Original Article

Optimization of K Value at K Nearest Neighbor for Classification and Prediction of Healing in Covid-19 Patient

T Taslim[1], Eka Sabna[2], Kursiah Warti Ningsih[3]

[1]Informatics Engineering, Faculty of Computer Science, Universitas Lancang Kuning, Pekanbaru, Indonesia

[2]Informatics Engineering, Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru

[3]Public Health, STIKes Payung Negeri Pekanbaru, Pekanbaru, Indonesia

Abstract

In March 2020 WHO declared the coronavirus outbreak that causes Covid-19 a global pandemic. This disease is an infectious disease caused by the SARS-CoV-2 virus which causes mild to moderate respiratory infections and recovers without requiring special treatment. However, some will become seriously ill and require medical attention. The use of information technology in data science and machine learning can help in the fight against this pandemic, one of which is by creating a method that can classify and predict the recovery period of Covid-19 patients. However, there is no symptom-based model to predict the recovery period of Covid-19 patients that can improve clinical decision-making and become an alternative for resource allocation for treating patients in hospitals. Here we propose to test a symptom-based model to classify and predict the recovery period of Covid-19 patients using the KNN algorithm. This algorithm is a simple algorithm and has been widely used in various fields. This algorithm works by classifying an object into a class based on the neighboring distance of the object. The experiment began with data cleaning of Covid-19 patient data, then the classification process was carried out using the KNN algorithm. The test is carried out with and without optimization on the value of k. The first test without optimization with a default value of k=5 obtained an accuracy value of0.77%, while testing by optimizing the value of k with Grid Search CV obtained an accuracy level of0.86% with a value of k=1. From the test results, it can be seen that optimization on the value of k can increase the level of accuracy by 0.09%. For the prediction of the test results are displayed in the confusion matrix. This research will only focus on efforts to predict the recovery period of Covid-19 patients based on medical record data for Covid-19 patients in Pekanbaru, Indonesia.

Keywords: Patient, classification, healing.

Introduction

The Covid-19 virus outbreak that began in 2019 has claimed many lives until March 11, 2020, the World Health Organization declared Covid-19 a pandemic.(Bhattacharjee et al., 2020). Covid-19 is a new corona virus that has spread throughout the world with initial reports originating from Wuhan, China, turning into a pandemic and causing huge casualties.(Rai et al., 2022)Until August 2021 the number of countries affected were 224 countries with a confirmed number of 211,730,035 and a number of those who died as many as 4,430,697 people, while in Indonesia the number of positive ones was 3,989,060 people recovered as many as 3,571,082 people and a total of 3,571,082 people died as many as 127,214 people(Covid-19 Distribution Map, 2021)

The surge in cases of COVID-19 patients has forced countries in the world to make policies related to handling

Address for correspondence: T Taslim, Informatics Engineering, Faculty of Computer Science, Universitas Lancang Kuning, Pekanbaru, Indonesia

Access this article online

Quick Response Code:

Website: www.pnrjournal.com

DOI: 10.47750/pnr.2022.13.04.235

How to cite this article Taslim T, Sabna E, Ningsih W K, Optimization of K Value at K Nearest Neigbor for Classification and Prediction of Healing in Covid-19 Patient, J Pharm Negative Results 2022;13(4):1699-1708

Covid-19 cases, such as implementing lock downs, implementing restrictions on community activities, providing assistance to affected communities and administering Covid-19 vaccines. Based on data released by the provincial government of Riau, Indonesia, as of September 30, 2022, there are152522patients confirmed positive for Covid-19, where the number of patients recovered was147927and patients died as much as4478. The highest cases were in the city of Pekanbaru with .The number of confirmed Covid-19 patients is64399, the patient recovered as much as62943and died as much1401(Covid-19 Riau Province Update, 2022).

The surge in cases of patients who have tested positive for Covid-19 has also resulted in difficulties for related parties such as the Health Service and Hospitals in dealing with the surge in Covid-19 patients. especially in terms of the provision of drugs, health facilities and infrastructure.The increase in Covid-19 patients has an impact on increasing demand for appropriate medical care and causing an increase in the burden of hospital capacity and resources(Ali et al., 2021). Decision making is very important in medical problems, a wrong decision will cause various problems such as taking unnecessary actions and the high cost of patient care, refusal of care for patients who should be treated and unnecessary procedures.(Sreejith et al., 2020). Regarding the above so that the handling of patients exposed to Covid-19 can be carried out more effectively, it is necessary to predict the patient's recovery period to assisting the health sector related to the priority scale of handling Covid-19 patients.

Many studies have been carried out in modeling predictions of mortality, recovery, and priorities for handling COVID-19 patients. Elgendy uses three machine learning algorithms to predict patients who have a higher probability of death where the evaluation of the model is done by calculating the value of precision, accuracy, recall and value of F(Elgendy et al., 2020). Ali uses Artificial Neural Networks (ANN) to predict the death and recovery of patients infected with Covid-19, uses the Fuzzy Interval Mathematical (FIM) model to select patient priorities and combines the ANN method and the FIM model for priority scheduling priorities.(Ali et al., 2021).Nemati built machine learning and statistical models to predict the length of stay of Covid-19 patients in hospitals, enabling decision makers to prepare for hospital overloads. Ebinger uses a machine learning algorithm to generate a predictive model for the length of stay of Covid-19 patients based on available data. These models are expected to help hospitals prepare for bed capacity requirements, as well as consideration and information for patients about the possibility of length of stay.(Ebinger et al., 2021).

K-nearest-neighbors (KNN) is a supervised machine learning algorithm that is used for classification and regression problems.(Guo et al., 2003), this algorithm is a popular classical classification algorithm(Cover & Hart, 1967)and has been widely applied in many fields(Y. Wang et al., 2022) because it is simple and efficient(Todeschini, 1989)(Chanal et al., 2022). KNN classifies data by calculating the distance between each data point to another point in the set based on the value of the closest neighbor, the data then assigns identical label classes from the data to find patterns in a dataset(Ray, 2019)(Ertuğrul & Tağluk, 2017). Acuration of kNN-based classifiers is highly dependent on the value of k and the type of distance metric(Saini et al., 2013)where the selection of the optimal value of k depends on the dataset or application(Jasmir et al., 2021)

In this study, a prediction of the length of stay for Covid-19 patients will be made based on data from Covid-19 patients at the Pekanbaru Regional General Hospital, Indonesia using the k nearest neighbor algorithm with optimal k selection.

Research Method

Data Collection

In this study, medical record data of Pekanbaru Indonesia General Hospital patients who were declared positive had Covid-19 were used. The data used consists of 399 data from patients who were confirmed positive for Covid-19 from January 2020 to March 2022. The raw data for Covid-19 patients can be seen in table 1 below.

Table 1. Raw data of Covid-19 patients

1	*47*	*0*	*1*	*1*	*1*	*0*	*0*	*0*	*130*	*80*	*80*	*36.8*	*22*	*2*	*category2*
2	18	0	0	0	0	0	0	0	125	85	108	36.5	20	2	category2
3	36	1	1	0	0	1	0	0	123	80	93	36.5	20	2	category1
4	40	0	1	1	1	1	0	1	118	82	96	37.2	98	2	category1
5	22	1	1	1	1	1	0	0	110	70	86	37	20	2	category1
6	35	1	1	0	1	0	0	0	136	34	90	36.1	20	2	category1
7	36	0	1	1	1	1	0	0	160	110	106	36.8	18	2	category1
8	2	0	1	1	1	0	1	0			110	37	22	0	category3
9	51	1	1	1	0	1	1	0	152	100	90	36.1	18	2	category1
10	32	1	1	1	1	1	0	0	147	89	98	36	20	2	category1
11	29	0	0	0	0	0	0	0							category2
12	28	1	1	0	0	0	1	0	150	93	72	36.6	18	1	category1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
399	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Before entering the modeling stage, an analysis of the attributes that will be used is first carried out. Furthermore, according to the research objective, which is to predict the patient's recovery period, all data of patients who died will be deleted so that the remaining 330 data. Furthermore, the data preprocessing process related to missing values and imbalance classification is carried out. The first step in data preprocessing is to delete the attributes of the patient's medical record number, patient name, patient address and several other attributes. For the external attribute in the form of a dependent variable consisting of 3 classes of patient length of treatment category, namely category1, category2 and category3 and other attributes are independent variables (input).

The next stage is to convert the data into numeric and ordinal forms. The attributes and the encode of each attribute can be seen in table 2 below.

Table 2. Attributes and decodes

No	*Attributes*	*Decode*
1	age	-
2	Sex	male=1, female = 0
3	fever	yes=1, no=0
4	out of breath	yes=1, no=0
5	cough	yes=1,no=0
6	afternoon throat	yes=1,no=0
7	nauseous vomit	yes=1, no=0
8	reduced sense of smell	yes=1,no=0
9	systolic blood pressure	-
10	diastolic blood pressure	-
11	pulse	-
12	temperature	-
13	breathing	-
14	nutritional status	good=0, bad=1, moderate=2
15	length of treatment category	Category1 (1 to 10 days)
		Category2 (11 to 20 days)
		Category3 (more than 20 days)

Missing Value

Missing values is a common phenomenon in modern medical research of complex diseases where data often contain numerical or categorical Variables(Faisal & Tutz, 2022). From the results of the initial data recapitulation of this study, there are several missing values, and the handling of these missing values is important in data science(Cheng et al., 2019), ignoring this missing value can cause bias and errors when extracting information from the data(Atiqur et al., 2015). In this study, the handling of missing values uses the k nearest neighbors imputation method, where this imputation method works by calculating the distance according to the Minkowski distance or its variants, and has been shown to be generally efficient for numerical variables(Zhang, 2012).

From the dataset used in this study, several missing values were found, both in numerical data and in categorical data and this needs to be addressed so that there is no bias in extracting information.

Balance class datasets

Unbalanced data classification often occurs in medical diagnosis data, therefore, how to improve patient identification without affecting the classification of other individuals is an urgent problem.(Z. Xu et al., 2020)and this unbalanced data classification is a very important topic in data mining and machine learning(Viloria et al., 2020)because an unbalanced classification will result in an unbalanced class, where the class that has many examples will be the majority class and the class that has fewer examples will be the minority class(Chen et al., 2020), this matter can have a major impact on the value and meaning of accuracy(Luque et al., 2019)(Elreedy & Atiya, 2019)(Buda et al., 2018)and can also lead to biased values against the majority class(Sáez et al., 2016)(Thejas et al., 2022).

Based on the results of initial data processing, it can be seen that there is an unbalanced class from the class labels used, namely class 0 (category1), class 1 (category2) and class 2 (category3), this can be seen in Figure 5 below.

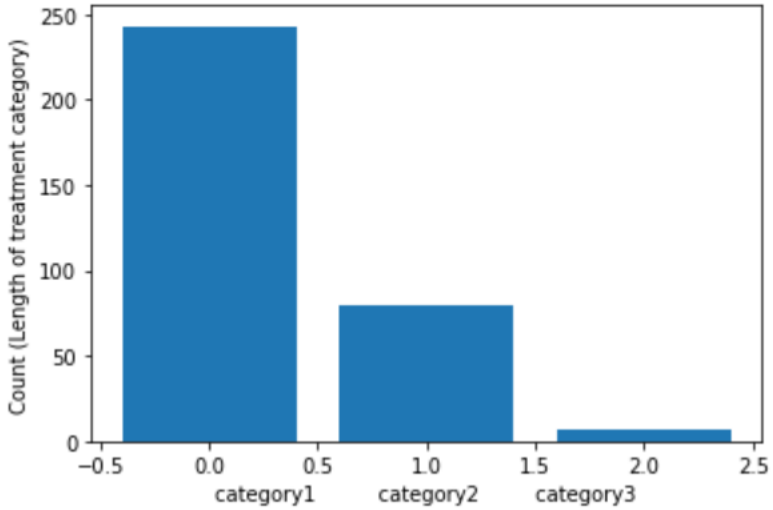

Figure 1. Imbalance class patient length of treatment category

To overcome the unbalanced class in this study, the SMOTE method was used, where this method is one of the most commonly used methods in class imbalance problems.(Chen et al., 2020). SMOTE(Chawla et al., 2002)is an over-sampling technique that is used to reduce imbalances in a dataset that works by creating and adding synthetic data on a minority class based on the nearest neighbor using the kNN algorithm by looking for one or more closest points in a search space.(Sreejith et al., 2020)(Pozzolo et al., nd). The main advantage of SMOTE lies in its ability to model a much larger and less discrete decision region where the number of instances belonging to the majority class remains the same.(Chawla et al., 2002).

Cross validation

Classification is an important task for predicting class values of new instances, and k-fold cross validation is one of the popular algorithms for evaluating the performance of classification algorithms.(Wong, 2015)(Celisse, 2010).

In the k-fold CV process, the original training data set is divided into k separate parts. Each k-part is alternatively used as a validation set and the other k-1 parts are combined to form a training set and each k-part is predicted exactly once. A common practice is to perform a 10-fold CV. For internal validation of a k-fold CV, the root mean squared error of cross-validation (RMSECV) can be used.(L. Xu et al., 2018)(Airola et al., 2011).

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{N}} \quad (1)$$

Test Result

The stages of data processing up to the KNN process can be seen in the following figure.

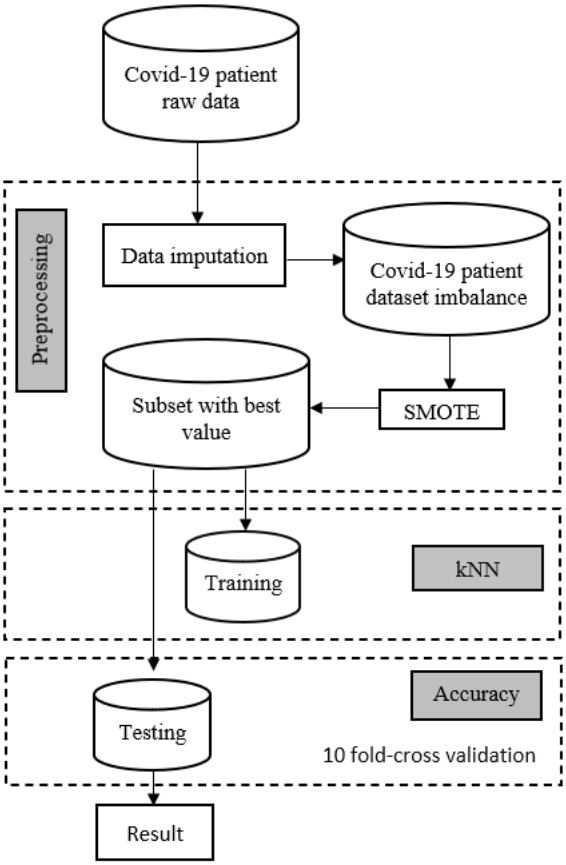

Figure 2. Stages of data processing

Data imputation

The first stage of this research is to preprocess the data starting from the data imputation process. The flow of the data imputation process can be seen in Figure 3 below.

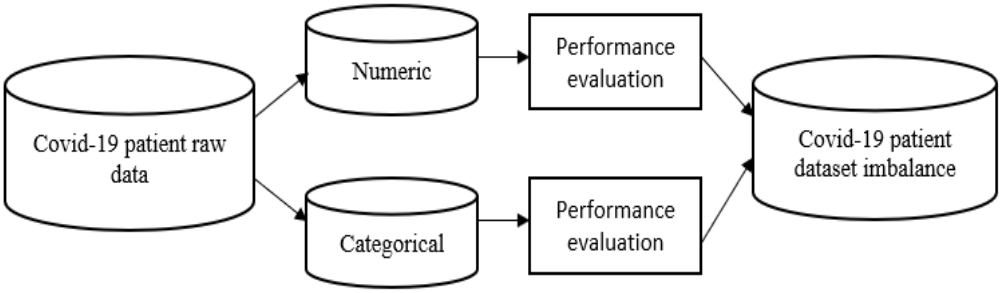

Figure 3. Data imputation process flow

In the imputation process, the data is separated into 2 groups of data, namely numerical data and categorical data where the imputation accuracy results are calculated using the root mean squared error. The accuracy of the imputation results can be seen in Figure 3 and Figure 4, where the optimal K value for numerical data imputation is 5 with an RMSE value of 0.946 and the optimal K for categorical data is 7 with an RMSE value of 4.952.

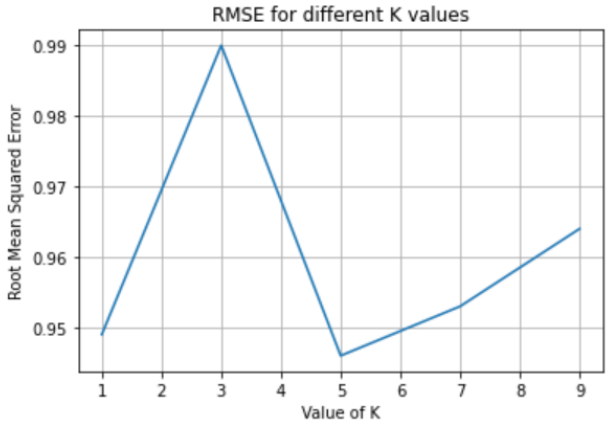

Figure 4. Value of nominal data imputation accuracy

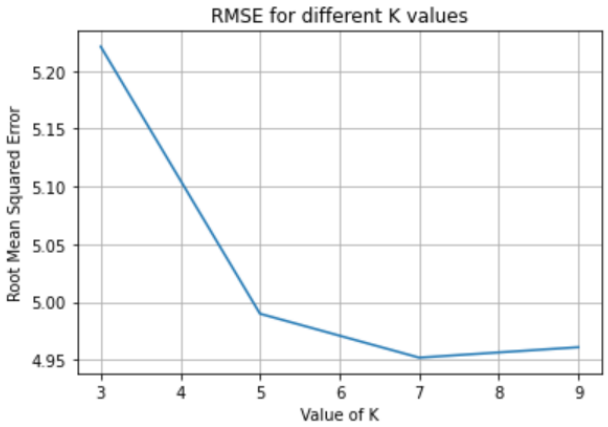

Figure 5. The value of categorical data imputation accuracy

Dataset Balance

After the data imputation process is continued with the imbalance dataset process, where the stages in the imbalance dataset process can be seen in Figure 6.

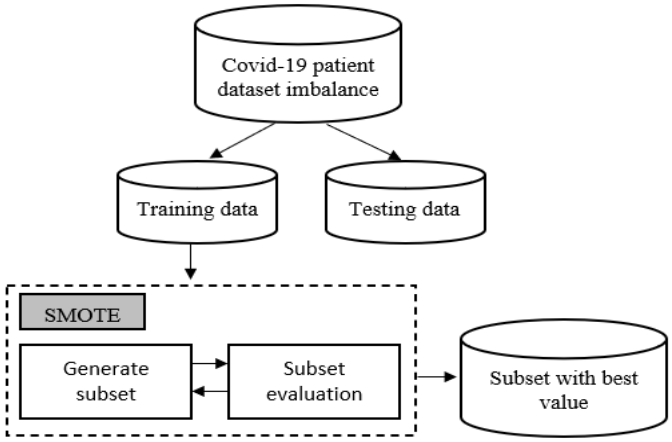

Figure 6. Process flow of imbalance dataset

The results of the initial data set processing are imbalanced datasets where the number of classes in class 0 is 243 (73.636%), class 1 is 80 (24.242%) and class 3 is 7 (2.121%) data. The first stage is to divide the imbalance dataset, 70% training data and 30% testing data. The second stage performs the SMOTE process with an auto sampling strategy and for evaluation of imbalance classification a grid search is used with cross validation to get the best hyperparameters where hyperparameter settings are an important part of any machine learning process to get better model performance(Mantovani, 2016).

After the grid search finds the best combination of hyperparameter values for each model, the set of values is used to adjust the training dataset with 10-cross validation. The test results obtained the best neighbor parameter value 1 with an accuracy value of 0.856%.Over sampling from SMOTE resulted in a class with a balanced distribution in each class as many as 243 (73.636%) data. Figure 7 below shows the distribution of the SMOTE results.

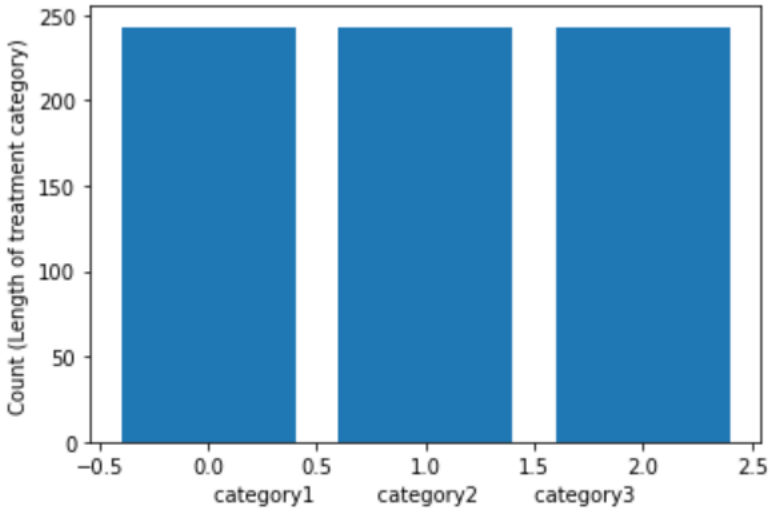

Figure 7. Class distribution of SMOTE results

KNN process

The main purpose of KNN is to find the closest neighbor distance between elements based on the value of k, then every k nearest element gives one vote for that element class.(Singh & Pandey, 2016). To calculate the distance betweenelements in the training data used Euclidean distance with equation (2). Where this equation is a general numerical equation used in multivariate analysis which is calculated through the equation(Elias et al., 2015).

$$Euclidean\ distance\ function = \sqrt{\sum_{j=1}^{k}(x_j - y_j)^2} \qquad (2)$$

The advantage of the kNN algorithm is that it works in a simple way with few influencing factors, but this algorithm also has some disadvantages, such as long processing time and the selection of the ideal k value.(H. Wang et al., 2022).In this study, hypertuning parameters to find the optimal value of k.

In this study, the kNN process will be carried out through several stages, namely, first the Covid-19 patient dataset (Subset with best value) will be divided into 70% training data and 30% testing data. The second stage is Building and training the modelwhere the starting point is k = 5, where this value is the default value of k in the kNN algorithm(Das et al., 2022). From the test results obtained an accuracy rate of 0.77%. The third stage is to evaluate the performanceof classification algorithm KNN with 10-fold CV, the original training data set is divided into k separate sections. Each k-part is alternatively used as a validation set and the other k-1 parts are combined to form a training set and each k-part is predicted exactly once. The results of this cross validation obtained an average value of 0.780%. The last step is hypertuning the parameters which aims to find the optimal parameters in the model used, in this study hypertuning the GridSearchCV parameter is used. GridSearchCV works by training the model several times over a predefined range of parameters. Thus, the model can be tested with each parameter and find out the optimal value to get the best accuracy results. The test is carried out with a neighbor value range of 1 to 25. From our test results, the results of the hypertuning parameter are obtained, namely, the number of neighbors is 1 with an accuracy of 0.856. As forCross validated Accuracy, cross-validated MSE and comparison of the level of accuracy of training data and testing data can be

seen in Figures 8, 9 and 10 below.

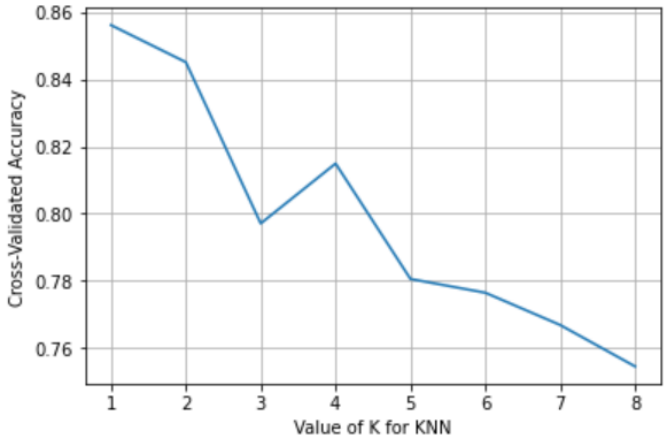

Figure 8. Cross validation Accuracy

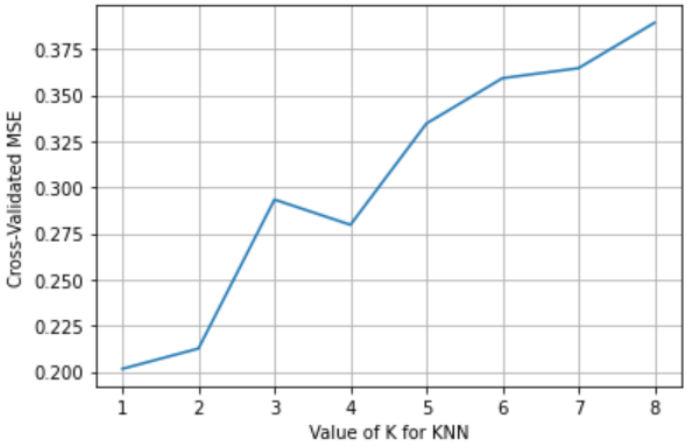

Figure 9. cross validated MSE

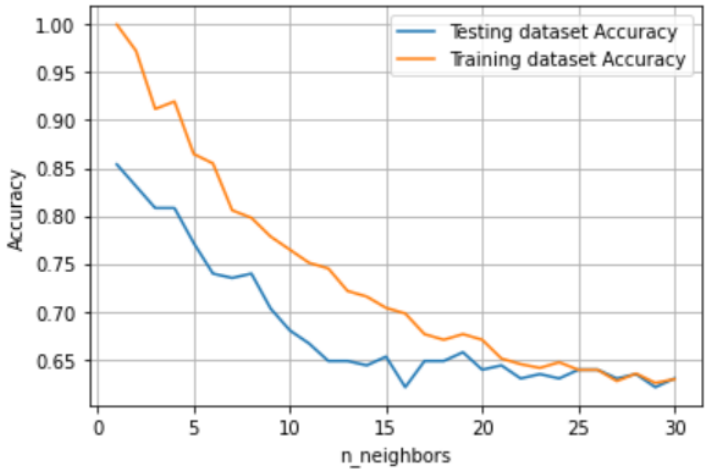

Figure 10. Comparison of the accuracy of testing data and training data

Table 3. Confusion Matrix

		Predicted		
		Class 0	Class 1	Class 2
actual	Class 0	41	21	5
	Class 1	5	72	1
	Class 2	0	0	74

Table 4. Classification Report Summary

Class	*Precision*	*Recall*	*F1-Score*
Class 0	0.89	0.61	0.73
Class 1	0.77	0.92	0.84
Class 2	0.93	1.00	0.96
Average	0.86	0.85	0.85

Conclusion

KNN is a simple classification algorithm and is widely used in various fields, but this algorithm has a weakness in determining the adjacency value (k). Testing with a default value of k = 5 produces an accuracy rate of 0.77%, while testing by optimizing the value of k using tuning hyper parameter with Grid Search to get the optimal neighbor value, the result is that the best number of neighbors is 1 with an accuracy rate of 0.856. The test results show an increase in the amount of accuracy of 0.09%. The prediction results of the classification performance measurement with the confusion matrix show that there are 41 elements of class 0 that were correctly detected as class 0, 21 elements of class 0 detected as class 1 and 5 elements of class 0 as class 2. There are 5 elements of class 1 detected as class 0, 72 class 1 elements detected were correct as class 1 and 1 class 1 element were detected as class 2. For class 2 there were 74 elements that were correct as class 2 and there were no elements detected in class 0 and class 1.

References

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., & Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. Computational Statistics & Data Analysis, 55(4), 1828–1844. https://doi.org/https://doi.org/10.1016/j.csda.2010.11.018

Ali, M., Ben, A., & Abdelhedi, M. (2021). Real-time prediction of COVID-19 patients health situations using Artificial Neural Networks and Fuzzy Interval Mathematical modeling. Applied Soft Computing, 110, 107643. https://doi.org/10.1016/j.asoc.2021.107643

Atiqur, S., Huang, Y., Claassen, J., Heintzman, N., & Kleinberg, S. (2015). Combining Fourier and lagged k -nearest neighbor imputation for biomedical time series data. JOURNAL OF BIOMEDICAL INFORMATICS, 58, 198–207. https://doi.org/10.1016/j.jbi.2015.10.004

Bhattacharjee, A., Kumar, M., & Kumar, K. (2020). When COVID-19 will decline in India ? Prediction by combination of recovery and case load rate. Clinical Epidemiology and Global Health, May, 0–1. https://doi.org/10.1016/j.cegh.2020.06.004

Buda, M., Maki, A., & Mazurowski, MA (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259. https://doi.org/https://doi.org/10.1016/j.neunet.2018.07.011

Celisse, A. (2010). A survey of cross-validation procedures for model selection. 4, 40–79. https://doi.org/10.1214/09-SS054

Chanal, D., Yousfi Steiner, N., Petrone, R., Chamagne, D., & Pera, M.-C. (2022). Online Diagnosis of PEM Fuel Cell by Fuzzy C-Means Clustering. In LF Cabeza (Ed.), Encyclopedia of Energy Storage (pp. 359–393). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-819723-3.00099-8

Chawla, N. V, Bowyer, KW, Hall, LO, & Kegelmeyer, WP (2002). SMOTE : Synthetic Minority Over-sampling Technique SMOTE : Synthetic Minority Over-sampling Technique. June. https://doi.org/10.1613/jair.953

Chen, B., Xia, S., Chen, Z., Wang, B., & Wang, G. (2020). RSMOTE : A self-adaptive robust SMOTE for imbalanced problems with label noise. Information Sciences. https://doi.org/10.1016/j.ins.2020.10.013

Cheng, C., Chan, C., & Sheu, Y. (2019). Engineering Applications of Artificial Intelligence A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. Engineering Applications of Artificial Intelligence, 81(October 2017), 283–299. https://doi.org/10.1016/j.engappai.2019.03.003

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Das, C., Sahoo, AK, & Pradhan, C. (2022). Chapter 12 - Multicriteria recommender system using different approaches. In S. Mishra, HK Tripathy, PK Mallick, AK Sangaiah, & G.-S. Chae (Eds.), Cognitive Big Data Intelligence with a Metaheuristic Approach (pp. 259–277). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-85117-6.00011-X

Ebinger, J., Wells, M., Ouyang, D., Davis, T., Kaufman, N., Cheng, S., & Chugh, S. (2021). Intelligence-Based Medicine A Machine Learning Algorithm Predicts Duration of hospitalization in COVID-19 patients. Intelligence-Based Medicine, 5, 100035. https://doi.org/10.1016/j.ibmed.2021.100035

Elgendy, O., Nasir, Q., & Nassif, AB (2020). Death / Recovery Prediction for Covid-19 Patients using Machine Learning. January 2021. https://doi.org/10.46300/91015.2020.14.25

Elias, J., Ferreira, V., Henrique, C., Miranda, RM De, & Figueiredo, AF De. (2015). Q uímica educacion the representative elements. Educación Química, 26(3), 195–201. https://doi.org/10.1016/j.eq.2015.05.004

Elreedy, D., & Atiya, AF (2019). PT USCR. Information Sciences. https://doi.org/10.1016/j.ins.2019.07.070

Ertuğrul, . F., & Tağluk, ME (2017). A novel version of k nearest neighbor: Dependent nearest neighbor. Applied Soft Computing, 55, 480–490. https://doi.org/https://doi.org/10.1016/j.asoc.2017.02.020

Faisal, S., & Tutz, G. (2022). Nearest neighbor imputation for categorical data by weighting of attributes. Information Sciences, 592, 306–319. https://doi.org/10.1016/j.ins.2022.01.056

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. In R. Meersman, Z. Tari, & DC Schmidt (Eds.), On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE (pp. 986–996). Springer Berlin Heidelberg.

Jasmir, J., Nurmaini, S., & Tutuko, B. (2021). Fine-Grained Algorithm for Improving KNN Computational Performance on Clinical Trials Text Classification. Big Data and Cognitive Computing, 5(4). https://doi.org/10.3390/bdcc5040060

Luque, A., Carrasco, A., Martín, A., & De, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Mantovani, RG (2016). Hyper-parameter Tuning of a Decision Tree Induction Algorithm. https://doi.org/10.1109/BRACIS.2016.62

Nemati, M. (2020). Article Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. Patterns, 1(5), 100074. https://doi.org/10.1016/j.patter.2020.100074

Covid-19 Distribution Map. (2021). https://covid19.go.id/peta-sebaran-covid19

Pozzolo, AD, Caelen, O., Johnson, RA, & Bontempi, G. (nd). Calibrating Probability with Undersampling for Unbalanced Classification.

Rai, N., Kaushik, N., Kumar, D., Raj, C., & Ali, A. (2022). Mortality prediction of COVID-19 patients using soft voting classifier. International Journal of Cognitive Computing in Engineering, 3, 172–179. https://doi.org/https://doi.org/10.1016/j.ijcce.2022.09.001

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 35–39.

Sáez, JA, Krawczyk, B., & Woźniak, M. (2016). Author ' s Accepted Manuscript Analyzing the oversampling of different classes and Reference : To appear in : Pattern Recognition. Pattern Recognition. https://doi.org/10.1016/j.patcog.2016.03.012

Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. Journal of Advanced Research, 4(4), 331–344. https://doi.org/https://doi.org/10.1016/j.jare.2012.05.007

Singh, A., & Pandey, B. (2016). An euclidean distance based KNN computational method for assessing degree of liver damage. 2016 International Conference on Inventive Computation Technologies (ICICT), 1, 1–4. https://doi.org/10.1109/INVENTIVE.2016.7823222

Sreejith, S., Nehemiah, HK, & Kannan, A. (2020). Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. Computers in Biology and Medicine, 126(September), 103991. https://doi.org/10.1016/j.compbiomed.2020.103991

Thejas, GS, Hariprasad, Y., Iyengar, SS, Sunitha, NR, & Badrinath, P. (2022). Machine Learning with Applications An extension of Synthetic Minority Oversampling Technique based on Kalman filter for imbalanced datasets. Machine Learning with Applications, 8(January), 100267. https://doi.org/10.1016/j.mlwa.2022.100267

Todeschini, R. (1989). k-nearest neighbor method: The influence of data transformations and metrics. Chemometrics and Intelligent Laboratory Systems, 6(3), 213–220. https://doi.org/https://doi.org/10.1016/0169-7439(89)80086-3

Riau Province COVID-19 Update. (2022). https://corona.riau.go.id/

Viloria, A., Bonerge, O., Lezama, P., & Mercado-caruzo, N. (2020). ScienceDirect ScienceDirect Unbalanced data processing using oversampling : Machine Learning Unbalanced data processing using oversampling : Machine Learning. Procedia Computer Science, 175, 108–113. https://doi.org/10.1016/j.procs.2020.07.018

Wang, H., Xu, P., & Zhao, J. (2022). Improved KNN algorithms of spherical regions based on clustering and region division. Alexandria Engineering Journal, 61(5), 3571–3585 . https://doi.org/https://doi.org/10.1016/j.aej.2021.09.004

Wang, Y., Pan, Z., & Dong, J. (2022). A new two-layer nearest neighbor selection method for kNN classifier. Knowledge-Based Systems, 235, 107604. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107604

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48(9), 2839–2846 . https://doi.org/https://doi.org/10.1016/j.patcog.2015.03.009

Xu, L., Hu, O., Guo, Y., Zhang, M., Lu, D., Cai, C.-B., Xie, S., Goodarzi, M., Fu, H.-Y. , & She, Y.-B. (2018). Representative splitting cross validation. Chemometrics and Intelligent Laboratory Systems, 183, 29–35. https://doi.org/https://doi.org/10.1016/j.chemolab.2018.10.008

Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. Journal of Biomedical Informatics, 107(May 2019), 103465. https://doi.org/10.1016/j.jbi.2020.103465

Zhang, S. (2012). The Journal of Systems and Software Nearest neighbor selection for iteratively k NN imputation. The Journal of Systems & Software, 85(11), 2541–2552 . https://doi.org/10.1016/j.jss.2012.05.073