Predicting Academic Performance of University Students Using Machine Learning: A Case Study in the UK

Titilayo Olabisi Soyoye
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
Titilayo.Soyoye@hud.ac.uk

Tianhua Chen
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
T.Chen@hud.ac.uk

Richard Hill
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK
R.Hill@hud.ac.uk

Keith Mccabe
Planning and Business Intelligence
University of Huddersfield
Huddersfield, UK
K.Mccabe@hud.ac.uk

***Abstract*—Students' performance is one of the key success factors in educational institutions. Understanding the performance of students can help identify issues, enabling real-time action where and when necessary. Further, early identification and improvement of students' academic performance at all levels has been a major challenge in these institutions. Students may experience some difficulties which can impair their study and can negatively impact their academic performance. These issues can be efficiently addressed if students' data is pre-analysed, and students' performance predicted early to allow immediate decisions on support. Early prediction by educators and policy makers can assist in improving student and class performance. This work applied machine learning algorithms to analyse significant factors that influence students' academic performance, which could be used to inform decisions on support, or to identify and notify the students who require assistance; thus, taking effective steps to improving their performance. We used the academic records of computer-science students at the University of Huddersfield from 2017-2022, which provided several features that useful in predicting the students' performance. Evaluation of the results showed that decision tree and Ada Boost regression have higher accuracy score.**

Keywords—data mining, machine learning techniques, students' academic performance, systematic review

I. Introduction

One of the success factors in higher education is students' academic performance. Students' academic performance measure is derived from the effect of students' endeavour in tests, quiz, assignments, assessments, and examinations. Likewise, students' capability to explain and apply what has been learnt in their studies. Predicting students' academic performance provides an opportunity to identifying issues with students' performance; thus, assisting educators and institutions to pre-emptively support students that require assistance.

Machine learning techniques have proven to play critical role in students' performance prediction; although, increasing the scope of research on machine learning based solution in monitoring and managing students' performance should be reviewed by researchers [7]. Likewise, regardless of the machine learning technique applied, similar machine learning gives different accuracy for several researchers, current systematic review has shown that different attributes and varying datasets is a significant factor in determining the absolute accuracy and effectiveness of the machine learning application [6]. Ahmad et al [1] concluded that previous studies are still limited in synchronisation between the existing contribution of various fields. Alsariera et al [2] concluded that CGPA, attendance, quiz, assignment, gender, family, and personal characteristics have impact in the prediction of students' performance. The findings generated from the review assisted in identifying how machine learning have been applied to predicting students' academic performance, the most frequent techniques employed and features that influences students' performance. However, applying the appropriate techniques with key features in identifying and predicting students 'performance is crucial to finding accurate result. The motivations of this work are to analyse students' data, predict students' academic performance by applying machine learning techniques to this bespoke dataset; also, to investigate students' performances of a local university.

II. Literature Review

Machine learning techniques with features have been applied in predicting students' academic performance to identify valuable insights from datasets and for predicting students' performance. Also, accurately predicting students' performance can lead to optimal learning goals.

Namoun & Alshanqiti [9] performed a systematic review by performing prediction on student academic performance, where the outcomes of learning were examined. The report recommended predicting the outcomes of program-level and noted that the highest three predictors on the outcomes of students' learning includes, scores from assessment, academic emotions, and students' online learning patterns.

Moonsamy et al. [8] performed a systematic review for predicting students' performance in programming. It was indicated from their results that the collective assessment is the most used approach and effective approach for predicting the students' performance offering computer programming.

Balaji et al. [3] carried out systematic review on the contributions of machine learning models towards the prediction of student academic performance. The study concluded that machine learning techniques can predict students' performance based on some categorised and specific features.

Yahia et al. [10] conducted a systematic review to discover the contemporary machine learning methods and attributes

used in predicting the student's performance. Their findings revealed that artificial neural network with features: attendance, cumulative grade point average (CGPA), gender, assessment have an impact in the prediction.

Baashar et al. [4] conducted a systematic review on the artificial neural network techniques to predict students' performance. It was indicated that artificial neural network has great exactness on its outcome with regards to students' academic achievement; likewise, methods like the decision tree and the support vector machine. The most frequent features identified are the cumulative GPA, demographic, and gender.

Bin Roslan & Chen [5] study discovered that the most common classification methods are decision tree, classifier, random forest, and Bayesian network; with student records and demographics being the most common primary features. The findings from the studies revealed that decision tree, random forest, artificial neural network and naïve bayes are mostly applied with students' grades, demographics, and internal assessment the most key features in predicting students' academic performance. Summary of review conducted is shown in Fig. 1 and Table1.

Fig. 1 shows the hierarchical representation of models' frequency applied in reviewed papers; with random forest, decision tree and artificial neural network being the topmost models frequently applied.

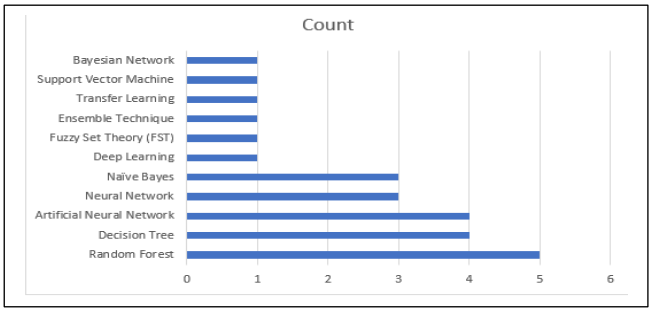

Fig. 1. Frequency of recommended models in review papers

Table 1highlights the most frequent and recommended features in the reviewed articles. The review showed that cumulative grade point average (CGPA) and internal assessment (test, quizzes, and assignment) are the most recommended attributes for predicting students' academic performance.

TABLE 1. CATEGORIES OF RECOMMENDED FEATURES IN REVIEW ARTICLES

S/N	Attributes	Authors	Count
1.	CGPA & internal assessment	(Shahiri et al., 2015), (Nawang et al., 2021), (Yahia et al., 2021), (Baashar et al., 2022)	4
2.	Periodic assignment submission	(Baashar et al., 2022)	1
3.	Family & personal preferences	(Sekeroglu et al., 2021)	1
4.	Demographic, academic & behavioural	(Balaji et al., 2021)	1
5.	Online learning patterns. Emotions, scores	(Namoun & Alshanqiti, 2020)	1

III. METHODOLOGY

Machine learning techniques were applied to a very recent real-world dataset, for the investigation of student performances. This helps identified valuable insights from datasets for predicting students' performance. The choice of machine learning techniques emanated from the review conducted, which revealed that the one of supervised learning algorithms with high accuracy for student performance prediction is the decision tree [2], [3].

Although, early identification and improvement of students' academic performance at all levels have been a major challenge in educational institutions [9]. This is due to the inadequate and reduced accessibility to student dataset and a dearth sophisticated machine learning method [9]; [11]. Hence, there is need for robust dataset and advanced machine learning techniques that would aid real time detection and prediction of student academic performance [3].

This guided the proposed method of applying machine learning techniques, classification algorithms and regression algorithms. Classification-Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Tree Classifier, KNeighbors Classifier; and Regression – Linear Regression, Lasso, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting, XGB Regressor, Cat Boosting Regressor and AdaBoost Regressor to identify students` academic performance.

The proposed machine learning techniques was applied to recognise and establish underlying relationships amongst the features in the academic dataset; this would help make inform decision especially for students that might need speedy support in their academic performance. Numerical data, categorical data, missing values, and null values found in the dataset were pre-processed. Standard scaler was applied on the numerical data to scale the features of the academic dataset, and one hot encoder was used to encode the categorical data.

The dataset was split into 80% training set and 20% testing set; to enhance the performance of the proposed classification and regression algorithms. Cross validation using grid search was initialised by fitting it with hyperparameters to get the optimal tuning parameters for the machine learning techniques. The efficacy of the proposed machine learning techniques was validated using classification metrics and regression metrics: precision, recall, f-score, root mean squared error, mean squared error, mean absolute error and r2 score to improve the predictive power of the machine learning techniques.

A. *Dataset*

An academic dataset from the University of Huddersfield students was used to identify significant features and predict students' academic performance. The analysis of student data for student support can be achieved by identifying students who require assistance in improving their performance. Also, the application of machine learning techniques to the student data allows identification of features that contribute to students' performance; hence, supporting the achievement of optimal learning goals. The dataset consists of 57,392 rows, 47 columns, 31 categorical and 10 numerical features. Null values and duplicated values were removed. The categorical dataset was encoded using One hot Encoding, with 2,161 rows of unique students and 41 columns (features). The datasets were split into 80% training dataset and 20% testing dataset. Numerical Features: clearing code, flag for foundation year,

expected course length, year spent on course, internal expected course length, year spent on course, internal reference on multiple records within a single year, disability code, average lateness to date, average hours attended to date, average activities attended to date, agreed mark.

Categorical Features: current year in program, student id, department name, course name, route name, month of entry, student status, last sandwich placement year, broad level of study, broad mode of domicile prior to entry, sex id, commuter, term time accommodation name, broad domicile, ethnicity, highest qualification on entry, parental education, socio economic name, last school name, last educational establishment name, foundation course, period code, occurrence code on a module, disability name, course year, module name, agree grade, overall result, broad qualification on entry attendance

B. Features Correlation in Dataset

Correlation among features in the dataset were analysed to gain more insights and identify features that influence students' performance. The result revealed that "Female" gender students that commute to university performed slightly better than "Male" gender and "Other" gender. But the male non-commuter and other male gender type performed slightly better than females and other type of gender that commute to university. It was also revealed that home students perform higher than overseas students. The dataset shows that female students perform higher that male and other students gender type. Also, for the disabled students in the dataset, it was revealed that students in the category: "blind or a serious visual impairment uncorrected by glasses" and "deaf or a serious hearing impairment" performed better than students' in other disability categories. From fig. 2; in the three plots, the first plot indicates that students' marks in the range of 60 to 80 occur more frequently while students' marks below 60 occur less frequently, the second plot shows that the students have a high attendance rate, and the third plot reveals that the students have little lateness to no lateness to classes. Fig. 3 shows that students' attendance and students' activities have higher correlation with agreed mark. Hence, students' attendance and students' activities have a positive influence on students' performance.

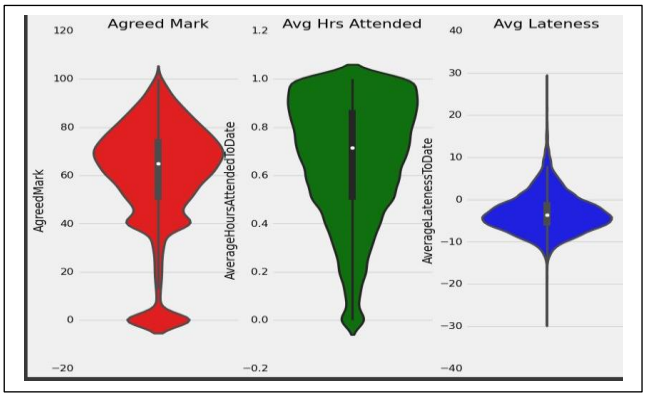

Fig. 2. Relationship between Agreed mark versus Hours attended for classes and lateness to classes.

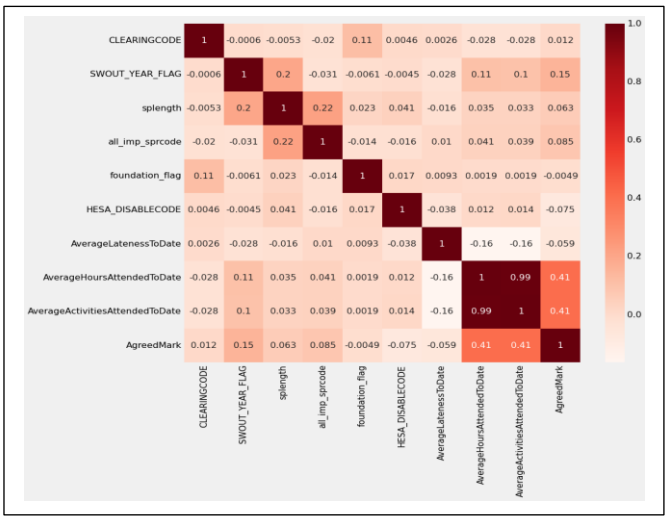

Fig. 3. Correlations among features

IV. Experimentation and Result

The proposed method was evaluated by applying sensitivity(recall), specificity(precision), F-score for the classification algorithms and applying Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE) and R2 score for the regression algorithms. The accuracy and performance of the proposed method was also assessed using a confusion matrix. Table 2 shows the resulting accuracy of each algorithm; with Decision tree having the highest accuracy score of 0.9846.

TABLE 2. CLASSIFICATION METRICS RESULT

S/N	Classification Algorithms	Accuracy Score	Precision Score	Recall Score	F1 Score
1.	Logistic Regression	0.9607	0.9599	0.9607	0.9630
2.	Random Forest Classifier	0.9148	0.9124	0.9148	0.9178
3.	Support Vector Machine	0.9777	0.9778	0.9777	0.9778
4.	KNeighbor Classifier	0.6195	0.5843	0.6105	0.6293
5.	Decision Tree Classifier	0.9846	0.9854	0.9846	0.9845

Fig. 4 shows the confusion matrix of the decision tree classifier, which compares the actual target values with the predicted values.

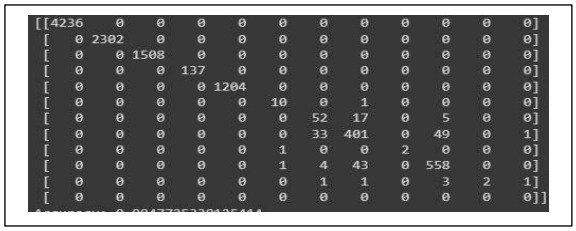

Fig. 4 Decision tree confusion matrix

Table 3a and Table 3b shows the scores obtained after training and testing. For the training stage, Random Forest Regressor have the highest training and testing R2 score: The R2 score during training is 0.9984 and the R2 score during testing is 0.9892.

TABLE 3A. REGRESSION MODEL PERFORMANCE TRAINING SET

S/N	Regression Algorithms	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	R2 Score
1.	Linear Regression	5.0525	25.5273	3.7469	0.9672
2.	Lassso	9.3993	88.3460	7.4818	0.8865
3.	Decision Tree	0.0000	0.0000	0.0000	1.0000
4.	Random Forest Regressor	1.1216	1.2580	0.5995	0.9984
5.	Gradient Boosting	5.4654	29.8702	3.9764	0.9616
6.	XGB Regressor	4..4231	19.5641	3.2531	0.9749
7.	CatBoosting Regressor	4.4883	20.1446	3.3447	0.9741
8.	AdaBoost Regressor	8.3062	68.9931	6.9608	0.9114

TABLE 3B. REGRESSION MODEL PERFORMANCE TEST SET

S/N	Regression Algorithms	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	R2 Score
1.	Linear Regression	5.3071	5.3071	3.9071	0.9637
2.	Lassso	9.3477	9.3477	7.4205	0.8873
3.	Decision Tree	3.5060	3.5060	1.0436	0.9841
4.	Random Forest Regressor	2.8980	2.8980	1.5667	0.9892
5.	Gradient Boosting	5.4298	5.4298	3.9103	0.9620
6.	XGB Regressor	4.6895	4.6895	3.3517	0.9716
7.	CatBoosting Regressor	4.7219	4.7219	3.4155	0.9712
8.	AdaBoost Regressor	8.2715	8.2715	6.9219	0.9118

A. Hyperparameter Tuning – Grid Serach Cross Validation

The model was tuned using grid search cross validation to give the most accurate predictions and improve prediction accuracy. This gives optimized values for hyperparameters and maximizes the model's predictive accuracy.

Table 4a shows that AdaBoost Regressor is the best model with hyper-parameters: {learning rate:0.1 and number of estimators: 128}.

TABLE 4A. MODEL PERFORMANCE TRAINING SET AFTER HYPER

Regression Algorithms	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	R2 Score
AdaBoost Regressor	8.1552	66.5075	6.8337	0.9146

Table 4b shows AdaBoost Regressor got an accuracy of 0.915082 in all the regression models; so, AdaBoost Regressor is chosen as the best model.

TABLE 4B. MODEL PERFORMANCE TESTING SET AFTER HYPER – PARSMETER TUNING

Regression Algorithms	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	R2 Score
AdaBoost Regressor	8.1145	8.1145	6.7905	0.9151

Fig. 5 shows the plot on the testing dataset and prediction, which show the plotted data and a linear regression model fit.

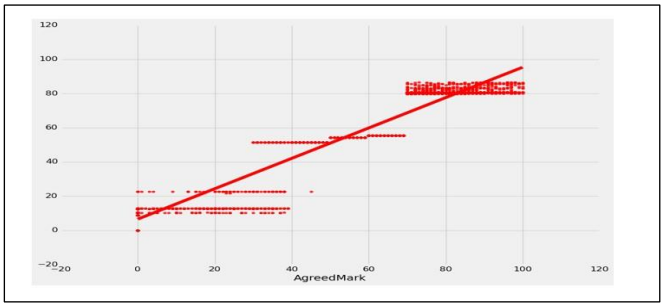

Fig. 5. Testing dataset and its prediction

V. CONCLUSION AND FUTURE WORK

The model incorporates several features with the machine learning algorithms to identify features that are likely to influence students' performance. The proposed model gives better performance on the prediction of students' overall performance. The highest accuracy score was obtained by AdaBoost after hyper tuning. Also, it has shown that lectures attendance and students' activities have a positive impact on the academic performance of students. The dataset will be extended in the future to improve on the accuracy of the result. Also, we will develop a more sophisticated model with significant features using a real-life dataset. Additionally, the proposed work would be implemented in the real-world to test its efficacy and hereby help in enhancing students' academic performance, improve employability rate of students' and enhancing the university reputation.

ACKNOWLEDGMENT

Appreciation goes to the University of Huddersfield Planning and Business Intelligence service for supporting this research with the accessibility to students' data.

REFERENCES

[1] Ahmad, S., El-Affendi, M. A., Anwar, M. S., and Iqbal, R., Potential Future Directions in Optimization of Students' Performance Prediction System. *Computational Intelligence and Neuroscience*, 1–26, May 2022.

[2] Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., and Ali, N., Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. *Computational Intelligence and Neuroscience*, 1–11, May 2022.

[3] Amelia, N., Abdullah, A. G., & Mulyadi, Y. (2019). Meta-analysis of Student Performance Assessment Using Fuzzy Logic. *Indonesian Journal of Science and Technology*, *4*(1), 74. https://doi.org/10.17509/ijost.v4i1.15804

[4] Balaji, P., Alelyani, S., Qahmash, A., & Mohana,M. (2021). Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review. *Applied Sciences*, *11*(21), 10007. https://doi.org/10.3390/app112110007

[5] Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., Alsariera, Y. A., Ali, A. Q., Hashim, W., & Tiong, S. K. (2022). Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs). *Applied Sciences*, *12*(3), 1289. https://doi.org/10.3390/app12031289

[6] Bin Roslan, M. H., and Chen, C. J., Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021). *International Journal of Emerging Technologies in Learning (IJET)*, Vol *17*(05), March 2022.

[7] Chakrapani, P., and D, C., Academic Performance Prediction Using Machine Learning: A Comprehensive & Systematic Review. *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, April 2022.

[8] Fahd, K., Venkatraman, S., Miah, S. J., and Ahmed, K., Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, Vol 27(3). 3743-775. ISSN: 1360-23572021, 2021.

[9] Kumar, S., & Janan, F. (2021). Prediction of Student's Performance Using Random Forest Classifier. http://www.ieomsociety.org/singapore2021/papers/1238.pdf

[10] Moonsamy, D., Naicker, N., T., T., & E., R. (2021). A Meta-analysis of Educa.tional Data Mining for Predicting Students Performance in Programming. *International Journal of Advanced Computer Science and Applications*, *12*(2). https://doi.org/10.14569/ijacsa.2021.0120213

[11] Namoun, A., & Alshanqiti, A. (2020). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences*, *11*(1), 237. https://doi.org/10.3390/app11010237

[12] Yahia, B., Gamal, A., Nor'ashikin, A., Hitham, A., & Hussein T, B. (2021). Predicting student's performance using machine learning methods: A systematic literature review. *Computer & Information Sciences (ICCOINS), 978-1-7281-7153-1*, 357–362. ¦ DOI: 10.11