An Implementation of C4.5 Classification Algorithm to Analyze Student's Performance

2020 3rd International Conference on Information and Communications Technology (ICOIACT) | 978-1-7281-7356-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICOIACT50329.2020.9332088

Latifaestrelita Indi Pramesti Aji
Magister of Informatics Engineering
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
latifaestrelita.aji@students.amikom.ac.id

Andi Sunyoto
Magister of Informatics Engineering
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
andi@amikom.ac.id

Abstract*—Due to the massive amount of information in educational database, predicting student's performance is more difficult. Thus, a comprehensive literature review to forecast student's performance utilizing data mining techniques—in particular using the C4.5 algorithm—is recommended to enhance the student's achievements. The main intention of this research is to provide a description of the data mining techniques used to predict student's results. This research paper also reflects as to how the classification algorithm may be used to classify the most significant characteristics in a student's information. In essence, we might increase student's performance and progress more effectively by utilizing educational data mining methods, particularly the C4.5 classification algorithm. Based on the results obtained by the dataset provided, the C4.5 algorithm had the accuracy of 71,9%. This could offer advantages and influence to students, lecturers and academic establishments.

Keywords*—performance prediction, student performance, C4.5 classification algorithm;

I. Introduction

In web-based academic environments, the capacity to forecast or identify the success of a student is essential. The usage of data mining is a very interesting way to accomplish this purpose [1]. In addition, classification is one of the most useful activities for e-learning. Different classification academic purposes occur, such as: separating students that are hinted-driven or failed-driven and discovering certain biases that pupils have [2], recognizing low-motivated learners and taking remedial steps to reduce drop-out levels [3], forecasting or classifying students through insightful tutoring programs [1], etc. Plus, there are various kinds of classification methods and artificial smart algorithms in order to project a student's performance, marks, or scores.

Academic performance of students is a key element in developing their future [4, 5]. The central focus of published literature works is on the performance of teaching, quality of teaching, and learning for students. However, other influences such as study patterns, school attendance, social events, the history of students' families and others may affect student's success. Understanding the effect of these factors will help enhance the performance of students in a subject as early as possible. Applying data mining in educational systems or Educational Data Mining (EDM) has been widely introduced to enhance the quality of student's performance. Early assessment and review of at-risk student recognition in classroom education can be useful for students and educators respectively [6]. Lecturers may have ample flexibility to conduct instructional activities to enhance student success [7]. The exploration of information on the implementation of machine learning in educational environments can be beneficial for teachers to use it to improve student efficiency, whilst the analysis of guidelines on the implementation of machine learning in the same area may be helpful for students to use them to boost their success in their learned topics. In addition, the usage of instructional data mining tools in the education sector may be useful to define instructional knowledge like student reports, learning habits, behaviors and progress in the classroom [8].

The functions of data mining can be explanatory or predictive. Explanatory data mining utilizes strategies like correlation rule mining, clustering, etc. to analyze the data embedded in large data sets and assist in smart decision solving. Predictive data mining builds models with rule collection, decision tree, neural nets, support vectors, etc. to determine the class of the current data set [9]. The key purpose of this paper is to evaluate the student's performance by using a certain data mining technique in the classification method, which is C4.5 algorithm. C4.5 is an improved variant of the standard algorithm of decision tree, which is ID3. Therefore, this paper propose the following method: C4.5 algorithm to predict and discover the statistical factors that are highly important on the academic outcomes of a student.

This research work discusses student causes whose academic success is not successful and increases school efficiency by recognizing slow learners in particular subjects (math, writing and reading scores) so that teachers can support them boost their output on their own. The precision of certain classification method for predicting student's performance is also explored in this paper. In general, the paper is categorized into five segments. The second segment discusses a variety of literature reviews or related works. The third segment explains the process of implementation using the C4.5 algorithm. The fourth segment discusses all the analysis and results. The conclusion is provided in the fifth and final segment of the paper. Classification is a simple process of discovering a prototype (or function) that recognize the salient features of data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown, and it forecast distinct and unordered labels in huge data sets. As with classification, the data set is used to build a predictor but an independent test set should be used to access its accuracy [10].

II. Related Works

This literary works report research studies examines the efficiency of classifying algorithms dependent on student datasets. The operating cycle of each algorithm is evaluated with the precision of the classification algorithms. It also explores the specific data mining methods used in the analysis of students performance academically.

Research works by Ryan Baker [11], Romero and Ventura [9, 12], which suggest success predictions as one of the new fields of Educational Data Mining, have been most frequently referred to in literature survey materials in Educational Data Mining. Pedro G. Espejo and César Hervás, alongside with Romero and Ventura [10] also stated that a comparative analysis for student classification based on student’s performance in some data sets across multiple data mining techniques may suggest that a successful classification model must be both reliable and comprehensible to educators.

On the other hand, Amirah Mohamed Shahiri, Wahidah Husain, and Nur’aini Abdul Rashid [13] presented their research by reviewing student’s performance prediction with five different data mining methods. Decision Tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine are among other algorithms that are used. An overview of the outcomes of recent research in the estimation of student’s performance reveals that the Artificial Neural Network has the best prediction precision of 98 percent among other four approaches.

In prior research, Abdeighani and Urthan conducted an overview of student’s performance using the student’s survival indicators utilizing data mining techniques. The researches tested three data mining techniques which are Naive Bayes, Back Propagated Neural Network, and C4.5 algorithms and noticed that the reliability of C4.5 algorithm is significantly better than the other two algorithms [14].

Nguyen [15] conducted an overview of the successful estimation of the academic success of undergraduate and postgraduate students at two seperate and distinct academic institutions: Can Tho University (CTU), a university in Vietnam, and the Asian Institute of Technology (AIT), a postgraduate college in Thailand. The researches also used various techniques of data mining to identify the Bayesian Networks and Decision Tree classification accuracy. The most effective predictive precision they have obtained is used to determined student outcomes. The findings of this research are very helpful in the quest for the right candidates to earn college scholarships. The outcome of this study shows that Decision Tree was reliably 3 to 12 percent more reliable than the Bayesian Network. The same can be inferred with the theoritical work carried out by Al Radaideh [20] using the Decision Tree model to estimate the fnal grade of students who studied C++ at Yarmouk University, Jordan, in 2005. Three specific classification techniques are used, including C4.5, Naive Bayes, with the introduction of ID3. The findings showed the Decision Tree model had a higher prediction than other simulations.

III. Prepare Your Paper Before Styling

Throughout this chapter, we explain the technique which was used to predict or to forecast specialized factors on student's performance. The prediction of students performance has two key variables, which are attributes and the technique to predict. The purpose of this article is to detect specialized parts by using the C4.5 algorithm and to help the institution optimize its efficiency and improve its reliability and availability.

The method suggested in this paper to improve prediction of students’ academic performance belongs to the process of Data Mining. There are four main stages in this method [16], which are data collection, preprocessing, classification, and interpretation (Figure 1). Data collection is gathering all information available on students considering factors affect the students’ performance.

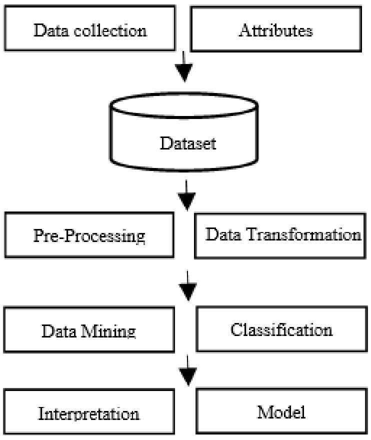

Fig 1. Method proposed for improving the prediction of student’s academic performance

This information can be collected from different sources of data available and combined into the dataset. In pre-processing stage, data cleaning, attributes selection, dimensionality reduction, and data partitioning are applied to get better prediction. Whereas, in classification stage, a data mining algorithm is executed with different variables and produce best results. Finally, in interpretation stage models obtained from previous stage are analyzed to predict students’ performance.

Pre-processing data is a necessary step for preparing the dataset before applying classification techniques. It is important to note that this task directly affect the result due to the quality and reliability of the available information. In this task, careful analysis of variable and their corresponding values is carried out to eliminate any abnormalities. We thoroughly analyze our dataset to identify attributes which have greater impact on our output variable. Although, we do not have large number of attributes in the dataset, still some features are not related to students’ performance. As stated above, at the end, a data mining technique which are C4.5 algorithms is applied on the transformed data and then obtain the expected result.

A. Dataset

The research presented in this paper is carried out based on a dataset in the Kaggle website, namely Student Performance which was carried out Asia Ahmed Abu Shawish. In order to predict student’s performance, the systematic measurement distinguishes essential attributes. The attributes commonly used are the Cumulative Grade Point Average or CGPA, and internal evaluation. The core reason of why most researchers are utilizing CGPA is that it has a real benefit for potential learning and job versatility. It may also be used as an indicator of the intellectual ability realization [16]. In the literary works courtesy of Christian and Ayub [17], Cumulative Grade Point Average is the most important element when deciding the performance of students throughout their studies, whether or not they will fulfill their studies [18]. In this report, the internal evaluation was categorized as an accomplished scholarship, quiz performances, and presence. All attributes would be combines under a single attribute called an internal evaluation.

Furthermore, demographic and external evaluations of students are the most frequently used for attributes. The demographics of students are including gender, age, and backgrounds of the family. Whereas external evaluations are known from the score that was obtained from measured quizzes. The explanation that most literature works use quantitative data, including gender in their learning processes is because they have specific types of female and male data [17]. A research by Borges in 2007 showed that most female students had different types and attitudes than male students [15]. Women are more organized, conscience-directed, often maintained and centered in their studies. On the other side, women have a strong learning approach [19]. They are self-motivated, coordinated and implemented by them consistently. Therefore, it has been seen that gender is one of the significant qualities that affect student's performance [17].

In another review, there are also some researchers who have predicted student's performance by using the neuropsychological element [17]. The student's participation, actions, commitment of time and familial support is described as a neuropsychological variable. These attributes have been used to make a framework appear very simple, transparent, and easy to use. It lets the instructor assess the performance of the students on the scale of their personal involvement and attitudes [20]. Though, these attributes are rarely used to determined student's performance by a number of studies because they focus more on statistical data and are also difficult to obtain valid information from participants in the study [17].

B. *C4.5 Algorithm*

Entropy is used to calculate the variability of a set of data samples. The entropy value can be estimated statistically with the following formula.

$$Entropy(S) = \sum_{i}^{c} - p_i \, log2p_i \qquad (1)$$

C = number of values in target attribute (number of classes)

Pi = number of samples in class i

Statistically, information gain of an attribute can be estimated as follows.

$$Entropy(S) - \sum_{Vevaluation(A)} \frac{|Sv|}{s} \, entropy(Sv) \qquad (2)$$

A: attribute

V: potential value for attribute A

Values (A): a set of potential values for attribute A

|Sv|: number of samples for value v

|S|: number of all data samples

Entropy (S): entropy for samples with value v

Gain ratio is estimates with the following formula.

$$gain\ ratio = \frac{Gain\ (S,A)}{split\ information\ (S,A)} \qquad (3)$$

$$split\ information = -\sum_{I=1}^{c} \frac{si}{s} log2 \frac{si}{s} \qquad (4)$$

S1 until Sc is c subset generated from separating S using attribute A with a variety of C values.

Pruning is a method for simplifying C4.5 algorithm tree layout. It is used to estimate the amount of rates and the scale of the forest in the tree system in order to establish a simplified tree model and to preserve class distribution [21]. A few aspects may be achieved to optimize a tree model, which involved of the addition of all instance values in any rule in a coherent distribution table and estimation of the degree of independence between parameters in a given regulation (i.e. between attribute and attribute target) and removing insufficient factors (which are conditions with high independence levels).

The program data may be analyzed by measuring the efficiency of the program depending on the combination of training data and test data. For this paper, the calculation in the classification which was carried out was accuracy [16]. Therefore, the calculation can be presented in the following table.

TABLE I. THE CALCULATION CLASSIFICATION

	defined as ineligible	defined as eligible
ineligible	a	b
eligible	c	d

Accuracy is the proportion of cumulative data collected which is properly defined. The formula is as follows.

$$accuracy = \frac{(a+d)}{(a+b+c+d)} \text{x } 100\% \qquad (5)$$

IV. DISCUSSION AND RESULTS

Identifying the student's academic performance requires a variety of factors to be examined. Data analysis that incorporate all detailed information such as gender, race or ethnicity and parental level of education then the social aspect, which if a student charges a regular rate for lunch or just for free. Information relating to student's literary comprehension of a specific topic in mathematics, their test preparation courses also their scores in writing and reading will also contribute in predicting their results.

A. *Data Preparations*

The data set used in this study was obtained from the site Kaggle.com, entitled "Student Performance" and prepared by Asia Ahmed Abu Shawish, which was published on Thursday, July 2nd 2020. For this phase, all data is obtained and arranged in the Microsoft Excel 2016 spreadsheet with 1000 records [22].

B. *Data Selection and Conversion*

Throughout this stage, certain fields that were needed for predictive analytics were selected. The data are gathered by way of a dataset that was previously provided. Students submit their personal details like gender, race/ethnicity, parental level of education, lunch rate, test preparation course, math score, reading score, and writing score. The building of the algorithm is derived on splitting nodes, which is absolutely essential during the design phase and largely determines the final structure of the decision tree.

TABLE II. VARIABLES, DESCRIPTION, AND VALUES

Variables	Description	Values
Gender	Student's Gender	{Male, Female}
Race/ethnicity	Student's race or ethnicity	{Group A, B, C, D}
Parental level of education	Parents' level of education	{Some High School, High School, Some College, Bachelor's Degree, Master's Degree, Associate's Degree}
Lunch	Lunch's rate card for the students	{Standard, Free/reduced}
Test Preparation Course	Student's preparation on examinations	{None, Completed}
Math Score	Student's score for the math subject	{A, B, C, D}
Reading Score	Student's score for the reading subject	{A, B, C, D}
Writing Score	Student's score for the writing subject	{A, B, C, D}

The efficiency of the machine learning method can be determined by measuring the precision of the algorithm estimation. Analyzing the validity and consistency of the data mining algorithm can be shown by the confusion matrix, the confusion matrix displayed the right and incorrect prediction details in both positive and negative contexts. Some meaning terms in confusion matrix used in this paper are accuracy (proportion or percentages of correct predictions) and precision (proportion or percentages of positive context predicted correctly) [11].

C. *The implementation of C4.5 Algorithm*

As stated before in the formula 1, we have to count the entropy first, and explore the entropy of each node. Below here is the sample of the estimation in entropy total.

$$ET = \left(-\frac{490}{1000} * log_2\left(\frac{490}{1000}\right)\right) + \left(-\frac{510}{1000} * log_2\left(\frac{510}{1000}\right)\right) = 0{,}9997$$

Then, we estimate the information gain for the algorithm as stated in formula 2. Below here is the sample of the estimation in information gain of the variable (gender).

$$IG = 0.9997 - \left(\frac{482}{1000} * 0.9634\right) - \left(\frac{518}{1000} * 0.6832\right) = 0{,}1814$$

Lastly, we count the gain ratio for the algorithm as stated in formula 3. Below here is the sample of the estimation in gain ratio of the variable (gender).

$$Gain\ Ratio = \frac{1.6701}{1.5738} = 1{,}0611$$

V. CONCLUSION

Based on the results obtained by the dataset provided, the C4.5 algorithm had the accuracy of 71,9% and an additional variable of 'status' that consisted Pass or Fail. Classification has been an intriguing topic for researches as it categorizes information extraction data effectively and precisely. Data mining techniques are becoming so widely known because they generate classification models that are simple to perform than some other classification techniques. As indicated above, this research is intended to recognize students who are presumed to find it difficult, so that these students should be deemed to be given sufficient guidance to enhance their outcomes.

This research could also be further significantly improved as a proposed development by observing data from a number of other universities, academies or schools and collecting more case studies to develop a model. Certain attributes may also be applied to the data set to improve the created model. Machine learning algorithms like the C4.5 decision tree algorithm, could learn efficient forecasting analytics from student data reported in earlier years. Scientific findings demonstrate that we can generate a quick but reliable forecast list for pupils by implementing predictive models to new students' data. In addition, some other classification techniques could be evaluated in this field.

REFERENCES

[1] Hämäläinen, W., Vinni, M. "Comparison of machine learning methods for intelligent tutoring systems". Conference Intelligent Tutoring Systems, Taiwan, 2006. pp. 525–534.

[2] Yudelson, M.V., Medvedeva, O., Legowski, E., Castine, M., Jukic, D., Rebecca, C. "Mining Student Learning Data to Develop High Level Pedagogic Strategy in a Medical ITS". AAAI Workshop on Educational Data Mining, 2006. pp.1-8.

[3] Cocea, M., Weibelzahl, S. "Can Log Files Analysis Estimate Learners' Level of Motivation?" Workshop on Adaptivity and User Modeling in Interactive Systems, Hildesheim, 2006. pp.32-35.

[4] Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004), "Detecting Student Misuse of Intelligent Tutoring Systems". Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540.

[5] Tang, T., McCalla, G. (2005), "Smart recommendation for an evolving e-learning system: architecture and experiment", International Journal on E-Learning, vol. 4, issue1, 105–129.

[6] Ching-Chieh Kiu. (2018), "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities", 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Malaysia.

[7] O. H. Lu, A. Y. Huang, J. C. Huang, A. J. Lin, H. Ogata, and S. J. Yang, "Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning", Journal of Educational Technology & Society, 21(2), 2018, pp.220-232.

[8] C. Romero, and S. Ventura, "Educational data mining: A Review of the state of the art", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2010, 40(6), pp.601-618.

[9] Tripti Mishra, Dr. Dharminder Kumar, and Dr. Sangeeta Gupta, "Mining Students' Data for Performance Prediction", Fourth International Conference on Advanced Computing & Communication Technologies, 2014, pp.255-261.

[10] Wati, M. Indrawan, W. Widians, J.A. & Puspitasari, N. "Data Mining for Predicting Students' Learning Result", 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 2017.

[11] R.S.J.D Baker and K.Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining 1, Vol 1, No 1, 2009.

[12] C. Romero and S. Ventura, "Educational data mining: a survey from 1995 to 2005," Expert Systems with Applications, no. 33, 2007, pp.135-146.

[13] Amirah Mohamed Shahiri, Wahidah Husain, and Nur'aini Abdul Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques", Procedia Computer Science, Vol. 72, 2015, pp.414-422.

[14] A Bellaachia, and E Guven, "Predicting the student performance using Data Mining Techniques", International Journal of Computer Applications, 2006, vol. 6.

[15] Anuradha, and T. Velmurugan, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance", Indian Journal of Science and Technology, Vol 8(15), IPL057, July 2015.

[16] U. Bin Mat, N. Buniyamin, P.M. Arsad, and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention", Engineering Education (ICEED), IEEE 5th Conference on IEEE, 2013, pp.126-130.

[17] T. M. Christian, and M. Ayub, "Exploration of classification using nbtree for predicting students' performance", Data and Software Engineering (ICODSE), 2014 International Conference on IEEE, 2014, pp.1-6.

[18] Amirah Mohamed Shahiri, Wahidah Husain, and Nur'aini Abdul Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques", Procedia Computer Science, Vol. 72, 2015, pp.414-422.

[19] Nguyen Thai Nghe, Janecek P., Haddawy P., "A comparative analysis of techniques for predicting acedemic performance", Frontiers In Education Conference – Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual, pp.T2G-7, T2G-12, Oct. 2007.

[20] A.L.Radaideh, Q.A.AI-Shawakfa, and E.M. AI-Najjar, "Mining student data using Decision Tree" International Arab Conference on Information Technology (ACIT 2006), Yarmouk University, 2006.

[21] Budi Santosa, "The Usage of Data using Data Mining Techniques for Business Requirements", Yogyakarta, 2007.

[22] Asia Ahmed Abu Shawish, Student Performance, Kaggle, July 3rd, 2020. Accessed on: July 20, 2020. [Online].