Prediction of Student Performance in Engineering Drawing Using Machine Learning Methods and Synthetic Minority Oversampling Technique (SMOTE).

Enughwure, Akpofure Avwerosuoghene (Chief Author)[1],
Ogbise, Ebitiminipre Mercy[2], Adia, Ogheneruno[3]

ABSTRACT

Engineering drawing courses are very crucial to engineering students forming the bases for more advanced design engineering courses. Seeing the high failure rate of engineering students in Nigerian higher institutes for various reasons. This research paper predicted the performance of students in engineering drawing courses at introductory level. Data was collected using paper-based questionnaire from engineering students in different departments. Logistics regression classifier and decision tree machine-learning algorithms were employed. SMOTE was introduced into the training phase in a bid to improve the prediction accuracy of the model. The machine-learning model was built on Kaggle with tools from Python Kernel. After running the models on a testing data, all the models were capable of classifying a successful outcome with accuracy between 67% - 78%. Logistics regression had the highest chance of prediction. The introduction of SMOTE clearly improved the prediction rate.

Keywords: Engineering drawing, Machine Learning, Student Performance.

INTRODUCTION

In recent times, the application of machine learning methods in various industry areas have become more widespread and sought after. In Nigeria, data mining which promotes the collection and use of data has been on the increase in so many sectors including education. Several educational problems can be tackled using data collected from students periodically. In a bid to tackle educational problems and proffer effective solutions, data mining uses the given data of students, lecturers and other staff to identify loopholes in the educational system, which also involve predicting students' performance in various course areas. Kucak, Juricic & Dambic (2018) rightly stated that, "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.

Considering the engineering department in Nigerian higher institutions, the engineering drawing course area is an aspect that has recorded great failure rate that demands ways to ameliorate the situation (Enughwure & Oluwafemi, 2020). In the engineering departments in Nigerian higher institutes of learning, students are required to take at least two engineering drawing courses at the foundational level. These courses introduce students to more advanced design engineering courses like highway and road design, engineering machine design, electrical wire design, chemical plant design, etc. Seeing that engineering drawing is vital to engineering students in Nigeria, the success rate in the courses is required to be high. However, many students fail these courses for various reasons such as faulty introduction to engineering drawing, infrequent attendance to lecturers, inadequate or poor drawing materials, lack of personal practice hours and other reasons. Adequate measures must be taken to increase the success rate amongst engineering students in engineering drawing. 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).

LITERATURE REVIEW

A great number of research works have been done in the field of using machine-learning techniques for educational purposes. Considering the other research works in education using machine-learning techniques, Ghada, Algobail, Almutairi & Almutery (2016) created a data-mining model for predicting student performance in a programming course based on their performance in English and mathematics courses. They used the Classification Based on Association rules (CBA) algorithm to build a classification model that investigated the poor performance of students in a programming course by identifying the association rules relating the programming course to mathematics, other courses and then using these rules to predict the students' grades in the programming course.

In their research, Altabrawee, Ali & Ajmi (2019) pointed out the various attributes used as input into the machine learning algorithm and even introduced two new attributes, which are the effect of using the internet to study and the amount of time students spend on social media. They used the four machine learning techniques, which are Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression to build a model. The result showed that the Artificial Neural Network had the highest level of accuracy. Similarly, Hamoud, Hashim & Awadh (2018) explored the decision tree algorithm to identify the factors that affect students' performance the most. The study will help both students and instructors find ways to improve the educational quality.

Oladokun, Adebanjo and Charles-Owoba (2002) approached their research from an interesting point of view where they employed an Artificial Neural system to predict the performance of students before they are even admitted into the university. By considering different pre-admission factors capable of affecting students' performance in the university, the study showed that an artificial neural network model could enhance the effectiveness of university admission system.

In addition, Buenaño-Fernández, Gil & Luján-Mora (2019) they applied machine learning techniques to predict the final grades of students using their academic historical data with the aim of reducing the dropout rate and improve the overall education system. They used computer-engineering students as a case study.

The various perspectives that researchers have approached predicting students' performance and using machine learning to improve education in different countries have shown the great extent to which machine learning can benefit education. This study sets out to predict the performance of engineering students in the introductory courses to engineering drawing by identifying the machine learning techniques that would work best. In addition, key variables that determine the performance of the students in engineering drawing will also be considered.

METHODOLOGY

Considering the solution, this research aims at proffering; suitable machine-learning techniques have to be employed. To effectively build a model to predict engineering students' performance in engineering drawing courses, a list of factors that can have a huge impact on students' performance are considered. These factors serve as input variables for the model. Enughwure & Ogbise (2020) also capture this idea in a review work where they discussed the various machine learning methods and variables that have been used by different researchers to predict students' performance in various course areas.

Data was collected for this research using a paper-based questionnaire. Students in engineering departments participated in filling the questionnaire. The variables used in the study are shown in table 1:

Table 1: Variable expression from the questionnaire

Attributes	**Value**
Age	16-30
Sex	Male, Female
State of Origin	At least one Nigerian state
Marital Status	Single, Married
Average Practice Hours per Week	1-10
Early Resumption	Early, Late
Department	Electrical, Petroleum and Gas, Civil, Mechanical, Marine
Possession of HB Pencil	Yes, No
Possession of 2B Pencil	Yes, No
Possession of Tee-Square	Yes, No
Possession of Set-Square	Yes, No
Possession of French Curve	Yes, No
Possession of Rotary Set	Yes, No
Attempted Technical Drawing in secondary school	Yes, No
Number of Tutors	1-5
Outcome	Pass, Fail

After the design and distribution of the questionnaires to the students in an unsupervised manner, we received 210 entries. The data was stored in a Common-Separated Values (CSV) file.

2.1 TOOLS AND TECHNIQUES

The machine-learning model for this research was built on Kaggle, which is one of the world's largest data science platforms with powerful tools and resources within a Python Kernel. Python programming language was also used. In order to create effective visualizations for the model, Python packages such as Panda, Numpy, Matplotib, Sklearn and SMOTE (Synthetic Minority Oversampling Technique) were used. The choice of machine-learning algorithm is critical to the performance of the model, so this research employed Logistics Regression Classifier and Decision Tree.

2.1.1 LOGISTICS REGRESSION CLASSIFIER

Logistics Regression is a non-black box model (a function that is too complicated for any human to comprehend) that can be in binary, multinomial or ordinal form function (Rudin C., 2019). In this study, we have only two states (failure and success) hence; we are using the binary logistics regression. The logistic regression takes the data inputs and predict which class the input belongs to. If the prediction is less than 0.5 then it takes the output as class 0 (failure) otherwise, it takes output as class 1 (success).

2.1.2 DECISION TREE CLASSIFER

Also known as, hierarchical classifiers, decision tree classifier need multi-level discrimination to determine which class a particular pattern belong. They are flexibly able to tackle multi-class as well as binary classification. In this work, we employed univariate decision trees since only a single feature participates in node splitting with information gain, Gini index and gain ratio as splitting criteria (Wang et.al, 2020)

2.1.3 SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

Imbalanced learning is one of the most common challenges in machine learning. This occurs when one or more classes has a higher amount of data points (majority) than other classes (minority). When a standard classifier is employed in the scenario above, the classifier performance is usually biased towards the majority class instances while sacrificing the minority class (Elreedy & Atiya, 2019). Hence, the chance of predicting the minority class is hampered and the majority class is increased.

With the use of K-means neighbour algorithms, SMOTE generates extra sample for the minority class thereby enhancing the chance of predicting the minority class in a learning session.

2.2 PREPROCESSING AND PARTITIONING DATA

The dataset collected for this study had 250 rows of data with blank data points in certain columns. The columns which were empty were the ones with the following questions: "Student's average number of personal practice hours per week", "Does Student have a rotary set?", "Does Student have a 2B pencil?", " Did Student offer technical drawing in secondary school?", "Student's Age", "Number of Lecturers/Tutors", and " Indicator for Student's resumption (Early/Late)". The blank spaces in these columns except the age column were filled with zero. The "Age" columns were replaced with the mean age of the students. The data replacements took place during the preprocessing phases. The data went through a cleaning process to eliminate irrelevant inputs. 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 is imbalanced because of the unequal distribution of classes within the dataset, since the successful outcome has an 80% chance of being picked in a random call than failed outcome with 20%. This imbalance created a need to employ Synthetic Minority Oversampling Technique (SMOTE). The SMOTE algorithm generates synthetic data by looking at the neighborhood of minority classes and generating new data points within the neighborhood.

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 was used to fit the model of interest afterwards, the built model is then applied to the testing set in a bid to assess the performance of the model. Certain parameters like precision score, recall score, and others were determined when the model was 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

This research work used the predictive modeling methodology via two different Machine Learning classification methods. Given the dataset size, Logistics Regression Classifier and Decision Tree were imputed to predict student performance in the engineering drawing course. These methods were used because they are relatively simple to implement. Although they do well in small-to-medium-sized datasets, they still ensure the model does not over fit the training data. The dependent variable was the **students' outcome** and the features were **state of origin, department, sex, age, average personal practice hours, possession of drawing tools like tee-square**. Other features were **french curve, drawing board, set-square, HB pencil, 2B pencil, attempted technical drawing in secondary school, attempted technical drawing exams in WAEC/NECO, resumption time and Number of tutors.**

2.4 MODEL QUANTITATIVE PERFORMANCE METRICS

It is imperative to monitor the performance of the models using certain performance metrics. This study measured the performance of the models by using the model accuracy, precision, recall, f1score, model accuracy after cross-validation, and the area under the receiver characteristics curve factor.

Model Accuracy

To get a model's accuracy, the function of its confusion matrix parameters, which are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) has to be carefully considered. TP predicts "Yes" as a "Yes"; FP predicts "Yes" as a "No"; FN predicts "No" as a "Yes" and TN predicts "No" as a "No". A typical confusion matrix and its parameters is shown below:

Table 2: A Sample Confusion Matrix

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

Precision

The precision accounts for the total number of cases that are correctly classified as positive or TP divided by the total number of cases in that prediction (that is, the total number of entries in the row, both correctly classified (TP) and wrongly classified (FP) from the confusion matrix). This is expressed in the mathematical form below:

Precision = tp/ (tp + fp)

Recall

Recall is gotten by dividing the total number of predictions that are true by the total number of predictions (both true and false) for the class. Simply put, it is the true positive divided by the sum of entries in the column. The equation is given as follow:

Recall = tp/(tp + fn)

F1 score

The F1 score is another important parameter that helps us to evaluate the model performance. It considers the contribution of both precision and recall using the following equation:

F1 score = (2 * precision * recall) / (precision + recall)

Receiver Operating Characteristics (ROC)

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 ROC 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 successes 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 were assessed by running them with the testing data. Recall on the independent variables, 0 and 1 represents failure and success respectively. We performed various combination of techniques in a bid to get the model with the best performance. The table shows the performance of all algorithms (with and without SMOTE) used for the study as well as their quantitative performance metrics used in this research:

Table 3: A comparative table between the two models used:

Machine Learning Algorithms	Class	Precision	Recall	F1-score	10 fold CV Model Accuracy	Area under ROC
Logistics Regression	No	0.75	0.38	0.5	77.46%	0.73
	Yes	0.87	0.97	0.92		

without SMOTE						
Decision Tree Classifier Without SMOTE	No	0.45	0.62	0.53	68.46%	0.72
	Yes	0.85	0.82	0.84		
Logistics Regression with SMOTE	No	0.57	0.50	0.53	77.46%	0.81
	Yes	0.89	0.91	0.90		
Decision Tree Classifier with SMOTE	No	0.50	0.62	0.56	67.24%	0.74
	Yes	0.91	0.85	0.88		

Without the implementation of SMOTE algorithm, Logistics Regression (LR) performed better with a 10 fold cross validation (CV) model accuracy and AUC of 77.46% and 0.73 with Decision Tree (DT) 68.46% and 0.72. The introduction of SMOTE, values became 77.46%, 0.81 and 67.24%, 0.74 respectively. The application of SMOTE to the model generally improve the prediction capacity of the model. This is evident in the improvement of the precision, recall and F1-score values particularly in the decision tree classifier.

CONCLUSION

This research paper has shown the performance of engineering students in engineering drawing courses at introductory level as poor, thereby using machine-learning methods with SMOTE boosts the model's prediction. The data collected from students contained the variables that served as input for building the machine-learning model. The model, which was built on Kaggle in Python Languages using packages like Pandas, Scipy, Numpy, Sklearn etc. In preprocessing stage, the data imbalance was corrected with SMOTE which created synthetic data by generating new data points within the neighborhood by looking at the minority classes. A prediction accuracy ratio of 0.67 to 0.77 shows the model holds great promise when deployed to predict students performance in Engineering Drawing.

Acknowledgements

We cannot express enough thanks to Professor (Mrs.) O.M.O Etebu for her continued encouragement and leadership. Our completion of this project could not have been accomplished without the support of our families. The words of possibilities you spoken into our lives are producing great tidings.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

Altabrawee, H., Ali J. A., & Ajmi, S. Q (2019). Predicting Students' Performance Using Machine Learning Techniques. *ResearchGate*, 193-205. https://www.researchgate.net/publication/332893829

Badra, G., Algobaila, A., Almutairia, H., & Almutery, M. (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. *Elsevier*, 80-89. Doi:10.1016/j.procs.2016.04.012

Buenaño-Fernández, D., Gil, D., & Luján-Mor, S. (2019). Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *MPDI*. doi.org/10.3390/su11102833

Elreedy, D., & Atiya, A. F. (2019). *A Comprehensive Analysis of Synthetic Minority Oversampling TEchnique (SMOTE) for Handling Class Imbalance. Information Sciences.* doi:10.1016/j.ins.2019.07.070

Enughwure, Akpofure A. & Ogbise, Mercy (2020). Application of Machine Learning Methods to Predict Student Performance: A Systematic Literature Review. *International Research Journal of Engineering and Technology*, 3405-3415, https://www.researchgate.net/publication/341674171

Enughwure, Akpofure A. & Oluwafemi John D. (2020). Predicting student performance in engineering drawing using supervised learning methods. *International Journal of Maritime and Interdisciplinary Research, Nigeria Maritime University*. Available at https://www.researchgate.net/publication/343658000_PREDICTING_STUDENT_PERFORMANCE_IN_ENGINEERING_DRAWING_USING_SUPERVISED_LEARNING_METHODS as at August 27th 2020

Hamoud, A. K., Hashim, A.S., & Wid Aqeel Awadh (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *ResearchGate*, 25-31. https://www.researchgate.net/publication/323371097

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, ISSN 1726-9679, Vienna, Austria DOI: 10.2507/29th.daaam.proceedings.059

Oladokun, V.O., Adebanjo, A.T., & Charles-Owaba, O.E (2002). Predicting Students Academic Performance Using Artificial Neural Network: A Case Study of an Engineering Course. ResearchGate, 71-79. https://www.researchgate.net/publication/228526441

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1,** 206–215. https://doi.org/10.1038/s42256-019-0048-x

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

Wang, F., Wang, Q., Nie, F., Li, Z., Yu, W., & Ren, F. (2020). *A Linear Multivariate Binary Decision Tree Classifier Based on K-means Splitting. Pattern Recognition, 107521*. doi:10.1016/j.patcog.2020.107521