REVIEW ARTICLE | UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Performance Analysis and Prediction Student Performance to Build Effective Student Using Data Mining Techniques

Sirwan M. Aziz[1], Ardalan H. Awlla[2]

[1]Department of Computer Science, Darbandikhan Technical Institute SPU, Darbandikhan, Kurdistan Region - Iraq, [2]Department of Information Technology, Kurdistan Technical Institute, Sulaimani Heights, Behind Kurdsat TV, 46001 Sulaimania, Kurdistan Region – Iraq

ABSTRACT

In this period of computerization, schooling has additionally remodeled itself and is not restrained to old lecture technique. The everyday quest is onto discover better approaches to make it more successful and productive for students. These days, masses of data are gathered in educational databases; however, it stays unutilized. To be able to get required advantages from such major information, effective tools are required. Data mining is a developing capable tool for examination and expectation. It is effectively applied in the field of fraud detection, marketing, promoting, forecast, and loan assessment. However, it is an incipient stage in the area of education. In this paper, data mining techniques have been applied to construct a classification model to predict the performance of students. For the classification model, the cross-industry standard process for data mining was used as the classification model, the decision tree algorithm used as the main data mining tool to build the classification model.

Index Terms: Classification, Data Mining, Decision Tree, Naïve Bayes, Student Performance

1. INTRODUCTION

A decade ago, the quantity of higher education universities and institutes has multiplied manifolds. Massive numbers of graduates and postgraduates are produced consistently. Universities and institutes can also comply with the quality of the pedagogies; but nevertheless, they face the problem of dropout students, low achievers, and jobless students.

Understanding and breaking down the variables for negative overall performance is a complex and unremitting procedure, hidden in beyond and present facts congregated from educational overall performance and college students' behavior. Effective tools are required to research and expect the performance of college students scientifically.

Although universities and institutions gather a huge amount of students' information, this fact remains unutilized and does not help in any decisions or coverage making to enhance the performance of college students.

If universities could distinguish the circumstance for low execution prior and can predict students' conduct, this knowledge can help them in taking genius dynamic activities, to enhance the execution of such students.

It will be a win circumstance for every one of the partners of universities and institutions, i.e. administration, educators, students, and parents. Students could be able to

Access this article online	
DOI:10.21928/uhdjst.v3n2y2019.pp10-15	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Ardalan H. Awlla, Department of Information Technology, Kurdistan Technical Institute, Sulaimani Heights, Behind Kurdsat TV, 46001 Sulaimania, Kurdistan Region – Iraq. E-mail: ardalan.awlla@kti.edu.krd

Received: 10-05-2019 Accepted: 10-06-2019 Published: 20-06-2019

pick out their shortcomings in advance and can enhance themselves.

Teachers could be in a position to plan their lectures as according to the need of students and can give better direction to such students.

Data mining includes a fixed set of methods that can be utilized to extract appropriate and exciting knowledge from data. Data mining has numerous responsibilities, for instance, prediction, classification, association rule mining, and clustering. Classification strategies are supervised learning procedures that classify data object into a predefined class name. It is a standout among the most helpful strategies in data mining to create classification models from an input data set. The utilized classification procedures usually construct models that are utilized to predict future data patterns. There is the various algorithm used for data classification, for instance, Naïve Bayes classifiers and decision tree. With class, the created model could be able to predict a class for given data relying on earlier learned data from historical data.

Decision tree is a standout amongst the most utilized methods since it makes the decision tree from the records given utilizing clear conditions depending principally on the calculation of the gain ratio, which gives naturally a type of weights to attributes utilized, and the researcher can certainly distinguish the best attributes on the anticipated target. Due to this procedure, a decision tree would be worked with classification rules created from it.

Another classification method is Naïve Bayes classifier that is utilized to predict a target class. It relies on in its calculations on probabilities, particularly Bayesian theorem. Due to this use, the outcome from this classifier is more precise and efficient, and more delicate to new data added to the dataset.

Investigation and prediction with the assistance of data mining systems have demonstrated imperative outcomes in the area of predicting consumer conduct, fraud detection, financial marketplace, loan assessment, intrusion detection, bankruptcy prediction, and forecast prediction. It may be extremely powerful in education system also. It is a very effective tool to uncover hidden patterns and valuable information, which otherwise may not be identified and hard to discover and recognize with the assistance of statistical techniques.

In general, this paper tries to use data mining ideas, especially classification, to assist the universities and institutions directors and decision makers by assessing student' data to think about the primary characteristics that may influence the student' performance. This paper is organized as follows in section 2; literature review is discussed, in section 3 an entire detail of the study is introduced, in section 4 modeling and experiments are discussed, and in section 5 results and discussion presented. Finally, section 6 presents our conclusions.

2. LITERATURE REVIEW

All researches conducted previously, discover some huge areas in the education sector, where expectation by data mining has gained benefits; like, finding some students with weak points [1], select the points that students such as the exact course [2] evaluation of college [3], overall student evaluation [4], [5], class teaching language behavior [6], expecting students' retraction [7], [8], plan for course registration [9], guessing the enrollment headcount [10], and cooperate activity evaluation [11].

Some researchers indicate that there have been strong relationships between the student's personality likings and their work characteristics [12]. It is detected that there is detailed expertise needed to have once graduates to gain occupation and that these expertise are important to academic education generally. Characteristics such as sensitive cleverness, self-management development, and life work experience also are significant reasons for work development [13]. Employers try to differentiate the highest and lowest importance with soft skill and academic reputation [14]. Using machine learning techniques to predict the performance of a student in upcoming courses [15]. Overview of the data mining techniques that have been used to predict students' performance [16]. The performance of the students is predicted using the behaviors and results of previous passed out students [17].

3. BUILDING THE CLASSIFICATION MODEL

The fundamental target of the planned methodology is to fabricate the classification model that tests certain attributes that may influence student performance. To achieve this goal, the cross-industry standard process for data mining was used to construct a classification model. It comprises five stages that include: Data understanding, preparing data, business understanding, modeling, assessment, and deployment, as seen in Fig. 1.

3. 1. Data Classification Preliminaries

In general, data classification consists of two-advanced process. In the initial step, which is known as the learning step, a model that describes planned classes or ideas is constructed by examining a set of training dataset instances. Each instance is pretended to have a place with a predefined class. Within the second step, the model is tested utilizing an alternate different dataset that is utilized to assess the classification accuracy of the model. If the accuracy of the model is viewed as adequate, the model can be utilized to classify future data instances for which the class label is not notable. Ultimately, the model goes about as a classifier within the decision-making process. There are many strategies which can be utilized for classification, for instance, Bayesian techniques, Neural Networks, rule-based algorithms, and decision tree.

Decision tree classifiers are very well known procedures because the development of tree does not need any parameter setting or domain knowledgeable data and is acceptable for exploratory data discovery. A decision tree can deliver a model with rules that are comprehensible and interpretable. The decision tree has the benefits of simple clarification and understanding for decision makers to match with their domain information for approval and justify their decision. A number of decision tree classifiers are C4.5/C5.0/J4.8, NBTree, etc.

The C4.5 method is a type of the decision tree families that can deliver decision tree and rule sets, and develop a tree to improve expectation accuracy. The C4.5, C5.0, and J48 classifier is among the most famous and effective decision tree classifiers. C4.5 makes an initial tree utilizing the partition and Conquer algorithm. The entire depiction of the algorithm can be discovered in data mining or machine learning books, for example, C4.5: Programs for Machine Learning.

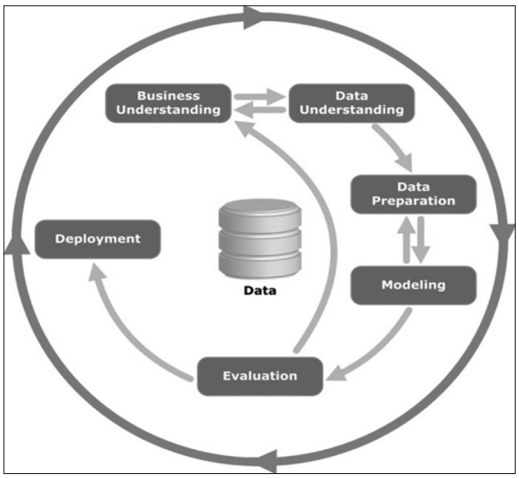

Fig. 1. Cross-industry standard process for data mining.

Weka contains a collection of machine learning and data mining algorithms for analyzing data and predicting modeling, together with the graphical user for simple access to these functions. It developed at the University of Waikato in New Zealand written in Java. WEKA contains tools for classification, regression, clustering, association rules, data pre-processing, and visualization.

3. 2. Data Collection Process and Data Understanding

While the concept of the paper came into mind, it means to apply a classification model for predicting performance relying on a dataset from a certain educational institute. With the goal that some other factors in regard to the studying environment, administration, conditions, and colleagues would have a comparable impact on all students, the impact of gathered attributes would be more evident and less difficult to classify. The data collected from three different educational institutes. To gather the necessary data, a questionnaire was organized and delivered either by email or manually to the students of all institutions. Then, it was additionally shared on the web, to be filled by students in any university or institutions. The survey was filled by 130 students, from the first, the second, the third institutions, and the rest from a few different institutions using the net questionnaire.

In the questionnaire, several attributes have been asked that may expect the performance class. The rundown of the gathered attributes is presented in Table 1.

3.3. Data Preparation

After the surveys were gathered, the method of preparing the data was completed. First, the information inside the questionnaires has been conveyed to (arff) to be appropriate with the WEKA data mining tool.

3. 4. Business Understanding

We have defined a classification model to predict if a student might show excellent performance. This issue is interesting since there are many universities/institutions interested in recognizing students with outstanding performance. For the input records for the prediction, the model use the data describing pupil conduct and the data defined student behavior as described in the previous table. The dataset includes 260 instances. Our model class label is a binary attribute, which separated students passed from first attempt exam (label value 1), for the students passed from second attempt exam (label value 0).

TABLE 1: Description of attributes used for predicting the student performance

Attribute	Description	Possible values
Gender	Student's Gender	Male, female
Time	Coming time to class	Never, once a week, twice a week, More than twice a week
Punishment	Number of punishment	I have never been punished, about twice, more than 3 times, very often
Family	Total number of family members	Between 3 and 5, between 6 and 10, more than 10
Parent live	My father and mother live harmoniously	Strongly agree, agree, natural, disagree
Parent education	Parent's education levels	Up to university, up to diploma, up to secondary school, up to primary, did not go to school
Parent financial	Parent's financial levels	Strongly agree, Agree, Natural, Disagree
Environment	The community around supports building of classrooms, library, toilets, etc.	Strongly agree, agree, natural, disagree
Encouragement	Our teachers inspire us to work hard	Strongly agree, agree, natural, disagree
Absent	Our teachers are never absent without a good reason	Strongly agree, agree, natural, disagree
Help	Our teachers are available and wiling to assist us in our studies	Strongly agree, agree, natural, disagree
Father	Does your father alive?	Yes, No
Mother	Does your mother alive?	Yes, No
Love	Do you have relationship love?	Yes, No
Accommodation	Are you stay at home or dormitory	Home, dormitory
Work	Are you working with your study?	Yes, No
Study	How many hours do you study per a day?	About 1 h, about 2 h, about 3 h, more than 3 h
Sleeping	Are you sleeping well?	Yes, NO
Pass	Are you passing in the first trial or second trial?	Yes, No

The fundamental usage of this model could identify well-performing students on a course. Individuals who ought to gain this model would be:

1. Instructors, for the qualification of students who can work together with;
2. Students, for checking if there is a requirement for more attempt to accomplish better outcomes;
3. Business people, for early attractive with students who are probably going to end up outstanding on a selected subject.

4. MODELING AND EXPERIMENT

After the data had been arranged, the classification model has been created. Utilizing the decision tree method on this technique, the gain ratio measure is used to signify the weight of influences of every attribute at the tested class, and thus, the ordering of tree nodes is specified. The results are discussed in the below section.

Referring to the analysis of earlier studies, and as defined in Table (1), a set of attributes has been selected to be tested against their influence on student performance.

These attributes consist of (1) personal information such as gender, love, sleeping, (2) education environment such as number of punishment, coming time to class, (3) parent information such as parent's education levels, and parent's financial levels. These attributes were used to predict student performance.

TABLE 2: Accuracy rate for predicting performance

Method	10-fold cross validation (%)	Hold-out (60%)
C4.5 (J4.8)	42.3	48.1
Naïve bayes	40.7	44.2

Three types of the technique have been applied to the dataset reachable to construct the classification model. The techniques are The Naïve Bayes classifier and decision tree with version ID3 (J4.8 in WEKA). The experiment, accuracy was assessed using 10-folds pass-validation, and hold-out technique. Table 2 shows the accuracy rates for each of those techniques.

The time attributes, which is student attendance to the class, have the maximum gain ratio, which made it the starting node and most efficient attribute. Other attributes cooperate inside the decision tree were parent lives, which is student's parent live, father, parent education, study, and accommodation. Rest of other attributes appeared in other parts of the decision tree.

The tree demonstrated that every one of these attributes has a type of impact on the student performance, but

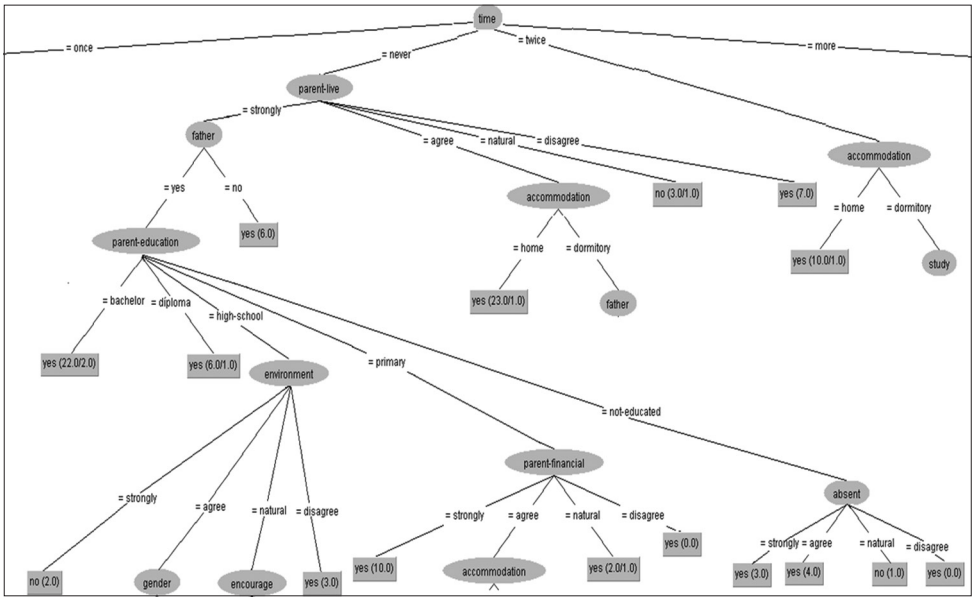

Fig. 2. A decision tree generated by the C4.5 algorithm for predicting performance.

the biggest attributes had been: Time, parent live, father, accommodation and parent education, as seen in Fig. 2, according to the dataset we collected in three different institutes and universities. It means if a student is never late to class, his or her parents live together harmoniously; their father is not dead, students stay at home not in a dormitory and their parents educated those students are passed in the first trial exam. The death of their mothers also impact student's performance, but in Iraqi Kurdistan, father's death affects students' performance more because fathers are the main financial providers for the family usually.

Wherever love (romantic relationship) is considered, students who do not fall in love have better performance than those having romantic relationship. Furthermore, the interesting attribute, which is home study (homework) does not have big effect, because if a student does not have a good environment no matter how many hours she or he studies, it does not have much effect to students' performance, as shown in a Fig. 3

The tree produced the use of the C4.5 algorithm showed that the time attribute is the most effective attribute. The Naïve Bayes classifier does not demonstrate the weights of every attribute incorporated into the classification; however, it has been used in comparison with the consequences generated from C4.5, as shown in Table 2, it can be seen that the efficiency percentage ranges about 36%–45%, which are low percentages.

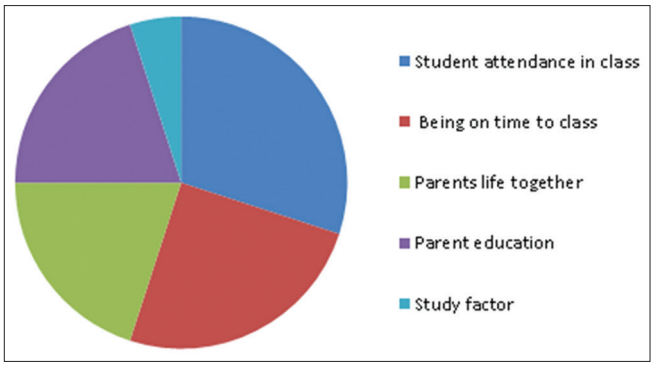

Fig. 3. High impact factors on students' performance.

Due to deep of the tree produced by J4.8 in WEKA, the visualization tree image is not clear here, we could show only a part of it, but if anyone is interested, they can download the dataset from the link [25] to do the experiment in WEKA.

5. RESULTS AND DISCUSSION

The study has shown that numerous elements may have a high impact on students' performance. A standout among the best is the student attendance in class. Various family factors also seemed to have an influence on the students' performance. A parent living together is one of the greatest positive factors in performance. It means if students' parents live within a good relationship, students' performance also increase because experiment

indicates that those students whose parents live together harmoniously are passed in the first trial exam. In addition, some other attributes after time and parents life attributes are father and mother education. If a student is never late to class, their parents are living together, their father is not dead and their parents have bachelor degree are in the second rank passed in the first trail, as explained before here in our community in Iraqi Kurdistan, fathers usually take financial responsibility of family not mothers, it means students do not need to work, otherwise students should work to pay for their life.

The rank attribute has shown an interesting influence on performance; it was not included as a high-efficiency factor. It was noticed in the experiment, the study factor. This is natural as no matter how long a student might study or prepare himself if they are not living in a good and secure house; they still perform very poorly in the exams.

6. CONCLUSION AND FUTURE WORK

This paper has focused on the probability of constructing a classification model for predicting student performance. Numerous attributes had been tested, and a number of them are found powerful on the performance prediction. The student attendance in class was the strongest attribute, then the parent living together harmoniously, father and mother education level, with the moderate impact of student performance.

The student punishment, sleeping hours, and family members did not show any clear effect on student performance while the no love relationship, parent strong financial status and student encouragement to study beside teachers have shown some effect for predicting the student performance.

For universities and institutes, this model, or an enhanced one, can be utilized in predicting the newly applicant student performance.

As future work, it is recommended to gather more appropriate data from several universities and institutions to have the right performance rate for students.

When the proper model is collected, the software could be created to be used by the universities and institutions, including the rules generated for foreseeing the performance of students.

REFERENCES

[1] A. Hicheur, A. Cairns, M. Fhima and B. Gueni. "Towards Custom-Designed Professional Training Contents and Curriculums through Educational Process Mining." IMMM; 2014. *The Fourth International Conference on Advances in Information Mining and Management*, 2014.

[2] B. N. A. Abu, A. Mustapha and K. Nasir. "Clustering analysis for empowering skills in graduate employability model." *Australian Journal of Basic and Applied Sciences*, vol. 7, no. 14, pp. 21-28, 2013.

[3] P. K. Srimani and Malini M. Patil. "A Classification Model for Edu-Mining". *PSRC-ICICS Conference Proceedings*, 2012.

[4] Y. He and Z. Shunli. "Application of Data Mining on Students' Quality Evaluation. Intelligent Systems and Applications (ISA)". *2011 3rd International Workshop on. IEEE*, 2011.

[5] S. Yoshitaka, S. Tsuruta and R. Knauf. "Success Chances Estimation of University Curricula Based on Educational History, Self-Estimated Intellectual Traits and Vocational Ambitions". Advanced Learning Technologies (ICALT). 2011 11th *IEEE International Conference on. IEEE*, 2011.

[6] P. U. Kumar and S. Pal. "A data mining view on class room teaching language." *International Journal of Computer Science*, vol. 8, no. 2, pp. 277-282, 2011.

[7] V. Dorien, N. De Cuyper, E. Peeters and H. De Witte. "Defining perceived employability: A psychological approach." *Personnel Review*, vol. 43, no. 4, pp. 592-605, 2014.

[8] A. S. Svetlana, D. Zhang and M. Lu. "Enrollment Prediction through Data Mining". *Information Reuse and Integration, 2006 IEEE International Conference on. IEEE*, 2006.

[9] P. A. Alejandro. "Educational data mining: A survey and a data mining-based analysis of recent works." *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432-1462, 2014.

[10] E. A. S. Bagley. "*Stop Talking and Type: Mentoring in a Virtual and Face-to-face Environmental Education Environment*." *Ph. D Thesis*. University of Wisconsin-Madison, Madison, 2011.

[11] J. Bangsuk and C. F. Tsai. "The application of data mining to build classification model for predicting graduate employment." *International Journal of Computer Science and Information Security*, vol. 10, pp. 1-7, 2013.

[12] M. Backenköhler and V. Wolf. "Student Performance Prediction and Optimal Course Selection: An MDP Approach" *International Conference on Software Engineering and Formal Methods*, pp. 40-47, 2017.

[13] A. M. Shahiri, W. Husainand and N. A. Rashid. "A review on predicting student's performance using data mining techniques." *Procedia Computer Science*, vol. 72, pp. 414-422, 2015.

[14] P. Shruthi and B. P. Chaitra. "Student performance prediction in education sector using data mining" *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 6, no. 3, pp. 212-218, 2016.

[15] P. L. Dacre, P. Qualter and P. J. Sewell. "Exploring the factor structure of the career EDGE employability development profile." *Education Training*, vol. 56, no. 4, pp. 303-313.

[16] S. Saranya, R. Ayyappan and N. Kumar. "Student progress analysis and educational institutional growth prognosis using data mining." *International Journal of Engineering Sciences and Research Technology*, vol. 3, pp. 1982-1987, 2014.

[17] A. E. Poropat. "A meta-analysis of the five-factor model of personality and academic performance". *Psychological Bulletin*, vol. 135, no. 2, pp. 322-338, 2009.