Utilizing Feature Selection in Identifying Predicting Factors of Student Retention

January D. Febro
Department of Information Technology
MSU–Iligan Institute of Technology, Iligan City, Philippines

Abstract—**Student retention is an important issue faced by Philippine higher education institutions. It is a key concern that needs to be addressed for the reason that the knowledge they gain can contribute to the economic and community development of the country aside from financial stability and employability. University databases contain substantial information that can be queried for knowledge discovery that will aid the retention of students. This work aims to analyze factors associated with student's success among first-year students through feature selection. This is a critical step prior to modelling in data mining, as a way to reduce computational process and improve prediction performance. In this work, filter methods are applied on datasets queried from university database. To demonstrate the applicability of this method as a pre-processing step prior to data modelling, predictive model is built using the selected dominant features. The accuracy result jumps to 92.09%. Also, through feature selection technique, it was revealed that post-admission variables are the dominant predictors. Recognizing these factors, the university could improve their intervention programs to help students retain and succeed. This only shows that doing feature selection is an important step that should be done prior to designing any predictive model.**

Keywords—Educational data mining; feature selection; data preprocessing; knowledge discovery; student retention

I. Introduction

Universities have continuously experience challenges in retaining students. Accordingly, about 40% of students in tertiary will not graduate on time [1]. This has been a pressing problem in universities around the world. As 'higher education enrolments have increased in recent decades, dropping out of university has become a common experience' [2]. Like in the Philippines, Commission on Higher Education (CHED) records show that there has been a 4.1 million to 3.6 million total number of dropped out students between academic year 2015-2016 and 2016-2017 [3]. Further, according to the survey, "only 23% of Filipinos finish college" [4]. Undergraduate college enrolments have grown increasingly but with less graduates. Yet, few researchers in Philippine-educational-community have addressed attrition and retention problems.

First-year is regarded in this study considering that it has high attrition rates [5]. It has been affirmed in the study of Garett, Bridgewater and Feinstein [6] that first year is vital in indicating academic success and considered very important at many educational institutions [7]. Thus, the assistance and monitoring of first-year students should be regarded because universities can respond to these students through intervention programs. According to Seidman [8], the "formula for student success is: Retention = Early Identification and Early Intensive Continuous Intervention".

Educational data mining (EDM) can be used to resolve this student retention problem. EDM 'refers to a method for extracting information from large collection of data in educational institutions through data mining (DM) techniques to extract useful knowledge to help decision makers' [12]. Records of students can be queried as an attribute dataset, such as admission test scores and socio-demographic attributes. These can be utilized as predictors for the prediction model for knowledge discovery in databases (KDD).

The two of the three most popular model used in extracting knowledge from data are KDD process model (shown in Fig. 1) and Cross-Industry Standard Process for Data Mining (CRISP-DM) model [9]. Both models contain data preprocessing phase, which is crucial and tedious. In fact, performing the tasks in this phase can consume considerable amount of time. This includes data cleaning, data transformation, and data reduction. An overview of common DM preprocessing steps will be discussed in details in the succeeding section.

However, this paper will only use filter selection feature methods: Correlation Feature Selection, Information Gain Ratio, and Chi-Square analysis. To sought if the results of these selection methods will vary, it will be tabulated and ranked according to feature importance, and will be compared.

This study also aims to cite evidence in support of feature selection method as part of preprocessing step to increase the classification accuracy of a predictive model which has been omitted in some DM studies; like in the following similar studies [10],[11], and [12]. In view of this, two predictive models using classification technique with different feature datasets is proposed– model 1 will used all the dataset attributes queried from the university database and model 2 will used the ranking of important features. Moreover, feature selection method in this study is utilized to identify the possible factors instrumental to student retention and as part of data reduction phase. The significance of this result affects the student and society, along with financial consequences for the institution.

The structure of this paper is as follows: Section 2 reviews some similar works of this study and presents feature selection methods used in this study. Then in Section 3, presents the

The author has been supported by the Commission on Higher Education, Individual Research Grants (CHED M.O. No. 14-2016).

methodology while Section 4 discusses the results. Finally, Section 5 provides conclusion and future work.

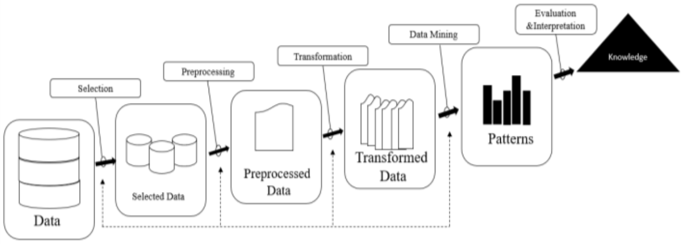


Fig. 1. KDD Process Model [8].

II. Literature Review

Because of the importance of student success in any universities around the world, there were a number of studies in EDM that studied student retention, attrition or academic performance. Most works presented were developing predictive models. The usual conceptual framework for EDM is shown in Fig. 2. Generally, student data were large datasets collected from university databases and normally, data attributes is selected manually.

A. Preprocessing Techniques

Data preprocessing steps is one of the major activities to perform to turn the collected data in an appropriate format for DM algorithms. Its purpose is to remove noise, handle missing data, normalize, select attributes or features, discretize and reduce dimensionality. This section will focus more on the feature selection methods while data cleaning, data transformation and data reduction are discussed briefly below, based on the review of [13].

1) Data cleaning. Row data can have incomplete or irrelevant records that needs to be done. To handle missing data, it can be ignored for large datasets but for small datasets values must be fill manually, fill with global constants or probable value. For noisy data, it can be handled through binning, regression or clustering methods.

2) Data transformation. Data are transformed to suitably fit DM algorithms. This may involve normalization, smoothing, aggregation and generalization.

3) Data reduction. To improve efficiency in large amount of data, data reduction technique is utilized. Certain steps may be performed, these are: data cube aggregation, attribute subset selection or feature selection, numerosity reduction, and dimensionality reduction.

And most frequently, feature selection is disregarded in data preprocessing phase despite that it is a substantial technique for it has been very effective to improve accuracy results. In the following studies presented, the researchers did not perform feature selection as part of data pre-processing step prior to modelling.

In the [11] study, the researchers used the dataset from Prince of Songkla University to predict dropout. They had collected four academic year of student data from Faculty of Science. Tree model and Rule-induction was used and compared in creating the model. The parameters used were pre-academic data and GPA. The model JRip rule induction has the highest accuracy result of 77.30%. Data transformation and Data cleaning were the only pre-processing steps made.

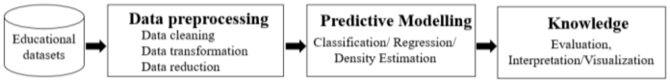


Fig. 2. A Typical EDM Conceptual Framework.

In the study of [12], researcher collected an 8-year-period record that contains student demographics, family background information and academic records for predicting academic performance of the Computer Science students. They used DT, Naïve Bayes and Rule based classification to create a predictive model. The best model in their study was Rule Based with 71.3% accuracy value. In their study, preprocessing steps performed were Data Selection, Data Cleaning and Normalization. The selected nine parameters on the study were based on their literature review.

In the study of [13], the researchers used regression in analyzing the academic performance using the academic subject data of the graduating students of Computer Science students of New Era University and in calculating accuracy they used Mean Absolute Percentage Error. In their study, they performed Data Selection, Data Cleansing and Data Transformation as a preprocessing step. The factors selected were the course subjects of students and the GPA.

B. Feature Selection Methods (FSM)

FSM has been proven to reduce dimensionality, remove noise and unimportant data from thus improving accuracy result [14]. Moreover, feature selection is substantial for data mining algorithms for many reasons such as generalization performance, running time requirements and constraints. There are three types exists for machine learning these are, filters, wrappers and embedded. The difference between filter method and wrapper methods is that the first calculates the number of features based on the common features of the data utilizing heuristics while the latter evaluate the number of features employing the learning algorithm [15]. Embedded methods on the other hand searches the best subset of parameters that is that is embedded in the classifier construction [16].

As a pre-processing step to DM, filter method has an advantage, for example filters execute faster than wrappers and it does not need re-execution on different learning algorithms. Hence, this study is focused on the filter method specifically Correlation Features Selection, Chi-Square Analysis, and Gain Ratio.

C. Correlation-based Feature Selection (CFS)

CFS, "ranks feature subsets as per correlation based heuristic evaluation function in which numeric features are first discretized to gauge correlation between nominal features". This algorithm seeks for features that are particularly correlated with the explicit class [15]. CFS formula is given in equation (1):

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \qquad (1)$$

where r_{zc} in the equation is the interrelatedness of scored features, k is the sum of features, r_{zi} is the mean of the correlations relating to the class variable, and is the mean of inter-correlation between features [15].

D. Chi-Squared

Chi-squared is a frequently used method for feature assessment by calculating the value of chi-squared statistic with regard to the class [17]. The formula is provided below.

$$X^2 = \frac{\sum (o-e)^2}{e} \tag{2}$$

where o is observed frequencies and e is expected frequencies. This method is used to identify whether a distribution of observed frequencies varies from the supposed expected frequencies.

E. Informaton Gain Ratio (IGR)

IGR method computes the importance of the features using information gain and give weights to them accordingly even if it applied to features that have dissimilar value using the equation below [18].

$$GR(att) = \frac{IG(att)}{H(att)} \tag{3}$$

where equation (4)

$$H(att) = \sum_j - P(v_j) \log_2 P(v_j) \tag{4}$$

where $P(v_j)$ corresponds to the chances of having v_j by providing general values for an attribute j.

III. Methodology

Fig. 3 illustrates the activity in this study. In the data pre-processing feature selection is emphasized.

A. Dataset Collected

The records used in this study were real records of five academic years queried from a university database. These records contain information about the entrance result, grades, and among others. The data for this research was inputted in a data mining tool. The dataset is comprised of 7, 936 records with 29 features.

The potential predictor variables queried fall into two categories: pre-college data and post-admission data. Pre-college data are records prior to admission, it includes admission test scores and socio- demographic attributes. The pre-college dataset features examined in this study is grouped into two: 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) and academic potential (admission test score in Math, Language Usage, Aptitude, and Science). On the other hand, post-admission data are educational achievement indicators such as course, scholarship status, grades in Math and English subjects, and grade point average of first semester.

B. Data Pre-Processing

Prior to modelling and to improve the input data quality and suitability, data pre-processing is needed. For this study, identifying noise data, missing values, irrelevant and redundant data and removing outliers were crucial steps. Data cleaning is done using a data mining tool.

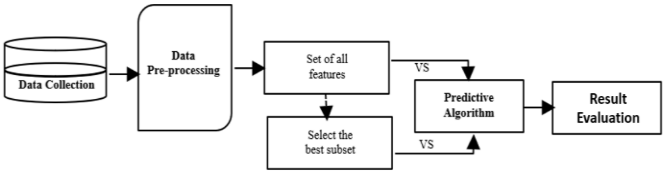


Fig. 3. Methodology Flow.

To remove the irrelevant data and noise from the dataset, the following steps were carried out.

1) Load data collected
2) Integrate collected data
3) Filter data by removing missing records
4) Remove duplicates
5) Do normalization
6) Detect Outlier

Careful data integration is done to reduce and avoid redundancies and inconsistencies. Redundant data were carefully examined; same attributes were not included in this study.

Data cleansing steps is performed to remove the incomplete data. A list-wise deletion method is adopted to delete the entire record from the analysis if any variable in the model has a missing value. Missing data is ignored to avoid adding bias and distortion to the dataset. Removing a few records will not impede the results of the model since this study contains large dataset. Finally, to handle outliers, local outlier factor (LOF) is executed.

C. Feature Selection

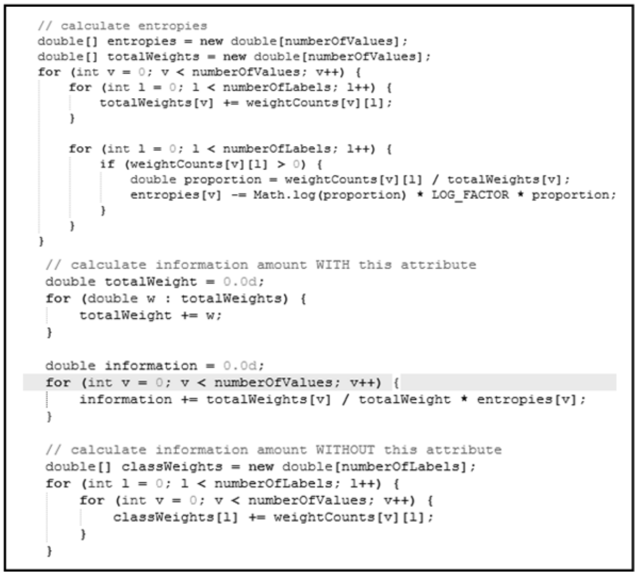
One of the main goals in this study is to identify what dominant variable or combination of variables collected can be used as predictors of first year student success. In this study, filter model using feature rankings are used, namely, the Info Gain Ratio, the Correlation Feature Selection, and the Chi Square, to identify the dominant variables. The significance of using filter model method is that it separates feature selection from learning [19]. Thus, no bias towards any learning algorithm.

During the feature selection process, no specific form of relationship is assumed. The outcome of the feature selection is list of predictors ranked according to their importance.

1) Information Gain Ratio (IGR): The first FSM employed is the IGR. It calculates the entropies in class and resolves the vulnerability of IG. Fig. 4 shows the code snippet of the method used in this study.

2) Correlation feature selection: CFS finds attribute that are highly related with the specific groups but still have at least inter-correlation amongst the attributes themselves. Fig. 5 shows the code snippet of the method used in this study.

3) Chi-Square: The chi-square statistics use nominal data and is utilizes to identify if a distribution of the stated frequencies varies from the actual expected frequencies. Fig. 6 shows the code snippet of the method used in this study.



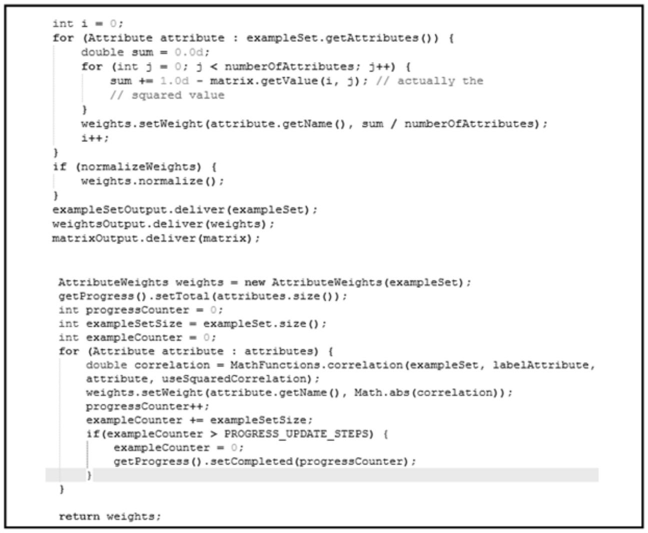
```java
// calculate entropies
double[] entropies = new double[numberOfValues];
double[] totalWeights = new double[numberOfValues];
for (int v = 0; v < numberOfValues; v++) {
    for (int l = 0; l < numberOfLabels; l++) {
        totalWeights[v] += weightCounts[v][l];
    }

    for (int l = 0; l < numberOfLabels; l++) {
        if (weightCounts[v][l] > 0) {
            double proportion = weightCounts[v][l] / totalWeights[v];
            entropies[v] -= Math.log(proportion) * LOG_FACTOR * proportion;
        }
    }
}
 // calculate information amount WITH this attribute
 double totalWeight = 0.0d;
 for (double w : totalWeights) {
     totalWeight += w;
 }

 double information = 0.0d;
 for (int v = 0; v < numberOfValues; v++) {
     information += totalWeights[v] / totalWeight * entropies[v];
 }

 // calculate information amount WITHOUT this attribute
 double[] classWeights = new double[numberOfLabels];
 for (int l = 0; l < numberOfLabels; l++) {
     for (int v = 0; v < numberOfValues; v++) {
         classWeights[l] += weightCounts[v][l];
     }
 }
```

Fig. 4. IGR Code Snippet.

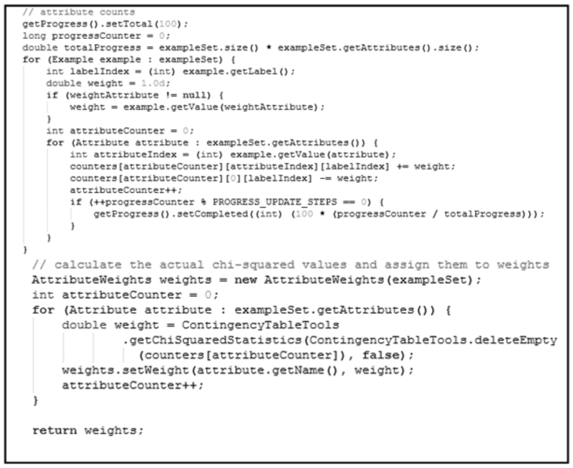


```java
int i = 0;
for (Attribute attribute : exampleSet.getAttributes()) {
    double sum = 0.0d;
    for (int j = 0; j < numberOfAttributes; j++) {
        sum += 1.0d - matrix.getValue(i, j); // actually the
        // squared value
    }
    weights.setWeight(attribute.getName(), sum / numberOfAttributes);
    i++;
}
if (normalizeWeights) {
    weights.normalize();
}
exampleSetOutput.deliver(exampleSet);
weightsOutput.deliver(weights);
matrixOutput.deliver(matrix);


 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. 5. CFS Code Snippet.



```java
// attribute counts
getProgress().setTotal(100);
long progressCounter = 0;
double totalProgress = exampleSet.size() * exampleSet.getAttributes().size();
for (Example example : exampleSet) {
    int labelIndex = (int) example.getLabel();
    double weight = 1.0d;
    if (weightAttribute != null) {
        weight = example.getValue(weightAttribute);
    }
    int attributeCounter = 0;
    for (Attribute attribute : exampleSet.getAttributes()) {
        int attributeIndex = (int) example.getValue(attribute);
        counters[attributeCounter][attributeIndex][labelIndex] += weight;
        counters[attributeCounter][0][labelIndex] -= weight;
        attributeCounter++;
        if (++progressCounter % PROGRESS_UPDATE_STEPS == 0) {
            getProgress().setCompleted((int) (100 * (progressCounter / totalProgress)));
        }
    }
}
 // calculate the actual chi-squared values and assign them to weights
 AttributeWeights weights = new AttributeWeights(exampleSet);
 int attributeCounter = 0;
 for (Attribute attribute : exampleSet.getAttributes()) {
     double weight = ContingencyTableTools
             .getChiSquaredStatistics(ContingencyTableTools.deleteEmpty
               (counters[attributeCounter]), false);
     weights.setWeight(attribute.getName(), weight);
     attributeCounter++;
 }

 return weights;
```

Fig. 6. Chi-Square Code Snippet.

D. *Data Modelling*

A prediction model for EDM can be developed using EDM techniques but will heavily depend on the type of datasets. In this study, logistic regression method is used.

The dataset is partitioned into training and validation subsets. Two predictive models were created, for the first model all the features will be inputted. On the second model, only the significant variables assessed by feature selection techniques were the final parameters in creating the model. 70% of the dataset is used in training and the remaining 30% is used as a test-set for both models and are tested for accuracy using 10-fold cross-validation.

E. *Result Evaluation*

The performance of the two models is evaluated by its accuracy and precision which are computed using the equation below.

$$Accuracy = \frac{\text{True Positive+True Negative}}{\text{True Positive+True Negative+False Positive+False Negative}} \quad (5)$$

The accuracy is computed by the actual instance of correct classification (True Positive + True Negative) over the total instances of that class.

$$Precision = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad (6)$$

Precision is computed by the positive predicted instances over the total predicted instances.

IV. Results and Discussions

A. *Results of Feature Selection*

Fig. 7 shows the result of IGR. The result is based on the upmost Gain ratio ranked by their importance. Any information gain above zero shows some type of significance. Factors like English status, Math status, family income and college entrance score for language usage, math, aptitude and science largely influence the result of student’s retention.

Fig. 8 shows the features and Correlation-based Feature Selection scores ranked in ascending order of importance.

Among the highest ranked by CFS are English status, gross income and math status.

Fig. 9 results show features that were highly influential or with high chi-square values. These values are displayed in ascending order.

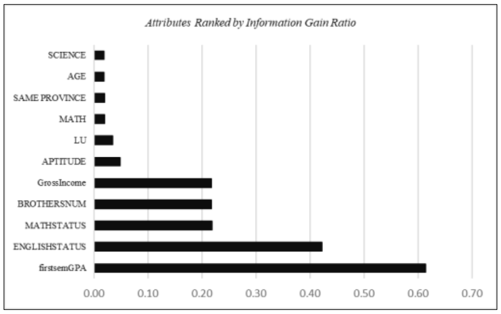


Fig. 7. Attributes Ranked by Information Gain Ratio.

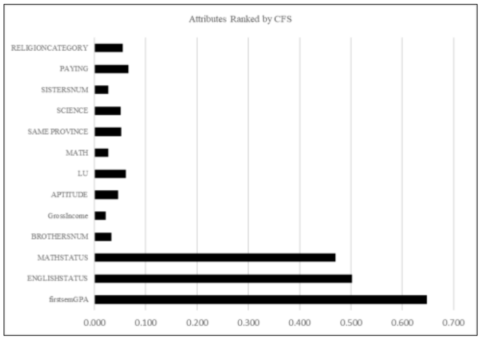


Fig. 8. Attributes Ranked by CFS.

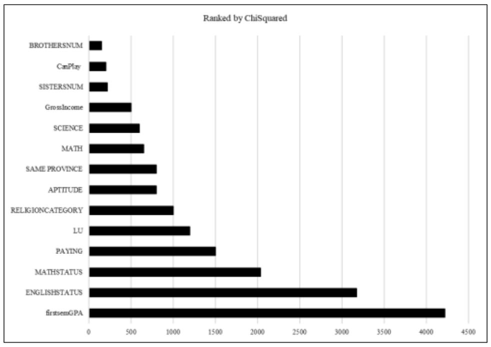


Fig. 9. Attributes Ranked by ChiSquared.

From the results, it can be concluded what dominant predictors affects student retention. These predictors in Table I are utilized in building a predictive model. From 29 possible predictor variables, only 14 predictor variables are used in the second model.

B. Modelling

In this study, logistic regression is used. As mention, the Model 1 used all 29 predictor variables and the Model 2 used only 14 predictor variables which were ranked by the filter model during the selection feature analysis. Both models tested for accuracy using 10-fold cross-validation.

TABLE. I. RANKED OF PREDICTOR VARIABLES IN ASCENDING ORDER

CFS	Info Gain Ratio	ChiSquared
firstsemGPA	firstsemGPA	firstsemGPA
ENGLISHSTATUS	ENGLISHSTATUS	ENGLISHSTATUS
MATHSTATUS	MATHSTATUS	MATHSTATUS
PAYING	BROTHERSNUM	PAYING
LU	GrossIncome	LU
RELIGIONCATEGORY	APTITUDE	RELIGIONCATEGORY
SAME PROVINCE	LU	SAME PROVINCE
SCIENCE	MATH	APTITUDE
APTITUDE	SAME PROVINCE	SCIENCE
BROTHERSNUM	SCIENCE	MATH
MATH	SISTERSNUM	GrossIncome
SISTERSNUM	RELIGIONCATEGORY	CanPlay
GrossIncome	PAYING	SISTERSNUM
CanPlay	CanPlay	BROTHERSNUM

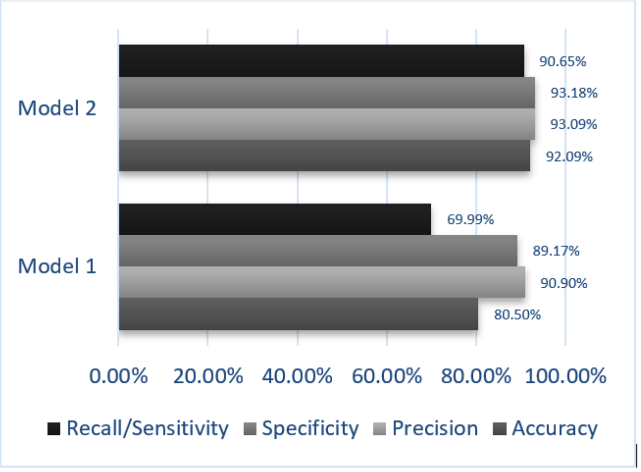


Fig. 10. Result Comparison of the Two Models.

Fig. 10 shows the results of the two predictive models. The results presented, indicates that the accuracy result of model 2 jumps to 92.09% from 80.50%. This only tells that doing feature selection is vital as part of preprocessing.

V. CONCLUSION AND FUTURE WORK

Early detection of potential student leavers is favorable for both students and institutions. In this paper, 14 features from 29 predictor variables have identified to have importance by performing filter model FSM. Based on the feature selection result, it was found that aside from first semester gap, students retaining in university was positively correlated with the following predictors, namely, college entrance exam score (math, language usage, aptitude and science category), number of siblings, family income, English grade, and math grade. The generated information will be quite useful for the university management to develop policies and strategies for better planning and implementation to increase the retention rate in HEIs.

In future, the study can be enhanced by applying few hybrid feature selection algorithms on student datasets in order to predict student retention. A web-based system will be developed that helps to monitor students and accurately predict student retention and attrition.

ACKNOWLEDGMENT

The author would like to thank the Commission on Higher Education (CHED) for supporting this study under the Individual Research Grants (CHED M.O. No. 14-2016).

REFERENCES

[1] National Center for Education Statistics, “The Condition of Education 2016.” (NCES 2016-144), Undergraduate Retention and Graduation Rates, 2016.

[2] A. Norton, and I. Cherastidtham, “Dropping out: the benefits and cost of trying university”, Grattan Institute, 2018.

[3] Commission on Higher Education (CHED), “2018 Higher Education Facts and Figures,” 2018.

[4] Philippine News Agency, “Only 23% of Filipinos finish college,” BusinessMirror, (April 27, 2017).

[5] Australian Government Department of Education and Training, “Improving retention completion and success in higher education,” Higher Education Standards Panel Discussion Paper, June 2017.

[6] N. Garett, M. Bridgewater and B. Feinstein, “How Student Performance in First-Year Composition Predicts Retention and Overall Student Success,” Retention, Persistence, and Writing Programs, Louisville, CO: University Press of Colorado, 2017.

[7] P. Van der Zanden, E. Denessen, A. Cillesen and P. Meijer, “Domains and predictors of first-year student success: A systematic review,” Educational Research Review, 23 57-77, 2018.

[8] A. Seidman, “College student retention: formula for student success,” Westport, CT: ACE/Praeger, 2005.

[9] U. Shafique, and H. Qaiser, “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA),” International Journal of Innovation and Scientific Research, Vol. 12 No. 1 Nov. 2014.

[10] P. Ramya, K. Gudlavalleru and M. Kumar, “Student Performance Analysis Using Educational Data Mining,” International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, 2016.

[11] J. Pattanaphanchai, K. Leelerpanyakul, & N. Theppalak, ”The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining: A Case Study of Faculty of Science, Prince of Songkla University,” Journal of University of Babylon for Pure and Applied Sciences, Vol.(27), No.(1): 2019.

[12] F. Ahmad, N. Ismail, and A. Aziz, “The Prediction of Students’ Academic Performance Using Classification Data Mining Techniques,” Applied Mathematical Sciences, vol. 9, no. 129, pp. 6415-6426, 2015.

[13] W. Bhaya, “Review of Data Preprocessing Techniques in Datamining,” Journal of Engineering and Applied Sciences, 12 (16): 4102-4107, 2017.

[14] A. Algarni, “Data Mining in Education,” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016.

[15] M. Paraiso, H. Torres, et al., “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. Issue 3, 2015.

[16] M. Hall, and L. Smith, “Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper,” Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, 1999.

[17] Sheena, K. Kumar, and G. Kumar, “Analysis of Feature Selection Techniques: A Data Mining Approach,” International Journal of Computer Applications, 2016.

[18] J. Novaković, P. Strbac, and D. Bulatović, “Toward optimal feature selection using ranking methods and classification algorithms,” Yugoslav Journal of Operations Research 21, 2011.

[19] M. Trabelsi, N. Meddouri, and M. Maddouri, “A New Feature Selection Method for Nominal Classifier based on Formal Concept Analysis,” Procedia Computer Science, 2017.