I.J. Modern Education and Computer Science, **2021, 3, 57-67**
Published Online June 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2021.03.05

Predicting Student Program Completion Using Naïve Bayes Classification Algorithm

Joann Galopo Perez
Bulacan State University, Philippines
Email: joann.galopo@bulsu.edu.ph

Eugene S. Perez
Bulacan State University, Philippines
Email: eugene.perez@bulsu.edu.ph

Received: 25 January 2021; Revised: 23 February 2021; Accepted: 14 March 2021; Published: 08 June 2021

Abstract: Data mining approaches provide different educational institutions opportunities to find hidden patterns from the data stored in the database. Many researchers have used these data to develop a model that would assist the institution administrators in decision-making. This study was performed to predict student program completion using the Naïve Bayes classifier technique. The dataset utilized in this study was obtained from Bulacan State University – Sarmiento Campus in the Philippines under BS Information Technology program from five-year graduates' data for Academic Year 2012-2016. This dataset was pre-processed, cleansed, transformed, and balanced before constructing the model. Ten predictors were used for predicting student completion. The feature selection technique was used to filter and evaluate the significance of each factor. The significant variables assessed by the feature selection technique (Weight by Correlation) were the final parameters in creating the model. The Naïve Bayes classifier was applied to predict the students' completion using the 70:30 ratios for training and testing dataset distribution. Correlation analysis identified the weight of individual attributes to the label attribute. From 10 possible predictor variables, only four (4) predictor variables were selected after correlation analysis. The identified significant attributes affecting program completion, namely (in order of significance): parents' monthly income, mother and father's educational attainment, and High School GPA attributes. The significant attributes identified in correlation analysis splitted into 70% training data or 447 records and 30% testing data or 191 records. There were 84 out of 191 data samples, or 44% of students were predicted to complete the program. On the other hand, 107 out of 191 data samples, or 56%, were predicted as not completing the program. The accuracy values performed an 84% rating with 80.46% class precision, and 83.33% class recall in the testing dataset (n=191). The outcomes of this study have a significant impact on HEIs, particularly on college completion rates. This study shall be highly significant and beneficial specifically to university administrators as this be a tool for them to identify students who will complete college based on variables included in the model.

Index Terms: Data Mining, Naïve Bayes Classification Algorithm, Predictive Model, and Program Completion

1. Introduction

One of the pressing and ongoing concerns for Higher Education Institutions (HEIs) is students' retention [1]. This undertaking is crucial and essential for an academic institution to highlight and focus on upholding its existence. Higher education institutions use learning analytics to enhance the programs they deliver and meet visible and achievable objectives such as grade point average and retention [2]. It is vital for the university to effectively make strategic adjustments leading to a high retention rate and program completion among its clientele - the students.

It is of interest to address the students' completion records of Bulacan State University (BSU) Sarmiento Campus under the Bachelor of Science in Information Technology (BSIT) program.

This study focused on developing a program completion model based on the graduate data of BSIT students from 2012 to 2016 through the classification process of organizing data into categories. Since the BSIT program takes four years to finish, the researchers retrieved the records of the students who enrolled from 2008 to 2012 from the registrar's office. These records were treated with the utmost security, considering that the final result will not contain any identifiable information to protect students' privacy. The names of the students were also removed before handed to the researchers.

To have the most relevant classifier and an effective and efficient model, the researchers utilized the Naïve Bayes Classification Algorithm, a simple technique for constructing classifiers or models based on Bayes Theorem of

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

conditional probability and strong independence assumptions. Many researchers have previously used it and observed that among other classification approaches, Naïve Bayes performs well [2, 3, 4, 5, 6, 7].

While existing studies on retention and attrition [8, 9, 10, 11, 12] have explored various attributes to affect completion rates, no study has integrated the specific attributes to be used; hence, the researchers employ all available attributes in the students' records. With the use of a more robust predictive model, while considering relevant factors, the entire procedure could identify the students who need help at the very early stage or even before a student realizes he or she is in trouble. Consequently, the institution can intercede, postulate specific support, and identify relevant resources that the students need to continue academically. Moreover, the school administrators could devise specific decision-making schemes for on-time program completion, thereby improving the retention rate, particularly among BSIT students at BSU Sarmiento Campus.

The data mining implementation and processing in this study was done using RapidMiner Studio, which offered numerous data mining methods such as data cleansing, data transformation, optimization, validation, and visualization [12]. Split Data operator in RapidMiner randomly split up the dataset into a training set (70%) and test set (30%). This study did not create an actual system that automatically identified students who will complete college. Instead, it developed or layout the model or a decision support system that will assist in readily identifying the possibility of completion of who will complete college or not.

2. Related Works

The literature of this paper, which served as references in the completion of this research, concentrated on predictive modeling using data mining, particularly the application of Naïve Bayes. It also presented the student-related attributes that contributed to student college completion based on research to select model building variables.

A. Application of Naïve Bayes Algorithm

According to [13], Naive Bayes is a basic probabilistic classifier based on the Bayesian theorem. It is based on the data set's independent functions or features. The existence or absence of any other aspect has no bearing on the attribute of a class. The key advantage of this classifier is that it takes little training data to measure the mean and variance of the variable for classification.

Several researchers have used Naive Bayes techniques to predict student performance in the field of Educational Data Mining, and it has also been applied to other fields. In their paper [14] it aims to predict the student's success using the concept of mining methods. This paper compared the percentage of accuracy with various data mining approaches such as Decision Tree, Neural Network, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine. From the result obtained, the Decision Tree and Neural Network are the most accurate of these approaches.

In this paper [15], three supervised data mining algorithms were applied to preoperative evaluation data to predict course completion (either passed or failed). The learning methods' performance was assessed based on their predictive accuracy, ease of learning, and user-friendly characteristics. The findings revealed that the Naive Bayes classifier outperforms decision tree and neural network approaches.

Research conducted by [16] on how data can be preprocessed using the Optimal Equal Width Binning discretization approach and the Synthetic Minority Over-Sampling (SMOTE) over-sampling methodology to increase the precision of the students' final grade prediction model for a specific course. It has been concluded in their study that the Naive Bayes classification algorithm is computationally faster than the Neural Network Backpropagation algorithm, making it the best option.

A study proposed by [17] states a framework for predicting first-year bachelor students' academic success in a Computer Science course. The variables used in the study included the students' demographics, past academic results, and family history information. To generate the best academic performance prediction model for students, Decision Tree, Naive Bayes, and Rule-Based classification techniques are applied to their results.

A study performed by [18] compared Bayes Net, Naive Bayes, and a hybrid of these two classifiers to see which one will yield the better results using diabetic patients' data accessible in the WEKA tool. The results suggest that combining Bayes Net and Naive Bayes produces better results than using these classifiers independently.

Another research [19] used the naive Bayesian classification to construct a learner classifier. It has been concluded that Bayesian classification theories, in general, are useful analysis forms for forecasting potential data patterns and making wise decisions in the distance education system.

This paper [20] provides an overview of the decision tree and the Naive Bayes classification. The research utilized the student's output classification dataset to demonstrate how a decision tree and naive can be constructed and the principle of both classification strategies. The analysis concludes that the Naive Bayes classifier is a basic algorithm that provides more reliable results, as shown in this paper. As extended to large data sets, the Naive Bayes algorithm outperforms Decision Trees.

Also, [21] performed an analysis in which they used Naive Bayes to forecast the viability of early-stage foreign development projects and found good projects that could impact its financial stability. Another study [22] used the Naive Bayes machine learning classifier to build a Feedback Validation Model in Education data mining. The model's

probabilistic approach will easily unearth the explanations for the student's academic success in the teaching-learning process individually. Another study [23] presents a method for predicting sales agents' success in a call center devoted solely to sales and telemarketing operations. A naive Bayesian classifier is used in this technique. The goal is to assess which levels of the attributes are representative of individuals that perform well.

Furthermore, [24] also presented a survey evaluating and investigating students' placement after doing MCA applying the three selected classification algorithms using the Weka tool. As compared to other classifiers, the naive Bayes classifier has the lowest average error at 0.28. These findings indicate that among the machine-learning algorithms studied, and the Naive Bayes classifier will vastly outperform traditional classification methods for placement. The current study used a Naïve Bayes classifier to develop the proposed prediction model.

The Naive Bayes Classifier outperforms all other algorithms presented in the literature. Aside from the result, there are many reasons why the Naive Bayes Classifier was chosen for this study. Naïve Bayes is quick and can be trained on all of the training data. It has a sound and straightforward foundation for modeling the data and is quite robust to outliers and missing values. A complex resulting classifier can be determined reliably from a limited amount of data. The Naive Bayes classifier is easy to introduce, and it is fast and can be used in real-time [25].

B. Factors that influence student performance

Many studies in the literature on various factors such as personal, socioeconomic, psychological, and other environmental variables affect students' success and the models used to predict performance. For further reference, a few basic studies are mentioned below.

This paper aims to define the factors affecting students' success in final exams and develop a suitable data mining algorithm to predict students' grades to provide a timely and accurate warning to students at risk. The findings showed that the type of school has little effect on student success and that parents' occupation plays a significant role in predicting grades [26]. Also, predictor variables as a basis for student completion included demographics such as gender, age, family background, and external assessments [27].

Features such as family and community background including socioeconomic status, parent educational level, and the size of the student's high school and hometown community. Socioeconomic status, often measured by family income, had been positively correlated with college persistence [28]. In a study conducted by [29] suggested factors like family background and family income should be considered essential attributes in this study. Since family background played an important role in student academic performance, it is found that students from a family with less affluent economic background have more chance of being unsuccessful in a course due to their other involvements to support the family and a higher tendency to discontinue education in the middle. Factors like high school GPA and socioeconomic status [30], [31] were utilized in the previous researches study, and also, residence and first-semester GPA were predictors of whether students will drop out or continue to program completion [32].

According to research conducted by [33] using 300 samples of students from 5 different degree colleges in India, there is a relationship between the attributes students' academic success with their grades in the senior secondary test, living place, the medium of teaching, mother's qualification, annual family income, and family status. Moreover, another study explored new factors associated with college success, including parents' education level and annual household income [34] .

This research [35] examined demographic variables, family backgrounds, pre-college and college academic success indicators, and the degree to which required placement in remedial courses predicts persistence at a public research institution. The findings revealed that High school GPA and first-semester college GPA were discovered as significant factors to college success. Another study by [36] also revealed that high school grades are a significant factor for college readiness than test scores.

3. Methods

The framework of the study adopts the Knowledge Discovery Process (KDP) as suggested by [37]. The KDP process scheme consists of Selection, Preprocessing, Transformation, Data Mining, and Interpretation. Input data were initially selected, and the target data were isolated. Pre-processing and transformation were performed to ensure the database reliability where data mining was the core analysis. The knowledge discovery process ended with interpreting the results. Based on this KDP scheme, this study came up with a framework, as shown in Figure 1. These processes are *Phase 1: Preparation of Data* – which involved data collection and production of dataset retrieved from the university student database; *Phase 2: Preprocessing* – which involved the removal of noisy and irrelevant data and further transformed into forms appropriate for mining; *Phase 3: Classification Process* – which involved model building where Naïve Bayes algorithm was applied to predict student program completion; *Phase 4: Result* – which involved interpretation and evaluation of the data and also identification and presentation of interesting patterns representing knowledge-based to the user.

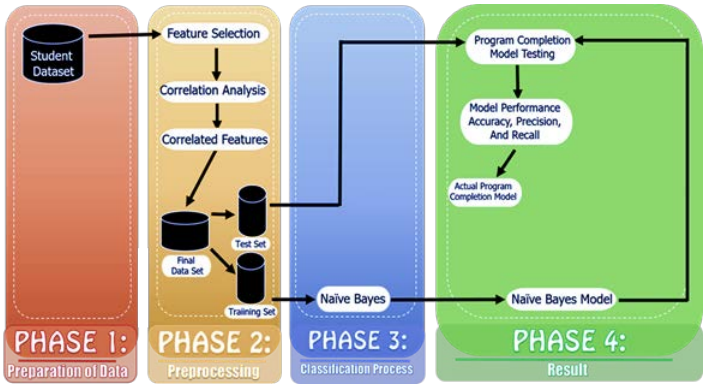

Fig. 1. Proposed Framework of the study

Phase 1: Preparation of Data

Comprehensive research of the literature and consultation with experts on student completion revealed various aspects that are believed to influence a student's success. In this study, the program completion is predicted based on the datasets of the graduates from 2012 to 2016. The framework of this study is presented in Figure 1. The dataset contained 768 records and eleven (11) attributes. The predictor variables in the dataset were included and were retrieved from the university registrar. The variables selected provided a significant impact on the studies presented in the related works. The list of attributes in the student dataset that originated from the registrar was presented in Table 1.

Table 1. Student List of Attributes

Attributes	Description/Assigned Values
Age during the student admission	Age admitted to the university.
Gender	Student's gender. It is the category of the student, whether male or female.
High School Type	Public/Private. Type of High School
High School GPA	High School Grade Point Average. It is the general weighted average of the student in the last year in high school.
High School Math Average	4th Year High School Mathematics average
1st-year college 1st semester GPA	Grade Point Average in the first-year first semester. This is the average grade of a student while in the first-year first semester. The values range from 1.0 to 5.0, where 1.0 is the highest grade a student can get and 5.0 is the lowest grade.
1st-year college 2nd semester GPA	Grade Point Average in first year second semester. This is the average grade of the student while in the first year second semester. The value ranges from 1.0 to 5.0, where 1.0 is the highest grade a student can get and 5.0 is the lowest grade.
Parents' Monthly Income	Monthly income of parents 1 '5000 or below' 2 '5001-10,000' 3 '10,001-15,000' 4 '15,001-20,000' 5 '20,001-25,000' 6 '25,001 or above'
Father's Educational Attainment	Highest educational attainment of the father 1 'Less than High School' 2 'High School Graduate' 3 'Two-Year Associate Degree' 4 'Four-Year Bachelor's Degree' 5 'Master's Degree' 6 'Doctorate Degree'
Mother's Educational Attainment	Highest educational attainment of the mother 1 'Less than High School' 2 'High School Graduate' 3 'Two-Year Associate Degree' 4 'Four-Year Bachelor's Degree' 5 'Master's Degree' 6 'Doctorate Degree'
Completed	The student completed the BSIT program in four (4) years (Yes/No)

Phase 2: Pre-processing

Figure 2 displayed the workflow of the pre-processing phase. It contained operators that were used to develop the model from retrieving the student dataset; a *Set Role* operator was applied to determine the attribute label to be predicted. Initially, the dataset was in an excel format that contained 768 records. In data pre-processing, the dataset was checked and cleaned to reduce the unnecessary attributes.

Fig. 2. Pre-processing

a) Data Balancing

Fig. 3. Data balancing process

The *Sample* method in Figure 3 determined how the amount of data is specified. The *Sample* parameter was set *'absolute'* to create an exactly defined number of records. This was the required number of samples as specified in the *Sample* size parameter. The entire dataset was divided into two parts with equal proportions of Yes and No classes. Yes and No contained 319 each absolute size.

After removing irrelevant attributes, data balancing was applied to the student dataset to achieve the model's higher accuracy result. Data balancing is necessary if the classification categories are not approximately equally represented as it is considered an imbalanced dataset [38]. The *Sample* method created a sample based on the original records. The sample size was an absolute basis and focused on the number of current records and class distribution in the resulting sample. This basis increased the accuracy and balance of the Yes and No classes [39]. Thus, from 768 data samples, 638 records were retained after applying the sampling method. The *Weight by Correlation* operator in Figure 4 calculates the weight of attributes for the *Completed* label attribute using correlation analysis. The higher the weight of the attribute, the more important it was to consider.

b) Application of Weight Attribute using Correlation Analysis

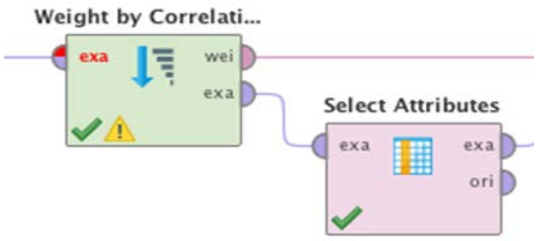

Fig. 4. Weight by a correlation process

The weight by correlation was used to determine the weight and relevance of a specific attribute, among other attributes. Pearson *r* correlation was the most widely used correlation statistic to measure linearly related variables' relationship. Pearson *r* correlation measured the degree of relationship between the two attributes[40].

Pearson's r formula:

$$r = \frac{n(\sum xy - (\sum x)\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (1)$$

r - Pearson r correlation coefficient
N - number of observations
∑xy - sum of the products of paired scores
∑x - sum of x scores

$\sum y$ - sum of y scores
$\sum x^2$ - sum of squared x scores
$\sum y^2$ - sum of squared y scores

The attributes with a high correlation coefficient among the variables were the final data inputs for building the Naïve Bayes classification model. *Select Attributes* parameter in Figure 4 selected the *Attribute Selection* filter; the method was a subset option: this option allows choosing multiple attributes based on weight by correlation. The significant variables assessed by the feature selection technique (Weight by Correlation) were the final parameters in creating the model to feed into the Naïve Bayes process.

Phase 3: Classification Process

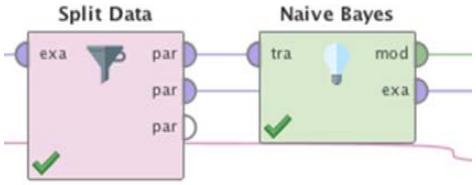

Fig. 5. Classification process

The third phase of the study was the classification process. In this phase, the significant attributes from Phase 2 were partitioned into training data and testing data. As presented in Figure 5 the operator *Split Data* was constructed to shuffle data into two subsets, sized 70% and 30% of data. In the Training phase, the *Naive Bayes* operator builds a Naive Bayes model upon the training set. The model is then fed into the testing phase for performance evaluation. A Naïve Bayes algorithm has been applied to the dataset for building the classification process. The following steps were also done:

Phase 4: Result

a) Model Testing and Evaluation

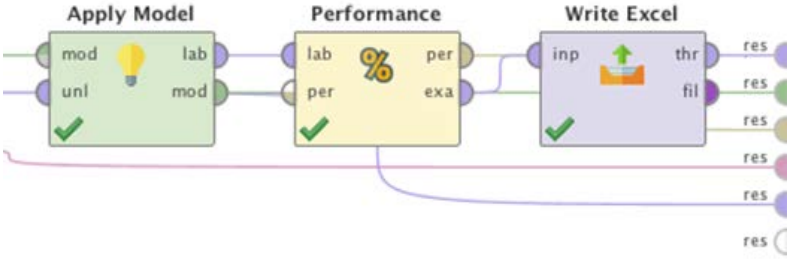

Fig. 6. Data model testing and evaluation process

The last phase of the workflow included the testing and evaluation of the model. The model was applied to the 30% test data and evaluated using confusion matrix cross-validation to get the model's accuracy percentage.

The *Apply Model* (Figure 6) operator predicted the label value for each test set. This prediction was added to a new column named prediction (completed). After applying the formula and generating the projection column in the test sample, the model's output was tested using the *Performance* operator. After selecting a suitable model, the data was exported into an excel file using the *Write Excel* operator.

Table 2. Confusion Matrix Table

		Yes	No
Predicted Class (Expectation)	**Yes**	TP(true positive) *positive Examples that have been correctly identified*	FN(false negative) *positive Examples that have been incorrectly identified*
	No	FP(false positive) *negative Examples that have been incorrectly identified*	TN(true negative) *negative Examples that have been correctly identified*

The model performance was tested and evaluated through the confusion matrix accuracy, precision, and recall. The confusion matrix showed prediction, either positive or negative, arranged as in the 2x2 matrix, as shown in Table 2.

Accuracy measures the percentage of the model's correct predictions.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (2)$$

Precision evaluates how precise a model is in predicting positive labels. It is calculated as a total number of true positive's divided by the total number of true positive's + total number of false positives.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (3)$$

Recall measures the percentage of actual positives a model correctly identified.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (4)$$

4. Results and Analysis

This section discussed the results and findings obtained based on the conceptual framework and the experimental design and methods. It discusses the result generated by the correlation analysis technique, classification process applying Naïve Bayes algorithm, and model performance evaluation.

1) Results of Correlation Analysis

Correlation analysis identified the weight of individual attributes to the label attribute. The initial data were loaded, and different attributes were identified and ranked using the correlation analysis method, as presented in Table 3. From 10 possible predictor variables, only four (4) predictor variables were selected after correlation analysis. Among the ten (10) attributes, the variables with above 0.2 weights were considered significant attributes and were used to classify students in predicting student completion. Four (4) attributes in Table 3 were considered: Parents' Monthly Income attribute got the highest weight among other attributes, followed by Mother's Educational Attainment, Father's Educational Attainment, and High School GPA. The other six (6) attributes were not considered since it contains minimal correlation to what is being predicted, which is the label *completed*.

Table 3. Correlation Analysis Output (Rank by Weight)

Rank	Attributes	Weight by Correlation
1	Parents' Monthly Income	0.648
2	Mother's Educational Attainment	0.451
3	Father's Educational Attainment	0.420
4	High School GPA	0.194
5	4th Year HS Mathematics Average	0.162
6	1st-year 1st semester GPA	0.140
7	High School Type (Public/Private)	0.053
8	Age during admission	0.017
9	1st-year 2nd semester GPA	0.009
10	Gender	0.004

2) Results of the Classification phase

The significant attributes identified in correlation analysis was splitted into 70% training data or 447 records and 30% testing data or 191 records. The primary variable used as a label/target was *Completed*. The classification model was built to predict the likelihood for program completion (Yes/No) output variable - Yes, students who were likely to complete college and No, students who were likely not to complete college.

Testing the model was used to make predictions using the summaries trained from the training phase. Prediction involved calculating the probability that a given data instance belongs to each class and selecting the class with the most significant probability as the prediction. To test the model, it was applied to a previously unseen record as a test set. Table 4 shows the resulting percentage of the prediction. There were 84 out of 191 data samples, or 44% of students were predicted to complete the program. On the other hand, 107 out of 191 data samples, or 56%, were predicted as not completing the program.

Table 4.Test result percentage

Prediction (Completed)		
		Prediction
NO	107	0.56
YES	84	0.44

The actual result of the prediction ***(a portion of the result from the RapidMiner)*** for the given records was shown in Table 5 (predicted Yes) and Table 6 (predicted No). It was observed the level of confidence for each instance. The higher the parents' monthly income supported with both parents' high educational attainment, the student was predicted to complete the program, having a confidence level for Yes of 1 or approximately 100%. On the other hand, the level of confidence for a student having low parents' monthly income and low educational attainment of each parent was observed to nearly 0 for Yes and almost 1 for No.

Table 5. Test Result with YES Prediction output in RapidMiner Studio

Completed	Prediction	**Confidence (YES)**	**Confidence (NO)**	HSGPA	Parents' Income	Father Educ attainment	Mother Educ Attainment
YES	YES	**0.97**	**0.03**	83.6	5	3	2
YES	YES	**1.00**	**0.00**	88.72	5	3	4
YES	YES	**0.97**	**0.03**	84.61	3	4	2
YES	YES	**1.00**	**0.00**	86.7	4	1	5
YES	YES	**0.71**	**0.29**	87.59	4	2	2
YES	YES	**1.00**	**0.00**	85.73	5	3	3
YES	YES	**1.00**	**0.00**	88.44	6	4	4
YES	YES	**1.00**	**0.00**	85.78	6	3	4
YES	YES	**0.91**	**0.09**	87.1	3	3	3
YES	YES	**1.00**	**0.00**	87	4	4	4

Table 6. Test Result with NO Prediction output in RapidMiner Studio

Completed	Prediction	**Confidence** (YES)	**Confidence** (NO)	HSGPA	Parents' Income	Father Educ Attainment	Mother Educ Attainment
NO	NO	0.03	0.97	89.34	1	2	2
NO	NO	0.24	0.76	88.66	2	1	3
NO	NO	0.04	0.96	86.4	2	1	1
NO	NO	0.01	0.99	80.8	2	1	1
NO	NO	0.04	0.96	85.57	2	1	1
NO	NO	0.02	0.98	85.29	1	2	2
NO	NO	0.01	0.99	98.75	1	1	2
NO	NO	0.01	0.99	87.99	1	1	1
NO	NO	0.01	0.99	83.52	1	1	2
NO	NO	0.01	0.99	80.01	1	2	2
NO	NO	0.03	0.97	87.14	1	2	2
NO	NO	0.02	0.98	85.87	1	1	2
NO	NO	0.04	0.96	81.95	1	2	3

3) Results of Model performance evaluation

Two sets of tests were carried out to determine the effects of the feature selection technique on models' predictive performance. The first test used the three (3) features in the dataset (Father and Mother Educational Attainment, High School GPA (Table 7), and the second test used the four best-ranked attributes (Table 8) (Parents' Monthly Income attribute, Mother's Educational Attainment, Father's Educational Attainment, and High School GPA according to their weights by correlation).

Table 7. Top 2,3,4 Attributes Confusion Matrix for Validation Set

Accuracy: 72.77%			
Predicted	true YES	true NO	class precision
YES	55	20	**73.33%**
NO	32	84	72.41%
class recall	**63.22%**	80.77%	

The result from Table 7, using the attributes Mother's Educational Attainment, Father's Educational Attainment, and High School GPA accuracy, achieved only 72.77%. This can be interpreted that removing parents' monthly income attribute greatly affects the model's accuracy result.

Table 8. Top Four (4) Attributes Confusion Matrix for Validation Set

ACCURACY: 83.77%			
Predicted	True Yes	True No	Class Precision
Yes	TP= 70	FN=17	**83.33%**
No	FP= 14	TN=90	84.11%
Class Recall	**80.46%**	86.54%	

Table 8 showed the confusion matrix's result in terms of the model's accuracy, precision, and recalls percentage of the model. There were 191 data samples for the test set from the result obtained; the accuracy was 83.77%, with 80.46% class precision and 83.33% class recall. The result was higher than the previous experiment utilizing all the top four (4) attributes.

The detailed computation of the model's percentage obtained in terms of accuracy, precision, and recall is presented below.

The result of the model in terms of accuracy is:

$$\mathbf{83.77}\% = \frac{\mathbf{70+90}}{\mathbf{70+90+17+14}} \tag{5}$$

The result of the model in terms of precision is:

$$\mathbf{80.46}\% = \frac{\mathbf{70}}{\mathbf{70+17}} \tag{6}$$

The result of the model in terms of recall is:

$$\mathbf{83.33}\% = \frac{\mathbf{70}}{\mathbf{70+14}} \tag{7}$$

5. Conclusion and Future Works

The study identified significant attributes affecting program completion, namely (in order of significance): parents' monthly income, mother and father's educational attainment, and High School GPA attributes. From the result obtained, the model identified 44% or 84 students out of 191 of the probability of student who was able to complete, and 56% or 107 out of 191 students was the probability of student who was not able to finish college.

Using test data samples, two sets of tests were carried out to determine the effects of the feature selection technique on models' predictive performance. The first test used the three (3) features in the dataset (Father and Mother Educational Attainment, High School GPA, and the second test used the four best-ranked attributes (Table 8) (Parents' Monthly Income attribute, Mother's Educational Attainment, Father's Educational Attainment, and High School GPA according to their weights by correlation). The first test result showed 73% accuracy, while the second experiment utilizing all the top four attributes achieved 84% accuracy with 80.46% class precision and 83.33% class recall; the said model yielded compelling results. Using the Naïve Bayes Classifier to develop a prediction model for program completion is feasible. It can be utilized to provide a powerful educational tool in determining students' probability for completion or retention.

The findings of this study have a significant impact on HEIs, specifically in combatting the retention problem. University administrators can use the vital yields regarding the best attributes that contribute to students' success to complete college. For instance, in the attribute Parent's Monthly Income, the university can provide scholarship programs or extend any possible assistance, such as earning additional income thru Student Assistantships in different offices. Teachers who have the data of that student at-risk should immediately call the identified students' attention and extend any possible help, such as guidance and counseling. On the other hand, the student-at-risk should not hesitate to seek help from their parents, teachers, and school authorities.

Future researchers may integrate this model into an actual system that automatically identifies students who will complete college. Furthermore, researchers can adopt other predictor variables related to attrition and program completion of students. Other researchers may also use the model by utilizing more complex students' records, which cover other university programs or courses.

References

[1] D. E. Azarcon, C. D. Gallardo, C. G. Anacin, and E. Velasco, "Attrition and Retention in Higher Education Institution: A Conjoint Analysis of Consumer Behavior in Higher Education," Asia Pacific J. Educ. Arts Sci., vol. 1, no. 107, pp. 2362–8022, 2014, [Online]. Available: www.apjeas.apjmr.com.

[2] H. K. Das and V. Janardhan, "Materials Today : Proceedings Machine learning approaches in education," Mater. Today Proc., no. xxxx, 2020, doi: 10.1016/j.matpr.2020.09.566.

[3] R. S. Agieb, "Machine learning models for the prediction the necessity of resorting to icu of covid-19 patients," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 5, pp. 6980–6984, 2020, doi: 10.30534/ijatcse/2020/15952020.

[4] J. A. A. Repaso and E. T. Capariño, "Analyzing and predicting career specialization using classification techniques," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1 Special Issue 3, pp. 342–348, 2020, doi: 10.30534/ijatcse/2020/5391.32020.

[5] S. Cui et al., "Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation," Injury, vol. 49, no. 10, pp. 1865–1870, 2018, doi: 10.1016/j.injury.2018.07.025.

[6] J. Wu, "A generalized tree augmented naive Bayes link prediction model," J. Comput. Sci., vol. 27, pp. 206–217, 2018.

[7] J. S. Aviles and R. A. Esquivel, "Mining social media data of Philippine higher education institutions using naïve bayes classifier algorithm," Proc. 2019 9th Int. Work. Comput. Sci. Eng. WCSE 2019, pp. 681–688, 2020.

[8] H. Shaziya, R. Zaheer, and G. Kavitha, "Prediction of Students Performance in Semester Exams using a Naïve bayes Classifier," pp. 9823–9829, 2015, doi: 10.15680/IJIRSET.2015.0410072.

[9] T. Barbé, L. P. Kimble, L. M. Bellury, and C. Rubenstein, "Predicting student attrition using social determinants: Implications for a diverse nursing workforce," J. Prof. Nurs., vol. 34, no. 5, pp. 352–356, 2018, doi: 10.1016/j.profnurs.2017.12.006.

[10] D. Delen, K. Topuz, and E. Eryarsoy, "Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition," Eur. J. Oper. Res., vol. 281, no. 3, pp. 575–587, 2020, doi: 10.1016/j.ejor.2019.03.037.

[11] J. M. Ryan, T. Potier, A. Sherwin, and E. Cassidy, "Identifying factors that predict attrition among first year physiotherapy students: a retrospective analysis," Physiotherapy, 2017, doi: 10.1016/j.physio.2017.04.001.

[12] Raheela Asif, Agathe Merceron, Mahmood K. Pathan, "Predicting Student Academic Performance at Degree Level: A Case Study", International Journal of Intelligent Systems and Applications, vol.7, no.1, pp.49-61, 2015.

[13] J. L. Wircenski and C. Membe, "Identifying factors that predict student success in a community college online distance learning course."

[14] L. Thamarai, L. Parthiban, and K. Mahalakshmi, "Comparison of classification techniques on data mining," no. April, 2019, doi: 10.12732/ijpam.v118i11.43.

[15] P. S. Performance, "www.econstor.eu," 2012.

[16] M. Rashedur, "www.econstor.eu," pp. 0–25, 2015, doi: 10.1186/s40165-014-0010-2.

[17] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The Prediction of Students ' Academic Performance Using Classification Data Mining Techniques," vol. 9, no. 129, pp. 6415–6426, 2015.

[18] Abhilasha Nakra, Manoj Duhan, "Comparative Analysis of Bayes Net Classifier, Naive Bayes Classifier and Combination of both Classifiers using WEKA", International Journal of Information Technology and Computer Science, Vol.11, No.3, pp.38-45, 2019.

[19] Ma Da, Wei Wei, Hu Hai-guang, Guan Jian-he,"The Application of Bayesian Classification Theories in Distance Education System", International Journal of Modern Education and Computer Science, vol.3, no.4, pp.9-16, 2011.

[20] Kirtika Yadav, Reema Thareja, "Comparing the Performance of Naive Bayes And Decision Tree Classification Using R", International Journal of Intelligent Systems and Applications, Vol.11, No.12, pp.11-19, 2019.

[21] W. Jang, J. K. Lee, J. Lee, and S. H. Han, "Naive Bayesian Classifier for Selecting Good/Bad Projects during the Early Stage of International Construction Bidding Decisions," Math. Probl. Eng., vol. 2015, 2015, doi: 10.1155/2015/830781.

[22] P. Butka, P. Bednár, and J. Ivančáková, "Methodologies for Knowledge Discovery Processes in Context of AstroGeoInformatics," Knowl. Discov. Big Data from Astron. Earth Obs., pp. 1–20, 2020, doi: 10.1016/b978-0-12-819154-5.00010-2.

[23] M. A. Valle, S. Varas, and G. A. Ruz, "Expert Systems with Applications Job performance prediction in a call center using a naive Bayes classifier," Expert Syst. Appl., vol. 39, no. 11, pp. 9939–9945, 2012, doi: 10.1016/j.eswa.2011.11.126.

[24] Ajay Kumar Pal, Saurabh Pal, "Classification Model of Prediction for Placement of Students", International Journal of Modern Education and Computer Science, vol.5, no.11, pp.49-56, 2013.

[25] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," Procedia Comput. Sci., vol. 179, no. 2020, pp. 498–506, 2021, doi: 10.1016/j.procs.2021.01.033.

[26] V. Ramesh and K. Ramar, "Predicting Student Performance: A Statistical and Data Mining Approach," Int. J. Comput. Appl., vol. 63, no. 8, pp. 975–8887, 2013.

[27] D. D. Pokrajac, K. R. Sudler, P. Y. Edamatsu, and T. Hardee, “Prediction of Retention at Historically Black College / University using Artificial Neural Networks,” 2016.
[28] S. Hall and M. Aryee, “College Students ’ Persistence and Degree Completion In Science , Technology , Engineering , and Mathematics (STEM): The Role Of Non- Cognitive Attributes Of Self-Efficacy , Outcome Expectations , And Interest,” 2017.
[29] A. A. Aziz, N. Hafieza, and I. Ahmad, “First Semester Computer Science Students ’ Academic Performances Analysis by Using Data Mining Classification Algorithms,” no. September, pp. 15–16, 2014.
[30] S. Geiser and M. V. Santelices, “VALIDITY OF HIGH-SCHOOL GRADES IN PREDICTING STUDENT SUCCESS BEYOND THE FRESHMAN YEAR : High-School Record vs . Standardized Tests as Indicators of Four-Year College Outcomes”, 2007.
[31] T. A. Cardona, E. A. Cudney, and J. Snyder, “Predicting degree completion through data mining,” ASEE Annu. Conf. Expo. Conf. Proc., 2019, doi: 10.18260/1-2--33183.
[32] C. Ernesto and L. Guarín, “Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia,” 2013.
[33] B. K. Baradwaj, “No Title,” IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 6, pp. 63–69, 2011.
[34] Muladi, U. Pujianto, and U. Qomaria, “Predicting high school graduates using Naive Bayes in State University Entrance Selections,” 4th Int. Conf. Vocat. Educ. Training, ICOVET 2020, pp. 155–159, 2020.
[35] D. Rozon, “FACTORS AFFECTING PERSISTENCE RATES AMONG COLLEGE FRESHMEN,” 2015.
[36] M. M. Chingos, “What Matters Most for College Completion? ACADEMIC PREPARATION IS A KEY PREDICTOR OF SUCCESS,” AEI Pap. Stud., p. 3A, 2018.
[37] O. M. Way, “Knowledge Discovery and Data Mining : Towards a Unifying Framework,” 1996.
[38] N. V. Chawla, “Data Mining for Imbalanced Datasets: An Overview,” Data Min. Knowl. Discov. Handb., pp. 875–886, 2009, doi: 10.1007/978-0-387-09823-4_45.
[39] J. D. Febro and J. Barbosa, “Mining student at risk in higher education using predictive models,” J. Adv. Technol. Eng. Res., vol. 3, no. 4, 2017, doi: 10.20474/jater-3.4.2.
[40] T. Fu, X. Tang, Z. Cai, Y. Zuo, Y. Tang, and X. Zhao, “Correlation research of phase angle variation and coating performance by means of Pearson’s correlation coefficient,” Prog. Org. Coatings, vol. 139, no. October 2019, p. 105459, 2020, doi: 10.1016/j.porgcoat.2019.105459.

Authors’ Profiles

Joann G. Perez earned her master’s degree in information technology at Technological Institute of the Philippines and is now pursuing her Doctor in Information Technology Program at the same institution. Currently, she is a full-time faculty member of Bulacan State University-Sarmiento Campus. Her research interests include data mining and software engineering. Ms. Perez is a member of the Philippine Society of IT Educators Central Luzon Chapter (PSITE R3).

Eugene S. Perez is a faculty member of Bulacan State University-Sarmiento Campus. He took up his Master in Information Technology at Angeles University Foundation. Mr. Perez is also a web-developer, his research interest includes Web development, Software Engineering, Cybersecurity and Information Systems. Mr. Perez is a member of the Philippine Society of IT Educators Central Luzon Chapter (PSITE R3).

How to cite this paper: Joann Galopo Perez, Eugene S. Perez, " Predicting Student Program Completion Using Naïve Bayes Classification Algorithm ", International Journal of Modern Education and Computer Science(IJMECS), Vol.13, No.3, pp. 57-67, 2021.DOI: 10.5815/ijmecs.2021.03.05