Education and Information Technologies (2023) 28:3027–3057
https://doi.org/10.1007/s10639-022-11299-8

Predictive modelling and analytics of students' grades using machine learning algorithms

Yudish Teshal Badal[1] · **Roopesh Kevin Sungkur**[2]

Received: 21 January 2022 / Accepted: 24 August 2022 / Published online: 8 September 2022

Abstract

The outbreak of COVID-19 has caused significant disruption in all sectors and industries around the world. To tackle the spread of the novel coronavirus, the learning process and the modes of delivery had to be altered. Most courses are delivered traditionally with face-to-face or a blended approach through online learning platforms. In addition, researchers and educational specialists around the globe always had a keen interest in predicting a student's performance based on the student's information such as previous exam results obtained and experiences. With the upsurge in using online learning platforms, predicting the student's performance by including their interactions such as discussion forums could be integrated to create a predictive model. The aims of the research are to provide a predictive model to forecast students' performance (grade/engagement) and to analyse the effect of online learning platform's features. The model created in this study made use of machine learning techniques to predict the final grade and engagement level of a learner. The quantitative approach for student's data analysis and processing proved that the Random Forest classifier outperformed the others. An accuracy of 85% and 83% were recorded for grade and engagement prediction respectively with attributes related to student profile and interaction on a learning platform.

Keywords Machine learning · Predictive analysis · Random forest · Online learning platform · Student engagement

✉ Roopesh Kevin Sungkur
r.sungkur@uom.ac.mu

[1] Mauritius Institute of Education, Reduit, Mauritius

[2] Department of Software and Information Systems, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Reduit, Mauritius

1 Introduction

With the Covid-19 pandemic putting the world in an unprecedented crisis, technology has played a vital role in maintaining continuity as far as possible (Dhawan, 2020). According to UNESCO, more than 990 million learners are affected by the crisis. The implementation of E-learning systems is the only sustainable solution. This change has caused multiple challenges and opportunities to the community that can be harnessed to improve the quality of service (UNESCO, 2020). Universities proved how E-Learning is beneficial for distance education.

The amount of data collected through E-learning platforms have massively increased over the last few years. As of now, more than 58 million students have registered for online courses worldwide with above 7000 courses offered (Moubayed et al., 2018). All the information collected can be utilised through machine learning and data analytics in the domain of online learning. Fields such as educational data mining and learning analytics are emerging with the aim to improve teaching/learning by making use of machine learning and visualisation techniques. Making efficient use of analysing and tracing data is still challenging. Machine learning is an efficient tool with the capacity to find hidden patterns of learner interaction. It can analyse complex non- linear relationships and has demonstrated to be a feasible approach in obtaining prediction of users on online learning platforms (Al-Shabandar et al., 2017).

Even though access is easier for developed countries, the dropout rate is higher compared to traditional modes of delivery. Consequently, assessing a student's performance is challenging. Academics are interested in forecasted results on assessments since they can direct their effort in improving the student's experience (Bakki et al., 2015). At present, universities administer courses using online learning platforms. Analysts are making use of input features such as time, activity, assessment, and online discussion forums to forecast student performance. Recently, academics have been focusing a lot on predicting a learner's performance and explaining the factors that affect the learning process (Sorour et al., 2015). The information collected can be used for decision making in terms of curriculum design, content, and mode of delivery. Students in universities are often unable to complete a course due to lack of understanding and engagement with the topics which is undoubtedly a matter of concern (Patil et al., 2018).

Being able to predict the grades of a learner is important in the learning process since it will help academics to understand the learner's full potential and give the academics enough time to take corrective measures. Indeed the role of the academic should be to accompany the learner throughout his/her learning process and to be able to take corrective measures well before the exams. In line with what has been discussed above, the main aim of this research is to provide a predictive model to forecast students' performance (grade/engagement) and to analyse the effect of online learning platform's features. Implementing a working predictive software can be a baseline to initiate other research opportunities in the field of predictive analysis and the tools used for online learning. The wider implication of the study involves opening new avenues in terms of research in

techno pedagogy. With predictive analysis in mind, academics can venture into new online learning tools, instructional design and adaptive learning techniques to revamp the content for students. The results of the analytics will be of assistance to students who are mostly likely to perform poorly. Corrective measures can therefore be implemented by academics and tutors.

1.1 Rationale and significance of research

Education plays an enormous role in what constitutes a society. Modern society is based on people who have high living standards and knowledge which allows them to implement solutions to challenging problems. Higher educational institutions are functioning under an increasingly convoluted environment. The competition among institutions, the response to local and global economic changes, politics and social changes are among a bunch of factors impacting the proportion of students, disciplines available and the overall quality. Institutions' management are intended to adapt their decision-making process with the rapid changes occurring. Those decisions are often made without recourse to the vast data sources that are generated by manual and digital systems. The data, coupled with a predictive analysis system, can bring to light innovative action plans (Daniel, 2014).

To impart new skills and knowledge, universities have been making use of online learning platforms to deliver content. The global challenge for education is not just about providing access, but to ensure learning is taking place. To assess the comprehension of the subject, academics are making use of features such as online quizzes and discussion forums. Thereafter, students are examined through handwritten exams or assignments. The performance on the quizzes, discussion forum and the background of the students can be used to determine the grade for the handwritten exam or assignment. The performance allows the evaluation if a student has grasped the knowledge imparted and provides a scientific approach to investigate gaps (Yin, 2021).

1.2 Research questions

Research questions act as a catalyst for research projects and help to focus on the steps that will be taken to produce the analysis, findings, and results. Kitchenham's approach was considered to create the research questions. This approach takes into consideration the Population, Intervention, Context and Outcome (PICO) (Shahiri et al., 2015). The criteria were defined below in Table 1 below.

The research questions were framed in the Table 2 below.

2 Literature review

2.1 E-learning and online learning platform

E-learning is the delivery of education and all related activities using various electronic mediums such as the internet. It has provided several benefits like

Table 1 Criteria for research questions

Criteria	Details
Population	University students
Intervention	Using machine learning algorithms for predicting performance
Context	Academic institution specialised in pedagogy Use of secondary data (Student's data, examination results, online learning platform & files)
Outcome	Predicting accuracy of machine learning algorithms and correlation analysis

Table 2 Research questions

#	Research Question & Description
RQ 1	How accurate are the machine learning algorithms at predicting students' performance (Grade & engagement)? The dataset set compiled will be fed to 7 machine learning algorithms which are at the forefront of the analytics community. The best one can be a deciding factor for an education analytic framework
RQ 2	What are the important attributes in predicting the students' grade? The students' dataset will consist of multiple attributes such as age, certificates obtained, experience, activities and so on. The attributes retained for prediction will be identified through analysis
RQ 3	Can an adaptable predictive modelling framework be developed for student performance and engagement? The framework should cater for new features in online learning and predict the performance and engagement of a student

learner's flexibility and an increase in interactions through both asynchronous and synchronous in the form of digital activities by using a learning management system (LMS) (Coman et al., 2020). Asynchronous e-learning is the most prevalent form of teaching/learning technique of E-learning due to its flexible methodology. LMS is used in an asynchronous environment to provide students with available learning objects in the form of video, document, audio, presentation and so on. The online learning platform provides the framework for students to view and communicate asynchronously and act as a repository for learning objects (Cohen & Nycz, 2006). Quizzes and assignments are also among the most helpful asynchronous activities for education (Perveen, 2016).

The use of LMS has increased tremendously in higher education. Discussion forum is a helpful asynchronous approach to initiate exchange of ideas and to participate on a particular topic. The students, at their own comfort, can contribute on the platform (Shida et al., 2019). Other tools such as quizzes can act as a self-assessment exercise to help students to improve understanding of concepts. LMS are usually able to capture student's data and activities. One area of research is the multi-faceted benefits when exploiting the data (Shida et al., 2019). Devising asynchronous e-learning policies can increase student motivation, participation,

problem solving, analytics and thinking skills (Adem et al., 2022; Chilukuri, 2020).

2.2 Student engagement

The interest in exploring learning analytics related to student engagement has been growing considerably lately. This has further expanded the research field for education. Higher education institutions have shown their interest in making use of analytics to support their engagement. This can act as an instrument that will help in mediating student/teacher information sharing resulting in effective learning, improve awareness, and a way to tackle current challenging situations (Silvola et al., 2021). Students who are engaged in their activities normally perform well and take pleasure in learning new content. Research has revealed that student engagement influences cumulative learning, long-term achievement, and promotes overall learner's well-being (Salmela-Aro & Read, 2017). (Dewan et al., 2019) reviewed the engagement detection techniques in an online learning environment with its challenges. The detection methods were classified as automatic, semi-automatic and manual. Techniques in the automatic category obtain data from various sources. Log-files have been an efficient way of extracting valuable information especially in an online learning environment. (Cocea & Weibelzahl, 2011) analysed logs generated in an online platform known as HTML-tutor. They were able to extract 30 attributes such as number of tests attended, correct answers given number of pages accessed and so on (Dewan et al., 2019).

2.3 Machine learning

The purpose of machine learning is to obtain information from data which is why it is closely related to statistics, AI, and computer science. There are 3 types of machine learning techniques namely supervised learning, unsupervised learning, and reinforcement learning (Müller & Guido, 2017). Machine learning algorithms that learn from inputs and respective output pairs are known as supervised learning algorithms. Their purpose is to be able to generalize from known examples to automate decision-making processes (Müller & Guido, 2017). Though it is difficult to mount and analyse a dataset, supervised learning algorithms are popular, and their performance is easy to calculate. Unsupervised learning looks for undetected patterns in an unlabeled data set and little human supervision. In reinforcement learning, an agent will observe an environment to learn and achieve a goal. The computer employs trial and error to solve a problem (Russell, 2018).

The efficiency of a machine learning solution relies on the nature of the dataset and performance of the algorithms. Selecting a proper learning algorithm that is suitable for an application in a particular domain is strenuous. The reason behind this is that the purpose of ML algorithms is different. Even the outcome of different learning algorithms in a similar category may vary depending on the data characteristics (Sarker et al., 2019). Many machine learning algorithms have been

implemented in the research community. Among the most important and famous techniques that figure in data science literature are listed below (Russell, 2018).

1. Logistic regression
2. K-nearest neighbors
3. Naïve Bayes
4. Decision trees
5. Random forests
6. Support vector machines
7. Deep Learning

2.4 Logistic regression

Basically, linear regression is performed to determine the relationships between two or more variables impacting each other, and to make predictions by making an analysis on the variations (Uyanık & Güler, 2013). Models requiring more than one independent variable are known as multiple linear models. The equation describes how independent variables affect the dependent variable (Petrovski et al., 2015).

$$Y = \beta 0 + \beta 1\ x1 + \beta 2\ x2 + ... + \beta n\ xn$$

Whereby x is an independent variable, y is the dependent variable, B1,β2,…… βn are unknown parameters (coefficients) and β0 is a constant to create a line of best fit. Using linear regression is not convenient for categorical output. Considering for example a two-class classification problem, linear regression is prone to plot inaccurate decision boundaries in the presence of outliers. Logistic regression was developed for classification problems. The objective of logistic regression is to map a function from the features of the dataset to the targets to calculate the probability that a new entry belongs to one of the target classes (Bisong, 2019).

2.5 K-nearest neighbors

It is among the most basic and straightforward classification techniques. This method is suitable when there is little or no information about the distribution of the data. KNN was developed when reliable parameters to estimate probability were unknown or hard to establish (Hall et al., 2008). A parameter named k determines how many neighbors will be selected for the algorithm (Zhang, 2016). The performance is mainly determined by the choice of *k* and the distance metric used. If *k* is small, the estimate tends to be poor because of sparseness in data. Large values of k cause over-smoothing, performance degradation and miss out on important patterns (Zhang, 2016).

The aim is to choose a suitable k value to balance out overfitting and underfitting. Some researchers suggest setting k equal to the square root of the number of observations in the dataset (Gil-García & Pons-Porrata, 2006).

The k-nearest-neighbor classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. By default, the knn() function use the Euclidean distance which can be determined with the equation:

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Whereby D is the Euclidean distance and p and q are subjects to be compared with n characteristics (Zhang, 2016).

2.6 Naïve bayes

It makes use of a simple probabilistic function for classification. It computes a set of probabilities by calculating the frequency and combination of values in a dataset. It allows all attributes to contribute to the final decision equally. (Wibawa et al., 2019)

$$P(Q|X) = \frac{P(X|Q).P(Q)}{P(X)}$$

With

X	Data with unknown class
Q	The hypothesis X is a specific class
(Q|X)	The probability of the Q hypothesis refers to X
(Q)	Probability of the hypothesis Q (prior probability)
(X|Q)	Probability X in the hypothesis Q
(X)	Probability X

Naïve Bayes works well with high-dimensional sparse data and is insensitive to irrelevant data or noises. Its simplicity and low execution time makes it an ideal choice for predictive analysis. (Müller & Guido, 2017).

2.7 Decision trees

A decision tree has a tree-like structure where each node shows an attribute, each link shows a decision (rule) and each leaf shows an outcome. It can be used for both continuous and discrete data sets (Patel & Prajapati, 2018). Decision tree begins with a root node. From this node, users split each node recursively according to a decision tree learning algorithm based on if-the questions (Yadav & Pal, 2012). The result is a decision tree in which each branch represents a possible scenario of

decision and its outcome (Sungkur & Maharaj, 2022). An example of a decision tree is shown in Fig. 1 below.

2.8 Random forests

Random forest is considered as an expert solution for the majority of problems and falls under the ensemble learning classifiers whereby weak models are combined to create a powerful one. Ensemble methods are among the most promising areas for research. It is defined as a set of classifiers whose predictions are brought together to forecast new instances. Ensemble learning algorithms have shown to be an efficient technique to improve predictive accuracy and dampen learning problem complexities into sub-problems (Krawczyk et al., 2017). Numerous decision trees are produced in random forests. To classify an object having attributes, every one of the trees gives a classification which is also considered as a vote. The forest is then given the ability to choose the classification with the maximum votes. This is shown in Fig. 2 below.

2.9 Support vector machines

The basic idea of SVM is to plot data in n-dimensional space with n number of features and apply a hyperplane to distinguish the classes which are used for classification and regression (Deepa & Senthil, 2020). The input space is mapped to a

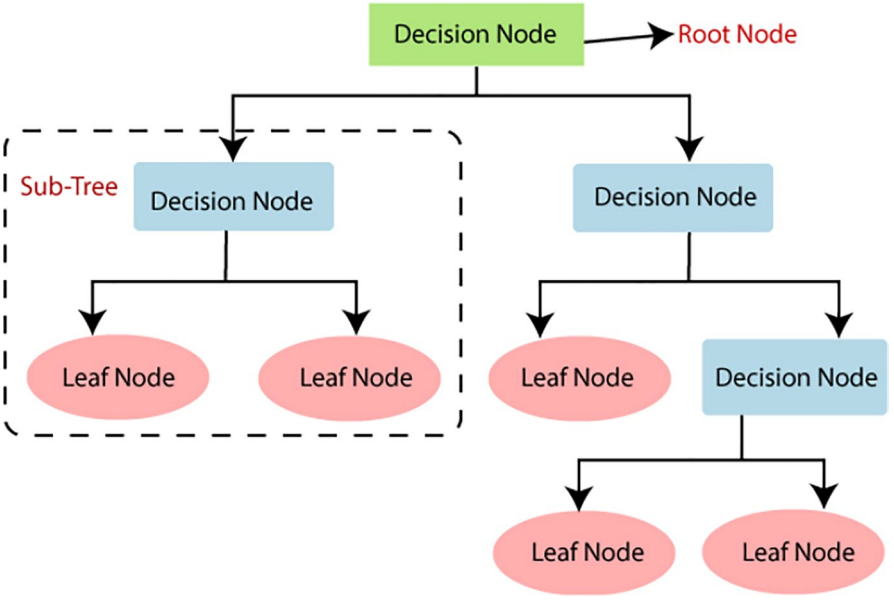

Fig. 1 Structure of decision tree algorithm (Hafeez et al., 2021)

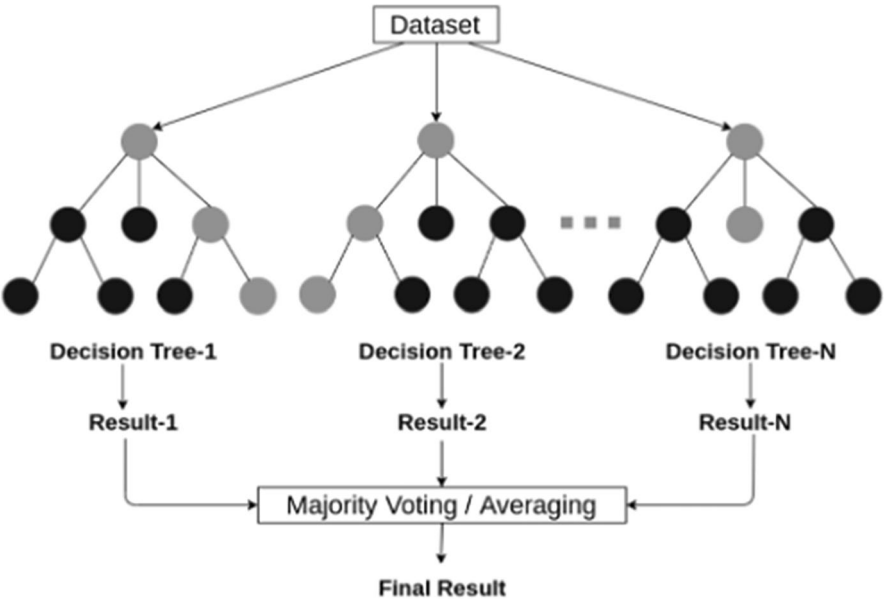

Fig. 2 An example of a random forest structure considering multiple

high-dimensional feature based on a transformation defined by a kernel function (Ø) (Theobald, 2017). This is shown in Fig. 3 below.

The hyperplane classifies the data separated by boundaries produced by the hyperplanes that separate classes of data points (Nayak et al., 2015). The optimization objective is to maximise the margin which is the distance between the separating hyperplane, decision boundary, and the training samples that are closest to this hyperplane (Raschka & Mirjalili, 2017). Using large margins tends to produce lower generalisation errors in models where a small margin is more likely to overfit. This is illustrated in the Fig. 4 below.

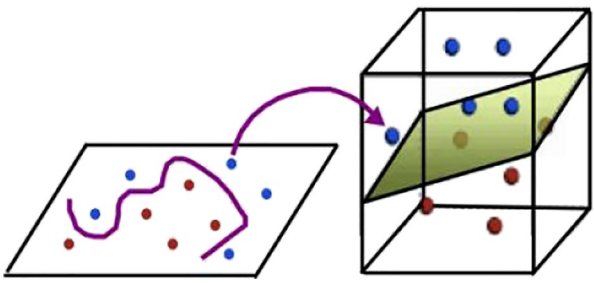

Fig. 3 Transformation of data into a higher dimension with the kernel function (Theobald, 2017)

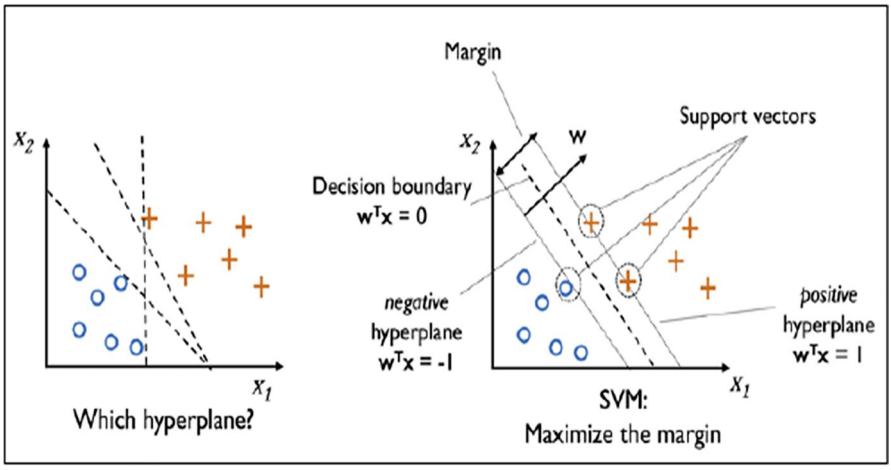

Fig. 4 Decision boundary distance in SVM (Raschka & Mirjalili, 2017)

2.10 Deep learning

Deep learning is a subfield of ANNs termed as such due to its use of multiple layered neural networks to process data. The idea is to have hidden layers (called hidden since they do not receive the raw data) combine the values in the preceding layer to learn the more complicated function of the input. (Sungkur & Maharaj, 2021) presents a research where ANN with Backpropagation Algorithm is used to provide personalised learning for cybersecurity professionals. This approach addresses the problem of 'one-size-fits-all' learning and makes the learning process more motivating, engaging and effective.

It is challenging for computers to understand raw data. This is where deep learning decomposes challenging problems into a series of nested concepts where each is described by a different layer of the predictive model (Di Franco & Santurro, 2020). Implementing typical machine learning algorithms is usually repetitive with lots of trial-and-error methods. Selecting different algorithms will produce different results which can be acceptable in several contexts. Nevertheless, with the limitations of different algorithms and the upsurge in machine learning theories and infrastructure, deep learning is, as technique, a more profound way of explaining high/low level of abstraction for a given dataset which typical machine learning algorithms are unable to do (Beysolow, 2017).

2.11 Machine Learning life cycle and methodology

Machine learning has its own life cycle that is the process the data undergo for the development and deployment of a predictive system. As compared to software development life cycle, the development of machine learning models involves experimenting on datasets to achieve the aims and objectives defined when applying fresh data after training (Ashmore et al., 2019). The basic workflow necessitates

extraction of data, training, testing, tuning and evaluating the model before deploying it to production (Landset et al., 2015). This is shown in Fig. 5 below.

Machine learning systems demand that data are in a certain format to be fed and processed. Essential processing activities are performed such as cleaning of unusual values, handling mistakes, formatting and normalisation.

2.12 Data sources & education database system

The data sources have to be identified for extraction and consolidation in a container to facilitate access. The education system is continuously expanding with an increase in the number of students. The student information has therefore increased considerably causing a lot of pressure in information organisations both at academic level (Yin, 2021). Research and academic institutes are working to uncover new theories from knowledge discovery. Modern educational institutions are constantly undergoing digital transformation both in terms of administrative and teaching/learning services. A comprehensive digitisation of the education process is at the root of all research. This will increase the attention of researchers in the field of data science and machine learning (Yin, 2021).

The core idea is to have the centralised database act as a data warehouse used to process and manage data (Jayashree & Priya, 2019). The data from multiple heterogeneous sources are put together in an organised and easily accessible manner. This enhances decision-making and provides greater insight in an organisation's operation. Data warehousing, mining and analytics are famous in the business world. Its usage is still low in educational institutions. However, different studies and research areas in educational data mining are motivated to have their analytical processes applied to a database. The need of a data warehouse is obvious for learning analytics and evaluation of teaching–learning techniques (Moscoso-Zea et al., 2018). Quality-wise, it can be an instrument in obtaining organisational knowledge (Moscoso-Zea & Lujan-Mora, 2017).

Educational institutions with centralised databases can improve information management. Strategy implementation by board of directors, recruitment decision,

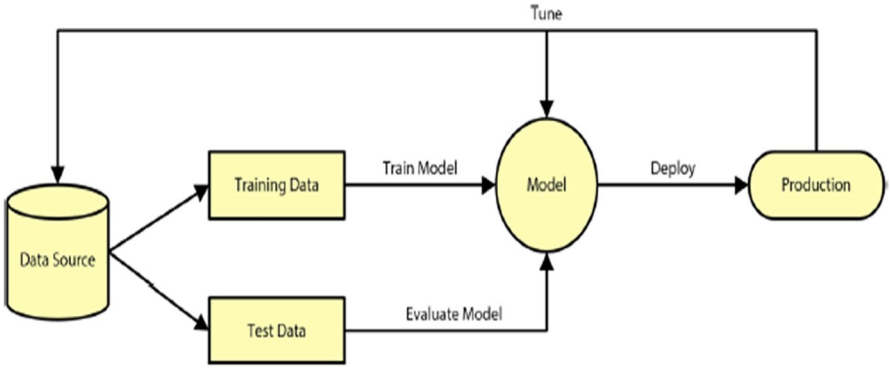

Fig. 5 Basic machine learning workflow (Landset et al., 2015)

retention and performance of students are some among several benefits that organisations can exploit (Williamson, 2018). Modern database system is emerging in the education domain leading to the emergence of Education Data Mining (EDM) as a field. It is playing a huge role in identifying patterns in learning principally performance. Predicting performance, success and retention rate with an e-learning environment as backdrop is becoming essential (Alyahyan & Düştegör, 2020). There are different approaches in building a database system/warehouse. Kimball's methodology uses a bottom-up approach which is convenient for projects having limited time and usually on a budget (Kimball & Ross, 2013). This is shown in Fig. 6 below.

A study by (Moscoso-Zea et al., 2016) suggested the design of Kimball's to be convenient in educational institutions. The main reason is that their units are not integrated and usually function individually. Implementation design is challenging but beneficial for data consolidation and analysis. (Moscoso-Zea, et al., 2018). The major benefit of centralising the data is the possibility of having multiple client applications retrieving data simultaneously. All data stored one place allows easier querying and benefits in terms of execution time. Other applications, for example analytics software, can plug into the database (Singh, 2011).

2.12.1 Performance metrics

Evaluating the predictive model is an essential part to determine the accuracy of the student's performance. To do so, it is important to quantify the quality of a system's predictions (Mourdi et al., 2019). Some important performance metrics to assess the machine learning techniques are:

Accuracy It is defined as the ratio of correct predictions to total number of sample input. It is a frequently use metric to assess the quality of a classifier's solutions. It is

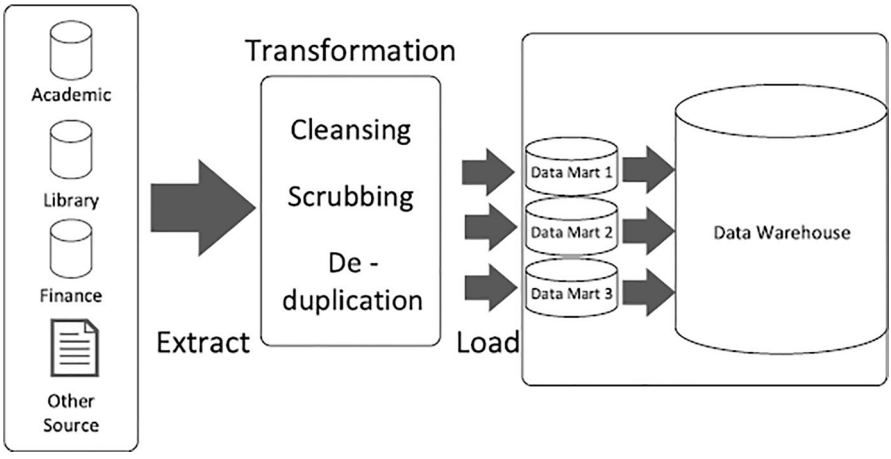

Fig. 6 Ralph Kimball's bottom-up approach to DWH design. (Kimball & Ross, 2013)

the most used evaluation metric for both binary and multi-class classification. It is a determining value for assessing the capability of an algorithm

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Sample}$$

Precision It is the number of correct positive results divided by all samples labelled as positive by the algorithm.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall (Sensitivity) It is the number of correct positive results divided by all samples that should have been labelled as positive by the algorithm.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F-measure (F_1-score) A model can have a high recall value with low precision. Those values alone are not enough for indicating a good classifier. F-measure represents a harmonic mean of precision and recall. A higher value designates a high classification performance (Table 3).

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

The variables used in the equation are defined as follows:

2.13 Related works

Considerable research has been conducted to forecast student performance. Researchers have gone through different methodology to showcase results to support their findings in terms of evaluation metrics. Lately, (Adnan et al., 2021) investigated the capabilities of seven algorithms such as the traditional SVM, KNN, ensemble techniques and deep learning. Though having recorded accuracy close to 91% with random forest, the attributes didn't cater for several online learning tools. The student profile and number of clicks are among the 13 attributes used for prediction on a large dataset of 35,593 imbalanced records. (Ko & Leu, 2021) happened to record 82.26% with Naïve Bayes on a small dataset of 215 students without any

Table 3 Variables for performance metrics (Michelucci, 2019)

Variables	Definition
True positives (TP)	Tests are predicted correctly
False positives (FP)	Test predicting a particular class but actually is not
True negatives (TN)	Test correctly predicting not belonging to a class
False negatives (FN)	Test predicted as not belonging to a particular class when in fact it is

balancing technique. The latter did not include online learning features as attributes. The e-learning aspects have been considered by (Mourdi et al., 2019) together with a dataset of 3585 students. The 25 attributes include information from quizzes, videos and forums. Though unbalanced, an accuracy as high as 99% was obtained for identifying a student as pass, fail or drop out. (Bujang et al., 2021) and (Costa et al., 2017) catered for imbalance classes during their analysis. (Bujang et al., 2021) indicated how the combination of SMOTE and feature selection influence accuracy of predictive models. (Costa et al., 2017) coupled SMOTE with fine tuning of algorithms. However, no comprehensive documentation was mentioned in terms of hyperparameters and values. (Tarik et al, 2021) opted to remove all missing data from its initial 142,110 students. With the remaining 72,010, accuracy of up to 70% were recorded with Random Forest.

Liu et al. (2022a) highlights on the importance of emotional and cognitive engagement as two prominent aspects of learning engagement. The authors further discuss about how emotional and cognitive engagement further share an interactive relationship and that these two factors thereafter jointly influence learning achievement. Liu et al. (2019) presents an unsupervised model, namely temporal emotion-aspect model (TEAM), modelling time jointly with emotions and aspects to capture emotion-aspect evolutions over time. Liu et al. (2022b) explores the relationship between social interaction, cognitive processing and learning achievement in a MOOC discussion forum. Liu et al. (2022c) discusses the relationship between discussion pacing (i.e., instructor-paced or learner-paced discussion), cognitive presence, and learning achievements. Emotion experiences, cognitive presence or social interactions in discourses as highlighted by the works of (Liu et al., 2019, 2022a, 2022b, 2022c) provide some deeper and implicit features that have a definite impact on the learning achievement of the learner.

3 Proposed solution

3.1 Research design

Experimental research is essentially the investigation of one or more variables (dependent variables) manipulated to assess the effect on one or more variables known as independent variables. It is based on the cause-and-effect relationship on a chosen subject matter to conclude the different relationships that a product, theory, or idea can produce (Jongbo, 2014). The nature among the variables is established with precise and systematic manipulation. This technique is suitable where testing theories and evaluation of methods are at the core of a study. Furthermore, the same set up and protocol can be replicated with the same variables. This can substantiate the validity of products, ideas, and theories. (Wabwoba & Ikoha, 2011). Additionally, this type of scientific approach can provide a set a guideline for evaluating and reporting information for research (Marczyk et al., 2005). Figure 7 below shows a popular general aspect of how experiments are conducted before reaching model evaluation.

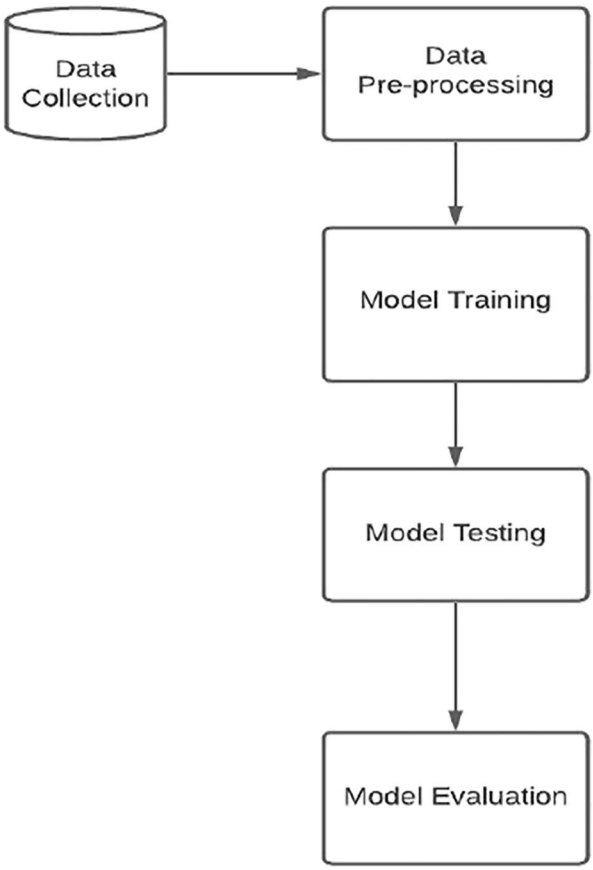

Fig. 7 General Research Approach for machine learning (Kamiri & Mariga, 2021)

To empirically assess the algorithms and interpret the research outcome, the criteria used for the experimental procedures will be set as:

- **The algorithms running**

  Type of supervised learning algorithm

- **The evaluation technique**

  The training and testing procedures (E.g Cross validation)

- **Predictive performance on unseen data**

  This involves estimation metrics such as percentage accuracy.

- **Model specific properties**

  Hyperparameters (E.g. depth of a decision tree)

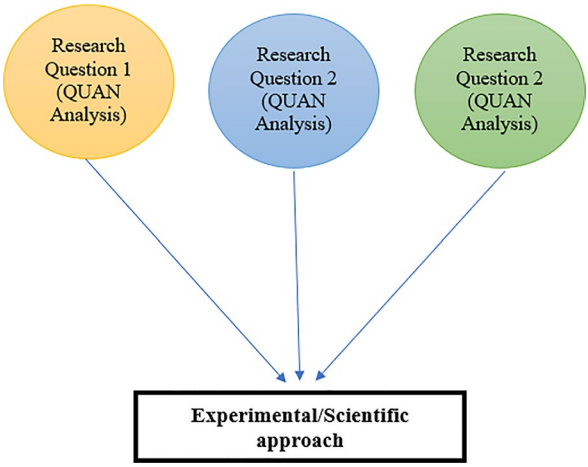

Fig. 8 Research design derived from research questions

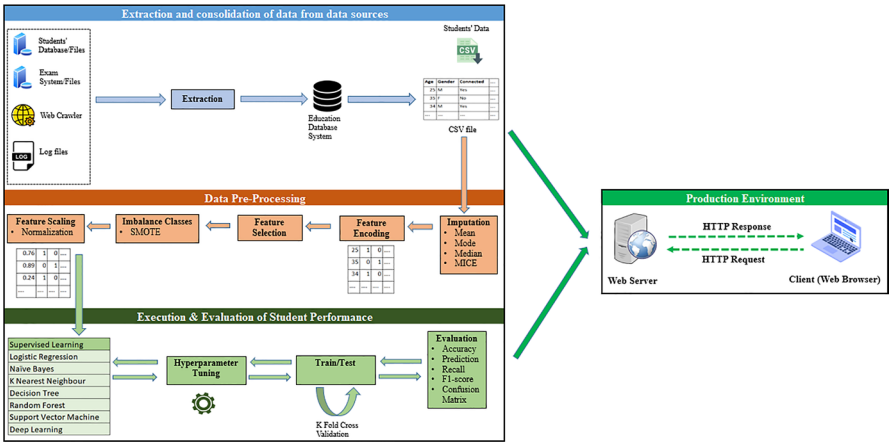

Fig. 9 Machine learning architectural design to evaluate student performance

In this research, the controlled experiment will be related to the machine learning algorithms used and the estimation of each model on unseen data. It will be a convenient method to discover the techniques chosen that work best for the dataset and under which specific conditions with systematic experimentation. A comparative model analysis will have the supervised learning algorithms as variable and the assessment criteria as dependent variable. This is shown in Fig. 8 below.

3.2 Machine learning architectural design

To interpret the processes, a new workflow with feature selection techniques and a way to handle imbalance classes have been incorporated in typical ML procedures (Fig. 9). It is represented in the architecture below.

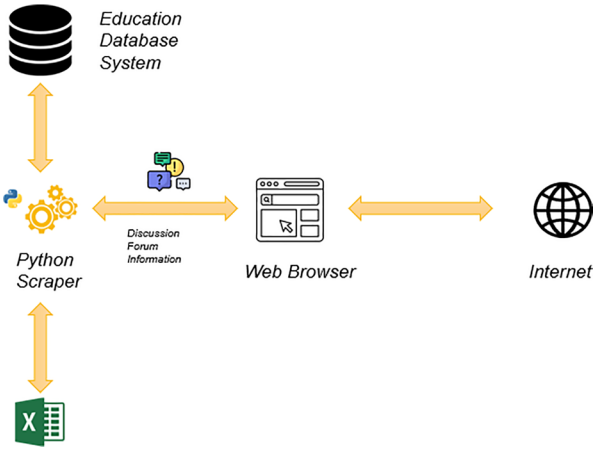

Fig. 10 Web Scraper Architecture for retrieve discussion forum information

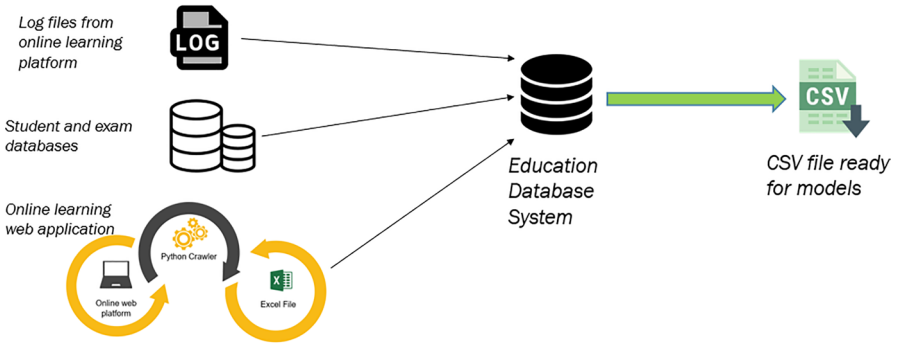

Fig. 11 CSV format generated from database

3.3 Web scraper

Due to the restriction on accessing the LMS database and limitations, a web scraper is required to retrieve data pertaining to discussion forums by sifting through the web pages of the internet-facing application. A web scraper was developed as shown in Fig. 10 below.

3.4 Data consolidation & database

The data from the LMS, examination and student section have to be consolidated into a database. It is advantageous to have all data under one umbrella prepared for data extraction. A csv file can then be generated. The tabular dataset will be fed to the models (Fig. 11).

3.5 Proposed framework

The practices underlying the concept of data consolidation, processing and evaluation of student performance were translated into a framework. The same concept can be re-applied in different educational contexts (Fig. 12).

3.6 Dataset

The data digitally available was extracted and cleaned. As per the cohorts available, a total of 1074 students' data was used (Table 4).

3.7 Student performance prediction software

The implementation of a software for predicting performance will address challenges in a systematic manner. The functional requirements describe the intended function of the Student Performance Prediction Software and are shown in Table 5 below.

A web application was developed to execute the machine learning algorithms as per the best instance. The rationale behind it is to have a front-end web interface where users will be able to upload a file with a student's data. The interface will then predict the student's performance and engagement. The application can be deployed on a production environment for the institution (Fig. 13).

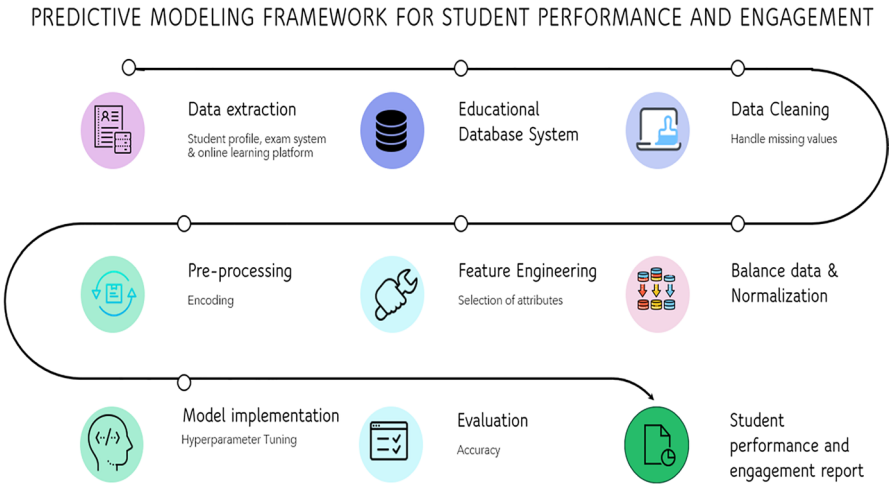

Fig. 12 Proposed framework for predicting student performance and engagement

Table 4 Dataset after feature encoding for predicting student's grade

#	Feature Name	Description	Data type
1	Age	Student's current age	Ordinal
2	Rural or urban	Student living in urban or rural area	Nominal
3	Gender	Male or Female	Nominal
4	Work	Student's work status (No, work in rural or urban area)	Nominal
5	Marital status	Married or single	Nominal
6	No of Attempt	Number of times student attempted exam	Ordinal
8	Work experience count	Number of work experiences	Ordinal
9	Student's total work days	Total number of work experiences in days	Ordinal
10	No of certificate in Training Centre	Number of certificated obtained in training centres	Ordinal
11	Masters at University	Number of master's degree obtained	Ordinal
12	Undergraduate at University	Number of bachelor degree or diploma awarded	Ordinal
13	Sum No of days at university	Total number of days spent in a tertiary institution as student	Ordinal
14	Connected	student has connected at least one time to LMS	Ordinal
15	Learning object menu	Page containing list of learning objcts	Ordinal
16	Learning object 1	Video, animation, Audio file, pdf file etc.…	Ordinal
17	Learning object 2	Video, animation, Audio file, pdf file etc.…	Ordinal
18	Learning object 3	Video, animation, Audio file, pdf file etc.…	Ordinal
19	Learning object 4	Video, animation, Audio file, pdf file etc.…	Ordinal
20	Learning object 5	Video, animation, Audio file, pdf file etc.…	Ordinal
21	Learning object 6	Video, animation, Audio file, pdf file etc.…	Ordinal
22	Learning object 7	Video, animation, Audio file, pdf file etc.…	Ordinal
23	Learning object 8	Video, animation, Audio file, pdf file etc.…	Ordinal
24	Learning object 9	Video, animation, Audio file, pdf file etc.…	Ordinal
25	Learning object 10	Video, animation, Audio file, pdf file etc.…	Ordinal
26	Number of post created	Number of posts initiated by student	Ordinal
27	Sum of Number of words	Total number of words submitted by student for all posts created	Ordinal

Table 4 (continued)

#	Feature Name	Description	Data type
28	Sum of Number of characters	Total number of characters submitted by student for all posts created	Ordinal
29	Participate in discussion	Total number of discussions participated	Ordinal
30	Sum of Number of words in discussion	Total number of words submitted by student in discussions	Ordinal
31	Sum of Number of characters in discussion	Total number of characters submitted by student in discussions	Ordinal
32	MCQ attempted	Student attempted MCQs(Yes or No)	Ordinal
33	Duration in hours	Time take to submit MCQs	Ordinal
34	MCQ Score	Score obtained in online MCQ	Ordinal

Table 5 Functional and non-functional requirements

Functional Requirements	
FR 1	The system should be able to accept a csv file with all data
FR 2	The system should allow the selection of ML algorithm for grade prediction
FR 3	The system should allow the selection of ML algorithms for student engagement level prediction
FR 4	The system should apply the Mean, Mode, Median and Mice imputation technique
FR 5	The system should be able to apply feature encoding
FR 6	The system should be able to remediate imbalance classes using SMOTE
FR 7	The system should be able to discard irrelevant attributes
FR 8	The system should evaluate the model's accuracy, prediction, recall and F1 score
FR 9	The system should be able to normalise the dataset
FR 10	The system should evaluate the best hyperparameters
FR 11	The system should display the best accuracy of the model in percentage

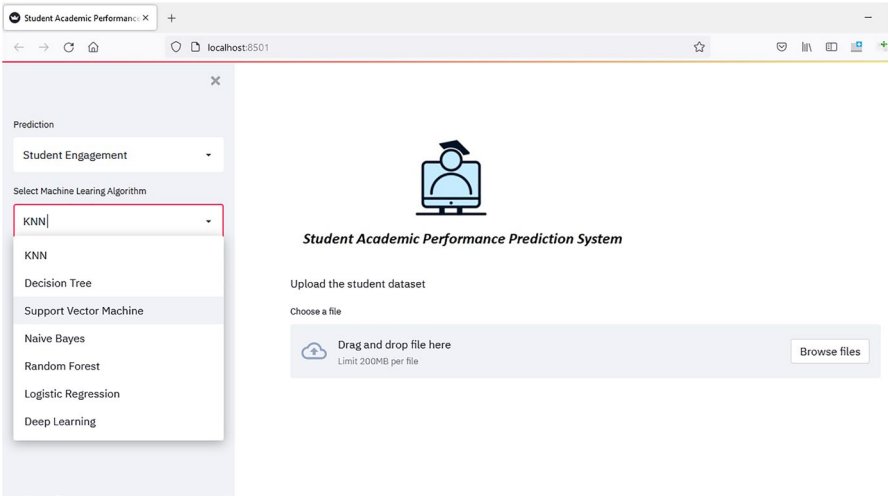

Fig. 13 Web application for student performance and engagement prediction

4 Results and discussions

4.1 Testing and evaluation

The accuracy of a model is the primary indicative factor to assess a model. The results obtained per algorithm were reported according to the highest accuracy observed. For the same configuration providing the best accuracy, the average total precision, recall and F1-scores per fold were recorded and illustrated in the sub-section. An average of the confusion matrix per fold was estimated. The machine learning algorithms that were compared and contrasted include Logistic regression,

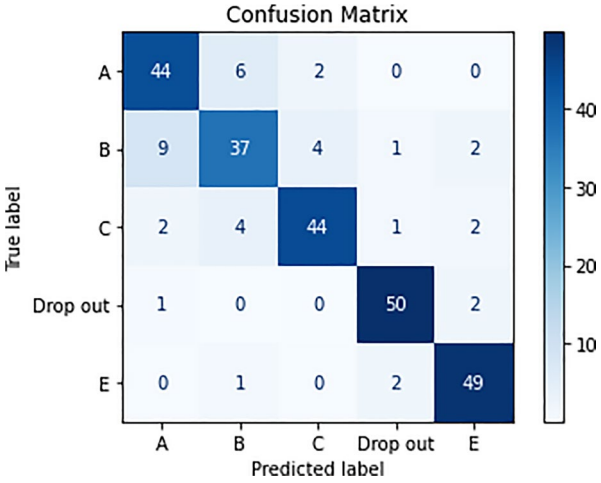

Fig. 14 Confusion matrix for grade prediction using RF

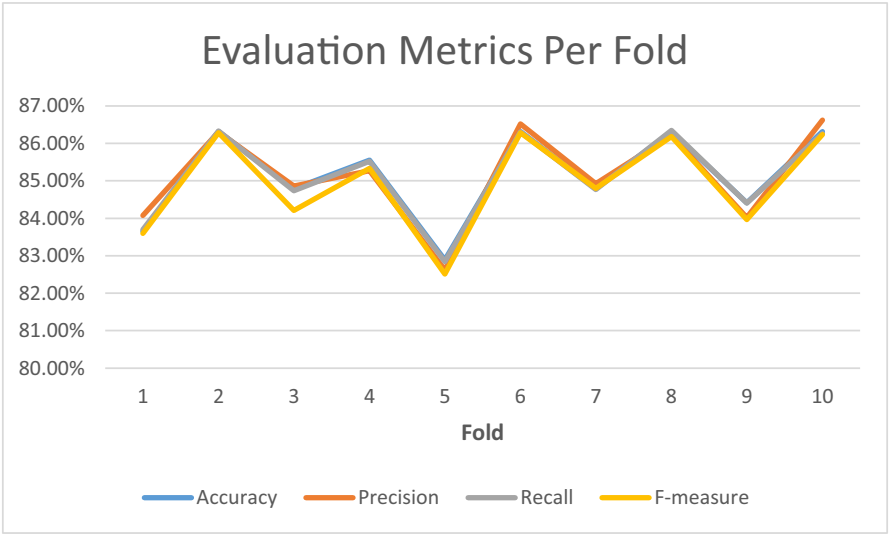

Fig. 15 Average of Evaluation metrics per fold—grade prediction using RF

K-nearest neighbors, Naïve Bayes, Decision trees, Random forests, Support vector machines and Deep Learning. It was observed that Random Forests yielded the best results. For the sake of simplicity, the diagrams of only Random Forests are shown below.

4.1.1 Random forests—Student grade prediction

Figure 14
Figure 15

Table 6 Performance metrics for grade prediction using RF

Metric	
Average Accuracy	85.13%
Average Precision	85.14%
Average Recall	85.12%
Average F-Measure	84.94%
Imputation Technique	
MICE	
Hyperparameters	
max_depth	None
max_features	auto
n_estimators	1000

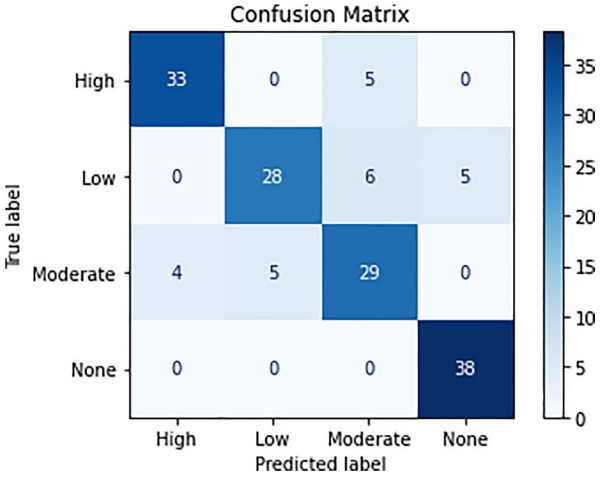

Fig. 16 Confusion matrix for engagement prediction using RF

Table 6

4.1.2 Student engagement prediction

Figure 16

Figure 17

Table 7

Following implementation and evaluation stages, all the functional and non-functional requirements have been achieved. The results were studied to answer the research questions initially set.

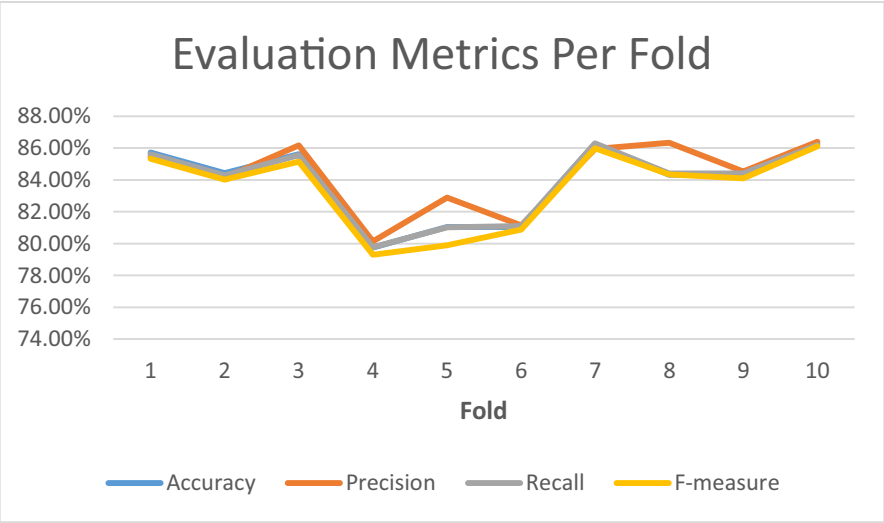

Fig. 17 Average of Evaluation metrics per fold—engagement prediction using RF

Table 7 Performance metrics for engagement prediction using RF

Metric	
Average Accuracy	83.88%
Average Precision	84.31%
Average Recall	83.86%
Average F-Measure	83.51%
Hyperparameters	
max_depth	10
max_features	auto
n_estimators	100

4.2 Research questions

This section discusses the answers to the research questions set earlier. This further helps to shed light on how machine learning algorithms can be used for predicting students' grades.

RQ 1. How precise are the machine learning algorithms at predicting students' performance (Grade & engagement)?

Figure 18

Figure 19

Evidence revealed Random Forest outperformed its counterparts in accuracy, prediction, recall and F1-score both for predicting grade and engagement level. The algorithm's properties seem to be effective for classification of such a peculiar dataset with the application of MICE as imputation technique, feature selection, SMOTE and normalisation. RF is particularly advantageous when dealing with high

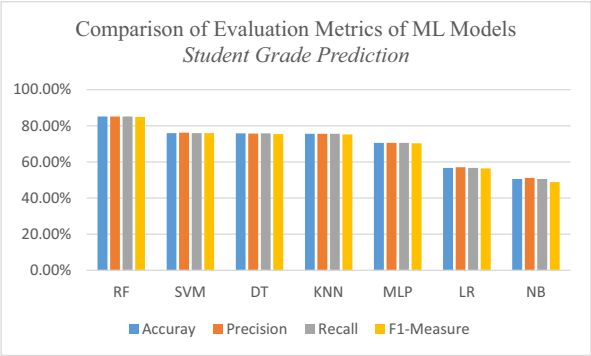

Fig. 18 Evaluation of ML Models for grade prediction

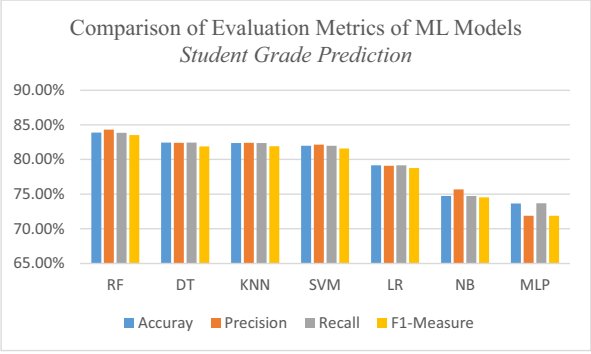

Fig. 19 Evaluation of ML Models for engagement prediction

dimensional attributes. Setting a high number of trees (n_estimators = 1000 and 100 trees in forest) as per the number of attributes unveiled an accuracy of 85% and 83% for both models respectively. The evaluation metrics for all classifiers oscillate in every fold. However, for the engagement prediction, MLP suffered a drastic drop in the second fold. It may have been exposed to data which is beyond its training configuration. Anomalies as such hurt a model as it is preferable to have a high metric value across all folds to ensure generalisation (Fig. 20).

Nevertheless, RF, SVM, DT, KNN & MLP obtained above 70% as average in all metrics. It can be concurred that the classifiers are applicable in an education-related context with multiple student attributes, both personal data and interaction in an online environment for grade/engagement classification.

RQ 2. What are the important attributes in predicting the students' grade?

Initially 42 features were identified for processing. Following imputation and category encoding, the feature selection technique discarded the redundant attributes that would not benefit the model. After MICE imputation, 16 features were removed. Collecting data can be expensive and since the redundant attributes can be ignored,

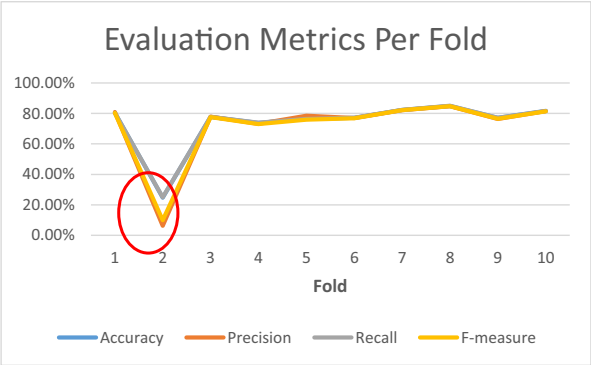

Fig. 20 Drop in 2nd Fold for Average of Evaluation metrics per fold—engagement prediction using MLP

the focus can be shifted on other new features. Among the 30 remaining attributes, 9 out of 10 learning objects were retained. A deeper understanding of the discussion forums and the way students tackle the MCQs can open promising directions. Only the total number of discussions, participation and the time taken to submit and MCQ was utilised. This can be a reference guide for potential research with respect to interactions and duration on a platform.

RQ 3. Can an adaptable predictive modelling framework be developed for student performance and engagement?

The architecture can accommodate new data and techniques for experimentation such as including new learning objects and going through different ML processes. The application can be deployed in different contexts as per the requirements and available data on or off premises. The prospect for new E-learning strategies and predictive analytics are enormous. The framework can eventually be transposed and adapted in different educational establishments (primary school, secondary school, training centres etc.) for experimentation and at production level. Having analytic tools incorporated in the educational system will allow institutions to provide a level playing field for scholars. The study can bring to the fore the importance of digital transformation and unification of data sources in a single container for analysis which is essential for E-learning analytics. The major achievement of this research and the difficulties encountered are outlined below.

4.2.1 Major achievements

1. A machine learning architecture and comprehensive life cycle was set up. The system implements all the machine learning phases such as imputation, feature selection, class balancing etc.…
2. The predictive framework for student grade and engagement level was translated into a working prototype through a web application. An engagement threshold was estimated.
3. The hyperparameter tuning was documented together with values and parameters.

4. The system was reworked to approximate confusion matrices and evaluation metrics per fold. A stratified cross validation of 10 folds were imposed and the results provide an average of the confusion matrix most likely to be generated from a set of data. The confusion matrix and charts are indicative with respect to the overall performance of classifiers.
5. Data consolidation and cleaning is an intricate process with hiccups at every turn. A web scraper has been developed to circumvent data retrieval issues which is common practice in data science projects. The centralised database was set up to assemble all relevant data.
6. Using latest releases of ML libraries (E.g., Keras) and resorting to release documents for guidance during development.

4.2.2 Difficulties encountered

1. The databases are not centralised. There are no unique identifiers present across all data sources making it difficult to extract, compare, and process data.
2. The log files are large causing the download process to often time out. The configuration setting on the server cannot be altered neither the settings on the web application. The log files also generate fields that are irrelevant to the. The log files were generated and downloaded in stages to avoid retrieving large files. The logs were then cleaned, and only pertinent data were retained.
3. Processing and debugging are tedious due to the execution time of the algorithm's constraint by the hardware.

5 Conclusion

This research can contribute to create actionable steps for growth to improve educational institutions' reputations and ranking both at national and international level. The software will be at the disposal of experts and will act as a device to help in reinforcing the learning process for existing or novel pedagogical interventions. Applying the framework as a magnifying glass on the education system can make way for innovative concepts that will undoubtedly bring waves of change in the learning process. From a scientific point of view, every phase of the machine learning life cycle can be further explored. Data scientists now researching new filtering algorithms, imputation techniques and normalisation procedures can measure their efficacy in the education context. Random Forest classifier outperformed the other classifiers. An accuracy of 85% and 83% were recorded for grade and engagement prediction respectively with attributes related to student profile and interaction on a learning platform. From an educational point of view, this research can help educators identify learners that are at risk as far as poor performance is concerned and can help the educators take timely corrective measures.

One of the limitations of this research is that external factors might be affecting the student performance when participating in discussion forums and quizzes, for example, the bandwidth and performance of computer or mobile devices might be impacting on the participation of the learner in certain learning activities. Future

studies can include the investigation and implementation of strategies in context with modern learning techniques such as personalised learning for isolated learners. Moreover, to supplement the quantitative approach, qualitative research methods can be combined to gather insights about the learning process and outcome of students. For example, analysing the students' feedback can be included in the existing experimental setup, subject matter experts for a given material can examine students' response and uncover other areas in teaching and learning when preparing the dataset. Future works also include multi-feature fusion since it would be interesting to feature out the causal relationship of emotion, cognition, behaviours and motivation behind learning performance and how to further improve it.

Data availability statement The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Adem, A., Çakıt, E., & Dağdeviren, M. (2022). Selection of suitable distance education platforms based on human-computer interaction criteria under fuzzy environment. *Neural Computing and Applications*, 1–13.

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access*, [online] 9, pp.7519–7539. Available at: https://ieeexplore.ieee.org/document/9314000 Accessed 10 December 2021.

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., & Radi, N. (2017). 'Machine learning approaches to predict learning outcomes in Massive open online courses', *Proceedings of the International Joint Conference on Neural Networks*.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1).

Ashmore, R., Calinescu, R., & Paterson, C. (2019). *Assuring the machine learning lifecycle: Desiderata, methods, and challenges*.

Bakki, A., Oubahssi, L., Cherkaoui, C., & George, S. (2015). Motivation and engagement in MOOCs: How to increase learning motivation by adapting pedagogical scenarios? *Design for Teaching and Learning in a Networked World*, pp. 556–559.

Beysolow, Y. (2017). *Introduction to deep learning using R: A step-by-step guide to learning and implementing deep learning models using R*. Ca Apress.

Bisong, E. (2019). *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners* (1st. ed.). Apress.

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. AMd. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access, 9*, 95608–95621.

Chilukuri, K. C. (2020). A novel framework for active learning in engineering education mapped to course outcomes. *Procedia Computer Science, 172*, 28–33.

Cocea, M., & Weibelzahl, S. (2011). Disengagement detection in online learning: Validation studies and perspectives. *IEEE Transactions on Learning Technologies*, [online] 4(2), pp.114–124. Available at: https://ieeexplore.ieee.org/abstract/document/5518758 [Accessed 26 Dec. 2021].

Cohen, E., & Nycz, M. (2006). Learning objects and e-learning: An informing science perspective. *Interdisciplinary Journal of e-Skills and Lifelong Learning, 2*, 023–034.

Coman, C., Țîru, L. G., Meseșan-Schmitz, L., Stanciu, C., & Bularca, M. C. (2020). Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective. *Sustainability, 12*(24), 10367.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256.

Daniel, B. (2014). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology, 46*(5), 904–920.

Deepa, B. G., & Senthil, S. (2020). "Constructive effect of ranking optimal features using random forest, support vector machine and naïve bayes forbreast cancer diagnosis." *Big Data Analytics and Intelligence: A Perspective for Health Care, First Edition, Emerald Insight*.

Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: a review. *Smart Learning Environments*, *6*(1). https://doi.org/10.1186/s40561-018-0080-z

Dhawan, S. (2020). Online learning: A Panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems, 49*(1), 5–22.

Di Franco, G., & Santurro, M. (2020). Machine learning, artificial neural networks and social research. *Quality & Quantity, 55*(3), 1007–1025.

Gil-García, R., & Pons-Porrata, A. (2006). *A new nearest neighbor rule for text categorization*.

Hafeez, M. A., Rashid, M., Tariq, H., Abideen, Z. U., Alotaibi, S. S., & Sinky, M. H. (2021). Performance improvement of decision tree: A robust classifier using Tabu search algorithm. *Applied Sciences, 11*(15), 6728.

Hall, P., Park, B., & Samworth, R. (2008). 'Choice of neighbor order in nearest-neighbor classification', *The Annals of Statistics*, 36.

Jayashree, G., & Priya, C. (2019). Design of visibility for order lifecycle using datawarehouse. *International Journal of Engineering and Advanced Technology, 8*(6), 4700–4707.

Jongbo, O. C. (2014). The role of research design in a purpose driven enquiry. *Review of Public Administration and Management, 3*(6), 87–94.

Kamiri, J. & Mariga, G. (2021). *Research methods in machine learning: A content analysis. international journal of computer and information technology* (pp. 2279-0764)

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit* (3rd ed.). Wiley, Cop.

Ko, C. Y., & Leu, F.-Y. (2021). Examining successful attributes for undergraduate students by applying machine learning techniques. *IEEE Transactions on Education, 64*(1), 50–57.

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion, 37*, 132–156.

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1).

Liu, Z., Yang, C., Rüdian, S., Liu, S., Zhao, L., & Wang, T. (2019). Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums. *Interactive Learning Environments, 27*(5–6), 598–627. https://doi.org/10.1080/10494820.2019.1610449

Liu, S., Liu, S., Liu, Z., Peng, X., & Yang, Z. (2022a). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement, Computers & Education, Volume 181. *ISSN, 104461*, 0360–1315. https://doi.org/10.1016/j.compedu.2022.104461

Liu, Z., Zhang, N., Peng, X., Liu, S., Yang, Z., Peng, J., Su, Z., & Chen, J. (2022b). Exploring the relationship between social interaction, cognitive processing and learning achievements in a MOOC discussion forum. *Journal of Educational Computing Research., 60*(1), 132–169. https://doi.org/10.1177/07356331211027300

Liu, Z., Kong, X., Liu, S., et al. (2022c). Looking at MOOC discussion data to uncover the relationship between discussion pacings, learners' cognitive presence and learning achievements. *Education and Information Technologies*. https://doi.org/10.1007/s10639-022-10943-7

Marczyk, G. R., Dematteo, D., & Festinger, D. (2005). *Essentials of research design and methodology*. John Wiley & Sons.

Michelucci, U. (2019). *Advanced applied deep learning*. Apress.

Moscoso-Zea, O., Paredes-Gualtor, J., & Lujan-Mora, S. (2018). A holistic view of data warehousing in education. *IEEE Access, 6*, 64659–64673.

Moscoso-Zea, O., & Lujan-Mora, S. (2017). Knowledge management in higher education institutions for the generation of organizational knowledge. In *2017 12th Iberian Conference on Information Systems and Technologies (CISTI).*

Moubayed, A., Injadat, M., Nassif, A., Lutfiyya, H., & Shami, A. (2018). 'E-Learning: Challenges and research opportunities using machine learning data analytics', *IEEE Access*.

Mourdi, Y., Sadgal, M., El Kabtane, H., & Berrada Fathi, W. (2019). A machine learning-based methodology to predict learners' dropout, success or failure in MOOCs. *International Journal of Web Information Systems, 15*(5), 489–509.

Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. O'reilly.

Nayak, J., Naik, B., & Behera, H. S. (2015). A comprehensive survey on support vector machine in data mining tasks: Applications & challenges. *International Journal of Database Theory and Application, 8*(1), 169–186.

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10).

Patil, A. P., Ganesan, K., & Kanavalli, A. (2018). 'Effective deep learning model to predict student grade point averages', *2017 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2017*.

Perveen, A. (2016). Synchronous and asynchronous e-language learning: A case study of virtual university of Pakistan. *Open Praxis*, [online] 8(1). Available at: https://files.eric.ed.gov/fulltext/EJ1093436.pdf [Accessed 20 November 2021].

Petrovski, A., Petruseva, S., & Zileska, P. .V. (2015). Multiple Linear regression model for predicting bidding price. *Technics Technologies Education Management, 10*(1), 386–393.

Raschka, S., & Mirjalili, V. (2017). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow* (2nd ed.). Packt Publishing.

Russell, R. (2018). *Machine learning step-by-step guide to implement machine learning algorithms with Python*. Editorial: Columbia, Sc.

Salmela-Aro, K., & Read, S. (2017). Study engagement and burnout profiles among Finnish higher education students. *Burnout Research, 7*, 21–28.

Sarker, I. H., Kayes, A. S. M., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, 6(1).

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

Shida, N., Osman, S., & Abdullah, A. H. (2019). Students' perceptions of the use of asynchronous discussion forums, quizzes, and uploaded resources. *International Journal of Recent Technology and Engineering, 8*(2S9), 704–708.

Silvola, A., Näykki, P., Kaveri, A., & Muukkonen, H. (2021). Expectations for supporting student engagement with learning analytics: An academic path perspective. *Computers & Education, 168*, 104192.

Singh, S. K. (2011). *Database systems: concepts, design and applications*. Dorling Kindersley, India.

Sorour, S., Mine, T., Goda, K., & Hirokawa, S. (2015). A predictive model to evaluate student performance. *Journal of Information Processing*, 23(2).

Sungkur, R. K., & Maharaj, M. (2022). A review of intelligent techniques for implementing SMART learning environments. In: Sikdar, B., Prasad Maity, S., Samanta, J., Roy, A. (Eds.), Proceedings of the 3rd International Conference on Communication, Devices and Computing. Lecture Notes in Electrical Engineering, vol 851. Springer, Singapore. https://doi.org/10.1007/978-981-16-9154-6_69

Sungkur, R. K., & Maharaj, M. S. (2021). Design and implementation of a SMART Learning environment for the Upskilling of Cybersecurity professionals in Mauritius. *Education and Information Technologies, 26*, 3175–3201. https://doi.org/10.1007/s10639-020-10408-9

Tarik, A., Aissa, H., & Yousef, F. (2021). Artificial intelligence and machine learning to predict student performance during the COVID-19. *Procedia Computer Science, 184*, 835–840.

Theobald, O. (2017). *Machine learning for absolute beginners: A plain english introduction* (2nd ed.). Scatterplot Press.

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, [online] 106, Available at: https://www.sciencedirect.com/science/article/pii/S1877042813046429 [Accessed: 20 November 2021].

Wabwoba, F., & Ikoha, A. (2011). Information Technology research in developing nations: Major research methods and publication outlets. *International Journal of Information and Communication Technology Research., 1*(6), 253–257.

Wibawa, A. P., Kurniawan, A. C., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes classifier for journal quartile classification. *International Journal of Recent Contributions from Engineering, Science & IT (iJES), 7*(2), 91.

Williamson, B. (2018). The hidden architecture of higher education: building a big data infrastructure for the "smarter university." *International Journal of Educational Technology in Higher Education*, 15(1).

Yadav, S. K., & Pal, S. (2012). 'Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification', 2(2), Available at: http://arxiv.org/abs/1203.3832 [Accessed 6 Nov. 2021].

Yin, X. (2021). Construction of student information management system based on data mining and clustering algorithm. *Complexity, 2021*, 1–11.

Zhang, Z. (2016). 'Introduction to machine learning: K-nearest neighbors', *Annals of Translational Medicine*, 4.