TEST Engineering & Management

March - April 2020
ISSN: 0193-4120 Page No. 4998 - 5010

College Dropout Prediction Model using Supervised Machine Learning

Diosdado C. Caronongan
University of Luzon, Philippines
sirscorpion01@yahoo.com

Article Info
Volume 83
Page Number: 4998 - 5010
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 27 March 2020*

Abstract

Analyzing corporate data to understand and find patterns in customer behavior is a critical exercise that should be undertaken by every business establishment in order to stay competitive. Accordingly, in the education sector, analyzing dropout data will help schools better understand student behavior. Education statistics show that foreign as well as local educational institutions are experiencing an alarmingly high dropout rates. Thus, a school equipped with a tool that can predict a dropout, will have a competitive advantage since it can timely prescribe needed programs of intervention that could prevent dropping out from happening.

This research work aimed to develop dropout prediction models using the two major types of supervised machine learning – Classification Method and Regression Method. The dataset used in the study are based on the academic records of 687 Information Technology and Computer Science students who are enrolled at University of Luzon, Dagupan City Philippines.

Using the combination of accuracy, precision, recall and F measure metrics, the study compared the prediction performances of the two supervised learning models and determined the better solution. Likewise, the effect of feature engineering on the performances of the prediction models were measured and determined. Moreover, a web-based dropout prediction system was developed and deployed using the Shiny package framework.

Index Terms; *College Dropout Prediction, Supervised Machine Learning, Classification Method, Regression Method.*

I. INTRODUCTION

The National Center for Education Statistics (NCES), the primary federal entity for collecting and analyzing data related to education in the U.S. and other nations, presented its study indicating that less than half of all students enrolled in a 4-year bachelor’s program will earn a degree. It further claimed that only 28.1% of full-time nontraditional learners, who comprise the majority of degree-seeking students, had earned an associate’s or bachelor’s degree after 6 years of study [1]. According to Tumapon [2], citing Commission on Higher Education’s (CHED) data on Higher Education Institution (HEI) enrollees from 2001-2012, the Philippine dropout rate reached an alarming 83.7 percent. From an academic institution’s standpoint, these high dropout rates demand a deeper examination on the reasons why many college students drop out from their studies.

Situations similar to what was discussed above, call for the application of machine learning to devise the prediction tool. According to Mitchell [3], the basic idea of machine learning is that a computer can learn from experience. Jou [4] presented similar understanding that in machine learning, software can be used to automatically build analytical models and make computers learn by observing company’s past historical data. This means that the computer finds patterns and rules hidden in a large amount of data that it analyzed, and eventually use these rules to meaningfully describe new data. The creation of rules from data is an automatic process, and it is something that continuously improves with newly presented data [4]-[6]. The emphasis is on making computers build models based on learning and perform tasks without any need for additional programming. In other words, machine learning is nothing more than algorithms that automate the same learning process that humans naturally do.

Machine learning algorithms can be divided into supervised and unsupervised learning. In supervised learning, the input data includes a known class structure [3], [7]. The algorithm is used in creating a model that can predict the dependent variable by using the independent variables. The model is then used to process data whose class structure is the same as the input data. On the other hand, in unsupervised learning, the algorithm is tasked to find a structure in the data whose input data have no known class structure [5]. This suggests that unsupervised learning is more complicated because no labels are given or doesn't require accurate examples.

This research work compared the performances of the two main types of supervised machine learning, namely Classification and Regression. Under Classification, Naïve Bayes algorithm was used while Logistic Regression algorithm was utilized for Regression. Naïve Bayes is a technique based on Bayes' theorem which assumes the independence of a particular variable to any other variable in a class [8]. Perceptive Analytics [9] claimed that Naïve Bayes is simple but very powerful that it is often known to outperform complex algorithms for very large datasets. Le [10] described Logistic Regression as a powerful statistical way of modeling that measures the relationship between a categorical dependent variable and one or more independent variables by estimating probabilities using the cumulative logistic distribution. Comparing Naïve Bayes and Logistic Regression is a good study between the two types of supervised learning since both algorithms operate with the same data types of independent and dependent variables.

The study measured the performances of the two supervised machine learning algorithms in predicting college dropout in the Philippines. Unlike most of the reviewed studies which used accuracy as the only performance metric, this research work made used of four different metrics, namely accuracy, precision, recall and F measure for more comprehensive comparison of the two models' prediction performances. The main participants were Information Technology and Computer Science students of an autonomous university in the Philippines.

The effects of feature engineering in the performances of the prediction models were also determined. To do this, two datasets – the raw dataset and the feature engineered dataset, were used. Feature engineered dataset contained variables with modified values. The study utilized four independent variables which contained the values of percentage of dropped subjects, general weighted average, residence location and gender of students. Previous grade and residence location were found by Agrawal et al. [11] in their research as among the most highly potential variables or predictors for student performance measurement.

The machine learning models were built using R platform. R is an integrated collection of software tools for data manipulation, analysis, and modeling [12], [13]. Furthermore, it has built-in functions and third party packages that automatically measure prediction performances of machine learning prediction algorithms including Naïve Bayes and Logistic Regression.

The identified need for a dropout prediction tool that can help schools better manage their enrollees and the rise in popularity of machine learning and its applications motivated the researcher to choose the topic for his project. The researcher was excited to learn more about the technology and explore its application in addressing school-related problems and issues like student retention.

II. REVIEW OF RELATED LITERATURE

A. Machine Learning Concepts

Machine learning is a field of study that encompasses concepts and techniques from several other areas. It uses concepts from artificial intelligence, statistics, pattern detection, optimization and learning theory to develop algorithms and techniques which can learn from and make predictions on data without being explicitly programmed [14].

Aside from these areas, earlier authors on machine learning like Nilsson [6] included the disciplines of brain models, psychological models and evolutionary models as having significant contributions to machine learning. According to Nilsson, several machine learning techniques are based on neural networks which are patterned after learning process of living brains. Further, on the psychological model discipline, he stressed the contribution of reinforcement, which is an important concept in machine learning research, in the learning process by showing the influence of reward in the learning behavior of goal-seeking animals. On the area of evolutionary models, Nilsson explained that species evolve to improve their individual performance. Techniques relative to biological evolution have been proposed as learning models in improving performance of computer programs since the dividing line between evolving and learning is not clear in computer systems.

Likewise, Mitchell [3] included computational complexity theory and control theory in the numerous disciplines that have influences on machine learning. Examples of computational complexity theory's influence on machine learning are the complexity of learning tasks as measured in terms of the required effort and number of training examples and mistakes, in order to learn. On the other hand, examples of control theory's influence are the procedural controls that learn to optimize defined objectives and predict the next process state [3].

Moreover, Gollapudi [7] added Data Science and Data Mining to the related domains of study where machine learning is closely associated. According to Gollapudi, the fields of machine learning and data mining are intertwined and that there is a significant overlap in their underlying principles and methodologies. He added that data science is a superset of machine learning and data mining. Data science covers the complete process starting from data loading until production.

Machine learning was described by Dey [15] as teaching computers how to efficiently handle data. With the abundance of datasets available, many industries use machine learning to find and interpret patterns and extract relevant information.

Hurwitz and Kirsch [16] explained further the concept of machine learning in their IBM paper dealing with the topic. They claimed that machine learning is a form of artificial intelligence (AI) that uses several types of algorithms to learn and predict outcomes from data rather than through explicit programming. It is a complex process of repeatedly learning from data to improve and make more accurate prediction models.

Another description about machine learning was given by Chao [17], stating that in machine learning, the data and the learning algorithm play a very important role in discovering and learning knowledge. Chao further claimed that the learning and prediction performance greatly depends on the quality or quantity of the dataset.

In a similar line of thought, Jou [4] opined that machines learn by continually observing data. Given enough data, machines can create patterns. Observation of the patterns can lead to generalizations, a process accomplished by taking examples and creating general statements or truths. Jou defined machine learning as algorithms that automate the same learning process that humans naturally do.

In his 2018 paper, Simeone [18] described machine learning as a good alternative to the engineering way of designing an algorithm. Instead of acquiring the needed domain knowledge, the machine learning approach collects enough number of examples of targeted behavior for the chosen algorithm. The training examples serve as input to a learning algorithm producing a model that performs the desired task. Learning is made possible by the selection made by the learning algorithm from a set of possible models during training.

In a nutshell, machine learning is synonymous to saying that a computer can learn from experience [3]. As precisely defined by Mitchell, learning happens if the performance P of a machine to a particular task T improves with experience E as measured by P. Because of this learning, the resulting model can make predictions on data even without additional programming. This is also the reason why machine learning models perform better in the future, added Bali and Sarkar [14].

The learning is obtained from observations on past historical data. This large amount of data is analyzed to find hidden patterns and rules which are then used by the computer in characterizing new sets of data. The process of rule creation continuously improves as more new data are presented, added Pojon [5] and Jou [4]. Moreover, there can be feedback mechanisms to the model to improve the results based on the output. Simply put, this whole system forms a machine learning model which can be used directly on completely new data or observations to get results from, without the need to write separate algorithm to work on same data.

There are primarily two types of datasets required in performing machine learning. The first dataset called the training dataset is used to learn or build the model. The second dataset which will help evaluate the performance of the model is referred to as the testing dataset [7]. A third dataset, referred by Hodeghatta&Nayak [13] as validation dataset, is usually a portion of the training dataset that is used to fine-tune the performance of the model. There are situations where the model performs poorly when deployed with new data. This is known as over-fitting, and is caused by absence of data or data not represented well in the training dataset [19]. Pojon [5] opined that the performance should be measured in multiple test partitions to avoid over-fitting and for possible improvement in the model’s performance.

B. Types of Machine Learning

Machine learning is of two main types: supervised machine learning and unsupervised machine learning [4], [13], [14]. Supervised machine learning is one in which the training dataset is already categorized and labeled into different classes. It requires accurate examples. After the learning, the model can classify new data of unknown labels. The main types of supervised machine learning algorithms follow:

Classification. According to Bali &Sarkar [14], the output classes of classification algorithms are discrete. Classification algorithms include Naïve Bayes, support vectors, random forests and decision trees among others. Bali and Sarkar further claimed that Naïve Bayes algorithm is the most popular classification algorithm used by the machine learning community.

Regression. The output classes in this type are continuous and not discrete. Examples of regression algorithm include logistic regression, linear regression and regression trees. Hodeghatta and Nayak [13] claimed that although logistic regression is still a linear regression, it can be used in cases where the output variable is a discrete value. In other words, Classification predicts categorical class (or discrete values), whereas Regression predicts continuous valued functions. Logistic regression, however, can also handle categorical class prediction [13].

In unsupervised machine learning, additional analysis is needed to fully understand the outputs of the model since the target class is not identified. The goal in this case is to decipher the structure in the data against the build mapping between input and output variables of data [4], [14], [17]. This means that the output variables are not defined in unsupervised machine learning. The output could be another associations or clusters between two variables. Accordingly, unsupervised machine learning is more complicated than supervised machine learning. The main types of unsupervised machine learning algorithms, according to Hodeghatta and Nayak [13], are:

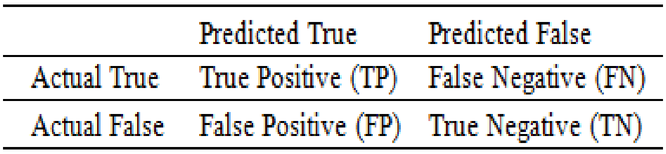
Clustering. Clustering models are built using the likes of hierarchies and means where data items are grouped into different categories based on the features of the input data. Popular clustering algorithms include hierarchical clustering, k-medoids, and k-means.

Association. It uses rules, explaining relationships of different variables, which are extracted from the datasets. It is used to mine and extract rules and patterns from datasets. It also show frequency of data items and depict patterns occurring in the data. Popular association algorithms include FP growth and Apriori.

C. Confusion Matrix and Performance Metrics

The performance of a prediction model can be evaluated by comparing the predicted values against the actual values which are contained in a table called confusion matrix [5], [20], [21].

Table 1Possible Prediction Results



	Predicted True	Predicted False
Actual True	True Positive (TP)	False Negative (FN)
Actual False	False Positive (FP)	True Negative (TN)

Table 1 shows an example of the confusion matrix, containing the possible results of a prediction:

According to Bali and Sarkar [14], Avati [21] and Jovanovic [20], the most frequently used performance metrics for evaluating predictive models include precision, recall, F measure and accuracy. The values of these metrics are computed from the prediction results summarized in the confusion matrix.

Accuracy, which is basically the percentage of correct predictions, describes how close the predictions are to the actual value. Juba and Le [22] claimed that accuracy is the simplest and most widely used performance metric. They cautioned, however, that accuracy is not advisable to use as the metric with imbalanced data. According to them, when the negative class is dominant for example, high accuracy can be achieved by predicting negative most of the time. Accuracy is measured using the formula [22]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision and Recall are related to each other as they measure positive predictions and are used together to make a better evaluation [5]. According to Powers [23], the two metrics capture little information about the number and types of errors made and focus only on the positive cases and predictions. He further claimed that neither precision nor recall includes any information about negative cases or takes into accounts the number of true negatives. A good predictive model must take into consideration both successful positive predictions and successful negative predictions. Precision and Recall are measured with the following formula [20]:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

The last metric, F measure, summarizes Precision and Recall into single value. It is represented by the formula, [5]:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Majority of the reviewed studies on student performance and dropout prediction, made use of accuracy as the only performance metric in evaluating the model. These include the study conducted by Sara et al. [24] on High School dropout prediction in Denmark, Kotsiantis et al. [25] on Distance Education dropout prediction in Greece, Agrawal and Mavani [11] on student performance in India and Kloft et al. [26] on online courses dropout prediction in Germany. Meanwhile, Er [27] made a study on predicting at-risk students in Turkey using accuracy, precision and recall as the measures in evaluating the performance of his prediction model. Two other studies that were reviewed made used of four measures – accuracy, precision, recall and F measure - in assessing the performance of the model. These were the studies conducted by Pojon [5] where he compared the effects of feature engineering and machine learning methods selection in evaluating the prediction model for student performance, and Rovira et al. [28] for their data-driven prediction model performing two different tasks, which are prediction of subsequent course grades of students and prediction of student dropout.

The researcher adopted the four performance measures namely accuracy, precision, recall and F measure in evaluating the prediction models for this study. Some related studies presented their conclusion as to the limitation of the accuracy metric in prediction performance evaluation especially when the class distribution is imbalanced. Precision and recall measures are included to have a good combination of successful positive predictions and successful negative predictions. Moreover, F measure was added to present a single value that takes into consideration both precision and recall values. The combinational use of the four metrics helped the researcher capture a more credible prediction performance measurement between the two supervised machine learning models. The measurements provided an intelligent basis for realistic comparative study outcomes by taking into consideration all possible prediction scenarios.

D. Machine Learning in School Dropout Prediction

Machine learning applications are extensively utilized in our society today. Machine learning is applied into many areas of computing including social media applications and even complex financial systems [16]. Machine learning can be used to improve customer experience, detect patterns and anomalies, predict results from vast complex data and transform business operations. There are numerous examples in almost every industry. Hurwitz and Kirsch [16] gave the following examples of how machine learning can be applied to solving complex business problems: selecting the most effective treatment with fewer side effects given a patient’s medical conditions, age, gender and medical history; building models to predict future mechanical problems and the potential for failure based on the enormous amount of data from Internet of Things (IoT) sensors; responding proactively to potential IT problems by feeding complex IT operations data of the organization to the model and also help identify patterns for better performance of the computing environment; and identifying an anomalous or an unauthorized actions and blocking intrusions before damage can occur.

Machine learning can be found in various sectors including retail, advertising, healthcare, cyber security, transportation and the academe. Some specific real-world machine learning applications given by Gollapudi [7] and Bali and Sarkar [14] include: (1) Face Detection – widely used in today’s social media websites and cameras that provides an ability to tag a person across many digital photographs; (2) Speech Recognition – used in automated call centers and cellphones where a user’s request on the phone is interpreted and mapped to one of the tasks for execution; (3) Spam Detection – used to categorized an email and marked as spam based on some rules that it builds using some key features of email data, (4) Customer Segmentation/ Product Recommendation – used by online shopping and social media websites like Amazon, Facebook and Google+ which identify the products the customer will most likely be interested in buying; (5) Credit Card Fraud Detection – identify any transaction that is not potentially made by the customer and marked them as fraudulent for necessary action to be taken based on the usage patterns of the credit card by the customer and the purchase behavior of the customer; (6) Sentiment Analysis – used by political strategists to gauge public opinion on policy announcements or campaign messages based on opinion shared by others; and (7) Self-driving vehicles.

In the academe sector, one of the areas where machine learning can be of big help is dropout prediction. Several studies related to the application of machine learning in dropout prediction were reviewed. Most of these studies compared several algorithms to find which algorithm offers the most accurate results.

Sara et al. [24] in 2015 conducted a study in Denmark comparing several algorithms to predict high-school dropout. The study made use of a considerably large sample size of 72,598 high-school students. The best results were obtained using Random Forest with an accuracy of 93.5%. The researchers concluded that machine learning can be used to accurately detect possible high-school dropout given the information available to the school.

In Germany, a study made use of collected click-stream data of participants from a psychology Massive Open Online Courses (MOOC) course to train the dropout prediction model. The whole course which lasted for 12 weeks has 11,607 participants in the beginning week and only 3,861 participants staying until the last course week. In the study, the researchers proposed Support Vector Machines algorithm for a machine learning model that will predict dropout between MOOC course weeks [26].

In 2012, ErkanEr [27] studied in Turkey, two semesters log data to accurately predict at-risk students in an online course. Three machine learning algorithms, instance-based classifier K-Star, Naïve Bayes and decision tree C4.5, were evaluated and compared in classifying students whether failure or successor. Results of the study indicated that K-star reached 82% rate, which is the highest accuracy among the three machine learning algorithms. Naïve Bayes achieved 81% while C4.5 achieved 79% accuracy.

A similar study was conducted in Greece with 354 participating students. The objective of the study was to determine if machine learning techniques can prevent student dropout in distance learning and to identify the best algorithm to solve the problem. It was reported that Naïve Bayes algorithm gave the best results, exhibiting prediction accuracies of 63% for the beginning of the academic period and 83% for the remaining period [25]. The researchers claimed that the study is the first work that predicts dropouts using machine learning techniques.

Another significant study was conducted in India where 80 Information Technology students were used to train two algorithms – Neural Networks and Naïve Bayes classification. In the study, Neural Network algorithm outperformed Naïve Bayes classification with an accuracy rate of 70.48%. Although the researchers were surprised by the results, they later realized that the results are justified by the continuous input data type used since Naive Bayes requires discrete data. In the study, it was also found that previous year's grade and residence location are the variables with the highest potentials for predicting student performance [11].

Likewise, a study was conducted in Spain in which five machine learning classifier methods namely, Random Forest, Linear Regression, Adoptive

Boosting, Naïve Bayes and Support Vector Machines, were compared in predicting student dropout. The research made use of a total 4,434 students studying Law, Mathematics and Computer Science at the University of Barcelona [28]. Results of the study showed that non-parametric model such as Naïve Bayes and Linear Regression perform better when using small data set. Further, it concluded that parametric models like Random Forest and Adoptive Boosting are recommended for projects using larger data sets.

Although one can readily conclude from the above-mentioned studies that machine learning can successfully predict dropout, this researcher cannot claim with certainty which among the algorithms (supervised – classification, regression; or unsupervised – clustering, association-rule) offers the most effective solution because of the contrasting results. This study focused on developing two prediction models using Naïve Bayes and Logistic Regression algorithms only, and subsequently evaluated and compared the models' performances. It was a good comparative exercise between two algorithms of same supervised learning category but different methodology and making use of the same dataset with the same categorical output type. The contribution of this study is the side-by-side comparison of the performances of the two types of supervised machine learning algorithms to determine which type of algorithm offers the better solution in college dropout prediction.

E. Predictors of College Dropout

Majority of related researches on dropping out made use of student academic performance and personal characteristics as predictors.

In the study conducted by Tan and Shao [29], the dataset used is composed of variables related with the students' academic performance and personal characteristics. Majority of values of these variables are extracted directly from the Academic Management System of the school while the values of the other variables are computed from data extracted from the same source. Although, several other factors associated with dropping out were identified in the study, these were not included as predictors because of difficulty in obtaining data related to the factors. These factors include the curriculum, academic support services, teacher-student relationship, family support and study motivation factors.

In India, a study was conducted by Agrawal and Mavani [11] where they proposed a model for predicting student performance. In the study, they considered the importance of several variables and further determined the top variables that are correlated with student performance. Results of the study showed that the top factors contributing to student's performance are previous grades and living location.

Meanwhile, Sara et al.'s [24] significant research on high school dropout prediction in 2015 used information from public online sources, including government statistics and travel planner and the MacomLexio study administration system. The goal of the study is to build a prediction system which can identify students at risk of dropping out and generate report for the teachers. The system was able to include information about students' performance during the previous semester. They were also able to classify the dropouts according to the reasons specified by the teachers. Student performance data was augmented with gender, travel time to school, average income per postal code, school and class size and teacher pupil ratio.

An interesting study whose overall goal was to propose a model that can reliably predict students at-risk of dropping out in an online course was conducted by Er [27]. In the study, only time-varying data specifically attendance, midterm exam, final exam and assignments were used and time-invariant attributes like gender and age are excluded. The study found out that using time-varying data only is enough to obtain reliable prediction, and that no significant change on overall prediction results were noted when time-invariant data were excluded.

Another study using an online university's student database was conducted by Niemi and Gitin [30] in 2012. The study was aimed to determine how students' academic performances and demographic characteristics affect dropout rates. Demographic variables included were age, status, gender, previous college education, number of transfer credits from other schools, military status, and estimated family financial income. Academic variables included student performance measures like final exam, assignment and discussion project scores. The study showed that student performance measures that declined over time were good predictors of possible dropping out. Some surprising findings include that females, unmarried and students with transfer credits are more likely to drop out.

Moreover, Rovira et al. [28] developed prediction systems for academic grades and dropout using student records at University of Barcelona. The information which was collected by a University personnel, consists of the student final grades covering all academic years. The study claimed that the prediction systems are successful while using only grades in training the systems. Using grades as the only predictors would allow the systems to easily adapt for other degree studies and universities, according to the developers.

The findings in the reviewed literatures indicated that the most widely used predictors for dropout prediction are factors that are classified under student demographic characteristics and academic performances. Availability of data was also identified as an important factor to consider when selecting dropout predictors. Specifically, the data which are found to be highly correlated to dropping out includes previous grade and residence location. The researcher realized that similar predictors could be utilized in this study due to the availability of the data in the enrollment system of the researcher's academic institution. However, this study is different from the other related works due to its use of general weighted average and percentage of dropped subjects as measures of student performance.

F. Feature Engineering

Feature engineering is one of the most important and time consuming tasks in machine learning projects. It is the process of selecting or creating features or variables in a dataset and transforming them into formats that are suitable for the model. Having the correct variables can make the modeling process easier and will produce more accurate results [31], [19]. Variable selection can include removing unnecessary or redundant variables and this requires assessing the relevance of the variable. Variable creation involves modifying existing variables and combining different ones to create new variables [5].

Variable selection is a subset of feature engineering and it is usually done after the data collection process. Sagar [32] defined variable selection as the process of identifying the best set of variables that should be fed to the training models for better prediction results. Brownlee [33] added that variable selection methods are used to remove irrelevant and redundant variables from the dataset that may decrease the accuracy performance of the model.

One of the methods used in variable selection is measuring the importance of the variables by identifying which variables are used by the model and their contributions towards solving the problem at hand.

The summary function in regression and the varImp() function of the caret package are suggested by Sagar [32] as good methods in implementing variable selection. The regression summary function can be used to calculate variable importance as it describes variables and how they affect the dependent variable through significance. It works on variance and marks all variables which are significantly important. Such variables usually have a p-value less than 0.05 indicating that confidence in their significance is more than 95%. The output gives the estimates and probability values for each of the variables. It marks the important variables with stars based on p-values. For variables whose class is a factor, the variables are broken down on the basis of each unique factor level. Similarly, the varImp() function of the caret package in R, provides additional insight on the variables with high weightage and used frequently by the model.

Another method is the wrapper method. The wrapper method selects variable sets like a search problem where different sets of variable combinations are evaluated and compared against other combinations. It assigns a score based on model accuracy. One of the best ways for implementing variable selection with wrapper method is to use the Boruta package that finds the importance of a variable by creating shadow features [34].

The use of feature engineering in this study is the modification of variables. The variables were categorized by the researcher such that the ranges of possible values are limited. An example of this modification was made with the location variable containing the name of town where the student resides. Instead of the town name, the variable can be assigned a value of “others” to represent the locations which are very far from the school or locations that contain just one or two values.

III. METHODOLOGY

A. Research Design

The study is a combination of causal comparative, descriptive and developmental type of research. It showed the cause-effect relationship between the independent variables (percentage of dropped subjects, general weighted average, residence location and gender) on one side and the dependent variable (drop or continue action of student) on the other side. It developed two models, using two types of supervised machine learning algorithms namely Naïve Bayes and Logistic Regression, and determined the better solution in predicting college dropout. A web-based prediction system was developed, with the chosen algorithm at the core, using RStudio Integrated Development Environment (IDE) and the Shiny web framework package.

The web-based system is straightforward. The system can calculate predictions of all students enrolled in a given semester during a given school year. This process can be performed after the user has inputted the identification number, general weighted average (GWA), percentage of dropped subjects, location value corresponding to the town or city name where the student resides and the gender value of each student. Likewise, the summary of the prediction results can be viewed on the screen or printed out on paper.

The application was deployed at shinyapps.io where shiny applications can be published for free.

B. Participants of the Study

The main participants of the study are the 687 Information Technology and Computer Science students of University of Luzon whose academic records are downloaded from the University’s School Automate System (SAS). The academic records of the students, who are enrolled in the school year 2015-2016, form part of the dataset used in the study. The aforementioned school year was selected since the K-12 program was implemented in the Philippines the following year, thereby anticipating a great decrease in the first year tertiary level enrollees during that period. Having fewer enrollees would translate to having lesser participants, and that scenario is not beneficial to the study.

C.Data Analysis

The quantitative data were organized, analyzed and interpreted using the following statistical tools:

Frequency Count (f) and Percentage (%). This was used to compute for the accuracy, precision, recall and F measure average values from 10 consecutive prediction runs using 70-30 training-test data partition.

T test. This was used to determine the significant difference between the prediction performance using the original dataset and the engineered dataset in terms of accuracy, precision, recall and F measure.

IV. SUMMARY OF RESULTS

A. Predictors of College Dropouts

General weighted average, percentage of dropped subjects, location of residence and gender of students are good predictors of college dropout. General weighted average and percentage of dropped subjects are computed values from the subject grades and subject units of the students. These indicators underwent several data cleaning activities and variable importance tests before using in the prediction model.

B. Supervised Machine Learning Algorithms to Predict Dropouts

Both Naïve Bayes and Logistic Regression algorithms made successful dropout predictions on new untested data surpassing the dataset baseline accuracy. Using the original dataset, Naïve Bayes posted an accuracy rate of 87.5% and Logistic Regression an accuracy rate of 86.4% as against the dataset accuracy rate of 85.9%. Likewise, using the engineered dataset, Naïve Bayes' accuracy rate of 89.4% and Logistic Regression's accuracy rate of 87.3% are higher than the baseline accuracy rate.

C. Application Tool for Dropout Prediction

An application tool for predicting college dropout can be developed using the Shiny application development package of RStudio. The application can interface smoothly with the Naïve Bayes prediction model since both operates on the same R platform.

D. Performance of Supervised Machine Learning for Dropout Prediction

Naïve Bayes performs dropout prediction better than Logistic Regression using the original dataset. Naïve Bayes' accuracy rate of 87.5% and F measure rate of 92.7% are higher than Logistic Regression's accuracy rate of 86.4% and F measure rate of 92.2%.

E. Performance of Feature Engineering for Dropout Prediction

Using the engineered dataset, Naïve Bayes performs dropout prediction better than Logistic Regression. Naïve Bayes' accuracy rate of 89.4% and F measure rate of 93.8% are higher than Logistic Regression's accuracy rate of 87.3% and F measure rate of 92.8%.

F. Significant Difference Between Supervised Machine Learning and Feature Engineering

There's enough evidence to show that there's significant difference in the prediction performance between using the original dataset and using the engineered dataset in terms of accuracy, precision, recall and F measure.

G. Enhancements to Improve the Prediction Application

The developed dropout prediction application could be improved by regularly adding new student records in the training dataset. The prediction performance of the model is expected to improve by adding related variables including incomplete grade and type of residence. Likewise, combination of variables as feature engineering approach should be used in tandem with the categorization of variable for a possible improvement in prediction performance.

V. CONCLUSION

What follows are the conclusions drawn from the results of this study:

Supervised machine learning could predict college dropouts with Naïve Bayes algorithm performing better than Logistic Regression. Feature engineering could improve the prediction performance of Naïve

Bayes and Logistic Regression algorithms. Moreover, a user-friendly and functional dropout prediction system could be easily developed and deployed in the R platform using RStudio's Shiny user interface application package.

REFERENCES

[1]. Orion Jr, H., Forosuelo, E., Cavalida, J. (2014). Factors Affecting Students' Decision to Drop Out of School. Retrieved on September 14, 2018 from https://rpo.cjc.edu.ph/index.php/slongan/article/download/4/5/.

[2]. Tumapon, T. (2017). Creating A Culture Of Student Engagement. Retrieved on August 4, 2018 from http://www.manilatimes.net/creating-culture-student-engagement-2/349379/.

[3]. Mitchell, T. (1997). Machine Learning. McGraw-Hill Science/Engineering/Math.

[4]. Jou, S. (2017). Machine Learning: A Primer For Security. ISSA Journal, January 2017.

[5]. Pojon, M. (2017). Using Machine Learning To Predict Student Performance. Retrieved on July 12, 2018 from https://tampub.uta.fi/bitstream/handle/10024/101646/GRADU-1498472565.pdf.

[6]. Nilsson, N. (1996). Introduction to Machine Learning. Department of Computer Science, Stanford University.

[7]. Gollapudi, S. (2016). Practical Machine Learning. Packt Publishing Ltd. Birmingham UK.

[8]. Ray, S. (2017). Essentials of Machine Learning Algorithms (With Codes in Python and R). Retrieved on August 5, 2018 from the Analytics Vidhya website: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/.

[9]. Perceptive Analytics, (2018). Understanding Naïve Bayes Classifier Using R. Retrieved on July 4, 2018 from https://www.r-bloggers.com/understanding-naive-bayes-classifier-using-r/.

[10]. Le, J. (2018). Logistic Regression in R Tutorial. Retrieved December 15, 2018 from https://www.datacamp.com/community/tutorials/logistic-regression-R.

[11]. Agrawal, H. &Mavani, H. (2015). Student Performance Prediction Using Machine Learning. International Journal of Engineering Research and Technology. Vol 4 Issue 3, March 2015.

[12]. Venables, W., Smith, D., R Core Team (2018). An Introduction To R. R Manual 3.5.0 (2018-04-23).

[13]. Hodeghatta, U. &Nayak, U. (2017). Business Analytics Using R – A Practical Approach. Springer Science Business Media New York.

[14]. Bali, R., Sarkar, D. (2016). R Machine Learning By Example. Packt Publishing Ltd.

[15]. Dey, A. (2016). Machine Learning Algorithms: A Review. International Journal of Computer Science and Information Technologies. Vol. 7, 2016, 1174-1179.

[16]. Hurwitz, J., Kirsch, D. (2018). Machine Learning for Dummies. John Wiley & Sons, Inc. NJ.

[17]. Chao, W. (2011). Machine Learning Tutorial. Graduate Institute of Communication Engineering, National Taiwan University, available at http://disp.ee.ntu.edu.tw/~pujols/Machine%20Learning%20Tutorial.pdf.

[18]. Simeone, O. (2018). A Very Brief Introduction to Machine Learning With Applications to Communication Systems. Retrieved on April 14, 2019 from https://arvix/pdf/1808.02342.

[19]. Domingos, P. (2017). A Few Useful Things To Know About Machine Learning. Retrieved on July 3, 2018 from http://cs.washington.edu/homes/pedrod/class.

[20]. Jovanovic, J. (2015). Classification. Retrieved on August 20, 2018 from http://jelenajovanovic.net/Classification-basic-concepts-2015.pdf.

[21]. Avati, Anand (2017). Evaluation Metrics (Classifiers). Retrieved July 5, 2018 from

http://cs229.stanford.edu/section/evaluation_metrics.pdf.

[22]. Juba, B. & Le, H. (2018). Precision-Recall versus Accuracy And The Role Of Large Data Sets. Retrieved on July 5, 2018 from www.aaai.org.

[23]. Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness& Correlation. Journal of Machine Learning Technologies. Volume 2, Issue 1, 2011, pp 37-63.

[24]. Sara, N., Halland R., Igel, C., Alstrup S. (2015). High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. Retrieved on July 4, 2018 from http://www.i6doc.com/en/.

[25]. Kotsiantis, S., Pierrakeas, C., Pintelas, P. (2003). Preventing Student Dropout In Distance Learning Using Machine Learning Techniques. International Conference on Knowledge-Based and Intelligent Information And Engineering Systems. pp267-274.

[26]. Kloft, M., Stiehler, F., Zhend, Z., Pinkwart, N. (2014). Predicting MOOC Dropout over Weeks Using Machine Language Methods. Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 60-65.

[27]. Er, E. (2012). Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100. International Journal of Machine Learning and Computing. Vol 2, No. 4, August 2012.

[28]. Rovira, S., Puertas, E., Igual, L. (2017). Data-Driven System To Predict Academic Grades And Dropout. Retrieved on July 4, 2018 from https://doi.org/10.1371/journal.pone.0171207.

[29]. Tan, M., Shao, P. (2015). Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. Retrieved on June 15, 2019 from http://dx.doi.org/10.3991/ijet.v10i1.4189.

[30]. Niemi, D., Gitin, E. (2012). Using Big Data to Predict Student Dropouts: Technology Affordances for Research. IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2012).

[31]. Lam, H., Thiebaut, J., Sinn, M., Chen, B., Mai, T., Alkan, O. (2017). One Button Machine for Automating Feature Engineering in Relational Databases. Retrieved on August 6, 2018 from https://arxiv.org/pdf/1706.00327.pdf.

[32]. Sagar, C. (2018). Feature Selection Techniques With R. Retrieved on June 5, 2019 from https://dataaspirant.com/2018/01/15/feature-selection-techniques-r/.

[33]. Brownlee, J. (2018). Machine Learning Mastery with R. Retrieved July 4, 2018 from http://machinelearningmastery.com/machine-learning-with-r/.

[34]. Kaushik, S. (2016). Introduction to Feature Selection Methods With an Example. Retrieved on May 7, 2019 from https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/.