Increase Accuracy of Naïve Bayes Classifier Algorithm with K-Means Clustering for Prediction of Potential Blood Donors

Chandra Kurniawan Putra Rukma [1,*], Alamsyah [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: kurniawanc13@students.unnes.ac.id

ARTICLE INFO

Article history
Received 7 Maret 2022
Revised 13 Maret 2022
Accepted 15 April 2022

Keywords
Data mining
Naïve Bayes Classifier
K-Means Clustering
RFMTC
Blood donors

ABSTRACT

Branch of computer science knowledge is data mining. Data mining help people to processing a big and irregular data. In public health, data mining can be used to manage blood donors data. Blood donors is a proses to take some blood from volunteer then given to other people who need. One of the ways to fill up blood requirement in Indonesia is organize blood donors event regularly, but some people didn't routine give they blood. Solution of that problems, a system needed to predic future blood donor behavior. Recency, Frequency, Monetary, Time Churn Probability (RFMTC) is a modification from Recency of purchase, Frequency of purchase, and Monetary value of purchase (RFM) used to predict a blood donors behavior. In this research implemented a Naïve Bayes Classifier to blood donors classification The classification result with 224 data from RFMTC dataset is 78.13% accuracy. Combination Naïve Bayes Classifier algorithm with K-Means Clustering increase accuracy to 80.80%.

1 Introduction

The concept of data mining is recognized as an important tool in information management. Clustering is an example of a data mining technique, which is a method that organizes several data into a cluster, so that the data in a cluster has the same value (Selviana, 2016). Clustering is the process of distributing several dataset objects to different groups or clusters and the classification process is carried out without supervision (Nurzahputra et al., 2017). Beside that, K-Means is an unsupervised classification method and an example of a data clustering method. K-Means Clustering is a non-hierarchical method that divides some data into several clusters or groups. This method allows cluster members who have the same characteristics to group into the same cluster (Agusta, 2007).

I-Cheng, King-Jang, and Tao-Ming Ting modified the RFM (Recency of purchase, Frequency of purchase, and Monetary value of purchase) method to a better method, namely RFMTC (Recency, Frequency, Monetary, Time, Churn Probability) method used to predict blood donor behavior in the future (Yeh et al., 2009). The Naïve Bayes Classifier algorithm is an effective classification algorithm with precise and efficient values. Utilization of existing input quickly becomes the reasoning process of the Naïve Bayes Classifier algorithm. The process of the Naïve Bayes Classifier algorithm is a selective data class classification measured by predictive accuracy (Zhang & Su, 2004).

2 Methods

2.1 Related Research

According to research conducted by Darwiche et al. (2010) in an article entitled "Prediction of Blood Transfusion Donations" using Multilayer Perceptron (MLP) and Support Vector Machine

(SVM) on the RFMTC dataset variable (Recency, Frequency, Monetary, Time, Churn Probability) with 748 RFMTC datasets from UCI Machine Learning, the final results are 65.8% sensitivity and 78.2% specificity using 600 training data and 148 test data. In another study by Susanto and Agustina (2016) with the title is "Komparasi Akurasi Algoritma C4.5 dan *Naïve Bayes* untuk Prediksi Pendonor Darah Potensial dengan *Dataset* RFMTC" which discussed about the comparison of the accuracy of the C4.5 and Naïve Bayes algorithms on the dataset RFMTC and get an accuracy of 70.30% with the Naïve Bayes algorithm and an accuracy of 67.12% with the C4.5 algorithm on the RFMTC dataset using the UDD PMI dataset in Bantul Regency with RFMTC variables and two class variables, namely "donor", "tidak donor", so that there are 600 training data and 165 test data processed with the C4.5 algorithm and the Naïve Bayes algorithm and then evaluated with a confusion matrix and ROC curve.

Nugroho (2018) also doing a study "Klasifikasi Pendonor Darah Menggunakan Metode *Support Vector Machine* (SVM) Pada *Dataset* RFMTC " which aims to predict the behavior of blood donors to anticipate blood stock deficits using the Support Vector Machine method and obtain classification results on the RFMTC dataset is 72.64% accuracy. In this study, the ratio of 50%:50% was used to divide 748 data into training data and test data, then a linear kernel was used to normalize the data and the Support Vector Machine (SVM) method to calculate it. The last is a study conducted by Shinde et al. (2015) "Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm" using a combination of Naïve Bayes and K-Means Clustering methods to create a system that aims to predict heart disease.

2.2 Data Mining

Data mining is a series of many processes to explore the value of knowledge that has not been known until now due to manual data processing (Bustami, 2013). Based on Sugiharti et al. (2017) explained that the characteristics of data mining are finding something hidden and certain unknown data patterns also most data mining uses big data, often used to make data values more accurate so that it is useful for making important decisions, especially on strategy.

2.3 Naïve Bayes Classifier

British scientist Thomas Bayes is the discoverer of the Naïve Bayes theorem. The main purpose of using Naïve Bayes is to predict future opportunities based on past experience (Prehanto et al., 2019). The flowchart of the dataset classification process using the Naïve Bayes Classifier is shown in Figure 1.

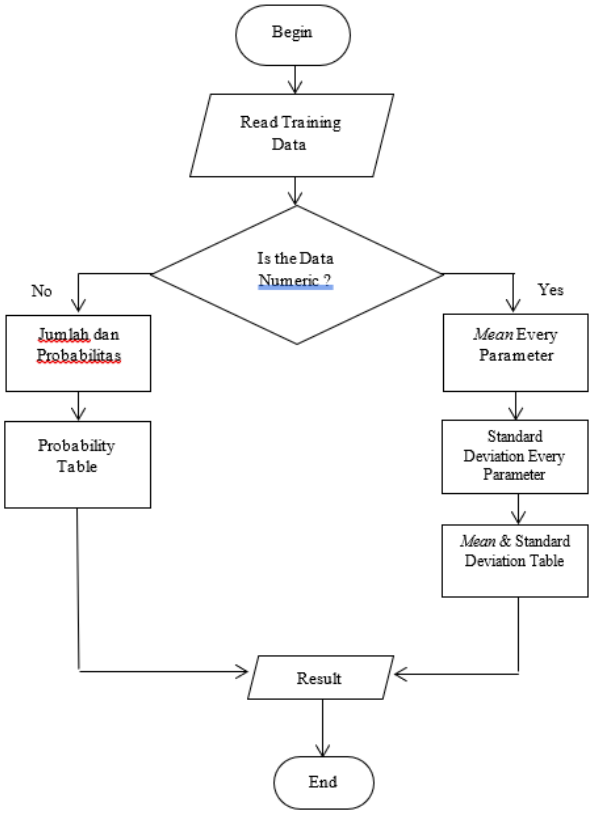

Figure 1. Naïve Bayes process flowchart

2.4 Clustering

Clustering is a process that distributes several dataset objects to several groups or clusters that match and without a supervised classification process (Ji et al., 2019). This method is also the most widely used method, especially for data grouping models (Windarto & Wanto, 2018).

2.5 K-Means

K-Means is a non-hierarchical data clustering method that manages objects in a dataset into one or more clusters with the desired number. The K-Means method tries to group data that have the same characteristics with each other into groups, while different characteristics will be grouped into other groups (Aldino et al., 2021). This method aims to find the best formation with the average distance of each node minimized (Hassan et al., 2020). So, it can reduce data variations in the same cluster and maximize variations between other clusters. Flowchart data clustering process used K-Means method shown on Figure 2.

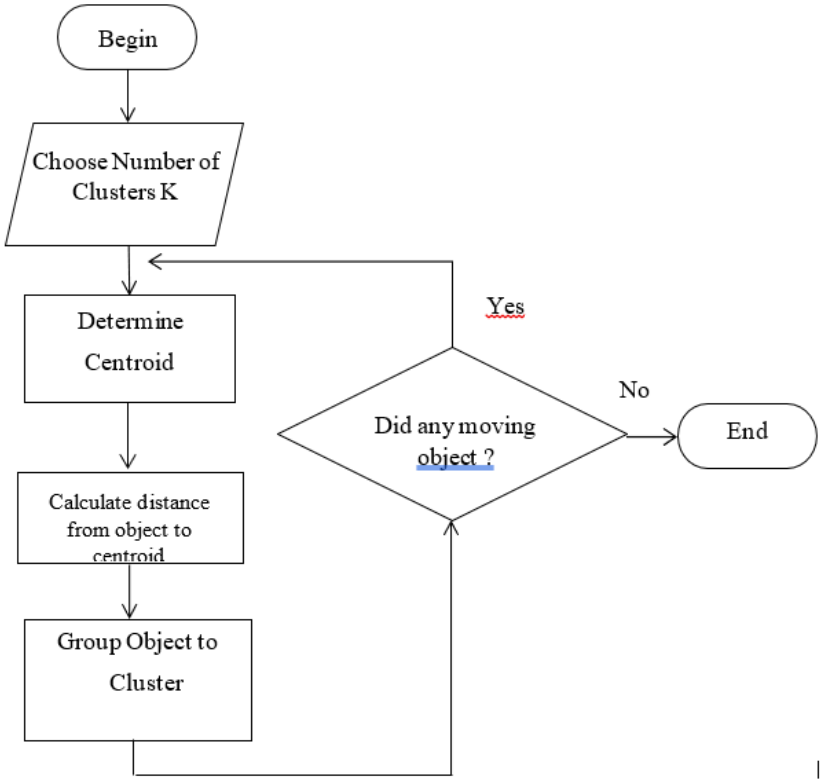

Figure 2. K-Means Clustering process flowchart

3 Method

The data used in this study is a Blood Transfusion Service Center Dataset taken from UCI Machine Learning with a total of 748 samples and 4 attributes of the RFMTC model, namely Recency, Frequency, Monetary, Time, and Churn Probability. The data were classified as “donor” or “tidak donor”, then separated into 224 test data and 524 training data. The flow chart of the combination of the Naïve Bayes Classifier and K-Means Clustering algorithm is shown in Figure 3.

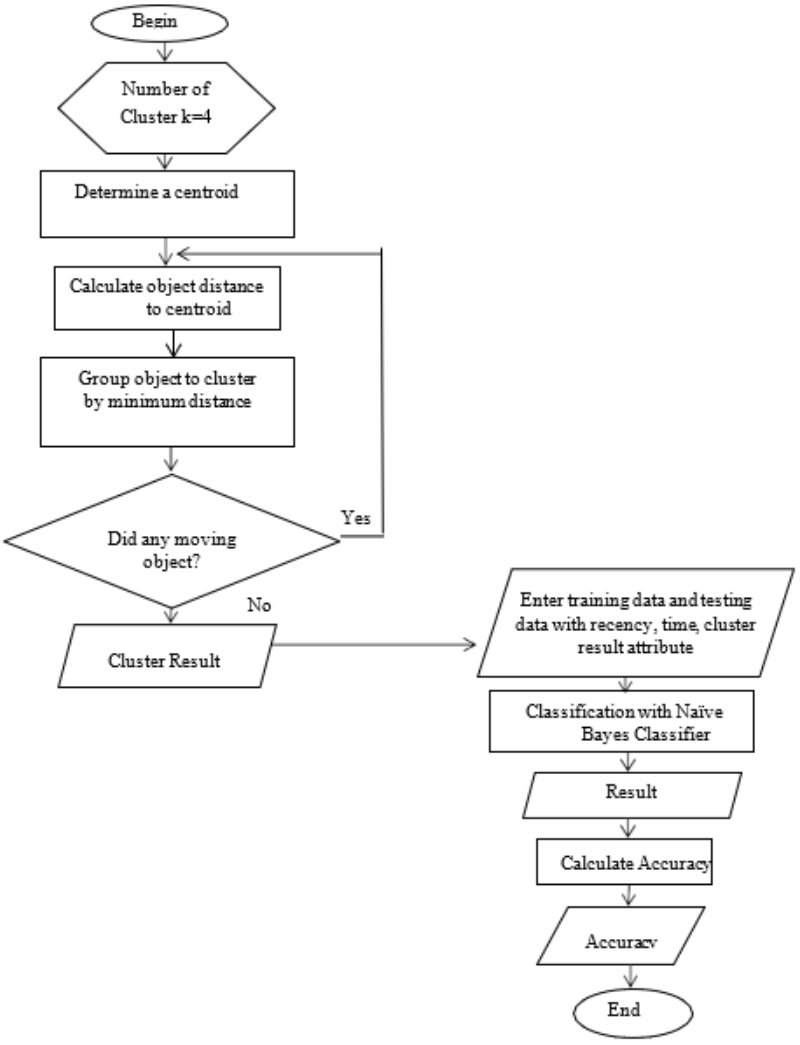

Figure 3. Flowchart Process combination of Naïve Bayes Classifier and K-Means Clustering

The process begins by classifying the RFMTC dataset using the Naïve Bayes Classifier. Then do clustering used K-Means Clustering method on RFMTC dataset and reclassification using the Naïve Bayes Classifier again. The addition of the K-Means Clustering method to the classification process is expected to improve prediction accuracy in the dataset of prospective blood donors.

4 Results and Discussion

First step do a classification on dataset used original Naïve Bayes Classifier algorithm. Classification result shown in Table 1.

Table 1. Final probability of Naïve Bayes Classifier

ID Pendonor	Recency		Frequency		Monetary		Time		Result		Aktual
	Tidak Donor	Donor	Tidak Donor	Donor	Tidak Donor	Donor	Tidak Donor	Donor	Tidak Donor	Donor	
525	-3.336	-2.588	-8.384	-8.614	-2.863	-3.093	-4.167	-4.087	-18.01	-18.124	Tidak Donor
526	-3.555	-2.768	-8.074	-8.748	-2.553	-3.226	-4.588	-4.518	-18.029	-19.001	Tidak Donor
527	-3.555	-2.768	-8.057	-8.695	-2.536	-3.174	-4.415	-4.336	-17.823	-18.713	Donor

528	-3.555	-2.768	-10.057	-8.882	-4.536	-3.361	-4.826	-4.948	-22.235	-19.699	Tidak Donor
529	-3.249	-2.556	-15.228	-10.159	-9.706	-4.637	-5.723	-6.018	-33.165	-23.11	Tidak Donor
530	-3.555	-2.768	-8.08	-8.656	-2.559	-3.134	-4.262	-4.178	-17.715	-18.476	Donor
531	-3.336	-2.588	-8.057	-8.695	-2.536	-3.174	-4.415	-4.336	-17.604	-18.534	Donor
532	-3.555	-2.768	-8.074	-8.748	-2.553	-3.226	-4.479	-4.403	-17.921	-18.886	Donor
533	-3.336	-2.588	-8.244	-8.615	-2.722	-3.094	-4.167	-4.087	-17.728	-18.125	Tidak Donor
534	-3.555	-2.768	-8.074	-8.748	-2.553	-3.226	-4.479	-4.403	-17.921	-18.886	Tidak Donor
535	-3.555	-2.768	-8.08	-8.656	-2.559	-3.134	-4.192	-4.11	-17.645	-18.408	Tidak Donor
…	…	…	…	…	…	…	…	…	…	…	…
748	-30.803	-87.055	-8.361	-8.984	-2.839	-3.462	-5.259	-5.466	-46.522	-104.708	Tidak Donor

From that Naïve Bayes Classifier final probability table, confusion matrix used to calculate accuracy. Confusion matrix from Naïve Bayes Classifier result shown in Table 2.

Table 2. Confusion Matrix from Naïve Bayes Classifier result

Prediction		Actual	
		Yes	No
	Yes	5	12
	No	37	170

From Table 2 obtained 5 TP, 170 TN, 37 FP and 12 FN. Formula 1 is used to calculate accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x100\% \quad (1)$$

Description:

TP = True Positive mean actual value is yes and nilai prediction value is yes

TN = True Negative mean actual value is no and prediction value is no

FP = False Positive mean actual value is no and prediction value is yes

FN = False Negative means actual value is yes and prediction value is no

$$Accuracy = \frac{5+170}{5+170+37+12} x100\% = 78,3\%$$

Next do a clustering for Frequency and Monetary attributes, because both attributes have directly proportional values. Number of cluster = 4. Clustering result shown in Table 3.

Table 3. Clustering result of Frequency and Monetery attribute

ID Pendonor	Frequency	Monetery	Distance to Centroid				Cluster
			C1	C2	C3	C4	

525	9	2250	8.88	44.03	36.07	48.20	C1
526	4	1000	28.13	7.03	0.93	11.20	C3
527	5	1250	25.13	10.03	2.07	14.20	C3
528	15	3750	21.13	14.03	6.07	18.20	C3
529	24	6000	17.13	18.03	10.07	22.20	C3
530	6	1500	37.13	1.97	9.93	2.20	C2
531	5	1250	34.13	1.03	6.93	5.20	C2
532	4	1000	29.13	6.03	1.93	10.20	C3
533	8	2000	32.13	3.03	4.93	7.20	C2
534	4	1000	4.88	40.03	32.07	44.20	C1
535	6	1500	18.13	17.03	9.07	21.20	C3
…	…	…	…	…	…	…	…
748	1	250	40.13	4.97	12.93	0.80	C4

Next do classification used Naïve Bayes Classifier with a result of clustering method on Frequency and Monetary attribute. Classification result shown in table 4.

Table 4. Combination result probability of Naïve Bayes Classifier and K-Means

ID Pendonor	Recency		Time		Cluster (F & M)		Result		Actual
	Tidak Donor	Donor	Tidak Donor	Donor	Tidak Donor	Donor	Tidak Donor	Donor	
525	-33.364	-258.849	0.47121	-0.19988	-416.679	-408.737	-629.153	-661.619	Tidak Donor
526	-355.538	-276.814	0.47121	-0.19988	-458.766	-451.841	-693.137	-722.688	Tidak Donor
527	-355.538	-276.814	0.47121	-0.19988	-441.511	-433.586	-675.882	-704.434	Donor
528	-355.538	-276.814	0.47121	-0.19988	-482.611	-4.948	-716.982	-765.647	Tidak Donor
529	-324.892	-255.575	-3.127.225	-811.003	-572.251	-601.764	-3.950.322	-1.642.388	Tidak Donor
530	-355.538	-276.814	0.47121	-0.19988	-426.169	-417.816	-660.541	-688.663	Donor
531	-33.364	-258.849	0.47121	-0.19988	-441.511	-433.586	-653.985	-686.469	Donor
532	-355.538	-276.814	0.47121	-0.19988	-447.925	-44.033	-682.297	-711.178	Donor

533	-33.364	-258.849	0.47121	-0.1998	-416.679	-408.737	-629.153	-661.619	Tidak Donor
534	-355.538	-276.814	0.47121	-0.1998	-447.925	-44.033	-682.297	-711.178	Tidak Donor
535	-355.538	-276.814	0.47121	-0.1998	-419.192	-41.102	-653.563	-681.867	Tidak Donor
…	…	…	…	…	…	…	…	…	…
748	-3.080.308	-8.705.517	0.47121	-0.1998	-525.868	-546.642	-348.501	-9.246.192	Tidak Donor

From the final probability table of Naïve Bayes Classifier combined with K-Means Clustering, a confusion matrix is used to calculate accuracy. The confusion matrix of the Naïve Bayes Classifier and the K-Means results are shown in Table 5.

Table 5. Confusion matrix from combination Naïve Bayes Classifier and K-Means Clustering

		Actual	
Prediction		Yes	No
	Yes	0	1
	No	42	181

From Table 5 obtained 0 TP, 181 TN, 42 FP and 1 FN. Formula 1 is used to calculate accuracy.

$$Accuracy = \frac{0+181}{0+181+42+1} x100\% = 80,80\%$$

Based on the research, the combination of the Nave Bayes Classifier algorithm with K-Means Clustering is more accurate than the original Nave Bayes Classifier without clustering. The results of the comparison accuracy between the original Nave Bayes Classifier and the combination of Nave Bayes Classifier with K-Means Clustering are shown in Table 6.

Table 6. Comparison accuracy result

Algorithm	Accuracy
Naïve Bayes Classifier	78,13%
Naïve Bayes Classifier + K-Means Clustering	80,80%

In Table 6, the accuracy of the Naïve Bayes Classifier algorithm increased by about 2.67% from the previous 78.13% to 80.80% when combined with K-Means Clustering. The K-Means Clustering method groups 2 attributes in the dataset into several clusters, so that the Naïve Bayes Classifier algorithm can work more effectively. Therefore, the combination of Naïve Bayes Classifier and K-Means Clustering can be used as an algorithm recommendation to determine predictions of prospective blood donors.

5 Conclusion

From the research and discussion about increasing the accuracy of the Naïve Bayes Classifier algorithm combined with the K-Means Clustering method in the prediction of prospective blood

donors, it can be concluded that the accuracy of the combination of Naïve Bayes Classifier and K-Means is 80.80%. This accuracy is more precise than the Naïve Bayes Classifier without attribute clustering using the K-Means Clustering method which produces an accuracy of 78.13%. Accuracy increased by about 2.67%.

References

Agusta, Y. J. J. S. d. i. (2007). K-means–penerapan, permasalahan dan metode terkait. *3*(1), 47-60.

Aldino, A., Darwis, D., Prastowo, A., & Sujana, C. (2021). Implementation of K-means algorithm for clustering corn planting feasibility area in south lampung regency. Journal of Physics: Conference Series,

Bustami, B. J. T.-J. T. I. (2013). Penerapan algoritma Naive Bayes untuk mengklasifikasi data nasabah asuransi. *5*(2).

Darwiche, M., Feuilloy, M., Bousaleh, G., & Schang, D. (2010). Prediction of blood transfusion donation. 2010 fourth international conference on research challenges in information science (RCIS),

Hassan, A. A.-h., Shah, W. M., Othman, M. F. I., Hassan, H. A. H. J. I. J. o. E., & Engineering, C. (2020). Evaluate the performance of K-Means and the fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks. *10*(2).

Ji, X., Henriques, J. F., & Vedaldi, A. (2019). Invariant information clustering for unsupervised image classification and segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Nugroho, E. B. (2018). *Klasifikasi Pendonor Darah Menggunakan Metode Support Vector Machine (SVM) pada Dataset RFMTC* Universitas Brawijaya].

Nurzahputra, A., Muslim, M. A., & Khusniati, M. J. T. C. (2017). Penerapan algoritma K-Means untuk clustering penilaian dosen berdasarkan indeks kepuasan mahasiswa. *16*(1), 17-24.

Prehanto, D., Indriyanti, A., Nuryana, K., Soeryanto, S., & Mubarok, A. (2019). Use of Naïve Bayes classifier algorithm to detect customers' interests in buying internet token. Journal of Physics: Conference Series,

Selviana, N. H. (2016). Analisis Perbandingan k-means dan fuzzy c-means untuk pemetaan motivasi balajar mahasiswa. Seminar Nasional Teknologi Informasi Komunikasi dan Industri,

Shinde, R., Arjun, S., Patil, P., Waghmare, J. J. I. J. o. C. S., & Technologies, I. (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. *6*(1), 637-639.

Sugiharti, E., Firmansyah, S., Devi, F. R. J. J. o. T., & Technology, A. I. (2017). Predictive evaluation of performance of computer science students of unnes using data mining based on naÏve bayes classifier (NBC) algorithm. *95*(4), 902.

Susanto, W. E., & Agustina, C. J. S. N. I. K., no. Snik. (2016). Komparasi Akurasi Algoritma C4. 5 Dan Naive Bayes Untuk Prediksi Pendonor Darah Potensial Dengan Dataset Rfmtc. 16-21.

Windarto, A. P., & Wanto, A. (2018). Data mining tools| rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia. IOP Conference Series: Materials Science and Engineering,

Yeh, I.-C., Yang, K.-J., & Ting, T.-M. J. E. S. w. A. (2009). Knowledge discovery on RFM model using Bernoulli sequence. *36*(3), 5866-5871.

Zhang, H., & Su, J. (2004). Naive bayesian classifiers for ranking. European conference on machine learning,