National College of Ireland

Predictive Analytics for Enhancing Student Success in the UK: A Machine Learning Approach

MSc Research Project
MSc Data Analytics

Oluwadamilare Adetuberu
Student ID: x18165125

School of Computing
National College of Ireland

Supervisor: Dr Anu Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Oluwadamilare Adetuberu……

Student ID: x18165125……..……

Programme: MSc Data Analytics…… **Year:** 2023……

Module: Research Project……..……

Supervisor: Dr Anu Shani……..……

Submission Due Date: 31/1/24……..……

Project Title: Predictive Analytics for Enhancing Student Success in the UK: A Machine Learning Approach……..……

Word Count: 6445…… **Page Count**…21……..…

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

Date: ……31/1/24……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the on-line project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Analytics for Enhancing Student Success in the UK: A Machine Learning Approach

Oluwadamilare Adetuberu

x18165125

Abstract

Student success is an essential part of human capital development in society.
This study is motivated by the desire to effectively predict students' success using Machine Learning (ML) algorithms based on educational data, thereby contributing to the enhancement of student success overall.
A case study approach of an on-line learning environment in the UK is adopted.
By analysing the dataset from the Open University Learning Analytics (OULAD), the most effective ML model in forecasting outcomes, based on student's academic record, demographic information and student behaviours records were investigated. Through an experimental approach, techniques including Logistic regression, Decision trees, Random Forest, and Gradient Boosting Machine were employed and evaluated using metrics such as Accuracy score, Precision, Recall, F1-score and ROC AUC, Log loss and a five-fold cross validation. The model ROC AUC was 0.790, 0.831, 0.798, 0.808, respectively. This research contributes to the field of Predictive Analytics, Learning Analytics and Educational Data Mining.

Keywords - Educational Data Mining (EDM), Predictive Analytics, Student Success, Machine Learning, Learning Analytics (LA)

1 Introduction

(Rufai, et al., 2021) studying the role of Machine Learning and Data Mining techniques in predicting student's academic performance, indicate that the prediction of student's academic performance is central to effective education and understanding a student's performance is important in creating a student focused learning environment by educational institutions. While student success is a significant contribution to human capital development, the definition of the term student success could be a subjective perception, however (York, et al., 2019) included course grades as one of the six main components of student success in their article on defining and measuring academic success.

In the incorporation of Machine Learning (ML) into educational settings, many possibilities and several challenges have been presented, one interesting possibility is the prediction of student performance using educational data. (Hendradi, et al., 2023) noted that this has become increasingly relevant in the aftermath of the COVID19 pandemic which resulted in a shift towards on-line education leading to a substantial increase in educational data, which in turn holds a potential for transforming student's educational outcomes by enabling informed and effective educational strategies through Predictive Analytics.

(Saxena, 2020) examined the use of ML classification algorithms and feature selection to identify attributes used for predicting student success in a systematic literature review.
The primary focus of this research was to examine the capabilities of ML algorithms in predicting student success and to explore their implementation using real-world educational data and identifying the factors that influence the predictions.

The earlier the better, for students at risk of failing to be identified, and given timely intervention, thereby enhancing their overall learning experience and academic performance. Machine learning models such as Distributed Random Forest and Gradient Boosting Machine are ensemble techniques that perform well in classification that require precision and avoids overfitting are applied in this research. Simpler and equally effective models such as the Decision tree classifier and Naïve Bayes are also applied to compare with the ensemble models.

(Badr, et al., 2016) The application of data mining techniques to educational data like student academic record is referred to as Educational Data Mining (EDM).

Research Question.
1. How effectively can ML algorithms predict student success based on their academic records, learning behaviours and demographic information?
2. Which ML model is most accurate in predicting student success in an on- line learning environment?
These are the fundamental questions answered in the implementation of this project.

Research Objectives.
Analysis the four MLs models and comparing their results to determine the most effective model.
Applying a stratified data sample approach to address class imbalance and choosing the appropriate evaluation metrics for classification tasks.

In summary, by utilising ML to analyse the digital learning data available this research enhances student success within the UK context while also having applications across diverse educational landscapes.
The finding of this study will further contribute to the growing literature of ML applications in an on-line learning environment.

The rest of this study is structured as follows: **Section 2**. A review of existing related works applying ML in educational settings. **Section 3**. Research Method, these are the phases involved in answering the research question. **Section 4.** Design specification describes the life cycle of the study's processes. **Section 5**. Implementation features the models applied. **Section 6**. Evaluation: results and discussions. **Section 7.** The main observations and insights from the study and recommendations for future work.

2 Related Work

2.1 Predictive Analytics using educational data

Applying ML in Education has been an emerging field of interest for more than three decades.

However, its effective application continues to evolve (Namoun & Alshanqiti, 2021) concluded from a systematic review of ML, that results are not generalizable across educational data. (Zeineddine, et al., 2021) using an ensemble AutoML model on preregistration student data, achieved a dropout accuracy of 83% and 75.9% of result accuracy. (Okoye, et al., 2022) achieved a prediction of 100% accuracy using KNN based on student interaction data, however this data contains gender features, which presents ethical considerations.
(Marwaha & Singla, 2020) running a hybrid ML Ensemble (Naive Bayes, Rule Induction, and random forest) performs best on academic, demographic, and social-behavioral student data with accuracy of 86.67% and 0.39 RMS error.
The performance accuracy of RF model:0.988 and SVM:0.985 on student log data on MOOC was found to be the best by (Al-Shabandar, et al., 2017) among models evaluated. While these results are efficient, the data does not include the time feature of student interaction which is included in this study.

(Jiang, et al., 2014) in applying Logistic regression to week one behaviour data obtained an accuracy of 92.6% and 79.6 in the two populations predicted (1. Distinction and Normal, 2. Normal and None earners).
From (Xu, et al., 2017) models were observed to have results of accuracy ranging up to 90% recall, 90% precision 74%, some of the models employed an accuracy of 70 5.9% ensemble model, support vector machine 73.8%, k means clustering 48.5%.
Analysing student behaviour data, (Hung, et al., 2020) found Deep Learning model (DL) performed best with 86% accuracy.
(Saa, et al., 2020) , in analysing a university dataset containing 34 attributes and 231,782 records of withing four attribute categories (Demographic, Course, Instructor information, Student general information and Student previous performance) found mixed high results among the seven ML models applied with Radom Forest Accuracy of 75.52%, GLM recall 76.10%, and DL precision of 88.44% being the highest.
(Bujang, et al., 2023) by employing a multiclass prediction model for student grade obtain a highest F-score of 99.5% for the Random Forest model.

2.2
Predictive Analytics using the OULAD data

While analysing dropout and result prediction using the OULAD dataset, (Jha, et al., 2019) conducted four experiments using four predictor categories transformed for the dataset, which are demographic info, assessments scores, VLE interactions, and all attributes. Four ML models DRF, GBM, DL, and GLM were run on each predictor category and the results indicate that the GBM model ran on student's interaction with VLE attribute, produced the highest AUC evaluation using the GBM model 0.91 dropout prediction and 0.93 for result prediction. This approach gives a comprehensive analysis of the factors that contribute to student success predictions within this dataset using the most effective ML model, however the study used aggregate features and excluded time related metric which could have improved the student success factor analysis.
(Aljohani, et al., 2019) focused on the time related variables of the dataset by transforming a subset of the OULAD dataset into a sequential format using the student engagement with the virtual learning environment (VLE) on weekly basis, creating a time series. Long short-term memory (LSTM) neural network model was applied in the prediction. A precision of 93.46%, and 75.79% recall was obtained.
In a related study by (Jawthari & Stoffa, 2022), assessment scores attribute was to the weekly

engagement data to predict student at risk of failing. Naïve Bayes, Random Forest and Logistic Regression models were applied, The RF model produced the best F-score of 0.78.

While these studies obtained high precision scores, this research incorporates all the attributes of the dataset in its analysis.

3 Research Methodology

This research employs a Knowledge Discovery in Database (KDD) methodology to harmonize the process of data mining and discovery of knowledge, from data sourcing, exploratory analysis, data processing to ML modelling and testing.
The method involved Data sourcing, Data merging, Exploratory data analysis, Data preprocessing, Feature engineering, Model building and training, Results and Evaluation.

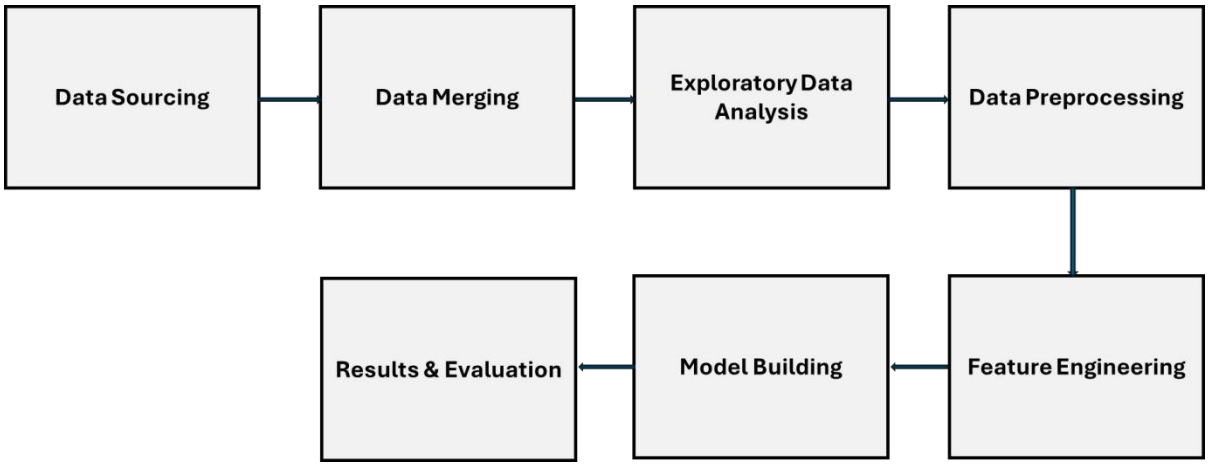

Fig1. Research methodology workflow

3.1 Data sourcing

This research utilizes the publicly available Open University Learning Analytics Dataset (OULAD), which contains a comprehensive academic, demographic, and detailed interaction data of students in on-line courses, released under CC-BY license (Kuzilek, et al., 2017).
The data was downloaded into Excel files stored securely on a computer hard drive.

3.1.1 Initial Exploration of the seven data files

The dataset contains records of 32,592 students, over 10 million rows for virtual learning environment (VLE) data, 22 different courses, and the data is available as 7 separate CSV files.
The overall structure of the dataset shows a high-level entity relationship, a lower-level entity relationship analysis and sample table structure in the various datafile are processed and visualized.

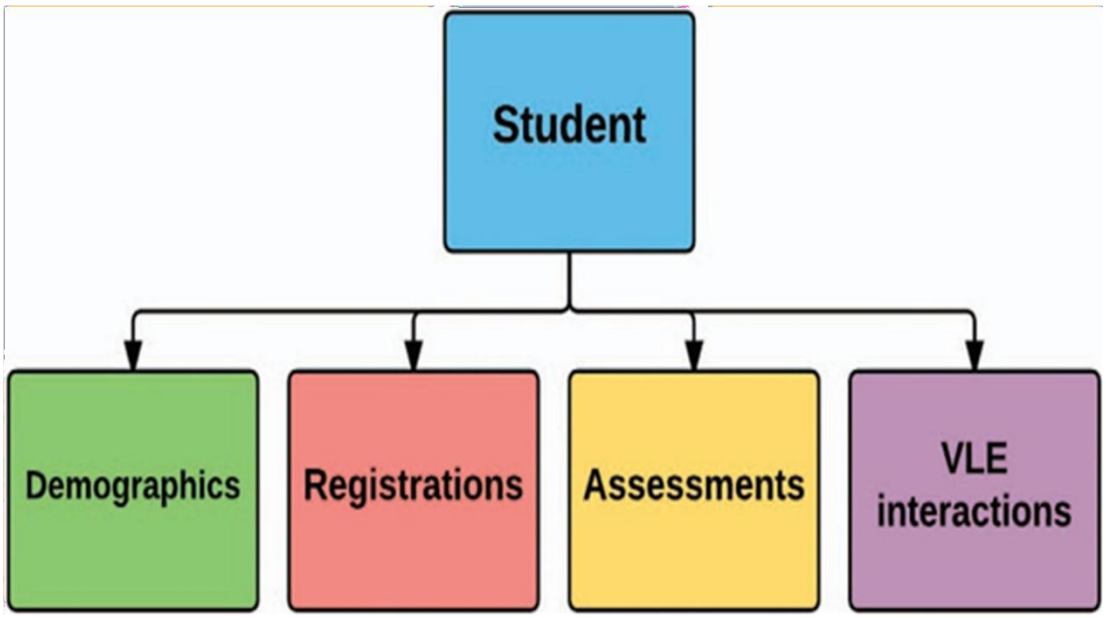

Fig 2. A high-level entity relationship diagram of the dataset

studentInfo.csv features:
1. code_module: The module student is enrolled in.
2. code_presentation: Course presentation code.
3. id_student: ID of the student.
4. gender: Student gender
5. region: Student region.
6. highest_education: The highest education attainment.
7. imd_band: Multiple Deprivation index of the student area lives.
8. age_band: The students age range.
9. num_of_prev_attempts: The number of attempt students have previously made on the course.
10. studied_credits: The number of study credits the student has obtained.
11. disability: Student has a disability status.
12. final_result: The final result achieved by students.

studentRegistration.csv features:
1. code_module: The module student is enrolled in.
2. code_presentation: Course presentation code
3. id_student: The unique ID of the student.
4. date_registration: Date the student registered for the course in relation to the start of the module.
5. date_unregistration: Date the student left the course.

courses.csv features:
1. code_module: The module student is enrolled in.
2. code_presentation: Course presentation code.
3. module_presentation_length: The duration of the module.

assessments.csv features:
1. code_module: The module student is enrolled in.
2. code_presentation: Course presentation code.
3. id_assessment: ID of the assessment.

4. assessment_type: The type of the assessment
5. date: date of the assessment.
6. weight: The weight of the assessment.

studentAssessment.csv features:
1. id_assessment: ID of the assessment.
2. id_student: ID of the student.
3. date_submitted: Date the student submitted the assessment.
4. is_banked: Whether the assessment result has been banked.
5. score: The scores of each student.

studentVle.csv
1 code_module: The module student is enrolled in.
2 code_presentation: Course presentation code.
3 id_student: ID of the student.
4 id_site: ID of the Virtual Learning Environment (VLE) site.
5 date: Date student interacted with the VLE.
6 sum_click : number of times students clicked on the VLE site on a particular date.

vle.csv features:
1. id_site: ID (VLE) site.
2. code_module: The module student is enrolled in.
3. code_presentation: Course presentation code.
4. activity_type: Type of activities on the VLE site.
5. week_from: The week activity started.
6. week_to: The week activity ended.

The columns across the seven dataset have academic record and demographic features in Assessment, Registration, Virtual Learning Engagement that support the modelling the prediction of student performance and drop-out risk as well as identifying the key factors that contribute to the student success. A comprehensive merging of the data sets aided the answering of the research question and achieving the research objectives.

3.2 Data Merging

Merging was done using primary and secondary key entity relationship between Student Info and Student Registration, Courses, Assessments, Student Assessment, Student Vle and Vle. This comprehensive merging provides all the available features in one dataset.

The new dataset has these columns:
1. code_module, 2. code_presentation, 3. id_student, 4. gender, 5. region, 6. highest_education, 7. imd_band, 8. age_band, 9. num_of_prev_attempts, 10. studied_credits, 11. disability, 12. final_result, 13. date_registration, 14. date_unregistration, 15. module_presentation_length, 16. id_assessment, 17 date_submitted, 18 is_banked, 19 score, 20 assessment_type, 21 **date_x : The date of the assessment, given as the number of days relative to the start of the module-presentation**, 22 weight, 23 id_site, 24 date_y The date of student interaction with the VLE, given as the number of days relative to the start of the module-presentation,
25. sum_click, 26. activity_type, 27. week_from, 28. week_to.
The time feature was incorporated in this research to achieve the future work objective from

(Jha, et al., 2019).

3.3 Exploratory data analysis of the merged dataset and visualization.

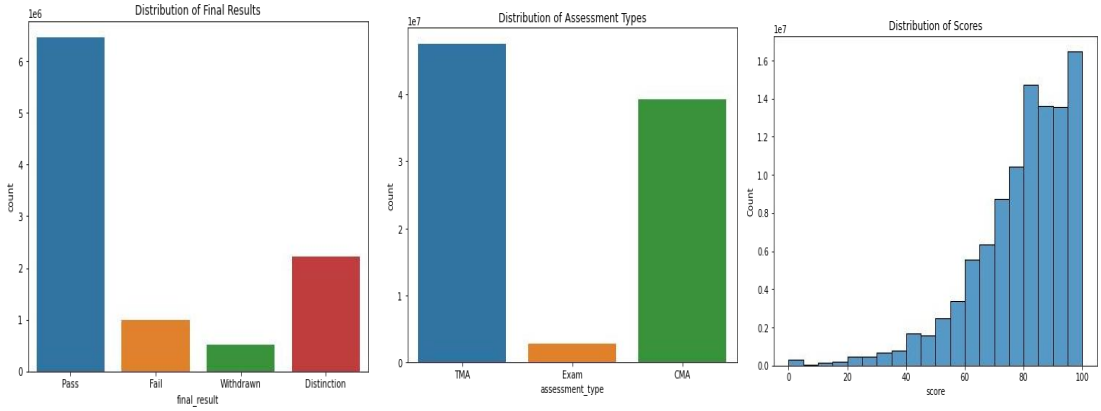

Fig 3. Distributions of Final results, Assessment Types and Scores

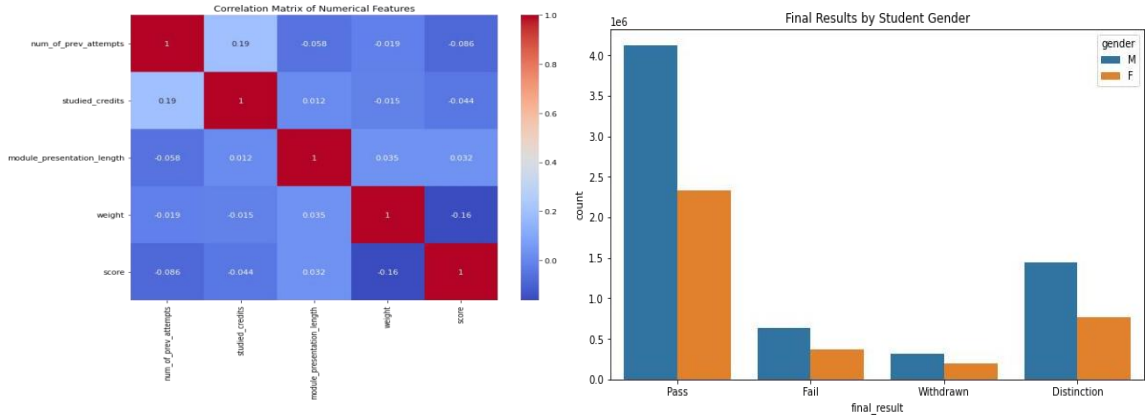

Fig4. A Correlation Matrix of Numerical Features and Distribution of Final Results by Gender

Final Results by Gender: The distribution of final results is similar for both genders, with a slightly higher number of males failing or withdrawing.

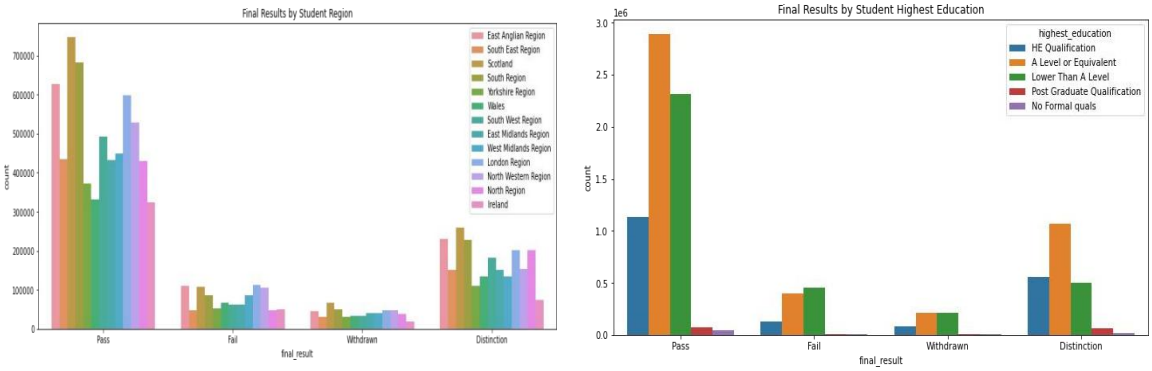

Fig 5. Distributions of Final Results by Region and Final Results by Highest Education

Final Results by Region: The distribution of final results varies by region. Some regions have a higher number of passes, while others have a higher number of failures or withdrawals. This could indicate that the region has an impact on student success, possible due to differences in educational resources or opportunities.

Final Results by Highest Education: Students with a Post Graduate Qualification or HE Qualification have a higher number of passes compared to other groups. Those with lower than A Level have a higher number of failures and withdrawals.

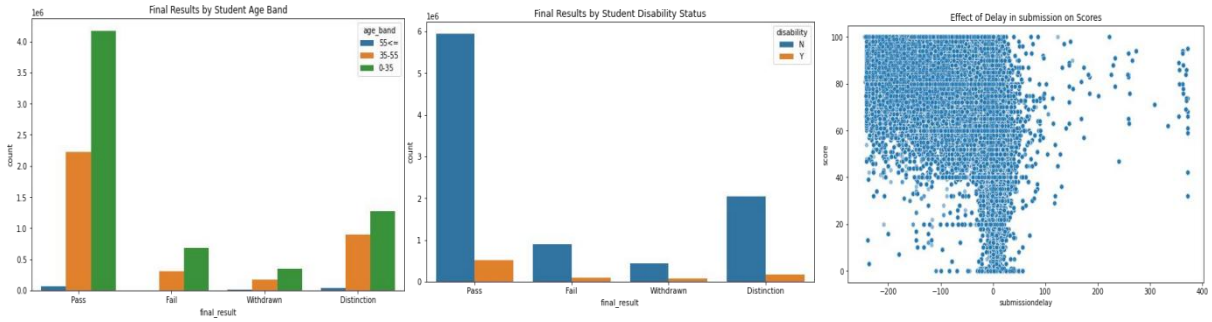

Fig6. Distributions of Final Result by Age Band, Final Results by Disability Status, and a Scatter plot of the Impact of Submission Delay on Scores

Final Results by Age Band: Older students (55<=) have a higher number of passes compared to other age groups. Younger students (0-35) have a higher number of withdrawals.
Final Results by Disability Status: Students without a disability have a higher proportion of passes, while those with a disability have a higher proportion of withdrawals.
Impact of Submission Delay on Score: The scatter plot shows that majority passed and submitted assignments before or around the deadline.

3.4 Data Preprocessing

The original dataset was too large for available computational resources, therefore a stratified sample of the largest StudentVLE which contained more than ten million records of students' interactions with the virtual learning environment was merged with the other six datafile to create a comprehensive merged dataset. The merged dataset contains about one million records.
The missing values were handled, numerical data was replaced by zero for score, assuming that student without scores dropped out from the course. Missing categorical variables where replaced the mode value of the distribution or dropped. Categorical variables were hot encoded into numerical variables for feature engineering. Standard scaling was used to standardize the numerical variables. Class imbalance was addressed using Hyperparameter tuning and stratification of train and test data split to mitigate model bias toward higher outcome features in the dataset.

3.5 Feature Engineering

A multi-class binary feature was engineered for Feature Selection, all the available features were used. The target variable for the binary classification is to predict if the student will pass or fail. The final_result column is converted into a multi-class binary variable as follows; Pass and Distinction = 1 (successful), Fail and Withdrawn = 0 (not successful).

3.6 Model Building

Four machine learning models are applied for predictions, logistic regression, distributed random forest (DRF), deep learning (DL) and gradient boosting machine (GBM). The performance of each model was evaluated using accuracy, f1-score, recall, precision, AUC-ROC as seen in previous works.

3.6.1 Logistic Regression (LR)

Logistic Regression classifier is a machine learning model for binary classification applications, using linear features as inputs, it applies a logistic function to predict that a data point belongs to a particular binary category, this is called the sigmoid function, and the outcome is either 0 or 1. This represents the probability of the binary class that a data point belongs to. It is a simple but effective model popularly used for classification tasks.
(Jiang, et al., 2014) applied Logistic regression in their analysis of week one behaviour data and achieved significant results with 92.6% and 79.6 accuracy in the two populations they analysed (1. Distinction and Normal, 2. Normal and None earners) respectively. The evaluation was done using Recall, Precision, F1 score, Accuracy and AUC-ROC.

3.6.2 Distributed Random Forest (DRF)

DRF classifier is an ensemble machine learning classification and regression model, it generates multiple decision trees in the model training process, which are known as forests, by random selection of data making sure there is adequate representation of the data points in the model. This process ensures overfitting is limited in the model result, the classification is achieved through a majority vote and regression achieved through calculating the average of the trees. This approach gives the model a better performance probability than Decision trees.

3.6.3 Deep Learning (DL)

DL is a robust model for training artificial neural networks, it is able to handle large and complex datasets by transforming the input into new features through pattern recognition. It can learn from unstructured data to improve accuracy. Tuning of the hyper-parameter further enhances the accuracy of the model.

3.6.4 Gradient Boosting Machine (GBM)

GBM is also an ensemble classification and regression model. It differs in the ability to build new series of decision trees to improve on the errors of the previous trees, using the residual errors. The process of continuous reduction of errors leads the model to produce higher accuracy in its predictions. However, it is prone to overfitting due to the process of error reduction, hyper-parameter tuning helps to mitigate the overfitting tendency.

3.7 Model Evaluation

The effectiveness of the various machine learning models is compared by the experimental approach and the most suitable models for predicting student's success is determined to be evaluating the performance of the models on the test set using appropriate metrics such as accuracy, precision, recall, and the F1 score.

4 Design Specification

The research design is a qualitative design which enables the use of ML models to analyse numerical data, using measurable evaluation metrics. The process flow diagram is designed to enable detailed exploration of the seven data files in the OULAD dataset.

The design also enables all relevant data files to be merged after creating entity relationship diagrams that could predict student success based on the available data.

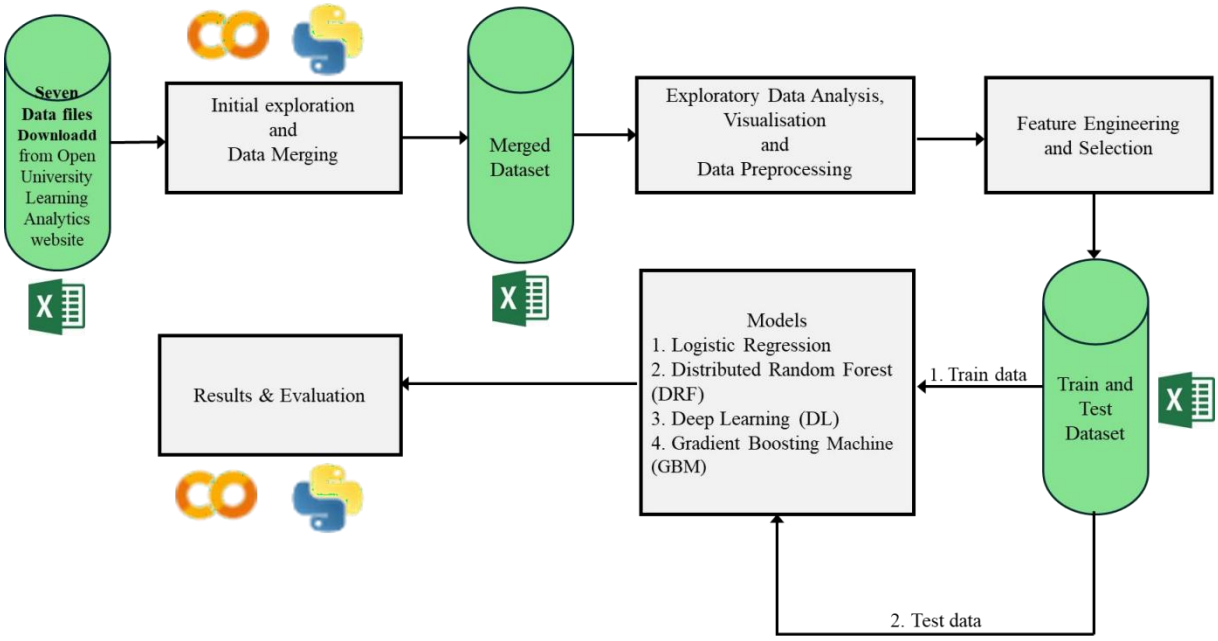

Fig 7. Research Design Specification

This research was undertaken as outlined in the Research Design Specification, Fig 7. The dataset was sourced from the Open University Learning Analytic website and downloaded into excel files in csv format. It contained seven data files.

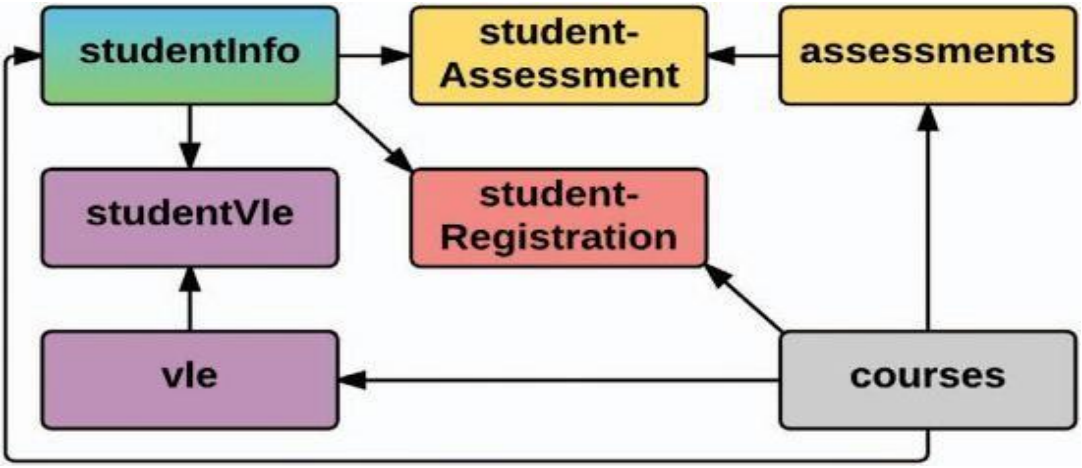

Fig 8. A lower-level entity relationship diagram of the dataset (Kuzilek, et al., 2017).

Using the Python programming language due its extensive data analytic packages and libraries such as sklearn, seaborn, and matplotlib which enabled the necessary data processing in implementation of the project. The dataset[1] was imported into the Python environment for analysis. The Python codes were run in Google Colab[2], an on-line development environment with built in code debugging link to the debug platform, stack-overflow[3], this augmented the local processing capacity of the windows system used. Initial data exploration was done to identify the variables present in all the seven datasets and to understand their relationship with each other through primary and

secondary keys, this was visualised to make the complex relationship easier to see.
This also allowed the features needed for the multi-class binary classification model to be identified.
The dataset set was then merged into a large Merged dataset for a comprehensive analysis. This was done by creating a stratified sample of the largest StudentVle dataset, to merge with all the other six datasets.
Exploration of the merged dataset shows the relationship between the target variable final result and score and other features of the dataset.
The data was preprocessed, filling missing values in the score variable with zero, assuming students without scores, withdrew from their course. The numerical features were standardised using standard scaling. Categorical features were encoded to numerical features and missing data was filled with the mode or dropped.
Class imbalance, particularly in the target variable, which is due to low fail outcomes compared to high pass outcomes in the dataset was addressed by Hyperparameter tuning and stratifying the test dataset.

Feature Engineering was done using a multi-class binary classification method, the four categories in the final result column were coded as Pass = 1, Distinction = 1, Fail = 0 and Withdraw = 0.
The dataset was split into 80% train and 20% test data.
The four machine learning models were trained using 80% of the merged dataset and their performances were evaluated using accuracy, f1-score, recall, precision, AUC-ROC as seen in previous works.

Ethical considerations

While the field of Data Analytic is discovering knowledge from data, it is important that is do ethically, as experience from other domains indicate that ethical intentions are not sufficient (Holmes & Tuomi, 2022) FATE ethical framework issues relating to fairness, accuracy, transparency, and ethical considerations must be adhered to in the interest of all stakeholders. The research problem and question are of a general nature, and it was ensured that there was not identifiable impact to a particular student in the research. Model bias was addressed by using a stratified training and testing data split, Hyperparameter tuning was also used to limit class imbalance.

Dataset consideration: the Oulad dataset is an anonymised, publicly available dataset under the creative commons license. The dataset creators stated their consent for its public use of it on their website. (Kuzilek, et al., 2017) . This research has no Ethical implications for all stakeholders.

[1] https://analyse.kmi.open.ac.uk/open_dataset.
[2] https://colab.google/.
[3] https://stackoverflow.com/.

5 Implementation

As discussed in section 3.1, the Oulad dataset that was used in the carrying out this study was downloaded from the Open University Learning Analytics website, into Excel files, the seven data file included in the dataset were merged after creating a stratified sample of the StudentVle file which contained over 10 million records, the final merged file contains 28 columns and 1048575 records. This was done using the Python programming language in the Jupyter notebook development environment.

5.1 Results of the initial exploratory data analysis

The pass or fail categories gave highest outcomes in the final result distribution, indicating that most students either passed or failed their courses, according to the metrics used.
The other lower outcome categories of withdraw and distinction showing comparable levels, were considered as either fail or pass for the purpose of binary classification, the wide difference between the variable outcomes created a class imbalance for the feature variable.

Most of the assessments are tutor marked (TMA) followed by computer marked assessments (CMA) and Exam assessment.
The score variable is negatively skewed indicating that most of the students got relatively higher marks to the minority who got low marks. This also created an imbalance in the dataset when analysing the scores as a measure of student performance.

Initial data exploration showing the distribution of the target variable and its relationship with various features was conducted and the visualisation displayed outputs as follows:

1. A significant majority of students passed in the final result distribution.
2. Most assessments were teacher marked, followed closely by computer marked assessments, exams were negligibly small, in the assessment types plot.
3. A vast majority of the students scored above 40 marks in the score distribution plot.
4. The correlation matrix of the numerical features shows no statistically significant correlation between any of the variables the highest positive correlation was between, number of previous attempts and studied credits = 0.19 and the highest negative correlation was between score and weight of the course = -0.16.
5. In the final result by gender visualisation, female distribution of result was about 65% of the male in the pass, fail and withdrawn categories, while it was about 60% for the distinction category.
6. The final result by student region plot indicated relatively similar trends in the pass, fail and distinction categories per region except the withdrawn category which relatively even despite the geographic region.
7. The final result by student highest education plot revealed that students with an A level or equivalent passed the most.
8. The 0-35 age band was the most in all the pass, fail, withdrawn and distinction categories of the column.
9. Students with disability had a relatively low representation in all the final result variables.
10. The time of submission indicated that most of the students that passed submitted their assignments well before the course deadline, while majority of those that failed submitted around the deadline date.

Feature engineering was done to create multi-class binary target variable with the following parameters: pass and distinction categories of the final result column = 1 and fail and withdraw categories = 0.
Due to the left skewed distribution of the column, the significant class imbalance was addressed by stratification of the minority class in the training and testing data split, class weight was also balanced for the random forest classifier. The use of SMOTE was avoided due to very high imbalance, which could introduce noise into the model training and testing.

The preprocessed dataset was imported into Google Colab using the import file function, through the pandas library, sklearn package libraries of matplotlib and seaborn for plotting were used for visualisations.
LabelEncoder, StandardScaler and SimpleImputer were used to encode the categorical variables, standardise the variable to scale, and input missing values, respectively.
Sklearn train_test_split was used to split the data into training and test sets.
Model libraries of Logistic Regression, DecisionTreeClassifier, RandomForestClassifier and GradientBoostingClassifier were used to initialise the model. Hyperparameter tuning was conducted to prevent overfitting of the models to the dataset.
The test results were evaluated using the accuracy_score, classification_report, roc_auc_score libraries.
The sklearn package was also used to import the cross_val_score library for a fivefold cross validation of the model result.

6 Evaluation

This research aimed to conduct comparative study of machine learning models to determine the most effective in predicting student success in an on-line learning environment. The outcomes of four models were trained and tested on an anonymized real-world dataset. The performance of each model was measured by the accuracy score, precision, F1-Score, Recall, ROC AUC score, Log loss and cross validated using a five-fold cross validation. A confusion matrix was created to help measure the model predictions efficiency.
These metrics provided insight into the strengths and weaknesses of each model; this could give indication of future work.

	Logistic Regression	Decision Trees	Random Forest	Gradient Boosting Classifier
Accuracy	0.863	0.901	0.765	0.901
Five-fold Cross validation average score	0.863	0.901	0.770	0.901
Precision	0.873	0.895	0.930	0.895
F1- Score	0.924	0.944	0.850	0.944
Recall	0.981	1.0	0.782	1.0

ROC AUC Score	0.790	0.831	0.798	0.808
Log Loss	0.345	0.280	0.589	0.312

Fig 9. Model Performance Result Table

6.1 Experiment 1

Logistic Regression Accuracy: 0.863, indicates the model performance is 86.3% accurate in predicting the outcome in 86.3% of cases. This shows that the model fairly accurately identified patterns overall. The precision score of 87.31% shows the model is slightly better than its overall performance in predicting positive cases specifically. Recall of 98.31% confirms that the model gets over 98% of its positive cases right, this is a very good performance. The F1 Score of 92.4% indicates there is a balanced distribution of positive and negative cases in its prediction.

The ROC AUC score of 79% shows that the model is good at distinguishing between the two classes.

A low Log Loss of 34.5% indicates relative accuracy and confidence in the model's prediction.

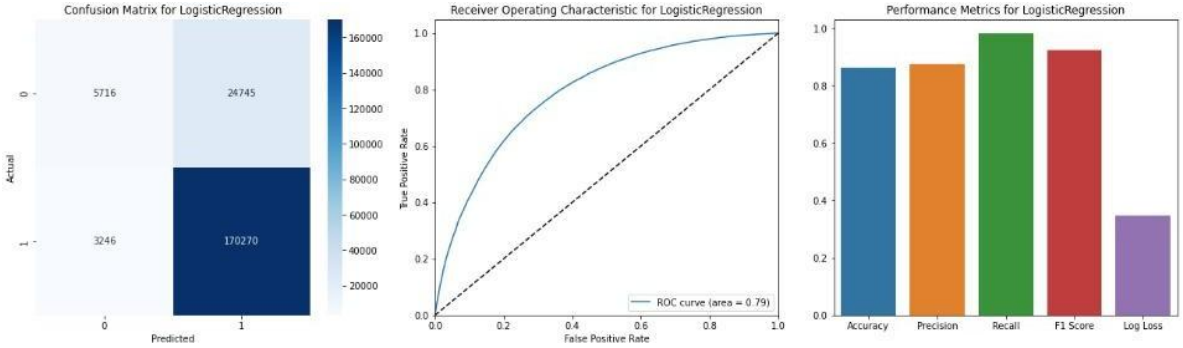

Fig 10. Logistic regression classifier evaluation visualization.

6.2 Experiment 2

DecisionTreeClassifier (DTC) Accuracy: 0.901, indicates the model performance is 90.1% accurate, this is an improvement on the Logistic model performance, and it could be due to DTCs ability to capture more complex and non-linear data points, however a more robust ensemble tree method could provide better performance than a single tree. Precision score of 89.57% also exceeds the Logistic regression model precision score, the model is better at predicting the positive class. The Recall score of 100% indicates a perfect identification of all positive cases however, this could be due to overfitting, which the model could be prone to. The high F1 Score of 94.5% shows the model a good balance between positive and negative cases but this could be linked to the Recall score. ROC AUC score exceeds the logistic regression score showing this model it better at knowing the difference between the classes. The Log Loss of 0.28 also exceeds logistic regression indicating the model's ability to make

prediction based on this study is better than the logistic regression model.

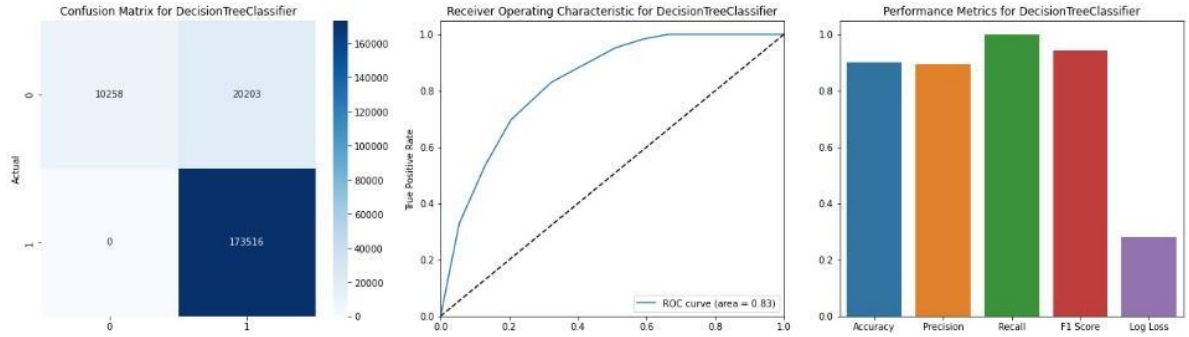

Fig 11. Decision Tree Classifier evaluation visualization.

6.3 Experiment 3

RandomForestClassifier Accuracy 76.52% this is lower than the two earlier models, for this study this model less accurate at make predictions overall, however, it scored a higher Precision score of 93% showing that it predicts fewer false positives compared to the two previous models. The lower Recall score than the two previous models indicates that it missed some positive cases compared to the 100% and 98% of Decision tress and Logistic regression, respectively. The lower F1-Score is linked to the lower recall score, the model has a comparable ROC AUC score of 79.87% to 79% and 83% of the previous models, its ability to separate between the positive and the negative classes is similar to theirs.
The confidence in the predictions of this model is much lower than the earlier two models at 0.589.

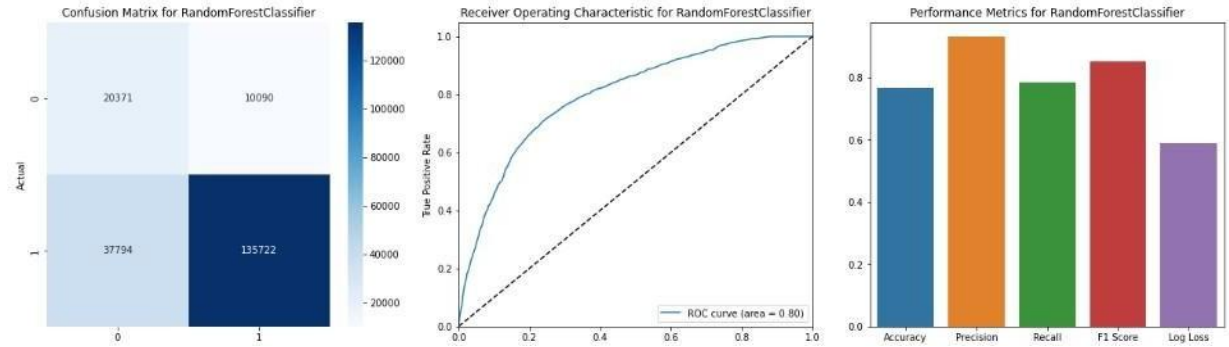

Fig 12. Random Forest Classifier evaluation visualization.

6.4 Experiment 4

GradientBoostingClassifier Accuracy 0.901, Precision 0.895, Recall 100%, F1 Score 0.944 these metrics are very close the those of the Decision tree classifier, the ROC AUC Score of 80.9% is second place after the Decision tree. The Log Loss of 0.31 is also second after the Decision tree.

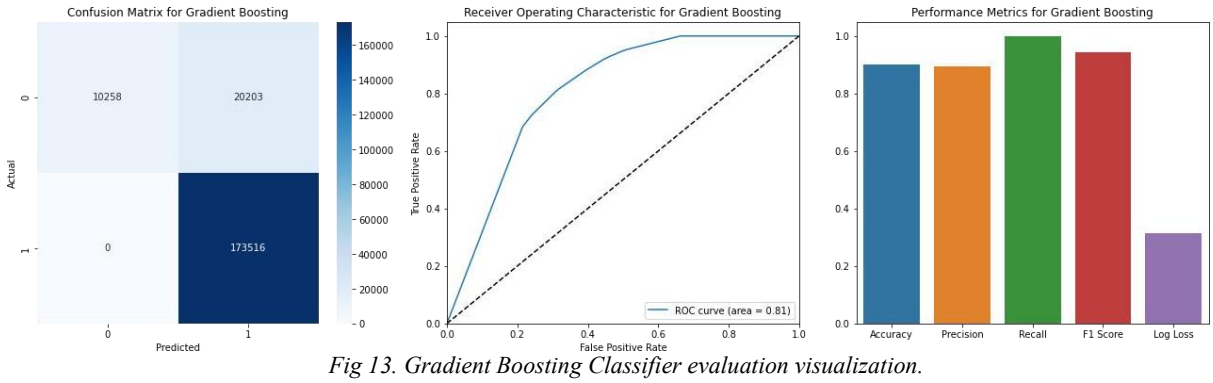

Fig 13. Gradient Boosting Classifier evaluation visualization.

6.5 Discussion

Of the four models studied in this research, the comparative performance is tabled in the model performance result table.
The Decision Tree Classifier and the Gradient Boosting models show very similar scores across F1 Score and Recall (which is highest value possible), the closeness of the values in two different models will require further investigation by Hyperparameter tuning or data preparation.
Random forest model exceeds all the other three models in its precision but is lower in other measurements. The Logistic regression model while presents a more balanced outcome on most of the performance metrics, the accuracy of Decision Tree Classifier and the Gradient Boosting Classifier are far superior.
The decision to limit the minority class generation by not using imbalance tools like SMOTE might have had an impact on the model prediction, a feature importance examination could give more insights into what features contribute the most to the model performance, less important feature could be removed to improve accuracy. Computational capacity and time constraints also limit the model iterations to convergence and well as further experimentation.

with Hyperparameter tuning. The creation of the Mergedata2 dataset, which was used for model training and testing, is a further stratification of the initially merged dataset, done due to the available RAM capacity on the local computer and the free version of Google Colab.

The results of the study indicate that the Decision Tree Classifier and the Gradient Boosting Classifiers are more applicable to datasets where accurate measurement of positive cases is essential as the OULAD dataset, where vast majority of the students either passed or got a distinction. The ROC AUC Score is comparable at .0808 to that achieved by (Jha, et al., 2019) at 0.93 from result prediction. While this study performed below the benchmark study by -0.05, the time related attributes included in this study provides a more comprehensive analysis of the dataset.

7 Conclusion and Future Work

This research project's aim was to study the use of predictive Machine Learning models in

predicting student success in an on-line learning environment.
Four machine learning models were applied to the Open University Learning Analytics Dataset which is a comprehensive dataset of student academic, demographic records, and records of student interactions with the virtual learning environment.
Logistic regression, Decision Trees, Random Forests and Gradient Boosting ML models were applied to the dataset yielding varying degrees of classification performance for each model.
Data processing was demanding as one of the seven data files contained over ten million records. This challenge was overcome, and a new comprehensive merged dataset was created that could assist future work requiring such dataset.
A lot of attention was paid to preventing overfitting the data due to high class imbalance as the vast majority of the students passed.
This project shows that data gathered across various aspects of an online educational entity can be unified to mine strategic information that assists decision making.
Further Hypertuning of parameters is needed to improve the model performance, deep learning model could also provide better performance.
The project expresses the value of comprehensive data analytic projects.

Acknowledgement:
I thank my daughters Misi and Ileri for their encouragement and errands they ran during my studies, and my sons, Ike, and Iyanu, you inspire me. I thank my dear friend Mr. Tunde Asaboro and finally my sister Mrs. Yetunde Babade for their goodwill and prayers.

I thank Dr Anu Shani for her guidance and support in carrying out this research project.

Sincerely,
Oluwadamilare Adetuberu.

References

Namoun, A. & Alshanqiti, A., 2021. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, Switzerland,* 11(237).

Aljohani, N., Fayoumi, A. & Hassan, S., 2019. Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment.. *Sustainability..*

Al-Shabandar, R., Abir, H. & Laws, A., 2017. Machine learning approaches to predict learning outcomes in Massive open online courses. *Proceedings of the International Joint Conference on Neural Networks.*

Badr, G., Algobail, A., Almutairi, H. & Almutery, M., 2016. Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. *Procedia Computer Science,* pp. 80-89.

Bujang, S. et al., 2023. Bujang, S.D., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E.E., FujiMulticlass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access.*

Hendradi, P., Ghani, M. K. A. & Mahfuzah, S., 2023. A Literature Review of E-Learning Technology in Higher Education. *Journal of Computer Science and Technology Studies,* Volume 5(1), pp. 01-07.

Holmes, W. & Tuomi, I., 2022. State of the art and practice in AI in education. European Journal of Education. *European Journal of Education,* 57(4), pp. 542-570.

Hung, J., Rice, K., Kepka, J. & Yang, J., 2020. Improving predictive power through deep learning analysis of K-12 online student behaviors and discussion board content.. *Information Discovery and Delivery,* 48(4), pp. 199-212.

Jawthari, M. & Stoffa, V., 2022. Predicting At-Risk Students Using Weekly Activities and Assessments.. *International Journal of Emerging Technologies in Learning (iJET),* Volume 59-73., p. 17.

Jha, N. I., Ghergulescu, I. & Moldovan, A.-N., 2019. OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. *Proceedings of the 11th International Conference on Computer Supported Education,* Volume CSEDU 2019, pp. 154-164.

Jiang, S., Williams, A. E. & Schenke, K., 2014. Predicting MOOC Performance with Week 1 Behavior. *Proceedings of the 7th International Conference on Educational Data Mining (EDM).*

Kuzilek, J., Hlosta, M. & Zdrahal, Z., 2017. *Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) 'Data Descriptor: Open University Learning Analytics dataset',.* [Online]
Available at: https://analyse.kmi.open.ac.uk/open_dataset
[Accessed 2 August 2023].

Marwaha, A. & Singla, A., 2020. A study of factors to predict at-risk students based on machine learning techniques. *Advances in Intelligent Systems and Computing.*

Okoye, K., Arturo, A.-P. & Claudia, C.-Z., 2022. Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification. *Education and Information Technologies,* 27(3).

Rufai, A., Suru, D. & Afrifa, J., 2021. The Role of Machine Learning and Data Mining Techniques in Predicting Students' Academic Performance.. *International Journal of Computer Applications Technology and Research..*

Saa, A., Al-Emran, M. & Shaalan, K., 2020. Mining Student Information System Records to Predict Students' Academic Performance.. *The International Conference on Advanced Machine Learning Technologies and Applications ,* Volume 921.

Saxena, M., 2020. Predictive Analytics of Education Data Based Learning Patterns - A Literature Review.. *Saxena, M. (2020). Predictive Analytics of EducatioInternational Journal for Research in Applied Science and Engineering Technology..*

Xu, J., Moon, K. & Van Der Schaar, M., 2017. A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal on Selected Topics in Signal Processing,* 11(5), pp. 742-753.

York, T. T., Gibson, C. & Rankin, S., 2019. Defining and Measuring Academic Success. *Practical Assessment, Research, and Evaluation,* 20(5).

Zeineddine, H., Udo, B. & Assaad, F., 2021. Enhancing prediction of student success: Automated machine learning approach. *Computers and Electrical Engineering,* Volume 89.