SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-1691 | e-ISSN 2964-8246

Classification of travel class with k-nearest neighbors algorithm using rapidminer

Dina Wachidah Septiana[1], Puan Bening Pastika[2]
[1, 2]Informatics Engineering Study Program, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received April 20, 2024
Revised July 4, 2024
Accepted July 5, 2024

Keywords:

Classification
K-nearest neighbors
Dummy dataset
Travel class
Rapidminer

ABSTRACT

The tourism industry in Indonesia plays an important role in the national economy. The selection of travel class according to the needs and budget of tourists is an important aspect in the tourism industry. This research aims to develop a travel class classification model using dummy datasets and the K-Nearest Neighbors (KNN) algorithm with RapidMiner software. The travel class dummy data set was obtained from the internet and modified according to research needs. The KNN algorithm was used to classify new travel classes based on previously classified dummy data. These dummy data were preprocessed and analyzed using RapidMiner software. The performance of the KNN model was evaluated using accuracy, precision, recall and F1-score. The results showed that the KNN algorithm with the values k = 1-2, k = 3-6, k = 8-10, k = 11-14 and k = 15 resulted in accuracy of 35.71%, 39.29%, 48.26%, 46.43% and 50.00%, respectively. This shows that the KNN algorithm with a value of k=15 produces the highest accuracy that can be effectively used to classify new travel classes based on dummy data.

1. Introduction

The tourism industry in Indonesia has continued to experience rapid growth in recent years. By 2023, Indonesia's tourism sector will contribute 5.03% to the Produk Domestik Bruto (PDB) and create 11.8 million jobs [1]. One of the crucial

[1] ***Corresponding Author:***
Dina Lusianti,
Faculty of Economics and Business,
Universitas Muria Kudus,
Jl. Lkr. Utara, Kayuapu Kulon, Gondangmanis, Kec. Bae, Kudus, Indonesia.
Email: dina.lusianti@umk.ac.id
DOI: https://doi.org/10.52465/josre.v2i2.357

things in the tourism industry is to choose a travel class that is suitable for the needs and budget of travelers [2].

Travel classes are usually divided into economy, business, and first class. Each class provides a variety of facilities and services at different prices. Choosing the right travel class can enhance the travel experience to the maximum [3]. A major problem in choosing travel classes for Indonesian travelers is the lack of comprehensive information [4] on the factors that influence travel class selection. These factors can include travel destination, travel duration, travel budget, traveler preferences, etc. Ignorance of these factors can lead travelers to choose a travel class that does not suit their needs, resulting in inconvenience and dissatisfaction during the trip.

This research aims to develop a travel class classification model using dummy datasets and the K-Nearest Neighbors (KNN) algorithm with RapidMiner software. The KNN algorithm is one of the popular classification algorithms and is easy to implement and has good performance in various data sets [5], [6]. RapidMiner is a data mining software that provides various tools for pre-processing, modeling, and evaluation of classification models [7].

Previous research has been conducted to develop a travel class classification model using the Naive Bayes algorithm [8]. The research involved 16 measuring variables and 1 response variable with a data set of 129,880 records. Data are divided into training data and testing data under four different conditions: 90%, 85%, 80%, and 75% for training data and the remainder for testing data. Research using the KNime program reveals that dividing training data by 90% and 10% test data yields the maximum accuracy of 81.466%.

This research proposes a KNN-based approach for classifying travel classes using dummy datasets. The KNN algorithm will be used to classify new travel classes based on pre-classified dummy data. These dummy data will be prepared and preprocessed using RapidMiner software. The KNN algorithm will then be trained with these dummy data and will be further used to classify the new travel class.

2. Method

In this research, several steps were used as stages of the research process. These stages are [9]: (1) collection of dummy data, (2) preprocessing data, (3) K-Nearest Neighbor, (4) result and analysis.

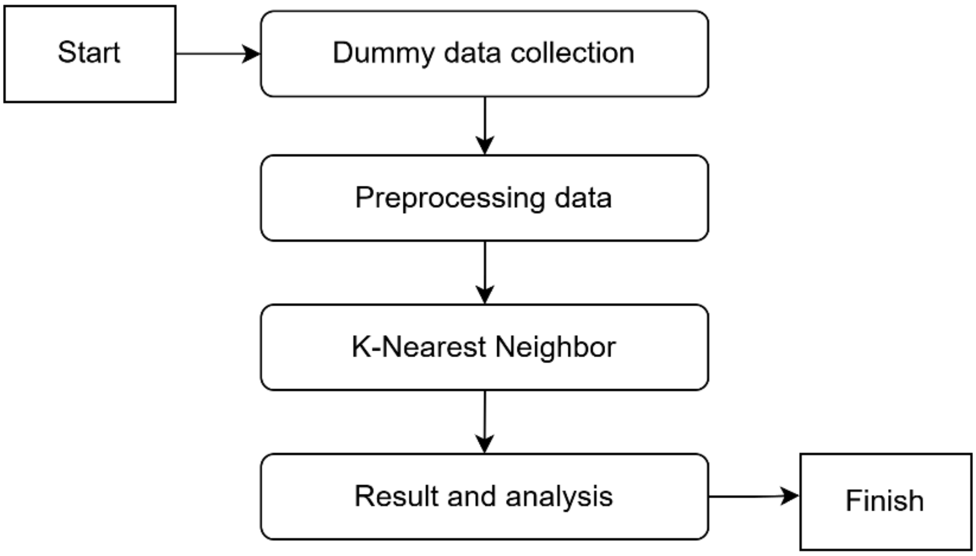


Figure 1. Research stage diagram

Dummy Data Collection

Dummy data are data that are artificially created to replace unavailable or irrelevant data [10]. The data utilized in this research are dummy data obtained from the results of generating Artificial Intelligence (AI), namely ChatGPT, based on predetermined prompts. From the prompts, the results of the attributes for the classification of the travel class in the form of tourist destinations, travel and travel budget are obtained. The results of the prompt data are shown in Table 1.

Table 1. The results of the prompt data

ID	Destination	Duration(days)	Price(USD)	Travel_class
1	Bali	5	500	Economy
2	Paris	7	1500	Business
3	New York	10	2000	First Class
4	London	6	1200	Business
5	Bali	4	400	Economy
6	Paris	8	1800	Business
7	New York	12	2500	First Class
8	London	5	1000	Economy
9	Bali	6	-	Business
10	Paris	-	1600	Business
11	New York	9	2200	-
12	London	7	1100	Economy
13	-	5	600	Economy
14	Paris	6	1700	-
15	New York	11	-	First Class
...	...	...	...	...
149	London	41	4600	Business
150	Bali	40	3900	Economy

Data Pre-processing

Data pre-processing is a data cleaning procedure that aims to detect and repair errors, inconsistencies, and incompleteness in raw data [11]. Furthermore, data cleansing is accomplished utilizing the RapidMiner application's filter tool. In this phase, the Custom Filter is used to manage attributes with missing values [12].

K-Nearest Neighbors

Classification is the process of grouping or organizing objects or data into categories or classes based on similarities or differences in certain characteristics [13]. It is a commonly used technique in various fields, such as computer science, statistics, and biology, to understand patterns, make decisions, or classify information [14]. In the context of travel class classification research, classification aims at categorizing travelers into appropriate travel classes, such as economy, business, or first class.

The K-Nearest Neighbors (KNN) algorithm is a machine learning approach used for classification and regression [15]. In classification, KNN classifies objects based on the majority of their nearest neighbor classes [16], [17]. An object is classified by the majority of votes from its nearest neighbors, i.e. those objects in the training dataset that are most similar to the object to be classified. The number of neighbors (k) is a parameter that must be predetermined [18].

Result and Analysis

The test is carried out by comparing the results of various k values in the KNN method, specifically k = 1-2, k = 3-6, k = 8-10, k = 11-14, and k = 15. The accuracy results of each k value will be analyzed to identify which k value produces the maximum accuracy, which can be used afterward to better classify new travel classes.

3. Results and Discussion

Data Collection

In the data collection stage, a search for dummy data is carried out that will be analyzed using ChatGPT according to the specified prompt. There are 150 data generated by ChatGPT as follows.

"Saya ingin membuat dataset dummy dengan tema travel, berikan dataset yang memiliki ketentuan sebagai berikut:
- Data terdiri dari 150 baris
- Terdapat missing values"

Figure 2. Data collection prompts using ChatGPT

Running Data Travel

After processing the travel data in the RapidMiner application, it can be seen below that there are some missing values.

(a)

Name	Type	Missing	Statistics	Filter (
destination	Nominal	3	Least New York (36)	
duration(days)	Integer	1	Min 4	
price	Integer	3	Min 400	
travel_class	Nominal	5	Least First Class (37)	

(b)

Row No.	destination	duration(day...	price	travel_class
9	Bali	6	?	Business
10	Paris	?	1600	Business
11	New York	9	2200	?
12	London	7	1100	Economy
13	?	5	600	Economy
14	Paris	6	1700	?
15	New York	11	?	First Class
16	London	8	1300	Business
17	Bali	7	700	?
18	Paris	9	1900	Business
19	?	10	2300	First Class
20	London	9	1400	Business
21	Bali	8	800	Economy
22	Paris	10	2000	?
23	New York	12	2400	First Class

ExampleSet (150 examples,0 special attributes,4 regular attributes)

(c)

Figure 3. Missing values on dummy data

Data Cleaning

Data cleaning, or data cleansing, or data scrubbing is an important process in the preparation of data for analysis [19]. Raw data often contain errors, inconsistencies, and incompleteness [11]. Data cleaning aims to identify and correct these problems so that data is ready to be used to generate accurate insights and conclusions [20]. The next stage is the data cleaning process using the filter example feature found in the RapidMiner application.

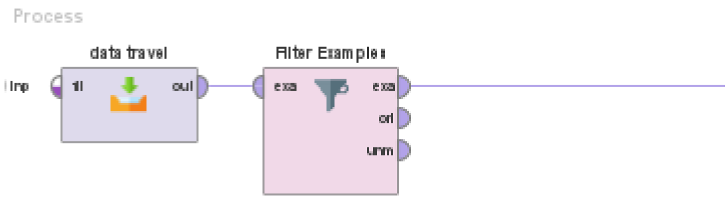


Figure 4. Data cleaning using the Filter Examples feature

At this stage, a custom filter is performed to set which attributes have missing values [12]. However, because all attributes have missing values, all attributes are included in the custom filter with the Is Not Missing setting. Then press the OK button.

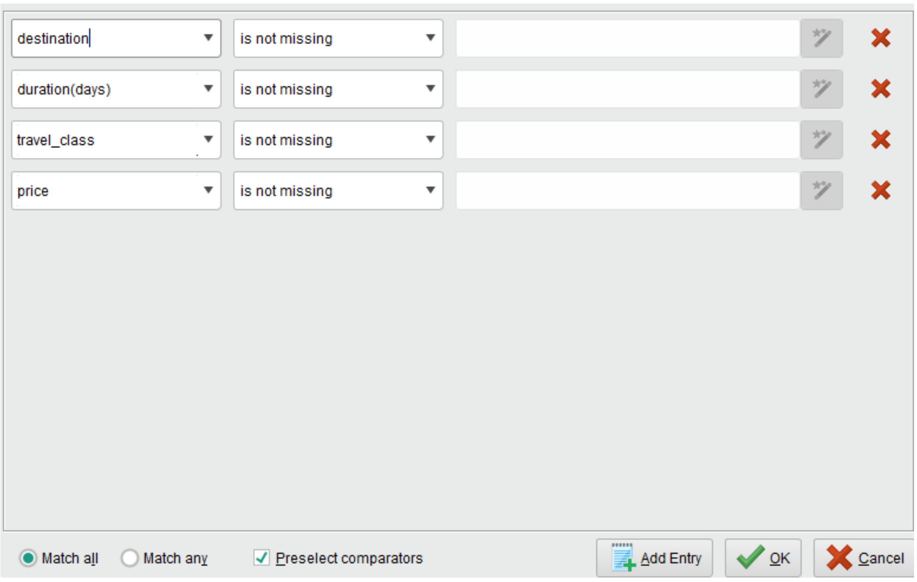
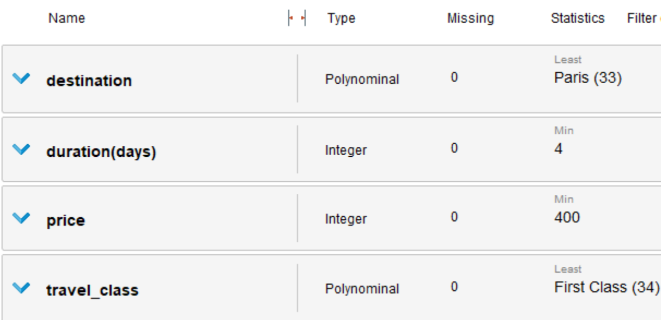


Figure 5. Custom filter to set attributes on missing values

The resulting data after cleaning are 138 data out of the total of 150 data given. This reduction in the amount of data occurred because missing values were removed to ensure better data quality.



(a)

Row No.	destination	duration(day...	price	travel_class
124	New York	40	8000	First Class
125	London	38	4300	Business
126	Bali	37	3600	Economy
127	Paris	39	7700	Business
128	New York	41	8200	First Class
129	London	39	4400	Business
130	Bali	38	3700	Economy
131	Paris	40	7900	Business
132	New York	42	8400	First Class
133	London	40	4500	Business
134	Bali	39	3800	Economy
135	Paris	41	8100	Business
136	New York	43	8600	First Class
137	London	41	4600	Business
138	Bali	40	3900	Economy

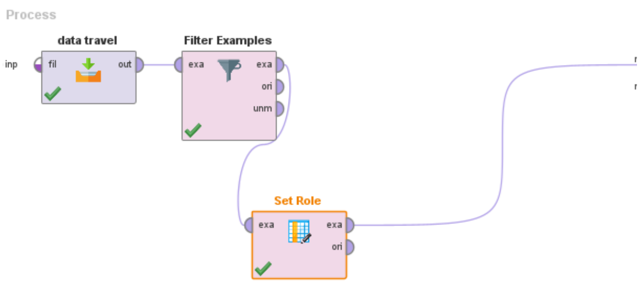
ExampleSet (138 examples,0 special attributes,4 regular attributes)

(b)

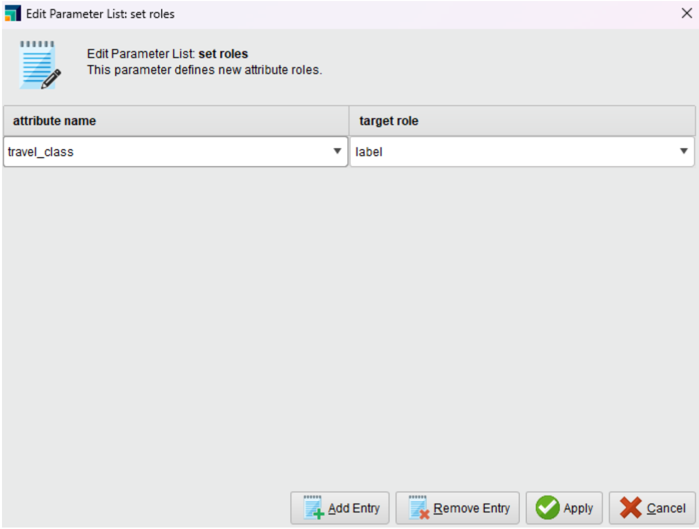
Figure 5. Data after the cleaning process

Implementation of the K-Nearest-Neighbors Algorithm

At this stage, the KNN algorithm is utilized for testing. However, before testing, the Set Role method is used to define which label will be tested, specifically the trip-class characteristic.



(a)

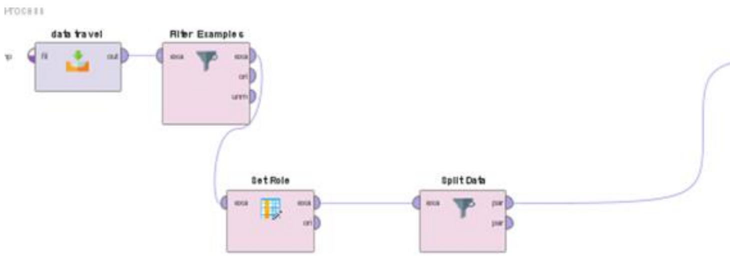


(b)

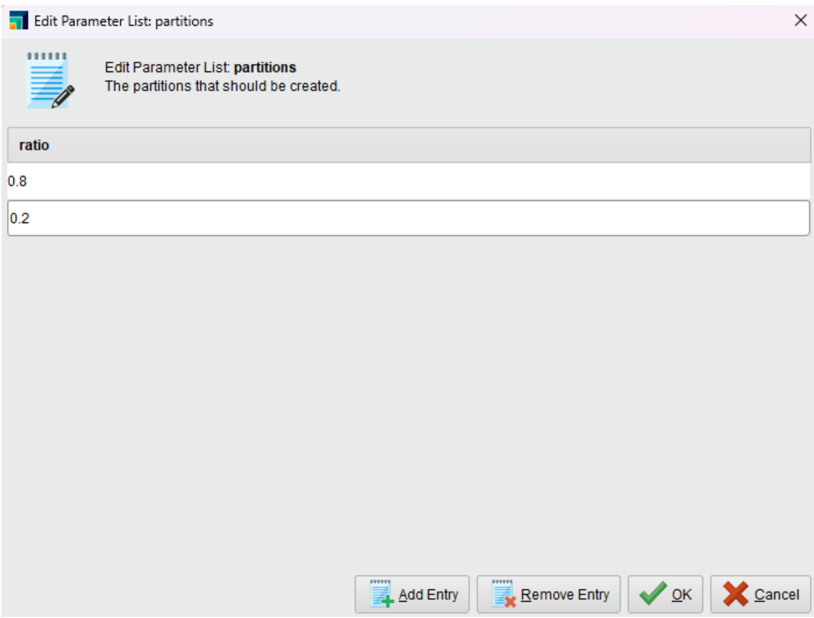
Figure 6. Set Role process for travel class attribute

Then press the Apply button on the RapidMiner application.

The next stage is to add the split data feature with the aim of dividing the data [21]. The data will eventually be separated into two sets: training data and testing data.



(a)



(b)

Figure 7. Split data process

The split data parameter has values 0.8 and 0.2. This amount will eventually serve as a reference for data division, with 80% for training data and 20% for testing data.

The next step is to utilize the KNN algorithm.

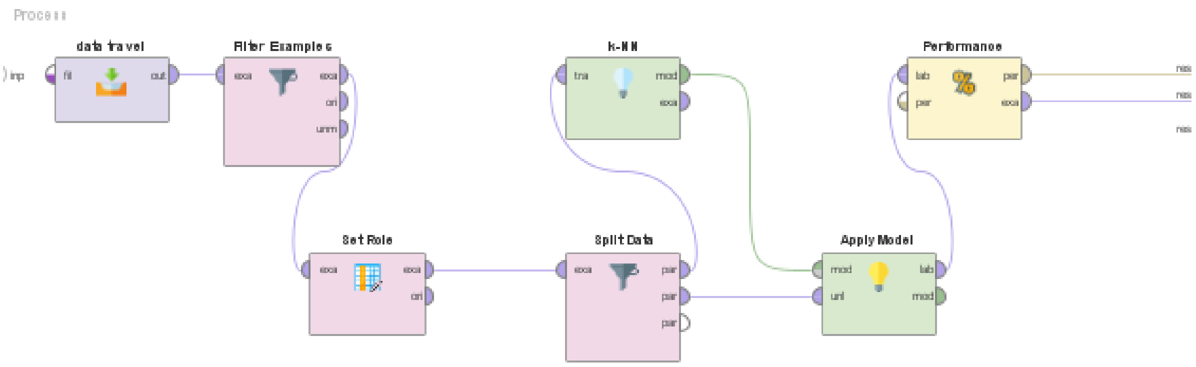


Figure 7. Implementation of the KNN algorithm

The results of running the KNN process with a value of k = 15 are shown in Table 2.

Table 2. Prediction result with k=15

Rows	Travel_ class	Prediction (travel class)	Confidence (economy)	Confidence (Bussines)	Confidence (first class)	destination	duration	Price
1	Business	First Class	0	0.397	0.603	Paris	35	6900
2	First Class	First Class	0	0.397	0.603	New York	37	7400
3	Business	Business	0	0.664	0.336	London	35	4000
4	Economy	Business	0.197	0.666	0.137	Bali	34	3300
5	Business	First Class	0	0.396	0.604	Paris	36	7100
6	First Class	First Class	0	0.397	0.603	New York	38	7600
7	Business	Business	0	0.601	0.399	London	36	4100
8	Economy	Business	0.129	0.668	0.203	Bali	35	3400
9	Business	First Class	0	0.397	0.603	Paris	37	7300

10	First Class	First Class	0	0.398	0.602	New York	39	7800
11	Business	Business	0	0.661	0.339	London	37	4200
12	Economy	Business	0.64	0.672	0.265	Bali	36	3500
	...	...	...	...	...	...	...	...
28	Economy	Businnes	0	0.731	0.269	Bali	40	3900

Table 3. KNN performance verctor testing

	True Economy	True Business	True First Class	Class Precicion
Pred.Economy	0	0	0	0.00%
Pred.Bussines	7	7	0	50.00%
Pred.First Class	0	7	7	50.00%
Class Recall	0.00%	50.00%	100.00%	
Accuracy	50.00%			

In the test mentioned above, k = 15 is used since it has a higher % accuracy than k = 1-14. Table 4 shows the test results for all k values.

Table 4. k-values testing

K-NN K 15	True Economy	True Business	True First Class
Class Recall	0.00%	50.00%	100.00%
Accuracy	50.00%		
K-NN K 11-14	True Economy	True Business	True First Class
Class Recall	0.00%	42.86%	100.00%
Accuracy	46.43%		
K-NN K 7-10	True Economy	True Business	True First Class
Class Recall	0.00%	35.71%	100.00%
Accuracy	42.86%		
K-NN K 3-6	True Economy	True Business	True First Class
Class Recall	0.00%	28.57%	100.00%
Accuracy	39.29%		
K-NN K-1-2	True Economy	True Business	True First Class
Class Recall	0.00%	21.43%	100.00%
Accuracy	35.71%		

Table 4 shows the results of testing all k values in the K-Nearest Neighbors (K-NN) algorithm to classify the classes of passenger travel (Economy, Business, and First

Class). For k=15, the K-NN algorithm produces an economy class recall of 0.00%, a business class recall of 50.00%, and a first class recall of 100.00%, with an overall accuracy of 50.00%. For k values between 11-14, the recall for the Economy class remains 0.00%, the Business class recall decreases to 42.86%, and the First class recall remains 100.00%, with an overall accuracy of 46.43%. For k values between 7-10, the recall for Economy class is still 0.00%, the Business class recall drops to 35.71%, and the First class recall remains 100.00%, with an overall accuracy of 42.86%. For k values between 3-6, the recall for the Economy class remains 0.00%, the Business class recall decreases again to 28.57%, and the First class recall remains 100.00%, with an overall accuracy of 39.29%. Finally, for k values between 1-2, the recall for the economy class remains 0.00%, the business class recall is the lowest at 21.43%, and the first class recall remains 100.00%, with an overall accuracy of 35.71%. From these results, it can be concluded that increasing the value of k generally improves the classification accuracy. The value of k=15 gives the highest accuracy of 50.00%, showing that the higher the value of k, the more effective the KNN algorithm is in classifying new travel classes.

4. Conclusion

In Indonesia's tourist business, selecting a travel class that suits travelers' needs and budget is critical to improving the overall trip experience. However, a lack of thorough knowledge of the elements influencing travel classes frequently causes inconvenience and unhappiness during travel. To address this problem, this study creates a travel categorization model using the K-Nearest Neighbors (KNN) method, dummy datasets, and RapidMiner software. The KNN technique was chosen because it is simple to develop and performs well on a variety of datasets.

The research process involved collecting chatGPT dummy data, preprocessing data, and analyzing the results. Dummy data were used to replace unavailable or relevant data and then cleaned the missing values on the data using the Filter Examples feature in RapidMiner. After that, the classification process is carried out using the KNN algorithm by determining the k parameter. The tests were carried out by dividing the data into training and testing data, and the results showed that the KNN algorithm with k = 1-2, k = 3-6, k = 8-10, k = 11-14 and k = 15 respectively produced an accuracy of 35.71%, 39.29%, 48.26%, 46.43% and 50.00%. This shows that the KNN algorithm with a value of k=15 provides the highest accuracy compared to other k values that can be used effectively to classify new travel classes based on dummy data.

REFERENCES

[1] N. Nasrulloh, E. M. Adiba, and M. N. Efendi, "Pengembangan Potensi Pariwisata Halal Pesisir Bangkalan Madura: Identifikasi Peranan Bank Syariah," *Muslim Herit.*, vol. 8, no. 1, pp. 79–

102, Jun. 2023, doi: 10.21154/muslimheritage.v8i1.4989.
[2] L. K. H. K. Yuni, "Analysis of Domestic Tourist Travel Preferences Post-Covid-19 Pandemic," *J. Appl. Sci. Travel Hosp.*, vol. 3, no. 2, pp. 80–88, Sep. 2020, doi: 10.31940/jasth.v3i2.2052.
[3] E. Sezgen, K. J. Mason, and R. Mayer, "Voice of airline passenger: A text mining approach to understand customer satisfaction," *J. Air Transp. Manag.*, vol. 77, pp. 65–74, Jun. 2019, doi: 10.1016/j.jairtraman.2019.04.001.
[4] E. Fernando, M. Irsan, D. F. Murad, S. Surjandy, and Djamaludin, "Mobile-Based Geographic Information System For Culinary Tour Mapping In Indonesia," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, IEEE, Jul. 2019, pp. 28–31. doi: 10.1109/ICOIACT46704.2019.8938511.
[5] LOUIS MADAERDO SOTARJUA and DIAN BUDHI SANTOSO, "PERBANDINGAN ALGORITMA KNN, DECISION TREE,*DAN RANDOM*FOREST PADA DATA IMBALANCED CLASS UNTUK KLASIFIKASI PROMOSI KARYAWAN," *J. INSTEK (Informatika Sains dan Teknol.*, vol. 7, no. 2, pp. 192–200, Aug. 2022, doi: 10.24252/instek.v7i2.31385.
[6] Sopiatul Ulum, R. F. Alifa, P. Rizkika, and C. Rozikin, "Perbandingan Performa Algoritma KNN dan SVM dalam Klasifikasi Kelayakan Air Minum," *Gener. J.*, vol. 7, no. 2, pp. 141–146, Jul. 2023, doi: 10.29407/gj.v7i2.20270.
[7] S. M. Dol and P. M. Jawandhiya, "Use of Data mining Tools in Educational Data Mining," in *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, IEEE, Jul. 2022, pp. 380–387. doi: 10.1109/CCiCT56684.2022.00075.
[8] A. Wijayanto, J. F. A. Bernardo, and S. Pamungkas, "Analisis Klasifikasi Kepuasan Penumpang Maskapai Penerbangan Menggunakan Algoritma Naïve Bayes," *J. Sains Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 97–103, May 2021, doi: 10.33084/jsakti.v3i2.2041.
[9] Q. A. A′ yuniyah and M. Reza, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru," *Indones. J. Inform. Res. Softw. Eng.*, vol. 3, no. 1, pp. 39–45, Mar. 2023, doi: 10.57152/ijirse.v3i1.484.
[10] X. Liu, Y. Song, and Z. Li, "Dummy Data Attacks in Power Systems," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1792–1795, Mar. 2020, doi: 10.1109/TSG.2019.2929702.
[11] C. N. Keiser and J. N. Pruitt, "Correction to 'Personality composition is more important than group size in determining collective foraging behaviour in the wild,' " *Proc. R. Soc. B Biol. Sci.*, vol. 287, no. 1928, p. 20201164, Jun. 2020, doi: 10.1098/rspb.2020.1164.
[12] H. Han, M. Li, J. Qiao, Q. Yang, and Y. Peng, "Filter Transfer Learning Algorithm for Missing Data Imputation in Wastewater Treatment Process," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12649–12662, Dec. 2023, doi: 10.1109/TKDE.2023.3270118.
[13] R. Baji Syadewo and N. Riza, "KLASIFIKASI PENERIMAAN DANA BANTUAN PADA DUSUN JATI BENING," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1220–1226, Sep. 2023, doi: 10.36040/jati.v7i2.6766.
[14] P. W. Sudarmadji, N. Fallo, and Y. S. Peli, "KOMPARASI ALGORITMA KLASIFIKASI UNTUK MEMPREDIKSI KELULUSAN MAHASISWA PROGRAM STUDI TEKNIK KOMPUTER JARINGAN," *J. Ilm. Flash*, vol. 8, no. 2, p. 109, Feb. 2023, doi: 10.32511/flash.v8i2.998.
[15] S. Anif, S. Sutama, H. J. Prayitno, and S. Sukartono, "EVALUASI PELATIHAN PENINGKATAN KOMPETENSI PROFESIONAL GURU SEKOLAH MENENGAH PERTAMA," *Manaj. Pendidik.*, vol. 14, no. 2, Jan. 2020, doi: 10.23917/jmp.v14i2.9966.
[16] M. Munir, E. Nababan, and T. Tulus, "Learning Vector Quantization with Local Mean Based to Determine K Value in the K-Nearest Neighbor Method," in *Proceedings of the Proceedings of the 1st International Conference on Management, Business, Applied Science, Engineering and Sustainability Development, ICMASES 2019, 9-10 February 2019, Malang, Indonesia*, EAI,

2020. doi: 10.4108/eai.3-8-2019.2290750.
[17] P. Kumar Sinha, "Modifying one of the Machine Learning Algorithms kNN to Make it Independent of the Parameter k by Re-defining Neighbor," *Int. J. Math. Sci. Comput.*, vol. 6, no. 4, pp. 12–25, Aug. 2020, doi: 10.5815/ijmsc.2020.04.02.
[18] R. Mulyani, D. Atmajaya, and F. Umar, "Klasifikasi Kematangan Buah Pala Menggunakan Metode K Nearest Neighbor (k-NN) Dengan Memanfaatkan Teknologi Citra Digital," *Bul. Sist. Inf. dan Teknol. Islam*, vol. 2, no. 3, pp. 140–146, Aug. 2021, doi: 10.33096/busiti.v2i3.826.
[19] L. Karlina and O. Nurdiawan, "PENERAPAN K- MEDOIDS DALAM KLASIFIKASI PERSEBARAN LAHAN KRITIS DI JAWA BARAT BERDASARKAN KABUPATEN/KOTA," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 527–532, Mar. 2023, doi: 10.36040/jati.v7i1.6348.
[20] W. D. Budimulia and F. Ridho, "PENERAPAN KOMPUTASI PARALEL PADA APLIKASI DATA CLEANING MULTIPLE DATA EDIT," *Semin. Nas. Off. Stat.*, vol. 2019, no. 1, pp. 7–14, May 2020, doi: 10.34123/semnasoffstat.v2019i1.120.
[21] W. Saputro and D. B. Sumantri, "Implementasi Citra Digital Dalam Klasifikasi Jenis Buah Anggur Dengan Algoritma K-Nearest Neighbors (KNN) Dan Data Augmentasi," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 5, no. 2, pp. 248–253, Dec. 2022, doi: 10.31539/intecoms.v5i2.4337.