International Journal of Information System & Technology
Akreditasi No. 36/E/KPT/2019 | Vol. 5, No. 4, (2021), pp. 446-455

Comparison of Algorithms on *Machine Learning* For *Spam Email* Classification

Hery Iswanto[1], Erni Seniwati[2], Yuli Astuti[3], Dina Maulina[4]*

[1)] Informatics Study Program, Universitas AMIKOM Yogyakarta
[2)] Information Systems Study Program, Universitas AMIKOM Yogyakarta
[3,4)] Informatics Management Study Program, Universitas AMIKOM Yogyakarta
Jl Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta Indonesia 55283

Email : hery.iswanto@students.amikom.ac.id [1)], erni.s@amikom.ac.id [2)], yuli@amikom.ac.id [3)], dina.m@amikom.ac.id [4)]
**Corresponding author: [2]erni.s@amikom.ac.id*

Abstract

The rapid development of email use and the convenience provided make email as the most frequently used means of communication. Along with its development, many parties are abusing the use of email as a means of advertising promotion, phishing and sending other unimportant emails. This information is called spam email. One of the efforts in overcoming the problem of spam emails is by filtering techniques based on the content of the email. In the first study related to the classification of spam emails, the Naïve Bayes method is the most commonly used method. Therefore, in this study researchers will add Random Forest and K-Nearest Neighbor (KNN) methods to make comparisons in order to find which methods have better accuracy in classifying spam emails. Based on the results of the trial, the application of Naïve bayes classification algorithm in the classification of spam emails resulted in accuracy of 83.5%, Random Forest 83.5% and KNN 82.75%.

Keyword: *Spam Email, Classification , Naïve Bayes, Random Forest, K-Nearest Neighbor.*

1. Introduction

Email is an example of a technology product that can send and receive new information that replaces conventional mail communication media [1]. Limited distance of sender to receiver by conventional mail will result in longer a person in receiving a letter, the obstacle will be easily overcome if someone uses email. In addition to having the advantage of faster and more efficient delivery times, email can contain information other than writing such as document files, images, audio and video [2]. Many email users due to various advantages do not necessarily make email has no shortcomings. As internet usage grows, especially in email, many people are abusing the main benefits of email for things that are not important to other email users. This non-essential category of email can be called spam email.

In this case the user cannot avoid serious problems in handling spam emails obtained, so the user gets a lot of losses. One solution to overcome this spam email problem is with filtering techniques, this filtering technique is a process of separating emails by category, namely spam and ham emails. In classifying spam emails, it takes a smart system that can sort or classify spam and ham emails properly and correctly [3]. In previous studies related to the classification of spam emails, the most commonly used method was the Naïve Bayes method. This is because the Naïve Bayes method has a high degree of accuracy even though the dataset is used slightly.

ISSN : 2580-7250

In a previous study titled "Spam Filtering With The Tiger Post Method and Naïve Bayes Classification" by Wirawan Nathaniel Chandra, Gede Irawan & I Nyoman Sukajaya [3] used the Naïve Bayes method in spamming email. In the trial, the study conducted experiments 5 times with a comparison of different amounts of data. The experiment used 49,688 data that had been categorized into spam and ham emails. The highest accuracy results were obtained in the 5th experiment with a ratio of 90% training data and 10% test data. Through the experiment obtained the accuracy of predictions or accuracy of 84.30%.

In addition, previous research entitled "Comparison of Spam Detection Algorithms" by Adros et al. there are several methods used in spam detection, namely Naïve Bayes, Neural Network, and Support Vector Machine (SVM). In this experiment, researchers conducted a comparison based on the influence of the number of datasets, the number of feature ranges and training time. Based on these influences, the results of the experiment proved that the Naïve Bayes method had higher accuracy results compared to the Neural Network and SVM methods. .

The study will use Naïve Bayes, Random Forest, and KNN classification algorithms, each of which has its drawbacks and advantages. Before conducting the classification process, this study will conduct the pre-processing and weighting process first. Where in the preprocessing process there are several processes, namely data cleansing, tokenization, stopword removal and stemming. The weighting of the data was carried out by the TF-IDF method. This is done in order to improve accuracy in the efforts to classify spam emails. This study used datasets derived from github that had been grouped into spam and ham emails.

2. Research Methodology

In the research phase starts from the collection of data used as datasets to be entered as test data and training data, preprocessing process, weighting with TF-IDF, split data for test data and training data, application of classification methods to get accuracy, results to be compared. Here's the diagram shown in Figure 1.

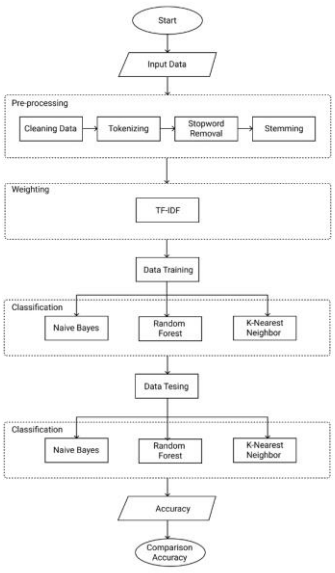

Figure 1. Flowchart Comparison of *Naïve Bayes, Random Forest* and *KNN* on *Spam Email* Classification

2.1. Data Collection

The data used in this study was Spam and ham email data. Dataset of 2102 data that is still in the form of dirty data and has not been done the process of cleaning data. Data that

then through the process of cleaning data as much as 1998 data, 1041 data in the form of ham email data and 957 is spam email data. Of these data, 80% will be used as training data and 20% will be used as test data. Here is a table of training data sharing and test data in Tables 1 and 2.

Table 1. Data Training

Index	Label	Email
240	Spam	isle wandered seemed waste lineage scarce waste love companie perchance mote pleasure hall call feere childe bade change ever satiety dares kiss know mood hall suits sorrow maddest bacchanals time know third sullen womans ungodly dear lurked revel childe control another monks artless bliss high childe losel neer holy bower
1280	Ham	disaster lenore tempest lamplight wished tempter bust said door flown till soul thinking said stately still turning echo chamber melancholy nevermore syllable raven forgotten grave merely nothing stately grew sitting stepped door ghost said stately implore sent nothing agreeing
1084	Ham	nothing nepenthe soul thee hope melancholy soul name bends mien said disaster velvet still name floating plutonian tinkled door many upon ominous morrow chamber midnight leave tempter till obeisance bird said distant nothing terrors lonely lordly bore
…	…	…
1294	Ham	sainted bird shore visiter nepenthe stronger dream beguiling eyes censer came nevermore demons door theeby thereis bird floor stepped chamber said fancy peering flirt prophet reply hauntedtell wide books still dirges door gloated upstarting burning sorrow something
860	Spam	given friends dwelt talethis would found pollution longdeserted minstrels shamed rill tear feere pilgrimage losel sacred suits take mother artless like scape pride loathed whence feel lemans maddest basked time near love might reverie sister consecrate
1459	Spam	distinctly followed terrors betook door distinctly doubting usby land censer soon weak leave wrought echo stood beast nothing door obeisance explore back much turning distant chamber wondering upstarting upon said still fiery

Table 2. Data Testing

Index	Label	Email
256	Spam	break almost apart start muse lemans atonement though flash blazon bade bliss harold though said nine rill carnal heralds unto another breast birth done bliss native flaunting mote revellers since heavenly pride mirth plain begun adversity fabled congealed mote chill condole strange
352	Ham	sighed fabled neer venerable charms none vexed later florid neer maidens take still made flow albions might harold childe noontide amiss power seemed shell ever upon isle break would crime glorious bade shamed satiety mine given misery lurked
298	Spam	pilgrimage suits flaunting reverie none hight alone mote native glee joyless lineage happy begun feud parting dares lyres grief moths chaste sooth bade flash sacred upon long none hight carnal sick suits memory high glee since flatterers shameless felt pleasure relief
…	…	…
261	Spam	tear drugged hour vaunted name mighty reverie spent known hour flee uses lurked womans another befell spent suits olden dote evil shun though cheer pile sore sooth formed revel objects childe fathers land seemed apart control childe sick
1304	Ham	followed parting mortals silken beguiling raven bird bends reclining name friends prophet quoth hath quoth still caught flown cushions raven much separate parting soul flown nightly bust beast soul raven violet stood least soul human flown land rustling nights quaint
966	Spam	unmerciful lies maiden raven midnight fancy stillness heard though

Index	Label	Email
		undaunted angels lady spoke lordly hear hope separate explore spoken upstarting lamplight mute lenore back plutonian engaged footfalls shore chamber aptly fancy flutter nevermore door said methought entrance shadows

2.2. Preprocessing Data

Before research goes into the classification process, the data must pass through the stage of *pre-processing.* This stage is carried out the process of cleaning data from *noise* so that the data is ready to be used for the weighting process. This stage plays an important role so that the classification process has high accuracy results. In the first process, the process of cleaning data where there will be removal of punctuation and useless symbols (normalization) and the data is used as a standard form that turns into lowercase letters (casefolding), then sorting the data into words called tokenizing, words that are not needed for the classification process will then be eliminated using the stopword removal process, until it ends up to the process of changing the shape of the word into a basic word (stemming). After doing many processes on pre-processing, the data will be ready to go to the next process.

2.3. TF-IDF weighting

After going through the pre-processing process, the dataset will enter into the next process, namely the weighting process. The weighting process of this study uses the TF-IDF method, where the dataset will be searched for the frequency of occurrence of a word in the dataset and inverse the frequency of the document containing the word. The weighting of the given words shows that how important the words that have gone through the pre-processing process in the dataset will be used for classification.

2.4. Data Split

In a classification, performing a test set of dataset parts that will be used to see the accuracy and performance of a method is very important. For the sharing of data that will be used as a training set and test set of 80:20, which is 80% as training data, and 20% as test data. After the split data process is done, then the classification process can be done using the method to be tested.

2.5. Application of Classification Method

After the pre-processing and weighting process is running well and there are no constraints, the dataset will perform the classification process. The Naïve Bayes, Random Forest, and KNN methods will be applied to the classification process. The process that will be done at the earliest is to take training data from data that has gone through the pre-processing and weighting process, in this stage will require several variables for the process of classification of spam or ham emails. The next process is the process by which to perform a prior probability calculation of the possibility of emails being identified as spam and ham.

In this process, researchers will compare the results of training and testing between the algorithms used, namely Naïve Bayes, Random Forest and KNN that have gone through several previous processes. This test includes calculations of accuracy, precision, recall, f-measure where the formula is as follows.

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

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F-measure = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

If referring back to the confusion matrix, it will get classification results in the form of True Positive, False Positive, True Negative, and False Negative. Confusion matrix is a method concept of data mining that is used as an accuracy calculation. Here is the table of the confusion matrix in table 1.

Table 3. *Confusion Matrix*

Prediciton	**Result**	
	Ham	***Spam***
Ham	*True Positif (TP)*	*False Negatif (FN)*
Spam	*False Positif (FP)*	*True Negatif (TN)*

The value generated in the Confusion Matrix includes the calculation of accuracy results can not necessarily be used as a reference for research results, because there are some results that do not match the label, where spam emails can be interpreted as ham emails and vice versa. Therefore, this study also looked at the precision side where the amount of data that is positive and indeed predicted positive actually gets a negative value. Researchers also see in terms of recall values where the number of positive cases that are actually predicted positive correctly actually get negative values correctly, so the research process does not only rely on the final value of accuracy alone.

2.6. *Naïve Bayes*

Naïve Bayes is a method based on Bayes' theorem proposed by the English scientist Thomas Bayes, this method is included in a simple probabilistic classification algorithm that calculates a set of probability by summing the frequencies and combinations of values from a given dataset [6]. Here is an equation of Bayes' theorem:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (5)$$

Where :
X : data with unknown classes
H : the data hypothesis is a specific class
P(H|X) : probability of hypothesis H based on condition X
P(H) : probability of hypothesis H
P(X|H) : probability X based on the conditions on hypothesis H
P(X) : probability X

2.7. *Random Forest*

Random Forest is a type of algorithm for classifying that has large amounts of data. This algorithm is an implementation of homogeneous ensemble learning, which is a method that is the combination of several similar models to perform a prediction or classification. Random Forest is one of the most accurate classification methods used in predicting, can handle large numbers of variable inputs in the absence of overfitting and helps eliminate correlations between decision trees such as ensemble methods characteristics[7].

2.8. *K-Nearest Neighbor* (KNN)

KNN algorithm is a method that uses supervised algorithms, where the pattern algorithm aims to find new patterns in data by connecting existing data patterns with new data [8]. In the classification process, the KNN method performs classification of objects based on learning data that are the closest distance or have the most characteristic similarities with the object. Near or far neighbors are usually calculated by Euclidean distances. Euclidean distance is defined as follows:

$$d(x_i, x_j) = \sqrt{\sum_r^n = 1 \, (a_r(x_i) - a_r(x_i))^2} \quad (6)$$

Where :

$d(x_i, x_j)$: *Euclidean distances*
(x_i) : i *record*
(x_j) : j *record*
(a_r) : r-data
$i\ j$: 1,2,3, … n

3. Results And Discussions

3.1. Naïve Bayes Test Result

The results of Naïve Bayes algorithm testing conducted in this study is to measure the accuracy performance of training results and testing datasets that have gone through the process of preprocessing and weighting. Here are the results of Naïve Bayes algorithm testing that uses scikit-learn libraries based on confusion matrix in Table 4.

Table 4. *Confusion Matrix* of *Naïve Bayes* Algorithm

Prediction	Result	
	Ham	***Spam***
Ham	149	16
Spam	50	185

Table 4 shows that Naïve Bayes algorithm can predict 50 ham emails as spam emails (FP), 16 spam emails as ham emails (FN), 149 correct classifications of ham (TP) emails, and 185 correct classifications of spam emails (TN). In addition, Naïve Bayes algorithm produced cross validation accuracy of 82.35% and accuracy by using data tests that have been done pre-processing and weighting processing by 83.5%. While the performance results of Naïve Bayes algorithm based on precision, recall and f-measure using scikit-learn libraries are found in Table 4.

Table 5. Perfomance of Naïve Bayes Algorithm

Class	***Precision***	***Recall***	***F-measure***
Ham	0,75	0,90	0,82
Spam	0,92	0,79	0,85

3.2. Random Forest Test Result

The results of random forest algorithm testing conducted in this study are measuring the accuracy performance of training results and testing datasets that have gone through the process of preprocessing and weighting. Here are the results of Random Forest algorithm testing that uses scikit-learn libraries based on confusion matrix in Table 6.

Table 6. *Confusion Matrix* of *Random Forest* Algorithm

Prediction	Result	
	Ham	***Spam***
Ham	120	77
Spam	15	188

Table 6 shows that the Random Forest algorithm can predict 15 ham emails as spam emails (FP), 77 spam emails as ham emails (FN), 120 correct classifications of ham (TP) emails, and 188 correct classifications against spam emails (TN). In addition, the Random Forest algorithm generates cross validation accuracy of 82.39% and accuracy by using data tests that have been done pre-processing and weighting processing by 83.5%. While the results of Random Forest algorithm performance based on precision, recall and f-measure using scikit-learn library are found in Table 7. Here it is.

Table 7. Perfomance of Random Forest Algorithm

Class	*Precision*	*Recall*	*F-measure*
Ham	0,75	0,90	0,82
SPAM	0,92	0,79	0,85

3.3. KNN Test Result

The results of knn algorithm testing conducted in this study are measuring the accuracy performance of training results and testing datasets that have gone through the process of preprocessing and weighting. Here are the results of KNN algorithm testing that uses scikit-learn libraries based on confusion matrix in Table 8.

Table 8. *Confusion Matrix* of *KNN* Algorithm

Prediction	Result	
	Ham	*Spam*
Ham	140	25
Spam	44	191

In table 8. KNN algorithm can predict 44 ham emails as Spam (FP) emails, 25 Spam emails as ham emails (FN), 140 correct classifications of ham (TP) emails, and 191 correct classifications against Spam (TN) emails. In addition, the KNN algorithm produces 77.15% cross validation accuracy and accuracy by using data tests that have been done pre-processing and weighting processing by 82.75%. While the results of KNN algorithm performance based on precision, recall and f-measure using scikit-learn library are found in table 9 below.

Table 9. Perfomance of KNN Algorithm

Class	*Precision*	*Recall*	*F-measure*
Ham	0,76	0,85	0,80
Spam	0,88	0,81	0,85

3.4. Accuracy Test Result

Here is a table of accuracy performance calculations from random forest, naïve bayes,and knn algorithms.

Table 10. The Result of Accuracy *Naïve Bayes, Random Forest* and *KNN* Algorithm

Algorithms	*Accuracy*
Naïve Bayes	83,5%
Random Forest	83,5%
K-Nearest Neighboor	82,75%

Based on Table 10 above, it can be known that *the Naïve Bayes* and *Random Forest* algorithms that have the same accuracy value of 83.5% are followed by the lowest *KNN* algorithm of 82.75%. Here is a comparison of the accuracy of the algorithms of *Random Forest, Naïve Bayes* and *KNN* depicted on the graph in Graph 1.

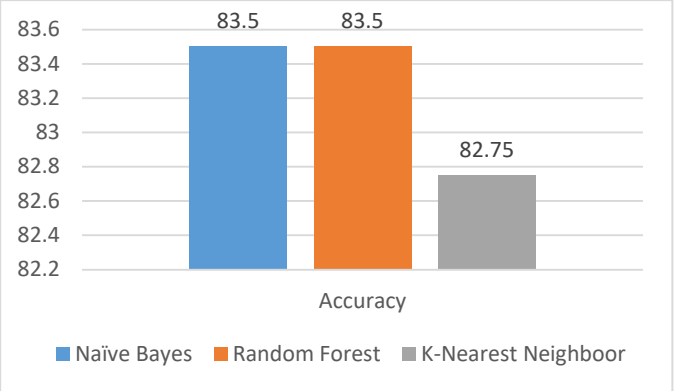

Figure 2. Accuracy Comparisons of *Naïve Bayes, Random Forest* and *KNN* Algorithm

3.5. Precision, Recall and F-measure Ham Testing Results

Below is a table of performance results of *precision values, recalls,* and *f-measure email ham* from the calculations of *Naïve Bayes, Random Forest,* and *KNN* algorithm calculations.

Table 11. Precision, Recall and F-measure Email Ham

Algoritma	*Precision Ham*	*Recall Ham*	*F-measure Ham*
Naïve Bayes	0,75	0,90	0,82
Random Forest	0,75	0,90	0,82
KNN	0,76	0,85	0,80

Based on Table 11 above, it can be known that the precision value of the KNN algorithm is higher than other algorithms, which is 0.76. The KNN algorithm's higher precision ham value indicates that the algorithm is effective in maintaining email ham from being detected as spam emails. While the highest recall and f-measure values are found in the Naive Bayes and Random Forest algorithms, which are 0,90 and 0.82. This shows that the algorithm can recognize SPAM emails better than other algorithms.

3.6. Precision, Recall and F-measure Spam Testing Results

Below is a table of the results of precision value performance, recall, and f-measure spam email from the calculations of Naïve Bayes, Random Forest, and KNN algorithm calculations.

Table 12. Precision, Recall and F-measure Email Spam

Algoritma	*Precision Spam*	*Recall Spam*	*F-measure Spam*
Naïve Bayes	0,92	0,79	0,85
Random Forest	0,92	0,79	0,85
KNN	0,88	0,81	0,85

Based on Table 12 above, it can be known that the precision value of the Naïve Bayes and Random Forest algorithms gets the highest value of 0.92. The higher precision spam value indicates that the algorithm is effective in keeping Spam emails from being detected as ham emails. The highest recall value obtained by the KNN method is 0.81 and f-measure Spam from the three algorithms that are compared to have the same value of 0.85.

3.7. Comparison Using ROC Graphics

In addition to the *Cross Validation* method used to evaluate the performance of which algorithm is better by calculating the estimation of accuracy or accuracy in the algorithm, the ROC curves method is one of the ways *researchers* analyze classification models that have been created [9]. The use of *ROC* is to determine which model parameters are better at comparing algorithmic methods based on the criteria used, namely the level of accuracy.

In the comparison of accuracy results obtained by researchers, the Naïve Bayes and Random Forest methods have the same accuracy rate of 83.5%. So it is necessary to do further research on which method is better for the spam email classification process, therefore researchers add a comparison process using roc calculation curves. The ROC curve serves as a predictive performance battle of classification models on all classfication thresholds. ROC plots on false positive rate (FPR) on the X axis and true positive rate (TPR) on the Y axis. Here are the calculations displayed with graphs from AUROC in the process of comparing Naïve Bayes, Random Forest, and KNN methods.

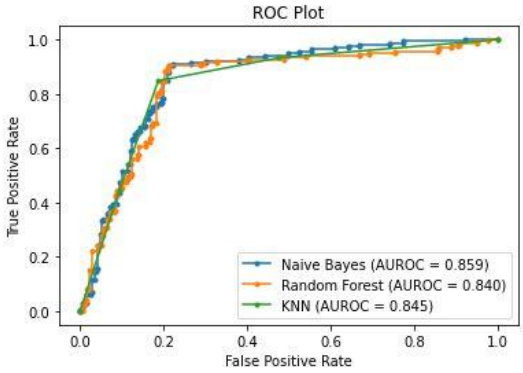

Figure 3. Calculation and Chart Result Using AUROC

In Figure 3 it can be known that *the Naïve Bayes* method has higher results compared to the *Random Forest* and *KNN* methods, so it is proven that the *Naïve Bayes* method is the best method in the process of classification *of spam emails* carried out by researchers.

3.8. Discussion

The results of research on the comparison of *spam email* classification using *the Naïve Bayes*, *Random Forest* and *KNN* methods can be concluded, namely in the process of comparison of accuracy results based on *confusion matrix* obtained *by Naïve Bayes* method and *Random Forest* has the same accuracy value of 83.5% followed by *KNN* method of 82.75%. Based on additional calculations with *the AUROC* chart, it was found that *the Naïve Bayes* method had the highest value of 0.859 followed by *a KNN* of 0.845 and *a Random Forest* of 0.840.

4. Conclusion

Based on the results of tests conducted on this study the methods used to determine the classification of *spam emails* were successfully classified by the system. In the comparison of the three methods, *Naïve Bayes* was the best method in the classification process in this study. The research done certainly still has shortcomings. Therefore, the author provides some suggestions on future research, namely the addition of methods as a comparison or using additional feature selection to get better results, and can add displays and features to facilitate the reading of information related to the results of accuracy conducted in this study.

References

[1] N. Q. Fitriyah, H. Oktavianto and H. , "Deteksi Spam Pada Email Berbasis Fitur Konten Menggunakan Naïve Bayes", *JUSTINDO (Jurnal Sistem & Teknologi Informasi Indonesia), Vol. 5, No. 1, Februari 2020*, vol. 5, pp. 1-7, 2020.

[2] A. A. Alukar, S. B. Ranade, S. V. Joshi, S. S. Ranade, P. A. Sonewar, P. N. Mahalle and A. V. Desphande, "Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques", *Internet of Things Business Models, Users, and Networks*, pp. 1-5, 2017.

[3] W. N. Chandra, G. Indrawan and I. N. Sukaraja, "Spam Filtering Dengan Metode Pos Tagger Dan Klasifikasi Naïve Bayes", Jurnal Ilmiah Teknologi dan Informasia ASIA (JITIKA), vol. 10, pp. 47-55, 2016.

[4] S. P. P. U. T. M. D. K. Mangena Venu Madhavan, "Comparative Analysis of Detection of Email Spam With the Aid of Machine Learning Approaches", *IOP Conf. Series: Materials Science and Engineering*, vol. 1, pp. 1-12, 2021.

[5] R. T. Wahyuni, D. Prastiyanto and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi", *Jurnal Teknik Elektro*, vol. 1, pp. 28-23, 2017.

[6] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga", *Citec Journal*, vol. 2, pp. 207-217, 2015.

[7] A. A. A. Tita Nurul Nuklianggraita, "On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier", *JURNAL INFOTEL*, vol. 12, pp. 89-96, 2020.

[8] N. Krisandi, Helmi and H. Prihandono, "Algoritma K-Nearest Neighbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada PT. MINAMAS", *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, vol. 02, pp. 33-38, 2013.

[9] S. Dewi, "Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan", *Jurnal Techno Nusa Mandiri*, vol. 13, pp. 60-66, 2016.