E-ISSN: 2980-3357

JOURNAL OF DATA APPLICATIONS

SAYI / ISSUE **2** • YIL / YEAR **2023**

İSTANBUL UNIVERSITY PRESS

E-ISSN: 3357-2980

JOURNAL OF DATA APPLICATIONS

ISSUE **2** • YEAR **2023**

İSTANBUL
UNIVERSITY
PRESS

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

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: 2, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

EDITORIAL MANAGEMENT BOARD

Editor-in-Chief

Co-Editors in Chief

Editorial Management Board Member

Language Editors

EDITORIAL BOARD

Journal of Data Applications
Issue: 2, 2023
E-ISSN: 2980-3357

CONTENTS

Araştırma Makaleleri / *Research Articles*

Journal of Data Applications
Issue: 2, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

Multi-Class Classification with the Gaussian Naive Bayes Algorithm

Ayşe ÇINAR[1]

ABSTRACT

Classification is a data mining technique involving supervised machine learning and is the process of predicting the class of data or dataset whose class is unknown using existing data with defined class. Supervised learning occurs during this classification process as a result of how this technique parses the data according to predetermined outputs. The Naive Bayes classifier is a type of machine learning algorithm and an approach that adopts Bayes' theorem by combining theoretically obtained preliminary information with new information. The most obvious advantages of this classifier are its simple algorithm and high accuracy rate. The aim of this study is to create a classification model using the Gaussian Naive Bayes algorithm and to evaluate the obtained prediction results. For this purpose, the study first theoretically considers within its scope the Naive Bayes classifier and then carries out an application on a dataset using the Gaussian Naive Bayes algorithm as one of the types of this classifier in order to create a classification model, which is the subject of the study. Operations were carried out for the classification model using Python, an open-source programming language. The dataset used within the scope of the study was obtained from the University of California Irvine (UCI) Machine Learning Repository website. The purpose for creating the dataset is to determine the different types of Erythemato-squamous disease (ESD). In line with developing technologies, the number of studies demonstrating the ability to make fast and reliable disease prediction using machine learning techniques are increasing daily.

Keywords: Gaussian Naive Bayes, Supervised Machine Learning, Classification, Naive Bayes Classifier

DOI: 10.26650/JODA.1389471

[1]Marmara University, Retired Faculty Member, İstanbul, Türkiye

ORCID: A.Ç. 0000-0001-7321-5959

Corresponding author:
Ayşe ÇINAR,
Marmara University, Retired Faculty Member, İstanbul, Türkiye
E-mail: ayse.cinar@gmail.com

Submitted: 11.11.2023
Accepted: 25.12.2023
Published Online: 07.03.2024

Citation: Çınar, A. (2023). Multi-class classification with the gaussian naive bayes algorithm. *Journal of Data Applications, 2*, 1-13. https://doi.org/10.26650/JODA.1389471

Introduction

The classification method known as supervised machine learning involves a process for distinguishing data classes in order to predict the class of an observation whose class label is unknown.

A dataset considered for classification purposes is generally divided into two sets: the training and test datasets. A dataset can also be divided into three sets to create a validation dataset in addition to these two. The training dataset, which is created with respect to a certain ratio of the whole data, is used to train classification algorithms. The training process occurs by using known values in the training dataset to predict a new observation or class of observations.

One of the popular methods used to classify data is the Naive Bayes classifier. The Naive Bayes algorithm is based on Bayes' Theorem and creates a probability set by calculating the frequency and combination of the values of the features in a given dataset (Wibawa et al., 2019). While determining the class of a new observation, the algorithm calculates conditional probabilities for each class label by taking into account the values of all the features of this observation. The highest calculated posterior probability and the class label with this value are selected and evaluated as the class of the new observation.

Naive Bayes algorithm has been proven to be effective and potentially good in many applications, including text classification, medical diagnosis, and system performance management (Aggarwal & Kaur, 2013; Wibawa et al., 2019; Salmi & Rustam, 2019). This study uses the Gaussian Naive Bayes classifier to process numerical data.

Literature Review

Many studies are found to have been conducted on the dermatology dataset using different techniques in order to accurately diagnose Erythemato-Squamous Diseases (ESD).

Bilgin and Çifçi (2021) used support vector machines (SVMs), ensemble learning algorithms (ELAs), decision trees (DTs) and k-nearest neighbors (k-NN) algorithm on the same dataset. According to their obtained results, they achieved the highest accuracy value of 99.73% with the SVM algorithm; their study did not include patient age in the model. Luukka (2011) proposed a classification method in which the data is first pre-processed using a new non-linear fuzzy robust principal component analysis (NFRPCA) algorithm to transform the data into a more suitable form; after the preprocessing step, Luuka obtained a prediction using a similarity classifier with a 99.02% accuracy rate. Verma et al. (2020) applied the bagging, adaptive boosting (AdaBoost), and gradient boosting (GBoost) methods and reached their highest accuracy rate (99.68%) with the gradient boosting method. Applying six different classification methods, Rashid et al. (2020) achieved their highest accuracy rate (97.54%) with the Naive Bayes and random forest methods. Putatunda's (2020) study on dermatology applied many methods to the same dataset, in which Derm2Vec as a hybrid deep learning model consisting of autoencoders and deep neural networks (DNNs) had the highest accuracy (96.92%), followed by the DNN (96.65%) and the extreme gradient boosting classifier (95.80%) methods. Shastri et al. (2021) reached a high accuracy rate of 99.45% with the GBoost method on the dermatology dataset.

Method

This study constructs a classification model using the Gaussian Naive Bayes algorithm, for which an application was implemented in Python. A dataset named *dermatology.data* was used to construct the model and contains 366 observations. The purpose of this dataset is to identify the types of ESD (İlter et al.,1998).

The application phase evaluates the first 10 observations in the dataset as validation data. Of the remaining 356 observations, 70% are considered the training dataset and 30% the testing dataset. Using this method for dividing the dataset enables the classification model to predict the class of 10 new observations it had never seen during the training phase for the 110 observations in the testing dataset.

Dataset

The dataset was downloaded from the University of California Irvine (UCI) Machine Learning Repository website and consists of 366 observations, 34 feature variables, and one target variable as shown in Table 1 (İlter & Güvenir, 1997). In the dataset, the feature named *family history* has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The feature named *age* represents the age of the patient. All other features are given a rating ranging from 0 to 3. For these features, 0 indicates that the feature is not present, 3 indicates the largest amount possible, and 1, 2 indicate relative intermediate values (İlter & Güvenir, 1997). Within the scope of this study, the feature named *class* was considered as the target (output), as shown in Table 2. The result from the classification model is determined by the class label of the target.

In the dataset, the classes of the target are represented by numerical values. The distributions of these classes in the dataset are shown in Table 3 along with their codes.

Table 1. *Information about the variables in the dataset.*

Name	**Explanation**	**Type**
Erythema	Erythema, reddening of the skin as a result of blood accumulation in the capillaries.	Integer
scaling	Skin flaking	Integer
definite_borders	strict borders	Integer
itching	Itching	Integer
koebner_phenomenon	Small, red blisters that form on the skin.	Integer
polygonal_papules	polygonal papules	Integer
follicular_papules	follicular_papules	Integer
oral_mucosal_involvement	Oral mucosal involvement	Integer
knee_and_elbow_involvement	Knee and elbow involvement	Integer
scalp_involvement	Skull skin involvement	Integer
family_history	Family history	Integer
melanin_incontinence	Inability to retain melanin	Integer
eosinophils_infiltrate	Eosinophil infiltration	Integer
PNL_infiltrate	PNL infiltration	Integer
fibrosis_papillary_dermis	Fibrous degeneration of blistered skin	Integer
exocytosis	Exocytosis	Integer
acanthosis	Acanthosis	Integer
Hyperkeratosis	Hyperkeratosis	Integer
parakeratosis	Parakeratosis	Integer
clubbing_rete_ridges	Clubbing of the rete protrusions	Integer
elongation_rete_ridges	Lengthening of rete processes	Integer
thinning_suprapapillary_epidermis	Thinning of the epidermis with high bubbles	Integer
spongiform_pustule	cancellous boil	Integer
munro_microabcess	Munro microabscess	Integer
focal_hypergranulosis	Focal hypergranulosis	Integer
disappearance_granular_layer	Disappearance of rough layer	Integer
vacuolisation_damage_basal_layer	Vacuuming and damage of the base layer	Integer
spongiosis	spongiosis	Integer
saw_tooth_appearance_retes	The rete has a sawtooth appearance	Integer
follicular_horn_plug	Follicular horn plug	Integer
perifollicular_parakeratosis	Perifollicular parakeratosis	Integer
inflammatory_mononuclear_infiltrate	Inflammatory mononuclear infiltrate	Integer
band_like_infiltrate	Band-shaped infiltration	Integer
Age	Age	Float
class	Type of ESD	Integer

Table 2. *The target.*

Name	Explanation	Type	Value
class	Type of ESD	Integer	1-6

Table 3. *Distribution rate of class labels of the target in the original dataset.*

Code	Class Labels	Frequency	Percentage
1	psoriasis	112	31%
2	seborrheic dermatitis	61	17%
3	lichen planus	72	20%
4	pityriasis rosea	49	13%
5	chronic dermatitis	52	14%
6	pityriasis rubra pilaris	20	5%

Data Preprocessing

The following data preprocessing steps were carried out before establishing the classification model.

As seen in Table 1, all features in the dataset have numerical values. The target (*class*) is also included in the dataset, with numerical codes representing class values. Within the scope of the study, these features are transformed into a categorical value by assigning the class names shown in Table 3.

The dataset has eight missing values that were found to belong to the feature named *Age*. This problem was resolved by assigning the mean value of the relevant feature instead of leaving the values blank. Moreover, when considering the class distribution ratio of the target (Table 3), the number of observations with the class named *psoriasis* was observed to be quite high compared to the other classes, while the number of observations with the class named *pityriasis rubra pilaris* was quite low. This situation reveals that the dataset under consideration is unbalanced. In order to balance the dataset, an oversampling process was applied using synthetic minority over-sampling technique (SMOTE), which enables the production of synthetic data.

Building a Classification Model

The Gaussian Navie Bayes algorithm, a type of Navie Bayes classifier, was applied to create a classification model.

Navie Bayes Classifier

The Naive Bayes classifier is a statistical classification method based on Bayes's theorem as developed by Thomas Bayes in the 18th century and works in accordance with the conditional probability principle discussed in Bayes' theorem (Meiriza et al., 2019). Based on Bayes' theorem, the resulting events (observations) are expected to be discrete events that are independent of one another (Khuda, 2021).

The Naive Bayes classifier reveals which class has the higher value in the form of a probability before making any classification. When estimating the class of new data with this method, a class with the highest probability value is considered as the class of new data. (Sarkar, 2023). Advantages and disadvantages of the Naive Bayes classifier can be listed as follows (Vadapalli, 2022):

Advantages:

It is a very fast working algorithm.

If the independence of the features assumption is valid, it performs better than other models with less training data.

It works effectively on multi-class predictions.

Disadvantages:

If a feature in the test dataset has a value that cannot be observed in the training dataset, it returns a probability value of 0, meaning it cannot make a prediction. This phenomenon is called zero frequency, and correction techniques such as Laplace estimation are used to resolve this issue (Wu et al., 2013).

All features are assumed to be independent. Although this seems possible in theory, finding a set of independent features is impossible in real life.

Bayes' Rule:

The Naive Bayes classifier is based on Bayesian Decision Theory . The probability equation of an observation whose class will be determined is shown below (Orhan & Adem, 2012).

- Posterior Probability:

$$P(C_i \backslash \mathrm{X}) = \frac{P(C_i\ ) * P(X \backslash \mathrm{C}_i)}{P(X\ )} \qquad \textit{Eq. 1}$$

$$P(X\ ) = \sum_{i=1}^{n} P(C_i\ ) * P(X \backslash C_i) \qquad \textit{Eq. 2}$$

Since the denominator *P(X)* will be equal for each class, only the numerator values are considered.

$$P(C_i \backslash \mathrm{X}) = P(C_i) * P(X \backslash \mathrm{C}_i) \qquad \textit{Eq. 3}$$

- Class-Conditional Probability:

The Naive Bayes classifier assumes the conditional independence of the feature values given the class as follows (Berrar, 2018):

$$P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i) \qquad \textit{Eq. 4}$$

- Classification Method for a New Data According to the Calculated Maximum of a Posteriori Probability (MAP):

Finally, by selecting the class with the highest probability value, the algorithm estimates the probability that a new observation with known feature values belongs to a class with this highest value. This expression is represented mathematically as:

$$Predicted\_Class = Argmax\, P(C_i) * \prod_{k=1}^{n} P(x_k \backslash C_i) \qquad \textit{Eq. 5}$$

$\mathrm{X} = \{x_1, x_2, \ldots., x_n\}$	Dataset with unknown class membership.
$x_1, x_2, \ldots., x_n$	The value of the features
$\mathrm{C} = \{C_1, C_2, \ldots., nC\}$	Class labels

There are three types of Naive Bayes classifiers (Ismail et al.,2020). Gaussian Naive Bayes should be used for features with continuous values. Gaussian Naive Bayes assumes that the features follow a normal distribution. The Gaussian Naive Bayes classification used within the scope of this study is discussed here. Multinomial Naive Bayes is mostly used to classify documents. Each document in the dataset should contain features with discrete positive integer values that express the frequency of words. Bernoulli Naive Bayes should be used for features with binary or Boolean values such as true/false or 0/1.

Classification with the Gaussian Naive Bayes Algorithm

Unlike with the Naive Bayes classification algorithm, normal (i.e., Gaussian) distribution curves are obtained for features with continuous values associated with each class using the Gaussian Naive Bayes algorithm. For this purpose, the distribution is summarized by estimating the mean and standard deviation values for each class from the training data. The obtained results obtained are considered as the conditional probability value for each class of features of an observation.

- Conditional Probabilities of the Features:

$$P\,(x_i \backslash C) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, exp\,(- \frac{(x_i - \mu_c)^2}{2\sigma_c^2}) \qquad \textit{Eq. 6}$$

- Class-Conditional Probability:

$$P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i) \qquad \textit{Eq. 7}$$

- Classification Method According to the Maximum a Posteriori Probability (MAP):

$$Predicted\_Class = Argmax\; P(C_i) * P(X \backslash C_i) \qquad i \in \{1, .., n\} \qquad \textit{Eq. 8}$$

Experimental Results

Confusion Matrix

Table 4 shows the confusion matrix obtained for the Gaussian Naive Bayes classification model created in this study. By considering the confusion matrix, various performance metrics are obtained for the classification models. One of these is the accuracy value, which is the correct prediction rate of the class label of the observations in the model’s testing dataset. The number of correct predictions for each class is shown in bold in Table 4. The accuracy value of the model constructed in this study is found to be 98.18%, as seen below.

$$\text{Accuracy} = \frac{\text{total number of correct predictions}}{total\ number\ of\ observations} = \frac{13+20+22+5+31+14}{107} = 0.9813$$

This classification model has a high accuracy rate and made incorrect predictions for the class of only two observations. Accordingly, although the class of both observations was *seb-dermatitis*, the model predicted it as *pityriasis rosea*.

Table 4. *Six-class confusion matrix.*

Actual **Predicted**	**chronic dermatitis**	**lichen planus**	**pityriasis rosea**	**pit_ rubra pilaris**	**psoriasis**	**seb_ dermatitis**
chronic dermatitis	**13**	0	0	0	0	0
lichen planus	0	**20**	0	0	0	0
pityriasis rosea	0	0	**22**	0	0	2
pit_ rubra pilaris	0	0	0	**5**	0	0
Psoriasis	0	0	0	0	**31**	0
seb_ dermatitis	0	0	0	0	0	**14**

Performance metrics apart from model accuracy, sensitivity (recall), specificity, precision, and harmonic precision-recall mean (F1 score) are calculated based on a positive class of the target. The calculation of these values was carried out with the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) frequency values included in the Confusion Matrix.

TP (True Positive): The number of observations that are positive and have been correctly classified as positive.

TN (True Negative): The number of observations that are negative and have been correctly classified as negative.

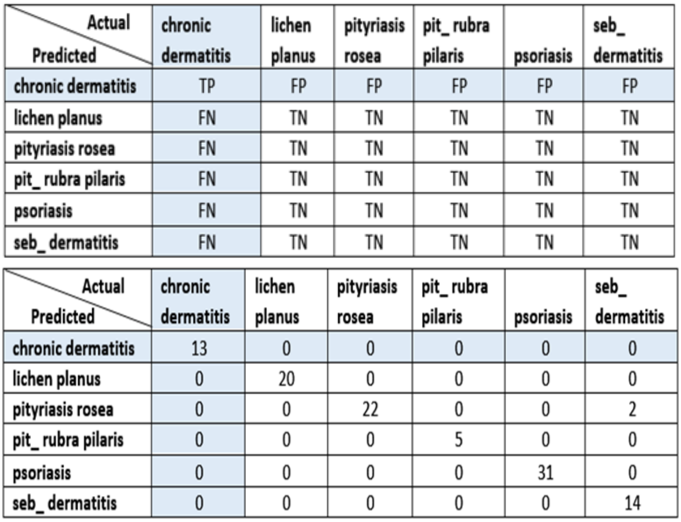
FP (False Positive): The number of observations that are negative but have been incorrectly classified as positive.

FN (False Negative): The number of observations that are positive but have been incorrectly classified as negative.

Within the scope of the study, performance metric values have been calculated by considering each class of the target as a positive class. The number of observations (T) of the test dataset considered in the calculation phase is 107.

The calculation method and obtained performance metric values of the positive classes called *chronic dermatitis* (Figure 1) and *pityriasis rosea* (Figure 2) are presented below as examples.

Positive class: « ***chronic dermatitis*** »



Actual Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	TP	FP	FP	FP	FP	FP
lichen planus	FN	TN	TN	TN	TN	TN
pityriasis rosea	FN	TN	TN	TN	TN	TN
pit_ rubra pilaris	FN	TN	TN	TN	TN	TN
psoriasis	FN	TN	TN	TN	TN	TN
seb_ dermatitis	FN	TN	TN	TN	TN	TN

Actual Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	13	0	0	0	0	0
lichen planus	0	20	0	0	0	0
pityriasis rosea	0	0	22	0	0	2
pit_ rubra pilaris	0	0	0	5	0	0
psoriasis	0	0	0	0	31	0
seb_ dermatitis	0	0	0	0	0	14

Figure 1. *Six-class confusion matrix, where the positive class is chronic dermatitis.*

The total TP, FN, FP and FN values obtained with the test dataset for *chronic dermatitis*, which is considered as a positive class, are as follows.

Total(TP)=13

Total(TN)=94

Total(FP)=0

Total(FN)=0

Performance Metrics:

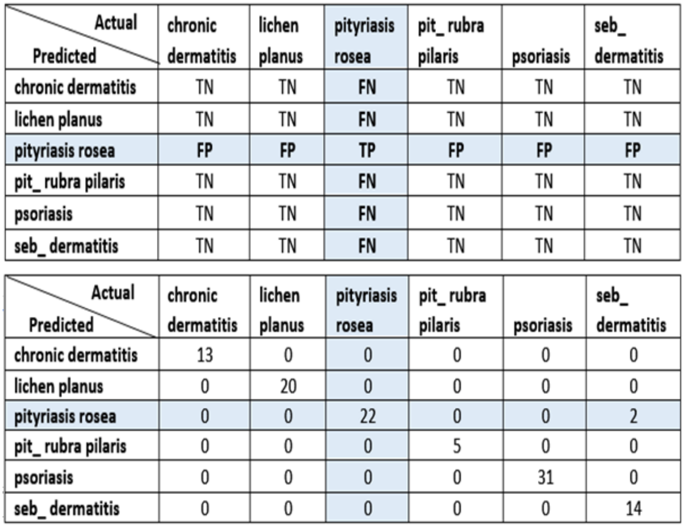
No Information Rate : $\frac{\text{Total(actual(chronic dermatitis))}}{T}$ = 0,12

Sensitivity(Recall) : $\frac{TP}{TP+FN}$ = 1

Specificity : $\frac{TN}{TN+FP}$ = 1

Precision : $\frac{TP}{TP+FP}$ = 1

Positive class: « ***pityriasis rosea*** »



Actual / Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	TN	TN	FN	TN	TN	TN
lichen planus	TN	TN	FN	TN	TN	TN
pityriasis rosea	FP	FP	TP	FP	FP	FP
pit_ rubra pilaris	TN	TN	FN	TN	TN	TN
psoriasis	TN	TN	FN	TN	TN	TN
seb_ dermatitis	TN	TN	FN	TN	TN	TN

Actual / Predicted	chronic dermatitis	lichen planus	pityriasis rosea	pit_ rubra pilaris	psoriasis	seb_ dermatitis
chronic dermatitis	13	0	0	0	0	0
lichen planus	0	20	0	0	0	0
pityriasis rosea	0	0	22	0	0	2
pit_ rubra pilaris	0	0	0	5	0	0
psoriasis	0	0	0	0	31	0
seb_ dermatitis	0	0	0	0	0	14

Figure 2. *Six-class confusion matrix, where the positive class is pityriasis rosea.*

The total TP, FN, FP and FN values obtained with the test dataset for *pityriasis rosea*, which is considered as a positive class, are as follows.

Total(TP)= 22

Total(TN)= 83

Total(FP)= 2

Total(FN)=0

Performance Metrics:

No Information Rate : $\frac{\text{Total(actual(pityriasis rosea))}}{T}$ = 0,21

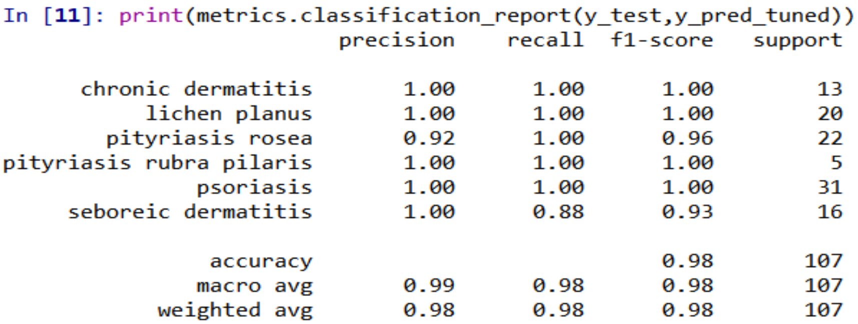
Sensitivity(Recall) : $\frac{TP}{TP+FN}$ = 1

Specificity : $\frac{TN}{TN+FP}$ = 0,98

Precision : $\frac{TP}{TP+FP}$ = 0,92

F1-test : $\frac{2*Precision*Sensitivity}{Precision+Sensitivity}$ = 0,96

The performance metrics obtained by the classification model created with the Gaussian Naive Bayes algorithm separately for each class of the target are shown in Figure 3.



```
In [11]: print(metrics.classification_report(y_test,y_pred_tuned))
                          precision    recall  f1-score   support

     chronic dermatitis       1.00      1.00      1.00        13
           lichen planus       1.00      1.00      1.00        20
       pityriasis rosea       0.92      1.00      0.96        22
pityriasis rubra pilaris       1.00      1.00      1.00         5
               psoriasis       1.00      1.00      1.00        31
     seboreic dermatitis       1.00      0.88      0.93        16

                accuracy                           0.98       107
               macro avg       0.99      0.98      0.98       107
            weighted avg       0.98      0.98      0.98       107
```

Figure 3. *Performance metrics obtained with Gaussian Naive Bayes algorithm.*

Conclusion

This study has constructed a classification model using the Gaussian Naive Bayes algorithm. A dermatology dataset was used for this model to make predictions about patients' types of ESD. Numerous studies have attempted to apply machine learning techniques to dermatology datasets. Some of those that have applied Naive Bayes and other classification algorithms are found within this study's Literature Review section.

The study carried out within its scope some data preprocessing on the dataset prior to building the model and getting the data ready for analysis. Additionally, the dataset was divided into three groups: the training, testing, and validation datasets. While the validation group includes the first 10 observations from the dataset, of the remaining observations, 30% made up the testing and 70% the training data. Training was carried out by constructing a classification model using the training data, and the accuracy and other performance metrics of the model were obtained using the testing data. The study then applied model tuning to improve the model's performance metrics. The most appropriate performance metrics for the dataset were achieved by building the model using the best parameter values obtained by applying this process. As a result, the accuracy rate of the model was 98.13%. According to this result, the classification model built with the Gaussian Naive Bayes algorithm was observed to have achieved results close to other studies using the same dataset.

After tuning the model, the tuned classification model predicted a class of 10 new observations in the validation dataset, and the model was seen to correctly predict the class of all new observations.

Ethics Committee Approval: Since the dermatology dataset used in my study was not obtained by me through a survey, I think that there is no need for ethics committee approval for my study. It is a public file published on the internet.

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

Aggarwal, S., & Kaur, D. (2013). Naive Bayes Classifier with Various Smoothing Techniques for Text Documents. *International Journal of Computer Trends and Technology (IJCTT)*, 4(4), 873-876. https://ijcttjournal.org/Volume4/issue-4/IJCTT-V4I4P187.pdf

Ahmed, M., Kashem, M. A., Rahman, M., & Khatun, S. (2020). *Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT),* Part of the Lecture Notes in Electrical Engineering book series (LNEE), volume 632, Springer.

Berrar, D.(January 2018), *Bayes' Theorem and Naive Bayes Classifier,* Reference Module in Life Sciences, https://www.researchgate.net/publication/324933572_Bayes'_Theorem_and_Naive_Bayes_Classifier#fullTextFileContent

Bilgin, G., & Çifçi A. (2021). Eritematöz Skuamöz Hastalıkların Teşhisinde Makine Öğrenme Algoritmaları Performanslarının Değerlendirilmesi. *Zeki Sistemler Teori ve Uygulamaları Dergisi*, 4(2), 195-202.

İlter, N., & Güvenir, H. (1997). Dermatology, https://archive.ics.uci.edu/dataset/33/dermatology

İlter, N., Güvenir, H., & Demiroz, G. (1998). Learning Differential Diagnosis of Erythemato-Squamous Diseases Using Voting Feature Intervals. *Artificial Intelligence in Medicine*, 13(3), 147-165.

Ismail, M., Hassan,N., & Bafjaish,S.S. (2020). Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task. *JSCDM-Journal of Soft Computing and Data Mining*, 1(2), 1-10.

Khuda, I. E. (2021). Innovative Teaching Pedagogy for Teaching and Learning of Bayes'Theorem. *Journal of Science and Engineering, CUSJE*, 18(1), 61-71, Çankaya University, 63-71.

Luukka P. (2011). A New Nonlinear Fuzzy Robust PCA Algorithm and Similarity Classifier in Classification of Medical Datasets. *International Journal of Fuzzy Systems*, 13(3), 153-162.

Meiriza, A., Lestari, E., Putra, P., Monaputri, A. & Lestari, D.A. (2019). *Prediction Graduate Student Use Naive Bayes Classifier*, Advances in Intelligent Systems Research Series. Volume 172, Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), Atlantis Press.

Orhan, U & Adem, K. (2012, October 29-December 01). *Naive Bayes Yönteminde Olasılık Çarpanlarının Etkileri* [Conference presentation]. Elektrik - Elektronik ve Bilgisayar Mühendisliği Sempozyumu, 29 Kasım - 01 Aralık 2012, Bursa.

Putatunda, S. (2020, July 2-4). *A Hybrid Deep Learning Approach For Diagnosis Of The Erythemato-Squamous Disease*[Conference presentation]. IEEE International Conference On Electronics, Computing And Communication Technologies, Bangalore, India, 1-6.

Rashid, A.N.M.B., Ahmed, M., Sikos, L.F., Haskell-Dowland, P. (2020). A Novel Penalty-Based Wrapper Objective Function for Feature Selection In Big Data Using Cooperative Co-Evolution. IEEE Access, 8, 150113-150129.

Salmi, N. & Rustam, Z. (2019). *Naive Bayes Classifier Models for Predicting the Colon Cancer*, 9th Annual Basic Science International Conference 2019 (BaSIC 2019), IOP Conf. Series: Materials Science and Engineering 546 (2019) 052068, IOP Publishing.

Sarkar P. (2023). *Naive Bayes in Machine Learning [Examples, Models, Types]*, https://www.knowledgehut.com/blog/data-science/naive-bayes-in-machine-learning

Shastri, S., Kour, P., Kumar, S., Singh, K., Mansotra, V. (2021). GBoost: A Novel Grading-AdaBoost Ensemble Approach for Automatic Identification of Erythemato-Squamous Disease. *International Journal of Information Technology*, 13(3), 959-971.

Vadapalli,P. (2022). *Naive Bayes Classifier: Pros & Cons, Applications & Types Explained*, https://www.upgrad.com/blog/naive-bayes-classifier/

Verma, A.K., Pal, S., Kumar, S. (2020). Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—A Comparative Study. *Applied Biochemistry and Biotechnology*, 190(2), 341-359.

Wibawa, A.P., Kurniawan, A.C., Murti, D.M.P., Adiperkasa, R.P., Putra, S.M., Kurniawan, S.A. & Nugraha, Y.R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. *iJES*, 7(2), 91-99, https://doi.org/10.3991/ijes.v7i2.10659

Wu, J., Cai, Z. & Zhu, X. (2013, August 4-9). *Self-adaptive probability estimation for Naive Bayes Classification*[Conference presentation]. International Joint Conference on Neural Networks (IJCNN), U.S.A.

Journal of Data Applications
Issue: 2, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

Forecasting Restaurant Sales with the Sensitivity of Weather Conditions and Special Days Using Facebook Prophet

Ali Kerem GÜLER[1], Ali MUSA[1], Mustafa TARIM[1], Osman SARAÇ[1], Mehmet GÖKTÜRK[2]

ABSTRACT

This article focuses on forecasting sales for restaurant businesses using the Prophet model developed by Facebook. A method is proposed to make more accurate forecasts by accounting for the effects external factors have on sales, including weather conditions and special days. The analyses conducted on the real-time sales data of the daily operations of a restaurant business (provided by PROTEL Inc.) reveal that the Prophet model can forecast the sales of different products based on daily sales and weather data. The prediction performance of the model was evaluated using four error metrics: Mean Absolute Error, Mean Absolute Percentage Error, Mean Squared Error, and Root Mean Square Error. The results revealed that the model produced more consistent and accurate predictions for some product categories. This study, which aims to contribute to the literature through an optimization of operational efficiency and decision-making processes related to the restaurant industry, highlights the importance of external factors in sales forecasting in the restaurant industry and provides a detailed analysis of incorporating these factors into the forecasting process. The findings may support restaurant businesses in obtaining more accurate sales forecasts by taking external factors into account. In particular, understanding the effects of weather changes and special days on sales can contribute significantly to operational decisions in such areas as personnel planning and inventory management. In this regard, the article proposes innovative approaches to the challenges faced by restaurant operations, presenting different approaches found in the literature and a detailed model evaluation process.

Keywords: Machine Learning, Time Series Forecasting, Facebook Prophet

DOI: 10.26650/JODA.1450459

[1]PROTEL A.Ş., Department of Software, İstanbul, Türkiye Turkiye
[2]Gebze Technical University, Department of Computer Engineering, Kocaeli, Türkiye

ORCID: A.K.G. 0000-0003-3405-005X;
A.M. 0009-0000-5829-5698;
M.T. 0000-0001-7079-5062;
O.S. 0000-0002-4870-9823;
M.G. 0000-0002-8030-8923

Corresponding author:
Ali Kerem GÜLER,
PROTEL A.Ş., Department of Software, İstanbul, Türkiye
E-mail: aguler@protel.com.tr

Submitted: 11.03.2024
Revision Requested: 28.03.2024
Last Revision Received: 28.03.2024
Accepted: 03.04.2024
Published Online: 24.06.2024

Citation: Güler, A. K., Musa, A., Tarım, M., Saraç, O., & Göktürk, M. (2023). Forecasting restaurant sales with the sensitivity of weather conditions and special days using facebook prophet. *Journal of Data Applications, 2*, 15-30. https://doi.org/10.26650/JODA.1450459

Introduction

Sales forecasting is the process of predicting future sales of businesses and is essential for planning business strategies and an effective use of resources. Accurate and effective sales forecasting is critical in a competitive business environment. Accurate sales forecasts impact many aspects of restaurant business operations, such as inventory management, supply chain optimization, financial planning, and customer satisfaction. A proper implementation of this procedure allows businesses to prevent stock overages and supply shortages, reduce costs, and obtain a competitive edge.

Traditional sales forecasting methods forecast future sales by taking into account various factors, such as historical sales data and seasonal patterns. In the restaurant industry, understanding and predicting the impact of external factors (such as weather changes and special occasions) on sales can directly impact the profitability of a business and improve customer satisfaction. Today, developments in artificial intelligence and big data analytics have greatly improved the sales forecasting process. Artificial intelligence-based methods can process more complex data sets and provide more accurate predictions.

Facebook's Prophet algorithm is employed in this study to forecast restaurant sales. Prophet (Taylor & Letham, 2018) is unique in that it may detect intricate patterns in time series data and has a flexible architecture. By addressing the sensitivity of external factors on sales forecasts, the research focuses on particular challenges encountered by the restaurant industry and seeks to optimize inventory management and business strategies. Additionally, the present research analyzes the effects that weather and special occasions have on sales, integrating these factors into sales forecasts.

The impact of sales forecasting appears across multiple domains, including inventory optimization, supply chain management, and satisfying customer needs. The restaurant industry has a complex structure which is marked by customer preferences changing overnight and numerous external influences. This study predicts future restaurant sales by incorporating such external factors. Thus, such an analysis enables restaurants to optimize their sales volumes and also the orders of necessary ingredients based on reliable forecasts. This approach allows restaurants to increase operational efficiency while reducing food waste and lowering costs.

The article is structured as follows: The Literature Review section covers relevant studies that have approached the topic from various angles. The Methodology section introduces the dataset, then explains the data processing and modeling techniques used. The Results and Discussion section presents the findings and their outcomes. The Future Work and Recommendations section provides information on potential subjects for further study. The article concludes with the Conclusion section.

Literature Review

Sales Forcasting

In their research, Patricia Ramos et al. (2015) focused on predicting coming sales, which is one of the most important factors of effective business operations. Accurate forecasting of demands is important for retail businesses in terms of production, purchasing, logistics, and workforce organization. The study also included time series of retail sales, which are a type of time series that is characterized by trending and seasonal patterns and presents several challenges in terms of implementing efficient forecasting models. In order to forecast consumer retail sales, the study analyzes the accuracy of state space and ARIMA models. Both single step and multistep forecasts were conducted, with the performance of the forecasts being illustrated by a case study of retail sales in five distinct groups of women's footwear (Boots, Ankle-length boots, Flats, Sandals, and Shoes). The prediction performances of the state space and ARIMA models were measured based on the MAE, MAPE, and RMSE metrics in both single step and multistep predictions. These analyses revealed that the error scores of both models were quite near to one another.

Sébastien Thomassey (2010) conducted a comprehensive analysis while detailed success in the textile-clothing industry, where consumer demands are constantly increasing. The study emphasizes the importance of complex information systems and logistics capabilities based on forecasting systems. Thomassey noted some of the unique challenges of the textile-clothing market, such as variable demand, strong seasonality, a large number of short-lived products, and a lack of historical data. The author provides information that, against these challenges, companies often develop simple but robust forecasting mechanisms. After evaluating current practices in the clothing industry, is the study proposed the use of different forecasting models that provide more accurate and reliable sales forecasts. Such models are built around different and innovative approaches, such as fuzzy logic, neural networks, and data mining techniques, which focus on the specific challenges faced by companies in the sector.

A study by Arunraj and Ahrens (2015) focused on providing an innovative approach to the daily sales forecasting problems encountered in the food retail industry. It is emphasized that food waste and stock outages in this sector are caused by incorrect sales forecasting and, as a result, ordering the wrong product. Time series sales data in the food retail sector can exhibit high variability and skewness. For this reason, it has been mentioned that interval forecasts should be made for retail companies to determine appropriate stock policies. This study demonstrates the effectiveness of a hybrid model called SARIMA-QR, produced by combining SARIMA and Quantile regression, for the difficulties encountered in daily sales forecasting in the industry. Using daily sales data of banana products from a German discount retail store, the authors developed a model to forecast future banana sales. Considering the variable sales dynamics and various factors affecting demand, the new model offers the potential to obtain more accurate and comprehensive forecasts.

It has been stated that the Bayesian model developed by Posch et al. (2022) for food and beverage sales forecasting in restaurants and cafeterias is more efficient than the existing methods of ARIMA, Seasonal ARIMA, and Exponential Smoothing. In other words, it provides a high accuracy rate and a low margin of error. It has been stated that the results obtained from sales forecasts contribute to reducing food waste and optimizing stock management. is the study pointed out that the model which they used has the ability to efficiently integrate different external factors and seasonal changes as a powerful forecasting tool in the restaurant industry.

Using Machine Learning for Sales Forecasting

A deep learning approach was used to predict the following season's product sales in the retail fashion industry in a study which compared traditional machine learning and deep learning models (Loureiro et al., 2018). Their models were constructed using real data which considered a variety of characteristics, including product physical characteristics and expert recommendations. The researchers revealed that deep learning techniques, as well as more traditional machine learning techniques, can be effective in this area. The study also serves as a reference for comparing deep and traditional machine learning techniques in retail sales forecasting.

Machine learning models were discussed in a study conducted by Tsoumakas (2019) with the aim of predicting the future sales rates of different businesses operating in the food industry (supermarkets, restaurants, and bakeries). The study emphasized that the model obtained from short-term sales forecasts can enable businesses to minimize both their stocks and expired products. In addition, the study also discussed important design processes, such as the temporal granularity of sales data, the input factors to be used in forecasting, and the representation of the output variable. Machine learning approaches used in sales forecasting and appropriate measurement metrics to evaluate the performance of these algorithms have also been included in detail.

Another important study was conducted by Shilong et al. (2021). In this research, which examines the role of the success of sales forecasts for large retail chains on the development, success, and failure of businesses, a sales forecasting model was put forward that allows companies to manage their resources more effectively and produce effective business plans. The basis of this model is to extract features from historical sales data through feature engineering and predict subsequent sales volumes using the Extreme Gradient Boosting (XGBoost) model. The model was shown to work well in terms of cost through experimentation conducted on Walmart retail products.

Using Facebook Prophet for Sales Forecasting

Prophet's model, based on time series decomposition, was applied to six different financial time series datasets with different input parameters in Yusof et al.'s (2020) study. The implemented datasets cover a variety of markets, such as the Hong Kong Hang Seng 300 Index (HS300), the Standard & Poor's 500 (SP500), and the Tokyo Nihon Keizai Shinbun Index (Nikkei). The study presents a new approach to the development of financial forecasting models. It argues that the Prophet model has the potential to effectively model the movements of financial markets with appropriate parameter settings and an average absolute percentage error of six percent.

A study conducted by Zunic et al. (2020) presented a comprehensive framework for using Facebook's Prophet algorithm for sales forecasting in the retail industry. Accordingly, the study purports that the Prophet model has the capacity to manage the effects of seasonality, as well as its effectiveness in monthly and quarterly sales forecasts. In addition, the research highlighted the importance of back testing for measuring forecast reliability and classifying product portfolios. It is emphasized that Prophet can be used effectively in retail forecasting By showing how Prophet can be applied to various product categories with different lengths of historical data, there is an emphasis on the effective use of the algorithm for retail forecasting.

A study conducted by Jha and Pande (2021) recommended using a time series sales rate forecasting model in supermarkets, which contains useful information about time that assists in statistical analyses. The researchers claimed that the model would also increase sales rates. The study reveals the usefulness of the Prophet model for supermarket data, suggesting a low prediction accuracy of this tool at 8.3, based on the MAPE result and other metrics. Additionally, the ARIMA, Holt-Winter, and Prophet models were compared using the same error metrics, concluding that the Prophet model performed better. Thus, the study indicates the benefits of using the Prophet model as a time series forecasting model for supermarket sales forecasting.

Methodology

Dataset

The data sets used in sales forecasting in this study were provided by PROTEL Inc. and include daily sales data of a business on a product basis. These data sets contain sales amounts of different products of businesses and historical information of sales. The data sets used in this study were collected on a daily basis through the Point of Sale (POS) system of PROTEL Inc. company and stored in a time series format. The collected data sets were prepared as .csv files in order to perform analytical processes and modeling.

Since data sets keep real-time sales data gathered from daily workflows of businesses, they present the sales trends of products, seasonal changes, and the impact of other time-related factors on sales in detail. Because these data sets are in a format suitable for time series operations, they have been a useful resource in developing and testing time series models for sales forecasting. Analyzing data sets in detail is very important in making sense of the sales performance of businesses and predicting future sales trends.

In this study, the weather data required for the model to produce precise predictions were obtained from the OpenWeatherMap platform. OpenWeatherMap (OpenWeatherMap, 2024) is an open-source platform that provides weather data via API. During the time period covered by the study, weather data were taken and recorded twice every day, at regular intervals between 9 am and 9 pm. This systematic approach enabled the holistic collection of all weather information required for the model since the inception date of the sales forecast data. The OpenWeatherMap platform contains a wide range of data, including important meteorological measurements, such as temperature, humidity, pressure, wind speed, and cloud information. This data helps the forecasting model better understand the impact of certain weather conditions on sales and make its forecasts more precisely.

This study covers detailed analysis of data sets and modeling processes to help businesses make their sales forecasts more consistent, improve their inventory management, optimize their operational planning, and create marketing strategies more effectively. Thanks to the regularly collected data, the forecast model has the capacity to produce more precise and reliable forecasts by efficiently analyzing daily sales dynamics.

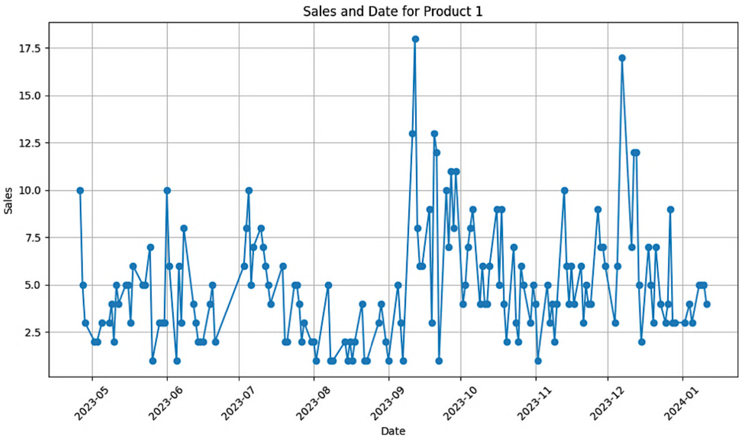


Figure 1. *Graph of historical sales for product 1.*

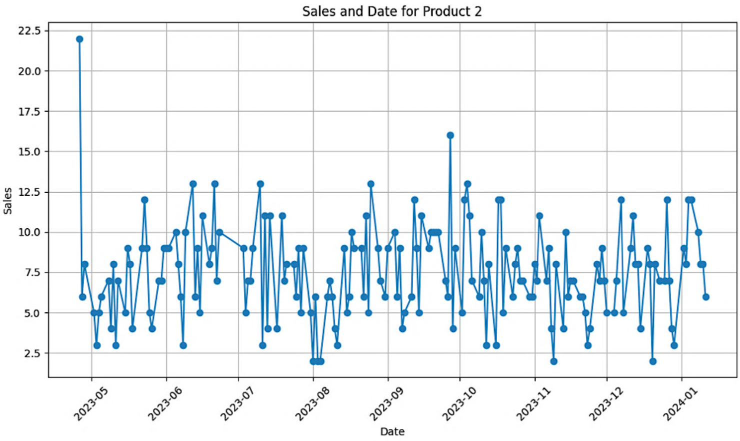


Figure 2. *Graph of historical sales for product 2.*

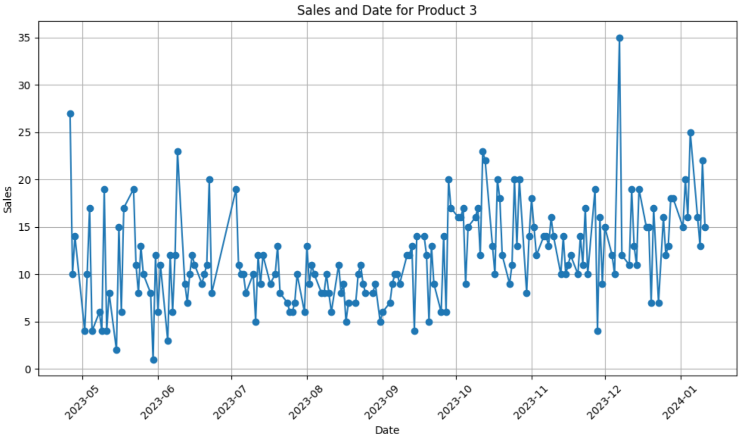


Figure 3. *Graph of historical sales for product 3.*

Eight months' worth of sales data for three different goods of a restaurant business are visualized in Figure 1, Figure 2, and Figure 3. The time series analysis conducted on the sales of each product has been observed within the framework of fundamental concepts such as trend, seasonality, cyclicality, and volatility. These concepts play a crucial role in enhancing the understanding of sales dynamics, thereby contributing significantly to the prediction of future sales. The initial examinations of the graphs indicate the challenges of identifying a clear and direct trend. However, the increases or decreases in sales volumes at certain intervals can be interpreted as potential signs of seasonality. Additionally, the fluctuations in sales data over time point to the influence of external factors. These fluctuations could stem from such environmental changes as economic, social, or weather conditions.

An analysis of the data forms the basis of the sales forecasting model and determines the data patterns that the model needs in order to learn. Additionally, such an analysis is crucial in terms of evaluating the performance of the model and implementing the necessary adjustments to forecast future sales. This holistic approach contributes to the continuous development of the model and will provide more concrete benefits to businesses.

Data Processing

For an accurate and efficient analysis of the time series data of the product sales of businesses, a data processing method, such as that used in this study, is crucial. The process begins by identifying missing data, before filling these gaps. The first area of focus relates to whether the weather data has been integrated into sales data. The weather data is divided into two different time periods of the day: morning and evening. This data includes various features, such as date, temperature, felt temperature, pressure, humidity, wind speed, cloud information, sunrise, and sunset.

A special function has been developed to detect and complete missing data points in case there is a break in the weather data for any reason. This function detects missing weather records for 9 am and 9 pm by using the date and time information in the data set. For each missing data point, an index number is kept along with the date and time information of the relevant record. This approach enabled the locations of missing data points in the data set to be found accurately and easily. Then, the detected missing data points were filled with the values of the weather records from the same time period of the previous day. When consecutive missing data points were detected, this was specifically handled by the function and the missing data were filled by taking into account the values of the last known data point.

The morning and evening weather datasets were equalized in terms of size after the missing data points were filled in. This process is important in terms of correcting the inconsistencies that occur in the start and end dates of the data set. If differences are detected in the start or end dates of the morning and evening, a new record is added to the smaller data set based on the first or last record of the other data set. With this method, morning and evening weather data were completed and checked for deficiencies. Following the verification procedure, the morning and evening weather data sets were combined in the correct order according to the dates.

When each data set used in sales forecasting was analyzed, it was observed that sales information was missing for some dates. These gaps were caused by missing sales records for the products on these dates. To solve this problem, a function that detects and fills in missing data has been developed. This function filled the sales information with a zero (0) value on the relevant dates, assuming that there were no sales on the dates determined for each data set

Following these processes, each sales data set was integrated with weather data on the relevant dates, which prepared everything to be used in sales forecasting.

Modeling

In the modeling part of the study, the open-source Prophet model developed by Facebook was used for time series forecasting. This model has a machine learning-based approach and has been effective in analyzing time series data, especially with its sensitivity to special days (such as holidays) and seasonal effects. The Prophet can model linear and nonlinear trends and automatically detect annual, weekly, and daily seasonality, which is why it was chosen for the modeling phase of this study.

Prophet models were trained separately for each sales data set, including weather information. By giving weather data to the models, they were able to learn the effects of weather conditions on sales (Badorf & Hoberg, 2020). The parameters for “weekly_seasonality” and “daily_seasonality” were adjusted to help the models learn weekly and daily patterns. In this way, the models are able to analyze the changes in sales according to hours of the day and days of the week. The “changepoint_prior_scale” parameter was used to adjust the sensitivity of the models to trend changes in the data sets. This value was determined as 0.09 for all models. In addition, with the “country_code” parameter, the models were enabled to automatically take country-based holidays into account, and thus, the effects of holidays on sales were included in the modeling process. Since the time series data considered belongs to a business in Turkey, the “country_code” was set to ‘TR,’ enabling the models to take into account holidays specific to the country.

The “make_future_dataframe” function of the Prophet model was utilized to produce future date data frames that would be similar to the training datasets. This function acquires information about the duration and frequency of the forecast through parameters. Since the study used daily sales data, the data frames were also generated with a daily frequency. In order to maintain a forecast period of 30 days, the data frames for the models’ predictions were set to 30 data points. Despite the models’ capability to generate future data frames, the inclusion of weather data into these frames was necessary for more precise forecasting. For this purpose, seven-day weather forecast data were obtained from the OpenWeatherMap platform via API using the Python programming language. These forecasts were produced with atmospheric measurements and mathematical models instead of machine learning approaches. Considering the possibility of inconsistencies in weather forecasts beyond a certain period, efforts were taken to ensure that the models used the extended data frames with weather forecast data for the first 7 days for their predictions. For the remaining 23 days, the models forecasted without weather data. Subsequently, the sales forecasts made with and without weather forecast data were merged with accuracy. This approach thoroughly assesses the effects of sales and weather data in a time series analysis, allowing the Prophet model to precisely model the impact of such significant factors as seasonality and trends.

Results and Discussion

Within the scope of this study, Facebook's Prophet model was applied to make forecasts based on daily sales data collected over approximately eight months for each product, provided by PROTEL Inc.

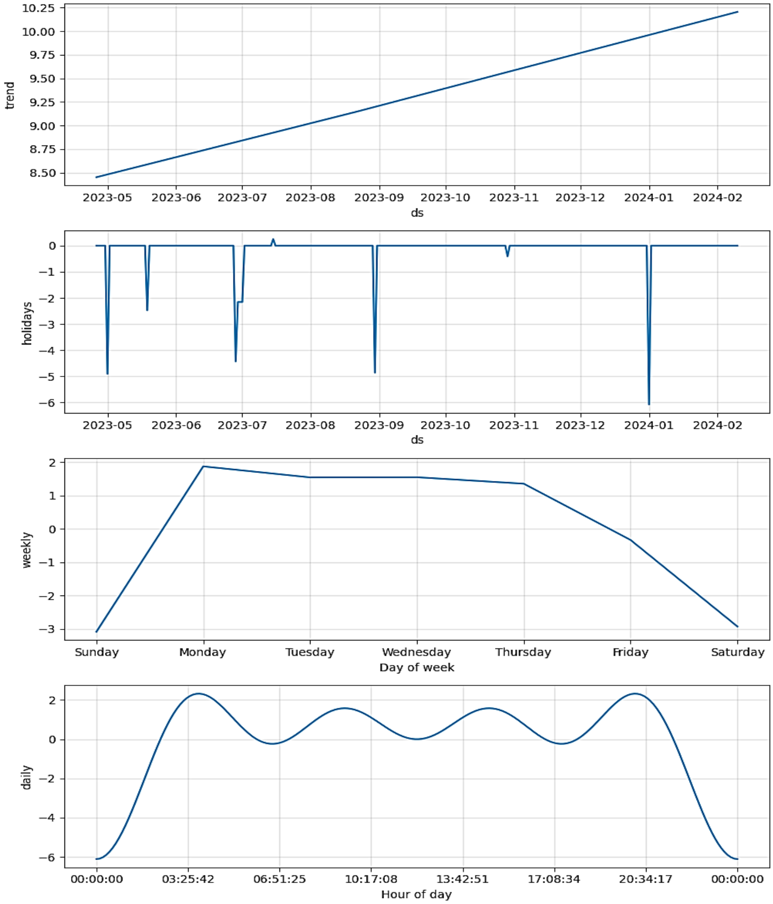


Figure 4. *Component graph from the Prophet model of the Product 1 sales data set.*

The time series analysis conducted using the Prophet model reveals an increasing trend in the sales of Product 1 from May 2023 to February 2024. The component graph provided by Prophet, shown in Figure 4, indicates that this linear and positive trend in sales could potentially point to factors such as an expanding customer base or an increase in market demand. The model's "holidays" graph reveals that special days have a negative impact on sales. A decrease in sales is observed particularly before and after holiday periods, with sales dropping to their lowest levels on the holidays themselves. This could be attributed to the influence of holidays on customer behavior and the business being closed during these periods. The weekly component analysis shows significant differences in sales across the days of the week, with the highest sales volume reached on Mondays and a gradual decline observed throughout the week until Friday. The daily component graph demonstrates how sales fluctuate during different hours of the day. All these component graphs provide valuable insights into the dynamics of sales for Product 1 over time, supporting decision-making processes in business management.

Figure 5, Figure 6, and Figure 7 below present the forecasts and confidence intervals for three separate products made using the Prophet model. The analyses conducted on the sales data of Product 1, Product 2, and Product 3 demonstrate the extent to which this model can capture different sales dynamics and manage the confidence intervals in making forward-looking predictions.

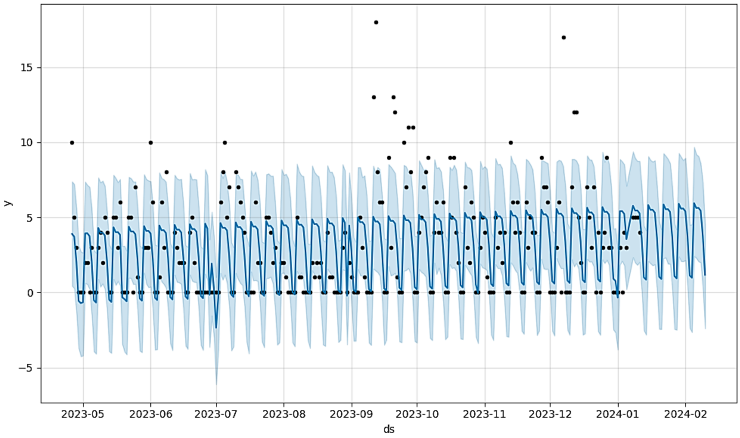


Figure 5. *Forecasts and confidence intervals for Product 1.*

The evaluation on Figure 5 for Product 1 shows that the Prophet model exhibits a good fit with the current sales (illustrated by the dark blue line along the plane with black dots) and future forecasts (continuing as a dark blue line beyond the black dots). The confidence interval, represented by the light blue area surrounding the model predictions, has remained almost constant in width throughout the forecasting period. This indicates that the model has efficiently adapted to past data and demonstrated consistency in future forecasts. This harmony between actual sales data and the model's predictions indicates that the model has the capacity to make sense of historical data and can produce reliable forward-looking forecasts.

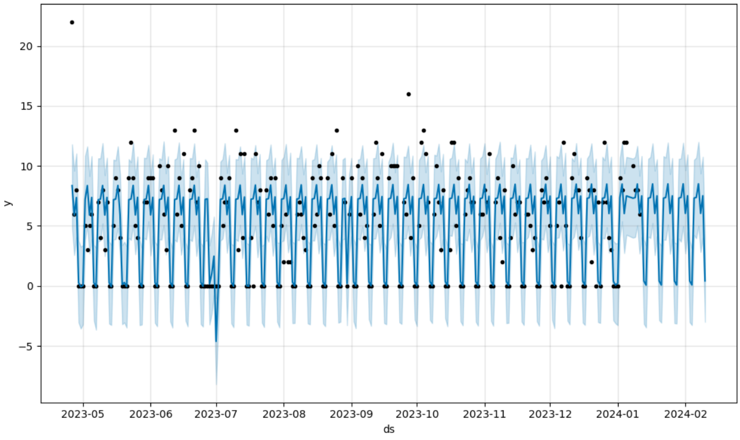


Figure 6. *Forecasts and confidence intervals for Product 2.*

In the analysis of Product 2 presented in Figure 6, it can be seen that the predictions of the model successfully follow the general trend of the actual sales data, with the confidence interval remaining at a regular area around the predicted trend line. However, the widening of the confidence interval in certain periods indicates that the model's predictions are less accurate in these periods and the data points show a greater variance. That said, the model was generally consistent in its predictions.

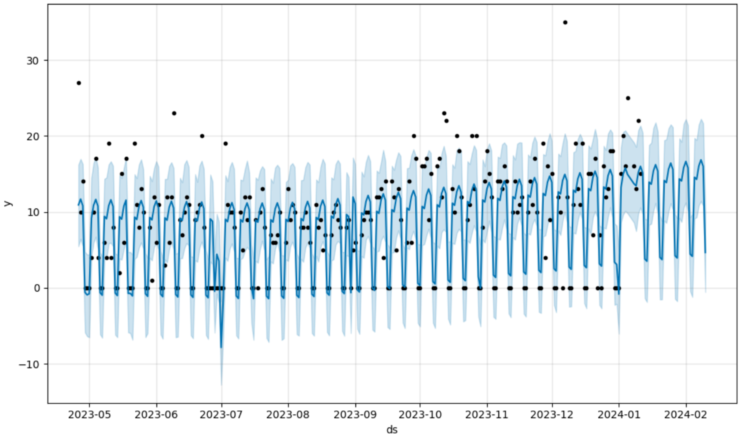


Figure 7. *Forecasts and confidence intervals for Product 3.*

As can be seen in Figure 7 above regarding Product 3, the prediction line produced by the model deviates from the real data from time to time. Although these deviations are not very large, they may indicate that the model is not always able to effectively filter out noise and high variability in the data set. The widening and narrowing of the confidence interval over time also confirms this situation. When the confidence interval widens, model predictions contain higher uncertainty and, accordingly, the reliability level decreases. The performance of the model for all three products contributed to understanding the model's strengths and weaknesses. This provided valuable insights regarding its interaction with real-world data.

To measure the prediction performance of the model, daily sales data for the last 30 days of each product in the data set were separated from the training data and used as the data set for testing. Following this approach made it possible to evaluate the degree to which the model's daily sales forecasts in each product deviated from the actual sales data for the following 30 days. Four distinct error metrics (Buitinck et. al, 2013) were implemented to evaluate the forecast's performance: Mean Absolute Error (MAE) (Eq. 1), Mean Absolute Percentage Error (MAPE) (Eq. 2), Mean Squared Error (MSE) (Eq. 3), and Root Mean Square Error (RMSE) (Eq. 4).

$$MAE = \frac{1}{n}\sum^{n}|yi - \hat{y}i| \qquad Eq.1$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{yi - \hat{y}i}{yi}\right| 100\% \qquad Eq.2$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2 \qquad Eq.3$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2} \qquad Eq.4$$

The rrror metrics reveal a detailed analysis of the performance of the model by considering the deviations of the predictions from real values across various dimensions. The RMSE and MAE metrics indicate how far the predictions are from the real values in terms of absolute and squared errors, whereas the MAPE metric uses a percentage technique to express this condition. The MSE error metric aids in observing the performance of the model from a different perspective by lending more weight to larger errors.

Table 1. *Evaluating the Deviations in the Prophet Model's Forecasts for Various Products in the Dataset Using Different Metrics.*

Data	MAE	MAPE	MSE	RMSE
Product 1	1.79	0.17	0.06	2.53
Product 2	4.33	0.19	0.09	5.98
Product 3	0.76	0.18	0.02	0.82

As can be understood from Table 1, reaching different error scores for three different products indicates that the model exhibits varying performances for different products. The varying daily sales quantities and sales trends of each product are also factors that affect the model's performance differently across products. For the first product, low MAE (1.79) and RMSE (2.53) scores demonstrate that the model generally performs well in predictions. The low MAPE (0.17) value for the same product indicates that the model's predictions are consistent in percentage terms. However, slightly higher MSE (0.06) and RMSE (2.53) scores suggest that deviations in predictions for some instances are larger. For the second product, somewhat higher MAE (4.33) and RMSE (5.98) scores indicate that the model makes more errors in predictions. A high MAPE (0.19) score further suggests that sales predictions for this product are less consistent compared to other products. Lastly, for the third product, lower MAE (0.76) and RMSE (0.82) scores reveal that the model produces robust predictions, and the obtained MAPE (0.18) score also shows that sales predictions for the product are consistent in percentage terms.

As a result, by using all error metrics together, the performance of the predictions produced by the model for different products has been analyzed in detail. For some products, the model predicts more reliably and accurately, while it has greater error rates for others. All of this information demonstrates how well the model can understand and predict each product's unique sales trends.

Future Work and Recommendations

This study used Facebook's Prophet model to predict restaurant sales, revealing the positive effects of external factors, such as weather and special days, on sales forecasts. Despite the success of the results, there are several recommendations for further improvement and expansion of the study and directions for future work.

The dataset used in this study is based on a specific enterprise and geographical location. In the future, the scope of the study could be expanded with data from different geographical regions and various businesses. This would improve the model's sensitivity to different market dynamics and customer preferences. Moreover, analyzing data gathered over years that include different seasons can help the model make better sense of long-term trends and seasonal variations. This research was limited in what was considered external factors, only weather conditions and special days (holidays, festivals, etc.). Future studies could integrate other external elements, such as business promotions, campaigns, local events, and economic indicators, into the model. Integrating these additional elements into the model can allow businesses to plan their marketing and operational strategies more efficiently by enabling more precise sales forecasts. As an alternative to Facebook's Prophet algorithm, other time series forecasting models and machine learning-based approaches can be tested by applying them to similar data sets. Comparing the forecasting performances of different models will reveal the strengths and weaknesses of each model and support the selection of the most appropriate model. A comparative analysis of deep learning models and traditional statistical approaches can provide a broader understanding.

Future studies in line with these recommendations may offer innovative solutions that will enable businesses to gain a competitive advantage in the market by increasing the efficiency of sales forecasting models. This study provides a basis for future research and development in the field of sales forecasting in the restaurant industry.

Conclusion

This study comprehensively analyzed the effectiveness of Facebook’s Prophet model in forecasting business sales and the effects of external factors on this process. The model provided insights into the specific challenges faced in the restaurant industry, particularly by accounting for the effects of external factors on sales, such as weather conditions and special days. The results of this research can be useful for businesses in terms of improving their sales forecasting strategies and optimizing their operational efficiency. Furthermore, the analyses provide valuable insights that may allow restaurants to develop future sales strategies in a more effective way.

Peer-review: Externally peer-reviewed.

Ethics Committee Approval: The corresponding author declared that this study falls under the category of research that does not require ethics committee approval for any reason.

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

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

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics, 170, 321-335. https://doi.org/10.1016/j.ijpe.2015.09.037

Badorf, F., & Hoberg, K. (2020). The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. Journal of Retailing and Consumer Services, 52, 101921. https://doi.org/10.1016/j.jretconser.2019.101921

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238.

Jha, B. K., & Pande, S. (2021, April). Time series forecasting model for supermarket sales using FB-prophet. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 547-554). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9418184

Loureiro, A. L., Miguéis, V. L., & Da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. Decision Support Systems, 114, 81-93. https://doi.org/10.1016/j.dss.2018.08.009

OpenWeatherMap. (2024). API documentation. OpenWeatherMap. Retrieved from https://openweathermap.org/api

Posch, K., Truden, C., Hungerländer, P., & Pilz, J. (2022). A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants. International Journal of Forecasting, 38(1), 321-338. https://doi.org/10.1016/j.ijforecast.2021.02.008

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. Robotics and Computer-Integrated Manufacturing, 34, 151-163. https://doi.org/10.1016/j.rcim.2014.12.001

Shilong, Z. (2021, January). Machine learning model for sales forecasting by using XGBoost. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 480-483). IEEE. https://doi.org/10.1109/ICCECE51280.2021.9342336

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45. https://doi.org/10.1080/00031305.2017.1380080

Thomassey, S. (2010). Sales forecasts in the clothing industry: The key success factor of the supply chain management. International Journal of Production Economics, 128(2), 470-483. https://doi.org/10.1016/j.ijpe.2010.07.007

Tsoumakas, G. (2019). A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1), 441-447. https://doi.org/10.1007/s10462-018-9656-1

Yusof, U. K., Khalid, M. N. A., Hussain, A., & Shamsudin, H. (2020, December). Financial time series forecasting using Prophet. In International Conference of Reliable Information and Communication Technology (pp. 485-495). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-33582-3_42

Zunic, E., Korjenic, K., Hodzic, K., & Donko, D. (2020). Application of facebook's prophet algorithm for successful sales forecasting based on real-world data. arXiv preprint arXiv:2005.07575.

Journal of Data Applications
Issue: 2, 2023
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

Bibliometric Analysis in Scientific Research Using R: A Review of Scopus and Web of Science Databases

Mehmet YILDIZ[1], Türkan KARAKUŞ YILMAZ[2]

ABSTRACT

The number of academic studies is increasing with the widespread use of digital tools. This increase makes it difficult to identify research trends, tendencies, themes, and other important points in various fields. Bibliometric analysis allows for the mapping of a field by analysing several academic studies and extracting meaningful conclusions. Therefore, bibliometric studies that are well conducted can build firm foundations for advancing a field in novel and meaningful ways. Bibliometric analysis, which is one of the approaches frequently used by researchers in the process of obtaining meaningful information from big data, has gained popularity especially in recent years. In this study, the bibliometrix package in the R programming language, which is frequently used in bibliometric analyses, is introduced and various analyses applied in the study are shown. Within the scope of this research, studies involving the R bibliometrix package were examined as sample applications. While bibliometric studies usually include the Scopus or Web of Science (WOS) databases, this study includes data obtained from both databases. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) steps were followed in the study. Accordingly, it provides an overview of bibliometric analysis, focussing in particular on its different techniques and providing step-by-step instructions on the procedures to be performed to perform bibliometric analysis with confidence. In the study, bibliometric analyses of 957 bibliometric studies obtained from the WOS and Scopus databases were also conducted. Because of the analysis, descriptive statistics such as year, author, journal, frequently used words, citations, etc. were presented.

Keywords: Bibliometric Analysis, Biblioshiny, Bibliometrix, R Studio, Science mapping

DOI: 10.26650/JODA.1462396

[1]Sakarya University of Applied Sciences, Distance Education Research and Application Center, Sakarya, Türkiye

[2]Atatürk University, Kazım Karabekir Faculty of Education, Erzurum, Türkiye

ORCID: M.Y. 0000-0002-9523-3805;
T.K.Y. 0000-0002-5809-3962

Corresponding author:
Mehmet YILDIZ,
Sakarya University of Applied Sciences, Distance Education Research and Application Center,
Sakarya, Türkiye
E-mail: mehmetyildiz@subu.edu.tr

Submitted: 31.03.2024
Revision Requested: 17.04.2024
Last Revision Received: 05.05.2024
Accepted: 23.05.2024
Published Online: 25.06.2024

Citation: Yıldız, M., & Karakuş Yılmaz, T (2023). Bibliometric analysis in scientific research using R: A review of Scopus and Web of Science databases. *Journal of Data Applications, 2*, 31-46.
https://doi.org/10.26650/JODA.1462396

Introduction

The development of electronic publishing with the widespread availability and accessibility of digital tools, has led to a significant increase in the number of academic publications. The increase in the number of acemetic publications has made it increasingly difficult to analyse these studies. The increasing number and desire to reveal the development of specific fields and map various fields has led to the development of various computer-based programmes. The number of citations, keywords, titles, abstracts, etc. from various databases has facilitated the analysis of studies using computer programmes. Bibliometric analysis (Donthu et al., 2021), which allows the examination of scientific development in various fields through publications, citations, keywords, authors, journals, countries, sources, etc. using mathematical and statistical methods, provides convenience for researchers in terms of evaluating a large number of studies. Bibliometric studies enable us the identification of trends in this field by quantitatively measuring and evaluating some aspects of research in a particular field (Ahmi, 2022). With bibliometric analysis, it is possible to follow the studies, researchers, institutions, and scientific flow related to the scientific subject (Martí-Parreño et al., 2016).

Bibliometric analysis, which applies quantitative techniques to bibliometric data, has gained popularity in recent years because it provides a systematic, transparent, and consistent review (Kasaraneni & Rosaline, 2022). Thanks to scientific databases, it is possible to access large amounts of data in line with the desired bibliometric data. For example, many databases such as Web of Science (WOS), Scopus, PubMed, and ProQuest offer data in various formats. In parallel with the data provided by databases, various computer software has started to be developed for bibliometric analysis. RStudio Biblioshiny (Aria & Cuccurullo, 2017), VOSviewer (van Eck & Waltman, 2010), CitNetExplorer (van Eck & Waltman, 2014), CiteSpace (Chen, 2006), and SciMAT (Cobo et al., 2012) are examples of these programmes that perform bibliometric or social network analysis.

The reasons why researchers benefit from bibliometric analysis are as follows: (1) gaining an overview at a glance, (2) identifying gaps in the field, (3) generating new ideas for research, and (4) positioning contributions to the field (Donthu et al., 2021). In bibliometric studies, the data used within the scope of the study should be large in order to analyse the data and reveal trends (Rashid, 2023). If the total number of academic studies used for analysis is small, satisfactory results may not be obtained. It is possible to manually analyse some data, and databases usually offer various statistical data.

The study provided a comprehensive guide on how bibliometric analysis, which offers important opportunities for researchers, is performed using the R programming language. This study provides important information on what bibliometric analysis is, what bibliometric concepts mean, and how to perform the analysis step by step. The study also provides explanations on how to conduct the analysis process by combining WOS and Scopus data, which are the most widely used databases in the field, as opposed to the analysis usually carried out from a single database.

This study also analysed bibliometric studies published in the WOS and Scopus databases. In this context, findings such as frequently used words, journals where bibliometric studies are commonly published, authors, countries, etc. are presented.

Research Questions;

1. How to conduct a bibliometric analysis?
2. What is the distribution of bibliometric studies by years?
3. What is the distribution of bibliometric studies by source?
4. What is the distribution of bibliometric studies by authors?
5. What is the distribution of bibliometric studies by country?
6. What are the frequently used keywords in title-abstract-keywords in bibliometric studies and how is the distribution of these words?
7. What is the distribution of frequently used keywords?

Literature Review

In this section, the researcher will briefly discuss the advantages of bibliometric analysis using the R bibliometrix package, the significance of bibliometric analysis, and various concepts used in bibliometric analysis.

Bibliometric Analysis

It is noteworthy that the emergence of scientific databases such as Scopus and Web of Science has made acquiring large volumes of bibliometric data relatively easy, and bibliometric software such as Gephi, Leximancer, and VOSviewer enable the analysis of such data in a very pragmatic way, thereby raising scholarly interest in bibliometric analysis in recent times.

Bibliometric research is a quantitative method used for the analysis, evaluation, and monitoring of published research articles (Župič & Čater, 2014). It involves the application of statistical and mathematical techniques to bibliographic data to identify patterns, trends, and relationships within a specific field of study (Liu, 2023). By utilising bibliometric methods, researchers can gain insights into the intellectual structure of a research field, track research trends, and evaluate the impact of publications (Ramos-Rodríguez & Ruíz-Navarro, 2004).

Bibliometric research plays a crucial role in academia by providing a systematic and transparent way to analyse scientific literature (Khoshroo & Talari, 2022). It allows for assessing research performance at various levels, such as journals, researchers, countries, and institutions (Singh & Arora, 2022). Through bibliometric analysis, researchers can identify research hotspots, visualise trends, and inform future research directions (Luo, 2023). Additionally, bibliometrics help in recognising the intellectual structure of a field of knowledge and assessing the scientific influence of research outputs (Ruiz–Real et al., 2018).

Moreover, bibliometric studies contribute to the advancement of knowledge by offering comprehensive overviews of research fields and highlighting key contributors and communities within those fields (Mahato et al., 2022). They aid in understanding the evolution of scientific areas over time and provide valuable insights for researchers to navigate through the vast landscape of scholarly publications. Bibliometric analysis also assists in identifying research gaps, shaping research agendas, and promoting evidence-based decision-making in various disciplines.

In conclusion, bibliometric research serves as a valuable tool for researchers to analyse, evaluate, and navigate the vast landscape of scholarly publications. It enables the systematic assessment of research performance, identification of trends, and shaping of future research agendas across diverse fields of study. By leveraging bibliometric methods, researchers can gain valuable insights that contribute to the advancement of knowledge and facilitate evidence-based decision-making in academia.

Bibliometric Analysis Using R

Bibliometric analysis provides an overview and in-depth analysis of research conducted in a particular subject area. In this study, we present a guide on how to perform bibliometric analysis using the R bibliometrix package in RStudio with the biblioshiny plugin. Biblioshiny produces research output to determine the analysis of annual scientific production, the most prolific authors, the most frequently used words, the most popular journals, cross-country collaborations, etc. related to the selected research topic (Rashid, 2023). Among the reasons for choosing the R programming language in the study; features such as making data analysis easier and faster, visualisations and data are clearer, there are many features that can be used for analysis, and Biblioshiny's web interface is constantly updated. In addition, the R programming language is an open source, free software. One of the important reasons for choosing the R programming language is that it allows analysis by combining data from multiple databases, as shown in this study.

Method

In this section, processes such as obtaining data from databases, transforming them, filtering them, making them ready for analysis, analysing them, and obtaining and interpreting the findings are included.

In addition, the processes of making bibliometric studies ready for analysis are mentioned in the study.

Research Design

In this study, a bibliometric mapping technique was used based on various criteria such as author, publication, keywords, country, and number of citations. Bibliometric mapping is a visual representation of the links among disciplines, fields, specific publications, and authors (Donthu et al., 2021). Bibliometric studies enable the detection of trends in the field by measuring and evaluating some aspects of research in a particular field (Ahmi, 2022). It is possible to follow the studies, researchers, institutions, and scientific flow related to the determined scientific subject using bibliometric analysis (Martí-Parreño et al., 2016).

Data Collection

In bibliometric analyses, it is necessary to determine which databases will be searched with which keywords and the inclusion and exclusion criteria of the studies accessed. The scope in which the keywords selected in the searches will be searched is also an important stage that affects the results of the study. Figure 1 shows a sample search conducted within the scope of this study. The search was conducted on 30.03.2024. Information about the database search query is shown in Figure 1.

- Web of Science (https://www.webofscience.com/wos/woscc/advanced-search)
- Scopus (https://www.scopus.com/search/form.uri?display=advanced)

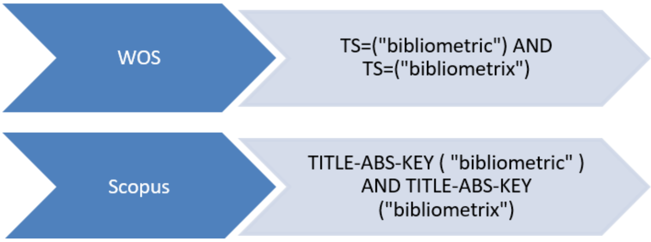


Figure 1. *Keyword search process based on databases.*

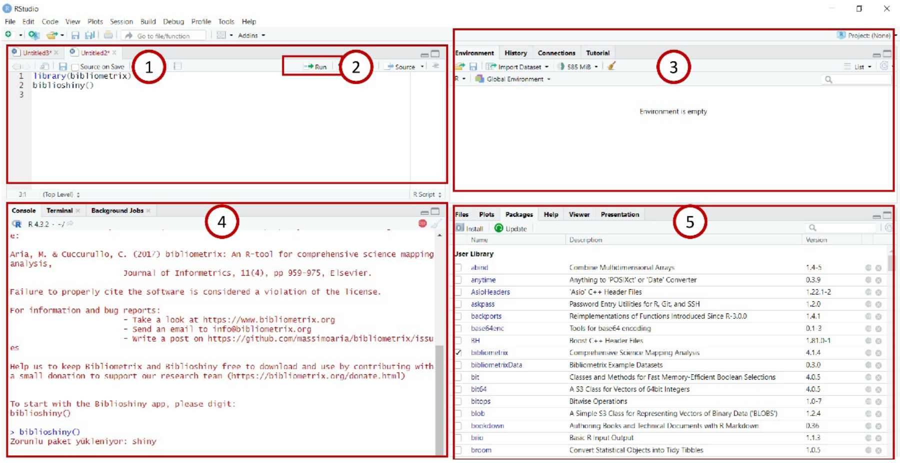
Figure 1 Keywords and criteria according to which these keywords were scanned. As in many bibliometric analyses, both databases were searched within the scope of title-summary-keywords. The data obtained from the study should be limited according to various criteria (e.g. language, study type, year, etc.) and downloaded in the appropriate format. While Scopus data were downloaded in BibTeX format, WOS data were manually merged into two parts in.txt format, as more than 500 data points were not allowed to be downloaded at the same time. The inclusion criteria were that the studies were in English and the document type was an article.

Data Transformation

In this section, the processes of converting the data obtained from the databases into Excel format using the biblioshiny plug-in of the R programming language, merging the data, and removing the same studies from the data pool are discussed. Through the R Studio programme;

Library(bibliometrix) ---> biblioshiny() codes are run to open the biblioshiny plugin.

The RStudio interface is shown in Figure 2.



```
library(bibliometrix)
biblioshiny()
```

Environment is empty

```
e:

Aria, M. & Cuccurullo, C. (2017) bibliometrix: An R-tool for comprehensive science mapping
analysis,
                        Journal of Informetrics, 11(4), pp 959-975, Elsevier.

Failure to properly cite the software is considered a violation of the license.

For information and bug reports:
                        - Take a look at https://www.bibliometrix.org
                        - Send an email to info@bibliometrix.org
                        - Write a post on https://github.com/massimoaria/bibliometrix/issu
es

Help us to keep Bibliometrix and Biblioshiny free to download and use by contributing with
a small donation to support our research team (https://bibliometrix.org/donate.html)


To start with the Biblioshiny app, please digit:
biblioshiny()

> biblioshiny()
Zorunlu paket yükleniyor: shiny
```

	Name	Description	Version
	User Library		
□	abind	Combine Multidimensional Arrays	1.4-5
□	anytime	Anything to 'POSIXct' or 'Date' Converter	0.3.9
□	AsioHeaders	'Asio' C++ Header Files	1.22.1-2
□	askpass	Password Entry Utilities for R, Git, and SSH	1.2.0
□	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.4.1
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.81.0-1
☑	bibliometrix	Comprehensive Science Mapping Analysis	4.1.4
□	bibliometrixData	Bibliometrix Example Datasets	0.3.0
□	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.5
□	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5
□	bitops	Bitwise Operations	1.0-7
□	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.4
□	bookdown	Authoring Books and Technical Documents with R Markdown	0.36
□	brio	Basic R Input Output	1.1.3
□	broom	Convert Statistical Objects into Tidy Tibbles	1.0.5

Figure 2. *RStudio interface.*

When the numbered sections in Figure 2 are examined in order;

1. Code writing screen
2. Command execution screen
3. Data loading and preview screen
4. Console screen
5. Can be expressed as file, output, or package preview screen

For the analysis, first, the files downloaded from the databases should be uploaded to the programme using the R bibliometrix package and then outputted as Excel files.

After the above steps are taken, the “wos.xlsx” and “scopus.xlsx” files obtained as excel files can be edited according to the needs of the researcher. The following steps are followed in the process of merging the data.

1. Select the folder containing the database files by following the Session->Set Working Directory->Choose Directory steps through the RStudio programme.
2. Adding libraries
   a. library(bibliometrix),
   b. library(openxlsx)
   c. library(xlsx)
   d. library(fBasics)
3. Writing code
   a. Web_data<-convert2df("wos.txt")

```
Scopus_data<-convert2df("scopus.bib",dbsource="scopus",format="bibtex")
combined<-mergeDbSources(web_data,scopus_data,remove.duplicated=T)
write.xlsx(combined,"combinedabs.xlsx")
```

When the above steps are followed, it will be seen that the programme will combine two separate excel files and remove the duplicate data and output as a single excel file. The data collection process of the bibliometric studies analysed within the scope of the study is shown in Figure 3.

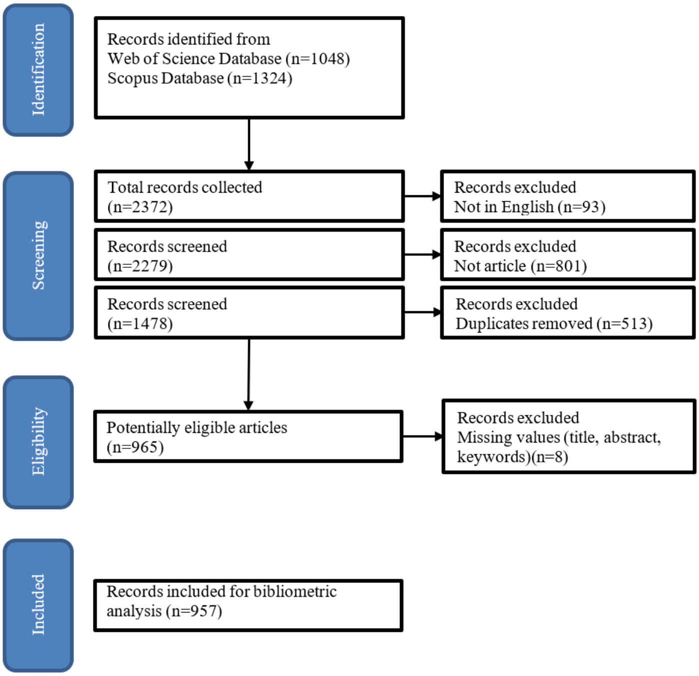


Figure 3. *Study data selection and filtering.*

Results

In this section, the sample findings are presented in parallel with the data obtained within the scope of the study. The main findings of the study are shown in Table 1. Table 1 shows the statistical information of the publications included in the study. Table 1 presents many findings such as the year range in which the studies were published, the number of sources in which the studies were published, the number of articles, the annual rate of increase in the number of studies, authors, and keywords.

When the findings are analysed, it can be seen that the studies conducted using the R bibliometrix package were first published in 2017. This is also the date of the first release of the R bibliometrix package. The 957 articles analysed within the scope of this research were published in 552 different sources. When the annual increase rate of bibliometric studies is analysed, an increase of 80.74% is observed. A total of 2757 different authors conducted the studies, and 41 persons published as a single author. While 17.66% of the studies had international co-authors, each study by approximately five authors. It was observed that 2337 keywords were included in the analysis. Although the articles are relatively recent, the average number of citations is 15.34.

The number of articles published by years is shown in Figure 4.

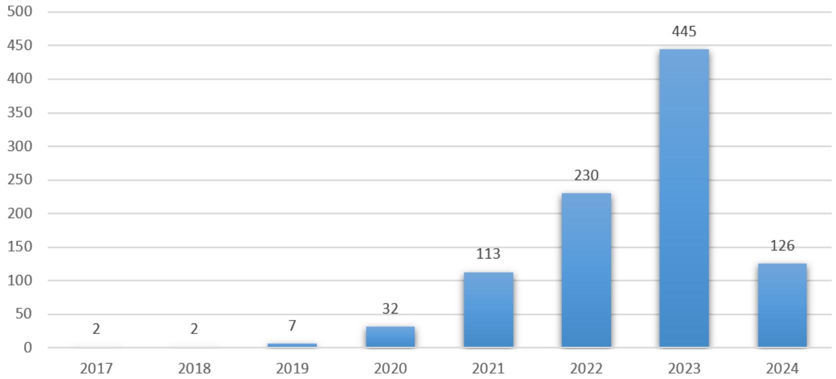


Figure 4. *Number of articles published by the year.*

Figure 4 shows that the number of studies is gradually increasing. The number of studies has increased, especially in recent years, with the highest number of studies published in 2023 (445). Although 2024 (126) is only the beginning of the year, the number of published studies is higher than that in 2021 (113). Since the R bibliometrix package was released in 2017, there have been no studies from previous years.

The sources in which the bibliometric articles were most frequently published in the study are shown in Figure 5.

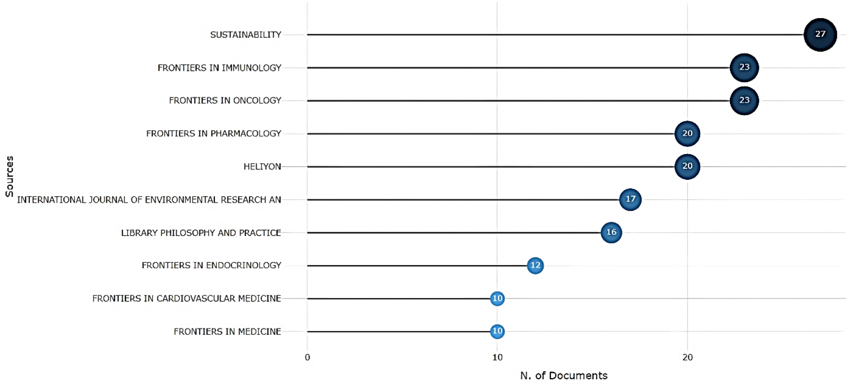


Figure 5. *Sources where bibliometric articles are frequently published.*

Figure 5 shows that most bibliometric studies were published in the journal "Sustainability" (27). In "Frontiers in Immunology" (23), "Frontiers in Ontology" (23), "Frontiers in Pharmacology" (20), and "Heliyon" (20) are similarly the sources where most bibliometric studies are published.

The authors with the most bibliometric articles are shown in Figure 6.

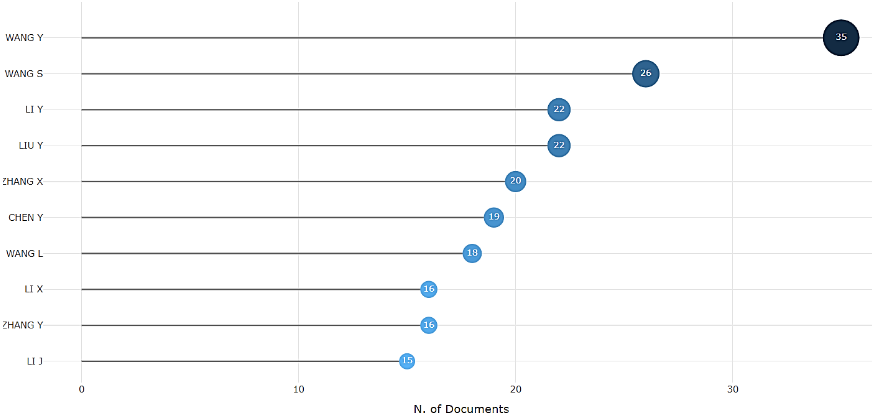


Figure 6. *Authors with the most bibliometric articles.*

Figure 6 shows the top 10 authors with the most bibliometric publications. When the number of publications is analysed, Wang, Y. (35), Wang, S. (26), Li, Y. (22), Liu, Y. (22), and Zhang, X. (20) are the authors with the highest number of publications.

The countries' total bibliometric article production is shown in Figure 7.

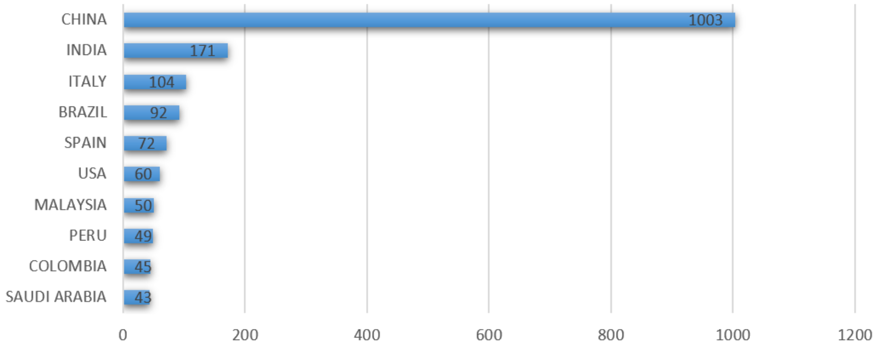


Figure 7. *Countries' total numbers of bibliometric publications.*

Figure 7 shows the total number of bibliometric publications in each country. Although only article-type publications were analysed in the study findings, publications other than articles were also included in the total number of publications. According to the findings, China (1003) is the country with the highest number of publications, and the number of publications is considerably higher than that of other countries. When we see other countries in India 171, Italy 104, Brazil 92, and Spain 72 articles published. In Türkiye, 26 articles were published.

The three-field plot of the words frequently used in the title-abstract-keywords is shown in Figure 8.

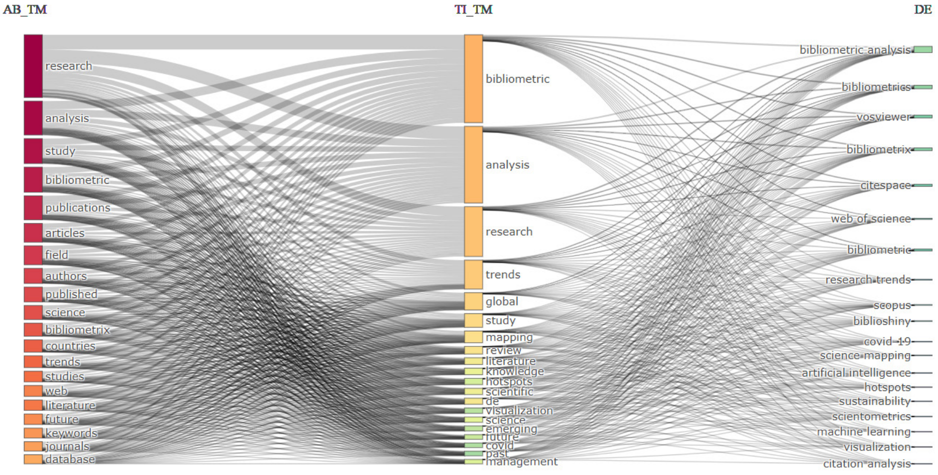


Figure 8. *Three-field plot (*AB_TM*: Title, TI_TM*: Abstract, DE*: Keywords).

Figure 8 shows that the most frequently used words in the title are research (12600), analysis (6830), study (5004), bibliometric (5000), and publications (4820). In the abstract, bibliometrics (17700), analysis (15400), research (10000), trends (5830), and global (3470) are the most frequently used words. Similarly, words such as bibliometric analysis (1260), bibliometrics (684), however (568), bibliometrix (545), and citespace (421) were frequently used as keywords.

The frequently used keywords in the study are shown in Figure 9.

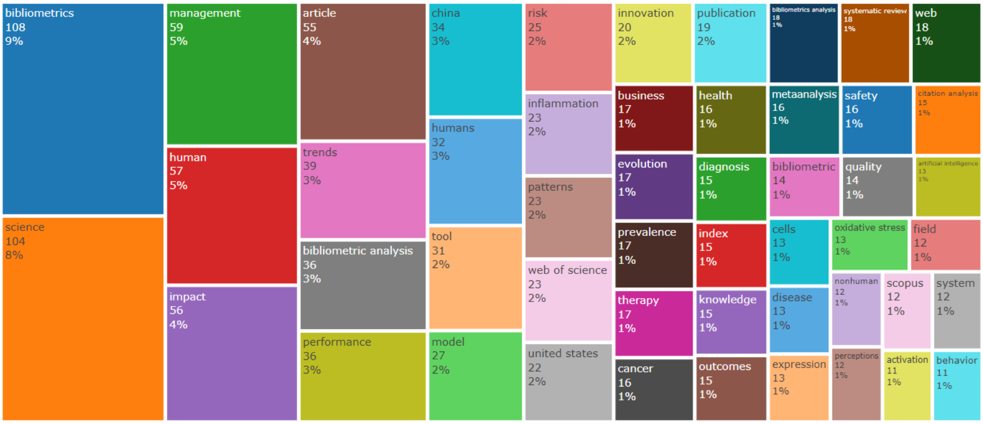


Figure 9. *Frequently used keywords in the study.*

Figure 9 shows the numbers and percentages of the 50 most frequently used words in the studies analysed within the scope of the study. The most frequently used words are bibliometrics (108), science (104), management (59), human (57), and impact (56).

Discussion and Conclusion

In this study, there are explanations about what bibliometric analysis is, how it is analysed findings are obtained. Within the scope of the study, we aimed to introduce the R bibliometrix package and biblioshiny interface in the R programming language, and a step-by-step bibliometric analysis was applied. Within the scope of the study, the stages of keyword selection, database selection, downloading, filtering, transforming, and preparing the data for analysis are shown with sample applications. The reporting process was also presented to the researchers during the data acquisition stages.

One of the most important points in data analysis in bibliometric studies is keyword selection (Yadav et al., 2023). The choice of keywords in bibliometric analysis plays a crucial role in the quality and usefulness of the analysis. Properly selected keywords make the literature review more effective, accurately reflect research trends, and help identify gaps in the field. Therefore, researchers must carefully select keywords and ensure that the words they choose accurately reflect the scope and focus of the research. Keyword selection not only improves the accuracy of bibliometric analyses but also facilitates the scientific community's access to and understanding of relevant literature, thus contributing to the advancement of knowledge. Otherwise, it may be difficult to interpret the findings of bibliometric studies, which by their nature present more superficial findings and do not have the opportunity for in-depth analysis.

The second important stage of bibliometric studies is database selection. Because various digital tools facilitate bibliometric analysis, it is often the case that the database is chosen for the tool rather than the tool for the database. However, the scope of the study should be determined in line with the focus of the research and the needs of the literature. While the tools developed for bibliometric analysis mostly work with a single database, in this study, the findings obtained from multiple databases were analysed together. In this process, the data downloaded from WOS and Scopus databases are used in the RStudio programme and both file type conversion and data merging are provided with the help of the codes mentioned in the method section. One of the important operations performed in this study is the removal of duplicate data during data merging. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process should be considered in the bibliometric analysis process. Accordingly, the keywords, publication type, year, etc. used in the screening process should be expressed step by step. After keyword selection, downloading, transforming, merging, and removing duplicate data, the analysis process can begin.

When the findings are examined, it is seen that since the R bibliometrix package was first published in 2017, the first study was published in 2017. The 957 articles analysed within the scope of this research were published in 552 different sources. Compared with the number of articles, there is variety in the number of sources.

When the annual increase rate of bibliometric studies is analysed, it is observed that there is a great rate of 80.74%. As a matter of fact, the fact that there are nearly 1000 articles in approximately 3-4 years is one of the most important indicators of this. One of the reasons why these studies are published so intensively is that they do not require processes such as long-term application and analysis. The fact that it does not have a high economic cost is also one of the main reasons why bibliometric studies are widespread. In addition to the fact that they can be written and published in a short time, the high number of citations can also be cited as a reason for the widespread use of such studies. The fact that 2757 authors conducted bibliometric studies supports this argument. At the same time, when we look at the average number of citations, it is seen that 15 citations are given for each article despite such a short time. This rate demonstrates why researchers attach so much importance to bibliometric studies.

When the journals in which bibliometric studies are most frequently published are examined, it can be seen that there are mostly journals in the field of health. There may be several reasons for this. The fact that the field of health has a dynamic structure with the use of new methods, techniques, treatments, and medicines, especially with technology, can be shown as a reason for this. The opportunity to work interdisciplinaryly and the high volume of publications with funding support may have prepared a suitable ground for bibliometric studies where a large amount of data is important. The COVID-19 pandemic may have also led to more bibliometric studies in the field of health (Korkmaz & Altuntaş, 2022).

When the number of publications is analysed, Wang, Y. (35), Wang, S. (26), Li, Y. (22), Liu, Y. (22), and Zhang, X. (20) are the authors with the highest number of publications. the number of publications for each author is quite high. Therefore, it can be said that the indicators that increase the number of bibliometric publications are valid for the authors.

When the number of publications by country is analysed, it is seen that China and the USA publish considerably more than other countries. The high number of publishers in the country can also be a reason for this. The one with the most publications is considered the most productive (Erdoğan, 2021). The United States and China are the most productive countries with the highest number of publications on nursing and COVID-19 (Korkmaz & Altuntaş, 2022). There is a relationship between the number of scientific publications published by countries, the duration of the Covid-19 pandemic, and the level of impact of these countries from the pandemic (Hao et al., 2020; Oh & Kim, 2020; Tao et al., 2020). In Turkey, 26 publications were made, and in this context, it is one of the 20 countries with the highest number of bibliometric studies.

One of the important findings of bibliometric studies is the three-field plot graph (Figure 7). this graph shows us the relationships between various topics. Since the frequency of use with other titles is also important in this graph, frequently used keywords can be seen differently in Figure 7 and 8 within the scope of the study. This finding shows that bibliometric analysis and similar words are frequently used in accordance with the keywords of the study.

When frequently used keywords are analysed, parallel results are obtained with the three-field plot graph. However, it is noticeable that more words are used in the field of health. The high number of bibliometric studies in the field of health confirms that the frequently used words are from the field of health.

In summary, this study provides comprehensive information on bibliometric data analysis. In this study, the bibliometric analysis process shows step-by-step all the processes such as data collection, filtering, organising, analysing and interpreting. This study can be a guide for researchers who want to conduct bibliometric analysis.

In bibliometric studies, where a large number of academic studies can be analysed at the same time, the emphasis on quantity and the presence of various errors can affect any analysis using such data (Rashid, 2023). To reduce errors, scholars should carefully select keywords and clean their bibliometric data, which involves removing duplicate and erroneous entries. Because only a limited number of databases were selected, there are limitations. In particular, the qualitative claims of bibliometrics can be highly subjective, given that bibliometric analysis is quantitative in nature and the relationship between quantitative and qualitative results is uncertain (Wallin, 2005). In this context, academics should be more careful when making qualitative claims about bibliometric observations and support them with content analysis when appropriate (Gaur & Kumar, 2018). Therefore, academics should avoid making overly ambitious claims about the research area and its long-term impact (Donthu et al., 2021). Despite these limitations, Bibliometric analysis helps academics gain a comprehensive perspective, identify knowledge gaps in the field, obtain research ideas, and find ways to make expected contributions to the field (Ellegaard & Wallin, 2015; Rashid, 2023).

Bibliometric studies using the separate R bibliometrix package were also analysed in the study. For beginners, utilising R packages is easier (Rashid, 2023). In this context, one of the results obtained within the scope of the study is that there is a great increase in the number of bibliometric studies, that they can be written and published in a short time, and that they can be highly cited.

Limitations

The research methodology used in this study has some limitations that should be considered when evaluating the results;

1. This study is limited to studies included in the WOS and Scopus databases.

2. Instead of publishing a large number of scientific articles, some authors contribute in different ways, for example, by taking active roles in important projects or initiatives, or by making a significant impact in their research field. Such contributions are often overlooked in bibliometric analyses.

3. This study focussed only on articles published in scientific journals; future studies may consider other sources such as books or proceedings.

4. This study focussed only on articles published in English; therefore, other languages can be considered in future studies.

5. In this study, only the R bibliometrix package was introduced with the RStudio programme and bibliometric studies published under the bibliometrix keyword were analysed. Different programmes and studies can be analysed in different studies.

Peer-review: Externally peer-reviewed.

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

Author Contributions: Conception/Design of Study- M.Y., T.K.Y.; Data Analysis/Interpretation- M.Y.; Drafting Manuscript- M.Y., T.K.Y.; Critical Revision of Manuscript- M.Y., T.K.Y.; Final Approval and Accountability- M.Y., T.K.Y.; Technical or Material Support- M.Y., T.K.Y.; Supervision- M.Y., T.K.Y.

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

Ahmi, A. (2022). *Bibliometric Analysis using R for Non-Coders: A practical handbook in conducting bibliometric analysis studies using Biblioshiny for Bibliometrix R package*.

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Chen, C. (2006). CiteSpaceII: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377. https://doi.org/10.1002/asi.20317

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609-1630. https://doi.org/10.1002/asi.22688

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W.M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res*, 133, 285–296.

Ellegaard, O., & Wallin, J.A. (2015), "The bibliometric analysis of scholarly production: how great is the impact?", Scientometrics, Vol. 105 No. 3, pp. 1809-1831.

Erdoğan, R. E. (2021). Analysis and mapping of computational design research in architecture with bibliometric methods. [Unpublished master's thesis]. Baskent University.

Hao, Y.-F., Peng, K., Mai, Q.-L., Meng, M.-Q., Wang, D., & Zhang, X.-Y. (2020). A bibliometric analysis of nursing research in COVID-19 in China. Journal of Integrative Nursing, 2(3), 116. https://doi.org/10. 4103/jin.jin_32_20

Gaur, A., & Kumar, M. (2018). A systematic approach to conducting review studies: An assessment of content analysis in 25 years of IB research. *Journal of World Business, 53*(2), 280–289.

Khoshroo, M., & Talari, M. (2022). Scientific mapping of digital transformation strategy research studies in the industry 4.0: a bibliometric analysis. *Nankai Business Review International*, 14(1), 3-34. https://doi.org/10.1108/nbri-03-2022-0021

Korkmaz, A. Ç., & Altuntaş, S. (2022). A bibliometric analysis of COVID-19 publications in nursing by visual mapping method. *Journal of Nursing Management*, *30*(6), 1892-1902.

Liu, N., Ji, Y., Liu, R., & Jin, X. (2023). The state of astragaloside iv research: a bibliometric and visualized analysis. *Clinical Pharmacology*;, 38(2), 208-224. https://doi.org/10.1111/fcp.12956

Luo, X., Yan, X., Yin, D., Xia, Y., Li, S., Shi, S., … & Zhou, J. (2023). A bibliometric systematic review of extracellular vesicles in eye diseases from 2003 to 2022. *Medicine*, *102*(33), e34831. https://doi.org/10.1097/md.0000000000034831

Martí-Parreño, J., Méndez-Ibáñez, E., & Alonso-Arroyo, A. (2016). The use of gamification in education: a bibliometric and text mining analysis. *Journal of computer assisted learning*, *32*(6), 663-676.

Oh, J., & Kim, A. (2020). A bibliometric analysis of COVID-19 research published in nursing journals. *Science Editing*, *7*(2), 118–124. https:// doi.org/10.6087/kcse.205

Ramos-Rodríguez, A. R., & Ruíz-Navarro, J. (2004). Changes in the intellectual structure of strategic management research: a bibliometric study of the strategic management journal, 1980–2000. Strategic Management Journal, 25(10), 981-1004. https://doi.org/10.1002/smj.397

Rashid, M.F.A. (2023). How to Conduct a Bibliometric Analysis using R Packages: A Comprehensive Guidelines. *Journal of Tourism, Hospitality & Culinary Arts, 15*(1), 24-39

Ruiz–Real, J. L., Uribe-Toril, J., Valenciano, J. d. P., & Gázquez–Abad, J. C. (2018). Worldwide research on circular economy and environment: a bibliometric analysis. International Journal of Environmental Research and Public Health, 15(12), 2699. https://doi.org/10.3390/ijerph15122699

Singh, N., & Arora, S. (2022). Recognizing the legacy of the tqm journal: a bibliometric analysis of scopus indexed publications (2008 - 2021). *The TQM Journal*, *35*(4), 946-963. https://doi.org/10.1108/tqm-01-2022-0002

Tao, Z., Zhou, S., Yao, R., Wen, K., Da, W., Meng, Y., Yang, K., Liu, H., & Tao, L. (2020). COVID-19 will stimulate a new coronavirus research breakthrough: A 20-year bibliometric analysis. *Annals of Translational Medicine*, *8*(8), 528. https://doi.org/10.21037/atm.2020.04.26

van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics, 8*(4), 802-823. https://doi.org/10.1016/j.joi.2014.07.006

van Eck, N. J., & Waltman, L. (2010). Software survey: VoSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Yadav, N., Luthra, S., & Garg, D. (2023). Blockchain technology for sustainable supply chains: a network cluster analysis and future research propositions. *Environmental Science and Pollution Research*, *30*(24), 64779-64799. https://doi.org/10.1007/s11356-023-27049-3

Župič, I., & Čater, T. (2014). Bibliometric methods in management and organization. *Organizational Research Methods, 18*(3), 429-472. https://doi.org/10.1177/1094428114562629

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*

		……../……../……………