APPLICATION OF THE C5.0 ALGORITHM TO DETERMINE GOOD OR BAD ON 5S AUDIT RESULTS

Oleh:
Indra Aliyudin [1)]
Ari Purno Wahyu [2)]
Universitas Widyatama, Bandung [1,2)]
E-mail:
indra.aliyudin@widyatama.ac.id [1)]
aripurnowahyu@gmail.com [2)]

ABSTRACT

Artificial Intelligence is currently growing and is widely used in various aspects of life in society. Likewise in today's corporate environment, we must be good at managing all activities so that AI can help in lightening and streamlining decision making at work. In terms of lightening this work, it is in the aspect of data management and data analysis. AI provides many methods and ways to analyze data so that the data can be used as a reference for employee self-assessment or even as a determinant of a company's business going forward. This study discusses the C5.0 algorithm which is implemented or tested against the 5S (Short, Set in Order, Shine, Standardize and Sustain) audit data set obtained from the company P.T. Bekaert Indonesia. This study uses two types of methods from the C5.0 algorithm model as a reference, namely the tree-based model and the rule-based model, besides that this study uses the cross fold validation method which is expected to increase the level of accuracy of the results of this study. This study was conducted aiming to find out whether the C5.0 algorithm can be implemented on the 5S audit result data set and has high accuracy or not. With the data collection method, analysis was carried out using RStudio software and the R programming language, this study shows that determining the good and bad 5S in an area can be done with the C5.0 algorithm with a tree-based model or a rule-based model and produces high accuracy.

Keyword: Artificial Intelegence; C5.0 Algorithm, Classification, Data Mining, Short, Set In Order, Shine, Standardize And Sustain

ABSTRAK

Kecerdasan Buatan saat ini semakin berkembang dan banyak digunakan dalam berbagai aspek kehidupan di masyarakat. Begitu juga di lingkungan perusahaan saat ini, kita harus pandai mengelola semua aktivitas agar AI dapat membantu dalam meringankan dan mengefektifkan pengambilan keputusan di tempat kerja. Dalam hal meringankan pekerjaan ini adalah pada aspek pengelolaan data dan analisis data. AI menyediakan banyak metode dan cara untuk menganalisis data sehingga data tersebut dapat digunakan sebagai referensi penilaian diri karyawan atau bahkan sebagai penentu bisnis perusahaan ke depan. Penelitian ini membahas algoritma C5.0 yang diimplementasikan atau diuji terhadap set data audit 5S (Short, Set in Order, Shine, Standardize and Sustain) yang diperoleh dari perusahaan P.T. Bekasi Indonesia. Penelitian ini menggunakan dua jenis metode dari model algoritma C5.0 sebagai acuan yaitu model tree-based dan model rule-based, selain itu penelitian ini menggunakan metode cross fold validation yang

diharapkan dapat meningkatkan tingkat akurasi. dari hasil penelitian ini. Penelitian ini dilakukan dengan tujuan untuk mengetahui apakah algoritma C5.0 dapat diimplementasikan pada dataset hasil audit 5S dan memiliki akurasi yang tinggi atau tidak. Dengan metode pengumpulan data, analisis dilakukan dengan menggunakan software RStudio dan bahasa pemrograman R, penelitian ini menunjukkan bahwa menentukan baik buruknya 5S pada suatu area dapat dilakukan dengan algoritma C5.0 dengan model atau rule berbasis pohon. model berbasis dan menghasilkan akurasi yang tinggi.
Kata Kunci: Kecerdasan Buatan, Algoritma C5.0, Klasifikasi, Penambangan Data, Pendek, Diatur Dalam Rangka, Bersinar, Standarisasi Dan Mempertahankan

1. INTRODUCTION

Technological developments continue to increase along with the times, one of which is Artificial Intelligence which is able to assist in facilitating human work.

Artificial intelligence or AI is a technology developed to be able to learn, think and work like humans based on data (Samek *et al*, 2017). This means that AI can learn from existing data for further data training as a learning process. Data Mining are some examples of AI that can help human work.

Data mining can be defined as the process of searching for unknown or unexpected data patterns. Data mining is one of the stages in the whole process of knowledge discovery in databases. In general, there are several data mining techniques, one of the data mining techniques is classification. One of the classification techniques is a decision tree.

These artificial intelligence techniques can be utilized and implemented by various parties, both individuals and companies. Many companies have used this technique to later be combined with other disciplines. In addition, in large companies, technology must also develop so that the company's image becomes better because it can keep up with the times. On the other hand, the use of technology such as AI can help reduce the burden on workers and of course can reduce the company's cost because it can be cheaper to use AI than having to pay employees. Things that are commonly used in the implementation of data mining are to analyze CRM, customer segmentation and to be used to detect fraud or fraud detection.

Along with the development of technology, more and more research is being carried out both for the benefit of individuals and companies. There are a lot of research currently available, call it research on data mining, one of the things that can be implemented in a daily activity or formal activity in the company, for example research to assess employee

performance, research to assist consumers in choosing clothes in a shops and many other examples. The two examples utilize data mining with the decision tree method and use the C4.5 algorithm as well as the C5.0 algorithm in their research. After comparing between the two studies, it turns out that there are shortcomings in the C4.5 algorithm (Fajri, M. *et al*, 2022). Although the C5.0 algorithm is better than the C4.5 algorithm, the accuracy of the research results can still be improved with several additional methods, one of which is adding the cross fold validation method which can increase the level of accuracy for the better (Setyaning *et al,* 2020).

In this case the author is interested in conducting research related to data mining in assessing or determining an area in the company environment whether it is good or bad in the application of 5S (Sort, Set in Order, Shine, Standardized and Sustain) using the decision tree method with deeper development than the previous method, namely using a decision tree with a tree-based and rule-based method with 10 cross validation which is expected to use this deeper method to produce better accuracy.

2. LITERATURE REVIEW

From other research that has been done by Kastawan (2018), the purpose of this

According to Bhatia (2019), data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so they may be used in an enterprise's decision making. Likewise, according to Witten (2016) who says that data mining is the process of analyzing data from different sources and summing it up into information or knowledge or patterns that are important to increase profits, reduce costs, or even both.

In this paper, we will implement one of the functions of data mining, namely classification, as Bhatia (2019) said, classification is a classical method which is used by machine learning researchers and statisticians for predicting the outcome of unknown samples. It is used for categorization of objects (or things) into given discrete number of classes. Classification problems can be of two types, either binary or multiclass. In binary classification the target attribute can only have two possible values. Which is in line with this research which aims to determine between two things from various existing data.

study is to get the results of employee performance appraisals from several existing

attributes. He explained the performance of the C5.0 algorithm can produce a fairly high accuracy. This supports this research so that it can be done.

Another research that has been done by Setyaning (2020) which aims to analyze the factors that affect the timely graduation of informatics engineering students, she conducted research with the C5.0 algorithm with the addition of the k-fold cross validation method. In her research resulted in a fairly good accuracy.

In another study that discussed the comparison of several classification methods that had been carried out by Hadiwandra (2019) the decision tree method was mentioned to be the most robust method among other methods, besides that it was also stated that all methods had good scalability and were able to increase accuracy when given a large number of records. bigger. Therefore, in this study, the decision tree and rule-based classification methods will be used because they are more suitable for the data to be processed in this study.

3. METHODS

Some of the methods used in this research are as follows:

1. Problem Analysis

   In determining whether an area is good or bad, a method that can work quickly and precisely is needed. This research was conducted to be able to produce a method that can help to make decisions about an area whether it is said to be good or bad.

2. Data Collection

   In this study, data collection was carried out by taking the results of the 5R audit conducted at PT Bekaert Indonesia, especially the Maintenance Department.

3. Data Analysis

   To determine whether an area is good or bad, in this study an analysis will be carried out using the C5.0 Algorithm. The following is a description of the data that will be used:

Table 1. *Predictor Attribute Description*

Number	Predictor Attributes	Remarks
1	Auditor	Name of the person conducting the audit

2	TotalR1	Total point value of "Sort"
3	TotalR2	Total point value of "Set in Order"
4	TotalR3	Total point value of "Shine"
5	TotalR4	Total point value of "Standardized"
6	TotalR5	Total point value of "Sustain"
7	Comment	Comments from the auditor on the audited area

3.1. C5.0 Algorithm

The C5.0 algorithm is a decision tree-based algorithm which is a refinement of the ID3 and C4.5 algorithms formed by Ross Quinlan in 1987. The C5.0 algorithm can handle continuous and discrete attributes. The selection of attributes in this algorithm will be processed using information gain. The attribute with the highest Gain value will be selected as the root for the next node.

3.2. Confusion Matriks

Confusion matrix is a method for evaluation that uses a matrix table. The results of the evaluation using the confusion matrix produce accuracy values, as well as the error rate. Accuracy states the amount of data that is classified correctly after the testing process is carried out, while the error rate is used to calculate identification errors. To calculate the accuracy is as follows:

$$Accuration = \frac{TP + TN}{TP + FN + FP + FN} X100$$

Where TP is true positive, namely the amount of positive data that is correctly classified by the system, TN is true negative, namely the number of negative data that is correctly classified by the system, FN false negative is the amount of negative data but is classified incorrectly by the system and FP is false positive, i.e. the number of positive data but classified incorrectly by the system. The error rate can be calculated as follows:

$$Error\ Rate = \frac{the\ number\ of\ data\ identified\ incorrectly}{total\ data} X100\%$$

3.3. Cross Fold Validation

Cross validation or sometimes also called rotation-estimation or out-of-sample testing is one of a variety of similar model validation techniques to assess how statistical analysis results will generalize to independent data sets. Cross validation is a re-sampling method that uses different pieces of data to test and train the model at different iterations.

4. RESULTS AND DISCUSSION

The implementation of this research is carried out using data mining software, namely RStudio version 2021.09.2 Build 382. The dataset with the required attributes and classes has been collected in the form of files with comma delimited or .csv formats, for programming using the R language version 4.1.2.

In this study the author will test using two methods of the C5.0 algorithm classification, namely tree-based and rule-based. But before that below are the core steps of this test.

1. Step 1 : Process running required library.
2. Step 2 : Process of cleaning, reading data set and changing data type.
3. Step 3 : Process of randomizing the dataset, creating a model (determining predictors and attributes) and performing cross fold validation (including process of separated data to training-data and testing data).

```
treec5predict Bad Good
         Bad  177    3
         Good   5  207
```

4. Step 4 : Process all data in tree-based and rule-based using C5.0 algorithm.

The first test is the application of the tree-based c5.0 algorithm. The process has been mentioned above, and below are the details and the results.

1. Tree-based C5.0 algorithm training-data processing and results.

```
Evaluation on training data (3520 cases):

          Decision Tree
        ----------------
        Size      Errors

          21   57( 1.6%)   <<


         (a)   (b)    <-classified as
        ----  ----
        1683    21    (a): class Bad
          36  1780    (b): class Good
```

Figure 1. Results of tree-based C5.0 algorithm training-data.

From the results of the training data obtained the level of accuracy calculated by the confusion matrix is 98.38%.

2. Tree-based C5.0 algorithm testing-data processing and results.

Figure 2. Results of tree-based C5.0 algorithm testing-data.

The results of processing the testing-data on the training-data above after being calculated using the confusion matrix method, it can be concluded that the accuracy of the testing-data is 97.96%.

Next is testing the application of the rule-based C5.0 algorithm, the steps are the same as the steps above, only the method is different.

```
Evaluation on training data (3520 cases):

            Rules
      ----------------
        No      Errors

        12    67( 1.9%)   <<


       (a)   (b)    <-classified as
      ----  ----
      1686    18    (a): class Bad
        49  1767    (b): class Good
```

1. Rule-based C5.0 algorithm training-data processing and results.

   Figure 3. Results of rule-based C5.0 algorithm training-data.

From the results of the training-data obtained the level of accuracy calculated by the confusion matrix is 98.10%.

2. Rule-based C5.0 algorithm testing-data processing and results.

   Figure 4. Results of rule-based C5.0 algorithm testing-data.

The results of the processing of the testing-data on the training data above after being calculated by the confussion matrix can be concluded that the accuracy of the testing-data is 98.21%.

5. CONCLUSION

From the results of the study, it can be concluded that in analyzing the results of the 5S audit, the C5.0 algorithm can be used and the resulting accuracy is 97.96% for tree-based testing data, while the rules-based algorithm has an accuracy of 98.21%. When compared between the two models of the C5.0 algorithm, the one with greater accuracy is the rules-based model. Thus, it is more advisable to use the C5.0 algorithm with a rules-based model. After this research is conducted, the authors hope that in the future this research can assist in the design

and developement of applications that can be used for 5S audits that can provide direct results with high accuracy.

6. REFERENCES

Samek, W., Wiegand, T., & Müller, K. R. (2017). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. arXiv preprint arXiv:1708.08296.

Bhatia, P. (2019). *Introduction to Data Mining. In Data Mining and Data Warehousing: Principles and Practical Techniques (pp. 17-27).* Cambridge: Cambridge University Press. doi:10.1017/9781108635592.003

Frank, E., Pal, C. J., Witten, I. H., Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Netherlands: Elsevier Science.

Kastawan, P., Wiharta, D., & Sudarma, M. (2018). *Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil*. Majalah Ilmiah Teknologi Elektro, 17(3), 371-376. doi:10.24843/MITE.2018.v17i03.P11

Setyaning Nastiti, V. R., Azhar, Y., & Pramudita, A. E. (2020). *Penerapan Algoritma C5.0 Pada Analisis Faktor-Faktor Pengaruh Kelulusan Tepat Waktu Mahasiswa Teknik Informatika Universitas Muhammadiyah Malang*. Jurnal Repositor, 1(2), 131-140. https://doi.org/10.22219/repositor.v1i2.545

T. Yudi Hadiwandra. (2019). *Perbandingan Kinerja Model Klasifikasi Decission Tree , Bayesian Classifier, Instance Base, Linear Function Base, Rule Base pada 4 Dataset Berbeda*. Sains Dan Teknologi Informasi, 5(1), 70–78. https://doi.org/10.33372/stn.v5i1.452

Fajri, M., Utami, I. T. ., & Maruf, M. . (2022). *Comparison of C4.5 and C5.0 Algorithm Classification Tree Models for Analysis of Factors Affecting Auction: Perbandingan Model Pohon Klasifikasi Algoritma C4.5 dan C5.0 untuk Analisis Faktor yang Mempengaruhi Keberhasilan Lelang*. Indonesian Journal of Statistics and Its Applications, 6(1), 13–22. https://doi.org/10.29244/ijsa.v6i1p13-22