# Decision Tree for Analyzing the Accuracy of Answer Sheet Data in Paper-based Test

Edy Suharto
Magister of Information System
Diponegoro University
Semarang, Indonesia
edys@undip.ac.id

Aris Puji Widodo
Magister of Information System
Diponegoro University
Semarang, Indonesia
arispw@undip.ac.id

Suryono
Magister of Information System
Diponegoro University
Semarang, Indonesia
suryono@fisika.undip.ac.id

***Abstract*—The accuracy of test data is important for education quality assurance. However, there is a problem due to the possibility of incorrect data filled by test taker during paper-based test. On the contrary, this problem is not found in computer-based test. A method was proposed in this study to analyze the accuracy of answer sheet filling out in paper-based test using data mining technique. A layer of data comprehension was added in the method instead of using raw data. The results of the study were a web-based program for data pre-processing and decision tree models. Among 374 instances which were analyzed, the accuracy of answer sheet filling out attained 95.19% and the accuracy of classification was 100%. This study could motivate the administrators for improvement of the test system since it preferred computer-based test to paper-based.**

***Keywords—data mining, decision tree, paper-based test, education.***

## I. INTRODUCTION

In higher education, quality assurance is essential for global competition [1]. Every educational system has at least three components, i.e. potential students as input, learning process, and graduated students as output. Admission test is crucial problem in every country [2]. Student only takes one test for entrance while one needs to take a number of tests during the learning process until graduation. These tests could be taken as indicators of learning process of a higher education institution [3].

Data mining implementation is essential for data analysis in an educational institution. Information taken from data mining is useful for improvement of educational method, such as learning process personalization, monitoring, and evaluation. Data mining is handful for learning approach modification [4]. It is also useful to explore students' characteristics which in turn leads to predict students' success [5].

As a natural mechanism of human to solve problem, decision tree is applicable to assist one to take a decision among a number of options. It has proved to be one of the strong and popular approaches to find pattern in data science [6]. Among their kinds, ID3 and C4.5 are the popular decision tree algorithms. Using information entropy and information gain, ID3 has three disadvantages, i.e. the tendency to choose attributes that have many values, the lack in antinoise handling, and the disability to handle missing values. As the revision of ID3, C4.5 algorithm is able to cope with continues attributes and missing values. It uses not only information gain ratio as separated criteria in order to get better partition but also pruning method for efficiency [7]. Furthermore, C4.5 is applicable to predict the fluctuative stock price [8].

Along with its capability, data mining should be advantageous for test data analysis. In this advanced information technology, there still exist paper-based tests while institutions tend to replace them with computer-based tests [9]. One problem faced in paper-based but not found in computer-based test is the data correctness. The examples of incorrect data content are blank value, different value compared to its reference, and unreadable content due to its blur pencil scratching. It is natural for human to do some mistake in filling out data in answer sheet. However, this incorrectness may lead to lengthy process to match one's sheet to its valid data content. This consolidation is important to make sure the scanned answer sheet is ready for computer-based test grading process [10]. Since data incorrectness did not happen in computer-based model, it led to a gap between both test models, although in a single test system.

This study explored the problem solving to measure and analyze the accuracy of answer sheet data in paper-based test. The proposed method to analyze the accuracy of answer sheet filling out was using data mining technique. In addition, this study also aimed to learn what components which denoted the most important in answer sheet data. However, the number of components were limited down to three, i.e. test taker's number, test taker's date of birth, and test set code. The results of this study should be taken into consideration by test administrators in order to improve the test model.

## II. LITERATURE REVIEW

### A. Computer-based versus Paper-based Test

For more than a decade, discussion about computer-based against paper-based test still continues. Study in [11] exposed comparative result between the two in English listening test. However, it was gender biased since female test takers tended to choose paper-based. Furthermore, study in [12] convinced that computer-based test results were more stable and consistent for a number of repetitions taken by the same test takers. Other study resulted that even the use of scratch papers in computer-based test were less than in paper-based [13]. Recent study observed that the respondents tended to choose computer-based test since they were more exposed to the advantage of information technology in daily life [14].

**ISBN 978-602-53524-0-9**

### B. Data Mining

Defined as an effort to explore uncovered knowledge in massive data using certain algorithms, data mining is a part of bigger knowledge discovery including pre-processing and post-processing tasks. Both data mining and knowledge discovery are iterative and interactive [15]. Data mining consists of six phases, i.e. business understanding, data understanding, data preparation, data modeling, model evaluation and model implementation. Business and data understanding are interactive processes to get insight the problem. Data preparation and modeling are also interactive to get the model fit the data. Evaluation is performed in order to ensure that the model fit the problem as in business. If it does not fit, the process of business understanding is repeated. On the contrary, the model could be implemented as it fits the business [16].

The application of data mining in education domain has been expanded. Based on analysis study of a number of articles, educational data mining (EDM) covered some themes such as orientation on learning interaction, learning evaluation, and educational media recommendation and recovery. The study presented perspective, trend identification, and potential research direction, such as behavior, collaboration, interaction, and performance in learning process interaction [4]. Another study was conducted in order to use data mining to predict the student achievement in learning process. The study focused on small number of education data instead of large number nature of data mining. Its promising result motivated the university to implement data mining as an important part in higher education knowledge management systems [5].

### C. Decision Tree

Techniques in data mining include classification, association, and clustering [16]. Classification is applicable to map an instance of information into a category which is defined based on certain attribute values of the instance [17]. A simple classification could be done with decision tree learning, especially where the target function is discrete. The function learned is represented as a decision tree or a set of if-then rules to enhance human readability. Decision tree classifies examples by sorting the tree from root node into possible leaf nodes. In every node is done a test for attributes of instance and every branch from the node based on a value of the attribute. Then focus continues to the branch based on the attribute value. This process is repeated for all sub-trees until it reaches leaf node. Each leaf node is the final decision taken from a set of rules i.e. a path of set of attribute values from the root node down to the terminal node [18].

As one of decision tree algorithm, C4.5 is applicable for classification. The first step is to build decision tree from training set, then expand the nodes so that the training data get fit and well defined. The second step is to perform tree conversion into a set of equivalent rules by tracing a rule for each path from root node down to leaf node. The third, prune the rules by deleting pre-conditions in order to improve accuracy. The fourth, sort the rules based on accuracy estimation when classifying sequential instances [7].

In order to build decision tree, it needs qualitative measurement in prioritizing which attributes to proceed subsequently as nodes. Information entropy $E$ of all $c$ partitions of sample $S$ is calculated as the negative sum of each information proportion $p_i$ multiplied by the 2-logarithm of the proportion itself (1) [19]. Entropy is used for calculating information gain $G$ of attribute $A$, which is defined as the entropy of sample subtracted by the negative sum of all proportions of certain $v$ valued sample $S_v$ multiplied by that certain valued entropy (2). Split information $SI$ is defined as the entropy of sample relative to a certain attribute (3). Then, the gain ratio $GR$ is defined as the ratio of information gain and the split information (4). The attribute choosing is repeated until all attributes were included in tree model, or the terminal node had the same target attribute value [18].

$$E(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

$$G(S,A) \equiv E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v) \quad (2)$$

$$SI(S,A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (3)$$

$$GR(S,A) \equiv \frac{G(S,A)}{SI(S,A)} \quad (4)$$

There are two kinds of classification test option commonly used. First, cross validation test uses two partitions of sample as training data and testing data. For example, 10-fold cross validation means the sample is divided into 10 partitions. Each partition is used as testing data for other nine partitions. The other option is percentage split. For example, 66% percentage split means there are 66% sample used as training data and the rest 34% used as testing data [16]. In evaluation, positive false and negative false should be considered [20]. Accuracy of classification is calculated by dividing the sum of positive true and positive false with total samples [17].

## III. Discussion

The discussion here were arranged using data mining steps as stated in section II.B. For efficiency, the first and the second phase were joined. The last phase, i.e. implementation was not applied here since this study was limited to analysis process. The information system framework for the analysis was shown in Fig. 1.

**ISBN 978-602-53524-0-9**

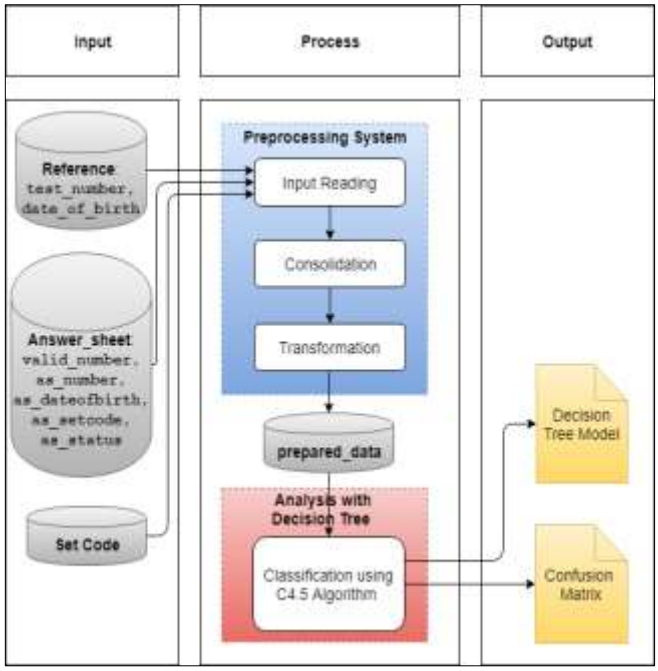


Figure 1. Information System Framework

### A. Problem and Data Understanding

In this first phase, it was resumed a basic process of answer sheet in paper-based test as follows. After collected, the answer sheets were then scanned in order to get a text file containing their associative data. Each answer sheet test data was then compared to its reference data by charaters. If it was matched then it was labeled correct. Otherwise, it was manually searched the test taker's actual number to put in as the correct test number. This searching process might be lengthy as the growing number of incorrect contents grew. The labeling process was handled by other system outside this study. That system supplied labeled data as a text file for analysis. However, this data format was not familiar for analysis tool. Therefore, in this study was also developed a pre-processing system in order to transform data supplied by that system into a format readable by common tool for analysis.

### B. Data Preparation

In this study was built a web-based program using PHP scripting language as pre-processing system. The program put a text file consisting of reference data, a text file consisting of labeled answer sheet data, and information of valid test set codes. The reference data consisted of test taker's number (*test_number*) and test taker's date of birth (*date_of_birth*). The labeled answer sheet data consisted of correct test number (*valid_number*), written test number (*as_number*), written date of birth (*as_dateofbirth*), written test set code (*as_setcode*), and correctness status (*as_status*). Of course, the term written here means what computer percept as filled out by a test taker on the answer sheet. Input would be read by the program, then they were consolidated to match answer sheet data with the reference. The program output was a comma-separated values (CSV) file.

Instead of using formatted labeled answer sheet data, in this study was added another upper layer to treat. Data were transformed into a format that represented the comprehension of each attribute value correctness. The program produced a CSV file consisting of four attributes of status, i.e. number status, date of birth status, test set code status, and answer sheet status. This was done in order to minimize variability of attribute values. Each attribute in answer sheet data was compared to its correct values in reference data as well as to the valid test set codes. Thus, the three kinds of incorrect values as stated in section I were treated as a same value, i.e. false. In the contrary, the correct values were treated as true. In other way, the possible values of each attribute were limited to two kinds, i.e. true or false.

### C. Modeling

The output of pre-processing system became the input for analytical tool. In this study was used Waikato Environment for Knowledge Analysis (WEKA). The data mining technique which was chosen was classification since this study attempted to map each instance of answer sheet into its status class (*as_status*) based on the attribute status values of test number (*no_status*), date of birth (*dob_status*), and test set code (*set_status).* The composition of attribute values was shown in Table I. The number of answer sheet instances was 374 records. Among the instances, 356 (95.19%) were labeled as true and 18 (4.81%) were labeled as false.

In order to produce decision tree model, the classification was employed using two options of test, i.e. 10-fold cross validation and 66% percentage split. The generate decision tree was shown in Fig. 2. In the first option, it was generated a decision tree consisting of five nodes, where three of them were leaf nodes. Based on the tree model, there were only two attributes which determined the answer sheet status. They were number status and test set code status. In the second option, it was also generated a tree model having five nodes, where three of them were leaf nodes. There were only number status and test set code status which appeared as determinative attribute for answer sheet status. Both tree models showed that the answer sheet status would be true if the values of number status and set codes were true, respectively. It could be inferred that the date of birth status did not give any contribution to the value of answer sheet status.

TABLE I. ATTRIBUTE VALUES COMPOSITION

| No | Attribute Name | Number of True | Percentage of True | Number of False | Percentage of False |
|---|---|---|---|---|---|
| 1 | no_status | 364 | 97.33% | 10 | 2.67% |
| 2 | dob_status | 367 | 98.13% | 7 | 1.87% |
| 3 | set_status | 365 | 97.59% | 9 | 2.41% |
| 4 | as_status | 356 | 95.19% | 18 | 4.81% |

**ISBN 978-602-53524-0-9**

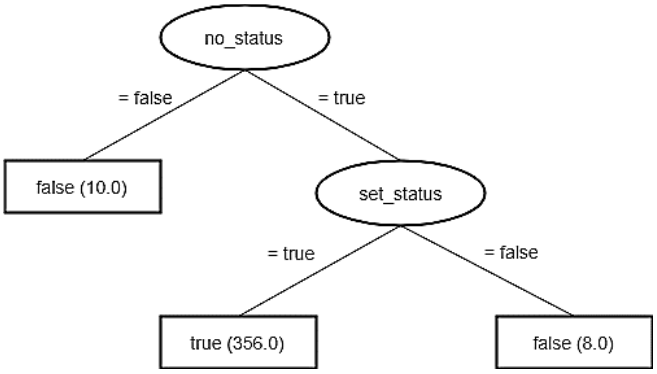


Figure 2. Generated Decision Tree

There was an interesting fact to expose, i.e. the difference between the total false value in answer sheet compared to those in Table I. Based on the tree model, there were 18 instances stated as false. Whereas, only two attributes which contributes to that number. But, if the number of false values in both number status and test set code then then it yielded 19. It could be inferred that there was one instance having incorrect data both in test number and test set code. After some investigation, it was found the instance. Although this instance had true value in date of birth attribute, the answer sheet status was still set to false. This convinced the statement that the status of date of birth did not determine the status of answer sheet.

Based on decision tree model and Table I, it was shown that the accuracy of filling out the answer sheet data attained 356 (95.19%) of 374. While it was known that the attributes which contributed to this number were number status and set code status. However, the number of true values in number status was 364 (97.33%) while the number of true values in set code was 365 (97.59%). It seemed there was about 2% deficit in accuracy measurement. Therefore, it should be done some subtraction between two opposite attributes.

First, the number of true values in number status (364 instances) was subtracted by the number of false values in set code status (8 instances), where this "8" value came from the false value in set code status (9 instances) minus an instance having false value in both number status and set code status. It yielded 356. Second, the true value in set code status (365 instances) was subtracted by the number of false values in number status (9 instances), where this "9" value was produced by the subtraction of false values in number status by an instance having false values in both number status and set code status. It also yielded 356. Hence, the accuracy calculation in Table I compared to the decision tree was convergent.

Another interesting fact to explore was why the decision tree took a such shape where the root node was number status instead of other attributes. In order to find out the answer, the gain ratio was calculated. The result was shown in Table II. The gain ratio of number status was the highest among the others as it reached 0.731666171. The number status attribute had two branches. When the value was false it came to leaf node which led to the false value of answer sheet status. Otherwise, it came to next candidate node.

TABLE II. TOTAL GAIN RATIO CALCULATION

| No | Attribute Name | Information Gain | Split Information | Gain Ratio |
|---|---|---|---|---|
| 1 | no_status | 0.130060505 | 0.177759353 | 0.731666171 |
| 2 | dob_status | 0.001821711 | 0.134172564 | 0.013577370 |
| 3 | set_status | 0.115560119 | 0.163688461 | 0.705975962 |

After number status was chosen as root node, the options of next candidate node were date of birth status and set code status. The gain ratio of true number status relative to other attributes was calculated as shown in Table III. The gain ratio of set code status reached 1.00. It meant that in this node, both values true or false led to final node, respectively. Hence, it did not need to increase some child nodes anymore. The decision tree which was built only had two attribute nodes, i.e. number status and set code status, and three leaf node, i.e. false number status, true set code status, and false set code status. It convinced that the generated decision tree in analysis process matched the calculation based on the theory.

TABLE III. TRUE NUMBER STATUS GAIN RATIO CALCULATION

| No | Attribute Name | Information Gain | Split Information | Gain Ratio |
|---|---|---|---|---|
| 1 | set_status | 0.152406999 | 0.152406999 | 1.000000000 |
| 2 | dob_status | 0.004473235 | 0.137099479 | 0.032627658 |

### *D. Evaluation*

The last step of data mining implemented in this study was evaluation. The quality of classification based on decision tree models was measured using confusion matrix. The accuracy was calculated as the ratio, i.e. the sum of correctly classified instances divided by total tested instances. The 10-folds cross validation matrix was shown in Table IV. There were 374 of 374 instances classified correctly. Hence, the incorrect classification of decision tree was 0%. On the other way, the 66% percentage split matrix was shown in Table V. There were 127 instances tested, representing 34% of total instances. Among the tested instances, there were 127 instances were correctly classified. This meant that both tests converged into same accuracy result, i.e. 100%.

TABLE IV. CONFUSION MATRIX OF CROSS VALIDATION

| Classified as | False | True |
|---|---|---|
| **False** | 18 | 0 |
| **True** | 0 | 356 |

TABLE V. CONFUSION MATRIX OF PERCENTAGE SPLIT

| Classified as | False | True |
|---|---|---|
| **False** | 9 | 0 |
| **True** | 0 | 118 |

**ISBN 978-602-53524-0-9**

## IV. Conclusion

Based on the descriptive analysis in section III, it came to three items of conclusion as follows:

*1) Data mining implementation* : data mining technique was applicable to solve the problem of data analysis in paper-based test. It resulted the accuracy measurement of the answer sheet filling out in this study as high as 95.19%. In other words, there were 4.81% test takers filled out the answer sheet incorrectly.

*2) The use of Algorithm* : C4.5 decision tree algorithm was proven to be suitable for the data in this study, as it produced tree models which were consistent in both cross validation and percentage split tests. In addition, it reached 100% accuracy of classification in both tests.

*3) The most important data* : there were two components which determined the status of answer sheet, i.e. test number and test set code. The content of date of birth did not contribute to the answer sheet status.

Lesson learned from this study led to suggestion that the test administrators should take it into their account in order to improve the test system. For example, the paper-based test could be more convincing using pre-printing answer sheet in order to minimize incorrect data filling out. Otherwise, the test administrators should also take into their consideration to substitute paper-based test with computer-based test. Other suggestion for the future work is the study done using big data or using the other comparable methods.

## References


[1] M. S. Kandil, A. E. Hassan, A. S. Asem, and M. E. Ibrahim, Prototype of Web2-based system for Quality Assurance Evaluation Process in Higher education Institutions, International Journal of Electrical & Computer Sciences IJECS-IJENS 10 (02), 2010.

[2] P. B. Ebrey, The Cambridge Illustrated History of China, Cambridge University Press, Cambridge, 2010.

[3] K. J. G. Leeuwenkamp, D. J. Brinke, and L. Kester, Assessment quality in tertiary education: An integrative literature review, Studies in Educational Evaluation 55, 2017, pp. 94-116.

[4] M. W. Rodrigues, L. E. Zárate, and S. Isotani, Educational Data Mining: A review ofevaluation process in the e-learning, Telematics and Informatics, 2018.

[5] S. Natek and M. Zwilling, Student data mining solution–knowledge management system related to higher education institutions, Expert Systems with Applications 41 (14), 2014, pp. 6400-6407.

[6] X. Guan, J. Liang, Y. Qian, and J. Pang, A multi-view OVA model based on decision tree for multi-classification tasks, Knowledge-Based Systems 138, 2017, pp. 208-219.

[7] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, 1993.

[8] Q. A. Al-Radaideh, A. A. Assaf, and E. Alnagi, Predicting Stock Prices Using Data mining Techniques, The International Arab Conference on Information Technology, 2013.

[9] H. R. Kim, M. Bowles, X. Yan, and S. J. Chung, Examining the comparability between paper- and computer-based versions of an integrated writing placement test, Assessing Writing 36, 2018, pp. 49-62.

[10] N. A. Karim and Z. Shukur, Proposed features of an online examination interface design and its optimal values, Computers in Human Behavior 64, 2016, pp. 414-422.

[11] D. Coniam, Evaluating computer-based and paper-based versions of an English-language listening test. Cambridge University Press 18 (2), 2006, pp. 193-211.

[12] C. Y. Piaw, Replacing paper-based testing with computer-based testing in assessment: Are we doing wrong?, Procedia - Social and Behavioral Sciences 64, November 9, 2012, pp. 655-664.

[13] A. A. Prisacari and J. Danielson, Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use, Computers in Human Behavior 77, 2017, pp. 1-10.

[14] S. Chan, S. Bax, and C. Weir, Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test, Assessing Writing 36, 2018, pp. 32-48.

[15] M. J. Zaki and W. Meira Jr., Data mining and analysis: fundamental concepts and algorithms, Cambridge University Press, New York, 2014.

[16] I. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data mining: practical machine learning tools and techniques Fourth Edition, Morgan Kaufmann, Cambridge, 2017.

[17] J. Han, M. Kamber, and J. Pei, Data mining: Concepts and Techniques, 3rd edition., Morgan Kaufmann, Amsterdam, 2011.

[18] T. M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[19] C. E. Shannon, Prediction and entropy of printed English. The Bell System Technical Journal 30, 1951, pp. 50–64.

[20] A. B. Downey, Think Stats: Probability and Statistics for Programmers, Green Tea Press, Needham, 2011.


**ISBN 978-602-53524-0-9**