WWJMRD 2023; 9(11): 09-14
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
Impact Factor SJIF 2017: 5.182 2018: 5.51, (ISI) 2020-2021: 1.361
E-ISSN: 2454-6615

Januarius Anongtop
Department of Statistics, IPB University, Bogor, Indonesia.

Hari Wijayanto
Department of Statistics, IPB University, Bogor, Indonesia.

Budi Susetyo
Department of Statistics, IPB University, Bogor, Indonesia.

Correspondence:
Januarius Anongtop
Department of Statistics, IPB University, Bogor, Indonesia.

Application of Variable Selection on K-Nearest Neighbors and Support Vector Machine for Classification of the Quality of Junior High Schools in Papua Province

Januarius Anongtop, Hari Wijayanto, Budi Susetyo

Abstract
Classification is a supervised learning algorithm that aims to define an object's class into a predetermined type or category. Data with many variables will affect the classification algorithm training. Boruta algorithm is a wrapper technique built around a random forest algorithm to select relevant variables. K-Nearest Neighbors (KNN) is a classification algorithm that determines a new class based on the nearest neighbors in the unknown sample. Support Vector Machines (SVM) is a classification algorithm that aims to find the optimal hyperplanes that can linearly separate classes. Boruta produces five contributing factors to the quality of Junior High Schools in Papua. The KNN model with relevant variables and SMOTE training data produces an accuracy of 70%, sensitivity of 69%, specificity of 100%, and F1-score of 83%. The SVM model with relevant variables and SMOTE training data produces an accuracy of 81%, sensitivity of 83%, specificity of 50%, and F1-score of 89%. The SVM model performs better than the KNN model.

Keywords: Classification, Quality of Junior High Schools, Boruta, KNN, and SVM.

1. Introduction

Classification is a technique in supervised learning that aims at determining the class of an object into a predetermined type or category (1). Variable selection is a key step in identifying the relevant variables from a dataset with many variables in classification model building (2). Relevant variables can reduce the processing time in model training but also minimize the level of complexity of the model, make the model more understandable and improve model performance evaluation (3).

Several classification algorithms have been developed in dealing with classification problems, such as Neural Networks, Naïve Bayes, K-Nearest Neighbor, Decision Trees, Logistic Regression, and Support Vector Machines (4). Data with many variables can affect algorithm training time (5), computational and model maintenance costs (6), and prediction accuracy (3).

Previous studies have focused on variable selection methods. Selvaraj et al.(2018) compared the Boruta and Chi-Square methods to investigate the effect of selecting important variables on several classification models. Their findings show that Boruta works more effectively on data with many variables and improves classification accuracy. Another study conducted by Yahaya et al. (2020) applied Boruta on mobile malicious software detection applications. They used selected variables as input for model training of several classification algorithms such as Random Forest, J48, and GLM. As a result, the prediction accuracy increases in detecting mobile malicious software. Sulastri et al. (2019) have researched comparing K-Nearest Neighbor (KNN), Naïve Bayes, and Neural Networks models for the prediction of Hepatitis disease, and their study shows that KNN algorithm has higher accuracy in predicting the disease than the other two algorithms. Novianti & Purnami (2012) have also researched comparing Logistic Regression and Support Vector Machines (SVM) algorithm to diagnose breast cancer patients. Their findings show that the SVM model can predict

better than the Logistic Regression model. Hence, this study will focus on selecting important variables using Boruta, and then its results will be used as input to compare the performance of the KNN and SVM algorithms in classification of the quality of Junior High Schools in Papua province based on the accuracy, sensitivity, specificity, and F1-score. Model validation and parameter tuning are done through k-fold cross validation and imbalanced data is handled using Synthetic Minority Oversampling Technique (SMOTE).

2. Theoretical Framework

2.1Boruta

Boruta is a variable selection algorithm that applies the random Forest algorithm in estimating the variables importance (11). The random Forest algorithm has the advantage of estimating the importance of a variable. Boruta creates random copies of the original data to add randomness to an extended system, and then asses the variables importance from that system (5). The Z-score in this algorithm provides an essential measure of the importance of each variable (3). The comparison of the Z-score of original variables to the scores of the random variables is obtained (12). The original variables with a higher Z-score than the maximum Z-score of all random (shadow) variables give the relevant variable. Boruta categorizes the variables into confirmed, rejected, and tentative. The confirmed variable is decided when an original variable has a Z-score higher than the maximum Z-score of all random variables. On the other hand, the rejected variable is a variable with lower Z-score than the maximum Z-score of all random variables. The tentative variable is all variables left undecided due to the algorithm reaching the predetermined maximum number of random forests runs.

2.2. Imbalanced Data

Imbalanced data in Machine Learning can be defined as the condition where the data has the proportion of one class larger than the other (13). Imbalanced in data can affect the performance of a classification model, specifically in classifying the minority class (14). As a result, the classification model will be subjective to the majority class (15).

Synthetic Minority Oversampling Technique (SMOTE) is a method used for handling imbalanced data. This method balances the data by adding data in the minority class from artificial data. Hence, the quantity of the minority and majority classes becomes more balanced.

2.3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a simple and widely used supervised learning algorithm for classification (16). It classifies an object based on the distance between the data and its k closest neighbors in its training data (17). The important parameter in KNN is the *k* parameter. The best value of k depends on the data itself. Larger values of k reduce the effect of noise on the classifications, but make margins between classes less distinct. Two techniques to determine a good value of *k* are calculating the square root of the number of training examples and cross validation(18). Suppose a line connects the points of test record and training record in n dimensional space, the length of the line is the Euclidean distance. Suppose a, b both have n variables with coordinates $(a_1, a_2, \dots, a_n)$ and $(b_1, b_2, \dots, b_n)$. The Euclidean distance is calculated as in Eq.(1)(19):

$$dist(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \quad (1)$$

2.4. Support Vector Machines (SVM)

Support Vector Machine (SVM) is a supervised learning technique for classification and regression. The SVM applies the principal of structural risk minimization (SRM) method and the statistical learning theory. SVM is one of the powerful classification algorithms due to its generalization of the solution to the classification problem compared to other classification algorithms (18). SVM aims to find the optimal hyperplane that can linearly separate the data. In n-dimensional input space planes, a hyperplane is in $p-1$ dimension. The best hyperplane is the hyperplane with the maximum margin. The solution to linear SVM can be expressed in Eq. (2)(20):

$$f(x) = sgn((\boldsymbol{w}.\boldsymbol{x}) + b) \quad (2)$$

The margin is the closest distance of the observation to the hyperplane as given in Eq. (3).

$$\frac{2}{\|\boldsymbol{w}\|} \quad (3)$$

The best hyperplane is obtained by minimizing the margin as given in Eq. (4).

$$\frac{1}{2}\|\boldsymbol{w}\|^2 \quad (4)$$

With constraint

$$y_i((\boldsymbol{w}.\boldsymbol{x_i}) + b) \geq 1 \quad (5)$$

Where **w** is a vector weight, $\boldsymbol{x_i}$ is the training set with i=1,…,m.

The solution to linear problem in Eq.(4) under the constraint in Eq. (5) is given by the Lagrange multiplier as seen in Eq. (6) (20):

$$L_p(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{w}^T x_i + b) - 1], \text{ where } \alpha_i \geq 0 \quad (6)$$

Where α_i is the Lagrange multiplier. The Lagrangian has to be minimized with respect to **w**, b and maximized with respect to $\alpha_i \geq 0$. Eq.(6) is the Lagrange primal.

The above explanation is for solving linear classification. However, it is difficult to find linearly separated data in real-life situation. The introduction of slack variables ($\xi_i \geq 0$) and a penalty cost ($C > 0$) as soft margin are the solution for non-linear data. The optimal hyperplane is

obtained by minimizing(20):

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + \mathrm{C}\sum_{i=1}^{m}\xi_i \quad (7)$$

With respect to

$$y_i\big((\boldsymbol{w}.x_i) + b\big) \geq 1 - \xi_i, \xi_i \geq 0 \quad (8)$$

The final solution of SVM for non-linear data becomes (21):

$$f(x) = sign(\sum_{i=1}^{N}\alpha_i y_i(\boldsymbol{x}.\boldsymbol{x_i}) + b) \quad (9)$$

Another trick for non-linear data is using the kernel function. The kernel function maps the data into high dimensional space as given in Eq. (10)(19).

$$K\big(\overrightarrow{x_i}, \overrightarrow{x_j}\big) = \phi(\overrightarrow{x_i}).\phi(\overrightarrow{x_j}) \quad (10)$$

The solution function of SVM now becomes:

$$f(x) = sign(\sum_{i=1}^{ns}\alpha_i y_i K(x, x_i) + b) \quad (11)$$

The following are different types of kernel in SVM (19):

1. Polynomial kernel with degree *d*

$$K\big(\overrightarrow{x_i}, \overrightarrow{x_j}\big) = (\overrightarrow{x_i}.\overrightarrow{x_j} + 1)^d \quad (12)$$

2. Gaussian Radial Based Function

$$K\big(\overrightarrow{x_i}, \overrightarrow{x_j}\big) = exp\left(-\frac{\|\overrightarrow{x_i} - \overrightarrow{x_j}\|^2}{2\sigma^2}\right) \quad (13)$$

Where x_i and x_j are the pair of training data, d, C, σ >0.

2.5. Cross Validation

Cross validation is a method in used for assessing the expected performance of a classification model (19). K-fold cross validation is technique in cross validation that divides the original sample into k equal sized subsamples or folds (22). A single subsample out of all k samples is kept as final testing data for the model validation. The remaining k-1 subsamples are used as training data. This process is repeated k times, corresponding to the k folds. Each of the k subsamples is used once as the validation data. At last, corresponding k results are averaged to generate the final best model estimation.

2.6. Classification Model Evaluation

The confusion matrix is a summary of prediction results on a classification problem. Table 1 shows the confusion matrix.

Table 1: Confusion Matrix.

		Predicted Class	
		Positive	***Negative***
Observed Class	***Positive***	TP	FN
	Negative	FP	TN

True Positive (TP) is the number of correctly classified positive classes as positive, False Positive (FP) is the number incorrectly classified negative classes as positive, True Negative (TN) is the number of falsely classified positive classes as negative, and the False Negative (FN) is the number of falsely classified positive classes as negative. Hence, the evaluation performance classification model accuracy, precision, recall, and F1-score can be calculated by the Eq. (14), (15), (16), and (17)(18).

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}} \quad (14)$$

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (15)$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (16)$$

$$F1 - score = \frac{2 \times precision \times recall}{recall + precision} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (17)$$

3. Methodology

3.1. Data

The data for this study is the data of the characteristics of Junior High Schools in Papua province from Basic Education Data (DAPODIK) of the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia and the Education Management Information System (EMIS) of the Ministry of Religion of the Republic of Indonesia in 2019. The data consists of 179 samples, 14 predictors, and a target variable. The fourteen numeric predictors are percentage of graduates per number of students (x1), percentage of students who drop out (DO) per number of students (x2), ratio number of students per study group (x3) , ratio number of teachers per number of students (x4), percentage of teachers who have certificates per number of teachers (x5), percentage of teachers with Bachelor's degrees per number of teachers (x6), Ratio of administrative staff per number of groups (x7), Ratio of the number of classrooms per group (x8), ratio of the number of complete classrooms per group (x9), percentage of the number of complete classrooms per number of classrooms (x10), ratio of the number of computers per number of students (x11), ratio of the number of students per number of toilets (x12), laboratory availability ratio (Availability of Science, Computer and Language Labs) (x13), and supporting space availability ratio (x14). The target variable has two categories (The quality: "Good" if the average National Examination score (UN) $>$ 70) and "Bad" if the average National Examination score (UN) $\leq$70)).

3.2. Analysis Procedure

This study is analyzed using R version 4.0.2 for Windows 64bit with classInt, Boruta, classification and regression training (CARET), and utility-based learning (UBL) packages. The steps in the analysis are as follows:

1. Create a new dataset from SMP/MTs data that consists of 14 predictors and one target variable (National

Examination score (UN)).
2. Handling missing data using multivariate imputation by chained equation (MICE).
3. Discretization of average National Examination score of Junior High Schools (target variable in two categories: "Good" for average national examination score (UN) > 70 and "Bad" for average national examination score (UN) ≤70).
4. Normalization of numerical predictors using the min – max normalization method.
5. Variable selection analysis using Boruta package (set the number of iterations (maxRuns) to 500).
6. Divide the data into 70% training data and 30% testing data.
7. Perform SMOTE using UBL package on training data and apply the method 'Euclidean'.
8. 10-fold cross validation for model validation and parameters searching.
9. KNN and SVM classification analysis are done in two scenarios: 1) Using all predictors and with SMOTE training data and 2) Using the Boruta results and with SMOTE training data.
10. Comparing the algorithm result.
11. Summarize the result of the Boruta variable selection and comparison of the KNN and SVM algorithms

4. Results

4.1. Data Description

The quality of Junior High Schools in Papua province was obtained by discretizing the average of national examination (UN) score. Figure 1 shows the quality of Junior High Schools in Papua province in 2019.

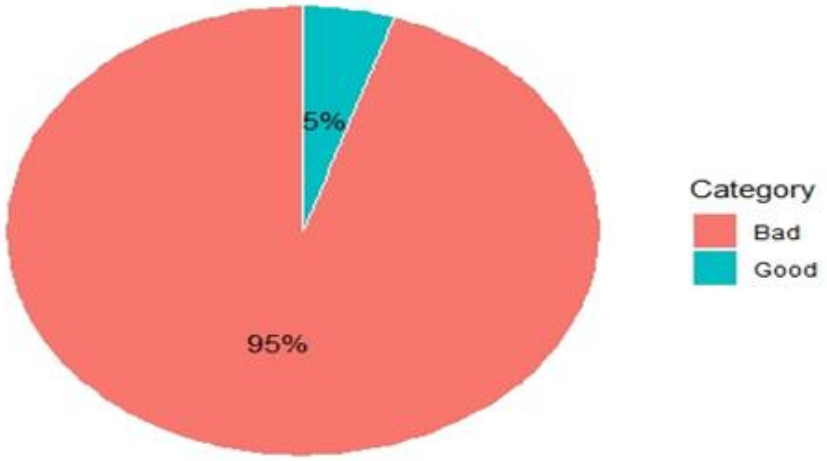

Fig. 1: Quality of Junior High Schools in Papua Province

4.2. Boruta Variable Selection Result

The data comprises of 179 samples and 14 variables. Boruta performs 290 random Forests runs in 24.44191 seconds. Among the 14 variables, nine were unimportant and five were important. The confirmed important variables are the percentage of graduates per number of students (x1), ratio of number of students per study group (x3), ratio of number of teachers per number of students (x4), percentage of the number of complete classrooms per number of classrooms (x10), and ratio of the number of students per number of toilets (x12). Figure 2 shows the Z scores variability among variables during the Boruta run.

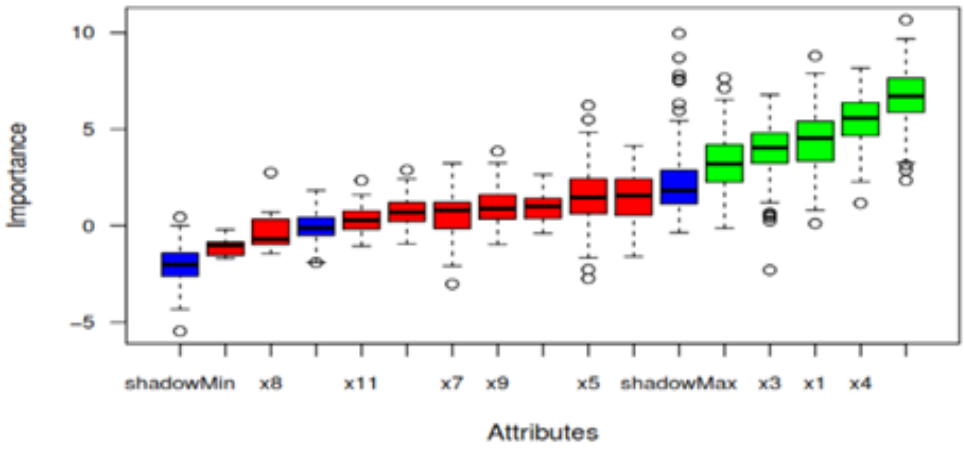

Fig. 2: Boruta result plot for the characteristics of Junior High School in Papua data

Blue boxplots represent the minimum, average, and maximum Z-score of a random (shadow) variable. Furthermore, Red and green boxplots show the Z-scores of the rejected and confirmed variables.

4.3. Classification Result

Before the classification analysis, model training employed 70% of the dataset and the remaining 30% of the dataset for testing the model. The training data consists of 126 samples

and 53 samples for the testing data. As seen in Figure 1, imbalanced data occurred. Therefore, SMOTE was applied to handle it. Table 2 shows the training data of balanced and imbalanced.

Table 2 Training Data.

Class	Imbalanced Data	Balanced Data
Bad	119	63
Good	7	63
Total	126	126

This classification analysis applied two scenarios, first scenario was used all variables and SMOTE for imbalanced data handling, and the second used selected important variables and SMOTE for handling the imbalance data. This study applied 10-fold cross validation for model validation and parameters searching. The parameters for k for KNN, C and gamma for SVM were set to the default value in the R system and then applied 10-fold cross validation to obtain the best parameter. The choice of kernel in SVM could affect the performance of SVM model. Therefore, the Gaussian Radial Based Function (RBF) kernel for SVM model was used since it gives the best performance for non-linear data (23). The best model and parameters' value are shown in Table 3.

Table 3 The best parameters of 10-fold Cross Validation.

Model	Best Parameter Values		
	k	Gamma	C
KNN + SMOTE	5	-	-
SVM + SMOTE	-	17	1
KNN + Boruta +SMOTE	5	-	-
SVM + Boruta + SMOTE	-	4	0.5

The optimal model of two scenarios were evaluated using testing data. Table 4 shows the performance evaluation of kNN and SVM model.

Table 4 Evaluation Performance Results of kNN and SVM model.

Model	Variables	Accuracy	Sensitivity	Specificity	F1-*Score*
KNN + SMOTE	14	0.70	0.73	0	0.82
SVM + SMOTE	14	0.91	0.94	0	0.95
KNN + Boruta + SMOTE	5	0.70	0.69	1	0.81
SVM + Boruta + SMOTE	5	0.81	0.82	0.5	0.89

5. Discussion

Boruta algorithm has been used as a method for variable selection in previous studies by Yang et al. (2017), Maya & Bhargavi (2018), and Yahaya et al. (2020). Their studies have shown that Boruta algorithm can identify all relevant variables and improve the classification model using relevant variable in training it. In this study, Boruta produces five important variables that affect the quality of Junior High Schools in Papua Province Indonesia. As seen in Figure 1, 95% of SMP/MTs in Papua province are with bad quality. There are five contributing factors to the quality of SMP/MTs in Papua province. They are the percentage of graduates per number of students, the ratio of number of students per study group, the ratio of number of teachers per number of students, the percentage of the number of complete classrooms per number of classrooms, and the ratio of the number of students per number of toilets. This study confirmed few contributing factors to the quality of Junior High Schools in Papua province as previously investigated by Sumule et al.(2022), such as the quantity and quality of teachers.

Yang et al. (2017) proved that there was improvement in prediction accuracy when applying relevant variables to a classification problem. As seen in table, the KNN model without using relevant variables and SMOTE for imbalanced data handling give the accuracy of 70%, sensitivity of 73%, specificity of 0%, and F1-score of 82%. The SVM model without using relevant variables and SMOTE for imbalanced data handling give the accuracy of 91%, sensitivity of 94%, specificity of 0%, and F1-score of 95%. On the other hand, when applying the relevant variables and imbalanced data handling in training the classification model, the KNN model give the accuracy of 70%, sensitivity of 69%, specificity of 100%, and F1-score of 81%, and SVM give the accuracy of 81%, sensitivity of 82%, specificity of 50%, and F1-score of 89%. Boruta algorithm improves the specificity of 0% to 100% and 50% of the KNN and SVM model respectively. However, the KNN algorithm does not perform well with or without applying SMOTE for imbalanced data. Moreover, SVM shows a high accuracy of 81% with only five relevant variables. The SVM model outperform the KNN model in accuracy, sensitivity, and F1-score. Therefore, SVM model is the best model for prediction of the quality Junior High School in Papua province.

6. Conclusion

This study presents some important factors affecting the quality of Junior High Schools in Papua province using Boruta. Furthermore, Boruta affects the performance of KNN and SVM classification algorithms. This result has been justified using 10-fold cross validation. SMOTE solved the presence of the imbalanced data in this study. Support Vector Machines model is preferable model for predicting the quality of Junior High Schools in Papua.

References

1. Parapat IM, Furqon MT, Sutrisno. Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. ptiik. 2018;2(10):3163–9.
2. Saran MS, Kiran NS, Saritha B. Feature Selection & Classification on Human Activity Recognition using Machine Learning Approaches. International Journal for Scientific Research & Development. 2018;6(07):390–3.
3. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on

machine learning methods. Jurnal of Big Data [Internet]. 2020; Available from: https://doi.org/10.1186/s40537-020-00327-4
4. Roihan A, Sunarya PA, Rafika AS. Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. IJCIT [Internet]. 2020 May 1 [cited 2023 Jun 27];5(1). Available from: https://ejournal.bsi.ac.id/ejurnal/index.php/ijcit/article/view/7951
5. Maya GPS, Bhargavi R. Feature Selection for Yield Prediction Using Boruta Algorithm. International Journal of Pure and Applied Mathematics. 2018;118(No. 22):139–44.
6. Rasheed AA. Feature Selection: An Assessment of Some Evolving Methodologies. TURCOMAT. 2021 Apr 16;12(7):1612–7.
7. Selvaraj C, Bhalaji N, Sundhara Kumar KB. Empirical study of feature selection methods over classification algorithms. IJISTA. 2018;17(1/2):98.
8. Yahaya CAC, Firdaus A, Ernawan F, Razak MFA. Automated Feature Selection using Boruta Algorithm to Detect Mobile Malware. International Journal of Advanced Trends in Computer Science and Engineering. 2020;9(No. 5).
9. Sulastri, Hadiono K, Anwar MT. Analisis Perbandingan Klasifikasi Prediksi Penyakit Hepatitis dengan Menggunakan Algoritma k-Nearest Negihbor, Naive Bayes, dan Neural Network. Jurnal DINAMIK. 2019 Jul 1;24(2):82–91.
10. Novianti FA, Purnami SW. Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi. Jurnal Sains dan Seni ITS. 2012;1(No 1).
11. Kursa MB, Jankowski A, Rudnicki WR. Boruta – A System for Feature Selection. Fundamenta Informaticae. 2010;101(4):271–85.
12. Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A. Feature Selection methods and genomic big data: a systematic review. Springer Open [Internet]. 2019;6(79). Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0241-0
13. Santoso B, Wijayanto H, Notodiputro KA, Sartono B. K-Neighbor over-sampling with cleaning data: a new approach to improve classification performance in data sets with class imbalance. ams. 2018;12(10):449–60.
14. Kumar P, Bhatnagar R, Gaur K, Bhatnagar A. Classification of Imbalanced Data:Review of Methods and Applications. IOP Conf Ser: Mater Sci Eng. 2021 Mar 1;1099(1):012077.
15. Kaope C, Pristyanto Y. The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance. matrik. 2023 Mar 1;22(2):227–38.
16. Khairi A, Ghozali AF, Hidayah AD. Implementasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Masyarakat Pra Sejahtera Desa Sapikerep Kecamatan Sukapura`. TRILOGI: Jurnal Imu Teknologi, Kesehatan, dan Humaniora. 2021 Dec 31;2(3):319–23.
17. Rivki M, Bachtiar AM. Implementasi Algoritma K-Nearest Neighbor dalam Pengklasifikasi Follower Twitter yang menggunakan Bahasa Indonesia. Jurnal Sistem Informasi. 2017 May 1;13(1):31.
18. Gorunescu F. Data mining: concepts, models and techniques. Berlin, Heilelberg: Springer-Verlag; 2011. 357 p. (Intelligent systems reference library).
19. Dinov ID. Data Science and Predictive Analytics - Biomedical and Health Application using R. Michigan, USA: Springer; 2018. 832 p.
20. Gunn S. Support Vector Machines for Classification and Regression. University of Southampton: University of Southampton; 1998 May p. 52.
21. Amelia OD, Soleh AM, Rahardiantoro S. Pemodelan Support Vector Machine Data Tidak Seimbang Keberhasilan Studi Mahasiswa Magister IPB. XPlore. 2018;2(1):33–40.
22. Azis H, Purnawansyah P, Fattah F, Putri IP. Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung. Ilk J Ilm. 2020 Aug 27;12(2):81–6.
23. Azies HA, Trishnanti D, Mustikawati EPH. Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI). IPTEK Journal of Proceedings. 2019;6:53–7.
24. Yang Z, Jin M, Zhang Z, Lu J, Hao K. Classification Based on Feature Extraction For Hepatocellular Carcinoma Diagnosis Using High-throughput Dna Methylation Sequencing Data. Procedia Computer Science. 2017;107:412–7.
25. Sumule DrAI, Moudow DrJ, Bauw DrL, Sidik DrM, Nurkholis, Chalil DrTM. Kajian Peningkatan Pelayanan Pendidikan Berkualitas pada Tingkat Pendidikan Dasar dan Menengah di Provinsi Papua dan Papua Barat. Kolaborasi Masyarakat dan Pelayanan untuk Kesejahteraan (KOMPAK); 2022. 109 p.