International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Jan-2022)

http://dx.doi.org/10.12785/ijcds/110127

Machine Learning Based Selection of Incoming Engineering Freshmen in Higher Education Institution

Mary Anne M. Sahagun[1]

[1]*Electronics Engineering Department, Don Honorio Ventura State University, Pampanga, Philippines*
mamsahagun@gmail.com

Received 08 Sep.2020, Revised 20 Jul. 2021, Accepted 24 Jul. 2021, Published 9 Jan. 2022

Abstract: The Accrediting Agency of Chartered Colleges and Universities of the Philippines recommends through the university testing unit, a system to interpret and analyse entrance test results that may help direct and guide students in choosing a Baccalaureate degree to take in the college. While the present system of manually evaluating each of the freshman applicants is used, there is a need to adopt technological tools for faster and accurate analysis. Thus, the study presents machine learning methods of classifying freshmen applicants if they are qualified or not in the college of engineering and architecture. Specifically, determining if a freshmen applicant may succeed in the five engineering program at university. The study used classifiers such as Decision Tree, K-Nearest Neighbor (KNN), Decision Tree, and Support Vector Machine (SVM). A cross-validation of ten-fold model was used better accuracy of classifiers. The predicted models performed well, however, the Decision Tree classifier outputs a higher average accuracy and F1-measure. The result shows that the classifier accurately classifies qualified and non-qualified engineering freshmen for program acceptance.

Keywords: : K-Nearest Neighbor, Decision Tree, Support Vector Machine, Data Mining

1. Introduction

Data mining is finding distinct patterns present in each dataset by a systematic processes. For the past couple of years, different data mining implementations have been applied in a wide variety of smart applications such as intelligent traffic systems (ITS) [1]–[4]; digital health [5]–[8]; disaster preparedness [9], [10]; and business economics [11], [12]. This type of technology can also be applied in education for improving learning such as distance learning [13] or predicting student achievement [14]. The rationale of data mining is to recognize data pattern and discover new and practical insights [15], known as Educational data mining (EDM).

Educational Data Mining describes a an area on research that uses data from academic settings to develop methods, gain relevant information, and knowledge to better understand the student, university environment further, and enable better education planning. The EDM can be used to optimize a school, college or any other learning institution, as well as automate managerial decision-making.

Analytical prediction is an EDM technique that predicts a future state [16] for possible implementation based on analysis. It's a way of predicting success rates, dropout rates, and devising retention strategies. It is particularly beneficial in aligning education's future with industry trends. An approach used in EDM is classification. It's a supervised technique that maps data attributes to targets. This technique is highly effective for predicting student performances, risk analysis, student monitoring systems, and error detection.

A. Problem Statement and Research Contribution

The method of education in different countries, including the Philippines, continues to anchor advancement in technology in learning delivery and evaluation. Classroom discussion has been geared toward smart delivery. However, data gathered in the university setting has not been thoroughly investigated and used for curriculum development and educational management decisions such as freshman acceptance. Admission System in State Colleges and Universities in the Philippines are different from those in private institutions. The manual method of the admission process for government institutions requires practical processes but consumes much time. The Admission and Testing Office of the Don Honorio Ventura State University serves incoming Grade VII, applicants for Senior High School, applicants in incoming college freshmen, transferees, returning students, cross-enrollees, shifters, foreign-students, and graduate school applicants. The university has six (6) extension campuses and the admission of students is centralized at the main campus. During the admission period, the applicant patiently stand in long-queue to have his or her turn to get an application form and proceed with admission procedures

E-mail address: mamsahagun@gmail.com

[17] as seen in Table I. The admission requirements for college freshmen require the following documents: senior high school report card, birth certificate, a colored picture, and an accomplished application form.

TABLE I. ADMISSION PROCEDURES FOR COLLEGE FRESHMEN

Step No.	Procedure
1	Obtain an application form from the Admission and Testing Office
2	Fill out the application form
3	Submit requirements directly to the Admission and Testing Office
4	Encode Information for Student Profiling at Admission and Testing Office
5	Obtain from Admission and Testing a Test Permit noting the date, time, and venue of examination.

After an applicant had completed the admission procedure, the person needs to come back again to the university for another scheduled date of an entrance exam. The exam is a paper and pencil-type of exam and is checked and computed manually by the testing unit personnel. A list of incoming freshmen applicant results of the college entrance exam and the general weighted average (GWA) of senior high school report card are forwarded to respective deans of each college as seen in Figure 1. It is at the discretion of the dean to conduct an interview or a college qualifying exam.

In April 2019, the first qualifying exam was conducted in the College of Engineering and Architecture. Historically, only interviews are conducted with incoming engineering freshmen and are interviewed by the respective department program chair. This activity lasts for several months and is not sufficient to measure the ability of the students for the chosen program applied for. The qualifying exam is a paper and the pen-type exam consists of different questions related to general engineering and is checked manually by assigned faculty. The applicants are notified through posting of announcements such as the schedule, venue, and what to bring during the exam proper through official Facebook of the university.

The results of the qualifying exam, entrance exam, and GWA are forwarded by the college dean to each respective program chair for manual assessment and evaluation of individual applicants in the program. On average, an incoming freshmen applicant waits for 1-3 months to know if qualified for the engineering program applied for, and need to visit the university four times between the admission and qualifying exam process. These lengthy periods of waiting for the qualifier results lead other students to find other course programs. The unsuccessful qualifiers who did not pass the applied specific engineering program also find difficulty in assessing oneself of finding other engineering programs that he or she can apply for. Since the conventional process only includes evaluation for the specific course program.

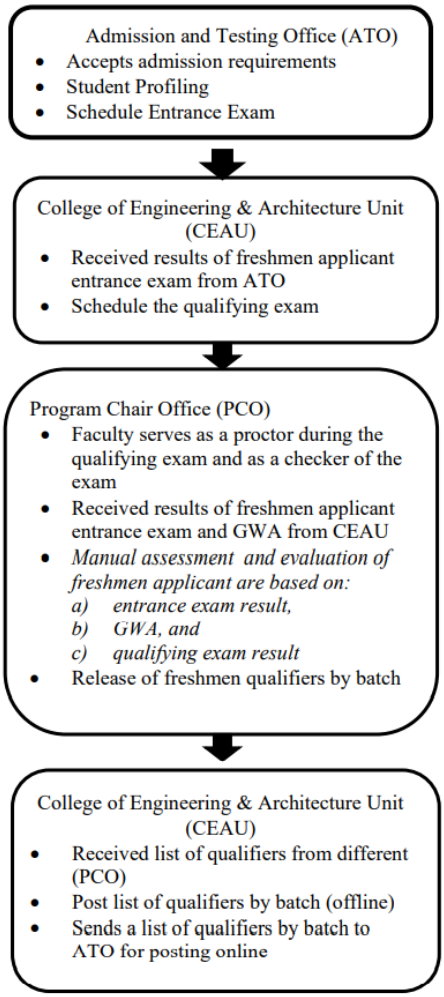

Figure 1. Flowchart of manual evaluation of incoming freshmen

While the existing method of manual evaluation of incoming freshmen is considered effective; there is a need to apply EDM technology to utilize available data and improve the evaluation of the freshmen applicants if qualified or not in the applied program or another engineering program. This research aims to contribute to the full implementation of freshmen classification for suitability before admission to the program.

B. Proposed Solution and Implementation of EDM

The College of Engineering and Architecture of Don Honorio Ventura State University has always been supportive of implementing a smarter system in terms of education in terms of its delivery, business decision, and implementation. Thus, the solution of implementing EDM in freshmen classification is another step towards achieving this goal. There are five engineering programs in the university, namely, Civil Engineering (CE), Mechanical

Engineering (ME), Electrical Engineering (EE), Electronics Engineering (EcE), and Industrial Engineering (IE). Among these programs, Electronics Engineering has the lowest passing rate in the national board examination in the country, programs like CE, ME, and EE always have a high passing rate, and IE is a non-board program. From these different trends of the national passing rate per program, still, the acceptable score in the entrance exam and GWA of the freshmen applicants are standardized to all engineering programs. However, with the qualifying exam prepared by the College of Engineering and Architecture, there is a separate acceptable score per program.

EDM will use data such as secondary education grades of students, standardized entrance exams, and interviews. While there is a clear boundary in accepting students based on their numerical scores; however, once accepted in the program some students tend to drop their course or subject and/or shifted to another program. This study aims to completely evaluate students before they enter college. This solution proposed in this paper applies classification techniques such as KNN, Decision Tree, and SVM.

2. System Overview

This section describes the dataset and feature description; the different models considered; and the metrics used for evaluation.

A. Dataset and Feature Description

As described in section 1-B, historically, qualifiers are admitted to the engineering program by considering three (3) inputs: entrance exam results, GWA, and short interview only. However, it was improved in 2019 when the college conducted its first qualifying exam.

Several factors need to be taken into consideration for the proper evaluation of students such as (i) the graded weight average (GWA) from secondary education, (ii) standardized entrance exam, (iii) qualifying exam, and (iv) senior highschool academic strand. However, problems involving low grades in their freshmen year; or the issue of program shift needs to consider other attributes that can lead to a better selection of incoming freshmen. The Accrediting Agency of Chartered Colleges and Universities of the Philippines recommends through the university testing unit, a system to interpret and analyze entrance test results may help direct and guide students in choosing a baccalaureate degree to take in college. This proposed system once implemented can address the problem of program mismatch.

For this study aside from the conventional scores of the entrance exam results, the Student Ability Index and equivalent Stanine value were further considered. The historical dataset used for the study was obtained from the Guidance and Testing Office, and the College of Engineering and Architecture. Considering the data privacy, no names were given but only a sample number of 745 freshmen with a record of GWA, School Ability Index (SAI), Stanine (S), senior high school strand, and the College Qualifying Exam (CQE) as seen in Table II. The permit to obtain these datasets was approved by the public information office of the university.

TABLE II. DATASET DESCRIPTION

Criteria	Type	Range
Grade Weighted Average (GWA)	Numeric	75 to 100
School Ability Index (SAI)	Numeric	0 to 150
Stanine (S)	Numeric	1 to 9
Strand	Numeric	1 STEM 0 non-STEM
College Qualifying Exam (CQE)	Numeric	1 to 15

TABLE III. REQUIRED MARK PER CRITERIA

Criteria	Acceptable Score
Grade Weighted Average (GWA)	≥ 85
School Ability Index (SAI)	≥ 104
Stanine (S)	≥ 6
College Qualifying Exam (CQE)	≥ 8

There are three academic strands in senior highschool these are Business, Accountancy, and Management (BAM), Humanities, Education, and Social Sciences (HEMSS), and Science, Technology, Engineering, and Mathematics (STEM). A student is accepted to any engineering program regardless of the strand. Excluded from the lists are students with incomplete records, returnee, and shifters.

Table II provides a description of the dataset used in the proposed model's training and testing, while the selection of qualifiers was based on the set mark by the College of Engineering and Architecture provided in Table III. An SAI of 104 has a percentile rank of 60 and Stanine of 6, this means that the qualifier performed as well as 60 percent of the applicants of this age who took the test, while Stanine of 6 indicates that the applicant's performance was slightly above average. The dataset was normalized to change the numeric values to a common scale without any distortions in the differences of range values. Also, the cross-validation of 5-fold was used to protect against over-fitting during the supervised learning process.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

The system was broken down into two distinct stages: (a) training and (b) t classification (testing). For the training and testing a ratio of 70:30 was used, respectively. For the training stage, the input features from the dataset will be used to train different models to be used for comparison. During the testing phase of the system, all of the classification models produced will be tested and the model with the highest accuracy will be optimized and be used for the

study's final classification model.

B. Machine Learning Models Used

Data mining as applied in education needs to be tested using several known possible machine learning models. These models were trained using the Classification Learner Application of Matlab R2020b. The predictors used were the criteria attributes GWA, SAI, S, and College Qualifying Exam (CQE) as seen in Figure 2. In this study, there are four categorical responses, namely, qualified for all engineering programs, not qualified, qualified for ME-CE-EE, qualified for IE only. These four response classifications were based on the actual results of successful qualifiers.

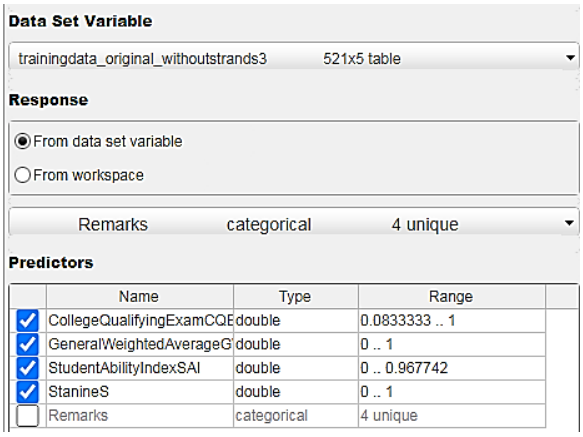

Figure 2. Predictors and response of the study

During the training phase, the training datasets were applied to Classifier Learner application of Matlab R2020b, and different machine models were trained. The performances of each model classifier are expressed in percentage accuracy. Three trained classifier models performed well and these are decision tree, KNN, and SVM. Thus, this section summarizes the models used that can answer the problem mentioned in the previous section.

C. KNN Classifier

The KNN [18], [19],often known as the lazy method. This simple yet effective tool uses the idea of varying the number of K-values and applying the formula for Euclidean distance (d), refer to equation 2, to determine a certain feature close that can map whether a student is qualified or not for a certain program.

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

D. Decision Tree Classifier

A decision tree is a machine learning model that is characterized as being simple, classic, and natural. It builds its model in the form of a tree where it breaks down a complicated dataset into smaller subsets referred to as leaf nodes. The core algorithm in building an effective decision tree is by using Entropy $H(T)$ and Information Gain, refer to equations (3(and (4).

$$H(T) = \sum_{i=1}^{C} -p_i \log_2 p_i \tag{3}$$

$$Gain(\mathrm{T}, \mathrm{X}) = H(T) - H(T, X) \tag{4}$$

For the entropy calculations in (3), decisions are made based on a measure of uncertainty or likelihood probability values. While equation (4), is all about finding the feature or attribute that returns the lowest gain until the terminal node or the leaf of the tree is achieved.

E. SVM Classifier

The support vector machine (SVM) is a classifier that uses the idea of support vectors defined by the different kernels as a reference to the study [3]. The different kernel used for this study is summarized in equation (5) to (7).

$$K(x_i, y_i) = (x_i * y_i) \tag{5}$$

$$K(x_i, y_i) = (x_i, y_i + 1)^d \tag{6}$$

$$K(x_i, y_i) = e^{\left\{\frac{-|x_i - y_i|^2}{2\sigma^2}\right\}} \tag{7}$$

The value of d determines the degree of the polynomial; and σ provides the Gaussian kernel's width.

F. Performance Metrics and Model Evaluation

The study used a machine learning algorithm that is supervised learning. Confusion matrix provides visualization of performance for each algorithm, on how well does it correctly classified the test dataset. It has four basic terms: the true positive (TP), the true negative (TN) correctly rejected prediction for specific class or a correctly predicted no, the false positive (FP) inaccurately forecast for specific class and the false negative (FN) is an inaccurately identified forecast for specific lcass. From the confusion matrix performance of the measure is computed. The following metric of success [18], [20] were used: Accuracy, Precision, Recall, F-Measure, and k-fold cross-validation.

Accuracy (A) is a metric in evaluating classification models that describe how often the classifier is correct. It's the ratio of the sum of true positive (TP) and true negative (TN) over the entire number of datasets, with a value of 1 (100%) giving the best result.

$$A_n(overall) = \frac{TP + TN}{Total\,number\,of\,datasets} \tag{8}$$

Precision (P) measure attempts to answer: When it predicts identification, how often is it correct? A P equal to "1" gives the best result.

$$P = \frac{TP}{TP + FP} \tag{9}$$

where: FP= false positive

Recall (R) is sometimes called "Sensitivity" or True

Positive Rate (TPR). These metric attempts to answer the question: When it is positive or yes, how often does it predict positive or yes? Mathematically:

$$R = \frac{TP}{TP + FN} \tag{10}$$

where: FN= false negative

F-measure (F) is sometimes called F-score or F-measure. In a statistic analysis of binary classification, it ranges from 0 to 1, where 1 gives the best result. It is mathematically defined as:

$$F = 2\frac{P}{P + R} \tag{11}$$

G. Cross-validation

Cross-validation in machine learning is estimating the competence of the machine learning models on a new dataset or unknown dataset. The purpose of this is to prevent over-fitting or high bias. The K-fold Cross-Validation process is a re-sampling that splits or divides dataset into equal sizes called K. If the data sample is split into ten, then k = 10 and is called 10-fold cross-validation. Each of these folds is considered as a validation set and the remaining are set for training. On the training dataset, a model classifier is fitted, and the model is evaluated on the validation set. A 10-fold cross-validation model was utilized in this study to verify that the system's accuracy is not simply relied on a single split of training and testing data; but rather provides an additional metric of model robustness using different training and testing for each k-fold. There is no rule in choosing the value of k, it is usually 5 or 10 [21].

3. Results and Discussion

This section summarizes collected data from the Admission and Testing Office and the dean of the College of Engineering and Architecture. It also displays results that can quantify the proposed solution for the data mining problem for DHVSU. From the 745 historical data, 58 are qualified for any engineering program, 12 are qualified for ME-EE-CE-IE, 195 are qualified for IE, and 480 non-qualified freshmen applicants. This dataset is summarized and presented by using a scatter plot in Figure 3. The scatter plot also summarizes the comparison of each of the attributes.

As shown in Figure 3, in row 1 column 2, the student's SAI are mostly above 80 and their GWA on average is 90. Referring to Table II, qualified students must have an SAI of equal or greater than 104 and a GWA of 85 or above. From these two attributes, SAI and GWA, the applicants mostly failed to achieve these requirements. The clustering patterns are present in features of SAI versus GWA is a good predictor in the output course decisions. The scatter plot on row 1 of column 1 in Figure 3 tells us that, most freshmen applicants failed to achieve to attain an SAI of 104 and CQE of 8 or more. In this plot, few are qualified for any engineering program or most of the applicants are not qualified for any engineering program.

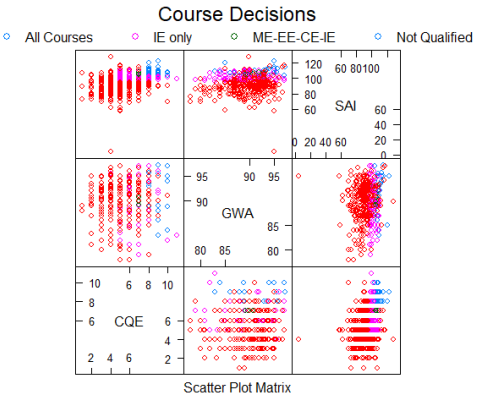

Figure 3. Scatter Plot Matrix of Attributes vs Course Decisions

Similarly, most accepted applicants are not qualified for any engineering program using the attribute CQE. By looking to row 3 column 3 scatter plot, most students have a good GWA of 85 and above but poor performance on CQE. Recalling that the required mark for CQE must be 8 and above. Clustering patterns are observable when SAI is paired with GWA and when SAI is paired with CQE. This means that SAI contributes or is a good predictor in the output course decisions when paired with either GWA or CQE in its analysis.

A. CQE and GWA Description

Unlike some schools where the qualifying entrance examination provides the decision on whether or not a student is accepted in the program, DHVSU analyzes possible concerns with regards to CQE. In Figure 5, there is a greater range of coverage in terms of the "not qualified" candidates as compared to the other three label courses. Another observation in Figure 4 is the data labeled IE qualifiers, the set mark on Table II states that the passing mark for CQE is 8 and above, however, the box plot tells us that the average score of CQE for IE program is 6 which is below the required mark.

This means that the conventional decision of admitting an "IE qualifiers" based on CQE acceptable score has deviated to a lower value of 5. Those who are qualified for the "All" engineering program have an average CQE above the minimum requirement of the college. Also, there are a minimum number of qualified freshmen for the courses"ME-EE-CE-IE" as seen in Figure 4.

The university also focuses on other evaluating attributes that may contribute to complete analysis of student evaluation. The GWA on the other hand provides a fairly distributed course decision as provided in Figure 5. Unlike on the results of CQE, where qualified applicants have distinguished scores, the GWA of qualified freshmen for IE, ME, CE, IE programs, and even those who were not qualified for any engineering program have an average GWA of 90. This means the data from the applicant's GWA

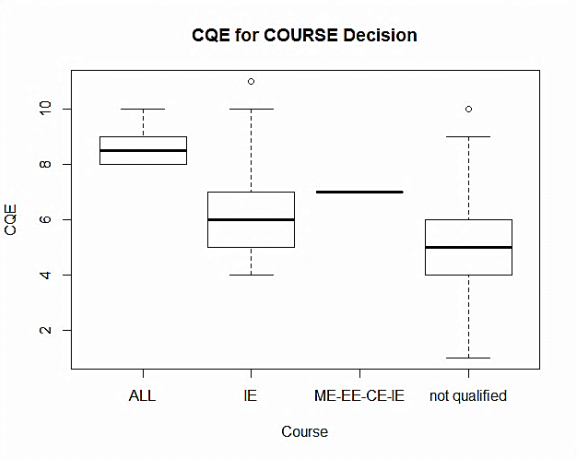

Figure 4. CQE Box plot Analysis

is not much reliable in assessing applicants as compared to those from CQE. For the past years, only in 2019 did the College of Engineering and Architecture had provided a standardized College Qualifying Exam and most of its assessments are based on the GWA.

B. Implementation of Dataset to EDM

Since the collection of data for this study is of a different scale as shown in Table II, there is a need for normalization before learning implementation. Tables IV and V present the parameters for normalization, and the sample normalize a database, respectively. The minimum and maximum values are based on the collected data. It can be seen that the maximum value of GWA is 97, whereas the maximum value of CQE is only 11 out of 15.

Table V displays the normalized value of attributes using equation (1 and Table II. The range of the normalized value is from 0 to 1.

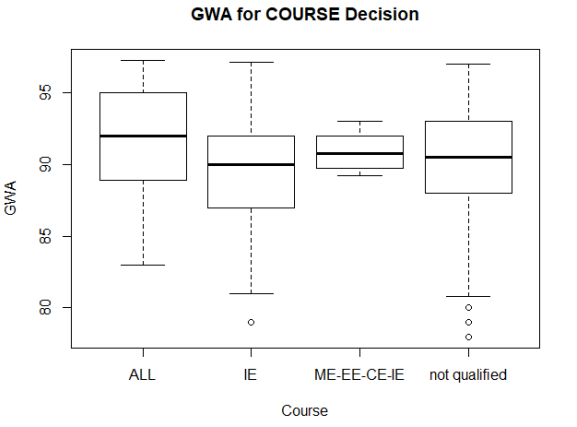

Figure 5. GWA Box plot Analysis

TABLE IV. PARAMETERS FOR NORMALIZATION

Features (Attributes/Predictors)	Minimum Value	Maximum Value
Grade Weighted Average (GWA)	78	97
School Ability Index (SAI)	57	128
Stanine (S)	1	9
College Qualifying Exam (CQE)	1	11

TABLE V. SAMPLE NORMALIZED DATA ENTRY

CQE	GWA	SAI	S
0.4	0.83116883	0.5352	0.375
0.3	0.57142857	0.4789	0.375
0.7	0.77922078	0.7465	0.625
0.5	0.81350649	0.5211	0.375
0.8	1	0.8169	0.75
0.6	0.99272727	0.5634	0.5
0.7	0.98701299	0.5915	0.5

C. Feature Correlation

In this research work, an investigation on the possible feature to feature correlation was performed and the results are shown in Figure 6. The normalized SAI shows a high correlation with normalized Stanine of r = 0.91, among other features. A linear plot slanting to the right depicts the association, which can be seen in row 3 of column 4 of the scatter plot in Figure 6.

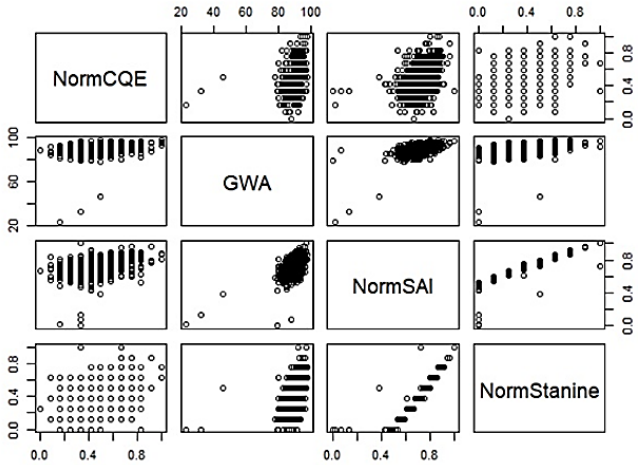

Figure 6. Scatter Matrix Features

D. Senior Academic Strand and Engineering Program Remarks

In addition to finding a correlation, the relationship between the academic strand in senior high school and the remaining four features was investigated. As a result, the strand feature does not correlate with the four given features as seen in Table VI.

TABLE VI. CORRELATION BETWEEN STRAND AND FEATURES

Features	Correlation, r
CQE	0.2770544
GWA	0.1706467
SAI	0.2647769
Stanine	0.2706938

Ideally, a freshman applicant for any engineering program should have taken the STEM strand as one of the requirements. However, this is not the case at present; students of any strand are welcome to apply for any engineering program. Further investigation was conducted to determine if the type of strand taken correlates with the results or remarks of the applicant.

From Figure 7, it can be seen that the type of strand a student took does not correlate with engineering program remarks. This means a student who took either a stem or non-stem strand can be qualified or not in any engineering program. For this reason, from Table V and Figure 7, the strand was eliminated as attributes to the study.

E. Machine Learning Model Comparison

The ten-fold cross-validation model outperforms the single fold technique in terms of accuracy. Table VII summarizes the accuracy of the three proposed models using the 521 × 5 training datasets using Matlab R2020B.

According to the performance in Table VII and Figure 8 training results, the decision tree classifier has the greatest average accuracy followed by SVM, and KNN, while Quadratic discriminant was unable to learn the training dataset.

F. Comparative Analysis of KNN, SVM, and Decision Tree during training

The Matlab R2020b was used to train the three machine learning models, KNN, SVM, and Decision Tree. The results of training are presented using the confusion

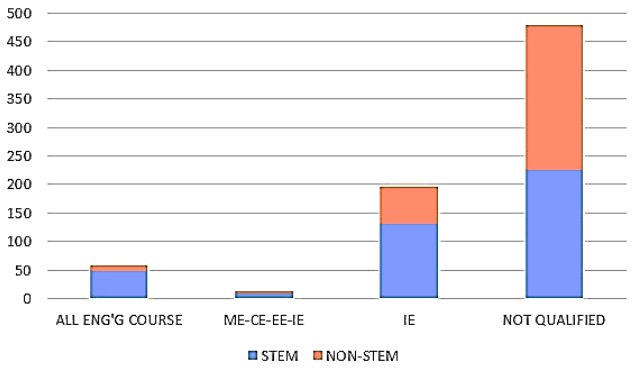

Figure 7. Geometric Graph of STEM and NON-STEM versus Remark.

TABLE VII. ACCURACY OF MACHINE LEARNING MODELS

Model	Accuracy
1. KNN	
a. Fine	95.20%
b.Medium	95.80%
c. Cubic	95.40%
2. Fine Decision Tree	**96.70%**
3. SVM	
a. Linear Kernel	93.70%
b.Quadratic Kernel	96.40%
c. Cubic Kernel	95.60%
d. Gaussian Kernel	95.40%

matrix as seen in Figure 9- 11. In this matrix, the algorithm classification performance is displayed under the predicted class concerning the true class. After the training, the algorithm performance will be evaluated on how it performs for each class. The misclassified remarks under the predicted class are counted against the true class, as well as the correctly classified ones during the training. The Medium KNN algorithm had 42 correctly classified remarks of "ALL Eng'g course" with 1 misclassified and was predicted as "IE" remarks. There were 131 correctly classified "IE," remarks, 3 misclassified as "ALL Eng'g course," remarks, and 2 misclassified as 'not qualified'. The Medium KNN also did not correctly classify the remarks "ME-EE-CE-IE", instead it misclassified the data as "ALL Eng'g course' and "IE". Lastly, there were 326 correctly classified as "not qualified' and 8 misclassified as "IE". The algorithm performance is 95.8% accuracy. Figure 10 shows the Quadratic SVM's confusion matrix this algorithm has a similar performance with KNN in classifying the "ALL Eng'g course" having 42 correctly classified classes with 1 misclassified. Similarly, the number of correctly classified "IE" and "not qualified" were also the same as the KNN algorithm. However, Quadratic SVM had 4 correctly classified the "ME-EE-CE-IE" remarks, as compared with KNN which unable to classify it correctly. The algorithm performance is 96.4% inaccuracy. Unlike KNN and SVM, Fine Decision Tree correctly classified all the "ME-EE-CE-IE" as seen in Figure 11. However, it has 2 misclassified with "ALL Eng'g course", 6 misclassified in "IE" remarks, and 9 misclassified for "not qualified" remarks. Among the three machine learning algorithms, the decision tree performance outstands in predicting the classification of the training dataset.

TABLE VIII. PERFORMANCES OF SVM, KNN, and TREE

Performance Metrics	KNN	SVM	decision tree
Accuracy	95.80%	96.50%	96.70%
Precision	0.67	0.88	0.96
Recall	0.73	0.86	0.97
F-measure	**0.7**	**0.87**	**0.96**

The confusion matrices result in the performances of

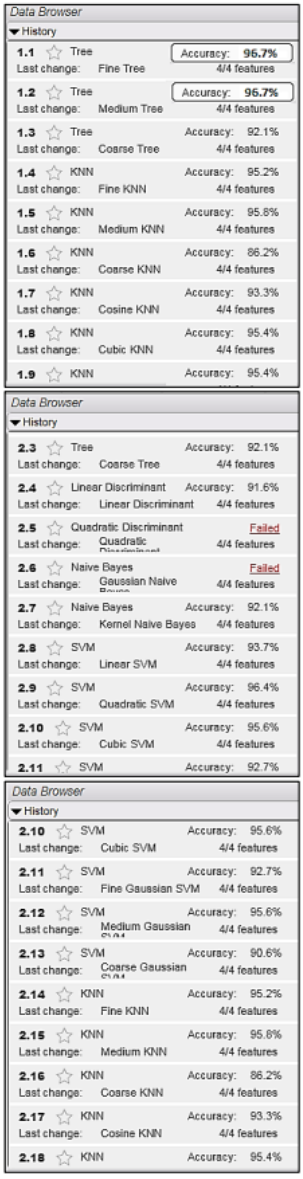

Figure 8. Training of different Machine Learning algorithm

each algorithm. The Table VIII shows that the decision tree classifier was able to predict correctly for "ME-EE-CE-IE" qualifiers. The equations (8) – (11) were used to determine the performance of metrics of each model according to four performance metrics as seen in Table VII. In terms of F-measure, Recall, Precision and Accuracy, the Decision Tree outperforms both SVM and KNN.

G. F-measure of the Tree Classifier using the test dataset

Having tested the decision tree classifier in terms of accuracy of 96.7%, there is a need to further analyze this model in terms of the 224 × 4 test dataset, and the results of the classification are presented in a confusion matrix as shown in Figure 12.

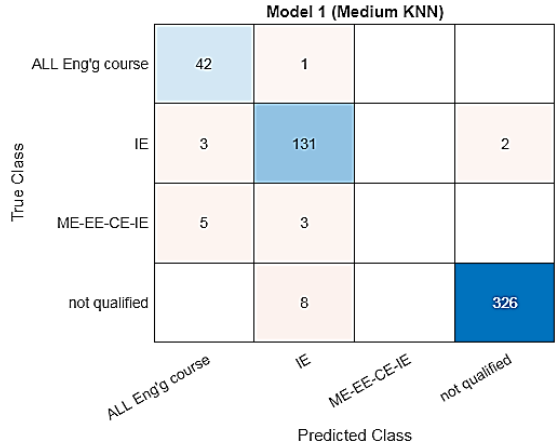

Figure 9. Medium KNN Confusion Matrix

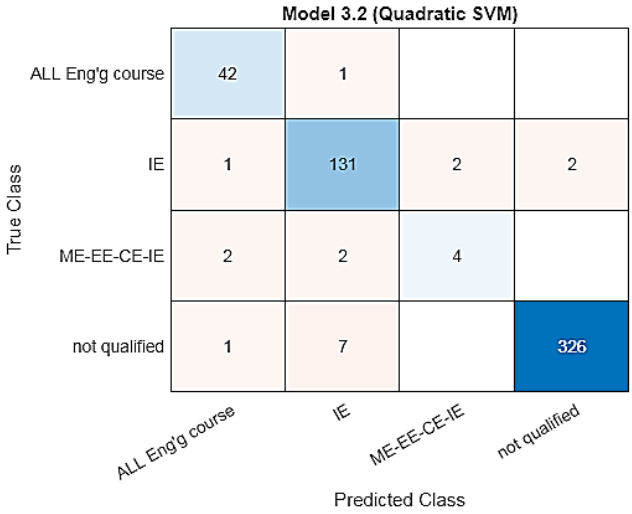

Figure 10. Quadratic SVM Confusion Matrix

TABLE IX. DECISION TREE PREFORMANCE USING TEST DATASET

Performance Metrics	Result
Accuracy	96.70%
Precision	0.96
Recall	0.97
F-measure	**0.96**

Thus, the resulting decision tree classifier as shown in Table IX has a classification rate of 96.7% when used in the 224 × 4 test datasets. It had correctly classified qualifiers of "ME-EE-CE-IE", has a true positive rate of 97%, the precision of 96%, and the F-measure of 0.96. The predictions of test dataset using trained dataset use the Matlab command as shown:

yfit=trainedModelFcn(T)

where: T is the test dataset

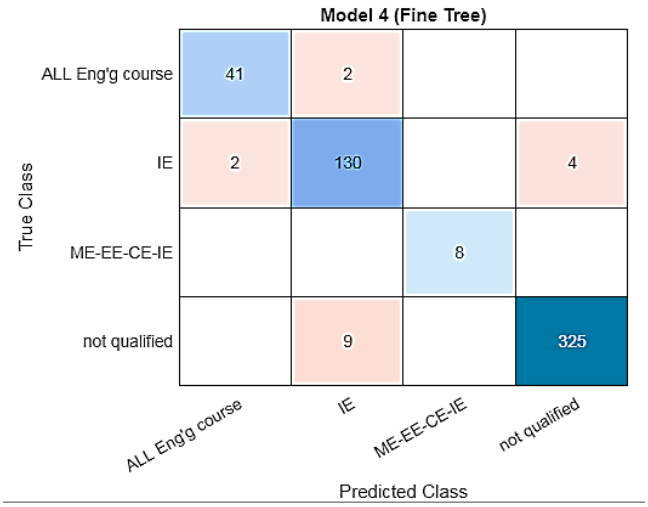

Figure 11. Fine Tree Classifier Confusion Matrix

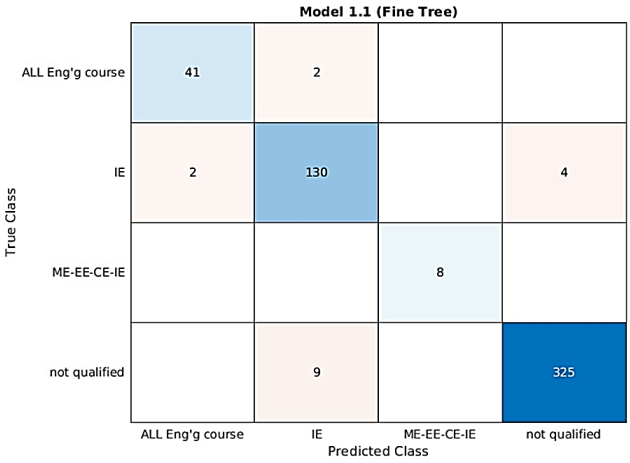

Figure 12. Decision Tree Confusion Matrix

The generated tree is a 5-level in height having Stanine as the principal root in the study. The decision rules on the succeeding nodes as shown in Figure 13 are mostly on the results of the college qualifying exam. The right edge of the decision tree shows the decision for qualified applicants, while the left edge is a decision for non-qualifiers. The decision tree was generated using the command *view(trainedmodel.ClassificationTree,'Mode','graph)*

4. Conclusions

This paper presented several machine learning models that can accurately classify qualified and non-qualified engineering freshmen for program acceptance. Based on the different simulation and metrics presented three machine learning classifiers can be used for this study namely, Support Vector Machine, K-Nearest Neighbor, and Decision Tree. Among these three, the best fits the need of the DHVSU is the decision tree classifier with an average accuracy of 96.7% and f –measure of 0.96. The principal

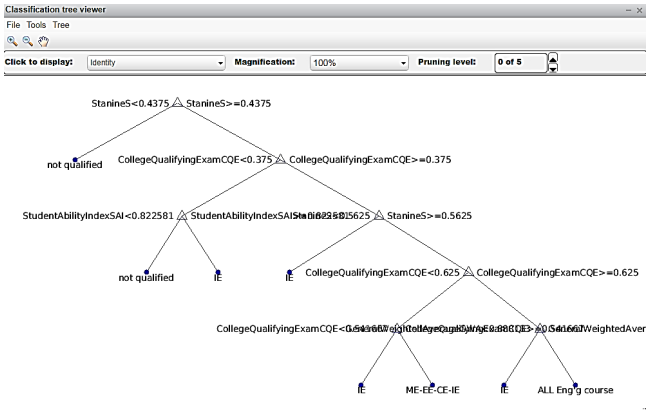

Figure 13. Decision Tree structure showing the selection qualified incoming freshmen

root of the decision tree of the study is the Stanine while the succeeding nodes are results of the college qualifying exam. This is a promising result considering the school's decision of using EDM for student evaluation. For the past years, the College of Engineering and Architecture had not considered qualifying exam as added attributes in assessing freshmen applicants. Based on this study, aside from the qualifying exam, school ability index, and Stanine value; the General Weighted Average is not enough basis in assessing an applicant as a practice by the college for many years. Results show that applicants of any senior-high-school strand have approximately the same General Weighted Average, whereas the use of the College Qualifying Exam in 2019 had provided a sufficient basis in evaluating an applicant. Using the Educational Data Mining technology, the study was able to see the pattern of acceptance in the college. Stanine and College Qualifying Exam results are good predictors in assessing freshmen applicants to any engineering program, while the school ability index is just secondary for this study. The use of Educational Data Mining greatly helps deans and program chairs in student assessment. Thus, it also avoids the deviating the passing mark for each criterion which were observed during the traditional assessment of freshmen applicant.

Acknowledgement

The researcher wishes to thank the Guidance and Testing Office, the Dean of College of Engineering and Architecture, and the Public Information Office of Don Honorio Ventura State University, Pampanga, Philippines in providing test results, data interpretation, and support during the study.

References

[1] J. C. McCall and M. M. Trivedi, "An integrated, robust approach to lane marking detection and lane tracking," in *IEEE Intelligent Vehicles Symposium, 2004*. IEEE, 2004, pp. 533–537.

[2] J. J. P. Belen, J. C. V. Caysido, A. B. Llena, E. J. O. Samonte, G. N. Vicente, and E. A. Roxas, "Vision based classification and speed estimation of vehicles using forward camera," in *2018 IEEE 14th*

International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, 2018, pp. 227–232.

[3] M. S. R. Sajib and S. M. Tareeq, "A feature based method for real time vehicle detection and classification from on-road videos," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2017, pp. 1–11.

[4] E. A. Roxas, R. R. P. Vicerra, L. A. G. Lim, J. C. D. Cruz, R. Naguib, E. P. Dadios, and A. A. Bandala, "Multi-scale vehicle classification using different machine learning models," in *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. IEEE, 2018, pp. 1–5.

[5] J. C. T. Mallare, D. F. G. Pineda, G. M. Trinidad, R. D. Serafica, J. B. K. Villanueva, A. R. Dela Cruz, R. R. P. Vicerra, K. K. D. Serrano, and E. A. Roxas, "Sitting posture assessment using computer vision," in *2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, 2017, pp. 1–5.

[6] O. Gencoglu, H. Similä, H. Honko, and M. Isomursu, "Collecting a citizen's digital footprint for health data mining," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 7626–7629.

[7] X. J. Cai, J. I. E. Ignacio, E. F. Mendoza, D. J. F. Rabino, R. P. G. Real, and E. A. Roxas, "Iot-based gait monitoring system for static and dynamic classification of data," in *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. IEEE, 2018, pp. 1–4.

[8] R. M. J. S. Bautista, V. J. L. Navata, A. H. Ng, M. T. S. Santos, J. D. Albao, and E. A. Roxas, "Recognition of handwritten alphanumeric characters using projection histogram and support vector machine," in *2015 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. IEEE, 2015, pp. 1–6.

[9] M. A. M. Sahagun, J. C. D. Cruz, and R. G. Garcia, "Wireless sensor nodes for flood forecasting using artificial neural network," in *2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. IEEE, 2017, pp. 1–6.

[10] M. A. M. Sahagun, J. C. D. Cruz, and R. G. Garcia, "Nonlinear autoregressive with exogenous inputsneural network for water level prediction," in *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, 2018, pp. 1–6.

[11] J. Ming, L. Zhang, J. Sun, and Y. Zhang, "Analysis models of technical and economic data of mining enterprises based on big data analysis," in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE, 2018, pp. 224–227.

[12] S.-M. Kim and Y.-G. Ha, "Automated discovery of small business domain knowledge using web crawling and data mining," in *2016 International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2016, pp. 481–484.

[13] Y. Zhou, "Design of intelligent guidance system for distance art education in colleges and universities based on the integration of current information literacy model," in *2018 International Conference on Robots & Intelligent System (ICRIS)*. IEEE, 2018, pp. 266–269.

[14] N. Ketui, K. Homjun, K. Poonyasiri, J. Deepinjai, and P. Luekhong, "Item-based approach for online exam performance and its application," in *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, 2016, pp. 1–5.

[15] V. Dahiya, "A survey on educational data mining," *International Journal of Research in Humanities, Arts and Literature*, vol. 6, no. 5, pp. 23–30, 2018.

[16] R. Jindal and M. D. Borah, "A survey on educational data mining and research trends," *International Journal of Database Management Systems*, vol. 5, no. 3, p. 53, 2013.

[17] Don honorio ventura state university - admission process. [Online]. Available: https://dhvsu.edu.ph/index.php/admission-menu/enrollment-guidecollege-admission

[18] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.

[19] Y.-l. Cai, D. Ji, and D. Cai, "A knn research paper classification method based on shared nearest neighbor." in *NTCIR*, 2010, pp. 336–340.

[20] C.-C. Kiu, "Data mining analysis on student's academic performance through exploration of student's background and social activities," in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, 2018, pp. 1–5.

[21] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

Mary Anne M. Sahagun is an Associate Professor at Don Honorio Ventura State University, Pampanga, PHILIPPINES. She is also an accreditor of the Accrediting Agency of Chartered Colleges and Universities of the Philippines, a member of the Institute of Electrical and Electronics Engineers (IEEE) Region 10. She has been a research leader in a Japan Funded Research Project under the Asia Pacific Telecommunity (APT) last 2015. She had presented and published researches related to Artificial Neural Network, Image Processing, Wireless Sensor Network, Telemedicine, Fuzzy System, IC Design.She had also produced modules for Communication Engineering, Power Electronics, Basic Electronics, and published Circuit Laboratory Manual. At present, she is the chairperson of the Electronics Engineering Department and currently taking her graduate program in Ph.D. in Electronics Engineering.