University of Ibadan
Journal of Science and Logics in ICT Research

Predicting Students' Academic Performance in Virtual Learning Environment Using Pearson Correlation Coefficient

[1]✉ Adelodun F.O., [2]Sakpere W., [3]Famurewa K. F. and [4]Oguns Y.J.
[1*,2,3] Department of Computer Science, Lead City University, Ibadan, Nigeria.
[1*,4]Computer Studies Department, The Polytechnic Ibadan, Nigeria

adelodunfelicia@gmail.com., sakpere.wilson@lcu.edu.ng, ogedengbekofo@ymail.com, oguns.yetunde@polyibadan.edu.ng

.

Abstract

Feature Selection involves selecting the most relevant features from a dataset during the prediction process. The selection method of features greatly influences how accurate, understandable, and effective predictive models are. Predicting students' academic success or struggle in a Virtual Learning Environment (VLE) is limited. Students who drop out of online courses are substantially more numerous than those who drop out of traditional courses [1,2]. The methodology followed in the study involved the use of two approaches: training and testing machine learning models with features selected from the dataset, and the second approach involved training and testing the machine models using all features in the dataset without feature selection. The Pearson Correlation Coefficient (PCC) feature selection method is used to select the features used for prediction. The two approaches were compared in terms of their impacts on the performance of the machine learning algorithms. The study was carried using nine classification models, which include Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest, Gradient Boosting, XGBoosting, LightGBM, MLP classifier (Neural Network) and Naïve Bayes. The result of the study showed that logistic Regression show highest accuracy mean of 0.7333 with feature selection and reduced accuracy mean of 0.7188 when all features were used in the prediction process. Without feature selection, the accuracy mean of Random Forest is 0.6813 and applying PCC feature selection to select the features for prediction, the accuracy mean of Random Forrest increased to 0.7333 revealing that feature selection method such as PCC is important for improving model performance.

Keywords**:** Feature Selection, Students' Academic Performance, VLE, Pearson Correlation Coefficient.

1.0 Introduction

Feature Selection (FS) is an important first step in creating a prediction model for students' academic achievement since it helps to discover the features that truly influence students' academic performance and boost prediction accuracy.

Massive Open Online Courses (MOOCs), Virtual Learning Environments (VLEs), and Learning Management Systems (LMSs) are just a few of the many types of online learning that are currently accessible [3, 4]. Students who drop out of online courses are substantially more numerous than those who drop out of traditional courses [1, 2]. This may be explained by the lack of direct communication between students and instructors as well as the absence of a classroom environment, which makes it challenging to forecast students' future performance in VLEs [5, 6]. Millions of features and enormous datasets are used in VLE learning, and it is computationally costly to extract the most pertinent features from these enormous datasets. Redundancy results from features in VLEs being highly correlated. Choosing the best features has been used extensively to forecast how well students would succeed on various online learning platforms [1].

Educational institutions can enhance their online services and offer effective learning materials by using predictive analytics to forecast students' academic success in a virtual learning environment [7]. This study offers a chance to look into the analytics for video

Adelodun F. O., Sakpere W. Famurewa K. F. and Oguns Y J. (2025). Predicting Students' Academic Performance in Virtual Learning Environment Using Pearson Correlation Coefficient. *University of Ibadan Journal of Science and Logics in ICT Research* (UIJSLICTR). Vol. 13, No.1, pp. 63 – 68.

learning offered by a VLE. Finding the most pertinent features that affect students' performance in the VLE can help predict their academic success. Applying the feature selection strategy to the dataset will help identify the most pertinent features that affect the academic achievement of the students. A feature is the data used as input for machine learning models to generate predictions [8]. Some of these features are redundant and irrelevant to the machine learning model. Irrelevant features are features that are unrelated to the intended outcome of the machine learning model. Redundant features are features that are duplicated [9]. The relevant features must be chosen to increase the accuracy of the machine learning algorithms used to predict students' academic performance.

The accuracy or performance of a machine learning algorithm does not only depend on the machine learning model alone but also on the feature selection method [10]. Feature selection is classified into four categories: filter, wrapper, embedded and information - theoretic methods [11]. Filter methods such as Pearson Correlation Coefficient (PCC) rank features according to their relevance using statistics such as correlation [12].

Virtual Learning Environment (VLE) is an interactive online space, where students and teachers can communicate. E-learning is often referred to as virtual, remote, and distance online learning. It is the process of describing how students can use electronic devices to access educational methods outside of the traditional classroom. Performance is the accomplishment of a task evaluated against predetermined benchmarks for speed, accuracy and completeness. Academic performance prediction is the process of predicting a student's future outcome such as grade using data, statistical models and algorithms.

2.0 Related works

A flexible predictive model, where raw data were used directly to construct a prediction framework for students' academic performance was investigated [13]. The proposed framework skipped the feature selection step and the framework was tested with Artificial Neural Network (ANN) and Random Forest (RF). Results of the experiment showed that the predictive models had an accuracy of 81%, a precision of 69% and a recall of 57%.

A predictive model for student performance in the classroom using student interaction with e-textbooks was conducted to predict students' performance [14]. Regression analysis was used to predict the final degree for the examination and classification algorithms were used to predict the performance of every student whether their performance will be good or bad, the classification algorithms were evaluated based on accuracy, precision, recall and F-measure. For the Regression analysis, their evaluation metrics were Mean Absolute error (MAE) Root Mean Square Error (RMSE) and Coefficient of determination (R^2).

The Pearson Correlation Coefficient (PCC), often referred to as 'r', is a statistical tool used to measure the direction and strength of a linear relationship between two continuous variables. The value of 'r' falls between -1 and +1. A value of -1 means a perfect negative linear correlation, while +1 indicates a perfect positive linear correlation. A value of 0 means there is no linear correlation between the two variables. The sign of the PCC value (+ or -) shows the direction of the relationship. A positive sign means that as the value of one variable increases, the value of the other variable also tends to increase. Conversely, a negative sign indicates that as one variable increases, the other tends to decrease.

It is important to note that the PCC specifically measures linear relationships, and may not be a good indicator of the strength of a relationship if the association between the variables is non-linear. Also, two strongly correlated variables are not automatically causative. . It is generally understood that correlation does not imply causation.

3.0 Methodology

The xAPL Edu dataset used for this research is an educational classification dataset designed to analyse and predict students' academic performance. It was retrieved from Alibaba Cloud Tranchi platform using the Kallboard 360 Learning Management System. The dataset includes student records from two semesters, encompassing various countries and gender.

The dataset consists of 480 instances with 17 features each, categorized into demographic, academic background, and behavioural features. The purpose of this study is to investigate the impact of feature selection techniques on the accuracy of the prediction of students' academic performance. This quantitative study employ, experimental and comparative designs.

For the experimentation, nine (9) classification models were used as the machine learning models which comprised of Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest, Gradient Boosting, XGBoosting, LightGBM, Neural Network (NN) and Naïve Bayes. The reasons for choosing Logistic Regression is because it is easy to implement and interpret. KNN is simple and suitable for small dataset. SVM is robust and effective for high dimensional dataset. Random Forest is robust and versatile. Gradient Boosting, XGBoosting and LightGBM helps in achieving high accuracy. NN is effective in capturing complex, non-linear relationship in dataset and Naïve Bayes is simple and computationally efficient.

The first step of the methodology was to pre-process the dataset. The following pre-processing steps were followed to pre-process the datasets: Exploratory Data Analysis to identify outliers and explore feature relationships. Data Cleaning and Missing Data: The missing values were manually found and removed. Outliers were detected and corrected using box plots and Z-scores, retaining only the values that appeared to contain meaningful information.

The second step of the methodology was to train the classification models without Feature Selection. This implies training and testing the machine learning models with the entire original feature set including irrelevant and redundant features. Ten-fold cross-validation was used to evaluate the performance of the models to ensure unbiased evaluation. This second step of methodology serves as the control experiment set.

The third step of the methodology was to apply Feature Selection (FS) techniques. The FS technique to be applied is the Pearson Correlation Coefficient (PCC) to select top correlated features. Then we trained the models with the selected features. This serves as our comparator against our control set. Again, the machine learning model's performance were evaluated by ten-fold cross-validation.

In this study, PCC is used as a feature selection method (filter feature selection method). It is used to measure the relationship between a feature and the target variable. When two features are highly correlated with each other, one may be removed to prevent redundancy. A feature may be considered very relevant if there is a high correlation between that feature and the target variable.

The PCC selector selected thirteen (13) features at the threshold of 0.1 and selected four (4) features at the threshold of 0.5. PCC selector at the threshold of 0.5 was used because it was proven to be effective [15]. The PCC selector at the threshold of 0,1 was chosen to serve as a comparator to PCC selector at threshold of 0.5. The PCC formula is given in Equation 1.

$$P_{X},y = \frac{Cov(X,Y)}{\sigma x \sigma y} \qquad (1)$$

Where,
Cov is covariance
σ_x is standard deviation of X
σy is standard deviation of Y

3.1 Model Evaluation

The study is a classification problem. The evaluation metrics were Accuracy, Precision, F-mean (F- score) and Recall based on the confusion matrix. Where,

TP = the proportion of positive cases that were correctly identified. For instance, predict as positive when actual positive.

TN = the proportion of negative cases that were classified correctly. Predict negative when actual negative.

FN = the proportion of positive cases that were incorrectly classified as negative. Predict negative when actual positive.

FP = the proportion of negative cases that were incorrectly classified as positive.

Predict positive when actual negative

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

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F-mean} = \frac{2(\Pr ecision * \operatorname{Re} call)}{\Pr ecision + \operatorname{Re} call} \quad (5)$$

3.2 Research Method

Research Design:	Quantitative Research
Area of Study:	Predicting Student Academic Performance
Data:	Student Educational Data
Programming tools:	Python
Model Evaluation Tools:	Confusion matrix

4.0 Results and Discussion

4.1 Results

Table 1: Result of Experiment 1 (Single Class Prediction without feature selection)

Model	Accuracy Mean	Accuracy Std	F1 Mean	F1 Mean Std	Recall Mean	Recall Std	Precision Mean	Precision Std
Logistic Regress-ion	0.7188	0.0598	0.7258	0.0621	0.7291	0.0597	0.7391	0.0674
k-NN	0.6000	0.0708	0.6020	0.0764	0.6082	0.0782	0.6266	0.0737
SVM	0.5917	0.0752	0.5855	0.0809	0.6032	0.0824	0.6397	0.0993
Random Forest	0.6813	0.0575	0.6798	0.0679	0.6831	0.0664	0.7142	0.0527
Gradient Boosting	0.6542	0.0711	0.6591	0.0700	0.6626	0.0663	0.6835	0.0803
XGBoost	0.6646	0.0628	0.6690	0.0617	0.6716	0.0617	0.6973	0.0679
LightGBM	0.6583	0.0660	0.6653	0.0667	0.6667	0.0703	0.6897	0.0679
Naive Bayes	+0.6896	0.0451	0.6996	0.0458	0.7175	0.0540	0.7202	0.0573
Neural Network	0.6688	0.0694	0.6714	0.0687	0.6763	0.0666	0.7080	0.0760

Table 2: Result of Experiment 2 (at Threshold 0.1)

Model	Accuracy Mean	Accuracy Std	F1 Mean	F1 Std	Recall Mean	Recall Std	Precision Mean	Precision Std
Logistic Regression	0.7333	0.0565	0.7411	0.0591	0.7427	0.0523	0.7529	0.0598
k-NN	0.6042	0.0801	0.6069	0.0879	0.6128	0.0895	0.6282	0.0852
SVM	0.5958	0.0769	0.5894	0.0837	0.6071	0.0845	0.6436	0.0996
Random Forest	0.7333	0.0657	0.7377	0.0694	0.7380	0.0713	0.7602	0.0638
Gradient Boosting	0.6792	0.0529	0.6866	0.0489	0.6876	0.0555	0.7013	0.0531
XGBoost	0.6688	0.0820	0.6740	0.0812	0.6766	0.0800	0.6947	0.0849
LightGBM	0.6667	0.0828	0.6733	0.0825	0.6735	0.0852	0.6942	0.0867
Naive Bayes	0.7188	0.0627	0.7303	0.0624	0.7446	0.0657	0.7453	0.0647
Neural Network	0.6896	0.0614	0.6913	0.0681	0.6963	0.0722	0.7149	0.0604

Table 3: Result of Experiment 3 (at Threshold 0.5)

Model	Accuracy Mean	Accuracy Std	F1 Mean	F1 Std	Recall Mean	Recall Std	Precision Mean	Precision Std
Logistic Regression	0.7146	0.0884	0.7212	0.0951	0.7273	0.0998	0.7426	0.0910
k-NN	0.5521	0.0819	0.5502	0.0970	0.5596	0.0985	0.5743	0.0908
SVM	0.6063	0.0920	0.6085	0.1012	0.6229	0.1067	0.6476	0.0925
Random Forest	0.6792	0.0946	0.6864	0.0942	0.6870	0.0991	0.7105	0.0961
Gradient Boosting	0.6604	0.0950	0.6690	0.0942	0.6754	0.0945	0.6920	0.0976
XGBoost	0.6354	0.0919	0.6478	0.0901	0.6476	0.0926	0.6646	0.0948
LightGBM	0.6333	0.0890	0.6417	0.0895	0.6427	0.0925	0.6711	0.0912
Naive Bayes	0.6896	0.0788	0.7023	0.0786	0.7224	0.0838	0.7212	0.0845
Neural Network	0.6563	0.0537	0.6606	0.0699	0.6793	0.0698	0.6841	0.0734

4.1 Results' Discussion

Discussion 1

The dataset obtained from public repository was pre-processed, cleaned and missing values were manually found and removed. Outliers were detected and corrected using box plots and Z-scores, three experiments were carried out. First experiment, experiment 1 used as a control set was carried with all features used to train and evaluate the performance of nine classification models. The second and third experiments used Pearson Correlation Coefficient (PCC) as a feature selection method to select features to train and evaluate machine learning models. The PCC selector at the threshold of 0.1 selected thirteen (13) features. The PCC selector at the threshold of 0.5 selected four (4) features.

Discussion 2

The highest-performing model in experiment 1 is the logistic Regression with an accuracy mean of 0.7188 and F1- Mean of 0.7258. This is followed by the Naïve Bayes classifier, which also shows robust performance with an Accuracy mean of 0.6896 and an F1 Mean of 0.6996. The Neural Network model also performs relatively well, indicating that it can capture complex relationships within the data, although it did not outperform Logistic Regression or Naïve Bayes. On the other hand, K-NN and SVM show comparatively lower Accuracy and F1-mean, suggesting that they might not be effective for this classification problem.

Discussion 3

Discussion on Experiment 2 selecting feature at the threshold of 0.1. The superior performance of the PCC threshold at 0.1 suggested that by retaining a more extensive set of features, even those with weaker correlations, the models were better equipped to capture complex, non-linear patterns within the data. The improved performance of models like Logistic Regression, Random Forest and Gradient Boosting with a broader feature set implies that these algorithms benefitted from the additional information provided by weakly correlated features.

Discussion 4

The decline in performance at the higher PCC threshold of 0.5, where fewer features were selected indicates that the dataset complexity cannot be adequately captured by a limited set of strongly correlated features. The standard deviations of Accuracy, F1-Mean, Recall and Precision metrics were also higher at the threshold of 0.5, highlighting that the models become less stable and more sensitive to data splits.

5.0 Conclusion

This study examined the influence of feature selection on the accuracy of machine learning models for predicting students' academic performance in virtual learning environment. Three experiments were carried out on an educational dataset retrieved from the Alibaba Cloud Tranchi platform using the Kallboard 360 Learning Management System. The dataset consists of 480 student records with 17 features each. The first experiment was carried out on the dataset without feature selection. The second and third experiments were carried out with feature selection utilizing the Pearson Correlation Coefficient (PCC) as the feature selection method at the threshold of 0.1 and 0.5 respectively. The experiments were carried out on Nine linear models. The results of the experiments showed that feature selection is important for the accurate prediction of student academic performance. Feature selection is dependent on the data is used, and machine learning algorithm employed.

When we compared the results of Experiment 1 (without feature selection) with Experiment 2 (using Pearson Correlation Coefficient (PCC)) feature selection method at a threshold of 0.1, we observed that PCC feature selection generally improved model performance compared to model performance without feature selection. For example, the accuracy mean for Logistic Regression increased from 0.7188 to 0.7333, and the F1 mean increased from 0.7258 to 0.7411. Similarly, Random Forest's accuracy mean rose from 0.6813 to 0.7333, and its F1 mean improved from 0.6798 to 0.7377. This trend suggests that retaining more features, even those with weak correlations enhances model performance by providing more comprehensive information for prediction.

Notably, Experiment 1 utilized all available features, whereas Experiment 2 with a PCC threshold of 0.1 slightly reduced the feature set. The better performance with a PCC threshold of 0.1 indicates that the dataset benefits from a reduced feature set, suggesting that some degree of multi-collinearity may have been present. Hence, the slight reduction in features helped eliminate redundancy without losing significant information, thereby enhancing the model's ability to generalize. For future studies increased data size is necessary for improved ability to generalize.

References

[1] P.S. Muljana & T.Luo, "Factors Contributing to student retention in online learning and recommended strategies for improvement: A Systematic literature review," J. Inf.Technol. Educ. Res., Vol.18 pp.19-57, 2019. https://doi.org/10.28945/4182.

[2] A.M.Mogus, I.Djurdjevie and N.Suvak, " The impact of student in a virtual learning environment on their final mark," Active Learning in Higher Education, Vol.13, no.3, pp.177-189,2012. https://doi.org/10.1177/1469787412.452985.

[3] A.Al-Azawei and M.A.Al-Masoudy, " Predicting learners' performance in virtual learning environment (VLE) based on demographic, behavioral and engagement antecedents," Int.J. Emerg. Technol. Learn., VOL 15.no.9 pp.60-75, 2020. https://doi.org/10/3991/ijet.v15j09.12691.

[4] M.Hussain, W.Zhu, W.Zhang and S.M.R. Abidi, " Student engagement and their impact on student course assessment scores," Hindawi, vol.6, pp.1-22,2018. https://doi.org/10.1155/2018/6347186.

[5] A.Al-Azawei, "Modelling e-learning adoption : The influence of learning sytle and universal learning theories," December, 2019.

[6] A. Al-Azawei, P.Parslow and K.Lundqvist, "The effect of universal design for learning(VDL) application on e-learning acceptance: A structural equation model," Int.Rev.Res. Open Distance Learn, Vol.18 no.6, pp. 54-87,2017. https://doi.org/10.19173.irrodiv/8i6.2880.

[7] M. A. A. Al-Masoudy & A. Al-Azawei "Proposing a Feature Selection Approach to Predict Learners' Performance in Virtual Learning Environments (VLE)" June 2023 International Journal of Emerging Technologies in Learning (IJET) Vol.18, no.11,2023 Doi:10.3991/ijet.v18i11.35405

[8] https://www.tecton.ai>Blog.

[9] M, Buykkececi and M. Cudiokar. (2022) "A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning." *GAZI University Journal of Science*,. DOI:10.35378/guys.

[10] P.Manikandaprablu. (2021) "Feature Selection Methods: A study" *Compliance Engineering Journal* Vol.12 issue 5.

[11] L.Gong, S.Xie, Y.Zhang, M,Wang & X.Wang "Hybrid feature selection method based on feature selection subset and factor analysis. IEEE ACCESS,10 (2022),pp. 120792-120803

[12] N.R.Beckham, I.J Akeh, G.N.P. Mitaart, & J.V "Moniaga "Determining Factors that affect Student performance using Various machine learning methods" Procedia Computer Science,216(2023) pp597-603

[13] A. Al-Zawqari; Dries Peumans and Gerd Vandersteen, (2022) "A flexible feature selection approach for predicting the students' academic performance in online courses". *Computers and Education: Artificial Intelligence* 3 100103.

[14] A. Abd Elrahman, Taysir Hassan A Soliman, Ahmed I. Taloba and Mohammed F. (2023) Farghally,"A predictive Model for Student Performance in classroom using student interactions with an eTextbook," *2022 10th International Japan-Africa conference, Research Square*: Inf.sci.Lett 12, no,1: 9 – 12.

[15] A.L. Hemdanou, M.L. Sefian, Y. Achtoun & I.Tahiri Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models. Computers and Education Artificial Intelligence Volume 7, December 2024, 100301