Comparison of Classification C4.5 Algorithms and Naïve Bayes Classifier in Determining Merchant Acceptance on Sponsorship Program

Normah¹, Ita Yulianti², Deny Novianti³, Monikka Nur Winnarto⁴, Ainun Zumarniansyah⁵, Safitri Linawati⁶

¹Teknik Informatika, STMIK Nusa Mandiri, Jakarta, Indonesia
²,³,⁴Sistem Informasi, ⁵Teknologi informasi, ⁶Teknologi Komputer, Universitas Bina Sarana Informatika, Jakarta, Indonesia

E-mail: normah.nor@nusamandiri.ac.id

Abstract. The large number of merchants that make sponsorship held by the Bank reaches thousands, data mining is used to classifying thousands of data. Naïve Bayes algorithm and C 4.5 are classification algorithms in data mining. The classification results are used as determinant where the merchant deserves to receive the sponsorship program, which potentially provides the source of funds and increase the brand awareness of the company by looking at the performance, transaction amount, total nominal, average daily transaction, average transaction nominal. Comparison results show that The C 4.5 algorithm is the best model for handling case of Merchant eligibility in the Sponsorship Program. This can be proved by looking at the level of accuracy generated on the testing and validation process of the model. Both models have the same AUC value but the C 4.5 algorithm produces a superior accuracy value with a difference of 0.45% compared to Naïve Bayes.

1. Introduction

At the moment the buying and selling transaction is changing from the choice of payment method. In order to make it easier to process the payment between buyers and sellers (goods / services) many parties present this non-cash payment option. Non-cash transactions themselves have several options ranging from ATM Cards, Credit Cards, E-Money, to digital money such as Go-Pay, Grab-Pay, OVO and Link Aja. Each has advantages that make it easier for people to make transactions without having to carry a lot of cash or withdraw cash from savings in advance just to make a transaction. For merchants they are not only facilitated and safe in accepting payments, but also the funds received from each transaction can be directly stored in a bank account and easy to monitor.

With the advancement in technology and information, the bank has made the payment service sector as one of the media to raise funds from the public aside from savings and as a source of revenue from fees for each transaction by becoming a provider of non-cash payment services. Banks are required to not only provide easy payment options for merchants but also increase transaction volumes for merchants who collaborate with them which will impact on increasing profits for the merchant, and on the side of the customer, the bank is also expected to provide...
additional benefits for customers when one of them is giving a discount. Sponsorship is a mutually beneficial agreement commercially covering the provision of resource funds, goods or services by individuals or entities (sponsors) to individuals or entities (rights owners) in exchange for a set of rights that can be used in communication activities [1]. But unfortunately, of the thousands of merchants who proposed cooperation in the sponsorship program, the determination was only based on the big name of the merchant, regardless of the volume of transactions generated by the merchant. Transaction data archives that have been neglected so far can be used as material for decision making if processed properly. In making sponsorship decisions, companies must have targets that sponsors must reach, coordinate sponsorship and measure their effects such as return on investment [2].

Data mining is an activity that includes collecting, using historical data to find regularity, patterns or relationships in large-sized data. Data Mining outputs can be used to assist in future decision making, [3]. Classification is a data processing technique that divides objects into several classes according to the number of classes desired, [4]. The C4.5 algorithm is one of the algorithms used to create decision trees and is an algorithm devoted to supervised learning [5]. Naïve Bayes is a learning algorithm which predicts the probability in the future based on past experience, used to calculate the odds for a hypothesis in statistics, calculates the probability of a class of each group attributes exist, and determine which one is the most optimal class [6]. Naïve Bayes classification is probability-based classification model is easy to use for discrete values, robust and effective in practice, has had a wide variety of applications in many complex domains such as text classification, document classification, and sentiment analysis [7].

Feature selection can make either good classification more efficient/effective by reducing the number of analyzed data or feature identification which is suitable to be considered in learning process, [8]. Filter method contains Document Frequency, Mutual Information, Information Gain, and Chi-Square. None of such four methods which is accepted broadly as the best feature selection method for sentiment classification or text categorization, however, Information Gain mostly more excellent compared to others filter methods, [8]. This study compares the accuracy generated from the two classification algorithms namely C4.5 Algorithm using Information Gain as a feature selection method with Naïve Bayes.

The scientific contribution in this research is that thousands of merchant data enrolling in sponsorship programs can be processed using data mining, and classified by volume of financial transactions, which can then be a reference in determining which merchant deserves to be sponsored. So the determination of merchant Acceptance for sponsorship program is no longer done manually and subjectively, just looking at the big name of merchant.

2. Related Work
C4.5 algorithm is use to find the pattern of linkage between soft competencies and teacher performance assessed based on hard competencies aspect. The data abtained from teacher result assessment and teacher performance appraisal. Gain Ratio for 10 attributes is Achievement Orientation (AO) 0.22, Planing and Organizing (PnO) 0.21, Concern for Order (CFO) 0.20, Customer Service Orientation (CSO) 0.20, Analitycal Thinking (AT) 0.19, Conceptual Thinking (CT) 0.18, Organizational Commitment (OC) 0.16, Communication (C) 0.15, Flexibility (F) 0.13, Interpersonal Understanding (IU) 0.09. Classification patterns that generated from C4.5 algorithm, can be known the accuracy levels by using X-Cross Validation technique. The generated accuracy is 70.83%, [9].

In the research analysis of the sentiment of football players, conducted comparisons of accuracy results against SVM and NBC methods, as well as implementing optimizations with the PSO algorithm. Calculation results of the SVM method get an accuracy value of 78.50% and AUC value 0.893 while the calculation result SVM (PSO) Get the accuracy value of 84.00% and the value of AUC 0.914 meaning there is a increase in the accuracy rate of 5.50% and the increase
in the AUC value of 0.021. The calculation of the NB method results in an accuracy value of 76.50% and AUC value 0.633, while the calculation result of NB (PSO) Gets the accuracy value 83.00% and the value of AUC 0.653 means that there is an increase in the accuracy of 6.50% and the increase in AUC value of 0.020. The accuracy of SVM and NB has a difference in value of about 2.00%, while SVM (PSO) and NB (PSO) have a difference in value of about 1.00%. The results obtained from data testing accuracy SVM (PSO) get higher accuracy compared to SVM, NB and NB (PSO) [10].

Classification model in the prediction of the success of reuse software is more accurate Using C4.5 algorithm compared With Naïve Bayes. It can be From the accuracy value and the AUC Generated for algorithm model C4.5 of 95% with AUC value 0.900 While the accuracy of the model Naïve Bayes of 91.67% with AUC 0.750 [11].

3. Research Methodology
In this study, the data used are primary data obtained from the results of data collection including observation techniques, conducting interviews with Junior Managers from the MCH (Merchant Chain) and MRT (Merchant Retail) teams related to transaction data and sponsorship programs offered, and studies literature by finding references from various sources of literature including books, scientific journals and the internet, while the Model implemented in this study is the Cross Industry Standard Process for Data Mining (CRISP-DM) model consisting of:

- **Bussines Understanding**
  In order to determine the suitable of the merchant in the sponsorship program, a comparison of C4.5 and Naïve Bayes algorithm models is performed to find the best algorithm model that produces the highest and the most optimal prediction level.

- **Data Understanding**
  The data used are primary data sourced from one of the Junior Managers of the MCH (Merchant Chain) and MRT (Merchant Retail) team at Bank Rakyat Indonesia. The data has a total of 4,397 data with attributes consisting of transactions amount, total nominal, average daily transactions and average nominal sales transactions.

- **Data Preparation**
  At this stage, select data is performed to simplify the next process by reconsidering which attributes will be used. The data that has been obtained previously is then converted from numerical types to categorical. The final step in data preparation is the process of cleaning data by launching detected noise, identifying outliers and correcting data inconsistencies.

- **Modelling**
  Modeling is carried out by implementing the C4.5 classification algorithm and Naïve Bayes as a comparison algorithm in determining Merchant Eligibility in the Sponsorship Program.

- **Evaluation**
  Evaluation is done using the Confusion Matrix and ROC curve. ROC curves are used to measure the AUC was calculated to measure the difference performances method was used. the larger the area under the curve, the better the prediction results, Witten & Frank [12].

- **Deployment**
  At the Deployment stage, the knowledge presentation process are perfome on the model produced for making the program in further research.

The novelty element in this research is by comparing the two classification algorithms, namely C4.5 Algorithm and Naïve Bayes using the latest data sources from BRI banks, this research is very useful in recommending merchants who are entitled to receive sponsorship programs, especially when digital payments are very widespread such as at present, by looking at several
criteria containing performance, transaction amount, total nominal, average daily transaction, nominal average transaction.

4. Result and Discussion
Datasets used in this research are primary data that has been processed based on the results of interviews and observations with experts in the field of sponsorship program in the selection of merchants. The Dataset has a total of 4,397 records obtained from the results of sales transactions across all Indonesian merchants with a period of time from 01 to 30 April 2019. Each record contains 8 attributes and 2 types of classes consisting of decent and unworthy. In this research, data preprocessing and implementation of the model is done using Rapidminer Software.

4.1. C4.5 Algorithm
The application of C4.5 Algorithm classification model in the RapidMiner Software is done by selecting the Decision Tree operator and the information gain criteria to produce the decision tree presented in the Figure 1 below:

![Figure 1. Results of The Decision Tree](image)

Based on the Figure 1 above, the Daily Transaction Average attribute becomes the root in determining Merchant Eligibility in the Sponsorship Program. Root is obtained based on the highest gain in each attribute. The calculation of the highest gain will continue until all cases in the branch have the same class so that a decision tree is formed.

Not only that, experiments on this model also resulted in an accuracy value of 99.95% which is explained in full with the details of class precision and recall in the following figure:

![Figure 2. Accuracy Results of C4.5 Algorithm Classification Model](image)
Evaluation results obtained from processing the ROC curve are proved that the C4.5 algorithm classification model generates AUC value of 1.000 and belongs to Excellent Classification category.

4.2. Naïve Bayes

Different from the C4.5 Algorithm model, the Naïve Bayes classification model does not produce a decision tree but a probability value. In RapidMiner Software, probability values for this research dataset can be seen in the simple distribution table as follows:

![SimpleDistribution](image)

**Figure 3.** Simple Distribution results on Naïve Bayes classification Model

The results of the probability value from the simple distribution table obtained are used as a determining parameter in the Naïve Bayes model to predict the Merchant Eligibility in the Sponsorship Program by using a new dataset. If the probability value generated from the calculation of the record is 0.979 then the prediction result is "Unsuitable", but conversely if the probability value is 0.021 then the prediction result is "Suitable" or selected as a Merchant in the Sponsorship Program.

Similarly, in the previous modeling, in the Naïve Bayes classification model is also implemented Confusion Matrix method to test and evaluate the model shown in the figure below:

![ConfusionMatrix](image)

**Figure 4.** Accuracy Results of Naïve Bayes Classification Model

In addition, evaluation on the model Naïve Bayes also carried out using the ROC curve expressed with a value of AUC of 1.000. It clearly shows that the modeling is also included in the Excellent Classification category.

4.3. Comparison Results

At this stage, the process of comparison accuracy is done based on the results of experiments obtained in the implementation of the algorithm classification model C4.5 and Naïve Bayes presented in the table below:
Table 1. Results of The Comparison Model Classification Algorithms C4.5 and Naïve Bayes Classifier

|                  | Accuracy  | AUC     |
|------------------|-----------|---------|
| C4.5 Algorithm   | 99.95%    | 1.000   |
| Naïve Bayes Classifier | 99.50%    | 1.000   |

5. Conclusion

The C4.5 algorithm is the best model for handling case studies of Merchant eligibility in the Sponsorship Program. This can be proved by looking at the level of accuracy generated on the testing and validation process of the model. Although both models have the same AUC value but the 4.5 algorithm produces a superior accuracy value with a difference of 0.45% compared to Naïve Bayes. The results of this research can be further developed in future studies, such as making a decision support system program to determine the acceptance of merchants in the sponsorship program, similar research can be done by experimenting using optimization techniques and trying to use other algorithms such as PSO, Genetic Algorithms, KNN, etc. Or use more complex primary data such as using more than one bank.

References

[1] Masterman G 2011 Sponsorship for a return on investment 2nd ed. New York: Routledge.
[2] Hans Mathias Thijmoœ, Erik L. Olson P S B, Decision-making Processes Surrounding Sponsorship Activities J. Advert. Res. 42, 6 p. 6–5.
[3] Nofitri R and Juna Eska 2018 Seminar Nasional Royal (SENAR) 2018 9986, September.
[4] Arifin M F and Fitrianah D 2017 Rekomendasi Penerimaan Mitra Penjualan Studi Kasus PT Atria Artha Persada, December.
[5] Algorithm Candra B P Rachmat E and Hidayat S 2018 Decision Support System for Stock Prediction and Supplier Selection Using Least Square 2018 3rd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. p. 241–246.
[6] Normah N 2019 Naïve Bayes Algorithm For Sentiment Analysis Windows Phone Store Application Reviews SinkrOn 3, 2 p. 13.
[7] Bagui S Devulapalli K and John S 2020 MapReduce Implementation of a Multinomial and Mixed Naive Bayes Classifier J. Intell. Inf. Technol. 16, 2 p. 23.
[8] Muthia D A 2014 Sentiment Analysis of Hotel Review Using Naïve Bayes Algorithm and Integration of Information Gain and Int. Semin. Sci. Issues Trends p. 25–30.
[9] Kadafi A and Gata W 2019 Particle Swarm Optimization Based C4.5 for Teacher Performance Classification p. 1–12.
[10] Hasan F N and Wahyudi M 2018 ANALISIS SENTIMEN ARTIKEL BERITA TOKOH SEPAK BOLA DUNIA MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE DAN NAIVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION Akrab Juara 3, 4 p. 42–55.
[11] Yulianti I Informatika B S and Reuse S, 2019 ANALISIS KOMPARASI KLASIFIKASI ALGORITMA C4.5 DAN NAIVE 7, 1.
[12] Kristiyanti D A Normah and Umam A H 2020 Prediction of Indonesia Presidential Election Results for the 2019-2024 Period Using Twitter Sentiment Analysis p. 36–42.