Yıl:2020 Cilt:4 Sayı:2 Year:2020 Vol:4 No:2

UYBİSBBD

ULUSLARARASI YÖNETİM BİLİŞİM SİSTEMLERİ VE BİLGİSAYAR BİLİMLERİ DERGİSİ

ULUSLARARASI YÖNETİM BİLİŞİM SİSTEMLERİ VE BİLGİSAYAR BİLİMLERİ DERGİSİ

INTERNATIONAL JOURNAL OF MANAGEMENT INFORMATION SYSTEMS AND COMPUTER SCIENCE

Cilt: 4 ● Sayı: 2 ● Aralık 2020
Vol: 4 ● No: 2 ● December 2020

e-ISSN: 2618 - 5954

ULUSLARARASI YÖNETİM BİLİŞİM SİSTEMLERİ VE BİLGİSAYAR BİLİMLERİ DERGİSİ

INTERNATIONAL JOURNAL OF MANAGEMENT INFORMATION SYSTEMS AND COMPUTER SCIENCE

Cilt: 4 ● Sayı: 2 ● Aralık 2020
Vol: 4 ● No: 2 ● December 2020

e-ISSN: 2618-5954

E-mail : ybsbb.info@gmail.com
Web : dergipark.gov.tr/uybisbbd

UYBİSBBD, uluslararası hakemli, uluslararası indeksli, açık erişimli bilimsel bir dergidir

ULUSLARARASI YÖNETİM BİLİŞİM SİSTEMLERİ VE BİLGİSAYAR BİLİMLERİ DERGİSİ

INTERNATIONAL JOURNAL OF MANAGEMENT INFORMATION SYSTEMS AND COMPUTER SCIENCE

Adres (Address)
Bandırma Onyedi Eylül Üniversitesi
Gönen Meslek Yüksekokulu
10900 Balıkesir / TÜRKİYE

E-mail : ybsbb.info@gmail.com
Web : dergipark.gov.tr/uybisbbd

HAKEM KURULU	
Prof. Dr. Abdulkadir YILDIZ **(Kahramanmaraş Sütçü İmam Üniversitesi)**	**Dr. Öğr. Üyesi Zülfiye BIKMAZ** **(Kırklareli Üniversitesi)**
Prof. Dr. H. Mustafa PAKSOY **(Kilis 7 Aralık Üniversitesi)**	**Dr. Öğr. Üyesi Ö. Fatih KEÇECİOĞLU** **(Kahramanmaraş Sütçü İmam Üniversitesi)**
Prof. Dr. Mustafa AKSU **(İstanbul Üniversitesi)**	**Dr. Öğr. Üyesi Cuma ERCAN** **(Kilis 7 Aralık Üniversitesi)**
Prof. Dr. Sadettin PAKSOY **(Kilis 7 Aralık Üniversitesi)**	**Dr. Öğr. Üyesi Mustafa Oğuz GÖK** **(Kahramanmaraş Sütçü İmam Üniversitesi)**
Doç. Dr. Deniz Mertkan GEZGİN **(Trakya Üniversitesi)**	**Dr. Öğr. Üyesi Sinan UĞUZ** **(Isparta Uygulamalı Bilimler Üniversitesi)**
Doç. Dr. İlhan UMUT **(Trakya Üniversitesi)**	**Dr. Öğr. Üyesi Muhammet ATALAY** **(Kırklareli Üniversitesi)**
Doç. Dr. İrfan Deli **(Kilis 7 Aralık Üniversitesi)**	**Dr. Öğr. Üyesi Mustafa Mikail ÖZÇİLOĞLU** **(Kilis 7 Aralık Üniversitesi)**
Doç. Dr. Nursal ARICI **(Gazi Üniversitesi)**	**Dr. Öğr. Üyesi Hasan Hüseyin ÇAM** **(Kilis 7 Aralık Üniversitesi)**
Doç. Dr. Yusuf Ekrem AKBAŞ **(Adıyaman Üniversitesi)**	**Dr. Öğr. Üyesi Ebru KÜLEKÇİ AKYAVUZ** **(Kilis 7 Aralık Üniversitesi)**
Prof. Dr. Kemal Delihacıoğlu **(Kilis 7 Aralık Üniversitesi)**	**Dr. Öğr. Üyesi Halil ARSLAN** **(Cumhuriyet Üniversitesi)**
Doç. Dr. Bengü HIRLAK **(Kilis 7 Aralık Üniversitesi)**	**Dr. Öğr. Üyesi Emrah AYDEMİR** **(Ahi Evran Üniversitesi)**
Dr. Öğr. Üyesi Edip Serdar GÜNER **(Kırklareli Üniversitesi)**	**Dr. Öğr. Üyesi Ayşe ÇİÇEK KORKMAZ** **(Bandırma 17 Eylül Üniversitesi)**
Dr. Öğr. Üyesi Hüseyin KOÇARSLAN **(Selçuk Üniversitesi)**	**Dr. Öğr. Üyesi Hüseyin AKAR** **(Kilis 7 Aralık Üniversitesi)**
Dr. Öğr. Üyesi Yasin ORTAKCI **(Karabük Üniversitesi)**	**Dr. Nilüfer VATANSEVER TOYLAN** **(Kırklareli Üniversitesi)**
Dr. Öğr. Üyesi Mehmet ÖZÇALICI **(Kilis 7 Aralık Üniversitesi)**	**Dr. Murat GEZER** **(İstanbul Üniversitesi)**
Dr. Öğr. Üyesi Melda AKBABA **(Kilis 7 Aralık Üniversitesi)**	**Dr. Serra Çelik** **(İstanbul Üniversitesi)**
Dr. Öğretim Üyesi Hayrettin TOYLAN **(Kırklareli Üniversitesi)**	**Dr. Feyzi KAYSİ** **(İstanbul Üniversitesi)**
Dr. Öğr. Üyesi Yasin SÖNMEZ **(Dicle Üniversitesi)**	**Dr. Öğr. Üyesi Hakan AÇIKGÖZ** **(Gaziantep İslam bilim ve Teknoloji Üniversitesi)**
Dr. Öğr. Üyesi Ramazan ASLAN **(Adıyaman Üniversitesi)**	**Dr. Fatma Önay KOÇOĞLU** **(İstanbul Üniversitesi)**
Dr. Öğr. Üyesi Sibel YAŞAR **(Kırklareli Üniversitesi)**	**Dr. Fatih AYDIN** **(Kırklareli Üniversitesi)**
Dr. Öğr. Üyesi Şebnem ÖZDEMİR **(İstinye Üniversitesi)**	**Dr. Öğr. Üyesi Cemal AKTÜRK** **(Gaziantep İslam bilim ve Teknoloji Üniversitesi)**

Dr. Öğr. Üyesi ALİ DURDU **(Ankara Sosyal Bilimler Üniversitesi)**	**Dr. Öğr. Üyesi Tarık TALAN** **(Gaziantep İslam bilim ve Teknoloji Üniversitesi)**
Dr. Öğr. Üyesi Hakan ÜSTÜNEL **(Kırklareli Üniversitesi)**	**Dr. Öğr. Üyesi Ahmet Çağdaş SEÇKİN** **(Adnan Menderes Üniversitesi)**
Dr. Öğr. Üyesi Mehmet BAKIR **(Bozok Üniversitesi)**	**Dr. Emre AKADAL** **(İstanbul Üniversitesi)**
Dr. Öğr. Üyesi Doğan ÜNAL **(Kırklareli Üniversitesi)**	**Dr. Feden KOÇ** **(Uşak Üniversitesi)**
Doç. Dr. Dilek AVCI **(Bandırma 17 Eylül Üniversitesi)**	**Dr. Öğr. Üyesi Oğuzhan AYTAR** **(Karamanoğlu Mehmetbey Üniversitesi)**
Dr. Öğr. Üyesi Kazım SARIÇOBAN **(Mehmet Akif Ersoy Üniversitesi)**	**Dr. Öğr. Üyesi Namık Kemal ERDEMİR** **(Karamanoğlu Mehmetbey Üniversitesi)**
Dr. Öğr. Üyesi Mehmet ARICI **(Gaziantep İslam bilim ve Teknoloji Üniversitesi)**	**Dr. Öğr. Üyesi Mehmet Sait VURAL** **(Gaziantep İslam bilim ve Teknoloji Üniversitesi)**
Dr. Ayşe KARADAŞ **(Balıkesir Üniversitesi)**	**Prof. Dr. Aysun COŞKUN** **(Gazi Üniversitesi)**
Dr. Hava GÖKDERE ÇİNAR **(Bursa Uludağ Üniversitesi)**	**Doç. Dr. Atilla YÜCEL** **(Fırat Üniversitesi)**
Dr. Öğr. Üyesi Emir Hüseyin ÖZDER **(Başkent Üniversitesi)**	**Dr. Öğr. Üyesi Pınar Miç** **(Tarsus Üniversitesi)**

YAYIN POLİTİKASI

Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi yılda iki kez Haziran ve Aralık aylarında yayınlanan uluslararası hakemli bir dergidir. Dergide yer alan yazılar kaynak gösterilmeksizin kısmen ya da tamamen iktibas edilemez. Bu dergide yayınlanan çalışmaların bilim ve dil sorumluluğu yazarlarına aittir.

Dergimize gönderilen çalışmalar, alanında uzman iki ayrı hakem tarafından incelendikten sonra uygun görülenler yayınlanmaktadır. Yazım kurallarına ilişkin bilgilere dergimizin web adresinde yer verilmiştir.

International Journal of Management Information Systems and Computer Science is an international peer-reviewed journal which is published two times a year in June and December. The articles cannot be cited partly or entirely without showing resources. The responsibility about scientific and grammatical issues is belong to authors.

The papers sent to the journal are reviewed by two referees and after their approval, they will be sent to edit before being published. Writing & Publishing Policies can be found in the journal's website.

Makaleler / Articles

GOOGLE TRENDS İLE YAPILAN AVIAN INFLUENZA SORGULAMALARINDA İLGİLİ KONULAR BAŞLIĞININ SEÇİLMİŞ ÜLKELERE GÖRE DEĞİŞİMİNİN DEĞERLENDİRİLMESİ

***

EVALUATION OF CHANGE OF RELEVANT TOPICS ACCORDING TO SELECTED COUNTRIES IN AVIAN INFLUENZA QUESTIONS WİTH GOOGLE TRENDS

DOI: 10.33461/uybisbbd.677637

Berrin ŞENTÜRK*

Öz

Bu çalışmada, google trend verileri kullanılarak salgın algısının sosyal ve toplumlar arası farklılıkları araştırılmıştır. Bu amaçla, kuş gribi hastalığı seçilmiş ve "Avian Infuenza, Bird Flu ve Kuş Gribi" terimleri için "ilişkili konular" terimi değerlendirilmiştir. Çalışmada 01.01.2004-11.10-2019 dönemi için Türkiye, İngiltere, Çin ve Vietnam gibi bazı ülkelere ait google trends eğilimleri araştırılmıştır. Seçilmiş ülke verilerinin değerlendirildiği çalışmada internet verileri kullanılarak farklı ülke insanlarının bu hastalığa ilişkin genel arama eğilimleri frekans dağılımı dikkate alınarak değerlendirilmiştir. Dört ülke verisine ait arama terimlerinin ilgili konu başlıklarının değerlendirme sonuçları, hastalığın yüksek frekansa sahip konu başlıkları %26.25 oranında hesaplanmıştır (21/80). Düşük frekansa sahip konu başlıkları %18.75oranındadır (15/80). Bu durum hastalık konusunda bu ülkeler arasında algı düzeyinin önemli ölçüde değiştiğini düşündürmektedir. Çalışmada yüksek frekansa sahip konu başlıklarının kamu tarafından doğru bilgiye ulaşımın sağlanmasında linklerle desteklenmesinin ve hastalığa ilişkin bilginin yer aldığı dikkat çekici reklamların bu konu başlıkları ile ilişkili sayfalarda yer almasının önemli olduğu kanaatine varılmıştır.

***Anahtar Kelimeler**: Avian Influenza, Google Trends, Toplumsal Bilinç, İlişkili Konu.*

Abstract

In this study, social and inter-communal differences of epidemic perception were investigated using the google trends data. For this purpose, the avian influenza epidemic was chosen and key terms of "Avian Influenza, Kuş Gribi, and Bird Flu" related topics were evaluated. Some countries like Turkey, United Kingdom, China, and Vietnamese google trends data in the period of 01.01.2004-11.10-2019 were searched. In the study where country data were evaluated, general search tendencies of people from different countries were evaluated by using internet data considering frequency distribution. The results of the evaluation of the relevant topics of the search terms of the four country data, the disease's high-frequency topics were detected at a rate of 26.25% (21/80), while 18.75% of the low-frequency topics (15/80) were determined. This suggests that the level of perception among these countries varies significantly. In the study, it was found that it is important to support the high frequency topics with links in order to reach the right information by the public and it is important that the prominent advertisements containing the information about the disease appear on the pages related to these topics.

***Keywords:** Avian Influenza, Google Trends, Related Topic, Social Consciousness.*

* Doç. Dr, Department of Livestock Economics and Management, Faculty of Veterinary Medicine, Samsun, Ondokuz Mayıs University, Samsun, Türkiye, ORCID: 0000-0002-2540-6491

Geliş Tarihi / Received: 20.01.2020
Kabul Tarihi / Accepted: 08.07.2020

1. INTRODUCTION

Avian Influenza (AI) is an infectious disease of different poultry species. Sometimes mammals, and therefore humans, may develop this disease. The disease has subtypes such as H5N1, H5N2, H5N8, H7N8 and H7N9. Of these, H5N1 and H7N9 are more widely known to people because of their serious and lethal consequences (OIE, 2019). Globally, 861 cases of Avian Influenza H5N1 Virus were reported in 17 countries between January 2003 and June 2019, of which 455 resulted in death (WHO, 2019). On the other hand, since the beginning of 2013, 1,568 H7N9 subtypes of human infections have been reported by different laboratories (WHO, 2019). This new type of disease has shown exits involving many different countries. It is reported that the emergence of people during and after their travel raises deep concerns about the virus and its transmission (http://www.who.int/influenza/human_animal_interface/influenza_h7n9/).

A high number of deaths have been identified in China and Vietnam in cases of Avian Influenza caused by H5N1 virus on a global scale. It has been reported that between 2003 and 10 October 2019, 180 of the human cases were confirmed by laboratory results and 95 resulted in death (WHO, Western Pacific Region, Avian Influenza Weekly Update Number 710). In these countries and Turkey in disease studies are planned to investigate as keywords in google trends related topics. The height of the work in Turkey and the United Kingdom as well as of human cases in the selection of a limited number of countries where human cases in this country have been preferred because of the country's election.

Internet facilities are data sources that allow the storage of large amounts of long-term information. This situation creates data opportunities that will eliminate the problem of sample size in statistical terms by using this scientific knowledge. One of these possibilities is google trends data from search engines. The use of internet data in diseases has been increasing for the last ten years (Dugasve et al., 2013; Althouse et al., 2011; Ginsberg et al., 2009). In other news, a survey conducted in 2018 in Turkey in the field of health was used google trends (Yıldız, 2018). Google trends provides 15-year time series data for the searched keywords for any disease.

This study aimed to identify human cases of Avian Influenza in the number of deaths is high on the basis of the number of countries with data in Turkey google trends emerged in the interrogation of Avian Influenza for the difference. In the study conducted by examining the related topic and related question headings, the awareness of the people was tried to be determined and by using these data, it was tried to be guided in the development of disease management plans.

Today, using computer technology, which provides processable and manageable information for a large number of people, it is of utmost importance not to identify people's disease awareness and to increase this awareness by managing them and to use these opportunities in the control of diseases. Disease awareness in selected societies can be evaluated by analyzing the search data made by people from different countries on the same disease, taking into account a certain time period. Thus, programs that will increase the effectiveness of disease protection and control can be developed and implemented in line with the deficiencies determined using the results obtained. To this end, countries where a high perception of human deaths in this study (China and Vetna), perception developed countries (United Kingdom) and Turkey data are evaluated.

2. MATERIAL AND METHODS

The research question was determined as “to what extent people's perception of disease about “Avian Influenza” disease in different countries are similar and / or different from each other”. Frequency distributions of Google trends data will enable us to learn about disease awareness of individuals from different countries. The low number of low-frequency off-topic questions will strengthen this judgment.

In this study, "Avian Influenza", "Kuş Gribi" and the "Bird Flu" have to google trends of 10.16.2019 dated Turkey, the United Kingdom, data containing China and Vietnam started from January 2004 to June 2019 ranges were used (https: // trends.google.com.tr/trends). In this study, human cases used in country selection were determined by using World Health Organization data (https://www.who.int/docs/default-source/wpro---documents/emergency/surveillance/avian-influenza/ai-20191010.pdf?sfvrsn=30d65594_38; Access: 16.10.2019).

In the google trends queries of selected countries dated 16.10.2019; In the study, the data obtained for the keywords “Avian Influenza”, “Avian Influenza, Kuş Gribi and Bird Flu” which express the same disease were analysed by using “related topic” headings. Selected from four countries (Turkey, United Kingdom, China and Vietnam) 80 titles in a total of 10 subjects were included in the study for two keywords. In the study, the common headings of 4 country data and the frequency values of different headings were calculated for Avian Influenza” and “Kuş Gribi” or “Bird Flu” search headings only. High and low frequency topics were determined and the results were interpreted as to what contribution could be made to future disease control. The first 5 rows for high frequency calls are given in the study, while all low frequency calls are presented in the study. The data were presented by making a distinction in the form of low-frequency topics that are technically related to the disease and the topics that are associated with the disease.

3. RESULTS

In the study conducted, the frequency distribution of the topics obtained by bringing together the related topics of Avian Influenza, "Bird Flu" or "Bird Flu" search terms is presented in table 1.

Table 1. Frequency distribution of the topics obtained by combining the relevant topics of the search terms “Avian Influenza”, “Kuş Gribi” or “Bird Flu” (21/80).

Related Topic	Frequency
Virus-infectious agent type	5
Influenza A virus subtype H5N1-subject	4
Disease- subject	4
Vaccine-subject	4
Symptom-subject	4

Source: Calculation with Google trends data

The frequency distribution of the searches related only to the Avian Influenza term in the related topics is presented in Table 2.

Table 2. Frequency distribution of related topics in Avian Influenza search term (14/40)

Related Topic	Frequency
Disease-subject	4
Influenza A virus subtype H5N1-subject	3
Influenza- subject	3
Influenza A virus-virus	2
Bird-Animal	2

Source: Calculation with Google trends data

The frequency distribution of the related topics in the term Bird Flu or Bird Flu is given in Table 3.

Table 3. Frequency distribution of related topics in the term “Kuş Gribi” or Bird Flu (14/40)

Related Topic	Frequency
Virus-infectious agent type	3

Disease-subject	3
Swine flu-subject	3
Symptom-subject	3
H1N1 virus	2

Source: Calculation with Google trends data

Finally, in the study of google trends data, the low frequency titles of the related topics of Avian Influenza, Avian Influenza or Bird Flu search terms and the topics that are technically relevant and considered off-topic are presented in Table 4.

Table 4. In the Google Trends data, the low-frequency headings of the relevant subject headings of the Avian Influenza, "Kuş Gribi" or "Bird Flu" search terms and the headings that are technically relevant and considered off-topic (15/80)

Related Topics	Frequency	Related	Irrelevant
Sudden outbreak-subject	1	X	
A country in China-East Asia	1	X	
World Health Organization-subject	1	X	
Hemaglutinin-subject	1	X	
HTML5-Video game engine	1		X
Cell-subject	1	X	
Characterization -subject	1	X	
Husband-subject	1		X
Comedy –sort of film*	1		X
Room-subject	1		X
Death-subject	1	X	
Pathogenicity -subject	1	X	
Van lake monster-subject*	1		X
Van lake- A lake in Turkey *	1		X
Van-subject*	1		X

Source: Calculation with Google trends data

In the study, when technically low-frequency subject headings are divided into two categories as subject-related and non-subject low-frequency searches, the proportional distribution of these searches constitutes 46.6% (7/15) of non-subject searches, and 57.1% (4/7) it was found that from Turkey. In the calls that make up 18.75% (15/80) of the total calls, the non-subject search rate is 8.75% (7/80).

4. CONCLUSIONS and RECOMENDATIONS

The study suggests that the internet opportunities in the information technologies of the age will create important opportunities for disease protection and control in the field of health in the future. The tools of search engines such as google trends provide big data opportunities in determining human tendencies in animal diseases and zoonoses, and good use of this field creates an important opportunity not only for countries but also for sustainable health practices on a global scale. With the help of multidisciplinary studies, this information source will be used more effectively and technology and health information will be brought together to ensure that human resources reach accurate and reliable information.

Disease control requires a high public cost. Access to accurate and reliable information is extremely important as the Internet is one of the first tools used to make information easy and accessible. The public has an important responsibility for information. Since the Internet is an important source of information, the public should consider this area well.

The data of this study showed that the general awareness of Avian Influenza, a global disease, was 26.25% (21/80), and 18.75% (15/80) had low perception levels. Firstly, there should be consensus on the value of an acceptable reasonable rate of awareness in epidemic diseases. This rate will be different for each disease, for the initial stage and for the subsequent processes. Achieving the desired level of these rates within certain time intervals and uploading more information to keyword pages related to advertising, games, or other methods to be determined will create an important opportunity for combating diseases.

On the other hand, the use of public links with accurate and important information about the disease, the development of different applications designed for age groups such as visual perception, and providing this information with the right content will provide important opportunities in disease control.

In the study, it was interpreted that low frequency ratio was not sufficiently comprehended the importance of the subject in the society. In such cases, multi-faceted assessments should be made by the public, and all necessary measures should be taken to raise awareness and be informed with all kinds of resources available.

As a result, in the study, the frequency distributions of google trends data and the related topics for epidemic diseases were calculated and ways to benefit from this opportunity to raise the awareness of the disease on selected country results were tried to be put forward.

KAYNAKÇA

Althouse BM, Ng YY, Cummings DAT (2011). Prediction of Dengue Incidence Using Search Query Surveillance. PLoSNegl Trop Dis 5(8): e1258. doi: 10.1371/journal.pntd.0001258.

Ginsberg J, Mohebbi MH, Patel R S, Brammer L, Smolinski MS, Brilliant L. (2009): Detecting Influenza Epidemics using Search Engine Query Data, Nature, 457, 1012 –1014.

Yıldız MS (2018). Google Search Trends: An Application for Health Services Related Queries ın Turkey, International Journal of Health Management and Strategies Research, 4(2): 168-179.

OIE (2019). Animal Health in the World, Avian Influenza portal https://www.oie.int/animal-health-in-the-world/web-portal-on-avian-influenza/) (Accessed: 16.10.2019).

WHO (2019). World Health Organization. [(http://www.who.int/influenza/human_animal_

interface/influenza_h7n9/); (Accessed: 16.10.2019)].

WHO (2019). Western Pasific Region, Avian Influenza Weekly Update Number 710.

Anon (2019). Google Trends [(https://trends.google.com.tr/trends); (Accessed: 16.10.2019)].

YÜKSEK RAFLI DEPOLAMA SİSTEMLERİNİN ENERJİ OPTİMİZASYONUNDA ANOMALİ TESPİTİ İÇİN SINIFLAMA ALGORİTMALARININ KARŞILAŞTIRILMASI

***

COMPARISON OF CLASSIFICATION ALGORITHMS FOR ANOMALY DETECTION IN ENERGY OPTIMIZATION OF HIGH RACK STORAGE SYSTEMS

Cihan BAYRAKTAR*
Hadi GÖKÇEN**

DOI: 10.33461/uybisbbd.790369

Öz

Birimler arasında sağlıklı veri akışının sağlanması ile dijitalleşen üretim sistemleri ve bu dijitalleşme süreci doğrultusunda otomatikleşen zeki fabrika yapıları gün geçtikçe üretim endüstrisinde kendisine daha fazla yer bulmaktadır. Bu tür sistemler, üretim önemli gelişmeler ve teknolojik ilerlemeler sağlamış olsa da çeşitli sorunları da beraberinde getirmektedir. Bunlardan bir tanesi de otonom çalışan üretim sistemlerinde gerçekleşen bir anormal durumun hızlı bir şekilde tespit edilerek, çözüme kavuşturulması sürecidir. Bu kapsamda son zamanlarda anomali tespiti için çeşitli çalışmalar yapılmaktadır. Anomali tespiti konusunda en çok destek alınan alanlardan bir tanesi de makine öğrenmesi algoritmalarıdır. Bu çalışmada, yüksek depolama sistemlerinin enerji optimizasyonu hakkında uygulanmış bir prototip çalışmadan elde edilmiş olan iki farklı veri seti üzerinde çeşitli makine öğrenmesi algoritmalarının performansları test edilmiştir. Sonuç olarak, Yapay Sinir Ağları, C4.5 Karar Ağacı, Rastgele Orman ve k En Yakın Komşu algoritmaları ile oluşturulan öğrenme modelleri, test edilen veri setleri içerisindeki anomalileri tespit etme konusunda yüksek başarım oranı elde etmişlerdir. Özellikle bu algoritmalar içerisinde Rastgele Orman algoritması yaklaşık %98 seviyesindeki doğruluk performansı ile dikkat çekmiştir.

***Anahtar Kelimeler**: Anomali Tespiti, Sınıflandırma, Makine Öğrenmesi, Zeki Fabrikalar.*

Abstract

The production systems digitized by ensuring healthy data flow between the units and the smart factory structures that are automated in line with this digitization process find more and more places in the production industry. Although such systems have provided important developments and technological advances in production processes, they also bring with it various problems. One of these is the process of quickly detecting and resolving an abnormal situation occurring in autonomous production systems. In this context, various studies have been carried out recently for anomaly detection. One of the most studied areas for anomaly detection is machine learning algorithms. In this study, the performances of various machine learning algorithms were tested on two different data sets obtained from a prototype study on energy optimization of high storage systems. As a result, learning models created with Artificial Neural Networks, C4.5 Decision Tree, Random Forest and k Nearest Neighbor algorithms have achieved a high performance rate in detecting anomalies within the tested data sets. Among these algorithms, the Random Forest algorithm has attracted attention with its accuracy performance of approximately 98%.

***Keywords:** Anomaly Detection, Classification, Machine Learning, Smart Factories.*

* Öğr. Gör., Karabük Üniversitesi, Eskipazar Meslek Yüksekokulu, Bilişim Güvenliği Teknolojileri, Karabük, Türkiye, e-posta: cihanbayraktar@karabuk.edu.tr, ORCID: 0000-0003-4321-5485
** Prof. Dr., Gazi Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği, Ankara, Türkiye, e-posta: hgokcen@gazi.edu.tr, ORCID: 0000-0002-5163-0008

Geliş Tarihi / Received: 04.09.2020
Kabul Tarihi / Accepted: 16.11.2020

1. GİRİŞ

Endüstri 4.0, 2011 yılında üniversiteler ve özel şirketler ile birlikte Alman Federal Hükümeti tarafından bir girişim olarak icat edilmiştir. Stratejik bir program olarak ortaya çıkan bu programın amacı, endüstrinin üretkenliği, etkinliği ve verimliliğini arttırmak ve gelişmiş üretim sistemleri ortaya çıkartmaktır. Bu yapı, ürün yaşam döngüsüne katkı sağladığı bilinen bir dizi teknolojinin bir çatı altında bir araya getirilerek, ortak bir yapı içerisine entegre edilmesini kapsamaktadır. Endüstri 4.0, gelişmiş üretim veya zeki üretim yapısında, esnek hatların oluşmasına imkan tanıyan ve bu sayede, çok çeşitli ürün türü ve değişen şartlar doğrultusunda üretim süreçlerinin otomatik ayarlandığı bir sistem olarak kullanılmaktadır (Frank vd., 2019).

Endüstri 4.0'ın diğer endüstri devrimleri içerisinde planlı olarak gerçekleşen ilk devrim olacağı belirtilmektedir. Yeni nesil endüstri, dijitalleşen üretim sistemlerinin çıktıları ve bileşenleri tarafından şekillendirilecektir. Bu durum sayesinde, üretim sistemlerinde kullanılan tüm fiziksel bileşenler, makineler arası iletişim sistemine entegre edilecektir. Özellikle üretim sistemlerinde uygulanacak olan dijitalleşme süreci, optimum çalışma, kişiselleştirilmiş ürünler ve esnek üretim yapılarının ortaya çıkmasını sağlayacaktır (Riordan vd., 2019).

Büyük veri, nesnelerin interneti (IoT), siber fiziksel sistemler gibi yenilikçi kavramların yükselişleri ve çeşitli teknolojilerin geliştirilmesi yeni dönemde endüstri yapısını üst seviyelere taşımıştır. Bunların sonucunda zeki üretim sistemlerinin oluşturulmasını hedefleyen endüstriyel nesnelerin interneti (IIoT) yapısı ile zeki fabrikalar ortaya çıkmıştır. IIoT sayesinde zeki fabrikalar, bilgilerin bağımsız olmadığı grup etkileşimlerini gerçekleştirmektedirler. Birimler arasında bilgilerin kaynaşması ve çarpışması aracılığı ile zeki üretim süreçlerinin geliştirilmesi sağlanabilmektedir (Wan vd., 2019). Zeki fabrikalar, ürünlerin kısa yaşam döngülerinden, deneyimli çalışan eksikliğinden, ülkelerce yürütülen çeşitli çevre düzenlemelerinden ve sürekli olarak artan müşteri taleplerinden kaynaklanan sorunların, hızlı ve hatasız bir şekilde çözüme kavuşturulabilmesi için geliştirilmektedir (Yoon vd., 2019).

IIoT, zeki fabrika sistemlerinde temel ekipmanların yapıya entegre edilmesinde kullanılan bir teknolojidir. Bu şekilde üretim sistemi, algılama, ara bağlantı sağlama ve verilerin entegrasyonunu gerçekleştirme yeteneğine sahip olur. Verilerin analiz edilmesi ve bilimsel karar verme süreçleri, zeki fabrikalar bünyesinde, üretim planlanmasını, ekipmanların verimli kullanımını ve kalite kontrol süreçlerinde kullanılmaktadır. Ayrıca sistem verilerinin yerel bir veri tabanından, bulut sistemine yüklenmesi için internet de ciddi anlamda kullanılmaktadır. Zeki fabrikalar, insan ve makinenin etkileşimi sayesinde, küresel iş birliğine dayalı zeki üretim sistemlerini inşa etmektedirler (Chen vd., 2017).

Zeki fabrikalar, gün geçtikçe daha da karmaşık hale gelen bir dünyada, dinamik ve hızlı değişen şartlara sahip bir üretim sistemi için, ortaya çıkabilecek sorunları çözebilecek esnek bir üretim sistemi çözümüdür. Bu çözüm aynı zamanda, gereksiz iş gücü ve kaynak israfının önüne geçebilmek amacıyla kullanılması gereken yazılım ve donanım birimlerinin kombinasyonunu da ilgilendirmektedir. Ayrıca zeki sistemlerin gerçekleştirmesi gereken görevlerden biri olan endüstriyel ortam ve çevre arasındaki bağlantının da sağlıklı bir şekilde kurulması ve yönetilmesi konuları ile de ilgilenmektedir (Radziwon vd., 2014).

Zeki fabrika sistemleri için önemli olan sorunlardan bir tanesi anomali tespiti aşamasıdır. Zeki fabrikaların sahip olduğu karmaşık yapı, sistemlerin istenmeyen durumların oluşmasına sebep olabilmektedir. Son zamanlarda, özellikle güvenlik açıkları ve anomali konusu, zeki fabrikalar için fenomen halini almıştır (Hasan vd., 2019). Anomali tespiti için çalışma yapılan sistemlerde, bakım sıklığını en düşük seviyeye çekebilmek için tahmini bakım çalışmaları yapılabilir ve üretim kaynaklarının verimli kullanılması kapsamında uygulanan çalışmalarda önemli ölçüde maliyet avantajı sağlanabilir. Ayrıca fabrika içi üretim kapasitesi ve sistemin karmaşıklık seviyesi arttıkça, karşılaşabilecek olan sorunların çeşidi ve miktarı da artış gösterecektir. Oluşması muhtemel bu

sorunların çözümlerinde, sorunu erken tespit etmek ve maliyetleri en aza indirebilmek için, üretim sistemi içerisinde çalışan cihazların anormal davranışlarının analizlerinin yapılması ve tespit edilmesi gerekmektedir. Böylece üretim sisteminde yaşanacak süreç gecikmeleri ve zararların daha hızlı önüne geçmek mümkün olabilecektir (Hsieh vd., 2019).

Endüstri 4.0 devriminde, zeki fabrikalar tarafından temsil edilen fiziksel sistemler ve bilişim teknolojilerinin etkileşimi, birbirine bağlı olan tüm sistemler arasında çok büyük miktarlarda gerçek zamanlı veri alışverişine imkan sağlamıştır. Bu sayede, üretim alanında anomali tespit sistemleri için gereklilik ortaya çıkmıştır. Anomali tespit sistemlerinde, üretim sistemi içerisinde çalışan cihazlardan toplanan veriler, bilişim sistemleri içerisine dahil edilmektedir. Sisteme aktarılan veriler üzerinden anlık değerlendirme yapılarak, herhangi bir anomali tespit edildiği takdirde, sistemde oluşacak düşük performansı ve yüksek hata ihtimallerini hızlı bir şekilde engellemek için operatörlere gerekli bilgilerin aktarım işlemleri sağlanmaktadır. Bu sayede yüksek oranda kesintisiz ve performanslı bir üretim sisteminin işleyişi mümkün olabilmektedir (Bagozi vd., 2017).

Bu çalışmada, zeki fabrikaların üretim sistemlerinde, anomali tespiti konusunda makine öğrenmesi algoritmalarının nasıl bir başarım gerçekleştirdiklerinin ölçülmesi ve uygun algoritmanın tespit edilmesi amaçlanmıştır. Bu kapsamda da kaggle sisteminden temin edilmiş olan açık kaynak kodlu ve yüksek depolama sistemlerinde enerji optimizasyonun sağlanması üzerine yapılmış bir çalışmadan elde edilmiş veri setleri kullanılmıştır. Çalışmanın ikinci bölümünde, anomali tespiti üzerine yapılmış olan çalışmalar hakkında bilgi verilmiştir. Üçüncü ve dördüncü bölümlerde ise, veri setleri üzerinde makine öğrenme algoritmalarının anomali tespiti konusundaki başarımları ölçülmüş ve birbirleri arasında kıyaslanarak, sonuçlar yorumlanmıştır.

2. İLGİLİ ÇALIŞMALAR

2016 yılında gerçekleştirilen çalışmada, hibrit üretim sistemlerinde kullanılabilecek olan otomatik öğrenme özelliğine sahip anomali tespit algoritması önerilmiştir. Sistem içerisinde uygulanan gözlemlerden tespit modelini oluşturmak için derin öğrenme teknikleri ve zamanlı otomata sistemlerinin birleşiminden yararlanılmıştır. İki adet gerçek sistem dahil olmak üzere çeşitli veri setleri üzerinde test ettikleri algoritmanın umut verici sonuçlar verdiği açıklanmıştır (Hranisavljevic vd., 2016). 2017 yılında yayınlanmış olan devam çalışmasında da hibrit üretim sistemlerinde anomali tespiti için, denetimsiz ve parametrik olmayan bir yaklaşım oluşturulmuştur. Bu yaklaşım ile normal şartlarda anomali tespit uygulaması mümkün olmayan üretim sistemlerinde, hibrit zamanlı otomata kullanımına izin vererek anomali tespiti yapılmasını sağlayan, kendi kendini düzenleyen haritalar ve havza dönüşümleri kullanılmıştır (Birgelen ve Niggeman, 2017).

2018 yılında yapılmış olan bir çalışmada, IoT sisteminde hizmet içi servislerin iletişim ve çalışma yapısını öğrenen ve kendini sürekli güncel tutan, kaynak verimli bir yaklaşım önerilmiştir. Önerilen bu yaklaşımın, düğümler arasındaki süreç iletişiminde akan verileri analiz ederek, öğrenmiş olduğu model doğrultusunda anomali tespiti yapabildiği belirtilmiştir. Çalışma sayesinde IoT sistemlerinin güvenlik seviyelerinin daha üst noktalara ulaşabildiği sonucuna varılmıştır (Pahl ve Aubet, 2018).

2019 yılında yayınlanmış olan bir çalışmada, nesnelerin interneti üzerine gerçekleşebilecek siber saldırılar dolayısı ile oluşabilecek anomalilerin tespit edilmesi noktasında çeşitli makine öğrenmesi algoritmalarının performans karşılaştırmaları yapılmıştır. Yapılan ölçümler sonucunda, Karar ağacı, rastgele orman ve yapay sinir ağları algoritmalarının %94 gibi yüksek başarıya ulaştığı, performans bakımından ise rastgele orman algoritmasının öne çıktığı tespit edilmiştir (Hasan vd., 2019).

Bir başka çalışmada, üretim sistemi üzerinden bulunan cihazlardan toplanan gerçek veriler kullanılarak, zeki üretim sistemlerinde anomali tespiti için kullanılabilecek bir algoritma

önerilmiştir. Üretim hattından elde edilen çok değişkenli sensor veri setlerinde bulunan sınırlı ve düzensiz anomali verilerinin ortaya çıkartılabilmesi için, otomatik kodlayıcıya dayalı denetimsiz gerçek zamanlı anomali algılama algoritması kullanılmıştır. Sonuç olarak önerilen bu algoritmanın anomali tespitinde %90 başarım oranı elde ettiğini belirtilmiştir (Hsieh vd., 2019).

Wang ve diğerleri tarafından anomali tespiti için derin öğrenme yapıları üzerine yapılan çalışmada araştırmacılar, derin öğrenme tabanlı oluşturulan anomali tespit sistemlerinin daha iyi anlaşılmasını amaçlamışlardır. Çalışmada ilk etapta derin öğrenme yapılarından önce uygulanan anomali tespit teknikleri açıklanmış ve sonrasında günümüzde uygulanan yüksek teknolojiye sahip derin öğrenme tabanlı anomali tespit tekniklerinin, öncesinde kullanılan geleneksel algoritmaların sorunlarını aşma konusunda kullandıkları teknikleri tartışmışlardır (Wang vd., 2020).

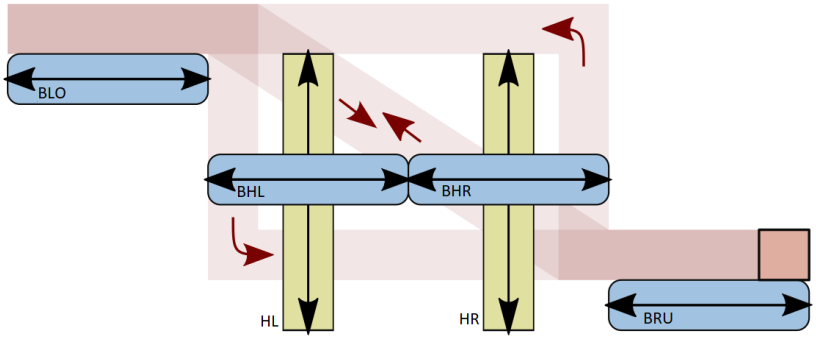
3. DENEYSEL ÇALIŞMA

Sistemin tamamı çeşitli bağımsız süreçlerden oluşmaktadır. Buradaki ilk aşama verilerin bir araya getirilmesidir. Veriler, dikkatli bir şekilde toplanıp incelenerek, uygun veri tipleri elde edilmeye çalışılmaktadır. Bir sonraki süreçte ise veri üzerinde önişleme adımları uygulanmaktadır. Önişleme adımları, verinin içerisinde bulunan gürültünün temizlenmesi, eksik verilerin tamamlanması, veri üzerinde dönüştürme ve birleştirme işlemlerinin uygulanmasından oluşmaktadır. Önişleme uygulamalarının ardından veri, artık sınıflandırma algoritmalarının testi için kullanılabilir duruma gelmiş olacaktır. Bu kısımda veri setlerinin analizinde 10 Fold Cross Calidation yöntemi kullanılmıştır. Test edilen her algoritma, öğrenme kümesini kullanarak kendi öğrenme modelini oluşturacak, sonrasında ise bu modeli test kümesi üzerinde sınayarak başarım derecesini ölçecektir. Bu çalışma kapsamında farklı sınıflandırma algoritmaları kullanılmıştır. Bunlar; Lojistik Regresyon, Naive Bayes, Destek Vektör Makineleri, Karar Ağaçları, Rastgele Orman, k En Yakın Komşu ve Yapay Sinir Ağları algoritmalarıdır.

3.1. Veri Setinin Oluşturulması ve Tanımlanması

Çalışmada kullanılan açık kaynak veri seti, Hranisavljevic ve diğerleri tarafından oluşturulmuş ve bu çalışma için kaggle ortamından çekilmiştir (Hranisavljevic vd., 2016; Hranisavljevic vd., 2018). İlgili veri setinin oluşturulabilmesi amacıyla, dört adet kısa konveyör bandından (BLO, BHR, BHL ve BRU) ve iki raydan (HL ve HR) oluşan bir yüksek raflı depolama sistemi oluşturulmuştur. Oluşturulan yüksek raflı depolama sisteminin görsel modeli, Şekil 1'de verilmiştir. Ortada bulunan BHL ve BHR konveyör bantları, raylar üzerinde dikey yönlü hareket gerçekleştirmektedirler. Diğer bantlar ise sabittir. Bu sistem, iki nokta arasında paket taşımak için oluşturulmuştur. Sistemin oluşturulan ilk versiyonunda orta konveyör bantları dikey hareket halinde iken yatay yönlü paket taşımayacak şekilde ayarlanmış ve bu modda çalıştırılarak ilk veri seti elde edilmiştir. İkinci versiyonda ise sistem optimize edilmiş ve orta bantlar dikey hareket ile aynı anda yatay yönlü paket taşıma işlemini de yapacak şekilde güncellenmiş ve ikinci veri seti oluşturulmuştur (Hranisavljevic vd., 2018). Bu çalışma bünyesinde, elde edilmiş olan veri setlerinin ikisi ile ayrı ayrı sınıflama algoritmaları uygulanmış, optimizasyon işlemi uygulanmış veri seti ile optimize edilmemiş veri seti arasındaki fark da gözlenmiştir.

***Şekil 1**: Yüksek Raflı Depolama Sisteminin Görsel Modeli* (Hranisavljevic vd., 2018).



Tablo 1'de verilmiş olan düzene göre, veri setlerinin her birinde 19 adet nitelik ve bir adette sınıf niteliği bulunmaktadır. Bu niteliklerden ilki işlem süresinin gösterildiği zaman damgası niteliğidir. Diğerleri ise yapı içerisindeki her bir elemanın (Dört adet Konveyör ve iki adet ray) mesafe, güç ve voltaj sinyallerinden oluşmaktadırlar (Birgelen ve Niggeman, 2017).

Tablo 1: Veri Setleri Nitelik Tanımları.

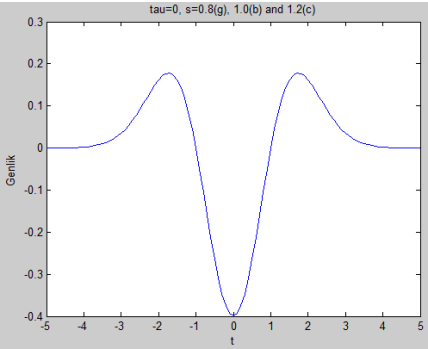
Nitelik Adı	Açıklaması	Standart Veri Seti Değer Aralığı	Optimize Edilmiş Veri Seti Değer Aralığı
TimeStamp	Saniye Cinsinden Süre	0 – 18,851	0 – 14,246
I_w_BLO_Weg	Sol Üst Konveyör Mesafe Bilgisi	(-315,836) – 739,2	(-3,4,264) – 1011,098
O_w_BLO_Power	Sol Üst Konveyör Güç Bilgisi	0 – 34817,661	(-82,014) – 30140
O_w_BLO_Voltage	Sol Üst Konveyör Voltaj Bilgisi	(-113,02) – 179,025	0 – 60
I_w_BHL_Weg	Orta Sol Konveyör Mesafe Bilgisi	(-548) – 1301,892	(-895,2) – 1130,4
O_w_BHL_Power	Orta Sol Konveyör Güç Bilgisi	(-4314,545) – 48719,681	(-8990,991) – 39838,352
O_w_BHL_Voltage	Orta Sol Konveyör Voltaj Bilgisi	(-22,111) – 66	(-43,87) – 105
I_w_BHR_Weg	Orta Sağ Konveyör Mesafe Bilgisi	(-1322) – 621,4	(-1322) – 1015,3
O_w_BHR_Power	Orta Sağ Konveyör Güç Bilgisi	0 – 32536	(-241,091) – 41507,7
O_w_BHR_Voltage	Orta Sağ Konveyör Voltaj Bilgisi	0 – 67,2	(-0,955) – 119,6
I_w_BRU_Weg	Sağ Alt Konveyör Mesafe Bilgisi	(-661,952) – 688,8	(-855) – 755,851
O_w_BRU_Power	Sağ Alt Konveyör Güç Bilgisi	0 – 62674	0 – 35008,471
O_w_BRU_Voltage	Sağ Alt Konveyör Voltaj Bilgisi	0 – 75,4	0 – 72,8
I_w_HL_Weg	Sol Ray Mesafe Bilgisi	(-1082,9) – 101,118	(-1151,204) – 186,85
O_w_HL_Power	Sol Ray Güç Bilgisi	0 – 41789,8	0 – 41940,6
O_w_HL_Voltage	Sol Ray Voltaj Bilgisi	0 – 596,4	0 - 279
I_w_HR_Weg	Sağ Ray Mesafe Bilgisi	(-1032,2) – 0	(-833) – 0
O_w_HR_Power	Sağ Ray Güç Bilgisi	0 – 42895,835	0 – 39060,793
O_w_HR_Voltage	Sağ Ray Voltaj Bilgisi	0 – 543	0 – 280
Sınıf	Anomali Bilgisi	0 veya 1	0 veya 1

Yüksek raflı depolama sisteminin çalıştırılması ile elde edilen veriler, anomali tespit işlemlerinde kullanılabilmesi amacıyla iki boyuta indirgenmiştir. Eğitim amaçlı yapılan ilk gözlemler, her değerlendirme gözleminde mesafeyi hesaplamak için referans değerleri sağlamaktadır. Mesafe değeri belirli bir eşik değerini aşan veriler anomali olarak işaretlenmiştir. Anomali durumunu gösteren eşik değerinin hesaplanmasında ise Meksika Şapkası Dalgacık yöntemi kullanılmıştır. Meksika Şapkası Dalgacığı, Gauss Fonksiyonunun normalizasyon

işleminden sonra elde edilen versiyonunun ikinci türevinin alınış halidir. Şekil 2'de görüldüğü üzere, eğri biçimi Meksikalıların giydiği şapkaya benzediğinden dolayı bu isimle anılmaktadır. Matematiksel ifadesi, aşağıdaki eşitlikte gösterilmektedir (Şeker vd., 2018).

$$\varphi(t,\tau,s) = \frac{[\left(\frac{(t-\tau)}{s}\right)^2 - 1]exp\{(\frac{(t-\tau)}{s})^2 * (-0{,}5)\}}{\sqrt{2\pi} * s^3} \quad (1)$$

Şekil 2. Meksika Şapkası Dalgacığı Grafiği



Eşik değeri için en uygun seviyesinin belirlenmesi önemlidir. Çünkü, eğer eşik değeri çok yüksek olursa, (Örn: %100) normal değerlerin anomali olarak algılanma olasılığı artacaktır. Eşik değeri çok düşük olursa, (Örn: %25) o zaman da algoritmalar anomaliyi tespit edemeyeceklerdir. Bu çalışmada kullanılan veri setlerinde işaretlenmiş olan anomaliler, %60 eşik değeri ile belirlenmişlerdir (Birgelen ve Niggeman, 2017).

Çalışma kapsamında kullanılan veri setlerinden birincisi, ortadaki konveyör bantlarının dikey hareket halinde iken yatay yönlü paket taşıma işlemi yapmadıkları standart çalışma süreçleri ile elde edilen verilerden oluşmaktadır. Bu veri setinde, toplam 23645 satır veri bulunmaktadır. Bu verilerden 5670 adedi anomali olarak işaretlenmiş, kalan 17975 adet veri ise normal süreç verileri olarak kaydedilmiştir.

İkinci veri seti ise ortada bulunan konveyör bantlarının dikey hareket halinde iken, yatay yönlü paket taşıma sürecini de gerçekleştirdikleri optimize edilmiş süreçlerden elde edilen verilerden oluşmaktadırlar. Bu veri setinde, toplam 19634 satır veri bulunmakta, bunlardan 4517 adedi anomali olarak işaretlenmiş, diğer 15117 adedi ise normal süreç verileri olarak kaydedilmiştir.

3.2. Veri Ön İşleme

Çeşitli çalışmalar kapsamında kullanılmak üzere elde edilen veri setlerinde, bazı verilerde eksiklikler, hatalar, tekrarlar veya anlamsızlıklar bulunabilmektedir. Bundan dolayı, veri setleri üzerinde çalışma yapmaya başlamadan önce, çeşitli veri düzenleme süreçlerinden geçirmek önem arz etmektedir. Bu süreçler, kayıp verilerin düzenlenmesi, gürültünün ortadan kaldırılması, bütünleştirme, dönüştürme ve azaltma şeklinde isimlendirilebilmektedir (Aydemir, 2019). Dolayısı ile araştırmacının elindeki veri setinin ihtiyaçları doğrultusunda, anılan düzenleme işlemlerinden uygun olanları değerlendirmesi gerekmektedir.

Kaggle ortamından temin edilen iki adet veri setleri üzerinde yapılan gözlemlerde, herhangi bir eksik ve anlamsız veriye rastlanmamıştır. Veri setlerinin her birinde bir adet sınıf etiketi, 19 adet ise nitelik etiketi bulunmaktadır. Nitelik etiketlerinin tamamı, nümerik verilerden oluşmaktadır. Sınıf etiketi ise verilerde anomali var olup olmadığını gösterdiği için, 1 veya 0 değerlerini içerecek şekilde nominal olarak işaretlenmiştir.

3.3. Teorik Kavramlar

Çalışma kapsamında birçok sınıflandırma algoritması kullanılmış ve karşılaştırma işlemleri yapılmıştır. Kullanılan algoritmalar ile ilgili açıklamalar alt başlıklarda ifade edilmiştir.

3.3.1. Naive Bayes (NB)

İstatistiksel bir sınıflandırma algoritması olan Naive Bayes (NB), arka planda istatistiksel değerlere göre farklılık gösterebilen bir çalışma sistemine sahiptir. Bu sebeplerden dolayı dinamik olarak çalışan sistemler üzerinde kullanımı esnasında, tekrar tekrar hesaplama işlemlerinin gerçekleştirilmesini gerektiren bir algoritmadır (Şeker, 2016).

NB sınıflandırma tekniği, Bayes kuralı ile birlikte karar ağaçları modeli birleştirilerek elde edilmiş bir tekniktir. NB Algoritması, örneği verilmiş olan her sınıfın olasılık değerini hesaplamak amacıyla Bayes kuralını kullanmaktadır (Akçetin ve Çelik, 2014). Makine öğrenmesi uygulamalarında sıkça karşılaşılan bir sınıflandırma tekniği olan NB, koşullu olasılık hesaplamaları üzerine çalışan ve Bayes kuralının en basit hali olarak nitelendirilen bir algoritmadır (İşçimen vd., 2014). NB, sınıflandırma teknikleri içerisinde en kısıtlayıcı alanda yer almaktadır. Bu algoritmada örnek verilerin hangi sınıfa ait oldukları bilinmektedir. Genel anlamda metin sınıflandırılmasında üstün başarı gösterdiği tespit edilmiştir (Nizam ve Akın, 2014). NB algoritması, koşullu sınıf bağımsızlığının varsayımı üzerinde durmaktadır. Yani, herhangi bir değer niteliğinin, diğer niteliklerin değerlerinden bağımsız olduğu düşünülür. Pratikte nitelikler arasında bir miktar bağımlılık olsa dahi, hesaplamaların kolaylaştırılması amacıyla teoride bu varsayım uygulanmaktadır. Eğer uygulanan varsayımlar doğru olursa, NB algoritması diğerlerine göre en iyi sonucu veren algoritma olacaktır. Ayrıca işleme tabi tutulan nitelikler arasında öneme göre bir derecelendirme yapılmaz ve sınıfın tahmin edilmesinde tüm niteliklerin aynı derecede önemli olduğu kabul edilir (Han ve Kamber, 2006).

Genel anlamda NB tekniği, X dizisi içerisindeki her bir verinin, C sınıfına ait olup olmama olasılığını hesaplamak için tercih edilmektedir (Hand vd., 2001).

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi|h1)P(h1)+P(xi|h2)P(h2)} \quad (2)$$

Eşitlik (2)'de P(h1) ifadesi, h1 hipotezi ile birlikte ön olasılık olduğu zaman, p(h1|xi) ifadesi sonraki olasılık olarak değerlendirilmektedir (Patil ve Sherekar, 2013).

Ayrıca eşitlik (3) ile P(h1|xi) değeri maksimize edilmektedir. Maksimize edilmiş olan P(h1|xi) değerinin C sınıfı, Bayes teoremine göre maksimum sonraki olasılık olarak ifade edilmektedir (Han ve Kamber, 2006).

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi)} \quad (3)$$

3.3.2. Yapay Sinir Ağları (YSA)

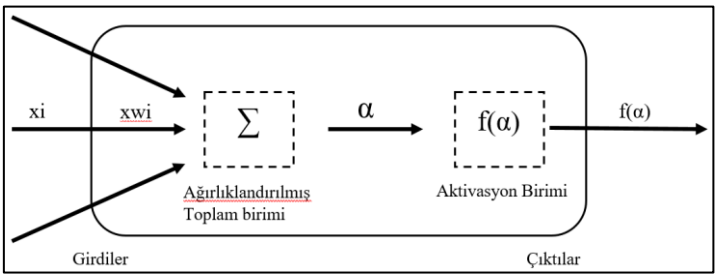
Yapay Sinir Ağları (YSA), ağırlıklı bağlantıları kullanarak birbiri ile bağlantı kurmuş olan elemanlardan oluşmaktadır. Ayrıca bu elemanların, her biri kendilerine ait, paralel ve dağıtılmış bilgi işleme yeteneğine sahip bellekleri bulunmaktadır. Farklı bir ifade ile YSA, biyolojik sinir sisteminin kopyası gibi çalışması amaçlanmış bir bilgisayar programıdır. YSA, bu özelliği sayesinde kendi kendine öğrenen bir yapıya sahiptir. Öğrenmenin yanında ezberleme, bilgiler arasındaki ilişkiyi ortaya çıkartma gibi yetenekleri de bulunmaktadır. Ayrıca tüm bunları yapabilmesi konusunda yazılımcının geleneksel yeteneklerine muhtaç değildir (Elmas, 2016).

YSA, yapay zeka çalışmalarının gelişmesine katkı sağlayan alanlardan bir tanesidir. Buna dayanarak YSA'nın, öğrenme yeteneğine sahip sistemlerin başında gelen yapay zeka teknolojilerinin bir parçası olduğu düşünülebilir. YSA için, insan beyninin temel elemanlarından olan nöronların çalışma prensiplerini kopyalamaya çalışarak, gerçek sinir sistemi yapısının bir simülasyonunu oluşturmaya yarayan programlar olduğu söylenebilir (Aydemir, 2019).

YSA, insan beyninde bulunan nöronlar gibi çalışan yapay nöronlar aracılığı ile örnekler üzerinde yeterli incelemeleri yaparak değerli olan bilgiyi ortaya çıkartmak için kullanılmaktadır. Bu yapay nöronlar, problemlerin çözüm aşamalarında kendi öğrendiklerini kullanarak karar verebilme yeteneğine sahiptirler. Kısaca YSA, çeşitli geometrik şekillere sahip yapay nöronların arasında kurulan bağlantı ile oluşan ağ yapıları olarak ifade edilebilmektedir. Bahsi geçen ağ yapıları oluşturulduktan sonra, yeni gelen verilerin sınıflandırılması için kullanılmaktadırlar (Staub vd., 2015). YSA sıklıkla, doğrusal olsun veya olmasın herhangi bir problem hakkında girdi olarak kullanılan veriler ile çıktı olarak elde edilmesi gereken veriler arasında gerekli bağlantıyı kurarak sonraki uygulamalar hakkında sonuçlar elde edebilmek amacıyla kullanılmaktadır (Yakut vd., 2014).

Şekil 3'te gösterilen, basit problemlerin çözümünde kullanılabilecek olan sinir ağı, girdi ve çıktı nöronlarından oluşan tek katmanlı ağlar olarak ifade edilmektedir.

Şekil 3: Tek Katmanlı Yapay Sinir Ağı Yapısı (Yakut vd., 2014)

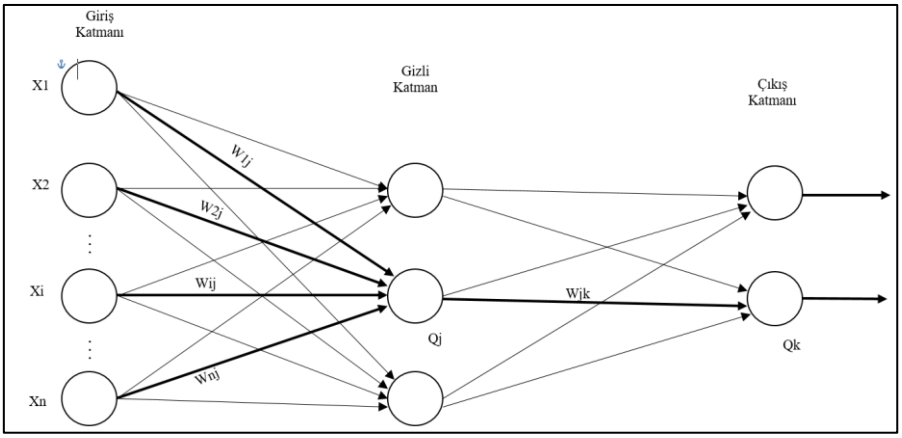


Bu tip ağlarda tüm girdi ve çıktı nöronları bir veya birden fazla olabilir ve tüm girdi nöronları, tüm çıktı nöronlarına ağırlıklandırılarak bağlanmaktadır. Ağırlıklandırılmış bağlantı, eşik değeri ile kontrol edilerek aktivasyon fonksiyonun çalıştırılması ve çıkış değerinin hesaplanması hedeflenmektedir. Bu olay matematiksel olarak şu şekilde ifade edilir (Yakut vd., 2014).

$$f(\alpha) = \sum_{i=1}^{m} wixi + \theta \quad (4)$$

Karmaşık problem çözümlerinde ise tek katmanlı ağlar yeterli olamayacağından dolayı girdi ve çıktı katmanlarının arasında gizli katmanların devreye girdiği, Şekil 4'te görünen çok katmanlı ağlar kullanılmaktadır.

Şekil 4: Çok Katmanlı Yapay Sinir Ağı Yapısı (Han ve Kamber, 2006)



Bu yapıda ağ, bir veri grubunun sınıf etiketini tahmin edebilmek için katmanlar arasında görev yapan ağırlıkların gerçek değerlerini öğrenme yoluna gider. Giriş katmanında bulunan nöronlardan gelen giriş verileri ağırlıklandırılarak, gizli katmanda bulunan nöronların değerleri hesaplanır. Sonrasında ise gizli katmanda bulunan değerler yine ağırlıklandırılarak, çıkış katmanı verisinin hesaplanmasına çalışılır (Han ve Kamber, 2006). Öğrenme modelinin oluşturulması aşamasında, elde edilen çıktı katmanı verisi ile beklenen çıktı katmanı verisi karşılaştırılarak hata miktarı hesaplanır. Ortaya çıkan hata miktarı doğrultusunda geri besleme yapılarak hata katsayıları hesaplanmaktadır ve sonrasında elde edilen katsayılar ile nöron bağlantılarında kullanılan ağırlık miktarları değiştirilmektedir. Sistem, hesaplanan çıktı katmanı verisi ile beklenen çıktı katmanı verisi arasındaki hata miktarı en aza indirilene kadar bu işlemleri tekrarlamaya devam eder (Arı ve Berberler, 2017).

3.3.3. Lojistik Regresyon (LR)

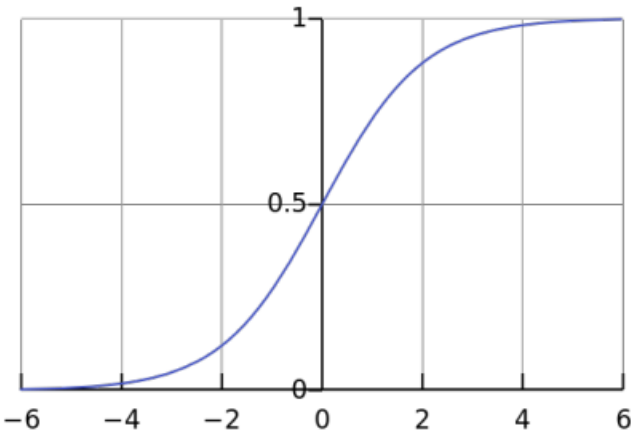
Lojistik Regresyon (LR) veri seti içerisinde bulunan tüm değişkenleri sayısal kabul eden ve normal bir dağılıma sahip olan, ikili sınıflandırma algoritması olarak tanımlanmaktadır. Bu varsayımda belirtilen noktaya rağmen, LR ile normal dağılım göstermeyen veriler üzerinde dahi iyi sonuçlar alınabilmektedir. LR algoritması, doğrusal olarak bir regresyon fonksiyonun içerisinde birleştirilmiş ve bir lojistik fonksiyon kullanılarak dönüştürülmüş her giriş değeri için bir katsayı öğrenmeye dayalı bir sistem üzerine kurulmuştur. Hızlı ve basit bir sisteme sahip olmasına rağmen, bazı problemler üzerinde son derece etkili sonuçlar vermektedir. LR sadece ikili sınıflandırma modellerini desteklemektedir (Brownlee, 2019).

LR, düzeltilmiş olasılık oranları hakkında çıkarım sağlamak için geliştirilmiş standart bir öğrenme algoritmasıdır (Mansournia vd., 2018). LR, Veri setinin kalitesine bağımlı olarak ayrıcı özelliğine sahip bir modeldir. Modelin belirlenmesinde, özellik değerleri (X1, X2,…,Xn), ağırlık değerleri (W1, W2,…,Wn), sapma değerleri (b1, b2,…,bn) ve sınıflar (1 / 0) dikkate alındığında eşitlik (5) kullanılabilmektedir (Hasan vd., 2019):

$$Tahmin\ Edilen\ Değer = p(y = C|X; W, b) = \frac{1}{1+\exp(-w^{transpose}\ X-b)} \quad (5)$$

LR, logaritmik ilerleme gerçekleştiren, logit bir fonksiyondur. LR, yapı itibari ile Şekil 5'teki grafikte gösterilen eğriye en iyi şekilde benzeyecek olan değerleri tespit etmeye çalışmaktadır (Şeker, 2016).

Şekil 5. Lojistik Regresyon Grafiği



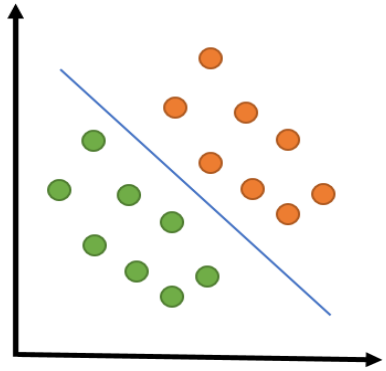
3.3.4 Destek Vektör Makineleri (DVM)

Destek Vektör Makineleri algoritması, LR algoritması gibi ayrımcı bir model olarak çalışmaktadır. Regresyon, aykırı veri tespiti ve sınıflandırma işlemleri için denetimli bir öğrenme sistemi sunmaktadır (Hasan vd., 2019). DVM, 1995 yılında Vladimir Vapnik tarafından tanıtılan, sınıflandırma ve örüntü tanıma süreçleri için kullanılabilen, basit ve verimli bir algoritmadır. DVM çalışma sisteminin ana amacı, hiper düzlemlerin ve sınırların ortaya çıkartacak fonksiyonların elde edilmesidir. Oluşturulan hiper düzlemler, istatistiksel öğrenmeyi kullanmak amacı ile belirli algoritmalar ile eğitilerek, giriş verisi noktalarının farklı kategorilere ayrıştırılması sağlamak için kullanılmaktadır (Jain vd., 2018).

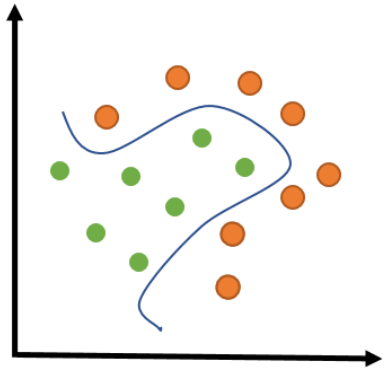
Destek Vektör Makineleri, prensip olarak istatistiksel öğrenme yöntemleri ve yapısal risklerin minimize edilmesine dayalı bir öğrenme algoritmasıdır. Hızlı öğrenme kapasitesine sahip büyük giriş verilerini kullanabilmesi, algoritmanın temel avantajını ortaya çıkartmaktadır. Bu sayede DVM algoritması ile doğrusal olmayan yüksek boyutlu veri modelleme sorunlarına çözüm getirilebilmektedir (Feizizadeh vd., 2017).

DVM, çoğunlukla sınıflandırma işlemlerinde kullanılan bir algoritmadır. DVM algoritmasının çalışma sisteminde, n bağımsız değişken sayısı olmak üzere tüm veri noktalarının n-boyutlu uzaydaki koordinatları değişken değerleri olarak belirlenmektedir. Bir üst aşamaya geçildiğinde ise iki sınıfı birbirinden ayıran iki boyutlu hiper düzlem tespit edilir ve sınıflandırma işlemi, Şekil 6'da gösterildiği gibi gerçekleştirilir. Fakat, elde edilen tüm verilerin doğrusal hiper düzlemler ile sınıflandırılması mümkün değildir. Doğrusal olmayan veri türleri, DVM sınıflandırılmasında kernel fonksiyonu kullanılarak yüksek boyutlu uzayda oluşturulan doğrusal olmayan bir hiper düzlem ile Şekil 7'de görselleştirildiği gibi birbirlerinden ayrılabilir duruma getirilebilmektedirler. Bu konumda uygulanan kernel fonksiyonu, doğrusal modda çalışan bir sınıflandırma algoritmasının, doğrusal olmayan bir problemi çözmesini sağlayacak olan kernel hilesi anlamında kullanılmaktadır (Gürsakal, 2018).

Şekil 6: Doğrusal Hiper Düzlem ile Sınıflandırma



***Şekil 7**: Doğrusal Olmayan Hiper Düzlem ile Sınıflandırma*



3.3.5. Karar Ağacı (KA)

Karar Ağaçları (KA) sınıflandırma ve regresyon problemlerinin çözümlerini destekleyebilmektedir. Veri örneklerini değerlendirmek amacıyla bir ağaç yapısının oluşturulması sistemine dayanmaktadır. Ters çevrilmiş bir ağacın kökü ile yapı başlar ve bir tahmin sonucuna ulaşılana kadar aşağıya doğru devam eder. KA, oluşturulması aşamasında, doğru tahminlere ulaşılabilmesi amacıyla en iyi ayırıcılık özelliğine sahip olan niteliğin tespit edilmesi süreçleri gerçekleştirilmektedir (Brownlee, 2019).

Dağılım tabanlı tahmin etme işlemlerinde, giriş verilerinin tamamı üzerinde bir modelin geçerli olduğu varsayılır ve veri setine ait parametrelerin öğrenilebilmesi için de bütün veri seti kullanılır. Öğrenme işleminin sonrasında, test verilerinde sistemin sınanması aşamasında da aynı yapının ve öğrenilmiş olan parametrelerin kullanımına devam edilir. Dağılıma bağımlı olmayan tahmin etme süreçlerinde ise belirlenmiş bir ölçüt (ör: Öklid) ile öğrenme seti yerel parçalara ayrılmakta ve giriş verisi için, kendi alanına denk gelen verilerle eğitilmiş lokal bir model kullanılmaktadır. KA, yapıları gereği dağılımdan bağımsız çalışmaktadır. Çünkü öğrenme süreçlerinin başında, sınıf dağılımları ile ilgili tahminlerde bulunmaz. Karar ağaçlarında, ağacın yapısı baştan belirli olmamaktadır. Veri setinin özelliklerine ve yapısına göre dallar ve yapraklar eklenerek oluşturulmaktadır (Alpaydın, 2017).

Karar ağacı oluşturulması konusunda, C4.5 ve ID3 gibi çeşitli algoritmalar kullanılmaktadır. Bu çalışma kapsamında C4.5 algoritması kullanılmıştır. C4.5, giriş verilerinin sıklıklarına göre sınıflandırma işlemini gerçekleştiren bir algoritmadır. Entropi hesabına dayalı olarak çalışmaktadır. C4.5 çalışma yapısı şu şekilde özetlenebilir (Aksu ve Karaman, 2017):

- (6) nolu eşitlik kullanılarak veri setinin sahip olduğu tüm nitelikler üzerinde entropi hesabı yapılır,
- Her bir nitelik için hesaplanmış olan entropi değeri, sınıfın entropi değerine bölünerek, niteliklerin bilgi kazanç değerleri ortaya çıkartılır,
- En yüksek bilgi kazancı değerine sahip olan nitelik kök olarak seçilir ve dağılım başlar,
- Kök olarak seçilen niteliğin sahip olduğu değerler haricinde kalan diğer nitelikler için aynı işlemler tekrarlanır,
- Tüm veriler işlenip yapraklara ulaşılınca ağaç tamamlanmış olur.

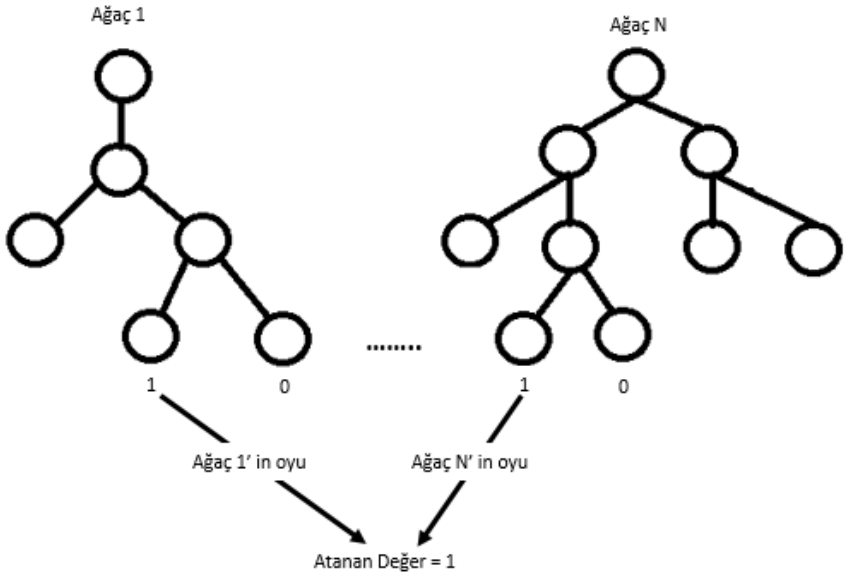
$$Entropi = -\sum_{i=1}^{m} p_i \, log_2(p_i) \quad (6)$$

3.3.6. Rastgele Orman (RO)

Rastgele Orman (RO), birden fazla karar ağacının birleşiminden oluşan bir yapıdır. Orijinal veri içerisinden rastgele çekilmiş birbirlerinden ayrı parçalar kullanılarak, orman içerisinde bulunan ağaçların eğitilmesi esas alınmaktadır. RO içerisinde ağaçların büyümeleri esnasında rastgele özellik seçim işlemi uygulanmaktadır. Bunun da sebebi olarak, çok büyük boyutlardaki veri setleri üzerinden tek bir karar ağacının yeteri kadar sağlıklı sonuç vermesi zor olacağından dolayı, büyük veri setinin parçalara ayrılıp ormandaki karar ağaçlarında ayrı ayrı öğrenme aşamalarının gerçekleştirilmesinin, doğruluk oranlarını arttırması olarak gösterilmektedir (Tanha vd., 2017). Rastgele özellik seçim süreci ile orman içerisindeki ağaçlar, değiştirme yöntemi ile eğitim özelliklerinin alt kümelerinin belirlenmesi ile oluşturulmaktadır. Bu durum, aynı özelliğin birkaç kez seçilebileceği, bazı özelliklerin ise hiç seçilemeyeceği anlamına gelmektedir (Belgiu ve Drăgut, 2016). Bu sayede daha hızlı ve gürültüye daha fazla dayanıklı öğrenme modelleri oluşturulabilmektedir. Buna ek olarak orman içerisindeki ağaçlar**,** modele esneklik kazandırdıkları için, sınıflandırma, kümeleme ve regresyon işlemlerinin performansını arttırmaktadır. Özellikle büyük veri setlerinin değerlendirilmesi aşamasında, iyi bir seçim olarak değerlendirilmektedir (Holzinger vd., 2017).

RO, ağaç yapılı sınıflandırıcıların birleşiminden oluşan bir sınıflandırma algoritmasıdır. Bahsi geçen ağaç yapılı sınıflandırıcılar, birbirlerinden bağımsız ancak, aynı şekilde dağıtılmış rastgele vektörler içermektedir ve yapısı içerisinde bulunan her ağaç, giriş verilerindeki en popüler sınıf değeri için oy verme süreci uygulamaktadır (Breiman, 2001). Ağaçların kullandıkları oylar sonucunda daha fazla ağaçtan oy almış olan sınıf etiketi, RO algoritması tarafından belirlenmiş etiket değeri olarak gösterilir (Belgiu ve Drăgut, 2016). Şekil 8'de RO görsel örneği verilmiştir.

Şekil 8: Rastgele Orman Algoritması, Sınıflandırma Şeması



3.3.7. k - En Yakın Komşu Algoritması (k-NN)

k-NN algoritması, sınıflandırma teknikleri altında benzerlik fonksiyonlarını çalıştıran ve bu şekilde tahmin süreçlerini çalıştıran bir algoritmadır. 1950'li yıllarda keşfedilmiş ve hala günümüzde popüler olarak kullanımı devam etmektedir. Çalışma mantığına, iki boyutlu bir düzlemde bakmak gerekmektedir. Bunun için Şekil 9'da görüldüğü gibi veriler iki boyutlu bir düzleme yerleştirilmekte ve dikey ve yatay eksen (x ve y) değerlerine göre, benzerlik değerleri hesaplanmaktadır (Şeker, 2016).

k-En Yakın Komşu Algoritması (k-NN), makine öğrenmesi algoritmalarının içerisinde en ılımlı çalışan sınıflandırıcı türüdür. k ile ifade edilen benzerlik vektörlerine dayanan çalışma sisteminde, bir nesnenin sınıflandırılması sürecinde k adet komşularının içerisindeki en çok oy alan sınıf değeri kullanılmaktadır. Sınıfı tahmin edilecek olan değerin komşularının adedini belirten k, genellikle çok küçük pozitif bir tamsayıdır (Harefa vd., 2016). k-NN algoritması, sınıf etiketi bilinmeyen yeni bir örnek veri giriş yapıldığında, bu verinin sınıflandırılma sürecinin yakın komşu konumunda bulunan çoğunluk tarafından gerçekleştirilmesi sağlayan denetimli bir algoritma olarak görev yapmaktadır. k-NN ile, test verileri ile algoritmanın sınanması süreçlerinde ve yeni girişi yapılan verinin sınıf etiketinin tespit edilmesi süreçlerinde, bu verilerin öğrenme kümesindeki veriler ile aralarındaki mesafeyi ölçerek, test edilecek veya sınıfı belirlenecek veriye en yakın olan k adet komşuyu bulmaya odaklanmaktadır. Burada bahsedilen mesafenin hesaplanması konusunda da genel anlamda Öklid eşitliğinin (7) kullanılması tercih edilmektedir (Indriani vd., 2017).

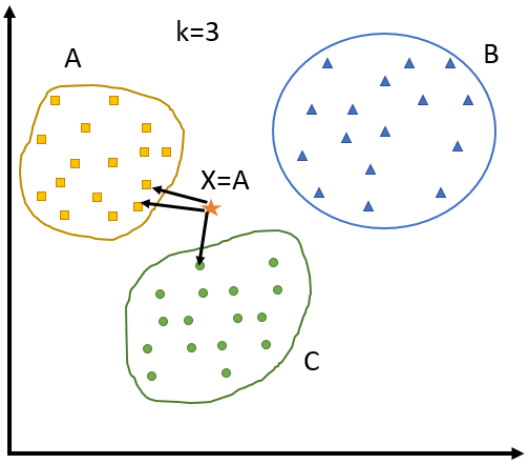
$$E(i,j) = \sqrt{\sum_{k=1}^{n}(i_{k-}\,_{j_k})^2} \tag{7}$$

k-NN, temel manada verilerin belirli bir özellik alanına ait olduğunu varsaymaktadır. Bu sebeple, yeni girişi yapılan veri noktasının, öğrenme kümesindeki tüm noktalara olan uzaklıkları tek tek dikkate alınmaktadır. Yeni girilen verinin sınıf değerinin tespit edilmesi konusunda da uzaklıkları hesaplanmış olan komşular için başta belirlenmiş olan k değeri baz alınır. Eğer k=1 olarak belirlenmişse, sadece en yakın mesafedeki komşunun sınıf değerine göre karar verilir. Uygulamadaki k değerinin optimum seviyede belirlenmesi önemlidir. Çünkü, k çok küçük olursa,

algoritma öğrenme kümesinde bulunan gürültüden çok fazla etkilenecektir. k değeri çok büyük olursa, aslında çok uzak olan komşularda yakın olarak değerlendirilecek ve bunun sonucunda verinin sınıfının tahmininde hata yapma ihtimali artacaktır. k-NN algoritmasının çalışması şu şekilde sıralanabilir (Jadhav ve Channe, 2016):

1. k değeri başlatılır,
2. Giriş verisi ve öğrenme kümesi verileri aralarındaki mesafeler ölçülür,
3. Ölçülen mesafe değerleri sıralanır,
4. k adet en yakın komşular belirlenir,
5. Şekil 9'da ifade edildiği gibi, basit çoğunluk sistemi ile komşular arasında en fazla bulunan sınıf değeri, giriş verisi için tahmin edilir.

Şekil 9: k-NN Algoritması için Örnek Şema (Jadhav ve Channe, 2016).



3.4. Değerlendirme

Çalışma kapsamında kullanılan veri setleri üzerinde makine öğrenmesi algoritmalarının sınanması sürecinde aşağıda belirtilen ölçüm birimleri kullanılmıştır. Bu ölçümlerden elde edilen sonuçlar doğrultusunda, veri setleri üzerinde en uygun değerlemeyi yapan öğrenme algoritmasına karar verilebilmektedir.

3.4.1. Karmaşıklık Matrisi

Karmaşıklık matrisi, öğrenme modellerinin performanslarının görselleştirildiği bir düzen olarak kullanılmaktadır. Herhangi bir algoritma ile sınıflandırma işlemi yapabilmek için oluşturulmuş olan öğrenme modellerinin, gerçek sınıf değerleri bilinen test verileri üzerinde sınanmaları sonucunda, modelin başarımını gösterir. Karmaşıklık matrisi tarafından model performansının belirlenmesi amacıyla yapılan tanımlamalar şu şekildedir (Hasan vd., 2019):

- True Positive (TP): Gerçek Pozitif- Gerçekte 1 olan sınıf etiketlerinin, 1 olarak tahmin edilme sayısı,
- True Negative (TN): Gerçek Negatif- Gerçekte 1 olmayan sınıf etiketlerinin, 1 olarak tahmin edilmemesi,
- False Positive (FP): Yanlış Pozitif- Gerçekte 1 olmayan sınıf etiketlerinin, 1 olarak tahmin edilmesi,
- False Negative (FN): Yanlış Negatif- Gerçekte 1 olan sınıf etiketlerinin, 1 olarak tahmin edilmemesidir.

Tablo 2 ve Tablo 3'te çalışma kapsamında kullanılan veri setlerine ait karmaşıklık matrisleri verilmiştir.

Tablo 2. Standart Veri Seti, Sınıflandırma Algoritmalarına Göre Karmaşıklık Matrisleri

NB	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 16283	FP= 1692
Anomali Var	FN= 4786	TP= 884

YSA	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 17818	FP= 157
Anomali Var	FN= 1527	TP= 4143

LR	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 17947	FP= 28
Anomali Var	FN= 5474	TP= 196

DVM	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 17975	FP= 0
Anomali Var	FN= 5639	TP= 31

C4.5	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 17723	FP= 252
Anomali Var	FN= 403	TP= 5267

RO	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 17871	FP= 104
Anomali Var	FN= 195	TP= 5475

k-NN	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 17834	FP= 141
Anomali Var	FN= 579	TP= 5091

Tablo 3. Optimize Veri Seti, Sınıflandırma Algoritmalarına Göre Karmaşıklık Matrisleri

NB	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 12032	FP= 3085
Anomali Var	FN= 2602	TP= 1915

YSA	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 14944	FP= 173
Anomali Var	FN= 1563	TP= 2954

LR	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 15041	FP= 76
Anomali Var	FN= 4376	TP= 141

DVM	Öngörülen Sınıf Negatif	Öngörülen Sınıf Pozitif
Gerçek Sınıf Negatif	TN= 15114	FP= 3
Gerçek Sınıf Pozitif	FN= 4376	TP= 141

C4.5	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 14894	FP= 223
Anomali Var	FN= 354	TP= 4163

RO	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 14999	FP= 118
Anomali Var	FN= 171	TP= 4346

k-NN	Öngörülen Anomali Yok	Öngörülen Anomali Var
Anomali Yok	TN= 14919	FP= 198
Anomali Var	FN= 553	TP= 3964

3.4.2. Ölçüt İfadeleri

Çalışma kapsamında, karmaşıklık matrisinden elde edilen değerlerin kullanıldığı ve model performansını değerlendirme amacıyla kullanılacak ölçüt ifadeleri ve eşitlikleri şu şekildedir:

- **Model Doğruluğu:** Öğrenme modelinin doğruluk derecesini belirler.

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

- **Model Kesinliği:** Öğrenme modelinin duyarlılık derecesini ölçer.

$$Kesinlik = \frac{TP}{TP+FP} \quad (9)$$

- **Modelin Duyarlılığı:** Test sonucunda gerçek pozitif değerlerin oranı olarak bilinmektedir.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (10)$$

- **F Ölçütü:** Duyarlılık ve Kesinlik değerlerinin harmonik ortalamasıdır.

$$F\ Ölçütü = \frac{2*Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \quad (11)$$

- **Hata Oranı:** Hatalı ölçümlerin, toplam değer sayısına olan oranının ölçümüdür.

$$Hata\ Oranı = \frac{FP+FN}{TP+TN+FP+FN} \quad (12)$$

3.5. Uygulama

Çalışmada belirtilen veri setleri üzerinde, ilgili algoritmaların uygulanması, öğrenme modellerinin oluşturulması ve sonuçların elde edilmesi süreçlerinde, Intel Core i5-3230M model, 2.6 Ghz frekansa sahip çift çekirdekli CPU, 6 Gbyte 1600 Mhz RAM bellek, Intel HD Graphics 4000 paylaşımlı ekran kartı özelliklerine sahip bir dizüstü bilgisayar kullanılmıştır. Öğrenme modellerinin oluşturulmasında, açık kaynak kodlu Weka yazılımı kullanılmıştır.

3.5.1. Ölçümlerin Analizi

Tablo 4, Tablo 5 ve Şekil 10' da gösterilen sonuçlar incelendiğinde, Yapay Sinir Ağları, C4.5 Karar Ağacı, Rastgele Orman ve k En Yakın Komşu algoritmaları tarafından oluşturulan öğrenme modellerinin yüksek başarı gösterdiği görülmektedir. Ayrıca bu modeller içerisinden, standart veri setinin test kümesinde bulunan 1148 adet anomali verisinin 1099 adedini, optimize edilmiş veri setinin test kümesinde bulunan 938 adet anomali verisinin 891 adedini tahmin ederek, en yüksek doğruluk oranı ile en iyi tahmin modelinin Rastgele Orman algoritması tarafından oluşturulduğunu söyleyebiliriz.

Bunların dışında kalan Naive Bayes, Lojistik Regresyon ve Destek Vektör Makineleri algoritmaları ise başarılı tahmin konusunda önemli ölçüde geride kalmış durumdadırlar. Özellikle her iki veri seti için de gerçekte anomali olan çok yüksek miktarda verinin, bu algoritmalar tarafından normal veri olarak tahmin edildiği görülmüştür. Sonuç olarak adı geçen bu algoritmalar, veri setlerinde belirtilmiş olan anomali verilerinin tespit edilmesi konusunda büyük oranda başarısız olmuşlardır.

Ayrıca algoritmalar üzerinde yapılan karşılaştırmalar sonucunda, standart çalışma düzeni ile oluşturulmuş olan veri seti ile optimize edilmiş çalışma düzeni ile oluşturulmuş olan veri setinin ölçümlerinde kayda değer bir fark gözlenmemiştir.

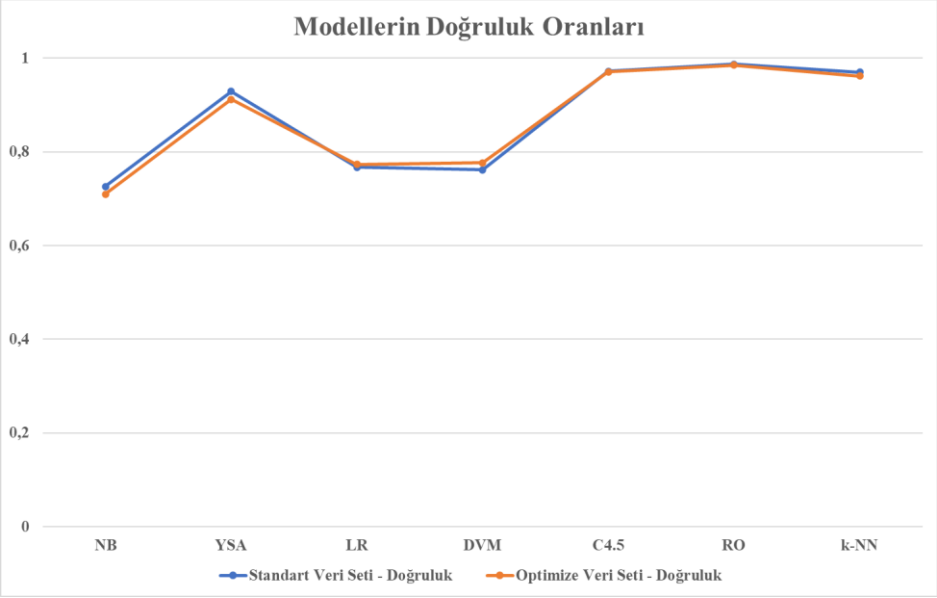
Tablo 4. Standart Veri Seti, Uygulama Sonuçları

	Doğruluk	**Kesinlik**	**Duyarlılık**	**F-Ölçümü**	**Hata Oranı**
NB	0,726	0,343	0,156	0,214	0,274
YSA	0,929	0,963	0,731	0,831	0,071
LR	0,767	0,875	0,035	0,067	0,233
DVM	0,762	1,000	0,005	0,011	0,238
C4.5	0,972	0,954	0,929	0,941	0,028
RO	0,987	0,981	0,966	0,973	0,013
k-NN	0,970	0,973	0,898	0,934	0,030

Tablo 5. Optimize Veri Seti, Uygulama Sonuçları

	Doğruluk	**Kesinlik**	**Duyarlılık**	**F-Ölçümü**	**Hata Oranı**
NB	0,710	0,383	0,424	0,402	0,290
YSA	0,912	0,945	0,654	0,773	0,088
LR	0,773	0,650	0,031	0,060	0,227
DVM	0,777	0,979	0,031	0,061	0,223
C4.5	0,971	0,949	0,922	0,935	0,029
RO	0,985	0,974	0,962	0,968	0,015
k-NN	0,962	0,952	0,878	0,913	0,038

Şekil 10: Öğrenme Modellerinin, Standart Veri Seti ve Optimize Veri Seti Üzerindeki Doğruluk Oranları Karşılaştırması



4. TARTIŞMA VE SONUÇ

Çalışma kapsamında, kaggle ortamında bulunan, enerji optimizasyonu üzerine hazırlanmış bir konveyör bant sistemi tarafından oluşturulmuş, çeşitli anomaliler içeren iki farklı açık kaynak veri seti temin edilerek kullanılmıştır. Birinci veri setinde, depolama sisteminin ortasındaki iki bant dikey yönlü hareket esnasında yatay yönlü hareket sergilemeden çalışmış ve bu şekilde enerji optimizasyonu değerlendirilmiştir. İkinci veri setinde ise ortadaki iki bant dikey ve yatay hareketleri aynı anda gerçekleştirmesi ile enerji optimizasyonu değerlendirilmiştir. Temin edilen bu veri setleri üzerinde, Naive Bayes, Yapay Sinir Ağları, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağacı, Rastgele Orman ve k En Yakın Komşu isimli sınıflandırma algoritmaları kullanılarak öğrenme modelleri oluşturuşmuş ve veri setlerinde var olan anomali verilerinin tespit edilmesi noktasında öğrenme modelleri test edilerek karşılaştırılmıştır.

Ayrıca ilgili veri setleri, Hranisavljevic ve ekibi tarafından önerilen Derin Ağ Zamanlı Otomat (DENTA) isimli bir modelin testinde kullanılmış ve sonucunda özellikle gerçek dünya veri kümeleri üzerinde anomali tespiti konusunda avantajlı sonuçların elde edildiği bildirilmiştir (Hranisavljevic vd., 2020). Kim ve ekibi tarafından önerilen Projeksiyon Yolu Boyunca Yeniden Yapılanma (RaPP) isimli yenilik tespit sistemi modeli kapsamında oluşturdukları otomatik kod çözücülerin (AE) değerlendirilmesinde ise çok çeşitli özelliklere sahip veri setleri sınanmıştır. Bu sınama sürecinde kullanılan veri setlerinden bir tanesi de bu çalışmada analiz edilen yüksek raflı depolama sistemlerinin enerji optimizasyonu için oluşturulmuş olan veri kümesidir. Genel anlamda önerilen modelin analizler sonucunda iyi bir performans gösterdiği ifade edilmiştir (Kim vd, 2019). Shin ve Kim tarafından yapılan çalışmada, yine RaPP için geliştirilen genişletilmiş otomatik kod çözücünün (XAE) performansının analiz edilmesi amacıyla, aynı veri setleri kullanılmış ve genel manada önceki çalışmaya göre bir miktar daha iyi doğruluk sonuçlarına erişildiği gösterilmiştir (Shin ve Kim, 2020). Bahsi geçen çalışmalar ile bu çalışmada elde edilen en yüksek doğruluk oranlarına sahip Rastgele Orman algoritmasının değerlerinin karşılaştırmaları Tablo 6'da verilmiştir.

Tablo 6. Farklı Çalışmalara Ait Sonuçların Karşılaştırılması

	RO (Standart Veri Seti)	**RO (Optimize Veri Seti)**	**DENTA**	**RaPP - AE**	**RaPP - XAE**
Doğruluk Oranı	0,987	0,985	0,812	0,650	0,631

Sonuç olarak, çalışma içerisinde yaptığımız analizler sonucunda, Yapay Sinir Ağları, Karar Ağacı, Rastgele Orman ve k En Yakın Komşu algoritmaları tarafından oluşturulmuş olan öğrenme modelleri anomali tespiti konusunda başarı elde etmiştir. Ayrıca bu algoritmalar içerisinde doğruluk oranı ve F-Ölçümü değeri ile Rastgele Orman algoritmasının öğrenme modeli hem bu çalışmada, hem de diğer çalışmalarda uygulanan modeller arasında en iyi anomali tespiti yapan model olmuştur.

Çalışmanın bundan sonraki süreçlerinde, zeki fabrikalar içerisinde çalışan siber fiziksel sistemler üzerinde, özellikle siber saldırılar tarafından meydana gelebilecek anomalilerin anlık olarak tespit edilmesi ile ilgili çalışmaların yapılması planlanmaktadır.

KAYNAKÇA

Akçetin, E., & Çelik, U. (2014). İstenmeyen Elektronik Posta (Spam) Tespitinde Karar Ağacı Algoritmalarının Performans Kıyaslaması. *İnternet Uygulamaları ve Yönetimi, 5*(2), 43-56.

Aksu, M. Ç., & Karaman, E. (2017). Karar Ağaçları ile Bir Web Sitesinde Link Analizi ve Tespiti. *ACTA INFOLOGICA, 1*(2), 84-91.

Alpaydın, E. (2017). *Yapay Öğrenme.* İstanbul: Boğaziçi Üniversitesi Yayınevi.

Arı, A., & Berberler, M. E. (2017). Yapay Sinir Ağları ile Tahmin ve Sınıflandırma Problemlerinin Çözümü İçin Arayüz Tasarımı. *Acta - Infologica, 1*(2), 55-73.

Aydemir, E. (2019). *Weka ile Yapay Zeka.* Ankara: Seçkin Yayıncılık.

Bagozi, A., Bianchini, D., Antonellis, V. D., Marini, A., & Ragazzi, D. (2017). Big Data Summarisation and Relevance Evaluation for Anomaly Detection in Cyber Physical Systems. *OTM 2017: On the Move to Meaningful Internet Systems* (s. 429-447). Rhodes, Greece: Springer.

Belgiu, M., & Drăgut, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*(114), 24-31.

Birgelen, A. v., & Niggeman, O. (2017). Using self-organizing maps to learn hybrid timed automata in absence of discrete events. *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (s. 1-8). Limassol: IEEE.

Breiman, L. (2001). Random Forests. *Machine Learning*(45), 5-32.

Brownlee, J. (2019). *Machine Learning Mastery With Weka.*

Chen, B., Wan, J., Shu, L., Li, P., Mukherjee, M., & Yin, B. (2017). Smart Factory of Industry 4.0: Key Technologies, Application Case, and Challenges. *IEEE Access*(6), 6505-6519.

Elmas, Ç. (2016). *Yapay Zeka Uygulamaları 3. Baskı.* Ankara: Seçkin Yayıncılık.

Feizizadeh, B., Roodposhti, M. S., Blaschke, T., & Aryal, J. (2017). Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping. *Arabian Journal of Geosciences, 10*(117).

Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics*(210), 15-26.

Gürsakal, N. (2018). *Makine Öğrenmesi.* Bursa: Dora Yayınevi.

Han, J., & Kamber, M. (2006). *Data Mining, Concepts and Techniques 2nd Edition.* San Francisco: Morgan Kaufmann Publishers.

Hand, D., Manila, H., & Smyth, P. (2001). *Principles of Data Mining.* London: Massachusetts Institute of Technology.

Harefa, J., Alexander, A., & Pratiwi, M. (2016). Comparison Classifier: Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) In Digital Mammogram Images. *Jurnal Informatika dan Sistem Informasi, 2*(2), 35-40.

Hasan, M., Islam, M. M., Zarif, M. I., & Hashem, M. (2019). Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. *Internet of Things*, 1-14.

Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., & Zatloukal, K. (2017). Machine Learning and Knowledge Extraction in Digital Pathology Needs an Integrative Approach. A. Holzinger, R. Goebel, M. Ferri, & V. Palade içinde, *Towards Integrative Machine Learning and Knowledge Extraction* (s. 13-50). Springer.

Hranisavljevic, N., Maier, A., & Niggeman, O. (2020). Discretization of hybrid CPPS data into timed automaton using restricted Boltzmann machines. *Engineering Applications of Artificial Intelligence*(95), 1-9.

Hranisavljevic, N., Niggemann, O., & Maier, A. (2016). A Novel Anomaly Detection Algorithm for Hybrid Production Systems based on Deep Learning and Timed Automata. *The 27th International Workshop on Principles of Diagnosis: DX*. Denver, USA.

Hranisavljevic, N., Niggemann, O., & Maier, A. (2018, 07 19). *High Storage System Data for Energy Optimization*. 03 15, 2020 tarihinde Kaggle: https://www.kaggle.com/inIT-OWL/high-storage-system-data-for-energy-optimization adresinden alındı

Hsieh, R.-J., Chou, J., & Ho, C.-H. (2019). Unsupervised Online Anomaly Detection on Multivariate Sensing Time Series Data for Smart Manufacturing. *IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA)* (s. 90-97). Kaohsiung, Taiwan: IEEE.

Indriani, O. R., Kusuma, E. J., Sari, C. A., Rachmawanto, E. H., & Setiadi, D. I. (2017). Tomatoes classification using K-NN based on GLCM and HSV color space. *International Conference on Innovative and Creative Information Technology (ICITech)* (s. 1-6). Salatiga: IEEE.

İşçimen, B., Kutlu, Y., Reyhaniye, A. N., & Turan, C. (2014). Balık tanınmasında görüntü analiz yöntemleri. *22nd Signal Processing and Communications Applications Conference.* Trabzon.

Jadhav, S. D., & Channe, S. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research*, *5*(1), 1842-1845.

Jain, M., Narayan, S., Pratibha, B., Bhowmick, A., & Muthu, R. (2018). Speech Emotion Recognition using Support Vector Machine. *International Conference on Informatics Computing in Engineering Systems (ICICES).* Chennai, India: IEEE.

Kim, K. H., Shim, S., Lim, Y., Jeon, J., Choi, J., Kim, B., & Yoon, A. S. (2019). RaPP: Novelty Detection with Reconstruction along Projection Pathway. *International Conference on Learning Representations (ICLR 2020),* (s. 1-10). Addis Ababa, Ethiopia.

Mansournia, M. A., Geroldinger, A., Greenland, S., & Heinze, G. (2018). Separation in Logistic Regression: Causes, Consequences, and Control. *American Journal of Epidemiology, 187*(4), 864-870.

Nizam, H., & Akın, S. S. (2014). Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı.* İzmir.

Pahl, M.-O., & Aubet, F.-X. (2018). All Eyes on You: Distributed Multi-Dimensional IoT Microservice Anomaly Detection. *14th International Conference on Network and Service Management (CNSM 2018)* (s. 72-80). Italy: Aconf.

Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications, 6*(2), 256-261.

Radziwon, A., Bilberg, A., Bogers, M., & Madsen, E. S. (2014). The Smart Factory: Exploring Adaptive and Flexible Manufacturing Solutions. *Procedia Engineering*(69), 1184-1190.

Riordan, A. O., Coady, J., Toal, D., Newe, T., & Dooly, G. (2019). Industry 4.0: Pillars for Smart Manufacturing - A Review. *no. February*.

Shin, S. Y., & Kim, H.-J. (2020). Extended Autoencoder for Novelty Detection with Reconstruction along Projection Pathway. *Applied Sciences, 10*(13), 1-14.

Staub, S., Karaman, E., Kaya, S., Karapınar, H., & Güven, E. (2015). Artificial Neural Network and Agility. *Procedia - Social and Behavioral Sciences, 195*, 1477-1485.

Şeker, H. İ., Tuna, M., & Koyuncu, İ. (2018). Gerçek Zamanlı Wavelet Dönüşümleri için FPGA-Tabanlı Meksika Şapkası Dalgacığının Tasarımı ve Gerçeklenmesi. *3rd International Conference on Engineering Technology and Applied Sciences (ICETAS)* , (s. 168-173). Skopje Macedonia.

Şeker, Ş. E. (2016). *Weka ile Veri Madenciliği.* İstanbul: Bilgisayar Kavramları Yayınları.

Tanha, J., Someren, M. V., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*(8), 355-370.

Wan, J., Li, J., Imran, M., Li, D., & Amin, F.-e. (2019). A Blockchain-Based Solution for Enhancing Security and Privacy in Smart Factory. *IEEE Transactions on Industrial Informatics, 15*(6), 3652-3660.

Wang, R., Nie, K., Wang, T., Yang, Y., & Long, B. (2020). Deep Learning for Anomaly Detection. *13th International Conference on Web Search and Data Mining* (s. 894-896). Houston, TX, USA: WSDM.

Yakut, E., Elmas, B., & Yavuz, S. (2014). Yapay Sinir Ağları ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeks Tahmini. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 19*(1), 139-157.

Yoon, S., Um, J., Suh, S.-H., Stroud, I., & Yoon, J.-S. (2019). Smart Factory Information Service Bus (SIBUS) for manufacturing application: requirement, architecture and implementation. *Journal of Intelligent Manufacturing*(30), 363-382.

İŞ SAĞLIĞI VE GÜVENLİĞİ ÖNLEMLERİNİN ETKİNLİĞİNİN GÖZ İZLEME CİHAZI İLE BELİRLENMESİ[+]

***

DETERMINATION OF THE EFFECTIVENESS OF OCCUPATIONAL HEALTH AND SAFETY MEASURES WITH EYE TRACKING DEVICE

Ömer Çağrı YAVUZ[*]
Pınar BAYKAN[**]
Ersin KARAMAN[***]

DOI: 10.33461/uybisbbd.777525

Öz

Bilişim sistemleri, iş sağlığı ve güvenliği alanındaki kontrol kayıtları, eğitimler ve çevre ölçümlerinde de kullanılmaktadır. Bununla birlikte, insan-bilgisayar etkileşimi alanında kullanılan kullanılabilirlik test yöntemlerinin iş sağlığı ve güvenliği (İSG) çalışmalarına uygulanması konusunda literatürde bir boşluk olduğu görülmektedir. Bu çalışmada iş sağlığı ve güvenliği önlemlerinin etkinliğinin kullanılabilirlik test yaklaşımı ile tespit edilmesi amaçlanmıştır. Kullanılabilirlik testi göz izleme cihazları yardımı ile gerçekleştirilmiştir. Bu deneysel araştırmaya 26 katılımcı dâhil edilmiştir. Çalışmada göz izleme cihazı takmış olan katılımcılardan yangın varmış gibi davranarak çıkışa yönelmeleri istenmiştir. İki farklı kurumda üç farklı deney yapılmıştır. Katılımcıların yangın ekipmanları, alarm butonları ve çıkış işaretlerine odaklanma süreleri ve odaklanma sayıları incelenmiştir. Ek olarak 12 İSG uzmanı ile görüşülerek uzman görüşleri çalışmaya dâhil edilmiştir. Sonuç olarak risk analizlerinin tek başına yeterli olmadığı ve risk analizi sonucunda alınması gereken önlemlerin tespitinde bilişim sistemleriyle yapılacak olan daha kapsamlı analizlere olan ihtiyaç ortaya koyulmuştur. Çalışmada önerilen yöntemin diğer acil durum ve afet bağlamında da değerlendirilebileceği önerilmiştir.

***Anahtar Kelimeler**: Göz Izleme Cihazı, Kullanılabilirlik, Insan-Bilgisayar Etkileşimi, Iş Sağlığı Ve Güvenliği.*

Abstract

Information systems are also used in control records, trainings and environment measurements in the field of occupational health and safety. Moreover, it seems to be a gap in the literature regarding the application of usability testing methods utilized in the field of human-computer interaction in to occupational health and safety (OHS) studies. In this study, it is aimed to determine the effectiveness of occupational health and safety measures via usability test approach. Usability testing was carried out with the help of eye tracking devices. In this experiment research, 26 people were participated. Participants wearing eye tracking devices were asked to find the exit assuming there is a fire. Three experiments were designed in two different institutions. The number of focuses and focus times for fire equipment, panic button and caution signs were examined. In addition, 12 OHS experts' opinions were included in the study. As a result imply that the risk analysis is not enough alone and more comprehensive analysis with information systems is required to improve the precautions to be taken as a result of the risk analysis. It is suggested that the method proposed in the study can also be employed in the context of other emergency cases and disasters.

***Keywords:** Eye Tracking Device, Usability, Human-Computer Interaction, Occupational Health And Safety.*

[+] Bu çalışma Atatürk Üniversitesi Yönetim Bilişim Sistemleri Anabilimdalı'nda tamamlanan "İş Sağlığı ve Güvenliği Önlemlerinin Etkinliklerinin Göz İzleme Cihazı ile Belirlenmesi" adlı yüksek lisans tezinden üretilmiştir.

* Arş. Gör. Karadeniz Teknik Üniversitesi, İİBF, Yönetim Bilişim Sistemleri Bölümü, omercagriyavuz@ktu.edu.tr, Trabzon,Turkiye, ORCID: 0000-0002-6655-3754

** Dr. Öğr. Üyesi Ağrı İbrahim Çeçen Üniversitesi, MYO, Mülkiyet Koruma ve Güvenlik, pbaykan@agri.edu.tr, Ağrı,Türkiye, ORCID: 0000-0001-5279-3872

*** Doç. Dr. Ankara Hacı Bayram Veli Üniversitesi, İİBF, Yönetim Bilişim Sistemleri Bölümü, ersinkaraman@atauni.edu.tr, Ankara,Türkiye, ORCID: 0000-0001-5459-4172

Geliş Tarihi / Received: 08.08.2020
Kabul Tarihi / Accepted: 29.11.2020

1. INTRODUCTION

It occurs that the methods, tools and concepts used in the information systems (IS) have started to be used and expanded in many areas. Various applications are developed in order to contribute to the decision making process which is one of the main objectives of information systems. Therefore almost all fields required to utilize IS tools and methods. In the scope of this study, occupational health and safety (OHS) is one of the field which is handled from this perspective.

The increase in the number of occupational accidents and occupational diseases arising from the technological development has increased the significance of OHS practices. A number of obligations have been introduced to the employers with the regulations related with OHS (Risk Assessment Regulation, 2012). One of the most important of these obligations is risk assessment. This new proactive regulations sway managers to take precautions before accidents occur in the workplace. Thus, dangers and risks in the work environment should be identified properly to minimize them. It is also an obligation that occupational health and safety professionals should be appointed when these processes are carried out. The employer should also receive required services from the authorized health and safety unit.

Occupational health professionals benefit from information systems in activities such as basic occupational health and safety (OHS) training, maintenance and periodic control records, health surveillance and work environment measurements. Computer software that can be used for managing these operations also helps OHS professionals and ensures that they do not make mistakes. However, there are limited IS research and applications that can be used a by OHS professionals other than such software or other basic operational level applications. This is also an evidence of originality of this study.

On the other hand, research methods and tools in information systems (IS) may provide important contributions to managerial decision making processes almost in all sector. For instance, eye tracking devices can be used to understand customers' attentions and focus in business environment such as retailing, digital interfaces and media studies. Such studies motivate us to study on understanding people awareness to the precautions taken by OHS context. So in this study it is focused on evaluating OHS precautions from usability testing perspectives.

In usability testing, eye tracking is one of the methods which is also used in different areas. Kupper, who made one of his first studies in this area, made it possible to watch his eye movements slowly in 1989. In Kupper's study, layout of titles and images evaluated based on eye movements in general (Ömür & Aydoğdu, 2017). In another similar study, the effect of local design factors on the visual behavior of readers was investigated (Holmqvist & Wartenberg, 2005). In another study, the availability of the Mazda company's website was tested with eye tracking technology (Centaur Communication, 2005, as cited in Özdoğan, 2008).

In the risk analyzes conducted within the OHS, the working environment is generally taken into consideration but the actual behavior of the employees, performance losses, psychological situation and individual faults kept in background. Risks can be analyzed by examining the physical and cognitive behaviours of the users are through usability test approaches, which are handled within the scope of human computer interaction. Through applications made with the help of eye tracking devices frequently encountered in usability studies, real behavior of users can be determined and more comprehensive data can be obtained. In this study, it is aimed to determine the OHS awareness of the employees by examining the real behaviors of the employees with the help of eye tracking device and to increase the efficiency of the risk analysis to be carried out with in this context.

As a result of the literature research, it can be said that the usability tests and eye tracking devices applied within the scope of human computer interaction have not been used in occupational

health and safety applications. However, the eye tracking device is used in various applications in different areas. Karaman et al. (2016) aimed to compare the reading performances of primary school students between normal writing and script. In another study, it was aimed to evaluate the usability of a library website with the tobii pro X2-30 eye tracker. As a result of the usability test consisting of seven tasks, suggestions were made to improve the usability of the website (Ritthiron & Jiamsanguanwong, 2017). Kaya (2007) conducted a sports science study which focuses on investigation of the influence of eye movements on multi ball training in table tennis players. İnce and Göktürk (2009) aimed to examine attention levels of the surveillance personnel to the changes. It is intended to offer a method to reduce negativity in case of a reduction in employee attention levels. In the study of Zambarbieri et al. (2008), the search behavior of the users and the reading status of the online newspaper pages over two online newspaper sites were investigated with the help of an eye tracking device. Kalaycı et al. (2011) have aimed to examine the usability of 3D virtual environments with eye tracking method.

In addition, the eye tracking device is one of the tools used in neuromarketing. In this context, studies using eye tracking devices in the field of neuromarketing by Yücel and Coşkun (2018) are discussed. For example, with the eye tracking device, the places focused on the product packaging and the places that are focused while walking around the market can be detected. According to a study conducted with an eye tracking device, the increase in sales from 28 percent to 44 percent by changing the location of the brand logo on the package reveals the importance of the eye tracker device (Girişken, 2015).

As can be seen, the focus time of the participants, number of focusing and striking objects were taken into consideration in general. Additionally, Rashid et al. (2013) aimed to examine the usability of occupational health and safety websites. Although the usability of websites in the field of OHS is examined, no study evaluating the usability of OHS measures has been found in the literature. Thus, this study is specific and original in terms of examining the usability of OHS measures.

As a result of risk assessment, the locations of emergency exit plates, emergency exit doors, emergency buttons and fire intervention equipment have investigated. The measurements have made within the educational institutions. In this study, two institutions providing associate degree and bachelor degree level education have taken into consideration. It is assumed that 26 employees included in the study acted in accordance with the working principles and behaved honestly. Occupational health and safety measures have only been used in the areas related to fire. Instead of a detailed risk assessment, the L type matrix risk analysis method was applied to identify the hazards that may occur only in the event of a possible fire and to take measures against the identified hazards. This study has also focused on the benefits of eye tracking technology in applied institutions taking into consideration spatial individuality.

2. MATERIAL AND METHOD

Usability is the process of measuring the time required to achieve the desired goals, the money spent to achieve the goals, the mental effort to achieve the goals, and the effectiveness of the system (Evcil ve İslim, 2012). Usability tests are one of the most effective methods used to identify usability issues that provide a realistic experience before a product is released. Usability tests provide researchers with direct information on how the system is used and help identify unforeseen problems during evaluation (Kaplan, 2015). Usability testing approaches are divided into four: design guide-based approach, expert approach, user-based approach and model-based approach (Çağıltay, 2011). In this study, it is aimed to determine the effectiveness of occupational health and safety measures with user-based (experimental) test approach. In this context, the applicability of eye tracking devices in the context of occupational health and safety has also been investigated. To

do so three experiments including same procedure were conducted. Before the experiments, not only primary risk analysis was carried out with the help of a checklist but also risk analysis was carried out using the L type matrix risk analysis method.

A total of 26 people, eight female an 18 male, have participated the experiments. They were between 24 and 56 years old. Those of six have PhD degrees, seven participants have master degree, seven people have under graduate degree whereas remaining six people have had compulsory schooling. All participants have normal vision and have no experience with eye-tracking studies. During the experiments participants who were wearing eye tracking devices were asked to go to the exit of the building and behave as if in the case of fire.

Experiments were conducted two different locations and with three different participants group. In the first experiment, 11 participants who are familiar to the location were asked to complete the task. Similarly, in second experiment conducted at another location, 10 participants who are familiar to that location were asked to complete the task. The third experiment were conducted in the location where first experiment conducted with the five participants who are un-familiar to the building. These buildings are educational institution.

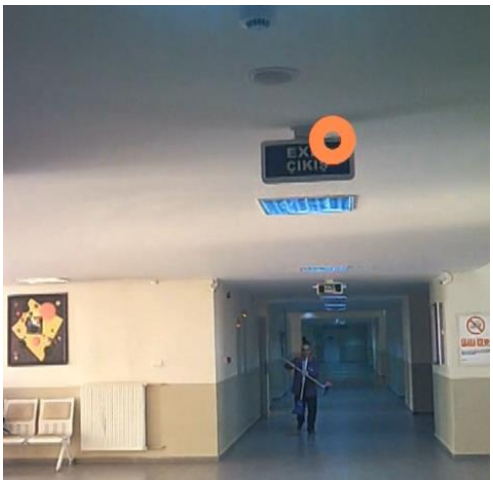
An eye tracking device was used to determine the awareness of occupational safety measures. SMI ETG 2W eye tracking device, Samsung S4 (SMI) and one powerbank were used in the study. The features of the eye tracking device used in this study are summarized in Table 1 and presented.

Table 1. The Features of the Eye Tracking Device

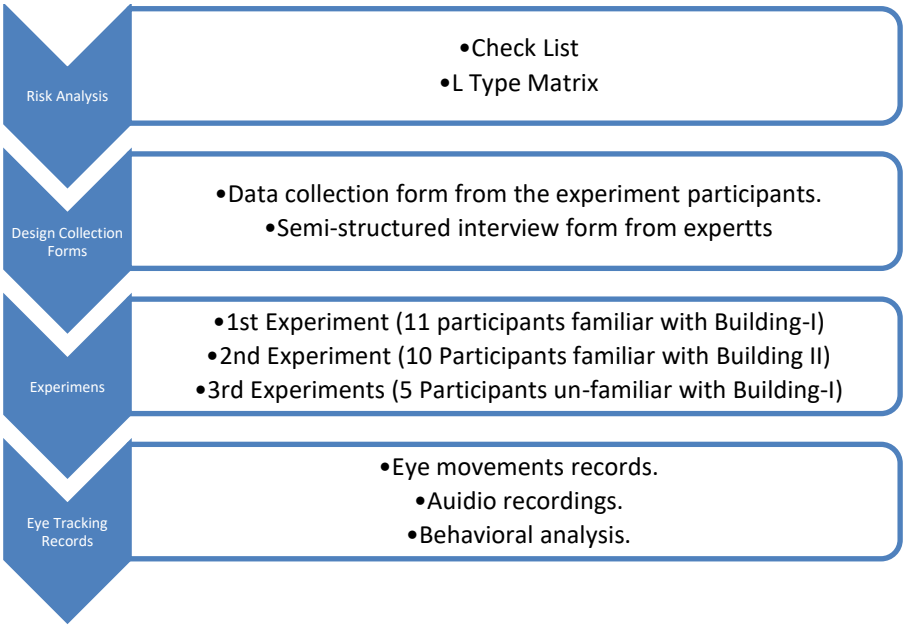
Weight	47g
Calibration	1-/3-Point Calibration
Sampling Rate	60Hz Binocular
Gaze Tracking Range	80° Horizontal, 60° Vertical
Resolution	1280x960p @24 fps, 1024x720p @30 fps
Scene camera field of view	60° Horizontal, 46° Vertical

The eye movements of the users have recorded and have analyzed using the eye tracking device. A sample image as an eye tracking output is shown below.

Figure 1. A Sample Image as an Eye Tracking Output



A data collection form has been created to detect demographic information of users. Users were asked to fill the form including age, gender, educational status, experience, position at the institution, body height and weight information. In order to support the results, a meeting was organized with OHS experts. A semi-structured form consisting of 10 open-ended questions has been prepared for the interview and 12 OHS experts opinions were asked. The institutions where the study conducted was selected according to the adequacy of the OHS measures. Finally, the data collected with the help of the eye tracking device was analyzed with the SMI BeGaze software. 26 people from two different institutions participated to the study. Data collection process summarized below.



3. RESULTS AND DISCUSSION

The first experiment was conducted at institution K and second experiment was conducted at institution E with institutions' own employees. The last experiment was conducted at institution E with employees of institution K. The purpose of the last experiment is to evaluate the OHS measures with participant who are unfamiliar the workplace. Assuming that the risk of fire in E and K institutions is at an equal level, it is aimed to determine the awareness of people in an institution they do not know about occupational health and safety measures.

The number and duration of focuses for each participant towards fire equipment (panic button, fire hose, and fire extinguisher) and caution signs were examined. The study composed of three different experiments. The participants' awareness of fire equipment and emergency exit signs has been shown in Figure 2.

Figure 2. Caution Sign (Focus)

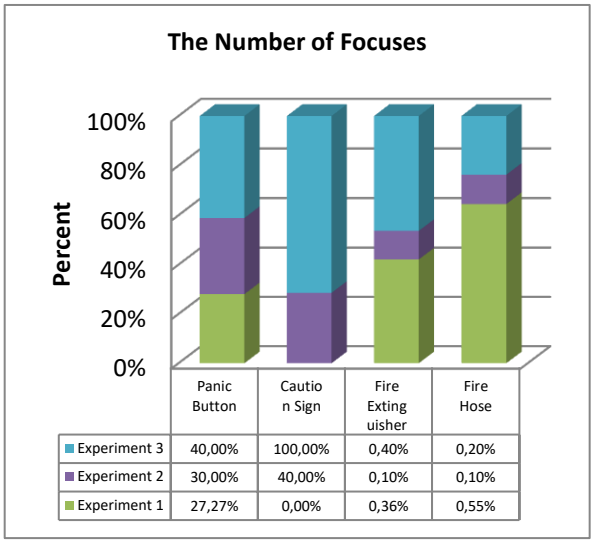


Figure 2 shows that the awareness of the employees for emergency exit signs and fire equipment is weak. For example, 10 % participants realized the "file extinguisher" in the experiment 2. According to the analysis, only "caution sign" were realized by all participants contributed to the experiment 3.

3.1. Experiment 1

This experiments were performed in two different faculties within a university in 2017. The institution where the first experiment was conducted named as "K" and the participants named as "K1, K2, K3 ...". Eleven staff members were included in the first experiment. In K institution, three out of 11 participants saw the panic button, and the average focus duration for the panic button was calculated to be 2250 milliseconds. Similarly, four of the participants saw the fire extinguisher, while the others did not see. It was observed that six of the participants saw the fire extinguisher. When the data were examined it was determined that none of the participants saw the caution sign. The focus periods of the participants are given in Figure 3-5.

Figure 3. Focus Time of the Participants to Panic Button (Experiment 1 – millisecond)

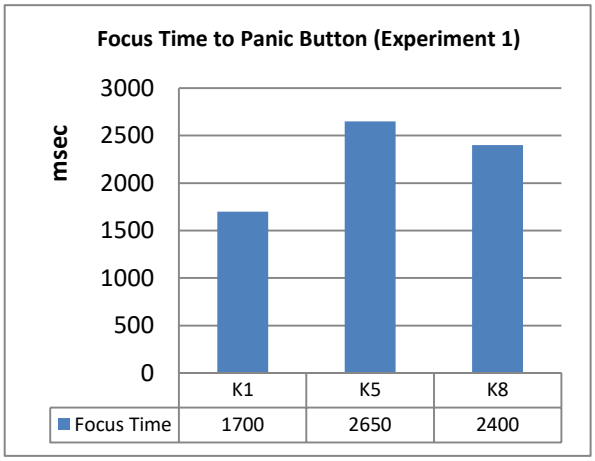


Figure 4. Focus Time of the Participants to Fire Extinguisher (Experiment 1 – millisecond)

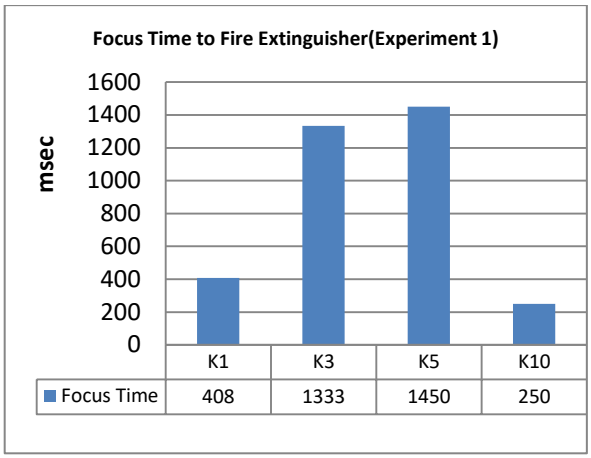
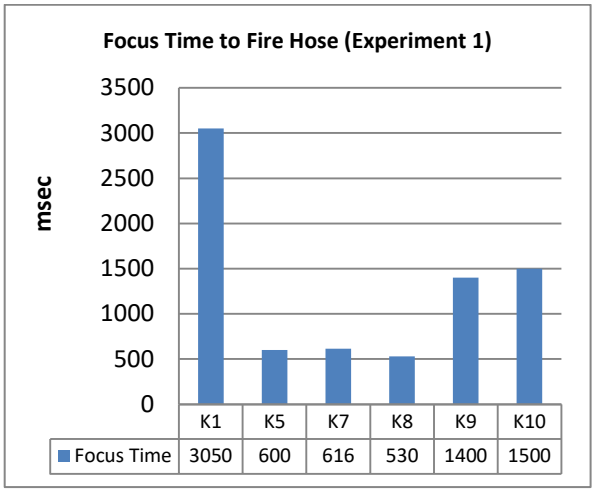


Figure 5. Focus Time of the Participants to Fire Hose (Experiment 1 – millisecond)



As can be seen in the Figure 4, four participants saw the fire extinguisher. The fire extinguisher is located on the side of the fire hose inside the fire cabinet.

3.2. Experiment 2

The institution where the second experiment was conducted named as “E” and the participants named as “E1, E2, E3 ...”. Ten staff members were included in the second experiment. In institution E, three out of 10 participants saw the panic button. Only one participant saw the fire extinguisher and the fire hose. Unlike the first experiments, four participants have seen the caution sign and while others have not seen it. The focus periods of the participants are given in Figure 6, Figure 7.

Figure 6. Focus Time of the Participants to Panic Button (Experiment 2 – millisecond)

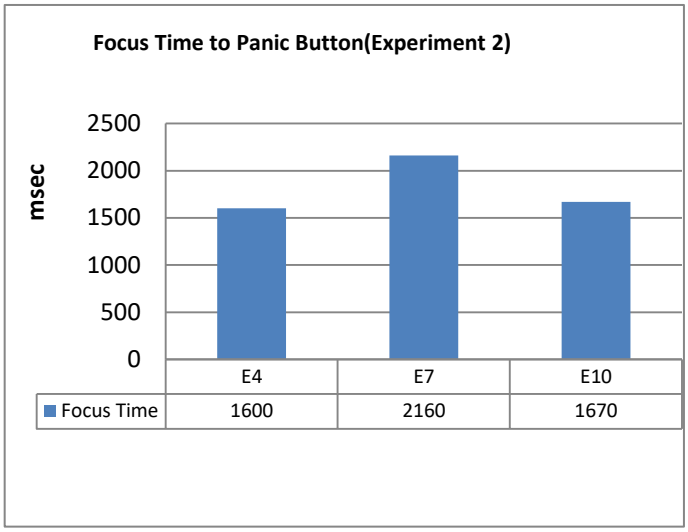
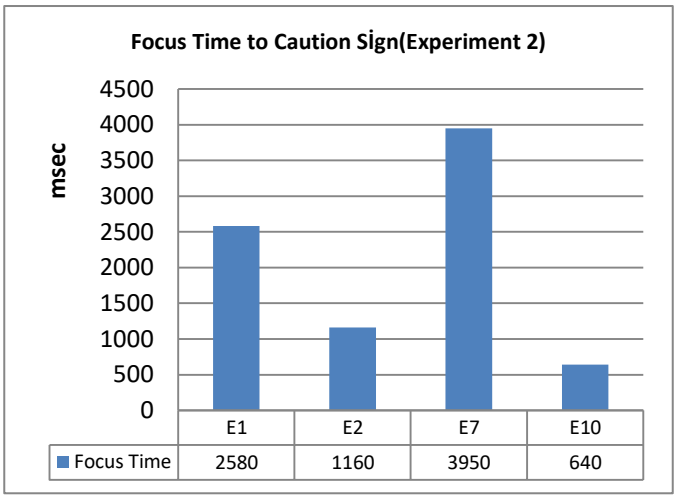


Figure 7. Focus Time of the Participants to Caution Sign (Experiment 2 – millisecond)



3.3. Experiment 3

In the third experiment, unlike other experiments, five employees in K institution were taken away to E institution. In this experiment, it is aimed to determine the awareness of the personnel who are not familiar to the building against measures. five participants included in the experiment were named D1, D2, D3, D4 and D5. In the third experiment, two out of five participants saw the panic button and all of the participants saw caution sign. Three out of five participants saw the fire extinguisher and two out of five participants saw fire hose. The focus periods of the participants are given in Figure 8-11.

Figure 8. Focus Time of the Participants to Caution Sign (Experiment 3 – millisecond)

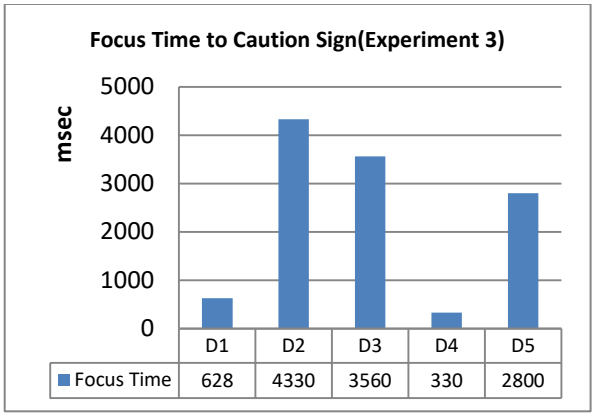


Figure 9. Focus Time of the Participants to Panic Button (Experiment 3 – millisecond)

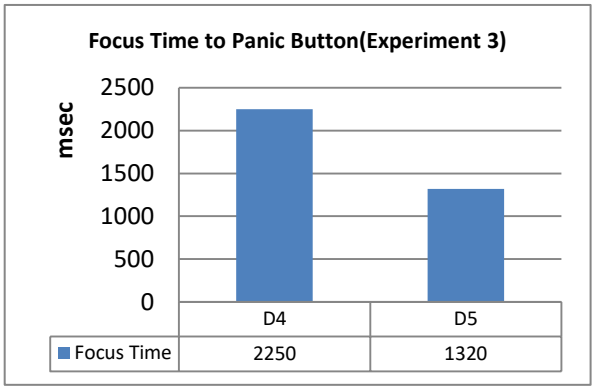


Figure 10. Focus Time of the Participants to Fire Extinguisher (Experiment 3 – millisecond)

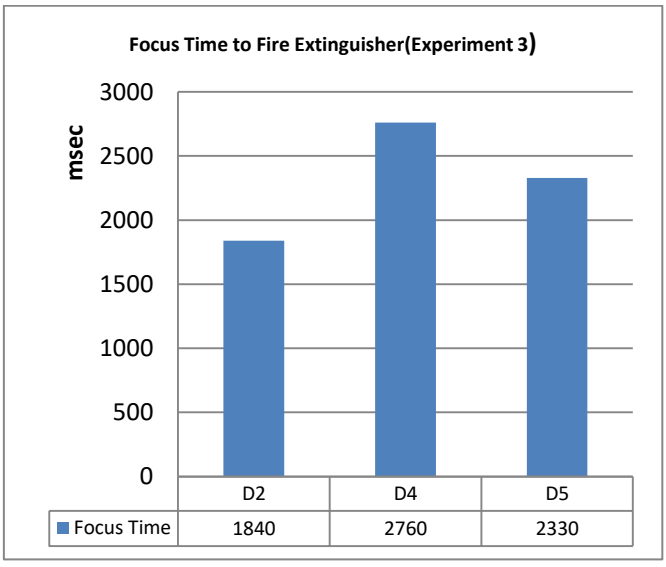
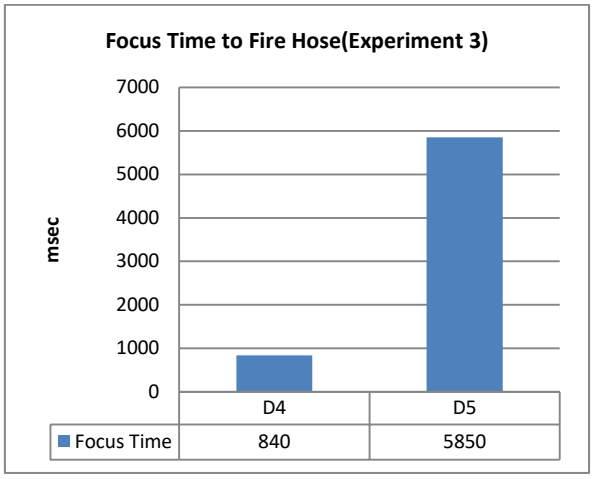
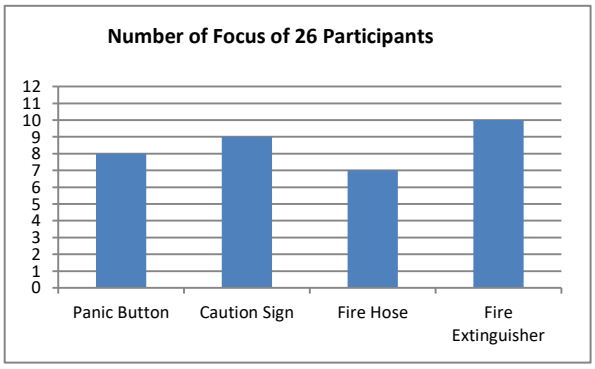


Figure 11. Focus Time of the Participants to Fire Hose (Experiment 3 – millisecond)



Although emergency exit signs and fire equipment are sufficed according to the previous risk analysis with the checklist, it can be said that employee awareness is not sufficient. Figure 12 shows the number of participants focused on to emergency exit sign and fire equipment.

Figure 12. Number of Focus of 26 Participants



4. OHS SPECIALIST OPINIONS

In order to examine the results, 12 occupational health and safety specialists were asked their opinions via a semi-structured form. Questions addressed to the participants and answers given by the participants to the questions are summarized below.

First, participants were asked questions about the use of technology in the work of OHS. Six participants stated that technology is being used effectively in occupational health and safety applications, while others disagree with this view. Seven participants stated that they are actively using technology in their studies.

When asked about the use of IT systems in OHS works, eight of the participants stated that IT systems were used effectively in OHS works and remaining six participants stated that they are actively using information systems in their past applications. All of the participants stated that eye tracking devices was not used in OHS works. Seven participants suggested that eye tracking devices will contribute to the field. Similarly, all participants stated that taking the actual behavior of employees into account in risk analysis would increase the efficiency. Finally, the exper asked whether OHS applications can be supported by usability tests within the scope of human computer

interaction. Nine participants stated that it could be supported. However, two participants did not agree about us contribution of usability tests. One participant did not report any opinions.

The expert opinions are summarized and presented in Table 2.

Table 2. OHS Specialist Opinions

Answer	**Yes**	**No**	**n/a**
Do you think technology is used effectively in today's occupational health and safety applications?	6	6	-
Have you benefited from technology in your past practices within the scope of occupational health and safety?	7	5	-
Do you think that information systems are used effectively in today's occupational health and safety applications?	8	4	-
Have you used information systems in your past practices within the scope of occupational health and safety?	6	6	-
Is an eye tracking device used in today's occupational health and safety applications?	-	12	-
Have you used an eye tracking device in your past applications within the scope of occupational health and safety?	-	12	-
Can the use of eye trackers within the scope of occupational health and safety increase efficiency in risk analysis?	7	2	3
Could taking into account the actual behavior of employees improve productivity?	12	-	-
There are 2 options in the checklists used for risk analysis (yes / no). In this context, would increasing the number of options and making ratings increase productivity?	9	1	2
Can occupational health and safety practices be supported by usability tests?	9	2	1

5. CONCLUSION AND RECOMMENDATIONS

Within the scope of this study, it is focused on evaluating OHS precautions from usability testing perspectives and the applicability of the eye tracking devices in terms of OHS measures. To do so, three experiments were designed to test the environment based on experimental testing methodology from human-computer interaction context. That is, the environment and precautions towards fire case (panic button, caution sign, fire hose and fire extinguisher) taken as a result of the risk analysis were assumed as an interface. In each experiments, participants wearing eye tracking glasses were asked to behave as if in a fire case and find the exit of the building. Data gathered from eye tracking device including number of focus and focusing time were analyzed to evaluate the precautions in terms of OHS measures. Results were also evaluated to understand the feasibility of using eye tracking glasses during risk analysis.

According to the results of the experiments in the institutions, it can be reported that participants may missing the signs and equipment including fire equipment and emergency exit warning signs although they are approved in terms of OHS according to the check list method used for risk analysis. For instance, in the risk analysis carried out with the help of the control list in terms of OHS, all equipment was approved as sufficient. However, almost half of the participants could not be able to focus on the equipment and signs. This may imply that, instead of making risk analysis based on any check list or other methods, eye tracking devices and usability testing methods should be applied to improve efficiency of OHS precautions.

In the first experiment made in institution K, it is found that participants went towards the exit without looking at emergency exit signs. This may be due to the fact that employees are familiar with the institution and know the points where the exit doors are located. Therefore, it is recommended that such practices should be carried out with employees who unfamiliar to the workplace. It has been observed that the employees in the second experiment look at the warning

signs more than the employees in the first experiment. In this case, it can be said that the positioning of emergency exit and warning signs in the second institution is better. Also, in the first experiment, it was observed that two of three participants who saw the alarm button used the student door and one used the staff door. It was determined that the alarm button on the route where the student door is located is at the top of the stairs. This may be the reason for providing more visibility.

In the second experiment, it is seen that the average focusing time on the alarm button is less than the first application. This may be due to the fact that the alarm button is placed in the corridor and its visibility is lower.

In the third experiment, it was observed that the employees did not head to the fire equipment, and they preferred to leave the building quickly. The reason may be because of the fact taht the employees are not familiar with the building or do not feel responsible.

According to the experts' opinion, it may be concluded that taking the actual behavior of employees into account via usability tests using ey-tracking device may increase the efficiency of the risk analysis.

In this study, employees were asked to act like there is fire in the institution. This is a limitation of the study since this may not provide a realistic environment. Moreover, the effectiveness of occupational health and safety measures were evaluated only in the context of fire. Possible further studies about evaluation of OHS measures and precautions can be conducted with the help of VR applications. The immersiveness of the environment can be increased by wearing virtual reality glasses to employees. The proposed method can also be applied in other disasters and emergency cases. In addition, such possible further studies in the class of hazardous workplaces may increase the widespread impact of usability testing in this context, and reveal the conformability and effectiveness of OHS measures.

REFERENCES

CENTAUR COMMUNİCATİON. (2005). "Mazda Turns To Eye Tracking To Assist Revamp Of European Site". New Media Age 3, 8.

ÇAĞILTAY K. (2011). "İnsan Bilgisayar Etkileşimi ve Kullanılabilirlik Mühendisliği: Teoriden Pratiğe (1. b.)". Ankara: ODTÜ Yayıncılık.

EVCİL E. S., ISLIM, Ö. F. (2012). "Kullanılabilirlik Kavramı ve Kullanılabilirlik Ölçümleri". In TH International Computer&Instructional Technologies Symposium (pp. 4-6).

GİRİŞKEN Y. (2015). "Gerçeği Algıla", İstanbul: Beta.

HOLMQVIST K., WARTENBERG C. (2005). "The Role of Local Design Factors for Newspaper Reading Behaviour- An Eye Tracking Perspective". Lund University Cognitive Studies 127, 1-21.

İNCE O., GÖKTÜRK, M. (2009). "Güvenlik Sistemi İzleyici Personelinin Görsel Tarama Davranışının Analizi". Akademik Bilişim. XI. Akademik Bilişim Konferansı Bildirileri. (699-704). Ankara: Nokta Matbaacılık.

KALAYCI E., TÜZÜN H., BAYRAK F., ÖZDİNÇ F., KULA A. (2011). "Eye-tracking Methods for Usability Testing in 3d Virtual Environments". Akademik Bilişim. XIII. Akademik Bilişim Konferansı Bildirileri (93-98). İnönü Üniversitesi, Malatya.

KAPLAN K. G. (2015). Prototype Fidelity and User Expertise in Usability Testing: A Study with Portable Navigation Device. (Master's Thesis). İstanbul: Istanbul Technical University Graduate School of Science Engineering and Technology.

KARAMAN G., ÇELİKER O., KARAMAN E., ÖZEN Ü. (2016). "Is normal handwriting or cursive hanwriting? A pilot study with eye tracking device". Journal of Management Information System, 2, 234-245.

KAYA Y. (2017). Analyzing the effects of multi-ball training on eye movements in table tennis players. Master's Thesis. Marmara University Graduate School of Health Sciences. İstanbul.

ÖMÜR S., AYDOĞDU A.G. (2017). "Eye tracking researches and new trends in the field of communication". International Journal of Social Sciences and Education Research 3, 1296-1307.

ÖZDOĞAN F.B. (2008). "Göz İzleme ve Pazarlamada Kullanılması Üzerine Kavramsal Bir Çalışma". Gazi Üniversitesi Ticaret ve Turizm Eğitim Fakültesi Dergisi 2, 134-147.

RASHID S., SOO S. T., SIVAJI A., NAENI H. S., BAHRİ S. (2013). "Preliminary usability testing with eye tracking and FCAT analysis on occupational safety and health websites". Procedia-Social and Behavioral Sciences, 97(6), 737-744.

RISK ASSESSMENT REGULATION. (2012). Information System of Regulations. 28512, 29 12 2012.

RITTHIRON S., JIAMSANGUANWONG A. (2017). "Usability Evaluation of the University Library Network's Website Using an Eye-Tracking Device". In Proceedings of the International Conference on Advances in Image Processing (pp. 184-188).

YÜCEL A., COŞKUN P. (2018). "Nöropazarlama Literatür İncelemesi". Firat University Journal of Social Sciences/Sosyal Bilimler Dergisi, 28(2).

ZAMBARBIERI D., CARNIGLIA E., ROBINO C. (2008). "Eye tracking analysis in reading online newspapers". Journal of Eye Movement Research 2, 1-8.

RASTGELE ORMAN ALGORİTMALARI İLE OTEL ÖZELLİKLERİ ANALİZİ

***

HOTEL FEATURES ANALYSIS WITH RANDOM FOREST ALGORITHMS

DOI: 10.33461/uybisbbd.756276

Sıla ŞİRİN*

Öz

Günümüzde insanlar otel yelpazesinin çok geniş olması nedeniyle otel seçimlerini, kendi tercihleri doğrultusunda filtreleyerek gerçekleştirmek istemektedirler. Farklı yaş grupları ve çocuklu ailelerin otellerden beklentileri değişmektedir. Örneğin, çocuklu aileler çocuklarının da eğlenerek vakit geçirebilecekleri, denize yakın, kumlu plaja sahip olan otelleri tercih etmektedirler. Daha ileri yaş gruplarında sessiz, spa özellikleri olan oteller tercih edilebilmektedir. İnsanların tercihleri mevsimlere göre de değişiklik gösterebilmektedir. Bu nedenle otel rezervasyonu yapan şirketler için; belirli müşteri gruplarına, bütçeye, yerleşim bölgesine ve mevsimlere göre tercihlerin yorumlanabilmesi önem arz etmektedir. Bu çalışmada müşteri tercihlerinin otel özellikleri bakımından yorumlanabilmesi için, ilk olarak otel özellikleri frekans analizi yöntemi ile azaltılmıştır. Kalan özellikler üzerinde Rastgele Orman Algoritmaları çalıştırılarak yaş gruplarına, mevsimlere ve çocuklu ailelere göre önemli otel özellikleri belirlenmiştir.

***Anahtar Kelimeler**: Rastgele Orman Algoritmaları, Otel Özellikleri, Otel Öneri Sistemi, Akıllı Seyahat.*

Abstract

Today people want to choose hotels by filtering them according to their different features, because the hotel range is very wide. The expectations of different age groups and families with children also vary. For example, today's families with children prefer hotels that are close to the sea, have sandy beach and features where their children can have fun. At older age groups, quiet, spa hotels can be preferred. The choices of people can also change according to the seasons. For this reason, hotel features that stand out according to certain customer groups, budget, residental area and seasons are important for companies that book hotels to form an accurate suggestion system. In this study, firstly, hotel features were reduced by frequency analysis method. The important features were determined for age groups, seasons and families with children by Random Forest Algorithms.

***Keywords:** Random Forest Algorithms, Hotel Features, Hotel Recommendation System, Smart Travel.*

* Türkiye, e-posta: silaasirin@gmail.com, ORCID: 0000-0002-4928-4046

Geliş Tarihi / Received: 22.06.2020
Kabul Tarihi / Accepted: 31.12.2020

1. GİRİŞ

Otel rezervasyonu yapan şirketler için, müşterilerin tercihlerini tahmin ederek otel rezervasyonu yapılma olasılığını artırmak en büyük amaçtır. Bu şirketlerin maksimum kâr elde edebilmeleri için otelleri doğru zamanda doğru müşteriye önermeleri gerekir (Boz ve ark., 2018). Günümüzde var olan rezervasyon uygulamaları, müşterilerin seçtikleri kriterleri kaydedebilmekte, en çok tercih edilen ve en çok memnun kalınan otellere göre otelleri filtreleyip, sıralayabilmektedir. Otellerin hangi özelliklerine göre ön plana çıktığı, farklı müşteri gruplarının hangi otelleri, hangi özelliklerine göre seçtiği analizinin yapılması gerekmektedir. Rezervasyon şirketlerinin müşterilere doğru otelleri önererek, satış sayılarını artırmaları gerekmektedir. Bu nedenle şirketlerin, otel tercihlerinde öne çıkan özellikleri belirleyebilecekleri ve sürekli yeni verilerle eğitebilecekleri bir modele ihtiyaçları vardır. Bu tür analizler için makine öğrenmesi yöntemlerine sıkça başvurulmaktadır. Bu çalışmada da Rastgele Orman Algoritmalarıyla cinsiyet, çocuklu aile, yaş aralıkları gibi müşteri özelliklerine ve mevsimlere göre otellerin öne çıkan özellikleri belirlenmek istenmiştir.

Çalışmayı yürütürken hangi otel özelliklerinin ön plana çıktığı vurgulanarak otel rezervasyonu yapan şirketler için satış sayılarını artırmak amaçlanmıştır. Elde edilen bilgiler ışığında çalışma farklı şekillerde detaylandırılarak; örneğin otel özelliklerine göre benzer otellerin belirlenip müşteriye sunulması gibi çalışmalar ile sürdürülebilir. Bu alandaki diğer çalışmalar incelendiğinde otel satışlarını artırmak için daha önceki müşteri yorumları üzerinde duygu analizleri yürütüldüğü gözlenmiştir. Bu çalışmada farklı olarak müşteri yorumları kullanılmadan satış ve otel verileri üzerinden satışın gerçekleşmesini sağlayan otel özellikleri belirlenmek istenmiştir. Çalışmayı sunarken ilk olarak veri özellikleri ve kullanılan metot anlatılacaktır, daha sonra ise elde edilen sonuçlar verilecektir.

2. LİTERATÜR TARAMASI

Tatil ayarlarken en önemli adımlardan birisi tatilcinin tercihlerine göre en uygun oteli bulabilmektir. Bu nedenle otel rezervasyonu yapan şirketler için uygun otelleri müşterilere önererek; müşterilerin konaklayacakları otellere bu şirketler üzerinden ulaşabiliyor olması önem arz etmektedir. Bu şirketler için başarılı bir gelir yönetim sisteminin önemli göstergelerinden biri, ne kadar çok rezervasyon ayarlayabildiğidir. Bu çalışmada bazı kriterlere göre otellerin öne çıkan özellikleri belirlendi ancak benzer çalışmalara bakıldığında genel olarak otel öneri sistemleri üzerinde çalışıldığı görüldü. Bizim çalışmamızdan farklı olarak otellerin ön plana çıkan özellikleri yerine, benzer otel grupları oluşturulması, satış tahminleri, maliyet hesaplamaları ve rezervasyon iptallerinin önlenmesi, müşteri yorumları üzerinden duygu analizi gibi konulara yoğunlaşan ve farklı metotların karşılaştırıldığı çalışmalar yapılmıştır. Otel öneri sistemlerinin çoğu otel puanlama ile ilişkilidir ve bu nedenle araştırmaların çoğunda otel sıralaması yapmak amaçlanmıştır (Sayar & Turdaliev, 2018).

Mavalankar ve ark. (2019)'nın ortak çalışmasında Expedia'nın veri seti kullanılmıştır. Çalışmada 100 farklı otel grubundan müşterinin hangisinde kalacağını tahmin etmek amaçlanmıştır. Bunun yanında ek bir çalışma ile her kullanıcının arama sorgusu için en olası ilk beş otelin sıralanması sağlanmıştır. Bizim çalışmamızdan farklı olarak bu çalışmada Stokastik Gradyan İniş sınıflandırıcısı, Rastgele Orman, XG Boost ve Naive Bayes metotları karşılaştırılmıştır. Stokastik Gradyan İniş, dışbükey kayıp fonksiyonları altındaki ayırt edici doğrusal sınıflandırıcılar için basit ama etkili bir yaklaşımdır. Naive Bayes, Bayes teoremini her özellik çifti arasında saf bağımsızlık varsayımıyla uygulamaya dayanan bir dizi denetimli öğrenme algoritmasıdır.

XGBoost, bir tahmin yapmak için zayıf öğrenici topluluklarını kullanmayı ifade eden Extreme Gradient Boosting'in kısaltmasıdır. Bu yöntem de Naive Bayes gibi bir denetimli öğrenme algoritmasıdır. XGBoost ve Rastgele Orman Algoritmaları, model olarak ağaç topluluklarını

kullanmaları bakımından benzerdir (Mavalankar ve ark., 2019). Çalışmanın sonucuna bakacak olursak; en iyi sonuca Rastgele Orman Algoritmaları ile ulaştıkları görülmektedir. Bizim çalışmamızda da Rastgele Orman Algoritmalarının otel verisi benzeri verilerde daha iyi sonuçlar elde edeceği öngörülmüştür; çünkü Rastgele Orman Algoritması aykırı değerlere karşı dirençlidir ve doğrusal olmayan verilerle iyi çalışır, aynı zamanda bizimki gibi büyük veri kümelerinde daha iyi sonuçlar elde edilir. Rastgele Orman Algoritmaları daha detaylı olarak Bölüm 3'te anlatılmıştır.

Aras ve ark. (2019) çalışmasında rezervasyon, otel özellikleri ve online seyahat acentesi verilerini kullanarak satış tahmini yapılması amaçlanmıştır. Yine bu çalışmada da bizden farklı, Mavalankar ve ark. (2019) çalışmasına benzer olarak 4 farklı algoritma sonuçları karşılaştırılmıştır. Yine aynı tür veri için XGBoost, Rastgele Orman Algoritmaları, Gradyan Artırma ve Genelleştirilmiş Doğrusal Model algoritmaları denenmiştir. Gradyan artırma Rastgele Orman Algoritmalarına benzer şekilde zayıf tahmin modellerinin bir araya gelmesiyle tipik olarak karar ağaçlarının oluşturduğu bir model oluşturur ve bunun üzerinde çalışır (Aras ve ark., 2019). Yine bu çalışmada da ağaç yapısı üzerine kurulu modellerin daha doğru sonuçlar verdiği gözlenmiştir. Çalışma sonucunda bizim çalışmamızdan farklı olarak geçmişteki satış verilerinden yararlanarak gelecekteki net toplam maliyet öngörülmüştür; bunun için bizim çalışmaya benzer olarak otel özelliklerinden ve satış verilerinden yararlanmışlardır.

Bir diğer çalışmada Kasper ve Vela (2012) farklı sitelerden müşteri yorumlarını toplayan otel yöneticileri için; kullanıcı yorumlarını analiz edip, sınıflandırmayı planlamışlardır. Müşteri yorumlarını pozitif, negatif, tarafsız ve birden çok konu olacak şekilde sınıflandırmışlardır. Bunun için genel olarak dilbilim analizleri yapılmış ve sınıflandırma algoritmaları kullanılmıştır. Bu çalışmada otel özelliklerinden hiç yararlanılmamış ve oteller için yapılan müşteri yorumlarının analizi üzerine durulmuştur.

3. MATERYAL VE YÖNTEM

Elimizde Setur grubunun kişisel veri olmayacak şekilde maskelenmiş, otel ve müşteri satış verileri bulunmaktadır. Bu bölümde elimizdeki tabloların öznitelik özellikleri ve veri üzerinde yapılan ön işlemler anlatılmıştır.

3.1. Veri Özellikleri ve Ön İşleme

Yapılan çalışmada otel özellikleri, satış ve müşteri bilgilerinin tutulduğu veri setlerinden yararlanılmıştır. TABLO 1, TABLO 2 ve TABLO 3'te verilerin hangi sütunlarda tutuldukları, veri tip ve nitelik bilgileri verilmiştir. (Varchar: farklı uzunluklarda veri girişi yapılacağı zaman kullanılan karakteristik veri tipi, Int: Tam sayı tutan nümerik bir veri tipi, Numeric: Ondalık ve tam sayı türünde veri saklayabilen nümerik bir veri tipi, Datetime: Tarih ve saat verilerini tutan tarih veri tipi (Anonim, 2017).)

TABLO 1 otel özelliklerinin tutulduğu tablo bilgisini içermektedir. Bu tabloda her bir otel için sahip olduğu özellikler verilmiştir. Otel özellikleri veri setinden Otel Numarası ve Özellik Adı nitelikleri kullanılmıştır. Analize başlamadan önce veri seti, her bir otel özelinde özelliklerin tek bir satırda gösterilecek şekilde değiştirilmiştir. Yani; otel özellikleri yeni birer sütun olarak tanımlanarak otel özelinde var/yok(1/0) değerleriyle tek satırda tutulmuştur.

TABLO 1. Otel Özellikleri Veri Seti

Nitelik	Veri Tipi
Otel Özellik Numarası	int
Özellik Numarası	int
Otel Numarası	int
Özellik Adı	varchar
Özellik Kategorisi	varchar

TABLO 2 satış verilerinin tutulduğu tablo bilgilerini içermektedir. Analizler sırasında Müşteri Numarası, Otel Numarası, Otel Satışındaki Toplam Çocuk Sayısı, Otel Satışındaki Toplam Yetişkin Sayısı, Hizmet Başlangıç Tarihi kolonlarından yararlanılmıştır. Satış sayısı verisi her otel özelinde verinin otel numarasına göre gruplanmasıyla elde edilmiştir ve bu bilgi yeni bir Satış Sayısı kolonu yaratılarak bu kolonda tutulmuştur. Bu şekilde otel bazında satış sayılarına ulaşılmıştır.

Ailelerin tercih ettiği otellerin belirlenebilmesi için Otel Satışındaki Toplam Çocuk Sayısı sütunu kullanılmıştır. Otel Satışındaki Toplam Çocuk Sayısı sütunu dolu olan satışlar çocuklu aile olarak sınıflandırılmıştır.

TABLO 2. Satış Veri Seti

Nitelik	Veri Tipi
Satış Numarası	int
Satış Tipi Numarası	int
Müşteri Numarası	int
Satış Yapan Firma Numarası	int
Satılan Otel Numarası	int
Satılan Cruise Numarası	int
Satılan Ekstra Servis Numarası	int
Satılan Tur Numarası	int
Kişi Adedi	int
Otel Satışındaki Toplam Yetişkin Sayısı	int
Otel Satışındaki Toplam Çocuk Sayısı	int
Tur Satışındaki Toplam Yetişkin Sayısı	int
Tur Satışındaki Toplam Çocuk Sayısı	int
Cruise Satışındaki Toplam Yetişkin Sayısı	int
Cruise Satışındaki Toplam Çocuk Sayısı	int
Verilen Hizmetin Toplamda Kaç Gece Sürdüğü Bilgisi	int
Verilen Hizmetteki Toplam Oda Sayısı	int
Satış Tarihi	datetime
Hizmet Başlangıç Tarihi	datetime
Hizmet Bitiş Tarihi	datetime
Kar	numeric
Satış Tutarı	numeric

TABLO 3 müşteri verilerinin tutulduğu tablodur. Müşteri Numarası ve Müşterinin Yaşı kolonları müşterilerin yaş aralığına göre otel özellikleri tercihlerinin belirlenmesi için kullanılmıştır.

TABLO 3. Müşteri Veri Seti

Nitelik	Veri Tipi
Müşteri Numarası	int
Müşteri Cinsiyet Bilgisi	varchar
Müşterinin Yaşı	int
Ülke Bilgisi	int
Şehir Bilgisi	int

Frekans analizi yöntemiyle frekansı 0.00050'ten düşük olan otel özellikleri çıkarılarak analiz yapılacak veri küçültülmüştür. Frekans analizi, otel özelliklerinin sıklığı incelenerek yürütülmüştür. Frekans analizinden önce veri satır sayısı 39559 iken frekans analizi ile 2263'e; aynı şekilde özellik sayısı da 207'den 122'ye düşürülmüştür.

Satış verileri içerisindeki mükerrer kayıtlar çıkarılmıştır. Mükerrer kayıtlar sonuçlarda sapmalara neden olacağı için veri ön işleme için önemli bir adımdır. Mükerrer kayıtların çıkarılması ile veri satır sayısı 300571'den 284329'a düşmüştür.

3.2. Rastgele Orman Algoritmaları

Rastgele Orman Algoritması, amaca uygun olarak; belirlenen ağaç sayısından meydana gelen sınıflandırma veya regresyon ağaç topluluklarından oluşmaktadır. Rastgele orman, her girdi vektöründen bağımsız olarak örneklenen rastgele bir vektör kullanılarak oluşturulan ağaçların kombinasyonundan oluşur. Bu çalışma için kullanılan rastgele orman, bir ağacı büyütmek için rastgele seçilen özelliklerin birleşiminden oluşturulmuştur. Bir sınıflandırıcı olarak, rastgele orman sınıflandırma için bir "güçlü değişkenler" alt kümesi kullanarak örtük bir özellik seçimi gerçekleştirir ve bu durum yüksek boyutlu veriler üzerinde üstün performansa yol açar (Breiman, 2001).

Rastgele orman algoritmaları hızlı, esnek ve yüksek boyutlu verileri analiz etmek için yeterince güçlüdür. Binlerce değişkeni silmeden veya doğrulukta bozulma olmadan işleyebilecek olağandışı yeteneğe sahip bir algoritmadır (Breiman, 2004). Algoritmanın, sınıflandırma ağaçlarına göre eksikliği, çıktı olarak bir ağaç vermemesidir (Korkmaz ve ark., 2018). Alternatif makine öğrenme algoritmalarına göre bir avantajı, ilgili özellikleri tanımlamak veya değişken seçimi yapmak için kullanılabilen değişken önem ölçütleridir. Gini endeksi gibi bölünmelerdeki kirliliğin azaltılmasına dayalı ölçütler yaygın olarak kullanılır, çünkü bu ölçütler basit ve hızlı sonuçlar verir. Gini endeksi, sınıflandırma ağaçlarında bölme kriteri olarak yaygın bir şekilde kullanıldığından, karşılık gelen safsızlık önemine genellikle Gini önemi denir.

Karar ağacının tasarımı için öznitelik seçim ölçüsü ve budama yönteminin seçilmesi gerekir. Bu çalışmada Sklearn kütüphanesinin tanımladığı Rastgele Orman Algoritması kullanılmıştır. Sklearn kütüphanesi Rastgele Orman Algoritması için öznitelik önemi(feature_importances_) özelliğini sunmaktadır. Öznitelik önemi bir öznitelik seçim ölçütüdür ve değerinin büyük olması özniteğin öneminin de fazla olması demektir. Öznitelik önemi, tahmin yaparken; her bir özelliğin göreceli önemini belirten bir tahmin modeli oluşturmak için, her bir özelliğe belli bir skor belirler. Bu skorlar sayısal bir değeri tahmin etmeyi amaçlayan regresyon veya sınıflandırma problemleri için hesaplanabilir. Rastgele Orman Algoritmasında öznitelik önemi, ağacın bölünme noktalarını seçmek için kullanılan kriterdeki (Gini endeksi) indirgeme duruma göre bu skorları belirler. Yani bu

özelliği kullanan ağaç düğümlerinin, ormandaki tüm ağaçlar arasındaki kirliliği ne kadar azalttığına bakarak, bir özelliği ön planda tutan bir araç sağlar ve her bir özellik için bu skoru otomatik olarak eğitir ve sonuçları ölçeklendirir, böylece tüm önemlerin toplamı 1'e eşit olur (Pedregosa, 2011). Rastgele Orman Algoritması öznitelik seçimi için Gini endeksi kullanmaktadır. Gini endeksi aşağıdaki şekilde hesaplanır. (Yılmaz, 2014)

$$I_G = 1 - \sum_{j=1}^{c} p_j^2 \qquad (1)$$

T : Tüm veri seti pj : Veri setindeki her bir verinin, kendisinden küçük ve kendisinden büyük eleman sayılarına bölüm karesi. c : Seçilen veri

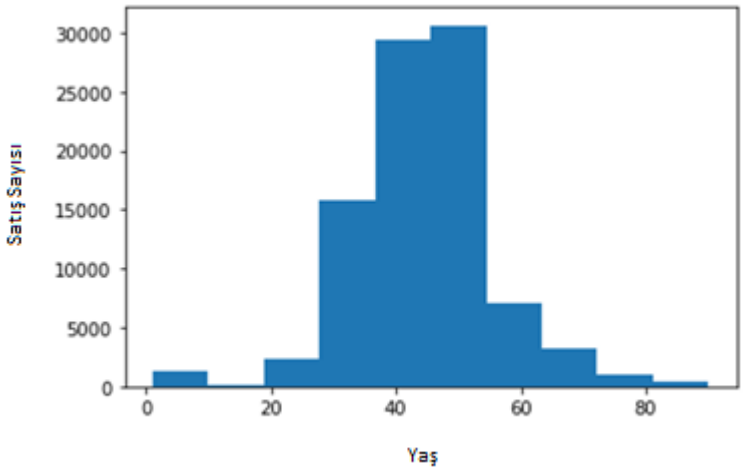
Bir karar ağacı, özelliklerin kombinasyonunu kullanarak yeni eğitim verileri üzerinde maksimum derinliğe kadar büyür. Bu çalışmada maksimum derinlik belirlenmemiştir, bu nedenle ağaçlar üzerinde budama yapılmamıştır. Daha doğru sonuçlar elde edebilmek için oluşturulacak ağaç sayısı 1000 olarak belirlenmiştir.

4. BULGULAR

Veri seti üzerinde istenilen amaca uygun olarak Rastgele Orman Algoritması kullanılmıştır. Öne çıkan özellikler; yaş aralığı, çocuklu aile ve mevsimlere göre ayrı ayrı belirlenmiştir.

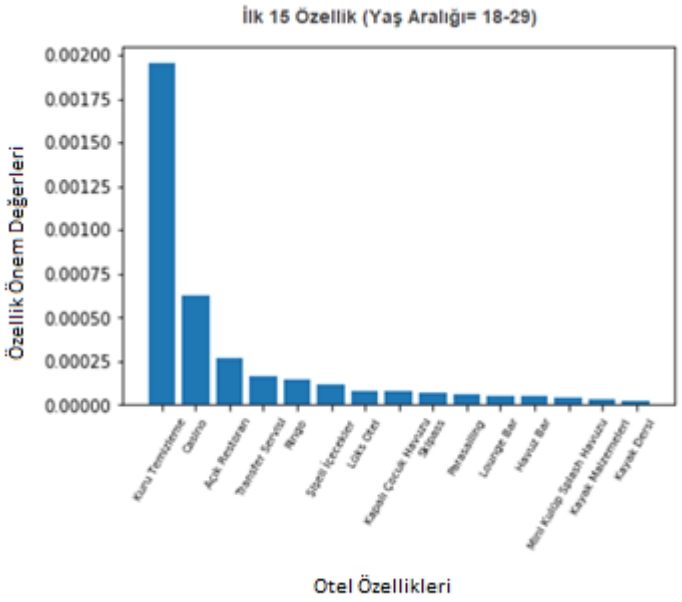
4.1. Yaş Aralığına Göre Öne Çıkan Otel Özellikleri

Bölüm 3'te bahsedildiği gibi müşterilerin verilerinin tutulduğu Müşteri veri seti ve otel satışlarının yer aldığı Satış veri setindeki verilerden yararlanılmıştır. Satış veri setindeki otel satışı olan verilerin müşteri numaraları ile müşteri veriş setindeki müşteri numaraları kullanılarak müşterilerin yaş bilgilerine ulaşılmıştır. Bu verilerden yararlanılarak öncelikle Şekil 1'de görüldüğü gibi satış verisinin yaşlara göre dağılımı gösterilmiştir.

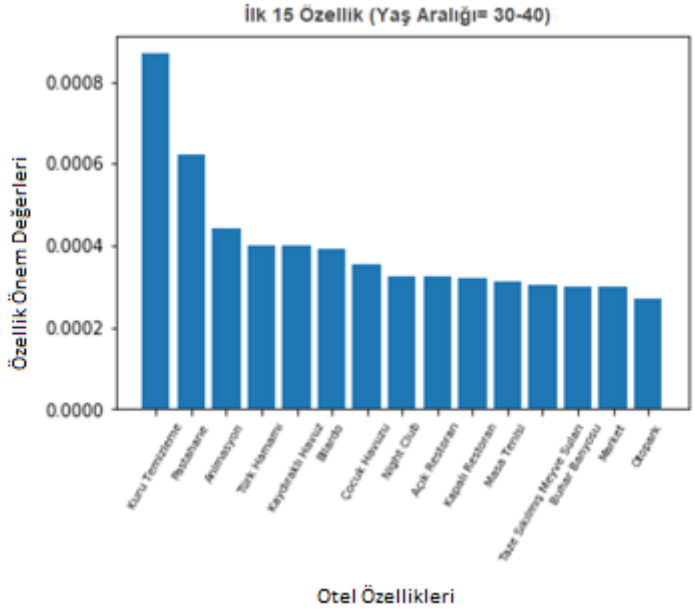


Şekil. 1. Satış verisinin yaşlara göre dağılımı

Şekil 1'deki dağılıma bakılarak müşteri yaş bilgilerinden 18-29, 30-40, 41-55, 56-74, 75+ yaş aralıkları belirlenmiştir. Her bir yaş aralığı sınıf olarak belirlenerek Rastgele Orman Algoritması çalıştırılmıştır. Özelliklerin önemini belirlemek için Bölüm 3'te bahsedilen Rastgele Orman Algoritmasının öznitelik öneminden yararlanmıştır, daha sonra her bir sınıf için özellikler ağırlıklandırılarak her sınıf için öne çıkan özellikler belirlenmiştir. Şekil 2 ve Şekil 3 üzerinde örnek olarak 18-29 ve 30-40 yaş aralıklarına göre belirlenen 15 en önemli özellik verilmiştir. İki yaş grubu için öne çıkan otel özelliklerinin değiştiği grafiklerde gözlemlenmiştir.



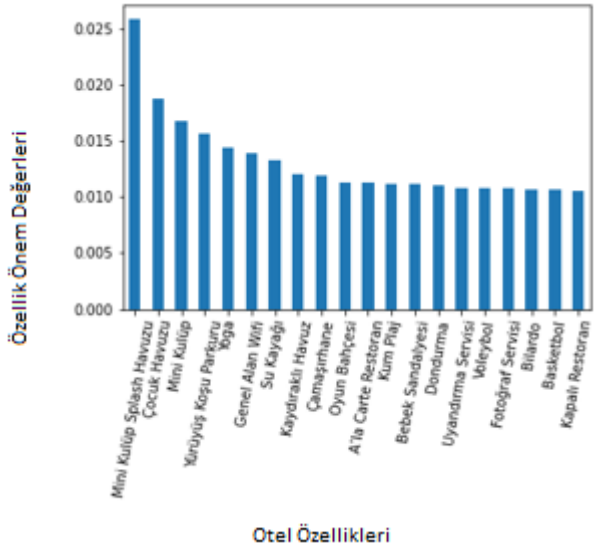
Şekil. 2. 18-29 yaş aralığı için öne çıkan 15 özellik



Şekil. 3. 30-40 yaş aralığı için öne çıkan 15 özellik

4.2. Çocuklu Ailelere Göre Öne Çıkan Otel Özellikleri

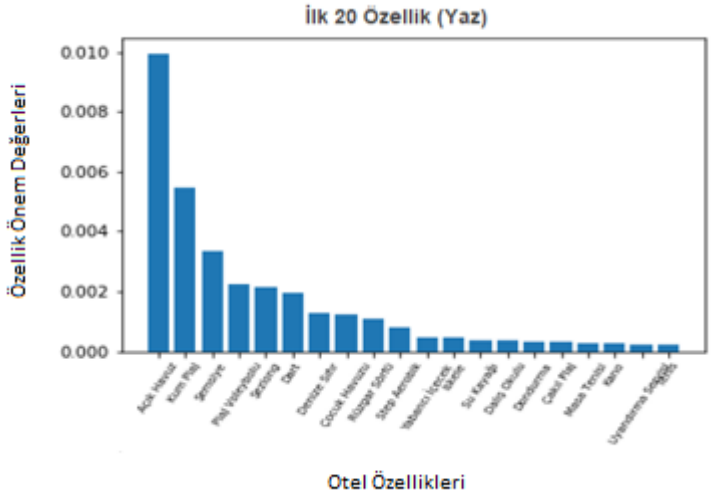
Satış verisi üzerinden Otel Satışındaki Toplam Çocuk Sayısı alanı kullanılarak çocuklu rezervasyonlar ve çocuksuz rezervasyonlar olarak 2 ayrı sınıf belirlenmiştir. Bu sınıflar için Rastgele Orman Algoritması çalıştırılarak yine her bir sınıf için öznitelik önemi sonuçları sınıf üzerinde ağırlıklandırılarak çocuklu ailelerin otel seçimlerinde hangi özellikleri tercih ettikleri çıkarılmıştır. Şekil 4 üzerinde çocuklu aileler için öne çıkan 20 özellik gösterilmiştir.



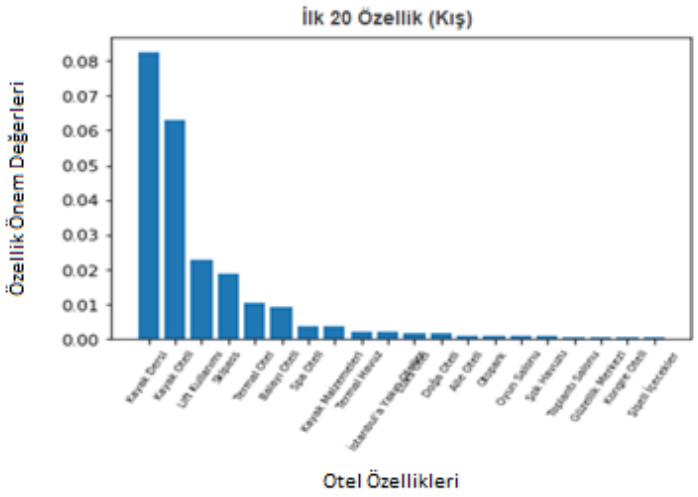
Şekil. 4. Çocuklu ailelere göre öne çıkan 20 özellik

4.3. Mevsimlere Göre Öne Çıkan Otel Özellikleri

Satış verisi üzerinden Hizmet Başlangıç Tarihi alanı kullanılarak satışların hangi tarihler için yapıldığı bilgisi elde edilmiştir. Nisan, Mayıs, Haziran, Temmuz, Ağustos, Eylül ayları yaz ayları olarak; Ekim, Kasım, Aralık, Ocak, Şubat, Mart ayları ise kış ayları olarak belirlenmiştir. Kış ve yaz mevsimleri için iki ayrı sınıf oluşturulmuştur. Satış verisi üzerinde kış ve yaz sınıfları için Rastgele Orman Algoritması çalıştırılarak, her iki sınıf için de öne çıkan özellikler belirlenmiştir. Bunun için yine Rastgele Orman Algoritmasının öznitelik öneminden yararlanılmıştır. Şekil 5 ve 6'da hem yaz hem de kış sınıfları için öne çıkan özellikler ayrı ayrı gösterilmiştir.



Şekil. 5. Yaz ayları için öne çıkan 20 özellik



Şekil. 6. Kış ayları için öne çıkan 20 özellik

5. TARTIŞMA

Otel rezervasyonu yapan şirketler için farklı müşteri gruplarına ve farklı zaman dilimlerine göre hangi otel özelliklerinin rezervasyonun tamamlanması için belirleyici olduğu üzerine çalışma yürütmek önem arz etmektedir. Bu çalışmada elde edilen bulgular üzerinden farklı analizler yürütülerek otel rezervasyonları artırılabilir.

Bu çalışmada veri, üç ana başlık için ayrı ayrı incelenmiştir. Yaş grupları, mevsim ve çocuklu aileler ayrı başlıklar olarak ele alınıp; her bir başlık için öne çıkan otel özellikleri Bölüm 4'te gösterilmiştir. Sonuçlar incelendiğinde; başlıklara göre belirleyici özelliklerin genel olarak makul olduğu görülmektedir. Örneğin; mevsimlere göre öne çıkan özellikler incelendiğinde kış ayları için kayak, termal, spa gibi kış sporları veya iç mekan olanaklarının, yaz ayları içinse havuz, kum plaj, plaj voleybolu gibi dış mekan olanaklarının otel rezervasyonu için daha belirleyici olduğu gözlenmiştir; ancak bunun yanında bazı durumlarda belirleyici olmaması beklenen özelliklerin de ön sıralarda listelendiği görülmüştür. Örneğin, yaş grupları için yapılan çalışma incelendiğinde "Kuru Temizleme" özelliğinin farklı yaş grupları için bile en ön sırada geldiği görülmüştür; ancak "Kuru Temizleme" özelliğin otel seçiminde en önemli kriter olması genellikle beklenmez. Bu nedenle bu tür sonuçların incelenmesi ve daha net sonuçlar elde edilmesi için çalışma farklı makine öğrenmesi yöntemleri kullanılarak ilerleyen zamanlarda devam ettirilebilir.

Literatürde yapılan çalışmalardan farklı olarak; bu çalışmada otel özelliklerinin farklı müşteri grupları ve farklı mevsimlere göre sıralanması sağlanmıştır. Literatürdeki diğer çalışmalar incelendiğinde, Mavalankar ve ark. (2019) otelleri gruplandırarak müşterinin hangi otelde kalacağını tahmin etmeye çalışırken; Aras ve ark. (2019)'da satış tahmini üzerine yoğunlaşmaktadır. Her iki çalışmada da farklı algoritmalar aynı veri seti üzerinde çalıştırılıp sonuçlar kıyaslanır. Bu çalışmada farklı olarak tek bir algoritma üzerinden öne çıkan otel özelliklerinin belirlenmesi sağlanmıştır. Literatürdeki diğer çalışmalar incelendiğinde benzer konuya eğilen çalışmalar bulunmamaktadır, bu nedenle bu çalışmanın sonuçları diğer çalışmalardan ayrışmaktadır.

6. SONUÇ

Konaklama hizmeti veren otellerin maksimum kâr elde edebilmesi için doluluk oranlarının yüksek olması gerekmektedir. Bu sebeple oteller rezervasyon sistemleri aracılığıyla sınırlı sayıdaki odalarını doğru zamanda, doğru müşteriye tahsis etmelidir. Bu araştırma makine öğrenmesi yöntemlerini kullanarak belirli hedef kitleleri ve belirli bölgeler için hangi otel özelliklerinin ön

plana çıktığını ve bu elde edilen sonuçlar ile daha iyi otel önerileri geliştirebilmek için otel rezervasyonu yapan şirketler ve aynı zamanda müşteriler tarafından kullanılabilecek daha başarılı öneri sistemleri geliştirmek hedeflenmiştir. Bu sayede otel rezervasyonu yapabilen şirketler tur şirketleri vb. her otel için sahip oldukları özelliklerden en çok tercih edilenleri ön plana çıkarıp satışlarını artırabilirler aynı zamanda müşteriler için de ihtiyaçlarına göre otel seçebilmeleri daha kolaylaştırılabilir. Diğer bir yandan şirketler için böyle bir bilgi, yönetimsel olarak da stratejilerini geliştirmelerine ve yönetim kalitesini artırmalarına imkân sağlar. Bu sayede şirketlerin elinde bulunan otel özellikleri, müşteri bilgileri, satış verileri arttıkça geliştirilen model ile çok daha kesin ve doğru analizlere ulaşılabilir. Farklı makine öğrenmesi yöntemleriyle bulunan sonuçlar test edilebilir.

KAYNAKÇA

Anonim, (2017). "Data types (Transact-SQL)". https://docs.microsoft.com/en-us/sql/t-sql/data-types/data-types-transact-sql?view=sql-server-2017, Erişim T: 13.12.2020.

Aras G., Ayhan G., Sarıkaya M., Tokuç A., Sakar C., (2019). "Forecasting Hotel Room Sales within Online Travel Agencies by Combining Multiple Feature Sets". 8th International Conference on Pattern Recognition Applications and Methods, 565-573.

Boz M., Canbazoğlu E., Özen Z., Gülseçen S., (2018). "Otel Rezervasyon İptallerinin Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi". Veri Bilimi. 1(1): 7-14.

Breiman L., (2004). "Consistency for a simple model of random forests". Technical Report 670, Technical report, Department of Statistics, University of California, Berkeley, USA.

Breiman L., (2001). "Random Forests". Machine Learning.

Kasper W., Vela M., (2012). "Sentiment Analysis for Hotel Reviews". Speech Technology, 4. 96-109.

Korkmaz D., Çelik E. H., Kapar M., (2018). "Sınıflandırma ve Regresyon Ağaçları ile Rastgele Orman Algoritması Kullanarak Botnet Tespiti". Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 23 (3): 297-307.

Mavalankar, A., Gupta, A., Gandotra, C., Misra, R., (2019). "Hotel Recommendation System". 10.13140/RG.2.2.27394.22728/1.

Pedregosa F., (2011). "Scikit-learn: Machine Learning in Python", JMLR 12, pp. 2825-2830.

Sayar A., Turdaliev N., (2018). "Makine Öğrenmesi ile Adaptif Otel Öneri Sistemi". 12th Turkish National Software Engineering Symposium, Istanbul, Turkiye.

Yılmaz H., (2014). "Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama". (Yüksek Lisans Tezi). Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı, Eskişehir, Türkiye.