International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 11 Nov. 2022, pp: 386-392 www.ijaem.net ISSN: 2395-5252

# Analysis of Course Specification in First Grade Students Using Data Mining and C4.5 Algorithm

Ratih Nurdiyani Sari(1), Syarifah Azharina Syafrudin(2), Indah Wahyuni(3), Eka Fitri Rahayu(4)
1, 2 ,3 ,4 *Gunadarma University*

Submitted: 05-11-2022 Accepted: 15-11-2022

**ABSTRACT**: Determining the courses by first grade students in the Information Systems department according to their wishes and interests is something that is expected. But it is not easy to choose the appropriate specialization due to their limited information. Various obstacles and confusing factors make it difficult for students to choose the specialization according to their wishes. Therefore, the researcher took the initiative to create a system to help students choose data mining-based specializations using the C4.5 Algorithm method. The selection of specializations is determined based on certain courses that have been carried out in the second semester or when they were in the first grade. The system through interest in courses using the C4.5 Algorithm method is expected to help students choose course interests as desired to optimize student academic achievement and become a note for the campus in utilizing the courses taught as the basis for the success of subsequent lectures. This study produces a decision tree that will show which subjects are in high demand or less in each class sampled in this study.

**KEYWORDS:**Course Specialization, Decision Tree, Data Mining, C4.5 Algorithm.

## I. INTRODUCTION

Along with the advancement of science and technology. Directed information will be needed to help make decisions. Decision taking is an activity of choosing an action to solve a problem. Educational institutions such as universities are also involved in problem-solving activities, as for the form of decisions that are carried out, namely in terms of determining the appropriate specialization of subjects for students so that learning is expected to be in accordance with their interests. Specialization in courses at the beginning of the lecture allows students to develop themselves and can study in depth certain subjects according to their respective interests and expertise. There are many obstacles in determining the interest in the courses that match the criteria, moreover some students just follow the others in determining the specialization of the courses and also lack confidence in their abilities.

Lectures which are usually taught at the university level or are called courses now vary depending on the campus and majors. One of the universities has even prepared courses to be taken by students on the campus according to their field groups to be taken in a certain semester[1]. In this study, the author analyses the specialization of courses in first grade students to get the final result in the form of which courses are the most desirable and less desirable by first grade students with a data mining approach and the C4.5 Algorithm method. The results obtained are expected to be a reference by the University in developing educational and informative courses to attract students' interest in studying these subjects. A limitation in this study in the form of data taken by the author of 75 cases and consists of three classes.

## II. THEORETICAL GROUNDS

Data Mining is a process or method used to find information or certain patterns that are useful and are taken from a large set of data (Meilani and Slamat, 2013)[2]. Data Mining has several techniques for classifying and one of them is Decision Tree. The data mining method used in this research and used in the formation of the decision tree is the C4.5 Algorithm. The C4.5 algorithm will later form a decision tree starting from the top and then down, where the top attribute is the root or called the node that will represent the attribute and the one at the bottom is called the leaf which is used to represent the class [3]. The C4.5 algorithm requires a calculation, namely the Entropy Value and Gain value of an attribute, an equation formula will be used which can be seen below [4]:

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

Description:
S: Sets of Cases
A: Attribute
n: Number of partition attribute A
|Si|: Number of cases on the 1st partition
|S|: Number of cases in S

The calculation of the Entropy value can be seen in the formula listed below.

$$\mathrm{Entropy}(S) = \sum_{i=1}^{n} - pi * \mathrm{Log}_2 pi$$

Description:
S: Sets of Cases
n: Number of partitions S
pi: Proportion of $S_i$ to S

## III. RESEARCH METHODS

The stages carried out in this study consisted of three phases, namely the Preparation Phase, Pre-Processing Phase and Post-Processing Phase. Each phase has a different task; the following is an overview of the research flow that will be discussed in this study that can be seen in Figure 1.

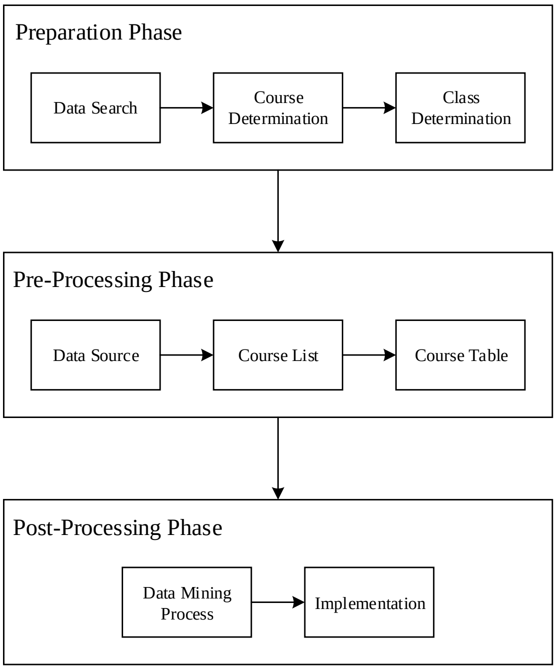


**Figure 1, Research Flow**

The diagram above shows the flow of the process that occurs in the research. Starting with the Preparation Phase for preparing the required data such as searching for files through a questionnaire, determining courses and determining the class that will be used to make a decision tree. Next is the Pre-Processing Phase, in this phase the data sources will be processed into a list of courses and course tables that will be used for the calculation process later. The last is the Post-Processing Phase; this last phase is processing the existing data to form a decision tree. After the decision tree, the last step is to look at its implementation in existing courses and classes, which ones attract students and which are less attractive.

## IV. DISCUSSION

The general steps for building a decision tree on a C4.5 algorithm are as follows [6], namely:

1. Select attribute as root
2. Create a branch for each value
3. Split cases in branches
4. Repeat the process until all cases on the branch have the same class. The selection of attributes as roots is based on the highest gain value of the existing attributes.

Calculations using the C4.5 algorithm to determine student interest in the courses they are undergoing will involve 75 data samples that have been taken from each of 3 class representatives. Filling in the node calculation table uses the number of samples contained in each category. The Class 01, Class 02 and Class 03 attributes contained in the node calculation table indicate that each sample application has a case in each category. Pay attention to Figure 2.

After the node calculation table is filled with the number of cases in each class, the next step is to calculate the entropy of all cases based on the category of courses and classes as well as the Gain calculation for each attribute. Row Total column entropy is calculated using the following entropy equation:

$$\mathrm{Entropy(Total)} = (-\frac{50}{75} * \log 2\left(\frac{50}{75}\right))) + (-\frac{25}{75} * \log 2\left(\frac{25}{75}\right)))$$

Entropy(Total)= 0,91829583

The total entropy of the number of course cases is 0.91829583, which will later become the root of the formula to calculate Gain. In this case, if you have created a formula at the beginning using Excel, then you only need to drag the formula from the top row to the bottom; the entropy value will automatically come out. Gain value calculation using the following formula equation.

$$\mathrm{Gain(Total, Programming\ Algorithm)} = 0{,}91829583 - \left(\frac{26}{75} * 0{,}89\right) + \left(\frac{24}{75} * 0{,}98\right) + \left(\frac{25}{75} * 0{,}85\right)$$

Gain(Total, Programming Algorithm) = 0,010883793

The calculation above is an example of calculating the Gain value where the programming algorithm course acts as an attribute. Gain value will be calculated for each subject tested and adjusted to the value of each entropy. After the calculation,the next step is to search for the largest Gain value, the course that has the largest Gain value will be the root or initial node of the research. After all entropy values are filled in, then the gain value is calculated and determined. After all the gains in the node research table are calculated, determine the largest gain that will be the root node and then make the first decision tree before moving on to the next stage. The decision tree for node 1 looks like in Figure 3 below.

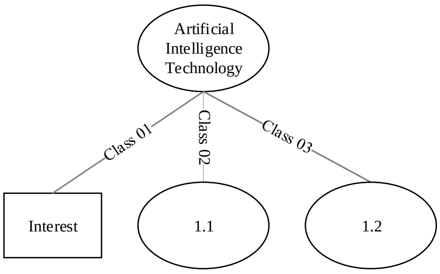


**Figure 3, Decision Tree Node 1**

In the decision tree above, it can be seen that the Artificial Intelligence Technology course is in demand by Class 01, so the course stops at that class. Meanwhile, in Class 02 and Class 03, it is still necessary to look for interest in other subjects other than the subjects that are the root of this research. Because at node 1 there is still no decision for class 02 and class 03, the research will continue where these classes will be the attributes of the next branch.

In the course column, the root node using the equations from data mining, namely true and false will also adjust the interest choice column and the uninterested choice column. The calculation method in the next node is the same as before, except that the number of cases decreases according to the root node that has been obtained. In the calculation of class 02 that will now be counted as node 1.1 and class 03 to become node 1.2, the filling of the research table will be adjusted according to the instructions previously described. The following are the results of the decisions obtained after calculating node 1.1 and node 1.2 that can be seen in the figure 4.

After the calculation of node 1.1, produces results where Basic Physics and Chemistry becomes the next branch attribute with Class 02 resulting in a decision that is Not Interested in the course, while Class 01 and Class 03 still have to look for values in other available courses. Then in the Node 1.2 section, the result is where Class 02 and Class 03 have an interest in Programming Algorithm courses and Class 01 will be the next branch because the results are still questionable.

Next, the research enters the calculation using branch attributes, namely Basic Physics and Chemistry for Class 01 and Class 03 of nodes 1.1.1 and node 1.1.2. The following decision tree is formed from the two nodes that have calculated the entropy and gain for the two nodes.

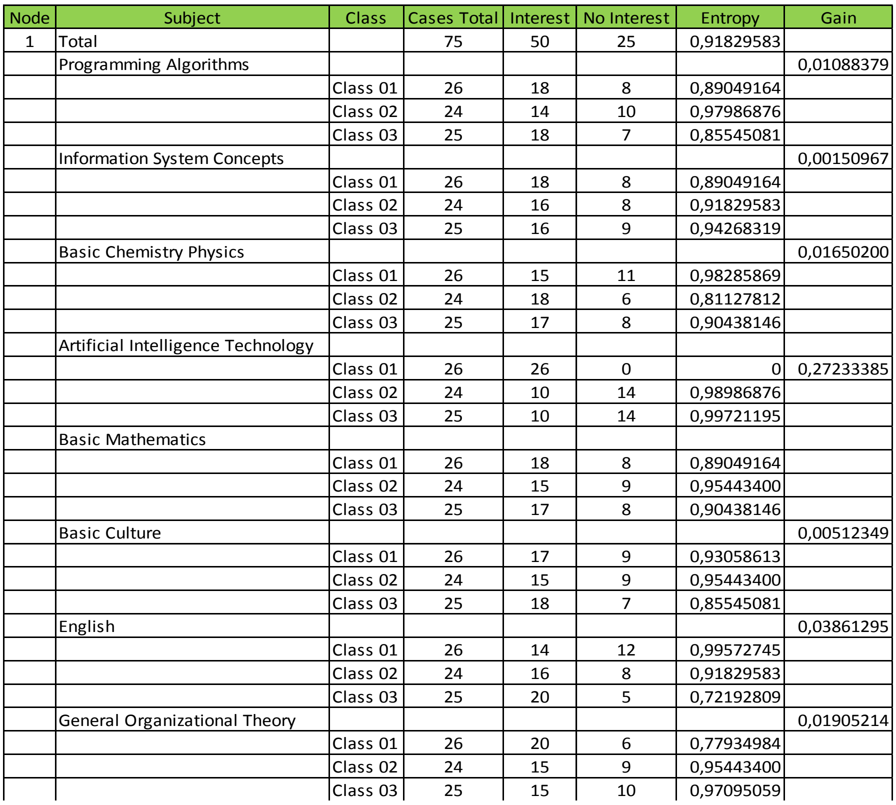
Figure 5 shows Node 1.1.1 produces an Information System Concept where the existing decisions already meet the three attribute values, such as Class 01 with the decision Not Interested, Class 02 with Interested and Class 03 also Interested. Because at the Information System Concept Node all attributes have been met, the decision tree at this node is complete. Node 1.1.2 produces General Organization Theory as a new branch because of the three attributes only Class 02 and Class 03 make decisions, while Class 01 must be searched for with the next node, namely 1.1.2.1.

In addition to Node 1.1.1 and Node 1.1.2 which have been searched previously, the next step is to calculate the value of 1.2.1 with the same entropy and gain formula. The search results for node 1.2.1 can be seen in the following decision tree (figure 6).

Node 1.2.1 produces English where the existing decisions already meet two attributes, namely Class 02 with the decision value Interested and Class 03 with the decision Not Interested. Attribute English - Class 01 must be searched again and a new node will be created, namely node 1.2.1.1. In the value search, it was found that the results for the three attributes in each node had met and the decision tree had reached its final form which can be seen in the image below. The search results for node 1.2.1.1 and node 1.1.2.1 can be seen in the following decision tree (figure 7).

The results obtained from the final decision tree above can be seen as follows:

1. The subject "Artificial Intelligence Technology" is considered interesting by Class 01
2. The subject of "Basic Chemistry Physics" is considered not interesting by Class 02
3. The subject "Programming Algorithms" is considered interesting by Class 02, but not by Class 03
4. The subject "Information System Concepts" was considered interesting by Class 02 and Class 03, but not by Class 01
5. The subject "General Organizational Theory" is considered interesting by Class 01 and Class 03, but not by Class 02
6. The "English" subject is considered interesting by Class 02, but not by Class 03
7. The subject of "Basic Culture" is considered interesting by Class 01, Class 02 and Class 03
8. The subject "Basic Mathematics" is considered interesting by Class 01, Class 02 and Class 03



| Node | Subject | Class | Cases Total | Interest | No Interest | Entropy | Gain |
|---|---|---|---|---|---|---|---|
| 1 | Total | | 75 | 50 | 25 | 0,91829583 | |
| | Programming Algorithms | | | | | | 0,01088379 |
| | | Class 01 | 26 | 18 | 8 | 0,89049164 | |
| | | Class 02 | 24 | 14 | 10 | 0,97986876 | |
| | | Class 03 | 25 | 18 | 7 | 0,85545081 | |
| | Information System Concepts | | | | | | 0,00150967 |
| | | Class 01 | 26 | 18 | 8 | 0,89049164 | |
| | | Class 02 | 24 | 16 | 8 | 0,91829583 | |
| | | Class 03 | 25 | 16 | 9 | 0,94268319 | |
| | Basic Chemistry Physics | | | | | | 0,01650200 |
| | | Class 01 | 26 | 15 | 11 | 0,98285869 | |
| | | Class 02 | 24 | 18 | 6 | 0,81127812 | |
| | | Class 03 | 25 | 17 | 8 | 0,90438146 | |
| | Artificial Intelligence Technology | | | | | | |
| | | Class 01 | 26 | 26 | 0 | 0 | 0,27233385 |
| | | Class 02 | 24 | 10 | 14 | 0,98986876 | |
| | | Class 03 | 25 | 10 | 14 | 0,99721195 | |
| | Basic Mathematics | | | | | | |
| | | Class 01 | 26 | 18 | 8 | 0,89049164 | |
| | | Class 02 | 24 | 15 | 9 | 0,95443400 | |
| | | Class 03 | 25 | 17 | 8 | 0,90438146 | |
| | Basic Culture | | | | | | 0,00512349 |
| | | Class 01 | 26 | 17 | 9 | 0,93058613 | |
| | | Class 02 | 24 | 15 | 9 | 0,95443400 | |
| | | Class 03 | 25 | 18 | 7 | 0,85545081 | |
| | English | | | | | | 0,03861295 |
| | | Class 01 | 26 | 14 | 12 | 0,99572745 | |
| | | Class 02 | 24 | 16 | 8 | 0,91829583 | |
| | | Class 03 | 25 | 20 | 5 | 0,72192809 | |
| | General Organizational Theory | | | | | | 0,01905214 |
| | | Class 01 | 26 | 20 | 6 | 0,77934984 | |
| | | Class 02 | 24 | 15 | 9 | 0,95443400 | |
| | | Class 03 | 25 | 15 | 10 | 0,97095059 | |

**Figure 2, Node Calculation Table**

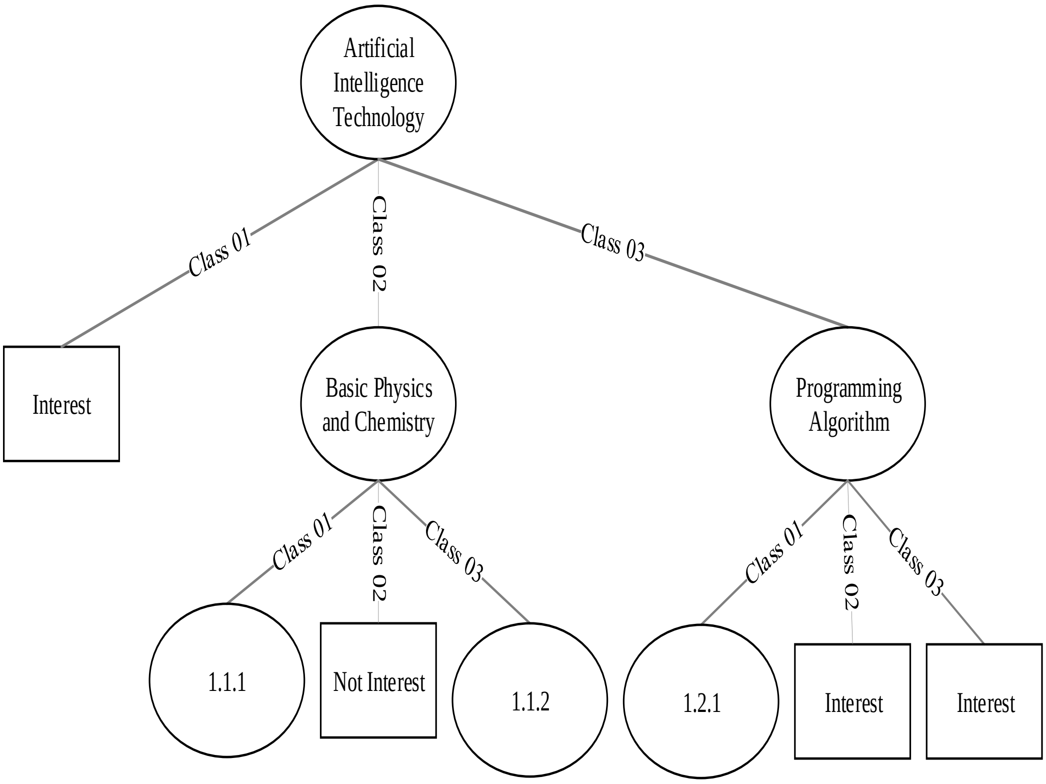


**Figure 4, Decision Tree Node 1.1 and Node 1.2**

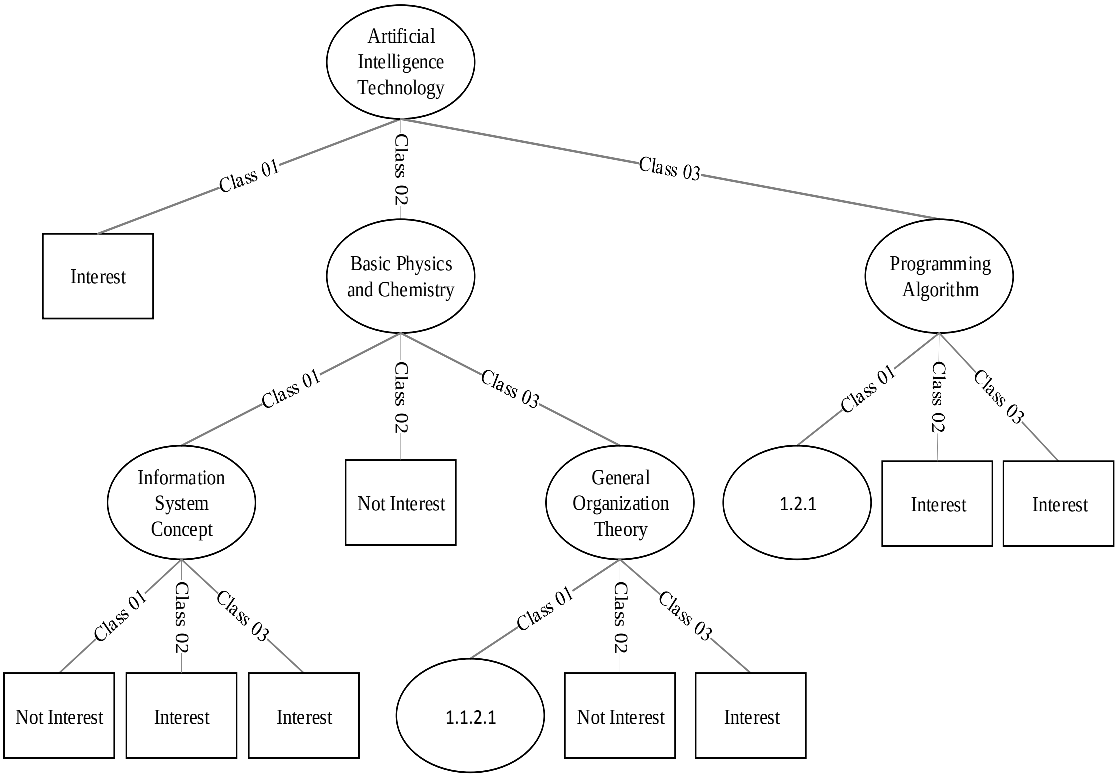


**Figure 5, Decision Tree Node 1.1.1 and Node 1.1.2**

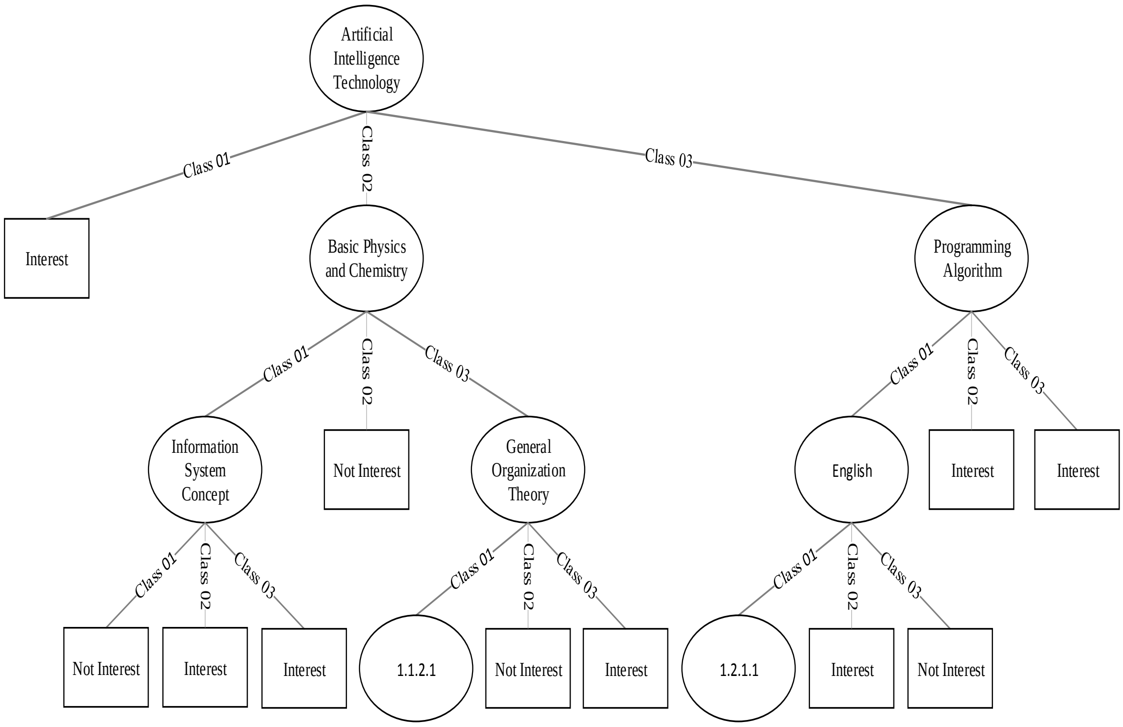


**Figure 6, Node Decision Tree 1.2.1**

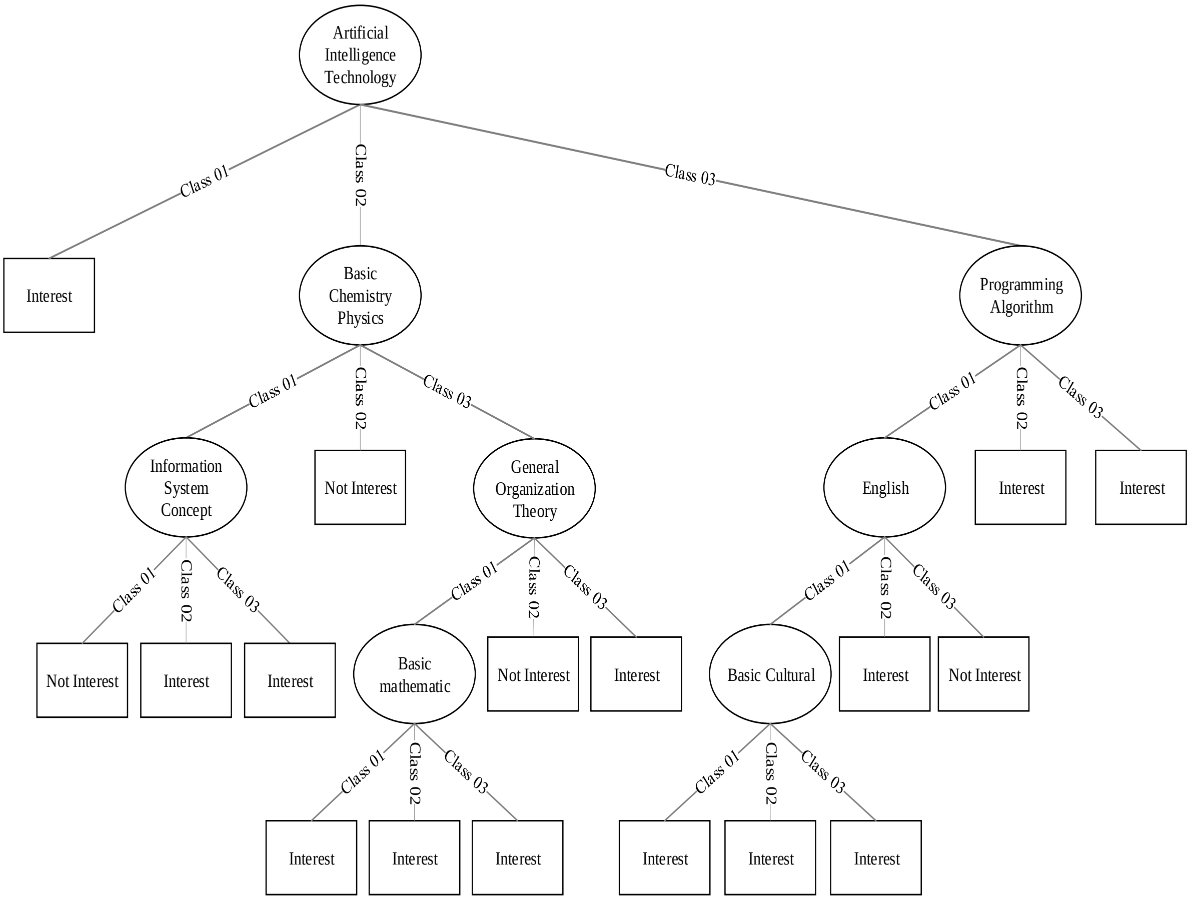


**Figure 7, Final Decision Tree**

## V. CONCLUSION

This research resulted in several conclusions

1. From the results of the discussion above, the university or institution can find out the pattern of value and interest in the courses taken by students at grade 2 in the department, later the University can provide appropriate direction for students in choosing the courses they are interested in either to increase their knowledge or to explore it.
2. Based on the results that have been obtained, it can be concluded that the university can review the less desirable courses so that lectures can run optimally.

## REFERENCES

[1]. Afuddin, R-N; andNurjannah, D.,2019, “SistemRekomendasiPemilihan Mata kuliah Peminatan Menggunakan Algoritma KmeansdanApriori (studikasus: Jurusan S1 TeknikInformatikaFakultas Informatika)” e-Proceeding of Engineering Vol.6, No.1 April 2019 - Page 2359

[2]. Meilani, B-D; andSlamat, A-F., 2013, “Klasifikasi Data KaryawanUntukMenentukanJadwalKerjaMenggunakanMetode Decision Tree”, Surabaya -InstitutTeknologiAdhi Tama Surabaya

[3]. Aldino, A-A; andSulistiani, H., 2020. “Decision Tree C4.5 Algorithm For Tuition Aid Grant Program Classification (Case Study: Department Of Information System, UniversitasTeknokrat Indonesia),” JurnalIlmiahEdutic, Vol. 7, No. 1, Pp. 40-50

[4]. Anestiviya, P; andPasaribu, A-F-O, 2021,“ANALISIS POLA MENGGUNAKAN METODE C4.5 UNTUK PEMINATAN JURUSAN SISWA BERDASARKAN KURIKULUM (STUDI KASUS : SMAN 1 NATAR)”,JurnalTeknologidanSistemInformasi (JTSI) Vol. 2, No. 1, Maret 2021, 80 - 85 E-ISSN: 2746-3699

[5]. Alifa, F; andUtami, A-W, 2017, “RANCANG BANGUN SISTEM PENDUKUNG KEPUTUSAN PEMINATANMATAKULIAHMENGGUNAKAN METODE WEIGHTED PRODUCT”, JurnalManajemenInformatika, Volume 08 Nomor 01 Tahun 2017, 20 - 28

[6]. Kusrini, 2009, Algoritma Data Mining, Yogyakarta