VOL 2 (2018) NO 3

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Illiteracy Classification Using K Means-Naïve Bayes Algorithm

Muhammad Firman Aji Saputra[#], Triyanna Widiyaningtyas[#], Aji Prasetya Wibawa[#]

[#] *Electrical Department, State University of Malang, Indonesia*
E-mail: firman.asbn@gmail.com, triyannaw.ft@um.ac.id, aji.prasetya.ft@um.ac.id

***Abstract*— Illiteracy is an inability to recognize characters, both in order to read and write. It is a significant problem for countries all around the world including Indonesia. In Indonesia, illiteracy rate is generally set as an indicator to see whether or not education in Indonesia is successful. If this problem is not going to be overcome, it will affect people's prosperity. One system that has been used to overcome this problem is prioritizing the treatment from areas with the highest illiteracy rate and followed by areas with lower illiteracy rate. The method is going to be a way easier to be applied if it is supported by classification process. Since the classification process needs a class, and there has not been any fine classification of illiteracy rate, there is needed a clustering process before classification process. This research is aimed to get optimal number of classes through clustering process and know the result of illiteracy classification process. The clustering process is conducted by using k means algorithm, and for the classification process is conducted by using Naïve Bayes algorithm. The testing method used to assess the success of classification process is 10-fold method. Based on the research result, it can be concluded that the optimal illiteracy classes are three classes with the classification accuracy value of 96.4912% and error rate value of 3.5088%. Whereas the classification with two classes get the accuracy value of 93.8596% and error rate value of 6.1404%. And for the classification with five classes get the accuracy value of 90.3509% and error rate value of 9.6491%.**

***Keywords*— Illiteracy, Clustering, K means, Classification, Naïve Bayes.**

I. Introduction

Illiteracy is mostly occurred in developing country, including Indonesia [1]. Central Bureau of Statistics recorded that 8.7 million Indonesian people suffered from illiteracy in 2009 [2]. In 2014, 6 million of Indonesian people suffered from it [3]. In 2017, 3.4 million of Indonesian people still suffered from illiteracy. Those amount of numbers have made Indonesia ranked in fourth place of world's largest illiterate population [4]. Although the illiterate people in Indonesia is significantly decreasing in numbers, the government still have an obligation to solve this problem thoroughly to the root. The illiteracy problem should be overcome immediately for being literate is a part of human rights. If it is not solved well, the illiteracy people will be continuously unproductive, since illiteracy is closely correlated to the ignorance, backwardness, and empowerment.

A strategy which the government has already used to overcome illiteracy problem in Indonesia is applying block system, which means setting priorities in giving treatment from areas with the highest illiteracy rate and followed by areas with lower illiteracy rate. That strategy should be supported by classification process so that this problem can be managed appropriately.

Classification is the most commonly-used learning model to manage data mining [5]. Classification is needed to determine the class of an object whose class has not been determined. Reference [6] show that in classification process, there are four key components. First, class which states a label or category of an object. Second, variable predictor that represents the characteristics of the data. Third, training dataset which is used to establish classification model. Lastly, fourth, testing dataset to examine the classification model that has been established. The classification process itself consists of two steps: learning process and testing process. Learning process is an algorithm classification process, managing an analysis about training dataset to form classification rules. Whereas testing process is a process using testing dataset to examine the accuracy of classification rules that have been established. With this process of illiteracy classification, the government and concerned districts could determine the priorities in overcoming illiteracy in a proper way.

Some algorithms that have ever been used for classification process are K-Nearest Neighbors (KNN) [7], Decision Tree C4.5 [8], and Naïve Bayes [9]. However, after some considerations between the upsides, downsides, and also the characteristics for each algorithm, it is figured out that Naïve Bayes is the best algorithm for this problem.

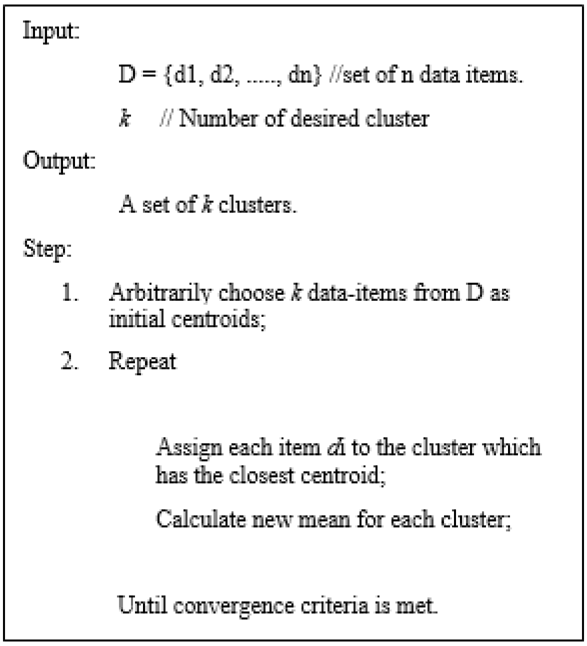
Naïve Bayes algorithm is a probability-based classification method that is often used, although this algorithm assumes all attributes are independent [9]. Naïve Bayes algorithm can work better in real life [10]. Besides that, Naïve Bayes algorithm does not need a big amount of training data to determine the parameter estimation of classification process.

Therefore, in this research, Naïve Bayes is chosen as classification algorithm of illiteracy in East Java. While in making illiteracy classes, it is used K Means algorithm, as in [11]. East Java was chosen since it is one of many provinces in Indonesia whose highest illiterate population in 2017. East Java province consists of 29 regencies and 9 cities with total illiterate population of 3.47% [4]. This fact is interesting to be inspected and investigated as East Java province is not a backward area, and since it also has good human resources in a large number.

II. The Proposed Algorithm

This research uses two kinds of algorithm, which are K means algorithm and Naïve Bayes algorithm. K means algorithm is used to establish illiteracy classes because there was no clear of illiteracy classes, while Naïve Bayes algorithm is used in the process of classifying illiteracy rate.

A. K means Algorithm



```
Input:
        D = {d1, d2, ....., dn} //set of n data items.
        k    // Number of desired cluster
Output:
        A set of k clusters.
Step:
    1.  Arbitrarily choose k data-items from D as
        initial centroids;
    2.  Repeat
            Assign each item di to the cluster which
            has the closest centroid;
            Calculate new mean for each cluster;
        Until convergence criteria is met.
```

Fig. 1 Pseudocode of k means algorithm

K means algorithm classifies objects in clusters according to their characteristics similarity [11]. The characteristics similarity can be formulated by using distance measure, one of them is Euclidean distance. The lower Euclidean distance's value, the more those two observations unit is similar. K means algorithm consists of two stages. The first stage is determining *k* based on the number of cluster. The next stage is choosing initial centroid from the dataset for each cluster. The classifying objects based on each object's distance towards the centroid by using Euclidean distance. If those objects are all classified already, it means the first iteration is done. Then, the thing should be done is getting new centroid for the next iteration process and classifying objects like what have been done in the first iteration process. Those steps are repeated all over again until the centroid does not change anymore [12]. The pseudocode of K means algorithm is shown in Fig. 1, as in [12].

B. Naïve Bayes Algorithm

Naïve Bayes algorithm is a simple probabilistic algorithm in classification technique which gets its probability value based on the calculation of frequency and value combinations from the associated collection [13]. This algorithm assumes that all attributes are independent [9]. The classification process of Naïve Bayes demands several clues or directions to determine the class of the data to be analyzed. Therefore, Equation 1 is applied [14].

$$P(C|F_1 \ldots F_n)=\frac{P(C).P(F_1 \ldots F_n|C)}{P(F_1 \ldots F_n)} \tag{1}$$

In Equation 1, variable *C* is a class, and variable $F_1 \ldots F_n$ represents characteristics required to do a classification. Therefore, the probability of matching data with a certain characteristic in C class (posterior) is C class probability emerged multiplied by the probability of samples characteristics in C class (likelihood), and then divided by the probability of samples characteristics globally (evidence).

```
Input:
        Training dataset T,
        F= ($f_1$, $f_2$, $f_3$,.., $f_n$)   // value of the predictor variable
        in testing dataset.
Output:
        A class of testing dataset.
Step:
    1.  Read the training dataset T;
    2.  Calculate the mean and standard deviation of the
        predictor variables in each class;
    3.  Repeat
            Calculate the probability of $f_i$ using the gauss
            density equation in each class;
        Until the probability of all predictor variables ($f_1$, $f_2$,
        $f_3$,.., $f_n$) has been calculated.
    4.  Calculate the likelihood for each class;
    5.  Get the greatest likelihood;
```

Fig. 2 Pseudocode of naïve bayes algorithm

Meantime, classification of continuous data uses Gauss density function as shown in Equation 2 [14].

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma^2_{ij}}} \quad (2)$$

While *P* is an probability, X_i is a predictor variable, x_i is the value of predictor variable, *Y* is the class to be looked for, y_i is sub class of *Y*, π is a constants valued 3.14, μ is mean that stated the average of all predictor variable in a certain class, σ is standard deviation that stated variants of all predictor variable in a certain class, and *e* is a constants valued 2.7183. The Naïve Bayes algorithm pseudocode with continuous data is shown as in Fig. 2.

III. Methodology

A. Research Data

The research data is obtained from Central Bureau of Statistics of East Java that can also be downloaded in its official website, www.jatim.bps.go.id. Data used as characteristics in this research consists of poor people percentage data, unemployment rate data, elementary school enrollment percentage, and junior high school enrollment percentage. Those data are selected for they are considered as factors that affect illiteracy in such kind of impact [2], [3]. The data of illiteracy rate is also required as a validation of clustering process. In this research, the research units are 38 regencies and cities in East Java in around 2013-2015, and there obtained 114 data record.

B. Pre-processing

Pre-processing stage in this research consists of data selection, data integration, and data transformation process.

1) Data Selection: Data selection is a kind of process of selecting data that is going to be involved in data mining. This step is done as the data table that has been obtained not only contains data required for research, but also contains other data as well.

2) *Data Integration:* Data integration is a process of combining data from multiple sources into a single unit of data that will be used for data mining. This step is necessary because the research data is obtained from different data tables.

3) Data Transformation: Data Transformation is a process of standardizing data so that it can be used for mining process.

C. Clustering and Classification Process

The sequences line of this research is shown in Fig. 3. In Fig. 3, the K means algorithm process will be repeated three times with different *k* values. The value of *k* = 2 to form two illiteracy classes, *k = 3* to form three illiteracy classes, and *k = 5* to form five illiteracy classes. The step is done to find the number of classes that are considered optimal and can provide the best results in the classification process using Naïve Bayes algorithm. After the entire clustering process is done, then followed by the process of classification using the Naïve Bayes algorithm. Naïve Bayes algorithm process will be repeated three times using the training dataset results from the clustering process that has been done before, the training dataset with two classes, training dataset with three classes, and training dataset with five classes. The step is taken to gain the best classification results based on the number of classes classified.

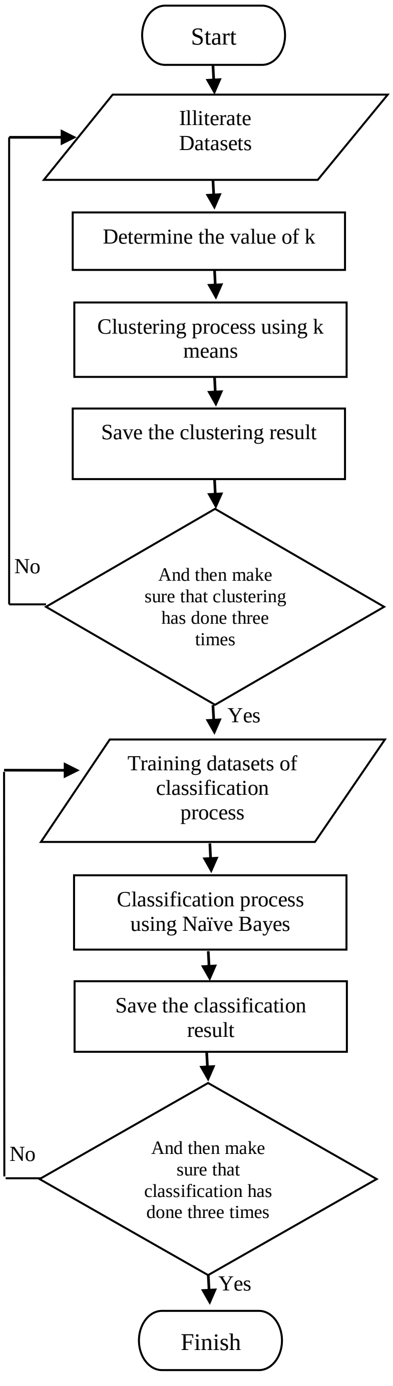

Start
Illiterate Datasets
Determine the value of k
Clustering process using k means
Save the clustering result
No
And then make sure that clustering has done three times
Yes
Training datasets of classification process
Classification process using Naïve Bayes
Save the classification result
No
And then make sure that classification has done three times
Yes
Finish

Fig. 3 Sequences line of research

D. Testing Method

Classification testing was performed using one cross-validation technique, which is the k-fold method [15]. In this research, a 10-fold method was chosen because this method is the best choice for obtaining accurate validation results based on extensive experimental results and theoretical proof [15]. This testing method is also recommended if the purpose of the test is to measure the classification error [16]. Then, the test results are displayed in the confusion matrix table, as in [17]. From the table of confusion matrix, there obtained the value of accuracy and error rate calculated using Equation 3 and Equation 4.

$$\text{accuracy}=\frac{\text{the number of tests is correct}}{\text{the number of tests}}\text{ x }100\% \qquad (3)$$

$$\text{error rate}=\frac{\text{the number of tests is wrong}}{\text{the number of tests}}\text{ x }100\% \qquad (4)$$

E. Evaluation

After the testing process has been done, there is an evaluation by looking for the best classification result. The best classification result is the one that has highest accuracy rate and lowest error rate.

IV. Result and Discussion

A. Pre-processing Analysis

At the pre-processing stage, the poor people percentage variable is initialized as F_1, the unemployment rate variable is initialized as F_2, the elementary school enrollment rate variable is initialized as F_3, and the junior high school enrollment rate variable is initialized as F_4. The results of data selection and data integration stages are shown in Table 1.

TABLE I
10 of 114 Samples of Data Selection and Data Integration

Name of Region	**Data Characteristics**			
	F₁	***F₂***	***F₃***	***F₄***
Pacitan	16.73	0.99	6.07	23.81
Ponorogo	11.92	3.25	4.31	18.71
Trenggalek	13.56	4.04	2.22	18.89
Tulungagung	9.07	2.71	0.95	18.21
Blitar	10.57	3.64	2.38	24.58
Kediri	13.23	4.65	2.00	27.54
Malang	11.48	5.17	1.56	28.24
Lumajang	12.14	2.01	4.98	31.85
Jember	11.68	3.94	2.13	27.22
Banyuwangi	9.61	4.65	6.56	24.81

The data in Table 1 has changed in value. In the variable of elementary school and junior high school enrollment rate obtained from the data source is a positive variable. Both variables need to be converted into negative variables to be the equal to the percentage of poor people and the percentage of unemployment rate that contributes negatively to the illiteracy class during the clustering process. The step is done since after conducting the first clustering process with the original data, it gets less appropriate results, where the areas that tend to have high illiteracy rates are classified as low-literate. The change in elementary school enrollment rate is done by 100% minus the enrollment rate of elementary school level. For example, the enrollment rate of elementary school is 93.93%, then the way to change it is 100% minus 93.93% to get the number 6.07%, the same way is done in the junior high school enrollment rate variable. For further explanation, the data transformation stage results are shown in Table 2.

TABLE II
10 of 114 Samples of Data Transformation

Name of Region	**Data Characteristics**			
	F₁	***F₂***	***F₃***	***F₄***
Pacitan	0.88581	-2.0218	1.37251	0.64015
Ponorogo	-0.0688	-0.6460	0.54576	-0.0307
Trenggalek	0.25662	-0.1651	-0.4359	-0.0070
Tulungagung	-0.6345	-0.9748	-1.0325	-0.0965
Blitar	-0.3368	-0.4086	-0.3608	0.74144
Kediri	0.19112	0.2061	-0.5393	1.13081
Malang	-0.1562	0.5227	-0.7460	1.22289
Lumajang	-0.0252	-1.4009	0.86049	1.69778
Jember	-0.1165	-0.2260	-0.4783	1.08872
Banyuwangi	-0.5273	0.20617	1.60268	0.77169

In Table 2, the negative value indicates that the data is below the total average, while the positive value indicates that the data is above the total average. The result of data transformation shown in Table 2 will be used as a clustering process dataset using K means algorithm to form illiteracy classes.

B. Clustering Analysis using K means Algorithm

Clustering process is performed three times with the values of *k=2*, *k=3*, and *k=5*. Initial cluster determining step with value of *k=2* is shown in Table 3.

TABLE III
Initial Cluster Centers

Data Characteristics	**Clusters**	
	Cluster 1	***Cluster 2***
F_1	2.17596	-1.52378
F_2	1.50281	-1.14525
F_3	-0.10717	1.03429
F_4	3.06717	-1.63692

In Table 3, where the initial cluster centers in cluster 1 is the 26th data and initial cluster centers in cluster 2 is the 76th data. After five times iteration, there obtained final cluster centers that is shown in Table 4.

TABLE IV
Final Cluster Centers

Data Characteristics	**Clusters**	
	Cluster 1	***Cluster 2***
F_1	0.94992	-0.45645
F_2	-0.64229	0.30863
F_3	0.17568	-0.08442
F_4	0.90804	-0.43633
Average	0.347838	-0.16714

In Table 4, it can be seen that the average of cluster 1 is 0.347838 and cluster 2 is -0.16714. So, the conclusion is that cluster 1 is higher than cluster 2, which means cluster 1 is high illiterate class and cluster 2 is low illiterate class. The final result of the clustering process using the K means algorithm with the value k = 2 is shown in Table 5.

TABLE V
10 OF 114 SAMPLES OF CLUSTERING RESULT

Name of Region	Label of Cluster	Value of Euclidean Distance
Pacitan	Cluster 1	1.84701
Ponorogo	Cluster 2	1.27409
Trenggalek	Cluster 2	1.02021
Tulungagung	Cluster 2	1.64115
Blitar	Cluster 2	1.41151
Kediri	Cluster 1	1.36255
Malang	Cluster 2	1.82393
Lumajang	Cluster 1	1.62836
Jember	Cluster 1	1.33075
Banyuwangi	Cluster 2	2.07874

From the results shown in Table 5, 37 regions were grouped in clusters 1 and 77 regions grouped in cluster 2. Then, the clustering process will be repeated in a same way for *k = 3* and *k = 5*. After all the clustering process has been done, it is obtained the number of members of each cluster as shown in Table 6.

TABLE VI
MEMBER IN EACH CLUSTER

Value of k	Label of Cluster	Number of Members	Percentage
k = 2	Cluster 1	37 regions	32%
	Cluster 2	77 regions	68%
k = 3	Cluster 1	25 regions	22%
	Cluster 2	31 regions	27%
	Cluster 3	58 regions	51%
k = 5	Cluster 1	12 regions	10%
	Cluster 2	27 regions	24%
	Cluster 3	27 regions	24%
	Cluster 4	19 regions	17%
	Cluster 5	29 regions	25%

C. Classification Analysis using Naïve Bayes Algorithm

After the illiteracy classes have been formed by clustering process with K means algorithm, then classification process can take place. The classification is done by three times based on the number of classes formed in the preceded clustering process. First classification process uses training dataset from the result of clustering process that has formed two illiteracy classes by using 10-fold testing method. The result of first classification process is shown in Table 7.

TABLE VII
FIRST CLASSIFICATION RESULT

Name of Region	Predicted Label	Label	Correction
Pacitan	Cluster 1	Cluster 1	Correct
Ponorogo	Cluster 2	Cluster 2	Correct
Trenggalek	Cluster 2	Cluster 2	Correct
Tulungagung	Cluster 2	Cluster 2	Correct
Blitar	Cluster 2	Cluster 2	Correct
Kediri	Cluster 1	Cluster 1	Correct
Malang	Cluster 2	Cluster 2	Correct
Lumajang	Cluster 1	Cluster 1	Correct
Jember	Cluster 1	Cluster 1	Correct
Banyuwangi	Cluster 2	Cluster 2	Correct

In Table 7, if the predicted label results give the same output results as the label, then the classification process performed is correct. Conversely, if the predicted label results give an output that is different from the label, then the classification process is wrong. Confusion matrix of the first classification results is shown in Table 8.

TABLE VIII
CONFUSION MATRIX OF FIRST CLASSIFICATION RESULT

Correct Classification	Classified as	
	Cluster 1	***Cluster 2***
Cluster 1	30	7
Cluster 2	0	77

In Table 8, it can be seen that there are 37 cluster members 1 and seven cluster members 1 are incorrectly classified as cluster 2 members. Meanwhile, there are 77 cluster members 2 and all members are correctly classified. From the test result, we can calculate the accuracy value using Equation 3 and error rate using Equation 4, so that the accuracy value is 93.8596% and the error rate is 6.1404%. Furthermore, the classification process is repeated by training a different dataset, which is by training a dataset consisting of three classes and five classes of illiteracy. So as to obtain the final result of accuracy and error rate shown in Table 9.

TABLE IX
CLASSIFICATION FINAL RESULT

Number of Class	Accuracy	Error Rate
2	93.8596%	6.1404%
3	96.4912%	3.5088%
5	90.3509%	9.6491%

Based on the final result of the classification process in Table 9, it can be seen that classification with three classes will give the best result. Classification with three classes obtained the highest accuracy value compared to other classification of 96.4912% and obtained the lowest error rate that is equal to 3.5088%. Three classes formed are: cluster 1 as high illiteracy class with 25 members, cluster 2 as medium illiterate class with 31 members, and cluster 3 as low literacy class with total number of 58 members.

V. Conclusions

Based on the results and discussion can be concluded:

- Naïve Bayes algorithm can be used for classifying illiteracy.
- The optimal number of illiteracy classes are three classes. Those three classes are: high illiteracy class with 25 members, moderate illiteracy class with 31 members, and low illiteracy class with 58 members.
- From the three classification processes that have been done, classification with three classes that obtain the best results with an accuracy of 96.4912% and error rate of 3.5088%. The second rank is a classification with two classes that obtained an accuracy of 93.8596% and error rate of 6.1404%. The third one is a classification with five classes that obtain an accuracy of 90.3509% and an error rate of 9.6491%.

In order to get better research result, it is highly suggested to:

- Augment data characteristics that can do such an impact to the number of illiteracy, so there will be a better clustering process.
- Increase the number of research unit, so there will be a more accurate classification process.

References

[1] Mariyono, “Strategi Pemberantasan Buta Aksara Melalui Penggunaan Teknik Metastasis Berbasis Keluarga,” *Pancaran*, vol. 5, no. 1, pp. 55–66, 2016.

[2] R. D. Bekti, E. Irwansyah, and Andiyono, “Analisis Faktor yang Mempengaruhi Angka Buta Huruf Melalui Geographically Weighted Regression: Studi Kasus Propinsi Jawa Timur,” *Comtech*, vol. 4, no. 1, pp. 443–449, 2013.

[3] R. Maharani and S. Winahju, “Pemodelan Angka Buta Huruf di Provinsi Sumatera Barat Tahun 2014 dengan Geographically Weighted Regression,” *J. Sains dan Seni ITS*, vol. 5, no. 2, pp. 361–267, 2016.

[4] Antara. (2017) Indonesia Peringkat Keempat Penduduk dengan Buta Huruf Terbanyak homepage on Media Indonesia. [Online]. Available: http://mediaindonesia.com/read/detail/121475-indonesia-peringkat-keempat-penduduk-dengan-buta-huruf-terbanyak.

[5] S. R. Ahmed, “Applications of Data Mining in Retail Business,” in *International Conference on Information Technology: Coding and Computing*, 2004, pp. 455 – 459.

[6] H. Leidiyana, “Penerapan Algoritma K-Nearest Neighbor untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor,” *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[7] E.-H. Han, G. Karypis, and V. Kumar, “Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2001, pp. 53–65.

[8] K. Polat and S. Güneş, “A Novel Hybrid Intelligent Method Based on C4.5 Decision Tree Classifier and One-Against-All Approach for Multi-Class Classification Problems,” *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1587–1592, 2009.

[9] R. Abraham, J. B. Simha, and S. S. Iyengar, “A Comparative Analysis of Discretization Methods for Medical Datamining with Naïve Bayesian Classifier,” in *Proceedings - 9th International Conference on Information Technology*, 2007, pp. 235–236.

[10] S. A. Pattekari and A. Parveen, “Prediction System for Heart Disease Using Naive Bayes,” *Int. J. Adv. Comput. Math. Sci.*, vol. 3, no. 3, pp. 290–294, 2012.

[11] C. Slamet, A. Rahman, M. A. Ramdhani, and W. Dharmalaksana, “Clustering the verses of the holy qur’an using K-means algorithm,” *Asian J. Inf. Technol.*, vol. 15, no. 24, pp. 5159–5162, 2016.

[12] K. a A. Nazeer and M. P. Sebastian, “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm,” in *Proceedings of the World Congress on Engineering*, 2009, vol. I, pp. 1–5.

[13] T. R. Patil and S. S. Sherekar, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” *Int. J. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 256–262, 2013.

[14] Bustami, “Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah Asuransi,” *J. Inform.*, vol. 8, no. 1, pp. 1–15, 2014.

[15] S. K. Lidya, O. S. Sitompul, and S. Efendi, “Sentiment Analysis pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN),” in *Seminar Nasional Teknologi dan Komunikasi*, 2015, pp. 1–8.

[16] J. D. Rodríguez, A. Pérez, and J. A. Lozano, “Sensitivity Analysis of Kappa-Fold Cross Validation in Prediction Error Estimation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, 2010.

[17] H. D. Masethe and M. A. Masethe, “Prediction of Heart Disease using Classification Algorithms,” in *Proceedings of the World Congress on Engineering and Computer Science*, 2014, vol. 2, pp. 22–24.