ILKOM Jurnal Ilmiah Vol. 15, No. 1, April 2023, pp. 101-106
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Classification of Engineering Journals Quartile using Various Supervised Learning Models

Nastiti Susetyo Fanany Putri [a,1]; Aji Prasetya Wibawa [a,2*]; Harits Ar Rasyid [a,3]; Anik Nur Handayani [a,4]; Andrew Nafalski [b,5]; Edinar Valiant Hawali [a,6]; Jehad A.H. Hammad [c,7]
[a] *Universitas Negeri Malang, Malang, 65145, Indonesia*
[b] *University of South Australia, 101 Currie Street, Adelaide, 5000, Australia*
[c] *Al-Quds Open University, Busia Rd, Palestine*
[1] *nastiti.susetyo.2005348@students.um.ac.id;* [2*] *aji.prasetya.ft@um.ac.id;* [3] *harits.ar.ft@um.ac.id;* [4] *aniknur.ft@um.ac.id;* [5] *nafalski@unisa.edu.au;*
[6] *edinar.valiant.2105348@students.um.ac.id;* [7] *jhammad@qou.edu*
** Corresponding author*

Article history: Received October 11, 2022; Revised October 11, 2022; Accepted December 20, 2022; Available online April 07, 2023

Abstract
In scientific research, journals are among the primary sources of information. There are quartiles or categories of quality in journals which are Q1, Q2, Q3, and Q4. These quartiles represent the assessment of journal. A classification machine learning algorithm is developed as a means in the categorization of journals. The process of classifying data to estimate an item class with an unknown label is called classification. Various classification algorithms, such as K-Nearest Neighbor (KNN), Naïve Bayes, and Support Vector Machine (SVM) are employed in this study, with several situations for exchanging training and testing data. Cross-validation with Confusion Matrix values of accuracy, precision, recall, and error classification is used to analyzed classification performance. The classifier with the finest accuracy rate is KNN with average accuracy of 70%, Naïve Bayes at 60% and SVM at 40%. This research suggests assumption that algorithms used in this article can approach SJR classification system.

Keywords: Quartile Journals; Classification; KNN; Naïve Bayes; SVM

Introduction

With the advancement of science, the demand for scientific literatures is constantly rising. Scientific articles, commonly known as journals, are compilations of various scientific articles which are published on a regular schedule [1]. Journals can be categorized according to how frequently and influential they feature in other studies. Q1, Q2, Q3, and Q4 belong to this category. The lower the Q level, the better the journal's quality and credibility [2]. The Scimago Journal publishes a ranking of these journals. However, there is still a disparity between the SJR indication and the quartile value in this quartile assessment.

Machine learning classification techniques are among approach can be used to cope with this inequality. Instance data classification is a process of predicting classes[3]. Classification is widely used in various fields including education [4] [5], banking [6], health [7] [8], stockbreeding [9], and even game development [10]. Naïve Bayes, K Nearest Neighbor, Decision Tree, Support Vector Machine (SVM), Convolution Neural Network (CNN), and hybrid approaches are some of the most commonly used classification methods.

In its operation, Naïve Bayes uses Bayesian concepts and performs effectively with high-dimensional data [11]. The Naïve Bayes classifier can generate the most likely output depending on the input, has no trouble incorporating new raw data during runtime, and is a stronger probabilistic classifier [12]. The simplest machine learning technique in terms of understanding and scalability is K-Nearest Neighbor (KNN). The process of this algorithm does not take quite a long time [13]. In this algorithm, the nearest neighbour is calculated based on the value of k which determines the number of nearest neighbours to be considered as sample data points [14]. The Support Vector Machine (SVM) is predicated on the concept of class margin. The margin is designed in such a way that the distance between it and the class is maximized, reducing the chance of misclassification [15].

Several studies have been conducted using various methodologies to classify journals or articles. Journal classifications, for example, [16] applied the inter-correlated Naïve Bayes method to achieve a classification accuracy 8.65% better than the ing usual Naïve Bayes. The accuracy results were 50.49% without inter-correlated and 59.14%

with inter-correlated. Other researchers employed Naïve Bayes, which resulted in 63% accuracy rate. This research also used KNN for missing data imputation. Classification study utilizing other methodology of SVM for journal classification obtained a 61,09% accuracy score.

The issue that arises, based on the above explanation, is the presence of errors between the SJR indicators and their quartile labels in various publications on the SJR website. Therefore, multiple classification algorithms were used in this study to identify quartiles of journals and compare them to see which classification algorithm performs best in article classification.

Methods

A. Research Design

There are four stages in this study. The first step is to locate the dataset that will be used in the study. Data preparation is the next stage. Data cleaning and feature selection are done on the dataset at this point. Data is then entered into the chosen classification algorithm to complete the process. KNN (k Nearest Neighbor), Naïve Bayes, and Support Vector Machine (SVM) are the classification algorithms utilized with a 50:50, 60:40, 70:30, and 80:20 ratio between training and test data. The investigation comes to an end with a confusion matrix evaluation. **Figure 1** describes the design of the research steps.

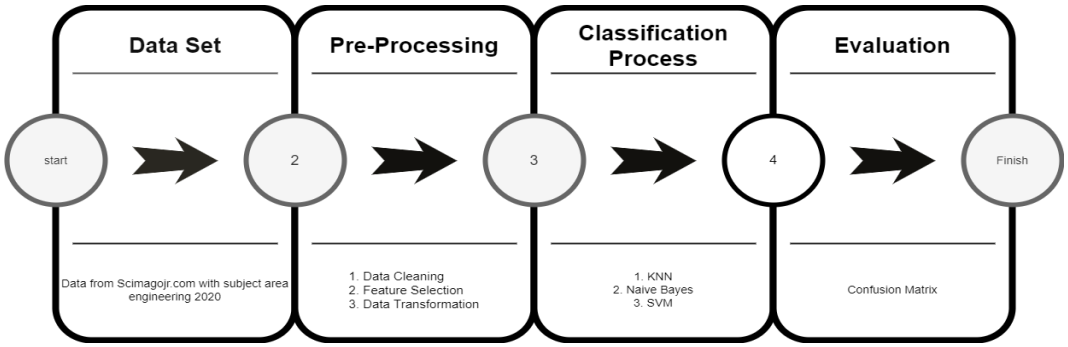

Figure 1. Research Stage Design

B. Research Data

Secondary information was collected from www.scimagojr.com. The information was obtained from a journal of engineering in 2020, which was downloaded on October 30, 2021. As seen in **Table 1**, it has 2674 data points with 20 properties. The best quartile SJR attribute is an attribute class label that includes Q1, Q2, Q3, and Q4, making this a multiclass categorization.

Table 1. Dataset attribute list

Attributes	Data Type	Range
Rank	Integer	1-2907
Sourceid	Real	12035-21100967264
Title	Nominal	Nature nanote, IEEE, etc
Type	Nominal	Journal, Prociding
Issn	Nominal	8756758X, 14602696, etc
SJR	Real	0.100-15.555
SJR Best Quartile	Nominal	Q1, Q2, Q3, Q4, NQ
H Index	Integer	0-489
Total Docs. (2020)	Integer	0-13729
Total Docs (3 Years)	Integer	0-11842
Total Refs.	Integer	0-558443
Total Cites. (3 years)	Integer	0-109926
Citable Docs. (3 years)	Integer	0-11720
Cite/ Doc (2 years)	Real	0-88
Ref.Doc.	Real	0-356.670
Country	Nominal	Saudi Arabia, USA,
Region	Nominal	Africa- Europe rope
Publisher	Nominal	Zhong gu, Springer New York, etc
Coverage	Nominal	2016-2018,2008-2020
Categories	Nominal	Water science, Civil engineering, Ai, etc

Only nine of the 20 variables listed above were used in this study to determine the journal's quartile class. SJR top quartile, H index, Total Cites (3 years), Total Refs, Total Docs (2020), Total Docs (3 years), Cites / Doc (2 years), Total Cites (3 years), Citable Docs (3 years) are the nine qualities (3 years). These properties were selected based on the fact that the SJR site uses them on its page. Therefore, it was assumed that these are the primary qualities employed in journal quartile classification. These characteristics were employed as independent factors to generate journal quartile predictive labels, which are also known as class labels [19]. The label of the journal quartile that acts as the label class was identified using this feature as an independent variable.

C. Preprocessing

This stage is critical for optimizing the classification algorithm's performance [20]. Data cleaning, data integration, data transformation, data reduction, and data resampling are some of the techniques used in preprocessing [21]. The preprocessing methods used in this investigation were data cleaning, feature selection, and data transformation.

1) Data Cleaning

Data cleaning is used to remove useless data (missing value) or data that has a lot of noise [22]. There is still a lot of data in the dataset that has no value, thus it is excluded because it would harm the algorithm's performance. The dataset contained 2674 data before cleaning. After data cleaning, the number of data is 2634. **Table 2** lists the specifics.

Table 2. Sample Data After Cleaning

SJR Best Quartile	Jumlah
Q1	704
Q2	704
Q3	680
Q4	546
Total	2634

2) Feature Selection

Feature selection is used to choose which qualities will be used as variables in the classification process. The attributes used in this study are SJR best quartile, H index, Total Cites (3 years), Total Refs, Total Docs (2020), Total Docs (3 years), Cites / Doc (2 years), Total Cites (3 years), and Citable Docs (3 Years).

3) Data Transformation

Data transformation is carried out to change the original data type into the required data type [23]. The SVM algorithm is the only one to which the data transformations are applied. Only numerical types are used by SVM [24].

D. Data Classification

The classification procedure includes numerous stages, including establishing the composition of the data training and testing using k-fold cross-validation and proceeding with journal quartile categorization using the KNN, Naïve Bayes, and SVM algorithms after the data have been cleaned. The cross-validation technique is also utilized as a model's performance value validation function. Because 10-fold cross-validation is a standard suggestion that is widely utilized in most validation approaches, it was chosen as the maximum option [25].

E. Output and Evaluation

The class variable of the journal quartile model is predicted by this study's output. The confusion matrix and classification performance were employed as evaluation models, with the One-Against-All technique being applied. Accuracy, precision, recall, and error state are all variables in this calculation. These four variables are used to assess the performance of the algorithm utilized in this model. The highest accuracy, precision, and recall levels, as well as the lowest error rate values, imply the best classification results. The Confusion Matrix is explained in **Table 3**.

Table 3. Confusion Matrix

Class	Positive Classified	Negative Classified
Positive	TP (True Positive)	FN (False Positive)
Negative	FP (False Positive)	TN (True Negative)

Where:

TP (True Positive) : The number of positive data with the truth value is true

FN (False Negative) : The amount of negative data that the system considers has a value of false

truth

FP (False Positive) : A lot of positive data that the system considers has a value of wrong truth

TN (True Negative) : A lot of negative data is considered by the system to have a value of true truth

The rate of the model's true value in classifying data is the amount of accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The level of sensitivity or recall shows the success of the model in finding information.

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

The degree of precision is the ratio of correctly predicted positive observations to the total predicted positive observations

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

Classification error is the possibility of an error in the classification.

$$Accuracy + Classification\ Error\ = 100\% \quad (4)$$

Results and Discussion

Several scenarios were tested during the data classification phase to determine which algorithm performed best in journal quartile classification. As stated in **Table 4**, the scenario is based on using training and testing data for each approach.

Table 4. Scenarios of Dataset Distribution

Scenario	Data Training	Data Testing
1	50	50
2	60	40
3	70	30
4	80	20

The results of the application of the algorithms are shown in Table 5. The first algorithm applied was KNN with algorithm value of k = 5. The second and the third algorithm were Naive Bayes algorithm and SVM with the linear kernel respectively. The linear kernel was chosen as it has a higher value than other kernels [26].

Table 5. Classification Result From Various Algorithms

Scenario	KNN				Naïve Bayes				SVM			
	Accuracy	***Precision***	***Recall***	***Error***	***Accuracy***	***Precision***	***Recall***	***Error***	***Accuracy***	***Precision***	***Recall***	***Error***
1	**74.87%**	**59.96%**	**59.93 %**	**25.13 %**	**68.79%**	**55.78%**	**55.86 %**	**31.21 %**	41.53%	30.24%	31.71 %	58.47 %
2	73.47%	58.50%	59.05 %	26.57 %	63.19%	52.38%	51.66 %	36.81 %	38.93%	28.41%	29.79 %	61.01 %
3	72.15%	57.61%	57.91 %	27.85 %	64.18%	52.67%	52.38 %	35.82 %	41.27%	29.80%	31.50 %	58.73 %
4	73.06%	58.26%	58.67 %	26.94 %	63.57%	57.99%	51.85 %	36.43 %	**43.07%**	**48.39%**	**33.24 %**	**56.93 %**

According to **Table 5**, the highest accuracy value from KNN was 74.87 percent, using a 50:50 ratio between testing data and a balanced test and a KNN value of 5. The same process occured when accuracy validation, recall, and classification error numbers were calculated. As seen in the table, there was no correlation between the number of sample ratios and the classification algorithm's accuracy. Naïve Bayes algorithm performed admirably, with an accuracy value of more than 60%. With the accuracy of 68.79 percent and inaccuracy of 31.21 percent, scenario one offers the best value. On the other hand, the second scenario has the poorest value with an error value of 36.81 percent. The results of journal quartile categorization with SVM follow a pattern that is nearly identical to that of the KNN approach. With a rating of 43.07 percent, the fourth scenario has the best accuracy. This scenario likewise has the greatest precision, with a difference nearly double that of the other ratios. Because all of the accuracy numbers are less than 50%, this algorithm works is the least optimal. Because SVM performs poorly in multiclass classification

and with a large number of data samples, it has the lowest accuracy value [27]. A summary of all the experiments carried out is presented in **Table 6**

Table 6. Comparison of each method's best results

Methode	Scenario	Accuracy
KNN	**1**	**74.87%**
Naïve Bayes	1	68.79%
SVM	4	43.07%

KNN, with an accuracy of 74.87 percent, is the best algorithm for journal quartile categorization, while SVM is less ideal.

Conclusion

The k-nearest neighbor (KNN) method has the highest accuracy in the journal quartile categorization by 70% for all data-sharing conditions from 50:50 to 80:20, according to the results of testing many techniques with different data-sharing scenarios. Additionally, it was found that the sample ratio distribution closely resembles the accuracy, precision, and recall values. While the correlation between the sampling ratio and the error classification value is inverse. SVM works poorly with accuracy values of less than 50%. It is strongly advised to use an advanced method like ensemble models for the same subject research in the future to improve performance.

Acknowledgement

This paper and the research behind it would not have been possible without the hard work of our team, as well as financial support from the Universitas Negeri Malang (UM) through a research grant, PPM 2022

References

[1] R. A. Agha *et al.*, "The PROCESS 2018 statement: Updating Consensus Preferred Reporting Of CasE Series in Surgery (PROCESS) guidelines," *Int. J. Surg.*, vol. 60, pp. 279–282, Dec. 2018, doi: 10.1016/j.ijsu.2018.10.031.

[2] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, p. 91, 2019, doi: 10.3991/ijes.v7i2.10659.

[3] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[4] A. Lomonosov, O. Lomonosova, and I. Nadtochii, "The systematization and classification of socio-economic problems in higher education," *Balt. J. Econ. Stud.*, vol. 5, no. 4, p. 137, Oct. 2019, doi: 10.30525/2256-0742/2019-5-4-137-147.

[5] A. Larasati, A. Miftahul Hajji, and A. N. Handayani, "Preferences analysis of engineering students on choosing learning media using Support Vector Machine (SVM) model," vol. 242, no. Icovet 2018, pp. 57–59, 2019, doi: 10.2991/icovet-18.2019.15.

[6] I. F. Jaramillo, R. Villarroel-Molina, B. R. Pico, and A. Redchuk, "A Comparative study of classifier algorithms for recommendation of banking products," 2021, pp. 253–263.

[7] B. T. Crabb *et al.*, "Comparison of international classification of diseases and related health problems, tenth revision codes with electronic medical records among patients with symptoms of Coronavirus Disease 2019," *JAMA Netw. Open*, vol. 3, no. 8, p. e2017703, Aug. 2020, doi: 10.1001/jamanetworkopen.2020.17703.

[8] B. Bozkurt *et al.*, "Universal definition and classification of heart failure," *J. Card. Fail.*, vol. 27, no. 4, pp. 387–413, Apr. 2021, doi: 10.1016/j.cardfail.2021.01.022.

[9] S. A. Braginets, O. V. Galanina, and V. S. Grachev, "Neural Networks and Artificial Intelligence in stockbreeding and forecasting dairy cattle productivity," in *AgroTech*, Singapore: Springer Nature Singapore, 2022, pp. 199–207.

[10] H. A. Rosyid, M. Palmerlee, and K. Chen, "Deploying learning materials to game content for serious education game development: A case study," *Entertain. Comput.*, vol. 26, no. January, pp. 1–9, 2018, doi: 10.1016/j.entcom.2018.01.001.

[11] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using

random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.

[12] D. N. Mhawi, A. Aldallal, and S. Hassan, "Advanced Feature-Selection-Based Hybrid Ensemble Learning Algorithms for Network Intrusion Detection Systems," *Symmetry (Basel).*, vol. 14, no. 7, p. 1461, Jul. 2022, doi: 10.3390/sym14071461.

[13] K. Taunk, "2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019," *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. Iciccs, pp. 1255–1260, 2019.

[14] D. Mustafi, A. Mustafi, and G. Sahoo, "A novel approach to text clustering using genetic algorithm based on the nearest neighbour heuristic," *Int. J. Comput. Appl.*, vol. 0, no. 0, pp. 1–13, 2020, doi: 10.1080/1206212X.2020.1735035.

[15] R. Khanjani-Shiraz, A. Babapour-Azar, Z. Hosseini-Nodeh, and P. M. Pardalos, "Distributionally robust joint chance-constrained support vector machines," *Optim. Lett.*, Mar. 2022, doi: 10.1007/s11590-022-01873-x.

[16] R. P. Adiperkasa, A. P. Wibawa, I. A. E. Zaeni, and T. Widiyaningtyas, "International Reputable Journal Classification Using Inter-correlated Naïve Bayes Classifier," *Proc. - 2019 2nd Int. Conf. Comput. Informatics Eng. Artif. Intell. Roles Ind. Revolut. 4.0, IC2IE 2019*, pp. 49–52, 2019, doi: 10.1109/IC2IE47452.2019.8940887.

[17] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," *Proceeding - 2019 5th Int. Conf. Sci. Inf. Technol. Embrac. Ind. 4.0 Towar. Innov. Cyber Phys. Syst. ICSITech 2019*, pp. 83–88, 2019, doi: 10.1109/ICSITech46713.2019.8987530.

[18] S. M. Putra, A. P. Wibawa, T. Widiyaningtyas, and I. A. E. Zaeni, "Performance of SVM in Classifying the Quartile of Computer Science Journals," *Proceeding - 2019 5th Int. Conf. Sci. Inf. Technol. Embrac. Ind. 4.0 Towar. Innov. Cyber Phys. Syst. ICSITech 2019*, pp. 13–17, 2019, doi: 10.1109/ICSITech46713.2019.8987453.

[19] A. P. Wibawa, "International Journal Quartile Classification Using the K-Nearest Neighbor Method," 2019.

[20] M. Tubishat, M. A. M. Abushariah, N. Idris, and I. Aljarah, "Improved whale optimization algorithm for feature selection in Arabic sentiment analysis," *Appl. Intell.*, vol. 49, no. 5, pp. 1688–1707, May 2019, doi: 10.1007/s10489-018-1334-8.

[21] I. Cordón, J. Luengo, S. García, F. Herrera, and F. Charte, "Smartdata: Data preprocessing to achieve smart data in R," *Neurocomputing*, vol. 360, pp. 1–13, Sep. 2019, doi: 10.1016/j.neucom.2019.06.006.

[22] T. H. Borkar and T. Ahuja, "Comparative Study of Supervised Learning Algorithms for Fake News Classification," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2022, pp. 1405–1411, doi: 10.1109/ICOEI53556.2022.9777118.

[23] C. Wang, Y. Feng, R. Bodik, I. Dillig, A. Cheung, and A. J. Ko, "Falx: Synthesis-Powered Visualization Authoring," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021, pp. 1–15, doi: 10.1145/3411764.3445249.

[24] M. A. Macmudi, "Uji Pengaruh Karakteristik Dataset," *J. Comput. Inf. Syst. Technol. Manag.*, vol. 1, no. 2, pp. 7–11, 2018.

[25] Y. R. Nugraha, A. P. Wibawa, and I. A. E. Zaeni, "Particle Swarm Optimization-Support Vector Machine (PSO-SVM) Algorithm for Journal Rank Classification," *Proc. - 2019 2nd Int. Conf. Comput. Informatics Eng. Artif. Intell. Roles Ind. Revolut. 4.0, IC2IE 2019*, pp. 69–73, 2019, doi: 10.1109/IC2IE47452.2019.8940822.

[26] N. Kalcheva, M. Karova, and I. Penev, "Comparison of the accuracy of SVM kernel functions in text classification," *Proc. Int. Conf. Biomed. Innov. Appl. BIA 2020*, pp. 141–145, 2020, doi: 10.1109/BIA50171.2020.9244278.

[27] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine," *J-Ptiik*, vol. 2, no. 2, pp. 802–810, 2018, [Online]. Available: http://j-ptiik.ub.ac.id.