Volume. 7 No 2 December (2024)

e-ISSN: **2615-3270**
p-ISSN: **2615-3599**

Eigen Mathematics Journal

Homepage jurnal: http://eigen.unram.ac.id/ index.php/eigen

Optimization of Classification Algorithms Performance with k-Fold Cross Validation

Moch. Anjas Aprihartha[a], Idham[b]

a. Program Studi PJJ Informatika, Universitas Dian Nuswantoro, Indonesia. Email: anjas.aprihartha@dsn.dinus.ac.id
b. Program Studi Sistem dan Teknologi Informasi, Universitas Muhammadiyah Mataram, Indonesia. Email: idham@ummat.ac.id

A B S T R A C T

Supervised learning is a predictive method used to make predictions or classifications. Supervised learning algorithms work by building a model using training data that includes both independent and dependent variables. Several methods for building classification include Logistic Regression, Naive Bayes, K-Nearest Neighbor (KNN), decision tree, etc. The lack of capacity of a classification algorithm to generalize certain data can be associated with the problem of overfitting or underfitting. K-fold cross-validation is a method that can help avoid overfitting or underfitting and produce a algorithm with good performance on new data. This study will test the Naive Bayes, K-Nearest Neighbor (KNN), Classification and Regression Tree (CART), and Logistic Regression methods with k-fold cross-validation on two different datasets. The values of k set for cross-validation are 2, 3, 5, 7, and 10. The analysis results concluded that each classification algorithm performed best at 10-fold cross-validation. In DATA 1, the Naive Bayes algorithm has the highest average accuracy of 0.67 (67%) and the error rate is 0.33 (33%), followed by the CART algorithm, KNN, and finally logistic regression. While DATA 2, the KNN algorithm has the highest average accuracy of 0.66 (66%) and an error rate of 0.34 (34%), followed by the CART algorithm, Naive Bayes, and finally logistic regressionbut can be a reference if you want to predict the growth direction of the accommodation and food service activities sector.

Keywords: CART, k-*fold cross validation*, KNN, Naive Bayes, Logistic Regression

Submitted: 14-05-2024; Accepted: 07-08-2024; Doi: https://doi.org/10.29303/emj.v7i2.212

1. Introduction

Data mining is the process of extracting valuable patterns, information, and knowledge from large data sets. Data mining plays an important role in discovering and minimizing risks in various aspects of life. In finding these patterns, it is necessary to use a data mining process using certain tools and techniques to analyze data from large data sources (Aprihartha *et al*., 2024). Based on the objectives, data mining is divided into two groups, namely descriptive methods and predictive methods (Gorunescu, 2011). The descriptive method is a method that describes patterns in data so that it is easy for researchers to understand, while the predictive method is a method that uses existing variables to predict future values.

Supervised learning is a predictive method used to make predictions or classifications. Supervised learning algorithms work by building models using training data that includes both independent and dependent variables. The most frequently used supervised learning algorithm that predicts output results in the form of numerical data is regression analysis (Han *et al*., 2022). Regression also includes identifying trend distributions based on available data. If the problem is related to classification, the output result will be data labeled categorical. Several methods for building classification models such as Logistic Regression, Naive Bayes, support vector machine, k-Nearest Neighbor (KNN), decision trees, etc.

Supervised learning algorithms attempt to estimate unknown mapping functions by given independent and dependent variables. The lack of capacity of a classification model in terms of generalization on certain data can be attributed to overfitting or underfitting problems. A model that experiences overfitting is manifested by a model that can learn the training data well but lacks the ability to detect different results in the testing data. On the other hand, underfitting is a condition where the model cannot learn the training data or testing data well.

* Corresponding author.
e-mail: anjas.aprihartha@dsn.dinus.ac.id

Data testing and cross validation approaches can help avoid overfitting and produce models that perform well on new data (von Neumann, 2016). K-fold cross validation is a method for assessing and testing the effectiveness of machine learning models (Anandan and Manikandan, 2023). This technique is often used in machine learning-based applications. This helps in comparing and selecting suitable models for a particular predictive analysis. K-fold cross validation is easy to use to calculate the relative efficiency of various models and has been applied in various case studies. For example, 10-fold cross validation was used to analyze the performance of the KNN, Naive Bayes, and support vector machine (SVM) algorithms in a study of Indonesian society's response to the Covid-19 pandemic on Twitter (Pamungkas, 2021). In the research of Firmansyach *et al*. (2022) adopted the k-fold cross validation technique to handle overfitting and underfitting using the Naive Bayes algorithm and decision tree.

In Ramaulidyah and Goejantoro's (2021) research which applied the Naive Bayes and KNN algorithms in classifying value added tax payment status, the researchers divided the data into 80% training data and 20% testing data. Researchers used the randomization technique five times so that the model with the lowest APER was selected, namely 17.07% for the Naive Bayes model and 19.51% for k-nearest neighbor. This shows that the Naive Bayes method provides better classification prediction accuracy. Research by Saputra (2022) applied the Classification and Regression Tree (CART) method in classifying patients suffering from dengue fever (DHF). Researchers used 230 samples with training data divided into 143 observations and testing data into 87 observations. The test results produced an accuracy of 56.32%.

This research will use data obtained from Ramaulidyah and Goejantoro (2021) and Saputra (2022) then combine the Naive Bayes method, k-Nearest Neighbor (KNN), Classification and Regression Tree (CART), and Logistic Regression with k-fold cross validation. The k values set for cross validation are 2, 3, 5, 7, and 10 which will provide variations in the performance and generalization of the model. This aims to determine the consistency of the model's capabilities in data classification from various k folds. In carrying out this test, the performance of the four classification methods with various k-fold configurations will be evaluated and compared, to understand which model can improve performance estimates to be more stable and accurate. The availability of information in the form of the necessary data is very helpful in this research.

2. Theoretical Foundation

3.1. Naive Bayes

The Naive Bayes (NB) algorithm assumes that all independent variables are independent of each other (Le *et al.*, 2022). Given a sample set, $X = \{x_1, x_2, \cdots, x_m\}$ which has m independent variables. Based on these variables, the sample can be partitioned into different categories, $Y = \{c_1, c_2, \cdots, c_N\}$. The goal of the Bayesian classifier is to maximize the posterior probability, so it can be stated as follows.

$$h^*(x) = \underset{c\epsilon Y}{\operatorname{argmax}} P\left(\frac{c}{x}\right) \tag{1}$$

According to Bayes' Theorem Equation (1) can be rewritten as follows.

$$h^*(x) = \underset{c\epsilon Y}{\operatorname{argmax}} \frac{P(x,c)}{P(x)} = \underset{c\epsilon Y}{\operatorname{argmax}} \frac{P(c)P(x/c)}{P(x)} \tag{2}$$

Because $P(x)$ is not relevant for all categories, equation (2) can be written as follows.

$$h^*(x) = \underset{c\epsilon Y}{\operatorname{argmax}} P(c)P\left(\frac{x}{c}\right) \tag{3}$$

When all independent variables are independent of each other, the Naive Bayes classification model equation is shown as follows.

$$h^*(x) = \underset{c\epsilon Y}{\operatorname{argmax}} P(c) \prod_{i=1}^{m} P\left(\frac{x_i}{c}\right) \tag{4}$$

3.2. K-Nearest Neighbour

Algorithm k-Nearest Neighbor (KNN) is one of the well-known classifiers applied in various research fields (Lin, 2024). The KNN method predicts categories by utilizing the distance relationship between the closest neighbors (Aprihartha *et al*., 2024). In addition, the distance metric chosen in the data experiment has an effect on the prediction results. Applied the KNN algorithm to identify the closest observation target using the Euclidean distance which is expressed as follows (Cholil *et al.*, 2015).

$$d(i,j) = \sqrt{\sum_{f\epsilon F} \left(v_{if} - v_{jf}\right)^2} \tag{5}$$

where $d(i,j)$ is the distance between the i-th and j-th observations, F is a set of candidate variables for classifying observations, v_{if} is the normalized value of the f-th variable and i-th observation before calculating the euclidean distance, $i = 1,2,\cdots,N$.

3.3. Classification and Regression Tree

Classification and Regression Tree (CART) is a classification or regression algorithm that produces a model in the form of a decision tree (Aprihartha *et al.*, 2024). CART classification is carried out in three stages, namely creating a classification tree with a formation procedure using recursive node separation, pruning the tree to obtain a simpler tree, and determining the optimal classification tree (Kuswanto and Mubarok, 2019). In forming the model, CART uses the Gini

index criterion because it can be expanded to include symmetric losses *(symmetrized cost)* (Aprihartha, 2024). Suppose D is a data set containing n samples and m independent variables with each sample represented as a vector $(x_1, x_2, x_3, \cdots, x_m)$ and one of the classes of the dependent variable is c. Gini index on variables j-th defined as follows.

$$Gini(j) = 1 - \sum (P(c_k|x_j))^2 \quad (6)$$

with $P(c_k|x_j)$ is the class probability of c_k on the x_j independent variable. The probability $P(c_k|x_j)$ can be calculated through the number of samples by class c_k divided by the total number of samples for each split. The Gini index is calculated for each independent variable and ranks them based on their importance.

3.4. Logistic Regression

The most frequently used statistical method for binary classification is Logistic Regression (RL) (Prasetya *et al*., 2024). This menthod is a derivative of linear regression involving a dependent variable labeled categorical. In his research, Logistic Regression is used to estimate the probability of a binary dependent outcome based on a set of independent variables X_s. The logit function is used by combining the probability p of the event of interest $P(Y = 1)$ with a linear combination of independent variables (Balboa *et al*., 2024). Logistic Regression uses odds ratios to represent the probability of an event occurring (p) by the probability of an event not occurring $(1 - p)$. The coefficients b_0 and b_i with $i = 1, 2, \cdots, n$ are the estimated parameters for the intercept and independent variable X_s. The odds ratio can be converted into a probability p using the exponential function.

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (7)$$

$$p = \frac{\exp\,(b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n)}{1 + \exp\,(b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n)} \quad (8)$$

3.5. K-Fold Cross Validation

K-fold cross validation is a technique used to optimize the performance of classification algorithms by partitioning the dataset into k folds with $(k - 1)$ as training data and the rest as testing data. Then analysis is carried out using a supervised learning algorithm for k iterations. The final stage of this process is to calculate the average of the results from all iterations to obtain a more stable model evaluation compared to just one data distribution. Figure 1 illustrates this process.

Determining the number of folds will affect the performance of the classification model. The most common k values are 5 and 10. The model performance becomes sensitive to k if the specified k value is very small (Chacón *et al*., 2023; Aprihartha, 2024). However, it takes a long time to compute when faced with k values that are too large.

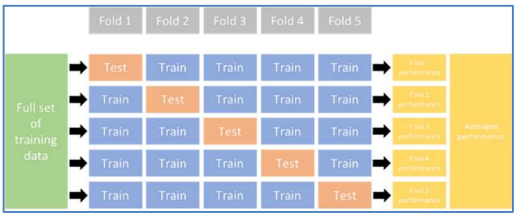

Figure 1. 5-*fold cross validation*

3.6. Model Performance Test

Confusion matrix is an effective tool in assessing how well the algorithm used is in identifying patterns from various classes. The confusion matrix consists of a set of rows and columns that display the results of the classification model test in tabular form. The rows reflect the actual classes or observation classes, while the columns represent the predicted classes. Accuracy is one of the indicators most widely used in class classification and in its calculations, it takes data from the confusion matrix (Chacón *et al*., 2023). In equation (9), the elements classified correctly by the model are expressed in the numerator while all cases studied by the model are expressed in the denominator.

$$\text{Accuracy} = \frac{\text{a} + \text{d}}{\text{a} + \text{b} + \text{c} + \text{d}} \quad (9)$$

$$\text{Error} = 1 - \text{Accuracy} \quad (10)$$

where a is true positive, b is true negative, c is false positive, and d is false negative (Aprihartha *et al*., 2024).

Table 1. *Confusion Matrix*

		Prediction Class	
		Yes	**No**
Observation Class	Yes	*a*	*b*
	No	*c*	*d*

2. Method

The data used in this research is secondary data obtained from research by Ramaulidyah and Goejantoro (2021) and Saputra (2022). To make naming easier, the research data from Ramaulidyah and Goejantoro (2021) was named DATA 1, while the data from Saputra (2022) was named DATA 2. DATA 1 consists of 205 observations and DATA 2 consists of 230 observations. The research variables in each data are presented in Table 2 and Table 3.

Table 2. DATA 1

Variables	**Catagories**
Tax Payment Status (Y)	1) Be obedient 2) Disobedience
Income (X_1)	1) Less than 100 million 2) 100 million – 500 million 3) 500 million – 1 billion

Variables	Catagories
	4) 1 billion – 10 billion
	5) More than 10 billion
Company Form (X_2)	1) PT
	2) CV
	3) BUMD/BUMN
	4) Cooperative
	5) Others
Reporting Status (X_3)	1) On time
	2) Not on time

Tabel 3. DATA 2

Variables	Catagories
Dengue Sufferers (Y)	1) Treated patients
	2) Untreated patients
Gender (X_1)	1) Man
	2) Woman
Thrombocytes (X_2)	1) $< 100000/\mu l$
	2) $\geq 100000/\mu l$
Hematocrit (X_3)	1) Decrease
	2) Normal
Length of Treatment (X_4)	1) $1-5$ days
	2) $6-10$ days

3. Results and Discussion

3.1. Evaluation of KNN Algorithm Performance

The underlying principle of the KNN algorithm is to find the shortest distance between the observed data and its k nearest neighbors. The k nearest neighbor values in this study are 1, 3, 5, 7, and 9. Meanwhile, the number of cross validation folds used in each algorithm is 2, 3, 5, 7, and 10.

Tabel 4. *K-fold* CV Algorithm KNN on DATA 1

	***k*-fold CV**				
***k*-NN**	**2**	**3**	**5**	**7**	**10**
1	0.63	0.64	0.64	0.6	0.64
3	0.64	0.61	0.62	0.64	0.66
5	0.6	0.57	0.6	0.64	0.68
7	0.54	0.54	0.54	0.62	0.65
9	0.53	0.48	0.50	0.57	0.57
Mean	0.59	0.57	0.58	0.61	0.64

In Table 4, the greatest accuracy is 0.68 in the KNN algorithm test with parameter $k = 5$ and 10th fold cross validation. From these results, it can be seen that the accuracy of the k-NN algorithm with k=9 tends to be higher and stable at larger folds (7 and 10), with the highest accuracy of 57%. Then the highest average accuracy was achieved when using 10 cross-validation folds, namely 0.64 for different k parameters. Meanwhile, the lowest average accuracy was when using 3rd fold cross validation.

Tabel 5. *K-fold* CV Algorithm KNN on DATA 2

	***k*-fold CV**				
***k*-NN**	**2**	**3**	**5**	**7**	**10**
1	0.55	0.63	0.66	0.60	0.65
3	0.65	0.63	0.63	0.63	0.64
5	0.63	0.65	0.65	0.65	0.67
7	0.62	0.63	0.65	0.64	0.67
9	0.63	0.60	0.67	0.66	0.67
Mean	0.61	0.63	0.65	0.64	0.66

Table 5 shows the KNN algorithm test with a value of $k = \{5,7,9\}$ and using 10th fold cross validation gives the highest accuracy of 0.67. From these results, it can be seen that the accuracy of the k-NN algorithm with k=9 tends to be more stable and high in larger folds (5, 7, and 10), with the highest accuracy of 67% for folds 5 and 10. This suggests that algorithm performance tends to be better when the data is validated with a larger number of folds, possibly due to more varied and more representative training and test data. Then the highest average accuracy obtained from testing with different k values in 10-fold cross validation was 0.66, while the lowest average accuracy when using 2nd fold cross validation was 0.61.

3.2. Evaluation of Algorithms Performance *k-fold* CV on DATA 1

Table 6 shows the accuracy of the four algorithms, namely Naive Bayes (NB), k-nearest neighbors (KNN), Classification and Regression Tree (CART), and Logistic Regression (RL) which were tested using cross validation on different numbers of folds. Accuracy plots for each classification algorithms are presented in Figure 2.

Tabel 6. *k-fold* CV Algorithms on DATA 1

***k-fold* CV**	**NB**	**KNN**	**CART**	**RL**
2	0.62	0.59	0.61	0.16
3	0.64	0.57	0.65	0.16
5	0.65	0.58	0.65	0.15
7	0.66	0.61	0.62	0.17
10	0.67	0.64	0.65	0.17
Mean	0.65	0.60	0.63	0.16

The Naive Bayes algorithm shows consistent improvement with increasing the number of cross-validation folds. Accuracy increases from 0.62 on the 2nd fold to 0.67 on the 10th fold. The average accuracy of Naive Bayes is the highest compared to other algorithms, namely 0.65 (65%) or error is 0.35 (35%).

In the KNN algorithm, although there are fluctuations, the highest accuracy is achieved at the 10th fold, namely 0.64. Meanwhile, the lowest accuracy occurred in the 3rd fold at 0.57. The average KNN accuracy for each fold is 0.60 (60%) or error is 0.4 (40%).

The CART algorithm obtained fairly stable accuracy, with slight variations. The lowest accuracy is 0.61 on the 2nd fold and the highest accuracy is 0.65 on the 10th fold. The average CART accuracy for each fold is 0.63 (63%) or error is 0.37 (37%).

The Logistic Regression algorithm has relatively lower performance compared to the other three algorithms. The resulting accuracy tends to be stable in the range of 0.15-0.17. The average Logistic Regression accuracy for each fold is 0.16 (16%) or error is 0.84 (84%) .

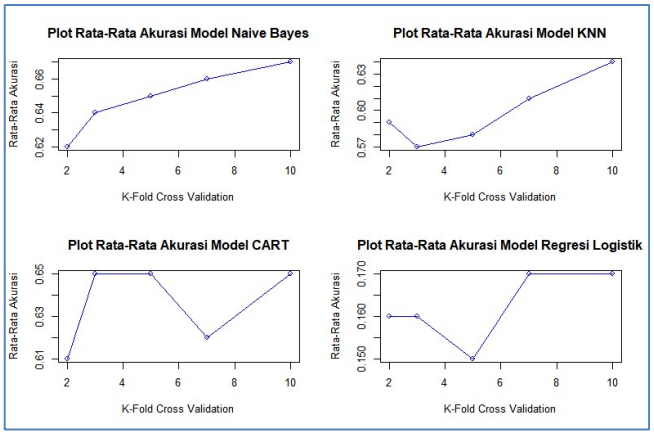

Figure 2. Plot of *k*-fold CV Classification Algorithms on DATA 1

In research conducted by Ramaulidyah and Goejantoro (2021), the error results obtained for the KNN and Naive Bayes algorithms were 19.51% and 17.07%, respectively. In other words, the resulting accuracy is 80.49% and 82.93% or 0.8049 and 0.8293, respectively. In this study, compared with the highest accuracy in the same algorithm, namely 0.67 and 0.64, there was a significant decrease. This is due to overfitting or underfitting in the algorithm applied. Therefore, k-fold cross validation is used to produce a algorithm with more stable and accurate performance.

3.3. Evaluation of Algorithms Performance *k-fold* CV on DATA 2

Table 7 shows that the Naive Bayes algorithm has stable algorithm performance without much variation. Accuracy tends to be consistent around 0.60 and there is a decrease when entering the 7th fold but increases again at the 10th fold. The average accuracy of Naive Bayes is 0.60 (60%) or error is 0.40 (40%).

Tabel 7. *K-fold* CV Algorithms on DATA 2

k-fold CV	NB	KNN	CART	RL
2	0.60	0.62	0.62	0.60
3	0.60	0.63	0.61	0.60
5	0.60	0.65	0.63	0.59
7	0.59	0.64	0.63	0.58
10	0.61	0.66	0.64	0.60
Mean	0.60	0.64	0.63	0.59

Figure 3 shows a significant increase in accuracy with increasing the number of cross-validation folds. Accuracy increases from the 2nd fold, namely 0.62 to 0.66 in the 10th fold. The average accuracy of KNN is the highest compared to other algorithms, namely 0.64 (64%) or error is 0.36 (36%).

The CART algorithm provides consistent performance with small improvements in accuracy. Accuracy improved slightly from 0.62 on the 2nd fold to 0.64 on the 10th fold. The average CART accuracy is 0.63 (63%) or error is 0.37 (37%).

The Logistic Regression algorithm has relatively lower and more stable algorithm performance compared to other algorithms. Accuracy ranges from 0.58 to 0.60. There was a decrease in the 7th fold then an increase in the 10th fold. The average accuracy of Logistic Regression is 0.59 (59%) or error is 0.41 (41%).

Research by Saputra (2022) applied the Classification and Regression Tree (CART) method to produce an accuracy of 56.32% or 0.5632. In this study, compared to the same algorithm with the highest accuracy, namely 0.64, there was an increase of 0.0768. This shows the CART algorithm with 10-fold cross validation provides better model performance.

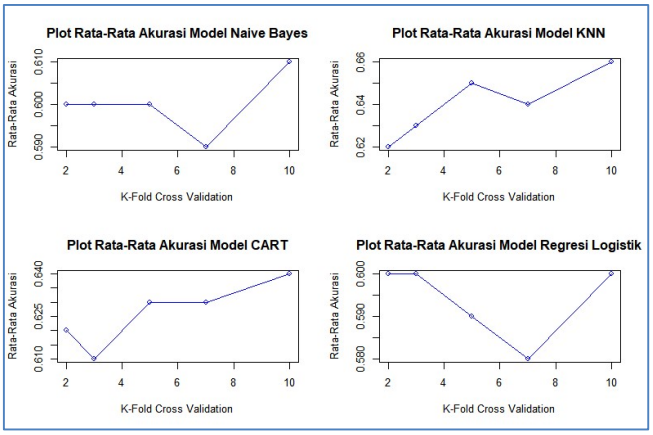

Figure 3. Plot of *k*-fold CV Classification Algorithms on DATA 2

4. Conclusions

Based on the results of data analysis and discussion in the application of the Naive Bayes algorithm, KNN, CART, and Logistic Regression, it can be concluded that each classification algorithm provides the best performance in 10th fold cross validation. The KNN test results on DATA 1 provide the highest accuracy of 0.68 with $k = 5$ in the 10th fold cross validation. In DATA 2, testing the KNN algorithm with a value of $k = \{5,7,9\}$ and using 10th fold cross validation gives the highest accuracy of 0.67. In DATA 1, the Naive Bayes algorithm has the highest average accuracy of 0.67 (67%) and the error rate is 0.33 (33%), followed by the CART algorithm, KNN, and finally logistic regression. While DATA 2, the KNN algorithm has the highest average accuracy of 0.66 (66%) and an error rate of 0.34 (34%), followed by the CART algorithm, Naive Bayes, and finally Logistic Regression.

REFERENCES

Anandan, B., & Manikandan, M. (2023). Machine learning approach with various regression models for predicting the ultimate tensile strength of the friction stir welded AA 2050-T8 joints by the K-Fold cross-validation method. *Materials Today Communications*, *34*, 105286. doi : https://doi.org/10.1016/j.mtcomm.2022.105286

Aprihartha, M. A. (2024). Implementasi CART-Real Adaboost dalam Memprediksi Minat Pelanggan Membeli Sepatu. *Jurnal EurekaMatika*, *12*(1), 35-46. doi: https://doi.org/10.17509/jem.v12i1.67808

Aprihartha, A. (2024). Penyelesaian Masalah Ketidakseimbangan Data Melalui Teknik Oversampling

dan Undersampling pada Klasifikasi Siswa Tidak Naik Kelas.*Jurnal Teknik Ibnu Sina (JT-IBSI)*, *9*(01), 43-52. doi: https://doi.org/10.36352/jt-ibsi.v9i01.807

Aprihartha, M. A., Alam, T. N., & Husniyadi, M. (2024). Perbandingan Metrik Euclidean dan Metrik Manhattan untuk K-Nearest Neighbors dalam Klasifikasi Kismis. *Jurnal Ilmu Komputer dan Informatika*, *4*(1), 21-30.

Aprihartha, M. A., Astutik, F., & Sulistianingsih, N. (2024). Comparison of Naïve Bayes, CART, dan CART Adaboost Methods in Predicting Tire Product Sales. *Jurnal Matematika, Statistika dan Komputasi*, *20*(3), 596-605. doi: https://doi.org/10.20956/j.v20i3.33187

Aprihartha, M. A., Putrawan, Z., Zulhan, D., & Nurfaizal, F. A. (2024). Algoritma Synthetic Minority Oversampling Technique dan C5. 0 dalam Mengatasi Ketidakseimbangan Data pada Klasifikasi Kelulusan Siswa. *UPGRADE: Jurnal Pendidikan Teknologi Informasi*, *2*(1), 1-10. doi: https://doi.org/10.30812/upgrade.v2i1.4148

Aprihartha, A., Putrawan, Z., Zulhan, D., & Nurfaizal, F. A. (2024). Klasifikasi Produktivitas Buah Nanas Menggunakan Algoritma Classification and Regression Tree (CART). *Diophantine Journal of Mathematics and Its Applications*, 64-70. doi: https://doi.org/10.33369/diophantine.v3i1.34193

Balboa, A., Cuesta, A., González-Villa, J., Ortiz, G., & Alvear, D. (2024). Logistic Regression vs machine learning to predict evacuation decisions in fire alarm situations. *Safety science*, *174*, 106485. doi: https://doi.org/10.1016/j.ssci.2024.106485

Chacón, A. M. P., Ramírez, I. S., & Márquez, F. P. G. (2023). K-Nearest Neighbor and K-fold cross-validation used in wind turbines for false alarm detection. *Sustainable Futures*, *6*, 100132. doi: https://doi.org/10.1016/j.sftr.2023.100132

Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi algoritma klasifikasi k-Nearest Neighbor (knn) untuk klasifikasi seleksi penerima beasiswa. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *6*(2), 118-127. doi: https://doi.org/10.31294/ijcit.v6i2.10438

Firmansyach, W. A., Hayati, U., & Wijaya, Y. A. (2023). Analisa Terjadinya Overfitting Dan Underfitting Pada Algoritma Naive Bayes Dan Decision Tree Dengan Teknik Cross Validation. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 262-269. doi: https://doi.org/10.36040/jati.v7i1.6329

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media. doi: https://doi.org/10.1007/978-3-642-19721-5

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

Kuswanto, H., & Mubarok, R. (2019). Classification of cancer drug compounds for radiation protection optimization using CART. *Procedia Computer Science*, *161*, 458-465. doi: https://doi.org/10.1016/j.procs.2019.11.145

Li, L., Zhou, Z., Bai, N., Wang, T., Xue, K. H., Sun, H., .& Miao, X. (2022). Naive Bayes classifier based on memristor nonlinear conductance. *Microelectronics Journal*, *129*, 105574. doi: https://doi.org/10.1016/j.mejo.2022.105574

Lin, K. Y. C. (2024). Optimizing variable selection and neighbourhood size in the K-Nearest Neighbor algorithm. *Computers & Industrial Engineering*, 110142. doi: https://doi.org/10.1016/j.cie.2024.110142

Pamungkas, F. S., & Kharisudin, I. (2021, February). Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter. In *PRISMA, Prosiding Seminar Nasional Matematika* (Vol. 4, pp. 628-634).

Prasetya, J., Fallo, S. I., & Aprihartha, M. A. (2024). Stacking Machine Learning Model for Predict Hotel Booking Cancellations. *Jurnal Matematika, Statistika dan Komputasi*, *20*(3), 525-537. doi: https://doi.org/10.20956/j.v20i3.32619

Rahmaulidyah, F. N., Hayati, M. N., & Goejantoro, R. (2021). Perbandingan Metode Klasifikasi Naive Bayes Dan K-Nearest Neighbor Pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Samarinda Ulu. *Eksponensial*, *12*(2), 161-164. doi: https://doi.org/10.30872/eksponensial.v12i2.809

Saputra, N. D. (2021). Penggunaan metode Classification and Regression Tree (CART) dalam mengklasifikasikan pasien penderita DBD di Rumah Sakit Anwar Makkatutu Kabupaten Bantaeng [Skripsi, Universitas Islam Negeri (UIN) Alauddin Makassar].

von Neumann, J. (2016). Model selection and overfitting. *Nat. Methods*, *13*, 703-704.