International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Feb-24)

http://dx.doi.org/10.12785/ijcds/150151

Forecasting Students' Success To Graduate Using Predictive Analytics

Jayrhom R. Almonteros[1], **Junrie B. Matias**[2] **and Joanna Victoria S. Pitao**[3]

[1,2,3]*College of Computing and Information Sciences, Caraga State University, Butuan City, Philippines*

Received 28 Jun. 2023, Revised 6 Jan. 2024, Accepted 21 Jan. 2024, Published 1 Feb. 2024

Abstract: Predictive analytics is the process of forecasting outcomes based on historical data. Execution of predictive analytics involves several phases, namely: data collection, analysis and massaging, identifying machine learning, predictive modeling, predictions, and monitoring. All phases play a vital role in the prediction's result, especially the data analysis and massaging or data preprocessing. This study aims to predict the students' probability of graduating on time using the students' demographic profiles, previous academic achievements (SHS track and grade point average), and college admission results (english, math, science, and abstract). The dataset was acquired from Caraga State University with 2207 samples of new entrants. This study implemented KNN to impute numerical data, while mode imputation was used for categorical values. Moreover, binary encoding was employed for nominal data to prevent the algorithm from ranking the values in order. Seven (7) algorithms were tested on the original dataset and compared to datasets integrated with LASSO Regressions (L1), Ridge Regression (L2), and Genetic Algorithm (GA) separately. The algorithms involved were Decision Tree, Random Forest, Ensemble, KNN, Logistic Regression, SVM, and Na¨ıve Bayes. The result shows that LASSO Regression (L1) with the Decision Tree classifier has the lowest accuracy (58%) and AUC score (50%). It also has the smallest number of features selected (5). Conversely, GA selected thirty-three (33) features with an AUC score of 71% and predicted 79% accurately using the Logistic Regression classifier. It exhibited a 21% increase in the AUC score compared to the no feature selected dataset (NFS) with the same classifier.

Keywords: feature selection, genetic algorithm, predictive analytics, prediction

1. INTRODUCTION

Predictive analytics is the procedure of extracting information from a data set to forecast future outcomes [1]. Various sectors may employ prediction in their procedure [2]. Insurance sectors may recognize clients with a high likelihood of attaining illness; through this, the target client to be offered insurance plans could be known. On the other hand, retail may study the customers' reaction towards a product and oil and gas to project the resources needed. Despite being applied to diverse sectors, it shares the exact purpose of "acquiring new information based on the historical data," bringing advancement to the company by undertaking necessary actions and interventions based on the prediction result. Another advancement, once integrated into software development, may improve service quality, which is identified as one of the motivating factors affecting users' intention to use the application [3]. In fact, [4] shows that about 80% expressed interest in engaging in predictive analytics for their three-years-establishment-plan as part of their operational process.

Execution of predictive analytics involves several phases: requirement collection, data collection, data analysis and massaging, statistics/machine learning, predictive modeling, and predictions and monitoring [5]. Requirement collection encompasses defining what data the client is involved in, the aim of prediction, and its benefits. The data were then collected, containing all the available variables defined in the first phase. Data analysis and massaging involve structuring the data, which addresses missing values and cleaning attributes to prevent possible erroneous data. After that, predictive modeling can be processed with the selected modeling technique; it could be statistical or machine learning techniques.

An evident application of machine learning techniques in predictive analytics that solves education-related problems also exists. Alipio [6] developed a model that predicts the academic performance of first-year college students in the Philippines using path analysis. His study concluded that academic adjustment and performance are affected based on the SHS strand taken by the student and was also supported by his follow-up study in the same year. Aside from that, the difficulty level of college subjects is also intensively related to the strand taken during senior high school [7]. The problem now is the presence of a

E-mail address: jralmonteros@carsu.edu.ph, jbmatias@carsu.edu.ph, jvcsaga@carsu.edu.ph

high number of mismatched SHS strands; this means that their preparatory education does not directly align with their college courses, thus defeating the objective of K12 implementation [8] [9]. Other vital predictors contributing to college academic performance are the admission test score and the high school GPA [10].

The different pronouncements from the present studies of which predictors best forecast the students' performance open an area of research using real-world data from Caraga State University. With this, the work aims to develop a web-based application that forecasts the success to graduate of a student. In achieving this goal, specific objectives are as follows:

- Identify the valuable predictors in forecasting the students' success to graduate through the implementation of feature selection methods: LASSO (L1) Regression, Ridge (L2) Regression, and Genetic Algorithm (GA);
- Develop predictive models using different classifiers, namely Decision Tree, Random Forest, Ensemble, KNN, Logistic Regression, SVM, and Naïve Bayes;
- Distinguish and implement the best-performing model through comparison of the accuracy and AUC score of the developed models in the developed application

This study intends to contribute to the body of research by adding new findings in the field of education and predictive analytics. Furthermore, since the data acquired contains the pioneer of the K-12 implementation in the Philippines, it is vital to inspect this together with other pre-admission data. The findings will also provide a basis for policy-making or modification of the present admission selection process in the university. At Caraga State University (CSU), the only criterion to be admitted as a new entrant is passing the entrance exam. However, admission has become more rigid since the passing of the [1]Republic Act No. 10931, known as the "Universal Access to Quality Tertiary Education Act." Due to this, the number of takers increased, requiring a higher entrance score and undergoing different procedures before the student was admitted. However, based on the dataset obtained, there is a low number of students admitted in 2018 who graduated on time. The low number of graduates suggests that admission score alone is not enough basis for the predictor in forecasting the "success to graduate" of a student in the CSU setting.

2. RELATED WORKS

A. *Understanding Predictive Analytics Process and its Challenges*

'What will happen?' is the central question concerning predictive analytics [11]. Predictive analytics uses historical data that represents future trends to forecast outcomes. Integrating a prediction could be done through supervised or unsupervised learning. The critical difference between the two is that the supervised uses a target variable while the latter does not; this means that the supervised has prior knowledge about the dataset through label [12]. According to Kumar & Garg [5], predictive analytics undergo seven stages explicitly: requirement collection, data collection, data analysis and massaging, statistics or machine learning, predictive modeling, and predictions and monitoring. [13] describes these phases as data collection, data cleansing, model generation, and evaluation. It differs only in terminologies, but both cover the same process. Each phase comprises several considerations and has a different technique to complete a phase successfully. To further understand predictive analytics, this section discussed the definition of each phase and how other researchers address the challenges encountered.

The first phase is called requirement collection, which analyzes what specific prediction is to forecast. The end goal of the prediction must be evident in the first place, and that must be defined in this phase. For instance, a prediction could be straightforward 'yes or no' only, such as predicting if a credit card is a fraud or not [14]. It could also forecast more than two classifications, such as [15] recognizing chronic kidney diseases or early detection of possible heart disease [16]. The latter study aims to classify if the patient may have either coronary artery, vascular disease, heart rhythm disorder, structural heart disease, or heart failure in the future. It is essential to state the goal of the prediction explicitly. Aside from classification, the prediction's output may also be numerical, such as employing this in sales forecasting [17]. The first phase in applying predictive is pigeonholing the goal of the prediction. With this, the possible data to be collected could be identified, leading to the data collection phase. Data collection is simply the process of acquiring a dataset required to develop the predictive model [5]. However, almost all of the raw data acquired needs to be structurally ready before feeding into developing a predictive model [18]. Hence, it needed to be cleaned and revised to correct errors and handle missing values [13]. This phase is called data analysis and massaging or data preprocessing. Thus, challenges and issues were discussed in the following sections.

Once the data is preprocessed and converted into a structural form that is ready for predictive modeling, the next phase is the election of either statistics or machine learning techniques to use in forecasting. All the predictive analytics models are based on statistical and/or machine learning; however, machine learning techniques have an advantage over the other [5]. Machine learning focuses on forecasting, while traditional statistics explains the relationship between variables [19]. Nevertheless, machine learning improved model discrimination compared to conventional statistical approaches [20]. The performance of prediction results also differs in relation to the dataset and the technique incorpo-

[1] Republic Act No. 10931, known as the "Universal Access to Quality Tertiary Education Act, provides free tuition and other school fees in state universities and colleges in the Philippines.

rated, especially if preprocessing was considered [21]. The study of Osisanwo et al [22] explored different algorithms to determine the most efficient classification algorithm. Two datasets were used containing 768 and 384 samples; though classifiers do not rank the same to both datasets, it could be seen that all classifiers increased in accuracy compared to the smaller dataset, thus showing that a larger data set is more effective in classifying.

Predictive Modeling is a process based on statistical or machine learning techniques that are tested by partitioning the dataset into training and test datasets [5]. This is done to evaluate the integrated machine-learning algorithm. Muraina [23] stated that most scholars' suggestion is to split the dataset with 100 – 1,000,000 into a 70/30 ratio; otherwise, 90/10. Randomized or cross-validation, on the other hand, is also the standard method of splitting the algorithm's performance [24].

After the predictive modeling, the last phase is the model's deployment or the ”prediction and monitoring” [5]. After developing the model, the best algorithm could be implemented using the Django web-based framework, the same as the stock market price prediction [25]. The Django framework offers an easy-to-use library and is scalable in rapid development. Indeed, automation is possible with the Django framework as a web-based application that automates the student schedule following a decision tree-based rule that was successfully implemented [26]. The monitoring takes place by evaluating its prediction using the new data. It is an unending task to ensure that the model is able to forecast effectively beneficial to the company's decision-making process and primarily used in marketing and sales [4].

B. Missing Data

Pre-processing the data may include treatment of missing data. Missing data may lead to inaccuracy of prediction [18]. It was found out by Nijman [27] through their literature review that no sufficient information for handling missing data was presented in most prediction models using machine learning.

Several strategies address missing data problems: listwise deletion, mean-mode substitution, and imputation. Listwise deletion is the easiest way among strategies but is the least recommended. Listwise deletion involves removing data that leads to data reduction that may affect the predictive model's performance [28]. Mean-mode substitution substitutes the mean for numerical and the most common value to missing categorical data. Imputation is substituting the estimated value for the missing data [29]. Imputation comes in many methods; MICE and KNN imputation are among the popular methods for handling missing data. These two strategies were also found to perform best [30], but MICE is a complex algorithm and works well in small datasets, which gives KNN a lead over MICE [31]. In addition, KNN imputation also performed best among other methods in a numeric dataset [29]. The comparison includes mean, median, predictive mean matching, and linear and Bayesian regression methods. However, for non-numeric and nominal, mode was used to replace missing data when comparing the performance of KNN (N=5) and Mean-median imputation train and both accuracy at 99 and above [32].

C. Feature Selection

Moreover, aside from solving missing data, choosing features efficiently will lead to better prediction results [16]. Feature selection is a tool that provides a list of significant features to prevent computational overload [33]. Features or columns not related to other features are considered noises, which causes a low prediction score. Choosing a feature selection technique depends on the problem. Supervised learning consists of three feature selection techniques – filter, wrapper, and embedded.

Filter-based feature selection is a technique that chooses the significant features. It is faster than a wrapper; however, its downfall is that it does not consider relationships between features. Also, it does not associate with the classifier algorithm, an advantage of an embedded technique. On the other hand, the wrapper and embedded look for a relevant subset that a filter-based solution lack [34].

The Least absolute shrinkage and selection operator (LASSO) and ridge regression, a filter-based feature selection both use regularization to prevent overfitting. LASSO uses L1 regularization to shrink the coefficients of less important features to zero, leading to a decrease of the total feature. On the other hand, ridge regression uses L2 regularization. It shrinks the coefficient of all the features but not to zero, unlike L1. LASSO regression is said to work better in small number features while the later on large predictors [35]. The study of Zhang et al. [36] used the LinearSVC algorithm with Lasso (L1) regularization as a feature selector and showed a high-accuracy prediction result. Though the said study dealt with a binary classification problem - which is the same as the problem that this work is trying to solve, it is also essential to consider that the wrapper chose more imperative features than the filter method applied in a classification problem [37]. Though Genetic Algorithm (GA), a wrapper feature selector, was not included in the mentioned research, the later year itemized that GA showed promising outcomes. The experimental study explored five (5) dataset classification problems and thus concluded that the dataset with feature selected by GA outperformed classifiers using the original feature and other feature selectors [38].

D. Data Encoding

The problem mostly with feature selection is that approaches were designed for numerical values [39]. We can assign a numerical value in each category; however, it is only acceptable if the data is ordinal [40]. Ordinal coding is practical when data implies order or ranking [41]. For instance, the salary range is from 1,000-501, 500-201, and 200-1, thus encoded as 3,2,1; this could be interpreted as 3 <

2 < 1, which is true since 1,000 < 500 < 200. In some cases, red, blue, and green will be numbered as 3,2,1, respectively. It will lead to an interpretation of red < blue < green, which is false. The findings showed that ordinal encoding gained the lowest accuracy rate of 81% compared to the other 7 encoding methods. With this, converting non-numerical and non-ordinal/nominal features is handled in another way.

According to Seger [42], one-hot can be used, feature hashing or binary encoding to convert a categorical feature to a numerical value. The most popular approach is One-hot encoding (OHE). Using the OHE approach, each category represents a dimension where the size of the dimension is equal to the number of categories, but only one space is equal to 1; the rest is zero, thus making each category unique [39]. The disadvantage of this approach is that the feature's dimension is also significant when there are many categories, leading to storage and efficiency problems [42]. Meanwhile, feature hashing can solve this issue. Feature hashing is implemented variously, but all use a hash function, thus reducing the encoding size of non-numerical data; however, it is primarily used in large-scale datasets. The last technique is binary encoding, where a number is assigned to a category first before converting to a corresponding binary value and then divided into columns. It is said to take the size of log2, which is smaller than one hot encoding.

E. Students' Predictors to Academic Success

Over the past years, several studies have been conducted concerning important features to forecast student academic success. In relation to this, Alyahyan & Düştegör [43] discussed the best practices in predicting academic success through a literature review, thus enumerating the predictors used by other studies. The use of prior academic achievement of the student has the highest number of reoccurrences as a predictor of the mentioned problem. It was also followed by students' demographics, environment, psychological, and e-learning activity.

The presence of the last three predictors is evidence of an underlying relationship between the student's environment, psychology, and e-learning activity. However, the data could only be captured once the student was admitted to the university. More likely, it could only be acquired after one semester, which defeats the primary purpose of this study since it aims to predict student success before its official admission. Therefore, the study will adopt the first two sets of predictors, which, according to the literature review conducted by [43]: prior academic achievement as predictors in a dataset was used at 44% of all existing research related to student success. It consists of pre-admission data such as test results, GPA, and high school background as influencing factors. Secondly, the student demographics (25%), which predictors include gender, age, residence, parent's education, occupancy, and family income.

Furthermore, more recent research has supported the importance of the student's prior academic achievement as a contributor to their college success, particularly the SHS track [7]. Nevertheless, admission test scores and high school GPA were highly studied, and it concluded that both are potent predictors as contributors to drop rate [10] [44] [45]. The existence of the mentioned predictors in the literature shows their relevance as factors contributing to a student's success. Moreover, these influencing factors were available in the pre-admission data at the university and, thus, will be included in this study. Table I summarized the goal, findings, and how these previous researches contribute to this paper.

F. Researches in Predictive Analytics Domain

Table II contains studies conducted by various researchers in the field of prediction. Each study compared several algorithms in developing a predictive model. The researchers of [46] [47] conducted an extensive review of related studies to identify the most frequent machine learning methods. Both researchers agreed that Random Forest, SVM, and Naive Bayes were among the most utilized algorithms; however, the latter concluded that Random Forest has the highest accuracy rate, while [46] declared Decision Tree.

The pronouncement of both studies with different algorithms as the best-performing algorithm is essential to this study since it also deals with prediction using student data; this shows, however, that results also depend on the preprocessing procedure, the data itself, and the algorithm. It also must be noted that studies included in the Table II deal with either binary problem classification [14] or school-related problems, such as the possibility of a student dropping out [45].

Moreover, Table II shows the ensemble model, logistic regression, and KNN as reoccurring methods. With this, it is relevant not to conclude that the algorithm declared in the mentioned studies will also perform well in the dataset used in this research. Hence, each reoccurring method will be included in the algorithms to be modeled and compared based on AUC and Accuracy. Nevertheless, there is no conflict in their findings since the subsequent studies on predictive analytics but in different datasets which structure is unidentical with each other.

3. METHODS

To carry out the needed process in developing a predictive model, necessary data and methods that will aid the completion of this study were identified. Figure 1 illustrates the experimental design conducted based on the findings of the review of literature conducted. The dataset was acquired from the university and underwent preprocessing, given that the dataset contains missing data and non-numerical values, as shown in Table III. Once it was handled, the preprocessed dataset went through feature selections separately, and each selected predictor was used in the modeling in seven different classifiers. The succeeding subsections will describe the experimental design in detail.

TABLE I. Researches Investigating Features Related to Student Success

Title	Research Goal	Findings/Conclusion	Contribution to this study
[43]	A literature review was conducted to provide guidelines, research methods, and access to data mining techniques involved in predicting student success.	Prediction of student's performance in the early stage improves student's success rate.	The investigated researchers from previous years understood the different possible datasets to collect. Prior academic achievement and demographics show high occurrence, suggesting interest among researchers. This study combined the mentioned features rather than investigating them separately.
[7]	Examine the implication to the college students who took the misaligned SHS track and their performance in the college performance.	A different level of difficulty for STEM and non-STEM students is present. Students from the STEM SHS Track outclassed the other.	The research focused alone on implementing the K-12 Curriculum and the effect of misalignment of track during college. The finding is evident that SHS Track must be included as a feature in the dataset. The fact that the dataset contains the first batch who graduated college after K-12 implementation will contribute to the body of research in the Philippines.
[10]	Investigated the relationship between admission scores, high school grade point average, and academic performance in business students.	A combination of high school GPA and standardized admission results is the best predictor for considering student admission.	The study focused only on a specific undergraduate business program. This study shares the same sentiment in finding features in considering student admission; however, it does not limit the scope to a specific program.
[44]	The paper reviewed academic preparation and college readiness, thus proposing a recommendation for policy-making for advancing college graduate rates.	High school grades are associated with college readiness than test scores. It was also part of the recommendation to screen students instead of admit students who volunteer to take particular coursework.	The study suggested enrolling in a program that fits them. Though K-12 had been implemented, the mismatch is present. In this study, the predictive model forecasts student success to graduate relating to the program it chooses. It could be a basis for which program the student fits in.
[45]	The study examines determinants of student likelihood to drop out; thus, it proposed a student dropout model.	The paper found out that student's average of 85 grades below is at risk of dropping out. In addition, gender and type of school is not a factor.	Limited variables were only included in the study: course, gender, high school grades in science, math, english, TLE, GPA, and type of school. Senior High School Track is not yet included. However, it was built on grades specific which is a limitation of this pursued study. Nevertheless, the study shows a comparison of accuracy result but is limited to tree classifiers and apply an ensemble approach.

A. Requirement and Data Collection

The dataset used was acquired from the Management Information System (MIS) of Caraga State University, Ampayon Butuan City, through the Office of Admission and Scholarships. A letter of intent addressed to the university's president was approved in October 2022, allowing the researcher to obtain the data as agreed not to disclose any sensitive data and strictly following Republic Act 10173 or the "Data Privacy Act of 2012". The data was received via email in .csv format with the following predictors shown in Table III. It contains the student records and admission scores from 2000 to 2022; however, most items in the previous years were empty, which was removed and narrowed down the data to 2,207, which is still highly relevant to the study. Moreover, data from 2000-2017 was removed because student data through the mentioned years does not include the senior high school strand, as K 12 was implemented in 2018. Hence, the study's goal is to forecast the new" success to graduate" of a new entrant student. The prediction classification is straightforward with binary value 1 for 'Yes' and 0 for 'No.'

B. Data Proprocessing

The acquired raw data has a high possibility of containing missing values. Missing values may lead to the ineffectiveness of the predictive model. Hence, the dataset in this study contains mixed data types; it needs to be handled accordingly with the use of the mode method for categorical values while KNN for numerical values. KNN

TABLE II. Researches in Predictive Analytics Domain

Research Description	Algorithms Used
Researchers stated that the exclusion of the protected attributes, namely gender, first-generation student, underrepresented minority (Asian or White), and high financial need, does not have a significant effect on the overall performance of the dropout prediction [48].	Logistic Regression, Gradient Boosted Trees
The research compared machine learning and one deep learning approach in detecting credit card fraud. ANN, a deep learning technique, ranked bottom among machine learning algorithms [14].	Decision Tree, SVM, Logistic Regression, Naive Bayes, Random Forest ANN
The study aims to identify the best time to predict a student's success -during admission, first semester, or second semester. The first two semesters were found to be significant, where grades during that semester were included as a predictor [49].	Ensemble Model, Logistic, Decision Tree, Bootstrap Forest, Boosted Tree
Linear SVC, followed by Logistic Regression, was among the methods that performed best when a model for employee attrition was developed. The study also identified factors that influence an employee to leave a company [50].	Linear SVC, Logistic Regression, Random Forest, KNN, SVC
Comparison with ensemble methods was conducted in this research. Features used were enrolment data, grades in science, english, and TLE to predict the student dropout [45].	Ensemble Model (Bagging+j-48), J-48, Forest Tree, Decision Tree
Forty-eight (48) articles about disease prediction were examined. SVM and Naïve Bayes were the most frequently used algorithms; however, Random Forest was found to have higher accuracy [47].	Random Forest, SVM, Naïve Bayes
One hundred twenty-one (121) articles were reviewed, particularly on the data source and variables, data handling, machine learning techniques, and accuracy evaluation. Findings show that most studies measure students' performance using scores, grades, and grades prior to graduation [46].	Decision Tree, Naïve Bayes, SVM, ANN, Random Forest, Logistics Regression
To obtain a high-quality dataset, it uses the KNN algorithm to impute the missing values of the data. The model was then trained using SVM and Naïve Bayes, where the first algorithm performed higher in predicting heart disease [13].	SVM, Naïve Bayes
The authors focused on developing predictive analytics to provide decision support to the administration during admission. Several algorithms were tested, and SVM overpowered the other algorithms [51].	SVM, Bayes, Logistic Regression, Neural Network, Chi-square Automatic Interaction Detector (CHAID)
Compared the number of machine learning approaches using the available student data before the start of the classes are used to predict the student performance. The Ensemble model is said to outperform the rest'[52].	Ensemble Model, Artificial Neural Network, k-Nearest Neighbors, K-Means Clustering, Naïve Bayes, SVM, Logistic Regression, Decision Tree

is an imputation technique that estimates the missing values based on the k-nearest neighbor method, replacing it with the 'N neighbors' mean value using the Euclidean distance metric. The mode method, on the other hand, will chiefly fill in the missing values with the most common value present in the particular feature of a dataset. The study of Sessa & Syed [32] found a significantly high accuracy rate prediction mode combining KNN and mode handling missing data. Though the execution of both methods will replace the missing values in the dataset, machine algorithms take numerical values to conduct a prediction. Thus, the presence of non-numeric values will need to undergo the process of binary encoding. For non-ordinal features, the said data could not be assigned with a numeric number, for this will cause the algorithm to think that value is based on ranking or hierarchy, giving bias to the result.

C. Identifying Predictors through Feature Selection

Moving forward, the dataset will undergo feature selection using Lasso Regression (L1), Ridge Regression (L2), and Genetic Algorithm. These three techniques will not be used together; however, it will be used separately to compare the feature it selected. Therefore, the study will test on one of the same datasets but with different feature selection methods. These are: 1) did not undergo the feature selection method or the dataset with the complete features, 2) the dataset applied with Lasso Regression, 3) the dataset applied with Ridge Regression, and 4) the dataset applied with Genetic Algorithm.

Though a study [53] found out that applying both methods L1 and L2, by running L1 first followed by L2, showed improvement in the result; it could be noted that it was employed to a dataset with 6,000 features reduced to

TABLE III. Features in the Acquired Dataset

Feature Name	Data Type	Number of Category	Missing Value
Demographics			
Sex	Categorical	2	0
Program	Categorical	30	0
Status	Categorical	5	0
Age	Numerical	-	137
Generation	Categorical	3	137
Civil Status	Categorical	2	0
Religion	Categorical	24	8
Municipality	Categorical	90	71
Province	Categorical	20	71
Father Occupation	Categorical	45	28
Mother Occupation	Categorical	45	36
Father Attainment	Categorical	11	90
Mother Attainment	Categorical	11	30
Father Income	Numerical	-	0
Mother Income	Numerical	-	0
Family Estimated Income	Numerical	-	0
Previous Academic Achievement			
SHS Track	Categorical	9	1702
Grade Point Average (GPA)	Numerical	-	593
Admission Exam Result			
NSAE Result (Total)	Categorical	2	0
NSAE Result (Total)	Numerical	-	3
English Score	Numerical	-	0
Math Score	Numerical	-	0
Science Score	Numerical	-	0
Abstract Score	Numerical	-	45

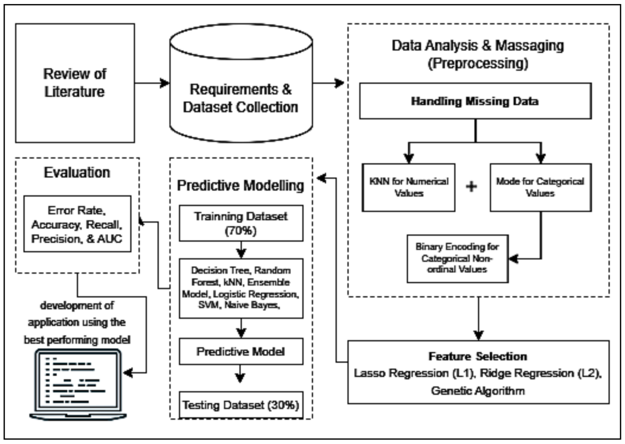

Review of Literature
Requirements & Dataset Collection
Data Analysis & Massaging (Preprocessing)
Handling Missing Data
KNN for Numerical Values
+
Mode for Categorical Values
Binary Encoding for Categorical Non-ordinal Values
Feature Selection
Lasso Regression (L1), Ridge Regression (L2), Genetic Algorithm
Predictive Modelling
Trainning Dataset (70%)
Decision Tree, Random Forest, kNN, Ensemble Model, Logistic Regression, SVM, Naive Bayes,
Predictive Model
Testing Dataset (30%)
Evaluation
Error Rate, Accuracy, Recall, Precision, & AUC
development of application using the best performing model

Figure 1. Experimental Design

50. In this study, however, only less than 100 features are present in the dataset, implying that the same experiment is not relevant and equivalent. And In contrast to the findings of Muthukrishnan & Rohini [54], which stated that LASSO works better than ridge regression, this study still wishes to verify by comparing both methods. With this, this work examined L1 and L2 separately as feature selection.

On the other hand, Genetic Algorithm (GA) will also be used as a feature selector in this study. Mweshi [38] in their literature review summarized that GA had been successful as a feature selector and shown promising outcomes. The genetic algorithm begins by generating an initial population of individuals. Each individual's fitness is then calculated on how it solves the given problem. Springs were produced through crossovers, reproduction, and mutation, which are responsible for creating new generations until they satisfy the termination condition before returning to the individual carrying the best solution. This work implemented 150 iterations before achieving the best-performing generation.

D. Classifiers and Predictive Modelling

With the help of Scikit-Learn and Google Collab, the preprocessed datasets will be split into 70-30, 70 for the training set and 30 for the test set. It will be modeled with seven (7) different algorithms: Decision Tree, Random Forest, Ensemble Model, KNN, Logistic Regression, SVM, and Naïve Bayes. The general definition of these algorithms is as follows [55].

A Decision Tree is a model that uses a tree-like structure to split data into smaller subsets based on different rules. It is utilized for both classification and regression tasks. Random Forest is an ensemble method that utilizes multiple decision trees. It trains each tree on a random subset of data and features and then combines their outputs to make a final prediction. The ensemble Model is a method that combines the predictions of multiple models to improve the overall performance. An example of this is using the bagging method with the J48 decision tree. The k-nearest Neighbor (kNN) method uses the k-nearest data points to make a prediction. It could be either by taking the majority class or the average value of the k-nearest neighbors. Logistic Regression is a model used for classification that uses a logistic function to model the probability of a binary outcome. It discovers the best linear combination of features that maximizes class separation. Support Vector Machine (SVM) finds the best boundary to separate different classes in the data by using a subset of the training examples called support vectors. Naive Bayes is a probabilistic method used for classification. It calculates the probability of a class given some features and assumes that the features are independent.

E. Evaluation Metrics

This study adopted the evaluation metrics used by Thabtah, et al. [56] in their study. There are five (5) evaluation metrics to use: error rate, accuracy, recall, precision, and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics were used in evaluating classification problems and will be derived based on the binary confusion matrix of the models.

Error rate or misclassification rate refers to the incorrect predictions produced by the model. The formula to get the error rate is shown below, where tp is the true positive, tn is the true negative, fp is the false positive, and fn is the false negative.

$$Errorrate = 1 - \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy denotes the correct classification made by the model. The formula to get the prediction accuracy is shown below, where tp is the true positive, tn is the true negative, fp is the false positive, and fn is the false negative.

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

Recall or sensitivity (true positive rate) is the actual positive occurrence that was correctly predicted positive. Models that resulted in higher recall mean that the model is good at identifying positive occurrences. The formula to get the recall is shown below, where tp is true positive, and tn is true negative.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Precision, or refers to positive predictive value, is the predicted positive instances that are actually positive. The formula to get the precision is shown below, where tp is the true positive, tn is the true negative, and fp is the false positive.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Area under the Curve (AUC) is a way to evaluate the model's prediction by measuring the area under the ROC curve. On the other hand, the ROC (Receiver Operating Characteristic) curve is a graphical representation illustrating the classifier's performance. It plots the curve of the true positive (tp) against the false positive (fp). This method of evaluation metric is famous because it is not affected by the dataset's class imbalance.

After finding the best-performing algorithm as accuracy and AUC as the final basis, the developed model will be translated into a web-based application using the Django Framework. Django is a Python-based framework suitable for scalable and maintainable application development. Moreover, it is capable of running most machine-learning algorithms, leading to easy implementation and maintenance. The application's front end will utilize HTML, CSS, and Bootstrap.

F. Software and Tools Used

Table IV summarizes the software used to implement the methods identified above. Microsoft Excel was used during the preprocessing phase, and Sklearn and Python scripts were run in Google Collab to handle missing data, binary encoding, develop and evaluate the models, and export the model to pkl file. The exported file was imported to the web application created using Bootstrap for UI and Django as the framework.

4. RESULTS

This chapter focuses on the discussion of the result using the methods mentioned in the previous chapter. This chapter is segmented into five (5) subsections: data preprocessing, predictive models' evaluation scores, features selected, developed web-based predictive application, and the implication of this study.

A. Data Preprocessing

The actual data was used; therefore, missing data is inescapable. The acquired data contains twenty-four (24) features. 'SHS Track' bears the highest amount of missing data, which is 1702, followed by 'grade point average' at 593, age (137), father attainment (90), municipality and province (both 71), mother occupation (36), mother attainment (30), father occupation (28), religion (8), NSAE

TABLE IV. Software Used in the Conduct of the Study

Software Name	Description and Usage
Microsoft Excel 2019	It is a spreadsheet software program that could be used as a visualization and analysis tool. This study used Microsoft Excel to pre-process the data acquired. With the aid of this software, the data was cleaned, categorized, and helped in the preprocessing phase of the dataset.
Sklearn	Short for scikit-learn - A python library used for machine learning algorithm and exported model to pkl file.
Google Collab	An online platform designed by Google to enable developer to execute codes using the browser.
Python 3.8.5	Python is a high-level programming language used in web development. It is also known for easy-to-implement machine learning-related problems. The development will use Python programming language.
Django 4.1.3	Django is a web framework that enables programmers to develop pragmatic design. It is a free, open-source web framework that follows the model-template-views (MTV) architectural pattern. The study findings will be translated into web applications using this web framework.
Bootstrap	Use for User Interface in the developed System

Result (3), and abstract score (45). Though the presence of high missing values in SHS track and GPA is prominent, these features were not dropped down to experiment on the effectiveness of KNN and mode imputation methods. After these codes were implemented, no missing values could be found in the dataset anymore.

Additionally, since the acquired dataset contains eleven (13) categorical features, specifically sex, program, status, generation, civil status, religion name, municipality, province, father and mother occupation and attainment, and SHS track, these nominal features undergo binary encoding before developing feature selection and predictive modeling. Initially, the dataset contained 24 features; however, after implementing the binary encoding, it expanded into sixty-four (64) features, excluding the target feature.

B. Predictive Models' Evaluation Score

Seven classifier algorithms were tested in four (4) datasets: The No Feature Selected (NFS) dataset and datasets employed with LASSO (L1), Ridge (R2), and Genetic Algorithm (GA) feature selection methods. Table V summarizes the metric scores for each classifier. No Feature Selected (NFS) dataset, which contains 64 features, shows that the Naïve Bayes accuracy score is the lowest (62%), together with logistic and SVM, with only 50% in the AUC score. Meanwhile, Random Forest showed the highest accuracy and AUC scores, 78% and 67%, respectively. However, NFS outperformed datasets applied with LASSO (L1) and Ridge (L2) Regression feature selection.

L1, selected only five (5) features and resulted in SVM with 70% accuracy and Naïve Bayes with only 62% AUC score as highest. L2, on the other hand, performed better than L1, with a greater number of features selected, and predicted 77% of test data correctly, 7% higher than the same classifier. Also, L2's SVM exhibited the highest AUC by 66%, which is only one point less than the random forest metric score from the NFS dataset. Moreover, Naïve Bayes scored a 70% accuracy rate as the lowest from the L2 dataset but is 8% higher when compared to the least accurate from the NFS dataset.

NFS performed best with Random Forest, followed by L2's SVM, but L1 resulted in significantly low accuracy; thus, the same findings were in AUC metric score. However, it is evident that the genetic algorithm FS increased the metric score of all the classifiers compared to the NFS dataset, L1, and L2. With logistic regression, the accuracy score increased by 10%, SVM – 7%, KNN and Ensemble – 6%, Decision tree – 3%, Naïve Bayes – 2%, and random forest by 1% compared to NFS. The upsurge in AUC score is also significant, with Logistic Regression on top surging by 21% more, SVM – 16%, KNN – 12%, Ensemble, and Naïve Bayes by 7%, and random forest and decision tree by 3% more than the AUC scored resulted in NFS dataset.

C. Features Selected

Since the data was binary encoded, it splits one feature into several sub-features based on its categorical values. Thus, it is reasonable to get the average of all the sub-features it created.

Table VI shows the result of the feature selection process; labeled as F means that the feature was not selected; otherwise, T for true. Among nine (9) feature selections employed in this study – L1, L2, and seven (7) classifiers from GA, the following were ordered based on times selected: NSAE Result (6), sex (5), mother income (5), English (5), math (5), program (4.6), father occupation (4.5), mother occupation (4.2), religion (4), mother attainment (4), family estimated income (4), science score (4), province (3.8), status (3.6), shstrack (3.25), municipality (3.14), age (3), generation (3), father attainment (3), father income (2), and

TABLE V. Evaluation Metric Score

Feature Selector	Classifier	Accuracy	Error	Precision	Recall	AUC	No. of Features
No Feature Selection (NFS)	RF	0.78	0.22	0.74	0.40	0.67	64
	EM	0.73	0.27	0.60	0.36	0.63	
	SVM	0.70	0.30	0.00	0.00	0.50	
	DT	0.69	0.31	0.49	0.54	0.65	
	LR	0.69	0.31	0.00	0.00	0.50	
	KNN	0.64	0.36	0.34	0.21	0.52	
	NB	0.62	0.38	0.41	0.57	0.61	
LASSO Regression (L1)	RF	0.67	0.33	0.43	0.27	0.56	5
	EM	0.68	0.32	0.44	0.28	0.56	
	SVM	0.70	0.30	0.00	0.00	0.50	
	DT	0.58	0.42	0.34	0.40	0.53	
	LR	0.69	0.31	0.00	0.00	0.50	
	KNN	0.63	0.37	0.33	0.22	0.51	
	NB	0.66	0.34	0.44	0.50	0.62	
Ridge Regression (L2)	RF	0.73	0.27	0.58	0.40	0.64	15
	EM	0.71	0.29	0.52	0.40	0.62	
	SVM	0.77	0.23	0.73	0.37	0.66	
	DT	0.71	0.29	0.53	0.40	0.62	
	LR	0.76	0.24	0.73	0.32	0.63	
	KNN	0.73	0.27	0.57	0.42	0.64	
	NB	0.70	0.30	0.52	0.30	0.59	
Genetic Algorithm (GA)	RF	0.79	0.21	0.73	0.49	0.70	29
	EM	0.79	0.21	0.72	0.48	0.70	37
	SVM	0.77	0.23	0.74	0.38	0.66	34
	DT	0.72	0.28	0.54	0.58	0.68	37
	LR	0.79	0.21	0.73	0.50	0.71	33
	KNN	0.70	0.30	0.51	0.49	0.64	35
	NB	0.64	0.36	0.44	0.82	0.68	23

abstract score (2).

Among features selected, admission result is the most selected feature appearing as 6/9 compared to GPA, which appears as 5/9, the same with sex, mother income, english, and math score. In contrast, the abstract and father income were selected the least, followed by age, generation, and father attainment based on their average occurrences in the feature selections employed as shown in Figure 2. The result suggests that the current admission selection based on the admission result is important among features; thus also showing abstract score in the exam is the least important area in the admission exam, english and math instead. It supports the findings [57], who declared that math skill is a better predictor of university performance. Meanwhile, [58] in which findings stated that english grade has a substantial correlation and are a strong predictor of students' year general point average toward non-english primary speaker, such as in the Philippines. Consequently, admission should consider the overall score in the admission exam and the specific score in math and english. Moreover, the literature review [59] stated that gender strongly influences the dropout rate, as well as parental background and status. This study, however, specified that a mother's income plays a role in the success to graduate of a student. Future work could explore these features to identify the underlying trend on sex and mother income.

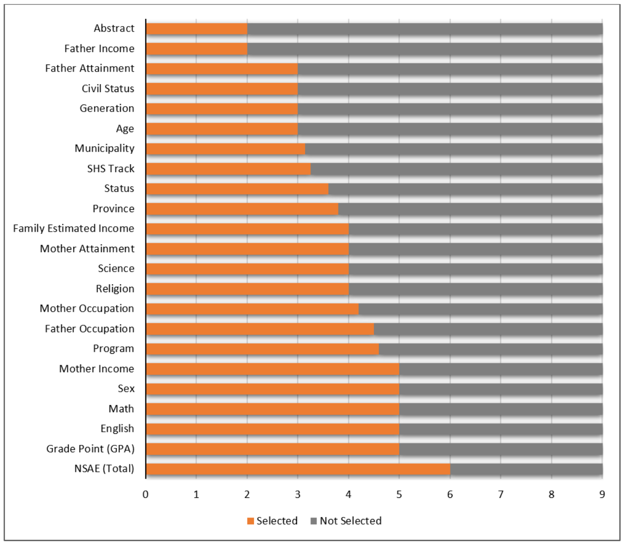


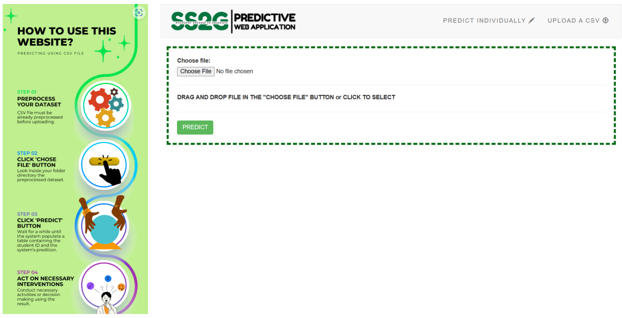
Figure 2. Selected Features of the FS and Classifiers

TABLE VI. Features Selected by L1, L2, and GA

	LASSO	Ridge	GENETIC ALGORITHM							Times	
FEATURES	(L1)	(L2)	EM	DT	KNN	LR	NB	RF	SVM	Selected	Average
Sex_0	F	F	T	F	T	T	T	T	F	5	5
Sex_1	F	F	T	T	T	T	F	F	T	5	
Program_0	F	F	T	F	F	F	T	F	T	3	4.6
Program_1	F	T	T	T	F	T	F	T	T	6	
Program_2	F	T	T	F	T	T	F	F	T	5	
Program_3	F	T	T	F	F	T	F	T	T	5	
Program_4	F	F	T	T	T	T	F	F	F	4	
Status_0	F	F	T	F	F	F	T	T	F	3	3.6
Status_1	F	F	T	T	F	T	F	F	T	4	
Status_2	F	F	T	T	T	F	F	F	T	4	
Age	F	F	F	T	T	F	T	F	F	3	3
Generation_0	F	F	T	T	T	F	F	T	F	4	3
Generation_1	F	F	F	F	F	T	F	T	F	2	
Civil_status_0	F	F	T	T	T	F	F	F	F	3	3
Civil_status_1	F	T	F	T	T	F	F	F	F	3	
Religionname_0	F	F	F	T	F	T	T	F	F	3	4
Religionname_1	F	T	T	T	F	T	F	T	T	6	
Religionname_2	F	F	F	F	F	F	T	F	T	2	
Religionname_3	F	F	F	T	T	T	F	T	T	5	
Religionname_4	F	F	F	T	T	F	T	T	F	4	
Municipality_0	F	F	F	F	F	F	T	F	F	1	3.14
Municipality_1	F	F	F	F	F	F	F	F	F	0	
Municipality_2	F	F	F	T	T	T	F	T	T	5	
Municipality_3	F	T	T	F	F	T	T	F	F	4	
Municipality_4	F	F	T	T	F	T	F	F	F	3	
Municipality_5	F	F	F	F	F	F	T	T	T	3	
Municipality_6	F	T	F	T	T	F	T	T	T	6	
Province_0	F	F	F	T	F	F	F	F	T	2	3.8
Province_1	F	F	F	T	T	T	F	F	T	4	
Province_2	F	F	F	F	T	F	F	T	F	2	
Province_3	F	T	T	F	T	T	F	T	T	6	
Province_4	F	T	T	T	F	T	T	F	F	5	
FatherOccupation_0	F	F	F	F	T	T	F	F	T	3	4.5
FatherOccupation_1	F	F	T	T	F	T	T	T	F	5	
FatherOccupation_2	F	F	T	T	F	F	T	T	T	5	
FatherOccupation_3	F	F	T	T	T	T	T	F	T	6	
FatherOccupation_4	F	F	T	F	T	F	T	F	F	3	
FatherOccupation_5	F	F	T	T	F	F	T	T	T	5	
Father_Income	T	F	F	F	F	F	F	F	T	2	2
Father_attainment_0	F	F	T	T	F	F	T	F	F	3	3
Father_attainment_1	F	F	F	T	F	T	F	F	F	2	
Father_attainment_2	F	T	F	T	T	F	T	F	F	4	
Father_attainment_3	F	F	T	T	F	F	F	F	T	3	
MotherOccupation_0	F	T	F	F	T	T	F	T	T	5	4.2
MotherOccupation_1	F	F	T	T	T	F	T	F	F	4	
MotherOccupation_2	F	T	F	T	F	T	F	T	T	5	
MotherOccupation_3	F	F	F	F	T	F	F	T	T	3	
MotherOccupation_4	F	F	T	T	T	T	F	F	F	4	
Mother_Income	T	F	F	T	T	T	T	F	T	6	5
Mother_Attainment_0	F	T	T	F	T	T	F	T	T	6	4
Mother_Attainment_1	F	T	T	T	T	T	F	F	T	6	
Mother_Attainment_2	F	T	F	F	F	F	F	F	F	1	
Mother_Attainment_3	F	F	T	F	F	F	F	T	T	3	
Family_est_income	T	F	T	F	T	F	F	F	T	4	4
Gradepoint	T	F	F	T	T	T	F	T	F	5	5
Shtrack_0	F	F	T	F	T	T	F	F	F	3	3.25
Shtrack_1	F	F	T	F	T	F	T	T	F	4	
Shtrack_2	F	F	T	F	F	T	F	T	T	4	
Shtrack_3	F	F	F	F	F	F	F	T	T	2	
NSAEResult	T	F	T	T	F	T	F	T	T	6	6
English	F	F	F	T	T	T	T	T	F	5	5
Math	F	F	T	T	T	T	F	T	F	5	5
Science	F	F	T	T	T	F	F	F	T	4	4
Abstract	F	F	T	F	T	F	F	F	F	2	2
TOTAL	5	15	37	37	35	33	23	29	34		

D. Developed Web-based Predictive Application

Logistic Regression with Genetic Algorithm Feature Selection as the model with the highest ACU, downloaded as a .pkl file and was loaded to the views.py of Django Framework. The developed web application has two ways to perform a prediction. The first approach is predicting by bulk using a CSV preprocessed dataset. The CSV file must already undergo preprocessed methods as described in the prior sections. However, the dataset must be added with a new column labeled as an ID number to identify which student needs the intervention. Figure 3 shows the interface of prediction by bulk approach. The 'choose file' allows opening a computer directory to locate the dataset as input. Afterward, the 'predict' button must be clicked for the web application to start its prediction.



HOW TO USE THIS WEBSITE?

PREDICTING USING CSV FILE

STEP 01 PREPROCESS YOUR DATASET
CSV file must be already preprocessed before uploading.

STEP 02 CLICK 'CHOSE FILE' BUTTON
Look inside your folder directory the preprocessed dataset.

STEP 03 CLICK 'PREDICT' BUTTON
Wait for a while until the system populate a table containing the student ID and the system's predition.

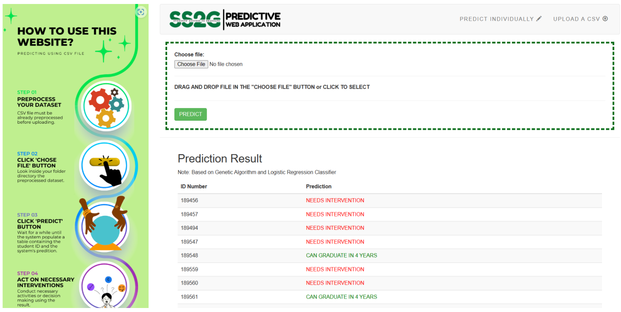
STEP 04 ACT ON NECESSARY INTERVENTIONS
Conduct necessary activities or decision making using the result.

Choose file: No file chosen

DRAG AND DROP FILE IN THE "CHOOSE FILE" BUTTON or CLICK TO SELECT

Figure 3. Upload CSV dataset

Once the prediction is made, the prediction will be displayed at the lower portion of the web application body, as shown in Figure 4. Two columns are presented, namely 'ID Number' and its prediction. The red text indicates the need for intervention for a specific student. In contrast, green text were students predicted to finish their program within four (4) years.



HOW TO USE THIS WEBSITE?

PREDICTING USING CSV FILE

STEP 01 PREPROCESS YOUR DATASET
CSV file must be already preprocessed before uploading.

STEP 02 CLICK 'CHOSE FILE' BUTTON
Look inside your folder directory the preprocessed dataset.

STEP 03 CLICK 'PREDICT' BUTTON
Wait for a while until the system populate a table containing the student ID and the system's predition.

STEP 04 ACT ON NECESSARY INTERVENTIONS
Conduct necessary activities or decision making using the result.

Choose file: No file chosen

DRAG AND DROP FILE IN THE "CHOOSE FILE" BUTTON or CLICK TO SELECT

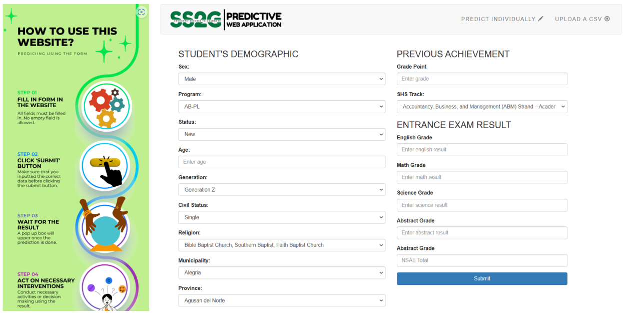
Prediction Result

Note: Based on Genetic Algorithm and Logistic Regression Classifier

ID Number	Prediction
189456	NEEDS INTERVENTION
189457	NEEDS INTERVENTION
189494	NEEDS INTERVENTION
189547	NEEDS INTERVENTION
189548	CAN GRADUATE IN 4 YEARS
189559	NEEDS INTERVENTION
189560	NEEDS INTERVENTION
189561	CAN GRADUATE IN 4 YEARS

Figure 4. Bulk Prediction Result

Another way to perform a prediction in the system is by providing individual data. Figure 5 shows the user interface. Each field has a designated value and will not allow empty inputs to prevent missing values. Once everything is filled in, the 'submit' button must be clicked to perform prediction. Finally, a pop-up window will appear, prompting whether it needs an intervention, as shown in Figures 6 and 7.



HOW TO USE THIS WEBSITE?

PREDICTING USING THE FORM

STEP 01 FILL IN FORM IN THE WEBSITE
All fields must be filled in. No empty field is allowed.

STEP 02 CLICK 'SUBMIT' BUTTON
Make sure that you inputted the correct data before clicking the submit button.

STEP 03 WAIT FOR THE RESULT
A pop up box will upper once the prediction is done.

STEP 04 ACT ON NECESSARY INTERVENTIONS
Conduct necessary activities or decision making using the result.

STUDENT'S DEMOGRAPHIC

Sex: Male

Program: AB-PL

Status: New

Age: Enter age

Generation: Generation Z

Civil Status: Single

Religion: Bible Baptist Church, Southern Baptist, Faith Baptist Church

Municipality: Alegria

Province: Agusan del Norte

PREVIOUS ACHIEVEMENT

Grade Point: Enter grade

SHS Track: Accountancy, Business, and Management (ABM) Strand – Acader

ENTRANCE EXAM RESULT

English Grade: Enter english result

Math Grade: Enter math result

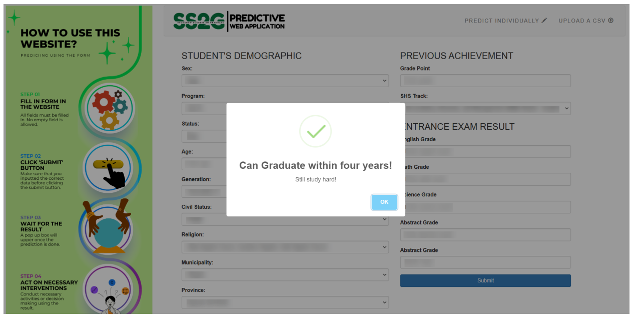
Science Grade: Enter science result

Abstract Grade: Enter abstract result

Abstract Grade: NSAE Total

Figure 5. Form for Predicting a Student

The message pops-up interface was created using sweet.js. When the model returns a value of 1, the message box will appear, as shown in Figure 6. The returned value '1' signifies that the student can graduate on time according to the model.



HOW TO USE THIS WEBSITE?

PREDICTING USING THE FORM

STEP 01 FILL IN FORM IN THE WEBSITE
All fields must be filled in. No empty field is allowed.

STEP 02 CLICK 'SUBMIT' BUTTON
Make sure that you inputted the correct data before clicking the submit button.

STEP 03 WAIT FOR THE RESULT
A pop up box will upper once the prediction is done.

STEP 04 ACT ON NECESSARY INTERVENTIONS
Conduct necessary activities or decision making using the result.

STUDENT'S DEMOGRAPHIC

Sex: / Program: / Status: / Age: / Generation: / Civil Status: / Religion: / Municipality: / Province:

PREVIOUS ACHIEVEMENT

Grade Point / SHS Track:

NTRANCE EXAM RESULT

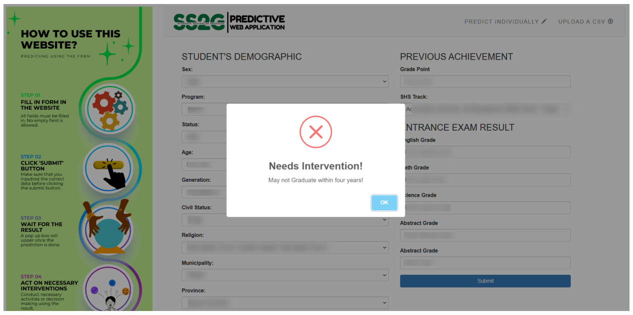
nglish Grade / ath Grade / cience Grade / Abstract Grade / Abstract Grade

Can Graduate within four years!

Still study hard!

Figure 6. 'Can Graduate on Time' Message Prompt

On the other hand, the 'Needs Intervention' prompt will appear once the prediction value is '0' shown in Figure 7, meaning the student probably will not graduate on time based on his/her data.



HOW TO USE THIS WEBSITE?

PREDICTING USING THE FORM

STEP 01 FILL IN FORM IN THE WEBSITE
All fields must be filled in. No empty field is allowed.

STEP 02 CLICK 'SUBMIT' BUTTON
Make sure that you inputted the correct data before clicking the submit button.

STEP 03 WAIT FOR THE RESULT
A pop up box will upper once the prediction is done.

STEP 04 ACT ON NECESSARY INTERVENTIONS
Conduct necessary activities or decision making using the result.

STUDENT'S DEMOGRAPHIC

Sex: / Program: / Status: / Age: / Generation: / Civil Status: / Religion: / Municipality: / Province:

PREVIOUS ACHIEVEMENT

Grade Point / SHS Track: Ac

NTRANCE EXAM RESULT

nglish Grade / ath Grade / cience Grade / Abstract Grade / Abstract Grade

Needs Intervention!

May not Graduate within four years!

Figure 7. 'Needs Intervention' Message Prompt

E. Implication

This study contributed new insight into the prediction domain by presenting a real-world problem in the education sector. The significant implication of this work can help the university craft or improve existing policy, thus positively affecting admission procedures—implementing a more holistic approach for a more efficient selection process. Lastly, for students, it could add guidance regarding career path selection after their senior high school. While the study focused on college students, these implications are relevant to any school and institute as the baseline in constructing admission plans, selection procedures, interventions, and student assessment. Moreover, these inferences are significant to practitioners in the field of predictive analytics and the like.

5. CONCLUSION AND FUTURE WORK

The study found that the Genetic Algorithm outperformed NFS, L1, and L2 feature selection in seven algorithms except for precisions of NFS' random forest, the accuracy rate of L2 using Naïve Bayes, and precisions of L2's KNN and Naïve Bayes. Among the features selected, admission result is the most selected feature compared to GPA, which was found to be second and equivalent substantial with sex, mother income, english, and math score. In contrast, the abstract score least selected featured followed by age, generation, and father income and attainment based on their average occurrences in the feature selections employed. Logistic Regression with genetic Algorithm as a feature selection method has the highest accuracy (79%) and AUC score (71%) among others; thus, it was selected as the best-performing predictive model. Moreover, it selected sex, mother income, grade point average, admission result, english and math exam scores. Also, it selected some portion of binary encoded features program, status, generation, religion, municipality, province, father occupation, father attainment, mother occupation, mother attainment, and shstrack. Further studies are recommended to continuously monitor the model's correctness, such as gathering the data of new entrants from 2019 and beyond, serving as the validation dataset to assess the implemented web application using the developed model. In addition, since the study is dependent on the available data in the university, there is a limit in features to feed into the model. It is recommended to explore other predictors such as internet connection, social media activity, hobbies, and peer influence in future work as it may affect student success to graduate.

References

[1] D. T. Larose, *Data mining and predictive analytics*. John Wiley & Sons, 2015.

[2] N. S. K. Mullapudi and B. P Sridhar, "An overview of trends and techniques in predictive analytics," *Journal of Contemporary Issues in Business and Government*, vol. 28, no. 4, pp. 952–959, 2022.

[3] J. V. S. Pitao, J. P. Nabas, J. B. Matias, J. Q. Timosan, and G. G. Rollorata, "Development and evaluation of enhanced national greening program monitoring and document archiving system using delone and mclean is success model," in *Proceedings of the 2022 11th International Conference on Networks, Communication and Computing*, ser. ICNCC '22. New York, NY, USA: Association for Computing Machinery, 2023, p. 334–340.

[4] A. Khasanov, "Impact of predictive analytics on the activities of companies," *Strategic decisions and risk management*, no. 3, pp. 108–113, 2018.

[5] V. Kumar and M. Garg, "Predictive analytics: a review of trends and techniques," *International Journal of Computer Applications*, vol. 182, no. 1, pp. 31–37, 2018.

[6] M. Alipio, "Predicting academic performance of college freshmen in the philippines using psychological variables and expectancy-value beliefs to outcomes-based education: A path analysis," 2020.

[7] M. Lumboy, "Senior high school strand choice: Its implication to college academic performance," *Ascendens Asia Journal of Multidisciplinary Research Abstracts*, vol. 3, no. 7, 2019.

[8] C. A. Quintos, D. G. Caballes, E. M. Gapad, and M. R. Valdez, "Exploring between shs strand and college course mismatch: Bridging the gap through school policy on intensified career guidance program," *CiiT International Journal of Data Mining and Knowledge Engineering*, vol. 12, no. 10, pp. 156–161, 2020.

[9] J. Santos, L. C. Blas, A. J. Panganiban, K. Reyes, and J. C. Sayo, "Alignment of senior high school strand in college course," *Jewel Christine, Alignment of Senior High School Strand in College Course (February 1, 2019)*, 2019.

[10] M. M. Sulphey, N. S. Al-Kahtani, and A. M. Syed, "Relationship between admission grades and academic achievement," *Entrepreneurship and Sustainability Issues*, vol. 5, no. 3, pp. 648–658, 2018.

[11] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, L. Halawi, R. McCarthy, M. McCarthy, W. Ceccucci, and L. Halawi, *Applying predictive analytics*. Springer, 2022.

[12] G. Lakshmi and M. Shang, *Hands-on Supervised Learning with Python ([edition unavailable*. BPB Publications, 2021.

[13] F. Khennou, C. Fahim, H. Chaoui, and N. E. H. Chaoui, "A machine learning approach: Using predictive analytics to identify and analyze high risks patients with heart disease," *International Journal of Machine Learning and Computing*, vol. 9, no. 6, pp. 762–767, 2019.

[14] M. Ashraf, M. A. Abourezka, and F. A. Maghraby, "A comparative analysis of credit card fraud detection using machine learning and deep learning techniques," in *Digital Transformation Technology: Proceedings of ITAF 2020*. Springer, 2022, pp. 267–282.

[15] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *2018 IEEE congress on evolutionary computation (CEC)*. IEEE, 2018, pp. 1–9.

[16] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020, pp. 302–305.

[17] B. M. Pavlyshenko, "Machine-learning models for sales time series forecasting," *Data*, vol. 4, no. 1, p. 15, 2019.

[18] H. Nugroho and K. Surendro, "Missing data problem in predictive analytics," in *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, 2019, pp. 95–100.

[19] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment," *Medicina*, vol. 56, no. 9, p. 455, 2020.

[20] S. M. Cho, P. C. Austin, H. J. Ross, H. Abdel-Qadir, D. Chicco, G. Tomlinson, C. Taheri, F. Foroutan, P. R. Lawler, F. Billia *et al.*, "Machine learning compared with conventional statistical models for predicting myocardial infarction readmission and mortality: a systematic review," *Canadian Journal of Cardiology*, vol. 37, no. 8, pp. 1207–1214, 2021.

[21] S.-A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *The Knowledge Engineering Review*, vol. 34, p. e1, 2019.

[22] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, J. Akinjobi *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[23] I. Muraina, "Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts," in *7th International Mardin Artuklu Scientific Research Conference*, 2022.

[24] J. Tan, J. Yang, S. Wu, G. Chen, and J. Zhao, "A critical look at the current train/test split in machine learning," *arXiv preprint arXiv:2106.04525*, 2021.

[25] A. Rajkar, A. Kumaria, A. Raut, and N. Kulkarni, "Stock market price prediction and analysis," *International Journal of Engineering Research & Technology (IJERT) Volume*, vol. 10, 2021.

[26] J. R. Almonteros, M. P. B. Pacot, and V. A. Pitogo, "Automation of curriculum-based student-subject encoding: A web application," in *Proceedings of the 2022 11th International Conference on Networks, Communication and Computing*, ser. ICNCC '22. New York, NY, USA: Association for Computing Machinery, 2023, p. 328–333.

[27] S. Nijman, A. Leeuwenberg, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, and T. Debray, "Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review," *Journal of clinical epidemiology*, vol. 142, pp. 218–229, 2022.

[28] C. A. Leke, T. Marwala, C. A. Leke, and T. Marwala, "Introduction to missing data estimation," *Deep Learning and Missing Data in Engineering Systems*, pp. 1–20, 2019.

[29] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019.

[30] K. Seu, M.-S. Kang, and H. Lee, "An intelligent missing data imputation techniques: A review," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 1-2, pp. 278–283, 2022.

[31] N. A. A. Wafaa Mustafa Hameed, "Comparison of seventeen missing value imputation techniques," *Journal of Hunan University Natural Sciences*, vol. 49, no. 7, 2022.

[32] J. Sessa and D. Syed, "Techniques to deal with missing data," in *2016 5th international conference on electronic devices, systems and applications (ICEDSA)*. IEEE, 2016, pp. 1–4.

[33] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022.

[34] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 703–715, 2019.

[35] A. Abdulhafedh, "Comparison between common statistical modeling techniques used in research, including: Discriminant analysis vs logistic regression, ridge regression vs lasso, and decision tree vs random forest," *Open Access Library Journal*, vol. 9, no. 2, pp. 1–19, 2022.

[36] D. Zhang, Y. Chen, Y. Chen, S. Ye, W. Cai, J. Jiang, Y. Xu, G. Zheng, and M. Chen, "Heart disease prediction based on the embedded feature selection method and deep neural network," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–9, 2021.

[37] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy." *Pertanika Journal of Science & Technology*, vol. 26, no. 1, 2018.

[38] G. Mweshi, "Feature selection using genetic programming," *Zambia ICT Journal*, vol. 3, no. 2, pp. 11–18, 2019.

[39] S. Solorio-Fernandez, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "A supervised filter feature selection method for mixed data based on spectral feature selection and information-theory redundancy analysis," *Pattern Recognition Letters*, vol. 138, pp. 321–328, 2020.

[40] Z. Gniazdowski and M. Grabowski, "Numerical coding of nominal data," *arXiv preprint arXiv:1601.01966*, 2016.

[41] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International journal of computer applications*, vol. 175, no. 4, pp. 7–9, 2017.

[42] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. degree project technology. published online 2018," 2018.

[43] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, pp. 1–21, 2020.

[44] M. M. Chingos, "What matters most for college completion," *Academic preparation is a key predictor of success. AEI Paper & Studies A*, vol. 3, 2018.

[45] F. F. Patacsil, "Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models," *Universal Journal of Educational Research*, vol. 8, no. 9, pp. 4036–4047, 2020.

[46] Y. Cui, F. Chen, A. Shiri, and Y. Fan, "Predictive analytic models of student success in higher education: A review of methodology," *Information and Learning Sciences*, vol. 120, no. 3/4, pp. 208–227, 2019.

[47] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–16, 2019.

[48] R. Yu, H. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected attributes?" in *Proceedings of the eighth ACM conference on learning@ scale*, 2021, pp. 91–100.

[49] X. Wang, H. Schneider, and K. R. Walsh, “A predictive analytics approach to building a decision support system for improving graduation rates at a four-year college,” *Journal of Organizational and End User Computing (JOEUC)*, vol. 32, no. 4, pp. 43–62, 2020.

[50] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, “Predicting employee attrition using machine learning techniques,” *Computers*, vol. 9, no. 4, p. 86, 2020.

[51] J. Cirelli, A. M. Konkol, F. Aqlan, and J. C. Nwokeji, “Predictive analytics models for student admission and enrollment,” in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, vol. 2018, no. SEP, 2018, pp. 1395–1403.

[52] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” *Computers & Electrical Engineering*, vol. 89, p. 106903, 2021.

[53] O. Demir-Kavuk, M. Kamada, T. Akutsu, and E.-W. Knapp, “Prediction using step-wise l1, l2 regularization and feature selection for small data sets with large number of features,” *BMC bioinformatics*, vol. 12, pp. 1–10, 2011.

[54] R. Muthukrishnan and R. Rohini, “Lasso: A feature selection technique in predictive modeling for machine learning,” in *2016 IEEE international conference on advances in computer applications (ICACA)*. IEEE, 2016, pp. 18–20.

[55] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[56] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, “Data imbalance in classification: Experimental evaluation,” *Information Sciences*, vol. 513, pp. 429–441, 2020.

[57] J. M. Delaney and P. J. Devereux, “Math matters! the importance of mathematical and verbal skills for degree performance,” *Economics Letters*, vol. 186, p. 108850, 2020.

[58] B. Waluyo and B. Panmei, “English proficiency and academic achievement: Can students’ grades in english courses predict their academic achievement?.” *Mextesol Journal*, vol. 45, no. 4, p. n4, 2021.

[59] A. Behr, M. Giese, H. D. Teguim Kamdjou, and K. Theune, “Dropping out of university: a literature review,” *Review of Education*, vol. 8, no. 2, pp. 614–652, 2020.

Jayrhom R. Almonteros received his MS degree in Information Technology at Caraga State University, Philippines in 2023. He is a faculty member at the same university, and was previously designated as the College Extension Coordinator; thus is presently the chairperson of the Department of Information Systems. His research interest includes data analytics, software development, and ICT for governance.

Junrie B. Matias is a faculty member at Caraga State University in Butuan City. He graduated from the same university with a bachelor’s Degree in Computer Science. He holds a master’s Degree in information technology from the University of Science and Technology Southern Philippines, Cagayan de Oro campus, and completed his Doctoral Degree in Information Technology at the Technological Institute of the Philippines. His research interests include artificial intelligence, technology adoption, software engineering, and programming languages. He has authored several research articles presented at international conferences and has several publications under his name.

Joanna Victoria S. Pitao received her BS in Computer Science in 2019 and taking her MS degree in Information Technology at Caraga State University. Currently, she is a faculty member in the College of Computing and Information Sciences. Her research interest includes software and machine learning.