ISSN 2278-3091

Volume 9, No.1, January – February 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse110912020.pdf

https://doi.org/10.30534/ijatcse/2020/110912020

Predicting Students' Dropout Indicators in Public School using Data Mining Approaches

Jay S. Gil*, Allemar Jhone P. Delima[2], Ramcis N. Vilchez[3]

[1,2]Professional Schools, University of Mindanao, Matina, Davao City, Philippines

[3]College of Computing Education, University of Mindanao, Matina, Davao City, Philippines

jaygil89@gmail.com*, allemardelima@umindanao.edu.ph[2], ramcis_vilchez@umindanao.edu.ph

ABSTRACT

Currently, one of the challenges of academic institutions is dropout student issues. It is critical to recognize which students are at risk of losing in school and what are the underlying factors of dropout. The study of students' dropout in Lawigan National High School enables teachers to identify the influential factors of dropout cases in school preemptively, to respond to it immediately, and to assist prospective dropout students in continuing studies and gaining knowledge for a better future. Lawigan National High School is a public school supervised with the Department of Education. The said school holds all the records of the students. The study aims to identify the underlying factors of dropout students that need to have an intervention to lessen the value of dropouts. The Weka Experiment Environment platform used to run simulations with machine learning algorithms from ready datasets. Two classification algorithms C4.5 and Naïve Bayes (NB) tested on a dataset containing student academic and demographic details, e.g., age, gender, marital status, written, performance, and quarterly examination grade, attendance percentage, home distance, parents' income, mothers' education, fathers' education, and dropout status, using a 10-fold cross-validation to estimate generalization accuracy. The experiment result shows that the top indicator for students' dropout cases is the academic performance with 98.9474% accuracy as perceived by the C4.5 algorithm, as shown in Table 2.

Key words: Algorithms, C4.5, Data mining, Indicators, Naïve Bayes, Risk of Dropping Out, Weka Experiment Environment Platform

1. INTRODUCTION

Educational Data Mining (EDM) has an exciting research area. Since the need for academic institutions to improve the quality of education has become vital, EDM was born. An enabling environment is an essential reason for educating individuals [1].

Predicting underlying indicators of dropout students in high school is an essential concern in education. It concerns too many students in every school and institution around the world. The common factors why the student failed their studies are financial losses, low academic performances, family issues, health concerns, and a learning environment that would result in low grades and worst is a dropout. The definition of dropout varies between researchers, but in any case, if the institution loses the student by anyway, the school has a lower retention rate. Effective prediction of the dropout indicators in school is crucial to the success of any school retention plan. It is vital to identify the factors that highly affects the students' dropout in school, to address and provide some care to prevent the risk of increasing dropout rate and create an intervention to facilitate student retention.

Student attrition in schools hurts all stakeholders: students, institutions, and the general public. Notwithstanding the academic gain of a student before leaving school, school attrition represents a misuse of public and personal resources. Additionally, to monetary losses, dropping out students may create feelings of inadequacy and cause one being socially stigmatized [2]. Conversely, student's failure will impact negatively on students and schools. Dropout determined as a consequence for students who cannot complete their education until the specified study period. It makes the skills and abilities of dropout students in their fields less than retention and significantly affects institution quality [3]. However, considering the importance of the subject, there is still a great deal of ignorance about the underlying indicators and consequences of dropout, as well as about the effective means of reducing student attrition [2]. The main reason for constructing a predictive model for the student in school using data mining classifiers is to effectively identify the great factors of student dropout in an educational institution.

This initiative will make it possible for the school to make appropriate decisions and develop better strategies to avoid academic risk as a phenomenon and its implications, thereby reducing the likelihood of withdrawal. In this context, the authors have applied the classification algorithms to manipulate the knowledge that the institution collects from students.

2. RELATED LITERATURE

A. S. On et al. [5] presented the study to predict factors of higher education students. In their proposed research, they assess that one of the significant challenges that higher education encounter today is predicting student performance. The school would like to know more about the performance of the student organization. He introduced a trial to examine student performance when an extensive student information system database (SIS) given. Generally, the problems of students categorized into different patterns based on the level of students as normal, average, and below average. In this subject, an effort made to analyze the SIS database using a rough set of theories to predict the future of students.

Gole et al. [6] suggested the implementation of Educational Data Mining (EDM) to predict dropout or cancellation of engineering admissions. The comparative study of techniques and their results is, therefore, studied to analyze the graph of such students and to identify model characteristics for further use. It would consider whether these admissions could be predicted and determined at the time of admission or after one or two years. Here, in this paper, they are trying to form a model in which its characteristics would tell about students who are likely to leave engineering education. This model will develop using Educational Data Mining and the tools and techniques involved with it.

Sunbok Lee and Jae Young Chung [4] introduced a study aimed at improving the efficiency of the early warning system: (a) dealing with the problem of class imbalance using synthetic minority oversampling techniques (SMOTE) and set-up approaches in machine learning; and (b) testing qualified classifiers with both receiver operating characteristics (ROC) and precision-recall (PR) curves. Towards this purpose, they trained random forest, improved decision tree, random forest with SMOTE, and upgraded decision tree with SMOTE using huge data samples from 165,715 high school students from the National Education Information System (NEIS) in South Korea.

Nindhia Hutagaol [3] designed a system to find the best modeling approach for the identification of dropout student predictors from 17,432 private university student data in Jakarta. They also evaluated and calculated the association between demographic variables and academic performance to predict student dropout using three single classifiers: K-Nearest Neighbor (KNN), Naïve Bayes (NB) and Decision Tree (DT). They considered indicators such as student attendance, homework, mid-test grade, final grade, overall credit, GPA, student location, parent income, parent education level, gender, and age as student dropout predictors.

Shubhangi Urkude and Kshitij Gupta [5] developed this study to predict the students' performances in the examination and also to predict whether or not the student will graduate. For this reason, they are using a statistical analytical technique, which is the F1 score. The F1 score or the F measure used to measure the accuracy of the prediction by taking into account the accuracy and recall of the score. The dataset used in this study contains 395 student records with attributes such as age, health, internet, school, father's job, mother's job, etc. Utilizing support vector machines (SVM), Decision Tree, and Naïve Bayes (NB) classification algorithms, the score of F1 determined for each algorithm. Based on the analysis, the F1 help vector machine score gives a better prediction compared to the rest of the two algorithms.

A. Cano and H. M. Fardoun [6] suggested a technique and a basic classification algorithm discover clear and understandable models of student dropout prediction as soon as possible. They have carried out several experiments to predict dropout at different stages of the course, to select the best dropout indicators, and to compare the proposed algorithm with some well-known classical and unbalanced classification algorithms. Results showed that the algorithm was able to accurately predict student dropout during the first four-six weeks of the program and was efficient enough to be used in the early warning system.

3. METHODOLOGY

The focus of this research is to predict the indicators of students' dropout using data mining approaches. In this section, the phases of this analysis are fourfold to come up with the predicted dropout indicators; data gathering, data preprocessing, applying the data mining algorithm, and analyzing the result, as shown in *Figure 1*. That data collected from Lawigan National High School Guidance Office has been reprocessed and selected attributes were used in classifying.

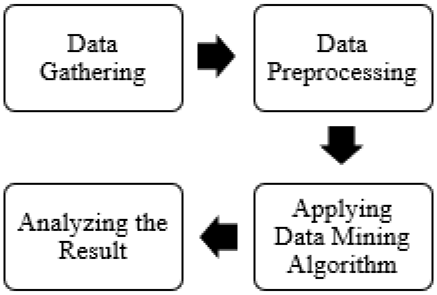

Figure 1: Research Method

3.1 Data Gathering

The dataset used for this research was collected from Lawigan National High School Guidance Office. Data attributes include students' grades, demographic, and academic performances, which were collected through school reports.

3.2 Data Preprocessing

Upon data collection, the data set was prepared to apply data mining algorithms.

Prior to the actual implementation of the specified model, preprocessing is handling the missing data by assigning the correct value and translating all values into numerical variables in order to improve the accuracy of the prediction, based on the algorithm's requirement. The preprocessing phase is used to measure the quality and appropriateness of the data. Only those attributes that were required for data mining were selected in this stage. This is achieved by extracting missing values, smoothing noise data, selecting the correct attribute from the database or eliminating irrelevant attributes, finding or removing external values from the dataset, and resolving data inconsistencies.

Some of the insignificant parameters have been removed from the database, such as student ID number and student name. A total of 12 variables correlated with student demographics and academic performance data have been listed from the dataset. The indexed variables in the database are in *Table 1*.

Table 1: Indexed Variables in Dataset

	Variables	Description	Values
1	age	Student's Age	15 yrs old & below = 1, 16-17 yrs old =2, 18 yrs old and above =3
2	gender	Student's Gender	male=0, female=1
3	marital.status	Student's Marital Status	single=1, married=2
4	written.grade	Student's quizzes and summative general grade	74 & below=1, 75-85 =2, 86-94=3, 95-100=4
5	performance.grade	Student's overall performance grade	74 & below=1, 75-85 =2, 86-94=3, 95-100=4
6	quarterlyexam.grade	Student's Examination Grade	74 & below=1, 75-85 =2, 86-94=3, 95-100=4
7	attendance	Student's Attendance Percentage	74 & below=1, 75-85 =2, 86-94=3, 95-100=4
8	home.location	Distance of School from Home	3Km below=1, 3Km-5Km=2, 5Km above=3
9	parents.income	Parent’s Income	5,000 below=1, 5,000-10,000=2, 10,001-20,000=3, 20,000 above=4
10	mother.education	Mother's Educational Attainment	no education = 0, elementary school = 1, secondary school = 2, tertiary = 3, master = 4, doctoral = 5
11	father.education	Father's Educational Attainment	no education = 0, elementary school = 1, secondary school = 2, tertiary = 3, master = 4, doctoral = 5
12	dropout	Student Dropout Status	Y/N

3. 3 Applying Data Mining Algorithm

The school has issues determining the variables that affect dropout events. While, by predicting the dropout factors, it can improve the performance of the school and help the academic system by giving early warning to students, by using a classification technique [7].

A. Hussain, et al. [8] claimed that the aim of data mining is to obtain relationships or trends that may provide useful information. Additionally, they argued that data mining is a process of getting significant relationships, patterns, and trends by analyzing a large number of stored data, using pattern recognition techniques such as statistical and mathematical techniques. One of the machine learning techniques is a classification algorithm, a learning technique used to predict the value of the target variables.

3.3.1 Decision Tree Classifier

J48 is an open-source JAVA implementation of the C4.5 algorithm in WEKA. C4.5 is a decision tree algorithm. The J48 Decision tree classifier uses a genetic algorithm [6]. When classifying a new item, it generates a decision tree that is based on the values of the attributes available in the training data. Therefore, whenever a new training set is identified, it defines the parameter that distinguishes the various instances with the utmost clarity. This means that these attributes are established which deliver optimal information and are used for better classification. This process takes place until all attributes with the highest to the lowest gain of knowledge are identified using the tree diagram. In this manner, the last classified attribute will bring us the result of the J48 technique prediction. Cross-validation of the dataset is also carried out to check if the result is correct. This is done by considering the ten-fold cross-validation training set.

3.3.2 Naive Bayes Classifier

Naïve Bayes Classifier is a classification based on Bayes Theorem, which assumes independent predictors. The classifier implies that if a particular feature found in a class, not linked to the existence of any other feature. The Naïve Bayes Classifier based on a simple definition. It utilizes variables separately included in the data sample by analyzing each of them. The Naïve Bayes classifier depends on the conditional probability obtained from the Bayes Law [9]. It utilizes all the variables included in the data, and then allows an individual analysis as they are equally important and independent of each other.

4. RESULTS AND DISCUSSION

In this analysis, the classification of dropout students was undertaken using data mining models based on C4.5 and Naive Bayes. 10-fold cross validation has been performed to verify each classifier. After performing the ten tests, the average test output was used to assess the accuracy of the model created.

Firstly, the authors do the data-cleaning process, such as resolving missing values in the dataset and facilitating datasets with the correct attributes. In this instance, four cells of under "student attendance" variable and five cells under "quarterly examination grade" variable having no value which are not related to the other variables have been excluded from the dataset. Next, the authors checked the standardization or distribution of data to determine whether the distribution of data was normal or balanced, and to minimize the effects of prediction errors during the modeling process. In addition, the impute value technique was implemented to fill missing value in the parents' income variable with its mean values in order to mitigate bias in the dataset. Lastly, the authors have 150 data with 12 variables as parameter inputs from the data cleaning process.

Based on the data, 57.33% of the student data was led by women, while men were only 42.67% of the total data. A percentage of 44.67 of student comes from three kilometers below from students' residence to school, 47.33% for three-five kilometers, and 8% for five kilometers above distance, which means that many of the students come from 3-5 kilometers away from school. In relation, 72% of students are mostly 16-17 years of age, followed by 15 years of age and below with 18.67% and 9.33 % for the students with18 years of age and above. The dataset also showed that 100% of students are single. The majority of mother's and father's educational attainment was "high school" and "elementary" level. The parent's financial income is mostly 5,000 to 10,000 a month. These data metrics, in demographic features, identify that the dataset has a relatively good variation to be used in the student dropout indicators prediction. Utilizing 150 input data, the authors observed, compared, and examined the efficiency of two different common classifiers, C4.5, and Naïve Bayes, as shown in Table 2 and Table 3.

Table 2: Accuracy Results of the Two Compared Algorithms

Classifier	Correctly classified instances	Incorrectly classified instances	Percent Accuracy of classification	Time Execution (seconds)
J48	94	1	98.95%	0.01
Naïve Bayes	93	2	97.89%	0.01

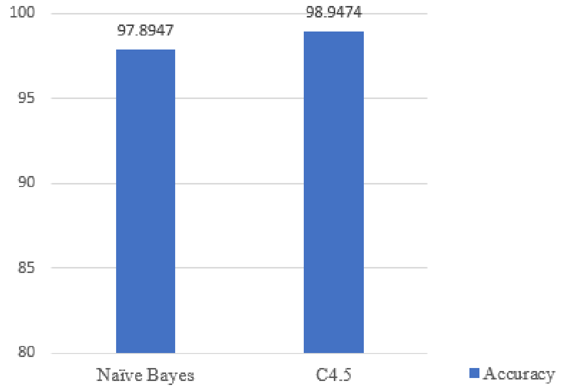

Figure 2: The Accuracy Rate of Two Algorithms

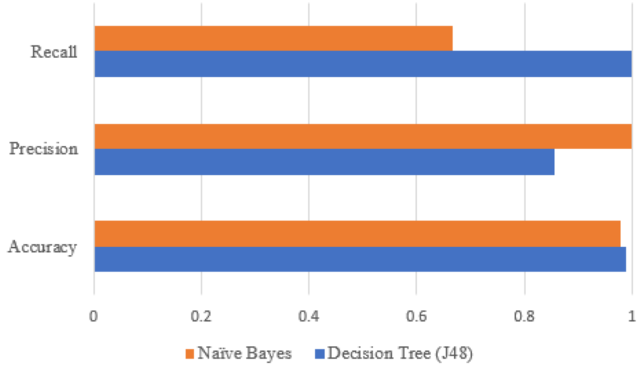

Figure 3: Comparison Performance Prediction between Two Models

Table 3: Performance Prediction Results

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class (Dropout Status)
J48	0.989	0.000	1.000	0.989	0.994	0.921	0.994	0.999	N
	1.000	0.011	0.857	1.000	0.923	0.921	0.994	0.857	Y
Naïve Bayes	1.000	0.333	0.978	1.000	0.989	0.807	1.000	1.000	N
	0.667	0.000	1.000	0.667	0.800	0.807	1.000	1.000	Y

Through a series of tests, the authors identified that the key attributes that affect student dropouts are mostly academic aspects, namely performance grade, quarterly exam rating, attendance percentage, and written score (e.g., quizzes, summative test). Such results are similar to the past data of Pereira and Zambrano [4], wherein the primary causes of student dropout attributed to the academic variables such as low academic grade. Also, the location or distance of the school from the students' residences has related factors of students' dropout. A similar study by Rahman and Dash, who have found out that the location of a student's residence, whether rural or urban, has been related to the attributes of dropout [5]. Moreover, the parent's income has also played a huge impact that affects the students' dropout, as shown in *Table 4*.

Table 4: Simulation Result for Rank Attributes using Gain Ratio

Rank	Attribute	Values
1	performance.grade	1
2	quarterlyexam.grade	0.7116
3	attendance	0.5135
4	written.grade	0.3925
5	home.address	0.1015
6	parents.income	0.0615

5. CONCLUSION/RECOMMENDATION

The main objective of this paper is predicting the student dropout indicators based on the academic and demographic data, running series of tests using the Weka Machine Learning Tool, and applying different approaches of data mining algorithm.

Predicting the causes of students' dropout is an important and challenging task for schools and educators. Accordingly, this study checked whether the use of data mining processes could be beneficial in addressing this issue in every school. Wherein the student dropout indicators successfully predicted using the Data Mining classification technique. To identify the dropout student indicators, the most commonly used data mining approaches were used based on C4.5 and Naive Bayes. These two different classification algorithms trained and tested using a ten-fold cross-validation approach. It alerts the educator to take appropriate action to improve student performance through specialized coaching and counseling.

Researchers have indicated that numerous student demographic indicators used to predict the incidence of dropouts [12]. With data mining, researchers may discover the attitudes and decisions related to student success, classify the student dropout indicators with identified ranks that causes dropout, personalize and customize learning material and instruction to suit individual needs, and enhance and maximize the use of educational support systems to facilitate learners [8]. Furthermore, researchers with limited technical experience may have trouble applying advanced data mining techniques and algorithms with their data. As [13] pointed out, it is also essential to create more user-friendly and practical tools for educators in the field of data mining.

In conclusion, the level of accuracy based on the prediction made by Naïve Bayes is 97.8947%, while the accuracy rate using C4.5 is 98.9474%. It shows that the C4.5 model is more accurate in predicting student dropout cases based on student data collection. Also, the gain ratio attribute evaluator found that the key attributes that profoundly affect student dropouts are in academic performance.

For future studies, researchers need to conduct research to synthesize and extrapolate meaningful knowledge from student data through data mining actions. The key limitations of this research were based only on a single experiment and a small sample of students. Many researchers can replicate this type of study with larger sets of student data from many schools. Additional studies may also include adjusting and adding variables, implementing certain algorithms, and altering preprocessing methods. Moreover, data collection from the qualitative data methodology (i.e., interviewing students' dropouts) may enable researchers to evaluate and analyze the data through the data mining technique by viewing a multifaceted picture of what happened. Also, these types of research could extend to other fields to achieve the greatest possible accuracy of predictions.

ACKNOWLEDGMENT

The heartfelt thanks to the staff and faculty of Lawigan National High School who have given educational data for this study, because of them, the objectives of the study successfully attained.

REFERENCES

1. J. Of and A. In, "JOURNAL OF ADVANCEMENT IN A Framework for Student Academic Performance Using Naive Bayes Classification Technique," vol. 6, no. 3, pp. 1–4, 2018.
2. J. Berens, "Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods †," 2018.
3. N. Hutagaol, "Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education," vol. 4, no. 4, pp. 206–211, 2019.
4. S. Lee, "applied sciences The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction," 2019, doi: 10.3390/app9153093.
5. S. Urkude and K. Gupta, "Student Intervention System using Machine Learning Techniques," no. 6, pp. 2061–2065, 2019, doi: 10.35940/ijeat.F1392.0986S319.
6. A. Cano and H. M. Fardoun, "Early Dropout Prediction using Data Mining : A Case Study with High School Students Early Dropout Prediction using Data Mining : A Case Study with High School Students," no. May, 2016, doi: 10.1111/exsy.12135.
7. S. K. Wanjau, G. Okeyo, and R. Rimiru, "Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions," in International Journal of Computer Applications Technology and Research, 2016, vol. 5, no. 11, pp. 698–704, doi: 10.7753/ijcatr0511.1004.
8. L. Jing, F. Shang-chun, A. Hussain, S. Gilani, and T. A. Al-aithan, "Drop out Estimation Students based on the Study Period : Comparisonbetween Naïve Bayes and Support Vector Machines Algorithm Methods Drop out Estimation Students based on the Study Period : Comparisonbetween Naïve Bayes and Support Vector Machines Algorit," 2016, doi: 10.1088/1757-899X/105/1/012039.
9. F. Razaque, N. Soomro, S. Ahmed, S. Soomro, J. A. Samo, and H. Dharejo, "Using Naïve Bayes Algorithm to Students' bachelor Academic Performances Analysis."
10. T. A. Cardona, S. Engineer-, C. S. Sigma, and B. Belt, "Predicting Degree Completion through Data Mining," 2018.
11. A. C. Study and P. S. Uni-, "The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining :," no. 1, pp. 356–368, 2019.
12. E. Yukselturk and C. Education, "PREDICTING DROPOUT STUDENT : AN APPLICATION OF DATA MINING METHODS IN AN ONLINE EDUCATION PROGRAM," vol. 17, no. 1, 2014, doi: 10.2478/eurodl-2014-0008.
13. S. Gole, P. Bajaj, and P. A. Thomas, "Prediction of Dropout Students from Engineering Education using Educational Data Mining (EDM)," 2017, doi: 10.15680/IJIRCCE.2017.