NEWTON: Networking and Information Technology
Vol. 2 No. 2 October 2022, Page. 79 - 83
E-ISSN: 2797-0728

Classification of Graduate Data Using The C4.5 Algorithm

Ratna Oktavia[1*], Agus Sifaunajah[2]
[1,2] Information Systems, KH. Abdul Wahab Hasbullah University
*Email: oktaviaratna93@gmail.com

ABSTRACT

Data mining is an activity that includes the collection, and use of historical data to find regularities, patterns, or relationships in large data sets. The pile of data in an information system gives rise to a data mining process. The Bahrul Ulum Islamic Boarding School Tambakberas Jombang is one of the oldest and largest Islamic boarding schools in East Java, with a large number of students each year. A large number of graduates in Islamic boarding schools requires the application of data mining processes. The c4.5 algorithm is one of the data mining algorithms that can be used to classify. The classification method is used to determine whether graduates are successful or not, with the parameter that becomes the root of the calculation results, namely work. The evaluation results show that the C4.5 algorithm has an accuracy rate of 84.00%, classification recall gets 93.33%, classification precision gets 85.00%, and an AUC score of 0.850.

***Keywords**: Data Mining, Klasifikasi, Algoritma C4.5*

INTRODUCTION

Data mining is an activity that includes the collection, and use of historical data to find regularities, patterns, or relationships in large data sets (Sularno & Anggraini, 2017). The pile of data in an information system gives rise to a data mining process. The c4.5 algorithm is one of the data mining algorithms that can be used to classify or predict. Classification is the process of finding a model that can distinguish data classes, one of the classification algorithm methods that can be used is the C4.5 algorithm method. The C4.5 method is a method with a decision tree technique (Ridwan, 2017). One of the benefits of a decision tree is that it breaks down complex decision-making processes into simpler ones so that the decisions taken can interpret the solution to the problem better. A large number of graduates in Islamic boarding schools require the application of data mining processes (Muhamad et al., 2019).

The Bahrul Ulum Islamic Boarding School Tambakberas Jombang is one of the oldest and largest Islamic boarding schools in East Java. The Bahrul Ulum Islamic Boarding School Tambakberas Jombang is under the auspices of the Bahrul Ulum Islamic Boarding School Foundation, this foundation was established in 1966. The Bahrul Ulum Islamic Boarding School Tambakberas Jombang has established 18 formal education units ranging from pre-school to tertiary levels. , with 18 graduates. more than 2000 students. In the development of this graduate data collection information system, the implementation of the c4.5 algorithm will be carried out to classify data with several specified parameters. This is to further maximize the potential of graduates for the Bahrul Ulum Islamic Boarding School Foundation.

METHOD

The research method used in this study is qualitative. The qualitative research method is a descriptive research and tends to use analysis. Process and meaning (subject perspective) are more highlighted in qualitative research. The theoretical basis is used as a guide so that the researcher's focus is on the facts on the ground. Qualitative research is known since the 1960s and is often called the alternative method. This method does not use detailed questions, but starts with general ones but then gets tapered and detailed.

Data Mining

Data mining is a series of processes to explore added value from a data set of data that has not been known manually, the process of data mining uses statistical techniques, mathematics, artificial intelligence, and machine learning to identify related information from large databases (Parapat & Sinaga, 2018).

C4.5 Algorithm

The C4.5 algorithm is an algorithm for classifying data that produces a decision tree (Novandya, 2017). Is a classification method that uses a tree structure representation, where each node represents an attribute, and a leaf represents a class, the topmost node of the decision tree is called the root (Anam & Santoso, 2018) . A decision tree is a tool that graphically depicts various activities that can be taken and associated with this activity with various future events that can occur (Bayu Febriyanto et al., 2018).
here are several stages in the C4.5 algorithm, namely (Asroni et al., 2018):

- Concept of Entropy
  Entropy (S) is the estimated number of bits needed to be able to extract a class (+ or -) from several random data in the sample space S. Entropy calculations are as in equation 1.

  $$Entropy\ (S) = \sum_{i=1}^{n} -pi * \log_2 pi \qquad (1)$$
- Gain
  Gain is obtained from the change in entropy in a data set. Calculation of gain as in equation 2.

  $$\text{Gain (S, A)} = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i) \qquad (2)$$
- Split Info
  Split info is a formula that expresses potential information or entropy. Calculation of split info as in equation 3.

  $$\text{Gain (S, A)} = -\sum_{i=1}^{n} \frac{Si}{S} log_2 \frac{Si}{S} \qquad (3)$$
- Gain Ratio
  A gain Ratio is a modification of the information gain which is used to reduce the bias of attributes that have many branches. The gain ratio has the following properties: large value if the data is evenly distributed and small value if all data enters one branch.

  $$\text{Gain Ratio (S, A)} = \frac{Gain\ (S,A)}{splitInfo\ (S,A)} \qquad (4)$$

RESULT AND DISCUSSION

The data used in this study is data based on the criteria used in the calculation, namely the alumni of Bahrul Ulum. The method proposed for the process as described above is a classification method with the algorithm used is the C4.5 algorithm with the following criteria used:

- Community activities
- Job
- Married or unmarried status
- Formal education

In the data criteria above, the target class is determined or called a label (target attribute). The target class used in the calculation is successful and not successful.

Result

At this stage, all the data is prepared the data to be entered into the calculation model, and the data is processed from the beginning. The data is processed in excel form by setting the target class first by using weighting. Calculate the entropy value from the training data with 27 unsuccessful alumni records and 23 successful alumni records so that entropy is obtained.

Tabel 1 Highest Gain Ratio

Attribute	**Number Of Cases**	**Has Not Succeeded**	**Succeed**	**Entropy**	**Gain**	**Split info**	**Gain ratio**
Total	50	23	27	0,9953			
Kegiatan masy					0,3609	3,0679	0,1176
Pbnu	3	0	3	0			
Pcnu	5	1	4	0,7219			
Mwcnu	3		3	0			
Pacnu	5	1	4	0,7219			
Ansor	7	3	4	0,9852			
Rantingnu	5	2	3	0,971			
Ippnu	4	2	2	1			
Takmir	4	1	1	0,8113			
tdk ada	11	11	0	0			
Pekerjaan					0,3357	0,795	**0,4223**
Sudah	38	11	27	0,868			
Belum	12	12	0	0			
Menikah					0,018	0,9988	0,018
sudah	26	10	16	0,9612			
belum	24	13	11	0,995			
Pendidikan					0,3971	2,6559	0,1418
Mts	5	5	0	0			
Ma	6	4	2	0,9183			
d1	10	7	3	0,8813			
D2	0	0	0	0			
D3	2	2	0	0			
S1	9	2	7	0,7642			
S2	12	3	9	0,8113			
S3	6	0	6	0			

It can be seen in the table above that the root node is located the job attribute has a value already and not. The value has not yet become a classification, getting a decision has not been successful. And for the value, it needs to be calculated again because it still has no and successful results, the calculation is carried out to determine the next root node, then the decision tree can be described from the table above as follows:

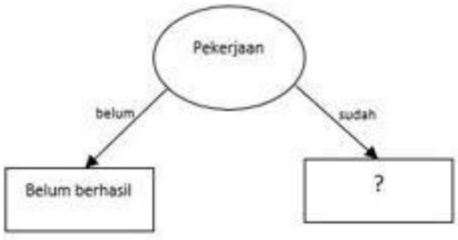

Figure 1. Decision Tree 1

The next step is to calculate the root node already, using the method above, by finding the gain, split info, and gain ratio of each attribute to find the next root node. In the picture below is the result of the final calculation of the decision tree as follows:

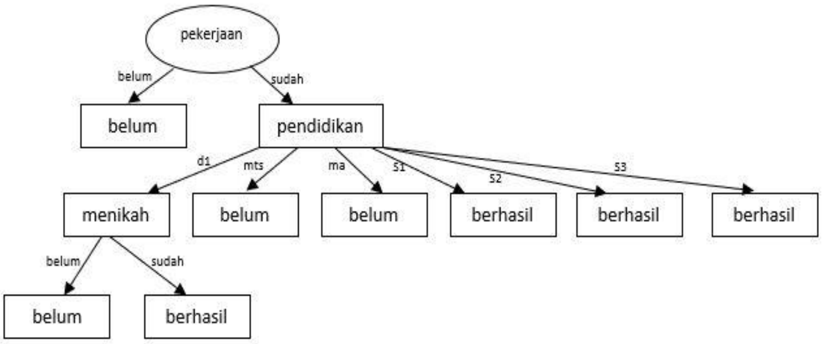

Figure 2. Final Decision Tree

The picture above shows the results of a complete description of the decision tree that has been formed using the C4.5 algorithm. the results of the description also show that the use of data mining algorithm C4.5 is good for use in the process of extracting data (data mining process) to draw conclusions that are visualized with a decision tree (decision tree). The following rules are generated from the decision tree:

Tabel 2 Generated Rules

Rule
1. *If* Pekerjaan = belum *Then* Hasil = Belum Berhasil
2. *If* Pekerjaan = sudah *And* Pendidikan formal = d1 *And* Menikah = belum *Then* Hasil = Belum Berhasil
3. *If* Pekerjaan = sudah *And* Pendidikan formal = d1 *And* Menikah = sudah *Then* Hasil = Berhasil
4. *If* Pekerjaan = Sudah *And* Pendidikan formal = Mts sederajat *Then* Hasil = Belum Berhasil
5. *If* Pekerjaan = Sudah *And* Pendidikan formal = MA sederajat *Then* Hasil = Belum Berhasil
6. *If* Pekerjaan = Sudah *And* Pendidikan formal = S1 *Then* Hasil = Berhasil
7. *If* Pekerjaan = Sudah *And* Pendidikan formal = S2 *Then* Hasil = Berhasil
8. *If* Pekerjaan = Sudah *And* Pendidikan formal = S3 *Then* Hasil = Berhasil

Discussion

Based on the results of the research above, it can be seen that there are 302 graduate data that have been successfully obtained. In the resulting decision tree, there is a new fact that the majority of alumni are already working, this is indicated by making work items as decision tree nodes. When it is known that the majority of graduates are already working, the foundation can begin to map out what jobs they do. From this mapping, the foundation can take advantage of the work map to then be processed into product segmentation produced by the foundation, so that it can be financially optimized to become a new source of finance for the foundation so that the foundation can establish and determine new financial sources with the new financial source, the foundation. can form scholarships that can be used by families of graduates who are in the poor category so that education equality can be created well at all levels.

CONCLUSION

With a large number of graduates in Islamic boarding schools and a large number of data hoards, data mining processes are urgently needed. Based on the results of research that has been carried out on pesantren graduates using the C4.5 algorithm classification method, it can be concluded that the problem of determining successful and unsuccessful graduates can be solved using data mining techniques with the C4.5 algorithm method by producing 8 rules. Based on the decision tree, the job becomes root because it has the highest gain ratio compared to other attributes. The level of accuracy produced by this method is 84.00% by doing 10 tests on the dataset.

With the application of this algorithm, the performance of the algorithm is very suitable for the classification of graduates, and with the application of the algorithm, it is hoped that it will be able to provide a solution to the Bahrul Ulum Islamic Boarding School Foundation to determine successful or unsuccessful graduates.

REFERENCES

Anam, C., & Santoso, H. B. (2018). Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa. *Energy - Jurnal Ilmiah Ilmu-Ilmu Teknik*, *8*(1).

Asroni, A., Masajeng Respati, B., & Riyadi, S. (2018). Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta. *Semesta Teknika*, *21*(2). https://doi.org/10.18196/st.212222

Bayu Febriyanto, D., Handoko, L., & Aisyah, H. (2018). Implementasi Algoritma C4.5 Untuk Klasifikasi Tingkat Kepuasan Pembeli Online Shop. *Jurnal Riset Komputer (JURIKOM)*, *5*(6).

Muhamad, M., Windarto, A. P., & Suhada, S. (2019). PENERAPAN ALGORITMA C4.5 PADA KLASIFIKASI POTENSI SISWA DROP OUT. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1). https://doi.org/10.30865/komik.v3i1.1688

Novandya, A. (2017). Penerapan Algoritma Klasifikasi Data Mining C4.5 pada Dataset Cuaca Wilayah Bekasi. *KNiST*.

Parapat, J. S., & Sinaga, A. S. (2018). Data Mining Algoritma C4.5 Pada Klasifikasi Kredit Koperasi Simpan Pinjam. *Jurnal Ilmu Teknik Elektro Komputer Dan Informatika (JITEKI)*, *4*(2).

Ridwan, M. (2017). Sistem Rekomendasi Proses Kelulusan Mahasiswaberbasis Algoritma Klasifikasi C4.5. *Jurnal Ilmiah Informatika*, *2*(1), 105–111. https://doi.org/10.35316/jimi.v2i1.460

Sularno, S., & Anggraini, P. (2017). Penerapan Algoritma C4.5 Untuk Klasifikasi Tingkat Keganasan Hama Pada Tanaman Padi. *Jurnal Sains Dan Informatika*, *3*(2).