A Naïve Bayes Students' Performance Prediction Model for Decision Support System

Jay C. Liza, Bobby D. Gerardo

jayliza@wvsu.edu.ph
bgerardo@nipsc.edu.ph

West Visayas State University- Janiuay Campus, Janiuay, Iloilo, Philippines
Northern Iloilo Polytechnic State College, Concepcion, Iloilo, Philippines

Abstract
Adequate assistance in the learning process is important using accurate estimation of students' academic performance based on new emerging techniques, and discover new knowledge, finding meaningful variables answers educational problems. Data contains hidden information for analyzing, extracting information and knowledge to find patterns, and using this for decision-making. WEKA (Waikato Environment for Knowledge Analysis) is used and Naïve Bayesian Classification Method is implemented, as the pre-processing mechanism for 10-fold cross-validation where classification models were generated, cross-validation method, and percentage split were used to evaluate the efficiency of this algorithm. Linear Regression is used to identify the significant predictors that affect the students' academic performance. The simulation results show that with Naïve Bayes Classification Method algorithm for classification with a correctly classified instance of 93.48% and incorrectly classified instance of 6.52%, this result indicates that with the classification method the accuracy level of prediction increases. To validate the generated model, the experiments were conducted using real data. The result is a good model intended to be used in the school decision support system in predicting learners' academic performance.

Keywords: classification, prediction, Naïve Bayes, student performance;

I. Introduction

In the advent of technology educational institutions must introduce a newly emerging technique to discover knowledge from data originating from the educational environment, and answer the educational problem. Educational institution stores huge amount of data and this may include students' academic records, their demographic profile and must be explored to have a strategic edge among the Educational Organizations. The available data, on the other hand, is frequently used to create simple queries and typical reports. Some techniques have to be introduced to effectively transform available data into information and knowledge to support decision-making. Prediction of learners' academic performance has long been regarded as an essential research topic in many academic disciplines for several reasons. Student academic performance can be predicted, then the instructor can use predictive models to help him or her forecast student academic success and take some proactive measures [7] and thus management can intervene timely and take essential intervention to help students to improve their performance. To successfully transform accessible data into information and knowledge to support decision-making, advanced information technologies must be deployed.

Educational data mining uses techniques and concepts from these different fields in researching, using computing technologies to find patterns in massive amounts of educational data, and putting them into practice. Without the aid of EDM techniques, humans can't find relevant patterns and data [18] is a new stream in the data mining research field that could give effective analyses and new instructional strategies to answer educational questions.

II. Related Studies

Many types of research had been done to identify the factors that influenced learners' academic performance. It is an initial step to align secondary education with higher education, where currently exists unquestionable impact against each other. Most of the prior works were beyond the data mining domain with a limited mechanism to transform institutional data into models that could be utilized for Decision Support System (DSS).

However, when it comes to learner's academic performance is rarely being studied especially within the scope of data mining, mainly due to this is due to the scarcity of reliable and accurate data sources. The data sources for previous works were obtained and compiled via a survey of a representative sample group, as well. Learner's academic performance issues have also been taken into consideration in other countries.

In recent years, there has been an increasing interest to investigate scientific questions with the use of data mining. Educational data mining is for scientific discovery about the learning of students and attempts to predict the student educational outcomes.

Han and Kamber (2012) describe data mining as a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing, and data visualization. Romero and Ventura, have a survey on educational data mining between 1995 and 2005 (Kabachieva, 2012). They came to the conclusion that educational data mining is a potential field of study with a unique set of requirements that aren't seen in other fields. Thus, work should be oriented towards the educational domain of data mining.

In the work, Bhardwaj and Pal (2012) found out that those students who performed well in their secondary school will perform well in their bachelor's studies. Furthermore, it was found that the living location, medium of teaching, mother's qualification, student other habits, family annual income, and student family status, all of which, highly contribute to the learner's academic performance, thus, it can predict a learner's grade or generally his/her performance if basic personal and social knowledge was collected about him/her.

Based on Chandra and Nandhini (2015) by applying the association rule mining analysis based on students' failed courses to identify students' failure patterns that will identify a hidden relationship between the failed courses and suggests relevant causes of students' performances. The extracted association rules reveal some hidden patterns of students' failed courses which could serve as a foundation stone for academic planners in making academic decisions and aid in the curriculum restructuring and modification to improve students' performance and reduce the failure rate.

Shahiri, et.al (2015) concluded that predicting student performance is mostly useful to help the educators and students improve their learning and teaching process and to monitor the student's performance systematically. Furthermore, this paper reviewed previous studies on predicting students' performance with various analytical methods. While on prediction techniques the classification method is frequently used in the educational data mining area. Under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting student's performance.

Asif (2017) used Decision trees and Naïve Bayes algorithms to predict the likelihood of success/failure of the undergraduate students from Ethiopia's Debre Markos University. His findings indicated that EHEECE (Ethiopian Higher Education Entrance Certificate Examination) result, gender, number of students It was the number of courses offered in a term, as well as the field of study, that influenced the most. The highest prediction accuracy was 92.34% obtained with the decision tree algorithm using 10-fold cross-validation

III. Methods and Proposed Model

This study developed the research methodology based on three phases of the data-mining techniques including data consideration, normalizing data, training data for model development, use of the model to predict student academic performance. We used the Waikato Environment for Knowledge Analysis to construct the classifiers (WEKA), an open-source data mining tool [12] that was developed at the University of Waikato New Zealand. It includes several learning algorithms that may be readily applied to the data. Only datasets in the Attribute-Relation File Format (ARFF) format are accepted by WEKA. Therefore, once the data preparation is done, we transform the dataset into an ARFF file with an extension of .arff. Using the Process of Knowledge Discovery in which data mining is a significant step Figure 1 depicts the iterative and sequential stages. It consists of four steps, which include data consideration, normalizing data, prepares training data.

1. 2 Data Consideration

Selecting data sources important considerations for choosing data include whether or not the key variables are available to appropriately define an analytic cohort and identify exposures, outcomes, covariates, and confounders [14]. The data should be granular enough, with enough historical data to define baseline variables, and a reasonable amount of time between visits.

As in any research or statistical analysis, one must start with the correct understanding of what needs to be done and why it needs to be done. This requires the correct understanding of the problem, the selection of the right set of actions to address the

problem, and the analyses of the results in light of the initial question(s). If data mining (or any other data analysis) is undertaken without this initial consideration, chances are that the final product will not solve any problems.

To prepare the data for the classification analysis, the first activity performed as data reduction to have a more compact, easily interpretable representation of the target concept. This is done by focusing on the variables most relevant to the scope of this study

1.3 Normalizing Data

The primary goal of data preparation is to change and convert raw data in order to reveal or make more accessible the information contained within the data collection. It's typical to go through a number of processes to change the original qualities or create new ones with superior properties to improve the prediction capability and the original attributes or to generate new attributes with better properties that will help the predictive power of the model [9], To avoid problems especially if the DM algorithm cannot correctly handle them, we normalize the collected data by transforming the nominal attributes into binary variables and then treated as numeric, typically starting from 0 or 1 onwards.

1.4 Training Data to use for Model Development

The training set sample corresponding to the selection of a candidate model development is given below. The table clearly shows the different attributes used for classification and the application-dependent nature. The test data is also similar to the training data without the class column which will be predicted with the help of the algorithm implementation

Table 1. Training Data for Student Performance

Gender	Family Type	Instruction	Extra-Curricular	G
Male	Joint	English	Yes	90
Female	Individual	Mixed	Yes	91
Female	Joint	Hiligaynon	Yes	87
Male	Individual	Filipino	No	78
Male	Joint	English	No	90
Female	Individual	English	Yes	77

1.5 Use of the model to predict student academic performance

In research, prediction models may assist in the design and analysis of randomized trials. Models are also useful to control for confounding variables in observational research, either in traditional regression analysis or with a modern approach.

1.6 Sampling

Researchers need to consider the type of information needed to obtain directly from the first-hand sources utilizing surveys, and questionnaires. The primary source of data used in this study is from Leonora S. Salapantan National High School and Bancal National High School which are the enrolled learners from the school year 2017 – 2018.

1.7 Data Collection

The data set used in this study was obtained through a survey questionnaire because the educational institution doesn't have a learner's database system where the data of the students were being stored. And another raw data was retrieved from the student form 137 where their GWA is located.

Table 2. Demographic Attributes of Students

No.	Attributes	Values
Independent Variables		
1	Gender	{ Male, Female}
2	Address	{Brgy. 1, Brgy. 2, …}

4	Family Type	{Joint, Individual}
5	Family Annual Income	{10.000-50.000,…}
6	Fathers Occupation	{Government, OFW, Farmer..}
7	Fathers Education	{Elem., HS, College..}
8	Mother Occupation	{Housewife, Government, OFW..}
	Mother Education	{Elem, HS, College..}
9	Internet access	{Yes, No}
10	Medium of Instruction	{English, Tagalog, Mix..}
11	Instructional Material	{Books, Encyclopedias,..}
12	Travel Time from residence to school	{1-10 min, 11-20 min,..}
13	Want to take higher education	{Yes, No}
14	In relationship	{Yes, No}
15	Extra-Curricular Activity	{Yes, No}
16	Medium of Transportation	{Walking, Tricycle, Jeep, Mix}
17	Weekly Study Time	{1-2 hours, 3-4 hours,5-6hours…}
Dependent variable		
1	GWA	{99, 98,97, 96, 95, 94…..}

To get better input data for data mining techniques, we did some preprocessing to prepare the dataset for the classification task. First, eliminate missing values in critical attributes, identify outliers, correct inconsistent data, as well as remove duplicate data, we normalize data before loading the data set to the data mining software. The attributes marked as selected as seen in Table 1 are processed via the Weka software to apply the data mining methods on them. The attributes like Students Name is not selected to be part of the mining process; this is because they do not provide any knowledge for the data set processing and they present personal information of the students, also they have very large variances or duplicates information which make them` irrelevant for data mining.

The following steps are performed as part of the preparation and preprocessing of the data set: The data set contains missing 58 values in various attributes from 550 records, the records with missing values are ignored from the data set since it doesn't consider a large amount of data. The number of records is reduced to 492 records.

The GWA attribute in the data set contains a large number of continuous values (GWAs). So for efficient later processing, simplified data description, and understanding for data mining results, we discretized this attribute to a categorical one. For example, we grouped all GWAs into five categorical segments; Excellent, Very good, Good, Average, and Poor.

To get insight and knowledge from the criteria of knowledge and skills towards student performance, eighteen (18) tables, which were related to the scope of this study, were selected. We relate these tables with each other and filter them in such a way that a single view on the same database was formed. In this view, there are a total of 492 columns.

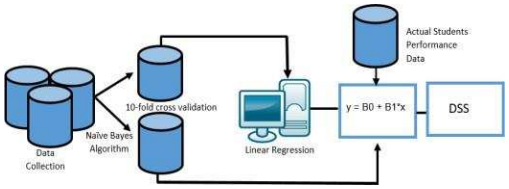


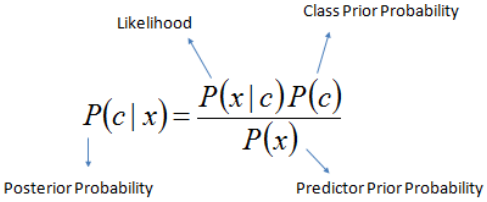
Figure 1. General Architecture for Classification using Naïve Bayes Algorithm for Students Performance

1.7 Prediction Model

1.7.1 Prediction using Classification Technique

This classification technique is for tracing or monitoring students at risk, performed in the information system and educational management. This model is intended to be used in the school decision support system in predicting learners' academic performance.

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

1.7.2 Linear Regression

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.
The linear equation assigns one scale factor to each input value or column, called a coefficient, and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.
For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation, therefore, is in the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example). It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

IV. Simulation and Result

Using the WEKA environment, the classification computation was rendered on a normalized dataset containing 492 records with 17 attributes as covariates and 1 attribute as dependent variables. The simulation results were based on normalized attributes from the student's data of Leonora Salapantan National High School and Bancal National High School. And this includes results on the classification accuracy using the Bayesian belief network and the Linear Regression algorithm for the prediction of student performance. For this study, performance prediction was only anchored on a single dependent variable on the "GWA" attribute.

The classification technique, which is called classification task, involves two stages in classification consisting of training and testing. The testing dataset is used to estimate the predictive accuracy. The classification phase implemented in WEKA software was computed, which accounted for 80% (N= 345) of datasets based on the original 492 instances.

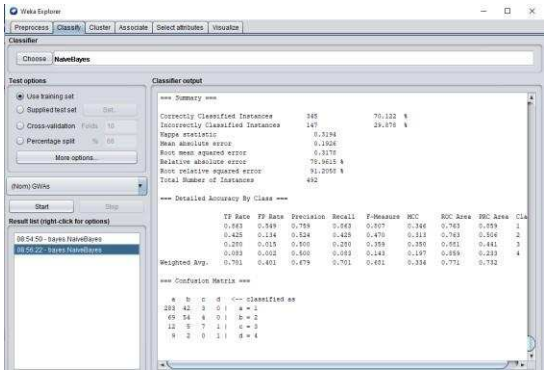


Figure 2. Classification report without Naïve Bayes Classification Application

The classification using the Naïve Bayes Algorithm rendered on WEKA yields 345 correctly classified instances while only 147 instances were incorrectly classified. The incorrectly classified instances are considered as outliers because we observed that there is an abnormal distance from other values in a random sample from datasets. The overall computation resulted in a Root Mean Squared Error of 0.16, which is relatively low or negligible. (Two classes were found labeled as Class 1 with the precision of 0.928 containing 345 correctly classified instances and Class 2 with a precision of 0.15 containing 25 incorrectly classified instances.)

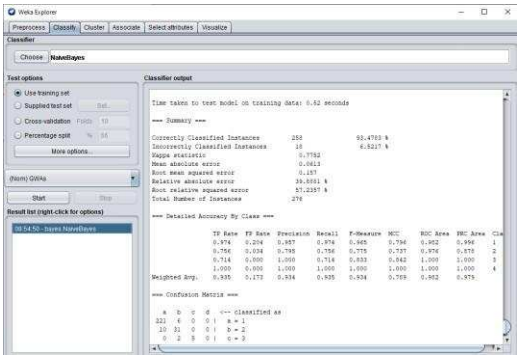


Figure 3. Classification report with Naïve Bayes Application

The comparison of results shows that the percentage accuracy without the application of the Naïve Bayes Method gained a percentage accuracy of 92.75% while the percentage accuracy of the data with the NB classification method is 93.48%. This shows that if you apply the classification method the accuracy level of prediction increases. And in comparison, the Relative Absolute Error generated a value of 40.21% without Naïve Bayes application, while with Naïve Bayes the relative absolute error gained 39.88%. This shows the decreasing percentage of error attributed to the pre-classification of the dataset.

Table 3. Linear Regression Analysis Summary for Prediction Student Performance

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.224	.299		7.434	.000
	Gender	-.168	.047	-.170	-3.560	.000
	Address	.006	.008	.039	.704	.482
	Fam_Type	.046	.047	.047	.981	.327
	Income	-.080	.020	-.217	-4.049	.000
	Father_Occu	.012	.010	.058	1.212	.226
	Father_Educ	-.076	.026	-.165	-2.900	.004
	Mother_Occu	-.010	.017	-.031	-.622	.534
	Mother_Educ	-.022	.026	-.047	-.866	.387
	Internet	-.004	.055	-.004	-.078	.938
	Instruction	.046	.021	.104	2.238	.026
	Material	-.030	.064	-.022	-.467	.641
	Travel_Time	.088	.030	.158	2.974	.003
	Higher_Ed	.154	.150	.048	1.023	.307
	Relationship	-.385	.057	-.314	-6.773	.000
	Extra_curr	-.044	.047	-.045	-.936	.350
	Transpo	.038	.039	.047	.965	.335
	Study_Time	-.049	.015	-.158	-3.235	.001

a. Dependent Variable: GWAs

The Linear Regression model result with was computed at 5 percent significance level, which showed 'Gender' (B=-.166 Std. Error=.046, Beta=-.172, t=-3.624, Sig<0.000) 'Family Income' (B=-.073, Std. Error=.019, Beta=-.202, t=-3.803, Sig<.000) 'Father Education' (B=-.072, Std. Error=.025, Beta=-.160, t=-2.827, Sig<0.005) "Medium of Instruction" (B=-.045, Std. Error=.020, Beta=-.104, t=2.247, Sig<0.025) 'Travel Time' (B=-.091, Std. Error=.029, Beta=-.168, t=3.189, Sig<0.002) 'Engagement in Intimate Relationship" (B=-.341, Std. Error=.063, Beta=-.283, t=-5.376, Sig<0.000) and 'Study Time" (B=-.348, Std. Error=.055, Beta=-.323, t=-7.010, Sig<0.000) as predictors of student performance. The other eleven (11) co-variates when entered into the equation were found non-contributors to the predictability of the student performance.

V. Conclusion and Recommendation

The result on linear regression model demonstrated factors such as 'Gender', 'Family Income', 'Father's Education', 'Medium of Instruction used inside the classroom', 'Travel Time spent going to school', 'Engagement in a relationship', 'Study Time' as predictors of student performance. The mathematical model could be used to forecast current and future data in the appropriate student performance and this demonstrated that the proposed data mining mechanism could mathematically model institutional data for decision support applications. Future enhancement of the proposed prediction scheme could be improved by incorporating other variables that may contribute to the variability of the equation, thus expecting a more enhanced model.

For both processing steps that are with NB and without NB, they indicated almost the same attributes for predicting learner's academic performance but the latter method "Language use inside the classroom" (Instruction, sig.0.089) become non-significant predictor, Over-all method showed a higher variance of predicting the learner's performance. All other attributes appeared to be non-significant predictors.

This mathematical model could be used to forecast current and future data in the appropriate students teaching-learning of student job and this demonstrated that the proposed data mining mechanism could mathematically model institutional data for

decision support applications. Future enhancement of the proposed prediction scheme could be improved by incorporating other variables that may contribute to the variability of the equation, thus expecting a more enhanced model.

Models for parametric data could be also considered in future works, where another variant of regression algorithms appropriate for the type of data could be applied and test its efficiency for developing a framework for students' performance.

References

[1] Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid, A Review on Predicting Student's Performance Using Data Mining Techniques, Procedia Computer Science, Volume 72, 2015, Pages 414-422, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.12.157. (http://www.sciencedirect.com/science/article/pii/S1877050915036182) Keywords: Student performance; educational data mining; performance prediction

[2] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.

[3] Bhardwaj, B.K. and Pal, S., 2012. Data Mining: A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.

[4] Chandra, E. and Nandhini, K. (2010) 'Knowledge Mining from Student Data', European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

[5] Garcia, S. et. al. (2015), Data Preprocessing in Data, Intelligent Systems Reference Library, Vol. 17. Using Classification

[6] Han, J., Kamber, M. (2012), Data Mining: Concepts and Techniques. Morgan Kaufman.

[7] Huang, S. (2011). Predictive modeling and analysis of student academic performance in an engineering dynamics course.

[8] Kornegay C, Segal JB. Selection of Data Sources. In: Velentgas P, Dreyer NA, Nourjah P, et al., editors. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013 Jan. Chapter 8. Available from: https://www.ncbi.nlm.nih.gov/books/NBK126195/

[9] Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, 500-508.

[10] Mense, Evan & Lemoine, Pamela & Richardson, Michael. (2020). Data Mining in Global Higher Education: Opportunities and Challenges for Learning. 10.4018/978-1-7998-0010-1.ch005.