# The Determination Analysis Of Telecommunications Customers Potential Cross-Selling With Classification Naive Bayes And C4.5

# The Determination Analysis Of Telecommunications Customers Potential Cross-Selling With Classification Naive Bayes And C4.5

**I Purnamasari[1*], F Handayanna[1*], E Arisawati[1], LS Dewi[1], E G Sihombing[1] Rinawati[1]**

[1]Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri

E-mail: indah.ihi@nusamandiri.ac.id*, frisma.fha@nusamandiri.ac.id*

**Abstract.** Every company has various marketing strategies. Marketing strategies are offered to sell the company's products to benefit if the company introduces new products. However, too many offers to customers that are not true, will only make marketing inefficient and ineffective. Data mining as a way to find patterns and relationships in data can be used to make valid predictions. To simplify the marketing strategy of PT. TELKOM then classifies the data that already exists in PT. TELKOM Jakarta. Telkom customers have the potential to become customers of new Indihome products, so marketing is carried out by PT. TELKOM has become more effective and efficient. By using data mining classification methods, namely the Naive Bayes Classifier algorithm and the C4.5 algorithm, patterns and relationships are obtained to simplify the marketing strategy of PT. TELKOM where in previous research, the results of classification of data mining research models with the Naive Bayes Classifier algorithm have an accuracy value of 85.08% and AUC 0.841 while in this study the C4.5 algorithm has an accuracy value of 88.61% and AUC 0.870. C4.5 is a model with good accuracy for customer classification data that has the potential for more effective and efficient in cross-selling marketing strategy..

## 1. Introduction

The cross-selling marketing strategy is carried out due to the rapid growth of telecommunication which is increasingly saturated where ordinary voice communication services are so common that telecommunications companies have difficulty increasing the number of their customers. Cross-selling is a marketing strategy for existing customers. Marketing for new customers will be more expensive than maintaining existing loyal customers. Therefore, the company sells new products to existing customers to increase company profits. SMS, telephone calls to customers, and product offer letters sent to customers together with monthly billing statements are to cross-selling marketing media to customers to offer products. One very important point is: when companies cross-sell during service delivery, companies often receive complaints [1]. At present the telecommunications market is always changing rapidly, following different developments in various periods. Telecommunications operators are always trying to release various types of new service packages. Therefore a reasonable and scientific predictive analysis is needed [2]. This is due to the promotion of new telecommunications service packages not only concerning revenue and product orientation for telecommunications operators but also the extent to which customers provide perceptions in receiving these new services. However, the level of telecommunications

service tariff management, as of now, cannot offset the level of development seriously, and analyze the way in which the effect of new services. One of the biggest mistakes in the implementation of cross-selling is to offer products that are random or products they don't need [3]. However, too many offers to customers that are not right, in addition to making marketing costs swollen, will also make saturation on the customer side and will have a negative impact where customers reject the products offered. For this reason, marketing of Indihome products can be more efficient and effective, then classification data mining is conducted to existing telephone and speedy customers of PT. Telkom in Jakarta, which has the potential existing customers to improve its services to become Indihome customers. Naive Bayes Classifier (NBC) as one of the data mining classification algorithms has been used in many studies including research which states that NBC requires less time to build the model [4]. Bayesian networks (BNs) More accurately referred to as machine learning technology [5].The C4.5 classification algorithm calculates all tests from the available data and selects the test with the best information (eg Test with the highest acquisition ratio). C4.5 classification algorithm can prune the resulting decision tree. This method can increase the error rate in the training data, but also can reduce the error rate in the test data that is not visible [8]. This research classifies telephone customers quickly for cross selling Indihome Indihome products using the Naive Bayes Classifier (NBC) and C4.5 algorithm to find a classification model that has a higher degree of accuracy, lower classification error and a good AUC value to be used in efficient and effective marketing strategies.

## 2. Literature Review

### *2.1. Data Mining*

The knowledge we have gained can be used to create systems that can provide information such as customer retention, service needs analysis, production control, marketing, market analysis and science exploration. Data mining techniques which include association, classification, prediction, and clustering. The data mining classification algorithm is used to classify large volumes of data and to provide interesting results [9]. Customer data can be used as a classification model to facilitate cross-selling. This model can predict whether existing customers will buy new products or services offered. Our model consists of stages that apply several data mining techniques [10]. Data mining produces models that can perform consumer trend analysis. A simple data mining process will produce models quickly, but its accuracy will not be quite sufficient. A complex process will produce models that take a long time to execute but will provide results with higher accuracy [11]

### *2.2. Naive Bayes Classifier*

One classification algorithm in mining data to obtain a model or function that can determine the class of a new data is the Naive Bayes classifier. Bayes' theorem was put forward by Reverand Thomas Bayes in the 18th century [7]

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{1}$$

Information :

X : Class yet is known data
H : Hypothesis Data X
P(H—X) : posteriori probability
P(H) : prior probability
P(X) : Probability X

$$P(H|X) = P(X|H).H \tag{2}$$

Classification with continuous data Density Gauss used the formula:

$$(X, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{-(x-\mu)^2}{2\sigma^2} \tag{3}$$

Information :
$\mu$ : Mean, average of all the attributes
$\sigma$ : Deviasi standar, expressed variants of all attributes
$\pi$ = 3,1416
exp = 2,7183

$$\sigma^2 = \frac{1}{N-1} \sum (X - \mu)^2 \tag{4}$$

*2.3. C4.5*
C4.5 algorithm derived from a simple divide-and-conquer algorithm that will form a decision tree. By calculating a way to deal with attributes that are numerically valuable after that, overcome missing values. A very important issue in pruning decision trees is that even though the decision trees made by the divide-and-conquer algorithm perform well on the training set, usually the divide-and-conquer algorithm is equipped with training data and does not generalize well on the test set freely. Then a brief method of how to change the decision tree into a classification rule and examine the options provided by the C4.5 algorithm itself. Finally, what is applied to the well-known CART system is used to study decision trees with classification and regression in decision tree pruning strategies[6].

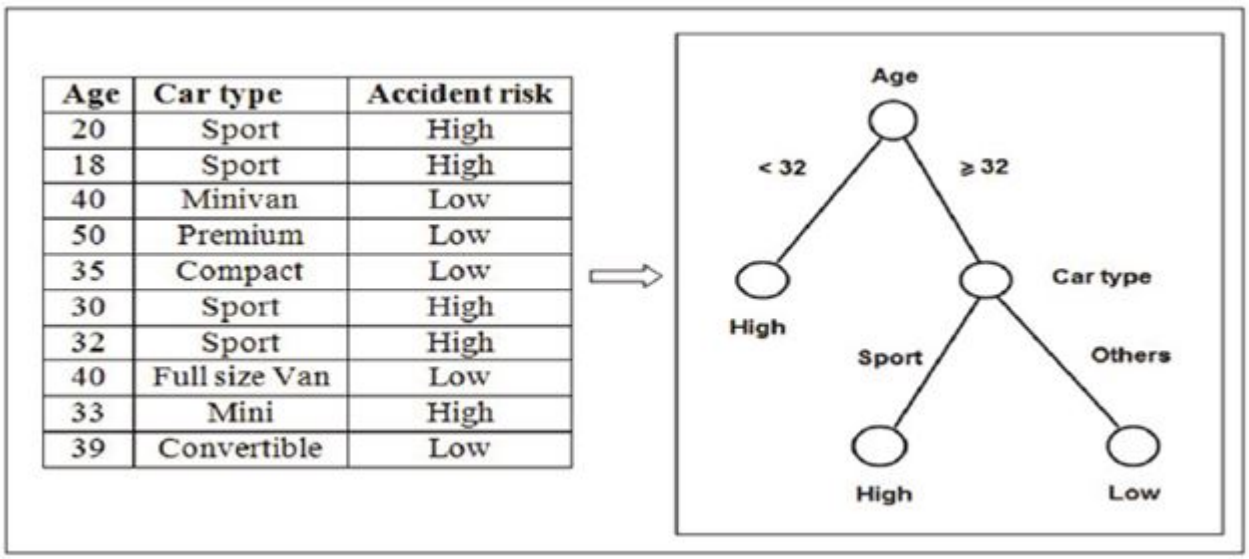


**Figure 1.** Decision Tree Schema

The decision tree is representative of the inductive process therefore the induction tree is a set term. As can be seen in Figure 1. It can be characterized as follows regarding the classification obtained in the induction of decision trees [12]:
• Each node (internal) in the tree is a test based on certain attributes;

• Each branch in the tree is a disclosure of the results of the test;
• A node is a leaf is a terminal node that represents a class (decision).
In principle, the decision tree model is to predict an object into various categories (classes), taking into account the value corresponding to its attributes (predictive variable). the flexibility of the decision tree method is one of the main data mining techniques). The decision tree method is very interesting, mainly because it presents the advantage of producing ('tree' which synthetically summarizes classification) with a highly suggestive visualization However, it must be emphasized that the decision tree method must always be strengthened by other old ways, especially when estimates of their work are examined (for example, Estimates about sending data) such as search retrieval techniques. But especially when the old-fashioned method does not exist, a decision tree will be used and will be preferred over other classification models. Although the classification method with a decision tree does not cover too broadly of the data factors it is likely in the field of recognition of the formation of patterns or structures, which are widely used in other fields of science such as the field of medicine (making diagnoses), the field of computer science (arrangement of information), the field of growth- herbs (classification techniques), the field of psychology (concerning behavior decision theory) and other fields of science [12]. One of the most popular classification methods is decision trees in various data mining applications and helps in the decision making process. Classification helps in doing other things such as choosing the right product to be recommended to certain customers and to predict possibilities. ID3, C4.5, and CART are among the most widely used decision tree algorithms.[13]. For handling data on the multi-label C4.5 algorithm that was adapted in. Several labels are permitted on special tree leaves, and with the modified entropy calculation formula as follows [12]:

$$Entropy(D) = -\sum_{j=1}^{q}(p(\lambda_j)\log_p(\lambda_j) + (q(\lambda_j)\log_q(\lambda_j)) \tag{5}$$

$$where(p(\lambda_j) = relativefrequencyofclass\lambda_j andq(\lambda_j) = 1 - p(\lambda_j) \tag{6}$$

*2.4. Indihome*
Indihome is a Triple Play services from PT. TELKOM consists of three services at once include fast Internet ( Internet on Fiber up to 100 Mbps ), Interactive TV (Usee TV) and Telephone using 100 fiber-optic means of fiber optic held up to the customer's home (Fiber To The Home / FTTH), Indihome also has some additional add-ons like wifi.id Seamless, IndiTravel, Indihome Telkomsel Mania, IndohomeCloud, Global Call, EduKids.id, Indihome view etc. Usee TV is Internet protocol-based television service (IPTV) is regulated by the Regulation of the Minister of Communication and Information Technology Number 15 of 2014 which represents a change of the Regulation of the Minister of Communication and Information Technology Number 11 / PER / M.KOMINFO / 07/2010 on the implementation of IPTV services

## 3. Methodology
This research uses a quantitative approach that can be called the discovery method because this method can be found and developed by various new sciences and technologies. The steps in this research are as follows in figure 2.

The picture above shows the stages of the research to be carried out, namely:
• Data Collection
• Preprocessing Data
• Data Mining Approach
• Experiments
• Measurement

• Analysis results the determination, analysis of the selection of telecommunications users who have cross-selling opportunities with the Naive Bayes algorithm and the C4.5 algorithm.

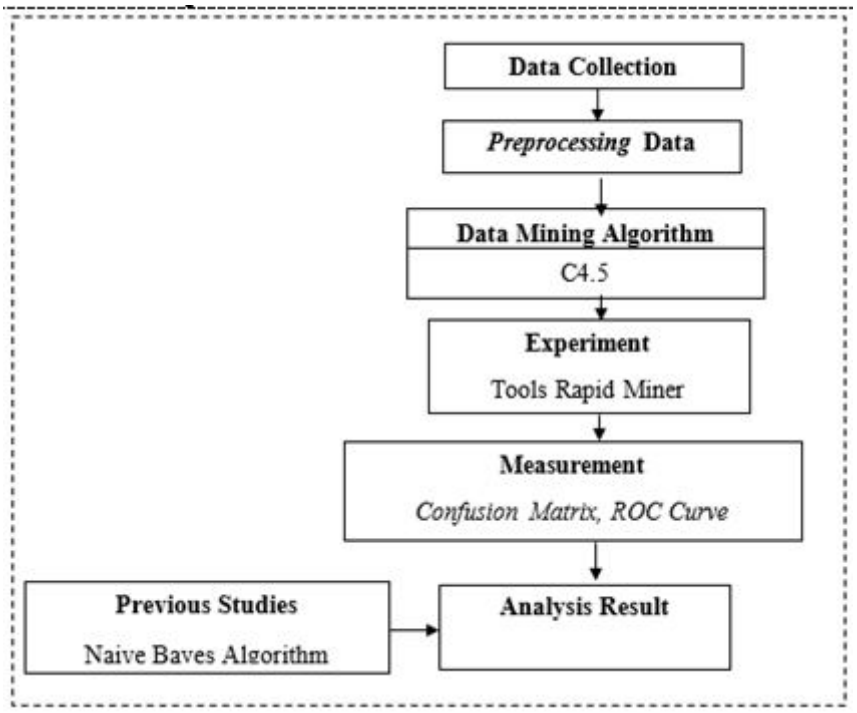


**Figure 2.** Research Design

## 4. Discussion

*4.1. Data Collection*

The data from the division of Business Planning and Performance of PT. Telkom Jakarta are 878 existing data and speedy telephone with 14 attributes with class label attributes (attribute output) which attributes Package. In the 878 telephone customer data, there are 587 existing customers Indihome (3P) and 291 telephone subscribers and speedy turn down the offer Indihome (Decline). Data consist the missing value will fill with the average value. 14 of these attributes is attributes is in table 1 :

*4.2. Data Mining Approach*

*4.2.1. Naive Bayes Classifier*

In previous research, new case classes can be determined by calculating the posterior probabilities of the previous probabilities above. For example in table 2.

Based on calculations using equation 1 above, the results of P(H—3P) = 0,000307061 and P(H—Decline) = 0,0001 so by using equation 2 above the results of P (H— paket = 3P) P (3P) = 0,000307061 x 0,66856492 = 0,00020529 and P(H— paket = Decline) P(Decline) = 0,0001x 0,33143508= 0,33288 So that P (H— paket = 3P) P (3P) ¡ P(H— paket = Decline) P(Decline) From the results of these calculations the P(H —3P) is smaller than the value of P

**Table 1.** Attribute telephone customer data

| Attribute | Information |
|---|---|
| NCLI | Customer identification number |
| Name | Name of the customer |
| Address | Customer address |
| Regional | Regional 2 |
| Witel | Region phone |
| POTS | Plain Old Telephone Service |
| Speedy | Speedy identification number |
| Quadrant | There is no use of the bill |
| Existing | Type of service package |
| R2BB | Ready to Broadband |
| Network Type | Network type in use |
| Zone | The area of marketing |
| Migration | Type of current customers (2P = customer phone and speedy) |
| Package | become customer (decline,3P) |

**Table 2.** Posterior Probability

| Attribute | Score | 3P | Decline |
|---|---|---|---|
| Regional | 2 | 1 | 1 |
| Witel | Banten Timur | 0,00170358 | 0,18556701 |
| Quadrant | 4 | 1 | 1 |
| Existing | INETR1M1 | 0,340715503 | 0,340206186 |
| R2BB | GIPTV | 0,182282794 | 0,41580756 |
| Network Type | Copper | 0,00170358 | 0,18556701 |
| Zone | Attackpromo | 0,01703578 | 0,020618557 |
| Migration | 2P | 1 | 1 |

(H — Decline), so it can be concluded that for the case included in the classification of Decline. This means that the customer will reject the offer of Indihome services.

*4.2.2. C4.5*

In this research prior probability calculation algorithm C4.5 in table 3. Entropy and gain calculations for all attribute, to get a value highest gain. The results of the entire calculation The attribute is seen in table 3

*4.3. Result*

Analysis of the selection of telecommunications users who have cross-selling opportunities with the Naive Bayes algorithm and the C4.5 algorithm based on the analysis of data mining, evaluation results can be seen in Table 4 as follows:

**Table 3.** Gain Ratio Algorithm C4.5

| Attribute | Gain | Split Info | Gain Ratio |
|---|---|---|---|
| Regional | 0 | 0 | 0 |
| Witel | 0,442 | 2,170 | 0,204 |
| Quadrant | 0 | 0 | 0 |
| Existing | 0,363 | -2,372 | -0,153 |
| R2BB | 0,041 | 1,278 | 0,032 |
| Network Type | 0,027 | 1,382 | 0,020 |
| Zone | 0,001 | 0,359 | 0,004 |
| Migration | 0,058 | 0 | 0 |

**Table 4.** Performance Comparison

| Attribute | NBC | C4.5 |
|---|---|---|
| Accuracy | 85,08% | 88,61% |
| AUC | 0,841 | 0,870 |

Table 4 explains that the Naive Bayes algorithm produces an accuracy value of 85.08% while the C4.5 algorithm produces an accuracy value of 88.61%. With the resulting value of ¡0.05 which indicates that there is an improvement of the two algorithms used where the C4.5 algorithm is better in cross-selling marketing strategies.

## 5. Conclusion

The research on cross selling marketing strategies with data mining methods using C4.5 algorithm has a good accuracy value of 88.61% and AUC 0.870 for the classification of customer data that has the potential to be more effective and efficient in cross-selling marketing strategies. However, this research does not use attribute selection and for the future research can be done by using attribute selection or other algorithms

## References

[1] T. Yu, K. de Ruyter, P. Patterson, and C. F. Chen, "The formation of a cross-selling initiative climate and its interplay with service climate," Eur. J. Mark., vol. 52, no. 7–8, pp. 1457–1484, 2018, doi: 10.1108/EJM-08-2016-0487.

[2] X. K. Jiang and X. Chen, "Research on prediction model of the impact of new telecom services tariff based on the customer choice behavior," Adv. Mater. Res., vol. 765–767, pp. 3249–3252, 2013, doi: 10.4028/www.scientific.net/AMR.765-767.3249.

[3] P. Surya Sumartha, F. Samopa, P. S. Sumartha, and F. Samopa, "Cross Selling Product Bundling Based On Customer Satisfaction Study Case Meat ; Food Suplier X," Int. J. Educ. Res., vol. 5, no. 1, pp. 241–252, 2017.

[4] M. Karim and R. M. Rahman, "Decision Tree and Naive Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," J. Softw. Eng. Appl., vol. 06, no. 04, pp. 196–206, 2013, doi: 10.4236/jsea.2013.64025

[5] A. Alkasem, H. Liu, and M. Shafiq, "Improving fault diagnosis performance using hadoop mapreduce for efficient classification and analysis of large data sets," J. Comput., vol. 29, no. 4, pp. 185–202, 2018, doi: 10.3966/199115992018082904015.

[6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques. 2016.

[7] M. Bramer, Principles of Data Mining. London: Springer, 2007.
[8] S. Singh and P. Gupta, "Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: a Survey," Int. J. Adv. Inf. Sci. Technol., vol. 27, no. 27, pp. 97–103, 2014, doi: 10.15693/ijaist/2014.v3i7.47-52.
[9] J. Han and M. Kamber, Mining Stream, Time-Series and Sequence Data, vol. 54. 2006.
[10] H. Ahn, J. J. Ahn, K. J. Oh, and D. H. Kim, "Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques," Expert Syst. Appl., vol. 38, no. 5, pp. 5005–5012, 2011, doi: 10.1016/j.eswa.2010.09.150.
[11] M. A. K. Anshary and B. R. Trilaksono, "Tweet-based target market classification using ensemble method," J. ICT Res. Appl., vol. 10, no. 2, pp. 123–139, 2016, doi: 10.5614/itbj.ict.res.appl.2016.10.2.3.
[12] F. Gorunescu, Data Mining Concepts, Models and Techniques. Berlin: Springer, 2011.
[13] N. Abdulla, Z. Ahmed, M. Gazzali, S. G. Nair, and A. Khade, "Measure Customer Behaviour using C4.5 Decision Tree Map Reduce Implementation in Big Data Analytics and Data Visualization," IJIRST - Int. J. Innov. Res. Sci. Technol., vol. 1, no. 10, pp. 228–235, 2015