Improving The Accuracy of Student Problem Identification Using Rule-Based Machine Learning

Budi Sulistiyo[1*], *Bayu* Surarso[2], *Wahyul* Amien Syafei[3]

[1] Graduate School of Information System Diponegoro University Semarang, Indonesia
[2] Faculty of Science and Mathematics Diponegoro University Semarang, Indonesia
[3] Department of Electrical Engineering Diponegoro University Semarang, Indonesia

Abstract. Adolescence are a period of development that is vulnerable to problems and often makes teens unable to control emotions. No exception for adolescents who are studying high school. Problems that do not need to be resolved immediately and bigger problems will arise later on. Many methods of solving students' problems are carried out in a conventional manner which takes time and costly. Therefore, teacher guidance and career guidance at school use the problem checklist method to identify student problems. One thing that promises to improve accuracy with time to identify problems by building information systems using intelligent technology such as machine learning. Machine learning offers sophisticated techniques built by automatic classification that can be utilized by students and teachers to improve accuracy and efficiency in identification. This article discusses issues related to problems faced by senior high school students and proposes a knowledge-based users (rules) machine learning to match the problems and alternative solutions. This system can be used by school counsellors to help students solving their problems and the students to access themselves without having to meet the school counsellor. The results of this research indicate that information system developed based on rule-based machine learning offer a student problem identification which is more accurate, faster, can be done anytime and anywhere, and requires less cost compared to existing conventional methods. Analysis of machine learning with rule-based models using WEKA gives 100% accuracy.

1 INTRODUCTION

The problems experienced by humans will be more complex as we get older, especially at the age of teenagers who are still in high school education. Problematic behaviour at school can be defined as behaviour that can interfere with student learning activities in class [1]. Problems experienced by students need to be addressed as they relate to achievement [2]. To prevent greater problems arising in the future, efforts are needed to detect students' problems from the start. Benefits in understanding the involvement of behaviour problems among

* Corresponding author: budisulistiyo@students.undip.ac.id

youth and informing the design of interventions to reduce behaviour problems [3]. Educators must be able to understand the behaviour and academic problems of students so that they can help the problems they face and prevent difficulties in the future [4].

The method used to identify students' problems is done in a conventional manner which is sometimes still subjective, must meet with the teacher's guidance, conducted in class, requires a relatively takes time and costly. Data collection techniques such as experiments [5], interviews [6], surveys [7], observations, and questionnaires [8]. The questionnaire is a list of questions in the form of problems that have been faced or are being experienced by students. The questionnaire uses behavioural indicators that can be measured. The items are given in the form of questionnaires and then filtered and classified. The classification process relies on rules made by humans with the appraisal function adding the scores of related items in the questionnaire in calculating the results. Therefore the quality of the results of the classification is determined on: (1) items used in this technique (2) Knowledge and user experience used in the classification process (3) rules in the valuation function [9]. The use of questionnaires to identify student problems is done in a conventional way that requires a relatively takes time and costly.

In designing rules for calculating the scores of questionnaire items requires knowledge and experience. Make rules with cases that have previously happened to remind identification results and also a better classification process. The results of the rules are automatic and not subjective like the results of human knowledge because they use knowledge such as machine learning thereby increasing the estimated time, reducing costs, and accuracy in identification. Machine learning is an algorithm that can learn from data and make predictions without being explicitly programmed to do something [10]. Classification in machine learning is divided into 3 categories [11], supervised learning where all data is labelled, unsupervised learning data that is not labelled [12], Reinforcement Learning wherein the agents learning how to act by being given a gift or being given a punishment, reward or punishment able to add data or reduce points [13].

Many studies have used machine learning to increase the classification time used to detect and diagnose, such as machine learning based on the induction of autism models [9], machine learning used to detect damage under operational and environmental variability [14], machine learning used to detect and evaluate heating loads on buildings [15], machine learning to detect false opinions on comments [16], machine learning implemented to detect the impact of turbine blades on wildlife and the environment [17], detect traffic jams automatically through sensors wireless traffic using the machine learning approach [18]. In the process of diagnosis or identification includes predictions whether included or not, in determining the class (classification) can use supervised learning. In this case making training data sets to create models using machine learning techniques. The model is then used to predict new cases (not yet classified) as accurately as possible.

This article proposes a method of classifying student problems based on an effective rule induction approach to help make decisions that classify them as "if-then" [18]. The induction rule classification method in machine learning rules is included in supervised learning. In this method, the system offers an automatic classification that is represented in the rules. The rules that are made in the classification can be used by the school counsellors to take further action in helping students solve their problems and in other side the system can be used by the students to access themselves, identify problems, and find an alternative solution without having to meet the school counsellor.

The results of this study indicate that the information system developed which is based on rule-based machine learning offers a classification that is more accurate, faster, can be done anytime, anywhere, and requires no cost compared to the existing conventional methods.

2 METHOD

2.1 Dataset

This research is based on data from one of the senior high schools in Karanganyar District. This school uses a sheet containing list of problems as a way to identify the problems of senior high school students. The problem checklist used has 12 categories consisting of health, economic situation, family life, religion and morals, recreation and hobbies, personal relationships, social life - organizational activities, youth issues, school adjustments, curriculum adjustments, study habits, future, and educational / service life. Each category has a number of attributes that have two labels. There are two types of labels in the question namely "0" which indicates the user is not experiencing problems, and "1" students are having problems. This attribute is used to determine the class of categories obtained through equation (1).

$$\sum_{j=1}^{k} Zj = Z_1+Z_2+Z_3\ldots+Z_k \quad (1)$$

The results of equation (1) will be correlated with the class of the problem shown in Table 1.

Tabel 1. Problem class

Scale (%)	Value	Description
0	A	Very Good
1 - 10	B	Good
11 - 25	C	Fair
26 – 50	D	Poor
51 - 100	E	Very Poor

2.2 Covering

One classification method in machine learning includes (rule-based). This method is widely studied to make a set of classification rules from examples[19]. The Covering technique is one way to directly classify extracted from the if-then rules. The method proposed in identifying students' problems is based on the classification of cover with the search method to find rules. The resulting rule includes some tuple data from the class, so one rule is used to present the class.

The step starts with processing the raw data, then the learning algorithm is applied to find which represents the variables in the training dataset with class variables (Very Good, Good, Fair, Poor, Very Poor). Data sets will only store rules that classify training instances. The final result is a classification system used to predict new classes of data that have not yet been classified as shown in Figure 1.

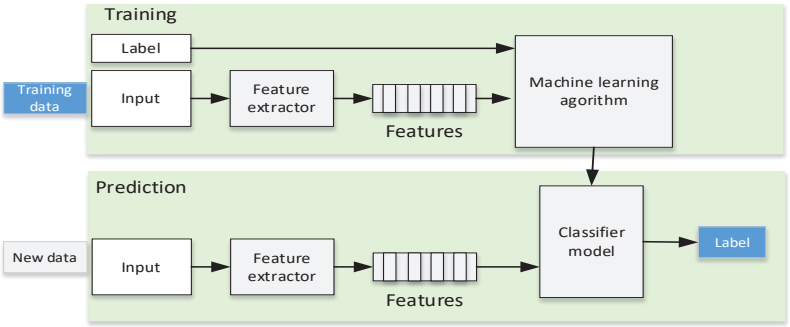

Fig. 1. Using a set training data to find the best rule (doing a good labelling task / predicting) results on new data (unseen data).

Sequential covering algorithms in the rule induction follow the following steps[19] :

1. Start with an empty rule.
2. Start to develop rules (R) for each class (C).
3. Remove tuples covered by rule (R).
4. Repeat steps 2 - 3 until the criteria for attributes are exhausted, to develop rules.

3 RESULTS

This research, the data preparation stage by analysing the data requirements needed. Data needed from the Problems Check List (PCL) includes data indicating problems, problem categories, suggestions as solutions to problems from counselling guidance teachers. The next step is to enter problem data, categories and recommendations into the system. After the data is entered into the system, the next step is the learning process in machine learning. The learning process in machine learning by labelling an indication of a problem to produce a set of rules for the problem category. The results of the rules will be saved as a training set or dataset to be modelled.

Figure 1. shows the new data input feature that will use the dataset as a model to be classified into problem categories. Machine learning will extract rules that connect target attributes with features. When there is new data that does not have a label using the classifier it will become a known label or class.

To test the machine learning with a rule-based model, Waikato Environment for Knowledge Analysis (WEKA) application software is used. WEKA is a platform for machine learning applications [21]. WEKA has many tools for data processing, i.e. ranging from pre-processing, classification, grouping, association, and visualization. This research uses 384 data instances in the PCL test consisting of 25 attributes as shown in Figure 2.

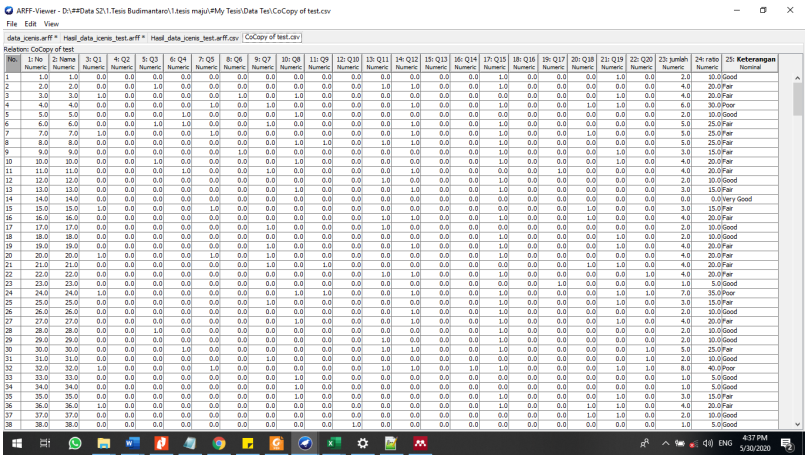

Fig. 2. Example of data on the PCL test.

Figure 2 shows the data used in testing the rule-based machine learning using WEKA. After opening the WEKA explorer and select the data file, then enter the classify menu, select the rule for the classifier type. In this case the training and test data are set to 66% as shown in Figure 3.

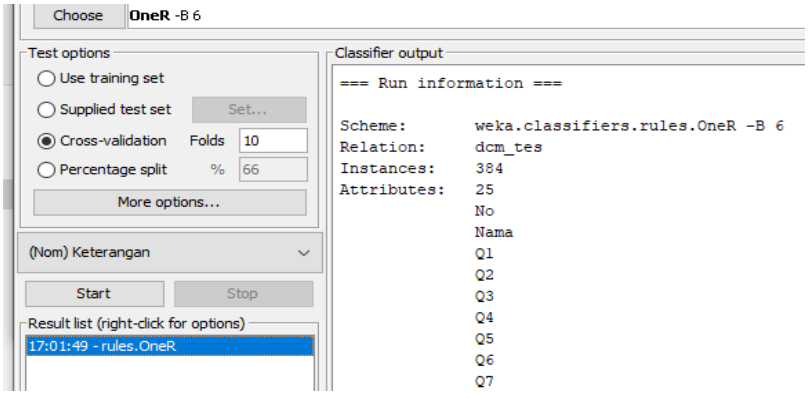

Fig. 3. Collection of training and split tests

Results of training and tests using WEKA obtained rules that were successfully extracted with a level of truth in the classification of 100%, which is indicated by red circle in Figure 4.

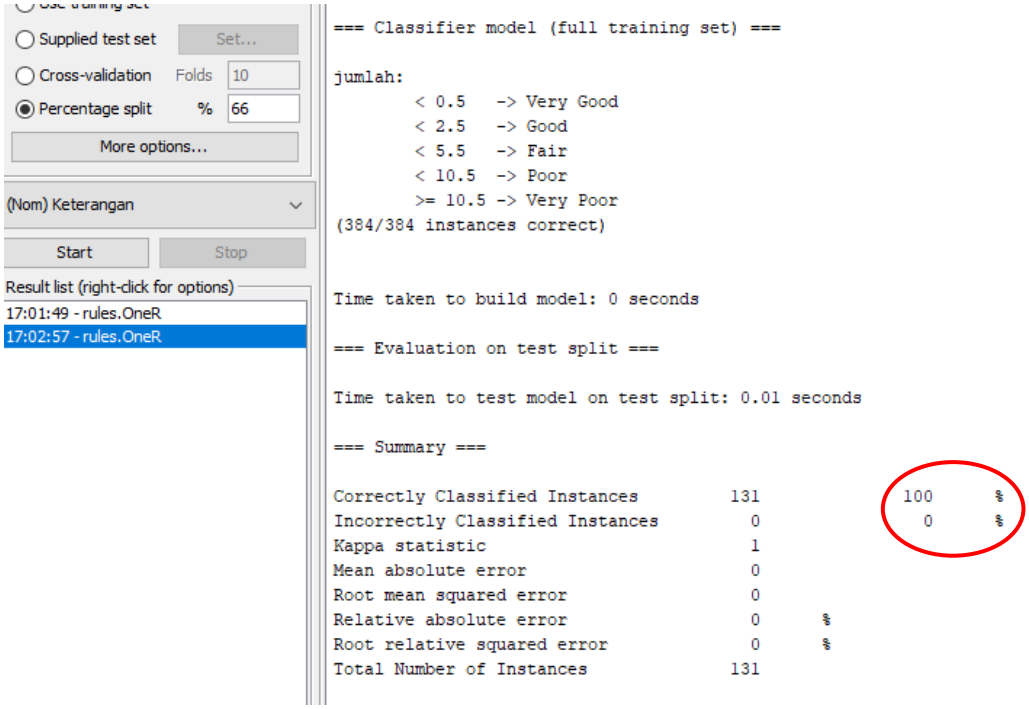

Fig. 4. Extraction of rules and classification accuracy obtain 100%.

Figure 4. Shows 131 total tested instances and the correct classification results then obtain the level of accuracy with the formula:

$$Accuracy = \frac{p}{t} \quad (2)$$

Where t is the number of all instances and p is the number of instances included in the rule and is positive, then the accuracy is obtained as100%.

4 CONCLUSION

Adolescence is a period of growth and development that experiences increasingly complex problems, especially when they are at the age of high school education where they begin to get to know the environment and the wider community. If a problem that is being faced by a student is not resolved immediately it will cause an ongoing problem, and will also affect student achievement. In understanding and identifying problems experienced by participants is the first step in handling further. The methods used in understanding students such as questionnaires, PCL and others. The assessment is done using the calculated answer score. During classification and assessment still use conventional methods which are sometimes still subjective. So that one of the important problems in research identifying students' problems is to improve the classification process so that the services in guidance and counselling are more accurate and faster. To achieve this, a rule machine-based information system is used that builds an accurate classification of data and previous cases. Test results show that the proposed a rule-based machine learning information system gives accurate and fast in identifying senior high school student problems. This system can be used by the student to access themselves and make it easy for school counsellors to help students solving their problems.

Acknowledgement

This work was supported by DRPM, Deputy for Strengthening Research and Development. Ministry of Research, Technology/ National Agency for Research and Innovation of the Republic of Indonesia through the LPPM Diponegoro University under the Magister Thesis Research Schema Year 2020 (Grant Number 225-55/UN7.6.1/PP /2020).

References

[1] K. Johnson and M. D. Hannon, "Measuring the Relationship Between Parent, Teacher, and Student Problem Behavior Reports and Academic Achievement: Implications for School Counselors," *Prof. Sch. Couns.*, vol. **18**, no. 1, p. 2156759X0001800, Sep. (2014).

[2] J. D. McLeod, R. Uemura, and S. Rohrman, "Adolescent Mental Health, Behavior Problems, and Academic Achievement," *J. Health Soc. Behav.*, vol. **53**, no. 4, pp. 482–497, Dec. (2012).

[3] O. E. El-Shenawy and A.-M. Shehata, "Applying Problem Behavior Theory in a Developing Arabic Country," *SAGE Open*, vol. **4**, no. 1, p. 215824401452181, Jan. (2014).

[4] A. K. Sanford and R. H. Horner, "Effects of Matching Instruction Difficulty to Reading Level for Students With Escape-Maintained Problem Behavior," *J. Posit. Behav. Interv.*, vol. **15**, no. 2, pp. 79–89, Apr. (2013).

[5] K. J. Mullinix, T. J. Leeper, J. N. Druckman, and J. Freese, "The generalizability of survey experiments," (2016).

[6] H. Alshenqeeti, "Interviewing as a Data Collection Method: A Critical Review," *English Linguist. Res.*, vol. **3**, no. 1, Mar. (2014).

[7] I. Krumpal, "Determinants of social desirability bias in sensitive surveys: a literature review," *Qual. Quant.*, vol. **47**, no. 4, pp. 2025–2047, Jun. (2013).

[8] B. Dash, "Methods of Data Collection," in *Essentials of Nursing Research and Biostatistics*, vol. **14**, no. 2, Jaypee Brothers Medical Publishers (P) Ltd., 2017, pp. 175–175.

[9] F. Thabtah and D. Peebles, "A new machine learning model based on induction of rules for autism detection," *Health Informatics J.*, vol. **I–23**, p. 146045821882471, Jan. (2019).

[10] W. Bleidorn and C. J. Hopwood, "Using Machine Learning to Advance Personality Assessment and Theory," *Personal. Soc. Psychol. Rev.*, vol. **23**, no. 2, pp. 190–203, May (2019).

[11] A. Awaysheh, J. Wilcke, F. Elvinger, L. Rees, W. Fan, and K. L. Zimmerman, "Review of Medical Decision Support and Machine-Learning Methods," *Vet. Pathol.*, vol. **56**, no. 4, pp. 512–525, Jul. (2019).

[12] S. Klassen, J. Weed, and D. Evans, "Semi-supervised machine learning approaches for predicting the chronology of archaeological sites: A case study of temples from medieval angkor, Cambodia," *PLoS One*, vol. **13**, no. 11, pp. 1–17, (2018).

[13] N. R. Ravishankar and M. V. Vijayakumar, "Reinforcement Learning Algorithms: Survey and Classification," *Indian J. Sci. Technol.*, vol. **10**, no. 1, pp. 1–8, (2017).

[14] E. Figueiredo, G. Park, C. R. Farrar, K. Worden, and J. Figueiras, "Machine learning algorithms for damage detection under operational and environmental variability," *Struct. Heal. Monit. An Int. J.*, vol. **10**, no. 6, pp. 559–572, Nov. (2011).

[15] K. M. H. Swhli, S. Jovic, N. Arsic, and P. Spalevic, "Detection and evaluation of heating load of building by machine learning," *Sens. Rev.*, vol. **38**, no. 1, pp. 99–101, Jan. (2018).

[16] Y. W. Oh and C. H. Park, "Machine Cleaning of Online Opinion Spam: Developing a Machine-Learning Algorithm for Detecting Deceptive Comments," *Am. Behav. Sci.*, p. 000276421987823, Oct. (2019).
[17] C. Hu and R. Albertani, "Machine learning applied to wind turbine blades impact detection," *Wind Eng.*, p. 0309524X1984985, May (2019).
[18] I. Qabajeh, F. Thabtah, and F. Chiclana, "A dynamic rule-induction method for classification in data mining," *J. Manag. Anal.*, vol. **2**, no. 3, pp. 233–253, Jul. (2015).
[19] M. Berry and M. Browne, *Lecture notes in data mining*. (2006).
[20] J. Brownlee, *Machine Learning Mastery With Weka*, V1.1.(2016).