Bit-Tech

Vol.1, No.2, Desember 2018
Available online at: http://jurnal.kdi.or.id/index.php/bt

Comparison of C4.5 Algorithm, Naive Bayes and Support Vector Machine (SVM) in Predicting Customers that Potentially Open Deposits

Yusuf Kurnia[1], Kuera Kusuma[2)]

[1) 2)]*Buddhi Dharma University*
Jl.Imam Bonjol No. 41 Karawaci Ilir, Tangerang, Indonesia
[1)]yusufkurnia.iuz@gmail.com
[2)] kuera.kusuma@gmail.com

Article history:

Received 01 December 2018;
Revised 4 December 2018;
Accepted 14 December 2018;
Available online 19 December 2018

Keywords:

Data Mining
C4.5
Naive Bayes
Support Vector Machine (SVM)
Supporting Application
Potential Customers

Abstract

This research is based on the application of data mining processing to produce information that is useful in helping decision making. In this study aims to determine the superior algorithm between C4.5, Naive Bayes and SVM algorithms in predicting which customers who have high potential to open deposits. The data used in this study is secondary data where its data is obtained from the UCI dataset. The comparison results of the accuracy value of C4.5 Algorithm 90.57%, accuracy of Naive Bayes 87.70% and SVM 89.29%. Based on the results of the comparison of accuracy values, it is found that the C4.5 algorithm has the highest level of accuracy. So that the application of supporting applications to predict customers who have the potential to open deposits uses the rules for establishing C4.5 data processing.

I. INTRODUCTION

Data Mining is a series of processes to explore any added values in the form of information that is not known manually from a database by extracting patterns from data with the aim of manipulating the data into valuable information obtained by extracting and recognizing important patterns or draw from data contained in the database. Data mining is also known by other names such as: Knowledge discovery (mining) in databases (KDD), extraction of knowledge (knowledge extraction) Analysis of data/patterns and business intelligence (business intelligence) and is an important tool for manipulating data for presenting information as needed user with the aim to assist in analyzing the collection of behavioral observations.

In general, the definition of data-mining can be interpreted as a process of finding interesting patterns of data which stored in large numbers, extraction from a useful or interesting information (non-trivial, implicit, as long as the potential uses are unknown) patterns or knowledge of stored data in large sums, exploration of the analysis automatically or semi-automatically on large amounts of data to find meaningful patterns and rules.

Data mining algorithm C4.5 is the algorithm that is most often used to process data mining, in this study, the authors would like to prove whether the C4.5 algorithm is superior to other algorithms for processing data mining by comparing accuracy and errors using the comparative algorithm Naive Bayes and SVM by using bank marketing dataset.

II. METHODS

Decision tree is a method that commonly used to classify data mining. As explained earlier, classification is a technique of finding a collection of patterns or functions that describe and separate one class of data from another to state that the object belongs to a certain category by looking at the behavior and attributes of the group that has been defined. This method is popular because it is able to classify and shows the relationship between attributes. A lot of algorithms can be used to build a decision tree, one of which is the C45 algorithm[1].

The C4.5 Algorithm is able to handle numerical and discrete data. C.45 Algorithm uses the gain of ratio. Before calculating the acquisition ratio, it is necessary to calculate the value of information in units of bits from a collection of objects, namely by using the concept of entropy.

ISSN 2622-2728 (online) 2622-271X (print)
doi: https://doi.org/10.32877/bt.v1i2.46

Entropy(S) is the estimated number of bits that is needed to be able to extract a class (+ or -) from a random amount of data in the sample space S. Entropy can be said as a bit requirement to express a class. The smaller the Entropy value, the more Entropy is used to extract a class. Entropy is used to measure the authenticity of S.

$$\boldsymbol{Entropy(S) \equiv -p_{\oplus}\, log_2\, p_{\oplus} - p_{\ominus}\, log_2 p_{\ominus}}$$

S: space (data) sample used for training.
P+: the number of positive or supportive solutions in the sample data for certain criteria.
P-: the number of negative solutes or does not support the sample data for certain criteria.
Entropi(S) = 0, if all examples in S are in the same class.
Entropi(S) = 1, if the number of positive and negative examples in S is the same.
0 > Entropi(S) > 1, if the number of positive and negative examples in S is not the same

Gain (S, A) is the acquisition of information from attribute A relative to the output of S data. Information obtained from output data or dependent variable S grouped by attribute A, denoted by gain (S, A).

$$\boldsymbol{Gain(S,A) \equiv Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)}$$

A: attribute.
S: Sample
n: Number of attributes set of A's attributes.
|Si|: Number of samples in perti to -i.
|S|: Number of sample in S

Bayes classification is also known as Naïve Bayes, has capabilities and comparable to decision trees and neural networks. Bayes Classification is the classification of statistics that can be used to predict the probability of membership of a class[2] Naïve Bayes can use kernel density estimators, which improve the performance if the assumption of normality is very incorrect, but also able to handle numeric attributes using supervised discretization. The Naïve Bayes (NB) technique is one simple form of Bayesian that is networked for classification. A Bayes network can be seen as directed as a table with a combined probability distribution of more than one discrete set and a stochastic variable. The following are the steps of bank marketing data processing with the naïve bayes method and RapidMiner tool, at the initial stage of the Read Excel module that contains excel data marketing data linked to the validation module, in the validation module there is a Baive Bayes module that is associated with the Apply Model module and Performance

Naive Bayes is the classification of statistics that can be used to predict the probability of membership of a class. Naive Bayes is based on the Bayes theorem which has classification capabilities similar to decision trees and neural networks. Naive Bayes proved to have high accuracy and speed when applied to databases with large data.

$$P\ H|X\ = \frac{P(X|H)xP(H)}{P(X)}$$

Bayes predictions are based on the Bayes theorem formula with formulas Meanwhile Naïve Bayes formula for classifying:

$$P\ Y|X\ = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)}$$

SVM (Support Vector Machine) main idea[3] is maximizing the hyperplane boundary, which is illustrated in the figure below, in the first picture there are a number of possible hyperplane options for the data set, while in the second picture is the hyperplane with the maximum margin. Even though the first image can actually use an arbitrary hyperplane, the maximum margin will give a better generalization to the classification method.

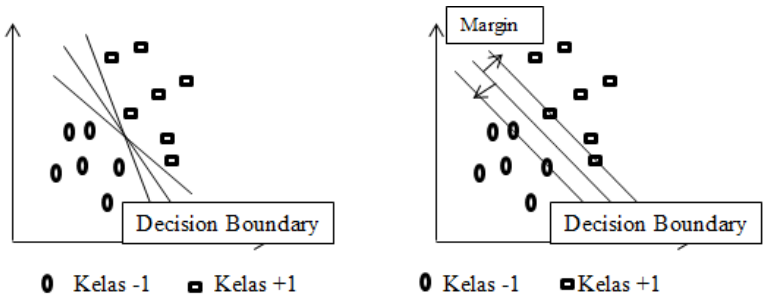

Fig 1. Hyperplane Metode SVM (Support Vector Machine)

III. RESULTS

Background to the use of Data Mining

Why do researchers use Decition Tree / C4.5 Algorithm is because the algorithm has been studied by examiners during the lecture period and the researchers also would like to prove whether the C4.5 algorithm to has higher accuracy than the Naive Bayes and SVM (Support Vector Machine) algorithms, and in the journal Application of C4.5 Algorithm to Predict Prospective New Employee Acceptance at PT WISE in this study, the C4.5 algorithm has 71% accuracy to predict the search for new recruits, so that testers want to prove whether the accuracy of the C4.5 algorithm can be superior for predicting customers who have the potential to open deposit accounts.

Business Understanding

To increase the customers who have the potential to open deposit deposits by designing an application that can predict customers who have the potential to open deposits.

Data Understanding

The data used in this study is secondary data, namely data that has been collected previously so that it no longer goes through the stages of retrieving data directly to the customer. Secondary data that us used in this study came from a web data provider that was demolished for further data processing purposes, which was taken from the uci repository site with the title of bank marketing data with the address. The data consists of Number of Instances: 45211 for bank-full.csv (4521 for bank.csv) and sixteen attributes and one attribute result, the data obtained from the bank marketing dataset is data that can be directly processed because the data has been provided by S. Moro, P. Cortez and P. Rita for data mining.

Data Preparation

The data preparation, the used data to conduct research are secondary data downloaded from public datasets that have a total of 45211 data and 4211 data in CSV format (Comma Separated Value) and have 17 attributes and 2 clasifications, and attributes used for processing data in rapid miner, for the formation of rules, and program design as many as 17 according to what was printed on the dataset that was downloaded

.age , job , martial status, education, default, balance, housing , loan, contact, day, last call based on day, month - last call based on month, duration - last call based on minutes, campaing – nome telephone number, pdays – days after direct marketing , previous – days before direct marketing, poutcome – last marketing's result dan y – current result of the marketing.

Modeling

Modeling is a phase that directly involves data mining techniques. Selection of data mining techniques, algorithms and determining parameters with optimal values, the following is a model to determine the algorithm to be used for design.

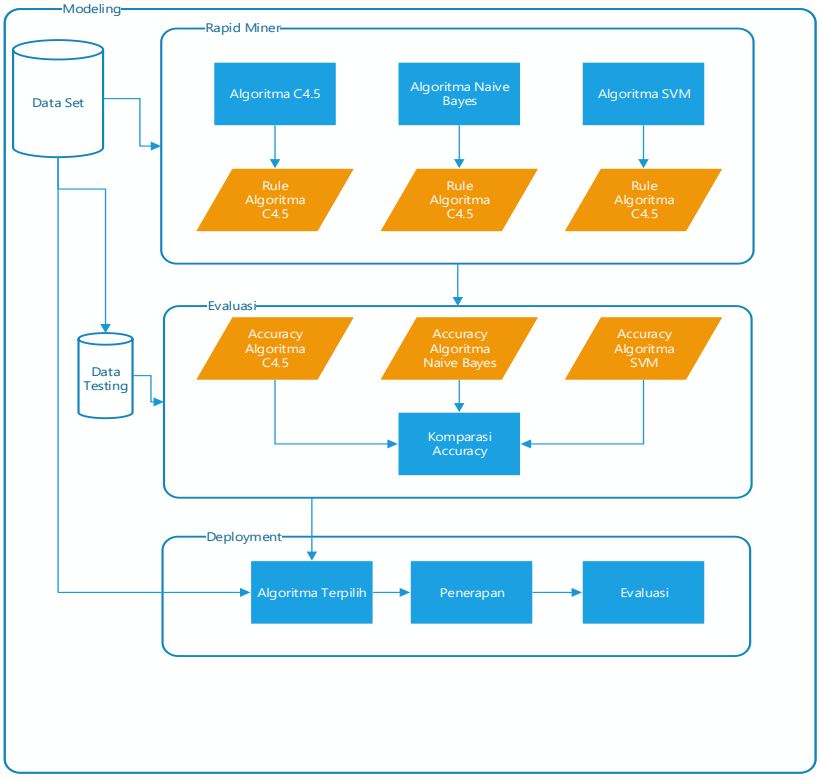

Fig 2. Modelling process

Evaluation

Evaluation is the phase of interpretation of the results of data mining. Evaluation is carried out in depth with the aim that the results in the modeling phase are in accordance with the goals to be achieved in the stage of the business understanding.

Evaluate Results

This stage is the comparative accuracy of the C4.5Algorithm, Naive Bayes, Support Vector Machine (SVM), to find out which of the three algorithms has the best accuracy.

Deployment

Deployment is the stage of making reports on the results of data mining activities. Final report regarding knowledge gained or pattern recognition on data in the process of data mining and presented in the form of graphs or descriptions that are easy to understand.

Validity evaluation

Validaty evaluation is a step to compare accuracy and error to three algorithms and to find the best algorithm that will be used to form proof applications, using data used by testers.

accuracy: 90.57%

	true no	true yes	class precision
pred. no	38717	3058	92.68%
pred. yes	1205	2231	64.93%
class recall	96.98%	42.18%	

Fig 3. Accuracy Algoritma C4.5

The picture above is a picture of accuracy results generated by rapidminer for the C4.5 algorithm and now manual calculations will be performed to find the accuracy value of the C4.5 algorithm.

Accuracy = (38717 + 2231) / (38717 + 2231 + 3058 + 1205) * 100
= 90,57%

accuracy: 87.70%

	true no	true yes	class precision
pred. no	36846	2484	93.68%
pred. yes	3076	2805	47.70%
class recall	92.29%	53.03%	

Fig 4. Accuracy Algoritma Naive Bayes

The picture above is a picture of accuracy results generated by rapidminer for the NaiveBayes algorithm and now manual calculations will be performed to find the accuracy value of the NaiveBayes algorithm.

Accuracy = (36846 + 2805) / (36846 + 2805 + 2484 + 3076) * 100
= 87,70%

accuracy: 89.29%

	true no	true yes	class precision
pred. no	39389	4311	90.14%
pred. yes	533	978	64.73%
class recall	98.66%	18.49%	

Fig 5. Accuracy Algoritma SVM

The picture above is a picture of accuracy results generated by rapidminer for the NaiveBayes algorithm and now manual calculations will be performed to find the accuracy value of the NaiveBayes algorithm.

Accuracy = (39389 + 978) / (39389 + 978 + 4311 + 533) * 100
= 89,29%

Comparation Table

This comparison table will compare 3 algorithms previously explained by using data used by testers.

Table 1. Algoritma C4.5 , Naive Bayes, dan SVM comparison table

	Naive Bayes	**SVM**	**C4,5**
Accuracy	87,70 %	89,29 %	90,57 %
Data	45211	45211	45211
Program	Rapid Miner	Rapid Miner	Rapid Miner

Based on the comparison of C4.5, Naive Bayes, and SVM Algorithms, it can be seen that the Naive Bayes algorithm has an accuracy rate of 87.70%, the SVM algorithm has an accuracy rate of 89.29% and the C4.5 algorithm has an accuracy rate of 90.57%. Then the C4.5 algorithm is the winner that will be used in the application design.

Designed applications uses rapidminer to get the C4.5 algorithm rules, Visual Studio Community 2017 as a tool for designing applications to be created, MySql as a database that will be used for the application created. Set Role this operator is used to change the role of one or more attributes, in the picture above the role set is used to determine the attributes that are the reference for the results of the data used, and to change the role of an attribute. Label: Is a special role, which acts as a target attribute for learning operators for example. Operator Decision Tree. Labels are also called 'destination variables'.

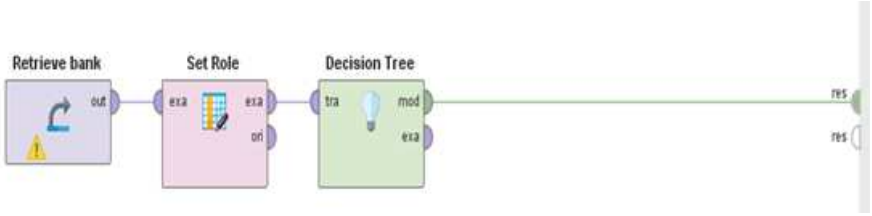

Fig 3. Retrieve Bank Set Role dan Decision Tree operator

After all operators are connected and then press the button, the process waits for a few moments, the rapidminer will display the decision tree in the view result.

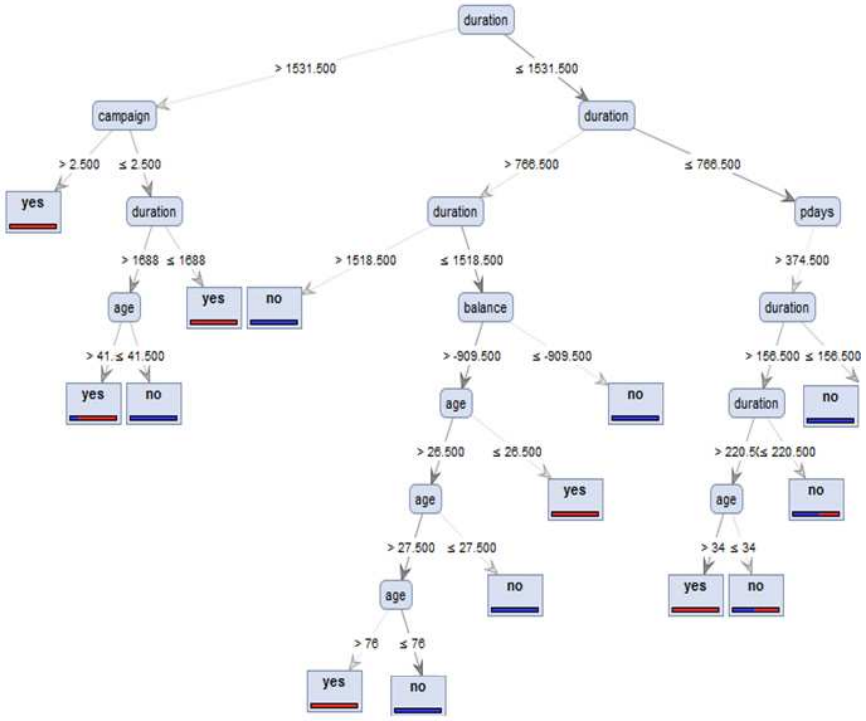

Fig 4. Decision tree

After knowing the process of forming the data processing process carried out by the C4.5 algorithm, the data processing process flow is poured into the programming language to produce applications that aim to facilitate predicting customers who have the potential to make deposits. The application used to design potential customer prediction applications is Visual Studio 2015 and uses the sqlserver 2008 database application. Following is the display of potential customer prediction applications.

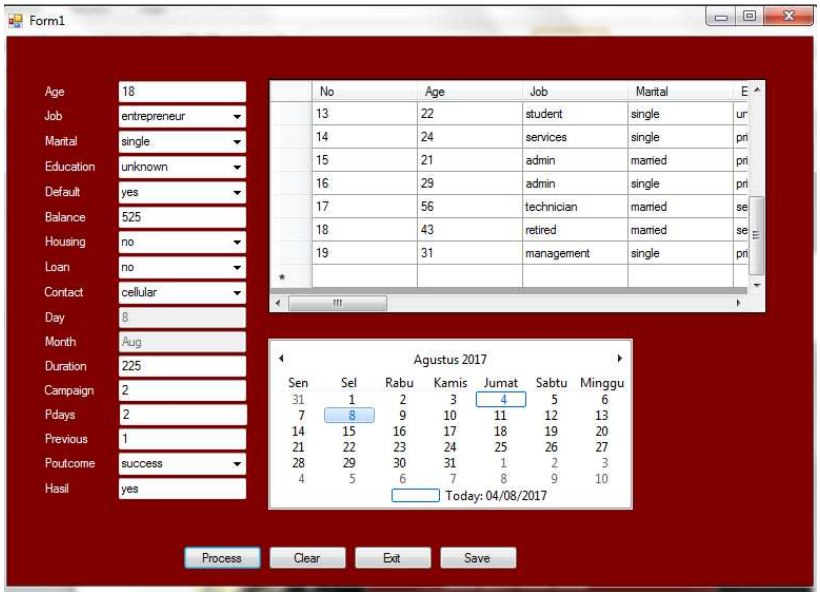

Fig 5. Display of Potential Customer Prediction Application

Flowchart[4] is a systematic presentation of the process and logic of information handling activities or graphical depictions of the steps and sequence of procedures of a program. Flowchart is a chart (chart) that shows flow in a program or system procedure logically. Flowchart is used mainly for communication aids and for documentation.

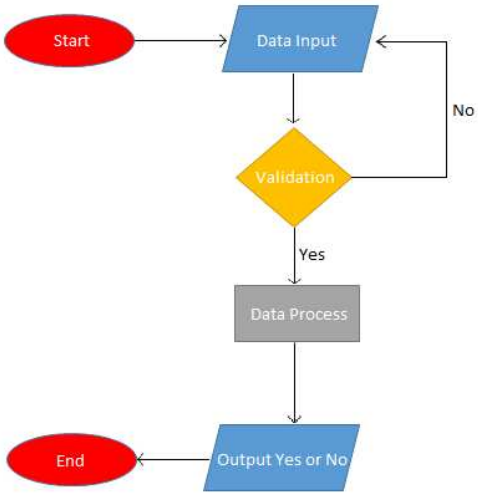

Fig 6. Potential customers prediction flowchart

IV. CONCLUSIONS

Based on the testing and evaluation that has been done, it can be concluded that the C4.5 Algorithm is superior compared to the Naive Bayes algorithm and SVM (Support Vector Machine) algorithm, C4.5 Algorithm can still be used to make predictions and the Rules used for programming are algorithms C4.5.

REFERENCES

[1] Dennis Aprilla C, Donny Aji Baskoro, Lia Ambarwati, I Wayan Simri Wicaksana (2013), Belajar Data Mining Dengan RapidMiner. Jakarta, pp 40-45.

[2] Kusrini, Andri Koniyo (2007), Tuntunan Praktis Membangun Sistem Informasi Akuntansi dengan Visual Basic dan Microsoft SQL Server. Andi Publisher: Jakarta, pp 28-32.

[3] Eko Prasetyo, Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. Penerbit ANDI, pp. 50-71.

Muslim Setyo Rejeki, Ali Tarmuji, 2013, MEMBANGUN APLIKASI AUTOGENERATE SCRIPT KE FLOWCHART UNTUK MENDUKUNG BUSINESS PROCESS REENGINEERING, Volume 1 Nomor 2, Oktober 2013, e-ISSN: 2338-5197, pp 2-3.