Utilization of Prediction Data for Prospective Decision Customers Insurance Using the Classification Method of C.45 and Naive Bayes Algorithms

Saruni Dwiasnati*, Yudo Devianto
Faculty of Computer Science, Universitas Mercubuana, West Jakarta, Indonesia

*saruni.dwiasnati@mercubuana.ac.id

Abstract. Prospective Insurance Customers are one or several who buy physical products with various variables used in their determination such as price, quality, quantity of the products, places and services that can be provided based on results to achieve the desired results. Customers are differentiated into individual customers or group customers who buy physical products with various variables used in their determination such as price, quality, quantity of the product, place and service that can be provided based on results to achieve the desired results. Data mining is based on the process of creating a patterned pattern in the future and for the information provided by the data provided by the data. The research we conducted uses the C4.5 Algorithm and the Naive Bayes Algorithm using the Sex, Marital, Age, Region, Job and Description Indicators. This research uses Rapidminner tools as a medium to test the data being carried out.

1. Introduction
Business growth in the era of globalization and advances in the field of information technology are rapidly giving considerable influence both in industry and services. It also brings a big change in the level of competition between companies, so that the perpetrators of the company must always create various techniques to survive.

Telemarketing can be defined as the systematic use of telephone to achieve your business goals.

Based on previous research to predict the decision of prospective insurance customers by Sunjana using the classification method with the C45 algorithm, the authors want to conduct research to estimate considerations for potential insurance customers using the classification method with the C 4.5 and Naive Bayes algorithms. Referring to the background of the problems in this study, the problems can be identified as follows:

- There has not been a more in-depth analysis of the data that has been contacted for other product offerings. So that the data can be maximally used.
- It is necessary to analyze data on prospective insurance customers using the C4.5 and Naive Bayes algorithm methodologies.
Based on the identification of the problems that have been described, several problems are formulated as follows: how to compare C4.5 and Naive Bayes algorithms in predicting the Decision of Prospective Insurance Customers?

Based on the formulation of the problem, the purpose of this research is: knowing the high level of accuracy by comparing two algorithms, namely C4.5 and Naive Bayes in predicting the decision of prospective insurance customers, this research is also expected to be used as reference material related to the subject of sales predictions.

From the results of this study it is expected that:

- Increase understanding and experience in using the Classification method with C4.5 and Naive Bayes algorithms in mapping or grouping data.
- Increase prediction accuracy by comparing 2 Classification methods as the output produced.
- Produce algorithms with more accurate accuracy to predict data on the decision of potential disaster insurance customers, so that further research can further improve its accuracy.

2. Theory basis and methods

2.1. Understanding Data Mining

Data mining is one part that is at the stage in Knowledge Discovery in Databases (KDD). Steps in generating new knowledge, among others, namely: first clarification of data you want to process (data cleaning), Combine processed data (data integration), data determination (data selection), data transparency (data transformation), evaluation of patterns that have been made beforehand (evaluation of patterns) and presentation of new knowledge needed (presentation of knowledge). Data mining is guided by the process of mining (mining) data from selected cases to produce new knowledge from very large data [16]. The stages of the process of extracting data consist of three stages, namely data collection (data collection), data transformation (data transformation), and data analysis (data analysis).

2.2. Algorithm C 4.5

The C 4.5 algorithm is the algorithm used to produce a decision tree developed by Ross Quinlan. The basic idea of this algorithm is the making of a decision tree based on the selection of attributes that have the highest priority or can be called having the highest gain value based on the entropy value of the attribute as the attribute axis classification. At the stage C4.5 algorithm has 2 working principles [9], namely: Making decision trees, and making rules (model rules). Rules of rules that are formed from decision trees will form a condition in the form of if then. [14]

The steps in generating a decision tree in the C 4.5 algorithm (Larose, 2005), namely:

1. Prepare training data first. Training data is usually taken from historical data that has occurred before or called past data and has been grouped in certain classes.
2. Calculate parents. parents will be taken from the selected attribute, by calculating the gain value of each attribute, the highest gain value that will be the first parent. Before calculating the attribute gain value, first calculate the value of the entropy. To calculate the entropy value, the formula is used:

\[
Entropy(S) = \sum_{i=1}^{n} - p_i \log_2 p_i
\]

Information:

\[
S = \text{case set}
\]

\[
n = \text{number of partitions } S
\]

\[
p_i = \text{proportion of } S_i \text{ to } S
\]

3. Then calculate the gain value using the formula:

\[
Gain(S, A) = entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{S} \times Entropy(S_i)
\]
4. Repeat step 2 and step 3 until all records are specified.
5. The decision tree partition process will stop when:
   a. all the records in N knots get the same class.
   b. There is no attribute in the partitioned record again
   c. There is no record in the empty branch

2.3. Naive Bayes
Naive Bayes algorithm is one of the algorithms found in classification techniques. Naive Bayes is a classification with probability method and statistics presented by British scientist Thomas Bayes, which predicts future opportunities based on past experience so that it is known as Bayes Theorem. The theorem is combined with Naive where it is assumed that the conditions between attributes are mutually independent. The Naive Bayes classification assumes that there is no particular feature of a class that has no relation to the characteristics of other classes. The equation of the Bayes theorem is:

\[
P(\text{H} | \text{X}) = \frac{P(\text{X}|\text{H}) \cdot P(\text{H})}{P(\text{X})}
\]

3. Results and Discussion
3.1. Calculation of C4.5 Algorithm
According to Kusrini in the Data Mining Algorithm which is used basically for a C 4.5 algorithm [14] in building a decision tree can be explained as follows:
   1. Select the attribute that has the highest result as root as a benchmark in making a decision tree
   2. Make the branches needed for each value that exists
   3. In the case of branches, repeat the process for each branch until all cases in the branch have the same class in accordance with the provisions.

In the first step of the C 4.5 algorithm calculation process, this is to determine the attributes first as the root. The selected attribute is the one that has the highest gain value from the gain-gain value of the other attribute, then becomes the root of the tree. The calculation to produce the entropy and gain values will be explained in table I as follows:
Table 1. Calculations determine Entropy and Gain

| Sample data is used to calculate entropy and gain as much as 21 and from the entropy and gain values obtained in table IV-I, then specify the next node, namely node 1.1, and calculate the entropy and gain of each attribute from the information. The number of cases calculated is from the information node and so on until all records in the node get the same class. When using training data to make a node, the following in Figure IV-1 is the test results with Rapid Miner tools for occupation so that the predictor attribute is determined directly by all other attributes. |

3.2. Calculation of Naive Bayes

The variables used in the classification of insurance customer data include:
1. Gender Gender is grouped into two categories [male | female].
2. Marriage status is grouped into two categories [Not Married | Married].
3. Customer age is grouped into three categories [20–29 | 30–40 | > 40].
4. Customer areas are grouped into five categories [North Jakarta, East Jakarta, South Jakarta, West Jakarta and Central Jakarta].
5. Customer work is grouped into three categories [PNS | private employees | self-employed].
6. Information Time is grouped into four categories [Morning, Afternoon, Afternoon and Evening].

Customer table data as follows:

3.3. Evaluation and Validation

The results of the model tests that have been carried out are the C 4.5 and Naive Bayes algorithms, testing the level of accuracy using the confusion matrix and ROC / AUC (Area Under Cover) curve.

3.3.1. Confusion Matrix algorithm C4.5

Confusion Matrix data training

Table 2 is a calculation of the accuracy of training data using the C4.5 algorithm which produces an accuracy of 75.27%. It is known that training data consists of 133 data records, 56 classified data NO and 77 data predicted Buy insurance.

| Table 2. Model Confusion Matrix algorithm C4.5 |
| --- | --- | --- |
| true No | true Buy | class precision |
| pred. No | 30 | 7 | 81.08% |
| pred. Buy | 26 | 70 | 72.92% |
| class recall | 53.67% | 60.01% |
Figure 1. Text view C4.5 algorithm Confusion Matrix model

Figure 1 is a calculation of data accuracy using C4.5 algorithm. The training data is known to consist of 133 data records, 30 data classified as No and 26 data predicted No but apparently Buy, 7 classified Buy data turns out that No and 70 data are correctly classified Buy.

3.3.2. Confusion Matrix algorithm Naive Bayes

Confusion Matrix data training

Table 4 is a calculation of the accuracy of training data using the Naive Bayes algorithm which produces 90.11% accuracy. It is known that training data consists of 133 data records, 56 classified data NO and 77 data predicted Buy insurance.

Figure 2. Text view of the Naive Bayes Confusion Matrix algorithm model
Figure 2 is a calculation of data accuracy using the Naive Bayes algorithm. It is known that training data consists of 133 data records, 46 data classified as No and 10 data predicted No but apparently Buy, 3 classified data Buy turns out that No and 74 data are correctly classified Buy.

**ROC curve**

The calculation results are visualized by the Receiver Operating Characteristic (AUC) curve. ROC has a level of diagnostic value (Gorunescu, 2011):

- a. Accuracy is 0.90 - 1.00 = excellent classification
- b. Accuracy is 0.80 - 0.90 = good classification
- c. Accuracy is 0.70 - 0.80 = fair classification
- d. Accuracy is 0.60 - 0.70 = poor classification
- e. Accuracy is 0.50 - 0.60 = failure

The results obtained from ROC processing using training data for C4.5 algorithm of 0.775 can be seen in Figure 3 with the level of diagnosis of fair classification.

The results obtained from the ROC processing for the Naive Bayes algorithm using training data of 0.961 can be seen in Figure 4 with a good diagnosis of classification.
3.4. Prototype of Selected Algorithm

From the results of the evaluation and validation above it can be seen that the Naive Bayes Algorithm has a good level of accuracy and performance, so that the rules generated by the Naive Bayes algorithm can be used as a rule for making prototypes that can facilitate predicting the decisions of prospective insurance customers.

![Flowchart Prototype Predicts Prospective Insurance Customers](image)

Figure 5 Flowchart Prototype Predicts Prospective Insurance Customers

4. Conclusion

Based on the discussion of the results of the research discussed in the previous chapter, in the study the Utilization of Predicted Data for Prospective Insurance Customers using the C4.5 and Naive Bayes Classification Methods can be concluded as follows:

1. The Naive Bayes algorithm has the highest accuracy rate of 90.11% while C4.5 is 75.27%, the difference between them is 15%. Naive Bayes algorithm model has AUC of 0.961 and C4.5 0.775, from the AUC value, the Naive Bayes algorithm belongs to the category of excellent classification and C4.5 fair classification, so the Naive Bayes algorithm can be implemented in determining potential insurance customers.

2. The rule generated by the Naive Bayes algorithm is applied in the prototype of the prediction of prospective insurance customers with the accuracy of prototype verification testing of 75.12%. Based on the accuracy generated by the prototype shows that the methods and prototypes that are applied are good in predicting prospective insurance customers.
5. References

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

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

[3] Vercellis, C., 2009. Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley & Sons, Inc.

[4] Sugiyono. Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta, 2012

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

[6] Larose, D. T. (2005). Discovering Knowledge in Data. New Jersey: John Willey & Sons, Inc. ISBN 0-471-66657-2.

[7] Liao, T. W., & Triantaphyllou, E. (2007). Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications. Series on Computers and Operations Research, Vol 6. USA: World Scientific. ISBN-13 978-981-277-985-4. ISBN-10 981-277-985-X

[8] Kusrnri 2009 Kusrmni. Algoritma Data Mining. Jakarta: Andi, 2009.

[9] Kotler 2009, 245 Batubara, Muhmad Hendri. Strategi Marketing Public Relation (MPR) Berupa Promosi Dan Sponsorsip Untuk Mempengaruhi Konsumer Dalam Keputusan Pembelian (Studi Produk Perawatan Bayi Johns ons Baby). Tes., Universitas Indonesia, 2010.

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

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

[12] Faizah Restu, Saputra Elvis, Adhiguna Dawam, Studi Identifikasi Mitigasi Bencana Gempa Pada Bangunan Sekolah Dasar Kaligondang Dan Rekomendasi Perbaikan, Jurnal REKAYASA SIPIL, Universitas Mercu Buana 2017, Vol 6 No. 2. p-ISSN: 2252-7699 e-ISSN: 2598-5051

[13] Muhammad Fauzul Arifin, Devi Fitrianah, “Penerapan Algoritma Klasifikasi C4.5 dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada”, IncomTech, Jurnal Telekomunikasi dan Komputer, vol.8, no.2, ISSN 2085-4811, eISSN: 2579-6089, 2018

[14] Dwiasnati Saruni, Yudo Devianto, " Naive Bayes Optimization Based On Particle Swarm Optimization to Predict the Decision of Insurance Customer Candidate " Vol. 5 - Issue 5 (08 - 14) - September - October 2018 , International Journal of Computer Techniques (IJCT) , ISSN: 2394 - 2231 , www.ijctjournal.org

[15] Firdaus Diky, “Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer” Jurnal FORMAT Vol 6 No. 2 Tahun 2017, ISSN : 2089 -5615 Hal 91-97

Acknowledgment

We say praise and gratitude to God for all His blessings and blessings so that the writing of scientific articles can be resolved properly. We do not forget to thank our family Faculty of Computer Science, Universitas Mercu Buana, and especially PUSLIT Universitas Mercubuana who have provided funding support in writing this scientific article so that it can be resolved properly.