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

Edy Suharto
Diponegoro University
Semarang, Indonesia
edys@undip.ac.id

Aris Puji Widodo
Diponegoro University
Semarang, Indonesia
arispw@undip.ac.id

Suryono
Diponegoro University
Semarang, Indonesia
suryono@fisika.undip.ac.id

***Abstract*—In education quality assurance, the accuracy of test data is crucial. However, there is still a problem regarding to the possibility of incorrect data filled by test taker during paper-based test. On the contrary, this problem does not appear in computer-based test. In this study, a method was proposed in order to analyze the accuracy of answer sheet filling out in paper-based test using data mining technique. A single layer of data comprehension was added within the method instead of raw data. The results of the study were a web-based program for data pre-processing and decision tree models. There were 374 instances which were analyzed. The accuracy of answer sheet filling out attained 95.19% while the accuracy of classification varied from 99.47% to 100% depend on evaluation method chosen. This study could motivate the administrators for test improvement since it preferred computer-based test to paper-based.**

Keywords-data mining; decision tree; paper-based test; education

I. Introduction

Quality assurance is essential for global competition in higher education [1]. An educational system has at least three components, i.e. potential students as its input, learning process, and graduated students as its output. Admission test especially in higher education is crucial problem in every country [2]. A student only needs to pass one test for entrance while he or she needs to take a number of tests during the learning process until graduation. These tests could be considered as indicators of learning process of a higher education institution [3].

Implementing data mining is essential for data analysis especially in an educational institution. Information derived from data mining is useful for improvement of educational method including learning process personalization, monitoring and evaluation. Data mining is also handful for learning approach modification effort [4]. It is also useful to explore students' characteristics which in turn it could be used to predict students' success [5].

Being a natural mechanism of human to solve problem, decision tree is applicable to assist someone to take a decision among a number of options. It was proven to be one of the strong and popular approaches to find pattern in data science [6]. ID3 and C4.5 are the popular decision tree algorithms among their kind. ID3 uses information entropy and information gain. It has three disadvantages, i.e. the tendency to choose attributes that have many values, the lack in antinoise handling and the disability to cope with missing values. Recognised as the revision of ID3, C4.5 algorithm is able to handle continues attributes and missing values. It employs 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 also applicable to predict the fluctuative stock price [8].

With its capability, data mining should be advantageous for test data analysis. In this very advanced information technology, there still exist some paper-based tests while some institutions tend to replace them with computer-based tests [9]. One problem appears in paper-based but being absent in computer-based test is the data correctness. Some examples of incorrect data content are blank value, unsimilar value compared to its reference, and unreadable content for its blur pencil scratching. Although it is natural for human to perform some mistake in filling out personal data in answer sheet, this incorrectness may lead to lengthy process in order to match one's sheet to its valid data content. Thus, consolidation is important to make sure the scanned answer sheet is ready for computer-based test grading process [10]. However, the process of matching test taker's answers with the correct answer is out of scope of this study. Since data incorrectness did not appear in computer-based model, it led to be a gap between both test models, although in a single test system.

This study explored the problem solving effort to measure and analyze the accuracy of answer sheet data in a paper-based test. The analysis was performed by measuring personal data which were manually filled out by the test taker in the answer sheet compared to its corresponding data taken from server database. The proposed method for analyzing the accuracy of answer sheet filling out was using data mining technique. Furthermore, this study also aimed to learn what components which represent the most important in answer sheet data. However, the number of components included here were limited down to three, i.e. test taker's number, test taker's date of birth and test set code. The output was decision tree model representing the most significant components of answer sheet data.

II. Literature review

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

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

B. *Data Mining*

Data mining is defined as an effort to explore uncovered knowledge in massive data using certain algorithms. It is a part of bigger knowledge discovery embrassing pre-processing and post-processing tasks. Both data mining and knowledge discovery are iterative and interactive processes [15]. There are six phases in data mining, i.e. business understanding, data understanding, data preparation, data modeling, model evaluation and model implementation. Both business and data understanding are interactive processes to get insight the problem. Data preparation and modeling are also interactive in order to ensure the model fit the data. Evaluation is performed to make sure that the model fit the problem as in business. Unless it fit, the process of business understanding is repeated. On the contrary, the model could be implemented if it fits the business [16].

Data mining application in education domain has been expanded. Based on analysis study of a number of articles, educational data mining (EDM) included some themes such as orientation on learning interaction, learning evaluation, and also 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 employ data mining to predict the achievement in student learning process. That study only focused on small number of education data, instead of large number as a nature of data mining. Its result was promising and could motivate the university to implement data mining as an essential part in higher education knowledge management systems [5].

C. Decision Tree

Techniques in data mining consists of classification, association, and clustering [16]. Classification technique is applicable to map an instance of information into a category defined based on certain attribute values of the instance [17]. A simple classification could be performed using decision tree learning, especially if the target function is discrete. The function learned is represented as a decision tree model or a set of if-then rules in order to enhance human readability. Decision tree classifies examples by sorting the tree from root node into possible leaf nodes. In every node, a test is done for attributes of instance and in every branch from the node based on the value of the attribute. The 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 set as the final decision taken from a set of rules, i.e. a path of set of attribute values from the root node pass down to the terminal node [18].

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

For building a 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 step of choosing the attribute is repeated until all attributes were included in tree model, or the terminal node had the same target attribute value [18].

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

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

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

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

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]. Both options were used in [20]. There is another test option, being supplied data test. This kind of classification test uses the other data content outside the training data in order to apply and evaluate the model generated using prior internal data set. Regarding evaluation, the classification result including positive false and negative false should be considered [21]. The accuracy of classification is calculated by dividing the sum of positive true and positive false with total samples in percentage [17].

III. Discussion

The discussion presented here were arranged using data mining steps as stated in section II.B. The first and the second phase were joined for efficiency. The last phase, i.e. implementation was applied here using supplied data test. The information system framework for the analysis was shown in Fig. 1.

A. Problem and Data Understanding

In this first phase, a basic process of answer sheet in paper-based test was defined as follows. After being collected, the answer sheets were then scanned to get a text file containing their associative data. The example part of that file was shown in Fig. 2. Each line corresponded to one answer sheet. The example contained some missing scanned values in line 5 and line 10 as explained in section I. The first five characters represents test number, followed by date of birth, a reserved character, then three characters of test set code.

Each answer sheet test data component was then compared to its reference data by characters. It then be labeled. If it was matched, then it was labeled correct. Otherwise, it was manually searched the test taker's actual number to be written down as the correct test number. This searching process might be lengthy as the growing number of incorrect contents grew. However, the labeling process was handled by the other system outside this study. That system supplied labeled data as a text file for analysis. Unfortunately, this data format in the file 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 which is readable by common tool for analysis.

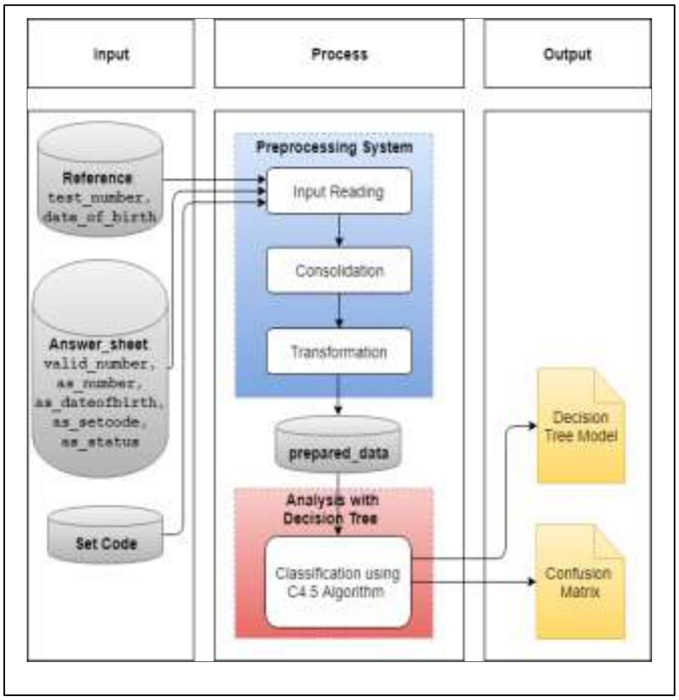
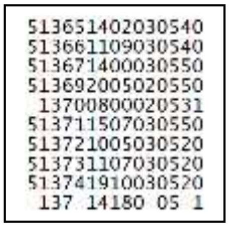


Figure 1. Information System Framework

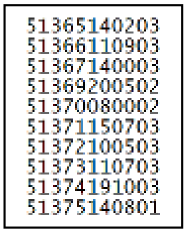


```
513651402030540
513661109030540
513671400030550
513692005020550
 13700800020531
513711507030550
513721005030520
513731107030520
513741910030520
 137 14180 05 1
```

Figure 2. Text data produced by scanning answer sheet

B. Data Preparation

In this study, a web-based program was built using PHP scripting language as pre-processing system. The program put a text file which consisted of reference data, a text file consisting of labeled answer sheet data, and information of valid test set codes. On one side, the reference data consisted of test taker's number (*test_number*) and test taker's date of birth (*date_of_birth*) as shown in Fig. 3.



```
51365140203
51366110903
51367140003
51369200502
51370080002
51371150703
51372100503
51373110703
51374191003
51375140801
```

Figure 3. Reference data

On the other side, the labeled answer sheet data consisted of correct test number (*valid_number*), scanned test number (*as_number*), scanned date of birth (*as_dateofbirth*), scanned test set code (*as_setcode*), and correctness status (*as_status*). Of course, the term "scanned" 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 data. The program output was a comma-separated values (CSV) file.

In this study, another upper layer to treat was added instead of using formatted labeled answer sheet data. Data were then transformed into a certain 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 then 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, Waikato Environment for Knowledge Analysis (WEKA) was used. The data mining technique which was chosen was classification because this study attempted to map each instance of answer sheet into its answer sheet 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. Based on correctness status attribute, there were two classes being produced in decision tree model, i.e. TRUE and FALSE.

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%

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 generated 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 represented the 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.

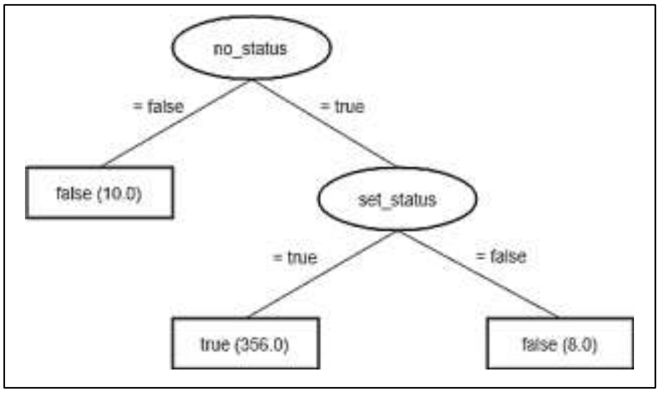


Figure 4. Generated Decision Tree

There was an interesting fact to discuss here, 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 labeled as false. Whereas, only two attributes which contributes to that number. In fact, if the number of false values in both number status and test set code status were added, 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 a certain investigation, the instance was found. Although this instance was labeled 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 both 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 the 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 next 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, being the sum of correctly classified instances divided by total tested instances. There were three options chosen for evaluation purpose, i.e. 10-fold cross validation, 66% percentage split and supplied data test.

The result of 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

Unfortunately, the result became different when a supplied data test was chosen as evaluation option. For this purpose, as many as 378 records were prepared as data test. These data were taken from other test held in another occasion. Among the tested instances, there were 376 instances correctly classified both as false and true classes. The accuracy was calculated as the sum of positive true (366) and positive false (10) divided by total samples (376). It resulted 99.47% accuracy of classification, which was less than prior test option results. There are 2 incorrectly classified instances, being 0.53% inaccuracy of classification. This small percentage was still acceptable.

TABLE VI. CONFUSION MATRIX OF SUPPLIED DATA TEST

Classified as	**False**	**True**
False	10	2
True	0	366

However, the decision tree model generated using supplied data test was matched to the one generated using data set. The tree model for supplied data could be seen in Fig. 3. Although there were some differences in number of instances classified, the tree branches and also leaves were still as same as the tree form shown in Fig. 2. The tree model consisted of five nodes, where the test number being the root node. The next branching node was test set code.

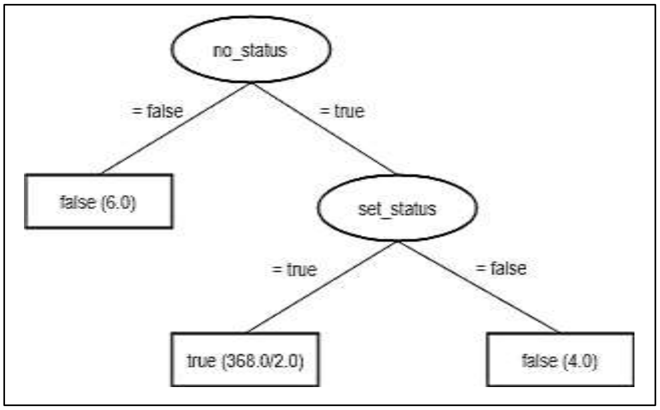


Figure 5. Decision Tree using Supplied Data Test

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 properly 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. Although the decision tree generated in supplied data test was as same as the tree generated by prior tests, the accuracy of classification decreased down to 99.47%.

3) The most important data. There were only two components which determined the status of an 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 some recommendation which the test administrators should take into consideration for test system improvement. For example, the paper-based test could be more convincing when it uses pre-printed answer sheet. This aims to reduce incorrect data which is written by the test taker. Otherwise, the test administrators should also take into their account to substitute paper-based test with computer-based test. The 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,” *Int. J. Electr. Comput. Sci. IJECS-IJENS*, vol. 10, no. 02, 2010.

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

[3] K. J. G. Leeuwenkamp, D. J. Brinke, and L. Kester, “Assessment quality in tertiary education: An integrative literature review,” *Stud. Educ. Eval.*, vol. 55, pp. 94–116, 2017.

[4] M. W. Rodrigues, S. Isotani, and L. E. Zárate, “Educational Data Mining: A review of evaluation process in the e-learning,” *Telemat. Informatics*, vol. 35, no. 6, pp. 1701–1717, 2018.

[5] S. Natek and M. Zwilling, “Student data mining solution–knowledge management system related to higher education institutions,” *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6400–6407, 2014.

[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 Syst.*, vol. 138, pp. 208–219, 2017.

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

[8] Q. A. Al-Radaideh, A. A. Assaf, and E. Alnagi, “Predicting Stock Prices Using Data mining Techniques,” *Int. Arab Conf. Inf. Technol.*, 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,” *Assess. Writ.*, vol. 36, pp. 49–62, 2018.

[10] N. A. Karim and Z. Shukur, "Proposed features of an online examination interface design and its optimal values," *Comput. Human Behav.*, vol. 64, pp. 414–422, 2016.

[11] D. Coniam, "Evaluating computer-based and paper-based versions of an English-language listening test," *Cambridge Univ. Press*, vol. 18, no. 2, pp. 193–211, 2006.

[12] C. Y. Piaw, "Replacing Paper-based Testing with Computer-based Testing in Assessment: Are we Doing Wrong?," *Procedia - Soc. Behav. Sci.*, vol. 64, pp. 655–664, 2012.

[13] A. A. Prisacari and J. Danielson, "Computer-based versus paper-based testing: Investigating testing mode with cognitive load and scratch paper use," *Comput. Human Behav.*, vol. 77, pp. 1–10, 2017.

[14] S. Chan, S. Bax, and C. Weir, "Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test," *Assess. Writ.*, vol. 36, pp. 32–48, 2018.

[15] M. J. Zaki and J. Wagner Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York: Cambridge University Press, 2014.

[16] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann Publishers, Inc., 2016.

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

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

[19] C. E. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, no. 1, pp. 50–64, 1951.

[20] E. Suharto, A. P. Widodo, and Suryono, "Decision Tree for Analyzing the Accuracy of Answer Sheet Data in Paper-based Test," *Proceeding 2nd Int. Conf. Informatics Dev.*, 2018.

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