PREDICTING STUDENT PERFORMANCE IN ENGINEERING DRAWING USING SUPERVISED LEARNING METHODS

By

Akpofure Avwerosuoghene Enughwure akpofure.enughwure@nmu.edu.ng
John Damilola Oluwafemi john.oluwafemi@nmu.edu.ng
Lecturer, Nigeria Maritime University, Okerenkoko, Delta State, Nigeria

ABSTRACT

The study examined the performance of engineering students in engineering drawing as a basic engineering course in Federal Nigerian University. Many students have been discovered to perform poorly in the course for reasons like poor orientation on the introduction to engineering drawing, little attendance to lectures, inadequate drawing materials, little or no personal practice hours, etc. This paper addresses identifying the chances of students' performing well in the introductory course, finding the best machine learning (ML) method between logistics regression and decision tree, and examining the variables that determine the performance of the students in engineering drawing. Data was collated using the paper-based questionnaire from undergraduate engineering students who offered the course in Civil, Electrical, Petroleum and Gas, Mechanical and Marine Engineering departments. The questionnaire was distributed in an unsupervised manner and 210 entries were received. The classification and clustering methodology was used alongside machine learning techniques like logistics regression and decision trees to predict students' performance in the course. Sklearn modules in Python Jupyter notebook and the training dataset were used to build the models. After running the models on a testing data, all the models were capable of classifying a successful outcome with accuracy between 67% - 81% and logistics regression having the highest chance of prediction. The prediction models developed could help both students, course instructors, and academic administrators take proactive steps in helping the students excel in the engineering drawing course.

Key Words: Machine learning, Engineering drawing, Predictive Model, Decision Tree, Logistics Regression.

Abbreviations

EDM	Educational Data Mining
CRISP-DM	Cross-Industry Standard Process for Data Mining
GPA	Grade Point Average
CGPA	Cumulative Grade Point Average
KNIME	Konstanz Information Miner
MATN	Student Matriculation Number
DEPT	Student's Department
SEX	Student's sex
STA	Student's state of origin
ANPP	Average number of personal practice hours per week
DBTS	Does Student own a tee-square
SETS	Does Student own a set square
FREC	Does Student own a French Curve
ROST	Does Student have a rotary set
HBPN	Does Student have an HB pencil
2BPN	Does Student have a 2B pencil

ATDWN	Did Student offer technical drawing in secondary school
AGE	Age of the student
NLT	Number of Lecturers/Tutors
REST	Indicator for Student's Resumption (Early/Late)
CSV	Comma-Separated Values
ML	Machine Learning
ROC	Receiver's Operating Characteristic Curve
AUC	Area under the receiver's operating characteristic curve
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
RMSE	Root Mean Square Error

1.0 INTRODUCTION

Every engineering student is required to take at least two courses in engineering drawing – a high-impact and high-enrolled course in Nigerian Universities. These courses are the vital bases and rudiments of advanced design engineering courses like highway and road design, engineering machine design, electrical wire design, chemical plant design, etc. However many students fail these courses because they do not have a good orientation on the introduction to engineering drawing, little attendance to lectures, inadequate drawing materials, little or no personal practice hours, etc.

The application of machine learning in the academic space offer students, lecturers, and even school administrators various possibilities including student retention improvement, assessing students, student performance prediction, and so on (Kucak, Juricic & Dambic, 2018). Prediction of student academic performance is an imperative research topic in various academic majors since it offers great gain to the tutors as well as students (Parneet, Manpreet & Singh, 2015). With the use of big data and machine learning in the university environment, engineering lecturers can assist students who are on the verge, struggling with engineering drawing (Shevtshenko, Karaulova, Igavens, Strods, Tandzegolskiene, Tutlys, Tavahodi & Kuts, 2017)

In the study, we attempted to address these questions:

1. What are the chances of identifying student performance in this introductory course?
2. Within this specific framework, which of the following machine learning techniques [Logistics Regression, Decision Tree] will perform best?
3. What are the key variables that determine the propensity of students to achieve certain academic performance levels in engineering drawing?

This study is different from other studies in more than a few ways. First, and in what regards the methodology, we proposed a classification method as well as clustering. The cluster approach is used to group students with similar attributes (age, average personal practice hours per week) in a bid to aid the tutors' development of more ways to help the students perform best, while the classification method for predicting the student performance. Supervised machine learning techniques like logistics regression, decision trees are used to predict students'

performance in this course. Finally, we will evaluate the important attributes of the students in this predictive model.

In various aspects and levels of education, advanced methods have been employed to predict the performance of students in schools with Educational Data Mining (EDM). For example, using the Cross-Industry Standard Process for Data Mining (CRISP-DM), predictions on students' performance were made on high school juniors in Brazil for the years 2015 and 2016. The data provided was used to understand the students and how they learn, to design educational policies aimed at improving the academic performance and reduce the failure rates at the end of the school year (Fernandes et al., 2018). Similarly, Burgos et al. (2017) used the logistic regression data mining technique to detect the dropout rate in e-learning courses. This data helped them launch a tutoring plan, which now serves as a model to control or reduce the students' dropout level.

Given attaining a smart campus in Nigeria, a research-based on engineering students in a private university was carried out to show the performance of the students in the various engineering departments. Through the use of Grade Point Average (GPA) of the students as the source of data and tools like Microsoft Excel, the performance variance of students from levels to levels yearly was shown and used to determine if there were significant differences in the GPA data of engineering students throughout their five-year study period (Popoola et al., 2018). In close relation to this study, Adekitan and Salau (2019) conducted a predictive analysis on the impact of the performance of engineering students in their first three years on their fifth or final year and Cumulative Grade Point Average (CGPA). In line with the aim of EDM, students who are likely to graduate with poor grades or not graduate at all were identified and early intervention can be deployed for such cases. The results were achieved through the use of Konstanz Information Miner (KNIME) and regression analysis in MATLAB.

In studying students' performance, various approaches can be considered. Helal et al. (2018) discovered that students' heterogeneity such as age, gender, and attendance type should be considered in building predictive models for determining their performance as students with different socio-demographic features or study modes may reveal varying learning motivations. Through the classification data mining method, sub-models were generated from students' subgroups to identify students at risk of academic failure. In contrast, Asif, Merceron, Ali, and Haider (2017) analyzed the performance of undergraduate students in Information Technology using admission marks from high school and final marks from first and second-year courses in the university without considering any other demographic or socioeconomic factors. Also, their research-derived courses that could serve as effective indicators of students performance in a degree program. Using the clustering technique, they investigated the progress of students in their academic performance over the four-year degree programme.

In a review of the techniques used in data mining to predict students' performance, Shahiri, Hussain, and Rashid (2015) addressed the attributes and methods mostly used in educational data mining. The commonly used attributes by researchers include CGPA for its tangible values, internal assessments such as assignment marks, quizzes, lab work, class tests, and attendance. Some researchers also consider demographics like gender, age, family background, and disability, as well as extra-curricular activities, high school background, and social interaction network. Only a few researchers have used psychometric factors such as student interest, study behavior, engagement time, and family support. On the other hand, the methods include tasks like classification, regression, and categorization. Some of these attributes and methods have used in the related research referred to in this study.

These various researches have explored various attributes and methodology of predicting students' performance. The performance of engineering students has been predicted in other

studies, but this study focuses on predicting students' performance in engineering drawing using supervised learning methods, an aspect that has not been researched.

The authors of this study observed that students usually struggle to perform well in engineering drawing courses hence the rationale of this study is to build a predictive model that will improve the students' performance in engineering drawing. Tutors and administrators will be able to access as well as predict student performance in this course based on student's work ethics as well as possession of certain drawing tools.

2.0 METHODOLOGY

After a brainstorming session with academic experts as well as extensive online literature research on student performance, a list of factors was considered to have a huge impact on students' performance in this course. These factors became the input variables for the model. A paper-based questionnaire was used to gather the students' data on their participation in the course. The variables used in this study shown in table 1:

Table 1: Variable expression from the questionnaire

Variable Name	Description	Domain	Data Type
MATN	Matriculation Number	Student's Unique number	Object
DEPT	Department	Electrical, Petroleum and Gas, Civil, Mechanical and Marine	Object
SEX	Student's Sex	Male, Female	Integer
STA	State	At least one Nigerian state	Object
ANPP	Number of student's practice hour per week	1-10	Integer
DBTS	Does Student own a tee-square?	Yes or No	Integer
SETS	Does Student own a set-square?	Yes or No	Integer
FREC	Does Student own a French Curve?	Yes or No	Integer
ROST	Does Student have a rotary set?	Yes or No	Integer
HBPN	Does Student have an HB pencil?	Yes or No	Integer
2BPN	Does Student have a 2B pencil?	Yes or No	Integer

ATDWN	Did Student offer technical drawing in secondary school?	Yes or No	Integer
AGE	The age of the student	16-50	Integer
NLT	Number of Lecturers/Tutors	1-5	Integer
REST	Indicator for Student's Resumption (Early/Late)	Early or Late	Binary

After the design and distribution of the questionnaires to the students in an unsupervised manner, we received 210 entries. The data were stored in a common-separated values (CSV) file.

2.1 TOOLS AND TECHNIQUES

In the paper, the tools used for the study were Python Jupyter Notebook was used for the data analysis and visualization with the use of various python modules like pandas, NumPy, Matpotlib, seaborn and others. The data was stored in a CSV format. Various machine-learning techniques like logistics regression classifier as well as decision tree algorithms employed in this experiment.

2.1.1 Logistics Regression Classifier

Logistics Regression is a statistical technique used for linear modelling based on the logistic function expressed in a mathematical form shown below:

$$P(x) = \frac{1}{1 + e^{-(\beta o + \beta i x)}}$$

Where βo is the bias term that helps the model to deal with any constant value offsets in the relationship between the input (x) and the output (y). βi is the parameter vector of the model that has been learned by training the model. P(x) is the output probability that a record (a combination of inputs 'x') belongs to a class 'y' (Mgala & Mbogho, 2015).

2.1.2 Decision Tree

A decision tree is a structure made out of nodes and arcs where an internal node presents a decision based on attribute values (input values), and the arcs represent the choice made in the node. It ends on a leaf node, which represents the class ('y') to be assigned (Guarin, Ernesto, Guzman, and Gonzalez, 2015). To classify a record with a decision tree, it starts by the root node and goes down one level at a time depending on the results of the conditions tested on every node; when it ends on a leaf node, the record is classified `according to the class of that leaf node.

2.2 Preprocessing and Partitioning Data

Dataset collected for this study had 250 rows of data with blank data points in the following columns: 'ANPP', 'ROST', '2BPN','ATDWN','AGE','NLT','REST'. In the preprocessing phase, blank data points in column 'AGE' were replaced with the mean age of the students. Columns like 'ANPP', 'ROST', '2BPN','ATDWN','NLT','REST' were filled with zero which means the student had no personal practice hours, rotary set, 2B pencil, did not attempt

technical drawing in WAEC/NECO or high school level and resumed late for the semester. After cleansing the data and removing duplicates, the dataset used had two hundred and ten (210) rows with one hundred and sixty-eight (168) successful outcomes and forty-two (42) failed outcomes.

The dataset used for this analysis was shared into random subsets: training set, and testing set, using the train-test split method in sklearn. 80% of the data point belong in the training set while the remaining 20% in the testing set. The training set is used to fit the model of interest. The built model is then applied to the testing set in a bid to assess the performance of the model. Parameters like precision score, recall score, and others are determined when the model is tried by the test set. The testing dataset will behave like a set of data imputed into the existing model by random users.

2.3 Classification and Prediction Modelling

The methodology in this research was predictive modeling via two different Machine Learning classification methods. Given the dataset size, Logistics Regression Classifier and Decision Tree Algorithms were imputed for grouping and predicting student performance in the engineering drawing course. This is because although these methods are relatively simple to implement and they do well in small-to-medium-sized datasets ensuring the model does not overfit the training data. The dependent variable was the student's outcome and the independent variables were sex, age, average personal practice hours, possession of drawing tools like tee-square, French curve, drawing board, set-square, HB pencil, 2B pencil, attempted technical drawing in secondary school, attempted technical drawing exams in WAEC/NECO, Time of Resumption and Number of Lecturers.

2.4 Model Quantitative Performance Metrics

Since this paper employed more than one ML classification method, there is a clear need to monitor the performance of the models using certain performance metrics. In this paper, the models were monitored by model accuracy, root mean square error, model accuracy after cross-validation, and the area under the receiver characteristics curve factor.

The accuracy of a model is determined by confusion matrix parameters. They are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). These parameters are used to check how well the model classifies an outcome of "success" as "success, "success" as "failure", "failure" as "failure and "failure" as "success" respectively. These parameters are obtained in a confusion matrix as shown below:

Table 2: A Sample Confusion Matrix

	Predicted Values		
		0	1
Actual Values	0	TP	FN
	1	FP	TN

Root Mean Square Error (RMSE) is an attribute used to measure the difference between the predicted values and the actual values of the model. The lower the RMSE, the better the model will perform when classifying a positive label for a positive outcome.

Mathematically, RMSE is given by

$$\sqrt{(\sum_{i=1}^{n} \frac{(X-Y)^2}{N})}$$

Where X = Predicted Values, Y = Observed Values and N = Number of Observations

After training a model, there is a chance that the model will correspond too closely to the trained dataset (overfitting) and fail to respond or predict future learning reliably. Whenever this happens, it is imperative to perform cross-validation on the model. Since the dataset used in the study is relatively small, models were cross-validated by dividing the training set by equal N-random parts. The model is trained with N-part and tested by the (1-N) parts. This is done for N times and the final model accuracy is the arithmetic mean of all the N- model accuracies.

Table 3: A comparative table between the two models using selected parameters

Machine Learning Algorithms	Model Accuracy on Train Data	Model Accuracy on Test Data	Root Mean Square Error	10 Fold Cross Validation on Test Data	AUC Factor
Logistics Regression	81.0%	81.0%	38.46%	79.86%	0.62
Decision Tree	100%	69%	57.74%	70.56%	0.52

The receiver operating characteristics (ROC) curve is a plot of the true positive rate of the y-axis and false positive rate on the x-axis. The area under the receiver operating characteristics curve is used to assess the model capability to distinguish between classes. The ideal point of the plot is the top left corner which implies that the model perfectly predicts all the success as success. In this condition, the area under the curve score is equal to one. A classifier whose ROC generates a curve closer to the top-left corner indicates a better performance. As the curve approaches the top-left corner, the area under the curve (AUC) increases hence the higher the value of the (AUC), the better the model can predict failure as failure and success as success.

3.0 RESULT

With the use of Sklearn modules in Python Jupyter notebook and the training dataset, the models were built. The models' performance was assessed by running them with the testing data. Recall on the independent variables, 0 and 1 represents failure and success respectively. Table 2 shows the Machine Learning algorithm used for the study as well as their quantitative performance metrics used in this research: All the models were capable of classifying a successful outcome with accuracy between 67% - 81% with logistics regression having the highest chance of prediction. The decision tree algorithm experienced overfitting with the train data. It had a 100% accuracy on the train data however performed at 67% hence performed poorly with the test data with a root mean square error of 57.74% between the predicted y-values and the test y-values. The overfitting that occurred in the decision tree algorithm prompted the need to carry out cross-validation of the various models. 10-fold cross-validation was used in this study. In this process, the trained data was divided into 10 random equal parts. The model ran on each part and resulted in an average accuracy of 70.56%. The area under the receiver operating characteristics factor is 0.52. The logistics algorithm performance is as follows: accuracy on

trained and test dataset 81% with a root mean square error of 38.46%. This had an AUC factor of 0.62 as well as an accuracy of 79.86% after 10 fold cross-validation on the test data.

Another interesting result we got from our analysis is the features' importance to the outcome. Feature importance is a set of techniques that allocate a score to input variables based on how useful they are at predicting the target variables. This means the impact of all the independent variables on the dependent variable. This process plays an imperative role in a predictive modelling project since it provides insight into the data, the model as well as an efficiency enhancer. The variable importance is based on a concept called Gini importance (Villar and Raya, 2015). The Gini importance is computed as the normalized total reduction of the criterion brought by the variable. With the use of the Decision Tree classifier method in the sklearn module in python, the variables importance was implemented. We were able to generate the variables important to the target variable and ranked them in ascending order as displayed in figure 1.

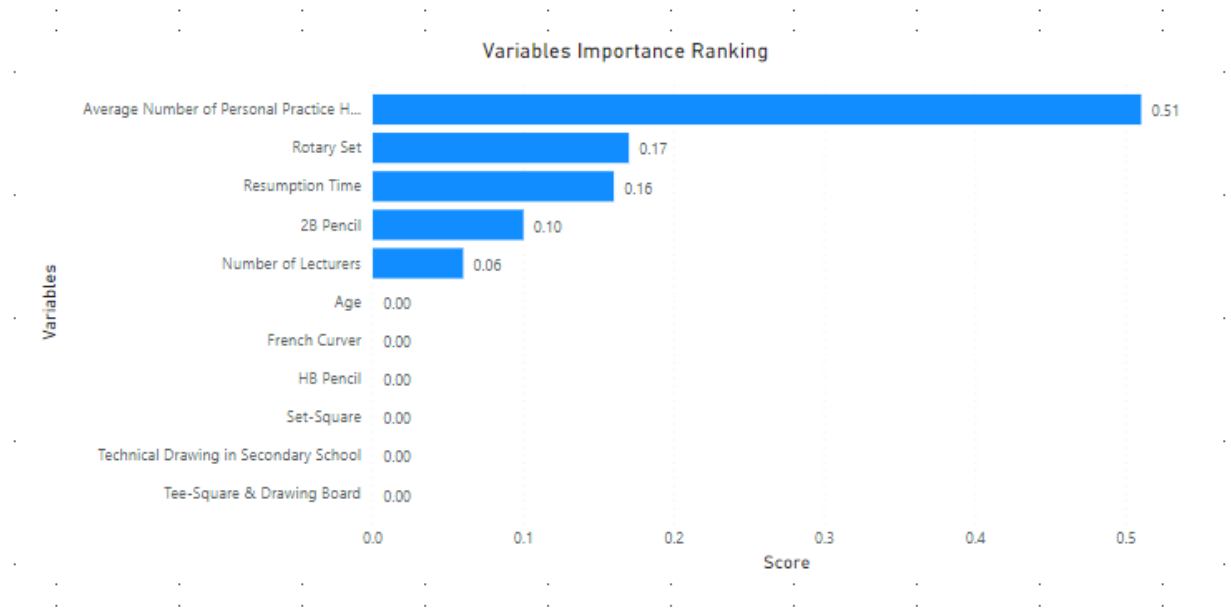

Figure 1: A bar chart that shows the variable importance ranking in the predictive model

4.0 DISCUSSIONS

Two classification models (logistics regression and decision tree) were built to predict the performance of students in introduction to engineering drawing. The models were compared to one other using the following parameters: model's accuracy on train and test data, RMSE value, model's accuracy after cross-validation as well as the AUC factor. Logistics Regression performed better than the Decision Tree. One interesting insight unraveled in this study is the variable importance in predicting student performance in engineering drawing. The performance of a student in engineering drawing heavily depends on the average number of personal practice hours. Given that the course is a practical one where tutors introduce various drawing topics and concepts, carry out a few drawing examples in class. There is a need for students' to make out time to review the examples in class as well as implement these concepts in other drawing problems. Another insight provided in this study is the zero impact of student participation in technical drawing in secondary school to their performance in engineering drawing in the university. Given that the questionnaire only asked the students if they offered technical drawing in secondary school not their performance in technical drawing in their Senior Secondary School

Leaving Exams as well as the drawing concepts they learnt during their time in secondary school hence there is a possibility that most students might not fully grasp sufficient understanding of drawing concepts.

Therefore, looking at the effect of the variable-technical drawing in secondary school, it can be seen that the students that offered it in secondary school did not acquire sufficient prior knowledge needed to give them an edge in the course.

The knowledge derived from the prediction models may be helpful for course instructors and academic administrators to help engineering students excel in their engineering drawing study so that proactive support strategies can be implemented promptly. The experimental results demonstrated the effectiveness of using students' participation in a course in predicting student academic performance. The study revealed that the students' practice hours play a vital role in their academic performance. The rotary set was ranked second on the chart. This is because, with the proper use of the rotary set, a student can comfortably draw various arcs and other circular lines.

There is a need to conduct further research to address the limitations of this paper. Firstly, we need to explore other machine algorithms like support vector machines, XG Boost to ascertain if a better prediction model accuracy will be obtained. Secondly, we need to check if other socioeconomic variables like parents' financial status, relationship status, and number of siblings can affect students' performance in engineering drawing.

REFERENCES

Adekitan, A., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon 5* (2019), 1 – 16. Doi: 10.1016/j.heliyon.2019. e01250

Asif, R., Merceron, A., Ali, S. & Haider, N. (2017). Analyzing undergraduate students' performance using educational data mining. Doi: 10.1016/j.compedu.2017.05.007

Burgos, C., Campanario, M., Pena, D., Lara, J., Lizcano, D. & Martinez, M. (2017). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering 000* (2017), 1-16. Doi: 10.1016/j.compeleceng.2017.03.005

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. & Erven, G. (2018). Educational data mining: Predictive analysis of academic performance of public schools in the capital of Brazil. *Journal of Business Research,* 1 – 9.Doi: 10.1016/j.jbusres.2018.02.012

Guarin, L., Ernesto, C., Guzman, E.L., Gonzalez, F.A.. A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. Tecnologias del Aprendizaje, IEEE Revista Iberoamericana de 2015;10(3):119–125.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. Knowledge Based System, 2 -33. Doi: 10.1016/j.knosys.2018.07.042

Kucak, D; Juricic, V. & Dambic, G. (2018). Machine Learning in Education - a Survey of Current Research Trends, Proceedings of the 29th DAAAM International Symposium, pp.0406-0410, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-20-4, ISSN1726-9679,Vienna,Austria DOI: 10.2507/29th.daaam.proceedings.059

Mvurya Mgala and Audrey Mbogho. (2015). Data-driven intervention-level prediction modeling for academic performance. In Proceedings of the Seventh International Conference on Information and Communication Technologies and Development. ACM, 2.

Parneet Kaura, Manpreet Singh, Gurpreet Singh Josanc "Classification and Prediction based Data Mining Algorithms to Predict Slow Learners in Education Sector" Science Direct Procedia Computer Science 57 (2015) 500 – 508 2015 (ICRTC- 2015).

Popoola, S., Atayero, A., Badejo, A., John, T., Odukoya, J. & Omole, D. (2017). Learning analytics for smart campus: Data in academic performances for engineering undergraduate in a Nigerian private university. *Data in Brief*, 1 -19. Doi: 10.1016/j.dib.2017.12.059

Shahiri, A., Husain, W. & Rashid, N. (2015). A review on predicting students' performance using data mining techniques. *Procedia Computer Science 72* (2015), 414 – 422.

Doi: 10.1016/j.procs.2015.12.157

Shevtshenko, E.; Karaulova, T.; Igavens, M.; Strods, G.; Tandzegolskiene, I.; Tutlys, V.; Tavahodi, S. & Kuts, V. (2017). Dissemination of Engineering Education at Schools and its Adjustment to Needs of Enterprises, Proceedings of the 28th DAAAM International Symposium, pp. 0044-0053, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-11-2, ISSN 1726-9679, Vienna, Austria, DOI: 10.2507/28th.daaam.proceedings.006

Villar, J., G. & Raya, J. M. (2015). Use of a Gini index to examine housing price heterogeneity: A quantile approach. Journal of Housing Economics Vol. 29, pp 59-71. Doi: 10.1016/j.jhe.2015.06.001