Prediction of national examination question using C4.5 algorithm

N Lukman1,*, B Subaeki1,2, H N Abdullah1, A R Atmadja1 and M Wildanuddin1

1 Department of Informatics, UIN Sunan Gunung Djati Bandung, Jalan A.H. Nasution 105 Bandung, Indonesia
2 Department of Information System, Sangga Buana YPKP University Bandung, Indonesia

*n.lukman@uinsgd.ac.id

Abstract. The National Examination in 2018 is no longer the only determinant of student graduation from their school, however the score of the national examination from 2015 has decreased. One of the reasons is that students found difficulties to predict the questions that probably appeared on the exam sheet. This research aimed to help students predict the questions that probably appeared, so that the average value hopefully will be increased. The prediction of the questions that probably appeared on the exam had been done by processing data of the questions used for national examinations several years ago using C4.5 Algorithm. C4.5 Algorithm is a Data Mining technique that can classify predictive questions and it is able to form a prediction of questions on the exam. The result showed that the prediction of questions that probably appeared on exams by using Algorithm C4.5 had good performance and its accuracy is 70%.

1. Introduction
The National Examination or commonly abbreviated as National Examination becomes one of the benchmarks for the graduation of compile students sitting in school with an absolute minimum value limit. With this National Examination, not a few participants failed to exceed the minimum score; It was recorded from 2015 the national exam scores decreased from an average value of 58.27% to 53.86% in 2016 [1]. This is caused by the level of ability of different students. We must realize that the difference in Indonesia consists of a lot of ethnic groups and cultures with varying patterns of life so that they can distinguish patterns of education and learn from the students themselves.

Technological development is very rapid, almost all lines all use technology so that technological developments can be used as a solution for various things [2,3], one of which is the use of technology in predicting or predicting something to solve a problem based on specific methods [4]. Whether or not the prediction of a problem is judged by the accuracy of the prediction with the actual results. Technology is a decision support system for the results of the adoption of human forecasting using artificial intelligence [5-8]. The purpose of developing this decision support is not to replace the role of humans permanently to help humans in this case students in predicting exam questions that will appear in the following year seen from the emergence of previous exam questions.

The emergence of previous exam questions is the basis of research using data mining techniques because the exam questions that appear on the exam in last years will be further processed using data
mining techniques which can then be used to obtain more accurate predictive results by decision support systems. Research in the field of decision systems has been done a lot [5], there are also other researchers Erwan Darmawan C4.5 Algorithm is implemented to predict the prospective new students in higher education can be done and can find out the number of candidates for new students who will register [4], other studies conducted by Erlin Elisa who implemented the C4.5 Algorithm to identify the factors causing work accidents and gave the conclusion that the C4.5 algorithm was more effective and flexible to be used for classification [9].

2. Methodology
The method that used in this study is C4.5 Algorithm for predict the emergence of national exam questions. C4.5 algorithm is a development of the Iterative Dichotomiser 3 (ID3) algorithm [9] which is a mathematical algorithm used to produce a decision tree that can classify an object [9-11]. ID3 represents concepts in the form of a decision tree [4,11]. The rules generated by ID3 have hierarchical relationships such as a tree (having roots, points, branches, and leaves). The illustration of the application of the decision tree in c4.5 algorithm is shown in Figure 1:

![Figure 1. C4.5 Algorithm flowchart.](image)

3. Result and discussion

3.1. National examination question analysis
The National Examination (UN) is an activity measuring the achievement of student competencies in certain subjects conducted in the world of education and adapted to national achievement standards. The National Examination or commonly abbreviated as National Examination becomes one of the benchmarks for the passing of students when they are in school with a specific minimum value limit [12]. With the existence of this National Examination, not a few students who fail to exceed the minimum value limit. This is caused by the level of ability of different students. We must realize that the difference in Indonesia consists of a lot of ethnic groups and cultures with varying patterns of life so that it can cause discrepancies in the patterns of education and learn from the students themselves.

3.2. The Implementation of C4.5 algorithm
The following is a description of the steps in the C4.5 algorithm for resolving cases of prediction of national exam questions based on the composition of the national exam questions for ten years.

| Content                  | Total Questions Per-Years |
|--------------------------|---------------------------|
|                          | 2008          | 2009          | 2010          | 2011          | 2012          | 2013          | 2014          | 2015          | 2016          | 2017          |
| Quadratic function       | 2             | 2             | 1             | 0             | 1             | 1             | 0             | 0             | 2             | 2             |
| Equation System          | 0             | 0             | 0             | 0             | 0             | 0             | 0             | 0             | 1             | 0             |
| Differential Equation    | 1             | 2             | 1             | 1             | 1             | 0             | 0             | 2             | 2             | 0             |
| Linear Program           | 1             | 1             | 0             | 1             | 1             | 0             | 2             | 0             | 1             | 0             |
| Line and raw             | 2             | 2             | 1             | 4             | 2             | 1             | 2             | 2             | 1             | 2             |
| Logic                    | 2             | 1             | 1             | 2             | 2             | 2             | 2             | 4             | 0             | 0             |
| Quadratic Equation       | 2             | 2             | 3             | 2             | 2             | 2             | 1             | 2             | 7             | 4             |
After obtaining question data for ten years, then we divide the questions based on the number of items, the level of difficulty of the problem and the consistency of the items, then Classify the item: Slight = less than 15 items, Much = 15 more questions;

**Table 2.** Case for determining national exam prediction questions.

| Content                              | Class | Question quantity | Category | Difficulty Level | Question Consistency |
|--------------------------------------|-------|-------------------|----------|------------------|----------------------|
| Quadratic function                   | 10    | 10                | Slight   | Difficult        | Permanent            |
| Equation System                      | 10    | 1                 | Slight   | Easy             | Vary                 |
| Differential Equation                | 11    | 9                 | Slight   | Difficult        | Permanent            |
| Linear Program                       | 12    | 7                 | Slight   | Easy             | Permanent            |
| Line and raw                         | 12    | 19                | Much     | Easy             | Vary                 |
| Logic                                | 10    | 19                | Much     | Difficult        | Vary                 |
| Quadratic Equation                   | 10    | 26                | Much     | Easy             | Vary                 |
| Three-dimensional space              | 10    | 38                | Much     | Easy             | Vary                 |
| Event probability                    | 11    | 19                | Much     | Difficult        | Vary                 |
| Integral                             | 12    | 41                | Much     | Difficult        | Vary                 |
| Logarithm                            | 10    | 17                | Much     | Easy             | Vary                 |
| Trigonometry                         | 11    | 33                | Much     | Easy             | Permanent            |
| Circle                               | 11    | 12                | Slight   | Easy             | Vary                 |
| Statistics                           | 10    | 24                | Much     | Easy             | Permanent            |
| Function, Composition, Invers       | 11    | 20                | Much     | Difficult        | Vary                 |
| Polynomials                          | 11    | 10                | Slight   | Difficult        | Permanent            |
| Limit Function                       | 11    | 19                | Much     | Easy             | Permanent            |
| Linear Equation System               | 12    | 12                | Slight   | Easy             | Vary                 |
| Matrix                               | 12    | 12                | Slight   | Easy             | Vary                 |
| Vector                               | 12    | 22                | Much     | Easy             | Vary                 |
| Geometry Transformation              | 12    | 12                | Slight   | Easy             | Permanent            |
| Exponent                             | 10    | 21                | Much     | Easy             | Vary                 |

Table 2 is used to calculate Entropy (Sten, Seleven, Stwelve) and Information Gain for the parameters of the Class Material Discussion.
3.3. Calculate Entropy Value for class material discussion parameters

\[
\text{Entropi}(S_{10}) = \left(- \left(\frac{2}{5}\right) \times \log_2 \left(\frac{2}{5}\right)\right) + \left(- \left(\frac{3}{5}\right) \times \log_2 \left(\frac{3}{5}\right)\right) = 0.95443456
\]
\[
\text{Entropi}(S_{11}) = \left(- \left(\frac{2}{7}\right) \times \log_2 \left(\frac{2}{7}\right)\right) + \left(- \left(\frac{5}{7}\right) \times \log_2 \left(\frac{5}{7}\right)\right) = 0.98417857
\]
\[
\text{Entropi}(S_{12}) = \left(- \left(\frac{2}{7}\right) \times \log_2 \left(\frac{2}{7}\right)\right) + \left(- \left(\frac{5}{7}\right) \times \log_2 \left(\frac{5}{7}\right)\right) = 0.86312056
\]

3.4. Calculating information gain for the number of questions:

\[
\text{Gain}(S,A) = 0.945660304 - \left(\frac{8}{22}\right) \times 0.95443456 + \left(\frac{7}{22}\right) \times 0.98417857 + \left(\frac{7}{22}\right) \times 0.86312056 = 0.04084278
\]

3.5. Calculating entropy and information gain for parameters number of predictions:

\[
\text{Entropi} (S_{\text{much}}) = \left(- \left(\frac{3}{13}\right) \times \log_2 \left(\frac{3}{13}\right)\right) + \left(- \left(\frac{10}{13}\right) \times \log_2 \left(\frac{10}{13}\right)\right) = 0.7793498374
\]
\[
\text{Entropi} (S_{\text{slight}}) = \left(- \left(\frac{4}{9}\right) \times \log_2 \left(\frac{4}{9}\right)\right) + \left(- \left(\frac{5}{9}\right) \times \log_2 \left(\frac{5}{9}\right)\right) = 0.9910760599
\]

3.6. Calculating information gain to discuss class materials:

\[
\text{Gain}(S,A) = 0.945660304 - \left(\frac{13}{22}\right) \times 0.7793498374 + \left(\frac{9}{22}\right) \times 0.9910760599 = 0.079695194487716
\]

3.7. Calculating entropy and information gain for parameters looking for number of prediction questions:

\[
\text{Entropi} (S_{\text{difficult}}) = \left(- \left(\frac{3}{7}\right) \times \log_2 \left(\frac{3}{7}\right)\right) + \left(- \left(\frac{4}{7}\right) \times \log_2 \left(\frac{4}{7}\right)\right) = 0.98417857
\]
\[
\text{Entropi} (S_{\text{easy}}) = \left(- \left(\frac{6}{15}\right) \times \log_2 \left(\frac{6}{15}\right)\right) + \left(- \left(\frac{9}{15}\right) \times \log_2 \left(\frac{9}{15}\right)\right) = 0.97095059
\]

3.8. Calculating information gain for difficulty problem:

\[
\text{Gain}(S,A) = 0.945660304 - \left(\frac{7}{22}\right) \times 0.98417857 + \left(\frac{15}{22}\right) \times 0.97095059 = 0.00052719
\]

Table 3 below is an analysis of each attribute and its values and calculates its entropy after getting the entropy from the whole case above.

| Node 1          | Attribute       | Number of cases | Permanent | Vary | Entropi  | Gain       |
|-----------------|-----------------|-----------------|-----------|------|----------|------------|
| Discuss Class   | Ten             | 8               | 5         | 3    | 0.95443456 | 0.04084278 |
| Material        | Eleven          | 7               | 3         | 4    | 0.98417857 |            |
|                 | Twelve          | 7               | 5         | 2    | 0.86312056 |            |
| Number of       | Many            | 13              | 3         | 10   | 0.7793498374 | 0.07969519 |
| questions       | a little        | 9               | 5         | 4    | 0.991076059 |            |
| Degree of        | Difficult       | 7               | 3         | 4    | 0.98417857 |            |
| difficulty      | Easy            | 15              | 6         | 9    | 0.97095059 | 0.00052719 |
The parameters of the number of questions are used as the root (root) at node 1 (root node) in the decision tree, the next node is analyzed further by separating data that has a large number of questions and the number of small questions so that a decision tree is formed with many numbers of parameters as follows:

**Figure 2.** Decision tree with parameters number of many questions.

From the results of the entropy calculation and gain produce the following rules or patterns:

- If (Number of Questions = Many) AND (Discuss Class Material = Twelve) then the consistency of the question = changes
- If (Problem Number = Lots) AND (Discuss Class Material = Eleven) then consistency of questions = change
- If (Number of Questions = Many) AND (Discuss Class Material = Eleven) then the consistency of the question = fixed
- If (Number of Questions = Many) AND (Discuss Class Material = Ten) AND (Level of Difficulty Problem = Easy) then consistency of questions = change
- If (Number of Questions = Many) AND (Discuss Class Material = Ten) AND (Level of Difficulty Problem = Difficult) then consistency of questions = fixed
- If (Number of Questions = Many) AND (Discuss Class Material = Ten) AND (Level of Difficulty Problem = Difficult) then consistency of questions = change
- If (Number of Questions = Few) AND (Level of Difficulty Problem = Difficult) then the consistency of the question = changes
- If (Number of Questions = Few) AND (Level of Difficulty Problem = Easy) AND (Discuss Class Material = Ten) then the consistency of the question = fixed
- If (Number of Questions = Few) AND (Level of Difficulty Problem = Easy) AND (Discuss Class Material = Eleven) then consistency of questions = change
- If (Number of Questions = Few) AND (Level of Difficulty Problem = Easy) AND (Discuss Class Material = Twelve) then consistency of questions = change

So as to produce a final decision tree and prediction of national exam questions:
Table 4. Prediction of national exam questions.

| Content                  | The appearance of questions Prediction | Content                  | The appearance of questions Prediction |
|--------------------------|----------------------------------------|--------------------------|----------------------------------------|
| Quadratic function       | 2                                      | Trigonometry             | 1                                      |
| Equation System          | 0                                      | Circle                   | 2                                      |
| Differential Equation    | 1                                      | Statistics               | 2                                      |
| Linear Program           | 1                                      | Function, Composition, Invers | 3                                      |
| Line and raw             | 4                                      | Polynomials              | 0                                      |
| Logic                    | 2                                      | Limit Function           | 2                                      |
| Quadratic Equation       | 2                                      | Linear Equation System   | 1                                      |
| Three-dimensional space  | 2                                      | Matriks                  | 1                                      |
| Event probability        | 2                                      | Vector                   | 2                                      |
| Integral                 | 4                                      | Geometry Transformation  | 0                                      |
| Logarithm                | 2                                      | Exponent                 | 4                                      |

4. Conclusion
The results showed that the accuracy of the predictions of the emergence of national exam questions carried out using the C4.5 Algorithm was successful and had a good performance with an accuracy of around 82% so that students could find out predictions about exam questions in the next few years.

References
[1] Pendidikan P P 2018 Perbandingan Hasil Ujian Nasional Antar Tahun [Online] Retrieved from: https://puspendik.kemdikbud.go.id/hasil-un/.
[2] Kurahman O T, Gerhana Y A, Syarifuddin U, Khosyi’ah S, Lukman N and Taufik I 2018 The Implementation of Breadth First Search in Determining of Waris IOP Conf. Ser. Mater. Sci. Eng. 434 1
[3] Munir 2009 Pembelajaran Jarak Jauh Berbasis Teknologi Informasi dan Komunikasi (Bandung, Indonesia: Alfabeta)
[4] Darmawan E 2018 C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education J. Online Inform. 3 1 22
[5] Septiana I, Irfan M, Atmadja A R and Subaeki B 2016 Sistem Pendukung Keputusan Penentu Dosen Penguji Dan Pembimbing Tugas Akhir Menggunakan Fuzzy Multiple Attribute
Decision Making dengan Simple Additive Weighting (Studi Kasus: Jurusan Teknik Informatika UIN SGD Bandung) J. Online Inform. 1 1 43

[6] Fauzan R, Indrasary Y and Muthia N 2018 Sistem Pendukung Keputusan Penerimaan Beasiswa Bidik Misi di POLIBAN dengan Metode SAW Berbasis Web J. Online Inform. 2 2 79

[7] Kamagi D H and Hansun S 2014 Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa J. Ultim. 6 1 15–20

[8] Uma J S S 2015 Human Interaction Pattern Mining Using Enhanced Artificial Bee Colony Algorithm S Int. J. Innov. Res. Comput. Commun. Eng. 3 10 10131–10138

[9] Elisa E 2017 Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT. Arupadhatu Adisesanti J. Online Inform. 2 1 p 36

[10] Rosminasih E, Uriawan W, Lukman N and Bandung J A H N 2017 Untuk Pengelompokan Surat Dalam Al-Qur’an Insight 1 1 32–37

[11] Rahmayuni I 2014 Perbandingan Performansi Algoritma C4.5 dan Cart Dalam Klasifikasi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang J. Teknol. 2 1

[12] Mukhlis H and Koentjoro 2016 Pelatihan Kebersyukuran untuk Menurunkan Kecemasan Menghadapi Ujian Nasional pada Siswa SMA Gadjah Mada J. Prof. Psychol. 1 3