Analysis of data mining classification by comparison of C4.5 and ID algorithms

R. Sudrajat¹, I. Irianingsih¹, D. Krisnawan²

¹Department of Computer Science, Padjadjaran University, Sumedang 45363, Indonesia
²Department of Mathematics, Padjadjaran University, Sumedang 45363, Indonesia

E-mails: r.sudrajat@unpad.ac.id, iinirianingsih@gmail.com, dillan.k@gmail.com

ABSTRACT. The rapid development of information technology, triggered by the intensive use of information technology. For example, data mining widely used in investment. Many techniques that can be used assisting in investment, the method that used for classification is decision tree. Decision tree has a variety of algorithms, such as C4.5 and ID3. Both algorithms can generate different models for similar data sets and different accuracy. C4.5 and ID3 algorithms with discrete data provide accuracy are 87.16% and 99.83% and C4.5 algorithm with numerical data is 89.69%. C4.5 and ID3 algorithms with discrete data provides 520 and 598 customers and C4.5 algorithm with numerical data is 546 customers. From the analysis of the both algorithm it can classified quite well because error rate less than 15%.

1. Introduction
Investment is currently growing very rapidly, public is interesting to do a form of investment such as deposits and mutual funds. On the other hand, the development of information technology has developed rapidly, many science informatics used in the banking sector, one of them is data mining. Data mining is used to explore large amounts of data into patterns/rules so as to simplify large data into parts more modest but has a precious value. C4.5 and ID3 algorithms are algorithms used in data mining to create decision tree, where decision tree is a classification method with an accurate prediction. Decision tree method to change the very large data into a decision tree that represents the rule.

1.1 Decision Tree
Classification by decision tree, a classification that is very reliable as a means of classification and prediction. In general, a decision tree is a modeling picture of a subject that is composed of a series of decisions that provide a solution. Classification rules can be easily presented in easy language so that users can understand it, or in its database as Microsoft Excel, Microsoft Access and SQL. In some applications, and prediction accuracy of classification is important, for example, predict potential customers in the business market. There are various algorithms to build a decision tree, the most popular is the ID3 and C4.5.

1.2 C4.5 Algorithm
C4.5 algorithm is an algorithm to form a decision tree by counting the value of the gain, where the biggest gains are to be used as an initial node or the root node. C4.5 algorithms step in building a decision tree as follows:
Select the attribute with the largest gain value as the root.
Create a branch for each value.
For the case of the branches.
Repeat the process for each branch until all cases the branches have the same class.

Entropy and gain search formula as follows:

\[
Gain(S) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i)
\]

where
- \( S \) = the set of cases
- \( n \) = number of partitions
- \(|S_i|\) = number of cases in the partition \( i \)
- \(|S|\) = number of cases in \( S \).

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

where
- \( S \) = The set of cases
- \( n \) = number of partitions
- \( p_i \) = proportion of \( S_i \) to \( S \).

1.3 ID3 Algorithm
ID3 (Iterative Dichotomiser3) algorithm is used to form a decision tree by calculating the entropy values, created by Ross Quinlan. Step ID3 algorithm to construct a decision tree as follows:
- Select all the existing attributes and calculate the entropy of each attribute.
- Select an attribute with the smallest entropy, then these attributes be the root node.
- Create an attribute node with the smallest entropy.
- Repeat the process for each branch until all cases the branches have the same class.

2. Analysis and System Design

2.1 Analysis Classification System
In the analysis stage of this system, which will be the topic of discussion is about what the problems faced as an analysis of the data and the accuracy of the analysis of algorithms C4.5 and ID3.

2.1.1 Data Analysis
The data used in the making of this program is the customer data a bank with attribute goal is PEP (Portfolio Evaluation Plan) or the planned investment, the investment data in the form of excel data which attributes the following selection \( id, age, sex, region, income, married, children, car, save act, current act and mortgage, \) and the attributes of the destination is pep. Total investment data as many as 600 data, the data is stored in csv (Comma Separated Values) format.

2.1.2 Analysis Software
Some of the software used in this study as follows: (1) The operating system used Microsoft Windows Vista Home Premium, (2) Net Beans Java programming as the software used to build the program, (3) Microsoft Excel as software that serves to change the data in .xls to .csv.

2.2 Designing Data Classification
Classification of data that will be discussed is divided into three parts, namely the classification of data using ID3 algorithm with discrete data, the algorithm C4.5 and C4.5 algorithms discrete data with numerical data. Data used in the calculation of investment manual 10 records.
### Table 1. Experiment Data 1

| ID | AGE   | SEX | REGION     | INCOME | MARRIED | CAR |
|----|-------|-----|------------|--------|---------|-----|
| 1  | MIDDLE| F   | INNER_CITY | LOW    | NO      | NO  |
| 2  | MIDDLE| F   | TOWN       | LOW    | YES     | NO  |
| 3  | MIDDLE| F   | RURAL      | LOW    | NO      | YES |
| 4  | YOUNG  | M   | INNER_CITY | LOW    | YES     | NO  |
| 5  | OLD   | F   | INNER_CITY | HIGH   | YES     | NO  |
| 6  | OLD   | F   | TOWN       | MID    | YES     | NO  |
| 7  | YOUNG  | M   | RURAL      | LOW    | NO      | NO  |
| 8  | OLD   | M   | TOWN       | MID    | YES     | NO  |
| 9  | MIDDLE| F   | SUBURBAN   | MID    | YES     | YES |
| 10 | OLD   | M   | INNER_CITY | HIGH   | YES     | NO  |

### Table 2. Experiment Data 2

| CHILDREN | SAVE ACT | CURRENT ACT | MORTGAGE | PEP |
|----------|----------|-------------|----------|-----|
| ONE      | NO       | NO          | NO       | YES |
| ONE      | NO       | YES         | NO       | YES |
| NOL      | NO       | YES         | NO       | YES |
| NOL      | YES      | YES         | NO       | YES |
| NOL      | YES      | NO          | NO       | NO  |
| TWO      | YES      | YES         | NO       | YES |
| NOL      | NO       | YES         | NO       | YES |
| NOL      | NO       | YES         | YES      | YES |
| TWO      | NO       | NO          | NO       | NO  |
| TWO      | YES      | YES         | YES      | YES |

### Table 3. Gain Value Calculation and Entropy

| Attribute   | Sub-attribute | Number of Cases (S) | NO (S1) | YES (S2) | Entropy          | Gain          |
|-------------|---------------|---------------------|---------|----------|------------------|---------------|
| Total       |               | 10                  | 2       | 8        | 0.7219280946     | 0.0729055950 |
| AGE         | MIDDLE        | 4                   | 1       | 3        | 0.8112781245     | 0.1709505941 |
|             | YOUNG         | 2                   | 0       | 2        | 0                | 0.1709505942 |
|             | OLD           | 4                   | 1       | 3        | 0.8112781245     | 0.2464393444 |
| SEX         | FEMALE        | 6                   | 2       | 4        | 0.9182958341     | 0.1177436965 |
|             | MALE          | 4                   | 0       | 4        | 0                | 0.0854752971 |
| REGION      | INNER_CITY    | 3                   | 0       | 3        | 0                |              |
|             | TOWN          | 3                   | 1       | 2        | 0.9182958341     |              |
|             | RURAL         | 3                   | 1       | 2        | 0.9182958341     |              |
|             | SUBURBAN      | 1                   | 0       | 1        | 0                |              |
| INCOME      | Low           | 5                   | 0       | 5        | 0                |              |
|             | Mid           | 3                   | 1       | 2        | 0.9182958341     |              |
|             | High          | 2                   | 1       | 1        | 1                |              |
| MARRIED     | YES           | 7                   | 2       | 5        | 0.8631205687     | 0.1177436965 |
|             | NO            | 3                   | 0       | 3        | 0                | 0.0854752971 |
| CHILDREN    | NOL           | 5                   | 1       | 4        | 0.7219280946     |              |
|             | ONE           | 2                   | 0       | 2        | 0                |              |
|             | TWO           | 3                   | 1       | 2        | 0.9182958341     |              |
### Table 4. Calculation of Node Root for C4.5 Algorithm

| Attribute     | Sub-attribute | Number of Cases (S) | NO (S1) | YES (S2) | Entropy   | Gain       |
|---------------|---------------|---------------------|---------|----------|-----------|------------|
| INCOME-MID    |               | 3                   | 1       | 2        | 0.9182958341 | -          |
| AGE           | MIDDLE        | 1                   | 1       | 0        | 0         |            |
|               | YOUNG         | 0                   | 0       | 0        | 0         |            |
|               | OLD           | 2                   | 0       | 2        | 0         |            |
| SEX           | FEMALE        | 2                   | 1       | 1        | 1         | 0.2516291674 |
|               | MALE          | 1                   | 0       | 1        | 0         |            |
| REGION        | INNER_CITY    | 0                   | 0       | 0        | 0         | 0.2516291674 |
|               | TOWN          | 2                   | 1       | 1        | 1         |            |
|               | RURAL         | 0                   | 0       | 0        | 0         |            |
|               | SUBURBAN      | 1                   | 0       | 1        | 0         |            |
| MARRIED       | YES           | 3                   | 1       | 2        | 0.9182958341 | 0          |
|               | NO            | 0                   | 0       | 0        | 0         |            |
| CHILDREN      | NOL           | 1                   | 0       | 1        | 0         | 0.2516291674 |
|               | ONE           | 0                   | 0       | 0        | 0         |            |
|               | TWO           | 2                   | 1       | 1        | 1         |            |
|               | THREE         | 0                   | 0       | 0        | 0         |            |

### Table 5. Calculation of Node Root for ID3 Algorithm

| Attribute     | Sub-Attribute | Number of Cases (S) | NO (S1) | YES (S2) | Entropy   |
|---------------|---------------|---------------------|---------|----------|-----------|
| CHILDREN-NOL  |               | 5                   | 1       | 4        | 0.7219280946 |
| Age           | Middle        | 1                   | 0       | 1        | 0         |
|               | Young         | 2                   | 0       | 2        | 0         |
|               | Old           | 2                   | 1       | 1        | 1         |
| Sex           | Female        | 2                   | 1       | 1        | 1         |
|               | Male          | 3                   | 0       | 3        | 0         |
| Region        | Inner_city    | 1                   | 0       | 1        | 0         |
|               | Town          | 1                   | 0       | 1        | 0         |
|               | Rural         | 3                   | 1       | 2        | 0         |
|               | Suburban      | 0                   | 0       | 0        | 0         |
| Married       | Yes           | 3                   | 1       | 2        | 0.9182958341 |
|               | No            | 2                   | 0       | 2        | 0         |
| Income        | Low           | 3                   | 0       | 3        | 0         |
|               | Mid           | 1                   | 0       | 1        | 0         |
|               | High          | 1                   | 1       | 0        | 0         |
From the calculation of Table 4 it is obtained the decision tree as follows:

![Figure 1](https://example.com/figure1.png)

**Figure 1.** C4.5 Decision Tree Algorithm

| Attribute         | Sub-Attribute | Number of Cases (S) | NO (S1) | YES (S2) | Entropy    |
|-------------------|---------------|---------------------|---------|----------|------------|
| CHILDREN-         | NO            | 3                   | 1       | 2        | 0.9182958341|
| MARRIED-          | YES           |                     |         |          |            |
| Age               |               |                     |         |          |            |
| Middle            |               | 0                   | 0       | 1        | 0          |
| Young             |               | 1                   | 0       | 1        | 0          |
| Old               |               | 2                   | 1       | 1        | 1          |
| Sex               |               |                     |         |          |            |
| Female            |               | 1                   | 1       | 0        | 0          |
| Male              |               | 2                   | 0       | 2        | 0          |
| Region            |               |                     |         |          |            |
| Inner_city        |               | 1                   | 0       | 1        | 0          |
| Town              |               | 1                   | 0       | 1        | 0          |
| Rural             |               | 1                   | 1       | 0        | 0          |
| Suburban          |               | 0                   | 0       | 0        | 0          |
| Income            |               |                     |         |          |            |
| Low               |               | 1                   | 0       | 1        | 0          |
| Mid               |               | 1                   | 0       | 1        | 0          |
| High              |               | 1                   | 1       | 0        | 0          |

From the calculation of Table 6 it is obtained the decision tree as follows:

![Figure 2](https://example.com/figure2.png)

**Figure 2.** Decision Tree ID3 Algorithm

### 3. Results and Discussion

In this experiment, use the data as much as 100 records, the data is tested again using ID3 algorithm with discrete data, the algorithm C4.5 and C4.5 algorithms discrete data with continuous data. The results of these experiments obtained a decision tree as follows:
Figure 3. Decision Tree ID3 Algorithm with Discrete Data

Figure 4. Decision Tree C4.5 Algorithm with Discrete Data
Figure 5. Decision Tree C4.5 Algorithm with Continuous Data

Table 7. Data classification results

|                  | C4.5 Algorithm | Total  | ID3 Algorithm | Total |
|------------------|----------------|--------|---------------|-------|
| YES              | 37             | 42     | 43            | 41    |
|                  | 42             | 39     | 244           | 41    |
|                  | 49             | 44     | 45            | 49    |
|                  | 44             | 44     | 272           |       |
| NO               | 53             | 50     | 52            | 40    |
|                  | 40             | 41     | 276           | 59    |
|                  | 52             | 56     | 56            | 52    |
|                  | 52             | 51     | 326           |       |
| Total            | 520            |        | 598           |       |
| Unclassified     | 80             |        | 2             |       |

Table 8. Classified Customers with Discrete Data

|                  | C4.5 Algorithm | Total |
|------------------|----------------|-------|
| YES              | 35             | 43    |
|                  | 44             | 43    |
|                  | 43             | 37    |
|                  | 245            |       |
| NO               | 56             | 46    |
|                  | 52             | 50    |
|                  | 52             | 45    |
|                  | 301            |       |
| Total            | 546            |       |
| Unclassified     | 54             |       |

Table 9. Classified Customers with Continuous Data

| Trial | ID3 with discrete data (%) | Failure (%) | C4.5/J48 with discrete data (%) | Failure (%) | C4.5/J48 with numeric data (%) | Failure (%) |
|-------|----------------------------|-------------|---------------------------------|-------------|-------------------------------|-------------|
| 1     | 100                        | 0           | 93                              | 7           | 93                            | 7           |
| 2     | 100                        | 0           | 91                              | 9           | 88                            | 12          |
| 3     | 99                         | 1           | 94                              | 6           | 95                            | 5           |
| 4     | 100                        | 0           | 80                              | 20          | 92                            | 8           |
| 5     | 100                        | 0           | 81                              | 19          | 84                            | 16          |
| 6     | 100                        | 0           | 84                              | 16          | 86                            | 14          |
| MEAN  | 99.83                      | 0.16        | 87.16                           | 12.83       | 89.69                         | 10.33       |

4. Conclusion

1. ID3 Algorithm excellent classifying investment data with a discrete form for 99.83% of the data can be classified by degree of error 0.16%, while the C4.5 algorithm to classify well enough to form discrete investment data for 87.16% of the data can be classified by
an error rate of 12.83% and 89.69% in numerical form with an error rate of 10.33%. Algorithm ID3 classifying investment data better than C4.5 algorithm.

2. ID3 algorithm to classify as many as 598 discrete data to unclassified 2, the algorithm C4.5 with as many as 520 discrete data classifying the unclassified 80 and C4.5 algorithms with continuous data to classify as many as 546 Unclassified 54 customers.

3. The decision tree obtained in this study, has a decision tree ID3 algorithm is more complex than the C4.5 algorithm, so it can be viewed accuracy of ID3 algorithm can better classify. The more complex a decision tree algorithm, the better classifying data.

References
[1] Berry, Michael J.A. and Gordon S. Linoff. 2004. Data Mining Techniques for Marketing. Second Edition. Wiley Publishing, Inc.
[2] Chen, M., Han, J. and Yu, P. 1996. Data Mining: An Overview from a Database Perspective. IEEE Trans. Knowledge and Data Engineering. 8(6): 866-881.
[3] Larose, Daniel T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Willey & Sons, Inc.
[4] Mobasher, Bamshad. 2005. Clasification via Decision Trees in WEKA. http://maya.cs.depaul.edu/classes/ect584/weka/classify.html.
[5] Moertini, Veronica S. towards the Use of C4.5 Algorithm for Classifying Banking Dataset. http://home.unpar.ac.id/~integral/Volume%208/Integral%208%20No.%202/C45%20Algorithm.
[6] Sudjana. 2010. Klasifikasi Data Nasabah Sebuah Asuransi Menggunakan Algoritma C4.5. http://journal.uii.ac.id/index.php/Snati/article/view/1923.
[7] Sunni, Ismail. 2011. Perbedaan Algoritma ID3, C4.5 dan J48. http://webcache.googleusercontent.com/search?q=cache:http://codemath.wordpress.com/2011/06/20/perbedaan-algoritma-id3-c4-5-dan-j48.
[8] Quinlan, Ross. 2000. ID3 Algorithm. http://www.soc.napier.ac.uk/~peter/vldb/dm/node11.html.