Journal Engineering Sciences (Improsci)
e-ISSN 3031-7088
p-ISSN 3032-3452

Implementation Of The Bagging-Based C4.5 Algorithm To Analyze Customer Satisfaction In Electronic Pulse

Muhamad Sulaeman[1], M. Setiawan[2], M. Taufiq Irpan[3].
[1]Universitas Sindang Kasih Majalengka, Indonesia, mslcrb@yahoo.com
[2]Universitas Muhammadyah Cirebon, Indonesia, setiawan.mohamat@gmail.com
[3]Universitas Sindang Kasih Majalengka, Indonesia, mochtaufikirfan@gmail.com
Corresponding Author : setiawan.mohamat@gmail.com

Abstract:
Background. Many users of Telkom mobile or gadgets in the world, especially in the field of mobile telecommunications technology, provide much convenience for the customer to make purchases of electronic top-up.
Aims. Competition among suppliers of telecommunication service providers requires an analysis to determine the level of customer satisfaction with electronic pulses.
Methods. In this study, the C4.5 and C4.5 Algorithm Based on Bagging will be used to analyze customer satisfaction, with data collected from 460 customers and 11 variables related to electronic top-up.
Conclusion. This research aims to generate customer satisfaction analysis using the C4.5 and C4.5 algorithms, based on the bagging algorithm, to determine an appropriate method for analyzing customer satisfaction with electronic top-up.

Keywords: Customer, Telecommunication, C4.5 Algorithm, C4.5 Base on Bagging

INTRODUCTION

Competition among mobile phone credit outlets, coupled with increasing demand for mobile telecommunications technology, requires analysis that can measure customer satisfaction to increase profits and retain customers. Retention is a key component of Customer Relationship Management (CRM) that fosters positive relationships between consumers and providers. Satisfaction varies from one consumer to another, as each has different needs. According to Kotler (1997), one way to measure customer satisfaction is to conduct a customer satisfaction survey. Bintang Pagi Cell has experienced both increases and decreases in customer numbers from year to year, due to many switching to other outlets and a small number of new customers. For this reason, an algorithm is needed to measure customer satisfaction. Bintang Pagi Cell is located at Jl. Kesambi No. 3 A, Cirebon City, which serves schoolchildren, boarding students, and the community around the Bintang Pagi Cell counter. The data used in this study were collected from 2010 to 2014, with each year spanning 3 months.

Customer satisfaction patterns cannot be predicted solely from the number of mobile phone credit sales transactions and consumer survey results. One method for predicting customer satisfaction is data mining. The data mining classification method used is the C4.5 algorithm, which uses decision tree classification to build accurate classifications without revealing personal information (Jin-Hui & Jian-Jun, 2009). The ensemble method builds a collection of classifiers from training data and then predicts class labels on test data by combining predictions from several classifiers. There are two methods in the ensemble method, namely Bagging and Boosting (Tan Pang Ning, 2005). The bagging method increases classification and prediction accuracy by generating multiple models and selecting the classification or prediction result that occurs most often across the models (majority vote). Meanwhile, boosting is achieved by combining models, and the selected classification or prediction result is the model with the highest weight. Using bagging and boosting techniques is often more accurate than a single classifier. Bagging works best when the classification procedure being used is volatile. That is, when small changes in the training samples can often result in significant differences in the predictions obtained using a particular method. Boosting is used with basic, somewhat unstable classification methods (Quinlan, 1993).

This study will use the Bagging-based C4.5 algorithm to analyze customer satisfaction. The final result of this study is a customer analysis prediction, calculated by evaluating the customer satisfaction prediction model using a lift curve, a Gini coefficient, and top-decile lift. In addition, an F-measure calculation will be performed to evaluate data imbalance.
Research on customer satisfaction analysis through data mining approaches, particularly C4.5 algorithms and ensemble methods such as Bagging, has grown rapidly in recent years. In general, previous studies:

1. Focusing on the use of the C4.5 algorithm in the classification of customer data, especially to predict churn, loyalty, or satisfaction levels, due to its advantages in producing an easy-to-interpret decision tree.

2. Strengthen prediction accuracy using ensemble techniques, such as Bagging and Boosting, to overcome the weaknesses of a single classifier against noise, data variants, and unbalanced data.
3. In the context of telecommunications, some previous research has focused more on:
   a. customer churn prediction,
   b. segmentation of customers based on the use of services,
   c. Simulation of customer loyalty and retention.

In this study, Sulaeman et al. contributed to the domain of the application of a special ensemble method for the customer satisfaction domain of electronic pulses, conducting a direct comparison between C4.5 vs Bagging-Based C4.5 using the evaluation measures of accuracy, precision, recall, ROC, and confusion matrix for the Bintang Pagi Cell customer dataset. Thus, the State of the Art of this study is at the intersection between: data mining, assembly algorithms, and Telecom customer satisfaction analysis.

LITERATURE REVIEW

Customer Satisfaction

According to Kotler (1997), customer satisfaction is a person's emotional state resulting from a comparison between a product's performance or outcome and customer expectations. Based on the definitions provided by the experts above, it can be concluded that customer satisfaction or dissatisfaction is the difference between pre-purchase expectations and the perceived performance or outcome after purchase. Customers are the recipients of a person's or organization's work, and therefore determine its quality, and they can communicate their needs and expectations. Everyone in a company must work with internal and external customers to determine their needs and collaborate with internal and external suppliers.

Data Mining

Data mining is a set of processes used to manually extract added value from a database, yielding previously unknown information. The resulting information is obtained by extracting and recognizing important or interesting patterns from the database [7]. Data mining is a process that uses statistical, mathematical, and artificial intelligence techniques, along with machine learning, to extract and identify useful information and related knowledge from large databases. According to the Gartner Group, data mining is the process of discovering meaningful relationships, patterns, and trends by examining large sets of stored data using pattern recognition techniques such as statistical and mathematical techniques (R. Kohavi, 1996).

C.45 Algorithm

The C4.5 algorithm is a well-known, favored data classification algorithm using decision trees. These advantages include: it can process both numeric (continuous) and discrete data, can handle missing attribute values, produces easily interpretable rules, and is the fastest among algorithms that use main memory on a computer [4]. The objective of the decision tree algorithm is to construct a tree data structure (called a decision tree) that can be used to predict the class of a case or a new record that does not yet have a class. C4.5 constructs a decision tree using a divide-and-conquer strategy.

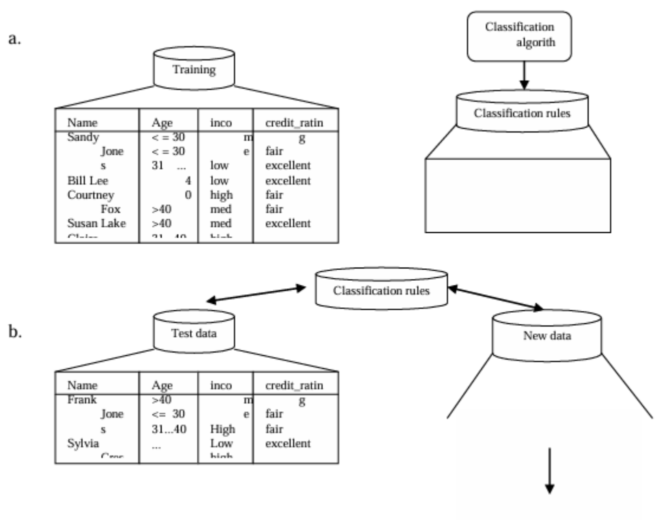

Figure 1. Decision Tree with C4.5 (Jin-Hui&Jian-Jun, 2009)

Introduction to Bagging

“Bagging” is a method that can improve the results of machine learning classification algorithms (Breiman, 1994). This method was formulated by Leo Breiman, and the name is derived from the phrase “Bootstrap Aggregating” (Breiman, 1994). Bagging is a method based on ensemble methods; therefore, the general steps of the Bagging method can be seen in the image described previously. There are several important aspects of this method, namely:

a. Data distribution (bootstrapping) is created using sampling with replacement.

b. Building a classifier on each bootstrap sample. How this method works is illustrated in the following image:

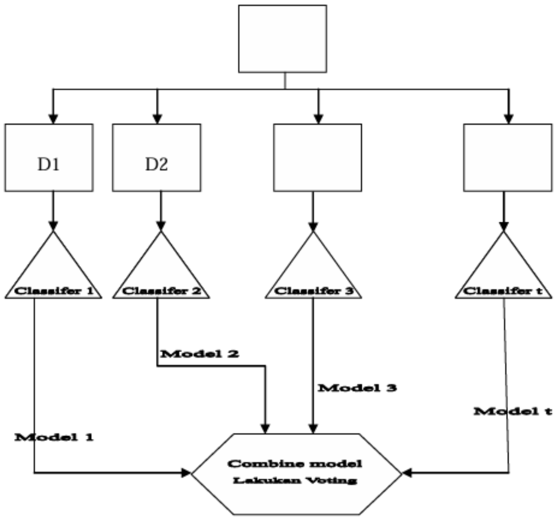

Figure 2. Bagging Process (Govindaraju, 2008)

The Bagging Process

From a data set, we can randomly select only a portion of observations (resampling) and then use them to construct a tree. Repeating this process will produce different portions of the data, followed by different trees. As an illustration, suppose we obtain three different trees from three partial data samplings, as depicted in Figure 3.

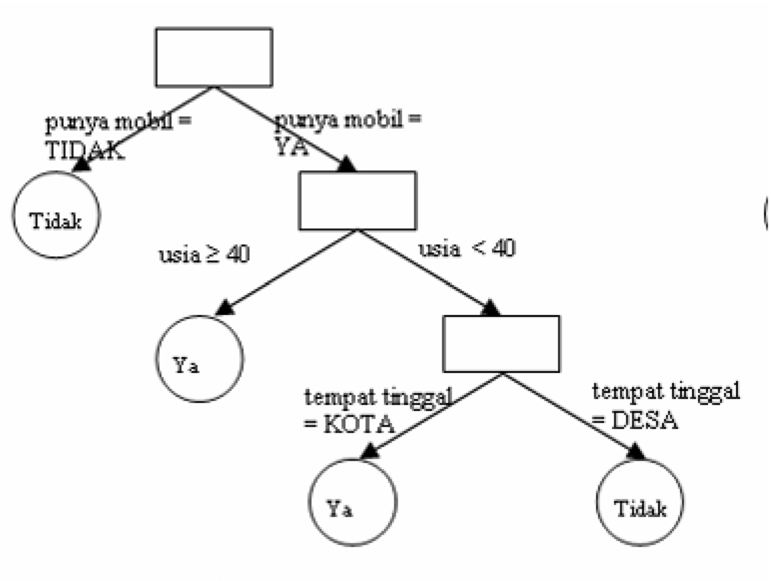

Figure 3. Bagging Process, tree 1 (Tse.O.K, 2005)

The research design used in this study is the experimental method. Experimental research is a study that answers the question "if we do something under strictly controlled conditions, what will happen?" To determine whether there is a change or not in a strictly controlled condition, we need a treatment in that condition, and this is what is done in experimental research. Therefore, experimental research can be said to be a research method used to find the effect of certain treatments on others under controlled conditions (Sugiyono, 2010). Experimental research is a study in which at least one variable is found to be manipulated to study cause-and-effect relationships (Solso, 2002). Therefore, experimental research is closely related to testing a hypothesis in order to find the effect, relationship, or difference in changes in the group that is given the treatment.

The elements of novelty in this article can be identified as:

1. Application of the C4.5 Bagging-Based algorithm specifically in the analysis of customer satisfaction of electronic pulses. Many studies discuss churn or loyalty, but the focus on customer satisfaction of local credit counters with internal datasets is still rare.
2. Systematic comparison of three attribute criteria: Gain Ratio, Information Gain, Gini Index, for both algorithms (C4.5 and Bagging-C4.5) simultaneously.
3. Evaluate the method using a combination of metrics: accuracy, precision, Remember, ROC (positive & negative), Confusion Matrix which is rarely done comprehensively in the context of credit customer satisfaction.
4. Original local-scale dataset (Morning Star Cell counter). Which reflects: Characteristics of Students/Students, certain regions, Electronic Pulse-specific services.

So that novelty can be formulated: "The use of Bagging-C4.5 to improve the predictive accuracy of local electronic pulse customer satisfaction with multi-criteria attribute comparison and multi-metric evaluation to determine the optimal classification method."

METHOD

The research method used in developing this herbal medicine expert system application for ear, nose, and throat (ENT) diseases uses the R&D (Research & Development) method. Research and development (R&D) is a research method used to produce specific products and test their effectiveness [20]. These products are not always in the form of objects or hardware, such as books, modules, classroom or laboratory learning aids, but can also be software, such as computer programs for data processing, classroom learning, libraries or laboratories, or models for education, learning, training, guidance, evaluation, management, and others (Wahana Komputer, 2013).

Descriptive analysis is a method that describes the data obtained according to the data to be collected and then examines the data conditions and then groups them according to data tabulation to illustrate the actual conditions that occur at the empirical level (Ade Johar Maturidi, 2012). In writing this thesis, the author used the research and development method or R&D (Research and Development) because the author developed a new application and collected data by directly observing conditions in the field so that it can be used as a consideration in decision making.

DISCUSSION

Electronic Credit Customer Dataset

The table below describes an example of a data set used in research to determine customer satisfaction using the C4.5 algorithm. Compared with the bagging-based C4.5 algorithm, this data set was used to determine the most suitable algorithm for predicting customer satisfaction.

Table 1. Bintang Pagi Cell Electronic Credit Customer Dataset

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
01092010	087829125833	Kesambi	Kota Cirebon	1 km	XL	Rp 5,000.00	Tunai	Komplain	Baru	Masih aktif
02092010	081909902909	Perum	Kota Cirebon	4 km	XL	Rp 5,000.00	Tunai	Tidak ada	Baru	Masih aktif
03092010	085224251777	Jagasatru	Kota Cirebon	2 km	AS	Rp 15,000.00	Kredit	Tidak ada	Lama	Masih aktif
04092010	08529480404	Kesambi	Kota Cirebon	4 km	AS	Rp 5,000.00	Tunai	Tidak ada	Lama	Masih aktif
05092010	08562136406	Kesambi	Kota Cirebon	5 km	IM3	Rp 25,000.00	Tunai	Tidak ada	Lama	Masih aktif
06092010	087829172360	Kesambi	Kota Cirebon	3 km	IM3	Rp 5,000.00	Tunai	Tidak ada	Lama	Masih aktif
07092010	081946904552	Kesambi	Kota Cirebon	9 km	XL	Rp 5,000.00	Tunai	Tidak ada	Lama	Masih aktif
08092010	081222151169	Kesambi	Kota Cirebon	2 km	SIMPATI	Rp 10,000.00	Tunai	Tidak ada	Lama	Masih aktif
09092010	087828163461	Kesambi	Kota Cirebon	500 m	XL	Rp 5,000.00	Tunai	Tidak ada	Lama	Masih aktif
10092010	02319299225	Kesambi	Kota Cirebon	300 m	ESIA	Rp 20,000.00	Tunai	Tidak ada	Lama	Masih aktif

Researchers use 11 variables which are described by normalization, the explanation of the attributes used is: X1: Customer Code X2: Customer Cell Phone Number X3: Customer

Sub-district Address X4: Customer Regency/City Address X5: Distance from Customer's Residence to the counter X6: Prime X7: Amount of Credit Top-Up X8: Payment X9: Satisfaction X10: Membership X11: Activation.

Entropy and Gain Calculations

Entropy and Gain calculations are required to find the highest node. They are used to determine the degree of the attribute with the highest value, which is placed at the root end. A description of the calculations can be found in the appendix.

Accuracy Comparison Model Testing

The samples will be tested on the training data to obtain comparative results between the C4.5 Algorithm and the Bagging-Based C4.5. In this test, other attribute selection methods, namely the Gini Index, Information Gain, and the Gini Index, were added to assess the accuracy of each attribute selection method using the same algorithm, C4.5 and the Bagging-Based C4.5. The results for each attribute selection method show differences, as shown in Table 2. Figure 4 below shows the experimental results for accuracy between the Gini Index, Information Gain, and the Gini Index.

Table 2. Accuracy Results for C4.5 and the Bagging-Based C4.5

Kriteria	AKURASI	
	C4.5 (%)	C4.5 berbasis Bagging (%)
Gain Ratio	90,65%	91,09%
Information Gain	91,09%	91,09%
Gini Index	89,13%	90,22%

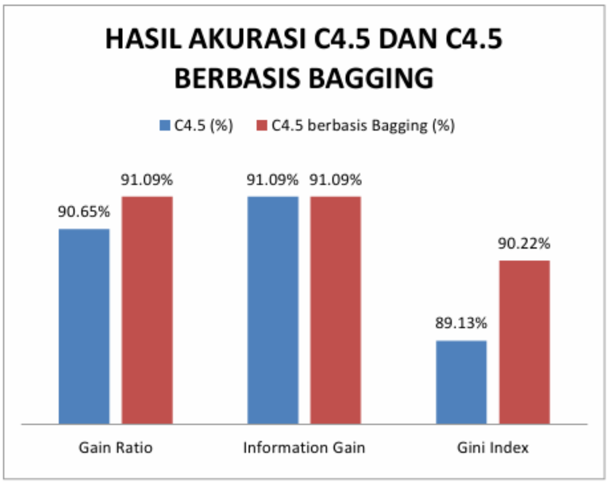

Figure 4: Accuracy Results Graph for C.45 and Bagging-Based C.45

Table 2 and Figure 4 show that, based on experimental criteria using Gain Ratio, Information Gain, and Gini Index, the bagging-based C4.5 algorithm performs better than the C4.5 algorithm for the Gain Ratio and Gini Index criteria. Meanwhile, for the Information Gain criterion, the accuracy results show the same value for both algorithms. Therefore, it can be concluded that the bagging-based C4.5 gain ratio criterion yields superior accuracy, with 91.09%.

Comparison Test of Class Precision of C4.5 and Bagging-Based C4.5

Table 3: Class Precision Results for C4.5 and Bagging-Based C4.5

Kriteria	CLASS PRECISION	
	C4.5 (%)	C4.5 berbasis Bagging (%)
Gain Ratio	94,12%	94,15%
Information Gain	94,14%	94,14%
Gini Index	94,02%	94,28%

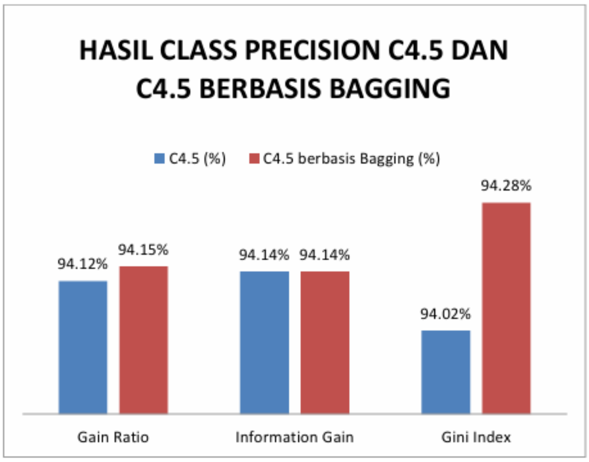

Figure 5. Graph of Class Precision Results for C.45 and C.45 Based on Bagging

Table 3 and Figure 5 show that, based on experimental criteria using Gain Ratio, Information Gain, and Gini Index, the bagging-based C4.5 algorithm performs better than the C4.5 algorithm for both Gain Ratio and Gini Index. Meanwhile, for the Information Gain criterion, Class Precision results show similar values for both algorithms. Therefore, it can be concluded that the Class Precision performed using the bagging-based C4.5 Gini Index criterion is superior, with a Class Precision value of 94.28%.

Comparative Test of Class Recall for C4.5 and C4.5 Based on Bagging

Table 4. Class Recall Results for C4.5 and C4.5 Based on Bagging

Kriteria	CLASS RECALL	
	C4.5 (%)	C4.5 berbasis Bagging (%)
Gain Ratio	96,07%	96,54%
Information Gain	96,54%	96,54%
Gini Index	94,46%	95,38%

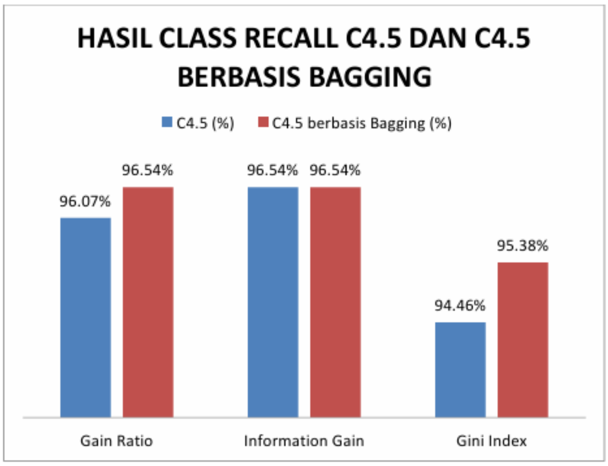

Figure 6. Graph of Class Recall Results for C.45 and C.45 Based on Bagging

Table 4 and Figure 6 show that, based on the experimental criteria of Gain Ratio, Information Gain, and Gini Index, the bagging-based C4.5 algorithm performs better than the C4.5 algorithm for the Gini Index criterion. Meanwhile, for the Information Gain and Gain Ratio criteria, the Class Recall results show similar values for both algorithms. Therefore, it can be concluded that the Class Recall using the bagging-based C4.5 Gini Index criterion performs better, with a Class Recall value of 95.328%.

ROC Curve Comparison Test (Positive Class)

Table 5. ROC Curve Analysis of Gain Ratio, Information Gain, and Gini Index Criteria for C4.5 and C4.5 Based on Bagging (Positive Class)

Kriteria	*Kurva ROC (Positif Class)*	
	C4.5 (%)	C4.5 Berbasis Bagging (%)
Gain Ratio	0.848	0.882
Information Gain	0.848	0.908
Gini Index	0.89	0.798

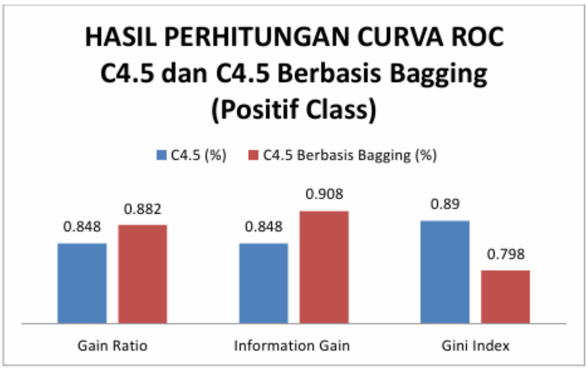

Figure 7. Comparison Graph of ROC Curves for C4.5 and Bagging-Based C4.5

Table 5 and Figure 7 show that, when evaluated using ROC calculations based on the Gain Ratio, Information Gain, and Gini Index criteria, the bagging-based C4.5 algorithm performs better than the C4.5 algorithm for both the Gini Index and Information Gain criteria. C4.5 performs better for the Gain Ratio criterion. Therefore, it can be concluded that the ROC using the positive class criteria, conducted using the Gini Index and Information Gain criteria for C4.5 based on bagging, is superior, with ROC values for the Positive Class of 0.882% and 0.902%, respectively.

ROC Curve Comparison Test (Negative Class)

Table 6. ROC Curve Analysis of Gain Ratio, Information Gain, and Gini Index Criteria for C4.5 and Bagging-Based C4.5 (Negative Class)

Kriteria	*Kurva ROC (Negatife Class)*	
	C4.5 (%)	C4.5 Berbasis Bagging (%)
Gain Ratio	0.848	0.882
Information Gain	0.848	0.908
Gini Index	0.89	0.798

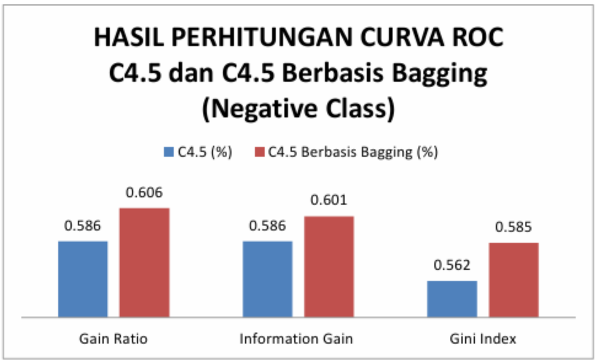

Figure 8. Comparison Graph of ROC Curves for Negative Class C4.5 and Bagging-Based C4.5

Table 6 and Figure 8 show that, based on experimental criteria, the bagging-based C4.5 algorithm outperforms the C4.5 algorithm for Gini Index, Information Gain, and Gini Ratio. Therefore, it can be concluded that the ROC for the negative class using the Gini Index and Information Gain criteria for bagging-based C4.5 is superior, with ROC values of 0.606%, 0.601%, and 0.585%, respectively.

Confusion Matrix Comparison Test

After analyzing the results of the Accuracy, Precision, Recall, ROC Curve (Positive), and ROC Curve (Negative) tests above, the final calculation was performed to determine the Confusion Matrix.

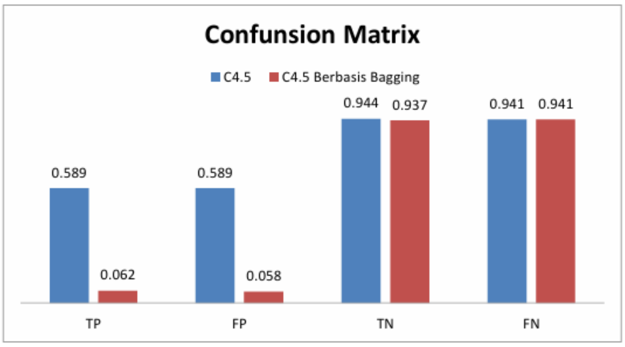

Figure 9. Comparison Chart of Confusion Matrix C4.5 and Bagging-Based C4.5

Figure 9 illustrates the experiment conducted in calculating the confusion matrix. The percentage of True Positives (TP) in C45 was 0.589, while Bagging-Based C4.5 was 0.062. The percentage of False Positives (FP) in C45 was 0.589, while Bagging-Based C4.5 was 0.058. The percentage of True Negatives (TN) in C45 was 0.944, while Bagging-Based C4.5 was 0.937. The percentage of False Negatives (FN) in C45 was 0.941, while Bagging-Based C4.5 was 0.941. Therefore, it can be concluded that the best confusion matrix value is found in C4.5.

From the article and the context of the research, several gaps can be seen:

1. The focus of the research is still on algorithm comparison, not yet on factors that cause satisfaction, interpretation of business decisions, and CRM strategy based on models.
2. Limited datasets: 460 customers, only 11 variables, 51-65 Sulaiman-Ade. Gap: not accounting for psychographic variables, competitor services, promotion, and network quality.

3. Haven't explored other ensemble techniques such as Random Forest, AdaBoost, and Gradient Boosting, which can produce a wider comparative effect.
4. Data imbalance is only mentioned through the F-measure; however, Imbalance management techniques are not followed (e.g., SMOTE, cost-sensitive learning).
5. Research applications still focus on one outlet. So that generalization: Region, Large telecommunications operators, Modern Digital E-Payment, It is still a further research opportunity.
6. External validation has not been done, such as K-fold cross-validation or split testing across datasets.

CONCLUSION

From the research results, it is proven that the C4.5 algorithm based on bagging is more appropriate than the C4.5 analysis based on the analysts' Accuracy results at BINTANG PAGI CELL management. This is supported by the research evaluation, which shows that the bagging-based C4.5 algorithm achieves a higher Accuracy than the original C4.5.

Implication

From the results of the evaluation in terms of accuracy, it was proven that C4.5 based on bagging was appropriate compared to the results of the C4.5 algorithm analysis carried out. With increasing accuracy in predicting behavioral patterns towards customer satisfaction, the company can provide better service to Bintang Pagi Cell electronic credit customers. In further research, it will be developed in retesting the research model by adding other attributes such as data on the percentage increase in the number of customers.

BIBLIOGRAPHY

Armstrong, Michael, Performance Management. Yogyakarta: Tugu Publisher, 2004.

Bond et.al., Integrated Impact Assessment for Sustainable Developmen: a Case Study Approach. World Development Vol.29 No.6, Pergamon, 2001, pp. 1011-1024.

Breiman, L., “Bagging Predictor”, Technical report No. 421, Departement of statistics University of California, 1994.

Bühlman, P. dan Yu, B., Analyzing Bagging, The Annals of Statistics, Vol. 30 no. 4, 2002, hal 927-961.

C.R. Kothari, Research Methodology, New Age International, 2004.

Chang, Chingpao-Pao, Software process improvement using multivariate statistical process control and action based defect prediction, Departement of computer science and information engineering, National Chen Kung University, Taiwan, 2008.

Clarke, R.T. dan Bintercourt, H. R , Use of Classification And Regression Trees (CART) to Classify Remotely_Sensed Digital Images, Research Report , Centro stadual de Pesquisas em Sensoriamento Remoto Universidade Federal do Rio Grande do Sul – UFRGS , Porto Alegre, Brazil, 2003.

Daihani, D. Umar, Komputerisasi Pengambilan Keputusan. Jakarta: PT Elekmedia Komputindo, 2001.

Dias, J.G dan Vermunt J.K., A Bootstrap based Aggregate Classier for Model based Clustering, Journal of Annals Statistics., 2005.

Efron, B. dan Tibshirani, R.J., An Introduction to the Bootstrap , New York: Chapman Hall, 1993.

Efron, B., dan Tibshirani, R. J., An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.

Engel. James F, Ronger D. Black Well dan Paul W. Miniard (1995). Perilaku Konsumen. Terjemahan Drs.F.X Budiyanto. Edisi Pertama. Binarupa Aksara.

Eni Efriani, Penerapan Algoritma Klasifikasi C4.5 Berbasis AdaBoost Untuk Prediksi Loyalitas Pelanggan Distributor Pulsa Elektronik, 2010.

Govindaraju, dkk., Perancangan Sistem Prediksi Churn Pelanggan PT. Telekomunikasi Seluler Dengan Memanfaatkan Proses Data Mining, 2008.

Han, J. dan Kamber, M., Data mining: Concepts and techniques (2nd ed,), Elsevier Inc., 2006.

I.S. Dhillon, Y. Guan, and J. Kogan, Iterative clustering of High Dimensional text data augumented by local search. In Proceedings of the 2002 IEEE International Conference on Data Mining Pages, 2002, pp.131-138.

Jeffery L. Whitten et al., Metode Desain dan Analisis Sistem.. Yogyakarta: ANDI, 2004.

Jin-Hui&Jian-Jun, The Comparative Analysis and Study of Mobile-based Customer Data Churn Prediction Model, 2009.

Kotler, Philip, Manajemen Pemasaran, Analisa perencanaan, Implementasi dan control, Edisi Kesembilan, Jilid 1 dan jilid 2. Terjemahan: Hendra Teguh S.E.,A.K., dan Ronny A. Rusli, S.E. Jakarta: Prehalindo, 1997.

Kraljevic & Gotovac, Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Service, 2010. [10]. Saaty, T.L., Decision Making For Leaders, 4th ed., University of Pittsburgh: RWS Publication, 2001.

Larose D, T., Data Mining Methods and Models, Jhon Wiley & Sons, Inc. Hoboken New Jersey, 2006.

Larose D, T., Discovering knowledge in data : an introduction to data mining, Jhon Wiley & Sons Inc., 2005.

Luan, J, Chapter 2: Data Mining and Its Application in Higher Education. Knowledge Management – Building a Competitive Advantage in Higher Education. Serban, A. & Luan, J. (eds.) Jossey-Bass, 2002.

Quinlan, J.R., C4.5: Programs for Machine Learning. Sam Mateo: Morgan Kaufmann, 1993.

R Kohavi, Scaling up the Akurasi of Naive-Bayes Classifiers: A Decision Tree Hybrid. Proceedings of the second international conference, 1996.

Sibaroni, Yuliant, Analisis dan Penerapan Metode Klasifikasi untuk Pembangunan Perangkat Lunak Penerimaan Mahasiswa Baru Jalur Non Tulis, ITB, Bandung, 2008.

Solso, R.L. & MacLin, M.K., Experimental Psychology : A Case Approach (5th Ed.). Boston: Allyn & Bacon, 2002.

Sri, Veronika, "Pengembangan Skalabilitas Algoritma Klasifikasi C4.5 dengan Pendekatan Konsep Operator Relasi" Bandung: ITB, 2007.

Sugiyono, Metode Penelitian Pendidikan: Pendekatan Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta, 2010.

Tan Pang Ning, Steinbach Michael dan Kumar Vipin, " Introduction to Data Mining [Laporan]." - United State of America : Pearson Education, 2005.

Tse. O.K., dan P.e. Wilton. 1988. "Models of Consumer Satisfaction Formation: AnExtention". Journal of Marketing Research. Vol. 25. May.

Veronica S. Moertini, Towards The Use Of C4.5 Algorithm for Classifying Banking Dataset, 2003.

Wijanarko Bambang dan Sumarmi, Bagging CART pada Klasifikasi Anak Putus Sekolah ITS, Surabaya, 2009.