suhartono , Dwi , Bambang, Sutikno ve Heri Endozya'nın Pujon ve Wagis bölgeleri için yaptıkları çalışmada aylık yağış tahmini için ANFIS (Adaptive Neuro Fuzzy Inference System) ve ARIMA (Autoregressive Integrated Moving Average) metotlarına dayanan bir grup (ensemble) metodu önermişlerdir. Çalışmada ARIMA, ANFIS ve bunların bileşkesinden oluşan grup (ensemble) metotlarının doğruluklarını karşılaştırmak amacıyla iki ampirik yağış veri seti kullanılmış ve sonuç olarak ANFIS ve ARIMA metotlarının tek başına grup (ensemble) metodundan daha başarılı sonuç verdiği görülmüştür. ANFIS metodu çalışma alanı olarak belirlenen iki alandan biri için daha başarılı sonuç verirken ARIMA metodu diğer alan için daha başarılı tahmin başarısı göstermiştir. 1975-2010 yılları arasına ait verilerin kullanıldığı çalışmada sonuç olarak ANFIS veya ARIMA metotlarının tek başına kullanılmasının daha başarılı sonuçlar verdiği ve karmaşık metotlar olan bileşke (ensemble) modellerin ise her zaman basit modellerden daha başarılı sonuç vermeyeceği ortaya konmuştur.

3. Araştırma Alanı ve Veriler (Study Area and Data)

3.1. Otomatik meteorolojik gözlem istasyonu (Automatic meteorological observation station)

Araştırmada kullanılan yağış verileri Meteoroloji Genel Müdürlüğü (MGM) 17130 no'lu yağış gözlem istasyonuna ait veriler olup MGM veri tabanından temin edilmiştir. Bu yağış istasyonu, Kalaba Mahallesi Haliç Sokak Keçiören/Ankara'da bulunan Meteoroloji 9. Bölge Müdürlüğü yanında ölçüm yapmaktadır. 39°58'21.0" kuzey enleminde ve 32°51'50.0" doğu boylamında yer alan istasyonun bulunduğu rakım 883m ve rasat türü mm'dir.



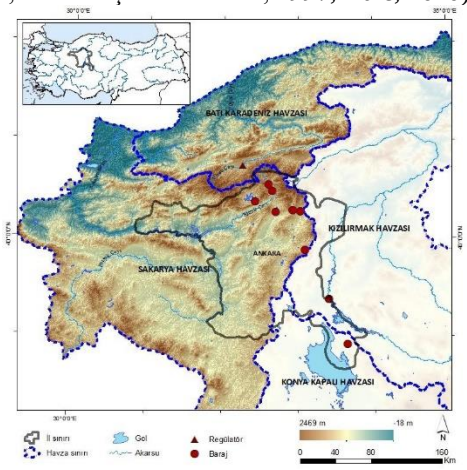
Şekil 1. Verilerin elde edildiği otomatik meteorolojik gözlem istasyonunun konumu (The location of meteorological observation station where the data were obtained from)

3.2. Çalışma bölgesi bilgileri (Study area)

Bu çalışmaya konu olan Ankara ili, Türkiye nüfusunun yaklaşık %7 sine ev sahipliği yapmaktadır. Köppen (1968) iklim sınıflamasına göre BSk iklim tipine sahip Ankara, yıllık ortalama sıcaklığı 18,0℃'den düşük, yaz kuraklığının bulunduğu yarı kurak bir iklime sahiptir (Akman, 1990). Ankara ili, yenilenebilir su potansiyelleri birbirlerinden farklı olan su havzaları üzerinde yayılım göstermektedir. Artan sanayileşme,

kentleşme ve tarımsal aktivite ile birlikte küresel iklim değişikliği Ankara ili su kaynakları üzerindeki antropojen baskıyı arttırmaktadır (Kale, 2020).

Ankara ili idari sınırları, alansal olarak %70,6 oranında Sakarya Havzası, %20,8 oranında Kızılırmak Havzası, %8,4 oranında Konya Kapalı Havzası ve %0,2 oranında Batı Karadeniz Havzası sınırları içerisinde yer almaktadır (Şekil 2). Söz konusu havzalar sırası ile Türkiye yenilenebilir su potansiyelinin %3.44, %3.48, %2.43 ve %5.34'üne ev sahipliği yapmaktadır (Öziş, Barant, Durnabaşı ve Özdemir, 1997; Kale, 2020).



Şekil 2. Ankara ili su havzası haritası (Ankara province water basin map)

3.3. Veriler(Data)

Çalışmada kullanılan veriler, Meteoroloji 9. Bölge Müdürlüğü'ne ait 17130 no'lu Otomatik Meteorolojik Gözlem İstasyonuna (OMGİ) ait 2010-2020 yılları arası günlük olarak kaydedilen ortalama aktüel basınç, ortalama sıcaklık, ortalama nispi nem ve toplam yağış parametrelerine ait 3992 adet gözlem verisinden oluşmaktadır. Veri setinde yer alan meteorolojik gözlem parametreleri basınç, sıcaklık, rüzgar hızı, rüzgar yönü, iki gün öncesine ait nem, bir gün öncesine ait nem, mevcut güne ait nem, iki gün öncesine ait yağış bilgisi (var/yok), bir gün öncesine ait yağış bilgisi (var/yok), mevcut güne ait yağış bilgisi (var/yok) şeklinde seçilmiş ve Weka aracında işlenmek üzere bu araca özgü *ARFF (Attribute-Relation File Format)* dosya biçimi şeklinde düzenlenmiştir. ARFF dosya biçimine dönüştürme işleminde Python dilinde kısa bir program yazılarak Meteoroloji Genel Müdürlüğü arşivinden her bir parametre (sıcaklık, basın, nem, yağış, vs.) için ayrı ayrı metin dosyası şeklinde elde edilen veriler bu program yardımıyla çeşitli işlemlerden geçirilmiş ve sadece gerekli bilgiler bu dosyalardan çekilerek tek bir ARFF dosyasında birleştirilmiştir. Bu parametrelerden basınç hPa cinsinden, sıcaklık °C cinsinden, rüzgar hızı m/sn cinsiden, rüzgar yönü olası tüm 16 yön isim kısaltmalarından (E, W, N, S, NW, NE, SW, SE, ENE, ESE, WNW, WSW, NNW, NNE, SSW, SSE) ve son olarak nispi nem % cinsinden MGM veri tabanında yer almaktadır ve bu birimler cinsinden veri tabanından çekilmiştir. Araştırmada kullanılan veriler bu araştırma için MGM arşivinden elde edilmiş veriler olup daha önce başka herhangi bir araştırmada kullanılmamıştır.

Veriler, sınıflayıcılara girdi olarak kullanılmadan önce veri ön işleme işleminden geçirilerek eksik veriler ayıklanmış ve geriye 3992 adet girdiden oluşan çalışma veri seti kalmıştır. Veri ayıklama işleminden sonra nümerik tipte olan basınç, sıcaklık, nem, rüzgar hızı verilerine verilerini 0-1 arasında değerlere dönüştüren min-max normalizasyon işlemi uygulanmıştır. Bu amaçla Weka aracında "Preprocess" sekmesinde yer alan Filtrelerden Unsupervise -> attribute başlığı altında yer alan "Normalize" filtresi bahsi geçen bu nümerik veriler üzerinde uygulanarak veriler normalize edilmiştir.

4. Ampirik Çalışma (Empirical Study)

Araştırmada ilgili analizleri yapmak için açık kaynak kodlu bir makine öğrenmesi yazılımı olan ve 1997 yılında Yeni Zelanda'daki University of Waikato tarafından geliştirilen WEKA (v3.8.5) yazılımı kullanılmıştır. WEKA, makine öğrenmesi konusunda en yaygın kullanılan yazılım araçlarından biridir. Başlangıçta C dili kullanılarak geliştirilen WEKA daha sonra Java dili kullanılarak yeniden yazılmıştır. WEKA, makine öğrenmesi problemlerinin çözümüne yönelik pek çok farklı makine öğrenmesi algoritması içerir. WEKA ayrıca veri ön-işleme, regresyon, sınıflama, kümeleme, birliktelik kuralları, nitelik değerlendirme ve görüntüleme amacıyla kullanılabilecek pek çok araca da sahiptir. Bununla beraber bu çalışmada Weka yazılımın sahip olduğu sınıflayıcı algoritmalardan sadece bazıları seçilerek kullanılmıştır.

Araştırmada kullanılan tüm makine öğrenmesi teknikleri performans açısından kolay ve doğru kıyaslanabilmeleri için sadece Weka aracı kullanılarak uygulanmışlardır. Weka aracılığı ile çalışma veri seti üzerinde araştırma için seçilen sınıflandırma algoritmaları uygulanma aşamasında öncelikle Weka "Classify" sekmesinde yer alan "Test options" kısmından veri setinin varsayılan oran olan %66 öğrenme (training) veri seti ve kalan %34'lük kısmının da test verisi olarak ayrılması ve sınıflandırmanın bu iki veri seti üzerinde uygulanması sağlanmıştır. Bu işlem, tüm 10 sınıflayıcı algoritma için algoritmalar uygulanmadan önce ayrı ayrı yapılmıştır.

Araştırmada mevcut girdi parametreleri olan basınç, sıcaklık, rüzgar şiddeti, nem ve yağış parametrelerinin birbirleri ile olan ilişkilerini kaybetmeyecek şekilde bu parametrelere ilave olarak 1 ve 2 gün önceki nem ve yağış verileri de girdi parametreleri setine eklenmiştir. Girdi parametreleri arasındaki korelasyon ilişkisi Tablo 1.'de verilmiştir.

Araştırmada sınıflandırma algoritmalarına girdi olarak verilen Weka ARFF dosyasındaki parametre isimleri ve birimleri şu sırasıyla şekildedir:

1. Basınç (hPa)
2. Sıcaklık (°C)

3. Rüzgar şiddeti (m/sn)
4. Rüzgar Yönü (E, W, N, S, NW, NE, SW, SE, ENE, ESE, WNW, WSW, NNW, NNE, SSW, SSE)
5. Nem-2 (2 gün önceki nem değeri (%))
6. Nem-1 (1 gün önceki nem değeri (%))
7. Nem (Mevcut güne ait nem değeri (%))
8. Yağış-2 (2 gün öncesine yağış bilgisi (var/yok))
9. Yağış-1 (1 gün öncesine yağış bilgisi (var/yok))
10. Yağış (Kestirimi yapılacak çıktı parametresi (var/yok))

Tablo 1. Parametreler arası korelasyon ilişkisi (Correlations among parameters**)**

	B	**S**	**R.Ş**	**N-2**	**N-1**	**N**	**Y-2**	**Y-1**	**Y**
B	1	-0,33	-0,15	0,09	0,05	0,08	-0,26	-0,26	-0,13
S	-0,33	1	0,27	-0,73	-0,69	-0,66	-0,11	-0,06	-0,05
R.Ş	-0,15	0,27	1	-0,25	-0,24	-0,25	-0,04	-0,03	-0,04
N-2	0,09	-0,73	-0,25	1	0,85	0,73	0,37	0,25	0,12
N-1	0,05	-0,69	-0,24	0,85	1	0,85	0,25	0,37	0,25
N	0,08	-0,66	-0,25	0,73	0,85	1	0,19	0,25	0,37
Y-2	-0,26	-0,11	-0,04	0,37	0,25	0,19	1	0,16	0,09
Y-1	-0,26	-0,06	-0,03	0,25	0,37	0,25	0,16	1	0,16
Y	-0,13	-0,05	-0,04	0,12	0,25	0,37	0,09	0,16	1

Tabloda 1.'de B basınç, S sıcaklık, R.Ş rüzgar şiddeti, N-2 iki gün önceki nem, N-1 bir gün önceki nem, Y-2 iki gün önceki yağış, Y-1 bir önceki gün yağış ve Y bulunulan günün yağış parametresini ifade etmektedir. Tablo 1. incelendiğinde nem parametrelerinin kendi aralarında en yüksek korelasyon değerlerine sahip olduğu görülmektedir. Tüm tablo içerisinde ise mevcut güne ait nem (N) ile bir önceki güne ait nem değerinin en yüksek korelasyon değerine (0,85) sahip olduğu görülmektedir. Yine aynı tablodan sıcaklık (S) ile nem parametreleri (N, N-1, N-2) arasında yüksek sayılabilecek oranda aynı yönlü bir ilişki olduğu görülmektedir. Bunun yanında basınç (B) parametresi ile yağış parametrelerinin (Y, Y-1, Y-2) diğer parametrelerle korelasyon ilişkisinin genel olarak düşük olduğu Tablo 1.'den görülebilir.

4.1. Sınıflandırma algoritmaları (Classification algorithms)

Yağış varlığı kestirimi için Weka aracında yer alan 5 adet ana sınıflayıcı kategorilerinin her birinden en az bir tane olmak üzere toplam 10 adet sınıflayıcı seçilerek çalışma veri seti üzerinde uygulanmıştır. Bu sınıflayıcı algoritmalar ve bulundukları kategoriler (parantez içinde) şunlardır:

- Bayes (BayesNet, NaiveBayes)
- Fonksiyonlar (MultiLayerPeceptron, SMO)
- Lazy (IBk)
- Kurallar (DecisionTable, JRip, OneR)
- Ağaç (J48, NBTree)

Aşağıda alt başlıklar halinde bu sınıflayıcılara dair özet bilgiler verilmiştir.

4.1.1. BayesNet

BayesNet, Bayes ağlarını özniteliklerin nominal olması ya da eğer nümerik iseler de önceden ayrıklaştırıldıkları (prediscretized) ve ayrıca eksik veri olmaması (eğer eksik veri varsa da bunlar global olarak değiştirilir) ön kabulü ile öğrenen bir algoritmadır. Ağın koşullu olasılık tablosunu tahminde iki farklı kısım bulunmaktadır. Bu araştırmada BayesNet SimpleEstimator kullanarak; K2 arama algoritması ise ADTree algoritması kullanılmadan çalıştırılmıştır (John ve Langley, 1995). Araştırmada BayesNet Algoritması şu Weka parametreleri ile uygulanmıştır: estimator=SimpleEstimator-A 0.5; searchAlgorithm=K2 -P 1 -S BAYES; useADTree=False

4.1.2. NaiveBayes

NaiveBayes sınıflandırıcısı, olasılıklı bilginin öğrenilmesi ve temsil edilmesi için temiz anlamsallığa (semantik) sahip basit bir yaklaşım sunar. Algoritmanın naive yani saf (naif) olarak adlandırılmasının sebebi; belirli bir sınıfa ait tahmin edici özniteliklerin (attribute) koşullu olarak (belirli bir sınıfa ait olmaları koşulu) bağımsız olması ve gizli ve örtük özniteliklerin tahmin işlemini etkilemeyeceği şeklindeki iki varsayıma dayanıyor olmasından kaynaklanmaktadır (John ve Langley, 1995). Araştırmada kullanılan Weka NaiveBayes algoritma parametreleri şunlardır: useKernelEstimator=False; useSupervisedDiscretization=False

4.1.3. MultilayerPerceptron (MLP)

MultilayerPerceptron, geri yayılım (backpropagation) algoritması kullanarak verileri sınıflandıran bir sınıflandırıcıdır. Bu yapay sinir ağı elle, algoritma ile ya da her ikisi kullanılarak oluşturulabilir. Oluşturulan ağ eğitim sırasında gözlemlenebilir ya da üzerinde değişiklikler yapılabilir. Bu ağda, sınıfın sayısal türde olmadığı durum haricinde tüm düğümler sigmoid fonksiyona sahiptir. Sınıf özniteliğinin sayısal (nümerik) olması durumunda ise çıktı düğümleri belirli bir eşiğe sahip olmayan lineer birimler olurlar (George-Nektarios, 2013). Araştırmada MLP algoritması için kullanılan Weka parametreleri şunlardır: hiddenLayers=a; learningRate=0.3; momentum=0.2; seed=0; trainingTime=500; validationSetSize=0; validationThreshold=20

4.1.4. SMO

SMO, bir destek vektör sınıflandırıcısının eğitimi için polinom veya Gauss çekirdeklerini (kernels) kullanarak sıralı minimal optimizasyon algoritmasını uygular. SMO sınıflandırıcısının değerlendirmesi şu parametrelerle yapılır: c = 1.0; epsilon = 1.0E; kernel = PolyKernel; numFolds = -1; randomSeed = 1 (Witten ve Frank, 2005). Araştırmada SMO algoritmasının uygulanmasında kullanılan Weka parametreleri şunlardır: c=1.0; calibrator=Logistic -R 1.0E-8 -M -1; epsilon=1.0E-12; kernel=PolyKernel -E 1.0; numFolds=-1; randomSeed=1;

4.1.5. IBk

IBk bir K en yakın komşu sınıflandırıcıdır. IBk sınıflandırıcısı uygun K değerini çapraz geçerlemeye dayanarak tayin edebilir. Bu sınıflandırıcı aynı zamanda uzaklık ağırlıklandırma da yapabilir (George-Nektarios, 2013). IBk algoritması araştırmada şu Weka parametreleri ile uygulanmıştır: KNN=1; crossValidate=False; distanceWeighting=No distance weighting; meanSquared=False; nearestNeighbourSearchAlgorithm=EuclideanDistance -R first-last; windowSize=0

4.1.6. DecisionTable

DecisionTable, karar tablosu çoğunluk sınıflandırıcısını oluşturur. Bu sınıflandırıcı, best-first arama metodunu kullanarak özellik (feature) alt kümelerini değerlendirir ve bu amaçla çapraz geçerleme (cross-validation) kullanabilir. Bu sınıflandırıcı için arama safhasında kullanılacak BestFirst, RankSearch, GeneticSearch, vs. gibi birçok metot bulunmaktadır. Ayrıca IBk algoritması da sonuca yardımcı olmak amacıyla işleme dahil edilebilmektedir. (Witten ve Frank, 2005). DecisionTable algoritması araştırmada şu Weka parametreleriyle uygulanmıştır: crossVal=1; evaluationMeasure=accuracy (discrete class); RMSE; search=BestFirst -D 1 -N 5; useIBk=False

4.1.7. JRip (RIPPER)

RIPPER en temel ve en popüler sınıflayıcı algoritmalardan biridir. Bu algoritmada sınıflar artan büyüklükte irdelenir ve sınıf için kademeli (incremental) azaltılmış hata budama yöntemi kullanılarak bir başlangıç kurallar seti oluşturulur (Witten ve Frank, 2005). JRip algoritması araştırmada şu Weka parametreleri ile uygulanmıştır: folds=3; minNo=2.0; optimazitions=2; seed=1; usePruning=True

4.1.8. OneR

OneR, Kural-tabanlı model kullanan bir diğer temel sınıflayıcı algoritmadır. Bu algoritma, her biri belirli bir özniteliği (attribute) test eden kurallardan oluşan tek seviyeli (one-level) bir ağaç yapısı oluşturur. OneR algoritması basit, hesaplama olarak ucuz ve çoğu zaman veri içindeki yapıları karakterize etmek için oldukça iyi kurallar üretebilen bir algoritmadır (Witten ve Frank, 2005). Araştırmada OneR algoritması şu parametrelerle uygulanmıştır: bathSize=100; minBucketSize=6

4.1.9. J48

Quinlan (Quinlan, 1993) tarafından geliştirilen C4.5 sınıflandırma algoritması günümüzde en popüler ağaç sınıflayıcılarından biri hatta belki de en popüleridir. J48 algoritması ise C4.5 algoritmasının optimize edilerek Weka'da uygulanmış halidir (Nguyen ve Choi, 2008). Araştırmada J48 algoritması şu Weka parametreleriyle uygulanmıştır: binarySplits=False; collapseTree=True; confidenceFactor=0.25; minNumObj=2; numFolds=3; reducedErrorPruning=False; seed=1; unpruned=False; useMDLcorrection=True

4.1.10. NBTree

NBTree, Karar Ağaçları ve Naive Bayes algoritmaları kullanılarak elde edilen melez bir sınıflayıcı algoritmadır. Bu sınıflayıcı, yaprak düğümlere ulaşabilen veri örnekleri için yaprakların birer Naive Bayes sınıflayıcısı olduğu ağaç yapıları oluşturur. NBTree algoritmasının performans açısından Naive Bayes sınıflandırıcısından daha iyi olmasını beklemek oldukça mantıklı olmakla beraber bu performansın elde edilebilmesi için hızdan fedakarlık yapılmış olmaktadır (Kohavi, 1996). NBTree algoritması araştırmada şu parametrelerle uygulanmıştır: batchSize=100; doNotCheckCapabilities=False

4.2. Performans Karşılaştırması

Her bir sınıflayıcının çalışma verileri ile Weka yazılımında uygulanmasından sonra elde edilen performans metriklerinden TP (True Positive – Doğru Pozitif), FP (False Positive – Yanlış Pozitif), TN (True Negative - Doğru Negatif) ve FN (False Negative – Yanlış Negatif) değerleri performans karşılaştırılması amacıyla kaydedilmiştir. Bu değerlerin kullanılarak

model sonuçlarına dair daha detaylı bilgiler veren Doğruluk (Accuracy), Hassasiyet (Recall veya Sensitivity) ve Kesinlik (Precision) gibi performans metrikleri aşağıdaki gibi hesaplanmaktadır (Gong, 2021):

$$\text{Doğruluk} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$\text{Kesinlik} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$\text{Hassasiyet} = \frac{TP}{TP+FN} \quad (3)$$

Bu formüllerde geçen parametreler aşağıdaki gibi tanımlanmaktadır (Bilgin, 2021):

- TP (DP): Gerçekte pozitif sınıf içinde, tahminde ise pozitif sınıf içinde yer alan değer sayısı.
- FN (YN): Gerçekte pozitif sınıf içinde, tahminde ise negatif sınıf içinde yer alan değer sayısı.
- FP (YP): Gerçekte negatif sınıf içinde, tahminde ise pozitif sınıf içinde yer alan değer sayısı.
- TN (DN): Gerçekte negatif sınıf içinde, tahminde ise negatif sınıf içinde yer alan değer sayısı.

Doğruluk (Accuracy), doğru olarak sınıflandırılan örneklerin yüzdesidir. Hassasiyet (Recall), pozitif olarak tahmin etmemiz gereken işlemlerin ne kadarını pozitif olarak tahmin ettiğimizi gösteren bir metriktir. "Gerçek pozitiflerin ne kadarı doğru bir şekilde tanımlandı?". Kesinlik (Precision), pozitif olarak tahmin edilen değerlerin gerçekten kaç adedinin pozitif olduğunu göstermektedir. (Ay, 2020).

Bunların yanı sıra sınıflandırma problemlerinde modelin performansını gösteren bazı metrikler de vardır; bu metriklerden bazıları ve en yaygın olarak kullanılanları F-Skoru (F-Measure) ve ROC (Receiver Operator Characteristic) performans metrikleridir.

F-Skoru, modelin doğruluğunun bir ölçüsüdür ve yukarıda formülleri verilen Kesinlik ve Doğruluk değerlerinin harmonik ortalaması alınarak hesaplanır. F-Skoru 1 ile 0 arasında bir değer alır; 1 değerine sahip olması modelin mükemmel olduğunu, 0 olması ise tamamen başarısız olduğunu gösterir. İyi bir F1-Skoru, yanlış pozitif ve yanlış negatif oranın az olduğunu ve örneklerin doğru biçimde sınıflandırıldığını gösterir. F-Skoru farklı düzeyde Doğruluk ve Kesinlik değerlerine sahip modelleri karşılaştırabilir hale getirebilmek için geliştirilmiş bir metriktir ve aşağıdaki şekilde hesaplanır (Nicholson, 2020):

$$\text{F-Skoru} = \frac{2x\text{Doğruluk}x\text{Kesinlik}}{\text{Doğruluk}+\text{Kesinlik}} \quad (4)$$

ROC değeri ise bir sınıflama modelin performansını değerlendirme konusunda kullanılan en önemli metriklerden biri olup farklı eşik değerlerinde modelin performansını gösterir. Bir olasılık eğrisi olan ROC sınıflama modelinin veri seti örneklerini farklı sınıflara ayırmada derecede başarılı olduğu hakkında bilgi verir. Bu eğrinin altında kalan alan AUC (Area Under Curve – Eğri Altında Kalan Alan) olarak adlandırılır ve bu alan ne kadar büyükse model veri seti örneklerini doğru sınıflara ayırmada o kadar başarılıdır (Narkhede, 2018).

ROC değerinin hesaplanmasında Gerçek Doğru Oranı (True Positive Rate - TPR) da denilen ve yukarıda açıklanan Hassasiyet metrik değeri ile Özgüllük denilen Yanlış Pozitif Oranı (False Positive Rate - FPR) metriklerinden yararlanılır. Özgüllük ya da diğer adı ile Yanlış Pozitif Oranı, Yanlış Pozitiflerin Yanlış Pozitif ve Doğru Negatiflerin toplamına oranını ifade eder. Bu iki metriğin dağılımına çeşitli eşit değerleri uygulanarak bu değerler için modelin sınıflandırmayı ne derece iyi yaptığını gösteren ROC eğrisi elde edilir (Narkhede, 2018).

Araştırmada ele alınan 10 adet sınıflayıcı algoritmasına ait yağış tahmini performans metrikleri Tablo 2'de verilmiştir. Tablo 2 genel olarak incelendiğinde en iyi doğruluk oranını %84,24 ile Fonksiyonlar kategorisinde yer alan SMO algoritmasının verdiği görülmektedir. Bununla beraber bu performans değerinin diğer sonuçlardan açık ara önde olmadığı, Kurallar kategorisinde yer alan JRip (%83.0633) algoritması ile Ağaç sınıflayıcılar kategorisinde yer alan NBTree algoritmasının (%83.4315) da bu sonuca yakın performans gösterdiği söylenebilir. Ayrıca Kurallar kategorisinde yer alan OneR sınıflayıcısının tüm kategoriler içerisinde en düşük ortalama Doğruluk Oranına (%75,5) sahip olduğu görülmüştür. Bayes kategorisinde yer alan BayesNet ve NaiveBayes algoritmalarının performanslarının da OneR algoritmasından sonra en düşük performansa (76,6 ve 76,7) sahip sınıflayıcılar olduğu ortaya çıkmıştır.

Kategori bazında baktığımızda ise Bayes Sınıflayıcılar kategorisinde sonuçların birbirine çok yakın olduğu ve BayesNet ve NaiveBayes sınıflayıcıların birbirine çok yakın performans gösterdiği görülmektedir. Fonksiyon Sınıflayıcılar kategorisinde yer alan iki adet sınıflayıcı algoritmasından en başarılısının tüm sınıflayıcıların en başarılısı da olan SMO algoritması olduğu; buna mukabil diğer sınıflayıcı MLP algoritmasının da performansının (%81,95) iyi bir seviyede olduğu görülmektedir. Lazy kategorisinde yer alan tek sınıflayıcı olan IBk sınıflayıcının performansının %78,05 ile fonksiyon sınıflayıcılardan biraz geride kaldığı görülmektedir. Kural tabanlı sınıflayıcıların ise genel olarak sınıflandırma performansının iyi olduğu, bu kategorinin en iyi performansına (%83.06) JRip algoritmasının sahip olduğu; DecisionTable algoritmasının buna yakın performans (%81.00) sergilediği, OneR algoritmasının performansının (%75,47) ise bunlardan biraz geride kaldığı söylenebilir. Son kategori olan Ağaç Sınıflayıcılar kategorisinde ise

J48 ve NBTree algoritmalarının performanslarının birbirine oldukça yakın olmakla beraber NBTree algoritmasının performansının J48 algoritmasının performansına nazaran biraz daha iyi olduğu görülmektedir.

Sınıflayıcı performanslarının Ortalama Doğruluk (OD) oranları haricinde tabloda yer alan Kesinlik, Hassasiyet, F-Skoru ve ROC Alanı değerlerine baktığımızda tabloda Doğruluk Oranı en yüksek olan sınıflayıcı SMO algoritmasının aynı zamanda en yüksek ortalama Doğru Pozitif (DP), Kesinlik, Hassasiyet ve F-Skoru değerlerine sahip olan sınıflayıcı olduğunu görmek mümkündür. Buna mukabil Kurallar kategorisinde yer alan JRip algoritması en düşük ortalama Yanlış Pozitif (YP) oranına ve Fonksiyonlar kategorisinde yer alan MLP (MultiLayerPerceptron – Çok Katmanlı Yapay Sinir Ağı) algoritmasının ise en yüksek ROC Alanına sahip olduğu ortaya çıkmıştır.

Tablo 2. On adet sınıflayıcı algoritmaya ait yağış tahmin performans metrikleri (Precipitation prediction performance metrics for the 10 classifiers**)**

Sınıflayıcı Kategorisi	Sınıflayıcı Algoritma	DP (%)	YP (%)	Doğruluk (%)	Kesinlik	Hassasiyet	F-Skoru	ROC Alanı
Bayes	BayesNet	76,7	26,0	76.730	0,789	0,767	0,774	0,827
	NaiveBayes	76,6	26,8	76.583	0,785	0,766	0,772	0,828
Fonksiyon	MLP	82,0	34,4	81.958	0,814	0,820	0,810	**0,858**
	SMO	**84,2**	27,7	**84.241**	**0,838**	**0,842**	**0,838**	0,783
Lazy	IBk	78,1	32,6	78.056	0,778	0,781	0,779	0,727
Kural	DecisionTable	81,0	27,8	81.001	0,809	0,810	0,810	0,854
	JRip	83,1	**25,7**	83.063	0,829	0,831	0,829	0,800
	OneR	75,5	35,2	75.478	0,755	0,755	0,755	0,701
Ağaç	J48	82,5	31,6	82.547	0,820	0,825	0,819	0,784
	NBTree	83,4	28,6	83.431	0,829	0,834	0,830	0,852

Araştırmada ayrıca farklı parametreler araştırma veri setinden çıkarılarak sınıflayıcıların performanslarındaki değişimler gözlemlenmiştir. Bu amaçla ilk önce veri setinden sadece rüzgar parametreleri (rüzgar yön ve şiddeti), sonrasında sadece önceki günlerin nem ve yağış parametreleri, en sonunda da hem rüzgar hem de önceki günlere ait nem ve yağış parametreleri çıkarılarak sınıflayıcı performanslarındaki değişimlere bakılmıştır. Tablo 3, sadece rüzgar parametrelerinin (rüzgar yön ve şiddeti), Tablo 4, sadece önceki günlere ait nem ve yağış parametrelerinin ve son olarak Tablo 5 ise hem rüzgar hem de önceki günler nem ve yağış parametrelerinin çıkarıldığı durumda 10 sınıflayıcının performanslarını göstermektedir.

Tablo 3. Rüzgar parametreleri çıkarıldığında sınıflayıcıların yağış tahmin performans metrikleri (Precipitation prediction performance metrics after wind parameters removed)

Sınıflayıcı Kategorisi	Sınıflayıcı Algoritma	DP (%)	YP (%)	Doğruluk (%)	Kesinlik	Hassasiyet	F-Skoru	ROC Alanı
Bayes	BayesNet	76,2	**26,5**	76.215	0,785	0,762	0,769	0,824
	NaiveBayes	76,4	26,6	76.362	0,785	0,764	0,770	0,825
Fonksiyon	MLP	**84,8**	32,4	**84.757**	**0,849**	**0,848**	0,836	**0,897**
	SMO	84,5	28,1	84.462	0,840	0,845	**0,839**	0,782
Lazy	IBk	78,5	32,5	78.497	0,782	0,785	0,783	0,730
Kural	DecisionTable	81,0	27,8	81.001	0,809	0,810	0,810	0,854
	JRip	84,0	29,5	84.020	0,836	0,840	0,834	0,778
	OneR	75,5	35,2	75.478	0,755	0,755	0,755	0,701
Ağaç	J48	82,0	31,7	81.958	0,813	0,820	0,813	0,818
	NBTree	83,1	26,7	83.063	0,827	0,831	0,829	0,867

Tablo 4. Önceki günlerin nem ve yağış parametreleri çıkarıldığında sınıflayıcıların performans metrikleri (Precipitation prediction performance metrics after humidity and precipitation parameters belonging to previous days removed)

Sınıflayıcı Kategorisi	Sınıflayıcı Algoritma	DP (%)	YP (%)	Doğruluk (%)	Kesinlik	Hassasiyet	F-Skoru	ROC Alanı
Bayes	BayesNet	79,2	32,2	79.234	0,788	0,792	0,790	0,833
	NaiveBayes	79,8	30,2	79.823	0,796	0,798	0,797	0,843
Fonksiyon	MLP	82,5	33,7	82.547	0,820	0,825	0,816	**0,860**
	SMO	83,3	30,5	83.284	0,828	0,833	0,826	0,764
Lazy	IBk	77,1	33,2	77.098	0,770	0,771	0,771	0,720
Kural	DecisionTable	81,4	28,4	81.443	0,812	0,814	0,813	0,846
	JRip	**83,9**	**27,8**	**83.873**	**0,834**	**0,839**	0,834	0,782
	OneR	75,5	35,2	75.478	0,755	0,755	0,755	0,700
Ağaç	J48	80,7	32,9	80.706	0,800	0,807	0,801	0,790
	NBTree	80,8	33,5	80.780	0,800	0,808	0,801	0,832

Tablo 5. Rüzgar ve önceki günlerin nem ve yağış parametreleri çıkarıldığında sınıflayıcıların yağış tahmin performans metrikleri (Precipitation prediction performance metrics after humidity and precipitation parameters belonging to previous days along with wind parameter removed)

Sınıflayıcı Kategorisi	Sınıflayıcı Algoritma	DP (%)	YP (%)	Doğruluk (%)	Kesinlik	Hassasiyet	F-Skoru	ROC Alanı
Bayes	BayesNet	80,8	29,7	80.780	0,804	0,808	0,806	0,843
	NaiveBayes	81,1	30,8	81.075	0,805	0,811	0,807	0,854
Fonksiyon	MLP	**83,4**	33,2	**83.431**	**0,832**	**0,834**	0,824	**0,873**
	SMO	**83,4**	30,6	**83.431**	0,829	**0,834**	0,827	0,764
Lazy	IBk	76,3	33,4	76.288	0,765	0,763	0,764	0,715
Kural	DecisionTable	81,4	28,4	81.443	0,812	0,814	0,813	0,846
	JRip	81,8	35,4	81.811	0,812	0,818	0,807	0,736
	OneR	75,5	35,2	75.478	0,755	0,755	0,755	0,701
Ağaç	J48	81,7	30,6	81.664	0,811	0,817	0,812	0,822
	NBTree	83,1	**26,3**	83.063	0,828	0,831	**0,829**	0,854

Tablo 3'ten de görüleceği üzere veri setinden rüzgar parametreleri çıkarıldığında da ortalama Doğruluk Oranı en yüksek olan algoritma yine Fonksiyonlar kategorisine ait bir algoritma olmuş ancak bu sefer SMO yerine MLP algoritması en yüksek doğruluk performansı göstermiştir. Bununla beraber SMO algoritmasının performansının (%84,5) MLP algoritması performansına (%84,8) çok yakın olduğu görülmüştür. Önceki seferde olduğu gibi veri setinden rüzgar parametreleri çıkarıldığında da Bayes kategorisine ait sınıflayıcılar en düşük performansa (%76,2 ve %76,4) sahip olmuştur. İlginç bir sonuç olarak, rüzgar parametrelerinin veri setinden çıkarılmasının aslında sınıflama performansını arttırdığı görülmüş; tam veri setinde en yüksek doğruluk oranı %84,2 iken rüzgar verileri çıkarıldığında doğruluk oranı %84,8'e çıkmıştır. Bununla birlikte doğruluk oranındaki bu artış önemli derecede bir artış değildir.

Rüzgar parametrelerini tekrar dahil edip veri setinden önceki günlere ait nem ve yağış parametrelerini çıkardığımızda ise Tablo 4'te görülen durum meydana gelmektedir. Tablo 4'ten de görüleceği üzere yeni durumda en yüksek performans sağlayan algoritma Kurallar kategorisinden JRip algoritması olmuştur. Ancak bu sefer yeni performansta (%83,9) öncekilere göre az da olsa bir düşüş olmuştur. Bu yeni senaryoda en düşük sınıflandırma performansı gösteren algoritma ise %75,5 oranı ile yine Kurallar kategorisinden OneR algoritması olmuştur.

Son durumda yani yağış veri setinden hem rüzgar hem de önceki günlerin nem ve yağış parametrelerinin çıkarıldığı durumda bakıldığında (Tablo 5); bu senaryoda en iyi performans gösteren algoritmanın kategorisinin değiştiği ve bir önceki durumdakinden farklı olarak Fonksiyonlar kategorisine ait MLP ve SMO algoritmalarının eşit (%83,4) ve en iyi sınıflama performansı sağladığı görülmektedir. Yeni performans değerinin diğerlerine çok yakın ve onlardan çok az daha düşük değerde olduğunu söylemek mümkündür. Bu son durumda en düşük performansı ise %75,5 ile Kurallar kategorisinden OneR sınıflayıcısının gösterdiği ortaya çıkmıştır.

5. Sonuç ve Tartışma (Results and Discussion)

Bu çalışmada 2010-2020 yılları arası için Ankara ili Keçiören ilçesinde yer alan Meteoroloji Genel Müdürlüğü'ne ait 17130 no'lu yağış istasyonundan elde edilen günlük basınç, sıcaklık, nem, rüzgar ve yağış verileri üzerinde çeşitli makine öğrenmesi algoritmaları uygulanarak yağışın varlığının kestirimi konusunda bu

sınıflayıcıların performanslarının karşılaştırılması amaçlanmıştır. Bu amaçla aynı veri setinde çeşitli değişiklikler uygulanarak dört farklı senaryoda 10 adet sınıflayıcı algoritma eldeki veriler üzerinde uygulanmış ve her senaryo sonucu elde edilen performans metrikleri performans karşılaştırması yapmak üzere kaydedilmiştir.

Çalışmada birinci senaryo olarak tüm verilere (basınç, sıcaklık, rüzgar, nem ve yağış) ilave olarak önceki iki günün nem ve yağış bilgileri de veri setine ayrı parametreler olarak dahil edilmiştir. İkinci senaryoda veri setinden sadece rüzgar parametreleri (rüzgar yön ve şiddeti) çıkarılarak bu yeni veri seti üzerinde 10 adet sınıflayıcı uygulanarak performansları kaydedilmiştir. Üçüncü senaryoda rüzgar parametreleri veri setine tekrar dahil edilmiş, ancak bu sefer önceki günlere ait nem ve yağış parametreleri veri setinden çıkarılmıştır. Son ve dördüncü senaryoda ise hem rüzgar parametreleri hem de önceki günlere ait nem ve yağış parametreleri veri setinden çıkarılmış ve bu haliyle sınıflayıcı performanslarındaki değişimler not edilmiştir. Elde edilen sonuçlara göre ilk senaryoda en iyi performans gösteren sınıflayıcı %84,2 doğruluk oranı ile Fonksiyonlar kategorisinden SMO algoritması olmuştur. İkinci senaryoda ise en iyi performans gösteren algoritma yine Fonksiyonlar kategorisinden %84,8 ortalama doğruluk oranı ile MLP (MultiLayerPerceptron – Çok Katmanlı Yapay Sinir Ağı) algoritması olmuştur. Üçüncü senaryoda en iyi doğruluk oranına sahip sınıflayıcı algoritma Kurallar kategorisinden JRip algoritması olmuştur. Son yani dördüncü senaryoda ise Fonksiyonlar kategorisinden MLP ve SMO algoritmaları birbirine eşit ve %83,4 ortalama doğruluk oranı ile en iyi performansa sahip algoritmalar olmuşlardır. Bu dört senaryoya ait sonuçlar Tablo 5'te özetlenmiştir.

Tablo 6. Dört senaryoya ait özet en iyi performans metrikleri (Summary of best performance metrics belonging to the four scenario)

	En iyi Sınıflayıcı	Kategori	DP (%)	YP (%)	DO (%)	Kes.	Hass.	F-Skoru	ROC Alanı
1. Senaryo	SMO	Fonksiyon	84,2	27,7	84.24	0,838	0,842	0,838	0,783
2. Senaryo	MLP	Fonksiyon	84,8	32,4	84.75	0,849	0,848	0,836	0,897
3. Senaryo	JRip	Kural	83,9	27,8	83.87	0,834	0,839	0,834	0,782
4. Senaryo	MLP	Fonksiyon	83,4	33,2	83.43	0,832	0,834	0,824	0,873
	SMO	Fonksiyon	83,4	33,2	83.43	0,832	0,834	0,824	0,873

Tablo 6'dan de görüleceği üzere en yüksek ortalama doğruluk oranları %83,4 ila %%84,8 arasında değişmektedir -ki bu da senaryolar arasındaki performans farkının çok fazla olmadığını göstermektedir. Dolayısıyla basınç, sıcaklık, rüzgar yönü, rüzgar şiddeti, nem ve yağış parametrelerinden oluşan ana veri setinden rüzgar parametrelerinin yahut önceki güne ait nem ve yağış parametrelerinin çıkarılması ya da bunların tekrar dahil edilmesi sınıflayıcı performanslarında çok büyük oranda bir değişime neden olmamaktadır. Bununla birlikte %80'ler civarında elde edilen bu performansların yağış tahmininde, diğer bir ifade ile yağışın varlığını kestirmede iyi derecede performanslar olduğu söylenebilir. Yine Tablo 6'dan görüleceği üzere yağışın varlığını tespit konusunda en başarılı kategori Fonksiyonlar kategorisi ve en başarılı algoritmalar ise bu kategorideki SMO (Sequential Minimal Optimization) ve MLP (Çok Katmanlı Yapay Sinir Ağı) algoritmaları olmuştur. Dört senaryo içerisinden bu durumun dışında kalan tek senaryo üçüncü yani rüzgar parametrelerinin dahil olduğu ama önceki günlere ait nem ve yağış parametrelerinin çıkarıldığı senaryo olmuştur. Bu senaryoda ise en başarılı kategori Kurallar kategorisi ve en başarılı sınıflayıcı algoritma ise bu kategorideki JRip algoritması olmuştur.

Elde edilen bu sonuçlar makine öğrenmesi algoritmalarının Meteoroloji biliminin bir konusu olan yağışın varlığının kestiriminde başarılı şekilde uygulanabileceğini göstermektedir. Elde edilen sonuçların hava tahmincilerinin, Meteoroloji bilimiyle ilgili kişiler ve uygulamacılar ve genel olarak makine öğrenmesi konusunda çalışan araştırmacılar için yararlı olacağı düşünülmektedir. Konu ile ilgili yapılacak gelecekteki çalışmalarda araştırmacılara yağışa etki eden farlı parametreleri de dahil ederek sınıflayıcı performanslarını karşılaştırmaları ve ayrıca yağışın varlığının kestiriminin yanında yağış miktarının kestirimi konusunda araştırma yapmaları önerilmektedir.

Kaynaklar (References)

Akman, Y. (1990). İklim ve Biyoiklim (Biyoiklim Metotları ve Türkiye İklimleri). Ankara: Palme Yayınları.

Ay, Ş. (2020). Model performansını değerlendirmek Metrikler. Erişim adresi: https://medium.com/deeplearning-turkiye/model-performans%C4%B1n%C4%B1-de%C4%9Ferlendirmek-metrikler-cb6568705b1

Bilgin, G. (2021). Makine öğrenmesi algoritmaları kullanarak erken dönemde diyabet hastalığı riskinin araştırılması. Zeki sistemler teori ve uygulamaları dergisi 4(1), 55-64. DOI: 10.38016/jista.877292

Dibike, Y.B., Solomatine, D.P. (2001). River flow forecasting using artificial neural networks. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere, Volume 26, Issue 1, Pages 1-7.

Gong, M. (February 2021). A novel performance measure for machine learning classification, International Journal of Managing Information Technology (IJMIT), Vol.13, No.1

Grabec, I. (1990). Emperical modelling of natural phenomena by a self-organizing system. Proc. Neural Network Conf. 90, Vol. 2, 529-532.

John, G. H., Langley, P (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence içinde.

Kale, M. M. (2020). İklim Değişikliği Çerçevesinde Ankara İli Ana Su Havzaları Gelecek Projeksiyonu: Sakarya ve Batı Karadeniz Havzaları, Coğrafi Bilimler Dergisi/ Turkish Journal of Geographical Sciences, 18(2), 191-215, doi: 10.33688/ aucbd.732831.

Kannan, M., Prabhakaran, S. and Ramachandran, P. (2010). Rainfall Forecasting Using Data Mining Technique. International Journal of Engineering and Technology, Vol.2 (6), 397-401

Kohavi, R. (1996). Scaling up the accuracy of naïve-bayes classifier: A decision-tree hybrid. Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining içinde (202–207). Menlo Park: AAAI Press.

Narkhede (2018). Understanding AUC-ROC curve. Erişim adresi https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nguyen, H. A., Choi, D. (2008). Application of Data Mining to Network Intrusion Detection: Classifier Selection Model. 11th Asia-Pacific Network Operations and Management Symposium (APNOMS 2008) içinde (399-408 ss.).

Nicholson (2020). Evaluation metrics for machine learning – Accuracy, precision, recall, and F1 defined. Erişim adresi https://wiki.pathmind.com/accuracy-precision-recall-f1

Öziş, Ü., Barant, T., Durnabaşı, İ., Özdemir, Y. (1997). Türkiye'nin su kaynakları potansiyeli. Meteoroloji Mühendisliği, 2, 40-45.

Quinlan, J. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann

Rani, B. K. and Govardhan, A. (2013). Rainfall prediction using Data Mining techniques-A Survey. Compututer Science and Information Technology, pp. 23-30.

Sangari, R. S. and Balamurugan, M. (2014). A Survey on rainfall prediction using Data Mining. International Jornal of Computuer Science and Mobile Applications, vol. 2, no. 2, pp. 84-88.

Shoba G (2014). Rainfall prediction using Data Mining techniques: A Survey. International Journal of Engineering and Computer Science, vol. 3, no. 5, pp. 6206-6211.

Soo-Yeon, J., Sharad, S., Byunggu, Y. and Dong, H. J. (2012). Designing a Rule-Based Hourly Rainfall Prediction Model. IEEE IRI, August 8-19.

Soucek, B. (1992). Prediction of Chaotic Dynamical Phenomena by a Neural Network. Dynamic, Genetic and Chaotic Programming içinde (471-500). U.S.A: John Wiley & Sons Inc.

Suhartono, S., R. Faulina, D. A. Lusia, B. W. Otok, Sutikno and H. Kuswanto (2012). Ensemble method based on ANFIS-ARIMA for rainfall prediction. 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), pp. 1-4, doi: 10.1109/ICSSBE.2012.6396564.

Theofilis, G. (2013). Weka classifier summary. Erişim adresi https://www.academia.edu/5167325/Weka_Classifiers_Summary

Uzunali, A. (2019). Yapay sinir ağlarına dayalı yağış tahmin ve analizi. (Yayımlanmamış yüksek lisans tezi). İstanbul Kültür Üniversitesi Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, İstanbul

Witten, I.H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. San Francisco: Morgan Kaufmann.

Xiao, R. And V. Chandrasekar. (1995). Multiparameter radar rainfall estimation using neural network techniques. Preprints, 27th Conf. on Radar Meteorology, Vail, CO, Amer. Meteor. Soc., 199–204.

Research Article
Journal of Intelligent Systems: Theory and Applications 5(1) 2022: 27-34
DOI: 10.38016/jista.878854

Prediction of Failure Categories in Plastic Extrusion Process with Deep Learning

Fatma Demircan Keskin[1*], Ural Gökay Çiçekli[2], İsmail Doğukan İçli[3]

[1,2] Ege University/Department of Business Administration, Izmir, Turkey
[3] Ege University/Graduate Faculty of Social Sciences, Izmir, Turkey
fatma.demircan.keskin@ege.edu.tr, gokay.cicekli@ege.edu.tr, dogukan.icli@gmail.com

Abstract
Today's manufacturing vision necessitates extracting insights from the data collected in real-time from manufacturing processes. Predicting failures with the predictive analysis of the collected process data and preventing these failures by taking necessary actions before they occur is a key factor in ensuring quality at the desired level, increasing productivity, and reducing costs in production systems. In the literature on predictive analysis of process data, machine learning and deep learning methods have attracted considerable attention, especially in recent years. This study has addressed a multi-class failure classification problem in the plastic extrusion process with a real case study. Classification models have been developed based on Long Short-term Memory (LSTM) as a deep learning method and Multilayer Perceptron (MLP) and Logistic Regression (LR) as machine learning methods to predict the failure categories. In the case study, real data taken from the extrusion process of one of the leading insulation companies operated in Izmir has been used. The final dataset includes actual measurements of seven parameters related to temperature and pressure and failure categories as the target variable. Three failure categories have been identified to define Category 0 (No failure), Category 1 (Filter change), and Category 2 (Feeding failures) states, and coded as 0,1 and 2 in the models, respectively. LSTM, MLP, and LR's performance to predict the failure categories have been evaluated and compared based on accuracy, precision, recall, and F1 Score measures. LSTM is the highest performing among the three methods, with 100% prediction accuracy for each failure category. On the other hand, LR and MLP have achieved considerable and close results except for Category 1.

Keywords: Deep learning, failure prediction, machine learning, plastic extrusion process.

Plastik Ekstrüzyon Sürecinde Derin Öğrenme İle Hata Kategorilerinin Tahmini

Öz
Günümüz üretim anlayışı, imalat süreçlerinden gerçek zamanlı olarak toplanan süreç verisinden kestirim yapabilmeyi gerektirmektedir. Toplanan süreç verilerinin kestirimci analizi ile hataların tahmin edilmesi ve gerekli aksiyonların alınmasıyla hataların ortaya çıkmadan önlenmesi, üretim sistemlerinde kalitenin istenilen seviyede sağlanması, verimliliğin artırılması ve maliyetlerin azaltılmasında kilit bir faktördür. Makine öğrenmesi ve derin öğrenme yöntemleri, süreç verilerinin kestirimci analizinde, özellikle son dönemlerde büyük ilgi görmektedir. Bu çalışmada plastik ekstrüzyon sürecinde çok sınıflı hata sınıflandırma problemi bir gerçek hayat örneğiyle ele alınmıştır. Problemin çözümü için derin öğrenme yöntemlerinden Uzun-Kısa Süreli Bellek (LSTM) ve makine öğrenmesi yöntemlerinden Çok Katmanlı Algılayıcı (MLP) ve Lojistik Regresyon (LR) kullanılmıştır. Çalışmanın uygulama kısmında, İzmir'de faaliyet gösteren Türkiye'nin önde gelen yalıtım firmalarından birinin plastik ekstrüzyon sürecinden alınan gerçek veriler kullanılmıştır. Nihai veri seti, süreçten alınan sıcaklık ve basınçla ilişkili yedi parametrenin gerçek ölçümlerini ve hedef değişken olarak hata kategorilerini içermektedir. Modellerde Kategori 0 (Hata yok), Kategori 1 (Filtre değişimi) ve Kategori 2 (Besleme hataları) durumlarını tanımlamak için üç hata kategorisi belirlenmiş ve sırasıyla 0,1 ve 2 olarak kodlanmıştır. LSTM, MLP ve LR'nin hata kategorilerini tahmin etme performansı, tahmin doğruluğu, kesinlik, duyarlılık ve F1 skoru metriklerine göre değerlendirilmiş ve karşılaştırılmıştır. LSTM, her hata kategorisi için %100 tahmin doğruluğu ile en yüksek performansa sahip olmuştur. LR ve MLP, Kategori 1 dışındaki hata kategorileri tahminlerinde başarılı ve birbirine yakın sonuçlar elde etmiştir.

Anahtar Kelimeler: Derin öğrenme, hata tahmini, makine öğrenmesi, plastik ekstrüzyon süreci

* Corresponding Author.
E-mail: fatma.demircan.keskin@ege.edu.tr

Recieved : 12 Feb 2021
Revision : 16 Sep 2021
Accepted : 27 Oct 2021

**This paper is an extended version of the paper published in the Proceedings Book of 11th International Statistics Congress, on 4-8 October 2019.

1. Introduction

Rapid developments in digital technologies have had transformative effects on manufacturing systems and turned them into smart systems. The current industrial vision, Industry 4.0, has structured an interconnected manufacturing environment and forced companies to reconsider their processes. One of the critical cornerstones of Industry 4.0 and smart manufacturing is collecting real-time data from the plant via sensors and networks and providing value by conducting a data-driven predictive analysis.

Failures may occur in the manufacturing environment due to many causes. Therefore, in smart manufacturing systems, it is critical to monitor manufacturing processes in real-time, predict failures, and take appropriate actions to prevent them from happening to ensure product quality (Tao et al., 2018).

Fault detection and prediction problems in the manufacturing environment have been extensively addressed through machine learning methods (Konar and Chattopadhyay, 2011; Jing and Hou, 2015) and, in particular, deep learning methods with increasing interest recently (Jing et al., 2017; Shao et al., 2017; Zhang et al., 2017a).

Neural networks (Hou, Liu, and Lin, 2003; Quintana et al., 2011) and LR (De Menezes et al., 2017) are among the most widely applied supervised machine learning methods for failure classification and prediction problems using process data.

LR is a supervised machine learning method with a wide range of application areas for prediction problems containing a categorical dependent variable and a set of independent variables (Caesarendra et al., 2010). When the dependent variable has multi-class, like the problem addressed in this study, multi-class LR needs to be employed. The conditional probability P(Y = y | X = x) in multi-class LR is calculated by using Equation (1) (Le Thi et al., 2020):

$$\mathrm{P}(\mathrm{Y} = \mathrm{y} \mid \mathrm{X} = \mathrm{x}) = \frac{\exp\left(b_y + W_{:,y}^T x\right)}{\sum_{k=1}^{Q} \exp\left(b_y + W_{:,y}^T x\right)} \quad (1)$$

where $\{(x_i, y_i): i = 1, \dots, n\}$ is a training set that includes observation vectors $x_i \in \mathbb{R}^d$ and labels $y_i \in \{1, \dots, Q\}$, Q denotes the number of classes, W is the dxQ matrix and $b = (b_1, \dots, b_Q) \in \mathbb{R}^Q$. It is aimed to find a (W,b) pair that maximizes the total probability of the correct class y to which x_i belongs. The negative log-likelihood function needs to be minimized to obtain (W,b) estimation (Le Thi et al., 2020).

MLP, one of the most employed neural network techniques, especially for the problems related to production control (Cadavid et al., 2020), contains input and output layers of units and hidden unit/units' layers between them (Fallah, Mitnitski and Rockwood, 2011). In MLP, the units are organized in a feed-forward layered topology (Venkatesan and Anitha, 2006). MLP uses various nonlinear functions to convert n inputs to l outputs. In Equation (2), the activation function used to determine the network output is given (Yilmaz and Kaynar, 2011):

$$x_0 = f\left(\sum_h x_h w_{ho}\right) \quad (2)$$

where f denotes the activation function, x_h is hth hidden layer node's activation and w_{ho} is the hth hidden layer node and oth output layer interconnection.

Deep learning, which has significant successful applications in many different areas such as text detection and classification, speech and image recognition, provides advanced analytical opportunities for analyzing big data obtained from manufacturing processes (Wang et al., 2018). There have been different deep learning approaches, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Auto-encoders, Deep Belief Network, Deep Boltzmann Machines, and each of them may have some sub-variants (Zhao et al., 2019). LSTM is an architecture of RNN which uses past sequences to forecast future data (Moghar and Hamiche, 2020). In LSTM architecture, information flows, including determining which information to remain and how long it persists, are regulated via input, forget, and output gates (Bandara, Bergmeir, and Smyl, 2020). The input, output, and forget gates have different abilities and tasks in the architecture. The input gate can choose information necessary to be stored in the internal state, the output gate has the capability of deciding the output information, and the forget gate can throw away the useless information (Zhang et al., 2017b). LSTM has widely preferred for the predictive analysis of sequential data and has a wide range of application areas, including failure prediction, remaining useful life prediction, voice recognition, time series analysis, document classification (Nabipour et al., 2020). LSTM stands out for its ability to recognize long-term dependencies and patterns in sequential data and provide more successful results of anomaly and failure detections than standard RNN in this data type (Greff et al., 2016; Meyes et al., 2019).

The mathematical expression of LSTM output of the jth cell (c_j) at time t is given in Equation (3) (Hochreiter and Schmidhuber, 1997; Smagulova and James, 2019):

$$y^{c_j}(t) = y^{out_j}(t) h\left(s_{c_j}(t)\right) \quad (3)$$

where $s_{c_j}(t)$ is an internal state:

$$s_{c_j}(t) = y^{\phi_j}(t) s_{c_j}(t-1) + y^{in_j}(t) g\left(net_{cj}(t)\right) \quad (4)$$

where y^{ϕ_j} is an output of forget gate:

$$y^{\phi_j}(t) = f_{\phi_j}\left(net_{\phi_j}\right) \quad (5)$$

The output values of the output gate (out_j) and input gate (in_j) are given in Equations (6)-(7) (Hochreiter and Schmidhuber, 1997; Smagulova and James, 2019):

$$y^{out_j}(t) = f_{out_j}(net_{out_j}) \quad (6)$$

$$y^{in_j}(t) = f_{n_j}(net_{in_j}) \quad (7)$$

Net inputs of a cell are expressed in Equations (8)-(10) (Hochreiter and Schmidhuber, 1997; Smagulova and James, 2019): (u: denotes units)

$$net_{out_j}(t) = \sum_u w_{out_j u} y^u(t-1) \quad (8)$$

$$net_{in_j}(t) = \sum_u w_{in_j u} y^u(t-1) \quad (9)$$

$$net_{c_j}(t) = \sum_u w_{c_j u} y^u(t-1) \quad (10)$$

This study aims to address the multi-class failure classification problem in the plastic extrusion process using the actual sequential process data of an insulation company by applying LSTM, one of the deep learning methods widely known for its successful performance in prediction problems for the sequentially formed datasets, and machine learning methods of MLP and LR, to evaluate and compare the class prediction performance of these approaches.

Even though there have been studies handling the determination of process parameters problem in the plastic extrusion process with machine learning methods (Huang and Liao, 2002; Al Rozuq and Al Robaidi, 2013; Cirak and Kozan, 2009), none studies reached addressing any problems in this process with deep learning. Therefore, this study aims to contribute to the related literature by addressing the failure classification problem in the plastic extrusion process and applying deep learning.

In the next section of this study, some of the relevant related works addressing similar problems by applying LSTM, MLP, and LR are introduced. Afterward, the problem is explained in detail. Following this section, application findings and their analysis are presented. Finally, the results are evaluated and discussed in the conclusion section.

2. Related Works

Failure classification and prediction problems have been attracted considerable attention in previous studies. In those studies, wide range of machine learning based methods, including Artificial Neural Networks (ANN) (Dreiseitl and Ohno-Machado, 2002; Gyimothy, Ferenc and Siket, 2005; Singh, Kaur and Malhotra, 2009), CNN (Janssens et al., 2016; Tan and Pan, 2019), Decision Tree (DT) (Dreiseitl and Ohno-Machado, 2002; Gyimothy, Ferenc and Siket, 2005; Singh, Kaur and Malhotra, 2009), Deep CNN (Razaviarab, Sharifi and Banadaki, 2019), k-nearest-neighbour (Dreiseitl and Ohno-Machado, 2002), LR (Dreiseitl and Ohno-Machado, 2002; Gyimothy, Ferenc and Siket, 2005; Singh, Kaur and Malhotra, 2009; Malhotra and Singh, 2011), LSTM (Malhotra et al., 2015; Zhang et al., 2017b; Morariu et al., 2018; Tan and Pan, 2019; Ye et al., 2019), MLP (Liukkonen et al., 2009; Kutyłowska, 2015; Hore et al., 2016; Orrù et al., 2020), Random Forest (Tan and Pan, 2019) and Support Vector Machine (SVM) (Dreiseitl and Ohno-Machado, 2002; Singh, Kaur and Malhotra, 2009; Zhang et al., 2017a; Oh et al., 2019) have been applied. This section presents indicators and findings related to LSTM, MLP, and LR models for the failure prediction problem used in previous studies.

Zhang et al. (2017b) employed the LSTM-RNN method to predict the battery's remaining useful life with deep learning capability. They compared the LSTM and SVM methods and noted that the LSTM-RNN method predictions are more accurate than SVM. Tan and Pan (2019) proposed a model to predict faults of wireless networks based on LSTM and CNN. This study compared CNN, CNN-LSTM, and Random Forest models' performances and showed that their CNN-LSTM hybrid prediction model had better performance than the other applied models. Malhotra et al. (2015) studied the fault prediction problem by applying the stacked LSTM. They used data sets, including power demand, multi-sensor motor, space shuttle. Their results indicated that normal time-series behavior could be modeled with the stacked LSTM. Morariu et al. (2018) used the LSTM approach to estimate energy consumption patterns in the production cycle accurately. They proposed a structure that processes the information flow in high-capacity production systems using map reduction algorithms and focuses on energy consumption with big data concepts collected in various layers. Ye et al. (2019) proposed the LSTM-RNN structure by making parameter estimates for a reasonable estimate of river water quality.

Hore et al. (2016) used the MLP-FFN classifier to predict failures of reinforced concrete buildings. They identified the possibility of failure of the handled buildings in the future. The experimental results obtained in this study indicated that the proposed model provides satisfactory performance. Kutyłowska (2015) developed MLP networks to model the damage frequency in the water supply systems. She noted that the plumbing could use the created model to determine the frequency of breakdowns and plan the replacement of broken pipes. Liukkonen et al. (2009) performed a wave soldering event study to predict product failures using the MLP neural network model. They focused on root causes in response to the number of failures they detected in their work. As the MLP algorithm's input, they accepted the types of failure as the output of the process parameters. Finally, Orrù et al. (2020) applied MLP and SVM for the fault prediction problem using

real-time collected sensor data from a refinery's production line.

Malhotra and Singh (2011) used the LR and seven other machine learning methods to predict faulty classes with object-oriented metrics in software testing. Singh, Kaur, and Malhotra (2009) compared LR, Artificial Neural Network (ANN), SVM, and DT methods for the fault proneness of object-oriented system classes by using Receiver Operating Characteristic analysis. Gyimothy, Ferenc, and Siket (2005) employed LR, neural network, and DT for fault prediction. The results showed that the logistic regression analysis was significant. Finally, Dreiseitl and Ohno-Machadob (2002) compared LR and ANN methods with other classification algorithms, such as SVM, k-nearest neighbors, and DT.

3. Problem and Data Description

Conducting predictive analysis based on process data is one of the prerequisites of today's manufacturing understanding. In this study, the classification problem of multi failure types occurring during the plastic extrusion process of an insulation company has been addressed. Plastic extrusion is a continuous process in which a solid plastic material is converted into a molten fluid; the flowable melt moves into the die and takes the desired shape. The temperature and pressure rollers fed from the top and bottom layers produce a double waterproofing sheet. The line is fed from top layers via extruders A and B, and from bottom layers, via extruder C. There are lower, central, and upper calenders at the end of the die. Finally, the calendered product is cooled and wound in rolls. Extruder C is used to reprocess granulated plastic waste. For this reason, a filter system is used in Extruder C.

The initial dataset received from the company covers the real measurements taken every 5 minutes sequentially and the failure categories at measurement times. In this process, a number of failure types, including edge tearing, die cleaning, die changing, filter changing, failures of material feeding can occur. However, only filter changes and material feeding failures to the line during the data collection period have occurred. So the failure categories have been labeled as "No failure," "Filter change," and "Feeding failures" in the initial dataset. These categories are coded to be used in the models as follows:

- No failure (0)
- Filter change (1)
- Feeding failures (2)

Some of the parameters' values do not change during the analysis period. Therefore, these parameters were excluded from the analysis. Also, there are some parameters with some missing values. After cleaning the dataset, the final dataset includes 7171 observations regarding seven parameters and failure categories as the target variable. The variables, their descriptions, and ranges are given in Table 1.

Table 1. Description of the variables in the dataset

No	Variable Name	Description	Range
1	Pane1-Temperature Central Roll (°C)	Temperature of the central roll	[9.500,57.089]
2	Pane1-Temperature Lower Roll (°C)	Temperature of the lower roll	[9.619,68.725]
3	Pane1-Temperature Upper Roll (°C)	Temperature of the upper roll	[10.363,58.945]
4	Pane1-Melt Temp. A (°C)	Melt temperature of extruder A	[8.766,196.181]
5	Pane1-Melt , Temp. B (°C)	Melt temperature of extruder B	[9.530,195.107]
6	Pane1-Melt Temp. C (°C)	Melt temperature of extruder C	[11.527,197.047]
7	Ext. C. melt pressure_difference (°C)	Difference of the two consecutive melt pressures in extruder C	[-157.937,223.702]

4. Application and Findings

This study has addressed the multi-class failure classification problem using actual measurement data taken from the plastic extrusion process of an insulation company with LSTM, MLP, and LR. The performance of LSTM depends on the values of its hyperparameters. Since there is no exact way of choosing which hyperparameter values work best, one of the most frequently followed methods is to use some combinations of parameters and test these combinations' performances with several experiments (Greff et al., 2016).

In this study, the analysis-ready data set was randomly divided into training, validation, and testing sets with the size of 70%, 20%, and 10% of the whole data set, respectively. Therefore, the fault categories' observations into the training, validation and testing sets are as equal as possible. In Table 2, features of the training, validation, and testing sets are presented.

Table 2. Features of the training, validation, and testing sets

Number (percentage) of failure categories in the training set	No failure: 4675 (94.79%) Filter change: 6 (0.12%) Feeding failure: 251 (5.09%)
Training set size	*4932*
Number (percentage) of failure categories in the validation set	No failure: 1441 (95.43%) Filter change: 3 (0.20%) Feeding failure: 66 (4.37%)
Validation set size	*1510*
Number (percentage) of failure categories in the testing	No failure: 695 (95.34%)

set	Filter change: 2 (0.27%) Feeding failure: 32 (4.39%)
Testing set size	*729*

Before implementing the analyzed methods, all inputs were normalized. In normalization, firstly, min-max and z-score normalization techniques are among the most widely used normalization techniques. The classification performance of MLP and LSTM have indicated that these models have better performance with the data normalized by the z-score technique. Therefore, the z-score normalization technique has been selected, and the results obtained with the z-score normalized data set have been presented in the rest of the study. This study followed the methodology of training models, running the trained models with multiple parameter settings several times by using the validation set and finally evaluating the models' performances in the testing set. Depending on the dataset's highly unbalanced structure, the performance of the methods is evaluated by employing evaluation metrics of precision, recall, and F1 score for each category. In addition to these metrics, the overall accuracy of the methods' predictions is also computed and compared. All employed metrics' formulas are given in Equations (11)-(14) (Orrù et al., 2020):

$$\text{Overall Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (11)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (12)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (13)$$

$$\text{F1 Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision+Recall})} \quad (14)$$

(TP: True positives, TN: True negatives, FP: False positives, and FN: False negatives)

In LSTM and MLP models, batch size, epoch, learning rate, dropout rate, and optimizer type combinations seen in Table 3 are run ten times in the validation set, and the combination that gives the best result among these combinations is employed in the testing set. All models are coded in R.

Table 3. Parameters in the experiments

Parameters	Value
Output Units	3
Batch Size	4,8,16,32,64
Epoch	10,20,50,100
Optimizer	RMSprop, adam
Learning Rate	0.001,0.0001
Dropout Rate	0.02, 0.2

Confusion matrices of the models are presented in Figures 1-3. Overall accuracies of the applied methods are given in the lower right corner of the matrices. Moreover, at the rightest column and in the bottom row, recall and precision values of the methods on each failure category are presented, respectively.

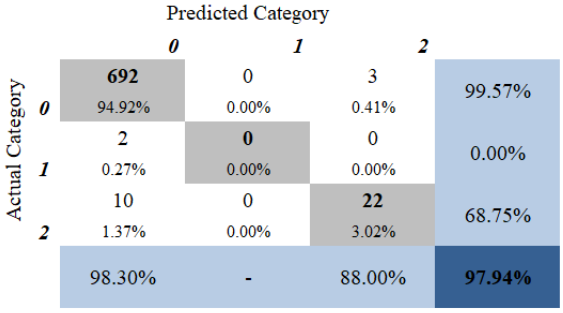


Figure 1. Confusion matrix of LR

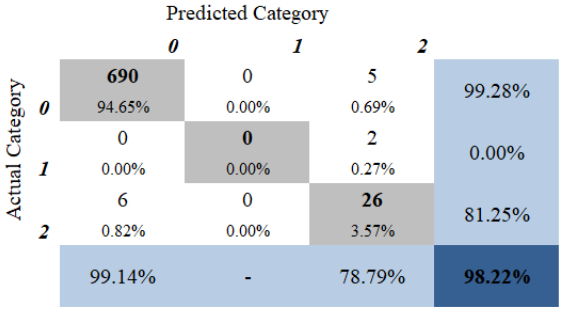


Figure 2. Confusion matrix of MLP

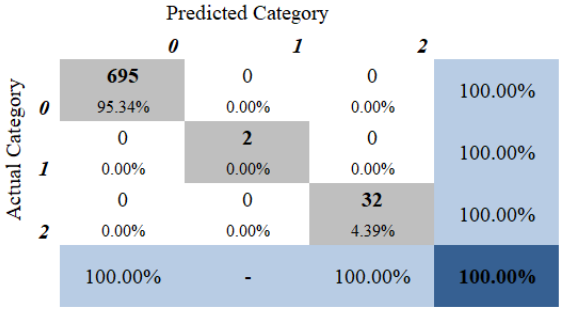


Figure 3. Confusion matrix of LSTM

All applied methods have reached a high overall accuracy. LSTM is the best with 100% accuracy, while LR is the last with 97.94% accuracy. Performances of the applied methods in terms of precision, recall, and F1 Score evaluation metrics are presented in Table 4. The results have indicated that LSTM has 100% performance for each category for each evaluation metric. The most important point revealing the success of LSTM is that it accurately predicts the class of Category 1 (Filter change), which occurs only twice in the whole test set. LR and MLP have been inadequate to predict the class of these two observations in Category 1. LR has predicted Category 1 as Category 0 (No failure), while MLP has predicted Category 2

(Feeding failure). The precision of LP and MLP for Category 1 is obtained as undefined from 0 divided by 0. Therefore, F1 Scores cannot be calculated.

Both LR and MLP have reached high precision, recall, and F1 Score over Category 0. However, MLP has better precision and F1 Score for Category 0 than LR, while LR has a higher recall value over Category 0 than MLP. Therefore, in parallel with this study's aim, to test and compare LR and MLP performances, it is more appropriate to emphasize Category 2 (Feeding failure) rather than Category 0.

LR has achieved a higher precision value over Category 2 than MLP. This result has revealed that the portion of the classes that LR predicts as Category 2 actually to be Category 2 is higher than MLP. On the other hand, MLP has yielded more successful results than LR in recall and F1 Score for Category 2. It implies that the ratio of actual "Feeding failure" detected correctly by MLP is higher than LR.

A 5-fold cross-validation procedure has been carried out to assess the validity of the models. As a result, the overall accuracy of LR, MLP, and LSTM methods have been obtained as 97.89%, 98.76%, and 99.26%, respectively.

In addition to overall accuracy, precision, recall, and F1 score for each category for each method have been calculated. The results presented in Table 5 indicate that LSTM has the best performance for all metrics. For example, while LR and MLP could not detect any observations of Category 1, LSTM has correctly predicted 66.67% of Category 1 observations.

The results of 5-fold cross-validation have also confirmed that LSTM has the most successful performance in all failure categories for the problem examined in this study.

Table 4 Prediction performances of the methods on the testing set

	LR			MLP			LSTM		
Category	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
0	98.30%	99.57%	98.93%	99.14%	99.28%	99.21%	100.00%	100.00%	100.00%
1	-	0.00%	-	-	0.00%	-	100.00%	100.00%	100.00%
2	88.00%	68.75%	77.19%	78.79%	81.25%	80.00%	100.00%	100.00%	100.00%

Table 5. Prediction performances of the methods with 5-fold cross-validation

	LR			MLP			LSTM		
Category	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
0	98.55%	99.25%	98.90%	99.16%	99.60%	99.38%	99.52%	99.77%	99.64%
1	-	0.00%	-	-	0.00%	-	66.67%	77.78%	68.89%
2	82.87%	72.69%	77.25%	91.31%	84.35%	86.87%	94.86%	90.24%	92.30%

5. Conclusions

This study addresses the problem of failure classification, which takes an important role in our age's manufacturing vision, based on the actual process data. The problem's application is conducted using the actual process data obtained from the plastic extrusion process of an insulation company. The study aims to contribute to the literature by addressing the failure classification problem in the plastic extrusion process and applying a deep learning method, LSTM, to the problem. In addition to LSTM, machine learning methods of MLP and LR are also applied, and performances of the models are compared based on accuracy, precision, recall, and F1 Score measures.

The models' class prediction accuracy has been obtained within a high range of 97.94% (LR) and 100.00% (LSTM). LSTM has classified all failure categories correctly. LR and MLP have reached a considerable and close performance in classifying Category 0 and Category 2, but they have been insufficient to predict the class of Category 1.

LSTM, as a deep learning method, has performed better than the considered machine learning methods and had 100% accuracy even though the problem dataset contains an extremely low number of Failure-1 observations. Further studies might test the models with larger datasets, including sufficient failure observations and more process parameters.

References

Al Rozuq, R. A. M. I., Al Robaidi, A. M. I. N. 2013. Application of neural network ANN to predict XLPE cable in extrusion processes. Journal of Materials Sciences and Applications, 2013.

Bandara, K., Bergmeir, C., Smyl, S. 2020. Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. Expert Systems with Applications, 140, 112896.

Cadavid, J. P. U., Lamouri, S., Grabot, B., Pellerin, R., Fortin, A. 2020. Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. Journal of Intelligent Manufacturing, 1-28.

Caesarendra, W., Widodo, A., Yang, B. S. 2010. Application of relevance vector machine and logistic regression for

machine degradation assessment. Mechanical Systems and Signal Processing, 244, 1161-1171.

Cirak, B., Kozan, R. 2009. Prediction of the coating thickness of wire coating extrusion processes using artificial neural network ANN. Modern Applied Science, 37, 52-66.

De Menezes, F. S., Liska, G. R., Cirillo, M. A., Vivanco, M. J. 2017. Data classification with binary response through the Boosting algorithm and logistic regression. Expert Systems with Applications, 69, 62-73.

Dreiseitl, S., Ohno-Machado, L. 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 355-6, 352-359.

Fallah, N., Mitnitski, A., Rockwood, K. 2011. Applying neural network Poisson regression to predict cognitive score changes. Journal of Applied Statistics, 389, 2051-2062.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., Schmidhuber, J. 2016. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 2810, 2222-2232.

Gyimothy, T., Ferenc, R., Siket, I. 2005. Empirical validation of object-oriented metrics on open source software for fault prediction. IEEE Transactions on Software engineering, 3110, 897-910.

Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory. Neural computation, 9(8), 1735-1780.

Hore, S., Chatterjee, S., Sarkar, S., Dey, N., Ashour, A. S., Balas-Timar, D., Balas, V. E. 2016. Neural-based prediction of structural failure of multistoried RC buildings. Structural Engineering and Mechanics, 583, 459-473.

Hou, T. H. T., Liu, W. L., Lin, L. 2003. Intelligent remote monitoring and diagnosis of manufacturing processes using an integrated approach of neural networks and rough sets. Journal of Intelligent Manufacturing, 142, 239-253.

Huang, H. X., Liao, C. M. 2002. Prediction of parison swell in plastics extrusion blow molding using a neural network method. Polymer testing, 217, 745-749.

Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., de Walle, R.V. Van Hoecke, S. 2016. Convolutional neural network based fault detection for rotating machinery. Journal of Sound and Vibration, 377, 331-345.

Jing, C., Hou, J. 2015. SVM and PCA based fault classification approaches for complicated industrial process. Neurocomputing, 167, 636-642.

Jing, L., Zhao, M., Li, P., Xu, X. 2017. A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. Measurement, 111, 1-10.

Konar, P., Chattopadhyay, P. 2011. Bearing fault detection of induction motor using wavelet and Support Vector Machines SVMs. Applied Soft Computing, 116, 4203-4211.

Kutyłowska, M. 2015. Neural network approach for failure rate prediction. Engineering Failure Analysis, 47, 41-48.

Le Thi, H. A., Le, H. M., Phan, D. N., & Tran, B. 2020. Stochastic DCA for minimizing a large sum of DC functions with application to multi-class logistic regression. Neural Networks, 132, 220-231.

Liukkonen, M., Hiltunen, T., Havia, E., Leinonen, H., Hiltunen, Y. 2009. Modeling of soldering quality by using artificial neural networks. IEEE Transactions on electronics packaging manufacturing, 322, 89-96.

Malhotra, P., Vig, L., Shroff, G., Agarwal, P. 2015., Long short term memory networks for anomaly detection in time series, Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Presses universitaires de Louvain. 22-24 April 2015, pp. 89-94.

Malhotra, R., Singh, Y. 2011. On the applicability of machine learning techniques for object oriented software fault prediction. Software Engineering: An International Journal, 11, 24-37.

Meyes, R., Donauer, J., Schmeing, A., Meisen, T. 2019. A Recurrent Neural Network Architecture for Failure Prediction in Deep Drawing Sensory Time Series Data. Procedia Manufacturing, 34, 789-797.

Moghar, A., Hamiche, M. 2020. Stock market prediction using LSTM recurrent neural network. Procedia Computer Science, 170, 1168-1173.

Morariu, C., Răileanu, S., Borangiu, T., Anton, F. 2018, June. A distributed approach for machine learning in large scale manufacturing systems. In International Workshop on Service Orientation in Holonic and Multi-Agent Manufacturing pp. 41-52. Springer, Cham.

Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., Mosavi, A. 2020. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access, 8, 150199-150212.

Oh, Y., Ransikarbum, K., Busogi, M., Kwon, D., Kim, N. 2019. Adaptive SVM-based real-time quality assessment for primer-sealer dispensing process of sunroof assembly line. Reliability Engineering System Safety, 184, 202-212.

Orrù, P. F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., Arena, S. 2020. Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. Sustainability, 12(11), 4776.

Quintana, G., Garcia-Romeu, M. L., Ciurana, J. 2011. Surface roughness monitoring application based on artificial neural networks for ball-end milling operations. Journal of Intelligent Manufacturing, 224, 607-617.

Razaviarab, N., Sharifi, S., Banadaki, Y. M. 2019. Smart additive manufacturing empowered by a closed-loop machine learning algorithm, In Nano-, Bio-, Info-Tech Sensors and 3D Systems III, International Society for Optics and Photonics, Vol. 10969 2009, p. 109690H.

Shao, S. Y., Sun, W. J., Yan, R. Q., Wang, P., Gao, R. X. 2017. A deep learning approach for fault diagnosis of induction motors in manufacturing. Chinese Journal of Mechanical Engineering, 306, 1347-1356.

Singh, Y., Kaur, A., Malhotra, R. 2009. Comparative analysis of regression and machine learning methods for predicting fault proneness models. International journal of computer applications in technology, 352-4, 183-193.

Smagulova, K., & James, A. P. 2019. A survey on LSTM memristive neural network architectures and applications. The European Physical Journal Special Topics, 228(10), 2313-2324.

Tan, Z., Pan, P. 2019. Network Fault Prediction Based on CNN-LSTM Hybrid Neural Network. In 2019 International Conference on Communications, Information System and Computer Engineering CISCE pp. 486-490. IEEE.

Tao, F., Qi, Q., Liu, A., Kusiak, A. 2018. Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157-169.

Venkatesan, P., & Anitha, S. 2006. Application of a radial basis function neural network for diagnosis of diabetes mellitus. Current Science, 91(9), 1195-1199.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., Wu, D. 2018. Deep learning for smart manufacturing: Methods and applications. Journal of Manufacturing Systems, 48, 144-156.

Ye, Q., Yang, X., Chen, C., Wang, J. 2019. River Water Quality Parameters Prediction Method Based on LSTM-RNN Model. In 2019 Chinese Control And Decision Conference CCDC pp. 3024-3028. IEEE.

Yilmaz, I., Kaynar, O. 2011. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert systems with applications, 38(5), 5958-5966.

Zhang, S., Wang, Y., Liu, M., Bao, Z. 2017a. Data-based line trip fault prediction in power systems using LSTM networks and SVM. IEEE Access, 6, 7675-7686.

Zhang, Y., Xiong, R., He, H., Liu, Z. 2017b, July. A LSTM-RNN method for the lithuim-ion battery remaining useful life prediction. In 2017 Prognostics and System Health Management Conference PHM-Harbin pp. 1-4. IEEE.

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., Gao, R. X. 2019. Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213-237.

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(1) (2022) 35-41
DOI: 10.38016/jista.966517

Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması

Ebru Aydındağ Bayrak[1]*, Pınar Kırcı[2], Tolga Ensari[3], Engin Seven[4], Mustafa Dağtekin[5]

[1] İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Bilimleri Bölümü, İstanbul, Türkiye

[2] Bursa Uludağ Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye

[3] Arkansas Tech University, Department of Computer and Information Science, Russellville, USA

[4,5] İstanbul Üniversitesi-Cerrahpaşa, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

ebruaydindag@gmail.com, pinarkirci@uludag.edu.tr, tensari@atu.edu, engin.seven@ogr.iu.edu.tr, dagtekin@iuc.edu.tr

Öz

Kanser dünya genelinde pek çok insanın ölümüne sebep olan en önemli hastalıklardan biridir. Özellikle göğüs kanseri kadınlar arasında en çok rastlanan hastalıkların başında yer almaktadır. Bu sebeple kanser hastalığının teşhisi ile alakalı herhangi bir gelişme insanların sağlıklı bir yaşam sürmesi açısından oldukça önemlidir. Günümüzde makine öğrenmesi yöntemlerinin kullanılması, kanser hastalığının erken teşhisi ve tahmini için yapılan çalışmalara büyük katkılar sağlamaktadır. Bu çalışmada da k-En Yakın Komşu, Destek Vektör Makinaları, Naive Bayes, Karar ağaçları ve Yapay Sinir Ağları gibi beş farklı makine öğrenmesi yöntemleri Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesi üzerinde uygulanmıştır. Elde edilen sonuçlar doğruluk değerleri ve karmaşıklık matrisi değerleri ile verilerek karşılaştırılmıştır. Birinci göğüs kanseri veri kümesi içinde %98,2456 doğruluk oranıyla ve ikinci göğüs kanseri veri kümesinde %93,8596 doğruluk oranıyla Yapay Sinir Ağları (YSA) yönteminde en yüksek doğruluk değerleri elde edilmiştir.

Anahtar kelimeler: Makine öğrenmesi, göğüs kanseri, sınıflandırma, erken teşhis.

Diagnosing Breast Cancer Using Machine Learning Methods

Abstract

Cancer is one of the most important diseases that cause the death of many people around the world. Especially, breast cancer is one of the most common diseases among women. For this reason, any development related to the diagnosis of cancer is critical for people to live healthy lives. Today, the use of machine learning methods makes great contributions to studies for the early diagnosis and prediction of cancer disease. In this study, five different machine learning methods such as k-Nearest Neighbor, Support Vector Machines, Naive Bayes, Decision Trees, and Artificial Neural Networks were applied on two other breast cancer datasets on the Kaggle platform. The obtained results were compared by giving accuracy values and confusion matrix values. The highest accuracy values were obtained in Artificial Neural Networks (ANN) method with an accuracy rate of 98.2456% in the first breast cancer dataset and 93.8596% in the second breast cancer dataset.

Keywords: Machine learning, breast cancer, classification, early diagnosis.

1. Giriş (Introduction)

Kanser dünya genelinde en çok insan ölümüne sebep olan hastalık türleri arasında ikinci sırada yer almaktadır ve 2018 yılında dünya genelinde yaklaşık olarak 9,6 milyon insan kanser hastalığından dolayı hayatını kaybetmiştir. Yapılan araştırmalarda dünyadaki her 6 ölümden 1 tanesinin kanser yüzünden gerçekleştiği görülmektedir. Özellikle, az ve orta gelişmiş ülkelerde meydana gelen ölümlerin %70'i de yine kanser hastalığından kaynaklanmaktadır (Cancer, 2021).

Kadınlar arasındaki en yaygın kanser türleri ise göğüs, akciğer ve kalın bağırsak kanserleri olup, bu üç kanser türü kadınlarda meydana gelen tüm kanser vakalarının yarısını oluşturmaktadır. Ayrıca meme kanseri kadınlarda teşhis edilen yeni kanser vakalarının %30'luk gibi büyük bir bölümüne karşılık gelmektedir (Siegel vd., 2018).

Bu derece yaygın olan bir hastalığın erkenden teşhis edilmesiyle alakalı yapılacak çalışmaların oldukça önemli olduğu açıkça görülmektedir. Özellikle kanser hastalığı üzerinde gerçekleştirilen makine öğrenmesi uygulamaları incelendiğinde, makine öğrenmesi

* Sorumlu Yazar.
E-posta adresi: ebruaydindag@gmail.com

Alındı : 08 Temmuz 2021
Revizyon : 04 Ekim 2021
Kabul : 30 Ekim 2021

tekniklerinin kanser hastalığının erken teşhis edilebilmesi ve öngörülebilmesi açısından oldukça kullanışlı olduğu açıktır. Makine öğrenmesi yöntemleri var olan verilerin analiz edilmesini ve veri kümesinde var olan ilişkileri ve önemli bilgilerin karakteristik özelliklerinin elde edilmesini sağlar. Ayrıca, verinin iyi şekilde tanımlanabilmesi sağlayan bir hesaplamalı bir model üretir. (Maity ve Das, 2017).

Poyraz (2012) çalışmasında Wisconsin göğüs kanseri veri seti üzerinde veri madenciliği metotlarını uygulayarak, sonuçları başarım ölçütlerine göre karşılaştırmıştır. J48 algoritması Karar ağacı algoritması, Naive Bayes, Lojistik Regresyon ve örnek tabanlı sınıflandırma algoritması olarak K-Star metotları WEKA çalışma ortamında kullanılmıştır. Çalışmanın sonucunda doğruluk değerleri açısından Lojistik Regresyon algoritmasının diğer algoritmalara oranla daha iyi sonuç verdiği görülmüştür (Poyraz, 2012).

Ahmad vd., (2013) çeşitli makine öğrenmesi tekniklerini göğüs kanseri hastalığının iki yıl içinde yeniden ortaya çıkabilme durumunun tahmin edilmesiyle ilgili kullanmışlardır. Çalışmalarında Tahran Ulusal Kanser Enstitüsü veri tabanında yer alan ve 1997-2008 yılları arasını kapsayan ICBC (Iranian Center for Breast Cancer) göğüs kanseri veri setini incelemişlerdir. Veri setinde eksik olan değerler Beklenti Maksimizasyonu algoritması kullanılarak düzenlenmiştir. Makine öğrenmesi yöntemlerinden Karar Ağacı (C4.5), Destek Vektör Makinaları (DVM) ve Yapay Sinir Ağları (YSA) uygulanmıştır. Göğüs kanserinin yeniden tekrarlanmasının tahmini üzerine yapılan bu çalışmada en yüksek doğruluk ve en az hata oranı DVM yönteminde elde edilmiştir.

İşeri (2014) mamogram görüntülerine makine öğrenmesi yöntemlerini uygulayarak göğüs kanserinin teşhis edilebilmesi üzerinde çalışmıştır. Çalışma mamogram görüntülerindeki mikro kireçlenme bölgelerinin tespiti ve bu bölgelerin kötü ya da iyi huylu olma durumlarına göre sınıflandırılması olacak şekilde iki aşamada gerçekleştirilmiştir. MATLAB ortamında göğüs kanseri tespit sistemi isimli (BCDS:Breast Cancer Detection System) bir yazılım geliştirilmiştir. Dört adet özellik çıkarım yöntemi ile Çok Katmanlı Yapay Sinir Ağı ve Destek Vektör Makinaları sınıflandırıcı olarak kullanılarak göğüs kanseri bulgularının tespiti amaçlanmıştır.

Şık (2014) çalışmasında kanser hastalığının erken teşhis edilebilmesinde veri madenciliği uygulamalarının etkisini araştırmıştır. Wisconsin Göğüs Kanseri veri setine WEKA ortamında Bayes Ağı, Naive Bayes, Çok Katmanlı Algılayıcı, Basit Lojistik, Olasılıksal Gradyan İniş, Sıralı Minimal Optimizasyon, IB1, K-Yıldız, PART, Lojistik Model Ağaçları ve Rassal Ormanlar gibi çeşitli sınıflandırma yöntemlerini uygulamıştır. Sınıflandırma sonuçlarını karşılaştırılırken Kappa istatistiği, doğruluk, kesinlik, duyarlılık, F-ölçütü ve ROC alanı gibi parametreler göz önüne alınmıştır. 0,94 Kappa istatistiği, %97,40 doğruluk, 0.97 kesinlik, 0,99 duyarlılık 0,98 F-ölçütü ve 1,00 ROC alanı sonuçlarına göre Basit Lojistik sınıflandırma yöntemi en iyi sonucu vermiştir.

Asri vd., (2016) UCI Makine Öğrenmesi Veri Havuzunda yer alan Wisconsin meme kanseri veri setine Destek Vektör Makinaları, Karar Ağacı (C4.5), Naive Bayes ve k-En Yakın Komşu makine öğrenmesi algoritmalarını WEKA ortamında uygulamışlardır. Yapmış oldukları çalışmada sınıflandırma modellerini değerlendirirken doğruluk, hassasiyet, duyarlılık ve özgüllük parametreleri kullanılmıştır. Uygulamanın sonucunda Destek Vektör Makinaları %97,13 gibi yüksek doğruluk oranı ve %0.02 hata oranıyla en iyi sonucu vermiştir.

Bazazeh ve Shubair (2016) göğüs kanserinin erken teşhis edilmesiyle ilgili yapmış oldukları bu çalışmada Destek Vektör Makinaları, Rastgele Orman ve Bayes Ağı yöntemlerini Wisconsin göğüs kanseri veri setine uygulamışlardır. WEKA yazılımının kullanıldığı bu çalışmada duyarlılık ve hassasiyet değerlerine göre Bayes Ağı en iyi performansı göstermiştir. ROC eğrisi parametresi dikkate alındığında ise Rastgele Orman yöntemi en iyi sonucu vermiştir. Doğruluk, özgünlük ve hassasiyet cinsinden ise en iyi performansı Destek Vektör Makinaları göstermiştir.

Anwer (2017) tez çalışmasında Python ortamında Wisconsin göğüs kanseri veri seti üzerinde çeşitli derin öğrenme algoritmalarını uygulayarak performans sonuçları üzerinden karşılaştırma yapmıştır. Tam bağlantılı sinir ağları, konvolüsyon sinir ağları, basit tekrarlayan sinir ağları, uzun kısa dönem yapay sinir ağları ve kapılı yinelenen birim sinir ağları gibi çeşitli derin öğrenme yöntemleri kullanılmıştır. Ayrıca çalışmada Naive Bayes, k-En Yakın Komşu, Lojistik Regresyon ve Karar Ağacı gibi klasik makine öğrenmesi yöntemleri de uygulanmıştır. Çalışmanın sonucunda derin öğrenme yöntemlerinin klasik makine öğrenmesi yöntemlerine göre daha üstün çalıştığı sonucu elde edilmiştir. Konvolüsyon sinir ağı yönteminde %99,30 ile en yüksek doğruluk değeri elde edilmiştir.

Maity ve Das (2017) Image J programını kullanarak göğüs kanseri hücre görüntülerinden özellik çıkarımı gerçekleştirmiş ve Yapay Sinir Ağı (YSA) algoritmasını uygulamışlardır. Çalışmanın sonucunda %90 doğruluk oranı ile göğüs kanseri hücre görüntüleri doğru şekilde sınıflandırılabilmiştir.

Turgut (2017) yapmış olduğu tez çalışmasında Python ortamında çeşitli makine öğrenmesi yöntemlerini iki farklı mikro dizi göğüs kanseri veri setlerine uygulamıştır. Çalışmada öznitelik seçimleri yapılarak makine öğrenmesi metotlarıyla yüksek doğrulukta tahmin yapılabilmesi amaçlanmıştır. Çalışmada DVM, YSA, k-EYK, Karar Ağaçları, Rastgele Orman, Lojistik Regresyon, Adaboost ve Gradyan Boosting Makina algoritmaları kullanılmıştır. İki göğüs kanseri veri kümesinde de öznitelik yöntemlerin uygulandıktan sonra en yüksek doğruluk DVM yönteminde, en düşük doğruluk Karar Ağaçları yönteminde elde edilmiştir. Ayrıca iki veri kümesinde de kullanılan aynı öznitelik yöntemleri çalışmada

uygulanan tüm makine öğrenmesi algoritmalarında birbirine yakın sonuçlar vermiştir.

Sherafatiyan (2018) çalışmasında göğüs kanseri hastalarının miRNA ekspresyon veri setlerini kullanıp minimal biyo-belirteçleri belirlemek için ağaç tabanlı sınıflandırma modellerinden yararlanmıştır. Önerilen biyo-belirteçlere ek olarak göğüs kanseri sınıflandırmasındaki en önemli mikro RNA'larda açıklanmıştır.

Turgut vd. (2018) çeşitli makine öğrenmesi yöntemlerini iki farklı mikro dizi göğüs kanseri veri seti üzerinde uygulayarak veri sınıflandırması yapmışlardır. Rastgele lojistik regresyon ve yinelemeli öznitelik eleme özellik seçim yöntemleri kullanılarak yüksek doğrulukta kanser teşhisinin yapılması amaçlanmıştır. İki farklı özellik seçim yöntemi uygulandıktan sonra iki mikro dizi göğüs kanseri veri seti içinde destek vektör makinaları en iyi performansı göstermiştir.

Aydındağ Bayrak vd., (2019) Wisconsin (orijinal) göğüs kanseri veri seti üzerinde yapmış olduğu çalışmada, WEKA ortamında Destek Vektör Makinaları ve Yapay Sinir Ağı makine öğrenmesi yöntemlerini uygulamışlardır. Algoritmaların sonuçları doğruluk, hassasiyet, duyarlılık ve ROC alanı gibi performans metriklerine göre karşılaştırıldığında Destek Vektör Makinaları (SMO algoritması) en iyi performansı göstermiştir.

Dhahri vd. (2019) makine öğrenmesi algoritmalarına dayanarak otomatik olarak göğüs kanserinin teşhis edilebilmesi üzerine çalışmışlardır. Özellik tabanlı ön işleme yöntemleri ve sınıflandırma algoritmalarının birleştirilerek kullanılmasının göğüs kanserinin teşhisi üzerinde daha iyi sonuç verebileceğini açıklamışlardır.

Ganggayah vd. (2019) makine öğrenmesi modellerini kullanarak göğüs kanseri hastalığından hayatta kalabilmek için önemli olan prognostik faktörleri belirlemeye çalışmışlardır. Destek Vektör Makinası, Rassal Ağaç, Yapay Sinir Ağları, Ekstrem Boost, Lojistik Regresyon ve Karar Ağacı gibi pek çok makine öğrenmesi algoritması çalışmada kullanılmıştır. Rastgele Orman yöntemi diğer yöntemlere oranla biraz daha yüksek performans sergilemiştir. Buna rağmen çalışmada kullanılan tüm algoritmalar birbirine yakın doğruluk değerleri vermiştir.

Tapak vd. (2019) göğüs kanseri hastaları üzerinde yapmış oldukları çalışmada makine öğrenmesi yöntemlerini kullanarak hayatta kalma ve metastaz tahmininde bulunmuşlardır. Naive Bayes, Rassal Orman, AdaBoost, Destek Vektör Makinası En Küçük Kareler Destek Vektör Makinası, Adabag, Lojistik Regresyon ve Lineer Diskriminant Analizi yöntemlerini uygulamışlardır.

Tseng vd. (2019) makine öğrenmesi teknolojilerini kullanarak göğüs kanseri metastazını belirlemek üzerine çalışma yapmışlardır. Rastgele Orman temelli makine öğrenmesi modelinin göğüs kanseri metastazını en az üç ay önceden tahmin etmek için en uygun yöntem olduğunu belirlemişlerdir.

Magna vd. (2020) hastaların tıbbi geçmiş bilgilerinden faydalanarak göğüs kanserinin sınıflandırmasında makine öğrenmesi, derin öğrenme ve kelime yerleştirme uygulamalarının kullanılması üzerinde çalışma yapmışlardır. Hekimin karar vermesini destekleyen bir öneri sistemi ortaya koymaya çalışmışlardır.

Reddy vd. (2020) destek değerine sahip derin sinir ağı (DNNS) yöntemini göğüs kanseri teşhisi için kullanmıştır. Deneysel sonuçlara göre, önerilen DNNS'nin mevcut yöntemlerden daha iyi sonuçlar verdiği kanıtlanmıştır.

Saxena ve Gyanchandani (2020) histopatolojiyi kullanarak bilgisayar destekli göğüs kanseri teşhisi yapabilmek için makine öğrenmesi yöntemlerini incelemişlerdir. Pek çok farklı yaklaşımı inceledikten sonra göğüs kanseri üzerine yapılan makine öğrenmesi çalışmalarının genellikle derin öğrenme konusunda yoğunlaştığı görülmüştür.

Bu çalışmada da popüler makine öğrenmesi yöntemlerinden olan k-En Yakın Komşu, Destek Vektör Makinaları, Navie Bayes, Karar Ağacı ve Yapay Sinir Ağları Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesine uygulanmıştır. Uygulamada elde edilen sonuçlar doğruluk performans metriği ve karmaşıklık matrisine göre karşılaştırılmıştır. Çalışmanın devamında, ikinci bölümde kullanılan veri kümelerinin ve makine öğrenmesi yöntemleri kısaca açıklanmaktadır. Ardından uygulamanın deneysel sonuçları, tartışma ve sonuç bölümleri ile çalışma sona ermektedir.

2. Materyal ve Yöntem **(Material and Method)**

2.1. Veri seti (Data set)

Bu çalışmada göğüs kanseri sınıflandırması için Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesi kullanılmıştır. Birinci veri kümesinde 30 adet özellik içermekte olup toplamda 569 kayıt yer almaktadır. İkinci veri kümesinde ise 569 kayıt yer alıp 5 adet özellik içermektedir. Bu niteliklerden bazıları ortalama yarıçap, ortalama doku, ortalama çevre, ortalama alan ve ortalama yumuşaklıktır. Veri seti Wisconsin-Madison Üniversitesi Hastanesinden elde edilmiştir (Kaggle, 2020).

2.2. Kullanılan makine öğrenmesi yöntemleri (Utilized machine learning methods)

Yapılan çalışmada iki farklı göğüs kanseri veri kümesi için kategorik verilerin nümerik verilere dönüştürülmesi, gereksiz verilerin atılması, tekrarlanan verilerin kaldırılması, normalleştirme gibi veri ön işleme basamakları gerçekleştirilmiştir. Veri ön işleme basamağından sonra elde edilen verilere k-En Yakın Komşu, Destek Vektör Makinaları, Karar Ağacı, Naive Bayes ve Yapay Sinir Ağları gibi popüler makine öğrenmesi yöntemleri Jüpiter notebook ortamında

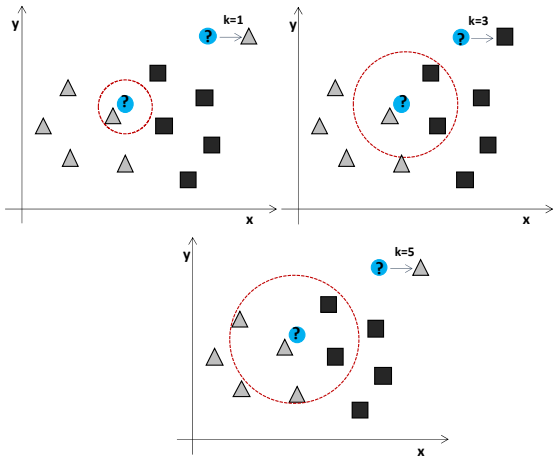
uygulanmıştır. Aşağıda çalışmada kullanılan makine öğrenmesi sınıflandırma yöntemleri kısaca açıklanmaktadır.

2.2.1 k-En Yakın Komşu (k-Nearest Neighbor)

k-En Yakın Komşu algoritması basit ve etkili, aynı zamanda da güçlü bir sınıflandırma yöntemidir. Verilerin sınıflandırılmasında, verilerin birbirleri arasındaki mesafe ölçümü kavramı kullanılmaktadır. Bu yöntem bir denetimli öğrenme yöntemidir, bu nedenle tüm veriler etiketlidir ve her bir veri parçasının hangi sınıfa girmesi gerektiği bilinmektedir. Etiketsiz yani yeni bir veri parçası bize verildiğinde ise, sınıflandırılması için yöntemde uygulanan adımlar aşağıdaki gibi özetlenmektedir (Harrington, 2012):

1. k parametresi belirlenir. k, yeni verilere en yakın olan komşuların sayısıdır.
2. Yeni veri (test) ile mevcut veri (eğitim) arasındaki mesafeler hesaplanır.
3. En yakın mesafe değerleri seçilir. (En yakın komşu bulunur.)
4. Hangi sınıfta en fazla sayıda benzer veri bulursa yeni veriler o sınıfa düşer.

Çalışmada k-en yakın komşu algoritması uygulanırken, en yakın komşu parametresi k=3 olarak belirlenmiştir.

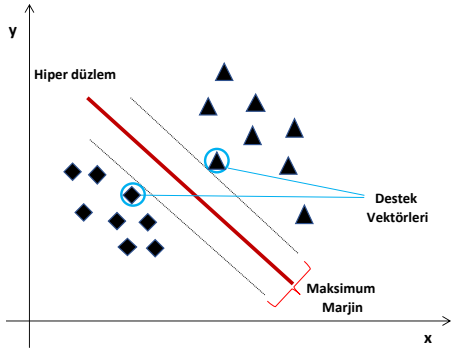


Şekil 1: k-en yakın komşu algoritmasının mimarisi (Ruan vd., 2017'den uyarlanmıştır) (Architecture of the k-nearest neighbor algorithm)

2.2.2 Destek Vektör Makinaları (Support Vector Machines)

Destek Vektör Makinaları (DVM) algoritması ilk kez Vladimir Vapnik (1995) tarafından geliştirilmiştir. DVM en temel haliyle destek vektörlerini kullanarak veri sınıflarını birbirinden ayırmak için en uygun hiper düzlemi bulmaya çalışan, sınıflandırma ve regresyon için kullanılabilen bir makine öğrenme yöntemi olarak açıklanmaktadır. Bu yöntemde öncelikle iki veri sınıfını ayırmak için en uygun hiper düzlemin bulunması amaçlanır ve ardından veri sınıflarının aralarındaki marjin maksimize edildiğinde iki sınıf birbirinden ayrılmaktadır. Eğer sınıflar basit bir hiper düzlemle birbirinden ayrılmazsa, veriler daha yüksek boyutlu yeni bir alana aktarılır ve verileri ayırabilmek için hiper düzlemin bulunması amaçlanır (Burakgazi, 2017).

Çalışmada DVM yöntemi uygulanırken C düzenleme parametresi 1 olarak belirlenmiştir. Ayrıca çekirdek olarak radyal tabanlı kernel fonksiyonu ve gamma parametresinin değeri ise otomatik olarak tercih edilmiştir.



Şekil 2: Destek Vektör Makinaları algoritmasının mimarisi (Alpaydın (2014)'ten uyarlanmıştır) (Architecture of the Support Vector Machines algorithm)

2.2.3. Sade Bayes (Naive Bayes)

Naive Bayes Sınıflandırıcı, Bayes teoremine dayanan en popüler sınıflandırma yöntemlerinden bir tanesidir. Çok basit bir yöntemdir, öyle ki sadece az miktarda eğitim verisiyle bile verilen örnekler sınıflandırabilmektedir. Mevcut ve geçmiş frekans olaylarını hesaplamak için kullanılan Naive Bayes algoritması aşağıdaki gibi açıklanabilir (Umadevi ve Marseline, 2017):

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

P(A): A'nın önsel olasılığı, yalnızca A'nın oluşumlarını saymaktadır.
P(A|B): B verildiğinde A'nın koşullu olasılığıdır. Ayrıca sonsal olasılık olarak adlandırılıp, A'nın B'den türetildiği anlamına gelmektedir.
P(B|A): A verildiğinde B'nin koşullu olasılığıdır.
P(B): B'nin önsel olasılığıdır.

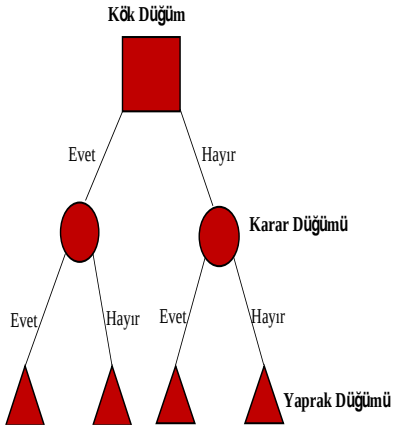
2.2.4. Karar Ağaçları (Decision Trees)

Karar ağaçları algoritmasının temeli böl ve yönet stratejisine dayanmaktadır. Karar düğümleri ve yapraklardan oluşan hiyerarşik bir yapıya sahiptir (Umadevi ve Marseline, 2017).

Karar ağacı yönteminde öncelikle verileri bölmek için hangi özelliğin kullanılacağına karar verilmelidir.

Bunun için her özellik ve ölçüm denenmeli ve ardından elde edilen en iyi sonuçlara göre veri kümelerini alt kümelere bölebiliriz. Yöntemin uygulama adımları aşağıdaki gibi özetlenebilir (Harrington, 2012):

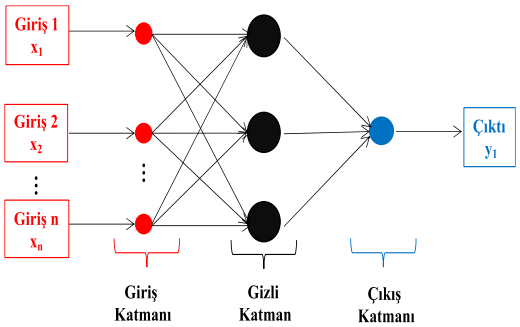
1. Öncelikle tüm veri seti kullanılır.
2. Veri kümesi, bir özelliğin değerine göre iki alt gruba ayrılır. (Bölünen en iyi özellik)
3. Özelliğin tümü aynı sınıfta olana kadar her alt kümeye aynı prosedür uygulanır. Aksi halde bölme işlemine devam edilir.



Şekil 3: Karar Ağacı algoritmasının yapısı gösterilmektedir (Alpaydın (2014)'ten uyarlanmıştır) (The structure of the Decision Tree algorithm)

2.2.5. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay Sinir Ağlarının (YSA) mühendislik çalışmalarındaki amacı sadece insan beynin modellenmesi değildir. Amaç yapay sinir ağlarını kullanarak daha iyi bilgisayarlar yapmak ve insanlara fayda sağlamaktır. İnsan beyni görüntü, konuşma, öğrenme ve tanıma gibi yetenekleri açısından bir mühendislik ürününden fazlasıdır ve bu yeteneklerin yapay zekâ ağları aracılığıyla bilgisayarlara uygulanması oldukça önemlidir (Umadevi ve Marseline, 2017).



Şekil 4: Yapay Sinir Ağlarının (YSA) temel yapısı (Atalay ve Çelik (2017)'den uyarlanmıştır) (Basic structure of Artificial Neural Networks (ANN))

Çalışmada YSA uygulanırken Keras'taki Sequential model tipi kullanılmıştır. Giriş ve çıkış katmanları dahil olmak üzere toplamda 5 katmanlı bir YSA modeli inşa edilmiştir. 3 gizli katmanda nöron sayısı sırasıyla 32, 16 ve 8 olarak belirlenmiş ve Doğrultulmuş Doğrusal Birim (ReLu-Rectified Linear Unit) aktivasyon fonksiyonu kullanılmıştır. Döngü sayısı (epoch) 50, küme büyüklüğü (batch size) ise 10 olarak belirlenmiştir. Ayrıca diğer katmanlarda Doğrultulmuş Doğrusal Birim (ReLu-Rectified Linear Unit) ile Sigmoid aktivasyon fonksiyonu da kullanılmıştır.

YSA modeli derlenirken optimizer için adam algoritması, loss fonksiyonu için binary_crossentropy algoritması ve metrik parametresi için doğruluk değeri kullanılmıştır.

3. Bulgular ve Tartışma (Results and Discussion)

Yapılan çalışmada iki farklı göğüs kanseri veri kümesi için yukarıda bahsedilen beş farklı makine öğrenmesi yöntemleri kullanılmıştır. Veri kümeleri için bazı veri ön işleme basamakları gerçekleştirilmiştir. Veri ön işlemenin ardından yapılan uygulamalarda, birinci veri kümesi için en yüksek doğruluk oranı ile Yapay Sinir Ağları (YSA) yönteminde elde edilmiştir. Tablo 1'de uygulanan beş makine öğrenmesi yönteminin doğruluk değerleri karşılaştırılmıştır. Sonuçlara bakıldığında, YSA yönteminde doğruluk değeri %98,2456 olarak hesaplanmıştır. YSA'dan sonra en yüksek doğruluk değeri ise Destek Vektör Makinalarında (DVM) elde edilmiştir. Bu sonuçları sırasıyla k-En Yakın Komşu, Naive Bayes ve Karar Ağacı algoritmaları takip etmektedir.

Tablo 1: Birinci veri kümesi için makine öğrenmesi yöntemlerinin doğruluk değerlerinin gösterilmesi (The accuracy values of machine learning methods for the first dataset)

Uygulanan Makine Öğrenmesi Yöntemleri	Doğruluk Oranı (%)
k-En Yakın Komşu	94,7368
Destek Vektör Makinaları	97,3684
Naive Bayes	92,1053
Karar Ağacı	86,8421
Yapay Sinir Ağları	**98,2456**

Tablo 2'de ise YSA yöntemine ait karmaşıklık matrisi (confusion matrix) gösterilmiştir. YSA yöntemi için yapılan testlerde ise performans metriklerinden f1 skoru 0,961538 olarak bulunmuştur. Ayrıca duyarlılık ve hassasiyet değerleri 0,961538 olarak bulunmuştur.

Tablo 2: Birinci göğüs kanseri veri kümesine uygulanan YSA yönteminin karmaşıklık matrisi (The complexity matrix of the ANN method applied to the first breast cancer dataset).

		TAHMİN		METRİKLER
		Pozitif	Negatif	
GERÇEK	Pozitif	TP=25	FP=1	Hassasiyet= $\frac{TP}{TP+FP}$=0.961538
	Negatif	FN=1	TN=87	Duyarlılık= $\frac{TP}{TP+FN}$=0.961538
Doğruluk $= \frac{TP+TN}{TP+TN+FP+FN}$=0.982456				F1 Skoru $= 2 * \frac{Hassayiyet * Duyarlılık}{Hassayiyet + Duyarlılık}$=0.961538

İkinci göğüs kanseri veri kümesi için en yüksek doğruluk oranı yine Yapay Sinir Ağları (YSA) yöntemi ile elde edilmiştir. Tablo 3'te de uygulanan beş farklı sınıflandırma yöntemlerinin doğruluk değerleri karşılaştırılmıştır. %93,8596 doğruluk değeriyle YSA yönteminde en yüksek doğruluk değeri hesaplanmıştır. YSA'dan sonra en yüksek doğruluk değeri birinci veri kümesinde olduğu gibi Destek Vektör Makinalarında elde edilmiştir. k-En Yakın Komşu, Naive Bayes ve Karar Ağacı yöntemleri sırasıyla doğruluk değerleri sıralamasında yer almaktadır.

Tablo 3*:* İkinci veri kümesi için makine öğrenmesi yöntemlerinin doğruluk değerlerinin gösterilmesi (Demonstrating the accuracy values of machine learning methods for the second dataset)

Uygulanan Makine Öğrenmesi Yöntemleri	Doğruluk Oranı (%)
k-En Yakın Komşu	90,3509
Destek Vektör Makinaları	92,1053
Naive Bayes	89,4737
Karar Ağacı	85,0877
Yapay Sinir Ağları	**93,8596**

İkinci göğüs kanseri veri kümesi için Yapay Sinir Ağları (YSA) ile yapılan testlerde f1 skoru 0,959537 olarak bulunmuştur. Ayrıca duyarlılık değeri 0,943181 ve hassasiyet değerleri 0,976470 olarak hesaplanmıştır. Tablo 4'te YSA yöntemine ait karmaşıklık (confusion) matrisi gösterilmiştir.

Tablo 4: İkinci göğüs kanseri veri kümesine uygulanan YSA yönteminin karmaşıklık matrisi (The complexity matrix of the ANN method applied to the second breast cancer dataset).

		TAHMİN		METRİKLER
		Pozitif	Negatif	
GERÇEK	Pozitif	TP=83	FP=2	Hassasiyet= $\frac{TP}{TP+FP}$=0.976470
	Negatif	FN=5	TN=24	Duyarlılık= $\frac{TP}{TP+FN}$=0.943181
Doğruluk= $\frac{TP+TN}{TP+TN+FP+FN}$=0.938596				F1 Skoru $= 2 * \frac{Hassayiyet * Duyarlılık}{Hassayiyet + Duyarlılık}$=0.959537

Uygulanan makine öğrenmesi yöntemlerinin performans sonuçları doğruluk değerlerine göre karşılaştırılmıştır. Bu sonuçlara göre Yapay Sinir Ağları yöntemi çalışmada kullanılan diğer makine öğrenmesi yöntemlerine göre bu problemin sınıflandırılmasında daha iyi performans sergilemiştir. İki farklı göğüs kanseri veri kümesi içinde yapılan çalışmada YSA yönteminde en yüksek doğruluk değerleri elde edilmiştir.

Kanser hastalığının erken teşhisi ve tanısı ile ilgili bilgisayar destekli çalışmaların yüksek doğruluk oranıyla gerçekleşmesi hastalığın tedavisi açısından önemli bir adımdır. Kanser hastalığından kaynaklı kayıplar düşünüldüğünde erken teşhis ile ilgili herhangi bir gelişme oldukça büyük önem taşımaktadır.

Kanser hastalığının erken teşhis edilebilmesine katkı sağlamak adına yapılan bu çalışmada Kaggle platformunda yer alan iki farklı göğüs kanseri veri kümesi kullanılarak, farklı makine öğrenmesi yöntemleriyle sınıflandırma işlemi gerçekleştirilmiştir. k-En Yakın Komşu, Destek Vektör Makinaları, Navie Bayes, Karar Ağacı ve Yapay Sinir Ağları iki farklı göğüs kanseri veri kümesine uygulanarak elde edilen sonuçlar doğruluk performans metriği ve karmaşıklık matrisine göre karşılaştırılmıştır. Uygulanan tüm makine öğrenmesi yöntemlerinin doğruluk değerleri genel olarak yüksek hesaplanmıştır. İki farklı göğüs kanseri veri kümesi içinde Yapay Sinir Ağları (YSA) yönteminde diğer sınıflandırma algoritmalarına kıyasla daha yüksek doğruluk değeri elde edilmiştir. Birinci göğüs kanseri veri kümesi için YSA yönteminde %98,2456, ikinci göğüs kanseri veri kümesi içinde %93,8596 doğruluk değerleri hesaplanmıştır.

Kaynaklar (References)

Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three

machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4*(124), 3.

Alpaydın, E. (2013). *Yapay öğrenme*, 2. Baskı, Boğaziçi Üniversitesi Yayınevi, ISBN-13: 978-6-054-23849-1.

Alpaydın, E. (2014). Introduction to Machine Learning. MIT Press.

Anwer, A. M. O., (2017). *Derin Öğrenme Yöntemleri ile Göğüs Kanseri Teşhisi.* Yüksek Lisans Tezi, Türk Hava Kurumu Üniversitesi, Fen Bilimleri Enstitüsü.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.

Atalay, M., & Çelik, Ö. G. E. (2017). Artificial Intelligence and Machine Learning Applications in Big Data Analysis. Mehmet Akif Ersoy University Journal of Social Sciences Institute, 9(22), 155–172.

Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, Nisan). Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1-3). IEEE.

Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *Electronic Devices, Systems and Applications (ICEDSA), 2016 5th International Conference on* (pp. 1-4). IEEE.

Burakgazi, Y., 2017, *Identification of Breast Cancer Sub-Types by Using Machine Learning Techniques*, M.Sc Thesis, Dokuz Eylül University, Graduate School of Natural and Applied Sciences.

Cancer, 2021, https://www.who.int/en/news-room/fact-sheets/detail/cancer, 01.04.2021.

Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2019*.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*, *19*(1), 48.

Harrington, P. (2012). *Machine learning in Action*, Vol. 5, Greenwich, CT: Manning.

İşeri, İ. (2014). *Mamogram Görüntülerinden Makine Öğrenmesi Yöntemleri ile Meme Kanseri Teşhisi,* Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü.

Kaggle, 2020, https://www.kaggle.com/yuqing01/breast-cancer, https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset, 01.04.2021.

Magna, A. A. R., Allende-Cid, H., Taramasco, C., Becerra, C., & Figueroa, R. L. (2020). Application of Machine Learning and Word Embeddings in the Classification of Cancer Diagnosis Using Patient Anamnesis. *IEEE Access*, *8*, 106198-106213.

Maity, N. G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. In *2017 IEEE Aerospace Conference*, pp. 1-9.

Poyraz, O. (2012). *Tıp'da Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi. Yüksek* Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri, Enstitüsü.

Reddy, A., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. *ICT Express*.

Ruan, Y., Xue, X., Liu, H., Tan, J., & Li, X. (2017). Quantum algorithm for k-nearest neighbors classification based on the metric of hamming distance. International Journal of Theoretical Physics, 56(11), 3496–3507. Doi:10.100710773-017-3514-4.

Saxena, S., & Gyanchandani, M. (2020). Machine Learning Methods for Computer-Aided Breast Cancer Diagnosis Using Histopathology: A Narrative Review. *Journal of Medical Imaging and Radiation Sciences*, *51*(1), 182-193.

Sherafatian, M. (2018). Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene*, *677*, 111-118.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, *Ca-a Cancer Journal for Clinicians*, *68* (1), pp. 7-30.

Şık, M. Ş., 2014, *Veri Madenciliği ve Kanser Erken Teşhisinde Kullanımı,* Yüksek Lisans Tezi, İnönü Üniversitesi, Sosyal Bilimler Enstitüsü.

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2019). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*, *7*(3), 293-299.

Tseng, Y. J., Huang, C. E., Wen, C. N., Lai, P. Y., Wu, M. H., Sun, Y. C., & Lu, J. J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *International journal of medical informatics*, *128*, 79-86.

Turgut, S. (2017). *Makine Öğrenmesi Yöntemleri Kullanarak Kanser Teşhisi,* Yüksek Lisans Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü.

Turgut, S., Dağtekin, M. and Ensari, T. (2018). "Microarray breast cancer data classification using machine learning methods," *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Istanbul, pp. 1-3, doi: 10.1109/EBBT.2018.8391468.

Umadevi, S., & Marseline, K. J. (2017, July). A survey on data mining classification algorithms. In *2017 International Conference on Signal Processing and Communication (ICSPC)* (pp. 264-268). IEEE

Research Article
Journal of Intelligent Systems: Theory and Applications 5(1) 2022: 42-56
DOI: 10.38016/jista.979582

Determination of the Focus Strategies Related to Renewable Energy For Turkey by Using the Fuzzy Sectional SWOT

Buket Karatop[1], Büşra Taşkan[2], Elanur Adar[3]*

[1] İstanbul University Cerrahpaşa, Department of Motor Vehicles and Transportation Technologies, İstanbul, Turkey

[2] Muş Alparslan University, Department of Industrial Engineering, Muş, Turkey

[3] Artvin Coruh University, Department of Environmental and Civil Engineering, Artvin, Turkey

buket.karatop@istanbul.edu.tr, b.taskan@alparslan.edu.tr, aelanur@artvin.edu.tr

Abstract

Determining the strategies that Turkey need to focus at the renewable energy field is aimed in this study. For this, an integrated method called the Fuzzy Sectional SWOT consisting of the Fuzzy AHP and the Sectional SWOT methods was used. Some disadvantages of the traditional SWOT analysis are eliminated with the method used. Firstly, the renewable energy field was divided into 6 sub-sections (hyropower, solar, wind, biomass, hydrogen and geothermal) according to the logic of Sectional SWOT analysis. Then strengths, weaknesses, opportunities and threats for each of these sub-sections were determined using the Sectional SWOT analysis. Weights were found with the Fuzzy AHP method for each of the renewable energy sources and they were prioritized according to these weights. Finally, focus strategies related to renewable energy field for Turkey were obtained with the creation of strategies related to renewable energy sources. Consequently, the focus strategies which should be primarily addressed are related to use of renewable energy potential, social awareness about the renewable energy, government supports and incentives related to the renewable energy, selection of suitable areas for renewable energy plants and domestic production of the constituent parts of renewable energy plants.

Keywords: Renewable energy, Sectional SWOT, Fuzzy AHP, Strategy

Bulanık Parçalı SWOT Kullanılarak Türkiye için Yenilenebilir Enerjiye İlişkin Odak Stratejilerin Belirlenmesi

Öz

Bu çalışmada, Türkiye'nin yenilenebilir enerji alanında odaklanması gereken stratejilerin belirlenmesi amaçlanmıştır. Bunun için Bulanık AHP ve Parçalı SWOT yöntemlerinden oluşan Bulanık Parçalı SWOT adı verilen entegre bir yöntem kullanılmıştır. Geleneksel SWOT analizinin bazı dezavantajları, kullanılan yöntemle ortadan kaldırılmıştır. Öncelikle, yenilenebilir enerji alanı Parçalı SWOT analizi mantığına göre 6 alt-bölüme (hidroelektrik, güneş, rüzgar, biyokütle, hidrojen ve jeotermal) ayrılmıştır. Daha sonra, bu alt bölümlerin her biri için güçlü, zayıf yönler, fırsatlar ve tehditler, Parçalı SWOT analizi kullanılarak belirlenmiştir. Bulanık AHP yöntemi ile yenilenebilir enerji kaynaklarının her biri için ağırlıklar bulunmuş ve kaynaklar bu ağırlıklara göre önceliklendirilmiştir. Son olarak, yenilenebilir enerji kaynaklarına ile ilgili stratejilerin oluşturulması ile Türkiye için yenilenebilir enerji alanına ilişkin odak stratejiler elde edilmiştir. Sonuç olarak öncelikli olarak ele alınması gereken odak stratejiler, yenilenebilir enerji potansiyelinin kullanımı, yenilenebilir enerji konusunda toplumsal farkındalık, yenilenebilir enerji ile ilgili devlet destekleri ve teşvikleri, yenilenebilir enerji santralleri için uygun alanların seçimi ve yenilenebilir enerji santrallerinin bileşenlerinin yerli üretimi ile ilgilidir.

Anahtar Kelimeler: Yenilenebilir enerji, Parçalı SWOT, Bulanık AHP, Strateji

* Corresponding Author.
E-mail: aelanur@artvin.edu.tr

Recieved : 24 Aug 2021
Accepted : 29 Nov 2021

1. Introduction

Due to reasons such as technological developments, industrialization, population growth, the need for energy in the world is constantly increasing and available fossil resources are insufficient to meet this need. Therefore, there has been an acceleration in the search for renewable energy resources (RES) in the world and this development has directed countries to renewable energy. The term "renewable" in RES indicates the main feature of these technologies; the unexplainable and renewable nature and the existence of the basic energy source in human dimensions (Eltrop 2013). The renewable energy is described by Henrik Lund (2010) as "energy that is produced by natural resources - such as sunlight, wind, rain, waves, tides, and geothermal heat - that are naturally replenished within a time span of a few years". The use of renewable energy goes back a long way. Renewable energy sources were largely used before the industrialization period and coal was used as the main energy source in the mid-19th century (Asif and Muneer 2007). Renewable energy and regenerative energy technologies are of great interest today. It is seen that new projects, new technologies and new energy stakeholder groups are emerging everywhere. Countries, regions and cities are competing to be the best in energy rankings. So where does this change in energy come from? (Eltrop 2013). Energy security, economic impacts and CO_2 emission reduction are three main factors which trigger the use RES (Heshmati et al. 2015).

There are enormous potential renewable energy sources worldwide that can provide clean energy such as hydro power, solar, wind and biomass and can increase the long-term supply of sustainable energy (Asif and Muneer 2007). In 2018, the share of RES in total electricity generation in the world was 26%. If it is looked the situation of energy supply world-wide, of the total 13.859.770 ktoe total energy supply in 2018, only 1.975.678 ktoe (hydro: 362.327 ktoe; wind, solar, etc.: 286.376 ktoe; biofuels and waste: 1.326.975 ktoe) belong to RES. If the overall situation of Europe which consists of 43 countries which also contain Turkey is looked, of the total 1.975.678 ktoe total renewable energy supply in 2018, only 303.893 ktoe belong to Europe (IEA, 2021).

According to data of the Republic of Turkey Ministry of Energy and Natural Resources (Online 2020), while %67.1 (from %37.3 coal, %29.8 natural gas) of Turkey electricity production in 2018 were supplied from nonrenewable energy sources, %31.5 (from %19.8 hydraulic energy, %6.6 wind, %2.6 the sun, %2.5 geothermal energy) were supplied from RES and the rest %1,4 were supplied from other sources. As of the end of September 2019, the installed power in electric energy was 90,720 MW and the distribution of this power according to the sources is as in Table 1.

Table 1. In 2019 distribution of installed power in electric energy according to the sources (Online 2020)

Sources of Energy	Installed Power Capacity
Hydropower	%31.4
Natural gas	%28.6
Coal	%22.4
Wind	%8.1
Solar	%6.2
Geothermal	%1.6
Other sources	%1.7

In this study, determining the focus strategies of Turkey related to renewable energy based on the increasing importance of renewable energy in the world is aimed. The study which is structured based on mentioned purpose, consists of 5 sections, including the introduction. In the second part of the study, related literature was mentioned. In the third part, the Sectional SWOT and the Fuzzy AHP methods that constitute the Fuzzy Sectional SWOT approach proposed in the study, were mentioned. In the fourth part, the proposed model was briefly mentioned and Turkey's focus strategies related to renewable energy were determined by making necessary analyses. In the final part, the results were interpreted and the study was completed.

2. Literature Review

Energy planning problems are in the category of issues that countries should consider carefully. Deciding on suitable energy alternatives and creating energy policies are two important ones of these problems. Since MCDM methods are suitable for the structure of these problems, they are frequently used in the literature to solve these problems. Since more sensitive, concrete and realistic results are obtained with Fuzzy MCDM, they are also widely used in solving these problems. When the current literature is examined, the most used method is Fuzzy AHP is seen. Likewise, one of the most used methods in the literature in strategy formulation is SWOT analysis.

2.1. Use of SWOT analysis or MCDM/fuzzy MCDM methods in energy planning problems

Bas (2013) proposed an integrated SWOT-fuzzy TOPSIS method combined with AHP for the analysis of the electricity supply chain in his study. With the SWOT analysis, the factors related to the electricity supply chain in Turkey were determined, the importance weights of these factors were determined with the AHP, and the SWOT factors evaluated were prioritized with the Fuzzy TOPSIS method. Okello et al. (2014) developed an approach integrating Desirability Functions into SWOT and AHP for participatory evaluation of technologies in their study. The application of the method was made through participatory evaluation of 4 bioenergy technologies in Uganda. The results showed that the method is

effective in assessing stakeholder priorities for bioenergy technology. In their study, Tasri and Susilawati (2014) tried to determine the most suitable renewable energy source for electricity generation in Indonesia. For this purpose, a method based on fuzzy AHP is proposed. When the results of the study were interpreted, it was finalized that hydropower is the best alternative.

Ervural et al. (2018) proposed a SWOT analysis based on ANP and fuzzy TOPSIS for Turkey's energy planning in their study. For solving the problem, SWOT analysis was preferred to determine the criteria and sub-criteria related to Turkey's energy sector, ANP was used to determine the weights of these criteria and sub-criteria, and Fuzzy TOPSIS was used to prioritize alternative energy strategies. Sensitivity analysis was used to confirm the results of the study. Gottfried et al. (2018) applied SWOT-AHP-TOWS analysis to increase private investment to China's biogas sector. For this purpose, investment criteria of private stakeholders have been defined in order to increase the active participation of private investors. SWOT analysis was used to define the investment criteria and then these criteria were prioritized using AHP. Finally, strategies related to investments were created using the TOWS matrix.

Erdin and Ozkaya (2019) discussed the location selection problem related to renewable energy sources for the example of Turkey in their study. Finally, the most suitable energy sources are presented according to the geography and energy potential of the regions. Kaya et al. (2019) conducted a comprehensive literature review in order to examine which fuzzy multi-criteria decision-making techniques are used in creating energy policy in the literature. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology was used as the literature review method. Finally, fuzzy AHP was determined as the most used method in the literature on this subject. Solangi et al. (2019) evaluated energy strategies for Pakistan's sustainable energy planning in their study. For this purpose, they proposed an integrated approach based on SWOT, AHP and Fuzzy TOPSIS. SWOT was preferred to determine the factors and sub-factors in sustainable energy planning, AHP was used to determine the weights of all factors and sub-factors, and Fuzzy TOPSIS was used to rank the determined energy strategies. Sensitivity analysis was performed to confirm the results of the study.

Papapostolou et al. (2020) proposed a method based on AHP-SWOT-Fuzzy TOPSIS to establish Cross-Border collaboration on renewable energy sources. SWOT analysis was used to determine the factors and sub-factors related to the subject, AHP was used to determine the weights of these factors and sub-factors, and Fuzzy TOPSIS was used to evaluate the alternatives. Wang et al. (2020) discussed the problem of selection of renewable energy sources that can be used for electricity generation in Pakistan's Sindh and Baluchistan cities in their study. To solve the problem, SWOT was preferred to identify the main factors affecting renewable energy technologies in Pakistan, and Fuzzy AHP was used to evaluate renewable energy alternatives. When the results were commented, it was found that wind energy is the best alternative.

As can be seen from the existing literature, studies on the subject have focused on solving energy planning problems using SWOT and MCDM or Fuzzy MCDM. Therefore, this study contributes to existing literature by making the right investment decisions by creating focus strategies for renewable energy investments using the Fuzzy Sectional SWOT. Because focus strategies were created by the fuzzy sectional SWOT method in this study.

3. Research Methodology

In this study, an approach called the Fuzzy Sectional SWOT which consists of the Fuzzy AHP and the Sectional SWOT, was preferred to decide the focus strategies for renewable energy sources in Turkey. This new approach eliminates some of the shortcomings of the traditional SWOT analysis. Firstly, the Sectional SWOT analysis was developed by extending the SWOT analysis. Then, by integrating the Fuzzy AHP method into the Sectional SWOT method, the Fuzzy Sectional SWOT approach used in the study was obtained.

3.1. A new approach to the SWOT analysis: The Sectional SWOT analysis

Strategies are very important to guide the future of an individual, an event or a company. In general, it is necessary to act from the their current situations in order to determine strategies related to the individual, event or company. The current situation refers to the conditions related to the individual, event or company itself and the environmen that it is located. These conditions are the strengths and weaknesses of the individual, event or company and the opportunities and threats in their environments. These mentioned conditions are investigated with a very famous analysis called as SWOT whose name consists from the combination of first letters of the Strengths, Weaknesses, Opportunities, Threats. The history of the SWOT analysis goes back to the 1960s (Learned et al. 1965) and it was popularized by Andrews (1965), who united the ideas of Peter Drucker, Philip Selznick and Alfred Chandler. The four groups of strategies created by the SWOT analysis are as follows (Sevkli et al. 2012);

1) Strengths and Opportunities (SO) Strategies: Using strengths to take advantage of opportunities.

2) Strengths and Threats (ST) Strategies: Using strengths to avoid from threats

3) Weaknesses and Opportunities (WO) Strategies: Overcoming weaknesses by taking advantage of opportunities

4) Weaknesses and Threats (WT) Strategies: Minimizing weaknesses to avoid from threats

In the literature, the criticisms of the traditional SWOT analysis are as follows; creating extremely long lists, no using weights to reflect priorities, using ambiguous words and expressions, conflicts have no solution, there is no obligation to validate ideas with data or analysis, only requires a level of analysis, no logical link to strategy implementation (Hill and Westbrook 1997). However, there is a more effective method, which is an extended variation of the SWOT analysis and was introduced into the literature as Sectional SWOT by Karatop (2015). In the Sectional SWOT analysis, the main topic is divided into sub-sections as different from the SWOT analysis. In this way, the SWOT analysis is converted to multi-criteria structure with the Sectional SWOT analysis. In addition to strengths and weaknesses, opportunities and threats in the environment are determined separately for each of the sub-sections obtained. Thus carring out a more detailed analysis, more appropriate results is obtained for the determination of strategies (Karatop 2015). The general structure of the sectional SWOT analysis is shown in the Table 2.

Table 2. Sectional SWOT (Karatop et al. 2018)

Main criteria	Sub-criteria	Sectional SWOT			
MC_1	**SC_{11}**	G_{111}	Z_{111}	F_{111}	T_{111}
		G_{112}	Z_{112}	F_{112}	T_{112}
		..	..	..	..
	SC_{12}	G_{121}	Z_{121}	F_{121}	T_{121}
		G_{122}	Z_{122}	F_{122}	T_{122}
		..	..	..	..
	..	..	..	..	..
	SC_{1a}	G_{1a1}	Z_{1a1}	F_{1a1}	T_{1a1}
		G_{1a2}	Z_{1a2}	F_{1a2}	T_{1a2}
		..	..	..	..
MC_2	**SC_{21}**	G_{211}	Z_{211}	F_{211}	T_{211}
		G_{212}	Z_{212}	F_{212}	T_{212}
		..	..	..	..
	SC_{22}	G_{221}	Z_{221}	F_{221}	T_{221}
		G_{222}	Z_{222}	F_{222}	T_{222}
		..	..	..	..
	..	..	..	..	..
	SC_{2b}	G_{2b1}	Z_{1b1}	F_{1b1}	T_{1b1}
		G_{2b2}	Z_{1b2}	F_{1b2}	T_{1b2}
		..	..	..	..
..	**..**	..	..	..	..
..	**..**	..	..	..	..
MC_n	**SC_{n1}**	G_{n11}	Z_{n11}	F_{n11}	T_{n11}
		G_{n12}	Z_{n12}	F_{n12}	T_{n12}
		..	..	..	..
	SC_{n2}	G_{n21}	Z_{n21}	F_{n21}	T_{n21}
		G_{n22}	Z_{n22}	F_{n22}	T_{n22}
		..	..	..	..
	..	..	..	..	..
		G_{nz1}	Z_{nz1}	F_{nz1}	T_{nz1}
	SC_{nz}	G_{nz2}	Z_{nz2}	F_{nz2}	T_{nz2}
		..	..	..	..

In the Table 2;
n = number of main criteria,
z = number of sub-criteria belong to nth main criterion,
MC_n = nth main criterion,
SC_{nz} = zth sub-criterion of the nth main criterion
The number of sub-criteria may be different for each main criterion. For example, the sub-criterion number of main criterion MC_1 is a (SC_{11}, SC_{12}, …, SC_{1a}) and the sub criterion number of main criterion MC_2 is b ((SC_{21}, SC_{22}, …, SC_{2b})). In the Table 4, G, Z, F and T respectively represent strengths, weaknesses, opportunities and threats. x, y and z in the expressions G_{xyz}, Z_{xyz}, F_{xyz} and T_{xyz} respectively represent the main criterion, related sub-criterion and sequence number. For example, the expression F_{243} refers to the 3rd opportunity of the 4th sub-criterion of the 2nd main criterion.

3.2. The Fuzzy AHP Method

The Analytical Hierarchy Process (AHP) was developed by Thomas L. Saaty in 1977 and is one of the most widely used multi-criteria decision making techniques in the literature. Although the aim of the AHP is to benefit from expert knowledge, the method may not reflect the way people think exactly. Therefore, the Fuzzy AHP method which is an extended version of the AHP method, was developed to solve hierarchical fuzzy problems. Several fuzzy AHP methods such as Van Laarhoven and Pedrycz Fuzzy AHP Method, Buckley Fuzzy AHP Method, Chang's Extended Analysis Method, are available in the literature (Kahraman et al. 2003). In this study, the Chang's extent analysis method was preferred because it requires less computation, follows the steps of traditional AHP and does not require additional processing (Toksari and Toksari 2011). The method was described in detail below (Chang 1996);

Let $X = \{x_1, x_2, \dots\dots, x_n\}$ be a criterion set and $G = \{g_1, g_2, \dots\dots, g_m\}$ be a objective set. In this method, each criterion is taken and extent analysis is performed for each purpose. Thus, m extent analysis values are obtained for each criterion. These are shown as $M_{g_i}^1, M_{g_i}^2, \dots\dots, M_{g_i}^m \quad i = 1,2, \dots\dots n$ and all the $M_{g_i}^j \ (j = 1,2, \dots\dots m)$ values are triangular fuzzy numbers (TFNs).

Step 1: According to criterion i, fuzzy synthetic extent values (S_i) are determined using Eq. (1);

$$S_i = \sum_{j=1}^{m} M_{g_i}^j * \left[\sum_{i=1}^{n}\sum_{j=1}^{m} M_{g_i}^j\right]^{-1} \quad (1)$$

Fuzzy addition operation is performed on M values by using Eq. (2) to obtain $\sum_{j=1}^{m} M_{g_i}^j$ in the Eq. (1);

$$\sum_{j=1}^{m} M_{g_i}^{j} = \left(\sum_{j=1}^{m} l_j \, , \sum_{j=1}^{m} m_j \, , \sum_{j=1}^{m} u_j\right) \quad (2)$$

To obtain $\left[\sum_{i=1}^{n}\sum_{j=1}^{m} M_{g_i}^{j}\right]^{-1}$ in the Eq. (1), Eq. (3) and Eq. (4) are used;

$$\sum_{i=1}^{n}\sum_{j=1}^{m} M_{g_i}^{j} = \left(\sum_{i=1}^{n} l_i \, , \sum_{i=1}^{n} m_i \, , \sum_{i=1}^{n} u_i\right) \quad (3)$$

$$\left[\sum_{i=1}^{n}\sum_{j=1}^{m} M_{g_i}^{j}\right]^{-1} = \left(\frac{1}{\sum_{i=1}^{n} u_i}, \frac{1}{\sum_{i=1}^{n} m_i}, \frac{1}{\sum_{i=1}^{n} l_i}\right) \quad (4)$$

Step 2: The possibility degree for $M_2 = (l_2, m_2, u_2) \geq M_1 = (l_1, m_1, u_1)$ is expressed as $V(M_2 \geq M_1) = sup_{y \geq x}\left[\min\left(\mu_{M_1}(x), \mu_{M_2}(y)\right)\right]$. If this equation is analyzed, the Eq. (5) is obtained;

$$V(M_2 \geq M_1) = hgt\,(M_1 \cap M_2) = \mu_{M_2}(d) = \begin{cases} 1, & if\ m_2 \geq m_1\,, \\ 0, & if\ l_1 \geq u_2\,, \\ \frac{l_1 - u_2}{(m_2 - u_2) - (m_1 - l_1)}, & otherwise, \end{cases} \quad (5)$$

In the Eq. (5), d is the ordinate of the highest intersection point between μ_{M_1} and μ_{M_2}. In order to compare M_1 and M_2 values both $V(M_2 \geq M_1)$ and $V(M_1 \geq M_2)$ values must be known.

Step 3: The possibility degree of a convex number being greater than k convex fuzzy numbers ($M_i \ i = \{1, 2, \dots, k\}$) must also be considered.

$$V(M \geq M_1, \dots, M_k) = V\left[(M \geq M_1)\ and\ (M \geq M_2) and \dots and\ (M \geq M_k)\right] = \min V\,(M \geq M_1) \quad (6)$$

In the Eq. (6), if $d'(A_i) = \min V\,(S_i \geq S_k)$ for i = $\{1,2, \dots, k\}$, weight vectors for k ≠ i are calculated using Eq. (7) as the following;

$$W' = \left(d'(A_1), d'(A_2), \dots, d'(A_n)\right)^T \quad i = \{1,2, \dots, n\} \quad (7)$$

Step 4: The weight vectors are normalized using Eq. (8);

$$W = \left(d(A_1), d(A_2), \dots, d(A_n)\right)^T, i = \{1,2, \dots, n\} \quad (8)$$

In the Eq. (8), W weight vector isn't a fuzzy number. The final alternative weights are found by hierarchically synthesizing obtained these weights.

4. The Proposed Fuzzy Sectional SWOT Approach and Its Application

The proposed model for the determination of Turkey's focus strategies related to renewable energy is as in the Figure 1. As mentioned also in the previous section, there are some criticisms directed the traditional SWOT analysis. Some disadvantages of the traditional SWOT analysis are overcome with the approach called the Fuzzy Sectional SWOT proposed in this study. Mentioned shortcomings and how they were eliminated were mentioned below;

- **No using weights to reflect priorities;** This disadvantage of the traditional SWOT analysis was eliminated by prioritizing renewable energy sources with the Fuzzy AHP method.
- **Using ambiguous words and expressions;** This disadvantage of the traditional SWOT analysis was eliminated by using the fuzzy logic.
- **There is no obligation to verify thoughts by data or analysis;** This disadvantage of the traditional SWOT analysis was eliminated by using the Fuzzy AHP method.
- **It requires only one level of analysis;** This disadvantage of the traditional SWOT analysis was eliminated with the use of Sectional SWOT analysis.
- **There is no logical link to strategy implementation;** This disadvantage of the traditional SWOT analysis was eliminated by determining the most important energy sources and creating the strategies related to them.

Application steps of the proposed model are as follows;

Step 1: Determination of the expert group

At this phase, the experts who are consulted to their opinions about the renewable energy sources addressed, were determined. The expert group consists of engineers working in the renewable energy sector and academicians who are expert in their fields.

Step 2: Application of the Sectional SWOT analysis

At this phase, the renewable energy field was divided into 6 sub-sections (hyropower, solar, wind, biomass, hydrogen and geothermal) according to the logic of Sectional SWOT analysis. Then strengths, weaknesses, opportunities and threats for each of these sub-sections (renewable energy sources) were determined in 3 ways using the Sectional SWOT analysis;

- Literature review,
- Examining the ministry reports,
- Asking questions which are seen in the Table 3 and were prepared to make Sectional SWOT analysis more systematic, to the experts.

Strengths, weaknesses, opportunities and threats related to the renewable energy sources addressed in this study express the following;

- Strengths contain issues that when the current situation for Turkey of addressed renewable energy source is examined, it can gain an advantage over other renewable energy sources.
- Weaknesses contain issues that when the current situation for Turkey of addressed renewable energy source is examined, it is

weak and deficient compared to other renewable energy sources.

- Opportunuties are issues that may create positive results for the related energy source when the current situation for Turkey of addressed renewable energy source is examined.
- Threats are issues that may create undesired results for the related energy source when the current situation for Turkey of addressed renewable energy source is examined.

The strengths, weaknesses, opportunities and threats determined for each of the RES addressed in the study as a result of the literature research, examination of the ministry reports and the informations received from the experts, are as seen in Table 4-Table 9.

Step 3: Prioritization of the renewable energy sources with the Fuzzy AHP

At this stage, firstly weights of the renewable energy sources were found by the Fuzzy AHP method and then they were prioritized according to their weights. For this purpose, firstly the criteria experts used when they compared RES, shown below were determined;

- ✓ Accessibility to renewable energy source,
- ✓ The effect of renewable energy source on the environment,
- ✓ The development of renewable energy source technology,
- ✓ The contribution ability of renewable energy source to technology

Then, the pairwise comparison matrix shown in the Table 11 was obtained with that the experts compared the renewable energy sources according to the above criteria using values in the Table 10.

Table 10. Values for expert evaluations (Karatop et al. 2018)

Importance Degrees	Linguistic Expressions
1	Both factors are equally important
2	1st factor is less important than the 2nd factor.
3	1st factor being averagely important with respect to 2nd factor
4	1st factor is more important than the 2nd factor
5	1st factor is very important than the 2nd factor

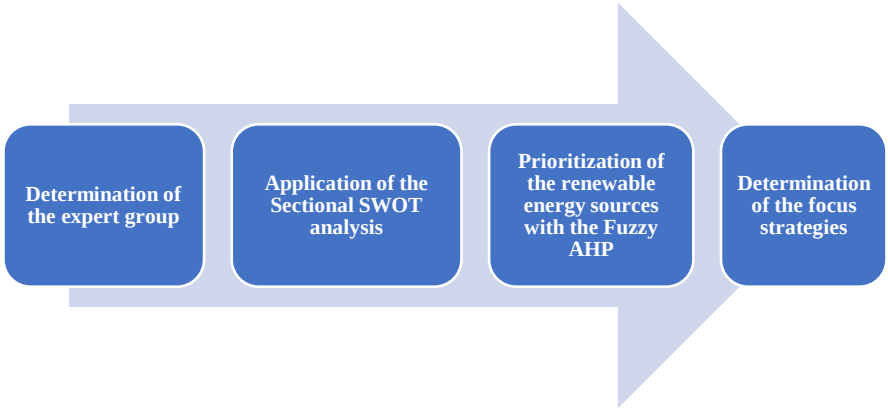


Figure 1. Application steps of the proposed model

Table 3. Questions related to renewable energy sources for the Sectional SWOT analysis

Sectional SWOT	Questions for the Sectional SWOT analysis
STRENGTHS	**1.** What are the advantages Turkey obtains in terms of its renewable energy sources? **2.** Which aspects of Turkey compared to other countries in terms of renewable energy sources are better / could be better? **3.** Do someones looking from the outside / stakeholders see powerful which aspects of the Turkey's renewable energy resources? **4.** In what aspects Turkey's renewable energy resources are in the leading position or have potential to can become leader? **5.** What are the advantages of Turkey's renewable energy resources / What can they happen?
WEAKNESSES	**1.** What are the disadvantages Turkey obtains in terms of its renewable energy sources? **2.** Which aspects of Turkey compared to other countries in terms of renewable energy sources are worse / are open to be improved? **3.** Do someones looking from the outside / stakeholders see weak which aspects of the Turkey's renewable energy resources? **4.** In what aspects Turkey's renewable energy resources aren't in the leading position? **5.** What are the weaknesses of Turkey's renewable energy resources / What can they happen?
OPPORTUNITIES	**1.** What are the opportunities standing in front of the Turkey's renewable energy sources? **2.** What sort of interesting developments take place in the environment related to the Turkey's renewable energy sources? **3.** Which developments in the technology are opportunity for the Turkey's renewable energy resources? **4.** Which opportunities similar practices taking place in other countries provide to the Turkey's renewable energy sources? **5.** Which sources in the environment are opportunity for the Turkey's renewable energy sources? **6.** What sort of opportunities wait the Turkey for renewable energy sources in the coming years?
THREATS	**1.** What are the threats standing in front of the Turkey's renewable energy sources? **2.** What sort of bad developments take place in the environment related to the Turkey's renewable energy sources? **3.** Which developments in the technology are threat for the Turkey's renewable energy resources? **4.** Which threats similar practices taking place in other countries provide to the Turkey's renewable energy sources? **5.** Which sources in the environment are threat for the Turkey's renewable energy sources? **6.** What sort of threats wait the Turkey for renewable energy sources in the coming years?

Table 4. The Sectional SWOT analysis for Hydropower at Turkey

	Strengths	Weaknesses	Opportunities	Threats
HYDROPOWER	Low operational cost (Heshmati et al. 2015;Online 2020; experts)	High initial investment cost (Heshmati et al. 2015, experts)	The Turkey is rich in terms of rivers and lakes (experts)	International problems may arise as a result of building power plants on streams which arise form Turkey and flow to other countries (Karali 2017; experts)
	Environmentally friendly (Heshmati et al. 2015; Online 2020; experts)	Causing displacement of local residents (Heshmati et al. 2015; experts)	Rough structure of the Turkey (experts)	-
	Ability to store energy (Heshmati et al. 2015; experts)	Damage to the environment during hydropower construction due to large engineering works (Heshmati et al.	Arising awareness about renewable energy (experts)	-

	2015; experts)			
	Suppling water for agriculture, household, and industrial use (Heshmati et al. 2015; experts)	-	-	-
	No fuel costs (Online 2020; experts)	-	-	-
	Highly productive (Online 2020; experts)	-	-	-
	Long-lived (Online 2020; experts)	-	-	-
	Reducing the occurrence of natural hazards such as soil erosion, flood etc. (Gedik 2015; Karali 2017; experts)	-	-	-
	Reducing external dependence and ensures supply security (Karali 2017; experts)	-	-	-
	Providing employment opportunities (experts)	-	-	-

Table 5. The Sectional SWOT analysis for Wind Energy at Turkey

	Strengths	Weaknesses	Opportunities	Threats
WIND ENERGY	Low maintenance and operational costs (Online 2020; experts)	High initial investment cost (Online 2020; experts)	Having adequate capacity for the plant installation of the Aegean and Marmara coasts (Karali 2017; experts)	Inadequate legal incentives in Turkey (Karali 2017; experts)
	Environmentally friendly (Online 2020; experts)	Low capacity factors (Online 2020; experts)	The constituent parts of the power plant are produced in Turkey (Karali 2017; experts)	May require outsourcing for power plant investment (Karali 2017; experts)
	Reaching of its cost to a level that can compete with today's power plants (Online 2020; experts)	Changing energy production (Online 2020; experts)	Avaliability of technological developments (Karali 2017; experts)	Inability to interfere with wind speed and it damage equipment at high speeds (Gedik 2015; experts)
	Installation and operating of its technology is relatively simple (Online 2020; experts)	Arising of additional cost from power transmission to residential areas (Ngô and Natowitz 2009; experts)	It is supported by European Union (EU) harmonization laws (Karali 2017; experts)	-
	Start-up of it can takes place in a short time (Online 2020; experts)	Intermittent structure (Ngô and Natowitz 2009; experts)	Arising awareness about renewable energy (experts)	-
	No risk of depletion and over time price increase (Online 2020; experts)	Causing bird deaths (Karali 2017; experts)	-	-
	Providing economic contributions related to renting, purchase etc. to the people in the power plant area (Karali 2017; experts)	-	-	-
	Providing employment opportunities (experts)	-	-	-

Table 6. The Sectional SWOT analysis for Solar Energy at Turkey

	Strengths	Weaknesses	Opportunities	Threats
SOLAR ENERGY	Ease of installation and use (Online 2020; experts)	Releasing of chemicals routinely and accidental (Tsoutsos et al. 2005; experts)	High solar energy potential of Turkey due to its geographical location (Online 2020; experts)	Due to climate changes the sunlight is not of good quality (Karali 2017; experts)
	Environmentally friendly (Online 2020; experts)	Use of land (Tsoutsos et al. 2005; experts)	Avaliability of appropriate lands for plant installation (Karali	Inadequate legal incentives (Karali 2017; experts)

			2017; experts)	
	No harmful waste (Online 2020; experts)	Impacts of large PV systems to ecosystems (Tsoutsos et al. 2005; experts)	With the rapid development of technology, it can be used more widely (Karali 2017; experts)	Interruption of the agriculture due to the use of agricultural land (Karali 2017; experts)
	Providing employment opportunities (experts)	Construction activities due to solar thermal energy (Tsoutsos et al. 2005; experts)	Arising awareness about renewable energy (experts)	-
	Low maintenance cost of power plants (experts)	Depending on weather conditions (Ngô and Natowitz 2009; experts)	-	-
	Possibleness to use it in every area where energy is needed (experts)	Availability just during daylight hours (Ngô and Natowitz 2009; experts)	-	-
	Increasing production power and efficiency with solar panels (experts)	Visual impact on buildings' aesthetics (Tsoutsos et al. 2005; experts)	-	-
	Sustainable (experts)	Intermittent structure (experts)	-	-
	-	Its investment cost is high (experts)	-	-

Table 7. The Sectional SWOT analysis for Geothermal Energy at Turkey

	Strengths	Weaknesses	Opportunities	Threats
GEOTHERMAL ENERGY	Environmentally friendly (Online 2020; experts)	High investment costs (Purkus and Barth 2011; experts)	Being rich in terms of geothermal due to geographical location (Online 2020; experts)	Inadequate legal incentives in Turkey (Karali 2017; experts)
	Cheap (Online 2020; experts)	Risk of insufficient heat (Purkus and Barth 2011; experts)	It is supported by EU harmonization laws (Karali 2017; experts)	May require outsourcing due to investment cost (Karali 2017; experts)
	Continuous structure (Heshmati et al. 2015; experts)	-	Arising awareness about renewable energy (experts)	-
	Independence from weather conditions (Fridleifsson and Freeston 1994; experts)	-	-	-
	Providing employment opportunities (experts)	-	-	-

Table 8. The Sectional SWOT analysis for Biomass Energy at Turkey

	Strengths	Weaknesses	Opportunities	Threats
BIOMASS ENERGY	Availability of it widely in all places (Tester et al. 2005; experts)	May environmentally hazardous (Evli 2018; experts)	Economic use in the countryside (Karali 2017; experts)	It is also raw material of other industries (Karali 2017; experts)
	Continuous structure (Gedik 2015; experts)	Increasing of food prices (Evli 2018; experts)	Arising awareness about renewable energy (experts)	Competition with agricultural products due to energy raw material (Karali 2017; experts)
	Easy storage (Gedik 2015; experts)	Low energy content compared to other non-renewable energy sources (experts)	-	-
	Providing employment opportunities (Gedik 2015; experts)	High transport and storage costs compared to other energy types (experts)	-	-
	Reducing dependence on fossil energy sources (Evli	-	-	-

	2018; experts)			
	Increasing of woodlands (Evli 2018; experts)	-	-	-
	Prevention of both damage to the environment and diseases stem from wastes (Evli 2018; experts)	-	-	-

Table 9. The Sectional SWOT analysis for Hydrogen Energy at Turkey

	Strengths	Weaknesses	Opportunities	Threats
HYDROGEN ENERGY	Environmentally friendly (Evli 2018; experts)	More expensive than other sources (Evli 2018; experts)	It is supported by EU harmonization laws (Karali 2017; experts)	Inadequate legal incentives in Turkey (Karali 2017; experts)
	Can be used for multiple purposes (Evli 2018; experts)	Environmentally harmful gases may ocur (Evli 2018; experts)	The presence of hydrogen stored chemically at the base of the Black Sea (Karali 2017; experts)	Potential users do not have sufficient information (Karali 2017; experts)
	More efficient than other fuels (Evli 2018; experts)	-	Arising awareness about renewable energy (experts)	Using the generated energy as an explosive weapon (Karali 2017; experts)
	Safer than other fuels (Evli 2018; experts)	-	-	-
	Providing employment opportunities (experts)	-	-	-

Table 11. Pairwise comparison matrix related to the renewable energy sources

	Hydropower	Wind	Solar	Geothermal	Biomass	Hydrogen
Hydropower	1	3	3	4	3	2
Wind	0.33	1	0.5	3	1	0.33
Solar	0.33	2	1	3	1	0.33
Geothermal	0.25	0.33	0.33	1	0.5	0.25
Biomass	0.33	1	1	2	1	0.25
Hydrogen	0.5	3	3	1	4	1

The obtained weight values and priority order determined according to the weights of RES as a result of the analyze are as in Table 12.

Table 12. Weight scores and priority order of the renewable energy sources

Renewable Energy Sources	Weight Scores	Priority Order
Hydropower	0.2710	1
Wind	0.2026	3
Solar	0.2186	2
Geothermal	0.0563	6
Biomass	0.1366	4
Hydrogen	0.1149	5
TOTAL	1	

Step 4: Determination of the focus strategies

At this stage, focus strategies have been created by using the strengths, weaknesses, opportunities and threats related to renewable energy sources determined by the Sectional SWOT analysis. The use of Sectional SWOT analysis enables to be created better focused strategies related to the renewable energy in Turkey. Strategies were generally structured on that using positive aspects (strengths and opportunities) to increase positive aspects and to turne negative aspects (weakneses and threats) to positive aspects or to neutralize. The focus strategies which were created related to the renewable energy for Turkey, are as follows;

- **Hydropower**

SO Strategy; The strategy created to further strengthen "Suppling water for agriculture, household, and industrial use", "Reducing the occurrence of natural disasters such as soil erosion, flood etc.", "Reducing external dependence and ensures supply security" and "Providing employment opportunities" strengths by taking advantage of "The Turkey is rich in terms of rivers and lakes" and "Rough structure of the Turkey" opportunities is as follows;

"Doing studies to can use whole of hydropower potential of the country"

WO Strategy; The strategy created to strengthen "Causing displacement of local residents" weakness by

taking advantage of "Arising awareness about renewable energy" opportunity is as follows;

"Doing activities to expand awareness even more related to renewable energy for the whole society and especially the society that is negatively affected by renewable energy investment".

- **Solar Energy**

SO Strategies;

Strategy 1: The strategy created to further strengthen "Providing employment opportunities" and "Possibleness to use it in every area where energy is needed" strengths by taking advantage of "High solar energy potential of Turkey due to its geographical location" and "Avaliability of appropriate lands for plant installation" opportunities is as follows;

"Giving support and incentive to increase the investments in solar energy"

Strategy 2: The strategy created to further strengthen "Possibleness to use it in every area where energy is needed" strength by taking advantage of "With the rapid development of technology, it can be used more widely" opportunity is as follows;

"Promoting R&D studies related to the solar energy"

ST Strategy; The strategy created to protect from "Due to climate changes the sunlight is not of good quality" and "Inadequate legal incentives" threat by using "Increasing production power and efficiency with solar panels" strength is as follows;

"Giving support and incentive for being used new technology in the solar power plants"

WO Strategies;

Strategy 1: The strategy created to strengthen "Depending on weather conditions", "Availability just during daylight hours" and "Intermittent structure" weaknesses by taking advantage of "High solar energy potential of Turkey due to its geographical location" opportunity is as follows;

"Doing studies to can use whole of the solar energy potential in the country"

Strategy 2: The strategy created to strengthen "Use of land" weakness by taking advantage of "Avaliability of appropriate lands for plant installation" opportunity is as follows;

"Allowing solar power plants to be installed only on suitable lands"

Strategy 3: The strategy created to strengthen "Depending on weather conditions", "Availability just during daylight hours" and "Intermittent structure" weaknesses by taking advantage of "With the rapid development of technology, it can be used more widely" opportunity is as follows;

"Supporting and encouraging the studies which increase production power and efficiency"

WT Strategies;

Strategy 1: The strategy created to protect "Depending on weather conditions" and "Intermittent structure" weaknesses from "Due to climate changes the sunlight is not of good quality" threat is as follows;

"Supporting and encouraging the use of technologies which increase production power and efficiency"

Strategy 2: The strategy created to protect "Its investment cost is high" weakness from "Inadequate legal incentives" threat is as follows;

"**Stressing the importance of solar energy, encouraging investors doing studies to produce the constituent parts of solar power plants cheaper and in the country**"

Strategy 3: The strategy created to protect "Use of land" weakness from "Interruption of the agriculture due to the use of agricultural land" threat is as follows;

"Using only suitable empty lands for installation of the solar energy plants"

- **Wind Energy**

SO Strategies;

Strategy 1: The strategy created to further strengthen "Providing employment opportunities" strength by taking advantage of "Having adequate capacity for the plant installation of the Aegean and Marmara coasts" opportunity is as follows;

"Building wind power plant in the areas with sufficient capacity"

Strategy 2: The strategy created to further strengthen "Start-up of it can takes place in a short time" strength by taking advantage of "The constituent parts of the power plant are produced in Turkey" opportunity is as follows;

"Using as much as possible domestic production parts while wind power plants are built"

Strategy 3: The strategy created to further strengthen "Providing employment opportunities" strength by taking advantage of "The constituent parts of the power plant are produced in Turkey" opportunity is as follows;

"Government's encouraging investors to produce the constituent parts of wind power plants"

Strategy 4: The strategy created to further strengthen "Reaching of its cost to a level that can

compete with today's power plants" strength by taking advantage of "Avaliability of technological developments" opportunity is as follows;

"Reducing the costs related to wind power plants by following the technological developments"

Strategy 5: The strategy created to further strengthen "Installation and operating of its technology is relatively simple" strength by taking advantage of "Avaliability of technological developments" opportunity is as follows;

"Implementing of the technological developments which facilitate the implementation of wind energy technology"

Strategy 6: The strategy created to further strengthen "Providing employment opportunities" strength by taking advantage of "It is supported by EU harmonization laws" opportunity is as follows;

"Promoting the establishment of wind power plants as part of the studies related to EU accession process"

ST Strategies;

Strategy 1: The strategy created to protect from "Inadequate legal incentives in Turkey" threat by using "Low maintenance and operational costs" strength is as follows;

"Providing consultancy service to support wind energy companies"

Strategy 2: The strategy created to protect from "Inadequate legal incentives in Turkey" and "May require outsourcing for power plant investment" threats by using "Reaching of its cost to a level that can compete with today's power plants" strength is as follows;

"Stressing the importance of wind energy, encouraging investors doing studies to produce the constituent parts of wind power plants cheaper and in the country"

WO Strategies;

Strategy 1: The strategy created to strengthen "Low capacity factors" weakness by taking advantage of "Having adequate capacity for the plant installation of the Aegean and Marmara coasts" opportunity is as follows;

"Establishment of wind power plants in the areas with adequate capacity"

Strategy 2: The strategy created to strengthen "High initial investment cost" weakness by taking advantage of "The constituent parts of the power plant are produced in Turkey" opportunity is as follows;

"Giving support or incentive for wind energy investments to accelerate R&D activities"

Strategy 3: The strategy created to strengthen "High initial investment cost" weakness by taking advantage of "It is supported by EU harmonization laws" opportunity is as follows;

"Giving support or incentive for R&D studies related to wind energy investments as part of the studies related to EU accession process"

WT Strategies;

Strategy 1: The strategy created to protect "High initial investment cost" weakness from "Inadequate legal incentives in Turkey" threat is as follows;

"Doing R&D studies of the government on wind energy investments"

Strategy 2: The strategy created to protect "High initial investment cost" weakness from "May require outsourcing for power plant investment" threat is as follows;

"Giving support or incentive for R&D studies related to wind energy investments"

- **Biomass Energy**

SO Strategy; The strategy created to further strengthen "Providing employment opportunities", "Reducing dependence on fossil energy sources", "Increasing of woodlands", "Prevention of both damage to the environment and diseases stem from wastes" strengths by taking advantage of "Economic use in the countryside" and "Arising awareness about renewable energy" opportunities is as follows;

"Giving support and incentive to investors to they invest in biomass energy"

ST Strategy; The strategy created to protect from "It is also raw material of other industries" and "Competition with agricultural products due to energy raw material" threats by using "Availability of it widely in all places" strength is as follows;

"**Ensuring to benefit from everything which can be raw materials for the production of biomass energy"**

WT Strategy; The strategy created to protect "Increasing of food prices" weakness from "It is also raw material of other industries" and "Competition with agricultural products due to energy raw material" threats is as follows;

"Separate production of common raw materials used by biomass energy and other sectors"

- **Hydrogen Energy**

SO Strategies;

Strategy 1: The strategy created to further strengthen "Providing employment opportunities"

strength by taking advantage of "The presence of hydrogen stored chemically at the base of the Black Sea" opportunity is as follows;

"Forming public opinion for processing of the hydrogen energy reserves in country"

Strategy 2: The strategy created to further strengthen "Providing employment opportunities" strength by taking advantage of "It is supported by EU harmonization laws" opportunity is as follows;

"Processing of hydrogen energy reserves in the country as part of the studies related to EU accession process"

WO Strategy; The strategy created to strengthen "More expensive than other sources" weakness by taking advantage of "The presence of hydrogen stored chemically at the base of the Black Sea" opportunity is as follows;

"Reducing costs by using existing hydrogen energy reserves in the country"

WT Strategy; The strategy created to protect "More expensive than other sources" weakness from "Inadequate legal incentives in Turkey" threat is as follows;

"Doing studies to reduce costs related to hydrogen energy"

- **Geothermal Energy**

SO Strategies;

Strategy 1: The strategy created to further strengthen "Providing employment opportunities" strength by taking advantage of "Being rich in terms of geothermal due to geographical location" opportunity is as follows;

"Using geothermal energy reserves in the country as much as possible"

Strategy 2: The strategy created to further strengthen "Providing employment opportunities" strength by taking advantage of "It is supported by EU harmonization laws" opportunity is as follows;

"Processing of geothermal energy reserves in the country as part of the studies related to EU accession process"

WO Strategy; The strategy created to strengthen "High investment costs" weakness by taking advantage of "It is supported by EU harmonization laws" opportunity is as follows;

"Doing studies to reduce investment costs by using supports and incentives given to geothermal energy which are a part of the studies related to EU accession process"

WT Strategy; The strategy created to protect "High investment costs" weakness from "Inadequate legal incentives in Turkey" threat is as follows;

"Doing studies to reduce costs related to geothermal energy"

5. Conclusions

Renewable energy has become a very popular type of energy all over the world due to its features such as providing security of energy supply, reducting the CO_2 emission and positive economic effects. Countries attach great importance to renewable energy and make their energy plans according to it. Based on the current importance of the subject, in the study it focused on determining of the Turkey's focus strategies related to renewable energy. This study contributes to existing literature by making the right investment decisions by creating focus strategies for renewable energy investments using the Fuzzy Sectional SWOT. For the aforementioned purpose, an integrated approach called the Fuzzy Sectional SWOT which consists of the Fuzzy AHP and the Sectional SWOT methods, is used. Some disadvantages of the traditional SWOT analysis are eliminated with the method used. Firstly, the renewable energy field was divided into 6 sub-sections (hyropower, solar, wind, biomass, hydrogen and geothermal) according to the logic of Sectional SWOT analysis. Then strengths, weaknesses, opportunities and threats for each of these sub-sections (renewable energy sources) were determined using the Sectional SWOT analysis. Weights were found with the Fuzzy AHP method for each of the RES and resources were prioritized according to these weights. Finally, focus strategies related to renewable energy field for Turkey were obtained with the creation of strategies related to renewable energy sources.

In the study, the order of importance of renewable energy resources from most important to less important was found as follows; hydropower, solar, wind, biomass, hydrogen, geothermal. Therefore, among the focus strategies obtained, the strategies to be taken into consideration at first are ones related to hydropower, solar and wind energies. However, this does not mean that the focus strategies related to biomass, hydrogen and geothermal energies can be neglected. For this reason, the focus strategies which should be primarily addressed related to renewable energy can be summarized as follows;

- Doing studies to use whole of the renewable energy potential in the country,
- Increasing social awareness related to the renewable energy,
- Providing necessary supports and incentives of the government,
- Selection of suitable areas for the installation of renewable energy plants,

- Producing the parts used in the installation of renewable energy plants, in the country.

The limitation of the study is the small number of experts who contributed to the study and therefore subjective judgment may predominate. Therefore, it would be good to increase the number of experts in future studies.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Asif, M., Muneer, T., 2007. Energy supply, its demand and security issues for developed and emerging economies. Renewable & Sustainable Energy Reviews, 11(7), 1388–1413.

Bas, E., 2013. The integrated framework for analysis of electricity supply chain using an integrated SWOT-Fuzzy TOPSIS methodology combined with AHP: The case of Turkey. International Journal of Electrical Power & Energy Systems, 44(1), 897-907.

Chang, D.Y., 1996. Applications of the extent analysis method on Fuzzy AHP. European Journal of Operational Research, 95(3), 649–655.

Eltrop, L., 2013. Renewable energy: resources and technologies. In Jenssen T (ed) Glances at Renewable and Sustainable Energy-Principles, Approaches and Methodologies for an Ambiguous Benchmark, London, England: Springer-Verlag, pp. 15-32.

Erdin, C., Ozkaya, G., 2019. Turkey's 2023 energy strategies and investment opportunities for renewable energy sources: Site selection based on ELECTRE. Sustainability, 11(7), 2136.

Ervural, B.C., Zaim, S., Demirel, O. F., Aydin, Z., Delen, D., 2018. An ANP and Fuzzy TOPSIS-based SWOT analysis for Turkey's energy planning. Renewable and Sustainable Energy Reviews, 82, 1538-1550.

Evli, S., 2018. Türkiye'de sürdürülebilir kalkinma ve yenilenebilir enerji kaynaklari (In Turkish). Dissertation, Tekirdağ Namık Kemal University, Turkey.

Fridleifsson, I.B., Freeston, D.H., 1994. Geothermal energy research and development. Geothermics, 23(2), 175–214.

Gedik-Torunoglu, O., 2015. Türkiye'de yenilenebilir enerji kaynaklari ve çevresel etkileri. Dissertation, Istanbul Technical University, Turkey.

Gottfried, O., De Clercq, D., Blair, E., Weng, X., Wang, C., 2018. SWOT-AHP-TOWS analysis of private investment behavior in the Chinese biogas sector. Journal of Cleaner Production, 184, 632-647.

Heshmati, A., Abolhosseini, S., Altmann, J., 2015. The development of renewable energy sources and its significance for the environment. Springer, Singapore.

Hill, T., Westbrook, R., 1997. SWOT analysis: It's time for a product recall. Long Range Planning, 30(1), 46-52.

International Energy Agency – IEA, 2021. Accessed: Feb. 12, 2021. [Online]. Available: https://www.iea.org/.

Kahraman, C., Cebeci, U., Ulukan, Z., 2003. Multi-criteria supplier selection using Fuzzy AHP. Logistics Information Management, 16(6), 382-394.

Karali, S., 2017. Yenilenebilir enerji kaynaklarinin Türkiye ve dünya ekonomisine katkisi (In Turkish). Dissertation, Bahçeşehir University, Turkey.

Karatop, B., 2015. Yerli otomotiv yatırımında odak strateji karar modeli: Bulanık AHP uygulaması (In Turkish). Istanbul, Turkey: Doğu Kütüphanesi.

Karatop, B., Kubat, C., Uygun, O., 2018. Determining the strategies on turkish automotive sector using Fuzzy AHP based on the SWOT analysis. Sakarya University Journal of Science, 22(5), 1314-1325.

Kaya, I., Colak, M., Terzi, F., 2019. A comprehensive review of fuzzy multi criteria decision making methodologies for energy policy making. Energy Strategy Reviews, 24, 207-228.

Learned, E.P., Christensen, C.R., Andrews, K.E., Guth, W.D., 1965. Business policy: Text and cases. Homewood, IL, USA: Irwin.

Lund, H., 2010. Renewable energy systems: The choice and modeling of 100% renewable solutions. Academic Press, Burlington, MA, USA.

Ngô, C., Natowitz, J.B., 2009. Our energy future: Resources, alternatives, and the environment. John Wiley & Sons, Hoboken, New Jersey, USA.

Okello, C., Pindozzi, S., Faugno, S., Boccia, L., 2014. Appraising bioenergy alternatives in Uganda using Strengths, Weaknesses, Opportunities and Threats (SWOT)-Analytical Hierarchy Process (AHP) and a desirability functions approach. Energies, 7(3), 1171-1192.

Online, 2020. Republic of Turkey Ministry of Energy and Natural Resources. Accessed: Jan. 13, 2020. Avaliable: https://www.enerji.gov.tr.

Papapostolou, A., Karakosta, C., Apostolidis, G., Doukas, H., 2020. An AHP-SWOT-Fuzzy TOPSIS approach for achieving a Cross-Border RES cooperation. Sustainability, 12(7), 2886.

Purkus, A., Barth, V., 2011. Geothermal power production in future electricity markets—A scenario analysis for Germany. Energy Policy, 39(1), 349–357.

Sevkli, M., Oztekin, A., Uysal, O., Torlak, G., Turkyilmaz, A., Delen, D., 2012. Development of a Fuzzy ANP based SWOT analysis for the airline industry in Turkey. Expert Systems with Applications, 39, 14-24.

Solangi, Y.A., Tan, Q., Mirjat, N.H., Ali, S., 2019. Evaluating the strategies for sustainable energy planning in Pakistan: An integrated SWOT-AHP and Fuzzy-TOPSIS approach. Journal of Cleaner Production, 236, 117655.

Tasri, A., Susilawati, A., 2014. Selection among renewable energy alternatives based on a Fuzzy Analytic Hierarchy

Process in Indonesia. Sustainable Energy Technologies and Assessments, 7, 34-44.

Tester, J.W., Drake, E.M., Driscoll, M.J., Golay, M.W., Peters, W.A., 2005. Sustainable energy: Choosing among options. The MIT Press, Cambridge, MA, England.

Toksari, M., Toksari, M.D., 2011. Bulanık Analitik Hiyerarşi Prosesi (AHP) yaklaşımı kullanılarak hedef pazarın belirlenmesi (In Turkish). ODTÜ Gelişme Dergisi, 38, 51-70.

Tsoutsos, T., Frantzeskaki, N., Gekas, V., 2005. Environmental impacts from the solar energy technologies. Energy Policy, 33(3), 289–296.

Wang, Y., Xu, L., Solangi, Y.A., 2020. Strategic renewable energy resources selection for Pakistan: Based on SWOT-Fuzzy AHP approach. Sustainable Cities and Society, 52, 101861.

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(1) (2022) 57-63
DOI: 10.38016/jista.993853

Makine Öğrenme Teknikleri Kullanılarak Kükürt Giderme İşleminde Kullanılan Malzeme Miktarının Tahmini

Esra Özcan[1], Emrullah Sonuç[2]*,

[1,2] Karabük Üniversitesi, Bilgisayar Mühendisliği Bölümü, Karabük, Türkiye

e.esracetinkaya@gmail.com, esonuc@karabuk.edu.tr

Öz

Bazik oksijen fırınlı bir demir çelik fabrikasının yüksek fırın tesislerinde kok, sinter ve diğer demir cevheri malzemelerinin kullanılmasıyla sıvı ham demir üretilmektedir. Bu üretimden sonraki adım çelik üretim süreci olup, hemen öncesinde sıvı ham demir, içerisindeki kükürt oranının belirli bir miktar düşürülmesi amacıyla kükürt giderme tesisinde işlem görmektedir. Desülfürizasyon olarak adlandırılan bu işlemin amacı bazı kükürt giderici reaktifler ilave edilerek hedef kükürt değerini yakalamaktır. İlave edilecek malzeme miktarlarını belirlemek için farklı yöntemler kullanılmaktadır. Genel olarak temel ve veriye dayalı modellerin uygulandığı çalışmalar görülmektedir. Ancak yapay zekâ tekniklerinin bu alandaki kullanımı oldukça kısıtlıdır. Bu çalışmada kükürt giderme işlemindeki malzeme (magnezyum, kireç, florit) oranları makine öğrenme teknikleri ile tahmin edilmiştir. Problem bir regresyon problemi olup altı farklı yöntem (Lineer Regresyon, K-En Yakın Komşu, Karar Ağaçları, Rastgele Orman, XGBoost, Yapay Sinir Ağları) veri seti üzerinde test edilmiştir. Çalışmada kullanılan veriler 2020 yılına ait olup kükürt giderme tesisinden alınmıştır. Verilerin %80'i eğitim, %20'si test için kullanılacak şekilde ayrılmıştır. Değerlendirme ölçütü olarak Doğruluk ve Ortalama Mutlak Hata Yüzdesi (Mean Absolute Percentage Error, MAPE) kullanılmıştır. Elde edilen sonuçlara göre Yapay Sinir Ağı modeli magnezyum, kireç ve florit için sırasıyla %85, %95,4 ve %80,14 doğruluk değerlerine ulaşmıştır. MAPE değerleri ise sırasıyla 14,99, 4,59 ve 19,85 olup bu da modelin başarılı bir tahmin gerçekleştirdiğini ortaya koymaktadır.

Anahtar kelimeler: Makine Öğrenmesi, Yapay Sinir Ağları, Regresyon, Kükürt Giderme.

Prediction of Amount of Material used in Desulfurization Process using Machine Learning Techniques

Abstract

Liquid raw iron is produced by using coke, sinter and other iron ore materials in the blast furnace facilities of an iron and steel factory with a basic oxygen furnace. The next step after the production, is the steelmaking process, and just before that, the liquid raw iron is processed in the desulfurization plant in order to reduce the available sulfur content by a certain amount. The purpose of this process, called desulphurization, is to achieve the target sulfur value by adding some desulfurization reagents. Different methods are used to determine the amount of material to be added. There are many studies in which basic and data-based models are applied in this desulphurization process. However, the use of artificial intelligence techniques in this area is quite limited. In this study, the material (magnesium, lime, fluorite) ratios in the desulfurization process were predicted by machine learning techniques. This problem is a regression problem and six different methods (Linear Regression, K-Nearest Neighbor, Decision Trees, Random Forest, XGBoost, Artificial Neural Networks) are tested on the dataset. The data used in the study belonged to the year 2020 and were taken from the desulfurization plant. 80% of the data is used for training and 20% for testing. Accuracy and Mean Absolute Percentage Error (MAPE) were used as evaluation criteria. According to the results obtained, Artificial Neural Network model obtained 85%, 95.4% and 80.14% accuracy for magnesium, lime and fluorite, respectively. The MAPE values are 14.99, 4.59 and 19.85, respectively, which shows that the model makes a successful prediction.

Keywords: Machine Learning, Artificial Neural Networks, Regression, Desulfurization.

* Sorumlu yazar.
E-posta adresi: esonuc@karabuk.edu.tr

Alındı : 13 Eylül 2021
Revizyon : 4 Kasım 2021
Kabul : 6 Aralık 2021

1. Giriş (Introduction)

Çelik sektöründeki farklı üretim yöntemleri ülkelerin gelişmişlik seviyeleri ve ekonomik durumuna göre farklılık göstermektedir (Ersöz vd., 2016). Bunlar arasında en yaygın olan yöntem ise demir cevherinden sağlanan sıvı ham demir kullanımıdır. Çelik üretimi iki farklı yolla yapılmaktadır. Bunlardan birincisi yüksek fırınlarda oksijen üflemesi yöntemi, diğeri ise elektrik ark ocaklarında hurdanın eritilmesi yöntemidir (Yıldız, 2017).

Bazik oksijen fırınlı entegre bir tesiste kömür ve cevher ham maddeleri ile üretim sürecinin ilk aşamasına başlanmaktadır. Kok üretimi için kok bataryalarında kömür işlem görmektedir. Cevher ise sinter tesisinde sinterleme için kullanılmaktadır. Elde edilen kok ve sinter ile diğer demir cevherleri, yüksek fırınlarda bir araya gelerek, sıvı ham demir üretimi gerçekleştirilmektedir. Sıvı ham demirin sıvı çeliğe dönüşümü ise hurda, kireç ve oksijen kullanımıyla bazik oksijen fırınlarında gerçekleşmektedir. Bu aşamadan sonra ise sıvı çeliğin kalitesini artırmak amacıyla pota fırınlarında alaşım malzemeler eklenir. Döküme hazır hale gelerek döküm tesislerinde slab, kütük vb. formlarına getirilir (Türkoğlu ve Özyıldırım, 2017). Bu yarı mamuller haddeleme işlemiyle de son ürün haline gelmektedir.

Entegre tesislerde çelik üretiminin en önemli girdisi sıvı ham demirdir (Özcan ve Köprü, 2020). Yüksek fırınlarda üretimi yapılan sıvı ham demir, çelik üretimine geçmeden önce desülfürizasyon işlemine tabii tutulmaktadır (Çetin, 2016). Amaç; sıvı ham demir içerisindeki kükürt, fosfor ve silis gibi değerlerin istenilen seviyeye getirilmesidir.

Bu maddelerden kükürdün seviyesi oldukça önemlidir; çünkü kükürdün çeliğin mekanik özellikleri üzerinde olumsuz etkileri mevcuttur. Ayrıca darbe mukavemetini azaltan zararlı bir element olup malzemenin sertleşebilirliğini ve kaynaklanabilirliğini de etkileyebilmektedir. Bu nedenle çelik içinde miktarının sınırlı olması beklenmektedir (Yıldız, 2017).

Kükürt oranını azaltmak için kireç, magnezyum, florit ve karpit gibi kükürt giderici reaktif malzemeler kullanılmaktadır. Bu malzemelerin kullanımı ve miktarları istenilen kükürt değerine ulaşmak için bir modele ya da yönteme göre kurgulanmalıdır (Visuri vd., 2020).

Literatürde yer alan bir çalışmada torpido arabasındaki sıcak metalde kükürt giderme işlemi için yapay sinir ağı kullanılmıştır (Deo vd., 1994). Modelde sıcak metal ağırlığı, ilk kükürt içeriği, işlem süresi reaktif enjeksiyon hızı gibi değişkenler kullanılarak bir tahmin işlemi gerçekleştirilmiştir. Başka bir çalışmada geri yayılımlı bir yapay sinir ağ modeli geliştirilerek kükürt giderme işlem parametrelerini tahmin etme hedeflenmiştir (Rong vd., 2005). Diğerinde ise geri yayılımlı bir sinir ağı ile sıcak metal ağırlığı, sıcaklık değeri, sıcak metaldeki başlangıç ile son kükürt içeriği gibi kriterler kullanılarak magnezyum tozunu tahmin etmek için bir model kurgulanmıştır (Zhan vd., 2010). Magnezyum bazlı ve kireç bazlı enjeksiyonların karşılaştırması için kükürt gidermede kinetik bir model önerilmiştir (Jin vd., 2006). Bir diğer yapılan çalışmada ise proses sonuçları regresyon yaklaşımı kullanılarak tahmin edilmiştir (Vino vd., 2007). Tahmin için sıcaklık faktörü baz alınarak regresyon modelleri oluşturulmuştur. Başka bir çalışmada kullanılan geri yayılımlı sinir ağında moment teriminin eklenmesi önerilmiş ve bu şekilde kükürt giderici ekleme miktarı tahmin modeli kurgulanmıştır (Liang vd., 2011).

İlgili literatür incelendiğinde kükürt giderici malzemelerin miktarına yönelik tahmin çalışmalarının kısıtlı olduğu görülmektedir. Son yıllarda makine öğrenme tekniklerinin gelişimi ve ülkemizdeki sanayi tesislerindeki üretim artışı bu çalışmanın motivasyon kaynağı olmuştur. Çalışmada, kükürt giderme işleminde kullanılacak malzemelerin miktarını tespit etmek için farklı regresyon modelleri test edilmiştir. Bu modellerden elde edilen sonuçlar kıyaslanarak literatüre bir katkı sağlanması amaçlanmıştır. Makalenin geriye kalan kısımları şu şekilde organize edilmiştir: İkinci kısımda kullanılan makine öğrenme tekniklerine kısaca değinilmiş ve veri seti tanıtılmıştır. Üçüncü bölümde kullanılan yöntemlerden elde edilen sonuçlar tablolarla sunulmuş ve yorumlanmıştır. Son bölümde ise çalışmadan elde edilen kazanımlara değinilmiştir.

2. Materyal ve Yöntem (Material and Method)

Çalışmada makine öğrenme teknikleri kullanılmıştır. Bu teknikleri uygulamak için yararlanılan veri seti, giriş ve çıkış parametreleri belirlenerek kullanıma uygun hale getirilmiştir. Tahmin edilecek veri sayısal bir değerdir. Bu şekilde sayısal bir değeri tahmin etmeyi içeren problemler regresyon olarak tanımlanmaktadır. Regresyon, bağımsız değişkenlerin bağımlı değişkendeki değişimini Eşitlik (1)'deki gibi ifade etmektedir.

$$Y = a + bX_1 + cX_2 + dX_3 + \varepsilon \quad (1)$$

Tahmine dayalı modelleme, yani veriler üzerinde tahmin yapmak için geçmiş verilerden yararlanmayı gerektirir.

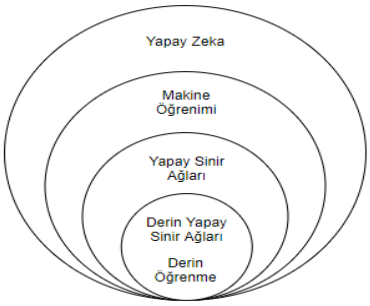
Regresyon problemleri, makine öğrenme mimarileri ile çözülebilmektedir. Bu kapsamda aşağıdaki yöntemler kullanılarak farklı modeller kurgulanmış ve tahmin işlemi gerçekleştirilmiştir:

- Lineer Regresyon,
- K-En Yakın Komşu,
- Karar Ağaçları,
- Rastgele Orman,
- XGBoost,
- Yapay Sinir Ağı (YSA).

Makine öğrenmesi, yapay zekanın bir alt dalı olup ilgili probleme ait verilerden faydalanarak bir modelin kurgulandığı teknikleri içermektedir. Şekil 1'de yapay zekâ ve alt dalları gösterilmektedir. Makine öğrenmesi

algoritmaları verileri kullanarak birtakım kurallar türetir ve bu sayede öğrenme işlemi gerçekleştirerek problemi çözme başarısının artırılmasını hedefler (Alan, 2020).

Makine öğrenmesinin temeli Alan Turing'in 1950'li yıllarda yaptığı çalışmalarla atılmıştır. Alan Turing'in bir makinenin zekaya sahip olabileceğini araştırması araştırmacıların ilgisi çekmiştir. Arthur Samuel'in hazırladığı bir dama programı ile 1959 yılında makine öğrenmesi kavram olarak ele alınmıştır (Görgün, 2020).



Şekil 1. Makine Öğrenme Metodolojisi (Machine learning methodology)

Makine öğrenme teknikleri günümüzde birçok alanda yaygın olarak kullanılmaktadır. Bunlardan başlıcaları; savunma sanayii, yüz tanıma, nesne ve ses tanıma, otonom sistemler, medikal ve tıp alanında hastalık teşhisi, bankacılıkta sahtelik tespitidir (Elen ve Avuclu, 2020, Aswad ve Sonuç, 2020, Dolapci ve Özcan, 2021). Makine öğreniminde kullanılacak modeller veriye göre değişiklik göstermektedir. Yine verinin yoğunluğunun yani veri setinin genişlik ve varyasyonlara sahip olma durumunun model başarısındaki rolü büyüktür (Baydilli, 2021).

2.1. Veri Seti (Data Set)

Çalışmada entegre bir demir çelik tesisinin kükürt giderme prosesinden elde edilen veriler kullanılmıştır.

Veri dosyası analiz edilerek tahmin işlemi için gerekli kriterler tespit edilmiştir. Bu kapsamda, 7 kriterin baz alındığı tamamı numerik değer içeren 4214 kayıttan oluşan veri seti elde edilmiştir. Tahmin edilecek değişkenler Tablo 1'deki gibi belirlenmiştir.

Tablo 1. Tahmin Edilecek Değişkenler (Predictable Variables)

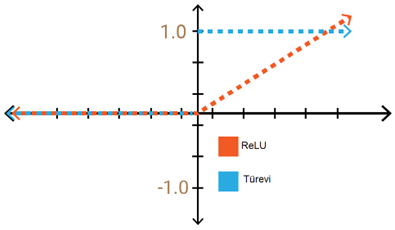
Çıkış Değişkenleri
Magnezyum (Mg)
Kireç (CaO)
Florit (CaF_2)

Üç malzeme için ayrı ayrı tahmin modelleri oluşturularak; sıvı ham demir tonaj miktarı, hedeflenen kükürt değeri, mevcut kükürt değeri ve mevcut silisyum değerleri sabit girdi değişkenleri seçilmiştir. Magnezyum miktarı tahmininde bu kriterler kullanılmış olup Kireç miktarı tahmininde girdiye Magnezyum; Florit miktarı tahmininde ise girdiye Magnezyum ve Kireç değişkenleri de ilave edilmiştir.

2.2. Yapay Sinir Ağları (Artificial Neural Network)

YSA makine öğrenimi yöntemlerinden biri olup insan beyninde bulunan nöronların görevinden esinlenerek geliştirilmiştir. Beyinde gerçekleşen öğrenme fonksiyonu simüle edilerek sınıflandırma, tahmin, kümeleme gibi problemleri çözmek için kullanılmaktadır (Karaatlı, 2012).

Sinir ağlarında doğrusal olmayan özellikleri tanıtmak amacıyla aktivasyon fonksiyonları kullanılmaktadır. Çalışmada, hız açısından avantajlı olması sebebiyle Şekil 2'de yer alan ReLU fonksiyonu kullanılmıştır. ReLU; 0 ile $+\infty$ arasında değer üreten bir aktivasyon fonksiyonudur.



Şekil 2. ReLU fonksiyonu (ReLU function)

2.3. Doğrusal Regresyon (Linear Regression)

Basit doğrusal regresyon, iki sürekli (nicel) değişken arasındaki ilişkileri özetlememize ve incelememize izin veren Eşitlik (2)'deki gibi istatistiksel bir yöntemdir. Değişkenlerden biri bağımsız değişken, diğeri bağımlı değişken olarak kabul edilir. Bağımlı değişkenin sürekli olması gerekirken, bağımsız değişkenler ya sürekli ya da kategorik özelliğe sahip değişkenler olabilir (Gök, 2017).

$$Yi = \beta 0 + \beta 1 \mathrm{X} i + \varepsilon i \qquad (2)$$

2.4. Karar Ağacı Regresyonu (Decision Tree Regressor)

Karar ağacı, regresyon veya sınıflandırma modellerini kurgularken ağaç yapısından faydalanır. İşlenen veri kümesi alt kümelere parçalanırken buna paralel olarak Eşitlik (3)'te verilen karar ağacı modeli geliştirilir. Sonuçta düğüm ve yapraklardan oluşan bir model elde edilir (Başer vd., 2021).

$$S(T,X) = \sum_{c \in X} P(c) S(c) \qquad (3)$$

2.5. K-En Yakın Komşu Regresyonu (Kneighbors Regressor)

Eşitlik (4)'te ifade edilen K-en yakın komşu yöntemi, eldeki tüm durumları hafızasında saklar ve mesafe fonksiyonu vb. benzerlik ölçüsü kullanarak tahmin işlemini gerçekleştirir. Bu yöntem, ilk olarak 1970'lerin başında örüntü tanıma ve istatistiksel tahminlemede kullanılmıştır (Salman ve Sonuç, 2021).

$$\hat{y} = \text{f(x)} = \frac{1}{k}\sum_{j=1}^{k} y_{i_j} \quad (4)$$

2.6. Rastgele Orman Regresyonu (Random Forest Regressor)

Karar ağaçlarının birden fazla kullanılmasıyla oluşmuş bir modeldir. Karar Ağacı Regresyonunun bazı durumlarda aşırı öğrenmeye neden olması beraberinde bir dezavantajı getirmektedir (Pekel, 2020). Bu sorunu ortadan kaldırmak için, Karar Ağacı Regresyonu yerine Rastgele Orman Regresyonu uygulanarak aşırı öğrenme sınırlandırılabilir (Görgün, 2020).

$$F(x) = \frac{1}{j}\sum_{j=1}^{j} c_{jfull} + \sum_{k=1}^{K}(\frac{1}{j}\sum_{j=1}^{j} contribution(x,k)) \quad (5)$$

2.7. Xgboost (Extreme Gradient Boosting)

Gradyan artırma algoritması; regresyon ve sınıflandırma problemlerinde kullanılan makine öğrenmesi yöntemidir. Veri setine farklı ağırlıklar verilmesi ile elde edilen ağaçlar topluluğundan tahminler yapılması ve zayıf tahminlerle modeli güçlendirmek temel amacıdır. Bu algoritmanın paralelleştirilmiş ve optimize edilmiş bir versiyonu ise Xgboost'dur. Veriler üzerinde mümkün olan en iyi modeli eğitmek yerine, eğitim veri kümesinin çeşitli alt kümelerindeki binlerce modeli eğitir ve ardından en iyi performans gösteren modelin çıktısı alınır (Dilwani, 2019). Verilerin parçalara ayrılarak analiz işlemine tabi tutulmasıyla gerçekleştirilmektedir. Buradaki amaç daha küçük parçalar kullanılarak daha iyi bir tahmin elde edilmesi ve başarının artırılmasıdır (Yangın, 2019).

$$\sum_{i=1}^{n} L(y_i, p_i) + \frac{1}{2}\alpha O^2{}_v \quad (6)$$

2.8. Model Performans Metrikleri (Model Performance Metrics)

Regresyon modellerinde, öğrenme işlemini başarılı kılmak için hata fonksiyonlarından çıkan değerin minimum olması hedeflenmektedir. Model için en uygun optimizasyon fonksiyonunu seçerek bu hedefi gerçekleştirmek mümkündür (Henrique vd., 2019).

Modellerin performansı ise hata fonksiyonları ile ölçülmektedir. Regresyon problemleri için kullanılan hata fonksiyonları Ortalama mutlak hata (MAE), Hata kareler ortalaması (MSE) ve Ortalama mutlak hata yüzdesi (MAPE) şeklinde sıralanabilir. Tahmin sonucunun gerçek değerden ne kadar uzak olduğu bu metrikler ile analiz edilebilmektedir. Uygulanan metotlarda değerlendirme ölçütü olarak Eşitlik (7)'de ifade edilen MAPE tercih edilmiştir. Eşitlikte *G* gerçek değer, *T* ise tahmin değeri ifade etmektedir. MAPE değerinin %10'un altında çıkması yüksek doğruluğa sahip modeli, %10 ile %20 arasında olması ise doğru tahmin modelini temsil etmektedir (Karabıçak vd., 2018).

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|(G_i - T_i)|}{|G_i|}}{n} \quad (7)$$

Modellerin başarısını ölçmek için kullanılan bir diğer metrik de doğruluktur. Eşitlik (8)'deki gibi modelde doğru tahmin edilen alanların toplam veri kümesine oranı ile doğruluğu hesaplamak mümkündür.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Burada TP (True Positive, Doğru Pozitif), TN (True Negative, Doğru Negatif), FP (False Positive, Yanlış Pozitif) ve FN (False Negative, Yanlış Negatif)'i ifade etmektedir.

2.9. Uygulama (Application)

Çalışma için *python* programlama dili ve model kurgusu için desteklediği kütüphaneler ile *spyder* programından yararlanılmıştır.

2.9.1. Python (Python)

Python okunabilirliği kolay, modüler ve yorumlanabilir bir script dildir. Son yıllarda bilimsel hesaplama yöntemlerinin başında gelen yapay zekâ alanında yaygın şekilde kullanılmaktadır. Sonuçları kolaylıkla okuma, analiz etme ve görselleştirme imkânı sayesinde araştırmacılar tarafından tercih edilmektedir. Platform bağımsız olması en büyük avantajlarından bir tanesidir (Ürün, 2019). Çalışmada *python 3.7.4* sürümü kullanılmıştır.

2.9.2. Spyder (Spyder)

Python için pek çok editör bulunmakta olup çalışmada *Spyder* programı tercih edilmiştir. Ücretsiz ve açık kaynak bir yazılımdır. *Spyder* programının değişken görüntüleme ekranından değişkenlerin isim, tür, boyut ve aldıkları değerlere ulaşmak mümkündür (Alkan, 2019).

2.9.3. Keras (Keras)

Python pek çok kütüphane desteği vermektedir. Bunlardan, makine öğrenme modellerini tanımlamak ve eğitmek amacıyla *Keras* kütüphanesi çalışmada kullanılmıştır.

2.9.4. Tkinter (Tkinter)

Ayrıca ara yüz geliştirmeye yarayan, form uygulamaları için olanak sağlayan *Tkinter* kütüphanesi de uygulamaya dahil edilerek tekrar tekrar kullanılabilmeye imkân sağlanmıştır.

3. Deneysel Sonuçlar (Experimental Results)

Tesiste kullanılan kükürt giderici üç reaktif malzeme olan Magnezyum, Kireç ve Florit için modelin tahmin yetenekleri test edilmiştir. Bu değişkenlerin tahminine yönelik her bir değişken için modeller oluşturulmuştur. Dolayısıyla model çıkışı tek bir değişkenden oluşmaktadır.

Kükürt giderme işlemi için tesiste kullanılan en önemli malzeme Magnezyum'dur (Özmen vd., 2018). Bundan dolayı çalışmada ilk hedef Magnezyum miktarının tahmin edilmesidir. Magnezyum tahmin işlemi gerçekleştirildikten sonra Kireç tahmini gerçekleştirilmiş olup bu tahminde Magnezyum değerlerinden de yararlanılmıştır. Son olarak Florit tahmin işlemi gerçekleştirilmiş olup burada da Magnezyum ve Kireç değerleri kullanılarak tahmin işlemi gerçekleştirilmiştir.

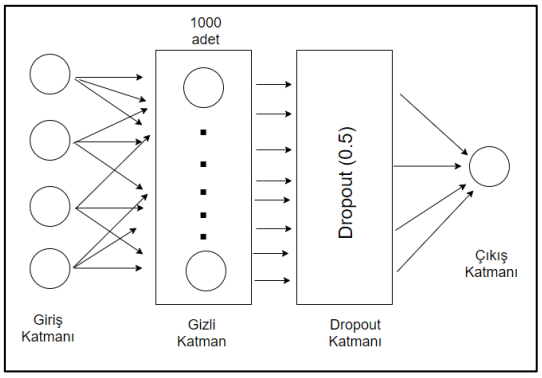
Bu kurgu dahilinde Linear Regression, Kneighbors Regressor, Decision Tree Regressor, Random Forest Regressor ve XGBoost metodları uygulanmıştır. Bu metodların algoritmaları *python* kütüphanelerinde tanımlı olup; LinearRegression, KNeighborsRegressor, DecisionTreeRegressor, RandomForestRegressor ve XGBRegressor fonksiyonları uygulamaya dahil edilmiştir. RandomForestRegressor metodunda n_estimators parametresi 100 olarak seçilmiştir. XGBRegressor için aşağıdaki şekilde parametreler belirlenmiştir. Diğer metodlar varsayılan hali ile kullanılmıştır.

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
    colsample_bynode=1, colsample_bytree=1, gamma=0,
    importance_type='gain', learning_rate=0.1, max_delta_step=0,
    max_depth=3, min_child_weight=1, n_estimators=100,
    n_jobs=1, nthread=None, random_state=0,
    reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
    silent=None, subsample=1, verbosity=1)
```

Şekil 3. XGBoost Modeli için Kullanılan Parametreler (Parameters Used for XGBoost Model)

Makine öğrenimi modellerinin başarısı, içerdiği parametrelere göre değişkenlik göstermektedir. Bu amaçla model farklı katman/nöron sayıları, maliyet fonksiyonları vb. parametreler kullanılarak test edilmiştir. Kullanılan farklı parametreler arasından en iyi performansa sahip mimari seçilerek ağın eğitimi tamamlanmıştır.

Geliştirilen YSA modeli Şekil 4'te gösterilmiştir (Özcan, 2021). Model sonuçlarının kıyaslanması için MAPE ve Doğruluk olmak üzere iki farklı metrik kullanılmıştır. Her bir çıkış değeri için sonuçlar Tablo 2, Tablo 3 ve Tablo 4'te gösterilmiştir.



Şekil 4. Önerilen YSA Modeli (Proposed ANN Model)

Tablo 2'ye göre MAPE ve doğruluk değerleri baz alındığında; Magnezyum için en iyi sonuçlar YSA modeli tarafından elde edilmiştir. Lineer Regresyon YSA'dan sonra başarılı olan ikinci yöntem olarak görülmektedir. Ayrıca K-En Yakın Komşu yöntemi de Lineer Regresyon'a yakın bir sonuç elde etmiştir. Bu üç yöntemin diğerlerine göre daha başarılı bir tahmin yaptığı görülmektedir.

Tablo 2. Mg Değişkeni için Metodların Sonuçları (Results of Methods for Mg Variable)

Yöntem	MAPE	Doğruluk
Lineer Regresyon	17,15	82,85
K-En Yakın Komşu	18,28	81,72
Karar Ağaçları	46,74	53,26
Rastgele Orman	44,70	55,29
XGBoost	44,37	55,63
YSA	14,99	85,00

Tablo 3'e göre Kireç için en iyi değerler Magnezyum da olduğu gibi yine YSA modeli tarafından elde edilmiştir. Yine Lineer Regresyon YSA'dan sonra ikinci sırada yer almakta ve onu K-En Yakın Komşu yöntemi izlemektedir. Diğer yöntemlerin elde ettikleri değerler ise kıyaslanacak seviyede değildir.

Tablo 3. CaO Değişkeni için Metodların Sonuçları (Results of Methods for CaO Variable)

Yöntem	MAPE	Doğruluk
Lineer Regresyon	6,03	93,96
K-En Yakın Komşu	7,81	92,19
Karar Ağaçları	48,53	51,47
Rastgele Orman	47,82	52,18
XGBoost	47,61	52,39
YSA	4,59	95,41

Tablo 4'e göre Florit için en iyi değerler diğer çıkış değişkenlerinde olduğu gibi YSA modeli tarafından elde edilmiştir. Lineer Regresyon ve K-En Yakın Komşu YSA'ya yakın değerleri elde eden iki yöntem olarak görülmektedir. Diğer yöntemlerin başarısı ise rekabet edecek seviyede değildir.

Tablo 4. CaF_2 Değişkeni için Metodların Sonuçları (Results of Methods for CaF_2 Variable)

Yöntem	MAPE	Doğruluk
Lineer Regresyon	25,05	74,94
K-En Yakın Komşu	25,48	74,51
Karar Ağaçları	65,86	34,14
Rastgele Orman	63,24	36,76
XGBoost	62,18	37,82
YSA	19,85	80,14

Tüm değerler ışığında tahmini gerçekleştirilen üç farklı malzeme için YSA modelinin diğer yöntemlere göre başarılı olduğu tespit edilmiştir. YSA Magnezyum için %85, Kireç için %95,4 ve Florit için %80,14 doğruluk değerlerini elde etmiştir. Yine bu değişkenler için YSA sırasıyla 14,99, 4,59 ve 19,85 MAPE değerlerine ulaşmıştır. Florit ve Kireç tahmininde

modele Magnezyum'un değişken olarak eklenmesi, bu malzemelerin miktarının belirlenmesinde Magnezyum miktarının da etkili olmasından kaynaklanmaktadır. Magnezyumun modele eklenmeden ve eklendikten sonra elde edilen sonuçlar karşılaştırmalı olarak Tablo 5'te verilmiştir.

Tablo 5. Giriş Değişkeni Olarak Mg'nin Model Sonuçlarına Etkisi (Effect of the Results of the Model for Mg as an Input Variable)

	Cao (Kireç)				CaF_2 (Florit)			
	Mg'siz Model		Mg'li Model		Mg'siz Model		Mg'li Model	
Yöntem	MAPE	Doğruluk	MAPE	Doğruluk	MAPE	Doğruluk	MAPE	Doğruluk
Lineer Regresyon	18,347	81,65	6,03	93,96	33,866	66,13	25,05	74,94
K-En Yakın Komşu	19,631	80,37	7,81	92,19	36,841	63,16	25,48	74,51
Karar Ağaçları	48,172	51,83	48,53	51,47	65,845	34,15	65,86	34,14
Rastgele Orman	45,543	54,46	47,82	52,18	59,809	40,19	63,24	36,76
XGBoost	45,075	54,92	47,61	52,39	59,492	40,51	62,18	37,82
YSA	16,507	83,49	4,59	95,41	36,005	63,99	19,85	80,14

4. Sonuçlar ve Tartışma (Conclusions and Discussions)

Çelik malzemelerin sanayinin gelişimiyle ülkemizde ve tüm dünyada birçok alanda kullanımı bu maddenin önemini artırmıştır. Sıvı ham demir kullanılarak elde edilen çeliğin kalitesi, bu aşamada gerçekleştirilen kükürt giderme işlemiyle yakından ilişkilidir. Kükürt giderme işleminde kullanılan başlıca malzemelerin en başında magnezyum, kireç ve florit gelmektedir.

Bu çalışmada, üretilen sıvı ham demirde kükürt giderme işlemi için yukarıda bahsedilen malzemelerin miktarının tahmini için farklı regresyon yöntemleri (Lineer Regresyon, K-En Yakın Komşu, Karar Ağaçları, Rastgele Orman, XGBoost, Yapay Sinir Ağı) test edilmiştir. Bu yöntemlerden elde edilen sonuçlar MAPE ve doğruluk metrikleri kullanılarak değerlendirilmiş ve YSA yöntemi diğer yöntemlere kıyasla başarılı olmuştur. MAPE değerleri Kireç için %5'in altında, Magnezyum ve Florit için ise %20'nin altında sonuçlanmıştır. Bu değerin %20'den daha düşük bir değerde olması, geliştirilen modeli doğru tahmin modeli sınıfına dahil etmektedir.

Veri setindeki verilerin sınırlı sayıda olması makine öğrenme modellerinin başarısını da sınırlı hale getirebilmektedir. Bu yüzden ileriki çalışmalarda veri artırma yöntemi kullanılarak makine öğrenme modelleri test edilerek başarı oranının daha da artırılabilmesi ve hata değerlerinin minimum hale getirilmesi mümkün olabilir.

Kaynaklar (References)

Alan, A., 2020. Makine Öğrenmesi Sınıflandırma Yöntemlerinde Performans Metrikleri ile Test Tekniklerinin Farklı Veri Setleri Üzerinde Değerlendirilmesi, Yüksek Lisans Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ.

Aswad, S.A., Sonuç, E., 2020. Classification of VPN Network Traffic Flow Using Time Related Features on Apache Spark. In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-8). IEEE.

Alkan, O., 2019. Parkinson Hastalığının Teşhisinde Derin Öğrenme Yöntemi ile Spect Görüntü Analizi, Yüksek Lisans Tezi, Ağrı İbrahim Çeçen Üniversitesi Fen Bilimleri Enstitüsü, Ağrı.

Başer, B.Ö., Yangın, M., Sarıdaş, E.S., 2021. Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25(1), 112-120.

Baydilli, Y.Y., 2021. Polen Taşıyan Bal Arılarının MobileNetV2 Mimarisi ile Sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi, (21), pp.527-533.

Çetin, Z., 2016. Sürekli Döküm Prosesinde Pota Nozulu Tıkanma Probleminin Analizi ve Azaltılması, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Deo, B., Datta, A., Haressh, M., Kalra, P.K., Boom, R., 1994. Adaptive Neural Net (ANN) Models for Desulfurization of Hot Metal and Steel. Steel Research International, 65(11), 466-471.

Dilwani, A.A.R., 2019. Makine Öğrenmesi ile Acil Triyaj ve Hastane Yatış Tahmini, Yüksek Lisans Tezi,Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Dolapci, B., Özcan, C., 2021. Automatic Ship Detection and Classification using Machine Learning from Remote Sensing Images on Apache Spark. Journal of Intelligent Systems: Theory and Applications, 4(2), pp.94-102.

Elen, A., Avuclu, E., 2020. A Comparison of Classification Methods for Diagnosis of Parkinson's. International Journal of Intelligent Systems and Applications in Engineering, 8(4), 164-170.

Ersöz, F., Ersöz, T., Erkmen, İ.N., 2016. Dünyada ve Türkiye'de Ham Çelik Üretimine Bakış. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen bilimleri Dergisi, 32(2), 1-12.

Gök, B., 2017. Makine Öğrenmesi Yöntemleri ile Akademik Başarının Tahmin Edilmesi. Gazi Üniversitesi Fen Bilimleri Dergisi, 5(3), 139-148.

Görgün, M., 2020. Makine Öğrenmesi Yöntemleri ile Kalp Hastalığının Teşhis Edilmesi, Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstanbul.

Henrique, B.M., Sobreiro, V.A., Kimura, H., 2018. Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of finance and data science, 4(3), pp.183-201.

Jin, Y., bi, X.G., Yu, S.R., 2006. Kinetic Model for Powder Injection Desulfurization. Acta Metallurgica Sinica, 19(4), 258-264.

Karaatlı, M., 2012. Yapay Sinir Ağları ile Otomobil Satış Tahmini. Uluslararası Yönetim İktisat ve İşletme Dergisi, 8(17), 87-100.

Karabıçak, Ç., Avcı, S., Akman, G., Aladağ, Z., 2018. Determination of Demand Estimation Methods by Values and Variability Measures for Stock Items in a Cleaning Paper Company. Journal of Current Researches on Engineering Science and Technology, 4(1), 47-68.

Liang, X.P, Ban, S.X., Wang, Y., Ge, W.S., Huang, Z.H., 2011. Research and Application of Desulfurizer Addition Amount Prediction Model Based on Neural Network. Metallurgical Industry.

Özcan C., Köprü, E.Y., 2020. Yapay Sinir Ağları ile Sıvı Ham Demir Tahmini ve 5. Yüksek Fırın Uygulaması. Avrupa Bilim ve Teknoloji Dergisi, Özel Sayı, 155-162.

Özcan, E., 2021. Kükürt Giderme İşlemi İçin Kullanılan Malzeme Miktarının Makine Öğrenme Yöntemleri İle Tahmini, Yüksek Lisans Tezi, Karabük Üniversitesi Lisansüstü Eğitim Enstitüsü, Karabük.

Özmen, K., Eskiyurt, T.G., Şahin, H., Erkal, H., Kocabaş, T., Çakır, M., Soysal Atan, B., 2018. Sıvı Ham Demir Kükürt Giderme Prosesinde Kullanılan Granüle Magnezyum Tüketiminin Seviye-2 Yazılımı ile Optimizasyonu. 19. Metalurji ve Malzeme Kongresi (IMMC 2018).

Pekel, E., 2020. Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology, 139(3), pp.1111-1119.

Rong, Z., Dan, B., Yi, J., 2005. A BP Neural Network Predictor Model for Desulfurization Molten Iron. International Conference on Advanced Data Mining and Applications, 22-24 July 2005, pp. 728-735.

Salman K., Sonuç, E., 2021. Thyroid Disease Classification Using Machine Learning Algorithms. In Journal of Physics: Conference Series (Vol. 1963, No. 1, p. 012140). IOP Publishing.

Türkoğlu, S., Özyıldırım, B. M., 2017. Developing Oxygen Amount Preediction Model of Basic Oxygen Furnace Steelmaking Process with Machine Learning Algortihms. Çukurova Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 39(12), 22-31.

Ürün, S., 2019. Python ile Programlamanın Temelleri.

Vinoo, D.S., Mazumdar, D., Gupta, S.S., 2007. Optimization and Prediction Model of Hot Metal Desulphurisation. Ironmaking & Steelmaking, 34(6), 471-476.

Visuri, V.V., Vuolio, T., Haas, T., Fabritius, T., 2020. A Review of Modeling Hot Metal Desulfurization. Steel Research International, 91(4), 1900454.

Yangın, G., 2019. Xgboost ve Karar Ağacı Tabanlı Algoritmaların Diyabet Veri Setleri Üzerine Uygulaması, Yüksek Lisans Tezi, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Ensitüsü, İstanbul.

Yıldız, K., 2017. Demir Çelik Metalurjisi. Sakarya.

Zhan, D.P., Zhang, H, Jiang, Z., 2010. Prediction Model of Magnesium Powder Consumption During Hot Metal Pre-Desulfurization. China Metallurgy.

Research Article
Journal of Intelligent Systems: Theory and Applications 5(1) 2022: 64-74
DOI: 10.38016/jista.951008

The Optimization of Routes Using Evolutionary Algorithms in Public Transportation Systems

Salih Serkan Kaleli[1]*, Mehmet Bayğan[2], Abdullah Naralan[3]

[1] Ardahan University, Department of Office Management and Secretary, Ardahan, Turkey

[2] Ardahan University, Department of Computer Engineering, Ardahan, Turkey

[3] Recep Tayyip Erdoğan University, Department of Research Information System, Rize, Turkey

salihserkankaleli@ardahan.edu.tr, mehmetbaygin@ardahan.edu.tr, abdullah.naralan@erdogan.edu.tr

Abstract
This study aims to examine, regulate, and update the land transportation of the Erzurum Metropolitan Municipality (EMM), Turkey using computerized calculation techniques. In line with these targets, some critical information has been obtained for study: the number of buses, the number of expeditions, the number of bus lines, and the number and maps of existing routes belonging to EMM. By using the information that has been obtained, this study aims at outlining specific outputs according to the input parameters, such as determining the optimal routes, the average travel, and the journey time. Once all of these situations were considered, various optimization algorithms were used to get the targeted outputs in response to the determined input parameters. In addition, the study found that the problem involved in modeling the land transport network of the EMM is in line with the so-called "traveling salesman problem," which is a scenario about optimization often discussed in the literature. This study tried to solve this problem by using the genetic algorithm, the clonal selection algorithm, and the DNA computing algorithm. The location data for each bus stops on the bus lines selected for the study were obtained from the EMM, and the distances between these coordinates were obtained by using Google Maps via a Google API. These distances were stored in a distance matrix file and used as input parameters in the application and then were put through optimization algorithms developed initially on the MATLAB platform. The study's results show that the algorithms developed for the proposed approaches work efficiently and that the distances for the selected bus lines can be shortened.

Keywords: Optimization, Evolutionary Algorithms, Traveling Salesman Problem, Genetic Algorithm, Management Information Systems

Toplu Taşıma Sistemlerinin Evrimsel Algoritmalarla Optimizasyonu

Özet

Bu çalışma, Erzurum Büyükşehir Belediyesi'nin (EBB) Türkiye kara ulaşımını bilgisayarlı hesaplama teknikleri kullanarak incelemeyi, düzenlemeyi ve güncellemeyi amaçlamaktadır. Bu hedefler doğrultusunda, çalışma için bazı önemli bilgiler: otobüs sayısı, sefer sayısı, otobüs hattı sayısı ve EBB'ye ait mevcut güzergâh sayısı ve haritaları elde edilmiştir. Bu çalışma, elde edilen bilgileri kullanarak, optimal rotaların belirlenmesi, ortalama seyahat ve yolculuk süresi gibi girdi parametrelerine göre belirli çıktıların ana hatlarını çizmeyi amaçlamaktadır. Tüm bu durumlar göz önüne alındığında, belirlenen girdi parametrelerine karşılık hedeflenen çıktıları elde etmek için çeşitli optimizasyon algoritmaları kullanılmıştır. Çalışma, EBB' nin ulaşım ağının modellenmesindeki problemin, literatürde sıklıkla tartışılan optimizasyonla ilgili bir senaryo olan "gezgin satıcı problemi" ile uyumlu olduğunu bulmuştur. Çalışmada genetik algoritma, klonal seçim algoritması ve DNA hesaplama algoritması kullanılarak bu problem çözülmeye çalışılmıştır. Çalışmada seçilen otobüs hatlarındaki her bir durak için konum bilgisi EBB'den alınmış ve bu koordinatlar arasındaki mesafeler bir Google API üzerinden Google Maps kullanılarak elde edilmiştir. Bu mesafeler bir mesafe matrisi dosyasında saklanmış ve uygulamada giriş parametreleri olarak kullanılmış daha sonra MATLAB platformunda geliştirilen optimizasyon algoritmalarına aktarılmıştır. Çalışmanın sonuçları, önerilen yaklaşımlar için geliştirilen algoritmaların verimli çalıştığını ve seçilen otobüs hatları için mesafelerin kısaltılabileceğini göstermektedir.

Anahtar kelimeler: Optimizasyon, Evrimsel Algoritmalar, Gezgin Satıcı Problemi, Genetik Algoritma, Yönetim Bilişim Sistemleri

* Corresponding Author.
E-mail: salihserkankaleli@ardahan.edu.tr

Recieved : 11 June 2021
Revision : 29 Nov. 2021
Accepted : 17 Dec. 2021

1. Introduction

The rapid and recent developments in computer and internet technology have enabled many processes to be done fully automatically and quickly. In addition, in light of these developments, although the problems encountered in daily life can occur rapidly, they can also be resolved much more easily using this technology. In parallel with this situation, it was inevitable that optimization processes would be used for constantly encountered problems in daily life. The optimization concept refers to a process developed every day and aims to facilitate the processes involved in everyday life. These can be used by human beings to use time more efficiently and solve problems in their work. In addition, by developing this concept every day and moving it to a higher level, the number of problems can be solved can be increased (Deng et al. 2017). At this point, the developments in computer technology and the continuous increases made in terms of processor power have downsized a large amount of computing time, which is a frequently encountered problem when utilizing optimization processes. Using this increase in processor power, problems that require a vast solution space can be calculated in a much shorter time and in a much more accurate way.

Optimization processes are encountered various fields, such as engineering, design, financial planning, holiday planning, computer science, and industrial computing (Sundararaghavan et al. 2010). Human beings have always aimed to maximize or minimize a purpose. For example, the effort at minimizing current expenses while trying to maximize the profit of an enterprise is something wanted by many human beings. For this reason, humankind has made it a goal to choose the most suitable solution for almost every problem. The aim of achieving the best price-performance balance when purchasing goods and when running or performing more than one business in parallel with each other over a specific period forms the basis of having an optimization process.

There are various optimization methods studied in the literature. These methods are generally based on numerical calculations, and the problems with a vast and possible solution area can be solved in a concise amount of time. Thanks to the development of these methods, completely tailored solutions can be produced and specific targets for success can be achieved. Examples of the best known of these methods include the genetic algorithm, the ant colony algorithm, particle swarm optimization, the differential development algorithm, the clonal selection algorithm, and the DNA computing algorithm. Although there are various algorithms in the literature, there are some basic steps involved in optimization algorithms. A block diagram showing these steps is given in Figure 1.

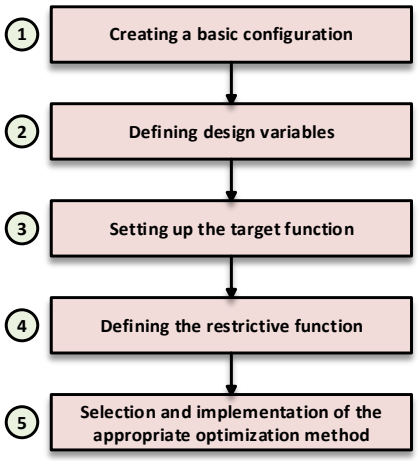


Figure 1. Steps to Follow for Optimization Problem

As shown Figure 1, the first step in an optimization problem is to create the basic configuration. After this step, it is necessary to determine the design variables and define them for the problem. In the third step toward the solution, the objective function should be determined and be suitable for the problem. After this process, the restrictive statements in the problem should be defined in the application, and the problem should be designed according to these constraints. In the last stage of the system, a suitable optimization method should be selected and then implemented. Optimization algorithms are used in many different fields (Guo et al. 2007). One of these studies sought to solve the traveling salesman problem by using the genetic algorithm. In the study, crossover and mutation steps were used and the mutation process was carried out in two stages in terms of scrolling and adding. After the crossover process was complete, the individuals were checked and an attempt was made to determine if any deterioration to their condition had occurred. The approach proposed within the scope of the study has been tested in two different cities, 40 and 100. It has been observed that the proposed method contributes significantly to solving the GSP (Chen 2013)-(Bolat and Cortés 2011) proposed an optimization approach that was based on the Genetic Algorithm and the Tabu Search algorithm for a group of elevator systems. In this study, the floor numbers and cabin numbers were grouped and tests were carried out in different combinations of these groups. In the proposed approach, it was observed that the average wait time and the average travel times of the passengers were lowered, and it was also found that the Genetic Algorithm yielded better results than the Tabu search algorithm. Another study on this subject, by (Groba et al. 2015), used the combination of a prediction technique and the genetic algorithm, which is an intuitive method, for the solution to the dynamic traveling salesman problem. In this study, route optimization was done for a scenario in which the targets were constantly moving. For this purpose, a genetic algorithm was created that feeds in Newton's motion equation. The study results proved that the trajectory approach using the prediction-

based genetic algorithm provides better and more effective results than the methods commonly used, as noted in the literature. A flow diagram summarizing the trajectory predictions based on the approach that used the genetic algorithm method developed within the scope of the study is shown in Figure 2.

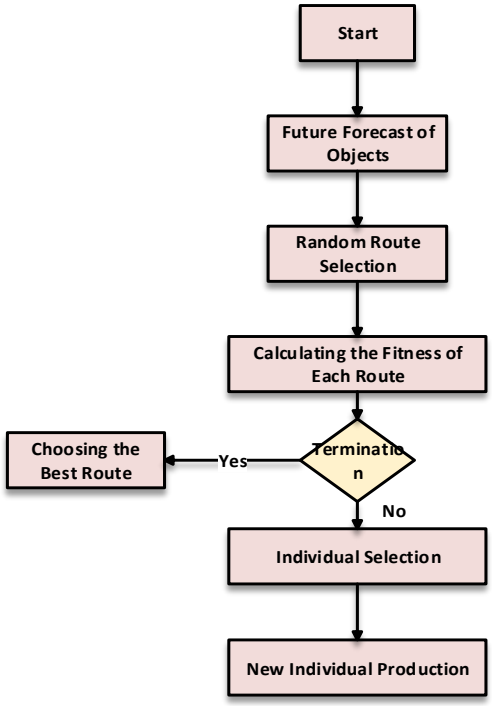


Figure 2. Genetic Algorithm Application for Dynamic TSP

In a study conducted by (Baygin and Karakose 2013), the genetic algorithm, the clonal selection algorithm, and the DNA computing algorithm were used, and the timing of a group of elevator systems was checked. For this purpose, an elevator system with 20 floors and five cabins was simulated and examined using the three different optimization methods. The study aimed to decrease the average wait time of the passengers, decrease the average amount of their travel time, and decrease the amount of energy that was consumed by the cabins. In another study that took up the traveling salesman problem, a hybrid approach was proposed and was based on the particle flock algorithm, the ant colony algorithm, and the 3-Opt algorithm. In this study, the particle flock algorithm was used to optimize the parameters affecting the performance of the ant colony algorithm. In addition, the 3-Opt intuitive method, the other method used in the study, was used to develop localized solutions. The experimental studies conducted in this research showed that the proposed approach yielded better results than many others found in the literature in terms of the quality and accuracy of the solution (Mahi et al. 2015). A block diagram that summarizes this study is shown in Figure 3.

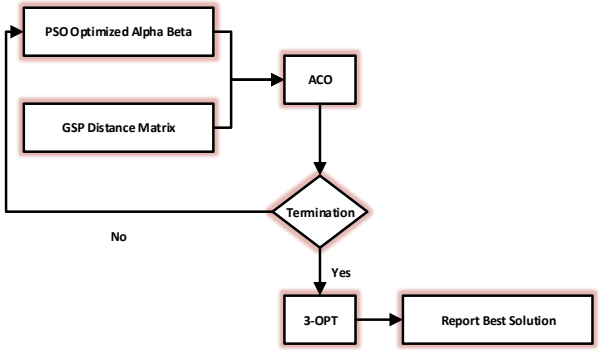


Figure 3. Ant Colony and 3-Opt Algorithm Based Optimization Process for TSP

This study, which was conducted in Turkey, aims at using optimization algorithms to optimize actively used bus lines in the city of Erzurum. In this context, certain lines have been optimized, and the purpose of the study was to find out how to provide a faster, better, and more effective service for bus passengers. First, six different line routes actively used by the municipality were selected and the GPS data for each of the stops on these lines were obtained. The GPS data were sent to Google Maps via a Java program using the Google API, and the distance matrix files were drawn from Google Maps. This distance was used to calculate the conformity function in the matrix file for the optimization algorithms. After these processes, the clonal selection algorithm, the genetic algorithms, and the DNA computing algorithms were developed from an artificial immune system in a MATLAB environment. These previously selected routes were subjected to optimization. After all of these processes, new routes were configured for each of the line routes and these were compared with the currently used routes. As a result of the simulations carried out, significant improvements were achieved on the six different routes. The distance for each of the routes currently in use has been significantly shortened.. In addition, the results have been marked on Google Maps and the new routes are visualized.

The traveling salesman problem is examined in the second part of this study. In the third section, the algorithms used for the optimization process are provided. The results of the simulations carried out are provided in the fourth part of the study and these results are given comparatively. In the fifth and last part of the study, further results and suggestions are included.

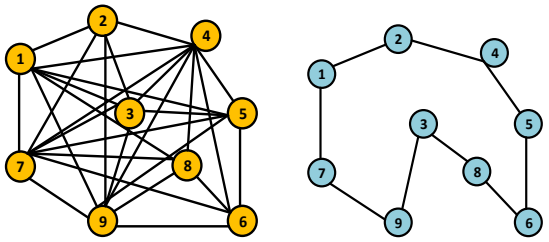
2. The Traveling Salesman Problem

The traveling salesman problem (TSP) is a scenario in which the purpose of a salesman is to sell his goods in an “n” different city. While performing this process, the traveller should visit the city only once and not visit the previously visited city. Determining the minimum length of the road route within the scope of these objectives and constraints is the primary purpose for solving the traveler salesman problem(Asadpour et al. 2010). A simple

version of the traveling salesman problem is summarized below(NARALAN et al. 2017);

- A traveler wants to sell his or her goods in "n" different cities. The salesperson wants to visit the cities as soon as possible and wants to visit each city alone. The problem aims to offer the seller the shortest route that has the least cost.
- In the first city, the seller has the right to choose between "n-1" different city roads.
- In the second city, the seller has the right to choose between "n-2" different city roads. In this way, the cities that the vendor has visited are subtracted from the total of the city "n," and the process is continued again in the same way. When the problem is considered in general, the total "(n-1)!" are the many situations or possibilities.

Consequently, if the seller is assumed to be going to these cities in the opposite direction, "(n-1)! / 2" has to be chosen from the different routes taken at each step. In Figure 4, an exemplary version of the traveling salesman problem is given (NARALAN et al. 2017).

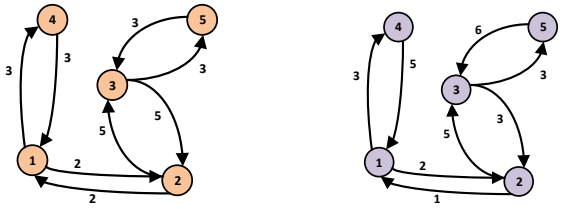


(a) (b)

Figure 4. An Example Traveling Salesman Problem and Solution

As shown in Figure 4, the numbers represent the cities and the lines represent the routes between the cities. For example, a seller has to choose between 29! / 2 different tours for a tour of the 30 cities. From this point of view, the travelling salesman problem is the type of problem that belongs in the NP-Tam class(Kovács et al. 2018). The possibilities of the routes to be followed by the salesman, who wants to go from the number 1 city to the number 9 city as soon as possible, is shown in Figure 4. If the total number of cities to be visited by the salesman is shown with "n," and the round-trip distances for these "n" cities that he visited may be the same or different (Wang et al. 2016). The main reasons for the different round-trip distances between the cities can be listed, including situations such as having to take a compulsory direction in traffic, the existence of traffic jams, and the compatibility of the round-trip lanes to be used. In other words, the cost calculated to go from "A" to "B" and the calculated cost to reach "A" from "B" may not always be the same. The literature defines this situation as the symmetrical / asymmetrical traveling salesman problem (Hasan Söyler 2007).

The symmetrical traveling salesman problem involves having a distance between the two cities equal to the round-trip length. Depending on the situation, the costs to be spent for the distances during the journey are also equal. The opposite of the symmetric traveling salesman problem is the asymmetrical traveling salesman problem(Nguyen et al. 2002). There are some cases in which symmetrical problems are not always valid. Especially in big cities, traffic jams, one-way roads, bad weather, bad road conditions, etc. For reasons such as these, the definition of the problem will need to be more elaborate(Saji and Riffi 2016). Considering all of these situations, the problem of asymmetrical traveling salesman involves a scenario in which the round-trip distance between two cities or points and the cost to be covered at these distances are not equal (Choong et al. 2019). A block diagram illustrating the symmetrical and asymmetrical traveling salesman problem is presented in Figure 5.



(a) Symmetrical traveler seller (b) Asymmetrical traveler seller

Figure 5. Symmetric / Asymmetric Traveler Salesman Problem Status

Figures 5 (a) and (b) present the symmetrical and asymmetrical traveling salesmen problem according to their distances. Figure 5- (a) shows that the round-trip distances between the same two-knot pairs are the same. On the other hand, in the symmetrical travelling salesman problem, the distances between the nodes remain the same for the round-trip and return. The opposite of this situation is given in Figure 5-(b). In this example, the round-trip distances between any two pairs of nodes are different from each other. When the data obtained within the scope of this study are examined, it can be observed that the optimization problem of the EMM public transportation network coincides with the asymmetric traveling salesman problem. For this reason, the study mainly aims to solve the problem of the asymmetric traveling salesman.

3. Proposed Method

Within the scope of this study, the optimization methods that can ascertain the optimal road routes for bus lines that are frequently encountered and used in daily life were researched and applied. For this purpose, data from the EMM Public Transportation Branch Directorate were provided, and the data were made ready for use in practice. After these steps, the distance between the coordinates of these stops was obtained using an application prepared using the Google API, and a distance matrix table was created. Three different

optimization methods were researched and applied in detail. In this context, the genetic algorithm, the DNA computing algorithm, and the clonal selection algorithm from an artificial immune system, which is one of the methods frequently used, as noted in the literature, were utilized. The approach proposed in this study was tested on six different routes and it aimed to shorten the routes that the buses actively follow. A block diagram that summarizes the system's flow in line with all of these objectives is shown in Figure 6.

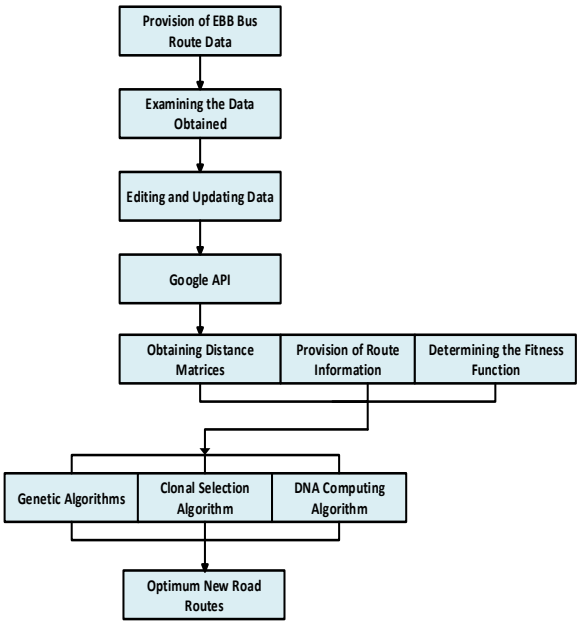


Figure 6. Steps of the Proposed Approach

As can be seen in Figure 6, three different optimization algorithms were used in the proposed method. These methods are the genetic algorithm, the clonal selection algorithm, and the DNA computing algorithm. The details for each of these optimization methods are presented below in the subsections.

3.1 The Genetic Algorithm

The genetic algorithm method consists of seven steps(Malhotra et al. 2011). A flow chart that summarizes the steps of the genetic algorithm is given in Figure 7, and the details for each of these steps are given below.

- Random Start Population Creation: The first step in using the genetic algorithm is to generate a random population based on the problem. Although there are various methods noted in the literature, e.g., binary coding and permutation coding, permutation coding was preferred for this study. Therefore, within the scope of this study, a 100-element population was created.
- Calculate Fitness: The next stage taken while using the genetic algorithm is to determine the purpose function and subject the individuals that were randomly created in the previous step to this function. The purpose function used in the application is presented in Equation 1.

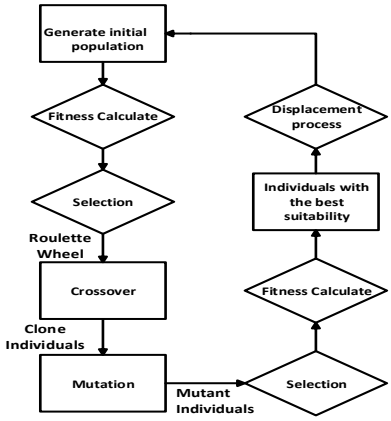


Figure 7. Genetic Algorithm Steps (Suman 2015)

$$\sum_{i=1}^{n-1} MT(P_i, P_{i+1}) \quad (1)$$
$$n = Total\ number\ of\ stops$$
$$P = Individual\ in\ the\ population$$
$$MT = Distance\ table$$

- Selection: At this stage in using the genetic algorithm, the individuals are ranked from the best to the worst, depending on their fitness values. After this ranking process, a selection process is carried out. There are various selection approaches taken up in the literature. The best known of these selection procedures are called elitism, roulette wheel, and tournament selection(Mohammed et al. 2017). In this study, the roulette wheel method was preferred.
- Crossover Process: In this step in using the genetic algorithm, the individuals selected in the previous step are subjected to the crossover process. The primary purpose here is to cross different individuals with the best suitability and to obtain child individuals with good suitability, but in a different structure (Hiassat et al. 2017). In this study, a single-point crossover was applied.
- Mutation Process: In this part of the genetic algorithm, a mutation process is applied to the child individuals obtained at the crossover stage. The mutation process is carried out in reverse proportion to the suitability of the individuals.
- Calculate Fitness and Change: The next step in using the genetic algorithm after the mutation phase is the calculation and change process. The re-eligibility of the new child individuals obtained is calculated and subjected to displacement with the individuals in the randomly generated population at the beginning(Priyo Anggodo et al. 2016).
- Termination: In this section, which constitutes the last stage in using the genetic algorithm, how long the algorithm will work is specified and, at the end of this period, the aim is to terminate the algorithm. In studies carried out for this purpose, the termination criteria of the genetic algorithm were determined in terms of 1000 iterations.

3.2 The Clonal Selection Algorithm

A flow diagram showing the steps involved in the clonal selection algorithm, taken from the artificial immune system, is provided in Figure 8. In addition, the steps for this method are as shown below (Muthreja and Kaur 2018).

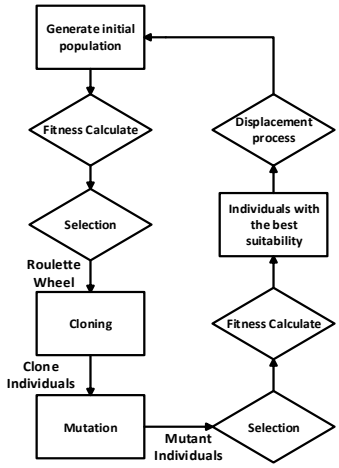


Figure 8. Clonal Selection Algorithm (Guney et al. 2007)

- Initially Population Create: The first step involved in the clonal selection algorithm is to generate random populations, just as in the genetic algorithm. In this step, in which permutation coding is preferred, again as in the genetic algorithm, random routes are created in terms of the population size specified by the user(Xu et al. 2016).
- Calculate Fitness: This stage involved in the clonal selection algorithm is the same process as in the genetic algorithm, and the calculation function given in equation 1 is used.
- Selection: At this stage in the clonal selection algorithm, the individuals with the best suitability are selected and these parent individuals are transferred to the next step, the crossover step. Again, in this algorithm, the roulette wheel method is preferred.
- Cloning: After the selection process is performed based on the suitability value, the next step is cloning, that is, the copying process of the selected "n" individuals(Shrikrishna et al. 2018). The primary purpose here is to copy the more well-suited individuals and fewer of the well-suited individuals. In this way, a new and temporary population can be obtained.
- Mutation: This is the process of replacing individuals in the temporary population by replicating the mutation process if certain conditions occur(Muthreja and Kaur 2018).
- Calculate Fitness and Change: The re-eligibility value of the temporary population that has been subjected to mutation is calculated and subjected to change.
- Termination: The last stage involved in the clonal selection algorithm is termination. This stage is the section that determines how long the clonal selection algorithm will operate.

3.3 The DNA Computing Algorithm

A flow diagram that summarizes the processes for the DNA computing algorithm is given in Figure 9.

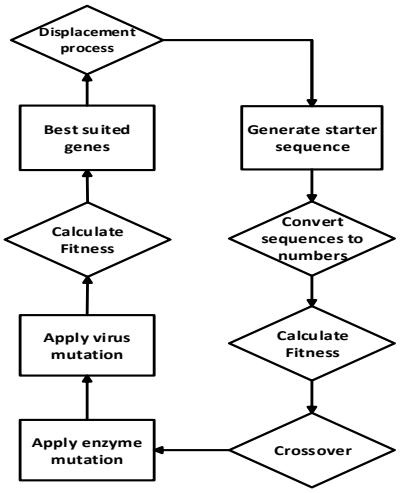


Figure 9. DNA Computing Algorithm Flow Diagram

- Generating Random DNA Sequences: The first step involved in the DNA computing algorithm is to generate random DNA sequences. DNA molecules are used in the creation of DNA sequences. These molecules are adenine, thymine, guanine, and cytosine. The numerical equivalents of these molecules are given below.

$$A = 0, G = 1, C = 2, T = 3$$
$$AAA = 0x4^0 + 0x4^1 + 0x4^2 = 0$$
$$TTT = 3x4^0 + 3x4^1 + 3x4^2 = 63$$

As can be seen from this example, by bringing the three molecules side by side, DNA sequences of up 63 stops can be obtained. In this study, the maximum number of stops in the selected lines is 60, and all of these stops can be expressed in terms of the three-string DNA sequences. In this context, when creating a random population, the DNA sequences of three are combined side by side and a DNA helix given in the total equation 2 is obtained. A sample DNA sequence randomly generated for an 18-stop route is given in Figure 10.

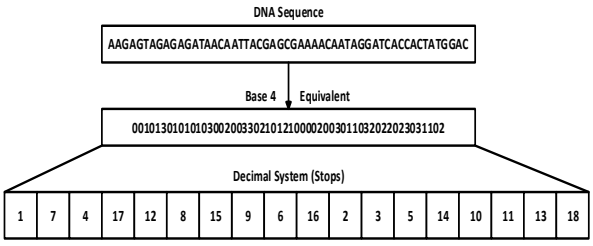


Figure 10: DNA Computing Algorithm Flow Diagram

- Numerical Value Cycle: The DNA helix obtained in the first stage of the application is converted into real numerical values in the second stage of the application(Zhang 2018).
- Calculate Fitness: A conformity calculation is carried out as in the other two optimization methods, and

Equation 1 is used for this purpose.

- Crossover Process: In the fourth stage used in the DNA computing algorithm, the crossover process is applied to the selected DNA sequences. The crossover is done on DNA strands, not on the numerical values.
- Enzyme and Virus Mutation: In the DNA computing algorithm, a two-stage mutation process is applied to the DNA sequences. The first of these processes is the enzyme mutation, which is the process of deleting any DNA sequences randomly selected from the DNA sequence. The second is a virus mutation, which involves replacing a randomly deleted DNA sequence with a newly produced DNA sequence(Dodge et al. 2020).
- Calculate Fitness and Change: This process is performed in the DNA computing algorithm and is applied just as in the other two methods. At this stage, The mutated DNA sequences are converted back into numerical data and their suitability is calculated(Ibrahim et al. 2018).
- Termination: A termination criterion is also used for the DNA computing algorithm. With this termination process, the DNA computing algorithm is stopped at a certain point and the result obtained is recorded as the solution. As with the other methods, 1000 iterations were preferred in using the DNA computing algorithm.

In this study, three different optimization methods were used for solving the traveling salesman problem. These methods were tested for the actively used routes, and the results obtained have been compared. The comparative steps for each of the algorithms used in the application are presented in Table 1.

Table 1. Comparison of the steps of the algorithms used

Step	GA	CSA	DNA
Start	- Determining the type of coding - Determination of fitness functions, population size and termination criteria		
1	Creating random populations	Creating random populations	Creating random populations
2	Calculate Fitness	Calculate Fitness	Calculate Fitness
3	Selection process	Selection process	Selection process
4	Crossover	Cloning	Crossover
5	Mutation	Mutation	-Enzyme mutation -Virus mutation
6	Calculate Fitness	Calculate Fitness	Calculate Fitness
7	Displacement	Displacement	Displacement
End	Repeat process from step 3 according to termination criteria		

4. Simulation Results

Six line routes, which the EMM public transportation administration actively uses, have been optimized using the steps outlined in this study. In the optimization processes, three different methods were used: clonal selection, genetics, and the DNA computing algorithm. In this study, the distance matrices are used in the calculation process, the results are then obtained, and then a comparison of the three optimization methods is made. The optimization methods take the distance matrix files as the input parameters. These distance matrix files were obtained through the Google API using GPS coordinates, and sample GPS data for one of the lines used is presented in Table 2.

Table 2. Comparison of the steps of the algorithms used

Stop	Coordinate	Stop	Coordinate
1	39.902638 41.274524	**10**	39.911979 41.266006
2	39.89879 41.270129	**11**	39.908586 41.265283
3	39.90003 41.267221	**12**	39.901406 41.266290
4	39.90942 41.265402	**13**	39.892800 41.248494
5	39.91174 41.265663	**14**	39.898357 41.262598
6	39.907539 41.278788	**15**	39.899936 41.266736
7	39.906147 41.286567	**16**	39.8981894 1.270431
8	39.905089 41.290898	**17**	39.897225 41.272914
9	39.911879 41.272602	**18**	39.898594 41.275351

As can be seen from Table 2, the D1 route consists of 18 stops in total. The GPS coordinates of these stops are given in Table 2. Within the scope of the application, a total of six different lines are used, and the properties for each of these lines are given in Table 3.

Table 3. The route information used within the scope of the application

Route No	D1	G1	G10	G3	G5	G7
Stop Number	18	42	60	45	57	28

The distances between the stops are different from each other. This is due to the problem of being asymmetrical. In other words, in this problem, the transportation distance from point “A” to point “B” and the distance from point “B” to point “A” may not be the same. For this reason, the distance matrix files previously obtained for each route are used for calculating the compliance function. As a result of touring the line routes made within the scope of this study, all of the bus stop coordinates for the six lines selected were obtained in a way that is closest to reality by considering the possible error margins (one to three meters) that may occur in the GPS fixing device. The real data that was obtained has been tested with the optimization methods and new routes that have shorter lengths than the existing routes have been identified. The conformity value changes obtained from the algorithms

for the selected routes are presented in Figure 11.

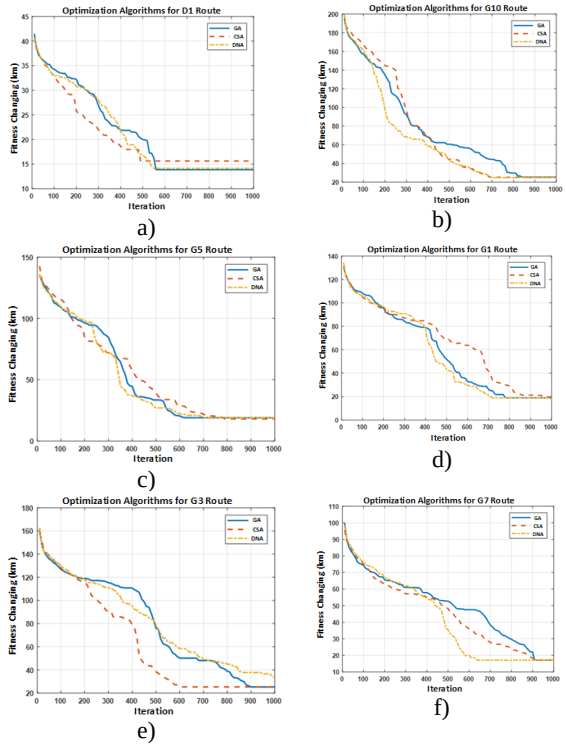


Figure 11. The changes of the suitability of algorithms for routes

In Figure 11, the compatibility changes of the optimization algorithms for each line are presented. As can be seen from the figure, each algorithm behaves differently for each of the lines. The final results obtained from these algorithms are presented in detail in Table 4.

As shown in Table 4, there has been an improvement on almost all of the lines. The genetic algorithm provides the optimum improvement for three lines, the clonal selection algorithm does so for two lines, and the DNA Computing Algorithm does so for three lines. The new optimized routes obtained from the algorithms, the daily voyages carried out on these lines, and the daily gains based on these voyages are given in detail in Table 5.

In addition, the new optimized routes that were obtained from the algorithms are presented in Table 6. As can be seen in Table 6, changes have been experienced in all of the routes. In the final stage of the application, the new routes obtained, along with the existing routes, are drawn on Google Maps, and the results for the application is visualized. An example image for route D1 is given in Figure 12.

Table 4. The change of distance based on the algorithms (km)

	Lines					
	D1	**G7**	**G1**	**G3**	**G5**	**G10**
Current Distance (Km)	16.2710	17.1850	18.8170	25.8130	20.1150	26.7470
Genetic Algorithm	**13.8350**	**17.0340**	18.8170	**25.3020**	18.7720	25.3110
Clonal Selection Algorithm	15.6110	17.1850	18.8170	**25.3020**	**18.0490**	25.2350
DNA Computing Algorithm	14.1270	**17.0340**	18.8170	**25.3020**	18.5810	24.8480

Table 5. The difference states varying by distance (km)

Routes	**Previous Distance (Km)**	**Next Distance (Km)**	**Percent (%)**	**Number of Voyages (Daily)**	**Daily Acquisition (Km)**
D1	16,2710	13,8350	14,9	16	38,976
G1	18,8170	18,8170	0,0	16	0
G10	26,7470	24,8480	7,09	14	21,168
G3	25,8130	25,3020	1,97	15	7,665
G5	20,1150	18,0490	10,27	15	30,990
G7	17,1850	17,0340	0,87	18	2,718
				Total	**101,517**
				Average	**16,920**

Table 6: The exit routes obtained as a result of the application

D1	1	7	8	6	9	10	11	4	5	2	3	12	13	14	15
	16	17	18												
G1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40	41	42			
G10	1	2	3	5	56	6	7	8	55	9	54	10	13	14	49
	16	47	18	19	20	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	17	46	48	15	50	51	52	12	53	11	4	57	58	59	60

G3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	16	17	35	19	20	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	18	36	37	38	39	40	41	42	43	44	45
G5	1	4	54	5	53	6	52	7	47	11	46	12	13	14	15
	16	17	18	19	20	21	22	35	23	25	26	27	28	29	30
	31	32	33	34	24	36	37	38	39	40	41	42	43	44	45
	10	48	9	49	8	50	51	55	3	2	56	57			
G7	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16
	14	17	18	19	20	21	22	23	24	25	26	27	28		

Figure 12. An example image for route D1

5. Results and Discussion

Today, public transportation systems are actively used in many countries around the world. In particular, railway, seaway, and road transportation are the services preferred by people when using public transport systems. One of the main reasons why these services are especially preferred by people is their desire to complete their daily work faster, easier, and in a shorter amount of time. As can be expected, people need to arrange these routines according to specific days and times in order to, do their daily routines as quickly as possible. When considering all of these situations, the concept of optimization emerges.

The optimization process is basically the selecting the best or near best solution from a vast solution space. The optimization process, which is used in many different fields such as financial planning, computer science, and industrial computing, is encountered even if people regulate their daily work. In this study, an application that can determine the optimal route for municipal bus lines, which is a service offered by the municipalities and is frequently encountered in daily life, is introduced. For this purpose, a problem based on real data was solved by using data provided by the EMM Public Transportation Branch Directorate. The data provided was used in applications made during this study using Google API and Google Maps. These routes, subjected to testing, have been examined, optimized, and visualized on Google Maps.

During this research, various restrictions were identified and some findings were identified. In this context, the research was carried out only for six of the line routes. Applying this method for all of lines the would provide significant results in terms of ascertaining the reliability of the applied methods. In addition, looking at the real-world experience of the test results obtained from the research on some pilot routes would confirm the applicability and performance of the results.

In addition, it was determined that some of the stations were located very close to each other and some of them were located very far from each other when the stops were physically visited in order to confirm the data provided by the Metropolitan Municipality. In addition, when considering the climatic structure of Erzurum, the distance between the stops is relatively far from each other, which significantly reduces the ease of access to the buses. For this reason, the stop locations need to be revised. During this research, it was observed that the majority of the passengers using the buses travel by foot. With the implementation of the optimization process on these lines, the return of the earnings obtained on km basis as an increase in the number of trips will allow the passengers to travel more comfortably and more efficiently.

When considering the energy consumption problems today, buses need to achieve the minimum amount of

fuel consumption. Depending on the situation, both the stops' locations and the buses' routes should be reviewed and updated periodically. In addition, the problem of global warming directly affects many living spaces. The biggest reason for this situation is the consumption of fossil fuels. At this point, minimizing the emissions of harmful carbon gases from these buses will provide many societal and environmental benefits.

ACKNOWLEDGEMENT

In this study, we thank the Erzurum Metropolitan Municipality Public Transportation Administration for the data it provided.

REFERENCES

Asadpour, A., Goemans, M. X., Mądry, A., Gharan, S. O., and Saberi, A. 2010. "An O(Log n/ Log Log n)-Approximation Algorithm for the Asymmetric Traveling Salesman Problem," *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* (June), pp. 379–389. (https://doi.org/10.1137/1.9781611973075.32).

Baygin, M., and Karakose, M. 2013. "Immunity-Based Optimal Estimation Approach for a New Real Time Group Elevator Dynamic Control Application for Energy and Time Saving," *The Scientific World Journal* (2013). (https://doi.org/10.1155/2013/805343).

Bolat, B., and Cortés, P. 2011. "Genetic and Tabu Search Approaches for Optimizing the Hall Call - Car Allocation Problem in Elevator Group Systems," *Applied Soft Computing Journal* (11:2), pp. 1792–1800. (https://doi.org/10.1016/j.asoc.2010.05.023).

Chen, P. 2013. "An Improved Genetic Algorithm for Solving the Traveling Salesman Problem," *Proceedings - International Conference on Natural Computation* (6), pp. 397–401. (https://doi.org/10.1109/ICNC.2013.6818008).

Choong, S. S., Wong, L. P., and Lim, C. P. 2019. "An Artificial Bee Colony Algorithm with a Modified Choice Function for the Traveling Salesman Problem," *Swarm and Evolutionary Computation* (44:December 2017), pp. 622–635. (https://doi.org/10.1016/j.swevo.2018.08.004).

Deng, W., Zhao, H., Zou, L., Li, G., Yang, X., and Wu, D. 2017. "A Novel Collaborative Optimization Algorithm in Solving Complex Optimization Problems," *Soft Computing* (21:15), Springer Berlin Heidelberg, pp. 4387–4398. (https://doi.org/10.1007/s00500-016-2071-8).

Dodge, M., MirHassani, S. A., and Hooshmand, F. 2020. "Solving Two-Dimensional Cutting Stock Problem via a DNA Computing Algorithm," *Natural Computing* (3), Springer Netherlands. (https://doi.org/10.1007/s11047-020-09786-3).

Groba, C., Sartal, A., and Vázquez, X. H. 2015. "Solving the Dynamic Traveling Salesman Problem Using a Genetic Algorithm with Trajectory Prediction: An Application to Fish Aggregating Devices," *Computers and Operations Research* (56), Elsevier, pp. 22–32. (https://doi.org/10.1016/j.cor.2014.10.012).

Guney, K., Babayigit, B., and Akdagli, A. 2007. "Position Only Pattern Nulling of Linear Antenna Array by Using a Clonal Selection Algorithm (CLONALG)," *Electrical Engineering* (90:2), pp. 147–153. (https://doi.org/10.1007/s00202-006-0056-9).

Guo, Z., Koehler, G. J., and Whinston, A. B. 2007. "A Market-Based Optimization Algorithm for Distributed Systems," *Management Science* (53:8), pp. 1345–1358. (https://doi.org/10.1287/mnsc.1060.0690).

Hasan Söyler, T. K. 2007. *Karınca Kolonisi Algoritması Ile Gezen Satıcı Probleminin Çözümğ*.

Hiassat, A., Diabat, A., and Rahwan, I. 2017. "A Genetic Algorithm Approach for Location-Inventory-Routing Problem with Perishable Products," *Journal of Manufacturing Systems* (42), The Society of Manufacturing Engineers, pp. 93–103. (https://doi.org/10.1016/j.jmsy.2016.10.004).

Ibrahim, G. J., Rashid, T. A., and Sadiq, A. T. 2018. "Evolutionary DNA Computing Algorithm for Job Scheduling Problem," *IETE Journal of Research* (64:4), pp. 514–527. (https://doi.org/10.1080/03772063.2017.1362964).

Kovács, L., Iantovics, L. B., and Iakovidis, D. K. 2018. "IntraClusTSP-An Incremental Intra-Cluster Refinement Heuristic Algorithm for Symmetric Travelling Salesman Problem," *Symmetry* (10:12). (https://doi.org/10.3390/sym10120663).

Mahi, M., Baykan, Ö. K., and Kodaz, H. 2015. "A New Hybrid Method Based on Particle Swarm Optimization, Ant Colony Optimization and 3-Opt Algorithms for Traveling Salesman Problem," *Applied Soft Computing Journal* (30), Elsevier B.V., pp. 484–490. (https://doi.org/10.1016/j.asoc.2015.01.068).

Malhotra, R., Singh, N., and Singh, Y. 2011. "Genetic Algorithms: Concepts, Design for Optimization of Process Controllers," *Computer and Information Science* (4:2), pp. 39–54. (https://doi.org/10.5539/cis.v4n2p39).

Mohammed, M. A., Abd Ghani, M. K., Hamed, R. I., Mostafa, S. A., Ahmad, M. S., and Ibrahim, D. A. 2017. "Solving Vehicle Routing Problem by

Using Improved Genetic Algorithm for Optimal Solution," *Journal of Computational Science* (21), Elsevier B.V., pp. 255–262. (https://doi.org/10.1016/j.jocs.2017.04.003).

Muthreja, I., and Kaur, D. 2018. "A Comparative Analysis of Immune System Inspired Algorithms for Traveling Salesman Problem," *2018 World Congress in Computer Science, Computer Engineering and Applied Computing, CSCE 2018 - Proceedings of the 2018 International Conference on Artificial Intelligence, ICAI 2018*, pp. 164–170.

NARALAN, A., KALELİ, S. S., and BAYĞIN, M. 2017. "Shortest Path Detection Using Clonal Selection Algorithm for Erzurum Metropolitan Municipality," *Mugla Journal of Science and Technology* (3:2), pp. 138–142. (https://doi.org/10.22531/muglajsci.357621).

Nguyen, H. D., Yoshihara, I., Yamamori, K., and Yasunaga, M. 2002. "Greedy Genetic Algorithms for Symmetric and Asymmetric TSPs," *IPSJ Trans. Mathematical Modeling and Its Applications* (43:10), pp. 165–175.

Priyo Anggodo, Y., Kartika Ariyani, A., Khaerul Ardi, M., and Firdaus Mahmudy, W. 2016. "Optimization of Multi-Trip Vehicle Routing Problem With Time Windows Using Genetic Algorithm," *Journal of Enviromental Engineering and Sustainable Technology* (3:2), pp. 92–97. (https://doi.org/10.21776/ub.jeest.2017.003.02.4).

Saji, Y., and Riffi, M. E. 2016. "A Novel Discrete Bat Algorithm for Solving the Travelling Salesman Problem," *Neural Computing and Applications* (27:7), Springer London, pp. 1853–1866. (https://doi.org/10.1007/s00521-015-1978-9).

Shrikrishna, K., B, N. V. N. K., and Shyamasundar, R. K. 2018. *Security Analysis of EMV Protocol*, (2:December), pp. 69–85. (https://doi.org/10.1007/978-3-319-72344-0).

Suman, S. K. 2015. "Genetic Algorithms: Basic Concepts and Real World Applications," *International Journal of Electrical, Electronics and Computer Systems* (November). (https://scholar.google.co.in/citations?view_op=view_citation&continue=/scholar%3Fhl%3Den%26as_sdt%3D0,5%26scilib%3D1&citilm=1&citation_for_view=dj9eYFMAAAAJ:UebtZRa9Y70C&hl=en&oi=p).

Sundararaghavan, P. S., Kunnathur, A., and Fang, X. 2010. "Sequencing Questions to Ferret out Terrorists: Models and Heuristics," *Omega* (38:1–2), Elsevier, pp. 12–19. (https://doi.org/10.1016/j.omega.2009.01.002).

Wang, J., Ersoy, O. K., He, M., and Wang, F. 2016. "Multi-Offspring Genetic Algorithm and Its Application to the Traveling Salesman Problem," *Applied Soft Computing Journal* (43), Elsevier B.V., pp. 415–423. (https://doi.org/10.1016/j.asoc.2016.02.021).

Xu, Z., Wang, Y., Li, S., Liu, Y., Todo, Y., and Gao, S. 2016. "Immune Algorithm Combined with Estimation of Distribution for Traveling Salesman Problem," *IEEJ Transactions on Electrical and Electronic Engineering* (11), pp. S142–S154. (https://doi.org/10.1002/tee.22247).

Zhang, Y. 2018. "The Image Encryption Algorithm Based on Chaos and DNA Computing," *Multimedia Tools and Applications* (77:16), Multimedia Tools and Applications, pp. 21589–21615. (https://doi.org/10.1007/s11042-017-5585-x).

Research Article
Journal of Intelligent Systems: Theory and Applications 5(1) 2022: 75-84
DOI: 10.38016/jista.996541

Classification of Stockwell Transform Based Power Quality Disturbance with Support Vector Machine and Artificial Neural Networks

Ezgi Güney[1]*, Ozan Çakmak[2], Çağrı Kocaman[3]

[1] Sinop University, Vocational High School, Department of Electrical And Energy, Sinop, Turkey

[2] Ondokuz Mayıs University, Vocational High School, Department of Electrical And Energy, Samsun, Turkey

[3] Turkish Airlines Flight Management, İstanbul, Turkey

eguney@sinop.edu.tr , ozan.cakmak@omu.edu.tr, ckocaman@thy.com

Abstract
The detection and classification of power quality events that disturb the voltage and/or current waveforms in the electrical power distribution networks is very important to generate electrical energy and to deliver this energy to the end-user equipment at an acceptable voltage. Various property extraction methods are used to determine the type of disturbances in the electrical signal. In this study, seven power distortions including voltage sag, voltage swell, voltage harmonics, voltage sag with harmonics, voltage swell with harmonics, flicker, transient signals and pure sine as a reference signal is used. Synthetic data are produced in MATLAB using parametric equations based on TS EN 50160 standard. Four kinds of feature extraction as frequency-amplitude, time-amplitude, geometric mean and standard deviation is made with Stockwell Transform (ST), which is one of the methods used for the feature extraction of the determined GKB. Detection of voltage distortions is interpreted through these properties. 640 simulation data is entered into the classifier by using Support Vector Machines (SVM) and Artificial Neural Networks (ANN) and their classification performance is compared.

Keywords: Power Quality Disturbance, Stockwell Transform, Support Vector Machine, Artificial Neural Network.

Stockwell Dönüşümü Tabanlı Güç Kalitesi Bozunumlarının Destek Vektör Makinası ve Yapay Sinir Ağları ile Sınıflandırılması

Öz
Elektrik enerjisi hizmetlerinin kesintisiz bir biçimde tüketiciye ulaştırılması büyük önem taşımaktadır. Sistemdeki bozulmaların tespiti ve alınması gereken önlemler bu açıdan önemlidir. Elektrik sinyalindekini bozulmaların türünün belirlenmesi için çeşitli özellik çıkarım yöntemleri kullanılmaktadır. Bu çalışmada, elektrik güç sistemlerinde meydana gelen Güç Kalitesi Bozunumlarından(GKB) gerilim yükselmesi, gerilim çökmesi, harmonikli gerilim, harmonikli gerilim düşmesi, harmonikli gerilim yükselmesi, flicker ve transient ile referans sinyali olarak saf sinüs sinyallerini içeren sekiz işaret toplam on dönem sürecek şekilde TS EN 50160 standartlarına göre MATLAB ortamında oluşturulmuştur. Belirlenen GKB'na ait özellik çıkarımı için kullanılan yöntemlerden biri olan Stockwell-Dönüşümü ile frekans-genlik, zaman-genlik, geometrik ortalama ve standart sapma olmak üzere 4 çeşit özellik çıkarımı yapılmıştır. Bu özellikler üzerinden gerilim bozulmalarının tespiti yorumlanmıştır. Toplam 640 benzetim verisi Destek Vektör Makinaları (DVM) ve Yapay Sinir Ağları(YSA) ile sınıflandırıcıya sokularak sınıflandırma başarımları karşılaştırılmıştır.

Keywords: Güç Kalitesi Bozunumları, Stockwell dönüşümü, Destek Vektör Makinaları, Yapay Sinir Ağları.

* Corresponding Author.
E-mail: eguney@sinop.edu.tr

Recieved : 16 Sep. 2021
Revision : 7 Feb. 2022
Accepted : 10 Feb. 2022

1. Introduction

Today, power quality problems are an important issue for electrical energy services. In an electrical power system, the power reaching the end consumer must be clean. That is, it must be completely sinusoidal and the basic parameters of the network must be acceptable and within the limits set by the standards (Elango et al.,2016). Equipment used in power distribution are very sensitive to malfunctions in the supply systems (Singh et al., 1999). Power Quality (PQ) is vital to the smooth operation of power systems (Singh et al.,2017). The majority of the loads in the system are non-linear loads and cause system failure. These distortions produce results such as system resonance, capacitor overload and decrease in efficiency and changes in voltage magnitude (Dharavath et al., 2017). The main reasons for the deterioration in power quality are malfunction, load switching, capacitor switching, high switching frequency electronic devices, power converters, arc furnaces and transformers. Timely reducing of these distortions requires quick and accurate classification. It is of great importance to process and extract the signals for a successful classification. Some popular methods used for feature extraction of power quality are Short-Time Fourier Transform (STFT)(Azam et al.,2004, Ingale, 2014, Yoo et al., 2015) , Hilbert-Huang Transform (HHT) (Tao et al., 2013, Saxena et al., 2014), Wavelet Transform (WT) (Poisson et al., 2000, Gaing, 2004) and Stockwell Transform (S-transform) (Mahela et al., 2016, Raj et al., 2016, Zhao et al., 2016, Shamachurn, 2019, Liang et al., 2021). FT and STFT are not sufficiently successful in feature extraction. Although frequency analysis is performed well with FT, time information cannot be obtained (Zhao et al., 2016). STFT, which allows Fourier analysis by windowing in short time intervals for time information, is also not successful enough (Karasu, 2016). WT has been extensively used in feature extraction of power quality impairment. However, this method is greatly affected by electrical noise in the signal. S-transform is a time-frequency spectral localization technique proposed by Stockwell, which combines the features of WT and STFT. The S-transformation uses a window whose width decreases with frequency and provides a frequency-dependent resolution (Elango et al.,2016).

For classification, Artificial Neural Networks (ANN) (Agarwal et al., 2018), Support Vector Machine (SVM) (Ozgonenel et al., 2013, Thirumala et al., 2018, Choudhary, 2021), Fuzzy Logic (FL) techniques (Chilukuri et al., 2004, Mishra et al., 2021), Deep Learning (DP) methods (Wang et al., 2019, Sindi et al., 2021) are extensively used.

The aim of this study is to determine and analyze the PQDs of power systems in a strong- reliable way with the S-transformation, and detect the most suitable classifier. For this purpose, firstly, seven different PQDs are created in MATLAB environment including voltage sag, voltage swell, transient at different amplitude, duration and angles, and voltage harmonic, voltage sag with harmonic, voltage swell with harmonic and flicker at different time and frequency. Sampling frequency is taken 25.6 kHz. Pure sine signal is selected as reference. The obtained waveform feature extraction is made by S-transformation. A total of 640 simulation data are obtained from the S-transformation of the signals for Amplitude-Time, Amplitude-Frequency, Geometric Properties, Standard Deviation Properties. Using these features, the success of classification of PQD is investigated with SVM and ANN.

The paper is organized in four sections. Section 1 gives a basic introduction to the topic. Section 2 describes the S-Transform and feature extraction technique according to types of power quality disturbances. Section 3 presents the achieved test results and discussion. In this section, classification techniques based on SVM and ANN are elaborated. Section 4 presents conclusion.

2. Materials and Methods

2.1. S-Transformation

First, The S-transform was defined by R. G. Stockwell and was derived from the continuous wavelet transform. This transformation includes both amplitude and phase spectrum information together (Raj et al., 2016). S-transform is a method that involves both short-time Fourier transform and wavelet analysis but falls into a different category (Cortes et al., 1995). The wavelet transform cannot yield significant results in noisy

environments, while the S-transform provides successful results in property extraction in the presence of noise. This makes S-transformation suitable for accurate detection and classification of power quality disturbances.

The S-transform uses an analysis window that decreases in width depending on frequency and provides a frequency-dependent resolution. The time-frequency spectrum of the modulated signal is focused. The time-frequency analysis technique provides a three-dimensional graph of a signal in terms of signal energy or magnitude of time and frequency (Zhao et al., 2016).

The general S-transform is defined by Equation (1).

$$s(\tau, f) = \int_{-\infty}^{\infty} x(t) g(\tau - t, f) e^{-j2\pi f t} dt \qquad (1)$$

x (t) is the signal and g (t) is the windowing function. The window function is a modulated Gaussian function expressed by Equation (2).

$$g(\tau) = \frac{|f|}{\sqrt{2\pi}} e^{-(t^2 f^2/2)} \qquad (2)$$

The general equation is;

$$(\tau, f) = \int_{-\infty}^{\infty} x(t) \frac{|f|}{\sqrt{2\pi}} e^{-((\tau-t)^2 f^2/2)} e^{-j2\pi f t} dt \quad (3)$$

After obtaining the S-transform, four different properties are extracted. The first property is the amplitude-time property obtained by taking the largest values of the lines of the S-transformation. This feature provides information about the amplitude of the signal. The second feature is the frequency-amplitude property obtained by taking the largest value of the columns of the S-transformation. This feature provides information about the frequency of the signal. The third feature is the geometric mean of the S-transformation. This feature helps to locate sudden amplitude changes in the signal. The last feature is obtained by taking the standard deviation. This provides the same information as the time-amplitude property, but is additionally used to detect the harmonics in the signal.

2.2. According to Disturbances Types of Signals S-Transformation and Feature Extraction

The healthy simulation model operates at 1 pu voltage amplitude. Sampling frequency is 25.6 kHz. PQD are produced as 5120 samples in 10 period's length.

2.2.1. Pure Signal

The pure sine signal and its S-transformation graphs are given in Figure 1. Frequency-amplitude, time-amplitude, geometric mean and standard deviation characteristics obtained from S-transform of pure sine signal are given in Figure 2.

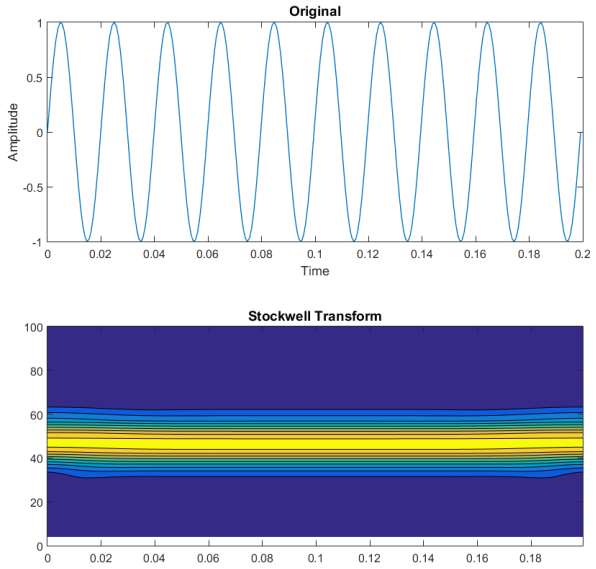


Figure 1. Pure Sine signal and S-transformation

The amplitude-time graph, which gives information about the amplitude of the signal, shows that the amplitude is constant and does not change over time. In the amplitude-frequency graph, it is seen that only 50 Hz network frequency is available. Since the periodicity of the signal is not disturbed, no change is observed in the graph of the geometric mean. In the standard deviation graph, we obtain information about the amplitude of the signal as in the amplitude-time graph, but unlike the first feature, we can also observe periodic fluctuations from this graph.

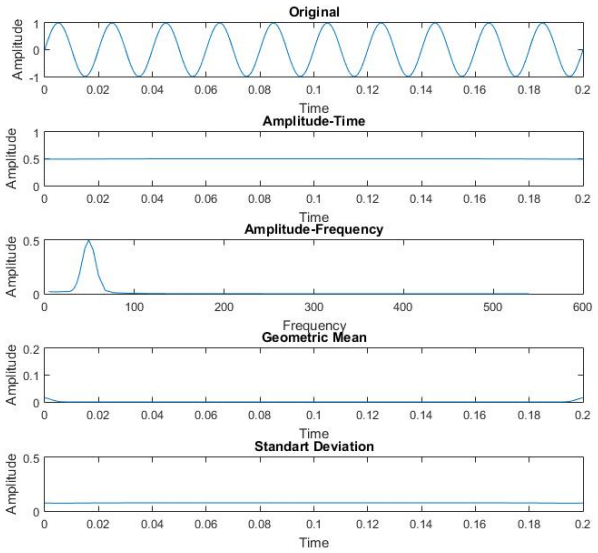


Figure 2. Properties derived from S-transformation of pure sine

2.2.2. Voltage Sag

Voltage sag is defined as the decrease in the mains voltage nominal value between 10-90%. The graph and S-transform of a 10-period mains voltage signal with a 50% voltage sag between the 2nd and 6th periods are given in Figure 3.

The frequency-amplitude, time-amplitude, geometric mean and standard deviation characteristics obtained from the S-transform of the voltage sag signal are given in Figure 4.

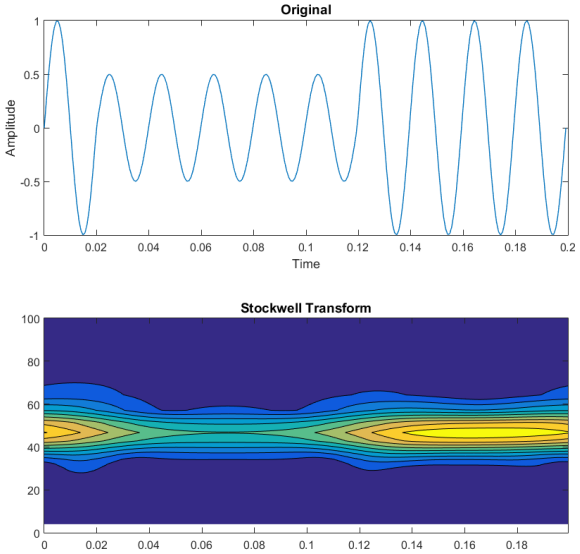


Figure 3. 50% voltage sag signal and S-transform

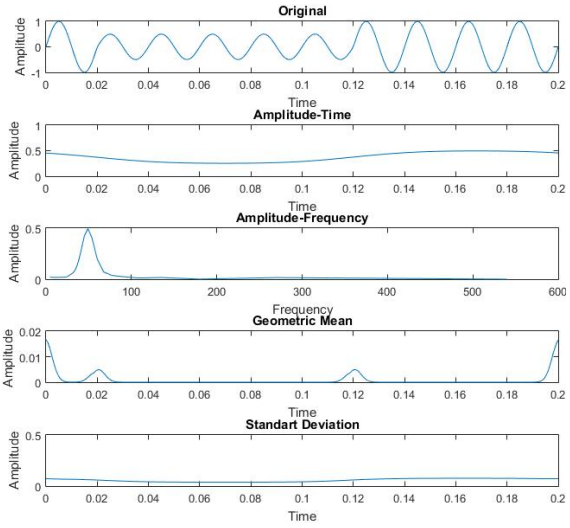


Figure 4. Characteristics obtained from S-transform of 50% voltage sag signal

As seen in the amplitude-time graph, which is the first feature, the amplitude decreased between the 2nd and 6th periods. This decrease is also seen in the standard deviation graph. Since there is no change in frequency, there is no difference in amplitude-frequency graph compared to pure sine. In the geometric mean graph, the location of the changes in the original signal on the time axis is revealed. The graph has changed at the beginning of the 2nd and 6th periods.

2.2.3. Voltage Swell

Voltage swell is defined as the increase of the mains voltage nominal value to 110-180%. The graph and S-transform of a 10-period mains voltage signal with a 150% voltage swell between the 2nd and 6th periods are given in Figure-5. The characteristics of the voltage swell signal are given in figure 6.

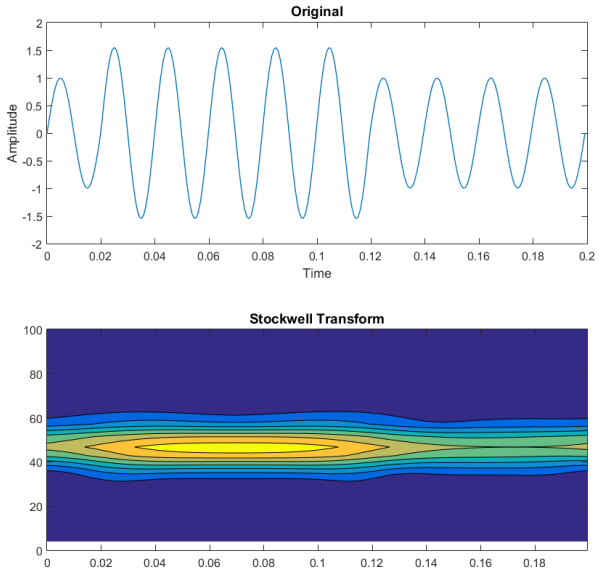


Figure 5. 50% voltage swell signal and S-transform

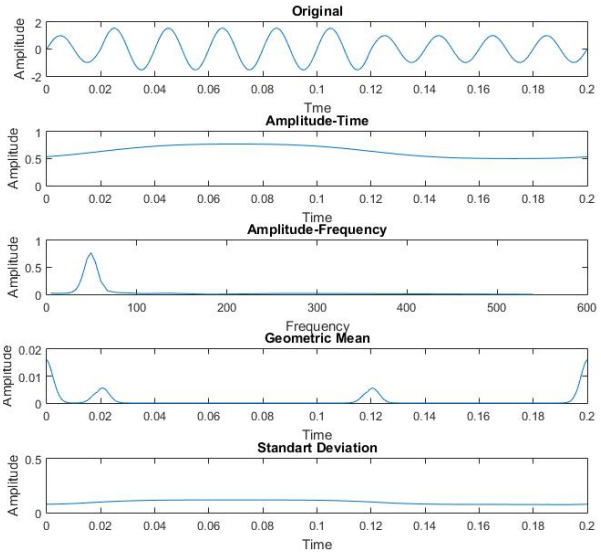


Figure 6. Characteristics obtained from the S-transform of 80% voltage swell signal

On the amplitude-time graph, the amplitude increased between the 2nd and 6th periods. The frequency-time graph shows that only a frequency of 50 Hz is present. On the other hand, the geometric mean graph shows the change in the original signal at the beginning of the 2nd and 6th periods, which are the place of the changes in the time axis.

2.2.4. Flicker

Voltage flickers under the frequency of 50 Hz, which are caused by fluctuations in the flicker load and cause flickering in lighting. In the standards, flicker intensity limit values are given between 0.8 and 1. In Figure 7, 0.8 flicker voltage signal and S-transform are given. The properties obtained from S-transformation for this disturbance are given in Figure-8.

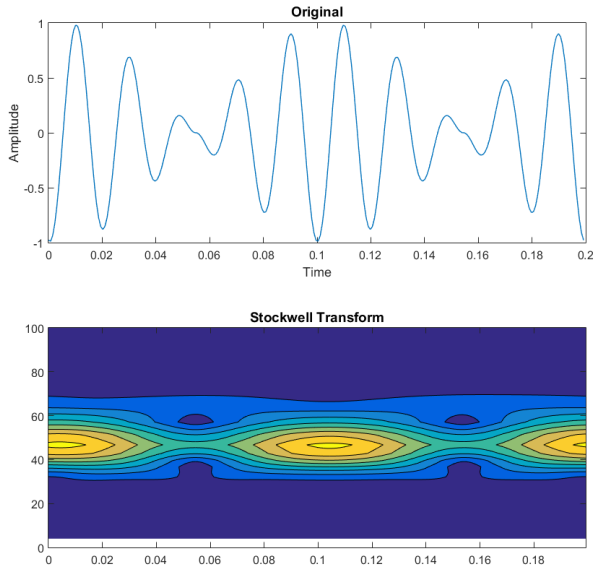


Figure 7. Voltage signal and S-transform with flicker intensity limit value 0.8

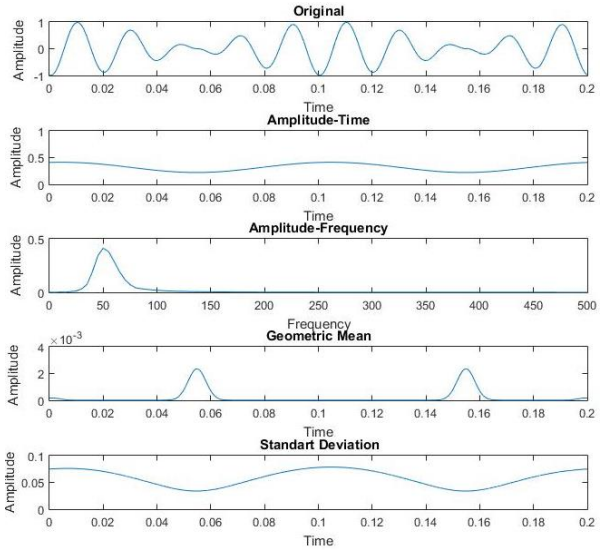


Figure 8. Properties of the S-transform of the voltage signal with a flicker intensity limit value of 0.8

The amplitude time graph shows the decrease in the voltage caused by the flicker. The amplitude frequency graph shows a 50 Hz sine signal. In the feature graph extracted from the geometric mean, the times when flicker is realized become apparent in the 3rd and 8th Periods. Since there are no different frequencies in the signal, only the voltage variation is observed in the standard deviation graph.

2.2.5. Transient

This is called a temporary change in the power system that takes place from 50 ns up to 50 ms. The distortion voltage of 4 kHz in the 3rd and 6th periods and the S-transform are given in Figure-9. Figure-10 shows the properties extracted from the S-transformation of this signal.

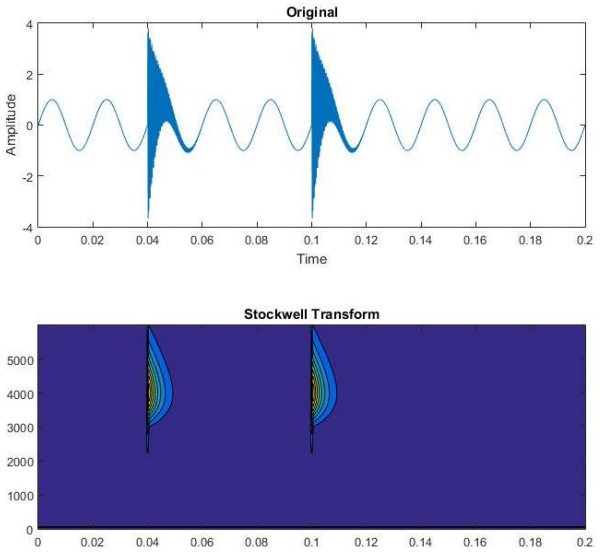


Figure 9. Transient signal and S-transform of power quality disturbance

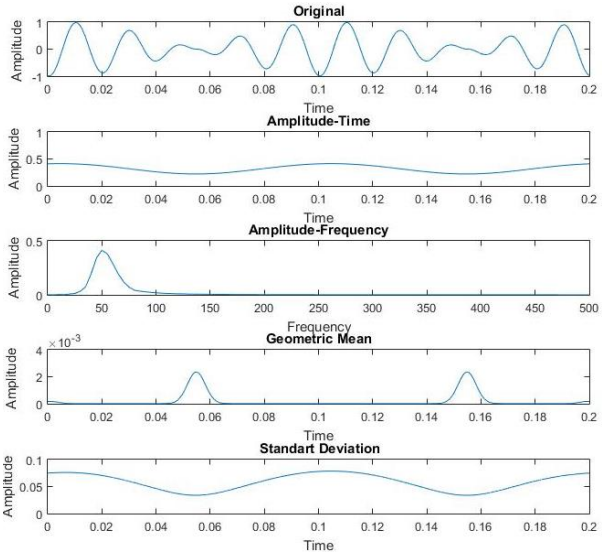


Figure 10. Features derived from S-transformation of transient power quality disturbance

The amplitude time graph shows the parts where the voltage rises. The 50 Hz signal and the transient signal present in the signal are plotted on the amplitude-frequency graph. The geometric mean graph gives the starting points of the transient signals. Finally, the standard deviation graph shows the change along the transient along with the amplitude change.

2.2.6. Voltage with Harmonics

It is the distortion of the voltage or current waveform from the ideal sine. The signal containing the 3rd and 5th harmonics in the 10-period signal and the S-transformation of this signal are given in Figure-11.

S-transformed properties of harmonic voltage are given in Figure-12.

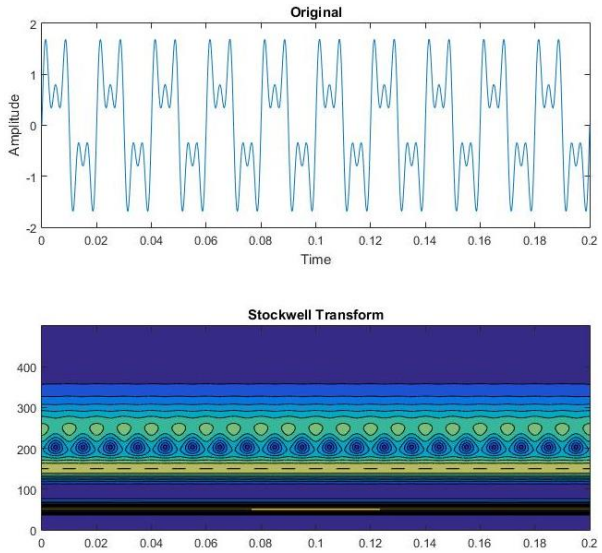


Figure 11. The 3rd and 5th harmonic added voltage signal and S-transform

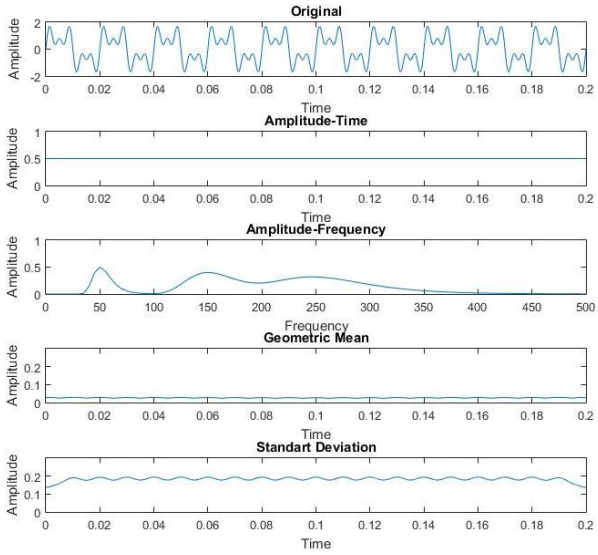


Figure 12. Properties obtained from S-transform of the 3rd and 5th harmonic added voltage signal

The amplitude-time plot is constant since no voltage decrease or increase in the periodic signal occurs. In the amplitude-frequency graph, the 3rd and 5th harmonics in the signal are seen as 150 kHz and 250 kHz. On the other hand, since the amplitude changes due to the 3rd and 5th harmonics are small, no significant change was observed in the geometric feature graph. In the standard deviation feature, fluctuations due to harmonic can be observed.

2.2.7. Voltage Sag with Harmonics

The 10-period signal contains the 3rd and 5th harmonics, as well as the voltage sag during the 2nd and 6th periods and the S-transform is given in Figure-13.

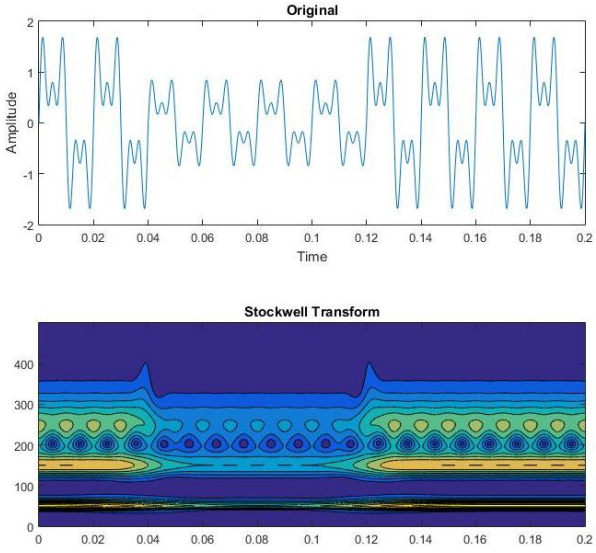


Figure 13. voltage sag signal with the 3rd and 5th harmonic added and S-transform

The properties of the voltage sag with harmonics signal derived from the S-transform are given in Figure 14. From this amplitude-time graph, the range in which the voltage drops are observed. The presence of harmonics of 150 Hz. and 250 Hz. is obtained from the frequency time graph. The range in which voltage starts to drop and ends is seen from the graph of the geometric property. In the standard deviation graph, which is the prominent feature in the detection of harmonics, harmonics can be observed.

Based on these properties, the range in which the voltage rises are clearly seen on the amplitude time graph. The frequency values in the signal are determined as 50 Hz, 150 Hz and 250 Hz in the frequency time graph.

As it can be seen in the graph of the geometric feature that helps us to determine the interval where the voltage starts to rise and ends, it started in the 3rd period and ended in the 6th Period. In the standard deviation graph, the observed fluctuations reveal the presence of harmonics in the signal.

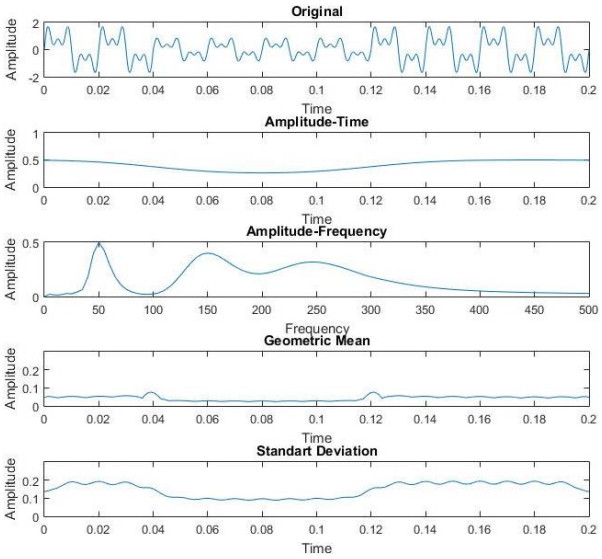


Figure 14. Properties obtained from the S-transform of the 3rd and 5th harmonic added voltage sag signal

2.2.7. Voltage Swell with Harmonics

The 10-period signal contains the 3rd and 5th harmonics, as well as the voltage swell during the 2nd and 6th periods and the S-transform is given in Figure-15.

The characteristics of the voltage swell with harmonics signal are given in Figure-16. Based on these properties, the range in which the voltage rises are clearly seen on the amplitude time graph. The frequency values in the signal are determined as 50 Hz, 150 Hz and 250 Hz in the frequency time graph. As it can be seen in the graph of the geometric feature that helps us to determine the interval where the voltage starts to rise and ends, it started in the 3rd period and ended in the 6th Period. In the standard deviation graph, the observed

fluctuations reveal the presence of harmonics in the signal.

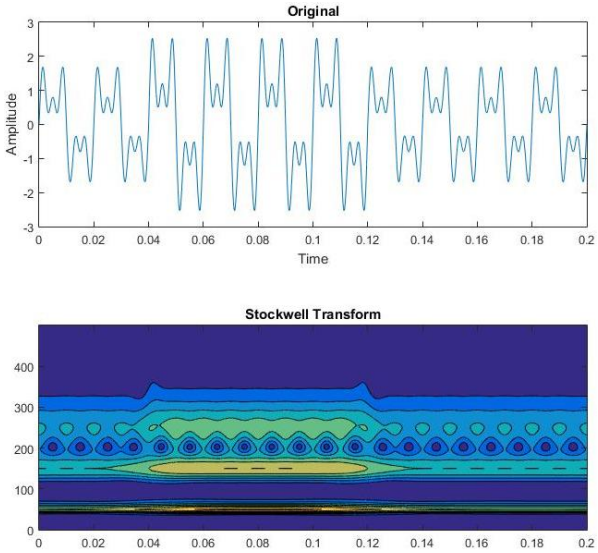


Figure 15. Voltage swell signal with the 3rd and 5th harmonic added and S-transform

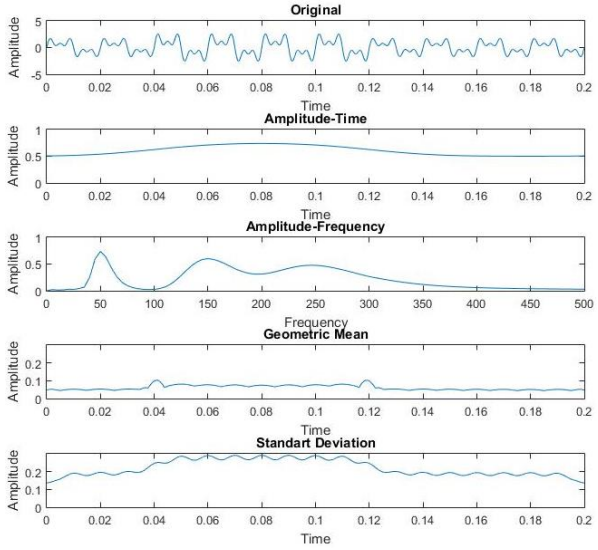


Figure 16. Properties obtained from the S-transform of the 3rd and 5th harmonic added voltage swell signal

As it can be seen in the graph of the geometric feature that helps us to determine the interval where the voltage starts to rise and ends, it started in the 3rd period and ended in the 6th Period.

In the standard deviation graph, the observed fluctuations reveal the presence of harmonics in the signal.

3. Test and Discussion

3.1. Classification of Signals with Support Vector Machines

Eight types of signal whose properties are obtained by using S-transform is classified with support vector machines. Classification is made using the Quadratic SVM and one-to-one method in the Classification Learner Toolbox in MATLAB. Table 1 shows the labels of the disturbances.

Table 1. Class of PQ Events

PQD	Class Label
Pure Sine	Distrubance1
Voltage Sag	Distrubance2
Voltage Swell	Distrubance3
Flicker	Distrubance4
Transient	Distrubance5
Voltage with Harmonics	Distrubance6
Voltage Sag with Harmonics	Distrubance7
Voltage Swell with Harmonics	Distrubance8

In Table 2, the success numbers and success percentages of the classification test data performed with SVM are given. Confusion matrix is given in Figure 17.

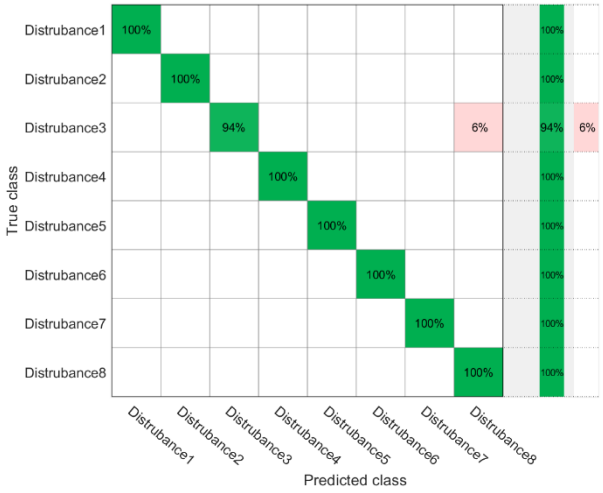


Figure 17. Confusion matrix for SVM.

Table 2. Numbers and percentages of success as a result of training

Classes	Total Tested	Successful	Unsuccessful	Performance Percentage
Distrubance1	16	16	0	%100
Distrubance2	16	16	0	%100
Distrubance3	16	15	1	%94
Distrubance4	16	16	0	%100
Distrubance5	16	16	0	%100
Distrubance6	16	16	0	%100
Distrubance7	16	16	0	%100
Distrubance8	16	16	0	%100

Total performance was 99.2%. Pure sine, voltage sag, flicker, voltage with harmonics, voltage sag with harmonics and voltage swell with harmonics can be estimated as 100%. Only one of the test data is perceived as voltage with harmonics while voltage swell.

3.2. Classification of Signals with Artificial Neural Networks

In this study, the architecture of the proposed ANN is selected from the MATLAB-Neural Network Toolbox and consists of an input layer with 900 inputs, a hidden layer with 10 neurons, and an output layer with a neuron as shown in Figure 18. The number of neurons and hidden layers depends on the problem and is determined by trial and error until a target performance is achieved (Greche et al., 2017).

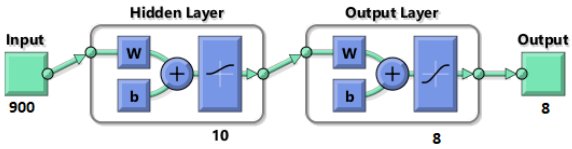


Figure 18. Two-layer feed-forward neural network architecture

The formulated network is trained with a total of 640 samples of 80 different samples per 8 disturbance classes. The number of samples of the ANN classifier for training, validation and testing is randomly selected and is given in Table 3.

Table 3. Number of samples for training, validation and testing

Method	%Samples	Sample number
Training	70	448
Validation	15	96
Testing	15	96
Method	%Samples	Sample number

As the training algorithm, Scaled Conjugate Gradient back propagation training algorithm is chosen. The reason is that this algorithm takes up less memory. The training automatically stops when generalization stops healing, as shown by the increase in error (MSE) in the mean square root of validation samples.

In Figure 19, 100% recognition performance in the confusion matrix of training, validation and test indicates that training performs well. ANN is trained with 100% accuracy and tested with 100% accuracy. This ratio is considered to be quite successful and sensitive. When the effectiveness of the applied method and the success of the results are taken into consideration, it can be concluded that the aim of this study has been achieved.

If the results are obtained less than 100%, during the training process, each layer's weight can be updated by a function called optimizer. The optimizer shows better performance in computational efficiency with advantages such as aigh precision solution and quick convergence (Ruder, 2016).

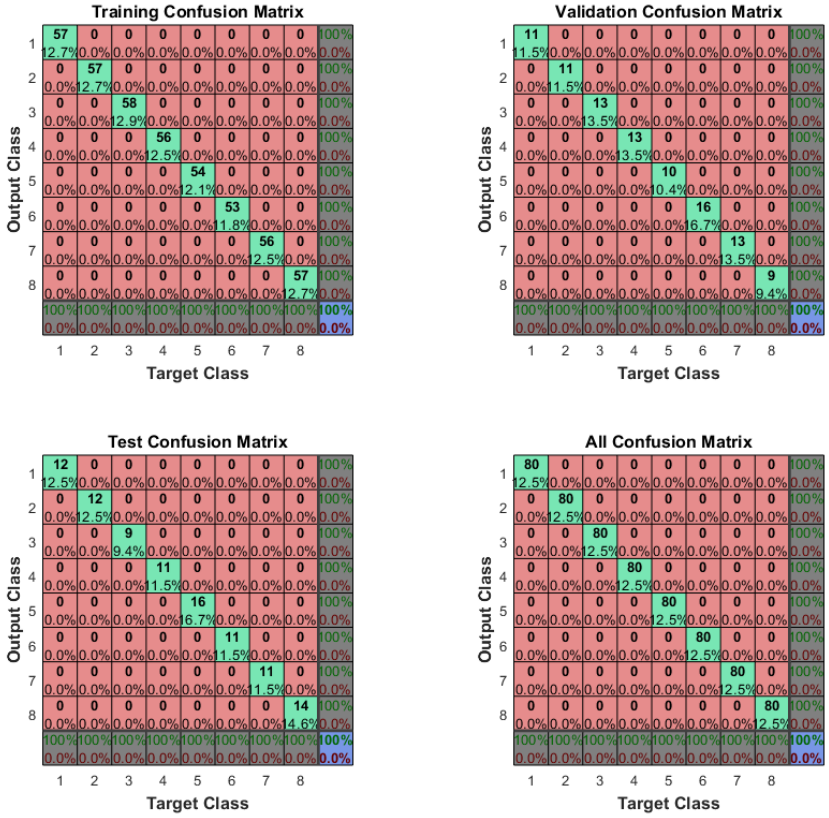


Figure 19. Confusion Matrix for ANN

4. Conclusions

In this study, seven signal distortions have been produced in MATLAB environment such as voltage sag, voltage swell, voltage with harmonics, transient, flicker, voltage sag with harmonics, voltage swell with harmonics for 1 signal of amplitude and frequency of 50 Hz. The stockwell transform of a total of 8 signals was taken with the pure sine signal. A total of 4 different properties were obtained from these signals from the Stockwell transform. These; time-dependent amplitude changes of the signal, amplitude of frequency components, geometric mean and standard deviation. The graphs of these features are presented and their roles in determining each signal type are interpreted. Changes in the amplitudes of the signals are determined from the time dependent amplitude property. The frequencies of the different components in the signal are determined from the frequency amplitudes. The properties of the changes in the amplitude of the signal are determined by using the geometric mean. In the standard deviation feature, periodic distortions such as harmonics are observed.

In the case of voltage sag and voltage swell disturbances, inferences are made from the amplitude-time graph about the voltage sag and swell. In addition, the geometric mean graph makes inferences about the time when fall and rise begin and end. In the flicker disturbance signal, the 10 Hz component is determined from the amplitude frequency graph. In addition, the voltage drops at the points where the flicker occurs can be observed from the amplitude time graph. When flicker occurs, it is determined by the geometric mean. In the transient distortion signal, besides the 50 Hz signal, the 4 kHz signal can be observed on the amplitude-frequency graph. As with other signals, the voltage variation of the signal and the location of the change can be observed in the amplitude-time and geometric average graphs. Unlike other disturbances in the voltage with harmonics signal, there is a fluctuation in the standard deviation graph, which is used to detect the harmonics in the signal. In addition, 50 Hz signal and 150 Hz and 250 Hz harmonic frequencies can be observed in the amplitude frequency graph. In contrast to the voltage with harmonic signal, the voltage decreases or increase in the amplitude-time graph and the points where the voltage starts to decrease and starts to increase in the voltage sag with harmonics and voltage swell with harmonics are determined.

High quality results can be obtained if training and testing is performed by using the classification methods obtained from this transformation. These properties are classified with SVM and ANN. As a result of the classification, decays can be estimated with 99.54% SVM and 100% ANN. Thus, both methods are successful for a strong feature extraction, and ANN gives a much better result.

Compared to other feature extraction methods, S-Transform contains both frequency and time information. This offers a great advantage in the feature extraction of the signals.

References

Agarwal, R. K., Hussain, I., Singh, B., 2017. Application of LMS-based NN structure for power quality enhancement in a distribution network under abnormal conditions. IEEE transactions on neural networks and learning systems, 29(5), pp. 1598-1607.

Azam, M. S., Tu, F., Pattipati, K. R., Karanam, R., 2004. A dependency model-based approach for identifying and evaluating power quality problems. IEEE Transactions on power delivery, 19(3), pp. 1154-1166.

Chilukuri MV, Dash PK., 2004. Multiresolution S-transform-based fuzzy recognition system for power quality events. IEEE Trans Power Delivery. 19(1), pp. 323-330.

Choudhary, B. (2021). An advanced genetic algorithm with improved support vector machine for multi-class classification of real power quality events. Electric Power Systems Research, 191, 106879.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp. 273-297.

Dharavath, R., Raglend, I. J., Manmohan, A., 2017. Implementation of solar PV—Battery storage with DVR for power quality improvement. In 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 1-5.

Elango, M. K., Loganathan,K., 2016.Classification of power quality disturbances using Stockwell Transform and Back Propagation algorithm. Emerging Technological Trends (ICETT), International Conference on. IEEE.

Gaing, Z. L., 2004. Wavelet-based neural network for power disturbance recognition and classification. IEEE transactions on power delivery, 19(4), pp. 1560-1568.

Greche, L., Es-Sbai, N., Lavendelis, E., 2017. Histogram of oriented gradient and multi-layer feed forward neural network for facial expression identification. In 2017 International Conference on Control, Automation and Diagnosis (ICCAD), pp. 333-337.

Ingale, R., 2014. Harmonic analysis using FFT and STFT. International Journal of Signal Processing, Image Processing and Pattern Recognition, 7(4), pp. 345-362.

Karasu, S., Başkan, S., 2016. Classification of power quality disturbances by using ensemble technique. In 2016 24th Signal Processing and Communication Application Conference (SIU), pp. 529-532.

Liang, C., Teng, Z., Li, J., Yao, W., Wang, L., He, Q., Hu, S., 2021. Improved S-Transform for Time-Frequency Analysis for Power Quality Disturbances. IEEE Transactions on Power Delivery.

Mahela, O. P., Shaik, A. G., 2016. Recognition of power quality disturbances using S-transform and rule-based decision tree. In 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), pp. 1-6.

Mishra, A. K., Ray, P. K., Mallick, R. K., Mohanty, A., & Das, S. R., 2021. Adaptive fuzzy controlled hybrid shunt active power filter for power quality enhancement. Neural Computing and Applications, 33(5), pp. 1435-1452.

Ozgonenel, O., Yalcin, T., Guney, I., Kurt, U., 2013. A new classification for power quality events in distribution systems. Electric Power Systems Research, 95, pp. 192-199.

Poisson, O., Rioual, P., Meunier, M., 2000. Detection and measurement of power quality disturbances using wavelet transform. IEEE transactions on Power Delivery, 15(3), pp. 1039-1044.

Raj, S., Phani, T. K., Dalei, J., 2016. Power quality analysis using modified S-transform on ARM processor. In 2016 Sixth International Symposium on Embedded Computing and System Design (ISED) (pp. 166-170). IEEE.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Saxena, D., Singh, S. N., Verma, K. S., Singh, S. K., 2014. HHT-based classification of composite power quality events. International Journal of Energy Sector Management.

Shamachurn, H., 2019. Assessing the performance of a modified S-transform with probabilistic neural network, support vector machine and nearest neighbour classifiers for single and multiple power quality disturbances identification. Neural Computing and Applications, 31(4), pp. 1041-1060.

Sindi, H., Nour, M., Rawa, M., Öztürk, Ş., Polat, K. (2021). An adaptive deep learning framework to classify unknown composite power quality event using known single power quality events. Expert Systems with Applications, 178, 115023.

Singh, B., Al-Haddad, K., Chandra, A., 1999. A review of active filters for power quality improvement. IEEE transactions on industrial electronics, 46(5), pp. 960-971.

Singh, U., Singh, S. N., 2017. Application of fractional Fourier transform for classification of power quality disturbances. IET Science, Measurement & Technology, 11(1), pp. 67-76.

Tao, W., Yin, S., Ding, M., Li, C., Yu, N., Bao, X., Guo, J., 2013. Classification of power quality disturbance signals based on S-transform and HHT. In Proceedings of the 32nd Chinese Control Conference, pp. 3639-3644.

Thirumala, K., Prasad, M. S., Jain, T., Umarikar, A. C., 2016. Tunable-Q wavelet transform and dual multiclass SVM for online automatic detection of power quality disturbances. IEEE Transactions on Smart Grid, 9(4), pp. 3018-3028.

Wang, S., Chen, H., 2019. A novel deep learning method for the classification of power quality disturbances using deep convolutional neural network. Applied energy, 235, pp. 1126-1140.

Yoo, J. H., Shin, S. K., Park, J. Y., Cho, S. H., 2015. Advanced railway power quality detecting algorithm using a combined TEO and STFT method. Journal of Electrical Engineering and Technology, 10(6), pp. 2442-2447.

Zhao, Z., Wang, S., Zhang, W., Xie, Y., 2016. A novel automatic modulation classification method based on Stockwell-transform and energy entropy for underwater acoustic signals. In 2016 IEEE international conference on signal processing, communications and computing (ICSPCC), pp. 1-6.

Research Article
Journal of Intelligent Systems: Theory and Applications 5(1) 2022: 85-91
DOI: 10.38016/jista.954098

Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches

Fatma BOZYİĞİT[1*], Onur DOĞAN[2], Deniz KILINÇ[3],

[1,3] Bakırçay University, Faculty of Engineering, Department of Industrial Engineering, İzmir/Turkey

[2] Bakırçay University, Faculty of Engineering, Department of Computer Engineering, İzmir/Turkey

fatma.bozyigit@bakircay.edu.tr, onur.dogan@bakircay.edu.tr, deniz.kilinc@bakircay.edu.tr

Abstract
Customer feedback is one of the most critical parameters that determine the market dynamics of product development. In this direction, analyzing product-related complaints helps sellers to identify the quality characteristics and consumer focus. There have been many studies conducted on the design of Machine Learning (ML) systems to address the causes of customer dissatisfaction. However, most of the research has been particularly performed on English. This paper contributes to developing an accurate categorization of customer complaints about package food products, written in Turkish. Accordingly, various ML algorithms using TF-IDF and word2vec feature representation strategies were performed to determine the category of complaints. Corresponding results of Linear Regression (LR), Naive Bayes (NB), k Nearest Neighbour (kNN), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) classifiers were provided in related sections. Experimental results show that the best-performing method is XGBoost with TF-IDF weighting scheme and it achieves %86 F-measure score. The other considerable point is word2vec based ML classifiers show poor performance in terms of F-measure compared to the TF-IDF term weighting scheme. It is also observed that each experimented TF-IDF based ML algorithm gives a more successful prediction performance on the optimal subsets of features selected by the Chi Square (CH2) method. Performing CH2 on TF-IDF features increases the F-measure score from 86% to 88% in XGBoost.

Keywords: customer complaints, complaint categorization, food industry, machine learning.

1. Introduction

Today's customers are more willing to complain about the products and services they purchase (Mahayudin et al., 2010). Organizations need to know and handle the increasing threat of online public complaints (Tripp and Gregoire, 2011). Academic and sector professionals regard customer satisfaction and complaints as critical input for companies' success in a competitive environment (Pinto and Mansfield, 2012). Besides, some researchers recognized successful complaint solving as a competing benefit (Fox, 2008).

In recent years, companies have collected customer complaints via social media platforms such as websites, Twitter, forums, and blogs, by benefiting from developing technology (Tax et al., 1998). Because social media enables businesses to communicate with their customers and get feedback from them (HaCohen-Kerner et al. 2019). Some organizations combine social media in their complaint handling process (Jin et al., 2013). These companies' customers are encouraged to participate in a survey by a direct message or email from one of the social media platforms after a service experience. One of the most popular social media platforms, Facebook, was mainly preferred by Generation Y (people born between 1981 and 1996) to express their complaints (Rossmann et al., 2017). Even though these surveys enable companies to receive feedback on purchased products and services with star ratings or points (Sohail et al., 2016; Bhole and Hanna, 2017). Customer complaint management has become an essential factor in optimizing the relationship between clients and companies in recent years. However, manual

* Corresponding Author.
E-mail: fatma.bozyigit@bakircay.edu.tr

Recieved : 18 June 2021
Revision : 31 Aug 2021
Accepted : 04 Oct 2021

analysis of complaints is ineffective and time-consuming, due to the notable rise in complaints in digital platforms. In this direction, more researchers have recently focused on the categorization of customer feedback using Machine Learning (ML) algorithms to handle complaints efficiently. Since many customer complaints are in textual form, techniques collected under Text Mining (TM) become necessary to handle the texts' implicit structure before execution of ML algorithms. TM includes two basic steps, such as pre-processing and creating feature sets for data representation. There are two sorts of research commonly applied for text representation, indexing and term weighting (Harish et al., 2010).

In this study, the customer complaints about package food products were categorized using Logistic Regression (LR) (Wright, 1995), Naive Bayes (NB) (Berrar, 2018), Support Vector Machine (SVM) (Noble, 2006), k Nearest Neighbour (kNN) (Peterson, 2009), Random Forest (RF) (Bioau&Scarnet, 2016), and Extreme Gradient Boosting (XGBoost) (Chen et al., 2015) ML algorithms. To the best of our knowledge, this study is the first attempt to analyze customer complaints for the food industry in Turkish text. It contributes to the literature by experimenting different ML classifiers with different feature vectorization strategies that categorizes the customer complaints written in Turkish. It applies TF-IDF and word2vec text representation methods and then uses the feature reduction technique since the high dimensional training space is produced after pre-processing textual input for ML algorithms. Then, it compares ML classifiers whose parameters are tuned with the use of grid-search algorithm (Kılınç et al., 2016). The remaining parts of the study are organized as follows. Section 2 discusses related works. Section 3 gives dataset, data pre-processing steps, feature engineering and ML methods used in the study. Section 4 presents the details of the experimental study with metrics and results. Finally, Section 5 concludes the study and Section 6 gives information about the threats of validity.

2. Related Works

Intensifying competition and developing technology force businesses to manage customers' complaints. Management of complaints is an effective tool in identifying shortcomings in service quality and creating customer loyalty. More researchers have recently addressed the analysis of customer comments for efficient complaint handling in this respect. For example, Hong and Wang (2021) proposed a framework to summarize customer opinions, including both positive and negative comments, from product reviews using neural networks. The effectiveness of the framework was tested with six datasets from real-world business scenarios. In other study, Chen et al. (2021) identified the affecting factors of customers satisfaction from unstructured online comments by lessening personal communication to collect these reviews. Lee and Choi (2020) studied public environmental complaints to investigate the factors that contribute to reducing ecosystem benefits. They performed statistical and spatial analyses on the complaints received by the Namyangju government. Finally, the citizens' comments were categorized as water, air pollution, electricity problems, etc. In another study, Onan et al. (2020) presented the categorization of service support requests using basic ML algorithms (NB, kNN, RF, C4.5, and SVM) on the dataset, including 17831 bug reports and service support requests. Before experimenting with classifiers, they built a TF-IDF scheme for feature representation. The experimental results showed that the classifiers achieved encouraging results in directing support requests to related services. Krishna et al. (2019) performed sentiment analysis of bank customers using the respective banks' online complaints platforms. They experimented with SVM, NB, LR, Decision Tree (DT), kNN, RF, XGBoost, and Multi-layer Perceptron (MLP) classifiers on data generated of TF-IDF, word2Vec and Linguistic Inquiry and Word Count (LIWC) vectors. Eventually, complaints were labelled as moderate or extreme. Experimental results indicated that the LIWC based RF and NB techniques achieved the best accuracy. The study of Stoica and Özyirmidokuz (2015) aimed to get meaningful data from unstructured customer feedbacks about a telecommunication firm in Turkey. After text processing techniques, k-Medoid was used to cluster documents according to the relevant categories. It was stated that the proposed method is advantageous since similar responses in a cluster could be answered by similar response mails.

Considering the literature, there are very few studies empirically examine the impact of customer complaints in food industry. Lemos et al. (2018) conducted a study to analyze the comments of moldy foods within the expiration date. In another study, Khan et al. (2013) analyzed customer complaints about fast-food products within the service, environmental conditions, price, and taste factors. Comments of four well-known fast-food brands were analyzed using multiple regression and correlation tests to determine which factors have a more critical impact on consumer satisfaction.

3. Materials and Methods

3.1. Dataset

The dataset which was obtained from a packaged food supplier in Turkey, includes 2217 customer complaints obtained through call center, e-mail, web pages, and social media (Facebook, Şikayetvar, etc.). The records were annotated as “unhygienic”, “foreign body”, “texture”, “package/label”, and “taste/smell” (Table 1). The language of the complaints in the dataset is Turkish.

Table 1. Sample complaints in the dataset (TR: Turkish, EN: English)

Complaint ID	Description	Category
1	**TR:** Nutella kavanozunun ağzı kırık çıkmış, içinde cam kırıkları varmış. Ürünü markete iade edecek. **EN:** Nutella jar's mouth was broken. There were glass breaks in it. He will return the product to the market.	Package/Label
2	**TR:** Daha önce de şikayette bulunmuş. xxxxx'nın tadı çok kötü bütün gramajları denedim hepsi aynı çocukluğumun nutellası değil dedi. Değiştirilmesini istiyor. **EN:** He had complained before. He said that he tried all xxxxx sizes, and they did not taste all that great. He wants it replaced.	Taste/Smell
3	**TR:** Ürün içerisinden kırmızı renkte plastik benzeri bir madde çıkmış. Çocuğum onu yutacaktı neredeyse dedi. **EN:** There was a red-colored object like plastic. He said that his child was almost going to swallow it.	Foreign body
4	**TR:** Merhaba, müşteri nutella kavonuzunda ilk aldıklarında kavanoz kapağının altında buhar bulunduğunu ve beyazımsı bir görüntü olduğunu söyledi. **EN:** Hello, the customer said that there was a steam under the jar lid, and it had a whitish appearance.	Texture
5	**TR:** xxxxx'nın dibinden son çatal aldım ekmeğe sürmek üzereyken içinde sinek olduğunu fark ettim. **EN:** I got the last bit of xxxxx from the jar, and when I was about to rub it on the bread, I realized that there was a fly in it.	Unhygenic

3.2. Data pre-processing

Text pre-processing makes raw textual data more useful for subsequent analysis. It forms the raw data by removing non-alphanumeric characters, numbers, punctuation marks, and stop words. In pre-processing, stemming is another important step that provides mapping of connected words into a base form.

In this study, first tokenization was performed to split phrases into tokens (meaningful elements) such as words, numbers, punctuation marks. Then, the customer complaints were evaluated in terms of spelling errors since there were many spelling mistakes in the unprocessed dataset. Therefore, Zemberek normalization module developed by Akın (2007) was employed on the raw data for pre-processing noisy text inputs. The stopwords which refers to commonly used words in a language (e.g., "and", "so", "the" in English) were filtered. Finally, the stemming was performed to reduce unnecessary diversity in the feature vectors. Figure 1 illustrates the pre-processing steps applied in this study.

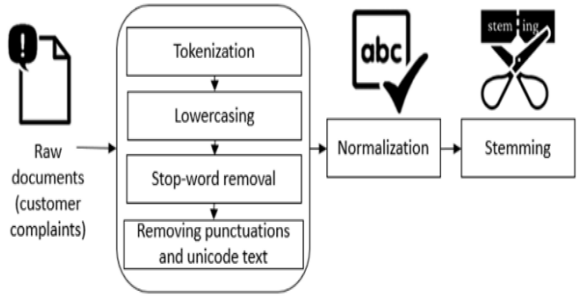


Figure 1. Set of applied pre-processing steps

3.3. Data representation and feature engineering

Unstructured textual data is challenging to process and needs to be described by term sets to represent their contents. The vector space model (Salton and Yang, 1973) is one of the most used text representation models to a host of information retrieval operations. This model also appeals to the underlying metaphor of practicing spatial proximity for semantic proximity (Zhang et al. 2011). There are two sorts of research commonly applied for text representation: indexing and term weighting. Indexing assigns indexing terms for documents, whereas term weighting assigns each term's weight to show its importance. This study uses word2vec method for indexing and the TF-IDF method to calculate each word's weights in the customer complaints.

Word2vec is a method of embedding words in a high-dimensional space. After an external neural network is trained for the word embedding, terms in the document are classified according to their similarities in the word2Vec space. There are two models for representing words in a multidimensional vector space namely skip-gram and Continuous Bag of Words (CBOW) (Onishi&Shina, 2020). In the skip-gram model, the surrounded representations of a context are predicted using the centre word. CBOW model predicts a target word by combining the distributed representations in its context. Due to their simple architecture, skip-gram and CBOW can be trained on a large dataset in a short time. The ability to train on very large datasets allows the model to learn complex word relationships such as vec(Turkish) + vec(food) ≈ vec(kebab).

TF-IDF is obtained by multiplying the term frequency (TF) and inverse document frequency (IDF) for a term in the text. While TF gives the occurrence frequency of a word in the document, the value IDF indicates this term's occurrence frequency in other documents. The main idea in TF-IDF is to classify terms as much as possible into the same category considering their high appearance in one document and high absence in other documents. When a term appears with a high TF frequency in a text document and rarely appears with low IDF frequency in other documents, it is accepted that the term has a good classification accuracy.

Since high dimensionality of textual data imposes high costs on model training and execution, feature set reduction on document representations may be necessary to optimize performance of ML algorithms. In addition to eliminate unnecessary terms, Feature Selection (FS) strategies also provide better model understand-ability, increase generalization capability of the model, and decrease in over-fitting risk (Tunalı&Bilgin, 2012). There are mainly three categories of FS strategies: filter, wrapper, and embedded methods. The most widely used methods in these groups are Information Gain (IG), Chi-Square (CH2), and Correlation Feature filtering (CF) (Özçift et al., 2019). These methods use a metric such as correlation, entropy, and mutual information to obtain the most valuable subset. In particular, CH2 filtering approach covers the relationship between two events (Howell, 2011). The filter tests the occurrence of specific word and occurrence of a complaint class to be independent or not. The rank of selected feature *t* for category c_i, x^2, is calculated using Equation 1.

$$x^2(t,c) = \frac{N \times (AD - BC)}{(\mathrm{A} + \mathrm{C})(\mathrm{B} + \mathrm{C})(\mathrm{A} + \mathrm{B})(\mathrm{C} + \mathrm{D})} \quad (1)$$

where *A* is observed frequency of *t* when it is included by category *c*, N is the total number of documents, *B* is the observed frequency of *t* when it is not included by category *c*, *C* is the observed frequency of *c* when it occurs without *t*, *D* is the number of documents that do not involve *t* and *c*.

3.4. Baseline machine learning (ML) algorithms

After the pre-processing and feature selection steps were utilized, some baseline ML algorithms, which are commonly used to classify the textual data, were performed.

Logistic Regression (LR): LR is used to analyze a data set within one or more independent features determining output class. It assigns a new sample to one of the determined discrete classes by utilizing a logistic function. Logistic regression is a statistical method used to analyze a data set within one or more independent features determining a result.

Naive Bayes (NB): NB depends on the common principle of Bayes Theorem, i.e., a distinct feature in a class is independent of any other feature's presence. It describes the probability of an event, based on prior knowledge of conditions using Equation 2.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

where P(A) and P(B) are prior probabilities and P(B|A) and P(A|B) are posterior probabilities of event A and B, respectively (Bozyiğit et al., 2019).

k-Nearest Neighbor (kNN): kNN is an instance-based ML algorithm that assigns a new sample's class according to the majority classes of its most similar *k* neighbours. There are mainly four distance metrics such as Euclidean, Manhattan, Minkowski, and Hamming to determine the k nearest neighbours of an instance to be classified.

Support Vector Machine (SVM): SVM aims to find a hyperplane that can separate the two classes of given samples by a maximal margin. The margin corresponds to the shortest distance between the nearest data points and any point on the hyperplane. The ability to generalize of SVM ensures a high classification accuracy.

Random Forest (RF): RF algorithm builds a multitude of individual decision trees using different training subsets (Bozyiğit et al., 2020). Each tree in the forest gives an output and the final class is determined by the majority vote of them.

Extreme Gradient Boosting (XGBoost): XGBoost is a gradient boosting framework, including an efficient linear model solver and tree learning algorithm. It maintains customized functions so that users are also allowed to define their objectives and evaluation easily. Its most important features are its ability to obtain highly successful predictive results, prevent over-learning, and manage null and noisy records (Kılınç et al., 2015).

4. Experimental Study

In this study, customer complaints about food products delivered to markets in Turkey were categorised under five categories such as Hygiene, Foreign body, Taste/Smell, Texture, and Package/Label. Figure 2 shows the general flow of complaint categorization task by using ML algorithms. First data preparation was utilized to transform the raw data in a useful and efficient format. After pre-processing, two feature representation models (TF-IDF and word2vec) were applied to observe their effects on categorization accuracy. Then, CH2 feature selection strategy was performed on experimented TF-IDF based ML methods to obtain best subset of features. The evaluation results of each ML method were obtained by dividing the data set into 10 pieces by cross-validation.

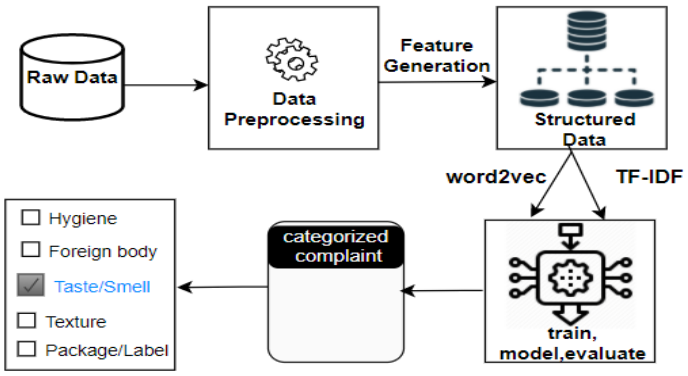


Figure 2. General flow of complaint categorization task by using ML algorithms

4.1. Experimental results

In the experimental studies, the real-world data including customer complaints were used to perform categorization task. Table 2 compares the F-measure scores obtained by performing ML algorithms with TF-IDF and word2vec (CBOW and skip-gram) representations. Considering the experimental results, the highest F-measure values were achieved by XGBoost classifier in all feature representations. On the other hand, NB classifier with both two representations had the lowest F-measure values among all classifiers.

Table 2. Evaluation of different ML algorithms and feature representations

ML algorithm	Feature representations		Precision	Recall	F-measure
LR	TF-IDF		0.80	0.83	0.81
	word2vec	skip-gram	0.66	0.64	0.67
		CBOW	0.54	0.48	0.46
NB	TF-IDF		0.75	0.71	0.73
	word2vec	skip-gram	0.53	0.54	0.53
		CBOW	0.62	0.63	0.62
kNN	TF-IDF		0.80	0.82	0.81
	word2vec	skip-gram	0.56	0.58	0.57
		CBOW	0.57	0.62	0.59
SVM	TF-IDF		0.81	0.81	0.81
	word2vec	skip-gram	0.69	0.74	0.71
		CBOW	0.59	0.66	0.62
RF	TF-IDF		0.82	0.81	0.81
	word2vec	skip-gram	0.53	0.54	0.53
		CBOW	0.62	0.63	0.62
XGBoost	**TF-IDF**		**0.83**	**0.84**	**0.84**
	word2vec	skip-gram	0.62	0.67	0.64
		CBOW	0.76	0.75	0.75

Another point to be noticed is that the ML algorithms with TF-IDF encoding method performed better than ones with the word2vec method (see Figure 3). The poor performance of word2vec representation can probably be based on the limited training data. To increase the accuracy of the word2vec, Pyhton NLP Aug library presenting augmenting for textual data was experimented. However, there was no significant effect on the word2vec performance since NLP Aug library did not perform well on Turkish texts. XGBoost with TF-IDF was evaluated as the most accurate classifier with 84% F-measure value. The closest performance result to XGBoost was achieved by SVM with TF-IDF and RF with TF-IDF with the 81% F-measure value. On the other hand, NB was the worst performing algorithm with 73% F-measure score among classifiers using TF-IDF representation model. There exists adequate evidence to show that the feature selection technique, CH2, improves the categorization accuracy by eliminating extraneous and unnecessary terms from the dataset. Figure 3 shows that the F-measure value increases in all classifiers after CH2 is applied on TF-IDF feature sets. It is obviously seen that TF-IDF feature with dimensionality reduction provides an improvement in prediction accuracy for all experimented classifiers. Performing CH2 on TF-IDF features achieved a statistically significant F-measure scores and increased it from 84% to 88% in XGBoost.

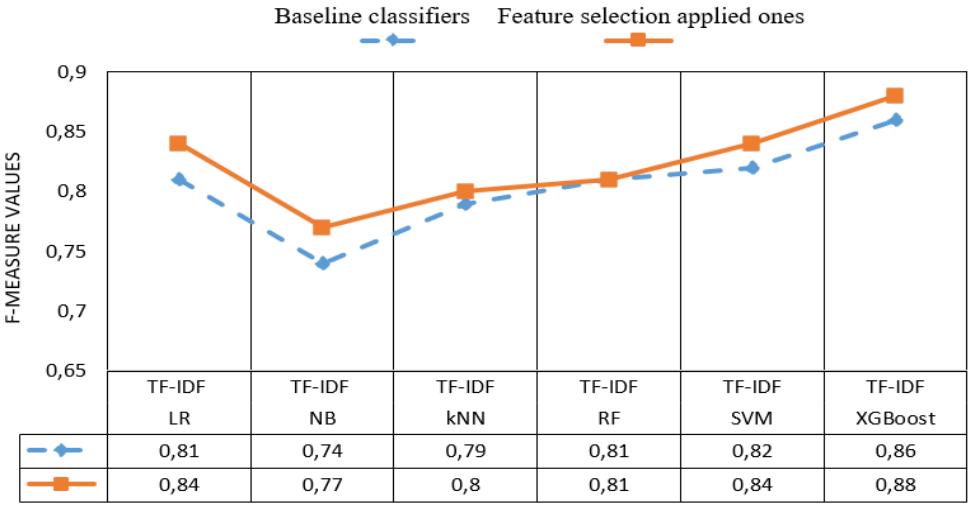


Figure 3. Performance results of the experimented methods after CH2

5. Conclusion

Customer satisfaction is specified as a primary factor for business success according to basic marketing theory. The benefits of the automatic complaint categorization are reducing the initial cost of labeling the complaints with the most appropriate tag, helping maintenance and keeping the efficient process for directing customer complaints to relevant departments, and removing the risk of depending on experts in the management of customer feedback. In this direction, feedback of consumers about the products are analyzed to determine possible problems and effective strategies to handle them. Analysis of customer comments can be challenging task for a human because it may be necessary to analyse high volume data during long time periods. An alternative is to automatically categorize the causes of customer dissatisfaction. In this study, six ML classifiers (LR, NB, kNN, SVM, RF, and XGBoost) with TF-IDF and word2vec feature representations were experimented to categorize customer complaint for food industry in Turkey. Accordingly, the complaints were categorized into "unhygenic", "foreign body", "texture", "package/label", and "taste/smell" categories. Experimental results showed that the best-performing method is XGBoost with TF-IDF weighting scheme and it achieved %86 F-measure score. The other considerable point is that word2vec based ML classifiers showed poor performance in terms of F-measure comparing to TF-IDF term weighting scheme. Since the performance of word2vec is directly related to training sample size, the data augmentation technique to generate new data was utilized to handle this problem. However, available Python libraries to produce synthetic data did not provide expected increase of F-measure. It was also observed that each experimented TF-IDF based ML algorithms showed more successful prediction performance on the optimal subsets of features selected by CH2 method. Performing CH2 on TF-IDF features increased F-measure score from 86% to 88% in XGBoost. Considering the results, it can be concluded that this study appears promising for future studies on complaint handling systems.

6. Threats to Validity

This section addresses the threats to validity that might have affected the complaint classification using the classification of Wohlin et al. (2000). The main threat to validity is the limited customer complaints analyzed in the experimental study. We examined only 2217 records obtained through call centres, e-mail, web pages, and social media platforms. For the future works, we will create large sized and comprehensive dataset to obtain better classification performances.

References

Akın, A. A., & Akın, M. D. (2007). Zemberek, an open source nlp framework for turkic languages. Structure, 10, 1-5.

Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands, 403-412.

Bhole, B., & Hanna, B. (2017). The effectiveness of online reviews in the presence of self-selection bias. Simulation Modelling Practice and Theory, 77, 108-123.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

Bozyiğit, A., Utku, S., & Nasiboğlu, E. (2019). Cyberbullying detection by using artificial neural network models. In 2019 4th International Conference on Computer Science and Engineering (UBMK) (pp. 520-524). IEEE.

Bozyiğit, F., Şahin, M., Gündüz, T., Işık, C., & Kilinç, D. (2020) Regression based risk analysis in life insurance industry. International Journal of Engineering and Innovative Research, 2(3), 178-184.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Chen, W. K., Riantama, D., & Chen, L. S. (2021). Using a Text Mining Approach to Hear Voices of Customers from Social Media toward the Fast-Food Restaurant Industry. Sustainability, 13(1), 268.

Fox, G. L. (2008). Getting good complaining without bad complaining. Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, 21, 23.

HaCohen-Kerner, Y., Dilmon, R., Hone, M., & Ben-Basan, M. A. (2019). Automatic classification of complaint letters according to service provider categories. Information Processing & Management, 56(6), 102102.

Harish, B. S., Guru, D. S., & Manjunath, S. (2010). Representation and classification of text documents: A brief review. IJCA, Special Issue on RTIPPR (2), 110-119.

Haron, S. A., & Fah, B. C. Y. (2010). Unpleasant market experience and consumer complaint behavior. Asian Social Science, 6(5), 63.

Hong, M., & Wang, H. (2021). Research on customer opinion summarization using topic mining and deep neural network. Mathematics and Computers in Simulation, 185, 88-114.

Howell, D. C. (2011). Chi-Square Test: Analysis of Contingency Tables.

Jin, J., Yan, X., Yu, Y., & Li, Y. (2013). Service failure complaints identification in social media: A text classification approach. In: Proceedings of the 34th International Conference on Information Systems.

Kılınç, D., Borandağ, E., Yücalar, F., Özçift, A., & Bozyiğit, F. (2015). Yazılım hata kestiriminde kolektif sınıflandırma modellerinin etkisi. Proceedings of IX. Ulusal Yazılım Mühendisliği Sempozyumu. İzmir, Turkey: Yaşar Üniversitesi.

Kılınç, D., Yücalar, F., Borandağ, E., & Aslan, E. (2016). Multi-level reranking approach for bug localization. Expert Systems, 33(3), 286-294.

Krishna, G. J., Ravi, V., Reddy, B. V., Zaheeruddin, M., Jaiswal, H., Teja, P. S. R., & Gavval, R. (2019). Sentiment Classification of Indian Banks' Customer Complaints. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 429-434). IEEE.

Lee, J. H., & Choi, H. (2020). An analysis of public complaints to evaluate ecosystem services. Land, 9(3), 62.

Lemos, J. G., Garcia, M. V., de Oliveira Mello, R., & Copetti, M. V. (2018). Consumers complaints about moldy foods in a Brazilian website. Food Control, 92, 380-385.

Noble, William S. “What is a support vector machine?” Nature biotechnology 24.12 (2006): 1565-1567.

Onan, A., Atik, E., & Yalçın, A. (2020). Machine learning approach for automatic categorization of service support requests on university information management system. In International Conference on Intelligent and Fuzzy Systems (pp. 1133-1139). Springer, Cham.

Onishi, T., & Shiina, H. (2020). Distributed Representation Computation Using CBOW Model and Skip–gram Model. In 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 845-846). IEEE.

Özçift, A., Kilinç, D., & Bozyiğit, F. (2019). Application of Grid Search Parameter Optimized Bayesian Logistic Regression Algorithm to Detect Cyberbullying in Turkish Microblog Data. Akademik Platform Mühendislik ve Fen Bilimleri Dergisi, 7(3), 355-361.

Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

Pinto, M. B., & Mansfield, P. (2012). Facebook as a complaint mechanism: An investigation of millennials. Journal of Behavioral Studies in Business, 5, 1.

Rossmann, A., Wilke, T., & Stei, G. (2017). Usage of social media systems in customer service strategies. In Proceedings of the 50th Hawaii International Conference on System Sciences.

Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. Journal of documentation.

Shahzad, K., Majid, H. S., & Fahad, Y. (2013). Determinants of Customer Satisfaction in Fast Food Industry A Study of Fast Food Restaurants Peshawar Pakistan. Studia commercialia Bratislavensia, 6(21), 56-65.

Sohail, S. S., Siddiqui, J., & Ali, R. (2016). Feature extraction and analysis of online reviews for the recommendation of books using opinion mining technique. Perspectives in Science, 8, 754-756.

Stoica, E. A., & Özyirmidokuz, E. K. (2015). Mining customer feedback documents. International Journal of Knowledge Engineering, 1(1), 68-71.

Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: implications for relationship marketing. Journal of marketing, 62(2), 60-76.

Tripp, T. M., & Grégoire, Y. (2011). When unhappy customers strike back on the Internet. MIT Sloan Management Review, 52(3), 37-44.

Tunali, V., & Bilgin, T. T. (2012, June). PRETO: A high-performance text mining tool for preprocessing Turkish texts. In Proceedings of the 13th International Conference on Computer Systems and Technologies (pp. 134-140).

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). Experimentation in software engineering. Springer Science & Business Media.

Wright, R. E. (1995). Logistic regression.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765.