Jurnal Mantik
Volume 4, Number 3, Nov 2020, pp. 2041-2048 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Implementation of C4.5 Agorithm in predicting the timeliness of student graduation (Case Study: Informatics Management Study Program, Labuhanbatu University)

Syaiful Zuhri Harahap[1], Musthafa Haris Munandar[2], Masrizal[3], Ibnu Rasyid Munthe[4], Meidy Putra Panusunan Siregar[5]

[1,2,3,5]Information System, Faculty of Science & Technology, Labuhanbatu University, Jln.S.M. Raja No.126 A Aek Tapa Rantauprapat, Kab. Labuhanbatu Sumatera Utara, Indonesia
[4]informatics Management, Faculty of Science & Technology, Labuhanbatu University, Jln.S.M. Raja No.126 A Aek Tapa Rantauprapat, Kab. Labuhanbatu Sumatera Utara, Indonesia

E-mail: syaifulzuhriharahap@gmail.com[1], harismunandaaar@gmail.com[2], masrizal120405@gmail.com[3], ibnurasyidmunthe@gmail.com[4], meidysiregar12345@gmail.com[5]

ARTICLE INFO

Article history:
Received: 12/07/2020
Revised: 22/08/2020
Accepted: 30/11/2020

Keywords:
C4.5 algorithm, Prediction, Student Graduation, Data Mining, Knowledge

ABSTRACT

The high success rate of college students and the low student failure rate represent the college's standard. College Student loss and its causal causes are interesting subjects for study. Colleges are actually in a very demanding climate. Any college aims to continually develop its management to improve the quality of its education and improve its accreditation. One feature of the college accreditation examination is time graduation. Also, timely graduation is an essential concern because graduation rates are the foundation for its effectiveness. One of the problems that are now the topic of talking about academic failure is students' dropout and graduation. Data mining is a technique of tracing current data to create a model and then using it to recognize other data patterns not contained in the database. One of the techniques used in data mining is the classification methodology using the C4.5 process.

1. Introduction

Colleges are actually in a very demanding climate. Any college aims to continually develop its management to improve its education quality and improve its accreditation. One feature of the college accreditation examination is time graduation. Also, timely graduation is an essential concern because graduation rates are the foundation for its effectiveness. One of the problems that are now the topic of talking about academic failure is students' dropout and graduation. Non-active students do not enroll at the beginning of the semester or do not attend lectures for at least one semester. Data mining is an organization law and is a business basket research technique. The market basket is characterized as an object that is purchased concurrently by the consumer in a transaction. Business basketball research is a valuable method for cross-selling marketing. The pattern is defined by two criteria, namely support and trust.

C4.5 is a decision tree classification algorithm commonly used since it has the key benefits of other algorithms. The benefits of the C4.5 algorithm can result in an easy-to-achieve decision tree, have a reasonable degree of precision, are useful in handling discrete-type attributes, and can handle discrete and numerical attributes.[1]

2. Method

The C4.5 algorithm is the decision tree algorithm. The decision tree is a very powerful and well-known classification and prediction tool. The system of decision tree converts a comprehensive reality into a decision tree that represents the rules. Rules of natural language can be readily understood. Moreover, they can also be represented in database languages like the standardized query language for searching for documents in a particular group.[2] In general, the C4.5 algorithm for building decision trees is as follows:

a) Select an attribute as the root
b) Create a branch for each value.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

c) Share the case in the branch.
d) Repeat the process for each branch until all cases on the branch have the same class.

It is dependent on the maximum benefit value of the current attributes to pick a root attribute. To measure the gain, see the entropy value first. A formula, as shown below, is used to measure the entropy value:

$$Entropy(S) = \sum_{j=1}^{n} - Pj \log 2 \ Pj$$

Description:
S = Case set (*Entropy)*
n = Banyaknya partition S
Pj = Probability that can be from the class divided by the total case

Then calculate the *gain value using* the following formula:

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

Description:
S: Case set
A: Attributes
n: Number of partition attributes A
|Si|: Number of cases on the partition to i
|S|: Number of cases in S

In general, the steps of the C4.5 algorithm to build the decision tree are as follows:

a) Calculate the benefit ratio of each current training data attribute, split information, and entropy.
b) Build the root node of the collection attribute with the highest benefit ratio.
c) Calculate the benefit ratio, separate info, and entropy by eliminating previously chosen attributes for each attribute.
d) Build an internal node from the collection of the most relevant attributes.
e) Check if all attributes on the tree have been created. If not, repeat process d and e, if this is already the next process.
f) Do tree cutting to remove unwanted trees.[3]

Centered on the study of the problem, the C4.5 program for data mining is intended as an alternative to knowledge presentation and consulting on forecasts and views for student graduations to forecast the timeliness of the created graduation, as an application that can define the prediction of time by IP value. Polytechnical campuses have made many attempts to know the predicted benefit in education, such as technical advice, testing, etc. Therefore an application must be made to help forecast student grades on campus with data collection methods.

A. C4.5 Algorithm

The C4.5 algorithm shapes the decision tree algorithm. The decision tree is a compelling and successful prediction process. The decision tree system transforms the decision tree into a decision tree, which reflects the rules. Natural language rules can be readily interpreted. Moreover, they can also be represented as database languages, such as Organized Query Language, to locate records within a specific group. It is dependent on the maximum benefit value of the current attributes to pick a root attribute. To measure the gain, see the entropy value first. The entropy value is determined using the formula as described below:[4]

$$Entropy(S) = \sum_{j=1}^{n} - Pj \log 2 \ Pj$$

Description:
S = Case set (*Entropy)*
n = Partition S
Pj = Probability that can be from the class divided by the total case

Then calculate the *gain value using* the following formula:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

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

Description:
S: Case set
A: Attributes
n: Number of partition attributes A
|Si|: Number of cases on the partition to i
|S|: Number of cases in S

The measures of the C4.5 algorithm for the decision tree are usually as follows:

a) Calculate the benefit ratio of each current training data attribute, split information, and entropy.
b) Build the root node with the highest benefit ratio in the attribute selection.
c) Calculate the benefit ratio, divide information, and entropy for each attribute by deleting previously chosen attributes.
d) Build an internal node from the collection of the most relevant attributes.
e) Verify that all tree attributes are created. If not, repeat process d and e if this is already the next process.
f) Do tree cutting to remove trees unwanted[5]

Flow Chart Algorithm Method C.45:

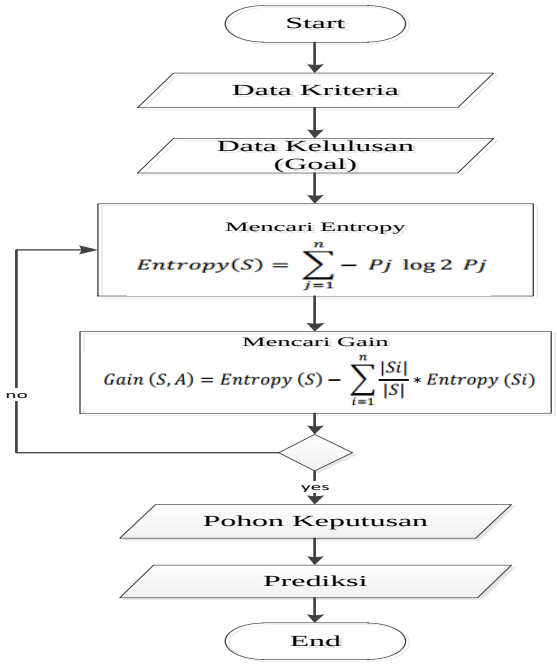

Fig 1. Flow Chart Algorithm Method C.45

3. Results and Discussion

A. Attribute: Gender

Total Men 234 Exactly 123 Late 111
Entropy (123,111) = [-(123/234) * log2(123/234)] + [-(111/234) * log2(111/234)] = 0.9981
Total Women 145 Exactly 97 Late 48
Entropy (97.48) = [(97/145) * log2(97/145)] + [-(48/145) * log2(48/145)] = 0.9160
Gain = 0.9812 - [(234/379) * 0.9981 + (145/379) * 0.9160] = **0.0145**

B. Attribute: Work

Total Work 133 Exactly 13 Late 120
Entropy (13,120) = [-(13/133) * log2(13/133)] + [-(120/133) * log2(120/133)] = 0.4618
Total Not Working 246 Exactly 207 Late 39
Entropy (207.39) = [-(207/246) * log2(207/246)] + [-(39/246) * log2(39/246)] = 0.6308
Gain = 0.9812 - [(133/379) * 0.4618 + (246/379) * 0.6308] = **0.4097**

C. Attribute: Age

Total Age ≤ 25 years 209 Exactly 158 Late 51

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Entropy (158.51) = [-(158/209) * log2(158/209)] + [-(51/209) * log2(51/209)] = 0.8017
Total Age>25 years 170 Exactly 62 Late 108
Entropy (62,108) = [(62/170) * log2(62/170)] + [-(108/170) * log2(108/170)] = 0.9465
Gain = 0.9812 - [(209/379) * 0.8017 + (170/379) * 0.9465] = **0.1146**

D. Attribute: Marriage

Total Married 8 Exactly 2 Late 6
Entropy (2.6) = [-(2/8) * log2(2/8)] + [-(6/8) * log2(6/8)] = 0.8113
Total Unmarried 371 Exact 218 Late 153
Entropy (218,153) = [(218/371) * log2(218/371)] + [(153/371) * log2(153/371)] = 0.9777
Gain = 0.9812 - [(8/379) * 0.8113 + (371/379) * 0.9777] = **0.0070**

E. Attribute: IPK

Total IPK ≤ 2.83 191 Exactly 75 Late 116
Entropy (75,116) = [-(75/191) * log2(75/191)] + [-(116/191) * log2(116/191)] = 0.9665
Total IPK>2.83 188 Exactly 145 Late 43
Entropy (145.43) = [(145/188) * log2(145/188)] + [-(43/188) * log2(43/188)] = 0.7758
Gain = 0.9812 - [(191/379) * 0.9665 + (188/379) * 0.7758] = **0.1093**

Table 2.
Attribute Assessment

Attribute	Information Gain
Gender	0.0145
Work	**0.4097**
Age	0.1146
Married	0.0070
IPK	0.1093

From the value of the above gain information, the job gain is the larger. Then the attribute is used as *the initial node.*

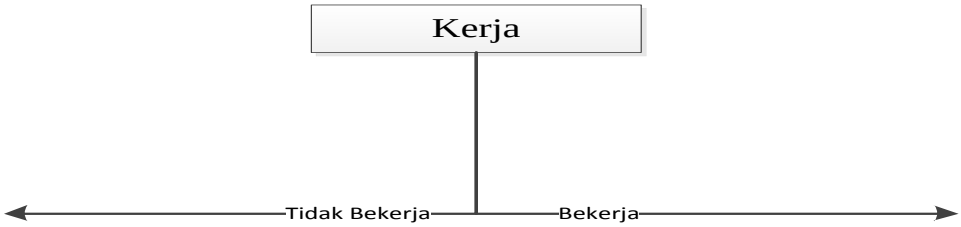

Fig 2. The decision tree on the initial node

So it is necessary to call the C.45 function with the sample set "working" with the target ="Precise" and "late." Once calculated, the biggest gain is IPK gain. Then produce a tree-like Figure.3 below:

Fig 3. Tree decision on level 0 recursion in 1st iteration

Next, it is necessary to call the C.45 function with the sample set "not working" with target="Precise" and "late." Once calculated, the biggest gain is IPK gain. Then produce a tree-like Figure.4 below:

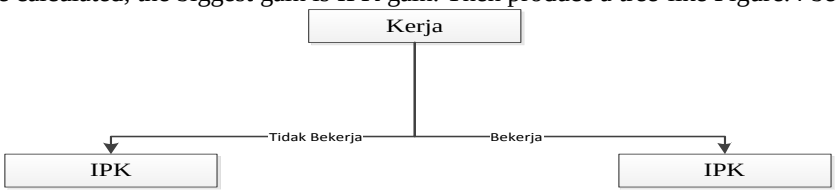

Fig 4. The decision tree on level 1 recursion in the first iteration

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Next set the function C.45 with the sample set "work" = " IPK" = " <=2.83" with targets = "Precise" and "late". Since the sample belongs to the "late" class, this function stops and creates a single node with the label "late." At this stage, produce a tree-like Figure.5 below:

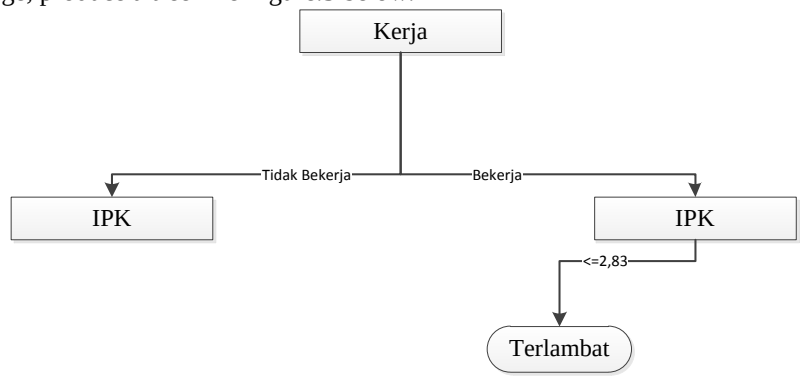

Fig 5. Tree decision on recursion level 0 in the second iteration

Next set the function C.45 with sample set "work" = "IPK" = " >2,83" with target ="Precise" and "late". Once calculated, the biggest gain is the "Gender" gain. Then draw a tree-like Figure 6. below:

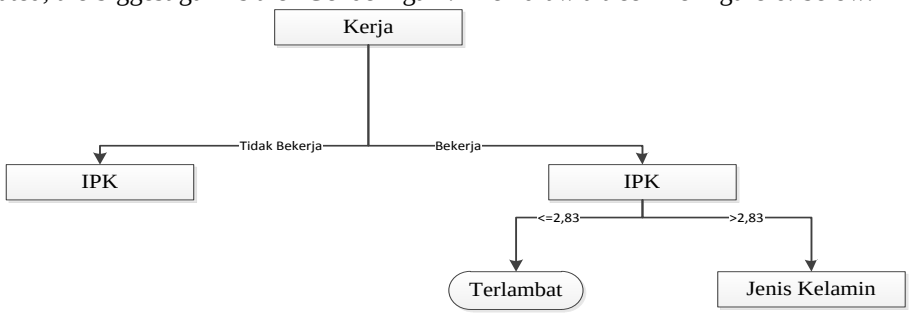

Fig 6. The decision tree on level 1 recursion in the second iteration

Next set the function C.45 with the sample set "Not working" = " IPK" = " >2,83" with targets = "Precise" and "late". Since the sample belongs to the "Right" class, this function stops and creates a single node with the label "Right". At this stage produce a tree like Figure.7 below:

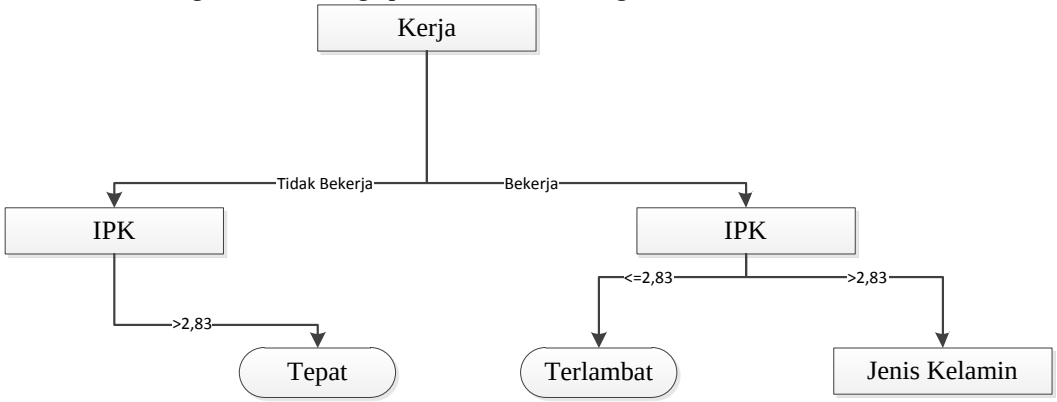

Fig 7. The decision tree on level 3 recursion in the second iteration

Next set the function C.45 with the sample set "Not working" = " IPK" = " <=2.83" with targets = "Precise" and "late". Once calculated, the biggest gain is the "Age" gain. Then produce a tree-like Figure.8 below:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

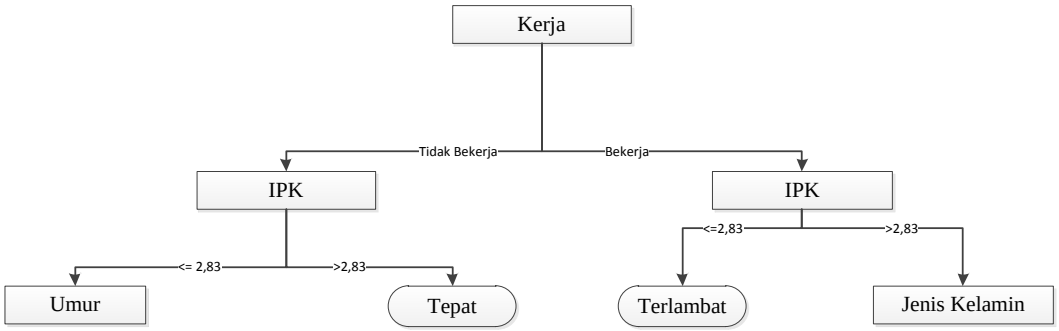

Fig 8. The decision tree on level 4 recursion in the second iteration

Next set the function C.45 with the sample set "work" = " IPK" = " >2,83" = "Gender" = "Male" with target = "Precise" and "late". Since the sample belongs to the "Late" class, this function stops and creates a single node with the label "Too Late". At this stage, produce a tree-like Figure.9 below:

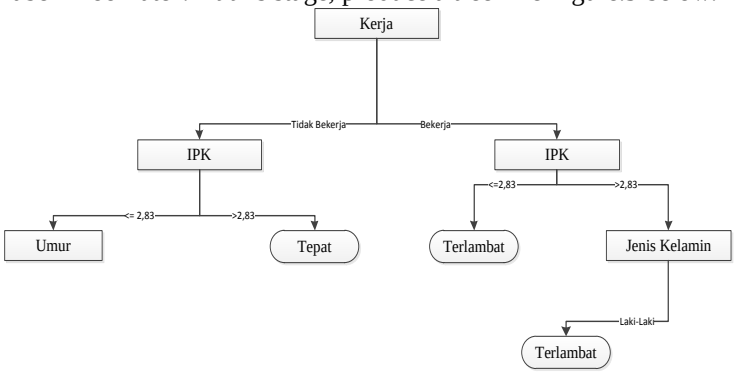

Fig 9. The decision tree on level 0 recursion in the third iteration

Next set the function C.45 with the sample set "work" = " IPK" = " >2,83" = "Gender" = "Female" with target = "Precise" and "late". Once calculated the biggest gain is the "Age" gain. Then produce a tree like Figure.I10 below:

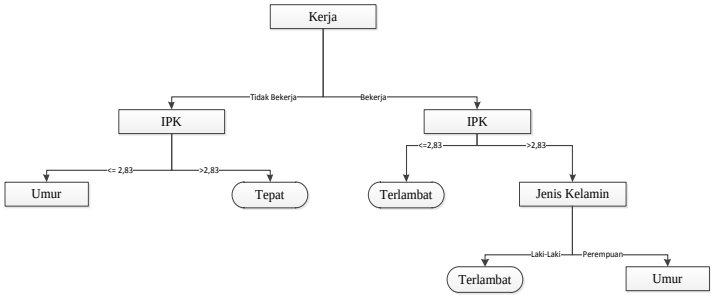

Fig 10. The decision tree on level 1 recursion in 3rd iteration

Next set the function C.45 with the sample set "Not working" = "IPK" = " <=2,83" = "Age" = "<=25" with targets = "Precise" and "late". Since the sample belongs to the "Late" class, this function stops and creates a single node with the label "Too Late". At this stage, produce a tree-like Figure.11 below:

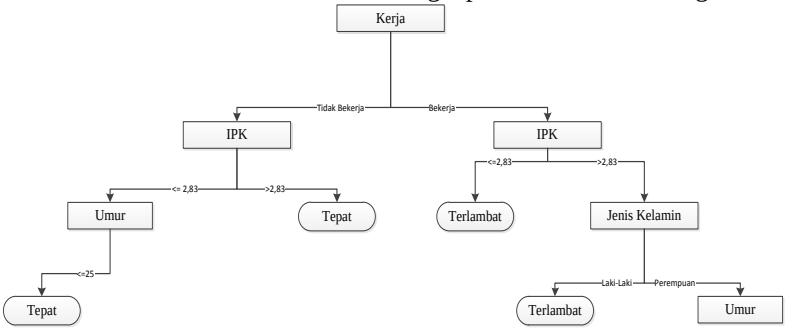

Fig 11. The decision tree on level 2 recursion in 3rd iteration

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Next set the function C.45 with the sample set "Not working" = " IPK" = " <=2,83" = "Age" = ">25" with targets = "Precise" and "late". Once calculated the biggest gain is the "Marry" gain. Then produce a tree like Figure.12 below:

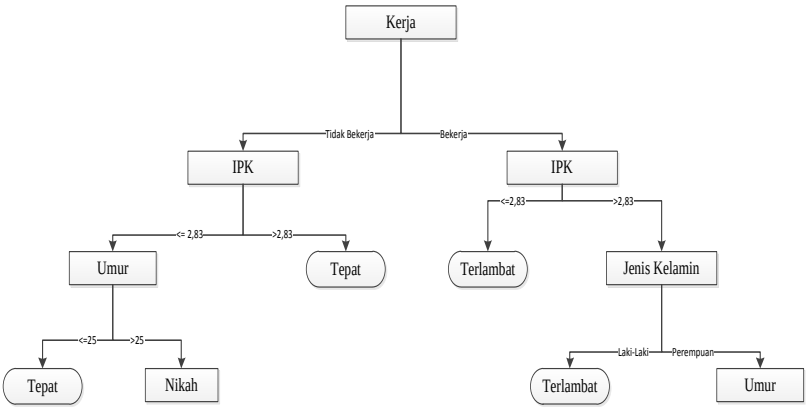

Fig 12. Tree decision on level 3 recursion in 3rd iteration

Next set the function C.45 with the sample set "work" = "IPK" = " >2,83" = "Gender" = "Female" = "Age" = "<=25" with targets = "Precise" and "late". Since the sample belongs to the "Right" class then this function stops and creates a single node with the label "Right". Then produce a tree like Figure.III.13 below:

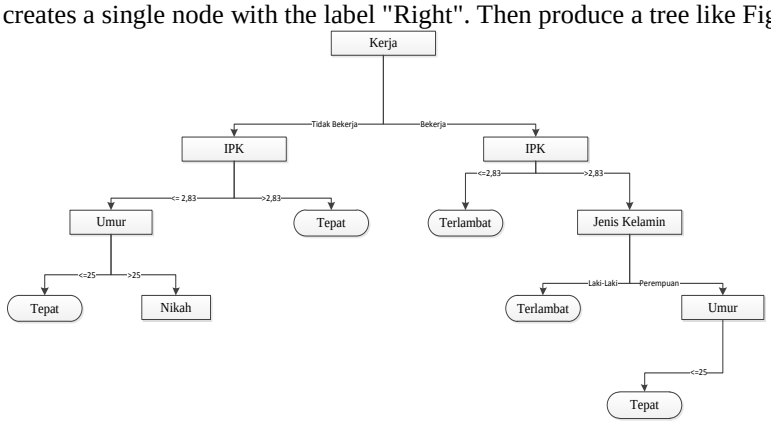

Fig 13. Tree decision on level 0 recursion in 4th iteration

In the previous function guidanc the previous sample call was "work" = "IPK" = " >2,83" = "Gender" = "Female" = "Age" = "<=25" returns "Appropriate" and no more nodes will be processed so as to generate the tree in Figure III.14 below:

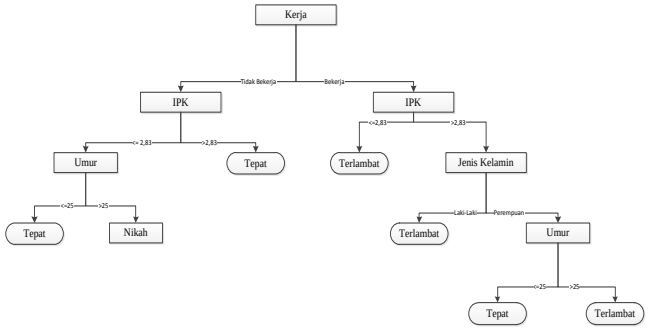

Fig 14. The decision tree on level 1 recursion in 4th iteration

Next set the function C.45 with the sample set "Not working" = " IPK" = " <=2,83" = "Age" = ">25" = "Marriage" = "Unmarried" with the target = "Exact" and "late". Since the sample belongs to the "Right" class then this function stops and creates a single node with the label "Right". Then produce a tree like Figure.15 below:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

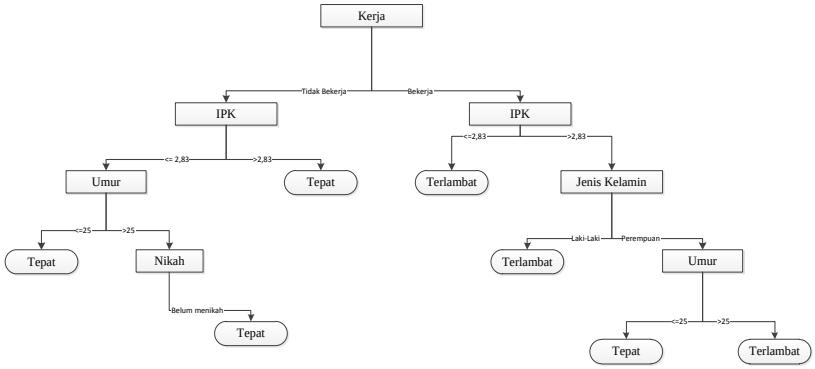

Fig 15. The decision tree on level 2 recursion in 4th iteration

In the previous function monitoring the previous sample call was "Not working" = "IPK" = " <=2,83" = "Age" = ">25" = "Marriage" = "Unmarried" returns "Appropriate" and no more nodes will be processed resulting in the tree in Figure 16. below:

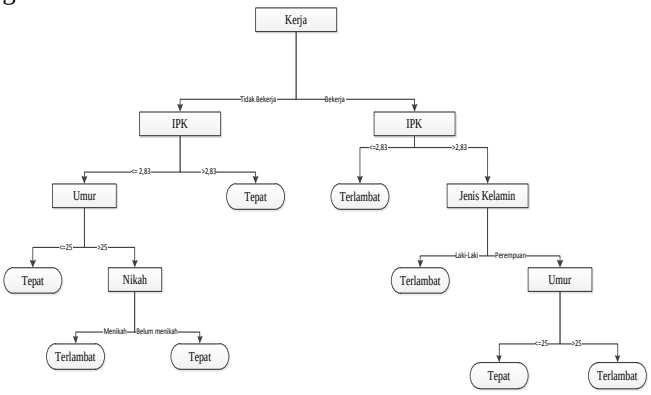

Figure 16. The decision tree on level 3 recursion in 4th iteration

4. Conclusion

Based on research performed during data mining using the C4.5 algorithm to predict the timeliness of graduation at Ganesha Polytechnic, the findings can be retrieved by the next researcher who will address the issue of the graduation value forecast method. It makes estimating graduation time simpler for students. The input data used for the device design are the total IPK value. The system will produce the performance for the predictive status system of students who are graduated promptly.

5. References

[1] A. Nastuti and S. Z. Harahap, "Amelia Nastuti, Syaiful Zuhri Harahap," Tek. DATA Min. UNTUK PENENTUAN PAKET HEMAT SEMBAKO DAN KEBUTUHAN Hari. DENGAN MENGGUNAKAN Algoritm. FP-GROWTH (STUDI KASUS DI ULFAMART LUBUK ALUNG), vol. 7, no. 3, pp. 111–119, 2019.

[2] R. Setiawan, "Analisis Kelayakan Pemberian Kredit Nasabah Koperasi Menggunakan Algoritma C4.5," Techno Xplore J. Ilmu Komput. dan Teknol. Inf., vol. 5, no. 2, pp. 74–78, Nov. 2020, doi: 10.36805/technoxplore.v5i2.1175.

[3] ibnu R. Munthe and V. Sihombing, "Klasifikasi Algoritma Iterative Dichotomizer (ID3) untuk Tingkat kepuasan pada Sarana Laboratorium Komputer," J. Teknol. dan Ilmu Komput. Prima, vol. 1, no. 2, pp. 27–34, Oct. 2018, doi: 10.34012/jutikomp.v1i2.237.

[4] Mochammad Yusa, E. Utami, and E. T. Luthfi, "Evaluasi Performa Algoritma Klasifikasi Decision Tree ID3 , C4,5 dan Cart pada Dataset Readmisi Pasien Diabetes," InfoSys J., vol. 4, no. 1, pp. 23–34, 2016.

[5] B. Hermanto, A. Sn, and F. P. Putra, "Analisis Kinerja Decision Tree C4.5 Dalam Prediksi Potensi Pelunasan Kredit Calon Debitur," Inovtek Polbeng Seri Inform., vol. 2, no. 2, 2017.