Evaluating Student Levelling Based on Machine Learning Model's Performance

Shatha Ghareeb (✉ S.R.Ghareeb@2019.ljmu.ac.uk)

Liverpool John Moores University

Abir Jaafar Hussain

Liverpool John Moores University

Dhiya Al-Jumeily

Liverpool John Moores University

Wasiq Khan

Liverpool John Moores University

Rawaa Al-Jumeily

Liverpool John Moores University

Thar Baker

University of Sharjah

Ahmed Al shammaa

University of Sharjah

Mohammed Khalaf

Al-Maarif University College

Research Article

Keywords: Machine learning, Student levelling datasets, Receiver operating characteristic (ROC), Performance evaluations, Accuracy, The Area Under Curve (AUC), school curriculum, student grade prediction, artificial intelligence

Posted Date: December 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1187994/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Evaluating Student Levelling Based on Machine Learning Model's Performance

Shatha Ghareeb¹, Abir Jaafar Hussain^{1 [2]}, Dhiya Al-Jumeily¹, Wasiq Khan¹, Rawaa Al-Jumeily¹, Thar Baker¹, Ahmed Al shammaa¹, Mohammed Khalaf¹

¹Faculty of Engineering and Technology, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK

²College of Computing and Informatics, University of Sharjah, United Arab Emirates

³Department of Computer Science, Al-Maarif University College, 31001, Iraq

⁴Electrical Engineering Department, University of Sharjah, United Arab Emirates

S.R.Ghareeb@2019.ljmu.ac.uk, A.Hussain@ljmu.ac.uk, D.Aljumeily@ljmu.ac.uk,
W.Khan@ljmu.ac.uk, R.Aljumeily@2019.ljmu.ac.uk, Tshamsa@sharjah.ac.ae,
M.i.khalaf@uoa.edu.iq, alshammaa@sharjah.ac.ae

Abstract. In this paper, a novel application of machine learning algorithms is presented for student levelling. In multicultural countries such as UAE, there are various education curriculums where the sector of private schools and quality assurance is supervising various private schools for many nationalities. As there are various education curriculums in United Arab Emirates, specifically Abu Dhabi, to meet expats' needs, there are different requirements for registration and success. In addition, there are different age groups for starting education in each curriculum. Every curriculum follows different education methods such as assessment techniques, reassessment rules, and exam boards. Currently, students who transfer to other curriculums are not correctly placed to their appropriate year group as a result of the start and end dates of each academic year as well as due to their date of birth, in which students who are either younger or older for that year group can create gaps in their learning and performance. In addition, pupils' academic journeys are not stored which create a gap for the schools to track their learning process. In this paper, we propose a computational framework applicable in multicultural countries such as United Arab Emirates in which multi-education systems are implemented. Machine Learning are used to provide the appropriate student' level aiding shoolds to provide a smooth transition when assigning students to their year groups and provide levelling and differentiation information of pupils for a smooth transition between one education curriculums to another, in which retrieval of their

progress is possible. For classification and discriminant analysis of pupils levelling, three machine learning classifiers are utilised including random forest classifier, Artificial Neural Network, and combined classifiers. The simulation results indicated that the proposed machine learning classifiers generated effective performance in terms of accuracy.

Keywords: Machine learning, Student levelling datasets, Receiver operating characteristic (ROC), Performance evaluations, Accuracy, The Area Under Curve (AUC), school curriculum, student grade prediction, artificial intelligence

1 Introduction

Pupil's levelling is a process that is based on age and lacks automation as well as a unified process amongst schools in UAE. In this case, based on the schools' curriculum, they could offer pass/fail or credit/no credit or better grades. To address the issue of family relocation, many countries such as UAE, Saudi Arabia, and Qatar offer a variety of school curriculums including British, American, Arabic to meet expat needs, yet this can create conflict when it comes to pupils levelling who are transferring curriculums. To determine student levelling, Machine Learnings (MLs) can offer an accurate levelling system for schools. Machine learning algorithms have been used in various fields such as health, financial time series prediction, genetics analysis, image analysis and online learning. Machine learning is transforming the way education is delivered and fundamentally changing the method of teaching, learning and research. Education institutes such as schools and colleges are using ML to identify struggling students earlier and develop action plans to improve success and retention [1]. ML is expanding the availability and effectiveness of online learning through localization, transcription, text-to-speech, and personalisation [2].

There are different methods of machine learning which include, supervised learning, unsupervised learning, deep learning, and reinforcement learning. Data preparation and training of a machine learning model is the foundation of the machine learning method, however, it is also vital to consider measuring the performance of the trained models. There are numerous machine learning methods used for classification which include, Random Forest Classifier (RFC), K-nearest Neighbours (KNN), Support Vector Machine (SVC), Naïve Bayes Classifier (NAIVEBC), Levenberg-Marquardt (LEVNN), and Voted Perception Classifier (VPC) [3].

This research aims to apply machine learning algorithms using datasets gathered from several education curricula to predict pupils' grades and levels when they transfer curricula [4]. To predict pupils' grades and levels, datasets with certain features have to be prepared. The datasets used in the study were classified into 3 classes based on student average level.

The remainder of this paper is organised as follows. Section II will discuss related work conducted in machine learning in education. Section II discusses the several

types of machine learning methods such as supervised, unsupervised and reinforcement learning. Section IV focuses on the method of supervised classifications, and how they have been implemented in this study. Section V mentions the model evaluation metrics that are required to quantify model performance. Section VI of this paper discusses the steps of machine learning implementation and statistical tools. Finally, the paper is summarised with section VII which discusses what we have learnt from this and what methods and techniques we can implement in the future to improve the study.

2 Related Work

In recent years, education sectors have faced many challenges to meet demands of e-Learning. The main motivation for researchers is to create a new system that is able to support education institutes in terms of student levelling and student transition between schools as well as performances [5]. There are a number of machine learning studies conducted on student levelling which is related to our research [6]. In this section we discuss influence gained from different studies and their limitations including online courses preferences, students levelling and students performances.

Masci et al. [7] developed and applied novel machine learning and statistical methods to analyse the reason for students PISA (Program for International Student Assessment) 2015 test scores in nine countries: Australia, Canada, France, Germany, Italy, Japan, Spain, UK and USA. The study was to find out which student characteristics are associated with test scores and which school characteristics are associated with school value-added (measured at school level). To address these issues, they applied a two-stage methodology using flexible tree-based methods. The researchers first run multilevel regression trees in the first stage to estimate school value-added. In the second stage, they relate the estimated school value-added to school-level variables using regression trees and boosting. Results show that while several student and school-level characteristics are significantly associated with students' achievements, there are marked differences across countries. This study has focussed on ways the school has impacted the student performance; however, it doesn't discuss the actual performance of the student just based on the input he/she is putting.

Shabandar et al. [8] researched Machine Learning approaches to predict learning outcomes in Massive Open Online Courses (MOOCs). The research is available within the area of MOOC data analysis, in particular considering the behavioural patterns of users. Based on learner behavioural patterns, two sets of features were compared in terms of their suitability for predicting the course outcome of learners participating in MOOCs. The research discovered that there is a strong correlation between click stream actions and successful learner outcomes. Various machine learning algorithms have been applied to enhance the accuracy of classifier models. Simulation results from the investigation showed that Random Forest achieved viable performance for the prediction problem, obtaining the highest performance of the models tested. This study has some relation to our study in terms of focusing on the student performance. However, the limitation in this study is that the researcher has

not put emphasis on the past student performance and applying a prediction based on that which is discussed in this paper.

Hsia et al. [9] conducted a study using data mining techniques to analyse course preference and the completion rates of enrollees in extension education courses in a university-based in Taiwan. Using their collected data, the researcher aimed to improve the target curriculum based on the student needs. The algorithms that the researcher used were Decision Tree Algorithms, Link Analysis Algorithms, and Decision Forest Algorithms. The data collected from the university was in the range of 5 academic years from the year 2000 to 2005, overall, 1408 records were collected. After testing 8 different algorithms and studying their capabilities in classification, prediction, clustering and description, the researcher selected Decision tree, Link Analysis and Decision Forest. Three separate variables were selected taking into consideration the desired outcome of the study. The variables selected were course category, completion status, and enrollee profession. Based on the results gained from the study, the Extension Education Center at CTU can plan for future courses based on the needs of the students. Masci study is closely related to our study because the researcher has used some features related to student needs such as student learning outcomes, their capabilities in different subject areas. Although the study is built around a concept that is similar to our research, yet there is a gap in the study in which a potential to be fulfilled. There's further testing that could have been implemented which would show if a student is suitable for a specific course or not based on past collected data.

Lykourantzou et al. [10] researched the dropout prediction method for e-learning courses based on three popular machine learning techniques and detailed student data. The machine learning techniques used in the study are feed-forward neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP. Since using a single algorithm may not provide accurate results, three different machine learning techniques were used to predict student dropout for e-learning courses. The method was examined in terms of overall accuracy, sensitivity and precision and its results were found to be significantly better than those reported in the relevant literature. This research provides vital information for the education sector which the dropout prediction, however, they focused on e-learning courses and not on face-to-face education in universities. Using the method conducted in this paper, the student level is predicted based on past and current exam marks in school, using those predictions we can predict if the student is going to be successful or not for different courses just based on his past performance.

In multicultural countries where many expats are found, there are several education systems available to accommodate parents and pupils needs. Switching from one system to another has become a critical issue as differences among systems could create gaps in education levels for students. Hence, it is extremely difficult to assign the right level for students when moving to a new curriculum. The challenge is to provide a smooth transition with minimum effects on pupils' performance. The age groups for each year group varies amongst curriculums. Nevertheless, there is a huge ambiguity in this area, which usually creates conflicts between parents, school, and the Ministry of Education which usually happens because different schools follow

different curricula and therefore there's no common levelling guide amongst those schools. Hence machine learning model can provide method for automating the levelling process for students when changing between education systems.

3 Related Work

This section of the study discusses machine learning algorithms and statistical tools. This section takes into consideration information about ML and the deployment of extracting useful information from the student and school data. In this section, we discuss supervised, unsupervised and reinforcement learning.

3.1 Supervised Learning

Most practical machine learning uses supervised learning. Supervised learning is where we have an input variable known as (x) and output variable (y), and then we use an algorithm to learn the mapping function from the input to the output [11].

The purpose of supervised learning is to make an approximate of the mapping function close to accurate so that when we introduce a new input data (x), we can predict the output variable (y) for that new data inserted. Supervised learning is anticipated in finding the patterns in the data which can then be applied to an analytics process [11]. The main aim of the training set is to learn from labelled instances in the training set to identify unlabelled instances during the testing phase with high potential accuracy as shown in Fig 1.

Fig. 1. Supervised learning workflow.

3.2 Unsupervised Learning

Unsupervised learning is where we only have input data (X) and without any corresponding output variables. With unsupervised learning there is no correct answer

or teacher, algorithms are left alone to discover and present the fascinating structure in the data.

Unsupervised learning can be grouped into two problematic scenarios:

- Clustering – clustering problem is where we want to the inherent groupings in the data, for example using purchasing behaviour of consumers to put them in groups.
- Association – An association rule learning problem is when we intend to find rules that describe large portions of our data.

Unsupervised learning is used when there is a use of large amounts of unlabeled data. An example of large amounts of unlabeled data is social media applications such as Twitter, Instagram and Snapchat. For us to understand the meaning behind this unlabeled data, we require an algorithm that classifies that data based on the patterns or clusters the system finds. Cluster analysis is the most common method in unsupervised learning that is used for exploratory analysis to find groupings or hidden patterns in datasets [12]. The main purpose of applying this kind of method is to discover the smallest group feature subset (clustering) from the datasets according to the chosen criteria [12]. Unsupervised learning conducts a repetitive process of analysing data without the use of human interaction. Unsupervised learning is used in email spam-detecting technology. For analysts to tag unsolicited bulk emails will be complicated because there are too many variables in legitimate and spam emails, therefore, with the use of applied clustering and association, machine learning is able to identify unwanted emails. Fig 2 shows the unsupervised learning workflow.

Fig. 2. Unsupervised learning workflow.

3.3 Reinforcement Learning

Reinforcement learning (RL) is a behavioural learning model where the system learns in an interactive environment by the use of trial and error [13]. The algorithm uses feedback from own actions and experiences, in other words receiving feedback from

the data analysis to guide the user to the best outcome. As the system learns through trial and error, having a sequence of successful decisions will help the process be reinforced [13].

4 Classification Algorithms

This research uses supervised classification because the datasets collected from the schools have been identified with relevant labels. Regression models are aimed at mapping instance (input) values to continue outcome values, while classification procedures aim to map instances (input) into discrete classes. For example, some studies aim to classify students that will drop out of their exams or otherwise won't. Within classification, the aim is to learn a decision-making platform that can correctly map an instance (input) space to an output. Within the education sector, ML researchers have researched ways to implement algorithms such as RFC, SVM, and ANN to predict student performance and requirements. The classification of student and education data has shown positive outcomes for student benefits and education institutes.

In terms of the classification process, object (x) is the input that has a set of features, while (y) is the class label that's assigned alongside x . The classification model is implemented to predict the class label for new samples. Classification techniques in education are important to improve student levelling and enhance education institute decision making in terms of grading and student abilities. There are various methods implemented for classification which are grouped into two, linear and nonlinear classifiers. As represented in Equation 1, the linear classifier is represented as a linear function (g) of the input (x). From the equation, (w) represents a set of weights, while (T) is the metrics response and (b) refers to the bias. Having two sets of classes k_1 and k_2 , therefore input vector x gets assigned to k_1 when $(x) \geq 0$. The boundary among k_1 and k_2 is linear.

$$g(x) = w^T x + b \quad (1)$$

4.1 Random Forest Classifier (RFC)

Random forest classifier (RFC) is one of the most successful ensemble learning techniques which has been proven to be an effective technique in pattern recognition and ML for high-dimensional classification and skewed problems. RFC does have a drawback which is tree classifiers have high variance. In practice, it is common for a slight change in the training dataset to be in a different tree. The reason behind this is the hierarchical nature of the tree classifiers.

Equations 2 and 3 describe the method of RFC.

$$f(x) = \frac{1}{m} \sum_{i=1}^m f(x, x_i p) \quad (2)$$

In equation 2, the variable that has a partial dependence is referred to as x and $x_i p$ takes into consideration any other variable, while other variables in the equation refer to feature from student data.

$$f(x) = \log t_j - \frac{1}{J} \sum_{k=1}^J (\log t_k(y)) \quad (3)$$

The number of classes is referred to as J in the equation while j is the class. The proportion for the total votes for the class is t_k .

Training of the RFC can be done by developing several B trees. The training sets for the classifier is $X = x_1 \dots x_n$ and the classes are $Y = y_1 \dots y_n$. When replacing B in $b = 1, \dots, B$: the training samples from X , Y have its place of B , which is then referred to as X_b , Y_b , where Y comes from the predicted class. To implement the classifier and select numerous training data sets, $M = \{(X_1, (X_n) \dots, (Y_1, Y_n))$, where $X_i, i = 1 \dots n$ is the descriptors vectors and depending on the classification outcome, Y_i can be referred to as the activity of interest or corresponding label.

4.2 Artificial Neural Network (ANN)

Artificial neural network (ANN) is an established numerical computerized system that can model complex relationships between random experimental inputs and their relative outputs[14], [15]. ANN works according to biological neural networks; biological neural networks are constructed from small parts that are called neurons and electrochemical signals are received through synapses to the neuron. Each synapse adds a special weight to the received signal and modifies weights during learning. A group of neurons form subsystem and subsystem collection form the brain. ANN consists of neurons, and input data to each neuron are weighted and these weights are corrected during learning. It can be said that ANN works like a brain, ANN layers work as subsystems and neurons have the same role at two networks. ANN generally contains three layers. The input layer that input data are entered into the network, hidden layer that connects input layer to the output layer, and an output layer that output data are recognized to the network[16], [17]. Between different types of ANN models, multilayer perceptron (MLP) neural networks are more common and used for nonlinear functions. MLP neural network is a feed-forward ANN and formed from an input layer, an output layer and some hidden layers. The activation function in the output layer at MLP ANN must be linear and hidden layers can be linear or nonlinear. A sigmoid function, Tansig or Logsig, is mostly used as a nonlinear function in MLP ANN models.

4.3 Combined Classifiers

Machine learning uses a training set combined with building a classifier that provides a reliable classification. This study used the multi-class classification problem where many classes are available in the datasets. This research combines multiple classifiers to improve the classification accuracy and performance in comparison to the single model. Study shows that combining different classifiers can generate better output [18]. The total information from multiple models is combined to generate better decisions. Using the bootstrap method, the training set is delivered to each model. Each model generates an outcome using the performance metrics method. To discover the classifiers that generate the highest performance, voting was used to select the classifier with the highest accuracy and performance. This study has taken the approach of combining the final classification results gained using N different feature sets $(f_1^1(x), \dots, f_N^N(x))$. To construct both classifiers, the training models need to be trained with the use of different feature sets. Where x refers to particular input and each model m^n generates its own output $y^n = (y^n(1), \dots, y^n(Z))$, where z is considered the class label, while $y^n(m)$ corresponds to the probability of c^n . Each classifier i generates L approximations to the probabilities $f_j^N(x), j = 1, \dots, L$. The purpose of combining multiple algorithms is to create better results. The research used the stacked and voting method. The stacked method involves a set of models that lead to the same space for them to be combined. Each classifier receives training from the same training set which receives 70% of the datasets, the validation receives 10%, while testing receives 20%.

5 Model Evaluation Metrics

Model evaluation metrics are required to quantify model performance. The choice of our evaluation was dependent on the task our model is required to perform.

5.1 Predictive Models

When discussing about predictive models, two models which are regression model (continuous output) and classification model (nominal or binary output) are considered. The evaluation metrics that are used in each model varies.

In our case with regards to classification, two types of algorithms can be used:

- Class output – includes algorithms like SVM and KNN
- Probability output – includes algorithms like Logistics Regression, Random Forest etc.
- **Confusion Matrix** involves:

True positives (TP) – Predicted positive and are positive

False positives (FP) – Predicted positive and are negative

True negatives (TN) – Predicted negative and are negative

False negatives (FN) – Predicted negative and are positive

- **Sensitivity** is the percentage of positive instances out of the *total actual positive* instances.

$$\frac{TP}{TP + FN}$$

- **Specificity** is the percentage of negative instances out of the *total actual negative* instances.

$$\frac{TN}{TN + FP}$$

- **Precision** is the percentage of positive instances out of the *total predicted positive* instances.

$$\frac{TP}{TP + FP}$$

- **F-Measure** is a mean of precision and recall, therefore the higher F1 score received the better the accuracy

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall}$$

- **Accuracy** is the most commonly used metric to judge a model, however, accuracy is not a clear indicator of the performance of the model, specifically when the classes are imbalanced.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

- The **Receiver Operator Characteristics (ROC)** curve is an evaluation metric for binary classification problems. It is the probability curve that plots the TPR against FPR at various threshold values and essentially separates the signal from the noise.
- The **area under ROC Curve** is a performance metric for measuring the ability of a binary classifier to discriminate between positive and negative classes.

6 Machine Learning Implementation and Statistical Tools

Machine learning models have been utilised to classify student academic levels and e-learning. Using machine learning models, schools can predict students' current levels and predict their academic path in the future. The main benefit of this model is to make use of recent technological development in ML algorithms to assist schools and teachers to develop their students more effectively with the use of past records. The

aim is to propose the prediction of the student level using classification methods based on the data sets collected previously from schools.

The proposed implementation comprises of many steps; data collection, data pre-processing, and then the data differentiates into three parts which are, building the model based on the training data, evaluating the model based on the testing sets, and then selecting the relevant model. The data collection process starts when the student is first registered into the school. To apply the ML model, the collected data must be cleaned (removing unwanted data and filling missing data), thereafter, various ML models are selected to evaluate the data sets. The holdout method is used to split the datasets into training, validation, and testing as shown in Fig 3.

Fig. 3. The methodology process.

6.1 Data Collection

The dataset presented comprising students' records for two academic years that included math, English, and science for 3 terms [4]. The selection of subject areas and some terms were based on influence from other researchers in a similar subject matter [8], [19]. The dataset comprises novel aspects specifically, in terms of student grading in diverse educational cultures within multiple countries – Researchers and other education sectors will be able to see the impact of having varied curriculums in a country. Furthermore, Dataset compares different levelling cases when students transfer from curriculum to curriculum and the unreliable levelling criteria currently set by the international schools. Fig 4 demonstrates the data collection process used in this study.

Fig. 4. Data collection process.

Table 1. Sample of student dataset [4].

Type	Number	Description
Records	1550	Data collection for 2 academic years
Classes	3	Class 1 is $85\% \leq 100\%$
		Class 2 is $75\% \leq 84.99\%$
Attributes	25	Class 3 is $< 75\%$
		Gender, age, proposed year/grade, year of admission, etc.

Table 2. Collected data description.

Input Attributes		
Attribute Name	Value	Description
Gender	Male/Female represented by 0 and 1	Gender of the student
Student Age (As of 2017/18)	6,7,8,9,.....etc.	Age of the student calculated from 2017 / 18 academic year
Age Group	-1, 0, 1	Based on the age of the student as per 2017; -1 group means 1 year below accepted age; 0 group is on the accepted age; and 1 group is 1 year above accepted age.
Year of Admission	Old BBS Student, Old GEMS Student, New Admission 18/19	The data collected is for two or more academic years; 2017/18 and before academic years + 2018/19 academic year.
Current Year (17/18)	FS1, FS2, Y1-12 / Grade 1-11	This is the year or grade group assigned to the student by the school
Proposed Year/Grade (18/19)	FS1, FS2, Y1-13 / Grade 1-12	This is the year or grade group assigned to the student by the school
Year of Admission	Old BBS Student, Old GEMS Student, New Admission 18/19	The data collected is for two or more academic years; 2017/18 and before academic years + 2018/19 academic year.
Previous School (17/18)	Many schools in UAE	Previous schools that the student was in before this study
Previous Curriculum	British / American / MOE / Canadian / Indian / Australian / CBSE / German	The curriculum the student transferred from.
Current School	GEMS, Belvedere	Name of the school that the data has been collected.
Current Curriculum	0, 1	The curriculum the student transferred to, 0 = American, 1 = British
Math-exam	Mark out of 100	Exam marks for school entry exam in math

Science-exam	Mark out of 100	Exam marks for school entry exam science
English-exam	Mark out of 100	Exam marks for school entry exam English
Math19-1	Percentage out of 100%	Term 1 student Maths Exam marks during academic year 2018/19
Science19-1	Percentage out of 100%	Term 1 student science Exam marks during academic year 2018/19
English19-1	Percentage out of 100%	Term 1 student English Exam marks during academic year 2018/19
Math19-2	Percentage out of 100%	Term 2 student Maths Exam marks during academic year 2018/19
Science19-2	Percentage out of 100%	Term 2 student science Exam marks during academic year 2018/19
English19-2	Percentage out of 100%	Term 2 student English Exam marks during academic year 2018/19
Math19-3	Percentage out of 100%	Term 3 student Maths Exam marks during academic year 2018/19
Science19-3	Percentage out of 100%	Term 3 student science Exam marks during academic year 2018/19
English19-3	Percentage out of 100%	Term 3 student English Exam marks during academic year 2018/19
Math20-1	Percentage out of 100%	Term 1 student Maths Exam marks during academic year 2019/20
Science20-1	Percentage out of 100%	Term 1 student science Exam marks during academic year 2019/20
English20-1	Percentage out of 100%	Term 1 student English Exam marks during academic year 2019/20
Math20-2	Percentage out of 100%	Term 2 student Maths Exam marks during academic year 2019/20
Science20-2	Percentage out of 100%	Term 2 student science Exam marks during

English20-2	Percentage out of 100%	academic year 2019/20 Term 2 student English Exam marks during academic year 2019/20
Math20-3	Percentage out of 100%	Term 3 student Maths Exam marks during academic year 2019/20
Science20-3	Percentage out of 100%	Term 3 student science Exam marks during academic year 2019/20
English20-3	Percentage out of 100%	Term 3 student English Exam marks during academic year 2019/20
<hr/> Output Attributes (Only class as it is a result of a formula applied to the average attributes) <hr/>		
Average 19/20	Percentage out of 100%	The average of 2019/2020 academic year extracted from the marks of the 3 subjects among the 3 terms
Class	1, 2, and 3	3 classes extracted from the Average value; class 1 from 85% to 100%; class 2 from 75% to 84.99%; and class 3 is below 75%

6.2 A Subsection Sample

Data preparation is an essential part when handling data which helps to check for correctness, meaningfulness, and the security of data to be used. Therefore, the data was prepared by unifying the outputs, filling in missing data, eliminating non-related features, and preparing headers. The data was cleaned using excel filtering, equalising the instances according to their categories. The date of birth for the students was converted to age for categorisation as per 2017. New attributes of age groups were created according to the age of the student and the standard age of the class into three categories as explained in Table 2. A new attribute is then created which is the average marks for the 2019/2020 academic year for three subjects over 3 terms. Then an additional attribute is created called class, which divides the average marks of the students for the academic year 2019/2020 for 3 subjects over 3 terms. Three classes are created, Class 1 is the average marks from 85% to 100%, class 2 is the average marks from 75% to 84.99%, and class 3 is the average marks below 75%. Any data that was deemed confidential is removed such as student name and ID. The selection

of a suitable classifier involves trial-and-error processes, but the use of statistical validation can guide that process [20].

6.3 A Subsection Sample

Feature selection is a method used in data mining and machine learning [21]. The purpose of applying feature selection is to select a subset by disregarding irrelevant features and unwanted information from the student levelling dataset. The technique of feature extraction can help in generating accurate results and removing any negative influence on the learning models. For example, in student levelling datasets, teachers usually ignore features such as exam board or gender because those features won't affect the results significantly. By using this technique in this study, we have reduced the number of unnecessary features before implementing the datasets. Extracting irrelevant features helped improve the performance of the models and improved the results generated from the datasets. Over-fitting can harm the performance of the models; therefore, this technique can reduce those risks. Feature selection decreases the search space determined throughout all the features, and therefore the models can process the data faster with less memory consumption [22], [23]. The irrelevant and redundant features can create confusion on the learning when dealing with a small number of training examples which will lead to overfitting and high dimensionality.

The high dimensionality of the extracted features will be reduced with the use of feature selection methods. This method is achieved by identifying spaces that have lower dimensions compared to those of the real data. The method of feature selection is divided into two forms:

- Feature transformations – dealing with lower dimensional space such as independent component analysis (ICA) and principal component analysis (PCA).
- Selecting the number of features for a given pattern which is based on the mean or standard deviation of the feature values.

Feature selection can reduce both the data and the computational complexity. In this study, various feature selection methods are utilised including Information Gain (IG), Symmetric Uncertainty (SU) and Relief-F. IG is an essential measure that can be used for ranking features. Given that entropy is a criterion of impurity in a training set S , a measuring defining further information about Y given X which represents the amount of decrease in entropy of Y . This measure of feature selection is known as IG which is given by equation 4.

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \quad (4)$$

IG is a symmetrical measure. The information gained about Y after observing X is equal to the information gained about X after observing Y . A weakness of the IG criterion is that it is biased in favour of features with more values even when they are not more informative [24]. SU is one of the most effective feature selection methods.

SU is a correlation measure between the features and the class which is given by the formula 5.

$$(5) \quad SU = \frac{(H(X)+H(Y)-H(X\backslash Y))}{(H(X)+H(Y))}$$

$H(X)$ and $H(Y)$ are the entropies based on the probability associated with each feature and class value respectively and $H(X, Y)$, the joint probabilities of all combinations of values of X and Y [24].

The feature selection of Relief-F is to have instances at random and compute their nearest neighbours and adjust the feature weighting depending on the features they represent the instances from neighbours of different classes. Relief-F works on discovering a high probability to assign as the weight of each feature f [25].

The performance of the proposed algorithms is evaluated using different benchmark datasets. The proposed algorithm is compared with the well-known feature selection methods; Info-Gain, Relief-F and Symmetrical Uncertainty in terms of the number of features reduced and learning accuracies as shown in fig 5 [26].

The feature selection method is applied to 25 different features that are believed to influence the outcome of the level of the student. These feature subsets assist teachers and schools to accurately predicting the level of the student in the future. The feature subsets assist schools and teachers to make a more precise prediction on the level of the students by removing irrelevant features, making the process more efficient.

Fig. 5. Feature selection method [26].

6.4 Experimental Procedure

The experimental part of the research covers the design of the test environment plus the formation of each model. In this research, the holdout method is used to evaluate how statistical analysis can transform into a dataset. The datasets are split into training, validation, and testing sets. Using this method allows us to find the average percentage of correct and incorrect classifications, therefore the training set receives 70%, validation 10%, and testing is 20%.

The algorithms used in this study is comprised of Random Forest Classifier (RFC), Artificial Neural Network (ANN), and combined classifiers. The main purpose of selecting those models is because they can deal with high-performance classifiers whilst also dealing with strong non-linear classifiers.

To obtain performance results for each algorithm used, the simulation was repeated 50 times and then the mean was calculated. A description of the models used in this study is described in Table 3.

Table 3. Description of the used models.

Model	Description	Architecture	Training Algorithm	Parameters	Role
RFC	Random forest	25 inputs, 400, 4 outputs	Random feature bagging	Number of generated decision trees 400;	Non-linear comparison model
BPXNC	The feed-forward neural network algorithm	Units: one context unit for each output unit	Trained with the use of mapping a set of input data to generate a computational modification for the whole weights	Size of feature: 1 Momentum coefficients between 0.01 and 1.0. Sigmoid function $f(x) = 1/(1 + e^{-x})$. Learning rate between 0.25 and 0.9. Performance: 0.0932	Non-linear comparison model

LEVNN	Multilayer perception, Levenberg- Marquardt algorithm	Units 25-2-4 tansig activations	Levenberg- Marquardt	Initialisation: Nguyen Widrow Adaptive learning rate settings: initial value: 0.001 coefficient for increasing LR: 10 coefficients for decreasing LR: 0.1 maximum learning rate: 1e10	Non-linear comparison model
-------	---	------------------------------------	-------------------------	---	-----------------------------------

6.5 Random Forest Classifier (RFC)

The performance evaluation techniques were achieved by using the collected student levelling dataset which had 1550 samples. The imperial study in decision trees was performed using random forest models. The classification performance was evaluated using the evaluation metrics which were discussed in previous chapters. During simulation, both training and testing datasets were selected randomly whilst also being repeated in every test run.

The results gained from our experiments show results gained from the training of the classifiers with feature selection which is shown in Table 4. The training of our proposed method is also evaluated in the form of visual performance evaluation with ROC charts as shown in Figure 6, and AUC as shown in Figure 7. The AUC is a performance metric for measuring the ability of a binary classifier to discriminate between positive and negative classes of the trained model. In our RFC model, we have used a different number of trees to evaluate the performance evaluation metrics and accuracy, which were 50, 100, 200, 400, and 500 trees. The RFC model was built and trained using multiple trees, trained using the student levelling dataset. During the training process, RFC with 100 trees generated the best accuracy with an average of 0.69 and an AUC average of 0.68. Although the AUC average was higher with 50 trees, it had lower accuracy. After running the training simulations, the sensitivity with RFC 400 trees performed better than with other trees with an average of 0.661.

Table 4. RFC performance with average of 3 classes (Training).

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC50/1	0.632	0.674	0.466	0.473	0.306	0.66	0.72
RFC100/2	0.592	0.673	0.434	0.460	0.265	0.69	0.68

RFC200/3	0.661	0.617	0.416	0.443	0.278	0.63	0.69
RFC400/4	0.524	0.657	0.405	0.416	0.181	0.65	0.61
RFC500/5	0.579	0.585	0.371	0.404	0.163	0.61	0.61

Fig. 6. ROC curve (training) for RFC.

Random forest classification combines the basics of decision trees with additional flexible parameters which increase the accuracy of the model. As stated previously, a new dataset that is similar in size to the original is created alongside bootstrap. The method of bootstrapping allows us to select the most important samples multiple times. Once the bootstrapped datasets are created, a random subset of variables is used to develop the random forest. For every single step, 25 attributes and 3 classes from the levelling of students are considered. Having taken into consideration the subset of variables for each step, a new bootstrapped dataset is developed alongside several RF trees, and this process was completed and repeated several times. Once the student data is ingested into all the trees in RF, a calculation is done to discover which model received the highest number of votes.

The bagging process takes several bootstrap samples from the original randomly generated datasets to construct an additional dataset. Making slight adjustments in the training or testing phase of RF can influence the accuracy and performance of the model, and this is achieved by bagging. The out-of-bag samples determine the label with the most votes. The out-of-bag samples with 3 classes in our student levelling datasets are accurately labelled by the RFC. With the method of bagging, we can estimate the accuracy of the RFC by finding the proportion of out-of-bag samples that were correctly classified by the RF model. The out-of-bag error is the proportion of out-of-bag samples incorrectly classified.

Fig. 7. AUC (training) for RFC

The RFC models are found to have similar performance; however, RFC/50 had the highest average in terms of specificity (0.674), precision (0.466), F1 score (0.4373), j score (0.306), and AUC (0.72). The AUC for RFC/50 averaged at 0.72 whilst also RFC/100 and RFC/200 had similar AUC averages of 0.68 and 0.69. The AUC for RFC/400 and RFC/500 was slightly lower than the other models. On further study of the results from all the models, the sensitivity from RFC/50 and RFC/200 fluctuated from the other models. It was discovered that although the RFC models generated reasonable AUC and accuracy values with 3 classes during the training phase, the number of trees may have had a slight influence on the total performance of the model with 3 classes.

During the testing phase, the performance of different RFC models with a different number of trees is investigated using multiple classification techniques. The testing sets results for the student levelling datasets are shown in table 5. The RFC/50 generated the highest AUC with an average of 0.587, whereas RFC/100 generated the highest accuracy of 0.634. On the other hand, RFC/200 generated the lowest AUC compared to all the other RFC models. Compared with the other tested machine learning models, RFC didn't generate the highest accuracy. RFC/50 generated the highest sensitivity of 0.709, whilst RFC/500 was the second-highest in terms of sensitivity. In terms of specificity, RFC/50 had the second-lowest specificity of 0.483, whilst RFC/100 generated the highest specificity of 0.656. Figure 8 shows a visual representation of the ROC curve for testing, and Figure 9 shows the AUC for testing of RFC. RFC/50 had the best performance in terms of ROC during training and testing.

Table 5. RFC average with 3 classes (testing)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
-------	-------------	-------------	-----------	----	---	----------	-----

RFC50/1	0.709	0.483	0.417	0.439	0.192	0.505	0.587
RFC100/2	0.493	0.656	0.388	0.393	0.149	0.634	0.554
RFC200/3	0.422	0.608	0.393	0.379	0.196	0.573	0.463
RFC400/4	0.508	0.55	0.368	0.381	0.03	0.556	0.527
RFC500/5	0.621	0.446	0.35	0.395	-0.52	0.515	0.531

Fig. 8. ROC curve (testing) for RFC

Fig. 9. AUC (testing) for RFC

Studying the received data from our models, the chosen student datasets for our experiment demonstrates non-linear relationships, which represents a challenge to the classifiers. The RFC classifiers performance is poor compared to the other classifiers, showing that RFC does not have the capabilities of handling the training data and unseen examples. It was found that the performance at the chosen rejection threshold

varied between the training and testing sets for classes 1 and 2 as shown in the AUC values. However, class 3 showed some performance similarities between training and testing, however, this was not consistent with all the evaluation factors. The performance varied amongst the models, this could be due to the classification settings, or the number of classes could have played a role in that. The number of features affects the evaluation of the RFC model, we have used 25 features, however, initially more than 25 features were proposed but 25 features were found to be the most crucial for teachers and schools to evaluate the student's level.

6.6 Artificial Neural Network (ANN)

Artificial neural network can perform multiple tasks such as classification, clustering, dimensionality reduction, and student datasets.

The evaluation of the model is achieved using the holdout method which is implemented on 1550 samples. The model received three sets of samples, which was 70% for training, 20% for testing, 10% for validation. The proposed research utilised two methods of NN models which included Levenberg Neural Network (LEVNN) and Backpropagation Network Classifier (BPXNC). The ROC curve and AUC for LEVNN and BPXNC are shown in Figures 10 and 11 for training, respectively. Although LEVNN obtained a higher accuracy average during training than BPXNC, they both generated the same AUC value of 0.962 which are good results when compared to the other models tested in this study as shown in Table 6. Overall LEVNN generated the best results compared to BPXNC which was since LEVNN did not use the entire training samples provided.

Table 6. ANN performance (training)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN	0.952	0.929	0.697	0.73	0.881	0.936	0.962
BPXNC	0.935	0.92	0.692	0.718	0.855	0.925	0.962

Fig. 10. ROC curve (training) for LEVNN and BPXNC

Fig. 11. AUC (training) for LEVNN and BPXNC

Although we have received good results during training for both LEVNN and BPXNC, the testing results performance for both NN models were low. For LEVNN, the hidden layers are amended using backpropagation links, whereas BPXNC were altered using feedback from the output layer. Table 7 shows the classification performance evaluation for LEVNN and BPXNC during testing. Figure 12 shows the ROC curves for LEVNN and BPXNC which is evaluated by calculating the outcomes based on the true positive rates and false-positive rates which is between 0 and 1. Figure 13 shows the AUC with an average of 3 classes for LEVNN and BPXNC.

Table 7. ANN performance (testing).

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN	0.801	0.815	0.568	0.602	0.616	0.816	0.732

BPXNC	0.792	0.789	0.571	0.577	0.581	0.774	0.767
-------	-------	-------	-------	-------	-------	-------	-------

Fig. 12. ROC curve for ANN (testing)

Fig. 13. AUC plot for ANN (testing)

6.7 Combined Classifiers

The combined classifier is a method that combines two or more classifiers to improve the performance and accuracy of the models. Combined classifiers were rarely utilised in other research on machine learning in education. This model of machine learning algorithms is found to generate great results when combining certain classifiers. Implementing combined classification in our study, the models

were able to learn non-linear components and generated accurate results combined with the other classifiers that have been used. To enhance the accuracy and performance of selected models, combined classifiers use a pattern recognition system combined with a bootstrap aggregating approach.

The proposed study implemented several classifiers combined which included NN with many features, LEVNN with a various number of features, and NN combined with RF. The classification performance evaluation method is done based on student levelling datasets with 25 features with 3 classes. Table 7 shows the outcome for combined classifiers in training. The prediction of the student levelling dataset during the training phase is achieved by taking the majority vote for RFC for several cycles. Alternatively, NN generated the predictions after several cycles by taking a weighted vote, plus having the proportional weights to the model's performance and accuracy.

In our simulations, the overall AUC for all the combined classifiers was 1 for training as shown in table 8, which is impressive results when comparing them to ANN and RFC classifiers. Although the precision for NN Com and LEVNN Com was 1 compared to 0.999 for NN combined with RF, yet still the overall performance for NN combined with RF was performing better. A strong combination between NN and RF is highlighted, which indicates that there's vital information is available in our selected datasets. NN Com and LEVNN Com were both capable of handling the datasets showing accuracy and AUC of 1 as shown in Figure 15.

Table 8. Combined classifiers (training)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
NN Com	0.997	0.995	1	0.997	0.992	1	1
LEVNN Com	0.997	0.995	1	0.997	0.992	1	1
NN and RF Com 1	0.999	0.999	0.999	0.999	0.998	0.999	1
NN and RF Com 2	0.999	0.999	0.999	0.999	0.998	0.999	1

Fig. 14. ROC curve for the combined classifier (training)

Fig. 15. AUC plot for the combined classifier (training)

During testing of combined classifiers, the performance of the algorithm was lower than that of training as shown in table 9. NN and RF obtained a sensitivity of 0.755, specificity 0.798, precision 0.565, F1 0.585, J score 0.552, accuracy 0.799, and AUC 0.824. We have stated previously that NN combined with RF performed excellently during training generating the highest performance, yet still, during testing, it didn't perform as well but it generated the highest AUC as shown in Figure 16. As shown in Figure 17, the ROC generated for classes 2 and 3 were performing much better than class 1 compared between a true positive rate of 100% and a false positive rate of 0%.

Table 9. Combined classifiers (testing)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
NN Com	0.769	0.76	0.538	0.549	0.529	0.748	0.712
LEVNN Com	0.773	0.656	0.528	0.533	0.429	0.649	0.663
NN and RF Com 1	0.755	0.798	0.565	0.585	0.552	0.799	0.824
NN and RF Com 2	0.776	0.722	0.542	0.547	0.498	0.712	0.787

Fig. 16. AUC for the combined classifier (testing) **Fig. 17.** ROC curve for the combined classifier (testing)

7 Conclusion and Future Work

In this study, a machine learning approach is used that utilised 25 features from 1550 samples for student-level prediction. The reason that combined classifiers performed well for such data is due to the training conducted and testing. The AUC for the best performing combined classifier (NN Com, LEVNN Com, NN and RF Com) produced 1 for the training sets as shown in Table 8, while testing generated an AUC of 0.824 as shown in Table 9. Overall, the type of results gained highlight the importance and potential of student data for student levelling. It has been proven that the choice of model is essential when needing great results, as shown in the classification performance tables for training and testing. Combined classification responded effectively to the student data and is therefore found to have potential in the education sector.

The performance evaluations are performed on data that don't have standard distributions. NN combined with RF is a powerful model for our data as it offered strong prediction accuracy and performance compared to other classifiers as shown in Figure 5.13. The datasets had 70% selected for testing, 10% for validation and 20% for testing. The process was repeated several times. In general, RFC was able to handle to simultaneously handle multiple descriptors much more efficiently compared to other classifiers. RFC handled redundant and irrelevant descriptors and dealt with regression and classification. The main reason that NN combined with RF generated the highest performance is due to the model not having to deal with over-fit and operated independently without guidance. The results received from the simulations show that the selected dataset presents a non-linear relationship which was challenging for several machine learning models.

Latest studies revealed that machine learning is found to be effective for pre-processing of student datasets to predict student suitability for certain careers and their capacity to learn in certain subjects. Our data can be analysed effectively with machine learning models when introduced to the classification of student past levelling marks and generate the student level in future. For future improvements, we need to consider using other machine learning techniques and increase the number of features used such as learning style of students, their motivation and interest, concentration level, family background, personality type, information processing ability and the way they attempt exams, however, the limitation is that the study has not considered students' performance in exams in their previous and current academic stage which can be an improvement to our study.

8 Data Availability

The data set generated during and analysed during the current study are available in the Mendeley Data repository, <https://data.mendeley.com/datasets/3g8dtwbjyy/1> (DOI: 10.17632/3g8dtwbjyy.1)

References

- [1] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using machine learning techniques," *Educ. Sci.*, vol. 11, no. 9, 2021.
- [2] M. W. Lee, S. Y. Chen, K. Chrysostomou, and X. Liu, "Mining students' behavior in web-based learning programs," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3459–3464, 2009.
- [3] A. L. and M. Wiener, "Classification and Regression by randomForest. R News 2," vol. 3, no. December 2002, pp. 18–22, 2003.
- [4] S. Ghareeb, A. Hussain, W. Khan, D. Al-Jumeily, T. Baker, and R. Al-Jumeily, "Dataset of student level prediction in UAE," *Data Br.*, vol. 35, p. 106908, 2021.
- [5] D. Kućak, V. Juričić, and G. Đambić, "Machine learning in education - A survey of current research trends," *Ann. DAAAM Proc. Int. DAAAM Symp.*, vol. 29, no. 1, pp. 0406–0410, 2018.
- [6] A. S. Ghareeb, R. Al-jumeily, and T. Baker, "A Machine Learning Based Framework for Education Levelling in Multicultural Countries : UAE as a Case Study," vol. 14, no. 3, p. 2020, 2020.
- [7] C. Masci, G. Johnes, and T. Agasisti, "Student and school performance across countries: A machine learning approach," *European Journal of Operational Research*, vol. 269, no. 3, pp. 1072–1085, 2018.
- [8] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, and N. Radi, "Machine learning approaches to predict learning outcomes in Massive open online courses," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 713–720, 2017.
- [9] T. C. Hsia, A. J. Shie, and L. C. Chen, "Course planning of extension education to meet market demand by using data mining techniques - an example of Chinkuo technology university in Taiwan," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 596–602, 2008.
- [10] I. Lykourantzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Comput. Educ.*, vol. 53, no. 3, pp. 950–965, 2009.
- [11] V. Nath and S. E. Levinson, *Machine learning*, no. 9783319056050. 2014.
- [12] J. C. Tong, "Cross-Validation," *Encycl. Syst. Biol.*, pp. 508–508, 2013.
- [13] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 4th ed. 2012.
- [14] A. N. Celik, "A techno-economic analysis of wind energy in Southern Turkey," *Int. J. Green Energy*, vol. 4, no. 3, pp. 233–247, 2007.
- [15] X. Zhuang, W. Zhang, Y. Wu, and Z. Zhao, "Comprehensive prediction method for die-roll height of fine-blanking components," *Int. J. Adv. Manuf. Technol.*, vol. 98, no. 9–12, pp. 2819–2829, 2018.
- [16] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharm. Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, 2000.

- [17] R. Haghbakhsh, H. Adib, P. Keshavarz, M. Koolivand, and S. Keshtkari, "Development of an artificial neural network model for the prediction of hydrocarbon density at high-pressure, high-temperature conditions," *Thermochim. Acta*, vol. 551, pp. 124–130, 2013.
- [18] A. Jain and S. Solanki, "An Efficient Approach for Multiclass Student Performance Prediction based upon Machine Learning," no. Icces, pp. 1457–1462, 2019.
- [19] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades," *Artif. Intell. Rev.*, vol. 37, no. 4, pp. 331–344, 2012.
- [20] M. Khalaf *et al.*, "Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models," *Neurocomputing*, vol. 228, pp. 154–164, 2017.
- [21] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.
- [22] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," no. April, 1999.
- [23] H. Vafaie and K. De Jong, "Genetic algorithms as a tool for feature selection in machine learning," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 1992-Novem, no. January 2013, pp. 200–203, 1992.
- [24] Z. Karimi, M. Mansour Riahi Kashani, and A. Harounabadi, "Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods," *Int. J. Comput. Appl.*, vol. 78, no. 4, pp. 21–27, 2013.
- [25] M. ROBNIK SIKONJA MarkoRobnik and friuni-ljsi IGOR KONONENKO IgorKononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, pp. 23–69, 2003.
- [26] Z. Cömert and A. F. Kocamaz, "A study of artificial neural network training algorithms for classification of cardiotocography signals," *J. Sci. Technol.*, vol. 7, no. 2, pp. 93–103, 2017.