RESEARCH ARTICLE OPEN ACCESS

Analysis Comparison of The Classification Data Mining Method to Predictthe Decisions of Potential Customer Insurance

Reza Avrizal[1], Arief Wibowo[2], Angger Styo Yuniarti[3], Deassy Ari Sandy[4], Kamal Prihandani[5]

[1]Faculty of Computer Science, Indraprasta PGRI University, Indonesia
[2345]Faculty of Computer Science, Budi Luhur University, Indonesia
E-mail :
[1]reza.avrizal@unindra.ac.id,[2]arief.wibowo@budiluhur.ac.id,[3]anggeryuniarti06@gmail.com,[4]deassy984@gmail.com, [5]k.prihandani@gmail.com

--*************************----------------------------------

Abstract:

Current insurance business increasingly developed due to the increasing public awareness to insure and provide protection against various aspects of his life. Insurance service companies whose customers come from bank customers are very concerned about the quality of service to customers. Constraints the companies are difficulties in determining potential customers. Buy or No if the company could identify levels-levels to determine the potential value of the customer so the customer data can be classified. As for some of the criteria that are considered important in determining the potential client that is based on area, age and job. For it to be developed an application of data mining to determine criteria for the customer. Data mining techniques applied is Classification while the method used is the classification Decision Tree (decision tree) and support vector machine (SVM). There are three parameters of the test were used as evaluation system i.e. accuracy, precision and recall. From the results of the comparisons, the decision tree has a higher percentage than the SVM i.e. 86.37% accuracy, precision and recall 86.25% 90.14%. ROC curve and decision tree classification of diagnosis rate is excellent whereas the SVM good classification.

Keywords — **Insurance, Data Mining, Decision Tree, C45, Support Vector Machine.**

--*************************----------------------------------

I. INTRODUCTION

The availability of lots of data and the need for information or knowledge as a support for decision making to make business solutions and infrastructure support in the field of informatics engineering are the forerunners of the birth of data mining technology. Data mining techniques is expected to help speed up the decision-making process, enabling companies to manage the information contained in customer data and become a new knowledge.

The application of data mining can help to analyze data obtained from transactions in information systems so that they can explore patterns that can be used as new knowledge for the process of consumer identification for companies. Data mining is the process of finding interesting and hidden patterns from a large collection of data stored in a database, data warehouse, or other data storage area. Data mining is also defined as part of the process of extracting knowledge in a database known as the Knowledge Discovery in Database (KDD).

The big companies in insurance services have other ways to market their products in addition to the conventional methods that we know so far, namely using telemarketing services. Telemarketing is remote marketing that uses telecommunications technology as part of a regular and structured marketing program.

According to [1], telemarketing helps companies to increase revenue, reduce sales costs and increase customer satisfaction. Offering insurance products through telemarketing services must also be supported by good data, good data plays a major role for a customer to follow or take products offered by the agent. The results of the data mining process that will be the basis of whether the customer has the potential to buy insurance or not.

The data used is data that will be offered only 1 (one) type of product. In this research will discuss about comparing the accuracy of the C45 algorithm with Support Vector Machine to predict potential customers by using Rapid Miner tools.

Three test parameters that are used as a system evaluation are Precision, Recall and F-measure, with an average result of each test parameter is above 75% [2]. Research [3] discusses the selection of work partners of transportation service providers using the C4.5, Naïve Bayes and Neural Network methods. Of the three methods obtained the Naïve Bayes method can provide the best solutions to choose the best partners that provide transportation services.

The purpose of this research is to improve accuracy by comparing C4.5 method with Support Vector Machine to predict potential customer data and produce algorithms with more accurate accuracy to predict sales, so that the next research can more improve its accuracy.

II. THEORETICAL BASIS

2.1. Data Mining

Data mining is finding information from a large amount of data. In other expert, data mining is the process of finding interesting knowledge from a large amount of data stored in a database, data warehouse, or other storage media [4]. It can be concluded that data mining is a process of finding information from large amounts of data stored in storage media.

Data mining can be created due to the desire to seek knowledge or information from data stored in big data. The so-called storage media include relational databases, data warehouses, transactional databases, advanced database systems, flat files, data streams, and the World Wide Web.

The Data mining method is diverse, according to [5] the 10's best method for data mining are the C4.5 method, k-means, support vector machine, priori, EM, PageRank, AdaBoost, KNN (K-Nearest Neighbor), Naive Bayes, CART (Classification and Regression Tree).

Whereas for each method adjusts to the function of data mining, i.e. the characterization function, discrimination function, association function, prediction function, classification function, and cluster function [4].

2.2. Decision Tree

One of the most popular classification techniques used in data mining processes is the decision tree [6]. Decision tree is a classification method that uses tree representation where each node represents an attribute, the branch represents the value of the attribute, and the leaf represents the class. The top node of the decision tree is called root.

The decision tree algorithm is based on a divide-and-conquer approach to classifying a problem.The algorithm works from top to bottom, looking at each stage of the attribute to divide it into the best part of the class, and recursively process the sub-problems resulting from the division. This strategy produces a decision tree that can be converted into a set of classification rules [7]. Data in decision trees are usually expressed in the form of tables with attributes and records. Attributes state a parameter that is made as a criterion in tree formation. Example, to determine playing tennis, the criteria to be considered are weather, wind and temperature.

2.3. C4.5

C4.5 is a decision tree which is a very strong and thick classification and prediction method. The decision tree method transforms a very large fact into a decision tree that represents the rules. Decision trees are also useful for exploring data, finding hidden relationships between a number of prospective input variables with a target variable. Because the decision tree combines data exploration and modeling, it is very good as a first step in the modeling process.

2.4. Support Vector Machine

Classification is one important task, in the classification of a classifier is made from a set of training data with classes that have been determined previously. The SVM classification is also known as Naïve Bayes, has the ability to be comparable to decision trees and neural networks [4].

SVM is easily, so that users who do not have expertise in the field of classification technology can understand it. Bayes Classification is a statistical classification that can be used to predict the probability of membership of a class [8].

The use of SVM in the Naïve Bayes algorithm is by combining prior probability and conditional probability in a formula that can be used to calculate the probability of each possible classification [9].

III. METHODOLOGY AND RESEARCH DESIGN

3.1. Research Method

According to [5] there are four research methods commonly used namely research, experiment, case studies and surveys. In the context of this study using experiments, namely a method carried out by referring to problem solving which includes collecting data, formulating hypotheses, testing hypotheses, interpreting results, and conclusions [6].

3.2. Data Collection

The author makes direct observations in the telemarketing division to collect data related to the prediction of sales of insurance products by observing and systematically recording the problems that are investigated and examined directly on the object to be studied.

3.3. Analysis Technique

A descriptive analysis of the techniques performed to analyze the data that will be made against results of data collection with the study of literature, interviews and observations in order to get the specification needs a system that will be developed. The analysis techniques will be performed using the methods of data mining, among others, Decision Tree and Support Vector Model.

3.4. Classification Process Design

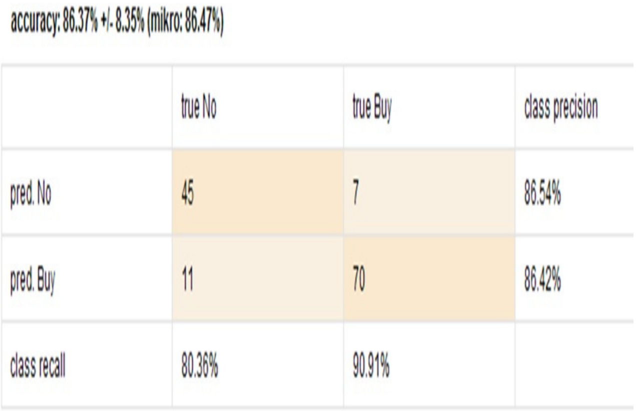
The process will be designed in a prototype of the system includes:

1. Import the excel data
   Import data done to enter data that will be predicted into the prototype will be designed. The format of the data in the form of a .csv or .xls.
2. Preprocessing
   Data import will check the entire contents by prototype to find out the feasibility of in processing. The process of checking may be checking the missing value, data format differences, and others.
3. The process of prediction
   After the data is clean then it will do the predictions against the test data using the methods with the best accuracy values have gone through the stages in the process of comparison analysis on research. The results of the predictions in the form of relevance between the field work to be taken and the fields have reached students, in a notation declared in the prototype of the Yes and No.

IV. RESULT AND DISCUSSION

4.1. Confusion Matrix of C4.5

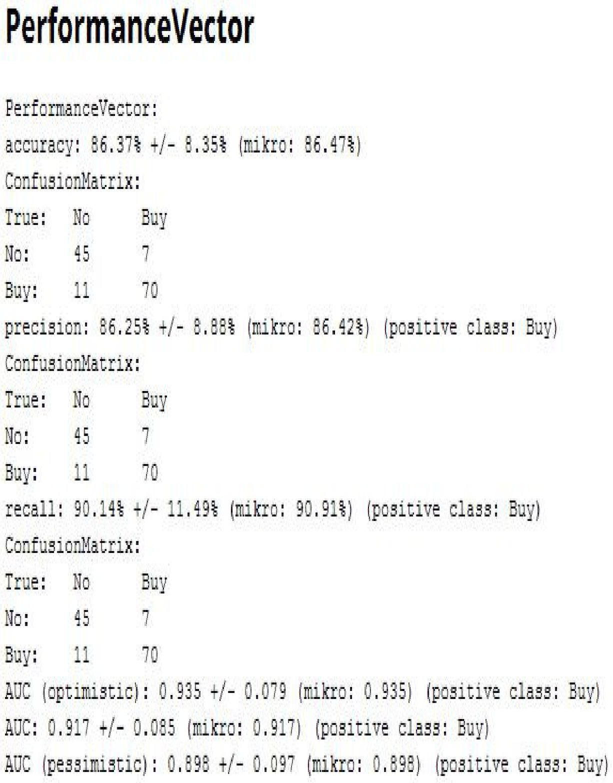
Figure 1 is a calculation of the accuracy of data training using an algorithm that yields 4.5 C 86.37% accuracy. Note data training consists of 133 record data, classified data NO. 56 and 77 data predicted Buy insurance.



accuracy: 86.37% +/- 8.35% (mikro: 86.47%)

	true No	true Buy	class precision
pred. No	45	7	86.54%
pred. Buy	11	70	86.42%
class recall	80.36%	90.91%	

Fig. 1. Confusion Matrix of C4.5



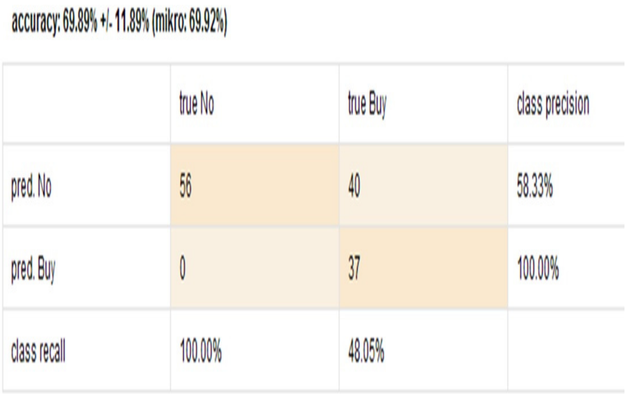
```
PerformanceVector

PerformanceVector:
accuracy: 86.37% +/- 8.35% (mikro: 86.47%)
ConfusionMatrix:
True:   No      Buy
No:     45      7
Buy:    11      70
precision: 86.25% +/- 8.88% (mikro: 86.42%) (positive class: Buy)
ConfusionMatrix:
True:   No      Buy
No:     45      7
Buy:    11      70
recall: 90.14% +/- 11.49% (mikro: 90.91%) (positive class: Buy)
ConfusionMatrix:
True:   No      Buy
No:     45      7
Buy:    11      70
AUC (optimistic): 0.935 +/- 0.079 (mikro: 0.935) (positive class: Buy)
AUC: 0.917 +/- 0.085 (mikro: 0.917) (positive class: Buy)
AUC (pessimistic): 0.898 +/- 0.097 (mikro: 0.898) (positive class: Buy)
```

Fig. 2.Text view Confusion Matrix of C4.5

4.2. *Confusion Matrix* of SVM

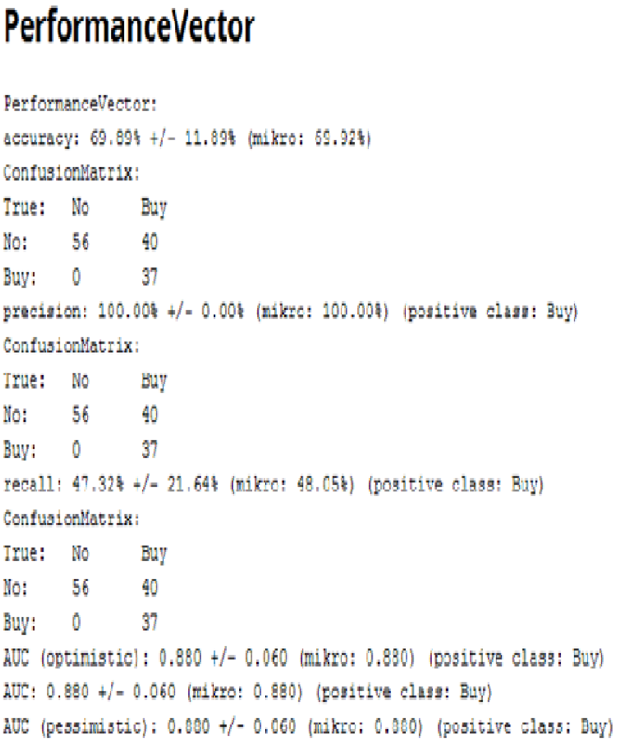
Below on figure 3 is the result calculation of the value of the confusion matrix against thealgorithms SVM with 133 records which resulted 69.89.24% accuracy rate.



accuracy: 69.89% +/- 11.89% (mikro: 69.92%)

	true No	true Buy	class precision
pred. No	56	40	58.33%
pred. Buy	0	37	100.00%
class recall	100.00%	48.05%	

Fig. 3. Confusion Matrix of SVM



```
PerformanceVector

PerformanceVector:
accuracy: 69.89% +/- 11.89% (mikro: 69.92%)
ConfusionMatrix:
True:   No      Buy
No:     56      40
Buy:    0       37
precision: 100.00% +/- 0.00% (mikro: 100.00%) (positive class: Buy)
ConfusionMatrix:
True:   No      Buy
No:     56      40
Buy:    0       37
recall: 47.32% +/- 21.64% (mikro: 48.05%) (positive class: Buy)
ConfusionMatrix:
True:   No      Buy
No:     56      40
Buy:    0       37
AUC (optimistic): 0.880 +/- 0.060 (mikro: 0.880) (positive class: Buy)
AUC: 0.880 +/- 0.060 (mikro: 0.880) (positive class: Buy)
AUC (pessimistic): 0.880 +/- 0.060 (mikro: 0.880) (positive class: Buy)
```

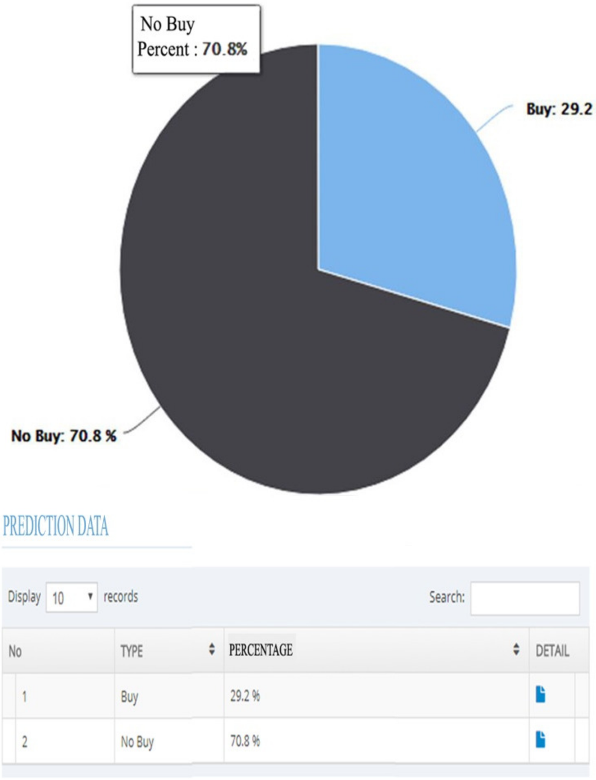
Fig. 4. Text View Confusion Matrix of SVM

From the results above, further confusion matrix calculation accuracy value is performed, the precision, and the recall. Comparison of the values of accuracy, precision, and the recall has been calculated for the method C 4.5 and Support vector Machine can be seen in table 1.

Table 1.
Comparison of Accuracy, Precision, dan Recall

	C4.5	SVM
	Training	Training
Accuracy	86.37%	69.89%
Precision	86.25%	100%
Recall	90.14%	47.32%

4.3. Applied of the Selected Algorithm

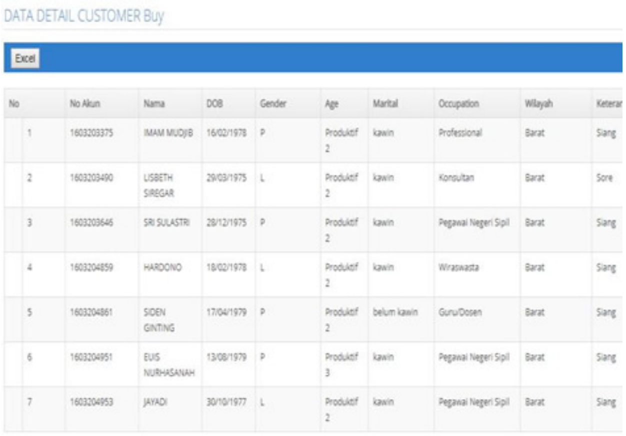


No	TYPE	PERCENTAGE	DETAIL
1	Buy	29.2 %	
2	No Buy	70.8 %	

Fig. 5. Graphic of Customer Prediction

4.4. Prediction result of Buy

From the results of the verification of the selected algorithm using web-based applications from 24 records obtained a predicted customer will Buy insurance by as much as 7 people with percentage of 29.2%.

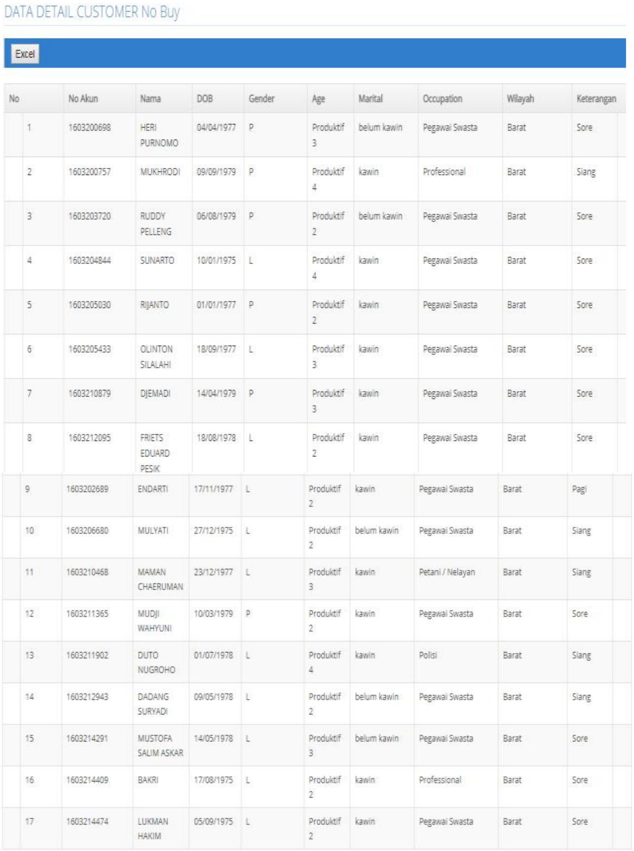


No	No Akun	Nama	DOB	Gender	Age	Marital	Occupation	Wilayah	Keterar
1	1603203375	IMAM MUDJIB	16/02/1978	P	Produktif 2	kawin	Professional	Barat	Siang
2	1603203490	LISBETH SIREGAR	29/03/1975	L	Produktif 2	kawin	Konsultan	Barat	Sore
3	1603203646	SRI SULASTRI	28/12/1975	P	Produktif 2	kawin	Pegawai Negeri Sipil	Barat	Siang
4	1603204859	HARDONO	18/02/1978	L	Produktif 2	kawin	Wiraswasta	Barat	Siang
5	1603204861	SIDEN GINTING	17/04/1979	P	Produktif 2	belum kawin	Guru/Dosen	Barat	Siang
6	1603204951	EUIS NURHASANAH	13/08/1979	P	Produktif 3	kawin	Pegawai Negeri Sipil	Barat	Siang
7	1603204953	JAYADI	30/10/1977	L	Produktif 2	kawin	Pegawai Negeri Sipil	Barat	Siang

Fig. 6. Form Details Buy

4.5. Prediction result of No

From the results of the verification of the selected algorithm using web-based applications from 24 records obtained a predicted customer will NO insurance by as much as 17 people, with the percentage of 70.8%.



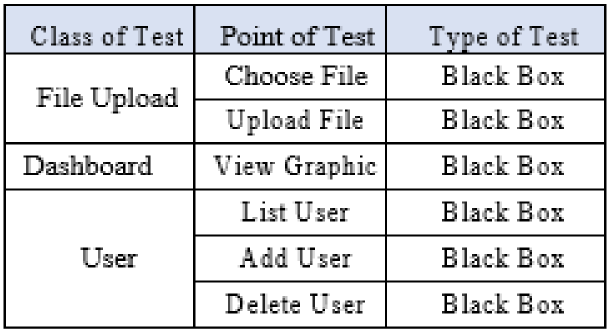
No	No Akun	Nama	DOB	Gender	Age	Marital	Occupation	Wilayah	Keterangan
1	1603200698	HERI PURNOMO	04/04/1977	P	Produktif 3	belum kawin	Pegawai Swasta	Barat	Sore
2	1603200757	MUKHRODI	09/09/1979	P	Produktif 4	kawin	Professional	Barat	Siang
3	1603203720	RUDDY PELLENG	06/08/1979	P	Produktif 2	belum kawin	Pegawai Swasta	Barat	Sore
4	1603204844	SUNARTO	10/01/1975	L	Produktif 4	kawin	Pegawai Swasta	Barat	Sore
5	1603205030	RIJANTO	01/01/1977	P	Produktif 2	kawin	Pegawai Swasta	Barat	Sore
6	1603205433	OLINTON SILALAHI	18/09/1977	L	Produktif 3	kawin	Pegawai Swasta	Barat	Sore
7	1603210879	DJEMADI	14/04/1979	P	Produktif 3	kawin	Pegawai Swasta	Barat	Sore
8	1603212095	FRIETS EDUARD PESIK	18/08/1978	L	Produktif 2	kawin	Pegawai Swasta	Barat	Sore
9	1603202689	ENDARTI	17/11/1977	L	Produktif 2	kawin	Pegawai Swasta	Barat	Pagi
10	1603206680	MULYATI	27/12/1975	L	Produktif 2	belum kawin	Pegawai Swasta	Barat	Siang
11	1603210468	MAMAN CHAERUMAN	23/12/1977	L	Produktif 3	kawin	Petani / Nelayan	Barat	Siang
12	1603211365	MUDJI WAHYUNI	10/03/1979	P	Produktif 2	kawin	Pegawai Swasta	Barat	Sore
13	1603211902	DUTO NUGROHO	01/07/1978	L	Produktif 4	kawin	Polisi	Barat	Siang
14	1603212943	DADANG SURYADI	09/05/1978	L	Produktif 2	belum kawin	Pegawai Swasta	Barat	Siang
15	1603214291	MUSTOFA SALIM ASKAR	14/05/1978	L	Produktif 3	belum kawin	Pegawai Swasta	Barat	Sore
16	1603214409	BAKRI	17/08/1975	L	Produktif 2	kawin	Professional	Barat	Sore
17	1603214474	LUKMAN HAKIM	05/09/1975	L	Produktif 2	kawin	Pegawai Swasta	Barat	Sore

Fig. 7. Form Details No

4.6. Functional Testing of Prototype

Testing conducted with the aim to find out whether the application is built in accordance with the functional in expected.

Table 2.
Functional test with Blackbox method



Class of Test	Point of Test	Type of Test
File Upload	Choose File	Black Box
	Upload File	Black Box
Dashboard	View Graphic	Black Box
User	List User	Black Box
	Add User	Black Box
	Delete User	Black Box

V. CONCLUSIONS

From performance measurement with the doing comparisons of the two algorithms has been done then it can be concluded that:

1. The C4.5 algorithm has the highest accuracy rate of 86.37% while SVM is 69.89%, the difference between them is 23%. The C4.5 algorithm model has AUC of 0.935 and SVM 0.88, from the AUC value, C4.5 algorithm is included in the category of excellent classification and SVM good classification, then the C4.5 algorithm can be implemented in determining potential insurance customers.
2. The rule generated by the C4.5 algorithm is applied in the prototype prediction of prospective insurance customers with the accuracy of prototype verification testing of 73.12%. Based on the accuracy produced by the prototype shows that the methods and prototypes applied are good in predicting prospective insurance customers.

From the results of this research it is expected that the selected algorithm is C4.5 algorithm in predicting the customer's decision to buy / have insurance more precisely and quickly, thus helping the achievement of company performance.

REFERENCES

[1] Kotler. 2009. 245. Batubara, Muhamad Hendri. Strategi Marketing Public Relation (MPR) Berupa Promosi Dan Sponsorsip Untuk Mempengaruhi Konsumen Dalam Keputusan Pembelian (Studi Produk Perawatan Bayi Johnsons Baby). Tes., Universitas Indonesia, 2010.

[2] Pratama, Ramadhan Wahyu. 2015. “Prediksi Calon Nasabah Gadai Potensial pada PT. Pegadaian (Persero) dengan Menggunakan Metode Support Vector Machine-Sequential Minimal Optimization (SVM-SMO)”. Skripsi Telkom University.

[3] Dhika, Harry. 2015. “Kajian Komparasi Penerapan Algoritma C4.5, Naïve Bayes, dan Neural Network dalam Pemilihan Mitra Kerja Penyedia Jasa Transportasi: Studi Kasus CV. Viradi Global Pratama”. Seminar Nasional Inovasi dan Tren (SNIT) 2015.

[4] Han, J., & Kamber, M. (2006). Data Mining Concept and Tehniques. San Fransisco: Morgan Kauffman. ISBN 13: 978-1-55860-901-3

[5] Wu, X. et al., 2008. Top 10 algorithms in data mining, A Chapman & Hall Book.

[6] Gorunescu, F. (2011). Data Mining Concept Model and Techniques. Berlin: Springer. ISBN 978-3-642-19720-8.

[7] Witten et al., 2011. Data Mining Practical Machine Learning Tools and Techniques 3rd ed., Burlington: Elsevier Inc.

[8] Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining. Yogyakarta: Andi Publishing.

[9] Bramer, Max. (2007). Principles of Data Mining. London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.

[10] Dawson 2009 Dawson, C. W. Projects in Computing and Information System A Student's Guide. England: Addison-Wesley, 2009.

[11] Berndtssom 2008 Berndtssom, M., Hansson, J., Olsson, B., & Lundell, B. A Guide for Students in Computer Science and Information Systems. London: Springer, 2008.