The Comparison of C4.5 and CART (Classification and Regression Tree) Algorithm in Classification of Occupation for Fresh Graduate

Febian Joshua Reynara[1], Sepriana Carolina[2], Iustisia Natalia Simbolon[3]

{febianjosuha@gmail.com[1], sepriana15@gmail.com[2], iustisia.simbolon@del.ac.id[3]}

Faculty of Informatics and Electrical Engineering, Del Institute of Technology, Laguboti, Indonesia - 22381

Abstract. The problem that college students face is the difficulty of determining the appropriate field of work after they graduate from college. In this study, a classification of the field of work was carried out using the data mining method based on the alumni field of work data. The data on the field of work of alumni contained information such as gender, study program, practical work topics, types of practical work companies, final project topics, and year of graduation. The classification on the field of work carried out was divided into three types of experiments, namely experiments in eight target categories (STQA Engineer, Software and Mobile Application Developer, Web Developer, UI/UX Designer, Software and Business Analyst, Lecturer and Researcher, AI Engineer, DevOps and Cybersecurity Practitioner), three target categories (SQA, Programmer, Data Manager, and Analyst) and two target categories (Programmer and Non-Programmer). The data mining algorithms used to classify were C4.5 and CART (Classification and Regression Tree). The accuracy obtained using the C4.5 algorithm was 42% in the eight categories experiment, 58% in the three categories experiment, and 75% in the two categories experiment. In comparison, the accuracy obtained using the CART algorithm was 43% in the eight categories experiment, 61% in the three categories experiment, and 77% in the two categories experiment. Based on the experimental results, it can be concluded that the best algorithm to classify the fields of work based on alumni data from the two algorithms used is the CART algorithm, even though the difference is not too significant.

Keywords: Field of Work, Alumni, Classification, C4.5 Algorithm, CART Algorithm

1 Introduction

Each study program at a university has a graduate profile and competence expected to be possessed by every student when they have completed their education in the study program. One of the problems often found was the number of students who still did not know and could not determine the field of work they would be involved in after graduating from lectures, even until they entered the final semester of the lecture period. Meanwhile, many positive impacts can be obtained once students can determine and choose the appropriate field of work as early as possible, one of which is to prepare and focus on developing the soft skills and hard skills needed in the field of work. There is a lot of knowledge and information that can be extracted

IConVET 2021, November 27, Singaraja, Indonesia

DOI 10.4108/eai.27-11-2021.2315527

from the alumni or graduate data available at a university, such as the classification of alumni's occupations, the estimated waiting period for alumni to get a job, the estimated absorption of graduates in the work field, predictions of the amount of graduates' income after getting a job, and so on[1].

In this study, the author aimed to classify the field of work according to the profile of graduates from several study programs studied as the classification target class. The classification results were expected to assist students in choosing the appropriate field of work based on the data of alumni who have worked. The classification was carried out using data mining techniques, namely the C4.5 algorithm classification method and the CART (Classification and Regression Tree) algorithm. These two algorithms were included in the ten best data mining algorithms, where the ten algorithms are C4.5, K-Means, KNearest Neighbor, SVM (Support Vector Machines), Apriori, Expectation-Maximization Algorithm, Page Rank, Naive Bayes Classifier, Classification and Regression Trees (CART), and Adaboost [2].

The researchers also compared the accuracy of the C4.5 algorithm and the CART algorithm in classifying work fields because these two algorithms use decision trees to classify the data. Although both produce decision trees, the process of finding the roots and branches of the two algorithms is different, but both produce a quite high accuracy [3]. The purpose of this comparison was to find out which decision tree classification algorithm was better in classifying fields of work according to the available data.

2 Literature study

This section describes the theoretical basis that underlines this research.

2.1 Previous studies

There have been a number of previous studies that used the same algorithm as the one in this study. In Asroni's research (2018), a classification of types of work has been carried out using the C4.5 algorithm, the target of which was divided into two types, namely private and government. It was concluded that the most influential attribute in the study was GPA, but the accuracy of the model obtained was not included [4]. In Monalisa's research, students' majors have been determined using the CART algorithm where the target consists of three classes, namely Science, Social Sciences, and Religion. The accuracy obtained was 88.61% [5]. These two studies can assist researchers in understanding the steps of working on the algorithm used to classify and as a benchmark to develop previous research.

2.2 Classification

Classification is a process or technique to find a function or model that is able to describe and distinguish concepts or data classes. The classification model is obtained from the results of training data analysis (for example, the data objects where the class label is known).

2.3 C4.5 algorithm

The C4.5 algorithm is used to form and produce a decision tree. In the decision tree formation method, a very large number of facts is converted into a decision tree that can represent several rules. The following are the stages of the C4.5 algorithm in forming a decision tree [6]:

- Specify the attribute used as the root node.

At this stage, the gain ratio value was calculated for each attribute. The root attribute was selected based on the attribute that has the highest gain ratio value. The gain ratio value for each attribute can be calculated using the following equation:

$$GainRatio\ (S, A) = \frac{Gain(S,A)}{SplitInfo(S,A)}\ . \tag{1}$$

Gain Ratio is a comparison of Gain with Split Info. Gain is the amount of information obtained from a variable and can be found using the following equation:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(s_i)\ . \tag{2}$$

Description:
S : Case collection
A : Attributes/features
n : Number of partitions on attribute A
| Si |: Proportion of Si to S / Number of cases in the i-th partition
| S | : Number of cases in the set S

The entropy value is used as a reference to find out how informative an attribute is. The entropy value in equation (2) can be calculated using the following formula:

$$Entropy\ (S) = \sum_{i=1}^{n} - p_i * log_2\, p_i\ . \tag{3}$$

Description:
S : Case collection
n : Number of partitions on S
p_i : The ratio of Si to S

While Split Info on the Gain Ratio formula can be obtained using the following equation:

$$SplitInfo\ (S, A) = - \sum_{i=1}^{n} \frac{s_i}{s} * log_2 \frac{s_i}{s}\ . \tag{4}$$

Description:
S : Case collection
A : Attributes/features
Si : Number of samples for attribute A

- From the root node attribute obtained in the first step, a branch is created for each value in the attribute of the split cases in one branch.
- Repeat the process for each existing branch until all cases on the branch are perfectly partitioned or have the same class. The partitioning process in the decision tree formation is terminated if.

  a). Each attribute on a record cannot be partitioned anymore.

  b). Every record in node N has got the same class.

  c). There are no records on the empty branch.

2.4 CART algorithm

The CART algorithm is one of the widely used classification methods. This method integrates various factors from different sources in classification and regression problems based on the binary recursive method [7]. Gini index is used in the application of the CART method. The Gini index can generate binary partitions on all attributes that have discrete values or continuous values, with the following equation:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \ . \tag{5}$$

Description:
Gini(*D*): the value of the impurity of partition *D*
m : index number
p_i : probability of tuple *D* at index-*i*

pi is said to be the probability that a tuple in *D* is in class *Ci*. The probability can be formulated using the calculation results |*Ci*,*D*|/|*D*|, where |*Ci*,*D*| is the number of tuples in *D* that have class *Ci* and |*D*| is the number of tuples in *D*. Checking for binary divisions is done by adding up the impurity values of each partition generated by the division. Assuming that the division performed on attribute *A* partitions *D* into *D1* and *D2*. Then the Gini index value of *D* can be obtained using the following equation [8]:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \ . \tag{6}$$

Description:
$Gini_A$ (*D*) : partition *D*'s impurity value in attribute *A*
Gini (*D1*) : the value of the first partition's impurity in the tuple'*D*
Gini (*D2*) : the value of the second partition's impurity in the tuple'*D*
D : tuple *D*
D1 : first partition of tuple *D*
D2 : second partition of tuple *D*

In discrete-valued attributes, the set in the subset that produces the smallest Gini index value or close to 0 for attribute A was then chosen as a splitting subset because the attribute at the top rank or splitting attribute is discrete and the decision tree formed is binary. All possible binary splits on an attribute should be checked. However, for continuous-valued attributes, all split points due to attributes that are at the top rank or continuous-valued splitting attributes must be checked. For the value of an attribute that has been sorted, the point or middle value between each pair of opposite values were taken as a split point. The point produces the smallest Gini index value for an attribute that will eventually be used as a split point. The decrease in the level of impurity obtained from a binary division of attribute A can be obtained using the following equation:

$$\Delta Gini(A) = Gini(D) - Gini_A(D) . \tag{7}$$

Information:
Δ*Gini* (*A*) : level of impurity
Gini$_A$ (*D*) : impurity of a partition *D* on attribute *A*
Gini (*D*) : first partition impurity on tuple *D*

2.5 SMOTE

SMOTE (Synthetic Minority Over-Sampling Technique) is a technique used on imbalance dataset problems for classification by balancing the distribution of the number of sample data in each minority class using a sample data selection technique so that the number of minority class sample data is balanced with the majority number of class data. The steps carried out in the SMOTE method were started by calculating each distance between data in the minority class. After that, the SMOTE percentage value was determined and then the nearest k number was determined until the last step, to create synthetic data with the following equation [9]:

$$Xsyn = X_i(X_{knn} - X_i) \times \delta . \tag{8}$$

Description:
Xsyn : Synthetic data to be created
X_i : Data to be replicated
X_{knn} : The data with the closest distance to the data to be replicated
δ : Random value between 0 to 1

2.6 Grid search optimization

Grid Search Optimization is a technique used to find the most optimal combination of hyperparameters to build a machine learning model. Searching for the best parameters can take a lot of time if done manually, especially if the algorithm used has many parameters [11]. The Grid Search method performs a complete search on every possible combination of parameters that have been determined.

2.7 Confusion matrix

The Confusion Matrix is used to evaluate the performance of a classification model. The intended performance evaluation is to determine how good a classification model is in classifying data [12]. The Confusion Matrix provides validation regarding the actual class and the class used for prediction. Validation refers to the process of making predictions using an existing model and then comparing the results with known actual results [13]. An example of a table for evaluating a classification model is as follows:

Table 1 Confusion matrix.

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Description:
TP: True Positive, where the prediction result is positive and the actual value is positive
FP: False Positive, where the prediction result is positive while the actual value is negative
FN: False Negative, where the prediction result is negative while the actual value is positive

TN: True Negative, where the prediction result is negative and the actual value is negative
Accuracy represents the accuracy of the model successfully performing the classification. We can calculate accuracy using the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} . \tag{9}$$

3 Research methodology

The stages carried out in this study can be seen in the following **Figure 1** Research flowchart below:

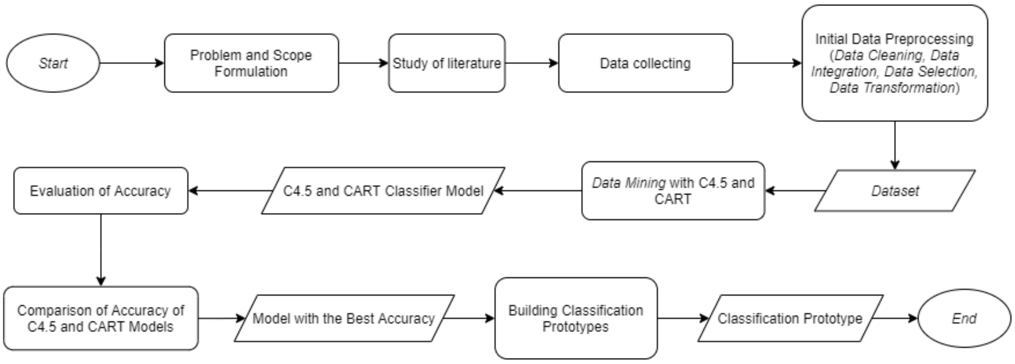
Start
Problem and Scope Formulation
Study of literature
Data collecting
Initial Data Preprocessing (Data Cleaning, Data Integration, Data Selection, Data Transformation)
Dataset
Data Mining with C4.5 and CART
C4.5 and CART Classifier Model
Evaluation of Accuracy
Comparison of Accuracy of C4.5 and CART Models
Model with the Best Accuracy
Building Classification Prototypes
Classification Prototype
End

Fig. 1. Research flowchart.

The initial stage in this research was to formulate the problem based on the research background that has been described. After that, is determining the scope of the research, the limitations, and the focus of the research. The next stage was conducting a literature study on data mining in previous research and the same case study with the problems discussed. In addition, an analysis of the data attributes needed for the classification process was carried out. The next stage is data collection, where at this stage, the data needed for this research was collected. All data collected was used as a candidate feature and then the initial data processing was carried out. This stage produced an alumni dataset used in research. After the data was ready to be processed, the data mining process was carried out by utilizing the C4.5 and CART algorithms to build a classification model. In each model formed from the two algorithms, an evaluation and analysis of the classification results was carried out to find out which model has the highest level of accuracy. The last stage was to build a prototype of job classification using models that have the best level of accuracy from each algorithm.

3.1 Data analysis and pre-processing

In this study, researchers used a dataset of alumni occupations consisting of 500 lines of data. The alumni dataset used is data from graduates of the Bachelor degree of Informatics study program, Bachelor degree of Information Systems, and Associate degree Informatics Engineering, collected from alumni graduation books from 2014 to 2020. The data on alumni occupations was obtained from the LinkedIn site. The dataset has gone through a pre-processing process. Data processing and analysis were carried out to prepare data to be processed in the job classification process. The initial processing steps carried out were:

1. Data cleaning to detect noise and inconsistent data.
2. Data integration to combine data obtained from various sources.
3. Data selection to select attributes to be used in this research.
4. Data transformation to convert data into new data or categorize existing data into more general categories to suit the algorithm method used.

After going through the initial data processing process, the data to be used for research consists of seven predictor attributes, namely Gender, Study Program, GPA, Practical Workplace, Job Training Topic, Final Project Topic, and Year of Graduation, as well as target attributes in the form of Alumni Jobs. The attributes of the data users have different characteristics to use in this study. The gender attribute is a supporting attribute with binary data type in taking classification decisions. In the research of Yang and Barth in 2015, it was stated that men are more likely to like technical fields of work than women. For example, in the IT field, men tend to prefer programming compared to women who prefer analysis and interface design. The study program attribute is one of the attributes to determine the decision-making in classification and the graduate competence based on the study program. The GPA attribute (Ordinal data type) is one of the determining attributes in determining the competence and quality of graduates. Therefore, the GPA will impact the selection of one's field of work. The Job Training topic (Nominal data type) is categorized into Service-Based and Product-Based companies. The Final Project Topic attribute (Nominal data type) refers to the measurement of ability and understanding of the skills and interests of alumni to determine the field of work when completing lectures. The attribute year of graduation is a supporting attribute in classifying the field of work. The researchers assumed that the year of graduation affects one's field of work. In the field of work, attributes based on alumni data will be used as the target class for classification.

The dataset that has been processed was then used to conduct an experiment on job classification using the C4.5 algorithm and the CART algorithm. In each algorithm, three experiments were carried out with different target classes in each type of experiment. The three types of experiments were experiments with eight target class categories (STQA Engineer, Software and Mobile Application Developer, Web Developer, UI/UX Designer, Software and Business Analyst, Lecturer and Researcher, AI Engineer, DevOps and Cybersecurity Practitioner), three class categories target (Programmer, SQA, Data Manager, and Analyst), and two categories of the target class (Programmer and Non-programmer).

3.2 Classification experiment stages

In each classification experiment process using the C4.5 and CART algorithms, the following steps were carried out:

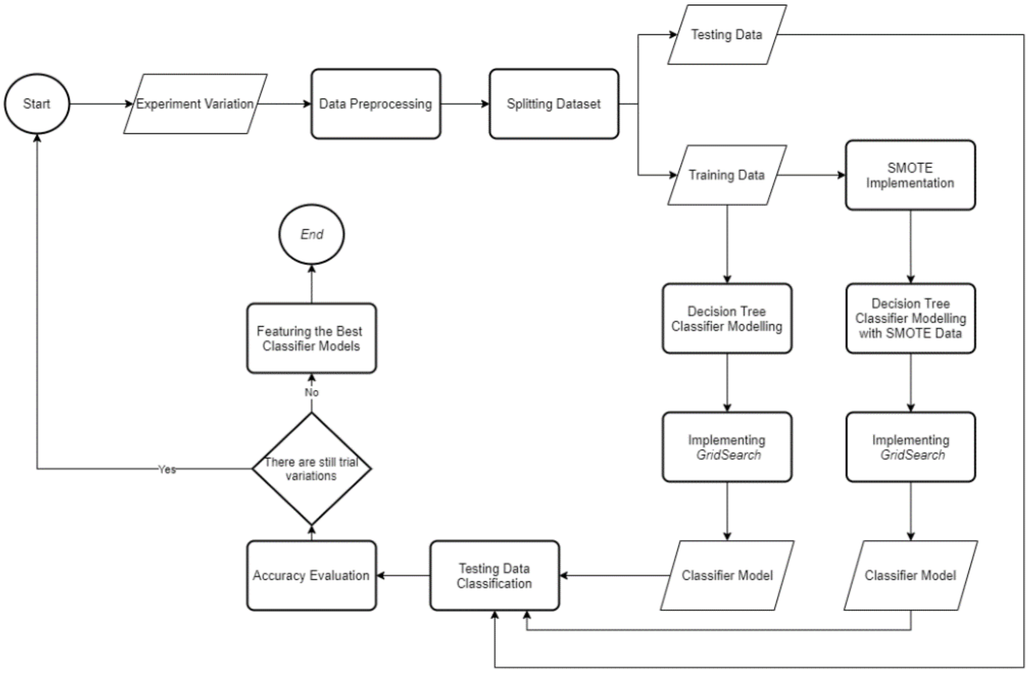

Fig. 2. Algorithm classification steps flowchart.

The explanation of the steps in the flowchart above is as follows:

1. Preparing experimental variations, with three experimental variations, namely experiments on data with 8, 3, and 2 target class categories.
2. Pre-processing data. At this stage, the encoding process was carried out using the One Hot Encoding method.
3. Split the dataset. The job field dataset based on alumni data was divided into training data and testing data using the Hold Out technique, where training data was used to generate models and testing data to test the resulting model. The comparison of training data and testing data used was 80:20.
4. SMOTE implementation was carried out on the training data to balance the amount of data in each class. Then the modeling was carried out twice, namely using the initial training data and the data from the SMOTE implementation.
5. Performs hyper-parameter tuning which used the Grid Search method to find the optimal parameters in the two models.
6. Performs the classification stage of the test data using the decision tree model that has been formed. An evaluation was carried out to find the accuracy of the resulting model.
7. If there were still variations in the experiment to be carried out, the process was returned to the initial stage to carry out the experiment. If all variations of the experiment have been carried out, the experiment has been completed.

4 Result and discussion

In this section, the evaluation results of the accuracy of the C4.5 and CART algorithms were described for each experimental category.

4.1 C4.5 algorithm experiment results

The results for each experiment using the C4.5 algorithm are shown in the following table.

Table 2. C4.5 algorithm experiment results.

Experiment Name	Initial Modelling Accuracy	Accuracy with Grid Search Result Parameter	Accuracy with SMOTE	SMOTE Accuracy with Grid Search Result Parameter
8 Categories	37 %	42 %	30 %	34 %
3 Categories	56 %	56 %	58 %	58 %
2 Categories	75 %	75 %	75 %	75 %

Based on the results obtained from several variations in the experiment, the reduction in the number of classes on the target attribute greatly affected the model's accuracy. The fewer classes on the target attribute, the better the accuracy would be. The most significant increase in accuracy occurred in the 2-category experiment, where the classification target only consisted of two classes. This increase in accuracy is due to the relatively small amount of data so that it is not sufficient to represent the data relationship to the target class, which consists of eight classes.

In each dataset, an experiment was also carried out by implementing SMOTE on the data train to overcome the problem of an unbalanced dataset. The results obtained show that the SMOTE method in the eight category experiment produces lower accuracy than the model built without the SMOTE method. This decrease in accuracy is caused by comparing the number of minority classes that are too far from the number of majority classes. For example, AI Engineer (9 data) and 10 UI/UX Designer (10 data) were very few compared to STQA Engineer data which totals 189 data lines, this caused the SMOTE method data to only have a small sample of data to create new synthetic data. In contrast to the 8-category experimental case, the SMOTE method produced higher accuracy in the 3-category experimental case where in this experimental variation the data distribution was indeed unbalanced but the difference was not too far away, namely the Programmer class with 199 data, SQA 189 data, and Data Manager and Analyst, totaling 112 rows of data.

4.2 CART algorithm experiment results

The results for each experiment conducted using the CART algorithm are shown in the following table.

Table 3. Experiment results CART algorithm.

Experiment Name	Initial Modelling Accuracy	Accuracy with Grid Search Result Parameter	Accuracy with SMOTE	SMOTE Accuracy with Grid Search Result Parameter
8 Categories	31 %	43 %	31 %	36 %
3 Categories	52 %	61 %	51 %	53 %
2 Categories	71 %	75 %	69 %	77 %

The results obtained in the CART algorithm experiment show a pattern almost the same as the experimental results of the C4.5 algorithm, where the reduction in the number of classes on the target attribute greatly affects the accuracy of the model built. The fewer classes on the target attribute, the better the accuracy will be.

4.3 Comparison of Experimental Results

Based on the results of the experiments that have been carried out, the comparison between the two algorithms used is based on the accuracy value obtained can be illustrated with the comparison graph as follows.

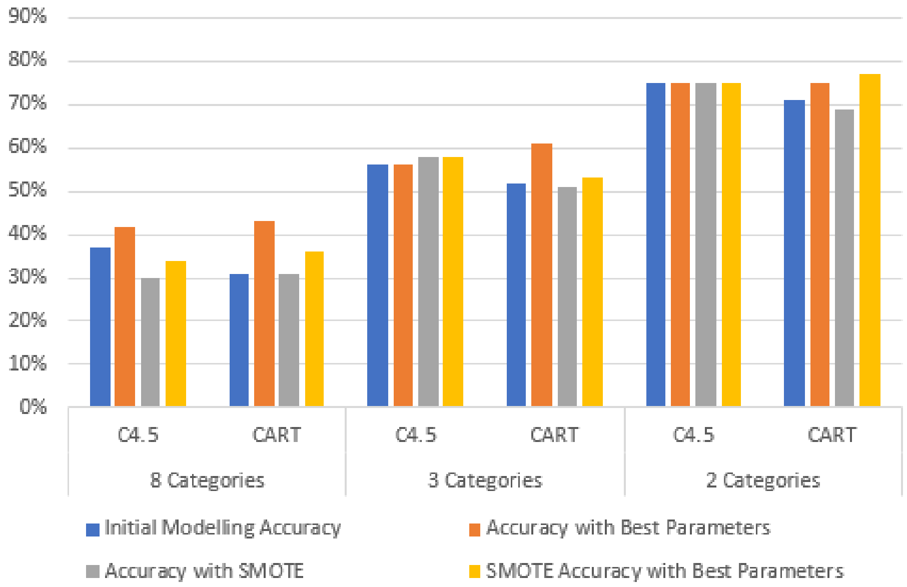

Fig. 3. Comparison of C4.5 and CART accuracy

Based on the experimental results in **Figure 4**, it is concluded that the number of classification target classes greatly affects the accuracy of the resulting model both when using the C4.5 and CART algorithms. Reducing the number of target classes can increase the model's accuracy generated in each experiment using both the C45 and CART algorithms.

To determine the best method for these two algorithms, a comparison was made using the highest accuracy of each algorithm in the three categories of experiments. The following is a summary table of the highest accuracy obtained by each algorithm in each experiment.

Table 4. Highest accuracy on every trial.

Experiment Name	C4.5 Highest Accuracy	CART Highest Accuracy
8 Categories	42%	43%
3 Categories	58%	61%
2 Categories	75%	77%

Based on the summary table of the highest accuracy, it can be concluded that the CART algorithm is the best algorithm for classifying alumni's occupations in this study. This is because all the highest accuracies of the CART algorithm produce better accuracy in each type of experiment compared to the C4.5 algorithm accuracies. The highest accuracy obtained is 77%, namely in the experiment of two categories of work. This accuracy is not quite satisfactory for the case of classification. This is due to the very limited number of rows and attributes of the dataset used. CART accuracy also has better accuracy when hyper-parameter tuning is performed using GridSearch. Based on this, the author also concludes that the hyper-parameters in the CART algorithm in this study are better and more optimal when tuning hyper-parameters are done.

5 Conclusions

Based on the experimental results obtained, the conclusions that researchers get are as follows: 1) this research has implemented the C4.5 algorithm to classify occupations based on alumni data and produce analysis and job classification models. The classification was carried out on three types of experiments, wherein the experiment eight categories of work, the highest accuracy was 42%. In the experiment of three categories of work, the highest accuracy was 58%. While in the experiment of two categories of work, the highest accuracy was 75%; 2) this research has implemented the CART (Classification and Regression Tree) algorithm to classify occupations based on alumni data and produce analysis and job classification models. The classification was carried out on three types of experiments, wherein the eight work category experiments, the highest accuracy was 43%. In the experiment of the three categories of work, the highest accuracy was 61%. While in the experiment of the two categories of work, the highest accuracy was 77%; 3) the best algorithm for classifying fields of work based on alumni data from the two algorithms used is the CART algorithm because all the highest accuracy of the model produced in the three experiments is better than the highest accuracy of the model generated by the C4.5 algorithm, although the difference in accuracy is not too significant.

Based on the results of research conducted, suggestions that can be made for further research are as follows: 1) further research can be developed using other classification algorithms included in the best data mining algorithm according to Wu et al. (2007) and Enriko (2019), such as Naïve Bayes, K-Nearest Neighbor or Support Vector Machine (SVM) to find out the difference in accuracy with the two algorithms used in this study; 2) According to Jason Brownlee in his book entitled Imbalanced Classification with Python (2020), there are still several methods that can be used to overcome imbalanced classification problems in addition to the two methods used in this study, such as Cost-Sensitive Learning, Gradient Boosting with XGBoost, Ensemble Algorithm, etc. Further research can apply these methods to overcome the imbalance of classification problems encountered.

References

[1] Nurwan, Resmawan. 2015. *Tracer Study: Kajian Profil Lulusan dan Relevansi Kurikulum Program Studi Pendidikan Matematika Tahun 2010-2014*. Jurusan Matematika FMIPA Universitas Negeri Gorontalo.

[2] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. Z. Michael, S. David, and J. H. Dan. 2007. Top 10 algorithms in data mining.

[3] Assiroj, P. 2016. *Kajian Perbandingan Teknik Klasifikasi Algoritma C4.5, Naïve Bayes dan CART untuk Prediksi Kelulusan Mahasiswa (Studi Kasus: STMIK Rosma Karawang)*. Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI.

[4] Asroni, Badrahini Masajeng Respati, Slamet Riyadi. (2018). *Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta*. Vol. 21 No.2, 158-165

[5] Siti Monalisa, Fakhri Hadi. 2020. *Penerapan Algoritma CART Dalam Menentukan Jurusan Siswa di MAN 1 Inhil*. Jurnal Sistem Informasi dan Komputer, Vol 09, no. 3.

[6] Daniel T. Larose. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, Inc.

[7] Siti Monalisa, Fakhri Hadi. 2020. *Penerapan Algoritma CART Dalam Menentukan Jurusan Siswa di MAN 1 Inhil*. Jurnal Sistem Informasi dan Komputer, Vol 09, no. 3.

[8] Joseph Wijaya. 2019. *Implementasi Algoritma Pohon Keputusan CART Untuk Menentukan Klasifikasi Data Evaluasi Mobil*. Yogyakarta.

[9] Brownlee, Jason. 2020. *Imbalanced Classification with Python*.

[10] N. V, Chawla, K. W. Bowyer & L. O. Hall. 2017. *Handling Imbalance Data Prediksi Churn menggunakan metode SMOTE dan KNN Based on Kernel*. e-proceeding of Engineering, vol. 4, no. 117, pp. 1-15.

[11] Iwan, S., Adam, P., & Gary, W. 2016. *SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance*. Telkomnika Vol.14, No. 4.

[12] Anam Choirul, Harry Budi Santoso. 2018. *Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa*. Vol.8 No.1.

[13] Gorunescu, F. 2011. *Data Mining Concepts, Models and Technique*. Berlin: Springer.