Classification and Regression Trees (CART) Algorithm for Employee Selection

Aulia Rahmawati*[1], Rizal Muhammad Affandi[2], Dea Debora Aprillia[3], Daffa Maulana[4]
Universitas Dian Nuswantoro, Penanggungan 41A, Kediri
E-mail : 611202100030@mhs.dinus.ac.id[1], 611202100029@mhs.dinus.ac.id[2]*
611202100021@mhs.dinus.ac.id[3], 611202100042@mhs.dinus.ac.id[4]
**Corresponding author*

Zudha Pratama[5], Moch. Sjamsul Hidajat[6]
Universitas Dian Nuswantoro, Penanggungan 41A, Kediri
E-mail : zudhapratama@dsn.dinus.ac.id[5] , moch.sjamsul.hidajat@dsn.dinus.ac.id[6]

Abstract – Recruitment is the main key in an effort to improve the quality of human resources in a company. Good or bad employees greatly affect the quality of the company. Therefore, it is necessary to be thorough and take a long time in screening applicants in order to get competent, professional and as expected prospective employees. The absence of professional staff to conduct employee selection is the background of this research. So the researcher uses the CART algorithm for the classification of employee recruitment, so it is hoped that it can help companies in conducting employee selection. The dataset was obtained from the selection of freelance daily workers at the Pati Regency Civil Service Police Unit in 2018, totaling 290 prospective employees. Based on calculations on 5-fold cross validation, the resulting accuracy is 98.27%, precision is 99.13% and recall is 96.88%.

Keywords – Employee, recruitment, classification, regression

1. INTRODUCTION

Labor competition in modern times is growing very rapidly. The imbalance between job growth and the number of productive age population requires job applicants to be able to improve their competencies in order to be able to compete. This may be difficult for job applicants, but not for company or office owners. The large number of applicants will make the company have many choices in recruiting new employees. With the competence of various applicants, it will make it easier for companies to choose which employee candidates are right to be recruited [1], [2]. So that in the future, it is hoped that these employees will be able to contribute well and increase the company's income financially.

Recruitment is the main key in efforts to improve the quality of human resources in a company. Good or bad employees greatly affect the quality of the company. Therefore, it is necessary to be thorough and take a long time in screening applicants in order to get competent, professional and as expected prospective employees [1]–[4]. Often the company is confused to determine which candidate is better because job applicants have their respective advantages and disadvantages. For the head office, it may not be a problem to determine whether or not a prospective employee is a problem, but there are still many branch offices in small cities that do not yet have professional staff to select prospective employees. Therefore, a data mining approach is needed to assist an office in classifying the quality of prospective employees so that the results are more accurate and efficient.

Data mining has various tasks, one of which is classification. Classification is a data mining task by grouping features into classes according to the circumstances. There are several algorithms in classification that have been proposed by many literatures such as Decision Tree, Naïve Bayes, K-Nearest Neighbor. In the Decision Tree algorithm, there are several types of algorithms including C4.5 [5], Classification and Regression Trees (CART) [2], [6], Credal C4.5, and Adaptive Credal C4.5, and ID3 [7], [8]. Previously, research showed that the CART algorithm has a performance advantage over K-Nearest Neighbor for predicting the area of rice harvested land. This happens because the data set used is large and CART is able to handle these problems better. In addition, K-Nearest Neighbor has a drawback in its performance because to determine the ranking that produces a predictive value, you must first calculate the distance between data lines.

Pahmi [4] has been diagnosed diabetes is more accurate using the CART algorithm than Naïve Bayes. Due to the Naïve Bayes algorithm there is a lack of probability that the prediction will be zero if the conditional probability is zero. Meanwhile, in this study, the classification of the length of study by students using the CART and C4.5 algorithms shows that CART has an accuracy advantage over C4.5. The binary sorting procedure in the CART algorithm can produce many variables with a mixed variable scale so that it affects the calculation results.

Based on the comparison results of several similar studies that have been described, the CART algorithm is able to process continuous and categorical data efficiently, easy to interpret and not too affected by noise data. Then the CART algorithm is used to classify employee acceptance so that the calculation process is more effective and able to produce good accuracy.

2. RESEARCH METHOD

2.1. Previous Study

Lots of research on data mining has been done. The following are some studies related to this research. In the research conducted [9], researchers have a background problem regarding the difficulty of accepting new organizational members manually. Frequent loss of data on prospective members, and it is difficult to determine whether or not the prospective member is eligible to join the background of this research. The CART algorithm was chosen by researchers in designing a system for accepting new members of the organization. This research is able to produce a system for accepting new members of the organization that is able to classify prospective members who are eligible and not eligible in the selection of new members of the organization. According to research [10] with the background of the problem of grouping prospective recipients of government social assistance so that assistance can be distributed according to the target. By using the CART algorithm, researchers were able to obtain an accuracy of 98.18% of the data ratio of social assistance recipients of 85%. According to research [11] which has a background problem of decreasing student achievement who are members of the organization so that they cannot graduate on time. By using the CART algorithm, researchers are able to design a decision system to classify the study period of students who join the organization. According to research [12] with the background of the problem of increasing open unemployment. By using the CART algorithm, researchers are able to classify the causes of open unemployment which continues to increase. So it is more effective to determine solutions to reduce unemployment. While the research [13], has a background problem about the possibility of private university students who do not continue their studies due to being accepted at state universities or other reasons. By using the CART algorithm, this research is able to classify new students who will continue their studies or not.

2.2. Data Mining

Larose argues that, data mining is a computational process in artificial intelligence to find patterns, relationships and trends from a set of similar large data. Data mining is about extracting related patterns from a set of data to create a data forecasting model [1], [11], [13], [14]. Kusnawi argues that data mining is a combination of technology that uses traditional analytical methods and sophisticated algorithms to process large amounts of data. Data mining is sourced from a collection of data that is processed and then produces information and knowledge. The purpose of data mining is to improve traditional methods so that they can be used to process large amounts of data and have different properties. Meanwhile, 5 important roles that can be utilized with the existence of data mining are estimation, forecasting, classification, clustering, and association.

2.3. Classification and Regression Tree (CART)

Conceptually, the decision tree is to group data into several decision rules. This method is able to interpret easy-to-understand and specific solutions to complex problems. The decision tree architecture consists of roots and leaf branches resembling a tree [1], [4], [6], [8], [10], [15], [16]. The types of decision tree methods include the Classification and Regression Tree (CART), C4.5, Credal C4.5, Adaptive Credal C4.5 and ID3.

The concept of CART was first proposed by Leo Breiman in 1984. CART is a decision tree algorithm with non-parametric calculations. CART consists of two types of trees, namely classification tree and regression tree. If the variable is of categorical type, it will produce a classification tree. Meanwhile, if the variable is numeric or continuous, it will produce a regression tree. The stages of CART calculation include:

1. Compile a classification tree
2. Selection tree classification
3. Determine the most optimal classification tree

Gini index is one way of selecting the optimal sorter in the decision tree. This index calculates the difference between the probability distributions of the target attribute values

$$Gini(i) = 1 - \sum (Pi)^2 \tag{1}$$

$$Gini_{split} = \sum \frac{n_i}{n} \times Gini(i) \tag{2}$$

K-Fold Cross Validation [2] is a method for evaluating the performance of a model by separating the data into two subsets, namely training data and evaluation data. In this method, each fold leaves a set of training data that will be used for the evaluation process.

2.4. Data Analysis

Confusion matrix [10] is one way to test the performance of a classification algorithm by comparing the calculations from the system with the classification results from the test data used. In the confusion matrix, there are several terms to represent the results of performance testing, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Positive is positive data that is classified correctly by the system. True Negative is negative data that is classified correctly. While false positive is data that should be negative but is classified as positive data. And vice versa, false negative is data that should be positive but is classified as negative data.

From the number of true positives, true negatives, false positives and false negatives that have been obtained, the accuracy of the system can be determined. In addition, the level of precision and recall can also be obtained. The level of precision is the percentage of positive data that is classified correctly from all data that is classified as positive either correctly or incorrectly. While recall is the percentage of positive data that is classified correctly. The following is the equation to find the level of accuracy, precision and recall as shown in equation (3) until (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \; x\; 100\% \tag{3}$$

$$Precisson = \frac{TP}{FP + TP} \; x\; 100\% \tag{4}$$

$$Recall = \frac{TP}{FN + TP} \; x\; 100\% \tag{5}$$

2.5. Dataset

At this stage the authors get a sample of data from the selection of daily freelance workers at the Pati Regency Civil Service Police Unit Office in 2018. This prospective employee data includes some information including test number, name, gender, Academic Potential Test (APT), psychological test, opportunities, interviews, work experience and direct decisions. The sample of 290 prospective employee data will then be processed using the CART algorithm as shown in Table 1 until Table 7.

Table 1. Sex Atribute

Sex	Accepted	Rejected	Total
Men	40	172	212
Woman	20	58	78

In table 1 it can be seen that 40 men were accepted and 172 were rejected from 212 job applicants.

Table 2. Academic Potential Test (APT) Atribute

Academic Potential Test (APT)	Accepted	Rejected	Total
Less	0	186	186
Good	22	44	66
Very Good	38	0	38

In table 2 the selection from the Academic Potential Test (APT) 186 applicants who were rejected had poor APT scores, 22 applicants were accepted and 44 applicants were rejected out of 66 applicants who had good APT scores, 38 applicants who were accepted had very good APT scores.

Table 3. Psychological test Atribute

Psychological test	Accepted	Rejected	Total
Less	0	99	99
Good	2	125	127
Very Good	58	6	64

In table 3 selection of psychological tests 99 applicants who were rejected had poor psychological test scores, 2 applicants who were accepted and 125 applicants who were rejected out of 127 applicants had good psychological test scores, 58 applicants who were accepted and 6 applicants who were rejected out of 64 applicants out of 64 have very good psychological test scores.

Table 4. Ability Atribute

Ability	Accepted	Rejected	Total
Less	2	190	192
Good	17	32	49
Very Good	41	8	49

In table 4 selection of ability tests 2 applicants were accepted and 190 applicants who were rejected out of 192 applicants had poor ability test scores, 17 applicants were accepted and 32 applicants were rejected from 49 applicants had good ability test scores, 41 applicants were accepted and 8 applicants who were rejected out of 49 applicants had very good ability test scores.

Table 5. Interview Atribute

Interview	Accepted	Rejected	Total
Less	0	201	201
Good	14	27	41
Very Good	46	2	48

In table 5 the interview test 201 applicants who were rejected had poor interview test scores, 14 applicants who were accepted and 27 applicants who were rejected out of 41 applicants had good interview test scores, 46 applicants who were accepted and 2 applicants who were rejected out of 48 applicants had test scores very good interview.

Table 6. Work Experience Atribute

Experience	Accepted	Rejected	Total
Yes	13	197	210
No	47	33	80

In table 6 history of work experience 13 applicants were accepted and 197 applicants were rejected from 210 applicants with work experience, 47 applicants were accepted and 33 applicants were rejected from 80 applicants without work experience.

Table 7. Last Education Atribute

Ability	Accepted	Rejected	Total
Senior High School	7	59	66
Diploma	15	111	126
Undergaduate	38	60	98

In table 7 recent educational history 7 applicants who were accepted and 59 applicants who were rejected out of 66 applicants had a history of senior high school education, 15 applicants who were accepted and 111 applicants who were rejected out of 126 applicants had a history of diploma education, 38 applicants who were accepted and 60 applicants who rejected from 98 applicants having undergaduate educational history.

Data that still has variables that are not needed will produce irrelevant information. Therefore, it is necessary to take steps first before the data is tested through the system. At this stage, some unused attributes are removed from the data to be used. Unused attributes are the test number and the name of the employee selection participant. The next step is to convert numeric data into nominal.

```
atribut=datalatih5(:,1:7);
keputusan_kasat=datalatih5(:,8);
cart=fitctree(a,b);
hasil_klasifikasi=predict(cart,datates);
```

Table 8. Last Education Atribute

Atribute	Value	Transformation
Academic Potential Test (APT)	<70	Less
	70-85	Good
	>85	Very good
Psychological test	<30	Less
	30-40	Good
	>40	Very good
Ability	<100	Less
	100-140	Good
	>140	Very good
Interview	<70	Less
	70-85	Good
	>85	Very good

3. RESULTS AND DISCUSSION

In this study the formation of a decision tree using the Gini index calculation. Calculate the Gini index of each attribute used. Based on the $Gini_{split}$ value that has been obtained, the attribute that has the smallest $Gini_{split}$ value will be the parent node. The psychological test attribute becomes the parent node because it has the smallest $Gini_{split}$ value, which is 0.0510. In the sub-data set in the 'less' psychological test leaf, only the decision is rejected. Then the 'less' psychological test leaves no more branching as shown in Figure 1. From the $Gini_{split}$ value that has been obtained, the attribute in the sub data that has the smallest $Gini_{split}$ value will be the next terminal. The psychological test attribute is 'very good' and the interview becomes terminal because it has the smallest $Gini_{split}$ value, which is 0. In the sub-data set in the interview leaf, 'not good' only has a decision to reject, 'good' interviews only have a decision to accept, while the 'very good' interview also only has a decision to accept. Then the interview leaves no more branching as shown in Figure 2.

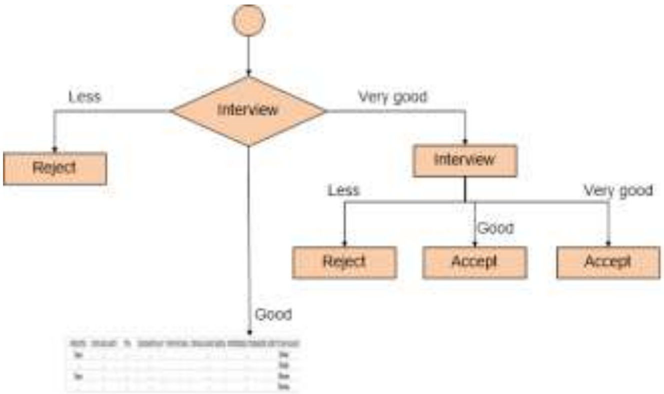

Figure 1. The last tree of the first decision

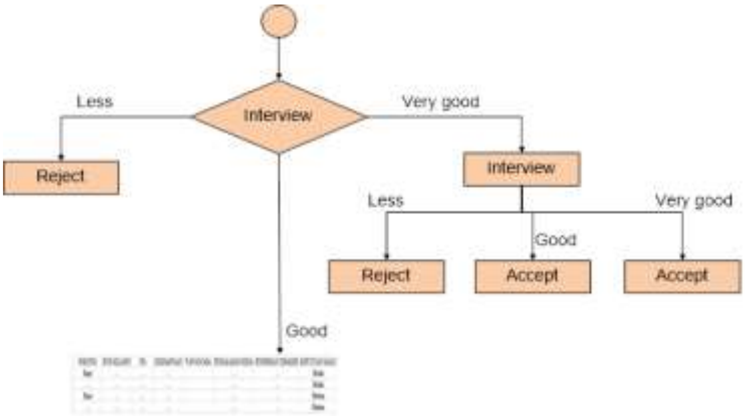

Figure 2. The last tree of the second decision

In the 'good' psychological test attribute, opportunity and interview become terminal because it has the smallest $Gini_{split}$ value, which is 0.0104. In the sub-data set in the interview leaf 'less' only has a rejected decision and the 'good' interview only has a rejected decision.

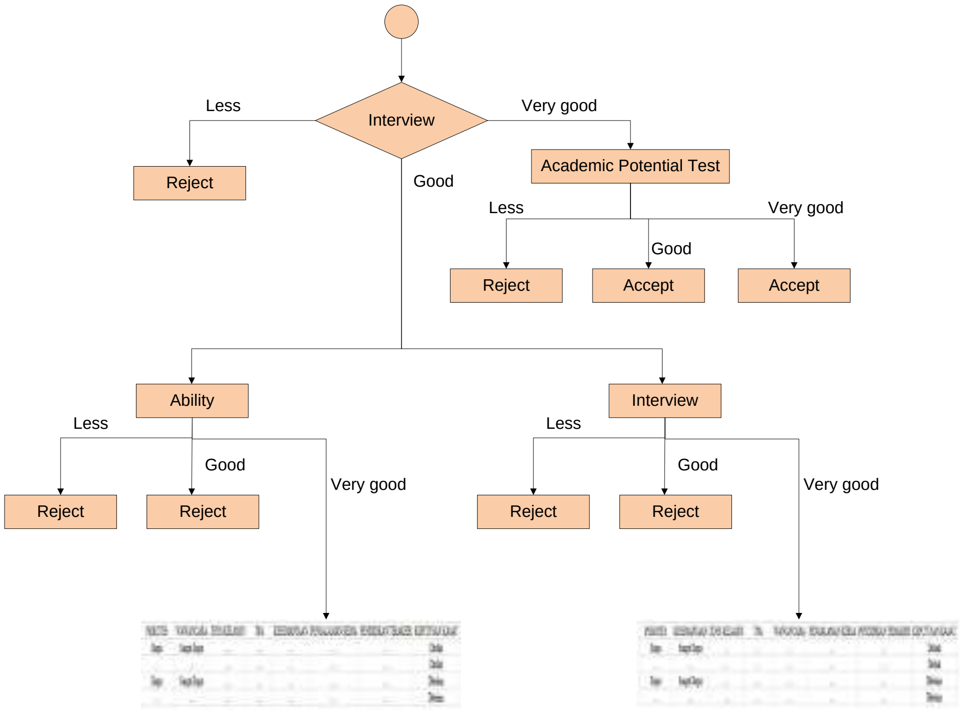

Figure 3. The last tree of the third decision

Meanwhile, 'less' opportunities only have a rejected decision and 'good' opportunities also only have a rejected decision as shown in Figure 3.

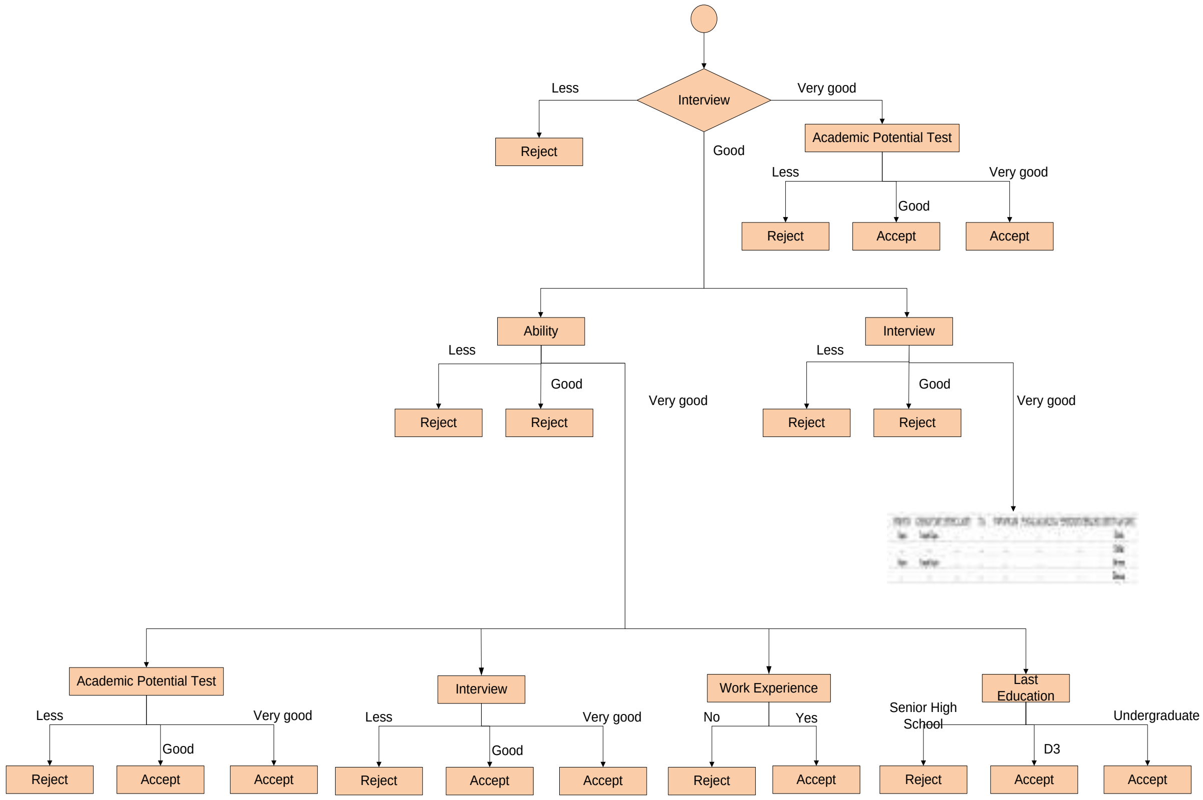

Figure 4. The last tree of the fourth decision

On the attributes of the psychological test 'good' and the opportunity 'very good', TPA, interviews, work experience and recent education have the smallest $Gini_{split}$ value, which is 0. So on the leaf there is no more branching as shown in Figure 4.

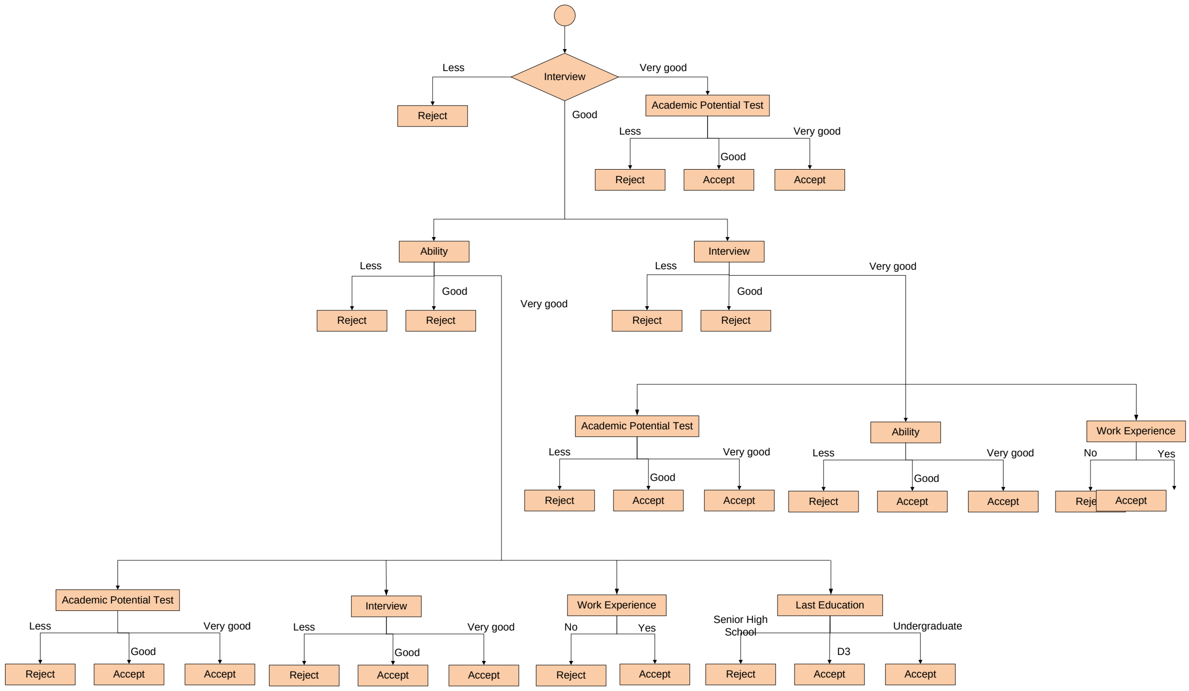

Figure 5. The last tree of the fifth decision

$$\text{Average Accuracy} = \frac{98{,}27 + 100 + 98{,}27 + 100 + 98{,}27}{5} = 98{,}27\%$$

$$\text{Average Precision} = \frac{95{,}65 + 100 + 100 + 100 + 100}{5} = 99{,}13\%$$

$$\text{Average Recall} = \frac{100 + 100 + 90 + 100 + 94{,}44}{5} = 96{,}88\%$$

On the attributes of the psychological test 'good' and the opportunity 'very good', TPA, interviews, work experience and recent education have the smallest $Gini_{split}$ value, which is 0. So on the leaf there is no more branching as shown in Figure 5.

4. CONCLUSION

Based on the results of research that has been carried out to classify the reception of freelance daily workers at the Pati District Civil Service Police Unit in 2018, it can be concluded using the CART algorithm to obtain an accuracy of 98.27%, a precision of 99.13 and a recall of 96.88%. Based on these conclusions, it is recommended for further research to be able to conduct tests using other testing techniques or other algorithms so that they can compare the results with this study.

REFERENCES

[1] A. Azar, M. V. Sebt, P. Ahmadi, dan A. Rajaeian, “A model for personnel selection with a data mining approach: A case study in a commercial bank,” *SA J. Hum. Resour. Manag.*, vol. 11, no. 1, hal. 1–10, Apr 2013.

[2] N. P. Wong, F. N. S. Damanik, C. -, E. S. Jaya, dan R. Rajaya, “Perbandingan Algoritma C4.5 dan Classification and Regression Tree (CART) Dalam Menyeleksi Calon Karyawan,” *J. SIFO Mikroskil*, vol. 20, no. 1, hal. 11–18, Apr 2019.

[3] C. Melina Taurisa dan I. Ratnawati, “ANALISIS PENGARUH BUDAYA ORGANISASI DAN KEPUASAN KERJA TERHADAP KOMITMEN ORGANISASIONAL DALAM MENINGKATKAN KINERJA KARYAWAN (Studi pada PT. Sido Muncul Kaligawe Semarang),” *J. Bisnis dan Ekon.*, vol. 19, no. 2, hal. 170187, 2012.

[4] S. Pahmi, S. Saepudin, N. Maesarah, U. I. Solehudin, dan Wulandari, “Implementation of CART (Classification and Regression Trees) Algorithm for Determining Factors Affecting Employee Performance,” in *2018 International Conference on Computing, Engineering, and Design (ICCED)*, 2018, hal. 57–62.

[5] R. K. Amin, Indwiarti, dan Y. Sibaroni, “Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region),” in *2015 3rd International Conference on Information and Communication Technology, ICoICT 2015*, 2015, vol. 0, hal. 75–80.

[6] S. Sarkar, R. Raj, S. Vinay, J. Maiti, dan D. K. Pratihar, “An optimization-based decision tree approach for predicting slip-trip-fall accidents at work,” *Saf. Sci.*, vol. 118, no. March, hal. 57–69, Okt 2019.

[7] X. Zhu, J. Wang, H. Yan, dan S. Wu, “Research and application of the improved algorithm C4.5 on decision tree,” in *Proceedings of the International Symposium on Test and Measurement*, 2009, vol. 2, hal. 184–187.

[8] B. HSSINA, A. MERBOUHA, H. EZZIKOURI, dan M. ERRITALI, “A comparative study of decision tree ID3 and C4.5,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 2, 2014.

[9] R. K. Dinata, F. Fajriana, dan K. Khairunnisa, “PENERAPAN ALGORITMA CLASSIFICATION AND REGRESSION TREES (CART) PADA PENERIMAAN ANGGOTA BARU UNIT KEGIATAN MAHASISWA (UKM) DI UNIVERSITAS MALIKUSSALEH BERBASIS WEB,” *TECHSI - J. Tek. Inform.*, vol. 10, no. 2, hal. 74, Okt 2018.

[10] H. S. Pakpahan, F. Indar, dan M. Wati, “Penerapan Algoritma Cart Decision Tree Pada Penentuan Penerima Program Bantuan Pemerintah Daerah Kabupaten Kutai Kartanegara,” *J. Rekayasa Teknol. Inf.*, vol. 2, no. 1, hal. 27, Jun 2018.

[11] N. Indah Prabawati, Widodo, dan H. Ajie, “Kinerja Algoritma Classification And Regression Tree (Cart) dalam Mengklasifikasikan Lama Masa Studi Mahasiswa yang Mengikuti Organisasi di Universitas Negeri Jakarta,” *PINTER J. Pendidik. Tek. Inform. dan Komput.*, vol. 3, no. 2, hal. 139–145, Des 2019.

[12] F. E. Pratiwi dan I. Zain, “Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara,” *J. Sains dan Seni ITS*, vol. 3, no. 1, hal. D54–D59, 2014.

[13] M. Mardiani, “Desain Model Data Mining pada Model SECI untuk Pemetaan dan Ekstraksi Pengetahuan Kompetensi Lulusan,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 3, hal. 1607–1614, Sep 2021.

[14] R. Rismayanti, “Implementasi Algoritma C4.5 Untuk Menentukan Penerima Beasiswa Di Stt Harapan Medan,” *J. Media Infotama*, vol. 12, no. 2, hal. 116–120, 2017.

[15] E. Y. S. Ritno, N. A. Hasibuan, dan Fadlina, “IMPLEMENTASIALGORITMA CLASIFICATION ANDREGRESSION TREES (CART) DALAM KLASIFIKASI EKONOMI KELUARGA PADA

DESADAGANG KELAMBIR TG . MORAWA,” *Maj. Ilm. INTI*, vol. 6, no. 1, hal. 66–72, 2018.

[16] A. B. Siregar, E. Buulolo, dan P. Ginting, “Pemanfaatan Algoritma Classification and Regression Tress (Cart) Untuk Memprediksi Omset Spanduk Pada Cv . Moeha,” *Konf. Nas. Teknol. Inf. dan Komput.*, vol. I, hal. 347–354, 2017.