SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Naive Bayes and KNN for Airline Passenger Satisfaction Classification: Comparative Analysis

Annisa Fitria Nurdina[1], Audita Bella Intan Puspita [2*]
[1,2]Department of Computer Science, Universitas Negeri Semarang, Indonesia

DOI: https://doi.org/10.52465/joiser.v1i2.167
Received 01 June 2023; Accepted 14 July 2023; Available online 14 July 2023

Article Info

Keywords:
Airlines passenger;
K-nearest neighbour;
Naive bayes;
Classification;
Data mining

Abstract

Air transportation is vital due to technological advancements and globalization. It is affordable and accessible worldwide, providing efficient services to reach destinations globally. This discussion focuses on full-service airlines that offer online-based services. Previous research indicates that available facilities and services influence passenger satisfaction. Previous research on customer satisfaction showed a correlation between satisfaction and services without accurate figures. In the present study, the customer satisfaction figure is measured using the Naive Bayes and K-Nearest Neighbour (K-NN) algorithm to obtain a tested level of accuracy. In this analysis, we will compare the effectiveness of Naive Bayes and K-NN algorithms in classifying airline passenger satisfaction. The results show that the accuracy of the Naive Bayes method of the two algorithms is higher than the K-NN method. The accuracy value of the Naive Bayes method is 84.48%, while the accuracy value of the K-NN method is 65.38%. From the test results, the precision value for Naive Bayes is 82.25%, and K-NN is 67.35%. Furthermore, the recall value for Naive Bayes is 82.43%, and K-NN is 74.33%.

1. Introduction

Air transportation has become an important part of today's modern life. Along with the development of technology and globalization, air transportation is becoming more affordable and easily accessible to many people worldwide. As a big business, the aviation industry provides efficient and effective air transportation services to reach destinations worldwide. In transportation, choosing air routes is the most popular alternative for those who want to get their goal quickly. Since users' needs differ, they can choose a flight that suits their schedule and budget to buy tickets [1].

Along with those aspects, customer satisfaction is the customer's response to evaluating the perceived discrepancy between previous expectations [2]. Service quality contributes positively to customer satisfaction, and customer loyalty provides added value to keep believing in the company's services [3]. Problems often occur in service quality and customer satisfaction because many companies

* ***Corresponding Author:***

Audita Bella Intan Puspita,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: auditabella829@students.unnes.ac.id

only consider the company's internal perspective rather than the customer's point of view. Therefore, companies need to create service strategies and programs that focus on customer interests by considering the service quality components to maximize the quality of passenger services following company goals [4].

This discussion will focus more on airlines in the full-service category. Airlines provide online-based services. In line with the development of the digital era, the services provided include online reservation, check-in, and e-ticketing. Previous research shows hat passenger satisfaction is associated with the facilities and available services [5].

If the company can ensure the implementation of several aspects according to standard operating procedures, the customer will be satisfied with the product or service provided. Customers will then compare the service with those received from other companies. If the company can provide satisfactory service, customers will feel very satisfied and tend to make repeat purchases and recommend the company to others. Therefore, companies need to seriously consider the importance of customer service to survive and compete in today's business market. Today's customer satisfaction has been recognized as vital to maintaining business and winning competition [6]. Data Mining is crucial for organizations as it is imperative to conduct thorough analyses determining the value associated with discovered patterns. This process ensures that the designs found are meaningful, valuable, and pertinent to the organization [7].

Various learning techniques have been developed, including supervised, unsupervised, and reinforcement learning. Among these, supervised learning methods teach computers to identify specific patterns within their data and associated labels or target values. Subsequently, when the trained model encounters data with similar designs, it is expected to make highly accurate predictions of the corresponding labels or target values while optimizing resource usage to the greatest extent possible [8].

Numerous research studies have been conducted to evaluate service quality and airline passengers' satisfaction levels. Some studies have employed conventional statistical testing, while others have utilized multiple-criteria methods to achieve similar objectives. Within the domain of machine learning, assessing airline passenger satisfaction commonly involves using sentiment analysis. This technique analyses text, tweets, or comments to identify whether the sentiment expressed is positive or negative regarding satisfaction [9].

In this comparative analysis, we will evaluate the effectiveness of the Naive Bayes and K-NN algorithms in classifying airline passenger satisfaction. We will analyze the dataset consisting of feedback from airline passengers and use both algorithms to classify the feedback as positive or negative. The accuracy of the classification results will be evaluated, and the advantages and disadvantages of each algorithm will be discussed [10].

2. Method

In this study, the authors used the K-Nearest Neighbour and Naive Bayes classification methods to classify the Airline Customer Satisfaction Dataset. This study aims to compare the two methods to determine which is better for predicting the accuracy of Airline Customer Satisfaction.

This study can be resolved through five process steps based on the description above. The flowchart model illustrating these steps can be seen in Figure 1 below.

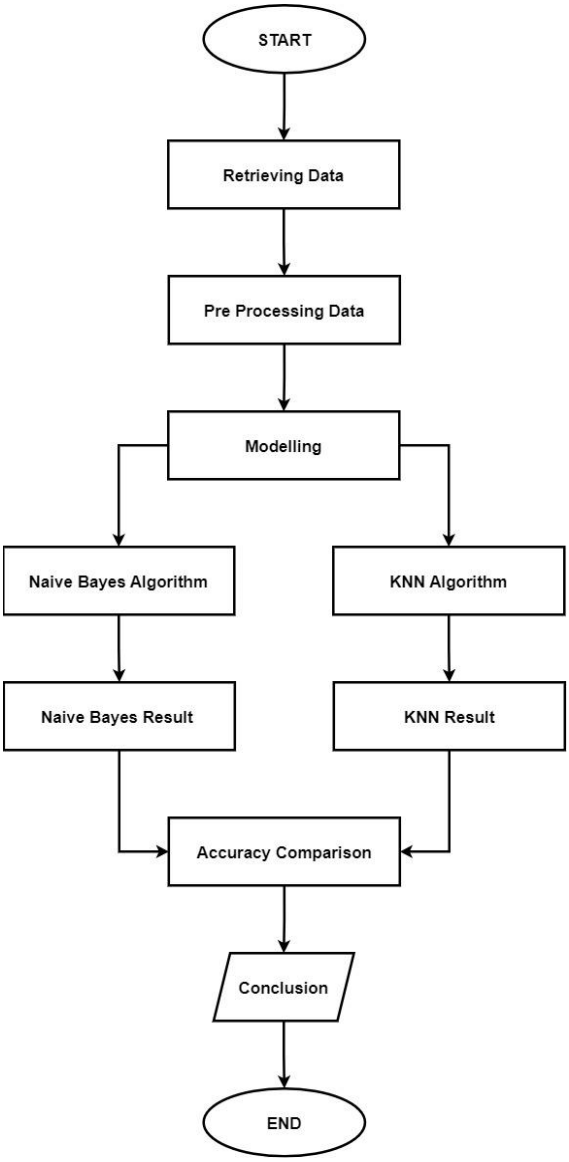

Figure 1. Flowchart of Research Framework

2.1 Data Mining

Data mining is a step in doing Knowledge Discovery in Databases (KDD) [11]. Many benefits can be obtained through data mining processing, which helps get valuable information and increase understanding of various data that can be analyzed using multiple algorithms[12]. Data mining is finding patterns contained in datasets with certain methods. This process is essential in creating new discoveries or knowledge from a dataset. One of the main roles of data mining is classification in classification utilizing data train to improve the model's quality and the analysis result [13].

Data mining places a strong emphasis on the precision of model predictions. A crucial indicator of effectiveness in data mining models is their capacity to make precise forecasts in practical scenarios. This focus on accuracy stems from the origins of data mining within the realm of Artificial Intelligence, which has always been concerned with developing applicable predictive models. These models have found utility in various real-world applications, including predicting insurance fraud, diagnosing illnesses, recognizing patterns, and more [14].

2.2 Dataset

Airline Customer Satisfaction is a dataset containing information about passenger satisfaction results after using airlines sourced from Kaggle.com. The number of data records in the dataset is 26 thousand, with a total of 9 attributes and one label. The attributes used in this dataset include flight distance, customer type, gender, age, class, type of travel, seat comfort, food and drink, and

departure/arrival time convenient. Some of the attributes mentioned refer to one label: the level of passenger satisfaction.

2.3 Classification

One of the goals that many generate in data mining is classification. Classification is a classification or grouping function that explains or distinguishes concepts or data classes to estimate the class of an object whose label is unknown or dividing something according to its classes [15]. Classification is the process of finding a data class so that it can estimate the class of an object whose label is unknown [16].

2.4 Naïve Bayes

Naïve Bayes is one of the most efficient and effective algorithms for machine learning and data mining that uses probability and statistical calculations proposed by British scientist Thomas Bayes. At this stage, the algorithm is implemented on the research dataset into a system that implements Naïve Bayes [17]. The main feature of this Naïve Bayes Classifier is a very strong (naïve) assumption of the independence of each condition/event [18]. Naive Bayes is proven to be accurate and fast when applied to large databases. The advantage of using Naive Bayes is that it requires little training information to determine the mean and variance parameters of the variables needed for classification. Bayes' theorem has the following universal form [19].

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \quad (1)$$

Information:
X: Data with unknown class
H: Hypothesis data is a specific class
P(H|X): Probability of hypothesis H based on condition X (Posteriors Probability)
P(H): Hypothesis Probability (Prior Probability)
P(X|H): Probability of X based on the condition of hypothesis H
P(X): Probability of X

2.5 K-nearest Neighbour

K-Nearest Neighbour (K-NN) is a step that needs to be accompanied by training data information to determine object classification to the closest distance. The prediction results with the KNN method are obtained by classifying the shortest distance from neighbours [8]. This method finds the shortest distance between the information to be evaluated with the value of K from neighbours in the training data [20].
Here is the universal K-NN theorem for calculating distances:

$$d_i = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - p_j\right)^2} \quad (2)$$

Information:
d_i = Sample Distance
x_{ij} = Knowledge sample data
p_j = Data Input var j
n = Number of Samples
The steps for implementing the K-NN method are as follows [20]:

1. Determine parameter k (number of nearest neighbours)
2. Calculates the square of the object's Euclidean distance to the given training data
3. Sort the squared results in step 2 from the highest to the lowest.
4. Collecting the nearest neighbour classification categories based on k values
5. Predict object categories using the nearest neighbour category with the highest value.

3. Result and Discussion

The performance of an algorithm in solving classification problems can be known by measuring. One of the most common ways is to calculate the algorithm's accuracy [21]. This study uses a dataset obtained from Kaggle.com which will then be classified using the KNN and Naive Bayes algorithms. The model in this study will be tested using RapidMiner Studio tools with version 10.1 to obtain Accuracy, Precision, Recall, and T-Test values [22]. Before starting the test, the dataset is processed to eliminate invalid and missing data.

3.1 Dataset Selection

This study uses Airline Passenger Satisfaction data obtained from the Kaggle dataset site and accessed via Kaggle.com. The data consist of 26,000 attributes, namely id, gender, customer type, age, type of travel, class, flight distance, inflight wi-fi service, departure/Arrival time convenience and one label, namely satisfaction.

3.2 Determination of Role Label

This process's goal label is passenger satisfaction, containing two data: satisfied and neutral or dissatisfied. Before analyzing and making conclusions based on the modelling done in the RapidMiner Studio application, it is important first to define the role title that will be applied to the resulting class attributes. In this context, RapidMiner Studio is a machine learning application used to process the data set. This role tag assignment is done when entering notes into the RapidMiner Studio application [23].

3.3 Data Selection

Data Selection minimizes the data used for the mining process while still representing the original data [24]. In the origin dataset, some data with missing values and invalid data are not included in the process. The data contained in the process are valid and according to the conditions of the process.

3.4 Testing Method

The testing process is carried out to determine the performance of the model built. Testing uses test data whose predictions will be searched through the RapidMiner Studio tool [25].

3.4.1 Naïve Bayes

This uses a model from Naïve Bayes, as shown in Figure 2.

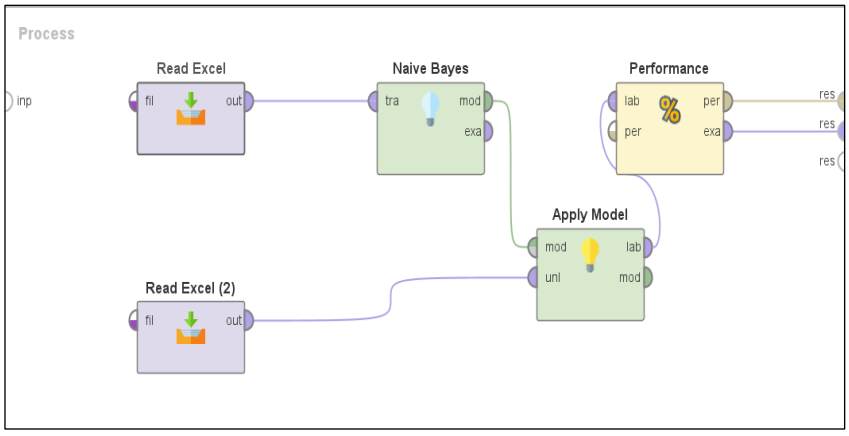

Figure 2. Naive Bayes Model Process

Figure 2 above is a model for testing the Naive Bayes algorithm using RapidMiner Studio, which starts with Read Excel, the operator used to read imported Excel files, and the Naive Bayes test using two datasets, Data Test and Data Train. Then, choose the Naive Bayes calculation model. After that, add the Apply Model and Performance operators.

3.4.2 K-nearest Neighbour

The K-NN test using RapidMiner Studio is carried out according to Figure 3 below.

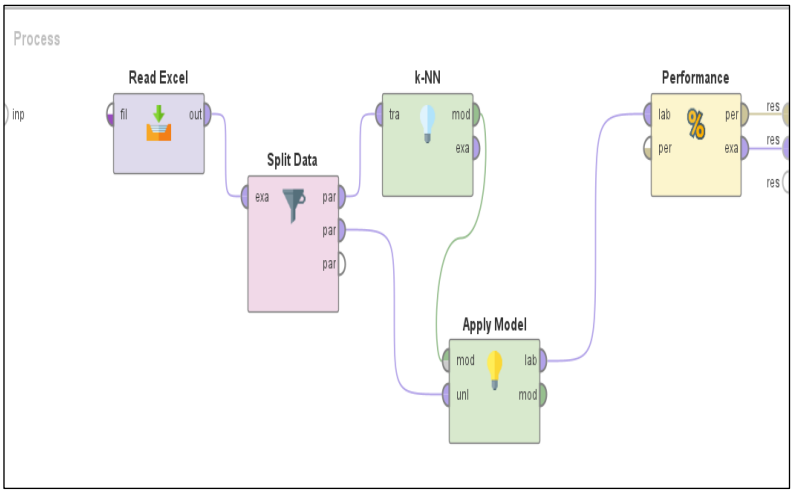

Figure 3. K-NN Model Process

In Figure 3, the K-NN test starts with importing the dataset file in Excel format. The operator used is the read Excel operator. Then the data split operator is added to the process. The ratio used in the divided data operator is 0.9 and 0.1, and then the operators added are applied model and performance.

3.5 Method Result

3.5.1 Naïve Bayes Result

Table 1. Accuracy Naive Bayes results

Accuracy: 84.48%			
	True Satisfied	True Neutral or Dissatisfied	Class Precision
Pred. Satisfied	9399	2028	82.25%
Pred. Neutral or Dissatisfied	2004	12545	86.23%
Class Recall	82.43%	86.08%	

Table 1 shows the result of the Naive Bayes method and states that the accuracy rate for this method is 84.48%. Where is Class Precision for prediction. Class precision for Satisfied is 82.25%, and for pred. Neutral or dissatisfied is 86.23%.

Figure 4 shows the plot view of the Naïve Bayes test result.

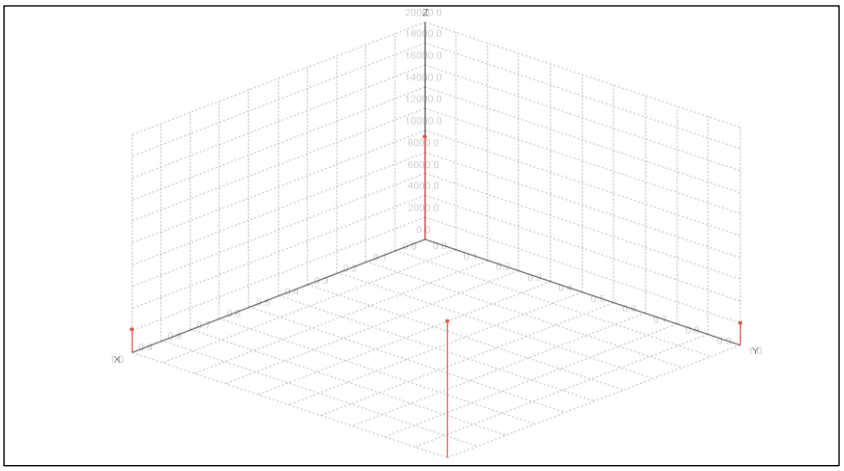

Figure 4. Plot View of the Naive Bayes Method

Figure 5 shows the Performance Vector of the Naïve Bayes test result.

PerformanceVector

```
PerformanceVector:
accuracy: 84.48%
ConfusionMatrix:
True:   satisfied       neutral or dissatisfied
satisfied:      9399    2028
neutral or dissatisfied:        2004    12545
precision: 82.25% (positive class: satisfied)
ConfusionMatrix:
True:   neutral or dissatisfied satisfied
neutral or dissatisfied:        12545   2004
satisfied:      2028    9399
recall: 82.43% (positive class: satisfied)
ConfusionMatrix:
True:   neutral or dissatisfied satisfied
neutral or dissatisfied:        12545   2004
satisfied:      2028    9399
AUC (optimistic): 0.912 (positive class: satisfied)
AUC: 0.912 (positive class: satisfied)
AUC (pessimistic): 0.912 (positive class: satisfied)
```

Figure 5. Data classification results using the Naive Bayes method

3.5.2 K-nearest Neighbour Result

Table 2. Accuracy K-NN results

Accuracy: 65.38%			
	True Satisfied	True Neutral or Dissatisfied	Class Precision
Pred. Satisfied	615	374	62.18%
Pred. Neutral or Dissatisfied	525	1038	67.35%
Class Recall	53.95%%	74.33%	

Based on Table 2, the results of testing the K-NN method with RapidMiner Studio, the accuracy value obtained is 65.38% with class precision value for pred. Satisfied by 62.18%, and pred. neutral or dissatisfied is 67.35%.
Figure 6 shows the plot view of the K-NN test result.

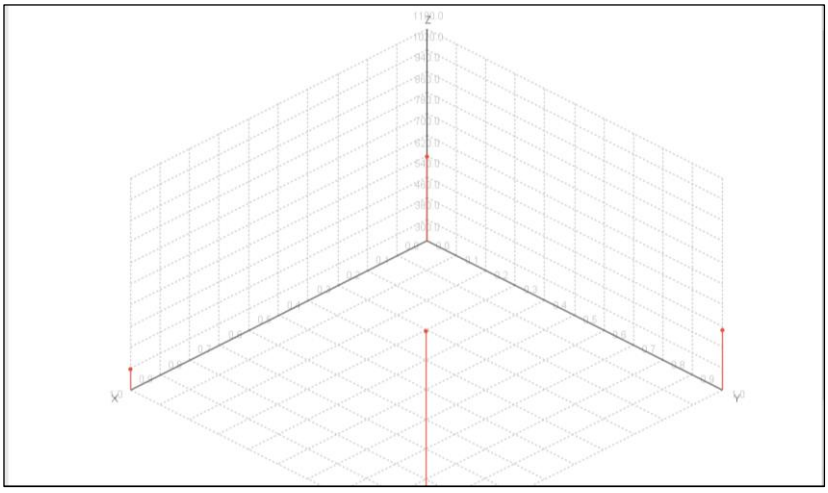

Figure 6. Plot View of the K-NN Method

Figure 7 shows the Performance Vector of the K-NN test result.

PerformanceVector

```
PerformanceVector:
accuracy: 65.38%
ConfusionMatrix:
True:   satisfied       neutral or dissatisfied
satisfied:      615     374
neutral or dissatisfied:        525     1083
precision: 67.35% (positive class: neutral or dissatisfied)
ConfusionMatrix:
True:   satisfied       neutral or dissatisfied
satisfied:      615     374
neutral or dissatisfied:        525     1083
recall: 74.33% (positive class: neutral or dissatisfied)
ConfusionMatrix:
True:   satisfied       neutral or dissatisfied
satisfied:      615     374
neutral or dissatisfied:        525     1083
AUC (optimistic): 0.695 (positive class: neutral or dissatisfied)
AUC: 0.690 (positive class: neutral or dissatisfied)
AUC (pessimistic): 0.685 (positive class: neutral or dissatisfied)
```

Figure 7. Data classification results using the K-NN method

3.5.3 Result Comparison

The result comparison of the Naive Bayes and K-NN methods is shown below in Figure 8.

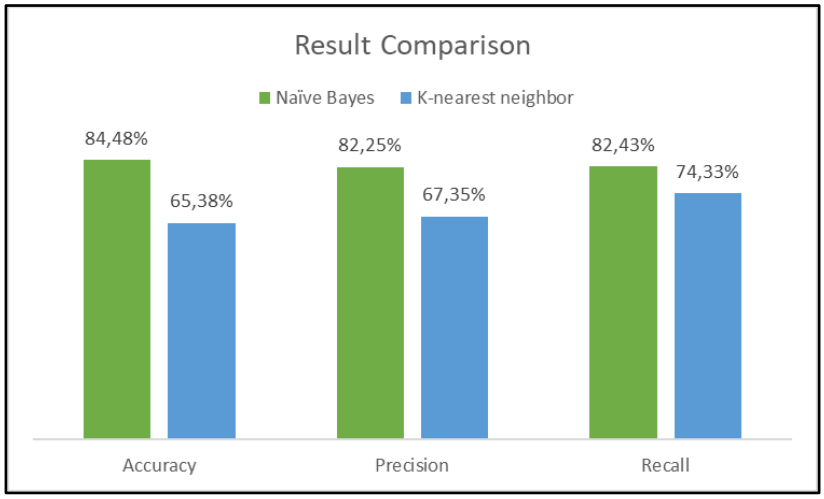

Figure 8. Result in Comparison of Naive Bayes and K-NN

4. Conclusion

Based on the results of testing the Naive Bayes and K-NN algorithms for the classification of airline customer satisfaction using the RapidMiner Studio version 10.1 application. The data are taken from the Kaggle.com Airline Passenger Satisfaction website. The results show that the accuracy of the Naive Bayes method of the two algorithms is higher than the K-NN method. The accuracy value of the naive Bayes method is 84.48%, while the accuracy value of the K-NN method is 65.38%. From the test results, the precision value for Naive Bayes is 82.25%, and K-NN is 67.35%. Furthermore, the recall value for Naive Bayes is 82.43%, and K-NN is 74.33%.

The Naive Bayes and KNN methods are used in this study to obtain a number that defines how effectively they predict customer satisfaction as it correlates with airline services. However, considering that the results of this study should be examined with caution because of the variety of attributes that determine passenger satisfaction and the method performed to evaluate the accuracy.

The level of passenger satisfaction may be assessed in the future using several techniques for the pre-flight, in-flight, and post-flight services. This can potentially be implemented to determine the reliability of airline services more precisely.

Acknowledgements

In this opportunity, the author would like to express his sincere gratitude to all those who have contributed significantly to this research. This research would not have been possible without their hard work, support and assistance.

First, the authors would like to thank the research supervisor, Jumanto, S.Kom., M.Cs., who provided valuable directions and insights throughout this research. The support and encouragement given mean a lot to the author.

Furthermore, the authors would also like to thank the research team members who were involved in the data collection and analysis process. Your contributions in collecting accurate data and conducting rigorous analysis have provided a strong foundation for this research. Lastly, the authors thank their family and friends who provided moral support and motivation during this research. Your help and understanding give strength and enthusiasm to the writer in facing challenges and obstacles.

Once again, the author would like to express his deep gratitude to all who contributed to this research. Hopefully, the results of this research can provide significant benefits and contributions in their field.

References

[1] Z. Gong, F. Zhang, W. Liu, and D. J. Graham, "On the effects of airport capacity expansion under responsive airlines and elastic passenger demand," *Transportation Research Part B: Methodological*, vol. 170, pp. 48–76, Apr. 2023, doi: 10.1016/J.TRB.2023.02.010.

[2] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc Sci Res*, vol. 110, Feb. 2023, doi: 10.1016/j.ssresearch.2022.102817.

[3] M. T. Alshurideh, A. Al-Hadrami, E. K. Alquqa, H. M. Alzoubi, S. Hamadneh, and B. Al Kurdi, "The effect of lean and agile operations strategy on improving order-winners: Empirical evidence from the UAE food service industry," *Uncertain Supply Chain Management*, vol. 11, no. 1, pp. 87–94, Dec. 2023, doi: 10.5267/J.USCM.2022.11.007.

[4] Y. Liu and K. Tahera, "A fuzzy decision-making approach for testing activity prioritization and its application in an engine company," *Appl Soft Comput*, vol. 142, Jul. 2023, doi: 10.1016/J.ASOC.2023.110367.

[5] L. Lopez-Valpuesta and D. Casas-Albala, "Has passenger satisfaction at airports changed with the onset of COVID-19? The case of Seville Airport (Spain)," *J Air Transp Manag*, vol. 108, p. 102361, May 2023, doi: 10.1016/J.JAIRTRAMAN.2023.102361.

[6] J. C. Weng, J. B. Yu, X. J. Di, P. F. Lin, J. J. Wang, and L. Z. Mao, "How does the state of bus operations influence passengers' service satisfaction? A method considering the differences in passenger preferences," *Transp Res Part A Policy Pract*, vol. 174, p. 103734, Aug. 2023, doi: 10.1016/J.TRA.2023.103734.

[7] J. Duque, F. Silva, and A. Godinho, "Data Mining applied to Knowledge Management," *Procedia Comput Sci*, vol. 219, pp. 455–461, 2023, doi: 10.1016/j.procs.2023.01.312.

[8] S. Adhikary and S. Banerjee, “Introduction to Distributed Nearest Hash: On Further Optimizing Cloud Based Distributed kNN Variant,” *Procedia Comput Sci*, vol. 218, pp. 1571–1580, 2023, doi: 10.1016/j.procs.2023.01.135.

[9] T. Noviantoro and J.-P. Huang, "Investigating airline passenger satisfaction: Data mining method," *Research in Transportation Business & Management*, vol. 43, p. 100726, Jun. 2022, doi: 10.1016/j.rtbm.2021.100726.

[10] Y. Zhang, L. Li, Y. Li, and Z. Zeng, "Machine learning model-based risk prediction of severe complications after off-pump coronary artery bypass grafting," *Adv Clin Exp Med*, vol. 32, no. 2, pp. 185–194, Feb. 2023, doi: 10.17219/ACEM/152895.

[11] S. Sharma, K. M. Osei-Bryson, and G. M. Kasper, "Evaluation of an integrated Knowledge Discovery and Data Mining process model," *Expert Syst Appl*, vol. 39, no. 13, pp. 11335–11348, Oct. 2012, doi: 10.1016/J.ESWA.2012.02.044.

[12] H. H. P. Nucci *et al.*, "Use of computer vision to verify the viability of guavira seeds treated with tetrazolium salt," *Smart Agricultural Technology*, vol. 5, Oct. 2023, doi: 10.1016/J.ATECH.2023.100239.

[13] T. T. Nguyen *et al.*, "Scalable maximal subgraph mining with backbone-preserving graph convolutions," *Inf Sci (N Y)*, vol. 644, Oct. 2023, doi: 10.1016/J.INS.2023.119287.

[14] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc Sci Res*, vol. 110, Feb. 2023, doi: 10.1016/j.ssresearch.2022.102817.

[15] A. Yazdinejad, A. Dehghantanha, R. M. Parizi, and G. Epiphaniou, "An optimized fuzzy deep learning model for data classification based on NSGA-II," *Neurocomputing*, vol. 522, pp. 116–128, Feb. 2023, doi: 10.1016/J.NEUCOM.2022.12.027.

[16] C. Singla and C. Jindal, "Comparison of Various Classification Models Using Machine Learning to Predict Mobile Phones Price Range," *Convergence of Cloud with AI for Big Data Analytics*, pp. 401–419, May 2023, doi: 10.1002/9781119905233.CH17.

[17] F. Carli, M. Leonelli, and G. Varando, "A new class of generative classifiers based on staged tree models," *Knowl Based Syst*, vol. 268, p. 110488, May 2023, doi: 10.1016/J.KNOSYS.2023.110488.

[18] A. M. Shanshool, E. M. H. Saeed, and H. H. Khaleel, "Comparison of various data mining methods for early diagnosis of human cardiology," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 3, pp. 1343–1351, Sep. 2023, doi: 10.11591/IJAI.V12.I3.PP1343-1351.

[19] K. V and S. P. S, "Adaptive boosted random forest-support vector machine based classification scheme for speaker identification," *Appl Soft Comput*, vol. 131, p. 109826, Dec. 2022, doi: 10.1016/J.ASOC.2022.109826.

[20] A. Ali, M. Hamraz, N. Gul, D. M. Khan, S. Aldahmani, and Z. Khan, "A k nearest neighbour ensemble via extended neighbourhood rule and feature subsets," *Pattern Recognit*, vol. 142, p. 109641, Oct. 2023, doi: 10.1016/J.PATCOG.2023.109641.

[21] W. Zhang, P. Li, L. Wang, F. Wan, J. Wu, and L. Yong, "Explaining of prediction accuracy on phase selection of amorphous alloys and high entropy alloys using support vector machines in machine learning," *Mater Today Commun*, vol. 35, p. 105694, Jun. 2023, doi: 10.1016/J.MTCOMM.2023.105694.

[22] A. M. Mariano, A. B. De Magalhães Lelis Ferreira, M. R. Santos, M. L. Castilho, and A. C. F. L. C. Bastos, "Decision trees for predicting dropout in Engineering Course students in Brazil," *Procedia Comput Sci*, vol. 214, no. C, pp. 1113–1120, Jan. 2022, doi: 10.1016/J.PROCS.2022.11.285.

[23] M. Z. Naser, "Machine learning for all! Benchmarking automated, explainable, and coding-free platforms on civil and environmental engineering problems," *Journal of Infrastructure Intelligence and Resilience*, vol. 2, no. 1, p. 100028, Mar. 2023, doi: 10.1016/J.IINTEL.2023.100028.

[24] İ. Aksangür, B. Eren, and C. Erden, "Evaluation of data preprocessing and feature selection process for prediction of hourly PM10 concentration using long short-term memory models," *Environmental Pollution*, vol. 311, p. 119973, Oct. 2022, doi: 10.1016/J.ENVPOL.2022.119973.

[25] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, "Top data mining tools for the healthcare industry," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 4968–4982, Sep. 2022, doi: 10.1016/J.JKSUCI.2021.06.002.