International Journal Information System and Computer Science (IJISCS)

Full Paper
eISSN : 2598-246X
pISSN : 2598-0793

IMPLEMENTATION OF C4.5 ALGORITHM TO ASSIST IN THE SELECTION OF FLOOR CONSTRUCTION PROJECTS

Aditya Roval Lendra[1], Diky Firdaus[2]
[1,2]Department of Computer Science, Mercu Buana University, Jakarta, Indonesia
[1,2]Street Meruya Selatan No.1, Meruya Selatan, Kembangan, Jakarta, Indonesia

* Corresponding author
41516110182@mercubuana.ac.id
diky.firdaus@mercubuana.ac.id

Article history:
Received: 14 September 2020
Revised: 17 October 2020
Accepted: 03 November 2020

Keywords:
Datamining;
Classification;
C4.5 Algorithm;
Construction;
Floor project;

Abstract
The country of Indonesia is a developing country. This is indicated by the improvement of the development process and business. So that, many development projects will appear. With the changing world in the industry with e-Government governance, project data is no longer in paper form. With the emergence of data - a lot of companies need to do project management activities in determining the project strategy to be taken so as not to affect the final results in determining the taking of the project. Then do a lot of research on the project data that appears in the construction world. The method used for this research is the C4.5 Algorithm which is one of the modern algorithms for data mining processes. C4.5 algorithm is also called a decision tree (decision tree) which is one of the classification methods with a tree structure representation. The concept is to collect data and be made into a decision tree based on the rules needed to get an outcome. The pattern and results obtained will be used for recommendations in determining which projects the company will take. The values generated using Rapidminer are Accuracy 97.18%, precision 100%, and recall 94.31%. With the result 1205 recommended floor construction projects were taken and 784 projects that were recommended not to be taken.

1.0 INTRODUCTION

The world has changed in the digital era, major changes can be seen in all fields, especially the delivery of information that is more systematic and informative. Likewise in the field of construction, government with e-Government and so on. Changes to the digital era have changed the form of data and storage media, so far, stored data is no longer in paper form. As one of the functions and activity processes in project management that greatly affects the final project results, control has an important role in minimizing any deviation that can occur during the project process. Inaccuracy in analyzing the possibilities that will occur often results in problems such as project delays that are not in accordance with the original plans and objectives. So a study / feasibility study of a project needs to be carried out. What is meant by a feasibility study is a research on whether or not a project (usually an investment project) can be implemented successfully. The first step that needs to be determined in a project feasibility study is the extent to which the aspects affecting the project will be studied, then for each of

these aspects it needs to be analyzed so that it has a picture of the feasibility of each aspect. Critical Success Factors or CSF's are factors or responses that are critical to the successful implementation of a project that must be done where without these factors the project will not be successful or successful in achieving certain targets or goals on a project or job. Critical Success Factors are very important to identify before the project starts. [1] [1] With so many construction projects developing throughout Java, and the amount of data that is being obtained, data mining techniques are needed to process and analyze these projects. In general, the uses of data mining are estimation, prediction, classification, clustering and association [2]. Data mining is one of the fastest growing fields due to the huge demands for the added value of large-scale databases that are in line with the growth of information technology and can extract large data sets into new knowledge. And also processing one or more machine learning techniques to analyze and extract automated work [3]. The classification model used in this study is the C4.5 algorithm. The C4.5 algorithm can explicitly describe the structure, model, and model structure of the C4.5 algorithm in the form of a root tree, so it is widely used in comparison with other algorithms (Thammasiri, Delen, Meesad, & Kasap, 2014) [2]. Classification is a data mining technique that can be used to predict group membership to data instances [4]. C4.5 Decision Tree is the first supervised fundamental machine learning classification algorithm to be widely applied and usually achieves very good performance in predictions [3]. It is hoped that processing with algorithms can make it easier to select existing projects based on the smooth running of existing projects.

2.0 THEORETICAL

2.1 Data Mining

The process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technology as well as statistical and mathematical techniques. [2] Data mining, often referred to as knowledge discovery in database (KDD), is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets [3]. Knowledge discovery in database (KDD) is essentially the process of finding useful knowledge from a data set. A. Berstein et al. [5]

2.2 Classification Data

Classification is the most commonly applied data mining technique, which uses a pre-classification set of examples to develop a model that can classify a typical population of records. [2] [4]. In classification, there are target categorical variables, such as income bracket, which, for example, can be partitioned into three classes or categories: high income, middle income, and low income. The data mining model examines a large set of records, each record containing information about a target variable as well as a series of input or predictor variables. [2]

2.3 Decision tree

Decision tree, a method of converting very large facts into decision trees that represent rules. Rules can easily be understood in natural language. The main benefit of using decision trees is their ability to break complex decision-making processes into simpler ones so that decision making will be better interpreted as problem solutions by converting the data form (tables) into a decision tree model, turning the decision tree model into a rule. Decision trees are also useful for exploring data, namely finding hidden relationships between a number of prospective input variables and the target variables. [6]

2.4 Algorithm C4.5

Algorithm C4.5 The algorithm was introduced by Quinlan to drive a classification model and is also called a decision tree whose data is based on training data provided by obtaining rations [3]. C4.5 is an efficient and effective approach to real-time classification and classification evaluation. The advantage of the C4.5 is that the model can be easily interpreted and implemented with continuous values and discrete values. It also works efficiently working

with electrocardiogram signals. The C4.5 algorithm is an improved ID3 algorithm. [7]. Some of the differences include:

- Able to handle attributes with discrete or continuous types.
- Able to handle missing value attributes
- Can prune branches.

A discrete attribute is an attribute that has a finite set of values or countably infinite values, which may or may not be represented as an integer. Attributes such as hair color, smoker, health test, and size drink size as in the above examples each have finite sum values, so the attributes are discrete. If an attribute is not discrete, it means that the attribute is continuous. The terms numeric attribute and continuous attribute are often used interchangeably in the literature. (This can be confusing because, in the classical sense, continuous values are real numbers, whereas numeric values can be either integers or real numbers.) In practice, real values are expressed in terms of numbers. Continuous attributes are usually represented as floating-point (decimal) variables.

2.5 Weight Information Gain dan Entropy

Weight information gain (WIG) is the most common method of weighting each variable from the evaluation attribute. To calculate information gain, one must first understand another rule called entropy. In the field of Information Theory, we often use entropy as a parameter to measure the heterogeneity (diversity) of a data sample set. The more heterogeneous the data sample set is, the greater the entropy value. After obtaining the entropy value for a sample data set, we can measure the effectiveness of an attribute. This measure of effectiveness is known as information gain. [2]

2.6 Rapid miner

Software for data mining processing. The work done by Rapid Miner text mining revolves around text analysis, extracting patterns from large data sets and combining them with statistical methods, artificial intelligence, and databases. The purpose of this text analysis is to obtain the highest quality information from the processed text. Rapid Miner provides data mining and machine learning procedures, which include: ETL (extraction, transformation, loading), data preprocessing, visualization, modeling and evaluation. The data mining process is composed of nestable operators, described in XML, and created in a GUI. The presentation is written in the Java programming language. [8].

3.0 METHODOLOGY

The research method used in the application of the C4.5 algorithm for recommendations for the acceptance of new sales partner candidates, uses a research design aimed at Figure 1 below [9]:

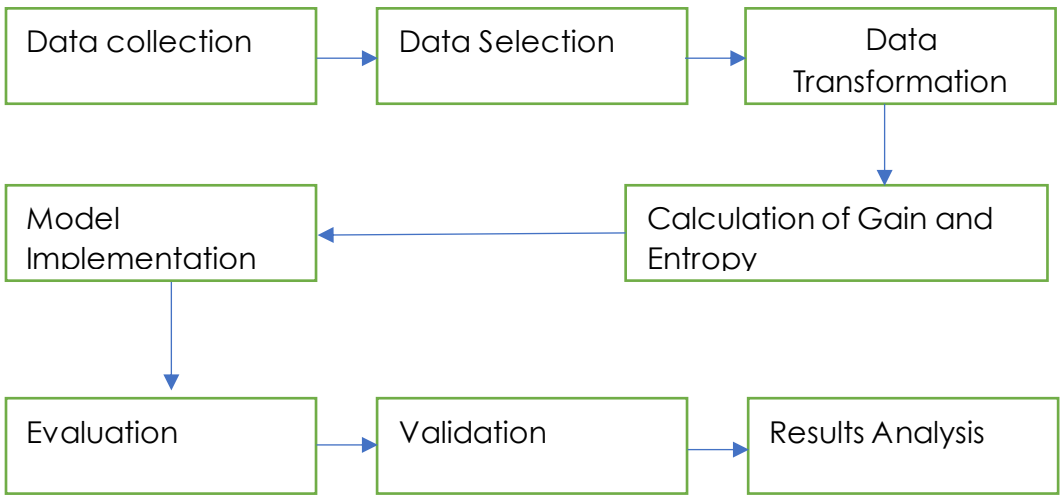

Figure 1. Stages

3.1 Data Collection

Data collection is done by requesting data from PT. PLP. This data includes several attributes that will be processed using the C4.5 Algorithm. The amount of data that will be used is 1989 rows.

No	Nama proyek	Region	Luas lantai proyek	Status proyek
1	WAREHOUSE	Jawa Tengah	6000	Deferred
2	FACTORY	Jawa Barat	1500	Contract Awarded, Builder Appointed
3	WAREHOUSE	Jawa Barat	400	Deferred
4	FACTORY	Jawa Barat	1000	Subcontractors Appointed
5	SHOPS	D.I. Jakarta	1000	Contract Awarded, Builder Appointed
6	SHOPS	D.I. Jakarta	1050	Construction Commenced
7	WAREHOUSE	Jawa Tengah	400	Construction Commenced
8	FACTORY	Jawa Barat	2500	Construction Commenced
9	SHOPS	Jawa Timur	4950	Design Application
10	SHOPS	Jawa Barat	800	Construction Commenced
11	SHOPS	Jawa Barat	800	Construction Commenced
12	HOSPITAL	D.I. Jakarta	4200	Contract Awarded, Builder Appointed
13	HOSPITAL	D.I. Jakarta	8400	Contract Awarded, Builder Appointed
14	HOSPITAL	D.I. Jakarta	8400	Contract Awarded, Builder Appointed
15	APARTMENTS	Jawa Barat	2685	Deferred
16	APARTMENTS	Jawa Barat	7142	Deferred
17	OFFICES	D.I. Jakarta	1800	Deferred
18	FACTORY	D.I. Jakarta	7800	Subcontractors Appointed
19	APARTMENTS	D.I. Jakarta	24000	Subcontractors Appointed
20	....	...	...	...
21	FACTORY	Jawa Barat	4000	Subcontractors Appointed

Figure 2. Data structure

3.2 Data Sellection

This stage is used to clean unused or duplicate variable / attribute data. From the data in Figure 2. Attributes / variables used in processing the decision tree are the project name, region, project floor area, and project status.

3.3 Data Transformation

Adjustment of data from each attribute in Figure 2, so that it can be calculated using the C4.5 algorithm.

Pengelompokan	
Luas proyek	
<1000	Kecil
1000 - 4999	Sedang
> 5000	Besar
Status Proyek	
Abandoned	tidak
Building Application	baik
Building Approval	baik
Construction Commenced	baik
Construction Completed/Nearing Completion	baik
Contract Awarded, Builder Appointed	baik
Deferred	tidak
Design Application	baik
Design Approval	baik
Design Competition / Architect tender	baik
Documentation in Progress	baik
Main Contractor On Site	baik
Site Works Commenced	baik
Sketch Plans	tidak
Subcontractors Appointed	baik
Tenderers Listed	baik
Tenders Called	baik

No	Nama proyek	Region	Luas Proyek	Status	Rekomendasi
1	WAREHOUSE	Jawa Tengah	besar	tidak	tidak diambil
2	FACTORY	Jawa Barat	sedang	baik	Diambil
3	WAREHOUSE	Jawa Barat	kecil	tidak	tidak diambil
4	FACTORY	Jawa Barat	sedang	baik	Diambil
5	SHOPS	D.I. Jakarta	sedang	baik	diambil
6	SHOPS	D.I. Jakarta	sedang	baik	diambil
7	WAREHOUSE	Jawa Tengah	kecil	baik	tidak diambil
8	FACTORY	Jawa Barat	sedang	baik	Diambil
9	SHOPS	Jawa Timur	sedang	baik	tidak diambil
10	SHOPS	Jawa Barat	kecil	baik	tidak diambil
11	SHOPS	Jawa Barat	kecil	baik	tidak diambil
12	HOSPITAL	D.I. Jakarta	sedang	baik	diambil
13	HOSPITAL	D.I. Jakarta	besar	baik	diambil
14	HOSPITAL	D.I. Jakarta	besar	baik	diambil
15	APARTMENTS	Jawa Barat	sedang	tidak	tidak diambil
16	APARTMENTS	Jawa Barat	besar	tidak	tidak diambil
17	OFFICES	D.I. Jakarta	sedang	tidak	tidak diambil
18	FACTORY	D.I. Jakarta	besar	baik	diambil
19	APARTMENTS	D.I. Jakarta	besar	baik	diambil
	...	...	...	...	...
1989	FACTORY	Jawa Barat	sedang	baik	Diambil

Figure 3. Data transformation

3.4 Calculation of Entropy and Gain

Entropy is a measure of information theory that can determine the characteristics of the impurity and homogeneity of a data set. From the Entropy value, the information gain value for each attribute is calculated [8]. The calculation of the Entropy value uses a formula as in Equation (1) [12].

$Entropy(s) = \Sigma -ni{=}0pi \ *\log 2(pi)$

Formula (1) is the formula used in entropy calculations to determine how informative the attribute is. The following is the description:
s : Case set
n : Number of partitions
pi : The number of cases on the partition

Information gain is information obtained from changes in entropy in a data set, either through observation or it can also be concluded by participating in a data set [8].

$GAIN(S,A) = Entropy(S) - \Sigma - |Si| / |S| * Entropy(S) ni = 1 \ldots..(2)$

Formula (2) is the formula used in calculating information gain after calculating entropy. The following is the description:
s : Case Set
n : Number of partitions attribute A
$|Si|$: The number of cases on the partition
$|S|$: Number of cases in s

By knowing the formulas above, the data that has been obtained can be entered and processed using the C4.5 algorithm for the decision tree making process. [9]

A complete calculation of gain and entropy is shown in Figure 4.

No	Atribut		Jumlah kasus	Diambil	Tidak	Entropy	Gain
	Total		1989	985	1004	0.9999	
1	Nama Proyek						-5.46462
		Apartment	283	112	171	0.9684	
		Carpark	32	11	21	0.9284	
		Condominium	11	3	8	0.8454	
		Factory	353	278	75	0.7462	
		Hospital	42	31	11	0.8296	
		hotel	171	63	108	0.9495	
		market	248	62	186	0.8113	
		Office	68	29	39	0.9843	
		Plant	6	2	4	0.9183	
		Shop	445	216	229	0.9994	
		Showroom	33	18	15	0.9940	
		Warehouse	297	160	137	0.9957	
2	Region						-4.96607
		DKI jakarta	640	492	148	0.7802	
		Jawa Barat	663	344	319	0.9990	
		Jawa tengah	291	61	230	0.7408	
		Jawa Timur	395	88	307	0.7652	
3	Luas Proyek						-5.268
		Besar	849	516	333	0.9662	
		Kecil	422	56	366	0.5648	
		Sedang	718	413	305	0.9836	
4	Status						-4.23541
		Baik	1554	985	569	0.9477	
		Tidak	435	0	435	0.0000	

Figure 4. Entropy and gain calculation results

3.5 Implementation and Testing

After data collection and processing, the model is tested using the Rapidminer application. An explanation of each operator used in the Rapidminer application will be shown in Figure 5

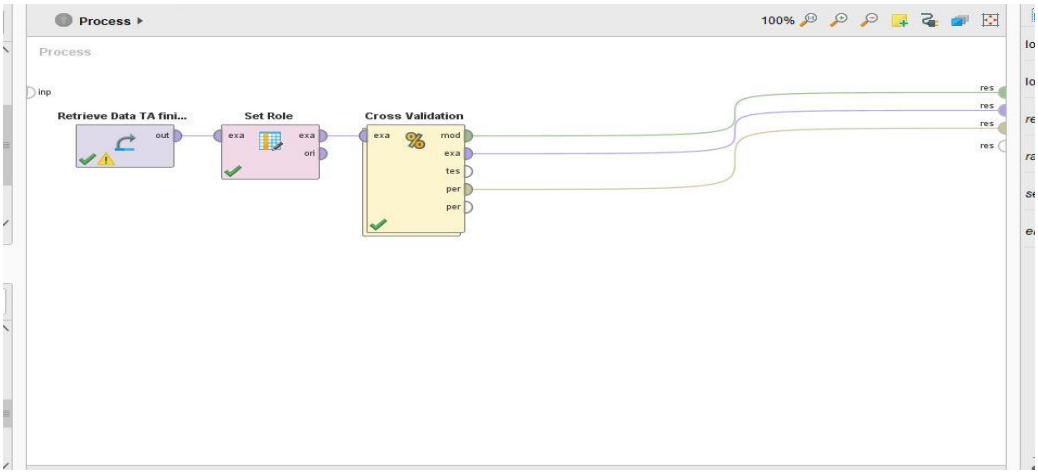

Figure 5. Testing Operators

- Read Excel operator: This operator is used to import data that will be used for testing from Microsoft Excel to Rapid Miner. After that, the user will be asked to specify a spreadsheet to use and select cells to import. [10]
- Operator Cross Validation: an operator that has two subprocesses that can be embedded in it and performs training and testing to train the model using the k-fold cross validation method.

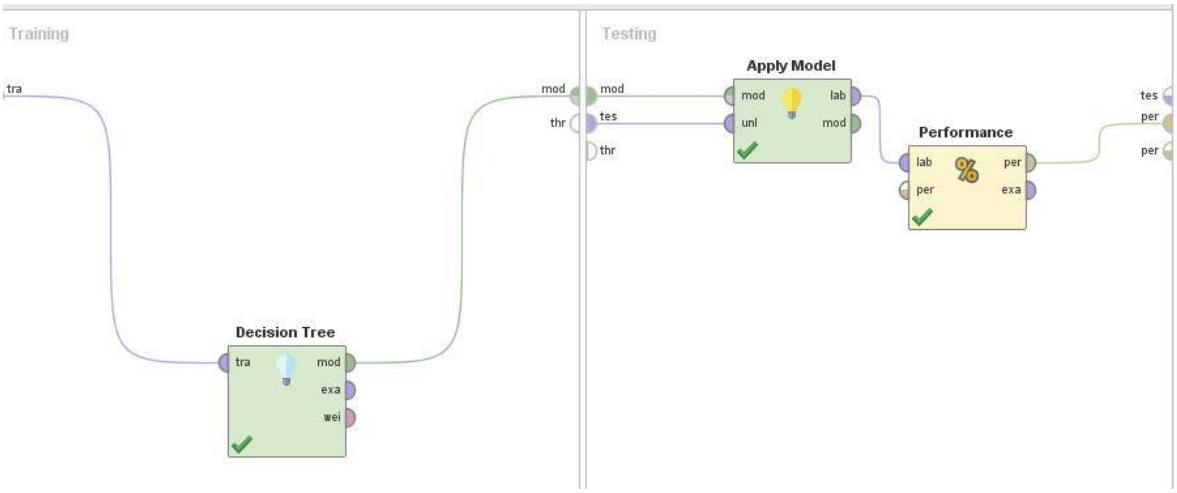

Figure 6. Operators in cross validation

- Operator Decision Tree: This operator is used to make decisions using the C4.5 algorithm. This operator is a tree of a collection of nodes. where each node represents a separation rule for one particular attribute and for a classification rule separates the values of different classes. In this operator, the results of the entropy and gain calculations will determine the criterion value, maximum depth, confidence, minimum gain, minimum leaf size, minimum size for split and number of pre pruning alternatives.
- Operator Apply Model: This operator is used to apply a trained model to the training data which can read the data to be estimated based on the training data.
- Operator Performance: This operator automatically determines all types of learning tasks. [10]

Evaluation is needed to analyze and measure the accuracy of the results obtained using the Confusion Matrix. Model validation is done using the Ten-Fold Cross Validation method. With validation it can be seen that all functions are working properly. Ten-fold Cross Validation is one of the K-folds recommended for selecting the best model because it can reduce

computation time while maintaining the accuracy of the estimation. Validation divides the data by dividing a data set into ten segments of equal size by randomizing the data. Validation and testing are carried out to determine the level of accuracy, precision, and recall of the classification prediction results. [10].

4.0 RESULANTS AND DISCUSSION

From the results of the gain and entropy calculations in Figure 4, the highest gain is -4.23451 on the "Status" label, it will be used as a node. Sequentially based on the gain from highest to lowest. Then get the following results in Figure 7.

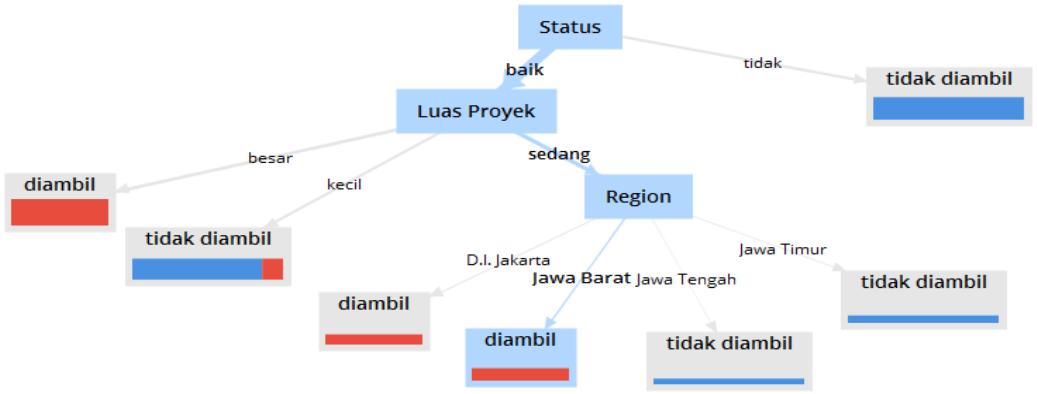

Figure 7. Decision Tree

From Figure 7, you get a role model in predicting project recommendations to be taken. In figure 8.

If Status = baik, Luas proyek = besar *then* Diambil
If Status = baik, Luas proyek = kecil *then* Tidak diambil.
If Status = baik, Luas proyek = sedang, Region = DKI Jakarta *Then* Diambil
If Status = baik, Luas proyek = sedang, Region = Jawa Barat *Then* Diambil
If Status = baik, Luas proyek = sedang, Region = Jawa Tengah *Then* Tidak Diambil
If Status = baik, Luas proyek = sedang, Region = Jawa Timur *Then* Tidak Diambil
If Status = Tidak *then* tidak diambil.

Tree

```
Status = baik
|   Luas Proyek = besar: diambil {tidak diambil=0, diambil=516}
|   Luas Proyek = kecil: tidak diambil {tidak diambil=349, diambil=56}
|   Luas Proyek = sedang
|   |   Region = D.I. Jakarta: diambil {tidak diambil=0, diambil=183}
|   |   Region = Jawa Barat: diambil {tidak diambil=0, diambil=230}
|   |   Region = Jawa Tengah: tidak diambil {tidak diambil=92, diambil=0}
|   |   Region = Jawa Timur: tidak diambil {tidak diambil=128, diambil=0}
Status = tidak: tidak diambil {tidak diambil=435, diambil=0}
```

Figure 8. Description

PerformanceVector

```
PerformanceVector:
accuracy: 97.18% +/- 1.26% (micro average: 97.18%)
ConfusionMatrix:
True:    tidak diambil    diambil
tidak diambil:  1004     56
diambil:        0        929
precision: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: diambil)
ConfusionMatrix:
True:    tidak diambil    diambil
tidak diambil:  1004     56
diambil:        0        929
recall: 94.31% +/- 2.56% (micro average: 94.31%) (positive class: diambil)
ConfusionMatrix:
True:    tidak diambil    diambil
tidak diambil:  1004     56
diambil:        0        929
AUC (optimistic): 1.000 +/- 0.000 (micro average: 1.000) (positive class: diambil)
AUC: 0.990 +/- 0.005 (micro average: 0.990) (positive class: diambil)
AUC (pessimistic): 0.980 +/- 0.010 (micro average: 0.980) (positive class: diambil)
```

Figure 9. Validation

	true tidak diambil	true diambil
pred. tidak diambil	1004	56
pred. diambil	0	929

Figure 10. Confusion Matrix

$$\text{Accuracy} = \left(\left(\frac{1004+928}{1989}\right) \times 100\%\right) = 97{,}18\%$$

$$\text{Precision} = \left(\left(\frac{929}{0+928}\right) \times 100\%\right) = 100\%$$

$$\text{Recall} = \left(\left(\frac{929}{56+928}\right) \times 100\%\right) = 94{,}31\%$$

Analysis

Based on the testing and analysis of the results of the tests carried out, with an accuracy rate of 97.18%, 100% precision, and 94.31% recall, it shows a value that is almost one hundred percent accurate, with very accurate precision and a recall that is still in the good category. With the result 1205 recommended floor construction projects were taken and 784 projects that were recommended not to be taken. Conclude that the researcher is successful in implementing the C4.5 classification algorithm properly and will assist the company in recommending whether or not a floor construction project will be taken.

5.0 CONCLUSION

5.1 Conclusion

The conclusions of the research conducted are as follows:

- The application of the C4.5 classification algorithm can be implemented in the process of determining the success of the project as seen from the accuracy level of 96.26% and a recall of 71.43%, which states that the calculations performed will be able to assist in the selection of a floor construction project.
- As a result, 1205 recommended floor construction projects were taken and 784 recommended projects not taken.
- With this result it can be done to determine which project can be taken based on the attributes that have been determined.

5.2 Advice

- Further studies from this research can try using other algorithms or the development of the C4.5 algorithm.
- GUI creation can be made based on the generated role model.

REFERENCES

[1] D. Ramayanti and U. Salamah, "Text Classification on Dataset of Marine and Fisheries Sciences Domain using Random Forest Classifier," *Int. J. Comput. Tech.*, vol. 5, no. 5, pp. 1–7, 2018.

[2] B. Ferdiansyah and L. Goeirmanto, "Prediksi Loyalitas dalam Keterikatan Karyawan terhadap Perusahaan Menggunakan Algoritma C4.5* (Studi Kasus PT.XYZ)," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 1, p. 87, 2020.

[3] M. F. Arifin and D. Fitrianah, "Penerapan Algoritma Klasifikasi C4.5 Dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada," *InComTech*, vol. 8, no. 2, pp. 87–102, 2018.

[4] A. Nursin, S. Wacono, and K. Kunci, "MODEL FOR ASSESSMENT OF PERFORMANCE CONSTRUCTION MANAGEMENT PROJECT," vol. 10, no. 1, 2011.

[5] J. Fondasi, M. Natalia, and Y. Partawijaya, "Analisis Critical Success Factors," vol. 6, no. 2, 2017.

[6] M. F. Sufa, "Identifikasi Kriteria Keberhasilan Proyek," *Performa*, vol. 11, no. 1, pp. 19–22, 2012.

[7] T. H. Setiawan and T. Ariadi, "Indikator Keberhasilan Proyek Pembangunan Bangunan Gedung Yang Dipengaruhi Faktor Internal Site Manager," vol. 11, no. 2, pp. 128–134, 2012.

[8] H. Hendrawan, "Faktor yang Mempengaruhi Keberhasilan Penerapan Teknologi Bidang Jalan dengan Kontrak Rancang Bangun," *Media Komun. Tek. Sipil*, vol. 24, no. 1, p. 45, 2018.

[9] Perdana, "済無No Title No Title," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2018.

[10] T. B. Santoso and D. Sekardiana, "PENERAPAN ALGORITMA C4.5 UNTUK PENENTUAN KELAYAKAN PEMBERIAN KREDIT (Studi Kasus : Koperia - Koperasi Warga Komplek Gandaria) Implementation of C4 . 5 Algorithm to Determine the Feasibility of Loan," *J. Algoritm. Log. dan Komputasi*, vol. ll, no. 1, pp. 130–137, 2019.