p-ISSN: 1979-5661
e-ISSN: 2622-321X

User Loyalty Prediction Using Naïve Bayes Method in "Udatari" an art Performance Marketplace

Ngurah Agus Sanjaya ER[a1], I Gusti Agung Gede Arya Kadyanan[a2]

[a]Informatics Department, Faculty of Mathematics and Natural Science University of Udayana
Bukit Jimbaran, Indonesia
[1]agus_sanjaya@unud.ac.id
[2]gungde@unud.ac.id(Corresponding author)

Abstract

Udatari is the first traditional dance platform in Indonesia which provides information about traditional events such as, dance tutorials, group dancer and dance attributes. The tight competition in the startup world, requires Udatari as a new startup to manage application users optimally. Knowing loyal users will help startups determine the right marketing strategy. In this study, the method used for clustering is the K-Means method where this method seeks to classify existing data into several groups provided that the data in one group have the same characteristics as each other. The model used for the clustering process is RFM, namely recency, frequency and monetary. The purpose of this clustering is to get the segmentation of users who have different Customer Lifetime Value. The second method for conducting classification is the Naïve Bayes method, where this method predicts future opportunities based on past experiences. The purpose of this classification is to predict new users into the user segmentation obtained from the clustering results.
From the results of this study, the optimum k value for K-Means are 3 clusters with the largest CLV value in the second cluster where testing on this method uses the Silhouette Index. Furthermore, for the test results of the Naïve Bayes method, the average accuracy value is 97.44% where the accuracy of each class is 92.31% for cluster 0 (first cluster), 100% for the second cluster and 100% for the third cluster.

Keywords: *K-Means, Naïve Bayes, Loyalty, Segmentation, RFM*

1. Introduction

From April 2018, Udatari startup have reached 400 users, 23 partners, and 1000 transactions. Until now, there so many transaction data with information including transaction times, the amount of payments and others that have not been processed properly to obtain information that can increase value for startups. In addition, there is no user mapping which causes the same effect on all users, for example in determining promos, prizes, advertisements and other services. Based on previous research by Hardiani et al[3] regarding the use of the K-Means method and the RFM model for segmentation of savings customers at Microfinance Institutions, the optimum k value using the Davies Bouldin Index are 3 with the results of the first cluster consisting of 100 customers, the second 69 customers and third 79 customers. The mean value of RFM score was between 441-544 and included in the superstar group. Then in the research of Kamila et al[5] regarding the grouping of loading and unloading transaction data in Riau Province where in the journal discussing the comparison between the K-Medoids and K-Means algorithms, it was found that the comparison between the two algorithms did not show any significant differences regarding

p-ISSN: 1979-5661
e-ISSN: 2622-321X

data grouping. The K-Means algorithm only takes an average of 1 second and the K-Medoids an average of 1 minute 38 seconds.

The result of optimum k value on K-Means is 3 lowers than K-Medoids where the optimum k result is search using Davies Bouldin Index. Then in research by Santosa[8] discusses how to optimize the accuracy of the prediction model using the Naïve Bayes algorithm based on Particle Swarm Optimization. From this study, the results obtained in the form of an accuracy of 98.54%, and using the ROC curve produces an AUC value of 0.99. The time used to process this test takes approximately 2 hours 45 minutes with 3333 records used.

In this study, the authors will conduct research on the prediction of user loyalty based on the results of user segmentation using the K-Means method and the RFM model. In the research that will be conducted, to predict the level of user loyalty using the Naïve Bayes method. The advantage of the Naïve Bayes method is that it can produce high accuracy with a small amount of training data[14]. To perform a classification using Naïve Bayes, several features are needed that are used as a basis for predictions, namely recency, frequency, and monetary (RFM). For training data and testing data used are clustering data where each cluster has been labeled. In this study, to perform clustering using the K-Means method. The reason for choosing this method is that it is an interactive method that is easy to interpret, apply, and is dynamic on scattered data. To facilitate the clustering process, it is necessary to have a model that describes the features used. The testing process uses the Silhouette Index (SI) method. The results of this study are expected to be useful for Udatari startups to perform different services to users and predict user satisfaction to anticipate a decrease in user satisfaction.

2. Reseach Methods

Here using Naïve Bayes method to predict which new user belongs to which user group. Where this method uses probability and statistical calculations. The results of the classification process will be tested using the Confusion Matrix with the aim of measuring the performance of the classification method. This analysis describes the system to be designed in general. The description of the system design of this study is as follows:

1. Retrieving user and transaction data using the REST API.
2. Performing the RFM extraction process, namely recency, frequency and monetary.
3. After carrying out the RFM extraction, the next step is to enter the data preparation stage, which is to prepare the data so that the data is quality and there are no defects that enter the method application stage.
4. Applying the K-Means method to generate user categories (clusters).
5. Conducting a testing process on the number of clusters obtained to find the optimum number of clusters.
6. After getting the optimum number of clusters, then look for the CLV value from the average RFM value in each cluster. The greater the CLV value, the higher the user loyalty in the cluster. After calculating the CLV value, the cluster is sorted according to the CLV value. In this study, cluster 0 is called the first cluster.
7. Perform denormalization of existing data in each cluster. This is done so that the data is back to normal, so that it can be used in the prediction process.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

RFM analysis is an analysis that can help distinguish influential parties from large data based on three variables, namely the interval of one transaction to the present, frequency and amount of money.

1. Recency (R)
   Vulnerable from a transaction is currently denoted by R or what is called recency. The shorter the interval, the greater the R value.
2. Frequency (F)
   F represents frequency, which is the number of transactions in a certain period in a certain period, for example, twice in one year or twice in one nine. The higher the frequency the greater the F value.
3. Monetary (M)
   M represents monetary, which is the value of the product in terms of money for a certain period. The more money in that period, the higher the value of M.

Classification using the Naïve Bayes Classifier is a classification using probabilistic and statistical methods, calculating the odds for a hypothesis, calculating the odds of a class from each of the existing attribute groups, and determining which class is the most optimal, known as the Bayes Theorem. The theorem is combined with Naïve where it is assumed that the conditions between attributes are mutually independent (Shukla & Naganna, 2014). Classification uses the Naïve Bayes Classifier based on the Bayes Theorem with Probability (B to A) equal to Probability (A and B) compared to Probability (A) based on equation 71[1].

$$P(H|X) = \frac{p(X|H)p(H)}{p(X)} \quad (1)$$

X represent data with an unknown class
H is a data hypothesis X which is a specific class
$P(H|X) =$ is the probability H based on condition X (*posteriori probability)*
$P(H) =$ is the probability H (prior probability)
$p(X|H)$ is the probability X based on the conditions in the hypothesis H
$p(X)$ is the probability of X

The classification process requires a number of clues to determine what class is suitable for the sample being analyzed. Therefore, Bayes' Theorem is adjusted to equation 2[1].

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \quad (2)$$

C is a class representation
$F1$ is a representation of the characteristics of the instructions for classification
P represent opportunity

The probability of a particular characteristic sample being included in class C (Posterior) is the chance that class C will appear (prior to the entry of the sample, it is called Prior), multiplied by the probability that the characteristics of the sample appear in class C (called Likelihood), divided by the probability that the characteristics of the sample appear globally (called Evidence). Therefore, the previous formula was written as equation 3[1].

p-ISSN: 1979-5661
e-ISSN: 2622-321X

$$Posterior = \frac{Likelihood * Prior}{Evidence} \quad (3)$$

The Evidence value is always fixed for each class in one sample. The value of the Posterior will be compared with the Posterior value of other classes, to determine what class a sample will be classified in [1].

3. Result and Discussion

The Naive Bayes Algorithm is a classification method using probability and statistics methods proposed by British scientist Thomas Bayes. The Naive Bayes algorithm predicts future opportunities based on prior experience and is known as Teorema Bayes. The main characteristic of Naïve Bayes Classifier is a very strong assumption (naive) of independence from each condition / event. The weighting is done by previous interviews with the company and the result is that the most important thing for the company is the number of users making transactions, namely frequency. Second is the amount of costs incurred by the user, namely monetary and finally the time span between the customer's last transaction and the current time, namely the recency. Table 1 is the result of RFM weighting.

Table 1. RFM Weighting Results

Number	Variable	Weight
1	*Recency*	0.2
2	*Frequency*	0.5
3	*Monetary*	0.3

The final probability is obtained by calculating the final probability value of the class. The final probability calculation is as follows:

$P(0 \mid X) = 0.3 * 5.293116711388974e^{-7}$
$P(0 \mid X) = 1.587935013416692e^{-7}$

We can also look at Table 2 for Probability result

Table 2. Probability result

Probability result			Cluster
0	1	2	
0.000000158794	0.000053303553	0.000000972157	1
0.000000689643	0.000011416772	0.000000015262	1
0.000000000000	0.000000000228	0.000006754702	2
0.000012335144	0.000000911465	0.000000000535	0
0.000000124629	0.000022276183	0.000002582014	1
0.000010278085	0.000000007746	0.000000000011	0

p-ISSN: 1979-5661
e-ISSN: 2622-321X

The Naive Bayes Accuracy Results is as table 3 below.

Table 3. Naive Bayes Accuracy Results

No	Class	Accuracy
1	0	92.31%
2	1	100%
3	2	100%

4. Conclusion

Based on the grouping of 200 users, the K-Means method shows that the optimum number of clusters is k = 3. Where the Sillhouette Index value at k-3 is 0.355. Furthermore, the results of the CLV calculation in each cluster show that cluster 1 (the second cluster) has the greatest CLV value, followed by clusters 0 and 2. The test results of the Naïve Bayes method show that the average accuracy is 97.44% where the accuracy in each class is 92.31% (Class 0), 100% (Class 1) and 100% (class 2). In this case, you could say that the system has a high level of accuracy for making user predictions based on the RFM variable and the segmentation results from clustering.

References

[1] ADRIYENDI, A. (2016). Prediksi Clustering, Calculation Dan Classification Fruit and Vegetable Consumption. Sainstek : Jurnal Sains Dan Teknologi, 7(2), 146. https://doi.org/10.31958/js.v7i2.135

[2] Dicky Nofriansyah, D. (2016). Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL (Studi Kasus di. Saintikom, 15(1978–6603), 81–92.

[3] Hardiani, T., Hartanto, R., & Mada, U. G. (2017). Segmentasi Nasabah Tabungan Menggunakan Model RFM (Recency , Frequency , Monetary) dan K-Means Pada Lembaga Keuangan Mikro. (May), 463–468.

[4] INFORMATIKALOGI. (2017). Algoritma Naive Bayes. Retrieved July 10, 2019, from https://informatikalogi.com/algoritma-naive-bayes/

[5] Kamila, I., Khairunnisa, U., & Mustakim. (2019). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau. Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi, 5(1), 119–125.

[6] Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc.

[7] Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, 1(12), 1750–1757. https://doi.org/10.1074/jbc.M209498200

[8] Santosa, S. (2017). Model Prediksi Pola Loyalitas Pelanggan Telekomunikasi Menggunakan Naive Bayes Dengan. 13, 154–169.

[9] Shofiani, N. (2017). Segmentasi Supplier Menggunakan Metode K- Means Clustering (Studi Kasus : Ptpn X Pg Meritjan).

p-ISSN: 1979-5661
e-ISSN: 2622-321X

[10] Shukla, S., & Naganna, S. (2014). A Review ON K-means DATA Clustering APPROACH. International Journal of Information & Computation Technology, 4(17), 1847–1860. Retrieved from http://www.irphouse.com

[11] Zamil, A. M. (2011). Customer relationship management: A strategy to sustain the organization's name and products in the customers' minds. European Journal of Social Sciences, 22(3), 451–459.