Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022r
DOI : https://doi.org/10.33395/sinkron.v7i3.11487

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison of Drug Type Classification Performance Using KNN Algorithm

Febri Aldi[1)]*, Irohito Nozomi [2)], Soeheri[3)]
[1)2)]Universitas Putra Indonesia Yptk, Padang, [3]Universitas Potensi Utama, Medan, Indonesia
[1)]febri_aldi@upiyptk.ac.id, [2)]irohito_nozomi@upiyptk.ac.id, [3)]soedjuli@gmail.com

Submitted : June, 26, 2022 | **Accepted** : July 27, 2022 | **Published** : July 31, 2022

Abstract: The error of decommissioning is a serious problem that is often faced in medicine. In the face of these problems, information technology has a very important role. One of the information technologies that can be used is to use the machine learning classification algorithm K-Nearest Neighbor KNN. KNN is a type of machine learning algorithm that can be applied to problems with classification and regression prediction. The classification of types of drugs for patients greatly affects the health of the patient. The patient data is processed and transformed to numbers, which are then divided into training data and test data from 90:10, 80:20, 70:30 and using the Cross Validation model. KNN works through the nearest neighboring value with a value of k = 3 calculated by the calculation of Euclidean Distance, and then evaluated using the Confusion Matrix. The performance of the KNN algorithm resulted in the highest Accuracy value of 98.33%, a Precision value of 98.8%, a Recall value of 96.2%, and an F-measure value of 97.48%. The performance is obtained from the sharing of training data and 90:10 test data. The data share results in high performance compared to other data shares, including using the Cross Validation model. And the lower the k value, the higher the value of the resulting performance. The results show that the performance of the KNN algorithm is working well.

Keywords: Drug; Machine Learning; KNN; Cross Validation; Confusion Matrix

INTRODUCTION

Medication error occurs at every stage of the drug administration process, including prescribing, copying, dispensing, administering, and monitoring (D, D, & D, 2009). According to reports, evidence shows about a quarter of all health care errors (Patel & Balkrishnan, 2010). Medication errors are the leading cause of death and loss worldwide (Makary MA and Daniel M). The World Health Organization (WHO) said that worldwide medical expenditure is around US$ 42 million per year, or 0.7% of the total amount spent on global health (Poulter & Lackland, 2017). One of the treatment errors is improperly administering medicine to patients. Medication error is still one of the trends in patient safety issues (Aprilia, Nursalam, & Panji, 2016). In addition, the Institute of Medicine (IOM) has stated that a prolonged treatment regimen, incorrect treatment, work schedule, family obligations, and conditions affecting health workers are all factors that contribute to a person's safety (Mao, Jia, Zhang, Zhao, & Chen, 2015).

Information technology has a very important use in this situation. Information technology is increasingly being used by healthcare organizations to meet the needs of doctors during their projects to articulate their operational goals (Amin & Ali, 2017). Machine Learning (ML) is one technology that can be utilized. ML makes the work in classifying diseases in the health field easy such as, knowing the type of disease and providing results in the form of more optimal images (Telaumbanua, Hulu, Nadeak, Lumbantong, & Dharma, 2019). One of the methods that can be used in ML for classification is K-Nearest Neighbor (KNN) (Yuliati & Sihombing, 2021). KNN is a type of machine learning algorithm that can be applied to problems with classification and regression prediction. You only need to specify the number of parallel tests and do not need any more parameters, so it is safe to use in the intended location (Xiong & Yao, 2021). Therefore, in this study the authors tried to use the KNN algorithm in classifying the types of drugs based on blood pressure levels, cholesterol levels, and the ratio of sodium to potassium in the blood, to overcome errors in determining the right type of drug for patients.

*Corresponding author

LITERATURE REVIEW

Selecting a value for k is often done during cross-validation. Larger values in k values help in lowering the effect of noise levels at the pixel level in training data sets (Murugan, Nair, & Kumar, 2019). Experimental findings suggest that the suggested algorithms have excellent classification performance and can significantly improve the classification efficiency of KNN algorithms in processing large data sets while maintaining the classification accuracy of KNN algorithms (Xing & Bei, 2019). With an accuracy of 96.76 percent, the KNN algorithm effectively identifies and detects the selected disease (Hossain & Rahaman, 2019). KNN can categorize tweet data related to three different vaccines into favorable, negative, and neutral sentiments (Shamrat & Chakraborty, 2021). The KNN classification algorithm can improve classification performance so that it can only classify four different diseases (Vaishnnave & Devi, 2019). It is recommended to classify Lao news texts using KNN-based techniques that use data normalization and data dimension reduction. Experimental findings suggest that the approach has produced positive (Zhou, Li, Zhang, Huo, & Chen, 2020).

The KNN algorithm has gained popularity as a method for classifying images. Results from KNN demonstrate its ability to accurately and successfully classify images while also having a high degree of resilience (Pavaloiu, Ancuceanu, Enache, & Vasilateanu, 2017). There is additional study on the use of KNN to image classification in the categorization of medicinal plants that focuses on the form characteristics of the plant (Vaishnnave, Suganya Devi, Srinivasan, & Arutperumjothi, 2019). KNN can give results with great accuracy, according to a number of earlier research. For instance, research on face categorization using KNN yields an accuracy of 81 percent when k = 1 (Wirdiani et al., 2019).

METHOD

The stages carried out by the author in the application of the KNN Classification algorithm include; Data Collection, Data Transformation, Data Sharing, Application of KNN Algorithms, and Evaluation.

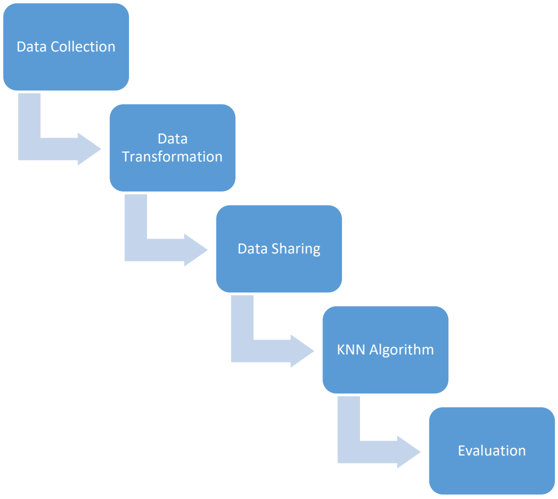

Fig. 1 Stages of KNN Classification

Data Collection

In this study using patient data on the type of drug based on blood pressure levels, cholesterol levels, and the ratio of sodium to potassium in the blood taken from kaggle. The attributes of the data include; age, gender, blood pressure level, cholesterol level, sodium to potassium ratio in the blood, and type of medication. The data collected was 200 data as in table 1.

Table 1
Patient Data

No	Age	Gender	Blood Pressure	Cholesterol	Sodium Potassium Ratio	Drug Type
1	23	P	HIGH	HIGH	25,355	Drug Y
2	47	L	LOW	HIGH	13,093	Drug C
3	47	L	LOW	HIGH	10,114	Drug C
4	28	P	NORMAL	HIGH	7,798	Drug X
5	61	P	LOW	HIGH	18,043	Drug Y
6	22	P	NORMAL	HIGH	8,607	Drug X
7	49	P	NORMAL	HIGH	16,275	Drug Y
8	41	L	LOW	HIGH	11,037	Drug C

*Corresponding author

9	60	L	NORMAL	HIGH	15,171	Drug Y
10	43	L	LOW	NORMAL	19,368	Drug Y
…	…	…	…	…	…	…
…	…	…	…	…	…	…
200	40	P	LOW	NORMAL	11,349	Drug X

Data Transformation

Data transformation is carried out to convert category data into nominal data on multiple attributes to facilitate the classification process (Kartini, Farmadi, & Nugrahadi, 2022). The process of transforming data into a certain format form so that the data is suitable for the data mining process (Putri, Purnamasari, Dikananda, Nurdiawan, & Anwar, 2021).

Data Sharing

The trainer data and the test data are separated from the preprocess data. The ratio of training data to test data is 90:10, 80:20, and 70:30, respectively, and is compared with the distribution of data using cross-validation methods (Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018). Cross Validation is a statistical technique for calculating the capabilities of machine learning models (Houcine, Mezache, & Oudira, 2019). The training set is divided into sections with cross-validation that are approximately the same size. The remaining sections are used as training data while each section is alternately selected as test data. Next, the class labels from the test data are predicted using a prediction model developed on the trainer data. Prediction accuracy in all blind tests is then collected to provide an estimate of overall performance after the process has been repeated until each area has been covered once (Baumann, 2003).

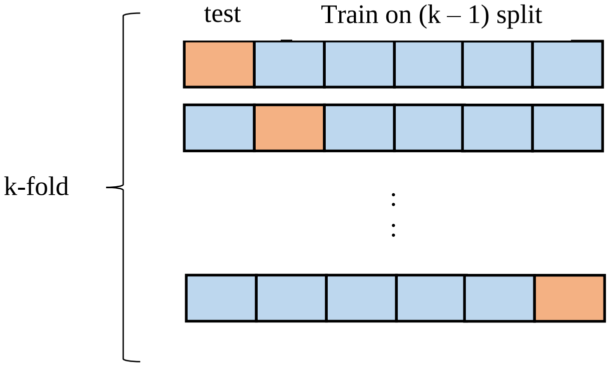

Fig. 2 Konsep Cross Validation

KNN Algorithm

An effective way to solve classification problems is through supervised machine learning methods (Arslan & Arslan, 2021). By calculating the distance between the observed sample and its closest neighboring sample K, the class in which the new observation data are located is identified during the prediction phase (Arslan & Arslan, 2021). To determine the distance of the nearest neighbor using the euclidean distance calculation on equation 1 with the value of k is an odd number.

$$dist_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \qquad (1)$$

Evaluation

The Confusion Matrix model was used to evaluate categorization in this study. The two-class confusion matrix is a contingency table that shows how many items were correctly predicted and how many were misclassified each time the researcher used an algorithm to separate elements from a data set containing two conditions (for example, positive and negative) (Anita & Bajusz, 2019). True positives (TP) refer to data that the algorithm correctly identifies as positive, while false negatives refer to data that the algorithm incorrectly identifies as negative (FN). False positives are elements that are mistakenly expected to be positive, while true negatives (TN) are negative elements that are accurately identified as negative (FP) (Chicco, Tötsch, & Jurman, 2021). Based on the confusion matrix, classification performance can be calculated such as accuracy, precision, recall, and F-measure (AmaliaLuque). Determine the accuracy value in equation 2, the precision value in equation 3, the recall value in equation 4, and look for the F-measure value in equation 5.

*Corresponding author

$$Akurasi = \frac{TP}{DS} \quad (2)$$

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

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F - measure = \frac{2(Recall\ x\ Presisi)}{Recall+Presisi} \quad (5)$$

RESULT

The data obtained from the Kaggle site on drug type data for patients has 6 attributes presented in table 2. The label on these attributes is medicine.

Table 2
Patient Data Attributes

Attribute	Information
Age	Age of Adolescence to Adulthood (Years)
Gender	M/F
Blood Pressure	High/Normal/Low
Cholesterol	High/Normal
Sodium Potassium Ratio	Ratio of Sodium to Potassium in the Blood (mmHg)
Drug	Drug Type Class

The data on the blood pressure attribute is transformed into data in the form of numbers, namely 0 for high blood pressure levels, 1 for normal blood pressure levels, and 2 for low blood pressure levels. Meanwhile, the data on the sex attribute was transformed into 0 for the male sex and 1 for the female sex. The data on the cholesterol level attribute is transformed into 0 for high cholesterol levels and 1 for normal cholesterol levels. It can be seen that the data becomes as in table 3.

Table 3
Patient Data Transformation

No	Age	Gender	Blood Pressure	Cholesterol	Sodium Potassium Ratio	Drug Type
1	23	1	0	0	25,355	Drug Y
2	47	0	2	0	13,093	Drug C
3	47	0	2	0	10,114	Drug C
4	28	1	1	0	7,798	Drug X
5	61	1	2	0	18,043	Drug Y
6	22	1	1	0	8,607	Drug X
7	49	1	1	0	16,275	Drug Y
8	41	0	2	0	11,037	Drug C
9	60	0	1	0	15,171	Drug Y
10	43	0	2	1	19,368	Drug Y
…	…	…	…	…	…	…
…	…	…	…	…	…	…
200	40	1	2	1	11,349	Drug X

The results of the data transformation in the table above are carried out a classification process using the KNN algorithm by dividing the data into training data and test data using a cross validation model with a fold number of 10. And the division of training data and test data manually as in table 4.

Table 4
Training Data Sharing and Test Data

Train Data : Test Data %	Train Data	Test Data
90 : 10	180	20
80 : 20	160	40
70 : 30	140	60

*Corresponding author

Furthermore, training data and test data testing were carried out using the KNN algorithm against a predetermined K value, namely k = 3 using Rapidminer software. From the tests that have been carried out using the KNN classification algorithm, it shows the confusion matrix of each training data and the test data can be seen in tables 5, 6, 7 and 8. The results of the calculation with training and test data 90: 10 k values = 3 can be seen in table 5. The confusion matrix of the overall predictions for the type of drug has a TP value of 177 data, an FP value of 3 data, and 0 data for FN. So that it is in accordance with the number of data tested, which is 180 data. The accuracy value of the test is 98.33%.

Table 5
Confusion Matrix Result 90 : 10

n = 200	True. DrugY	True. DrugC	True. DrugX	True. DrugA	True. DrugB
Pred. DrugY	82	0	0	0	0
Pred. DrugC	0	14	0	0	0
Pred. DrugX	0	0	49	1	2
Pred. DrugA	0	0	0	20	0
Pred. DrugB	0	0	0	0	12

The results of the test with training and test data 80: 2 k values = 30 can be seen in table 6. The Confusion Matrix of the overall predictions for the type of drug has a TP value of 157 data, an FP value of 3 data, and 0 data for FN. So that it is in accordance with the number of data tested, which is 160 data. The accuracy value of the test is 98.12%.

Table 6
Confusion Matrix Result 80 : 20

n = 200	True. DrugY	True. DrugC	True. DrugX	True. DrugA	True. DrugB
Pred. DrugY	73	0	0	0	0
Pred. DrugC	0	13	0	0	0
Pred. DrugX	0	0	43	1	2
Pred. DrugA	0	0	0	17	0
Pred. DrugB	0	0	0	0	11

The results of the test with training and test data 70: 30 k values = 3 can be seen in table 7. The Confusion Matrix of the overall predictions for the type of drug has a TP value of 137 data, an FP value of 3 data, and 0 data for FN. So that it is in accordance with the number of data tested, which is 140 data. The accuracy value of the test is 97.86%.

Table 7
Confusion Matrix Result 70 : 30

n = 200	True. DrugY	True. DrugC	True. DrugX	True. DrugA	True. DrugB
Pred. DrugY	64	0	0	0	0
Pred. DrugC	0	11	0	0	0
Pred. DrugX	0	0	38	1	2
Pred. DrugA	0	0	0	15	0
Pred. DrugB	0	0	0	0	9

The test results using the cross validation model can be seen in table 7. The Confusion Matrix of the overall predictions for drug types contained 13 TP values, 28 FP values, and 32 data for FN. Meanwhile, the number of data tested was 198 data. So there are 2 unclassified data. The accuracy value of the test is 70.00%.

Table 8
Test Results K-Folds 10

n = 200	True. DrugY	True. DrugC	True. DrugX	True. DrugA	True. DrugB
Pred. DrugY	84	0	3	1	1
Pred. DrugC	0	4	2	3	0
Pred. DrugX	1	7	34	10	8
Pred. DrugA	2	3	7	9	0
Pred. DrugB	2	2	8	0	7

*Corresponding author

The results of all tests carried out showed the highest accuracy values with a value of k = 3 were on the training data and test data of 90 : 10, the highest precision at 90 : 10 and 80 : 20, the highest recall at 90 : 10, and the highest F-measure at 90 : 10. It can be seen in the table 9.

Tabel 9
Comparison of Performance Value Results

Train Data : Test Data %	Accuracy	Precision	Recall	F-measure
90 : 10	98,33%	98,8%	96,2%	97,48%
80 : 20	98,12%	98,8%	95,8%	97,27%
70 : 30	97,86%	98,6%	95,4%	96,97%
Folds 10	70,00%	52,5%	53,2%	54,18%

DISCUSSIONS

The KNN algorithm in classifying the type of drug for patients based on the parameters of cholesterol level, level, blood pressure level, and the amount of Sodium to Potassium in the blood, shows excellent classification performance. The data testing using Rapidminer software, starting with data import, data sharing through 4 stages, determining the K value, and measuring algorithm performance, shows that the larger the training data, the better the performance of the KNN algorithm. Evidenced by the highest accuracy value of 98.33%, precision 98.8%, recall 96.2%, and F-measure 97.48% on the number of training and test data 90: 10. Then the second highest is the training data and test data 80: 20, with an accuracy value of 98.12%, precision of 98.8%, recall of 95.8%, and F-measure of 97.27%. And ranked third by the test training data 70: 30, with an accuracy value of 97.86%, precision of 98.6%, recall of 95.4%, and F-measure of 96.97%. The comparison of these results did not experience a significant decrease in performance. However, it is much different from the performance resulting from the Cross Validation model. The results showed that the values of Accuracy, precision, recall, and F-measure decreased by approximately 20 to 30%.

Meanwhile, testing the K value with each data share that has been carried out proves that if the K value is increased, the level of accuracy, precision, recall, and F-measure decreases. All tests performed with varying amounts of data percentages showed consistent matrix values. This proves that the KNN classification algorithm works well. So that the comparison of training data and different test data results in different values of accuracy, precision, recall, and F-measure. Only experienced a decrease of approximately 1% to 2%.

Then the authors tried to do the test using a cross validation model in the division of test data. The value of the folds is determined 10 times repeatedly tested. This results in low performance compared to manual data sharing. However, if the model cross validation is carried out by increasing the Value of folds, it will result in higher levels of accuracy, precision, recall, and F-measure.

1. CONCLUSION

The conclusion that the author can convey in this study is that the KNN classification algorithm can work well for classifying drug types for patients based on blood pressure levels, cholesterol levels, and the amount of sodium to potassium in the blood. The results of this study showed good performance when the value of k = 3 and the value of training data and test data were 90 : 10. With the highest accuracy value of 98.33%, precision of 98.8%, recall of 96.2%, and F-measure of 97.48%. And experience a decrease in performance when using the cross validation model. By using the KNN algorithm, it is hoped that it can help doctors to determine the right type of medicine for patients based on the patient's condition. Then the author suggests that for the next writing, the development of a cross validation model with the Electre algorithm is used.

2. REFERENCES

Amin, M. Z., & Ali, A. (2017). Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions. *Wavy AI Research Foundation*, (January). https://doi.org/10.13140/RG.2.2.26371.25127

Anita, R., & Bajusz, D. (2019). Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. *Molecules*, *24*, 1–18.

Aprilia, Nursalam, & Panji, A. C. (2016). OBAT DI RSUD SIDOARJO (Right Medication Related to Drug Centralized in RSUD Sidoarjo). *Jurnal INJEC*, *1*(2), 187–196.

Arslan, H., & Arslan, H. (2021). Engineering Science and Technology , an International Journal A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. *Engineering Science and Technology, an International Journal*, *24*(4), 839–847. https://doi.org/10.1016/j.jestch.2020.12.026

*Corresponding author

Baumann, K. (2003). Cross-validation as the objective function for variable-selection techniques. *Trends in Analytical Chemistry*, *22*(6), 395–406. https://doi.org/10.1016/S0165-9936(03)00607-1

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy , bookmaker informedness , and markedness in two-class confusion matrix evaluation. *BioData Mining*, *14*(13), 1–22.

D, T. F. P., D, R. S. P., & D, S. B. P. (2009). Transcription Errors Observed in a Teaching Hospital Fanak. *Arch Iranian Med*, *12*(12), 173–175.

Hossain, E., & Rahaman, M. A. (2019). A Color and Texture Based Approach for the Detection and Classification of Plant Leaf Disease Using KNN Classifier. *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 7–9.

Houcine, T., Mezache, A., & Oudira, H. (2019). ScienceDirect ScienceDirect Model Selection of Sea Clutter Using Cross Validation Method. *Procedia Computer Science*, *158*, 394–400. https://doi.org/10.1016/j.procs.2019.09.067

Kartini, D., Farmadi, A., & Nugrahadi, D. T. (2022). *Perbandingan Nilai K pada Klasifikasi Pneumonia Anak Balita*. *10*(1), 47–53.

Mao, X., Jia, P., Zhang, L., Zhao, P., & Chen, Y. (2015). An Evaluation of the Effects of Human Factors and Ergonomics on Health Care and Patient Safety Practices : A Systematic Review. *PLOS ONE*, 1–19. https://doi.org/10.1371/journal.pone.0129948

Murugan, A., Nair, S. A. H., & Kumar, K. P. S. (2019). Detection of Skin Cancer Using SVM , Random Forest and kNN Classifiers. *Journal of Medical Systems*, *43*, 269.

Patel, I., & Balkrishnan, R. (2010). Medication Error Management around the Globe : An Overview. *Indian Journal of Pharmaceutical Sciences*, 539–545.

Pavaloiu, I. B., Ancuceanu, R., Enache, C. M., & Vasilateanu, A. (2017). Important shape features for Romanian medicinal herb identification based on leaf image. *2017 E-Health and Bioengineering Conference, EHB 2017*, 599–602. https://doi.org/10.1109/EHB.2017.7995495

Poulter, N. R., & Lackland, D. T. (2017). Medication Without Harm : WHO ' s Third Global Patient Safety Challenge. *The Lancet*, *389*(10080), 1680–1681. https://doi.org/10.1016/S0140-6736(17)31047-4

Putri, H., Purnamasari, A. I., Dikananda, A. R., Nurdiawan, O., & Anwar, S. (2021). Penerima Manfaat Bantuan Non Tunai Kartu Keluarga Sejahtera Menggunakan Metode NAÏVE BAYES dan KNN. *Building of Informatics, Technology and Science (BITS)*, *3*(3), 331–337. https://doi.org/10.47065/bits.v3i3.1093

Shamrat, F. M. J. M., & Chakraborty, S. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *23*(1), 463–470. https://doi.org/10.11591/ijeecs.v23.i1.pp463-470

Telaumbanua, F. D., Hulu, P., Nadeak, T. Z., Lumbantong, R. R., & Dharma, A. (2019). Penggunaan Machine Learning. *Jurnal Teknologi Dan Ilmu Komputer*, *3*(1), 57–64.

Vaishnnave, M. P., & Devi, K. S. (2019). Detection and Classification of Groundnut Leaf Diseases using KNN classifier. *Proceeding of International Conference on Systems Computation Automation and Networking*, 1–5.

Vaishnnave, M. P., Suganya Devi, K., Srinivasan, P., & Arutperumjothi, G. (2019). Detection and classification of groundnut leaf diseases using KNN classifier. *2019 IEEE International Conference on System, Computation, Automation and Networking, ICSCAN 2019*. https://doi.org/10.1109/ICSCAN.2019.8878733

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision : A Brief Review. *Computational Intelligence and Neuroscience*.

Wirdiani, N. K. A., Hridayami, P., Widiari, N. P. A., Rismawan, K. D., Candradinata, P. B., & Jayantha, I. P. D. (2019). Face Identification Based on K-Nearest Neighbor. *Scientific Journal of Informatics*, *6*(2), 150–159. https://doi.org/10.15294/sji.v6i2.19503

Xing, W., & Bei, Y. (2019). Medical Health Big Data Classification Based on KNN Classification Algorithm. *IEEE Access*, *8*.

Xiong, L., & Yao, Y. (2021). Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm. *Building and Environment*, *202*(December 2020), 108026. https://doi.org/10.1016/j.buildenv.2021.108026

Yuliati, I. F., & Sihombing, P. R. (2021). Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia Implementation of Machine Learning Method in Risk Classification on Low Birth weight in Indonesia. *Matrik: Jurnal Manajemen, Teknik Informatika, Dan Rekayasa Komputer*, *20*(2), 417–426. https://doi.org/10.30812/matrik.v20i2.1174

Zhou, L. J., Li, X. Da, Zhang, J. N., Huo, W. J., & Chen, Z. (2020). ScienceDirect The Lao Text Classification Method Based on KNN. *Procedia Computer Science*, *166*, 523–528. https://doi.org/10.1016/j.procs.2020.02.053

*Corresponding author