TAF Journal of Advances in Technology and Engineering Research JATER
2017, **3**(4): 117-132

ORIGINAL CONTRIBUTION

Mining student at risk in higher education using predictive models

January D. Febro [1*], **Jocelyn Barbosa** [2]

[1] Mindanao State University, Marawi, Philippines

[2] University of Science and Technology of Southern, Cagayan de Oro, Philippines

Index Terms
Data Mining
Educational Data Mining
Knowledge Discovery in Databases
Student Prediction

Received: 5 January 2017
Accepted: 20 February 2017
Published: 21 August 2017

***Abstract*—** This research uses an approach in Data Mining techniques to analyze historical data of students. The goal is to predict and investigate factors affecting student leavers using university admission (pre-admission variables) and educational achievement indicators (post-admission variables) on at-risk freshmen students in higher institution. Feature selection as preprocessing methods are utilized for 30 potential predictors to identify the most relevant factors. With the aim to evaluate the significant predictors contributing to at-risk students, this research compares predictive models. The models were implemented using various data mining algorithms and it is found that kNN gives 80.50% accuracy, CART with 89.70%, Adaptive Boosting using Decision Tree with 92.20% and Logistic Regression with 92.09%. The models were trained using records from student dataset collected from the Mindanao State University – Marawi, academic year 2010-2015. In addition, this research also verified the precision of the models through 10-fold cross validation, which can give veracity about what kind of data mining models works best in HEI data mining analysis. It would predict the class label 'Result' as categorical value, AtRisk or NotAtRisk. Benefits of the prediction model includes improving student retention and graduation rates.

I. INTRODUCTION

In today's technology-driven society, Higher Education Institutions (HEIs) are data-rich but information poor. It has a centralized database which contains large volumes of data only for the purpose of record keeping or process support. What HEI do not realize, lying within those datasets are patterns and potentials. This stored data can be collected to determine the relationships among variables to generate useful information for decision making that can greatly help HEI in retention problem. Data Mining (DM) refers to the "practice of efficiently examining valuable, non-obvious information from a large collection of data so as to draw out new information" thus helping HEIs make better informed decisions [1]. Educational Data Mining (EDM) on the other hand is the application of data mining in the academe. EDM refers to techniques, implements, and research designed for extracting knowledge from immense repositories of data that benefitted the institution [2]. From this, it is possible to classify, cluster or develop a prediction model. As reported by National Center for Education Statistics [3], around 40% of students seeking bachelor's degrees do not complete their degree within 6 years and as claimed by [4] about 30% of full-time first-year students seeking baccalaureate degrees "do not return for a second year".

Undergraduate college enrolments in the Philippines have grown increasingly for the past years. In Academic Year (AY) 2003-2004, HEI enrolled nearly 1,591,675 students. That figure increases nearly 28% number of students who enrolled in AY 2013-2014 to 4,104,841 students. As enrolment rate increases, students not graduating on time increase as well see Figure 1 - Commission on Higher

[*] Corresponding author: January D. Febro
[†] Email: january.febro@gmail.com

Education (CHED) data. This increase places a financial burden on students, parents, institutions, and the government. This issue is of such national importance.

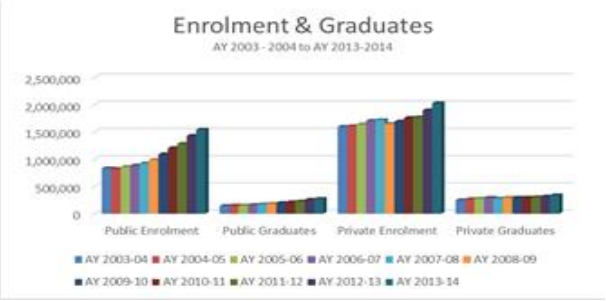

Fig. 1 . Higher Education Data 2014 (Public and private HEIs) enrolment and graduates

Retention and graduation should be a key concern and should be addressed by HEI. The advantages of a college degree are vital in many ways: first, it is apparently perceived that those students who finished college have a higher income compared to those with a secondary school diploma [5]; second, on top of much likely high financial income gain contributing to academic degree, bachelor's degree holders likely have a higher chance of being selected or employed [6]; and third, is the vast amount of knowledge gain experienced by college graduates who then go on to contribute to the economic and community development of our country [7].

Some institutions especially private HEIs, were dependent upon tuition revenue from their student. A decrease in enrollment means a financial strain to them. The focal question of modern HEI is how to thoroughly examine historical data of admitted students that will provide a better perspective on student retention and degree completion. In the Philippines, lack of research on the factors affecting student retention and the exigency of current system to evaluate and supervise student retention is not being regarded. The prediction model of student at-risk can be used as guidelines to prevent students from leaving HEIs. The motivation of this study is on the freshman students. Several researchers asserted that first year to second year is fundamental in indicating graduation rate [8, 9]. Freshmen come in college with worries and concerns over a new educational undertaking. They also bring complex educational and personal issues that require support service [10]. Securing students on the right tract commences with foreseeing their transition and meeting their needs as they begin. The researcher aims to answer question, which factors affect at-risk students of not retaining in HEIs and to discover if patterns can be found in the student data through predictive models.

This study will adapt the Systems theory input-output model of organization by Daft [11]. This theory, according to Owojari and Asaolu [12], postulates that an organized enterprise does not exist in a vacuum but is dependent on its external environment. Thus the enterprise receives inputs, transforms them and exports the output to the environment. In this study the university admits students (inputs) and then transforms them through teaching and learning which is reflected by the students' academic performance (output).

II. LITERATURE REVIEW

A. Selected Data Mining Methods for the Research Problem

The general objective of this study is to develop a predictive model for at-risk students. The researchers focused on the use and comparison of the following data mining algorithm for (Figure 2) predictive modeling.

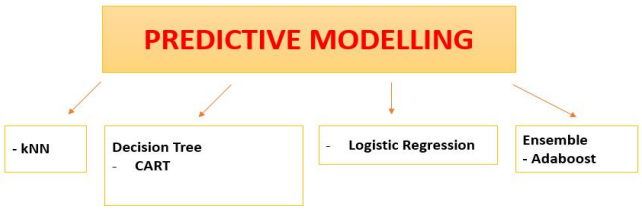

Fig. 2 . Predictive modeling used in this study

The k-Nearest Neighbor algorithm is established on "learning by analogy", that is, by comparing a given test example with training sets that are alike. It constitutes three vital methods: a set of labeled objects, a correlation metric to calculate how far or near the distance among the separating objects and the value of *k*. To categorize an unlabeled object, the distance between two object is calculated, its "k-nearest neighbors" are known, and the class labels of these nearest neighbors are then employed to decide the class label of the object [13]. Classification and Regression Tree (CART) algorithm was Leo Breiman's work in the early 1980s. It is an algorithm for data exploration analysis and predictive analytics. CART belongs to classification method which in order to construct DT uses past data [14]. Logistic Regression is a traditional statistical method for evaluating data sets using Logit model.

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

This technique is used to analyze independent variable/s that determine an outcome. The outcome is measured with two possible outcomes. The desired result of logistic regression is to find most fitting model to describe the relationship between dependent variable and the outcome variable and predictor variables. AdaBoost (Adaptive Boosting) is a meta-algorithm which can be used in conjunction with many other learning algorithms to improve their performance. It is formulated by Freund and Schapire, as cited in the work of [15], "is one of the most important ensemble methods, since it has solid theoretical foundation, very accurate prediction, great, and wide and successful applications".

B. Review of Studies Related to Enrolment, Retention and Graduation

The attribute is that often used in research studies as one of the predictors is cumulative Grade Point Average (GPA) [8, 16, 17, 18]. High GPA can be weighed as a notion of probable academic potential that leads student to retain and survive in university whereas low GPA in first semester indicates student attrition [19, 20]. Next to GPA is students' demographic and external assessments. Students' demographics comprise of gender, age, and family background [1, 21, 22]. As stated by [23], "family background in terms of family type, size, socio-economic status and educational background play important role in student's educational attainment and social integration". Herzog [24] stated that the best predictors of freshman's persistence is the first year grades acquired. Stinebrickner [25] concluded that the dropout that arises during the first-three years in HEI can be accredited to low academic performance.

Certainly, several studies have been conducted in educational and other types of research where these techniques used past data as attributes. However, EDM is still a new research discipline, and with little or no EDM research being done in the Philippine HEIs, it is fitting to conduct this research. The goal of the study, was to answer the call of the HEIs by bridging the research gap between the literature on predictors of first year student academic performance and the recent discussions about the emerging discipline EDM techniques and tools that will aid to solve student retention problem. Specifically, this undertaking focused on predicting the at-risk students of first-year enrolled in HEI.

Tino [26] suggests that success in the first semester of college is critical to student's persistence and that HEIs can make a difference. It is therefore crucial to determine variables that play a role in the first-year transition to college for students in HEIs. On the foundation of an extensive review of the literature, it was hypothesized that past data of students would be useful in predicting student retention. The result of this model can be very valuable and useful to HEIs in aiding decision making and driving change.

III. METHODOLOGY

A. Research Design

The data mining aspect of the work relied on historical data of student's which were collected between the academic year 2010 and 2015. Data mining is employed to derive significant knowledge concerning a particular dataset and to generate important relationships between variables stored in dataset. This research makes use of data mining approach to discover likely hidden valuable and unknown pattern from a collection of data. Through the aid of software data mining tool, data can be analyzed, preprocessed and summarized to identify relationships.

When using data mining procedure, no previous ideas of foreseeable relationships amongst variables can be obtained or have any ideas about the predictable results from the data set. Exploratory data analysis is performed to explore into the dataset and study the relationships between attributes. Through data mining techniques, patterns can be found from institutional datasets. In order to evaluate the significant variables contributing to at risk students, this research used selection feature as part of preprocessing phase.

Four different data mining predictive models: kNN, simple CART, logistic Regression, and AdaBoost are used and compared in this study. The accuracy and precision of the four data mining models are tested and validated, which can deliver meaningful information about what kind of data mining models is the best fit in higher education data mining analysis. The predictive model with highest accuracy will be used in prototype development phase to create a web-based warning system that could identify at-risk students.

B. Research Procedure

It is only suitable that dataset requirements be examined carefully. Dataset used in data mining models should be precise and consistent. Data mining process is necessary on account of the fact that the preciseness of results

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

acquired from data mining models is directly dependent on the accuracy of data. The detailed explanation about all phases is described below.

C. Phase 1 Data Collection

The dataset analyzed in this study were enrolled in Mindanao State University-Marawi (MSU); freshmen, full-time students from academic year 2010 to Academic Year 2015. The records were queried from the databases at the Information Systems Department (ISD) of MSU. ISD serves as the university's official source of information such as student enrollment, grades, employee profile, and payroll. The records obtained, contain information about the different courses of students, entrance result, GPA, etc. Information regarding student pre-university data and college data was collected from a total of 7,936 student records in five academic years. To select the risk variables, factors presented in the literature review were considered: first semester GPA [27], socio-economic status [26, 28], student demographics [1, 22] and critical basic courses that might hinder students' performance during first year such as Math and English, course status is also included. The dataset is delimited to first-time, full time (15 units or more) students who commence first semester of their academic career. Variables selected for the study were based on the availability of data from university records related to the literature on parameters for predicting academic performance. The 30 potential predictor variables selected fell into two categories: pre-college data and post-admission data. Pre-college data are data prior to admission. They include admission test scores and some socio- demographic attributes.

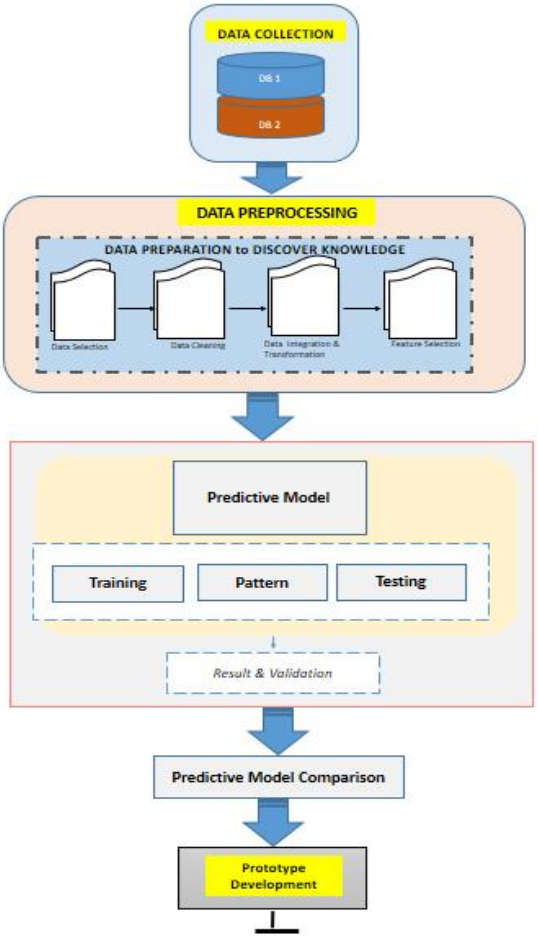

Fig. 3 . Methodological framework

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

Table 1 includes a detailed description of dependent variables and independent variable. The pre-college dataset fields investigated in this study can be grouped into two: academic potential (admission test score in Math, Language Usage, Aptitude, and Science) and demographic and socio-economic (Gender, blood type, skills, sports, musical instrument, province of origin, parents educational background, parent's income, parent's tribe, religion, number of brothers, number of sisters and rank in family). On the other hand, post-admission data are educational achievement indicators such as course, scholarship status, grades in Math and English subjects, and first semester GPA.

TABLE 1
SUMMARY OF PREDICTOR VARIABLES

Univid	Unique Identification Number of Student	Assigned Number
Gender	Student gender	Male, Female
Age	Age	0 - > 20
Blood type	Blood type of student	A, B, O, AB
Height	Height of student	<100 - >200
Weight	Weight of student	<100 - >200
Home add	Home address of student	home address
Prov	Home province of student	Lanao del Sur, Lanao del Norte, Mis Occ, …
SPECIALSKILLS	Special skills	Painting, Debater, …, NA
SPORTS	Sports	Basketball, Table Tennis,…,NA
Musicins	Musical instrument played	Piano, Guitar, Drums,…,NA
Mother education	Mother's highest degree of education	Elementary Graduate, Secondary Graduate, College Graduate, …, NA
Fathers education	Father's highest degree of education	Elementary Graduate, Secondary Graduate, College Graduate, …, NA
Gross income	Parent's income in a year.	0 - > 600, 000
Brothers Num	Number of brothers	0 - > 8
Siaters Num	Number of sisters	0 - > 8
Aptitude	Entrance exam aptitude score	<20 - > 50
LU	Entrance exam Language Usage score	<20 - > 50
Math	Entrance exam Math score	< 20 - > 50
Science	Entrance exam science score	< 20 - > 50
Genrating	General rating in entrance examination	< 20 - > 50
Camp pref	Campus preference of student	Marawi, IIT, Naawan, Sulu, Tawi-Tawi, …, NA
Program code	Program code	Program Code
Program	Program enrolled by student	BS Agriculture, BS Animal Science, …
Scholarship type	Paying or Scholar	Paying, Scholar (SPL, Tuition Privilege, Academic, CHED, DOST, Academic, CBCP)
Religion catagory	Religion category	Muslim, Non-Muslim
English status	English grade in first semester	1.0-5.0, 0 for not taken
Math status	Math grade in first semester	1.0-5.0, 0 for not taken
GPA	First semester GPA of freshman year	(1.0 – 5.0) 1.0-5.0
Enrollment status	Academic year (first year to second year) retention in a bachelor degree program	Enrolled, Not-Enrolled

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

IV. ASSUMPTIONS

It was postulated that the records in the university database were accurate. The following statement is logical considering that the information was exploited customarily by the university in generating results and producing reports.

Fig. 4 . Data processing

Phase 2 data preprocessing: Before running data mining algorithms, to improve the input data quality and suitability of data for mining data preprocessing was applied. For this, identifying noise data, missing values, irrelevant and redundant data, identifying or removing outliers, and data transformation are very crucial. This phase consists of data integration and transformation, data cleaning, and feature selection. At the end of this phase, the final dataset has been constructed. Careful data integration is done to reduce and avoid redundancies and inconsistencies. Redundant data are carefully examined during this phase. Same attributes and derived attributes in different databases are not included.

Some data are being transformed to suit a data mining technique. New attributes are constructed inferred from existing attributes. Feature type conversion, some algorithms only handle numeric features while some can handle only nominal features. In this study, features have to be converted to satisfy the requirements of the learning algorithms. (for instance, Numeric to Nominal Discretization is used in handling "parents' educational attainment, first generation students?–Yes/No.

TABLE 2
CONVERSION OF INCOME

1	2	3	4	5	6	7
A	B	C	D	E	F	G
Poor	Low income (but not poor)	Low middle income	Middle class	Upper middle income	Upper income (but not rich)	Rich
< 94,680	94,680-189,360	189,360-378729	378,729-946,800	946,800-1420,200	1420,200-1893,600	> 1893,600

Some algorithms require the data to be normalized to increase the efficacy of the algorithm. In this way, it helps prevent attributes with large ranges outweighing one with small ranges. In this study, proportion transformation method is performed to rescale gross income field. Dataset collected from the university repository may be missing. The data cleaning process will remove the missing data or incomplete, noisy data. Although there are several statistical methods for replacing values, list-wise deletion method will be adopted.

A list-wise deletion method deletes the entire record from the records if any variable in the model has a missing value. Filling in the missing data is not applied to avoid adding bias and distortion to the data just to make the information available. Removing a few records will not impede the results of the model. To handle outliers, local Outlier Factor (LOF) is performed. It works by associating the local density of the data occurrence to that of its neighbors [29].

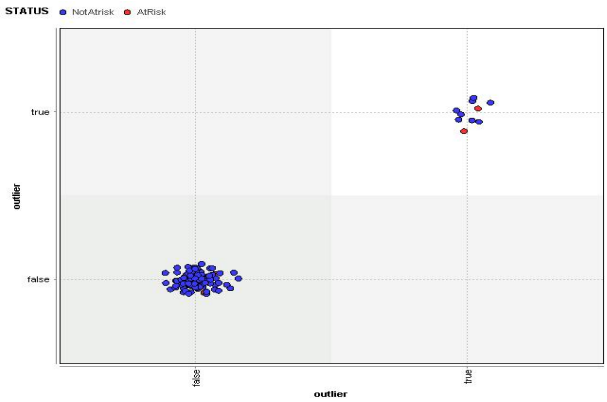

Fig. 5 . Outliers in the data

The local density of a data occurrence is proportional to the reciprocal of the mean distance to its k-nearest neighbors. The LOF score is set to the proportion of the local density of the data occurrence to the average local density of its neighbors. Suchlike returns in standard data

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

having an LOF score of relatively equal to 1, in contrast to outliers with scores higher than 1. This is explained by the fact that if the data lies in a dense region, their local density would be similar to that of their neighbors leading to a ratio of 1. An LOF with large dataset having a score of up to 2 demonstrated that the data instance is normal. After handling missing values and removing outliers, from 7, 936 records to 7, 860 records.

Feature selection: Part of this study was to identify what dominant variable or combination of variables collected can be used as predictors of academically at-risk students from the HEI. Feature selection is of great consequence in which the most important attributes or variables that have the direct effect on label attribute can be ascertained. There are different methods that are used for feature extraction. In this study, feature selection is performed as a preprocessing step. It has been considerably efficient in lessening proportion, eliminating unnecessary data or noise from dataset refining result coherently. In this study, filter model using feature ranking was used - Info Gain Ratio and Correlation Feature Selection. Filter model had devoid bias with regards to whatever learning algorithm.

• Information gain ratio resolves the flaw of information gain. Gain ratio is computed for each of the attribute using equation as cited in [30].

$$IGR(Ex, a = IG/IV) \tag{1}$$

$$IF(Ex, a) = H(Ex) - \sum_{v \varepsilon values(a)} \left(\frac{|x \varepsilon Ex | values(x, a) = v}{|Ex|} . H\left(\left\{ x \varepsilon Ex | value(x, a) = v \right) \right\} \right) \tag{2}$$

$$IV(Ex, a) = - \sum_{v \varepsilon values(a)} \frac{|x \varepsilon Ex | values(x, a) = v}{|Ex|} . log_2 | \left(\frac{\left\{ x \varepsilon Ex | value(x, a) \right\}}{|Ex|} \right) \tag{3}$$

```
protected double getSplitInfo(double[][] weightCounts) {
    double[] splitCounts = new double[weightCounts.length];
    for (int v = 0; v < weightCounts.length; v++) {
        for (int l = 0; l < weightCounts[v].length; l++) {
            splitCounts[v] += weightCounts[v][l];
        }
    }

    double totalSplitCount = 0.0d;
    for (double w : splitCounts) {
        totalSplitCount += w;
    }

    double splitInfo = 0.0d;
    for (int v = 0; v < splitCounts.length; v++) {
        if (splitCounts[v] > 0) {
            double proportion = splitCounts[v] / totalSplitCount;
            splitInfo -= Math.log(proportion) * LOG_FACTOR * proportion;
        }
    }
    return splitInfo;
}

protected double getSplitInfo(double[] partitionWeights, double totalWeight) {
    double splitInfo = 0;
    for (double partitionWeight : partitionWeights) {
        if (partitionWeight > 0) {
            double partitionProportion = partitionWeight / totalWeight;
            splitInfo += partitionProportion * Math.log(partitionProportion) * LOG_FACTOR;
        }
    }
    return -splitInfo;
}
```

Fig. 6 . Info gain ratio code snippet

• Correlation-based Feature Selection (CFS) looks for features that are particularly correlated with the explicit class. CFS is defined as follows:

$$r_{zc} = \frac{k r_{zi}^{-}}{\sqrt{k} + k(k-1) r_{ii}^{-}} \tag{4}$$

where: *rzc* is the correlation between the scored features, *k* is the sum of features, r_{zi} is the mean of the correlations relating to the class variable, and r_{ii} is the mean of inter-correlation between features [31].

```
public AttributeWeights calculateWeights(ExampleSet exampleSet) throws OperatorException {
    Attributes attributes = exampleSet.getAttributes();
    Attribute labelAttribute = attributes.getLabel();
    boolean useSquaredCorrelation = getParameterAsBoolean(PARAMETER_SQUARED_CORRELATION);

    AttributeWeights weights = new AttributeWeights(exampleSet);
    getProgress().setTotal(attributes.size());
    int progressCounter = 0;
    int exampleSetSize = exampleSet.size();
    int exampleCounter = 0;
    for (Attribute attribute : attributes) {
        double correlation = MathFunctions.correlation(exampleSet, labelAttribute, attribute,
        useSquaredCorrelation);
        weights.setWeight(attribute.getName(), Math.abs(correlation));
        progressCounter++;
        exampleCounter += exampleSetSize;
        if(exampleCounter > PROGRESS_UPDATE_STEPS) {
            exampleCounter = 0;
            getProgress().setCompleted(progressCounter);
        }
    }

    return weights;
}
```

Fig. 7 . CFS code snippet

The significant variables assessed by feature selection techniques were the final parameters used in creating a predictive model feed into a data mining tool. Through this a pattern can be extracted to predict student at-risk/not at-risk.

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

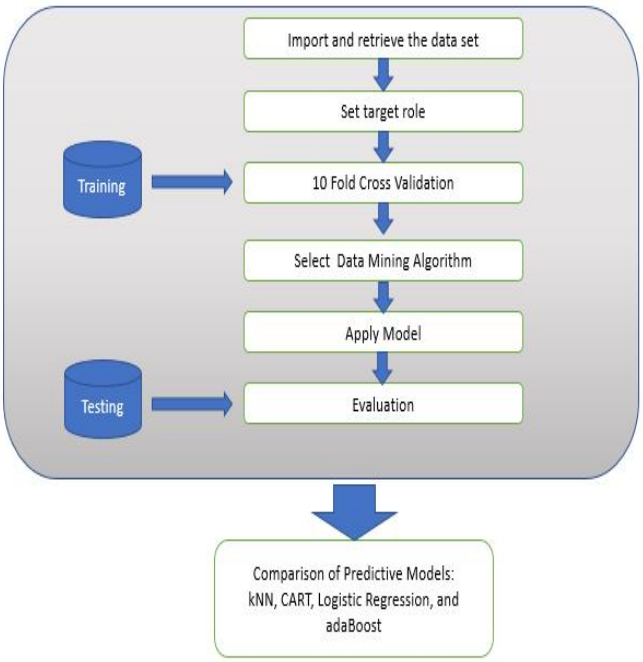

Fig. 8 . CFS code snippet

Phase 3 modelling: This study compared predictive data mining methods: kNN, CART, logistic regression, and AdaBoost based on the literature review in Chapter II. Algorithms that fall under predictive modelling algorithm are used since our goal is to learn how to assign class label -AtRisk (1), NotAtRisk (0) to the unseen data based on models built on the training data to create a prediction model. Each of these algorithms was optimized to fit the student at risk data and then compared to conclude the best data mining model. The model with highest precision is used in prototype development.

There is a close connection between data preprocessing phase and modeling phase in consideration of specific methods that require exact data formats. In addition, data problems might occur while modeling or a need for constructing new data. About two thirds (70%) of the data are set aside as training dataset and are split randomly. The remaining one third (30%) is assigned as a test set and is used for evaluating the model. In order to create a predictive model by means of DM method, the data need to be trained.

The result received from the testing data process will be evaluated by assessing the accuracy. At the evaluation phase, models were built from a data analysis perspective and were systematically evaluated. The procedures were reviewed in executing to construct model ascertaining the objectives are properly achieved. This study similarly verified the accuracy of the four data mining models used in order to provide additional information and understanding about which data mining models perform precisely in higher education data mining analysis. 10-fold cross-validation will be used to test the accuracy of the models. All the methods were tested by using a 10-fold cross validation. The test divides the dataset into 10 equal subsamples. One subsample is kept for validating the data, while the remaining 10–1 subsamples are used for training. This process is repeated until all subsamples have been used for validation.

Stratified sampling method is used to divide the entire dataset into ten different parts with equal proportions of at-risk students and not at-risk students in each set. Nine out of 10 sets are used as training data to build models and the data are run through the remaining one dataset. A classification error rate is calculated for the model and stored as an independent test error rate for the first model. Next, a second model is constructed with a different set of nine samples and then a test error rate is calculated. The same process is repeated ten times resulting in ten individual models.

The classification error rates for all ten models are then averaged. The prediction model which has the highest precision will be used in the prototype implementation. At the end of this phase, a decision on which of the predictive models should be used for prototype development is concluded.

A. Predictive Modeling Algorithm

kNN: The k-Nearest Neighbor algorithm compares a given test instance with training instances that are alike. It is grounded on learning by analogy. The algorithm initially calculates the test set like in this study the student being predicted then it finds similarity to all examples in our training set and retrieves the k most alike. Likeness is computed with a Euclidean distance between the features of the test data and matching features of each example in the training set. This study used standardized features for distance calculation as an instance of extending kNN. The standard score (z-score) is computed as:

$$Z_i = \frac{x_i - \mu}{\sigma} \quad (5)$$

where $x-i$ is the feature student initial value, μ is the average value, and σ is the *SD*.

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

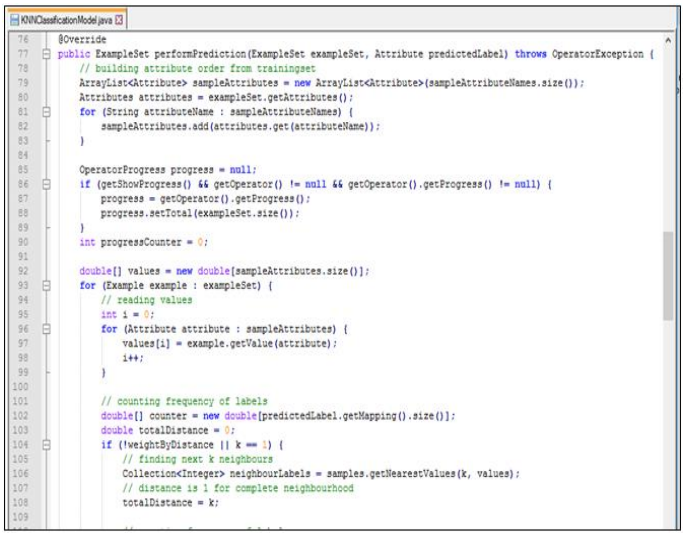

Fig. 9 . KNN code snippet

CART: Classification and regression tree (CART) is a decision tree is and learning technique, which gives the results as either classification or regression trees, depending on categorical or numeric data set [14]. CART's algorithm was Breiman's work [14]. To build a DT, CART uses Gini Index as an attribute selection measure:

$$1 - \sum jp^2 i \tag{5}$$

Fig. 10 . CART pruning method code snippet

B. Logistic Regression

Logistic regression uses Logit model for evaluating data sets. This technique is used to analyze independent variable/s that determine an outcome. This study uses binary logistic regression method since our target variable status response is AtRisk(0)/NotAtrisk(1) which is binary. Logistic regression will evaluate the likelihood of a student's status – AtRisk or NotAtRisk. The generalized linear model [32] formula is given by:

$$logit\left(E|Y_i|X_i|\right) = logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \beta.X_i \tag{6}$$

where β is grouping of $\beta 0$, $\beta 1$, ..., βm (regression coefficients) into a single vector of size *m + 1*; X_i that is a group of $x_{0,i}$, $x_{1,i}$, ..., $x_{m,i}$ (resulting explanatory variables) single vector of size *m + 1*.

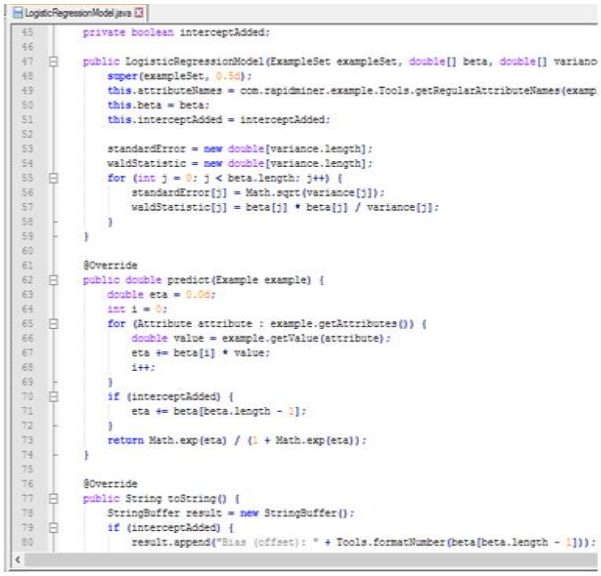

Fig. 11 . Logistic regression method code snippet

AdaBoost: Recall that AdaBoost is a meta-algorithm, a DT is used as learning algorithm conjunction to improve their performance. This study used the version of Friedman's Ada Boost algorithm [33]. It is done by lessening the exponential loss function.

$$\ell_{esp}(h|D) = E_x \sim D\left[e^{-f(x)h(x)}\right] \tag{7}$$

Using additive weighted combination of weak learners as;

$$H(x) = \sum_{t-1}^{T} a_t h_t(x). \tag{8}$$

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

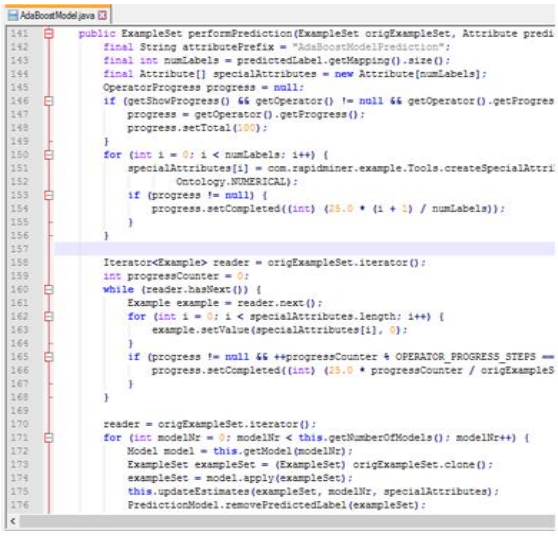

Fig. 12 . AdaBoost method code snippet

D. Phase 4 Result Evaluation

We will evaluate the models through its performance matrix. One of the performance matrices used for classifiers in machine learning is confusion matrix. It shows actual and predicted instances by classifiers. A confusion matrix is a table generally used in supervised learning to show the performance of an algorithm. Its columns denote the number of occurrences of a projected class, and its rows denote the number of occurrences of an exact class [34].

TABLE 3
CONFUSION MATRIX

		Predicted	
Actual	Negative	TN	FP
	Positive	FN	TP

All the entries with description are given below
- "TN" denotes correctly set as negative cases
- "FP" denotes incorrectly set as negative cases
- "FN" denotes incorrectly set as positive cases
- "TP" denotes correctly set as positive cases

Other parameters that can be assessed from matrix entries are accuracy, precision and recall: Accuracy per class can be defined as instances correctly classified as their actual class divided by total instances of that class; Precision can be defined as positive predicted instances divided by total predicted instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

Phase 5 prototype development: In prototype development phase, we aim to develop a prototype model that will predict at-risk students based on Logistic Regression. Figure 13 shows the linear input-output-process of the prototype. The user enters student information in the input form or through uploading an excel file and then the system will process the user's input or request. The system will display the predicted result.

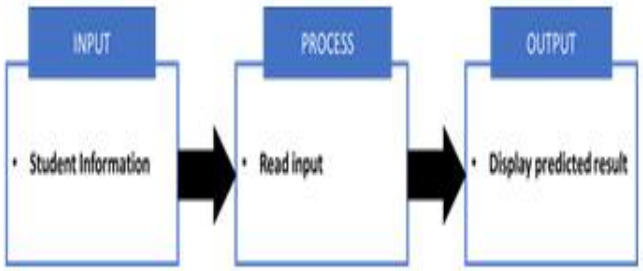

Fig. 13 . Input-process-output

IV. RESULTS

In Chapter 3 a detailed research methodology was presented and discussed. This chapter focuses on the work that was done and presents the results obtained. After measuring the quality of data, data mining techniques were applied to develop a prediction model and study factors affecting students at-risk.

A. Analysis of the Collected Data

We stated in first chapter that our goal was to analyze the factors affecting at risk students. For this study, data were collected by querying the university database of MSU-Marawi. The population was gathered between AY 2010 and 2015 with a total of 7,859 freshmen full-time students. The dataset collected can be categorized as pre-college variables and college variables. The pre-admission variables include socio-demographic information and college variables comprised of student data recorded at the end of the first semester (Math Status, English Status, first sem GPA). The population of at-risk students in MSU-Marawi was used in this research. The dataset had the target variable AtRisk.

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

TABLE 4
FREQUENCY AND PERCENT OF SOCIO-DEMOGRAPHIC STUDENTS DATASET

Status	Percent
At Risk	20.11%
Not at risk	79.89%

The socio-demographic issues explored in this record were Gender, Religion Category, Province, Number of Brothers and Sisters and Income Class. Tables 3 and 4 shows the demographic profile of the student's dataset. The age range of the students was from 13 to 29 years with the average of 17 years. Majority of the records belong to female population with 4,960 (63%) records while Male population is 2,899 (37%). In terms of religion category, 62% belongs to Muslim group and 38% belongs to Non-Muslim group. More than a half (57%) of the students were residing within Lanao del Sur.

TABLE 5
FREQUENCY AND PERCENT OF NOT RETAINING STUDENTS

Gender	Frequency	%	Religion Category	Frequency	%	Province	Frequency	%
Male	2899	37%	Muslim	4866	62%	Lanao del Sur	4466	57%
Female	4960	63%	Non-Muslim	2993	38%	Other Province	3393	43%

Most of the students belong to poor, low income and low middle income class. The poor income class makes up more than half of the collected data (67% or 5,230) of total population in the income class (see Table 4). The poor income class has an annual income of less than Php 94,680.00 while the rich class has an annual income of more than Php 1,893,600.00.

TABLE 6
FREQUENCY AND PERCENTAGE OF INCOME CLASS

Income Class	Frequency	%
Poor	5230	67%
Low income (but not poor)	1172	15%
Lower middle income	989	13%
Middle class	424	5%
Upper middle income	34	0%
Upper income (but not rich)	5	0%
Rich	5	0%

Table 5 shows the relationship between variables collected and at risk students. Of the 1,581 at risk samples, average age was 17. There were more male (51%) at-risk students than female (49%). 68% of the at-risk students were living in Lanao del Sur and 88% with no scholarship. Table 6 shows the comparison between their college entrance exam average scores in Aptitude, Language Usage, Math and Science.

TABLE 7
STUDENT DEMOGRAPHIC AND AT RISK POPULATION

Age	Gender			Religion Catagory			Province			Paying		
Average		F	N		F	N		F	N		F	N
17	Male	813	51%	Muslim	1145	72%	Living in Lanao del Sur	1079	68%	Paying	1393	88%
	Female	768	49%	Non-muslim	436	28%	Other province	502	32%	Scholar	188	12%

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

The Table shows that at-risk population gets lower scores on all the categories than non-at-risk students.

TABLE 8
AVERAGE SCORES IN COLLEGE EXAM

Average Score	At Risk	Not at Risk
Aptitude	15.68	18.76
LU	29.22	36.5
Math	12.67	16.1
Science	10.02	12.48

B. Predictors of at-risk students

Several researchers cited about how factors such as race and gender may or may not be relevant as predictors of student's academic performance [21, 22, 35, 36, 37]. In some countries like India, socio-economic and educational context, race and gender have been used as key determinants of student retention. In this research these factors were used to look into their effect or influence, if any, on at-risk students.

Feature selection was an important step to select the most important attributes or variables that have had direct effects on label attribute. The methods used were filter model using Info Gain Ratio and Correlation Feature Selection. Though the results of the feature selection suggested dominant predictors, we have included all available predictors in predictive modelling to assess variables with little significance that may affect the overall prediction outcome.

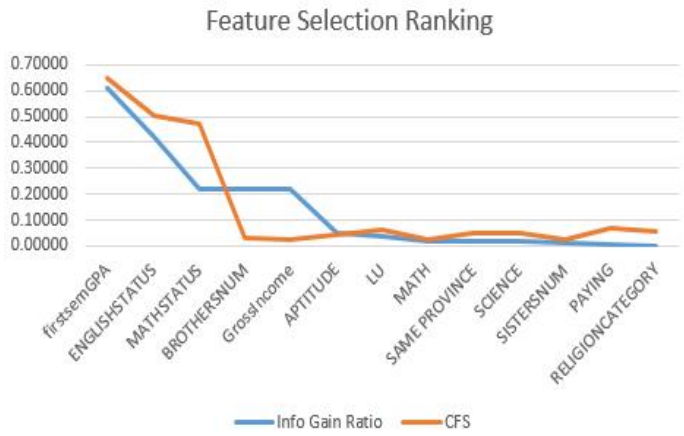

Fig. 14 . Feature selection

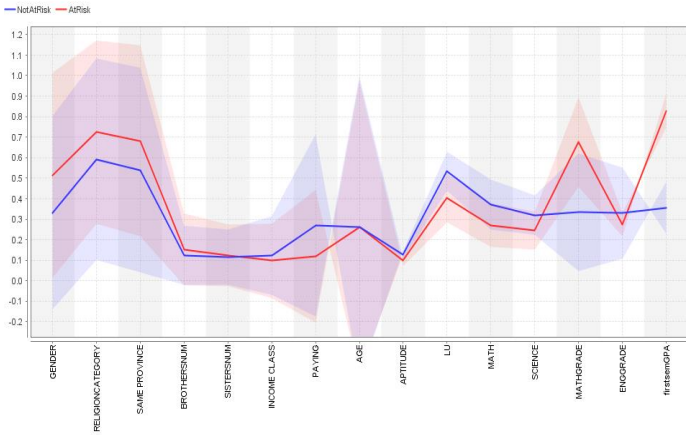

Fig. 15 . Standard deviation of predictors

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

C. Data Mining Technique Comparison and Selection

From 30 possible predictor variables, only 14 predictor variables were chosen after Multivariate Analysis.To address objective no. 2 of this study, identify and select data mining techniques for developing predictive models, we compared four predictive modelling techniques based on their accuracy, precision and recall. The main objective of this study focused on the At Risk students. To improve the accuracy and to balance the data of the at-risk students random under sampling was done to the NotAt Risk dataset. [38, 39, 40, 41].

The overall accuracy for kNN algorithm is 80.50% with precision of 90.90%. While CART accuracy is 89.70% with precision of 93.82%. For Logistic Regression algorithm, the accuracy is 92.09% with precision of 93.18%. Lastly, AdaBoost with DT as subprocess gave us overall accuracy of 92.20% and a precision of 90.15%.

TABLE 9
PREDICTIVE MODEL PREDICTION PERFORMANCE OVERALL RESULT

Accuracy	Precision	Specificity	Recall or Sensitivity	f-Measure	False Positive	False Negative	True Positive	True Negative	
kNN	80.50%	90.90%	89.17%	68.99%	79.02%	15.4%	37.1%	99.6%	132.2%
CART	89.70%	93.82%	94.66%	92.41%	93.10%	8.7%	10.8%	131.5%	133.6%
Logistic Regression	92.09%	93.09%	93.18%	90.65%	91.83%	9.7%	13.3%	129%	132.6%
AdaBoost	92.20%	90.15%	89.60%	94.80%	92.41%	14.8%	7.4%	134.9%	127.5%

Figure 14 shows the result of "true AtRisk result" of the four-data mining predictive models. Logistic Regression gives the highest precision result of 92.66%. Although other models showed a good considerable result we consider the model with highest positive predictive value and sensitivity [42, 43, 44, 45]. Moreover, execution time of logistic regression model takes 9 seconds while AdaBoost model runs in 15 seconds. Among the predictive models, result indicates that logistic regression is the best classifier.

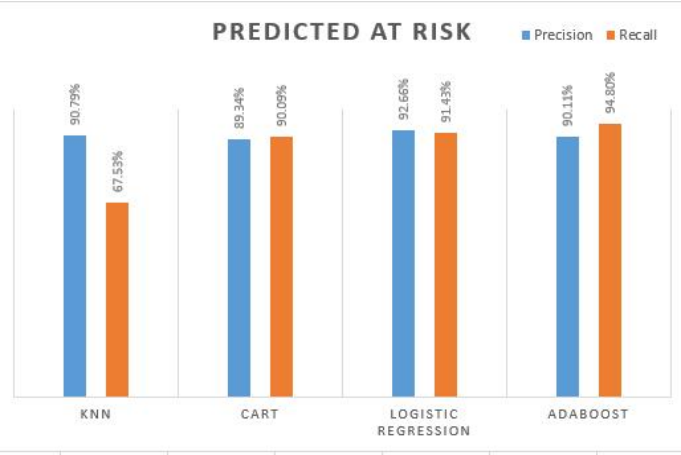

Fig. 16 . Result comparison for predictive models–True AtRisk

D. Scoring

In order to test the model in a more realistic way, the model is then applied to a previously unseen record as our test set. The data predicted 298 out of 316 data samples. That means 94.30% was precisely predicted using the model.

V. DISCUSSION & CONCLUSSION

In this paper, we attempted to determine whether there were relationships among at-risk students in terms of the attributes queried from the university database. We also tried to determine whether any of these attributes could predict at risk students [46, 47, 48]. Based upon that we found that past data of the students through educational data mining techniques can be utilized to develop predictive models.

As a result, having the information generated through our study: predicting at-risk students before they leave the institution is possible. This could help HEI to possibly intervene and prevent possible at-risk students from leaving.

We intend to extend the study in developing a system that will cluster students based on their courses as Arts or Science to find out further, if students who may be at risk in Science courses will not be at risk when they will shift to Arts courses.

ACKNOWLEDGMENT

We thank Prof. Ronald Silvosa, the Head of Information System Department and Prof. Maimona M. Asum, the Dean of the College of Information Technology in Mindanao State University – Marawi for giving us the data we used in realizing this study.

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2

REFERENCES

[1] M. R. Paraiso, H. V. Torres and A. A. Vinluan, "Data mining approach for analyzing graduating students' academic performance of new era university bachelor science in computer science," *International Journal of Conceptions on Computing and Information Technology,* vol. 3, no. 3, pp. 308-314, 2015.

[2] A. Padmapriya, "Prediction of higher education admissibility using classification algorithms," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, no. 11, pp. 330-336, 2012.

[3] National Center for Education Statistics, "The condition of education 2016 (NCES 2016-144)," 2016 [Online]. Available: goo.gl/6j7FYT

[4] M. Schneider, "Finishing the first lap: The cost of first year student attrition in America's four year colleges and universities," *American Institutes for Research,* vol. 12, no.3, pp. 23-27, 2014.
DOI: 10.1037/e537522012-001

[5] C. Goldin and L. F. Katz, "*The race between education and technology: The evolution of U.S. educational wage differentials 1890 to 2005*," National Bureau of Economic Research, Cambridge, MA, 2007.
DOI: 10.3386/w12984

[6] A. P. Carnevale, S. J. Rose and B. Chea, "*The college payoff: Education, occupations, lifetime earnings*," 2011 [Online]. Available: goo.gl/Cdmk3t

[7] National Center for Education Statistics, "Students whose parents did not go to college: Postsecondary access, persistence, and attainment," 2001 [Online]. Available: goo.gl/mTvn2G

[8] A. A. Azwa, N. H. Ismail, F. Ahmad and S. Fadhila, "Mining students' academic performance," *Journal of Theoretical and Applied Information Technology,* vol. 53, no. 3, pp. 306-310, 2013.

[9] B. O. Barefoot, J. N. Gardner, M. Cutright, L. V. Morris, C. C. Schroeder, S. W. Schwartz, M. J. Siegel and R. L. Swing, "*Achieving and sustaining institutional excellence for the first year of college*," San Francisco, CA: Jossey-Bass, 2005.

[10] L. A. Atkinson, "*Factors impacting student retention on the regional campuses and centers of Ohio university*," Ph.D. Dessertion, Department of Electical Engineering, Ohio university, Ohio, OH, 2008.

[11] R. L. Daft, M. Kendrick and N. Vershinina, " Management," *Cengage Learning EMEA*, vol. 2, no. 5, pp. 10-220, 2010.

[12] A. A. Owojori and T. O. Asaolu, "Critical evaluation of personnel management problems in the Nigerian school system," *International Journal of Educational Sciences*, vol. 2, no. 1, pp. 1-11, 2010.

[13] X. Liu, J. Wu, Z. Zhou, "Exploratory under sampling for class imbalance learning," *in Proceedings of the Sixth International Conference on Data Mining (ICDM)*, Hong Kong, Hong Kong, 2006.

[14] S. Kalmegh, "Analysis of wake Data mining algorithm reptree, simple cart and random tree for classification of Indian news," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 2, pp. 35-42, 2015.

[15] Y. Freund and R. E. Schapire, "A decision theoretic generalization of online learning and an application to boosting," *Journal of Computer and Systematic Sciences*, vol. 55, no. 1, pp. 119-139, 1997.
DOI: 10.1006/jcss.1997.1504

[16] S. T. Jishan, R. I. Rashu. N. Haque and R. M. Rahman, "Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decision Analytics,* vol. 2, no. 1, pp. 1–25, 2015.
DOI: 10.1186/s40165-014-0010-2

[17] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Economic Review,* vol. 10, no. 1, pp. 23-27, 2012.

[18] S. Natek and M. Zwilling, "Student data mining solution knowledge management system related to higher education institutions," *Expert Systems with Applications,* vol. 41, no. 14, pp. 6400-6407, 2014.
DOI: 10.1016/j.eswa.2014.04.024

[19] S. Moon and E. Sullivan, "High impact educational practices as promoting student retention and success," *in the Proceedings of the 9th Annual National Symposium*, Oklahoma, OK, 2013.

[20] Griffth University, "*Student retention strategy*," 2012 [Online]. Available: goo.gl/2yq9kT

[21] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and Information Technologies*, vol. 13, no. 13, pp. 61-72, 2013. **DOI:** 10.2478/cait-2013-0006

[22] R. Alkhasawneh and R. H. Hargraves, "Developing a hybrid model to predict student first year retention and academic success in STEM disciplines using neural networks," *Journal of STEM Education: Innovations and Research*, vol. 15, no. 3, pp. 35-42, 2014.

[23] M. A. Ushie, G. I. Onongha, E. O. Owolabi and J. O. Emeka, "Influence of family structure on students academic performance in agege local government area," *European Journal of Educational Studies*, vol. 4, no. 2, pp. 177-187, 2012.

[24] S. Herzog, "Measuring determinants of student return vs. dropout/stopout vs transfer: A first to second year analysis of new freshmen", *Research in Higher Education*, vol. 46, no. 8, pp. 883-928, 2005. **DOI:** 10.1007/s11162-005-6933-7

[25] R. Stinebrickner and T. Stinebrickner, "Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model," *Journal of Labor Economics,* vol. 32, no. 3, pp. 601-644. **DOI:** 10.1007/s11162-005-6933-7

[26] V. Tinto, " Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, vol. 45, no. 1, pp. 89-125, 1975. **DOI:** 10.2307/1170024

[27] T. M. Christian and M. Ayub, "Exploration of classification using N-tree for predicting students' Performance," *in Proceeding of Data and Software Engineering (ICODSE)*, Bandung, Indonesia, 2014. **DOI:** 10.1109/icodse.2014.7062654

[28] C. Romero, B. Cerezo and M. Sanchez-Santil´ lan´, "Clustering for improving educational process mining," *in Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, Indianapolis, IN*, 2014. **DOI:** 10.1145/2567574.2567604

[29] M. M. Breunig, H. P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying density based local outliers," *in Proceedings of the International Conference on Management of Data and Symposium on Principles of Database Systems*, Dallas, TX, 2000.

[30] J. David and S. B. S. Raja, "Prediction of learning disabilities in children: Development of a new algorithm in decision tree," *International Journal of Recent Advances in Engineering and Technology,* vol. 2, no. 5, p. 22-27, 2014. **DOI:** 10.1007/978-3-642-14493-6_55

[31] M. Hall, "Correlation-based feature selection for discrete and numeric slass machine learning," *in Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, 2000.

[32] A. Dobson and A. Barnett, "*Introduction to generalized linear models*," Boca Raton, FL: Chapman and Hall/CRC, 2008.

[33] J. Friedman, T. Hastie and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, p. 337-407, 2012. **DOI:** 10.1214/aos/1016120463

[34] S. Visa, B. Ramsay, A. L. Ralescu and E. Van Der Knaap, "Confusion matrix-based feature selection," *in Modern Artificial Intelligence and Cognitive Science*, Cincinnati, OH, 2012.

[35] T. Tamara, A. Singleton, D. Pope and D. Stanistreet, "Predicting students academic performance based on school and socio-economic characteristics," *Studies in Higher Education*, vol. 41, no. 18, pp. 1424-1446, 2016. **DOI:** 10.1080/03075079.2014.974528

[36] M. J. R. Quinlan, "Introduction of decision tree", *Journal of Machine Learning,* vol. 3, no. 5, pp. 81-106, 1986.

[37] N. Ugtakhbayar, B. Usukhbayar, S. H. Sodbileg and J. Nyamjav, "Detecting TCP based attacks using data mining algorithms," *International Journal of Technology and Engineering Studies*, vol. 2, no. 1, pp. 1-4, 2016. **DOI:** 10.20469/ijtes.2.40001-1

[38] T. Wang and A. Mitrovic, "Using neural networks to predict student's performance," *in proceedings of Computers in Education, International Conference on IEEE*, Auckland, Newzeland, 2002. **DOI:** 10.1109/cie.2002.1186127

[39] R. A. Thompson and B. L. Zamboanga, "Prior knowledge and its relevance to student achievement in introduction to psychology," *Teaching of Psychology*, vol. 30, no. 2, pp. 96-101, 2003. **DOI:** 10.1207/s15328023top3002_02

[40] V. Ramesh, P. Parkavi and K. Ramar, "Predicting student performance: A statistical and data mining approach," *International Journal of Computer Applications,* vol. 63, no. 8, pp. 35-39, 2013.

[41] J. Ruby and K. David, "Prediction accuracy of academic performance of students using different datasets with gigh influencing factors," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 5, no. 2, pp. 23-30, 2016.

[42] Z. J. Kovačić, "Early prediction of student success: Mining students enrolment data," *in Proceedings of Informing Science & IT Education Conference*, Cassino, Italy, 2010.

[43] M. Naseriparsa, A. Bidgoli and T. Varaee, "A hybrid feature selection method to improve performance of a group of classification algorithms," *International Journal of Computer Applications*, vol. 69, no. 17, pp. 111-116, 2013. **DOI:** 10.5120/12065-8172

[44] E. Uma, A. Kannan, "Self-aware message validating algorithm for preventing XML-based injection attacks." *International Journal of Technology and Engineering Studies*, vol. 2, no. 3, pp. 60-69, 2016. **DOI:** 10.20469/ijtes.2.40001-3

[45] C. L. S. Tablatin, F. F. Patacsil and P. V. Cenas, "Design and development of an information technology fundamentals multimedia courseware for dynamic learning environment," *Journal of Advances in Technology and Engi-*

neering Studies, vol. 2, no. 5, pp. 202-210, 2016. **DOI:** 10.20474/jater-2.6.5

[46] M. T. Dorak, "*Common concepts in statistics*," 2016 [Online]. Available: goo.gl/VVYptK

[47] Rapidminer, "*Rapidminer studio manual*", 2014 [Online]. Available: goo.gl/y88h5W

[48] S. S. Keerthi, O. Chapelle and D. De Coste, "Building support vector machines with reduced classifier complexity," *Journal of Machine Learning Research*, vol. 7, no. 8, pp. 1493-1514.

— This article does not have any appendix. —

ISSN: 2414-4592
DOI: 10.20474/jater-3.4.2