E-ISSN: 3357-2980

JOURNAL OF DATA APPLICATIONS

ISSUE **1** • YEAR **2023**

İSTANBUL UNIVERSITY PRESS

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

Owner
Assoc. Prof. Zeki ÖZEN
Istanbul University, Faculty of Economics,
Department of Management Information Systems, Istanbul, Turkiye

Responsible Manager
Assoc. Prof. Zeki ÖZEN
Istanbul University, Faculty of Economics,
Department of Management Information Systems, Istanbul, Turkiye

Correspondence Address
Prof. Dr. Ümit Yaşar Doğanay Cad.,
İktisat Fakültesi Ek Bina-2 No: 6 Kat:1 Vezneciler / Fatih, Istanbul, Turkiye
E-mail: joda@istanbul.edu.tr
https://iupress.istanbul.edu.tr/en/journal/joda/home

Publisher
Istanbul Üniversitesi Yayınevi / Istanbul University Press
Istanbul Üniversitesi Merkez Kampüsü, 34452 Beyazıt,
Fatih / Istanbul, Turkiye
Telefon / Phone: +90 (212) 440 00 00

Printed by
İlbey Matbaa Kağıt Reklam Org. Müc. San. Tic. Ltd. Şti.
2. Matbaacılar Sitesi 3NB 3 Topkapı / Zeytinburnu,
Istanbul, Turkiye
www.ilbeymatbaa.com.tr
Sertifika No: 17845

Authors bear responsibility for the content of their published.

The publication language of the journal is English.

This is a scholarly, international, peer-reviewed and open-access journal published biannually in April and October.

Yayın Türü / *Publication Type*: Yaygın Süreli / *Periodical*

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

EDITORIAL MANAGEMENT BOARD

EDITORIAL BOARD

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

CONTENTS

Research Articles

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

EDITORIAL

Dear Researchers,

We are thrilled to present the first issue of the *Journal of Data Applications*, a prestigious journal dedicated to advancing the understanding and application of data in various domains such as statistics, artificial intelligence, machine learning, deep learning, and data mining. As the field of data science continues to evolve at an unprecedented pace, this journal aims to serve as a central hub for researchers, practitioners, and enthusiasts to share their insights, discoveries, and innovations. The *Journal of Data Applications* looks to contribute to the development of applied data science studies that aim to obtain meaningful information from data and reveal hidden patterns and designs in data, thus contributing to the development of studies in this field.

In this inaugural issue, we present six articles that highlight the diversity and breadth of data applications across different sectors. These articles have undergone a rigorous peer-review process and been selected based on their significance, originality, and contribution to the field. Each article showcases unique perspectives and approaches, underscoring the multidisciplinary nature of data science and its real-world implications.

The first article, "Istanbul's Community Mobility Changes During the COVID-19 Pandemic: A Spatial Analysis" by Ahmet Okan Arık and Gülsüm Çiğdem Çavdaroğlu, explores the impact the COVID-19 pandemic has had on community mobility in Istanbul, Türkiye. The study investigates the changes in mobility patterns in different locations such as parks, transit stations, workplaces, grocery stores, pharmacies, and residential areas in Istanbul, examining both the spatial and non-spatial independent variables that influenced these mobility changes. The study finds that variables such as altitude had no overall impact on mobility changes, whereas latitude, longitude, and proximity to the sea did affect mobility patterns.

The second article is "Modeling Automobile Sales in Turkey with Regression-Based Machine Learning Algorithms," authored by Merve Babaoğlu, Ahmet Coşkunçay, and Tolga Aydın, and delves into the realm of automotive sales forecasting. The authors focus on demand forecasting in the automobile sector, and their study examines 10 independent variables, including gross domestic product, sector confidence index, capital expenditures, household consumption expenditures, inflation rate, consumer confidence index, deposit percentages, and prices of oil, gold, and the dollar. Various predictive models, including linear regression, decision tree, random forest, ridge, AdaBoost, elastic net, and lasso regression algorithms, are applied to construct a sales forecast model. The study aims to provide an approach to allow industries that are directly or indirectly influenced by automotive sales to gain foresight and make informed decisions in their respective sectors.

The third article, "The Ethical Dimensions of Artificial Intelligence" by Gökçe Karahan Adalı, discusses the influence of artificial intelligence (AI) on various aspects of our lives and highlights its potential to enhance productivity and decision-making processes. However, the study emphasizes that ethical considerations will be inevitable alongside these technical advancements. Like any technology, AI brings both benefits and risks. One significant aspect is machine ethics, which aims to create machines that adhere to ethical frameworks. The widespread use of AI has raised ethical dilemmas and concerns, leading to the need for controlling the behavior of intelligent machines within certain ethical

frameworks. The article provides an overview of AI, introduces fundamental concepts, and outlines strategies and key themes in AI ethics. It contributes to the ongoing debate on the ethical use of AI by discussing issues such as the legal identity of autonomous machines, legal problems related to AI, and proposed solutions with examples of applications.

The fourth article, "An Application with Python Software for the Classification of Chemical Data" authored by Oğuz Akpolat and Gonca Ertürk, presents the use of data mining algorithms in the field of chemistry to analyze and predict the relationships between various parameters measured in wastewater treatment. The focus is on wastewater treatment processes that aim to eliminate biodegradable organic matter, suspended solids, harmful substances, nitrogen and phosphorus compounds, and pathogens. The monitoring and control of these processes relies on continuous measurement of wastewater and activated sludge characteristics. Parameters such as biochemical oxygen demand (BOD5), chemical oxygen demand (COD), total organic carbon (TOC), and dissolved oxygen (DO) are crucial for assessing wastewater properties. The study aims to establish mathematical associations between BOD5 and other parameters to estimate BOD5 quickly.

The fifth article, "A Study of Ransomware, Understanding and Protection" discusses the escalating threat of ransomware to computer systems and data and presents a comprehensive study on the subject. The research aims to analyze the impact of ransomware attacks on organizations and individuals while evaluating existing countermeasures. It includes a review of literature and security provider sources, as well as analysis of real-world ransomware incidents, revealing the increasing complexity and broad targeting of such attacks, leading to significant financial and reputational risks for victims. The research emphasizes that traditional security measures might not be sufficient, advocating the use of Artificial Intelligence and a multi-layered defense approach to effectively mitigate ransomware incidents.

In the sixth and the final article, "A Comparative Assessment of Frequentist Forecasting Models: Evidence From S&P 500 Pharmaceutical Index", the authors utilize ARIMA, GARCH, and Neural Network Autoregression (NNAR) to forecast the S&P 500 Pharma Stock Index. They aim to determine the most accurate forecasting method based on Mean Average Forecasting Error (MAFE). The research provides valuable insights for health and financial econometricians focusing on pharmaceutical indexes and investment companies using forecasting methods to make investment decisions.

The *Journal of Data Applications* can be accessed online at IUPress and DergiPark. Our journal can also be followed on LinkedIn. We express our gratitude to all the authors who have selected our journal as the platform to publish their scientific research. We extend our appreciation to the reviewers who have accepted our referee invitations and provided their valuable feedback. Our heartfelt thanks also go to our readers, IU Press, and DergiPark for their interest in our journal and their dedicated efforts.

We hope you find this issue insightful and look forward to receiving your submissions for future issues of the *Journal of Data Applications*.

Assoc. Prof. Zeki ÖZEN
Editor-in-Chief

Assoc. Prof. Elif KARTAL
Co-Editor-in-Chief

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

Istanbul's Community Mobility Changes During the COVID-19 Pandemic: A Spatial Analysis

Ahmet Okan ARIK[1], Gülsüm Çiğdem ÇAVDAROĞLU[2]

ABSTRACT

COVID-19 was the most recent pandemic to strike humanity. Moreover, this pandemic occurred during the most active period of global interaction and mobility, unlike pandemics like cholera, plague, and flu in earlier centuries. Many countries restricted domestic mobility after suspending international mobility to prevent the pandemic from spreading. Although these policies differ from nation to nation, they have affected the mobility of communities. This study examined spatial and non-spatial independent variables that affected how the community's mobility patterns changed in various locations, including parks, transit stations, workplaces, grocery and pharmacies, and residential areas in Istanbul, Türkiye. The impact of the independent spatial variables on the mobility changes was examined after identifying the non-spatial independent variables influencing the mobility changes in 6 different areas. It was determined that the altitude variable, expected to impact how mobility changed, had no overall impact on the dependent variable. On the other hand, the dependent variables representing the mobility changes were affected by the independent variables representing the county center's latitude and longitude values and whether the county is located near the sea. Regression analysis across Türkiye will be performed in upcoming studies using an updated version of the methodology used in this study.

Keywords: Urban Mobility, Mobility Analysis, Spatial Analysis, COVID-19, Google Community Mobility Report

DOI: 10.26650/JODA.1215566

[1]Istanbul University, Institute of Science, Department of Informatics, Istanbul, Turkiye
[2]Işık University, Faculty of Economics, Administrative and Social Sciences, Information Technologies, Istanbul, Turkiye

ORCID: O.A.A. 0000-0002-4875-4800;
G.Ç.Ç. 0000-0002-6572-1605

Corresponding author:
Ahmet Okan ARIK,
Istanbul University, Institute of Science, Department of Informatics, Istanbul, Turkiye
E-mail: aokanarik@gmail.com

Submitted: 07.12.2022
Revision Requested: 05.01.2023
Last Revision Received: 19.01.2023
Accepted: 23.01.2023

Citation: Arik, A. O., & Cavdaroglu, G. C. (2023). Istanbul's community mobility changes during the COVID-19 pandemic: a spatial analysis. *Journal of Data Applications, 1*, 1-17. https://doi.org/10.26650/JODA.1215566

Introduction

Due to the new coronavirus's rapid rate of spread and potential danger, the disease was recognized as a global pandemic on March 11, 2020 (WHO, 2020). A highly contagious disease known as COVID-19 is thought to have started in the Chinese city of Wuhan and spread worldwide (Saha et al., 2020). The Omicron variant is currently the dominant variant in rapid circulation, whereas Alpha, Beta, Gamma, and Delta were the previous dominant variants. In addition, the Omicron variant, which spreads more rapidly than the Delta variant, has been found to be a milder form of the disease (Vitiello et al., 2022).

The fact that the vaccines are taken with reminder doses against the dominant variant Omicron and that the total number of vaccine doses administered as of May 7, 2022, was 11,579,263,039 may indicate that the end of the pandemic has recently been approached (Vitiello et al., 2022; WHO, 2022a). However, COVID-19, with its many variants, caused 513,955,910 confirmed cases of COVID-19 and 6,249,700 related deaths globally as of 5:31 pm CEST, May 6, 2022, globally (WHO, 2022a). From January 3, 2020, to May 9, 2022, at 5:13 pm CEST, in Türkiye, 15,043,379 cases and 98,846 deaths were reported, and 147,426,248 doses of the vaccine were administered (WHO, 2022b).

The COVID-19 pandemic has disrupted several areas of life, such as the education sector, which has transformed itself into the online domain, the business life that started with hybrid and remote working, and the changing of shopping habits with the widespread use of online shopping and contactless payment. Another issue that had to undergo changes and restrictions during the pandemic was mobility. Mobility is an inseparable part of people's daily life for required reasons for workplaces, hospitals, and activities such as entertainment and shopping. However, Istanbul has intense mobility as its population is nearly 20 million, and it is an international transit hub due to its geographical location. Therefore, after the start of the pandemic, the Republic of Türkiye, like many countries, restricted both domestic and international mobility with measures such as lockdowns, quarantines, and flight bans in all its provinces, including Istanbul, and banned many community activities such as sports, education, and cultural activities.

This study aims to explain the differences in mobility change rates using the mobility changes dataset enriched with socioeconomic and geographical features. Mobility changes include changes in retail and recreation visit changes, market and pharmacy visit changes, park visit changes, transit station usage statistics, and changes in residence time at workplaces and residences in Istanbul districts between March 2020 and March 2022. By explaining community mobility changes, it is aimed to indicate how the COVID-19 measures taken affected community mobility and the course of the pandemic and to provide policymakers with insight into developing a roadmap for potential pandemics.

Literature Review

Rice & Pan examined the factors affecting park visits at the county level for the early stages of COVID-19. The study depended on Google Community Mobility Reports (GCMR) (Google, 2020). The study revealed that for 97 districts in the western United States, changes in park visits in the spring of 2020 were caused by climate differences due to altitude and latitude. In addition, they point out that counties with older populations and longer stay-at-home orders may be some of the likely reasons for reduced park visits, fear of symptoms among seniors, and county-level travel restrictions. Due to decreased park visits among older people and long-term stay-at-home orders, they recommend examining the relationship between age and stay-at-home orders and park visitation and well-being (Rice & Pan, 2021).

Saha et al. examined the percentages of change in community mobility in India from February 15 to April 30. The dataset for this study was obtained from GCMR. Based on baselines created across India, the results show that workplace mobility decreased by 56.7%, grocery and pharmacy mobility decreased by 51.2%, park visits decreased by 46.3%, and transit station mobility increased by 66%. In addition, retail and recreation mobility decreased by 73.4%. However, lockdowns caused a 23.8% increase in mobility in residential areas (Saha et al., 2020).

Wen et al. examined community mobility and preferred mode of transport at various alert levels using GCMR and Apple maps data. According to the results, the highest-level set, alert level 4, is vital in reducing the quarantined person's mobility and diversity of transport modes. It is stated that the studies carried out to prevent contamination significantly negatively impact retail and recreation mobility. While the use of public transport decreased significantly, it is stated that this was relatively lower in the state of Wellington. It is emphasized that the recovery rate in retail and leisure mobility in the Otago state lags behind other areas. Another contribution of the study to the literature is that the GCMR and Apple maps data used were tested for consistency with the New Zealand Transport Agency data, confirming that they represented the entire population (Wen et al., 2022).

Sulyok & Walker investigated the relationship between the volume of COVID-19 cases and GCMR data and social activity and community mobility. It is stated that after COVID-19 became a global situation, mobility decreased. This decrease may be due to legal restrictions or people's fear of the disease. Sulyok & Walker stated that the decrease in mobility in some countries before legal restrictions reveals the importance of personal infection risk and behavior change. It is stated that it can be understood by cultural, social, and economic factors (Sulyok & Walker, 2020).

Many datasets made available during the COVID-19 pandemic have made it possible to analyze community mobility dynamics and spatial distribution throughout the quarantine. Beria & Lunkar used the "Italy Coronavirus Disease Prevention Map package" data produced

by Facebook in their studies. The first of the questions the authors seek to answer is to what extent people stay at home. The findings show that people with mobility ratio and range of motion are significantly reduced. In the second research question, the mobility of the people before the quarantine period was examined. According to the results, it is stated that some of the population preferred to go abroad to avoid restrictions most of the non-local travel is to neighboring provinces and long travel is below the usual statistics. In the final stage, the position of people during the lockdown was examined. According to the results, it is stated that the population has decreased starting in the northern provinces, especially in the big cities, and the population has been directed toward rural areas (Beria & Lunkar, 2021).

Aloi et al. examined the effects of quarantine measures, which went into effect on March 15, 2020, on urban mobility in Santander, a city in northern Spain. The study collected data from traffic meters, intelligent transport systems data, traffic camera records, and environmental sensors compared with the travel flows and durations before and during the quarantine. According to the study's findings, mobility, the use of public transportation, NO2 emissions, and traffic accidents all decreased by, respectively, 76%, 93%, 60%, and 67% (Aloi et al., 2020).

Bonaccorsi et al. examined the effects of the measures taken to combat the COVID-19 pandemic in Italy on the socioeconomic circumstances of Italian citizens. Mobility restrictions are modeled as an exogenous shock akin to a natural disaster, and a large dataset of human mobility can be examined in real time. It has been observed that the effects of lockdown measures are higher in municipalities with high financial capacity. Furthermore, it is stated in the study that the decrease in mobility is more substantial in municipalities where inequality is higher than in others and per capita income is lower. According to these results, the authors state the necessity of fiscal policies targeting poverty and inequality (Bonaccorsi et al., 2020).

Chan analyzed which features are associated with decreases in mobility for Canada during the COVID-19 outbreak based on Facebook's data (Movement Range Maps). According to the study, there were significant differences in the degree of social distancing in April compared to before February. Another socioeconomic finding is that people who live in multi-flat buildings are less mobile than those who do not. Those with more challenging living conditions are less likely to stay home during a pandemic (Chan, 2020).

Chang et al. (2020) examined human mobility and connectivity during the COVID-19 outbreak in Taiwan in collaboration with Facebook's "Data for Good". Different provinces were determined as density points. The study on these points states that urban travel discounts have more impact than intercity travel discounts due to the risk of pandemics. Furthermore, it is stated that the findings can direct future disease surveillance and travel restrictions in-laws after the controls are eased (Chang et al., 2020).

Wielechowski et al. conducted province and voivodeship level analysis of the COVID-19 pandemic's impact on Poland's public transportation mobility**.** Data sources included the Oxford COVID-19 Government Response Tracker, GCMR, and Polish Ministry of Health data. The study's findings, which account for March 2 and July 19, 2020, demonstrate that Poland's public transportation mobility barely changed. Moreover, there is a strong, negative, and significant correlation between changes in mobility in public transportation and the COVID-19 measures implemented by the Polish government. In conclusion, it has been demonstrated that the government's efforts to stop the pandemic's spread decreased Poland's mobility and increased social distance (Wielechowski et al., 2020).

Orro et al. examined the impact of COVID-19 on urban mobility. The authors examined the change in the number of passengers on a line basis, the use of stops, public transportation supply, duration, and reliability of the A Corua city bus network in Spain, stating that this effect varies depending on the type of public transportation used. The data set includes information on bus boarding, intelligent card usage, and automatic vehicle location. The study's findings indicate that the pandemic significantly impacts public transportation more than other types of traffic. According to the statement, this could result from reduced public transportation, reduced traffic, suspension of street parking fees, easy parking, or fear of contamination. The number of passengers decreased to between 6% and 20% of the reference values during the quarantine period, while the use of public transportation stations near shopping centers and universities was almost eliminated during periods of high cases, while the use of stations nearby in other commercial areas was at a low level (Orro et al., 2020).

Materials and Methodology

Study Area

Istanbul is represented as Türkiye's political, economic, and cultural hub. The study region of this study includes 39 districts in Istanbul. The study area is shown in Figure 1.

Figure 1. *Study area.*

Data Collection and Preprocessing

Several datasets were used to investigate the factors affecting the changes in the mobility areas in the counties of Istanbul during the COVID-19 pandemic. Since the study's primary purpose is to examine the spatial and non-spatial independent variables affecting the dependent variables (community mobility variables), the data sets are grouped according to this situation and explained in the following sections. The dataset, source codes, and regression results are also available on GitHub (datastd-dev, 2022). Table 1 summarizes the study's twelve independent and six dependent variables. The dependent variables will be referred to by the names given in this table in the rest of the study.

Table 1. *Data sets and attributes (D: dependent, I: independent, SV: spatial variable, NSV: Non-spatial variable).*

Type	Name	Data Set Group	Spatial	Explanation	Attributes / Units
D	GCMR	Google community mobility reports	No	GCMR show how visits to places are changing in each geographic region.	Attributes: GP, P, T, RR, R, W. Unit: percentage.
I	SV_1	Elevation	Yes	The elevation of the county center in meters.	Type: numeric. Unit: meter.
I	SV_2	Seaside	Yes	Whether the county is by the sea.	Type: boolean. Values: 0:no, 1: yes.
I	SV_3	Latitude	Yes	Latitude of the county.	Type: numeric.
I	SV_4	Longitude	Yes	Longitude of the county.	Type: numeric.
I	NSV_1	Average Monthly Income (2019-2020)	No	Monthly average household income.	Type: numeric. Unit: TL.
I	NSV_2	Number of Illiterate People (2020)	No	The number of illiterate people aged 6 and over.	Type: numeric. Unit: number of persons.
I	NSV_3	Number of Shopping Centers (2022)	No	The number of the shopping malls.	Type: numeric. Unit: piece.
I	NSV_4	Average Number of Persons in the Household (2022)	No	Average number of people in households.	Type: numeric. Unit: number of persons.
I	NSV_5	Number of Undergraduate and Graduate Graduates (2022)	No	Number of undergraduate and graduate graduates.	Type: numeric. Unit: number of persons.
I	NSV_6	Elderly Population Ratio (2022)	No	Ratio of elderly population to county population.	Type: ratio.
I	NSV_7	Middle Aged Population Ratio (2022)	No	Ratio of middle-aged population to county population.	Type: ratio.
I	NSV_8	Young Population Ratio (2022)	No	Ratio of young population to county population.	Type: ratio.
I	NSV_9	Population Density	No	Population density in the county.	Type: numeric. Unit: percentage.

Figure 2 shows an example view of the dataset.

	name	SV_1	SV_2	SV_3	SV_4	NSV_1	NSV_2	NSV_3	NSV_4	NSV_5	NSV_6	NSV_7	NSV_8	NSV_9	rere	grophar	parks	transit	work	resi
2	Fatih	29	1	41.0166667	28.9333333	9714	7101	1	3.04	59627	16	53	31	24187	-31.1202	-7.8046	-10.5792	16.5327	-21.4768	6.8907
3	Zeytinburnu	24	1	40.990635	28.89614	6703	4576	2	3.72	32967	10	51	38	23638	-24.7555	21.9448	-16.0055	7.5473	-15.2117	6.2617
4	Güngören	45	0	41.0166667	28.8833333	6232	4395	1	3.42	41836	13	52	35	40042	-18.8974	16.3428	28.3451	14.0693	-18.4372	6.6685
5	Bakırköy	21	1	40.968155	28.8228	16269	1390	8	2.99	75719	21	52	27	7540	-35.918	13.9183	-18.6011	5.9621	-31.0988	8.0785
6	Bahçelievler	68	0	40.9975	28.8505556	8597	9268	3	3.46	99363	11	54	35	34845	-33.8825	10.4852	3.3114	20.5874	-25.1585	7.5423
7	Bağcılar	92	0	41.0455556	28.8405556	5881	13562	3	3.94	69904	8	51	41	17552	-22.082	-7.8537	2.2076	36.295	-18.1038	6.4836

Figure 2. *An example view from the dataset.*

Dependent Variables

Google Community Mobility Reports

GCMR shows the mobility changes in communities throughout the COVID-19 pandemic. Mobility changes in public stations, parks, or specific locations could be detected using these reports. Places, where mobility change could be examined in GCMR are grocery & pharmacies (GP), parks (P), transit stations (T), retail & recreation (RR), residential (R), and workplaces (W).

GP are areas such as markets, pharmacies, and food shops. P provides mobility change in local parks, national parks, dog parks, and public gardens. T provides mobility changes at public stations such as buses, subways, and trains. RR provides mobility changes in restaurants, cafes, shopping malls, libraries, and movie theaters. R denotes the change in mobility in the seats. Finally, W provides mobility change in workplaces. The dates used to determine the reasons for the changes are between March 2020 and March 2022 in the dataset. It is aimed to reveal to what extent the mobility data presented under six different titles are affected by spatial and non-spatial independent variables in this study. For this reason, these six different mobility data will be handled one by one, and the relationship between them and the independent variables explained in the following section will be revealed.

Independent Variables

Community mobility changes based on the counties of Istanbul vary greatly. This study aims to reveal the reasons for these differences. Community mobility can be affected by many independent variables. Therefore, first, it was investigated to what extent the changes were affected by spatial and non-spatial variables. For this reason, independent variables are discussed under two main headings, spatial and non-spatial variables.

Spatial Variables

Spatial variables that affect community mobility changes based on Istanbul counties will be discussed in this section. Some literature studies have revealed that community mobility changes are more affected by spatial parameters than non-spatial parameters. Therefore, county elevation data, the information on whether the county is located on the sea coast, and the

latitude-longitude information of the county were used to investigate this situation specific to the province of Istanbul.

Elevation data is the first spatial variable predicted to affect community mobility change data. The province of Istanbul, with a surface area of 5,343 km², is a province where the altitude varies based on counties. For example, there is a difference of 318 meters between the lowest elevation value (Adalar county, 6m) and the highest elevation value (Maltepe county, 324m). In order to determine whether the change in community mobility is affected by this situation, the altitude values of the counties' centers were collected *(*İstanbul İlçeleri Haritası*, 2022)*.

Counties with a seacoast are assigned a value of 1, and counties that do not have a value of 0. This information was manually added to the data set by examining the Istanbul map.

Istanbul counties' latitude and longitude information was obtained from GitHub (NovaYear, 2019).

Non-spatial Variables

Non-spatial variables that affect mobility changes based on Istanbul counties will be discussed in this section.

The current number of shopping malls in the counties, population density (person/km2), the number of individuals with undergraduate and graduate degrees, and the ratio of elderly, young, and middle-aged individuals are obtained from the estimating real estate data analytics and insight platform Endeksa (*Endeksa*, 2017).

The monthly average household income data based on Istanbul counties, obtained from the report published by the Istanbul Governorship Open Door Branch Directorate (*Açık Kapı-İstanbul'un Sosyo Ekonomik Analizi*, 2021), was also added to the study as an independent variable. The purpose of adding the variable is to measure the contribution of monthly income amounts of individuals on a county basis to changes in community mobility. For example, in counties with high-income levels, individuals may have stopped going to their workplaces and chose to stay at home. However, individuals may have had to continue working in counties with low-income levels even if they were banned.

The number of illiterate individuals aged six and over on an Istanbul county basis was also obtained from the report published by the Istanbul Governorship Open Door Branch Directorate (*Açık Kapı-İstanbul'un Sosyo Ekonomik Analizi*, 2021) and added to the study as an independent variable. The purpose of adding the variable is to determine whether the education level of the individuals on a county basis affects compliance with the rules and measure its contribution to the community mobility changes.

Methodology

The first step was to create a dataset to identify the spatial and non-spatial factors influencing the COVID-19 based mobility changes in the counties of Istanbul. When choosing the features for the dataset, it was taken into account that the geographic characteristics of the counties and the demographic characteristics of their inhabitants could best explain the change in mobility at the specific places during the COVID-19 period. Then, to determine how well the spatial and sociodemographic features provided by the dataset explain the change in mobility, the variable selection method was applied in the second stage using the R package programming language. As a result, it was intended to quantify how independent variables affected changes in mobility.

The regression analysis determined the probability of producing the highest adjusted R2 value for non-spatial parameters. Then, the effect of spatial parameters on adjusted R2 value was measured for these values.

Due to the large number of spatial and non-spatial variables included in the study, it required a complex calculation to determine which variables affect the target variable and to find the differences between the effect levels of the influencing variables. The effects of 13 spatial and non-spatial variables on mobility changes under six different headings were investigated. The effect of 13 variables on six target variables will require many calculations considering different combinations. For this purpose, it was decided to use the "Best Subsets Regression Essentials" method of the R package programming language.

The study's primary purpose is to calculate the extent to which spatial variables affect target variables. For this reason, when non-spatial variables affecting the target variable were included in the regression, the extent to which the four considered spatial variables affected the regression was investigated. In order to reveal this detail, the extent to which the obtained quality measures (Adjusted R2) changed when four spatial variables were included in the regression or not. In order to select the best model, the overall performances of the models were compared, and some statistical metrics and strategies were determined to select the best one. The estimation error of each model was measured, and it was decided to choose the one with the lower estimation error.

"Best Subsets Regression Essentials" is a method for choosing a model that uses best subsets regression to test every possible combination of the predictor variables and then chooses the best model based on some statistical criteria. This method is also known as "all possible models" and "all possible regressions." The "leaps" library should be imported to use this methodology in the R package programming language environment.

A distinct least squares regression best subset should be constructed for each possible variable combination in order to carry out best subset selection. This means that all $\binom{p}{2} = p(p-1)/2$ models with exactly two variables should be fitted to all p models with exactly one variable. Now, the objective is to determine which of the resulting models is the best. The problem of choosing the best model from the 2^p options that are taken into account by best subset selection is not simple. Typically, there are two stages to this. The required statistics can be calculated with the help of the functions in the "leaps" library. These statistics are metrics such as Adjusted R2, Cp, and BIC. Using one of these criteria, the best model can be determined. For example, the Adjusted R2 criterion was used to determine the best model within the scope of the study. The adjusted R2 represents the proportion of variation in the outcome that is explained by the variation in predictors values. The higher the adjusted R2, the better the model. Algorithm 1 describes best subset selection methodology.

Algorithm 1 Best subset selection.

1: M_0 denotes the null model. M_0 contains no independent variables. This model makes a prediction about the sample mean for each observation.

2: For t = 1, 2, …p:

2.1: Fit all $\binom{p}{t}$ models that contain exactly t independent variables.

2.2: pick the best among these $\binom{p}{t}$ models (M_t). Here best is defined as having the largest Adjusted R2.

3: Select a single best model from among M_0, M_1, …, M_t using Adjusted R2.

Results and Discussion

The effect states and effect levels of the spatial and non-spatial parameters affecting the six dependent variables included in the study are different. The non-spatial parameters and the effects of these parameters on the regression specific to the dependent variables are given in Table 2. Expressions marked with X mean that the relevant independent variable affects the relevant dependent variable.

Table 2. *Non-spatial parameters and their effects on dependent variables.*

Parameter	GP	P	T	RR	R	W
NSV_1			X		X	X
NSV_2					X	
NSV_3		X	X	X		X
NSV_4				X		X
NSV_5	X	X	X	X	X	X
NSV_6	X				X	X
NSV_7	X		X	X		X
NSV_8	X					
NSV_9		X	X		X	X

Spatial parameters and the effects of these parameters specific to dependent variables on regression are given in Table 3.

Table 3. *Spatial parameters and their effects on dependent variables.*

Parameter	GP	P	T	RR	R	W
SV_1						
SV_2	X	X	X		X	
SV_3		X	X		X	
SV_4	X	X			X	X

In the following section, the information in the table is summarized on the basis of community mobility change parameters.

- *Spatial and non-spatial variables affecting the GP parameter*: The independent non-spatial variables affecting the community mobility change related to markets and pharmacies (GP) were determined as "number of undergraduate and graduate graduates," "elderly population ratio," "middle-aged population ratio," and "young population ratio." The independent spatial variables that positively affect the community mobility change related to markets and pharmacies (GP) were determined as "seaside" and "longitude." Accordingly, the county is located by the sea, and the longitude value of the county center increases the mobility in the market and pharmacies. At the same time, the elevation and latitude variables decrease the visits to the market and pharmacies.
- *Spatial and non-spatial variables affecting the P parameter*: The independent non-spatial variables affecting community mobility change in parks (P) were determined as "number of shopping centers," "average number of persons in the household," and "number of undergraduate and graduate graduates." Accordingly, the county's household size, the number of shopping malls, and the number of undergraduate and graduate graduates affected the park visits. The independent spatial variables that positively affect the community mobility change in the parks were determined as "seaside," "latitude," and "longitude." Accordingly, the fact that the county is located by the sea, while the latitude and longitude values of the county center increase the park visits, the elevation variable decreases the park visits.
- *Spatial and non-spatial variables affecting the T parameter*: The independent non-spatial variables affecting community mobility change in public transportation were determined as "average monthly income," "number of shopping centers," "number of undergraduate and graduate graduates," "elderly population ratio," and "population density." Accordingly, the household income level, the number of shopping malls in the county, the county's population density, the number of undergraduate-graduate graduates, and the middle-aged population ratio affected mobility in public

transportation. The independent spatial variables that positively affect the community mobility change in public transportation were determined as "latitude" and "longitude." Accordingly, while the latitude and longitude values of the county center increased the mobility in public transportation, the elevation and seaside variables decreased the mobility in public transportation.

- *Spatial and non-spatial variables affecting the RR parameter*: The independent non-spatial variables that affect the retail and recreation community mobility change were determined as "number of shopping centers," "average number of persons in the household," "number of undergraduate and graduate graduates," and "middle-aged population ratio." Accordingly, the number of shopping malls, household size, number of undergraduate-graduate graduates, and middle-aged population ratio affected retail and recreation mobility in the county. The independent spatial variable that positively affects the mobility of the retail and recreation community could not be determined.
- *Spatial and non-spatial variables affecting the R parameter*: The independent non-spatial variables affecting the residential community mobility change were determined as "average monthly income," "number of illiterate people," "number of undergraduate and graduate graduates," "elderly population ratio," and "population density." Accordingly, the average household income level, the number of illiterate people, the number of undergraduates and graduates, the proportion of the elderly population, and population density affected residential mobility. The independent spatial variables that positively affect the residential community mobility change were determined as "seaside," "latitude," and "longitude." Accordingly, the latitude and longitude values of the county center and the fact that the county is located by the sea increase the duration of staying at home, while the altitude variable decreases the duration of staying at home.
- *Spatial and non-spatial variables affecting the W parameter*: The independent non-spatial variables affecting workplace visits are "average monthly income," "number of shopping centers," "average number of persons in the household," "number of undergraduate and graduate graduates," "middle-aged population ratio," and "population density." Accordingly, the average household income level, the number of shopping malls in the county, the household size, the number of people with undergraduate and graduate degrees, the ratio of the middle-aged population, and population density affected workplace visits. The independent spatial variable that positively affects workplace visits has been determined as "longitude." Accordingly, while the longitude value of the county center increased the workplace visits, the variables of altitude, seaside, and latitude decreased the workplace visits.

Table 4 shows how much different combinations of non-spatial variables affect the adjusted R2 parameter in the regression model in which all spatial variables are included.

Table 4. *Change in Adjusted R2 according to different combinations of non-spatial variables.*

Parameter	Combination	Adjusted R2
GP	Add Seaside	Increases from 0.1531 to 0.1586
	Remove Elevation	Increases from 0,1378 to 0.1586
	Remove Latitude	Increases from 0.1325 to 0.1586
	Add Longitude	Increases from 0.1077 to 0.1586
	Remove Elevation & Add Seaside	Increases from 0.1316 to 0.1586
	Remove Latitude & Add Seaside	Increases from 0.1280 to 0.1586
	Add Longitude & Add Seaside	Increases from 0.1190 to 0.1586
	Remove Elevation & Remove Latitude	Increases from 0.1106 to 0.1586
	Add Longitude & Add Seaside & Remove Elevation	Increases from 0.1081 to 0.1586
	Add Seaside & Remove Elevation & Remove Latitude	Increases from 0.1053 to 0.1586
	Add Longitude & Remove Elevation	Increases from 0.0996 to 0.1586
P	Remove Elevation	Increases from 0.5096 to 0.5240
	Add Seaside	Increases from 0.4798 to 0.5240
	Add Longitude	Increases from 0.4385 to 0.5240
	Add Seaside & Remove Elevation	Increases from 0.4697 to 0.5240
	Add Seaside & Add Longitude	Increases from 0.4176 to 0.5240
	Add Seaside & Add Latitude	Increases from 0.3739 to 0.5240
	Add Latitude & Add Longitude	Increases from 0.3612 to 0.5240
	Add Seaside & Add Latitude & Add Longitude	Increases from 0.3501 to 0.5240
T	Remove Elevation	Increases from 0.7044 to 0.7065
	Remove Seaside	Increases from 0.6980 to 0.7065
	Add Longitude	Increases from 0.6612 to 0.7065
	Remove Elevation & Add Longitude	Increases from 0.6524 to 0.7065
	Add Latitude & Add Longitude	Increases from 0.5956 to 0.7065
RR	Remove Elevation	Increases from 0.3213 to 0.3358
	Remove Seaside	Increases from 0.3160 to 0.3358
	Remove Latitude	Increases from 0.3189 to 0.3358
	Remove Longitude	Increases from 0.3187 to 0.3358
	Remove Elevation & Remove Longitude	Increases from 0.3060 to 0.3358
	Remove Elevation & Remove Latitude	Increases from 0.3039 to 0.3358
	Remove Latitude & Remove Longitude	Increases from 0.3024 to 0.3358
	Remove Seaside & Remove Elevation	Increases from 0.3002 to 0.3358
	Remove Seaside & Remove Longitude	Increases from 0.2984 to 0.3358
	Remove Seaside & Remove Latitude	Increases from 0.2981 to 0.3358
	Remove Evl. & Remove Latitude & Remove Longitude	Increases from 0.2906 to 0.3358
R	Add Seaside	Increases from 0.7914 to 0.8039
	Add Longitude	Increases from 0.7755 to 0.8039
	Add Seaside & Remove Elevation	Increases from 0.7922 to 0.8039
	Remove Elevation & Add Longitude	Increases from 0.7851 to 0.8039
	Add Seaside & Add Longitude	Increases from 0.7746 to 0.8039
	Remove Elevation & Add Seaside & Add Longitude	Increases from 0.7805 to 0.8039
	Add Seaside & Add Latitude & Add Longitude	Increases from 0.6787 to 0.8039
W	Add Longitude	Increases from 0.8220 to 0.8498
	Remove Latitude & Add Longitude	Increases from 0.8195 to 0.8498
	Add Longitude & Remove Seaside	Increases from 0.8166 to 0.8498
	Add Longitude & Remove Elevation	Increases from 0.8161 to 0.8498

Limitations

Seasonality in the data can be misleading, as a baseline needs to be provided for analyzing activity in park visits. The report may be more valid when working with data covering other mobility activities less affected by seasonality.

GCMR does not provide mobility for those who are not mobile or whose location services are disabled. Therefore, the data shared by Google only represents some smartphone users or the entire population in Istanbul.

Conclusion

This study examined spatial and non-spatial independent variables affecting community mobility change parameters in 6 different domains presented by Google. Afterward, the impact rates and directions of the independent spatial variables on the regression were analyzed.

The results showed that the non-spatial independent variable "average monthly income" impacted mobility in public transportation, the duration of stay at home, and the amount of time spent at the workplace. The duration of stay at home was the only factor impacted by the non-spatial independent variable "number of illiterate people." The "number of shopping centers" non-spatial argument impacted mobility in parks, public transportation, retail, and recreation. The non-spatial independent variable "average number of persons in the household" impacted workplace length of stay and mobility in retail and recreation. The non-spatial independent variable "number of undergraduate and graduate graduates" impacted all six areas' mobility changes. The mobility in markets and pharmacies, the amount of time spent at home, and the amount of time spent at work were all impacted by the non-spatial independent variable known as the "elderly population ratio." The non-spatial "middle-aged population ratio" independent variable affected mobility in markets and pharmacies, public transportation, retail and recreation, and length of stay at workplaces. The non-spatial "young population ratio" independent variable only affected the mobility in markets and pharmacies. Finally, the non-spatial "population density" argument affected mobility at park visits, public transport, length of stay at home, and length at work.

According to the findings, the spatial "elevation" independent variable did not affect the mobility in any area positively or negatively. The spatial "seaside" independent variable positively affected mobility in markets and pharmacies, park visits, mobility in public transport, and duration of stay at home, and negatively affected activity in retail and recreation and duration of stay at work. The spatial variable of latitude positively affected park visits, mobility in public transport, length of stay at home, and negatively affected activity in grocery stores and pharmacies, retail and recreation, and length of stay at workplaces. The spatial "longitude"

independent variable positively affected mobility in markets and pharmacies, park visits, stay at home and stay at workplaces, and negatively affected mobility in public transportation and mobility in retail and recreation.

Table 5 indicates the directions and impact rates of independent spatial variables' effects on changes in mobility.

According to the findings obtained from the table, the independent spatial variable that most positively affects mobility in markets and pharmacies and duration of stay at the workplace is "longitude." The independent spatial variable that most positively affects park visits, mobility in public transport, and length of stay at home is "latitude." A spatial variable that positively affects mobility in retail and recreation was not found. The independent variable that was affected the most negatively was found to be "seaside."

Table 5. *The directions and impact rates (SV: spatial variable, D: direction, N: negative, P: positive).*

GMP	**GP**		**P**		**T**		**RR**		**R**		**W**	
SV	Rate	D	Rate	D	Rate	D	Rate	D	Rate	D	Rate	D
SV_1	14.91874	**N**	2.818391	**N**	0.288948	**N**	4.520502	**N**	0.108913	**N**	0.726555	**N**
SV_2	3.579245	**P**	9.202555	**P**	1.210322	**N**	**6.266106**	**N**	1.582957	**P**	0.65951	**N**
SV_3	19.75203	**N**	**28.32589**	**P**	**11.01983**	**P**	5.291971	**N**	**13.36279**	**P**	0.31097	**N**
SV_4	**47.25765**	**P**	19.48304	**P**	6.843956	**P**	5.373901	**N**	3.665533	**P**	**3.38267**	**P**

Ethics Committee Approval: Ethical approval is not applicable, because this article does not contain any data with human or animal subjects.

Peer-review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- A.O.A., G.Ç.Ç.; Data Acquisition- A.O.A., G.Ç.Ç.; Data Analysis/Interpretation- A.O.A., G.Ç.Ç.; Drafting Manuscript- A.O.A., G.Ç.Ç.; Critical Revision of Manuscript- A.O.A., G.Ç.Ç.; Final Approval and Accountability- A.O.A., G.Ç.Ç.;

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Açık Kapı-İstanbul'un Sosyo Ekonomik Analizi. (2021). http://www.istanbul.gov.tr/acik-kapiistanbulun-sosyo-ekonomik-analizi

Aloi, A., Alonso, B., Benavente, J., Cordera, R., Echániz, E., González, F., Ladisa, C., Lezama-Romanelli, R., López-Parra, Á., Mazzei, V., Perrucci, L., Prieto-Quintana, D., Rodríguez, A., & Sañudo, R. (2020). Effects of the COVID-19 Lockdown on Urban Mobility: Empirical Evidence from the City of Santander (Spain). *Sustainability*, *12*(9), Article 9. https://doi.org/10.3390/su12093870

Beria, P., & Lunkar, V. (2021). Presence and mobility of the population during the first wave of Covid-19 outbreak and lockdown in Italy. *Sustainable Cities and Society*, *65*, 102616. https://doi.org/10.1016/J.SCS.2020.102616

Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A. L., Valensise, C. M., Scala, A., Quattrociocchi, W., & Pammolli, F. (2020). Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences*, *117*(27), 15530–15535. https://doi.org/10.1073/pnas.2007658117

Chan, J. (2020). The geography of social distancing in Canada: Evidence from facebook. *Canadian Public Policy*, *46*(1), S19–S28. https://doi.org/10.3138/

Chang, M.-C., Kahn, R., Li, Y.-A., Lee, C.-S., Buckee, C. O., & Chang, H.-H. (2020). Variation in human mobility and its impact on the risk of future COVID-19 outbreaks in Taiwan. *MedRxiv*, 2020.04.07.20053439. https://doi.org/10.1101/2020.04.07.20053439

datastd-dev. (2022). *Spatial Regression*. GitHub. https://github.com/datastd-dev/SpatialRegression

Google. (2020). *COVID-19 Community Mobility Report*. https://www.google.com/covid19/mobility?hl=tr

Endeksa. (2017). https://www.endeksa.com/tr/

İstanbul İlçeleri Haritası. (2022). https://www.haritatr.com/istanbul-ilceleri-s22

Orro, A., Novales, M., Monteagudo, Á., Pérez-López, J. B., & Bugarín, M. R. (2020). Impact on City Bus Transit Services of the COVID–19 Lockdown and Return to the New Normal: The Case of A Coruña (Spain). *Sustainability 2020, Vol. 12, Page 7206*, *12*(17), 7206. https://doi.org/10.3390/SU12177206

Rice, W. L., & Pan, B. (2021). Understanding changes in park visitation during the COVID-19 pandemic: A spatial application of big data. *Wellbeing, Space and Society*, *2*, 100037. https://doi.org/10.1016/J.WSS.2021.100037

Saha, J., Barman, B., & Chouhan, P. (2020). Lockdown for COVID-19 and its impact on community mobility in India: An analysis of the COVID-19 Community Mobility Reports, 2020. *Children and Youth Services Review*, *116*, 105160. https://doi.org/10.1016/J.CHILDYOUTH.2020.105160

Sulyok, M., & Walker, M. (2020). Community movement and covid-19: A global study using google's community mobility reports. *Epidemiology and Infection* https://doi.org/10.1017/S0950268820002757

NovaYear. (2019). *Türkiye İl-İlçe Enlem ve Boylam Koordinatları*. GitHub. https://gist.github.com/NovaYear/4fe0fd530aca8b0e0bc0992b356fa32a

Vitiello, A., Ferrara, F., Auti, A. M., Di Domenico, M., & Boccellino, M. (2022). Advances in the Omicron variant development. *Journal of Internal Medicine*, *292*(1), 81–90. https://doi.org/10.1111/joim.13478

Wen, L., Sheng, M., & Sharp, B. (2022). The impact of COVID-19 on changes in community mobility and variation in transport modes. *New Zealand Economic Papers*, *56*(1), 98–105. https://doi.org/10.1080/00779954.2020.1870536

WHO. (2020). *Coronavirus Disease (COVID-19) — Events as they happen*. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen

WHO. (2022a). *Coronavirus (COVID-19) Dashboard*. https://covid19.who.int

WHO. (2022b). *Türkiye—WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data*. https://covid19.who.int

Wielechowski, M., Czech, K., & Grzęda, Ł. (2020). Decline in Mobility: Public Transport in Poland in the time of the COVID-19 Pandemic. *Economies 2020, Vol. 8, Page 78*, *8*(4), 78. https://doi.org/10.3390/ECONOMIES8040078

Supplementary: Istanbul districts

name	seaside	rakim_metre	enlem	boylam	rere	grophar	parks	transit	work	resi	income_m	oyb_2020	avm_sayisi	pop_dens (kisi/km²)	hane_2022	yfin_202281, yl,dr)	yasli_orani	genc_orani	ortaYas_orani
Fatih	1	29	41,0166667	28,9333333	-31,1202	-7,8046	-10,5792	16,5327	-21,4768	6,8907	9714	7101	1	24187	3,04	59627	16	31	53
Zeytinburnu	1	24	40,990635	28,89614	-24,7555	21,9448	-16,0055	7,5473	-15,2117	6,2617	6703	4576	2	23638	3,72	32967	10	38	51
Güngören	0	45	41,0166667	28,8833333	-18,8974	16,3428	28,3451	14,0693	-18,4372	6,6685	6232	4395	1	40042	3,42	41836	13	35	52
Bakırköy	1	21	40,968155	28,8228	-35,918	13,9183	-18,6011	5,9621	-31,0988	8,0785	16269	1390	8	7540	2,99	75719	21	27	52
Bahçelievler	0	68	40,9975	28,8505556	-33,8825	10,4852	3,3114	20,5874	-25,1585	7,5423	8597	9268	3	34845	3,46	99363	11	35	54
Bağcılar	0	92	41,0455556	28,8405556	-22,082	-7,8537	2,2076	36,295	-18,1038	6,4836	5881	13562	3	17552	3,94	69904	8	41	51
Bayrampaşa	0	69	41,0481503	28,9004553	-30,2049	20,9448	-20,0055	4,7034	-19,612	6,2024	6401	3310	1	26995	3,4	39114	13	34	54
Esenler	0	127	41,0794133	28,8538545	-33,7787	9,4031	6,5409	20,4503	-19,8511	6,0983	5237	8768	0	40570	3,81	38584	8	41	51
Küçükçekmece	1	10	41,008658	28,775342	-14,9617	17,5845	-7,8579	20,892	-19,1544	7,3306	6561	12380	2	20779	3,43	123539	10	37	53
Avcılar	1	9	41,0153479	28,7314618	-19,5369	26,8684	17,3838	11,7076	-20,5615	7,4672	6736	5869	1	11497	3,4	66008	11	36	53
Beylikdüzü	1	109	40,9910381	28,6498144	-20,6995	30,5983	17,2482	10,5396	-20,2746	8,58	7959	2682	2	9620	3,35	78912	11	36	53
Esenyurt	0	55	41,0342806	28,6801194	-23,8456	72,3705	9,7513	23,1789	-11,0355	7,1939	5562	14834	4	22265	3,55	91613	6	43	52
Beyoğlu	1	58	41,0382864	28,9703304	-39,8798	-8,9439	-30,612	-19,0386	-20,0847	7,1936	8779	5069	2	25155	3,08	32417	12	33	55
Eyüp	1	71	41,1871598	28,8829816	-33,5451	6,6563	-16,1339	14,9131	-21,8839	8,3975	8590	5641	4	1591	3,28	70538	11	35	54
Gaziosmanpaşa	0	104	41,0759477	28,9004553	-13,1694	32,3069	18,4276	23,362	-21,1803	7,1885	5553	8768	1	40648	3,57	56793	11	38	52
Sultangazi	1	84	41,1255794	28,8713314	-12,1434	-0,8232	12,6776	41,4922	-14,9426	6,7377	4023	11177	0	14930	4,08	38979	7	43	50
Kağıthane	0	42	41,071	28,97	-26,8142	3,0141	11,4153	23,6104	-26,015	8,433	7703	6734	2	29494	3,26	70086	9	35	56
Şişli	0	99	41,06	28,987	-45,2527	-10,8373	-30,3607	-10,6872	-27,4781	9,0372	14388	3692	8	24253	2,64	72257	16	26	58
Beşiktaş	1	21	41,068616	29,0285355	44,9126	23,6407	-12,2787	-19,7322	-36,7654	9,9332	19424	950	2	9290	2,51	76761	22	23	55
Başakşehir	0	110	41,077895	28,812551	-27,4577	36,0254	0,1898	12,403	-14,9563	8,4795	8301	5155	4	4895	3,77	78063	6	44	50
Büyükçekmece	1	8	41,034133	28,590003	-18,8115	24,7835	25,7553	5,1889	-19,1926	7,5939	6752	2812	1	1851	3,29	45019	13	35	52
Sarıyer	1	59	41,166328	29,04995	-32,0806	-2,1042	-11,8634	2,9836	-30,179	9,6188	13442	4573	3	2265	3,09	81717	14	31	55
Arnavutköy	1	163	41,2	28,733333	-10,3479	43,8968	-32,9959	41,1527	-4,2103	7,8002	3734	6291	1	685	4,06	20569	7	46	48
Silivri	1	14	41,080158	28,26829	-11,3593	6,66	61,5049	54,4257	-5,3264	8,065	4363	2189	1	240354	3,56	26530	13	32	55
Çatalca	1	105	41,148239	28,46773	-3,7393	-31,6987	34,0056	31,8009	-9,5176	5,2048	3914	876	0	64081	2,82	9406	18	31	51
Kadıköy	1	60	40,980141	29,08227	-40,2295	-13,6995	-27,9112	-4,101	-33,7063	10,6068	16601	2711	2	19279	2,47	209465	26	21	53
Adalar	1	6	40,8763772	29,095444	0	-32,8378	54,2476	0	-8,2649	0	12236	206	0	1457	2,55	3860	27	23	50
Üsküdar	1	30	41,032236	29,031938	-36,3019	20,5028	-17,3251	12,2062	-29,2322	8,9754	12852	6755	3	14879	3,09	142141	16	31	54
Ataşehir	0	110	40,9833333	29,1166667	-35,1762	3,3239	-8,9672	12,0587	-25,5984	9,7117	12098	6684	7	16903	3,15	102141	12	33	55
Ümraniye	0	152	41,0303	29,1065	-30,2309	-1,6985	-13,6831	15,7322	-20,1571	8,5204	6690	10382	6	15862	3,41	134935	9	36	55
Sancaktepe	0	184	41,0287028	29,2901829	-16,056	43,6322	11,4303	22,1947	-14,7131	8,4644	4843	8083	1	7368	3,69	53667	6	42	52
Maltepe	1	324	40,949047	29,174109	-37,862	73,3889	-5,5901	6,9494	-26,7213	9,6174	10617	6543	4	9717	3,06	131449	15	30	55
Beykoz	1	124	41,132719	29,10569	-18,2525	22,6449	-1,7814	8,7553	-26,0806	8,3196	6793	4466	0	750	3,3	43338	15	33	53
Çekmeköy	0	16	41,104235	29,3177272	-16,5516	42,1046	4,5186	28,9009	-16,5342	8,7252	6443	3527	0	2105	3,42	48764	8	38	54
Sultanbeyli	0	135	40,9611123	29,2669438	-16,2008	16,4449	18,2062	24,2885	-7,3524	6,6949	3995	7705	2	11838	4,17	24389	6	46	48
Kartal	1	65	40,899651	29,193649	-30,1844	63,734	-5,5969	7,9494	-23,7678	8,5874	7578	7372	3	11862	3,16	101877	13	32	55
Pendik	1	39	40,879326	29,258135	-25,8661	37,8718	-22,2855	11,2691	-19,5109	7,5683	5619	10726	5	4058	3,42	115130	9	38	53
Tuzla	1	108	40,842	29,295	-14,4945	0,5014	-7,9248	11,6452	-13,4631	7,7264	6267	3045	1	2171	3,33	51207	8	38	54
Şile	1	29	41,1763889	29,6127778	20,1286	43,265	-29,9898	56,5628	6,6562	4,1602	4565	609	0	48	2,66	5493	25	27	48

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

Modeling Automobile Sales in Turkiye with Regression-Based Machine Learning Algorithms

Merve BABAOĞLU[1], Ahmet COŞKUNÇAY[2], Tolga AYDIN[2]

ABSTRACT

The automobile sector is the locomotive of industrialized countries. The employment opportunities it creates are of great value because of its interconnectedness with other industries and the value it adds. Demand forecasting studies in such an important sector are one of the main drivers for the provision of raw materials and services needed in the future. In this study, 10 independent variables are used that directly or indirectly affect the level of car sales, which is our dependent variable. These variables are gross domestic product, real sector confidence index, capital expenditures, household consumption expenditures, inflation rate, consumer confidence index, percentage of one-year term deposits, and oil barrel, gold, and dollar prices. The dataset used consists of annual data between 2000 and 2021. To examine the sales forecast model, two variables that affect minimum sales are first extracted from the model using the least squares method. Linear Regression, Decision Tree, Random Forest, Ridge, AdaBoost, Elastic-net, and Lasso Regression algorithms are applied to build a predictive model with these variables. The Mean Squared Error (MSE), Mean Absolute Error (MAE), and coefficient of determination (R^2) are used to compare the performance of the predictive models. This study proposes an approach for sectors affected directly or indirectly by automotive sales to gain foresight on this issue.

Keywords: Automobile Sales, Regression, Demand Forecasting

DOI: 10.26650/JODA.1242645

[1]Artuklu University, Vocational Higher School of Mardin, Department of Computer Technologies, Mardin, Turkiye
[2]Ataturk University, Faculty of Engineering, Department of Computer Engineering, Erzurum, Turkiye

ORCID: M.B. 0000-0003-3030-8690;
A.C. 0000-0002-7411-310X;
T.A. 0000-0002-8971-3255

Corresponding author:
Merve BABAOĞLU,
Artuklu University, Vocational Higher School of Mardin, Department of Computer Technologies, Mardin, Turkiye
E-mail: mervebabaoglu@artuklu.edu.tr

Submitted: 26.01.2023
Revision Requested: 15.02.2023
Last Revision Received: 13.03.2023
Accepted: 15.03.2023

Citation: Babaoglu, M., Coskuncay, A. & Aydin, T. (2023). Modeling automobile sales in Turkiye with regression-based machine learning algorithms. *Journal of Data Applications, 1*, 19-33.
https://doi.org/10.26650/JODA.1242645

Introduction

The automotive sector is one of the sectors that has contributed the most to the transformation of the economy around the world since the twentieth century. This sector is also one of the main consumers of many branches of the manufacturing industry, from iron and steel to petrochemicals, and is an important trigger for their development. Especially for Turkiye, it is an indispensable sector as it creates high added value, brings high income, has a positive impact on employment and has a structure that allows bringing developed technologies to our country (Görener & Görener, 2008).

Although the automotive sector in Turkiye has a 50-year history, the sector, which started only with assembly in the mid-1950s, has reached its current position with increased investments after the Customs Union Agreement with the EU in 1996 (Dikmen, 2016).

Automobile production accounts for about 70% of the total mass production of motor vehicles in the world. This ratio is also true for Turkiye. Automobile production takes place on a much larger scale than the production of other motor vehicles. For this reason, automobile production also supports the production of other vehicles by creating a strong ancillary industry (Mmpvizyon, 2023).

Turkiye is a country with a majority young population.. The number of motor vehicles per capita of this population is increasing. The players in the automotive sector are also aware of this great strength. It is imperative for every player in this field to estimate how many sales there will be in the industry. Sales forecasting plays an important role in many areas, from purchasing raw materials to providing the necessary personnel for the industry, and from R&D investments to determining advertising expenditures.

The per capita rates of cars and motor vehicles in Turkiye are shown in Figure 1 (Euronews, 2022).

Forecasts are referred to when conclusions can be drawn about the future based on data from the past. These conclusions can be drawn using mathematical methods as well as subjective interpretations. Demand forecasting is the ability to predict how many sales will be made of a product using various methods. Demand forecasting is a very challenging problem because it is influenced by many factors. GDP, population, inflation, imports, exports, consumer confidence index, fuel prices, and gold and foreign exchange prices are just a few of these factors that affect automobile sales.

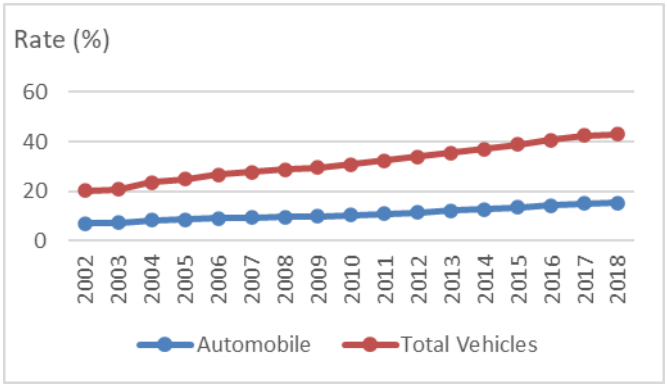

Figure 1. *Rates of cars and motor vehicles per capita in Turkiye.*

Machine learning methods, which are becoming increasingly popular in multivariate forecasting studies, continue to produce successful results. Choosing the right algorithms for a forecasting study is as important as choosing the right variables. The data must be properly analyzed, the algorithms must be understood, and the appropriate model must be selected for the area of focus.

The aim of the study is to make inferences about the future by looking at the history of automobile sales, which directly or indirectly affect many sectors. Thus, these sectors will be able to get ideas from these inferences to determine the steps they will take.

In this study, 10 different variables that directly or indirectly affect sales are used to model automotive demand. It is also noteworthy that so many variables are used in this study to predict automobile sales because the more variables used for the prediction, the greater the amount of explained variance in the dependent variable.

In order to process numerical data using machine learning methods and create the most accurate predictive models possible, regression algorithms are preferred in this study. Linear, Ridge, Lasso, ElasticNet, Random Forest, Decision Trees, and AdaBoost regression algorithms are preferred for this study. Models are built for sales forecasting and comparisons are made with some commonly used performance criteria.

In the second part of the study, studies in the field of automotive sales forecasting are presented from the literature. In the third part, regression algorithms, which are the most appropriate machine learning methods for the multivariate structure preferred for predictions, are described. The fourth part is about the methods used for prediction.

In the fifth part of the study, the findings obtained by the variables selected for research and machine learning methods are shown. In the sixth section, the findings obtained and the comparison of their results are carried out.

Literature Review

Numerous studies have been conducted in Turkiye and around the world on the future sales of motor vehicles. While mathematical or statistical methods have often been used for this purpose in the past, nowadays machine learning methods, artificial neural networks, or meta-heuristic algorithms are preferred, which provide faster and more reliable results.

Karaatlı et al. (2012), created a prediction of car sales using artificial neural network methods. Monthly data between January-2007 and June-2011 were used in the study. Gross domestic product, real sector confidence index, investment spending, consumer spending, consumer confidence index, dollar exchange rate, and time were used as independent data. Since the MAPE value in the study was 16.82%, the estimate made falls into the class of "correct estimates".

Sharma and Sinha (2012), used fuzzy neural back sales of cars of one brand in India. They tried to predict the results using a propagation algorithm and compared the results with multiple regression algorithm results.

Hulsmann et al. (2012), conducted a study to evaluate German and US car market prediction models. In their study, they argue that decision trees should be used as the most accurate and explainable method.

Kuvvetli et al. (2015), aimed to estimate monthly vehicle sales for different segments and brands considering economic and environmental parameters. The Levenberg-Marquadt algorithm was chosen to train the model. The results were compared with the linear regression results.

Topal (2019), attempted to estimate the sales volume of a specific car brand using online consumer integration and search engine data. The data was analyzed using artificial neural networks and the Bayesian backpropagation method. The correlation found was 74%, which is above the acceptable value.

Kaya and Yıldırım (2020), proposed an 8-layer Deep Neural Network (DSA) model for predicting automobile sales. The inputs of the model were composed of various economic indicators such as exchange rate, gross domestic product, consumer confidence index, and consumer price index. Forecasts for vehicle sales were made based on the model's outputs. Between 2011 and 2018, a total of 90 data were collected and analyzed on a monthly basis.

Civelek (2021), used artificial neural networks to predict the sales of tractors.

Forecasting Models

Machine learning, a subfield of artificial intelligence, is a discipline that aims to make the best decision by learning from available information or experience.

The more important it is to find the data that most affects the dependent variable in prediction studies, the more important it is to find the most appropriate algorithm for that data. The optimal algorithm shows the most successful performance.

Multiple regression is a statistical tool used to predict the outcome that depends on several other independent predictors or variables. It combines several factors to find out how and to what extent they affect a particular outcome (Lin & Wu, 1999).

Using regression to establish a relationship between the dependent variable and many independent variables is a suitable method for prediction. However, the selection of independent variables is a crucial first step to obtaining accurate predictions (Shahabuddin, 2009). Ridge, lasso, and elastic net regression methods can also be good options in such cases (Bulut, 2018).

In our study, the machine learning methods that are best suited for our preferred multivariate structure are regression algorithms. Regression algorithms are controlled algorithms used to find out how much independent variables affect the dependent variable and to find out the possible relationships between these variables.

Multiple Linear Regression

Multiple linear regression describes the linear relationship between two or more independent variables and a dependent variable. In this regression, there is a relationship between dependent and independent variables. Eq. 1 is used for linear regression. The independent variables are denoted by x and the dependent variables are denoted by Y.

$$Y = \beta_0 + \beta_1 x_1 + \epsilon \qquad \textit{Eq. 1}$$

Y: Dependent variable observation vector.

x: The independent variable observation matrix.

β: The vector of coefficients.

ϵ: The random error vector.

Representation of Eq. 1 by matrix:

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \dots & x_1 \\ \dots & \dots & \dots \\ 1 & \dots & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{bmatrix} \qquad \textit{Eq. 2}$$

In this case β:

$$\beta = (X'X)^{-1}X'Y \qquad \textit{Eq. 3}$$

Classical regression analysis is performed with the above Eq. 3.

Decision Tree

Decision Trees have a form that can be configured for both classification and regression.

If our data are numerical, the regression structure is used, if categorical, the classification structure is used. By decoding the data we have, calculations based on the variables are made and a rule tree is created by establishing a relationship between these variables.

This structure is used to divide a large amount of data into very small groups.

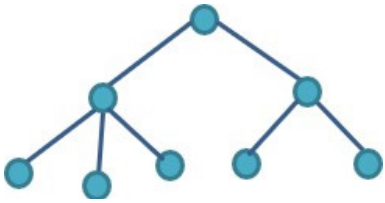

Figure 2. *Simple Decision Tree.*

Random Forest

Random Forest is a supervised learning algorithm. As the name implies, it creates multiple decision trees, translates them into a forest, and combines them to obtain a more stable prediction.

Random Forest adds additional randomness to the model as the trees grow. Instead of looking for the most important feature when decoupling a node, it looks for the best feature from a random subset of features. This creates a large variety that often leads to a better model (Devhunter, 2022).

Random Forest has two main processing parameters. These parameters are the number of randomly selected estimators at each node (m), and the number of trees in the forest (J). In classification, the default value of m is m=√p, where p denotes the total number of predictors (Cutler et al., 2012).

Ridge

Ridge regression has a similar structure to the least squares method. It creates a model with all variables in the data set but provides a solution by bringing the coefficients of the unrelated variables closer to zero.

The Ridge estimator formula is as shown in Eq. 4:

$$\hat{\beta}(ridge) = \arg\min||y - x\beta||^2 + \gamma\,||\beta||^2 \qquad \textit{Eq. 4}$$

As can be seen in Eq. 4, ridge regression applies a quadratic correction in addition to classical regression. Here $\gamma \geq 0$ is expressed as a penalty correction or complexity coefficient. At the same time, the magnitude of this value means that the correction will also be large.

Lasso

Lasso regression attempts to minimize the error by using the least squares method, as Ridge does. Unlike Ridge, however, it equates the coefficients of unrelated variables to zero. Thus, it has the great property of excluding irrelevant variables.

$$\hat{\beta}(lasso) = \arg\min||y - x\beta||^2 + \gamma\,||\beta|| \qquad \textit{Eq. 5}$$

In this algorithm, the correction is made according to the absolute value. Therefore, it is important that the margin of error determined by the least squares method is kept as small as possible.

Elastic-Net

Elastic-Net is a middle ground between Ridge and Lasso regression. Elastic-Net performs a punishment operation like Ridge regression and makes a variable selection like Lasso regression. Penalization is in the style of Ridge regression and variable selection is in the style of Lasso regression.

$$\beta(elasticnet) = (argmin\big||y - x\beta|\big|^2 + \gamma_2\big||\beta|\big|^2 + \gamma_1\big||\beta|\big|) \qquad \textit{Eq. 6}$$

When $\gamma=0$, Elastic-Net corresponds to Ridge regression and $\gamma=1$ corresponds to Lasso regression.

AdaBoost

In the AdaBoost algorithm, the training process continues by increasing the relative weight of the training data belonging to the incorrectly estimated data in the first regression while the next regression operation is performed. The regression process continues until the weights are updated and the stop condition is created (Freund & Schapire, 1997).

Method

Dataset

The study attempts to estimate the level of automobile sales in Turkiye using machine learning methods. The automobile sales data used in this study include all domestic and imported automobile sales. Some commonly used independent variables that directly or indirectly affect automobile sales are: Real Sector Confidence Index, Capital Expenditure, Consumer Confidence Index, and Oil Barrel (Karaatlı et al., 2012), GDP, Dollar Rate (Lin & Wu, 1999), Household Consumption Expenditure, Inflation Rate, Percentage of One-Year Time Deposits (Alper & Mumcu, 2000), and Gold Price. The sources from which these data are obtained are listed in Table 1.

Table 1. *Parameters measured in the experiments.*

Variable	**Source**
Car sales units	ODD
GDP	World bank
Oil barrel prices	Investing.com
Real sector confidence index, dollar price, percentage of 1 year term deposits	TCMB
CPI	Legal bank
Investment expenditures	Department of strategy and budget
Household consumption expenditures, Consumer confidence index,	TUIK
Gold Prices	Altinpiyasa.com

In total, we have 10 independent variables. To determine which of these variables is the one that most affects the level of auto sales, the least squares method is used.

The results of this method are shown in Figure 3. According to these results, the variables GDP, Real sector confidence, CPI and Maturity rate with a large p-value (significance) are excluded from the model. This is because the smaller the p-value, the more power these independent variables have to explain the dependent variable.

In Figure 3, you can see the p-value values of the variables.

	coef	std err	t	P>|t|
GDP	24.8794	267.360	0.093	0.928
Oil Barrel	-2844.3759	1203.190	-2.364	0.038
Real Sector Confidence	-3878.3620	5272.276	-0.736	0.477
CPI	1353.4736	4053.247	0.334	0.745
Investment Expenditures	0.0067	0.002	3.139	0.009
Household Consumption Expenditure	-56.6254	60.220	-0.940	0.367
Consumer Confidence Index	1.71e+04	5589.799	3.059	0.011
USD	-2.852e+05	8.14e+04	-3.504	0.005
Maturity Rate	-5381.1566	5904.687	-0.911	0.382
Gold	2743.8419	740.760	3.704	0.003
const	-2.528e+05	4.16e+05	-0.608	0.556

Figure 3. *p-values of independent variables.*

After the variables to be used in the model are determined, the data are preprocessed and the average of the series is taken and it is assigned to the lost data.

In the next section, the preferred methods for comparing the performance of the algorithms in the study will be described.

Model Performance Measures

Performance evaluation criteria R^2, MSE, and MAE are preferred to compare predictive modeling studies.

The R^2 value determines how well the data will fit into the regression model. MSE gives you an absolute number of how much your predicted results differ from the actual number. Finally, MAE sums the absolute error value, a more direct representation of the sum of the error terms.

The formulas of these criteria are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - \sigma)^2} \qquad \textit{Eq. 7}$$

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(o_i - p_i)^2 \qquad \textit{Eq. 8}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|o_i - p_i| \qquad \textit{Eq. 9}$$

In the above formulas, n is the number of samples, o_i is the true value of the observation, p_i is the estimated value of the observation, and σ is the average of the true observation values.

Findings

For Random Forest and Decision Tree Algorithms, 75% training and 25% test data are selected. For all other algorithms, 80% training and 20% test data are selected. In this way, the

best R^2 values are obtained. When data is divided in this way, Random Forest and Decision Tree regression algorithms are tested for 6 years and the others are tested for 5 years. In addition, since Ridge, Lasso and Elastic-Net regression algorithms give a better R^2value when the random_state value is determined and constant years are tested, the graph of these three algorithms is created together and it is observed that the prediction results are very close to each other.

Forecasting with Linear Regression Model

Figure 4 shows the actual sales values and the forecasted values that the algorithm has given us.

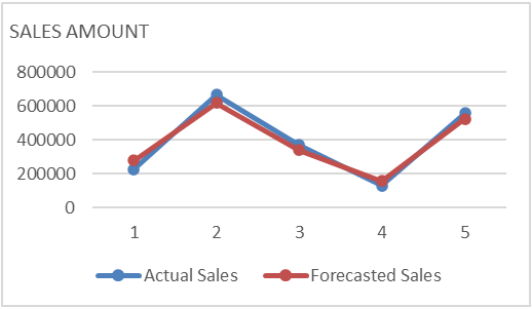

Figure 4. *Forecasted and actual sales by Linear Regression.*

In the Linear Regression algorithm, the performance criterion values are realized as follows:

$R^2_{linear} = 96\%$

$MSE_{linear} = 0.001794$

$MAE_{linear} = 0.02651$

Forecasting with Decision Tree Regression Model

Figure 5 shows the actual sales values and the predicted values that the algorithm gave us.

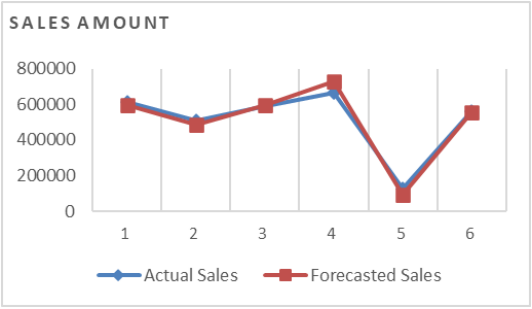

Figure 5. *Forecasted and actual sales by Decision Tree regression.*

The values of the performance criteria in the Decision Tree regression algorithm are realized as follows:

$R^2tree = 97\%$

$MSE_{tree} = 0.004795$

$MAE_{tree} = 0.052606$

Forecasting with Random Forest Regression Model

Figure 6 shows the actual sales values and the predicted values that the algorithm gave us.

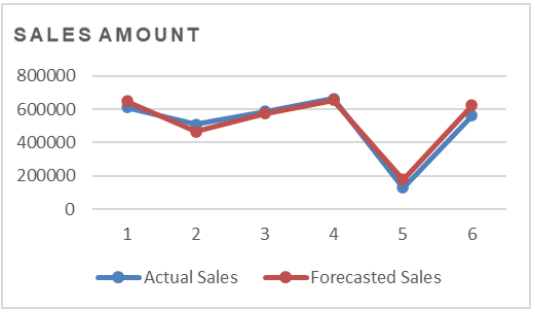

Figure 6. *Forecasted and actual sales by random forest regression.*

The values of the performance criteria in the Random Forest Regression Algorithm are realized as follows:

$R^2_{rf} = 94.67\%$

$MSE_{rf} = 0.004058$

$MAE_{rf} = 0.044487$

Forecasting with AdaBoost Regression Model

This algorithm is found to give better results when n_estimators=10 is chosen. Figure 7 shows the actual sales values and the predicted values that the algorithm gave us.

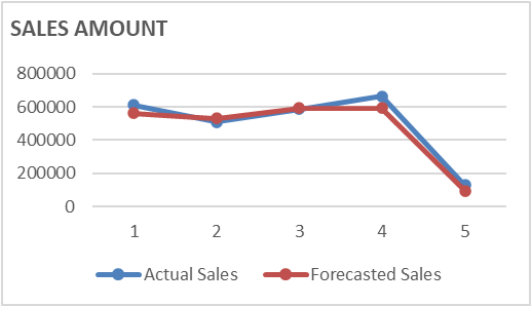

Figure 7. *Forecasted and actual sales by AdaBoost regression.*

The values of the performance criteria in the AdaBoost Regression Algorithm are realized as follows:

$R^2_{adaboost} = 95.14\%$

$MSE_{adaboost} = 0.009622$

$MAE_{adaboost} = 0.06866$

Forecasting with Lasso, Ridge and Elastic-Net Regression Models

Figure 8 shows the actual values of car sales and the estimated values obtained from the Ridge, Lasso, and Elastic Net algorithm models.

In the Lasso algorithm, it has also been observed that it gives better results when the alpha value in the algorithm is set to 0.05. In the Lasso regression algorithm, the values of the performance criteria are realized as follows:

$R^2_{lasso} = 90.04\%$

$MSE_{lasso} = 0.01452$

$MAE_{lasso} = 0.089073$

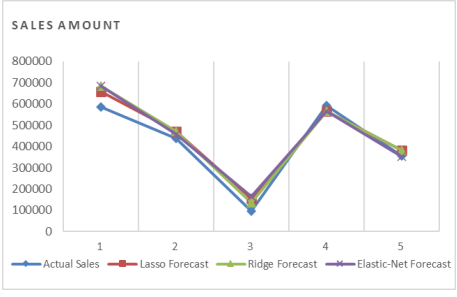

Figure 8. Forecasted and actual sales by Lasso, Ridge, and Elastic-Net regression models.

The Ridge algorithm was also found to give better results when the alpha value was chosen to be 0.01. The values of the performance criteria in the Ridge regression algorithm are realized as follows:

$R^2_{ridge} = 94.78\%$

$MSE_{ridge} = 0.014396$

$MAE_{ridge} = 0.088564$

The Elastic-Net algorithm was also found to perform better when the alpha value is 0.01, l1_ratio=0.5, and normalized=False. The values of the performance criteria in the Elastic-Net regression algorithm are realized as follows:

$R^2_{elastic} = 94.85\%$

$MSE_{elastic} = 0.014414$

$MAE_{elastic} = 0.088639$

After the prediction studies of the models, the performances of the algorithms are compared using the values of R^2, MSE, and MAE. The summaries of these results can be found in Table 2.

Table 2. *Performance measurement values of algorithms.*

Algorithms	**R^2**	**MSE**	**MAE**
Decision Tree	97	0.004795	0.052606
Linear	96	0.001794	0.02651
AdaBoost	95.14	0.009622	0.06866
Elastic-Net	94.85	0.014414	0.088639
Ridge	94.78	0.014396	0.088564
Random Forest	94.67	0.004058	0.044487
Lasso	90.04	0.01452	0.102048

Figure 9 shows the R^2 values of the results of the algorithms.

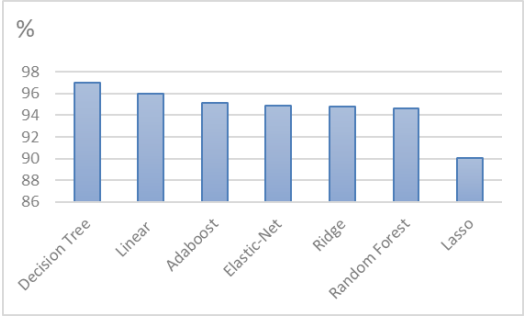

Figure 9. *R^2 values of algorithms.*

Conclusion

For this study, annual automobile sales data for 2000-2021 in Turkiye are taken from the ODD (Automotive Distributors Association) website. In order to perform prediction studies, modeling is performed using regression algorithms from machine learning methods and predictions are made. In performing this modeling, 10 different independent variables are used that most influence car sales. The variables are preprocessed and a value is assigned to the missing data.

Of these data, 4 variables that least affect the dependent variable using the least squares method are excluded from the model.

R^2 is a value that indicates how well the predicted values obtained explain the actual values. In other words, the obtained observed values are around or away from the regression line. The distance to this line is best determined by the value R^2. An R^2 value of 90% or more indicates that the predictive model gives a good result. Figure 9 shows the R^2 values of the regression algorithms whose performance is from best to worst.

Accordingly, the Decision Tree algorithm performed best with R^2 values of about 97% in the car sales model estimation study. Moreover, the MSE value of this algorithm is 0.004795. The worst modeling algorithm is Lasso regression with 90.04%.

There may be many variables that affect car sales. The fact that the vast majority of these variables are used in a study complicates predictive studies. In this study, the use of 10 separate independent variables is preferred. However, the least squares method is used to facilitate modeling and remove unnecessary variables. As a result, four variables with a large p-value are excluded from the study. Estimation studies are conducted with the remaining 6 variables. Despite the large number of independent variables selected, the R^2 values of 90% and higher obtained with all regression algorithms indicate that this study is successful and achieved its objective.

Looking at the automobile sales data, it is seen that the highest sales are made towards the end of the year. There may be different reasons for this. Thinking that prices will increase even more in the new year or entering the winter season may be the reasons that increase these sales. If automobile companies take advantage of this situation and make year-end campaigns, sales will increase substantially.

Ethics Committee Approval: Authors declared that this study does not require ethics committee approval.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- M.B., A.C., T.A.; Data Acquisition- M.B.; Data Analysis/ Interpretation- M.B.; Drafting Manuscript- M.B.; Critical Revision of Manuscript- A.C., T.A.; Final Approval and Accountability- M.B., A.C., T.A.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Alper, C. E., Mumcu, A. (2000). *Türkiye'de otomobil talebinin tahmini*. Research Report, Boğaziçi University. http://ideas.econ.boun.edu.tr/content/wp/ISS_EC_05_01.pdf.

Bulut, H. (2018). *R uygulamaları ile çok değişkenli istatistiksel yöntemler.* Nobel Academy.

Civelek, Ç. (2021). Tractor Sales Forcasting for Turkey Using Artificial Neural Network. *European Journal of Science and Technology ,31* (1), 375-381.

Cutler, A., Cutler, D. R. And Stevens, J. R. (2012), Random forests*, Ensemble Machine Learning Methods and Applications*,*45*(1), 157-176.

Devhunter. (2022, May 05), *Rastgele Orman (Random Forest) Algoritması,* Https://Devhunteryz.Wordpress.Com/2018/09/20/Rastgele-Ormanrandom-Forest- Algoritmasi/Comment-Page-1/

Dikmen, I. (2006). Otomotiv Sektörü Ve Rekabet Değerlendirme. Access Date: 10.12.2011, Http://Www.Kalder.Org.Tr/Genel/15kongre/Sunumlar/Isik_Dikmen.Doc.

Euronews. (2022, June 04). *Türkiye'de 100 kişiye düşen otomobil sayısı 14, AB'de ise 51,* Https://Tr.Euronews.Com/2019/12/30/Turkiye-De-100-Kisiye-Dusen-Arac-Say-S-28-Ab- De-Ise-51.

Freund, Y., Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139.

Görener A., Görener Ö. (2008). The contributions of automotive industry in Turkish economy and sectoral-expectances about the future. *Journal Of Yaşar University*, *7*(26), 306-319.

Hulsman, M., Borscheid, D., Friedrich, C.M., Reith, D. (2012). General sales forecast models for automobile markets and their analysis. *Transactions On Machine Learning and Data Mining*, *5*(2), 65-86.

Karaatlı, M., Helvacıoğlu, Ö., Ömürbek, N., Tokgöz, G. (2012). An artificial neural network based automobile sales forecasting . *International Journal of Management Economics and Business*, *8*(17), 87-100.

Kaya, K.S., Yildirim, Ö. (2020). A prediction model for automobile sales in Turkey using deep neural networks. *Journal Of Industrial Engineering*, *31*(1), 57-74.

Kuvvetli, Y., Dağsuyu, C., Oturakçı, M. (2015). A prediction approach based on artificial neural networks with consideration of environmental and economic indicators for car sales in Turkey. *Journal of Industrial Engineering*, *26*(3), 23-31.

Lin, Z.C., Wu, W.J. (1999). Multiple linearregression analysis of the overlay accuracy model zones. *IEEE Trans. On Semiconductor Manufacturing*, *12*(2), 229 – 237.

Mmpvizyon 2023 Otomotiv Sektör Raporu," Https://Tubitak.Gov.Tr/Tubitak_Content_Files/Vizyon2023/Mm/Ek3.Pdf, 22.05.2022.

Shahabuddin S. (2009). Forecasting automobile sales, *Management Research News*, 32(7), 670-682.

Sharma, R., Sinha, A.K. (2012). Sales forecast of an automobile industry. *International Journal of Computer Applications*, *53*(12), 25-28.

Topal, İ. (2019). Çevrimiçi tüketici bütünleşmesi ve arama motoru verileri kullanılarak yapay sinir ağları ile otomobil satış tahmini, *Nevşehir Hacı Bektaş Veli Üniversity Journal of ISS*, *9*(2), 534-551.

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

The Ethical Dimension of Artificial Intelligence

Gökçe KARAHAN ADALI[1]

ABSTRACT

Artificial Intelligence (AI) is based on technologies that have the potential to dramatically enhance productivity and ,-facilitate decision processes for decision makers. AI influences almost every aspect of our lives. However, when this potential for AI is accepted, not only technical but also ethical consequences are inevitable. Technology always brings benefits and risks. AI has the same benefits and risks. An innovative and significant aspect of AI is machine ethics. The creation of a machine that abides by the standards specified in a perfect ethical framework is the ultimate objective of machine ethics. These concepts serve as a guide for the potential course of action. The spread of artificial intelligence applications, which have been used in many areas, has caused ethical dilemmas and some concerns. Studies have have started to keep the behavior of intelligent machines under control and to put them intoto a certain framework. This article aims to provide an overview of artificial intelligence, a high-level conceptual treatment of the subject by introducing fundamental ideas, and to outline strategies, and key themes in AI ethics. This study provides an insight into the critical debate on the ethical use of AI. Some key ethical issues identified in this paper include the legal identity of autonomous machines, the legal problems arising from AI, and the solution proposals with application examples.

Keywords: Artificial Intelligence, Artificial Intelligence Ethics, Machine Ethics, Bias

DOI: 10.26650/JODA.1253475

[1]Halic University, Istanbul, Turkiye

ORCID: G.K.A. 0000-0001-8567-4626

Corresponding author:
Gökçe KARAHAN ADALI,
Halic University, Istanbul, Turkiye
E-mail: gokceadali@halic.edu.tr

Submitted: 20.02.2023
Revision Requested: 21.03.2023
Last Revision Received: 31.03.2023
Accepted: 02.04.2023

Citation: Karahan Adali, G. (2023). The ethical dimension of artificial intelligence. *Journal of Data Applications, 1*, 35-48. https://doi.org/10.26650/JODA.1253475

Introduction

The interaction of people with society has continued its existence by taking on different silhouettes since the existence of humanity. Machine learning and automation are deeply integrated into all aspects of our lives and work. Through the development of technology, it has become inevitable to come across artificial intelligence (AI) and its applications, which have taken their place in almost every aspect of our daily flow to make our lives easier. It takes place in our daily lives among voice assistants, language translations, suggestion systems, navigation applications that take you to your destination as fast as possible, social security systems that are authorized to allow login by an identified ID, voice, or face-id. There are also purchase forecasters with algorithms that recommend products to buy and assistant robot applications in customer services that answer your calls. It has been observed that these applications are widely used in many areas that you can think of, from smart home systems to driverless vehicles.

The potential of AI to improve itself by learning only through the examples presented to them and the successes achieved in this direction have led to concerns about whether intelligent systems will begin to act against human control over time. For this reason, studies have been started in this area to keep the behavior of intelligent machines under control and to put them into a certain framework.

The main aim of this article is to provide a high-level conceptual treatment of the subject by introducing fundamental ideas, outlining strategies, and key themes in AI ethics. This study tries to provide insight into the critical debate on the ethical use of AI. The legal identity of autonomous machines, the legal problems arising from artificial intelligence, the solution proposals and application examples in this field, can be counted among these titles.

Artificial Intelligence

The concept of AI was first introduced to computer science by John McCarthy in 1955 at a conference held at Dartmouth College in New Hampshire. John McCarthy, Martin Minsky (founder of the MIT AI Laboratory), Claude Shannon (IBM), Allen Newell (the first president of the USA Artificial Intelligence Association), and Herbert Simon (a Nobel Prize-winning economist), attended this meeting and , suggested investigating the possibility of realizing the development of intelligent computers. With the proposal in question, the work to be carried out in artificial intelligence gained momentum.

AI is described as the intelligent actions of a machine, such as reasoning, learning, communicating, etc. It is one of the most contentious topics of today. The basic goal of artificial intelligence research is to create a machine that is sentient, capable of thought, and morally

compatible in a manner akin to humans. In this regard, it is vital to incorporate artificial ethics to prevent harm to both humans and other living things if artificial intelligence systems have the capacity to think and become conscious.

Since the use of artificial intelligence applications has spread to various fields over time, artificial intelligence has taken on an umbrella identity that includes sub-disciplines in terms of structure. The sub-disciplines include machine learning and data mining, robotics, expert systems, fuzzy logic, natural language processing, machine vision and optimization (Özen, 2021). On the other side, the study of good and evil, as well as how they relate to morality and human behavior, is called ethics. A unique concept in abstract terms, ethics is an idea, structure, or model of thought and conduct having a flexible scope and substance. The explanation is that morality, good and evil, and models of human behavior are all notions that change across time and space rather than being fixed, inflexible, or static (Robles Carillo, 2020).

Artificial intelligence ethics is a field that has emerged as a response to the growing concern regarding the impact of AI. Today, artificial intelligence has been accepted as one of the most important research and areas of interest that shape the future of humanity. Just like every rapid technological development, this unstoppable and rapid rise of artificial intelligence has also led to various concerns and problems for human beings, who often have problems in keeping up with technological changes. The potential of artificial intelligence to improve itself by learning only from the examples presented to them and the successes achieved in this direction have caused concerns about whether intelligent systems will start to act against human control over time (Köse, 2018).

Therefore, to control the behavior of intelligent machines and to maintain the ethical dimension of artificial intelligence, studies have been started in this area. The legal identity of autonomous machines, legal problems arising from artificial intelligence and solution proposals in this field and practice proposals can be presented among these titles. Human beings have been in interaction with their environment throughout history. More precisely, man has tried to become the subject of the world he was born into, to change and transform it in line with his own wishes, desires and often ambitions. This effort has sometimes been in the form of taking steps that may be beneficial for humanity, and sometimes it has led to the disaster of humanity. However, in any case, humanity adds something new to its own history with every step it takes.

Although every invention made and every new step taken causes pain from time to time, in general terms and in the integrity of time, it has facilitated the development of humanity, the shaping of social structures, the establishment of systems, and the civilization of people (Çelebi, 2018). AI has come to the stage of being a decision maker on its own, rather than being in interaction with human beings. At this point, even the decisions made by human

beings are questioned in an ethical framework, while the decisions made or to be made by a machine open a window to brand new and never-ending discussions.

Literature Review

According to John Mc McCarthy who is considered as one of the founding fathers of the AI discipline, expressed AI as "science and engineering of making intelligent machines. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable" (McCarthy, 2007). Minsky defined AI as "the science of making machines do things that would require intelligence if done by men. It requires high-level mental processes such as perceptual learning, memory, and critical thinking".

The categorization for classification of artificial intelligence was carried out by Kaplan and Haenlein according to three different working systems of artificial intelligence. These are "analytical", "human-inspired" and "humanoid" artificial intelligence (Kaplan & Haenlein, 2019). In this context, "analytical artificial intelligence" is a type of artificial intelligence that constantly updates its future data autonomously and can use this existing data flexibly by making algorithmic comments. The working principle of this type of artificial intelligence is also called "machine learning".

On the other hand, human-inspired artificial intelligence, includes elements of emotional intelligence (EQ) as well as cognitive elements (IQ) such as algorithms in decision-making processes; at least, it can use the detected external emotional data in its autonomous decisions. This technology is also called "deep learning" in the literature (Akkurt, 2019). "Humanoid artificial intelligence" is a type of artificial intelligence". For example, artificial intelligence used in weather forecasts and games such as chess and "Go" can compete with humans at a high level and is evaluated in this category. Hence it can interact autonomously with other fellows or humans, as well as having cognitive and emotional functions, and in this respect, it also displays social intelligence features. Self-driving cars, virtual assistants, military, medical, industrial robots/bots, which can act together like a social community (interactive) can be included in this category.

A study of the global AI ethics landscape concluded that at least five principle-based approaches of an ethical nature; transparency, justice and fairness, non-harmful use, responsibility, and integrity/data protection are broadly accepted as having ethical value. The study also shows that there are significant discrepancies in how these principles are understood, why they are valued, what problems, areas, or actors they apply to, as well as how they ought to be put into practice. The most prevalent principle is "transparency," which appears to have several different meanings (Jobin et al., 2019).

Larsson analyzes the unique characteristics of AI development that have caused ethical issues to take on such a prominent role. It also emphasizes the interaction with legal mechanisms for governance (Larsson, 2020). According to a study, the moral use of AI has so many tensions that cannot be easily resolved. They suggest there are some advantages to state them more clearly. Principles must be codified in standards, codes, and finally regulations to be useful in practice (Whittlestone et al., 2019).

In the study of Vakkuri et al., (2019) the keywords' classified in the field of AI to guide and provide directions to future studies. It seems to be three main variations: Even if keywords may be categorized under the well-known themes, there is a lack of various branches of AI in keywords, technology-based keywords play a minimal role, and there is a great deal of variation in how keywords are phrased. Besides, no main papers in AI ethics were acknowledged because the focus was solely on the keywords.

According to Owe and Baum, the field of artificial AI ethics should pay more attention to the values and interests of nonhumans, including other biological species, the natural environment and the AI itself. They deserve moral consideration, which means that they should be actively valued for their own sake (Owe & Baum, 2021).

Huang et al. categorise ethical theories into three branches. Meta ethics; investigate nature, normative ethics; establish moral principles and guidelines that define what is right and wrong conduct. The third branch, applied ethics, is interested in application fields like animal rights, war, etc. Human beings never apply just one ethical theory, but instead alternate between them depending on the circumstance or context. It is confirmed by the study as strict adherence to any moral doctrine can have negative effects (Hang et al., 2022).

In Kumar et al. (2023), the authors emphasize the difficulties in integrating AI on a large scale into medical systems. They give an overview of the ethical, legal, trust, and long-term consequences of AI in healthcare. They highlight that we have a responsibility to carefully consider ethical values to ensure appropriate responses and the implications of applying AI.

Adams et al. argue the ethical principles currently guiding AI ethics policy development for children and K-12 education. Artificial Intelligence in K-12 Education ethics policy must be responsive not only to the diverse abilities, cultural backgrounds, and developmental requirements of children, but also to what they are becoming as human-AIs or post-humans (Adams et al., 2023)

The Ethical Dimension of Artificial Intelligence

Robots and smart software have an increasing impact on our daily lives, and the machines that are created by us will make decisions that can have profound effects on our lives. Some

of these decisions have a moral dimension. Therefore, we need to consider whether we want them to make such decisions and, if so, how to endow machines with "moral sensibility" or "moral decision-making abilities".

We are afraid of falling behind the artificial intelligence systems that we produce with our own natural intelligence. The question that comes to mind is, as Kaku stated, although scientists produce artificial intelligence systems, are these artificial intelligence systems as natural as the biologically existing human brain? It focuses on whether it will be successful in being thoughtful and experienced (Kaku, 2016). Therefore, the focus of the studies produced in the field of artificial intelligence is to bring the artificial intelligence systems working through programming to a level that will imitate the human brain by making it like the human brain.

The human brain has the most complex structure known in the universe, consisting of billions of neurons (Taslaman, 2015). The most striking characteristic of such a complex structure is the ability to offer creative solutions based on one's own and someone else's experiences, as they have learned the non-programmable/non-imitable feature. Compared to the computer with artificial intelligence, the human brain with natural intelligence is a very complex and superior structure of the human brain. We can think of it as having a system, and it creates itself without a program. The human brain can produce a creative, fast, and new solution in the face of a new or unexpected situation (Adalı, 2017). On the other hand, the solutions that artificial intelligence can produce are limited to the solutions taught to it.

Although it is said to be limited, just like every rapid technological development, the unstoppable and rapid rise of artificial intelligence has led human beings, who often have problems keeping up with technological changes, to having various concerns and problems. The potential for self-development by learning from the examples presented and the successes achieved in this direction cause concerns about whether intelligent systems will begin to act against human control over time (Köse, 2018). Some groups, including Anderson & Anderson and Kaku, think that it would be useful to gather the ethical rules that are thought to be applied to robots under the name of machine ethics. The goal of machine ethics is to create a machine that follows an ideal ethical principle or set of principles.

To introduce the subject of machine ethics, it is necessary to mention another point. Ethics can be seen as both easy and difficult (Vakkuri et al., 2019). The reason it seems easy is that we all make generally accepted decisions based on social norms under the umbrella of ethics in our daily routine (for example, exceeding the speed limit while trying to get a patient to the hospital). On the other side, let's consider that there are many patients who need urgent intervention at the same time in the context of work and profession, and we can use an artificial intelligence-based algorithm that intervenes in this process.

In the functioning of the decision mechanism in the healthcare system, when the decision maker decides which patient should intervene, he will enter into an ethical paradox. When you treat person A, person B will left be out, and when you treat person B, person A will be left out. Therefore, even the person her/himself can be inside an ethical paradox at this decision stage. Or, let's imagine a person who helps the poor by committing a crime, and while an artificial intelligence-based judge robot should judge the person in question, this situation contains ethical contradictions (Köse, 2018).

Therefore, it is necessary to integrate ethically correct and critical decisions into artificial intelligence-based robots against these paradoxical situations. According to the World Health Organization (WHO), health systems, and public health organizations should regularly publish information about the decision-making process that has gone into adopting AI technology, how it will be evaluated going forward, as well as its applications, known drawbacks, and the decision -making role (WHO, 2021). However, this does not mean that we are all experts in ethics. Ethics is a field that requires a lot of study and experience. Machine ethics is also an interdisciplinary field by nature (Anderson & Anderson, 2007).

There are three different approaches in the literature to teaching ethical values that can be applied to artificial intelligence. The first one is the determination of rules containing moral values or the teaching of normative moral rules that can be taken from traditional moral philosophy to the machine through algorithms. These normative rules should be chosen in such a way that there is no space for uncertainty and controversy for the behaviour of the robot. Utilitarian and duty-based morality can be used as examples of traditional thought. The other approach is for the machine to grasp right and wrong on its own, without any externally dictated set of rules.

There are examples of this approach, such as using evolutionary algorithms, such as genetic algorithms, or adapting game theory for moral actions. There is also a third approach in the literature that proposes the use of combining these two approaches. In this approach, the machine should first start with a set of rules and use it by changing over time (Wallach & Allen, 2009). Artificial intelligence-based algorithms developed to date are not considered a threat if they are under human control. However, recently, there are those who think that it has started to pose a threat to humanity, and there are differences of opinion because of dilemmas over this issue.

There are different opinions by the centreres that can be considered as authorities on how the use of artificial intelligence should be carried out within the framework of the proposed laws and the ethical rules. To list some of them: Joseph Weizenbaum, the creator of the ELIZA[1]

[1] ELIZA is a program which makes natural language conversation with a computer possible (Weizenbaum, 1996).

program, is seen as an advocate of the restrictive use of artificial intelligence. He claimed that some tasks, such as nursing or adjudicating, should never be done by artificial intelligence, and that these professions need compassion and intuition, and according to him that machines could not achieve these qualities.

Weizenbaum argued that the human brain is much more than a brain and that a simple copying of the brain could never achieve realistic human behavior. Pamela McCorduck, an American author interested in the history and philosophical significance of artificial intelligence, stated that she would take a chance with an impartial computer, unlike Weizenbaums, on why computers should not be judges. A properly programmed computer should be neutral towards minority issues such as gender, that people may be prejudiced against. Ben Schneiderman argues that how to assign responsibility to autonomous systems should be followed closely and repeated with plenty of examples of how to do it. The underlying logic is learning by feeding from the training set of autonomous systems.

It proposes a step-by-step approach, in which people initially monitor the system intensively and as they gain confidence in the system, the monitoring is reduced (Schulze, 2012). The famous trolley problem, created by the moral philosopher Philippa Foot, would be the best example of this situation. The problem is a collection of ethical and psychological thought experiments that simulate ethical decisions on whether to sacrifice one person to save more people. A runaway train or trolley is on course to hit and kill several people (often five) down the track, but a driver or bystander can intervene and reroute the vehicle to kill just one person on a different track. This is how most episodes of the series start. Then, different iterations of the runaway car are presented, along with comparable life-and-death situations (medical, legal, etc.). In each case, there is a choice between doing nothing, in which case several people will perish, and intervening and sacrificing one initially "safe" person to save the others. Most people prefer to save the others. This morality is a product of "teleological ethics". What is this teleological ethics? Moral theories that argue what determines the value of moral action is the result produced by the action. However, teleological ethics, including the phenomenon of social utilitarianism, is based on the basis that the criterion of goodness is to make the maximum number of people happy. Rather than watching four people die, it is preferable to intervene in the system and let another innocent person die. The reason for this is that the basic decision unit of our choice is utility. Rather than saving one life, people are in favour of providing maximum benefit by saving four lives.

At this point, the question that comes to mind is: Will robots behave like humans, and most thinkers answer that they cannot think like humans in instant and intuitive decisions. When we look at the trolley example, if the robot thinks like a human, a situation arises that contradicts Asimov's three robot laws. In contrast to this law, which is based on the principle that a robot

can never harm a human, the robot has harmed a human being, and with the decision it has made, it has changed the fate of an innocent human being who will not die.

It is possible to encounter similar studies on the "Moral Machine" Platform developed by the Massachusetts Institute of Technology (MIT). Visitors to the "Moral Machine" website are asked to decide, through an autonomous system, what to do when faced with certain scenarios (moralmachine.net, 2022). An online test leaves you with some moral dilemmas over driving behaviour. It compares your decisions against the answers given by others and is also accumulated to train autonomous machines.

For example, if a driverless car is forced to harm pedestrians, it would sacrifice two children to save three adults. Could it ? Could it sacrifice an old man to save a pregnant woman? As AI development progresses, experts have managed to figure out the best way to give an AI system an ethical or moral backbone. The basic idea aims to teach artificial intelligence behaviours because of the decisions taken by the average person. In these and similar cases, when the decision mechanism is human, it does not cause much speculation, but when a robot gets involved, it drags different questions and situations behind it. Based on the experiences of people, we need to integrate artificial ethics into artificial intelligence to minimise the damage when faced with this and similar critical situations.

There are participants from more than 200 nations, and the website gathered nearly 40 billion decisions. The findings show that visitors from various nations or areas frequently have diverse moral standards. Three geographical groups with distinct decision-making inclinations can be used to categorise countries. For instance, Eastern nations had the lowest propensity to preserve the young at the expense of the elderly, Southern nations preferred to save women, while Western nations had the highest propensity for inaction. Also there were observable values divergences between two nations related to variations in decision-making. People from nations with more economic disparity, for instance, were more inclined to protect the wealthy. In contrast to the conventional view of morality as a sharp distinction between right and wrong, this study demonstrates how cultural norms can shift the line (Awad et al., 2018).

Bias in Artificial Intelligence

A controversial ethical and well-observed issue for AI is bias. Bias can be defined statistically and socially. In terms of statistics, bias describes situations where a dataset's distribution does not accurately reflect the true distribution of the population. On the contrary, social bias refers to inequality that may lead to less desirable outcomes for specific groups of the human population (Norori et al., 2021). Statistically, bias refers to cases in which the distribution of a given dataset is not reflecting the true distribution of the population. Social bias, by contrast, refers to inequalities that may result in suboptimal outcomes for given groups of the human population.

According to Kartal's study, the numerous types of bias that appear in AI are categorized in various ways in the literature like systematic bias, human based bias, algorithmic bias. When we examine factors that lead AI systems to make biased decisions; data collecting, data set construction, and data preprocessing are the three types of them. (Kartal, 2002). These systems are built of data and the systems are influenced by the data they are fed (Carter et al., 2020). For example, a hiring algorithm, due to the high levels of predictive accuracy, an AI system can predict the likelihood of depression by simply analyzing the candidate's social media before symptoms appear.

The system can predict the probability of a potential candidate becoming pregnant or select a highly aggressive person to fit a corporate culture. This and similar examples can prove that hidden bias is also an important ethical problem in recruitment. Therefore, transparency is an important issue. The more powerful a predictive system is, the less transparent it becomes (Tüfekci, 2016). Building customer trust in and acceptance of AI-enabled products depends on effectively incorporating ethical concepts into these products and making sure that the user and the product are on the same ethical page (Due & Xie, 2021).

Biases are frequently determined by who funds and develops AI technology. AI-based technologies have historically been created by a single demographic group and gender. Thus, the first versions of the Apple Health Kit, which allowed specialized monitoring of some health risks, lacked a menstrual cycle tracker, possibly due to the lack of women on the development team. (WHO, 2021).

Artificial Intelligence and Responsibility Debates

By the production and use of artificial intelligence for many purposes, the possibility of harmful consequences from interactions with humans becomes inevitable day by day. In this context, artificial intelligence may violate personal values such as life, bodily integrity, health, privacy, personal data, honouur, and dignity because of autonomous decisions and behaviouurs, as well as cause moral and material damage as a result of some ethical and economic mistakes. For example, the fact that artificial intelligence for military purposes, especially robotics, terminates the right to life or harms health and/or bodily integrity, is one of the issues that has been discussed a lot in recent years.

The same risk is valid for the examples of artificial intelligence for automotive purposes as well as for the example of self-driving cars causing traffic accidents due to their wrong decisions. In the case of deciphering personal data by artificial intelligence used in the shopping or banking sector, violation of personal rights and moral/economic (material) damages to a large extent may occur (Akkurt, 2019).

These and similar unfortunate situations have recalled the questions of who will be responsible if you have an autonomous vehicle or an autonomous device with artificial

intelligence that moves with taught algorithms and makes a wrong decision due to a malfunction. We as the owner of the vehicle? The company that programmed the vehicle? Or the artificial intelligence product that commands the vehicle? Of course, many more can be derived from these questions, some of which are highly speculative and exaggerated. These are the law issues that will probably start to matter more and more. An increase in product liability claims is possible in the future. Such regulations are already starting to be passed in some USA states (Stout, 2022). In this instance, the manufacturer may occasionally be considered as the "driver." It is argued that by producing a successful moral algorithm in response to each question, behavioral patterns that can satisfy the majority can be developed.

The Legal Position of Artificial Intelligence

The long-standing development of robots and artificial intelligence continues today. Their increasing use in social life brings with it some legal problems. These problems are linked to various branches of law. For instance, in terms of the effects of robots used in the online environment in electronic commerce, Electronic Trade Law. In case the data excavated by "crawling" is personal data, Personal Data Protection Law. Criminal Law and Liability Law in terms of accidents involving autonomous vehicles; Law of Obligations in terms of automatic trading bots that enter a contractual relationship on the internet; in terms of analysing and influencing voter behaviouur through robots in the online environment.

Personality Recognition for Artificial Intelligence

The legal personality of artificial intelligence is important in terms of determining the liability issue. Whether abstract elements such as artificial intelligence can have abilities such as will, consciousness, logic and thinking are both philosophical and legal questions (Çetin, 2019). There are some opinions and/or suggestions in the doctrine about the legal position of artificial intelligence. These are generally listed as "goods", "slave", "legal person" and "electronic person" approaches. Legal systems, taking into account some realities of daily life, have given personality to some beings other than humans; For example, some individuals or groups of goods have been accepted as people. These structures, called legal people, refer to entities that have the title of person, apart from the real person.

Legal systems have accepted that these assets also have a a legal capacity. According to this view, it may be plausible to model the "company" structure for intelligent software. Since companies have similarly scattered and complex activities, a registry system has been developed to overcome this problem. Autonomous robots can also be registered in such a registry system and even their assets can be introduced; in difficult situations. It is emphasised that the damages can be collected from this pool. It is claimed that this pool of money can be created by the stakeholders who will use the system. The "Electronic Person" approach

for advanced autonomous robots was introduced by the European Parliament's Legal Affairs Commission in the Report with Recommendations to the Commission on Civil Law Rules on Robotics dated January 27, 2017. According to this idea, it is argued that a system includes various parties, such as the user, manufacturer and seller and would be especially beneficial in terms of liability. Similarly, in this approach, it is envisaged to design a system similar to a commercial registry system in which artificial elements such as robots are registered in an official registry. Then they will become a personality as soon as they are registered, and be able to apply for the funds allocated to robots under the responsibility of compensation.

Discussion and Conclusion

Machine ethics is a new and important dimension of AI. The goal of machine ethics is to create a machine that follows the rules established in an ideal ethical framework. These principles guide the possible actions that are taken. AI needs to be a part of human development to help increase human creativity and create a collaborative culture. There is a consensus on social norms or ethical principles that AI should now follow. However, different solutions are increasing day by day. Countries have accelerated their work in this area and started to publish artificial intelligence strategies. However, until ethics becomes a vital part of human behaviour, it is more than just autonomous machines produced with artificial intelligence technologies. It would be unfair to expect them to always remain faithful to ethical behaviour.

The general opinion is that artificial morality should be integrated into machines. For this, the machine will need to be trained with exemplary behaviour. It should be considered not only by ethics committees or research departments of future practice, but also by the government, industry, international institutes and institutions at the initial stage. To create and maintain an artificial intelligence-friendly environment, a culture of responsibility needs to be developed on a global scale (Pavaloiu & Köse, 2017).

Ethics Committee Approval: Authors declared that this study does not require ethics committee approval.

Peer Review: Externally peer-reviewed.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. *Computers and Education: Artificial Intelligence*, *4*(April 2022), 100131. https://doi.org/10.1016/j.caeai.2023.100131

Adalı, E. (2017). Yapay Zekâ, İTÜ Vakfı Dergisi, Ocak-Mart 2017 (75), s.8-13.

Akkurt, S. (2019). Yapay Zekânın Otonom Davranışlarından Kaynaklanan Hukukî Sorumluluk, Uyuşmazlık Mahkemesi Dergisi, 0(13), s.39-59.

Anderson, M. & Anderson S. (2007). Machine Ethics: Creating an Ethical Intelligent Agent, AI Magazine, 28(4), s.15-26.

Carter, S. M., Rogers, W., Win, K. T., Frazer, H., Richards, B., & Houssami, N. (2020). The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast*, *49*, 25–32. https://doi.org/10.1016/j.breast.2019.10.001

Çelebi, Ö. G. (2018). Tarih Öncesi Dönemlerde İletişim, Üsküdar Üniversitesi İletişim Fakültesi Akademik Dergisi, 2, s.142-156.

Çetin, S. (2019). Yapay Zekâ ve Hukuk ile İlgili Güncel Tartışmalar. İstanbul, Ankara ve İzmir Baroları Çalıştay Raporu.

Du, S., & Xie, C. (2021). Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. *Journal of Business Research*, *129*(February 2019), 961–974. https://doi.org/10.1016/j.jbusres.2020.08.024

European Parliament, (2017, January). Report with Recommendations to The Commission on Civil Law Rules on Robotics.http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html.

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, *PP*, 1–21. https://doi.org/10.1109/TAI.2022.3194503

Kaku, M. (2016). Geleceğin Fiziği. (Çev. Yasemin Saraç Oymak & Hüseyin Oymak). Ankara: ODTÜ Geliştirme Vakfı Yayıncılık.

Kaplan, A., Haenlein, M. (2018). Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence, Business Horizons, 32(1), s.15-25.

Kartal, E. (2022). A Comprehensive Study on Bias in Artificial Intelligence Systems. *International Journal of Intelligent Information Technologies*, *18*(1), 1–23. https://doi.org/10.4018/ijiit.309582

Kumar, P., Chauhan, S., & Awasthi, L. K. (2023). Artificial Intelligence in Healthcare: Review, Ethics, Trust Challenges & Future Research Directions. *Engineering Applications of Artificial Intelligence*, *120*(May 2022), 105894. https://doi.org/10.1016/j.engappai.2023.105894

Köse, U. (2018). Güvenli Yapay Zekâ Sistemleri İçin Denetimli Bir Model Geliştirilmesi, Mühendislik Bilimleri ve Tasarım Dergisi, 6(1), s.93-107.

Larsson, S. (2020). On the Governance of Artificial Intelligence through Ethics Guidelines. Asian Journal of Law and Society, 7(3), 437-451. doi:10.1017/als.2020.19

McCarthy J. (2007, November 12). *What Is Artıfıcıal Intellıgence?* Retrieved January 2023, from http://jmc.stanford.edu/articles/whatisai/whatisai.pdf.

Moralmachine.net. (2022). https://www.moralmachine.net/

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

Owe, A., & Baum, S. D. (2021). Moral consideration of nonhumans in the ethics of artificial intelligence. *AI and Ethics*, *1*(4), 517–528. https://doi.org/10.1007/s43681-021-00065-0

Özen, Z. (2021). Güncel Bilişim Teknolojileri. In N. Bozbuğa & S. Gülseçen (Eds.), Tıp Bilişimi (pp. 335–349). Istanbul University Press. https://doi.org/10.26650/B/ET07.2021.003.17.

Pavaloiu, A. & Köse, U. (2017) Ethical Artificial Intelligence - An Open Question, Journal of Multidisciplinary Developments. 2(2), s.15-27.

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, *44*(6), 101937. https://doi.org/10.1016/j.telpol.2020.101937

Schulze, C. (2012). Ethics and AI. University of Maryland. Retrieved December 12, 2022, from https://www.cs.umd.edu/class/ fall2012/cmsc828d/oldreportfiles/schulze1.pdf.

Stout, H. (2022, October 28). What Happens When Self-Driving Cars Crash? The Legal Ramifications of Automation. Entrepreneur. https://www.entrepreneur.com/living/what-happens-when-self-driving-cars-crash-the-rise-of/436942

Taslaman, C. (2015). Modern Bilim, Felsefe ve Tanrı. 10.Basım. İstanbul: İstanbul Yayınevi.

Tüfekci, (2016, June 1). Machine Intelligence makes human morals more important.TED.https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important.

V. Vakkuri and P. Abrahamsson. (2019). "The Key Concepts of Ethics of Artificial Intelligence," *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, Stuttgart, Germany, 2018, pp. 1-6, doi: 10.1109/ICE.2018.8436265.

Wallach, W. & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. Oxford Press.

Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: towards a focus on tensions. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 195-200).

WHO. (2021). Ethics and Governance of Artificial Intelligence for Health: WHO guidance. In World Health Organization. http://apps.who.int/bookorders.

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

An Application with Python Software for the Classification of Chemical Data

Gonca ERTÜRK[1], Oğuz AKPOLAT[2]

ABSTRACT

Nowadays, much data can be generated and stored by chemical analyses. It is possible to evaluate these data, to reveal the relationships between them, and to make predictions with new data measured based on these relationships thanks to data mining algorithms. Monitoring the treatment processes and providing the necessary controls for environmental studies are based on the continuous determination of wastewater and activated sludge characteristics. The main criteria for determining the properties of wastewater are biochemical oxygen demand (BOD5), chemical oxygen demand (COD), total organic carbon (TOC), and dissolved oxygen (DO). Among these parameters, BOD5 measurement takes 5 days, while the others can be measured within 1-2 hours at most. Since BOD5 values can be mathematically correlated with other parameters, estimating them in a short time will provide a great advantage in terms of process control. In this study, a data set was created by measuring the specified parameters from 334 samples taken from a treatment plant for statistical evaluation, and the interactions of the parameters in this data set with each other were analyzed by the decision tree method. Thus, by considering the weighted effects of the parameters, it was tried to predict the probable BOD5 value of an unknown sample. The algorithm selected for this data mining study was modeled with PYTHON software and the performance of the algorithm in the estimation of the BOD5 parameter depending on other parameters was examined by extracting decision tree rules.

Keywords: Wastewater, Analysis, Data Mining, Classification, Decision Trees

DOI: 10.26650/JODA.1264915

[1]Muğla Sıtkı Koçman University, Graduate School of Natural and Applied Sciences, Chemistry Department, Mugla, Turkiye
[2]Muğla Sıtkı Koçman University, Science Faculty, Chemistry Department, Mugla, Turkiye

ORCID: G.E. 0000-0002-8821-0330;
O.A. 0000-0002-6623-4323

Corresponding author:
Oğuz AKPOLAT,
Muğla Sıtkı Koçman University, Science Faculty, Chemistry Department, Mugla, Turkiye
E-mail: oakpolat@mu.edu.tr

Submitted: 17.03.2023
Revision Requested: 13.04.2023
Last Revision Received: 01.05.2023
Accepted: 03.05.2023
Published Online: 25.07.2023

Citation: Erturk, G., & Akpolat, O. (2023). An application with python software for the classification of chemical data. *Journal of Data Applications, 1*, 49-68. https://doi.org/10.26650/JODA.1264915

Introduction

One of the areas of chemistry in which a large number of data are produced is environmental chemistry. When examined from an environmental point of view, the largest part of the pollution in wastewater consists of detergents, organic substances, and oils. The analyses used to determine the properties of wastewater is based on chemical ones, where quantitative results can be obtained, rather than biological and physical ones, where qualitative measurements can be performed. Measurements based on quantitative analysis are based on gravimetric, volumetric, or physicochemical methods. The Aeration pool is of great importance for the activated sludge process, determining the characteristics of wastewater and activated sludge refers to analyses such as acidity, temperature, conductivity, dissolved oxygen, oxygen saturation, salinity, electrical conductivity, chemical oxygen demand, suspended solids, total nitrogen, total phosphorus and biological oxygen demand to be performed in samples taken from raw wastewater and treated wastewater coming to treatment plants, as well as measurements made on the design parameters of the sludge samples used for biological treatment. These are ventilation time, suspended solids concentration (MLSS), solid retention time parameter including outlet water quality, temperature and biokinetics as design variables related to sludge production rate and oxygen requirement factors; final precipitation pool surface hydraulic load and solid load, sludge volume index, solid retention time and temperature as design variables related to sludge production rate factors; the MLSS concentration, which includes the sludge recycling rate and the MLSS recycling concentration, as well as the sludge volume index and the final settling pool, are the recycling rate as design variables related to the factors of surface hydraulic load and solid load. The properties of wastewater are classified as physical, chemical, and biological as follows: (Toprak, 2018; Tchobanoglous and Burton, 1991; Eltem, 2001; Wikipedia, 2016; amazon, 2016).

1. Physical Properties of Wastewater: It consists of total solids, odor, temperature, and color.

2. Chemical Properties of Wastewater: Organic substances in wastewater are mostly composed of benzene derivatives, such as proteins, carbohydrates, fats and oils, urea, soap, detergents, and volatile components. Biochemical oxygen demand (BOD_5) is a measure of the amount of dissolved oxygen used by microorganisms for the biochemical oxidation of organic compounds and the most widely used. It is performed by chemical oxygen requirement (COD) test of wastewater to measure the organic matter content. The COD value of wastewater is often higher than the BOD value. In particular, the total organic carbon test (TOC) is applied to measure the total organic carbon content of wastewater at low concentrations. Acidity is important in determining the quality of the inorganic content of wastewater, and pH is most commonly used for this. The others are chloride, alkalinity, nitrogen, phosphorus, sulfur, heavy metals, and toxic compounds and gases.

3. Biological Properties of Wastewater: The apparent group of organisms in domestic wastewater are plants, animals, and microorganisms such as bacteria, algae, fungi, protozoa, and viruses. Coliform bacteria are an indicator of contamination from human waste. Algae also cause taste and smell problems. During the treatment of wastewater, they decompose organic substances with bacteria.

4. Determination of the Properties of Wastewater: Determination of Biochemical Oxygen Demand (BOD_5), Chemical Oxygen Demand (COD), Total Organic Carbon (TOC), and Dissolved Oxygen (DO) amounts are the most basic measurement criteria for the characterization of wastewater.

As in the rest of the world, all wastewater treatment plants in Türkiye are operated in accordance with the Environment Law and the Water Pollution Control Regulation implemented by the Ministry of Environment and Urbanization. In domestic biological wastewater treatment plants, domestic wastewater from households is treated and restored to nature, and it is also aims to protect the water mass in the basin. Monitoring the treatment processes and providing the necessary controls is only possible by continuously measuring the characteristics of wastewater and activated sludge. Analyses to be made with samples taken from raw wastewater or treated wastewater coming to treatment plants are acidity (pH), temperature (T), conductivity (C), dissolved oxygen (DO), oxygen saturation (SO), salinity (SA), electrical conductivity (EC), chemical oxygen demand (COD), suspended solids (LSS), total nitrogen (TN), total phosphorus (TP) and biological oxygen demand (BOD_5), and similarly analyzes made for activated sludge samples can be listed as aeration time (AT), sludge suspended solids concentration (MLSS), suspended volatile solids concentration (MLVSS), temperature (T), sludge production rate (ASPR) and retention time (RT) including bio-kinetics (BK) and the recycling rate (FBR). The determination of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total organic carbon (TOC) and dissolved oxygen (DO) amounts are the most basic measurement criteria for the characterization of wastewater in determining its properties.

As 11 of the wastewater parameters counted in one of the studies in recent years can be measured in a one-day study conducted in the laboratory, it is stated that the measurement of the BOD_5 parameter takes a minimum of 5 days. In a laboratory study done for samples taken from a treatment facility for statistical evaluation, a dataset was created by measuring 12 parameters from 334 samples. Over the BOD_5 parameter, the effects of the other parameters in this data set were examined using the decision tree method by the KNIME data mining package. Thus, it was attempted to estimate the possible BOD_5 value of a sample whose result is unknown by considering the weighted effects of parameters whose effects on the BOD_5 parameter are known. This shows us that environmental measurement data can be re-

evaluated by data mining methods. From this and similar studies, it is understood that statistical evaluations regarding the measured values and estimates can be made between the parameters from the studies conducted for the activated sludge quality (Güller et al. 2019; Born, 2017; Weka, 2019; Synder and Wyant, 2018; Mukhtarov, 2020).

Data mining is the acquisition of valid and applicable information from data stacks by a dynamic process. In these processes, many different techniques are used, such as classification, clustering, data summarization, learning classification rules, finding dependency networks, variability analysis and abnormal detection. In data mining, classification and curve fitting are defined as prediction methods, while methods such as clustering and association analysis are described as descriptive. Data mining techniques are divided into two different categories such as supervised learning and unsupervised learning. The difference between supervised learning and unsupervised learning is unsupervised learning learns from the data but without reference. Therefore, it is not necessary to create a prior model in unsupervised learning. As classification is an supervised learning technique, clustering is one of the unsupervised. It separates data into some groups called clusters in which objects are similar to each other. The classification method which is one of the main methods of data mining is based on a learning algorithm. It is applied in order to discover hidden patterns in large-scale data. The main classification methods are decision trees, Bayesian classification, artificial neural networks and support vector machines. Classification is done by quickly examining the attributes of a new object and assigning that object to a predefined class. The important thing here is that the characteristics of each class are determined in advance. Clustering is the grouping of data according to their proximity or distance to each other, and there are no predetermined group boundaries, but it can be optimized by giving the number of groups. Data mining software is divided into two groups as commercial and open source and, data mining algorithms are operated directly in some software without coding, or they can be modeled in software that can be coded, such as Python. Python is a widely used, high-level, general-purpose, interpreted, and dynamic programming language. The design philosophy emphasizes the readability of the code, and the syntax allows programmers to express concepts in fewer lines of code than is possible in languages such as C++ or Java. (Silahtaroğlu, 2016; Alan and Karabatak, 2020; Çelik, 2009; Çınar, 2019; Kacur, 2020; Sampaio and Landup, 2022; Robinson, 2022).

During the recovery of wastewater, some tests are performed for the quality of activated sludge by characterizing the wastewater and the acquired water, and only the treatment process can be controlled by these tests. It is clear from the studies conducted that the data stacks obtained by all the physical, chemical and biological analyses performed during these processes can only be examined in detail by data mining techniques that are related to each other. Based on this, estimates of measurement parameters can also be made. Data mining algorithms make it possible to identify the relationships between these data by evaluating them and making

predictions with the help of new data measured based on these relationships. In this study, a suitable algorithm for Decision Trees will be selected from the data analysis methods to be applied in the study of wastewater characteristics with this data set, modeled by coding with Python software, and the performance of the algorithm in estimating the BOD_5 parameter depending on other parameters will be examined by extracting decision rules. Pandas, NumPy, Pyplot, and scikit-learn packages will be used as the basis for programming with Python (Nelson, 2022; activestate, 2022; Li, 2017; Sampaio and Landup, 2022; Robinson, 2022; anaconda, 2022).

Material and Method

Data mining models can be grouped under four main headings: prediction, clustering, connection analysis, and difference deviations. Predicting and clustering investigate each record's relationship to others, while objective and temporal connections can be examined in connection analysis. The most well-known classification techniques used for prediction are decision trees, statistical-based algorithms such as Bayesian and Regression, distance-based algorithms, and artificial neural networks. Of these, the classification can be mathematically defined as:

$D=\{t_1,t_2,\ldots,t_n\}$

Let's have a database and let each t_i show a record.

$C=\{C_1,C_2,\ldots,C_m\}$

Let m denote the set of classes consisting of classes.

$f:D\rightarrow C$ and each t_i should belong to a class. Here C_j is a separate class, and each class contains its own records. So, it can be shown in the form:

$C_j=\{t_i/f(t_i)=C_j,1\leq i\leq n,vet_i\in D\}$

Classification can also carry a class (discrete) and continuous value of the dependent variable with the class we have or its statistical definition. In this respect, it approaches regression or multi-term regression. Classification can also be defined as a supervised learning approach in which hidden patterns within a certain range are revealed. The most common of these algorithms are ID3 and C4.5 (Silahtaroğlu, 2016). Normalization is one of these algorithms' most frequently used data transformation processes. With Min-Max normalization (Eq. 1), which is the most used of data normalization techniques, the original data are converted to the new data in range by a linear transformation. This data range is usually 0-1.

$Newdata=\{(Rawdata\text{-minRawdata})/(maxRawdata\text{-minRawdata})\}$ *Eq. 1*

The principles of the decision tree method and the steps of the decision tree algorithm are given below:

Basics of Decision Tree Method

1. Identification of the problem.
2. Drawing/structuring the decision tree.
3. Assigning the probabilities of the occurrence of events.
4. Calculation of the expected return (or benefit) for the corresponding chance point-backward, the transaction.
5. Assignment of the highest expected return (benefit) to the relevant decision point-backward, comparison.
6. The submission of the proposal is based on its principles.

The Steps of The Decision Tree Algorithm

1. The learning set T is created.
2. The attribute that best separates the samples in the set T is determined.
3. A node of the tree is created with the selected attribute, and child nodes or leaves of the tree are created from this node. Determine the instances of the subset of child nodes.
4. for each sub-dataset created in step three:

· If the samples all belong to the same class

· If there is no qualification to divide the samples

· If there is no sample with the remaining attribute value, the process is terminated. In the other case, the process is continued from the second step to separate the subset of data.

The decision tree can be easily encoded in any programming language using IF-ELSE expressions. Decision tree classification is a classification method that creates a model in the form of a tree structure consisting of decision nodes and leaf nodes according to the property and goal. The decision tree algorithm is improved by dividing the data set into small and even smaller pieces. A decision node may contain one or more branches. The first node is called the root node. A decision tree can consist of both categorical and numerical data. The randomness, uncertainty, and probability of an unexpected situation occurring in the formation

of any situation are defined by entropy, and if all the samples are regular/homogeneous, their entropy becomes zero. Here entropy is defined as in Eq. 2:

$$E(S) = \sum_{i\ 1}^{c} -p_i log_2 p_i \qquad \textit{Eq. 2}$$

Entropy is not calculated only on the target. In addition, entropy can also be calculated on properties. But when calculating entropy on properties, it is taken into account in the target. In this case, entropy is defined as in Eq. 3:

$$E(T,X) = \sum_{ceX} P(c)E(c) \qquad \textit{Eq. 3}$$

Information gain (Gain) is based on subtracting all entropy after dividing a dataset on a feature (Eq. 4). If the entropy is small, the importance of the feature increases for the Decision Tree algorithm ID3. On the other hand, as it gets closer to 1 the importance of the feature decreases. However, in information gain, the situation is the opposite, and in this respect, it can be thought of as the inverse of entropy. While constructing the Decision Tree, the feature with the highest information gain is selected.

Gain(*T,X*)=*Entropy*(*T*)-*Entropy*(*T,X*) *Eq. 4*

Overfitting is an important problem for decision tree models and many other prediction models. Overfitting occurs when the training set continues to reduce errors in a way that affects the learning algorithm. To avoid overfitting in a decision tree construction, two approaches are usually used:

- Pre-pruning: Stopping the growth of the tree before the containment process.
- Post-pruning: first creating the whole tree and then removing the unnecessary parts from the tree.

Due to the difficulty in determining when pruning will be done in practice, the first approach is hardly used. The second approach is much more successful. Attention should be paid to the following steps in this approach.

- A different dataset than the training data is used to decide on the pruning process. This data set is called the validation dataset. The validation dataset is used to decide on unnecessary nodes.
- After obtaining a decision tree, using statistical methods such as error estimation and significance testing (Chi-Square Testing), it is decided whether there will be pruning

and expansion (expanding – adding new nodes to the tree) on the training data.

- The Minimum Distance Description Principle is a measure between the Decision tree and the training dataset. When the size (tree) + Size (non-classifiable tree) is minimized, the tree growth is stopped.

The data tested in the decision tree modeling performed using the Python programming language in this section were measured for 334 days in a wastewater treatment plant and presented in Table 1 (Güller et al., 2019**)**. The measured values were chemically defined as acidity (pH: -), Temperature (TemperC: °C), Total Phosphate (TotalphosmgPL: mg/L), Suspended Solids (CVLmgPL: mg/L), Chemical Oxygen Demand (CODmgPL: mg/L), and Biological Oxygen Demand (BOD_5mgPL: mg/L).

Table 1. *Analytical values of chemical substances measured in wastewater samples.*

Label	Ph	TemperC	TotPhosmgPL	CVLmgPL	CODmgPL	BOD_5mgPL
1	7.3	8.7	17.7	310	920	19.41
2	7.55	9.7	15.9	150	495	169
3	7.47	10.3	11.6	180	401	209
4	8.03	9.7	5.2	130	433	272
Experimental data file in dimension of 334*7 (**# Data: EnvirodataI.txt)**						
331	8	17	1.95	154	474	120
332	7.77	17.2	0.55	8.4	142.65	36.4
333	7.76	30	0.31	42	162.12	45.6
334	7.41	24.4	3.43	33	153.5	38.4

The Python codes required for the analysis of the algorithm written for decision trees in data evaluation (Environment_Classification_00_Decission_Tree_Performance.py) and its numerical output are given as **Appendix 1**. The graphs of the program outputs are given too in Figure 1, Figure 2, Figure 3, and Figure 4 as "Distribution of samples according to BOD_5 values", "Display of sample distributions in box graphics", "Histograms of distributions related to chemical measurement values", "Correlation of chemical measurement values to each other" and "Decision tree of samples classified according to BOD_5 values", respectively.

Results

In this section, the outputs and results of the decision tree application selected for solving classification problems have also been examined, and the numerical and % distribution of BOD_5 samples are given in Table 2.

Table 2. *BOD5 percent distributions of all data in the specified intervals.*

BOD_5 Intervals	**Numeric value**	**Distribution (%)**
0-250	279	72
251-500	70	18
501-750	30	8
751-999	5	2
0-999	384	100

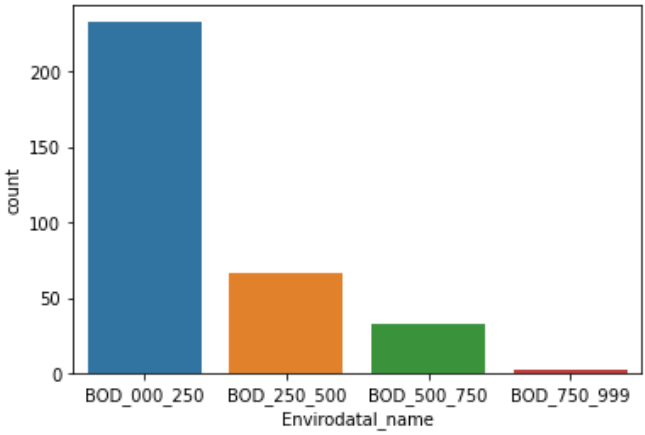

Figure 1. *Distribution of samples according to BOD_5 values.*

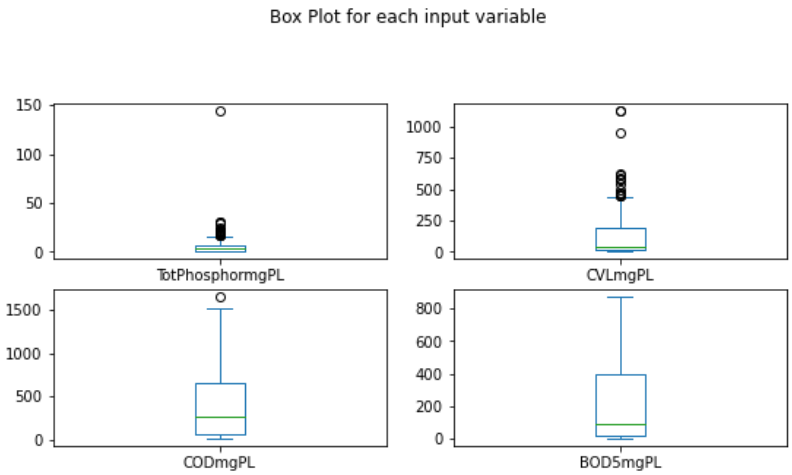

Figure 2. *Showing the sample distributions in box graphics.*

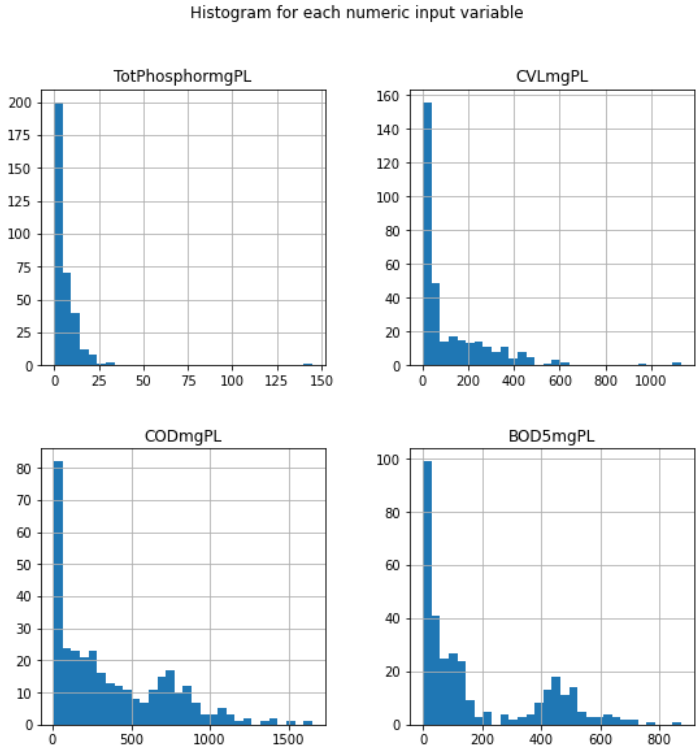

Figure 3. *Histograms of distributions related to chemical measurement values.*

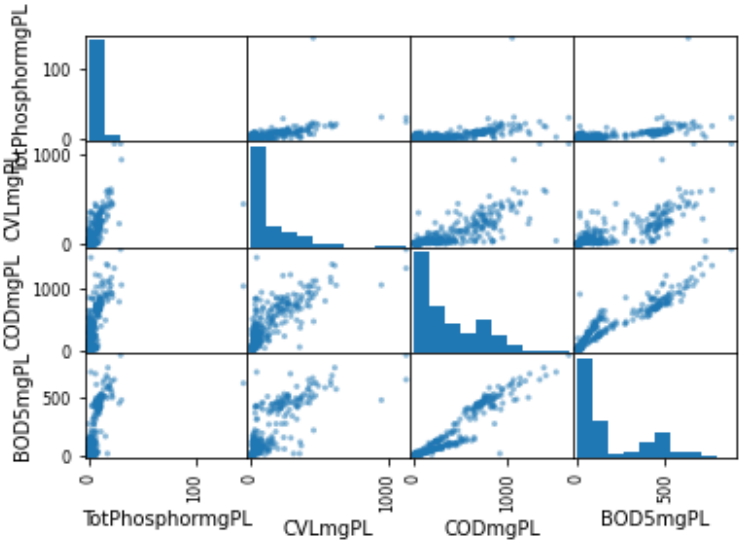

Figure 4. *Correlations of chemical measurement values with each other.*

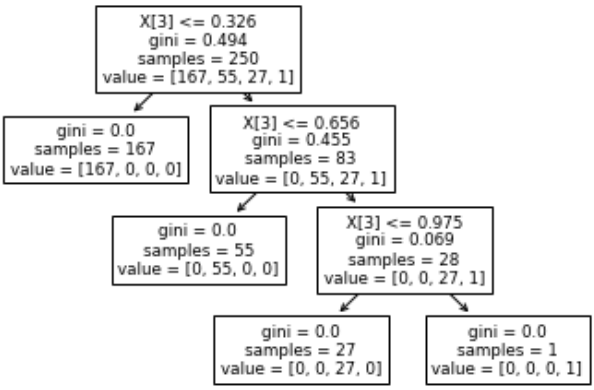

Figure 5. *Decision tree of samples classified according to BOD_5 values.*

The distribution of the Confusion matrix related to the test data taken from the decision tree is given in Table 3, and the accuracy value of this distribution is calculated as 100%.

Table 3. *The Confusion matrix for the test data, which accounts for 20% of the total data.*

	Bin1	**Bin2**	**Bin3**	**Bin4**
Bin1	66			
Bin2		11		
Bin3			6	
Bin4				1

Conclusion

Monitoring the treatment processes and providing the necessary controls is only possible by continuously measuring the characteristics of wastewater and activated sludge. Analyses to be made with samples taken from raw wastewater or treated wastewater coming to treatment plants are acidity (pH), temperature (T), conductivity (C), dissolved oxygen (DO), oxygen saturation (SO), salinity (SA), electrical conductivity (EC), chemical oxygen demand (COD), suspended solids (LSS), total nitrogen (TN), total phosphorus (TP) and biological oxygen demand (BOD_5), and similarly analyses made for activated sludge samples can be listed as aeration time (AT), sludge suspended solids concentration (MLSS), suspended volatile solids concentration (MLVSS), temperature (T), sludge production rate (ASPR) and retention time (RT) including bio-kinetics (BK) and recycling rate (FBR). Determination of biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total organic carbon (TOC) and dissolved oxygen (DO) amounts are the most basic measurement criteria for the characterization of wastewater in determining its properties.

As 11 of the wastewater parameters counted in one of the studies in recent years can be measured in a one-day study conducted in the laboratory, in a study conducted by Güller et al. (2019), it is stated that the measurement of the BOD_5 parameter takes a minimum of 5 days. For this work, in a laboratory study done for samples taken from a treatment facility for statistical evaluation, a dataset was created by measuring 12 parameters from 334 samples. Over the BOD_5 parameter, the effects of the other parameters in this data set were examined using the decision tree method by the KNIME data mining package. Thus, it was attempted to estimate the possible BOD_5 value of a sample whose result is unknown by considering the weighted effects of parameters whose effects on the BOD_5 parameter are known. This shows us that environmental measurement data can be re-evaluated by data mining. These studies in this area are quite new and show promise in the evaluation of environmental data stacks.

In this study, the four selected parameters considered to be much more effective from the data set given above, were chemically defined as acidity (pH), Temperature (°C), Total Phosphate (mg/L), Suspended Solids (mg/L), Chemical Oxygen Demand (mg/L), and Biological Oxygen Demand (mg/L). For this study a suitable algorithm was selected as a decision tree from the data analysis methods to be applied and modeled by coding with Python software, and the performance of the algorithm in estimating the BOD_5 parameter depending on other parameters was examined by extracting decision rules. Pandas, NumPy, Pyplot, and scikit-learn packages were used as the basis for programming with Python.

When the results obtained from the data set showing the analysis results of 4 parameters related to 334 domestic qualified wastewaters taken from an earlier study and evaluated by the decision tree method encoded by the Python algorithm were examined in this study, it was found that the BOD_5 value distribution of 334 samples was below 250 by 72%. The proportion of those with a BOD_5 value between 251-500 is 18%, while the proportion of those between 500-750 is 6%. As Figure 4 is examined, it is understood that the variable that affects the BOD_5 value the most is COD (Chemical Oxygen Demand). As compared to those of the study published by Güller et al. (2019) performed by KNIME software it is understood that the results of this study is very close to theirs, as expected.

Ethics Committee Approval: Authors declared that this study does not require ethics committee approval.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- O.A., G.E.; Data Acquisition- O.A.; Data Analysis/Interpretation- O.A., G.E.; Drafting Manuscript- O.A., G.E.; Critical Revision of Manuscript- O.A.; Final Approval and Accountability- O.A., G.E.; Material and Technical Support- O.A.; Supervision- O.A.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Activestate. (2022). *How to Classify Data In Python using Scikit-learn.* Retrieved May 3, 2023, from https://www.activestate.com/resources/quick-reads/how-to-classify-data-in-python/

Alan, A., & Karabatak, B. (2020). *Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi*, *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 32(2), 531-540.

Amazon, (2016). Retrieved May 3, 2023, from https://www.amazon.com/Hach-8505700-Measurement-Luminescent-Dissolved/dp/B00R3EGHJ4

Anaconda. (2022). *anaconda/packages/python.* https://anaconda.org/anaconda/python/anaconda/packages/(python3.10.6)

Çelik, M. (2009). *Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama* [Yüksek Lisans Tezi]. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Ana Bilim Dalı.

Çınar, A. (2019). Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi ve R Dili ile Bir Uygulama, *Marmara Üniversitesi Öneri Dergisi*, 14(51), 90-111.

Doğan, O. (2017). Ücretsiz Veri Madenciliği Araçları ve Türkiyede Bilinirlikleri Üzerine Bir Araştırma, *Ege Stratejik Araştırmalar Dergisi*, 8(1), 77-93.

Eltem, R. (2001). *Atık Sular ve Arıtım*, Ege Üniversitesi Fen Fakültesi Yayınları, 172

Güller, S., Silahtaroğlu, G. ve Akpolat, O. (2019). Analysis waste water characteristics via data mining: A Muğla province case and external validation. *Communications in Statistics Case Studies Data Analysis and Applications*, 5(3), 200-213. https://dx.doi.org/10.1080/23737484.2019.1604192

Jiawei, H., Kamber, M., & Pei, J. (2012). *Data Mining; Concepts and Technics*, Morgan Kaufmann Publishers, Elsevier Inc.

Kacur, T., M. (2020). *Atık Su ve Aktif Çamur Karakteristiklerinin Tahmininde Karar Ağaçları ve Yapay Sinir Ağlarının Karşılaştırılması* [Yüksek Lisans Tezi]. Muğla Sıtkı Koçman Üniversitesi Çevre Bilimleri Ana Bilim Dalı.

Li, S. (2017). *Solving A Simple Classification Problem with PYTHON — Fruits Lovers' Edition*. Retrieved May 3, 2023, from https://towardsdatascience.com/solving-a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2

Meyers, D.N., & Wilde, F. D. (2012). *USGS TWRI Book 9–A7* (Third Edition), http://water.usgs.gov/owq/FieldManual/Chapter7/NFMChap7.pdf

Mukhtarov, M. (2020). *Atık Su ve Aktif Çamur Karakteristiklerinin Sınıflandırılması ve Uygulanan Analiz Yöntemlerinin Değerlendirilmesi* [Yüksek Lisans Tezi]. Muğla Sıtkı Koçman Üniversitesi Çevre Bilimleri Ana Bilim Dalı.

Nelson, D. (2022). *Overview of Classification Methods in PYTHON with Scikit-Learn*. Retrieved May 3, 2023, from https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/

Qiao, J., Li, W., & Han, H. (2014). Soft Computing of Biochemical Oxygen Demand Using an Improved T–S Fuzzy Neural Network, *Chinese Journal of Chemical Engineering*, 22, 1254–1259.

Robinson, S. (2022). *Decision Trees in PYTHON with Scikit-Learn.* Retrieved May 3, 2023, from https://stackabuse.com/decision-trees-in-python-with-scikit-learn/

Sampaio, C., & Landup, D. (2022). *Linear Regression in PYTHON with Scikit-Learn*. Retrieved May 3, 2023, from https://stackabuse.com/linear-regression-in-python-with-scikit-learn/

Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları*, (2. Baskı). Papatya Yayıncılık.

Synder, R., & Wyant, D. (2018). *Activated Sludge Process Control Training Manuel, DEO, Water Resources Division*. Retrieved May 3, 2023, from https://www.michigan.gov/documents/deq

Tchobanogluos, G., & Burton, F. L. (1991). *Wastewater Engineering Treatment, Disposal, and Reuse*, McGraw-Hill Book Co.

Toprak, H. (2018). *Aktif Çamur Sürecinin Tanımı*. Retrieved May 3, 2023, from http://web.deu.edu.tr/atiksu/ana58/aktifkurs.doc

Weka. (2019). *Weka*. Retrieved May 3, 2023, from https://www.cs.waikato.ac.nz/ml/weka/index.html

Wikipedia, (2016). *Biochemical oxygen demand*. Retrieved May 3, 2023, from https://en.wikipedia.org/wiki/Biochemical_oxygen_demand

Appendix 1. Python Commands for Classification and Performance Evaluation

```
print('# Environment_Classification_00_Decission_Tree_Performance')
# Data: EnvirodataI.txt
attributes = ["EnvirodataI_label", "EnvirodataI_name", "pH", "TemperC",
"TotPhosphormgPL", "CVLmgPL", "CODmgPL", "BOD5mgPL"]
target_attribute = ["BOD5mgPL"]
# ('Downloading Required Libraries ')
import pandas as pd
import numpy as np

envirodata = pd.read_table('envirodataI.txt')
print(envirodata.head())
print(envirodata.shape)
print(envirodata['EnvirodataI_name'].unique())
print(envirodata.groupby('EnvirodataI_name').size())
import seaborn as sns

sns.countplot(envirodata['EnvirodataI_name'], label="Count")
# Distribution Measures
import matplotlib.pyplot as plt

envirodata.drop('EnvirodataI_label', axis=1).plot(kind='box', subplots=True,
layout=(4, 2), sharex=False, sharey=False, figsize=(9, 9), title='Box Plot
for each input variable')
plt.savefig('envirodata_box')
plt.show()

import pylab as pl

envirodata.drop('EnvirodataI_label', axis=1).hist(bins=30, figsize=(9, 9))
pl.suptitle("Histogram for each numeric input variable")
plt.savefig('envirodata_hist')
plt.show()

import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix

# Selection of three numerical features
attribute = ["TotPhosphormgPL", "CVLmgPL", "CODmgPL", "BOD5mgPL"]
# Plot the scatter matrix
# Depending on the features
scatter_matrix(envirodata[attribute])
plt.show()
# Statistical Study
X = envirodata[attribute]
y = envirodata['EnvirodataI_label']
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
print("X_train", X_train, "X_test", X_test, "y_train", y_train, "y_test",
y_test)
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
print("X_train", X_train, "y_train", y_train)
# Models
# Decission Tree
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier().fit(X_train, y_train)
print('Accuracy of Decision Tree classifier on training set: {:.2f}'
      .format(clf.score(X_train, y_train)))
print('Accuracy of Decision Tree classifier on test set: {:.2f}'
      .format(clf.score(X_test, y_test)))
# Confussion Matrix for Decission Tree
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

pred = clf.predict(X_test)
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
# Creating a Decision Tree
print('# Preparation of Decision Tree ')
X = [[0, 0], [1, 1]]
Y = [0, 1]
clf = clf.fit(X, Y)
clf.predict([[2., 2.]])
print('# Transforming the Data Set)')
X, y = X_train, y_train
# print("X",X,"y",y)
from sklearn import tree

print('# Creating a Decision Tree ')
clf.fit(X, y)
tree.plot_tree(clf)
print("#Performance for Decision Tree ")
CMAT = [[66, 0, 0, 0],
        [0, 11, 0, 0],
        [0, 0, 6, 0],
        [0, 0, 0, 1]]
print('CMAT', CMAT)
print('Interpretation of the Confussion Matrix')
print(CMAT)
print('Lines Prediction(T) Columns Real(G)')
```

```
print('High:"H", Low:"L", Middle:"M", VeryLow:"V"')
print('TOTAL:"TOT"')
CMAT00_TH_GH = CMAT[0][0]
CMAT01_TH_GL = CMAT[0][1]
CMAT02_TH_GM = CMAT[0][2]
CMAT03_TH_GV = CMAT[0][3]
CMAT10_TL_GH = CMAT[1][0]
CMAT11_TV_GL = CMAT[1][1]
CMAT12_TL_GM = CMAT[1][2]
CMAT13_TL_GV = CMAT[1][3]
CMAT20_TM_GH = CMAT[2][0]
CMAT21_TM_GL = CMAT[2][1]
CMAT22_TM_GM = CMAT[2][2]
CMAT23_TM_GV = CMAT[2][3]
CMAT30_TV_GH = CMAT[3][0]
CMAT31_TV_GL = CMAT[3][1]
CMAT32_TV_GM = CMAT[3][2]
CMAT33_TV_GV = CMAT[3][3]
print(CMAT00_TH_GH, CMAT01_TH_GL, CMAT02_TH_GM, CMAT03_TH_GV)
print(CMAT10_TL_GH, CMAT11_TV_GL, CMAT12_TL_GM, CMAT13_TL_GV)
print(CMAT20_TM_GH, CMAT21_TM_GL, CMAT22_TM_GM, CMAT23_TM_GV)
print(CMAT30_TV_GH, CMAT31_TV_GL, CMAT32_TV_GM, CMAT22_TM_GM)
#
TOT_TH_G = CMAT00_TH_GH + CMAT01_TH_GL + CMAT02_TH_GM + CMAT03_TH_GV
TOT_TL_G = CMAT10_TL_GH + CMAT11_TV_GL + CMAT12_TL_GM + CMAT13_TL_GV
TOT_TM_G = CMAT20_TM_GH + CMAT21_TM_GL + CMAT22_TM_GM + CMAT23_TM_GV
TOT_TV_G = CMAT30_TV_GH + CMAT31_TV_GL + CMAT32_TV_GM + CMAT33_TV_GV
TOT_T = TOT_TH_G + TOT_TL_G + TOT_TM_G + TOT_TV_G
N = TOT_T  #
TOT_GH_T = CMAT00_TH_GH + CMAT10_TL_GH + CMAT20_TM_GH + CMAT30_TV_GH
TOT_GL_T = CMAT01_TH_GL + CMAT11_TV_GL + CMAT21_TM_GL + CMAT31_TV_GL
TOT_GM_T = CMAT02_TH_GM + CMAT12_TL_GM + CMAT22_TM_GM + CMAT32_TV_GM
TOT_GV_T = CMAT03_TH_GV + CMAT13_TL_GV + CMAT23_TM_GV + CMAT33_TV_GV
TOT_G = TOT_GH_T + TOT_GL_T + TOT_GM_T + TOT_GV_T
#
Total_Accurate_Forecast = CMAT00_TH_GH + CMAT11_TV_GL + CMAT22_TM_GM +
CMAT33_TV_GV
# Accuracy= Total_Accurate_Forecast /N
Accuracy = Total_Accurate_Forecast / N
print("N=", N, "Accuracy=", Accuracy)
```

Appendix 1. Output:

```
Environment_Classification_00_Decission_Tree_Performance.py the outputs of
the named program are given below:

EnvirodataI_label EnvirodataI_name ... CODmgPL BOD5mgPL
0 1 BOD_000_250 ... 449.50 107.6
1 1 BOD_000_250 ... 393.38 100.0
2 1 BOD_000_250 ... 371.90 108.0
3 1 BOD_000_250 ... 560.41 155.0
4 1 BOD_000_250 ... 350.00 165.0
Selected: [5 rows x 7 columns] Total:(334, 7)
['BOD_000_250' 'BOD_250_500' 'BOD_500_750' 'BOD_750_999']
EnvirodataI_name
BOD_000_250 233
BOD_250_500 66
BOD_500_750 33
BOD_750_999 2

X_train TotPhosphormgPL CVLmgPL CODmgPL BOD5mgPL
278 5.100 245.00 575.00 341.0
92 0.800 66.00 227.71 59.6
312 11.900 413.00 807.00 519.0
234 4.785 76.20 950.00 455.0
216 2.940 15.85 230.00 122.8
.. ... ... ... ...
323 144.700 453.00 1051.00 629.0
192 1.110 15.00 45.00 10.0
117 0.790 5.00 63.00 12.0
47 1.420 35.60 308.29 58.0
172 1.280 17.00 51.00 16.0
Training set (%80): [250 rows x 4 columns]

X_test TotPhosphormgPL CVLmgPL CODmgPL BOD5mgPL
166 0.830 13.00 46.00 12.00
78 0.790 12.00 284.81 70.80
15 5.795 17.95 402.00 195.20
221 1.140 27.00 26.00 19.00
194 2.260 98.00 260.00 61.20
.. ... ... ... ...
171 0.320 13.60 43.61 8.92
8 2.145 81.10 420.00 195.12
223 2.975 29.80 274.00 125.20
236 6.950 38.60 604.00 315.00
156 1.090 7.00 32.00 10.00
Test set (%20): [84 rows x 4 columns]
```

```
for Training set y_train
278 2
92 1
312 3
234 2
216 1
..
323 3
192 1
117 1
47 1
172 1

for Test set y_test
166 1
78 1
15 1
221 1
194 1
..
171 1
8 1
223 1
236 2
156 1

X_train
[[3.52453352e-02 2.57112750e-01 3.73918829e-01 4.46833054e-01]
[5.52868003e-03 6.84931507e-02 1.42854291e-01 7.79677013e-02]
[8.22391154e-02 4.34141201e-01 5.28276780e-01 6.80159396e-01]
[3.30684174e-02 7.92413066e-02 6.23419827e-01 5.96266779e-01]
[2.03178991e-02 1.56480506e-02 1.44377911e-01 1.60811661e-01]
[9.60608155e-03 1.28556375e-01 3.76067864e-01 1.80736158e-01]
[4.07740152e-02 1.47523709e-02 3.06054558e-02 1.81942114e-02]
[1.03662751e-01 3.88830348e-01 5.88157019e-01 6.77537752e-01]
.....
[2.14236351e-03 4.32033720e-02 9.92149035e-02 5.96161913e-02]
[3.04077402e-02 2.95047418e-02 1.26413839e-02 2.08158557e-02]
[4.90670352e-03 2.25500527e-02 8.67198935e-02 2.93361997e-02]
[7.53282654e-02 2.46575342e-01 5.58882236e-01 6.35591443e-01]
[4.90670352e-02 1.95995785e-01 4.73719228e-01 6.13307466e-01]
[7.87836904e-02 3.85669125e-01 5.54890220e-01 6.44767198e-01]
[5.25915688e-02 9.35721812e-02 1.00465735e-01 9.81543624e-02]
[6.91085003e-03 2.55005269e-02 1.76533599e-01 9.46151426e-02]
[4.56116102e-03 5.75342466e-02 1.66573520e-01 9.02894295e-02]
[7.46371804e-02 2.23393045e-01 5.30272788e-01 6.04131711e-01]
[1.00000000e+00 4.76290832e-01 6.90618762e-01 8.24349832e-01]
[7.67104354e-03 1.47523709e-02 2.12907518e-02 1.29509228e-02]
```

```
[5.45957153e-03 4.21496312e-03 3.32667997e-02 1.55725671e-02]
[9.81340705e-03 3.64594310e-02 1.96467066e-01 7.58703859e-02]
[8.84588804e-03 1.68598525e-02 2.52827678e-02 2.08158557e-02]]

y_train
278 2
92 1
312 3
234 2
216 1
..
323 3
192 1
117 1
47 1
172 1

Accuracy of Decision Tree classifier on training set: 1.00
Accuracy of Decision Tree classifier on test set: 1.00
[[66 0 0 0]
[ 0 11 0 0]
[ 0 0 6 0]
[ 0 0 0 1]]
precision recall f1-score support
1 1.00 1.00 1.00 66
2 1.00 1.00 1.00 11
3 1.00 1.00 1.00 6
4 1.00 1.00 1.00 1
accuracy 1.00 84
macro avg 1.00 1.00 1.00 84
weighted avg 1.00 1.00 1.00 84
CMAT [[66, 0, 0, 0], [0, 11, 0, 0], [0, 0, 6, 0], [0, 0, 0, 1]]
Interpretation of the 'Confussion Matrix print(CMAT)
Rows Estimate (T) Columns Real (G)
High:"H", Low:"L", Middle:"M", VeryLow:"V"
Rows Estimate (T) Columns Real (G)
TOTAL:"TOT"
66 0 0 0
0 11 0 0
0 0 6 0
0 0 0 6
N= 84 Accuracy= 1.0
```

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

An Improved Protection Approach for Protecting from Ransomware Attacks

Ferhat GUVÇİ[1,2], Ahmet ŞENOL[3]

ABSTRACT

Ransomware is a type of malicious software that has become a significant threat to the security and availability of computer systems and data. Ransomware has found a special place in the world of malware and is the subject of many scientific studies, as it is a malicious software designed to benefit the user directly by using sensitive data of individuals or institutions. This research provides an in-depth study of ransomware, including its history and evolution. The primary objective of this research is to analyze the impact of ransomware attacks on organizations and individuals and to evaluate the effectiveness of existing countermeasures and mitigation strategies.

To achieve this objective, a comprehensive review of the literature and security provider sources on ransomware was conducted and data analyzed from real-world ransomware incidents. The findings indicated that ransomware attacks are becoming more sophisticated and complex, targeting a wide range of industries and geographical regions, which poses a significant financial and reputational risk to victims.

Moreover, this research showed that traditional security measures such as antivirus software, firewalls, and backups may not be sufficient to prevent or recover from ransomware attacks. Instead, artificial intelligence applications and a multi-layered defense approach that combined technical, administrative, and legal measures is necessary to reduce the likelihood and impact of ransomware incidents.

Overall, this article provides a valuable contribution to the understanding of ransomware threats and the development of effective countermeasures, and contributes to the literature especially on defense methods by explaining how to apply defense methods against ransomware attacks in light of field experience.

Keywords: Encryption, Machine Learning, Malware, Ransomware

DOI: 10.26650/JODA.1312412

[1]Uskudar University, Cyber Security Masters Degree Program, Istanbul, Turkiye
[2]İHS Kurumsal Teknoloji A.Ş., 34718, Istanbul, Turkiye
[3]Uskudar University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Istanbul, Turkiye

ORCID: F.G. 0009-0005-4329-8550;
A.Ş. 0000-0001-9891-4596

Corresponding author:
Ferhat GUVÇİ,
Uskudar University, Cyber Security Masters Degree Program, Istanbul, Turkiye
E-mail: ferhat.guvci@st.uskudar.edu.tr

Submitted: 09.06.2023
Revision Requested: 09.06.2023
Last Revision Received: 07.07.2023
Accepted: 09.07.2023

Citation: Guvci, F., & Senol, A. (2023). An improved protection approach for protecting from ransomware attacks. *Journal of Data Applications, 1*, 69-82. https://doi.org/10.26650/JODA.1312412

Introduction

Ransomware is malicious software that locks your computer or blocks access to your data using private key encryption until a ransom is paid. It can spread through malicious links or attachments in emails, downloads from malicious websites, or drive-by downloads. Once installed, ransomware encrypts files on the computer and demands payment in exchange for a decryption key. If the ransom is not paid, the files remain encrypted and inaccessible. Ransomware attacks can be highly costly as they can lead to significant data loss and disruption in business operations (Richardson & North, 2017).

The secure storage and protection of information against constantly changing and evolving cyber-attack vectors are becoming increasingly important. Cyber-attacks are on the rise worldwide, causing significant financial damage to both individuals and organizations, amounting to millions of dollars. Among malicious software, ransomware holds a significant position in terms of financial harm. Ransomware, a customized and specialized form of malware, is designed to threaten the integrity and accessibility of data by encrypting and stealing it, aiming to profit through extortion.

This study focused on examining ransomware and its impact, explaining the working principles of ransomware, exploring how artificial intelligence and machine learning algorithms could be utilized for protection against ransomware, and providing information on methods to safeguard against such malicious software.

Literature Review

Numerous studies have been conducted in the world on the status of ransomware activities and methods of protection. Maurya et al. (2017), examined the historical development of ransomware and the general characteristics of popular types of ransomware. They studied the attack principles of different types of ransomware and conducted research on infection types and revealed how systems were affected by ransomware attacks. Richardson and North (2017), presented a brief history of ransomware, the arguments for and against paying the ransom by detailed research on the payment methods and the best practices to prevent an infection, how to specifically deal with an infected machine, and recover from an infection should one happen. Askarifar et al. (2018) did in-depth research on WannaCry, one of the most popular ransomware applications, and discussed how cybercriminals bypass computer defensive systems, and how WannaCry, as well as other types of ransomware, affected computer systems overall.

While traditional threat-based databased defense methods were often used for this purpose in the past, today we see that these methods are supported by artificial intelligence and machine learning Fernando et al. (2020) conducted a survey about how machine learning and deep

learning algorithms contributed to detection of ransomware attacks. They investigated the contributions of research into the detection of ransomware malware using machine learning and deep learning algorithms and attempted to identify weaknesses in machine learning approaches and how they could be strengthened.

Kapoor et al. (2021) presented Detection Avoidance Mitigation (DAM). They devised a framework including tools, strategies, techniques to avoid, detect and mitigate Ransomware since there was no such consolidated framework. Gwozdenko (2023) did a comprehensive study on how artificial intelligence could be used against ransomware and made important recommendations for future AI-shaped protection systems. In addition, the study took a visionary approach to predict future ransomware attacks with the help of machine learning algorithms and how to take necessary precautions before an attack occurs.

History of Ransomware

The history of ransomware dates back to 1989 when the first known ransomware, called the AIDS Trojan, was created by Joseph Popp. The initial ransomware was distributed on floppy disks to participants at the World Health Organization's International AIDS Conference. It used simple symmetric cryptography to encrypt file names, and tools to decrypt and the encryption key were released shortly afterward. In May 2005, the first modern ransomware, known as the GPCoder Trojan, spread through spam email attachments. In 2013, CryptoLocker emerged as notorious ransomware and spread in an unprecedented manner. It gained momentum by disguising itself as an email from well-known courier companies like UPS and FedEx. The original CryptoLocker version could encrypt 67 different file types. With subsequent versions, the software made it difficult to trace the attacker by accepting payments in Bitcoin. It demanded two Bitcoins from its victims, later increasing to ten Bitcoins. By December 2013, over 250,000 machines had been infected by this malicious software (Teodoro et al., 2021). By the end of 2015, the total ransom payments made by CryptoLocker victims reached around $27 million, according to estimates from the FBI. CryptoLocker used asymmetric cryptography for encrypting the files it targeted, directly attacked and encrypted the 'My Documents' folders by design. In January 2016, a ransomware named KeRanger emerged, which was the first ransomware attacks targeting Apple systems. KeRanger was primarily distributed through a compromised version of the Transmission BitTorrent client, a popular macOS application used for downloading files through the BitTorrent protocol. Attackers managed to infiltrate the official website of Transmission and replace the legitimate software installer with a malicious version containing the KeRanger ransomware. KeRanger needed three days to be activated and successfully encrypted more than 300 file types (Maurya et al., 2017). In April 2016, a ransomware named Petya emerged, which encrypted the entire hard disk and denied access without paying the ransom. Petya worked by overwriting the Master

Boot Record (MBR), rendering the operating system unable to recreate unencrypted files and leaving victims with the option to either pay the ransom or replace the disk. In 2017, the Bad Rabbit and WannaCry attacks gained global attention and affected numerous international companies. WannaCry exploited a vulnerability in the Windows operating system called EternalBlue, which was allegedly developed by the U.S. National Security Agency (NSA) and later leaked by a hacker group called The Shadow Brokers. The ransomware propagated through the network by scanning for vulnerable systems and exploiting the EternalBlue vulnerability to gain unauthorized access. The WannaCry ransomware attack impacted more than 150 countries and resulted in damages reaching up to $1 billion within a week, with at least 100,000 organizations worldwide being affected (Askarifar et al., 2018).

In the years 2020 and 2021, another prominent ransomware attack software was NetWalker, also known as Mailto. NetWalker used asymmetric cryptography for the encryption process. This malicious software made a name for itself by incorporating the COVID-19 pandemic into its attacks, reaching a large audience through phishing emails related to the coronavirus. One example which targeted K-Electric, Pakistan's largest private power utility company, demanded $3.85M initially, $7.7M after a week. NetWalker collected around 2,795 Bitcoin (roughly $30M as of mid-September 2020 Bitcoin value), purportedly making it one of the most profitable active variants of ransomware (Health Sector Cybersecurity Coordination Center, 2020).

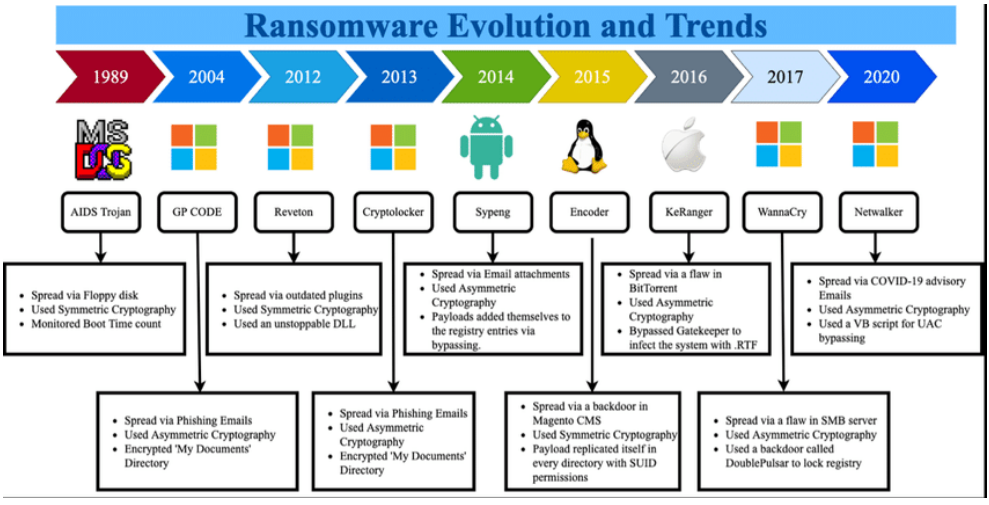

Figure 1. *Ransomware Evolution and Trends (Kapoor et al., 2021).*

How Ransomware Attacks Happen

The attack process is very similar across ransomware types. Attackers gain unauthorized access to a system or network through various methods, such as phishing emails, exploiting vulnerabilities, social engineering, or compromising weakly protected credentials. Any malicious website enters the victim's machine via email attachment or any malicious link and uses it as a base. When it starts running on the victim's machine, it establishes a connection to the Command and Control server. It sends the victim's machine information, reconnaissance information of other machines in the vicinity to the attack center and receives a randomly generated symmetric key from there. After obtaining the encryption key, it searches for specific files and folders to be encrypted (Monika et al., 2016). In some cases, it searches all disk drives, network shares, and removable drives to encrypt its data, without looking for the file path for encryption (Monika et al., 2016). Meanwhile, the malware deletes backup folders, all restore points and shadow copies. When the encryption process is completed, a message about what happened to the victim is displayed or the victim is directed by showing the directory where this message is located (Monika et al., 2016).

Stages of a Ransomware Attack

It is possible to classify the attack stages for ransomware attacks as follows (Dwyer, 2021).

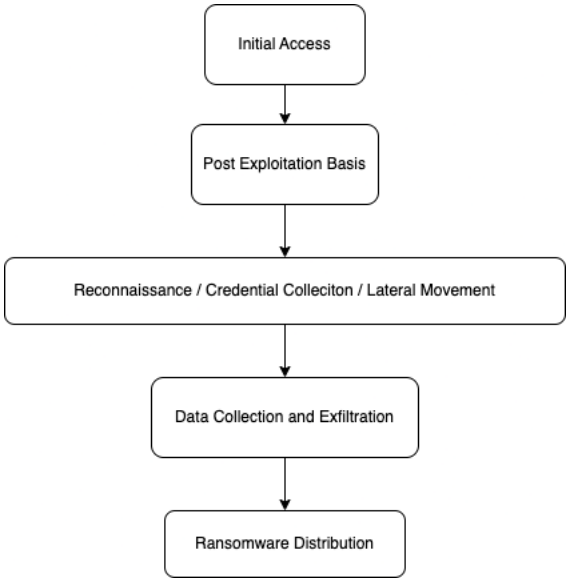

Figure 2. *Ransomware attack stages.*

Initial Access

The initial access stage of a ransomware attack is the phase where attackers gain their first unauthorized entry into a targeted system or network. It is the crucial point where the attackers exploit vulnerabilities, employ social engineering techniques, or use other means to establish their access to the targeted system.

The most common initial access vectors for ransomware attacks are email attachments, flash drives, malicious advertising (malvertising), social media, and SMS (Kapoor et al., 2021). The vast majority of phishing attacks that result in a ransomware event are carried out using known trojans and derivatives such as Bazar, TrickBot, QakBot or Valak, and the next stage is passed after the first access to the system through these pests (Dwyer, 2021).

Post Exploitation

Affected by the initial access vector, a remote access agent (RAT) or malware may be used in the next stage after which interactive access will be gained with an offensive security tool such as Cobalt Strike or Metasploit. It's important to note that the post-exploitation phase can vary depending on the specific attack and the objectives of the attackers. The activities mentioned above are common in many ransomware attacks, but the exact methods and techniques employed can differ.

Reconnaissance / Credential Collection / Lateral Movement

In the third phase of the attack, the attackers focus on understanding the local system and domain they are currently accessing and obtaining credentials to enable lateral movement. Local system discovery is usually done with built-in tools such as net, whoami, and tasklist.

Ransomware attacks use open-source utilities like AdFind to facilitate domain discovery. Such programs basically work as a command line Active Directory query tool, which were created by blending the features of ldapsearch, search.vbs, ldp, or dsquery tools.

The attackers enter the commands they want to run in a file in the script (batch) format they have created to collect information and save the outputs of these commands in different text files and use them to expand their knowledge about the environment. They usually forward this information to command and control servers. Some general discovery commands are shown below.

Table 1. *Adfind commands and outputs.*

Command	Action
adfind.exe -f "(objectcategory=person)" > ad_users.txt	Finds all person objects and saves in ad_users.txt
adfind.exe -f "objectcategory-computer" > ad_computers.txt	Finds all person computers and saves in ad_computer.txt
adfind.exe -sc trustdmp > trustdmp.txt	Finds all person computers and saves in trustdmp.txt
adfind.exe -subnets -f (objectCategory-subnet)> subnets.txt	Finds all subnets and saves in subnets.txt
adfind.exe -gcb -sc trustdmp > trustdmp.txt	Downloads all trust objects and saves in trustdmp.txt
adfind.exe -sc domainlist > domainlist.txt	Lists all domain naming contexts and saves in domainlist.txt
adfind.exe -sc dcmodes > dcmodes.txt	Lists the modes of all Domain Controllers and saves in dcmodes.txt
adfind.exe -sc adinfo > adinfo.txt	Shows Active Directory information and saves in adinfo.txt
adfind.exe -sc dclist > dclist.txt	Lists the names of the Domain Controllers and saves in dclist.txt
adfind.exe -sc computers_pwdnotreqd > computers_pwdnotreqd.txt	Lists the users who are not forced to use passwords and saves pwnotreq.txt

Although credentials can be collected by many access trojans, Mimikatz, ZeroLogon, and PrintNightmare are generally popular and used to obtain credentials that are then used for the rest of the attack (Dwyer, 2021).

In ransomware attacks, after Network and Active Directory reconnaissance movements, lateral movement is usually carried out via a server message block (SMB-Server Message Block) or remote procedure call (Remote Procedure Call) protocols. Additional systems may continue to collect credentials as necessary to obtain domain admin privileges (DFIR Reports, 2020).

Data Collection and Exfiltration

The focus at this stage of the attack is primarily to identify valuable data and extract it. Even if the institutions/organizations or individuals exposed to the attack are performing backup operations before the attack, the attack has achieved its purpose by threatening disclosure of data that should be kept confidential.

At this stage, it often moves laterally to additional systems to determine data what to leak, via SMB, RPC, and remote desktop protocol (RDP). They use the method of accessing and extracting data to collect before leaking it, which they can access via a RDP connection, but

data collection is mostly done over the SMB protocol (Dwyer, 2021). Tools that are often used in good faith by IT teams such as WinSCP and RClone to avoid attracting attention are the most common tools used to leak data.

Ransomware Distribution

When the above-mentioned stages of the attack are completed and this stage is reached, the ransomware is distributed to as many machines as possible to make the incident irreversible, with the aim of putting pressure on the victim to pay the ransom they want. During the ransomware distribution stage, attackers want to infect as many systems as possible to maximize the impact and increase the likelihood of victims paying the ransom.

For this purpose, they target Domain Controllers centrally where they can distribute ransomware quickly, and to load the ransomware, attackers usually use SMB from a share and payload with PsExec, WMIC, RunDll32 or tools like CrackMapExec. It is then distributed by creating scheduled tasks, a service that is provided by the Windows operating system (Dwyer, 2021).

Ransomware Detection and Prevention with Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) plays a crucial role in preventing ransomware attacks through early detection and blocking mechanisms. By leveraging machine learning algorithms, potential threats can be identified and prevented from infiltrating a system. This is particularly beneficial in the case of zero-day vulnerabilities, which are unknown to software vendors and lack patches. AI can swiftly detect these vulnerabilities, enabling security professionals to proactively safeguard against attacks. Furthermore, AI enhances the accuracy and speed of threat detection. Traditional security systems rely on predefined rules and signatures, often overlooking new or unfamiliar threats and necessitating time-consuming processes.

In contrast, AI can analyze vast amounts of data, discern patterns, and identify behaviors that indicate a potential ransomware attack. Consequently, AI outperforms conventional security systems, expediting threat detection and fortifying ransomware prevention efforts. Virtual assistants powered by AI are also instrumental in ransomware prevention. These assistants diligently monitor user behavior, scrutinizing anomalies that may signify an impending ransomware attack. For instance, sudden access to numerous files or attempts to download suspicious software trigger alerts to the security team, who then intervene promptly. Virtual assistants offer real-time guidance and recommendations to users, empowering them with knowledge on how to evade ransomware attacks and fortify their organizations. AI greatly enhances incident response during ransomware attacks. Assessing the extent of damage and

identifying encrypted files can pose challenges for security professionals (Gvozdenko, 2023). AI aids in this process by analyzing the affected system, swiftly pinpointing impacted files. This expedites the determination of the attack's scope, allowing the security team to take appropriate action promptly. Consequently, the time and resources required for responding to a ransomware attack are significantly reduced, minimizing disruption and mitigating damages caused by an attack.

As ransomware behaviors can be detected in different ways a sudden increase in the amount of encrypted data being transferred across a corporate network can be a strong indication of a potential ransomware infection. Encrypted files tend to have more random byte sequences compared to unencrypted files, leading to differences in statistical measures of randomness and information density. Therefore, performing statistical tests can be useful in determining whether a file has been encrypted or not. Following tests can be conducted to identify the encrypted files.

- Chi-square Test
- Entropy
- Arithmetic Mean
- Monte Carlo Value for Pi
- Serial Correlation Coefficient

These tests will most likely provide lots of false positive results. To minimize the occurrence of incorrect positive results in individual statistical tests, a classification machine learning (ML) model was created by security product providers. This ML model is designed to determine whether a file is encrypted or not. It considers various features, including statistical tests and other file characteristics, based on a vast dataset comprising millions of real and synthetic files of diverse types. LightGBM, a machine learning algorithm similar to decision trees, is employed by the model to autonomously discern the disparities between encrypted and unencrypted files.

Methods of Defense Against Ransomware Attacks

Most ransomware attacks have commonalities, and defenses based on these commonalities are vital in repelling attacks. The recommendations conveyed below include countermeasures against ransomware attacks, given what we know about the ransomware attack flow.

Ransomware attack prevention methods are listed as follows,

• Keeping all software and operating systems up to date by applying the latest security patches.

• Using strong passwords and two-factor authentication

• Regularly back up data and store it in an offline location.

• Using reliable antimalware/antivirus software and updating antivirus databases to always have the most current version.

• Disabling macros in Microsoft Office documents.

• Educating employees on cybersecurity best practices.

• Restricting access to sensitive data and systems.

• Monitoring of network traffic to detect suspicious activity

• Filtering of executable applications in email attachments (Mohurle & Patil, 2017).

Limitation of Privileged Access

The number of administrative accounts such as Domain Admin, Enterprise Admin, Schema Admin should be limited to a minimum and unnecessary members of the Domain Admins group, which include Domain Admin accounts, which are frequently used accounts in the execution of business processes, should be removed. Likewise, groups such as Enterprise Admin and Schema Admin, which are critical groups, should be constantly monitored, and actions such as adding, removing, or changing account passwords in these groups should be reported to the security teams as alarms and should be checked immediately.

Ordinary users should not have Local Admin rights, if there are any, they should be removed. For service accounts, Local Admin rights should be kept to a minimum. Since the backbone of ransomware attacks is the process of distributing the malware through the Domain Controller, changes in the Domain Admins and Enterprise Admins groups mentioned above should be monitored continuously and carefully. Since attackers usually prefer working outside of normal working hours, these and similar changes should be immediately reported to system administrators by the Security Operation Center, and every action to be taken by authorized users outside of working hours should be considered suspicious and followed up.End-user devices should be protected from unknown and untrusted executable files by controlling the use of removable devices and by using Application Control security solutions within the organization.

Protection of Privileged Accounts

Privileged accounts have elevated permissions and access to critical systems and data, making them prime targets for attackers. Privileged accounts should be added to the Protected Users Security Group to reduce the risk of internal credential disclosure.

The use of PAM - Privileged Access Management (Privileged Access Management) solutions within the organization is necessary for monitoring privileged accounts and performing access controls. With the use of these applications, stealing privileged accounts and taking unexpected actions within the knowledge of the system administrator can be prevented.

Management of Active Directory Structure

Unnecessary domain trusts between domains should be audited and, if unnecessary, this trust relationship should be removed. The purpose of the trust relationship is to ensure that users authorized by a domain on more than one domain can be used. If this structure is necessary and has to be used, it should be constantly monitored, and possible disasters can be prevented by creating alarms through security applications.

A group policy must be configured to allow the Domain Admin to log on only to domain controllers and to prohibit access to other domain-joined Windows systems. All systems within the organization should be configured to reject authentication attempts via legacy protocols. These authentication methods are insecure as both the username and password information are transmitted over the network and in some cases stored on the machine, so this critical information can be easily accessed by attackers.

Lateral Movement Restriction

Network segmentation policies should be strictly enforced. Access to high-risk resources should only be through specially designated management networks or pre-made jump servers should be used to access these resources.

The Network Configuration of the Institution must be carefully designed, the risk of using applications that allow file sharing and remote management with other machines on the network such as SMB (Server Message Block), RPC (Remote Procedure Call) and RDP (Remote Desktop Protocol) can be avoided through network segmentation. The more subnets are defined, and the communications of the subnets are connected to rules designed according to needs, the more restricted the horizontal movements of the attackers.

Defense Against Phishing Threats

An e-mail software security solution should be used that can detect phishing attack e-mails by checking reputation and content before it reaches the end user which plays a critical role in preventing the infiltration of ransomware into systems. Infiltration of data belonging to employees within the organization should be checked by using CTI tools, domain names that are similar to the corporate domain name should be constantly checked to prevent a possible fraudulent action against the company that could result in a ransomware attack.

Only technical measures fall short of ransomware attacks. It is very important to plan and run Awareness Training within an organization against Phishing attacks to ensure that end users are prepared for this type of attack. Using Phishing simulations with real-life scenarios, the attention and awareness levels of the end users can be increased, and measures taken against possible attacks.

Security Awareness Trainings

Ransomware attacks exploit human vulnerabilities by enticing them to click on malicious links or opening infected email attachments. By raising awareness and providing education on common attack techniques, security awareness training helps employees recognize potential threats and make informed decisions to protect themselves and their organization. Security awareness training helps employees understand the nature and scope of ransomware attacks. They learn about different attack vectors, such as phishing emails, social engineering techniques, and unsafe browsing habits. This knowledge empowers employees to be prepared, identify potential risks, and take appropriate precautions. Training sessions provide employees with practical guidelines and best practices for maintaining good cybersecurity hygiene. This includes tips on creating strong passwords, regular software updates, safe browsing habits, data backup strategies, and incident reporting procedures. These practices can significantly reduce the likelihood of a successful ransomware attack.

Patch Management

The application of security patches shared by manufacturers of the applications used on all devices without wasting time prevents possible security vulnerabilities and reduces the risk. By following the results of vulnerability scans carried out within the organization, using resources with vulnerability management standards such as the National Vulnerability Database (NVD), updates shared by cyber intelligence services, and the attack area is narrowed when updated versions of applications are used within an organization.

Using Antimalware Software

Direct defense against ransomware that will reach the target system is done by antimalware-antivirus software. Today, there is defense software that can perform advanced behavior analysis and a high level of protection can be provided against ransomware. Even if it cannot completely prevent a ransomware attack, it can ensure that the danger is noticed earlier through generated alarms (Furnell & Emm, 2017). In this way, security teams can focus on alarms and take actions to detect and reduce the damage by neutralizing the pest before later stages of the attack can occur.

Conclusion

Ransomware has become a method of cyber-attack that threatens almost every institution and individual, and it has evolved into a form that can cause serious harm to individuals and institutions today, where information is digitized and stored in digital media. Ransomware is a threat that can cause political crises at a high level by exceeding individuals and institutions through the theft of sensitive information, not only as attacks for non-material purposes. In today's cyber world, where new methods and attack techniques are added every day, defense mechanisms and individual cyber awareness needs to develop in the same way and with the same momentum.

Cyber-attacks, including ransomware attacks, have become a part of daily life in our age. Every individual needs to increase their cyber literacy, always be careful against Phishing attacks, where the first access is the most intense in malicious attacks and implement basic security controls. Apart from the measures taken individually, inspecting institutions to which an individual entrusts their data as data controllers and subjecting them to sanctions is a very important factor in making relevant investments. In this context, institutions at all levels should provide cyber security awareness in the education of each employee, increasing the time and budgets that are allocated to develop their cyber security infrastructures, inspections and exercises carried out before attacks occur, and preparing for a possible ransomware attack and disaster recovery in case of a successful attack. The readiness and viability of attach scenarios should be monitored continuously in a disciplined manner, and it is expected that the content and forms of the sanctions will be deterrent that will enable the institutions to take the necessary measures.

Acknowledgment: This work was supported in part by İHS Kurumsal Teknoloji A.Ş.

Ethics Committee Approval: Authors declared that this study does not require ethics committee approval.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- F.G.; Data Acquisition- F.G.; Data Analysis/Interpretation- F.G., A.Ş.; Drafting Manuscript- F.G.; Critical Revision of Manuscript- A.Ş.; Final Approval and Accountability- A.Ş., F.G.; Material and Technical Support- A.Ş., F.G.; Supervision- A.Ş.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Askarifar, S., Rahman, N. A. A., & Osman, H. (2018). A review of latest wannacry ransomware: Actions and preventions. *J. Eng. Sci. Technol*, *13*, 24-33.

DFIR Report, Reports. (2020, August 31). *NetWalker Ransomware in 1 Hour*. https://thedfirreport.com/2020/08/31/netwalker-ransomware-in-1-hour/

Dwyer, J. (2021, November 30). *Understanding the Adversary: How Ransomware Attacks Happen.* https://securityintelligence.com/posts/how-ransomware-attacks-happen/

Fernando, D. W., Komninos, N., & Chen, T. (2020). A study on the evolution of ransomware detection using machine learning and deep learning techniques. *IoT*, *1*(2), 551-604.

Furnell, S., & Emm, D. (2017). The ABC of ransomware protection. *Computer Fraud & Security*, *2017*(10), 5-11.

Gómez-Hernández, J. A., Sánchez-Fernández, R., & García-Teodoro, P. (2022). Inhibiting crypto-ransomware on windows platforms through a honeyfile-based approach with R-Locker. *IET Information Security*, *16*(1), 64-74.

Gvozdenko, A. (2023, April). *How AI will Revolutionize Ransomware Prevention.* https://www.cynergy.app/cyber-research/how-ai-will-revolutionize-ransomware-prevention/4765/#:~:text=AI%2C%20on%20the%20other%20hand,overall%20effectiveness%20of%20ransomware%20prevention

Health Sector Cybersecurity Coordination Center (2020, September) U.S. Department of Health and Human Services,

https://www.hhs.gov/sites/default/files/netwalker.pdf

Kapoor, A., Gupta, A., Gupta, R., Tanwar, S., Sharma, G., & Davidson, I. E. (2021). Ransomware detection, avoidance, and mitigation scheme: a review and future directions. *Sustainability*, *14*(1), 8.

Maurya A.K, Kumar N., Agrawal A., Khan R.A(2017). Ransomware: Evolution, Target and Safety Measures. International Journal of Computer Sciences and Engineering, Volume-6,Issue-1.

Mohurle, S., & Patil, M. (2017). A brief study of wannacry threat: Ransomware attack 2017. *International Journal of Advanced Research in Computer Science*, 8(5), 1938-1940.

Richardson, R., & North, M. M. (2017). Ransomware: Evolution, mitigation and prevention. *International Management Review*, *13*(1), 10.

Zavarsky, P., & Lindskog, D. (2016). Experimental analysis of ransomware on windows and android platforms: Evolution and characterization. *Procedia Computer Science*, *94*, 465-472.

Journal of Data Applications
Issue: 1, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

A Comparative Assessment of Frequentist Forecasting Models: Evidence from the S&P 500 Pharmaceuticals Index

Christian MUNEZA[1], Asad Ul İslam KHAN[1], Waqar BADSHAH[2]

ABSTRACT

This paper compares three forecasting methods, the autoregressive integrated moving average (ARIMA), generalized autoregressive conditional heteroscedasticity (GARCH), and neural network autoregression (NNAR) methods, using the S&P 500 Pharmaceuticals Index. The objective is to identify the most accurate model based on the mean average forecasting error (MAFE). The results consistently show the NNAR model to outperform ARIMA and GARCH and to exhibit a significantly lower MAFE. The existing literature presents conflicting findings on forecasting model accuracy for stock indexes. While studies have explored various models, no universally applicable model exists. Therefore, a comparative analysis is crucial. The methodology includes data collection and cleaning, exploratory analysis, and model building. The daily closing prices of pharmaceutical stocks from the S&P 500 serve as the dataset. The exploratory analysis reveals an upward trend and increasing heteroscedasticity in the pharmaceuticals index, with the unit root tests confirming non-stationarity. To address this, the dataset has been transformed into stationary returns using logarithmic and differencing techniques. Model building involves splitting the dataset into training and test sets. The training set determines the best-fit models for each method. The models are then compared using MAFE on the test set, with the model possessing the lowest MAFE being considered the best. The findings provide insights into model accuracy for pharmaceutical industry indexes, aiding investor predictions, with the comparative analysis emphasizing tailored forecasting models for specific indexes and datasets.

Keywords: Forecasting Accuracy, Pharmaceutical Industry Indexes, S&P 500, NNAR, Comparative Analysis

DOI: 10.26650/JODA.1312382

[1]Ibn Haldun University, Faculty of Humanities and Social Sciences, Department of Economics, Istanbul, Turkiye
[2]Istanbul University, Faculty of Economics, Department of Management Information Systems, Istanbul, Turkiye

ORCID: C.M. 0000-0001-6419-1940;
A.U.İ.K. 0000-0002-5131-577X;
W.B. 0000-0001-5009-8745

Corresponding author:
Asad Ul İslam KHAN,
Ibn Haldun University, Faculty of Humanities and Social Sciences, Department of Economics, Istanbul, Turkiye
E-mail: asad.khan@ihu.edu.tr

Submitted: 11.06.2023
Revision Requested: 12.06.2023
Last Revision Received: 14.07.2023
Accepted: 17.07.2023

Citation: Muneza, C., Khan, A.U.İ., Badshah, W. (2023). A comparative assessment of frequentist forecasting models: evidence from the s&p 500 pharmaceuticals index. *Journal of Data Applications, 1*, 83-94. https://doi.org/10.26650/JODA.1312382

Introduction

For the last two decades and especially after the 2008 global financial crisis, financial markets have grown to represent a sizable share of national incomes (Ross, 2021). To keep up with the trend and make sense of vast amounts of available data, financial economists have devised methods ranging from algorithmic trading to forecasting methods in order to anticipate the market and maximize profits (Organisation for Economic Cooperation and Development (OECD, 2021). While these methods have grown in relevance as toolkits investors use to gauge evidence-backed predictors of markets, they are still far from perfect, with the case in point being their widespread association with the 2008 financial crisis due to how overreliance on them blinded many from seeing what was about to come. The past failures of financial forecasting methods and the need to find ways they can be more accurate are what fuel this academic exploration.

Despite the increase in academic research on financial forecasting models, most leading academic papers have focused on forecasting aggregate market indexes, such as the S&P 500 (Niaki & Hoseinzade, 2013), Dow Jones (Nasr Ben & Lux, 2016), Nasdaq Composite (Sunarya, 2019), Financial Times Stock Exchange (FTSE 100; Niu et al., 2020), National Stock Exchange of India (NIFTY 50; Mahajan et al., 2022), and the Shanghai Stock Exchange (SSE) Composite (Lin, 2018) indices, among others. This has contributed to the increased understanding of the behavior of the market as an aggregate, but the current scholarship remains insufficient at helping understand the behavior of particular industries. Though the gap is present in many industries, it is at its greatest in pharmaceutical and healthcare-related stock indexes (Harris, 2018).

After identifying the current gap in the literature as it relates to forecasting industry-specific indexes, the study aims to contribute to addressing the gap by testing forecasting accuracies on the pharmaceutical industry indexes of the S&P 500. This paper has significance for both academics and policies. Academic-wise, health stocks remain the least studied when compared to other stocks. Despite the huge data potential, the field as a whole remains the least explored, especially in terms of the use of data analytics (Harris, 2018). The significances policy-wise are also paramount. With the COVID-19 pandemic, climate change, increasing aging population, rise of noncommunicable diseases, research on longevity and better treatments, and the other host of global disease burdens caused by urbanization and lifestyle changes, the public sector will clearly face a need to invest more in pharmaceutical R&D and improved healthcare solutions (Simpkin et al., 2019). The public sector alone definitely won't meet the funding needed to address the emerging global disease burden, hence the clear need for increasing the role of the private sector in healthcare financing and funding (Simpkin et al., 2019). Both the individual and institutional investors that are currently investing in such

industries as oil and gas, infotech, automobiles and manufacturing, real estate, and others will need to familiarize themselves with and increase their share in healthcare investments. For this to happen, more study is needed to show investors that investment in healthcare is not only safe but also profitable.

The next sections of the study will be divided as follows: Literature Review, Methodology, Results and Discussion, and Conclusion and Recommendation. The Literature Review will provide a window into the works and findings of preliminary studies on the subject matter, while the Methodology section will detail the processes this study follows, such as data collection, data examination, and forecasting. The Results and Discussion section will present the outcome of each stage and provide commentary and interpretation, and the Conclusion and Recommendation section will provide the final takeaways from the study.

Literature Review

Sizable academic research is found to have compared the accuracy of forecasting methods on several indexes, usually with differing conclusions. Mahajan and Thakan's (2022) study "Modeling and Forecasting the Volatility of NIFTY 50 Using GARCH and RNN Models" sought to evaluate the forecasting accuracy of the generalized autoregressive conditional heteroscedasticity (GARCH) and neural network autoregression (NNAR) families of models. Motivated by the great volatilities within the Indian stock market such as the 2000s tech advancements that led to the boom of the Indian stock market and the crash that followed, they analyzed NIFTY 50 to identify the behavior of the Indian market's volatility and then evaluated the forecasting abilities of the above-mentioned models. They concluded that NIFTY 50 volatility is asymmetric and concluded the exponential GARCH, or EGARCH (1,1), and threshold autoregressive conditional heteroscedasticity, or TARCH (1,1), models to be the best at forecasting.

In a hybrid stock price index forecasting model based on variational mode decomposition (VMD) and long short-term memory (LSTM) network, Niu and Xu (2020) introduced a new hybrid model using VMD-LSTM to study the FTSE 100 Index. Their study is advantageous in that VMD decomposes the original complex series into a limited number of series with simpler fluctuation modes, thus overcoming the shortcomings of mode mixing found in the typically used empirical decomposition method (EDM). Another advantage is that LSTM filters out the critical previous information, thus making it better for financial time series forecasting than traditional recurrent neural networks. They concluded their VMD-LSTM hybrid model to be a better forecaster than single models.

Sunarya's (2019) study "Modelling and Forecasting Stock Market Volatility of Nasdaq Composite Index" evaluates the best models for both autoregressive integrated moving average

(ARIMA) and GARCH regarding the Nasdaq returns from March 1971-April 2019. They employed the standard data analytics methodology of data cleaning, manipulation, and model estimation. For ARIMA, they find ARIMA (8,0,6) model to be the best due to its lowest Akaike information criterion (AIC) value. They also determine the ARIMA-GARCH model combination and the best model for this emerges as ARIMA (8,0,6)-EGARCH (1,1). while modelling and forecasting the stock market volatility of SSE Composite Index using GARCH models, Lin (2018) examined the econometric features of the Shangai Stock Exchange (SSE) Composite Index and compared the forecasting ability of the GARCH family of models. Lin's results found SSE to have significant properties regarding time variance and clustering due to rapid information dissemination, fast capital flow, and undulating prices. Due to these phenomena, Lin's forecasting experiments concluded the EGARCH (1,1) model to outperform the GARCH (1,1) and TARCH (1,1) models.

Yadav and Sharma's (2018) study "Statistical Analysis and Forecasting Models for Stock Market" evaluated the accuracy of ARIMA, exponential smoothing, naive, seasonal naive, neural network, mean, and BoxCox transformation forecasting methods to predict the Bombay Stock Exchange (BSE) SENSEX opening, high, low, and closing prices from January 1997-January 2016. Their methodology utilized standard data analytical techniques for data cleaning and manipulation, as well as model estimation. They set the mean error as the accuracy criteria, and the exponential smoothing and neural network models emerged as the best ones. Islam and Nguyen's (2020) study "Comparison of Financial Models for Stock Price Prediction" evaluated the accuracy of ARIMA, artificial neural network (ANN), and stochastic process-geometric Brownian motion for forecasting the S&P 500 using daily adjusted closing prices from April 1, 2015-December 31, 2019. They set the standardized residuals as the accuracy criterion, with the ARIMA and stochastic process-geometric Brownian motion models emerging as the best ones for predicting short-term next-day prices. Their findings agree with those from Merh et al. (2010) on ARIMA predicting stock prices better than ANN but contradict those from Khashei and Bijari (2010), who had concluded ARIMA to be no better than ANN.

Sharaff and Choudhary's (2018) study "Comparative Analysis of Various Stock Prediction Techniques" evaluated the accuracy of ARIMA, ANN, Holt-Winters, and NNARs for forecasting the S&P Bombay Stock Exchange using monthly closing prices from 2007-2012. Their methodology also involved the standard data analytics process of visualization, stationarizing, finding optimal parameters for models, and making predictions. They set the mean absolute percentage error (MAPE) as the accuracy criterion, with ANN emerging as the best model. Niaki and Hoseinzade's study (2013) "Forecasting S&P 500 Index Using Artificial Neural Networks and Design of Experiments" compared the predictive ability of ANN to traditional logit models. Their study included 27 financial and economic variables that tend to

influence the S&P500 movements and compared ANN and logit in terms of how they respond to these variables. They concluded ANN to be better at integrating influential variables and forecasting the index compared to the traditional logit model. Overall, the literature suggests that no decisive one-size-fits-all model exists that can be applied to stock predictions and that specific comparative scrutiny should be applied to each different index and dataset.

Methodology

The paper uses the ARIMA, GARCH, and NNAR models, and its methodology consists of four steps: data collection, data cleaning, exploratory analysis, and model building and forecasting in order to arrive at the best model. This section lays out the mathematical notions for the models, as well as the commentary on each of the methodological steps.

ARIMA Model

The ARIMA model is based on the following equation:

$$yt = c + \emptyset 1yt - 1 + \emptyset 2yt - 2 \ldots + \emptyset pyt - p + \theta 1\varepsilon t - 1 + \theta 2\varepsilon t - 2 \ldots \theta q\varepsilon t - q + \mathrm{e}t \quad \textit{Eq. 1}$$

where *yt* is the variable explained at time *t*; *c* is the constant; $\emptyset$(*i*=1,2..*p*) and θ(*j*=1,2,…*q*) are the model parameters; *p* and *q* are integers with *p* representing the autoregressive (AR) part and *q* representing the moving average (MA) part; and e*t* is the error term.

GARCH Model

The GARCH model is derived from the following equation:

$$\sigma t = \sqrt{\omega + \sum_{i=1}^{p} \alpha^2 t - i + \sum_{i=1}^{q} \beta\sigma^2 t - i} \quad \textit{Eq. 2}$$

where *σt* is the conditional standard deviation and its past values σt1,2,… are fed back into the process. $\sum_{i=1}^{p} \alpha^2 t - i$ represents the AR part of the model, and $\sum_{i=1}^{q} \beta\sigma^2 t - i$ represents the conditional heteroscedastic part of the model.

NNAR Model

The NNAR model comes from Eq. 3 as follows:

$$yt = w_0 + \sum_{j=1}^{q} w_j . g\left(w_{0,j} + \sum_{i=1}^{p} w_{ij} . y_{t-1}\right) + \mathrm{e}_t \quad \textit{Eq. 3}$$

where w_j ($j = 0,1,2,3\ldots, q$) and w_{ij} ($i = 0, 1, 2\ldots, p; j = 0, 1, 2, \ldots, q$) are the connection weights or model parameters, p is the number of input nodes, and q is the number of hidden nodes.

Methodology Steps

i. Data Collection. The paper collected daily Pharmaceutical S&P500 closing prices from January 4, 2010-December 31, 2019 from Market Watch.

ii. Data Cleaning. This process involves transforming the dataset into a time-series format readable by the program R by checking for the presence or absence of missing values and labeling columns appropriately.

iii. Exploratory Data Analysis. This process involves first plotting the dataset to see elements such as trends, seasonality, heteroscedasticity, and stationarity, then running stationarity tests, and finally applying logarithm and differencing techniques to transform the closing prices into stationary returns that will be used to build the forecasting models. The resultant dataset will be called "pharmareturns" and will be used in the following modeling stages based on the following formula:

$$z_i = y_t - y_{t-1} \qquad \textit{Eq. 4}$$

where z_i represents the returns, y_t represents the closing price at time t, and y_{t-1} represents the closing price at time t - 1.

iv. Model Building and Forecasting: This step involves the process of determining the best fit model for each of the ARIMA, GARCH, and NNAR models. The dataset on the returns (adjusted closing prices) will be divided into a training set constituting 70% of the dataset and a test set constituting the remaining 30%. The best fit model for each forecasting method will be found using the training set. Then the best fit model will be applied to the forecast, with its results compared to the test set. The model with the lowest mean average forecasting error (*MAFE*) will emerge as the best forecasting method.

Results

Exploratory Data Analysis

The dataset used in the study contains the S&P 500 Pharmaceuticals Industry daily closing prices from January 4, 2010-December 31, 2019. The S&P 500 Pharmaceuticals Index was used for a number of reasons, such as its focus on industry leaders, relevance in the investment space, and benchmark comparison. Specifically, the 20 largest pharmaceutical companies tracked by the S&P 500 account for 78.7% of the global prescription market and are therefore a representative sample for studying global trends in pharmaceutical stock indexes (Mikulic,

2022). After the processes of cleaning and transforming the data into a time series format, the dataset was analyzed for trend, seasonality, heteroscedasticity, and stationarity.

Step 1: Plotting the Dataset

The data have been plotted to observe its long-run behavior (Figure 1). As can be seen from Figure 1, a clear upward trend is present, as well as heteroscedasticity that continuously increases over time.

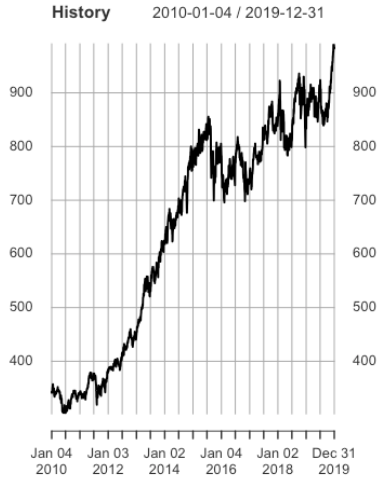

Figure 1: *S&P 500 Pharmaceutical Closing Prices from 2010 to 2019.*

Step 2: Unit Root Tests

Next, unit root tests were run to determine the stationarity of the data (see Table 1). A *p*-value of 0.5053 was obtained for the augmented Dickey-Fuller (ADF) test, indicating the null test is rejected, and the data are concluded to be nonstationary. For both the Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) level stationarity and trend stationarity tests, *p*-values of 0.01 were obtained, indicating the null test for KPSS to again be rejected at a 5% significance and the dataset to be similarly concluded as nonstationary. After this analysis, the need is seen to exist for transforming the dataset into stationary data before applying the forecasting models to it.

Table 1. *Unit Root Test Results.*

TEST	Null Hypothesis	P-value
ADF Test	Unit Root	0.5053
KPSS Test	Stationary	0.01

Transforming the Dataset to Stationary Data

Due to the presence of heteroscedasticity and unit roots, logarithms will first be applied to the data to contain the heteroscedasticity and then it will be differentiated to remove the trend. The resulting dataset will represent the historical returns and is expected to exhibit white noise behavior. The resultant dataset will henceforth be referred to as "pharmareturns" and will be used for the rest of the modeling.

Figure 2: *S&P 500 Pharmaceutical Returns from 2010 to 2019.*

The graphs in Figure 2 compare the pretreatment and posttreatment datasets. While the pretreatment dataset exhibits unit roots, the posttreatment dataset exhibits white noise, which is in agreement with the financial literature on the historical nature of returns. The ADF and KPSS tests are rerun on the treated dataset to check for stationarity (see Table 2).

Table 2. *Unit Root Test Results on Returns.*

TEST	Null Hypothesis	*p*-value
ADF Test	Unit Root	0.01
KPSS Test	Stationary	0.1

The ADF test confirms the expected stationarity of the pharmareturns dataset at a 5% significance of p = 0.01. The KPSS tests also resulted in $p = 0.1$, thus the null hypothesis is no longer rejected and the pharmareturns dataset is concluded to indeed be stationary.

Model Building and Forecasting

This section determines the best model for each method (i.e., the best ARIMA model, best GARCH model, best NNAR model). The pharmareturns dataset is first divided into a training set containing 70% of the data and a test set containing 30% of the data. Once the best model for each method has been determined, the best forecasting method will then be identified.

Determining the Best Models

ARIMA

Different *p* and *q* levels for the ARIMA(*p*, 0, *q*) model were experimented with in order to see which model has the least error based on the AIC value. This stage of the study experiments with *p* levels between 0-5 and *q* levels between 0-5. Table 3 lists the AIC values for some of the models.

Table 3. *ARIMA Models Estimation Results.*

ARIMA Model	Mean Specification	AIC Value
ARIMA (2, 0, 2)	non-zero mean	-11,182.78
ARIMA (0, 0, 0)	non-zero mean	-11,182.66
ARIMA (1, 0, 0)	non-zero mean	-11,182.94
ARIMA (0, 0, 1)	non-zero mean	-11,183.19
ARIMA (0, 0, 0	zero mean	-11,181.15
ARIMA (1, 0, 1)	non-zero mean	-11,183.89
ARIMA (2, 0, 1)*	non-zero mean*	-11,184.25*
ARIMA (2, 0, 0)	non-zero mean	-11,181.56
ARIMA (3, 0, 1)	non-zero mean	-11,181.05

* best model

As Table 3 shows, the ARIMA (2, 0, 1) model has been identified as the best model due to having the lowest AIC value. The models were applied to the forecast and compared with the forecasts from the test set, with Table 4 showing ARIMA (2, 0, 1) to have been determined as the best forecasting ARIMA model with the lowest *MAFE* value of 0.0587138.

Table 4. *ARIMA Model Selection.*

Set	*ME*	*RMSE*	*MAE*	*MFE*	*MAFE*
Training set	2.016731×10^{-6}	0.007305894	0.06173646	0.0654321	0.6745051
Test Set	-8.167452×10^{-5}	0.009037684	0.006500222	0.05746563	0.0587138

ME = margin of error; RMSE = root mean square error; MAE = mean absolute error; MFE = maximum favorable excursion

GARCH

Both the mean equation and the variance equation were found for the pharmareturns dataset. Table 5 shows the GARCH (1, 1) model with the mean model used being ARIMA (2, 0, 1). The model has a *MAFE* value of 0.000328.

Table 5. *GARCH Model Selection.*

GARCH Best Model	AIC	*MAFE*
GARCH Model (1, 1) Mean Model (2, 0, 1)	-6.5107	0.000328

NNAR

Similarly, different levels for *p* and *q* were experimented with for the NNAR model, with *p* representing the number of lagged values and *q* representing the number of hidden layers. Table 6 shows the NNAR (10, 6) model was obtained as the one with the lowest *MAFE*.

Table 6. *NNAR Model Selection.*

NNAR Best Model	*MAFE*
NNAR(10, 6)	8.147859×10^{-9}

The NNAR(10, 6) model was applied to the forecast and compared to the forecasts from the test set, and we obtain a mean average forecasting error of 8.147859e-09.

Best Forecasting Model Selection

This section determines the best forecasting model for the S&P 500 Pharmaceutical Index based on *MAFE*, with the results shown in Table 7.

Table 7. *Best Model Selection.*

Forecasting Model	*MAFE*	*RMSE*	*MAE*
ARIMA(2, 0, 1)	0.0587138	0.009037684	0.006500222
GARCH(2, 1) & (1, 1)	0.000328	0.025586	0.0185646
NNAR(10, 6)	8.147859×10^{-9}	7.951×10^{-5}*	6.546×10^{-5}*
*best model			

Table 7 shows the NNAR model to have conclusively emerged as the best model for the dataset, possessing a significantly lower *MAFE* than the other two models and also based on *RMSE* and *MAE*. The final conclusion is based on *MAFE* due to its wide acceptance in the academic literature as a measure of forecasting accuracy (Tofallis, 2017).

Conclusion and Recommendations

The aim of the study has been to compare some of the most commonly used forecasting models (i.e., ARIMA, GARCH, and NNAR) and to determine the best one regarding a dataset derived from the S&P 500 Pharmaceuticals Index. *MAFE* was used as the accuracy metric for determining the best forecasting model. The study involved rigorous processes for data cleaning, exploratory data analysis, model building, and best model selection, with the NNAR

model being determined as the best one due to it having the lowest *MAFE* value, which is widely used as a measure of forecasting accuracy. The study recommends NNAR be used when forecasting the S&P 500 Pharmaceuticals index as it forecasts are more reliable compared to the other models examined in this study.

The study has been limited to normal forecasting methods and did not leverage machine learning tools such as supervised and unsupervised learning or more robust cross-validation techniques that use multiple training and testing datasets. As such, future studies can use machine learning tools and robust cross-validation techniques in order to obtain results with higher confidence. Moreover, future research should emphasize on exploring less-studied indexes, such as those from emerging and developing countries.

Ethics Committee Approval: Authors declared that this study does not require ethics committee approval.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- A.U.İ.K., W.B., C.M.; Data Acquisition- C.M.; Data Analysis/Interpretation- A.U.İ.K., W.B., C.M.; Drafting Manuscript- C.M.; Critical Revision of Manuscript- A.U.İ.K., W.B.; Final Approval and Accountability- A.U.İ.K., W.B., C.M.; Material and Technical Support- C.M.; Supervision- A.U.İ.K., W.B.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Harris J. (2018). *The opportunities and challenges of data analytics in health care*. Brookings. https://www.brookings.edu/research/the-opportunities-and-challenges-of-data-analytics-in-health-care/

Islam, M. R., & Nguyen, N. (2020). Comparison of Financial Models for Stock Price Prediction. *Journal of Risk and Financial Management, 13*(8), 181. https://doi.org/10.3390/JRFM13080181

Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with Applications*, *37*(1), 479–489. https://doi.org/10.1016/J.ESWA.2009.05.044

Lin, Z. (2018). Modelling and forecasting the stock market volatility of SSE Composite Index using GARCH models. *Future Generation Computer Systems*, *79*, 960–972. https://doi.org/10.1016/J.FUTURE.2017.08.033

Mahajan, V., Thakan, S., & Malik, A. (2022). Modeling and Forecasting the Volatility of NIFTY 50 Using GARCH and RNN Models. *Economies, 10*(5), 102. https://doi.org/10.3390/ECONOMIES10050102

Merh, N., Prakash Saxena, V., & Raj Pardasani, K. (2010). *A comparison between Hybrid Approaches of ANN and ARIMA for Indian Stock Trend Forecasting Computational Aspects of I-Function View project*. https://www.researchgate.net/publication/45602108

Mikulic, M. (2022). *Market share top pharma companies Rx drugs sales globally 2026 | Statista*. Statistica. https://www.statista.com/statistics/309425/prescription-drugs-market-shares-by-top-companies-globally/

Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International*, *9*(1). https://doi.org/10.1186/2251-712X-9-1

Niu, H., Xu, K., & Wang, W. (2020). A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network. *Applied Intelligence*, *50*(12), 4296–4309. https://doi.org/10.1007/S10489-020-01814-0

OECD. (2021). Sustainable and Resilient Finance OECD Business and Finance Outlook 2020 9HSTCQE*diefgj+ Artificial Intelligence, Machine Learning and Big Data in Finance Opportunities, Challenges and Implications for Policy Makers. *OECD Journal*.

Sean Ross. (2021). *What Percentage of the Global Economy Is the Financial Services Sector?* Investopedia. https://www.investopedia.com/ask/answers/030515/what-percentage-global-economy-comprised-financial-services-sector.asp

Sharaff, A., & Choudhary, M. (2018). Comparative Analysis of Various Stock Prediction Techniques. *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, 735–738. https://doi.org/10.1109/ICOEI.2018.8553825

Simpkin, V., Namubiru-Mwaura, E., Clarke, L., & Mossialos, E. (2019). Investing in health R&D: where we are, what limits us, and how to make progress in Africa. *BMJ Global Health*, *4*(2). https://doi.org/10.1136/BMJGH-2018-001047

Sunarya, I. W. (2019). Modelling and Forecasting Stock Market Volatility of Nasdaq Composite Index. *Economic and Accounting Journal*, *2*(3), 181–189. https://doi.org/10.32493/EAJ.V2I3.Y2019.P181-189

Tofallis, C. (2017). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, *66*(8), 1352–1362. https://doi.org/10.1057/JORS.2014.103

Yadav, S., & Sharma, K. P. (2018). Statistical Analysis and Forecasting Models for Stock Market. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 117–121. https://doi.org/10.1109/ICSCCC.2018.8703324

DESCRIPTION

Journal of Data Applications is open access, internationally refereed, scientific journal published electronically twice a year (in April and October) within the body of Istanbul University Faculty of Economics Management Information Systems. The journal follows the double-blind peer-review process. The publication language of the journal is English. No processing fee or publication fee is requested for the articles sent to the journal.

AIM AND SCOPE

In parallel with the developing and widespread use of information and communication technologies, the amount of data produced daily is increasing. The Journal of Data Applications aims to contribute to the development of applied data science studies, which aim to obtain meaningful information from data and reveal hidden patterns and patterns in data, thus contributing to the development of studies in this field.

Journal of Data Applications accepts computational and applied scientific studies such as original research, compilation, and report on the field of data collection, storage, transmission, preprocessing, analysis, visualization, and interpretation of data, especially statistics, artificial intelligence, machine learning, deep learning, and data mining applications. In this context, the Journal of Data Applications has no discipline and application restrictions.

Scope of the Journal of Data Applications collapses from all disciplines in various fields such as information retrieval and extraction, clustering, predicting and forecasting applications, decision support systems, recommendation systems, image, sound and pattern recognition, and processing, natural language processing, signal processing, computer vision, big data processing, time series analysis includes various application studies from all disciplines in areas such as sentiment analysis, social media analysis, fraud and anomaly detection.

EDITORIAL POLICIES AND PEER REVIEW PROCESS

Publication Policy

The subjects covered in the manuscripts submitted to the journal for publication must be in accordance with the aim and scope of the journal. The journal gives priority to original research papers submitted for publication.

General Principles

Only those manuscripts approved by its every individual author and that were not published before in or sent to another journal, are accepted for evaluation.

Submitted manuscripts that pass preliminary control are scanned for plagiarism using iThenticate software. After plagiarism check, the eligible ones are evaluated by editor-in-chief for their originality, methodology, the importance of the subject covered and compliance with the journal scope.

The editor hands over the papers matching the formal rules to at least two national/international referees for evaluation and gives green light for publication upon modification by the authors in

accordance with the referees' claims. Changing the name of an author (omission, addition or order) in papers submitted to the journal requires written permission of all declared authors. Refused manuscripts and graphics are not returned to the author.

Open Access Statement

The journal is an open access journal and all content is freely available without charge to the user or his/her institution. Except for commercial purposes, users are allowed to read, download, copy, print, search, or link to the full texts of the articles in this journal without asking prior permission from the publisher or the author. This is in accordance with the BOAI definition of open access.

The open access articles in the journal are licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

Copyright Notice

Authors publishing with the journal retain the copyright to their work licensed under the Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0) (https://creativecommons.org/licenses/by-nc/4.0/) and grant the Publisher non-exclusive commercial right to publish the work. CC BY-NC 4.0 license permits unrestricted, non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article Processing Charge

All expenses of the journal are covered by the Istanbul University. Processing and publication are free of charge with the journal.

There is no article processing charges or submission fees for any submitted or accepted articles.

Peer Review Process

Only those manuscripts approved by its every individual author and that were not published before in or sent to another journal, are accepted for evaluation.

Submitted manuscripts that pass preliminary control are scanned for plagiarism using iThenticate software. After plagiarism check, the eligible ones are evaluated by Editor-in-Chief for their originality, methodology, the importance of the subject covered and compliance with the journal scope. Editor-in-Chief evaluates manuscripts for their scientific content without regard to ethnic origin, gender, citizenship, religious belief or political philosophy of the authors and

ensures a fair double-blind peer review of the selected manuscripts.

The selected manuscripts are sent to at least two national/international referees for evaluation and publication decision is given by Editor-in-Chief upon modification by the authors in accordance with the referees' claims.

Editor-in-Chief does not allow any conflicts of interest between the authors, editors and reviewers and is responsible for final decision for publication of the manuscripts in the journal.

Reviewers' judgments must be objective. Reviewers' comments on the following aspects are expected while conducting the review.

- Does the manuscript contain new and significant information?

- Does the abstract clearly and accurately describe the content of the manuscript?

- Is the problem significant and concisely stated?

- Are the methods described comprehensively?

- Are the interpretations and consclusions justified by the results?

- Is adequate references made to other Works in the field?

- Is the language acceptable?

Reviewers must ensure that all the information related to submitted manuscripts is kept as confidential and must report to the editor if they are aware of copyright

infringement and plagiarism on the author's side.

A reviewer who feels unqualified to review the topic of a manuscript or knows that its prompt review will be impossible should notify the editor and excuse himself from the review process.

The editor informs the reviewers that the manuscripts are confidential information and that this is a privileged interaction. The reviewers and editorial board cannot discuss the manuscripts with other persons. The anonymity of the referees is important.

PUBLICATION ETHICS AND PUBLICATION MALPRACTICE STATEMENT

Journal of Data Applications is committed to upholding the highest standards of publication ethics and pays regard to Principles of Transparency and Best Practice in Scholarly Publishing published by the Committee on Publication Ethics (COPE), the Directory of Open Access Journals (DOAJ), to access the Open Access Scholarly Publishers Association (OASPA), and the World Association of Medical Editors (WAME) on https://publicationethics.org/resources/guidelines-new/principles-transparency-and-best-practice-scholarly-publishing

All parties involved in the publishing process (Editors, Reviewers, Authors and Publisher) are expected to agree on the following ethical principles.

All submissions must be original, unpublished (including as full text in conference proceedings), and not under the review of any other publication synchronously. Each manuscript is reviewed by one of the editors and at least two referees under double-blind peer review process. Plagiarism, duplication, fraud authorship/denied authorship, research/data fabrication, salami slicing/salami publication, breaching of copyrights, prevailing conflict of interest are unnethical behaviors.

All manuscripts not in accordance with the accepted ethical standards will be removed from the publication. This also contains any possible malpractice discovered after the publication. In accordance with the code of conduct we will report any cases of suspected plagiarism or duplicate publishing.

Research Ethics

Journal of Data Applications adheres to the highest standards in research ethics and follows the principles of international research ethics as defined below. The authors are responsible for the compliance of the manuscripts with the ethical rules.

- Principles of integrity, quality and transparency should be sustained in designing the research, reviewing the design and conducting the research.
- The research team and participants should be fully informed about the aim, methods, possible uses and requirements of the research and risks of participation in research.
- The confidentiality of the information provided by the research participants and the confidentiality of the respondents should be ensured. The research should be designed to protect the autonomy and dignity of the participants.
- Research participants should participate in the research voluntarily, not under any coercion.
- Any possible harm to participants must be avoided. The research should be planned in such a way that the participants are not at risk.
- The independence of research must be clear; and any conflict of interest must be disclosed.
- In experimental studies with human subjects, written informed consent of the participants who decide to participate in the research must be obtained. In the case of children and those under wardship or with confirmed insanity, legal custodian's assent must be obtained.
- If the study is to be carried out in any institution or organization, approval must be obtained from this institution or organization.
- In studies with human subject, it must be noted in the method's section of the manuscript that the informed consent of the participants and ethics committee approval from the institution where the study has been conducted have been obtained.

Author Responsibilities

It is authors' responsibility to ensure that the article is in accordance with scientific and ethical standards and rules. And authors must ensure that submitted work is original. They must certify that the manuscript has not previously been published elsewhere or is not currently being considered for publication elsewhere, in any language. Applicable copyright laws and conventions must be followed. Copyright material (e.g. tables, figures or extensive quotations) must be reproduced only with appropriate permission and acknowledgement. Any work or words of other authors, contributors, or sources must be appropriately credited and referenced.

All the authors of a submitted manuscript must have direct scientific and academic contribution to the manuscript. The author(s) of the original research articles is defined as a person who is significantly involved in "conceptualization and design of the study", "collecting the data", "analyzing the data", "writing the manuscript", "reviewing the manuscript with a critical perspective" and "planning/ conducting the study of the manuscript and/or revising it". Fund raising, data collection or

supervision of the research group are not sufficient roles to be accepted as an author. The author(s) must meet all these criteria described above. The order of names in the author list of an article must be a co-decision and it must be indicated in the **Copyright Agreement Form**.

The individuals who do not meet the authorship criteria but contributed to the study must take place in the acknowledgement section. Individuals providing technical support, assisting writing, providing a general support, providing material or financial support are examples to be indicated in acknowledgement section.

All authors must disclose all issues concerning financial relationship, conflict of interest, and competing interest that may potentially influence the results of the research or scientific judgment.

When an author discovers a significant error or inaccuracy in his/her own published paper, it is the author's obligation to promptly cooperate with the Editor to provide retractions or corrections of mistakes.

Responsibility for the Editor and Reviewers

Editor-in-Chief evaluates manuscripts for their scientific content without regard to ethnic origin, gender, citizenship, religious belief or political philosophy of the authors. He/She provides a fair double-blind peer review of the submitted articles for publication and ensures that all the information related to submitted manuscripts is kept as confidential before publishing.

Editor-in-Chief is responsible for the contents and overall quality of the publication. He/She must publish errata pages or make corrections when needed.

Editor-in-Chief does not allow any conflicts of interest between the authors, editors and reviewers. Only he has the full authority to assign a reviewer and is responsible for final decision for publication of the manuscripts in the journal.

Reviewers must have no conflict of interest with respect to the research, the authors and/or the research funders. Their judgments must be objective.

Reviewers must ensure that all the information related to submitted manuscripts is kept as confidential and must report to the editor if they are aware of copyright infringement and plagiarism on the author's side.

A reviewer who feels unqualified to review the topic of a manuscript or knows that its prompt review will be impossible should notify the editor and excuse himself from the review process.

The editor informs the reviewers that the manuscripts are confidential information and that this is a privileged interaction. The reviewers and editorial board cannot discuss the manuscripts with other persons. The anonymity of the referees must be ensured. In particular situations, the editor may share the review of one reviewer with other reviewers to clarify a particular point.

MANUSCRIPT ORGANIZATION

LANGUAGE

The publication language of the journal is English.

Manuscript Organization and Submission

The manuscript is to be submitted online via DergiPark System.

1. The manuscript has a minimum of 5000 words and a maximum of 20 pages without a References Section.

2. The manuscript should be in A4 paper standards: having 2.5 cm margins from right, left, bottom, and top, Times New Roman font style in 11 font size, and single line spacing. Due to double-blind peer review, the main manuscript document must not include any author information.

3. A Title Page must be submitted with the manuscript, including the followings:

The title of the manuscript.

All authors' names and affiliations (institution, faculty/department, city, country), e-mail addresses, and ORCIDs.

Information of the corresponding author (in addition to the author's information e-mail address, open correspondence address, and mobile phone number).

Financial support.

Conflict of interest.

Acknowledgment.

4. Submitted manuscripts must have an abstract between 200 and 250 words before the introduction, summarizing the scope, the purpose, the results of the study, and the methodology used. Under the abstracts, a minimum of 3 and a maximum of 5 keywords that inform the reader about the content of the study should be specified.

5. The manuscripts should contain mainly these components: Abstract (and Keywords), Introduction, Literature Review, Discussion and Conclusion, Acknowledgment (if it exists) (Conflict of Interest (if it exists), Financial Support (if it exists)), References, Appendix (if it exists). The authors may add necessary sections such as Method, Findings, etc.

6. Tables and figures can be given with a number and a caption. Every Figure or Table must be "called out" within the text of your article in numerical order with no abbreviations.

7. References should be prepared by American Psychological Association (APA) 7 reference system.

8. Authors are responsible for all statements made in their work submitted to the journal for publication.

9. If the Ethics Committee Report is required, it should be submitted, and the date and number of the ethics committee report should be stated in the manuscript. Otherwise, a word file containing a statement explaining why the Ethics Committee Report was not submitted must be uploaded to the system for the study. This statement will then be added to the end of your article.

REFERENCES

Reference Style and Format

Journal of Data Applications complies with **APA (American Psychological Association) style 7th Edition** for referencing and in-text citations. References should be listed in alphabetical order. Accuracy of citations is the author's responsibility. All references should be cited in the text. It is strongly recommended that authors may use Reference Management Software such as Zotero, Mendeley, etc.

Ensure that the following items are present:

- The title page is prepared according to the journal rules.
- The study has not been submitted to any other journal.
- The study was checked in terms of English.
- The study was written by paying attention to the full-text writing rules determined by the journal.
- The references are arranged by the APA-7 reference system.
- The Copyright Agreement Form is uploaded.
- The Author Contribution Form has been uploaded.
- Permission of previously published copyrighted material (text-picture-table) if used in the present manuscript.
- Ethics Committee Report (if necessary) is uploaded, and the ethics committee report date and number are given in the study text. Otherwise, please upload a word file to the system explaining why the ethics committee report was not submitted for this study. This statement will then be added to the end of your article.
- Reviewing journal policies.
- All authors have read and approved the latest version of the manuscript.

COPYRIGHT AGREEMENT FORM / TELİF HAKKI ANLAŞMASI FORMU

İstanbul University
İstanbul Üniversitesi

Journal name: Journal of Data Applications

Copyright Agreement Form
Telif Hakkı Anlaşması Formu

Responsible/Corresponding Author *Sorumlu Yazar*	
Title of Manuscript *Makalenin Başlığı*	
Acceptance date *Kabul Tarihi*	
List of authors *Yazarların Listesi*	

Sıra No	**Name - Surname** *Adı-Soyadı*	**E-mail** *E-Posta*	**Signature** *İmza*	**Date** *Tarih*
1				
2				
3				
4				
5				

Manuscript Type (Research Article, Review, etc.) *Makalenin türü (Araştırma makalesi, Derleme, v.b.)*	

Responsible/Corresponding Author:
Sorumlu Yazar:

University/company/institution	*Çalıştığı kurum*	
Address	*Posta adresi*	
E-mail	*E-posta*	
Phone; mobile phone	*Telefon no; GSM no*	

The author(s) agrees that:
The manuscript submitted is his/her/their own original work, and has not been plagiarized from any prior work,
all authors participated in the work in a substantive way, and are prepared to take public responsibility for the work,
all authors have seen and approved the manuscript as submitted,
the manuscript has not been published and is not being submitted or considered for publication elsewhere,
the text, illustrations, and any other materials included in the manuscript do not infringe upon any existing copyright or other rights of anyone.
ISTANBUL UNIVERSITY will publish the content under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license that gives permission to copy and redistribute the material in any medium or format other than commercial purposes as well as remix, transform and build upon the material by providing appropriate credit to the original work.
The Contributor(s) or, if applicable the Contributor's Employer, retain(s) all proprietary rights in addition to copyright, patent rights.
I/We indemnify ISTANBUL UNIVERSITY and the Editors of the Journals, and hold them harmless from any loss, expense or damage occasioned by a claim or suit by a third party for copyright infringement, or any suit arising out of any breach of the foregoing warranties as a result of publication of my/our article.
I/We also warrant that the article contains no libelous or unlawful statements, and does not contain material or instructions that might cause harm or injury.
This Copyright Agreement Form must be signed/ratified by all authors. Separate copies of the form (completed in full) may be submitted by authors located at different institutions; however, all signatures must be original and authenticated.

Yazar(lar) aşağıdaki hususları kabul eder
Sunulan makalenin yazar(lar)ın orijinal çalışması olduğunu ve intihal yapmadıklarını,
Tüm yazarların bu çalışmaya asli olarak katılmış olduklarını ve bu çalışma için her türlü sorumluluğu aldıklarını,
Tüm yazarların sunulan makalenin son halini gördüklerini ve onayladıklarını,
Makalenin başka bir yerde basılmadığını veya basılmak için sunulmadığını,
Makalede bulunan metnin, şekillerin ve dokümanların diğer şahıslara ait olan Telif Haklarını ihlal etmediğini kabul ve taahhüt ederler.
İSTANBUL ÜNİVERSİTESİ'nin bu fikri eseri, Creative Commons Atıf-GayrıTicari 4.0 Uluslararası (CC BY-NC 4.0) lisansı ile yayınlamasına izin verirler. Creative Commons Atıf-GayrıTicari 4.0 Uluslararası (CC BY-NC 4.0) lisansı, eserin ticari kullanım dışında her boyut ve formatta paylaşılmasına, kopyalanmasına, çoğaltılmasına ve orijinal esere uygun şekilde atıfta bulunmak kaydıyla yeniden düzenleme, dönüştürme ve eserin üzerine inşa etme dâhil adapte edilmesine izin verir.
Yazar(lar)ın veya varsa yazar(lar)ın işvereninin telif dâhil patent hakları, fikri mülkiyet hakları saklıdır.
Ben/Biz, telif hakkı ihlali nedeniyle üçüncü şahıslarca vuku bulacak hak talebi veya açılacak davalarda İSTANBUL ÜNİVERSİTESİ ve Dergi Editörlerinin hiçbir sorumluluğunun olmadığını, tüm sorumluluğun yazarlara ait olduğunu taahhüt ederim/ederiz.
Ayrıca Ben/Biz makalede hiçbir suç unsuru veya kanuna aykırı ifade bulunmadığını, araştırma yapılırken kanuna aykırı herhangi bir malzeme ve yöntem kullanılmadığını taahhüt ederim/ederiz.
Bu Telif Hakkı Anlaşması Formu tüm yazarlar tarafından imzalanmalıdır/onaylanmalıdır. Form farklı kurumlarda bulunan yazarlar tarafından ayrı kopyalar halinde doldurularak sunulabilir. Ancak, tüm imzaların orijinal veya kanıtlanabilir şekilde onaylı olması gerekir.

Responsible/Corresponding Author; *Sorumlu Yazar;*	**Signature** / *İmza*	**Date** / *Tarih*
		……../……../……