Journal of KIIT. Vol. 20, No. 11, pp. 9-21, Nov. 30, 2022. pISSN 1598-8619, eISSN 2093-7571 9
http://dx.doi.org/10.14801/jkiit.2022.20.11.09

Comparative Study on Predicting Student Grades using Five Machine Learning Algorithms

HeeJeong Jasmine Lee*

Abstract

Accurate prediction of student academic performance is important because it provides early identification of the at-risk students who may need extra help to prevent them from being discouraged or dropping out of the education program. The objective of this study is to compare several machine learning(ML) classification algorithms on three student performance data sets of different sizes(i.e. small, medium and large-sized datasets) and structures to identify which algorithms are best able to generalize across different data sets in this field and provide reliable predictions of academic achievement. The three different datasets were: HarvardX-MIT, Open University Learning Analytics, and xAPI-Educational Mining dataset. A comparison of ML model performance metrics revealed that random forest tended to score highly across the tested data sets and metrics. This finding suggests that the random forest algorithm represents a useful ML tool for predicting student academic performance.

요 약

학생 학업 성취도에 대한 정확한 예측은 과목에서 낙제할 학생을 조기에 식별할 수 있으므로 중요하다. 이러한 예측은 학생, 교육자 또는 학교가 학생 유지와 학생의 학업 성취도를 높이려 하는 목표에 도움이 되는 중요한 정보를 제공할 수 있다. 본 연구의 목적은 크기와 구조(예: 소형, 중형 및 대형)가 다른 세 가지 학업 성취도 데이터 세트에 여러 머신러닝 분류 알고리즘을 이용하여 비교 분석하여 어떤 알고리즘이 학업 성취도 예측에 신뢰할 수 있는 예측을 제공하여 사용할 수 있는지에 대한 비교 연구결과를 제공하는 것이다. 세 가지 다른 데이터 세트는 HarvardX-MIT, Open University Learning Analytics 및 xAPI-Educational Mining 데이터 세트를 이용하였다. 머신러닝 모델 성능을 비교한 결과 랜덤 포레스트가 다양한 크기의 데이터 세트에 높은 정확도를 달성한 것으로 나타났으며 학업 성취도를 예측하는 유용한 머신러닝 알고리즘으로 확인되었다.

Keywords
prediction grades, educational data mining, learning analytics, machine learning, predictive modelling

* Assistant professor, Pierson College, PyeongTaek University
- ORCID: http://orcid.org/0000-0001-5153-756X

· Received: Aug. 27, 2022, Revised: Sep. 21, 2022, Accepted: Sep. 24, 2022

· Corresponding Author: HeeJeong Jasmine Lee
Pierson College, PyeongTaek University, Seodong-daero, Pyeongtaek-Si, Gyeonggi-Do, 3825, South Korea;
Tel.: +82-31-658-8781, Email: hjlee@ptu.ac.kr

Ⅰ. Introduction

The provision of quality and comprehensive education is one of the fundamental aspects of the Sustainable Development Goals(SDG) advanced by the United Nations toward the realization of global sustainability[1]. Notably, the concept of sustainable development is anchored in offering equal opportunities across all sectors, particularly the educational field, to ensure that every individual is being granted adequate and fair opportunities to access and complete his/her studies[1].

Retaining students in higher learning educational institutes and degree programs is one of the massive challenges for stakeholders such as educators, academics, and policymakers. A high rate of student dropout in higher learning institutions like colleges and universities harms both students and educational institutes as they both invest resources to attain their respective goals. In this regard, student retention remains a significant objective for the administration to ensure successful graduation rates and academic achievement. There is a lack of a dedicated framework that tackles the aspects behind attrition. One of the potential reasons behind an excess of students' dropouts could be improper methods to assist students in successfully navigating their respective disciplines. An inadequate understanding of what a specific course entails may result in withdrawals from the course. Appropriate strategies are therefore required to help students quickly adapt to the new environment to minimize dropout rates.

Grade prediction at the early stages of a student's academic degree can be utilized as an important measure to identify the chances of a student dropping out. Thus, institutes can plan to overcome these challenges accordingly. The integration of information and communication(ICT) technologies in educational institutions can play a pivotal role in identifying the aforementioned aspects. For the past several years, there have been a boom in machine learning(ML) techniques in predicting some desired aspects of an entity based on historical information. Similarly, ML methods can reveal insights into the reasons causing dropout.

Technology-enhanced learning(TEL) or Computer Assisted Learning are essential components of modern education that involve the application of digital technology to foster the overall teaching-learning experience. TEL, for instance, encompasses a wide array of tools to strengthen the process of knowledge construction such as learning management systems, social media networking and mobile devices learning applications, data mining, video teaching, and Artificial Intelligence(AI). This trend has largely been propelled by the rampant development of telecommunication infrastructure and growing access to the internet across the globe. This has enabled people to own mobile phones, tablets, and other personalized gadgets. Thus, sustainability in education can be achieved by assimilating technology as compared to the conventional methods of using textbooks and printed learning materials.

Massive Online Open Courses(MOOCs) provide a channel for analyzing the online learning behavior of a large, unevenly distributed group of learners through the extensive amounts of data collected. Subsequently, learning analytics are prepared, which entails the interpretation and optimization of the collected data[2][3]. A vast majority of these learning analytics revolved around predicting the performance of students to determine the characteristics of an appropriate learning environment as well as the preconditions for learning. This information can be utilized by both the instructors, to align the course materials with specific requirements of the learners, and students to reflect upon the teachings and to improve their learning process.

Online learning environments are capable enough to offer a variety of detailed data related to the behavior of different students which cannot be found in the case of a traditional learning environment. Predicting performance in such environments is typically pursued to predict the grades of students in exams, grades and homework of a specific course. MOOCs have gained widespread popularity over the last few years by providing a platform for students for registering different courses offered by some of the highest-ranked universities worldwide.

Although MOOCs have gradually evolved over the years to offer degrees from recognized universities, the rate of completion is relatively low at 7% with further criticism on the quality of education. The principal contributing factors to this diminishing number are a low teacher to student ratio, the nature of interactions that do not occur at the same time or place and varying educational backgrounds. The module of operation under this application involves the use of recorded video lectures, a set of questions, assignments that are graded and discussion forums to enable the sharing of information between learners[4].

Researchers have noted that although a high number of students express interest in these programs, only a limited fraction is involved in viewing video lectures, completing quizzes, and the submission of homework-based assignments. By focusing on self-reported surveys, the authors of previous research[5] reported the rate of dropping out from MOOCs could be determined by examining the factors behind the enrollment. People are guided by different motivations to undertake online courses such as to gain knowledge on a subset of a topic within the broader curriculum, to acquire college credits, future career progression, and as a social experiment on the experiences of online education. An evaluation of these aspects is an indication that students with similar interests will exhibit equal levels of inspiration towards the number of hours invested in the course. The rate of student motivation in MOOCs is, therefore, dependent on goals with statistics indicating that the rate of retention is generally below 20%[6]. This crisis can be reduced by utilizing machine learning algorithms to trace the data left behind by students to offer timely intervention and additional support to prevent the discontinuation of the course.

Researchers have presented different approaches to explore similar aspects to minimize the dropout rate. For example, Harb and El-Shaarawi examined the determinants of student performance across all aspects of life ranging from self-motivation, family income, the previous form of schooling, levels of parent's education to diligence and discipline[7]. The study established that these factors are correlated to the student’s grades. For instance, prior experiences with aptitude tests have been linked to improved cognitive functions and learning. Undergraduate economics students in the introductory stage tend to record better performance based on their overall achievement and knowledge in Calculus and regular class attendance is associated with a higher GPA. Other features that correlate with positive performance in students are age, efforts exerted towards studying, and homogeneity between the student’s learning styles and the tutor’s teaching methods. Besides, students with individual financial support for their education, rather than institutional funding, have been identified as having better performance in academics[7][8].

This study aims to gather information about which ML methods can generate accurate predictions of student grades across different real-world education data sets which differ in their size and structure. Such ML methods can subsequently be used for the learning recommendation system and intelligent tutor system.

Ⅱ. Related Work

Educational data mining(EDM) and AI have become important components of education in terms of learning analytics to guide all relevant parties in decision making.

The data derived from historical performance based on past academic information are analyzed by statistical tools and methods in predicting the final grade of students, as well as data mining and machine learning methods to assess the final results. By evaluating this information, educators can design strategies to reduce the dropout rate. For example, a predictive analysis of the final grades of students can allow the formulation of strategies like specific recommendations for students on how to improve their performance and feedback to the instructors to improve teaching methodology.

ML algorithms gather and integrate knowledge from real world data and use it to make quantitative predictions. Nowadays, ML algorithms are widely being used in various fields like Computer Science, Medicine and Marketing[9]. Numerous types of research projects have been carried out using interesting ML approaches to discover knowledge, make a decision and then provide recommendations. This section elaborates on some of the state-of-the-art ML approaches which could be suitable for predicting student academic performance.

Two types of data analysis approaches are used in ML, i.e., predictive modelling approaches and descriptive modelling approaches. Predictive modelling is a commonly used statistical technique that works by analyzing historical and current information and creating a model to predict future behavior. To find existing patterns, data are segmented along with demographics, behavior, expressed needs and other important factors. Supervised learning functions are used in predictive approaches for estimating unknown or new values of the dependent variables[10]. By comparison, descriptive modelling describes the similarities in real-world events and the relationships between factors responsible for them. Descriptive approaches are unsupervised methods that identify patterns for defining the structure of data[11]. Many ML techniques, such as matrix factorization, collaborative filtering Support Vector Machines(SVMs), Decision trees(DTs), Random Forest(RF), Naive Bayes, Gradient Boosting Machine(GBM), Neural Networks, and simple parametric regression are being used to predict students' grades[3]. Table 1 summarizes previous research which used the ML method to predict student grades.

Predictive models tend to be applied at the end of the MOOC as a post-hoc evaluation, therefore they cannot identify which students are at risk of low grades or dropout in the future. The article by Moreno-Marcos et al.[12] proposed the best time to predict dropout within a self-paced MOOC. Their analysis demonstrated that 25-33% of the theoretical period time of the MOOC is enough time to predict with very good predictive.

In this study, we compared the performances of several ML algorithms for predicting student grades in three real world education data sets of different sizes and structures. The accuracy of ML models generally rises when the size of the dataset increases, but does not do so linearly: the trend is closer to logarithmic [20].

Small datasets contain fewer instances to reflect the population distribution, so they might produce less accurate models. However, they can be loaded into memory and generate results rapidly. Large datasets on the other hand can generate various models and analyses. In this study we used a data set with more than 500,000 instances to represent our large dataset and conversely, less than 500 instances to represent our small dataset. A medium dataset is in between the two to find out how the dataset size affects the overall system accuracy for predicting student grades. The objective of this study was to compare ML classification algorithms on these data sets to identify which algorithms are best able to generalize across different data sets in this field and provide reliable predictions of academic achievement.

Table 1. Dataset description

Authors	ML method	Description
Koren et al.[13]	Matrix factorization(MF)	The model predicted the performance in Algebra and Bridge to Algebra courses. The factorization techniques were useful in the scenarios wherein the dataset is sparse and the student's background or previous information is unknown. The log file showed the students' and the computer-aided system's interactions. This approach further extended research by using tensor-based factorization for predicting the future performance of the students that further added the temporal effects to the performance of students. This model records the success and failure logs of the students on different exercises.
Bydžovská [14], Rechkoski et al.[15]	Collaborative filtering(CBF)	Collaborative filtering(CBF) is where the scores and performance can be predicted depending on the grades from the history of all the courses. CBF based method was proposed by Bydžovská wherein students' performance was predicted at the beginning of their academic period[13]. Other CBF methods that were based on Bayesian probabilistic methods and probabilistic matrix factorization models were proposed for performance prediction[14]. The experiments were performed on the data set collected from Masaryk University. The results depicted that the CBF method was very useful for predicting and recommending future performance.
Yang, T. et al.[4]	Matrix factorization, collaborative filtering, and Restricted Boltzmann Machines	This study focused on student retention by introducing technology to the learning process in universities by using Matrix factorization, collaborative filtering, and Restricted Boltzmann Machines. These systems are critical for evaluating the risks of students failing a particular course so that appropriate policies could be implemented. The study argued that performance during the first semester of university is highly dependent on the accomplishments of the entry-level test and high school achievements.
Khan et al.[16]	Naïve Bayes and Decision Trees	Researchers used the academic data collected from the secondary schools in the district of Kancheepuran, India. The decision tree produced promising results in predicting students' performance. A recommendation system was presented that extracts the educational data for predicting the future performance of the students. To verify this system, techniques for recommendation systems are compared with traditional methods, i.e., linear or logistic regression models.
Zimmermann et al.[17], Elbadrawy et al.[18]	Regression	Researchers combined both variable aggregation and variable selection approaches for predicting the performance of students along with their aggregates. The experiments were performed on 171 records of students obtained from ETH Zürich, Switzerland[17]. Another research proposed a solution that was based upon a multilinear regression model for the prediction of students' grades in a traditional university setup. For this purpose, they used a variety of data that includes a learning management system (LMS) and grades. Here, the incorporation of the LMS data allowed the prediction of grades at the "activity" level and individual evaluations in a specific course[18].
Bhardwaj and Pal[19]	Bayesian classification	The Bayesian classification model has been proposed wherein important factors are harnessed to predict the performance of students.

III. Method

This section encompasses details about the proposed methodology, illustrated in Figure 1. The proposed methodology has four modules: 1) Data Acquisition 2) Pre-processing, 3) Classification and 4) Evaluation.

3.1 Dataset description

The proposed study employs three data sets taken from HarvardX MITx[21], Open University Learning Analytics(OULA) dataset[22] and the xAPI-Educational Mining dataset[23][24]. The datasets' statistics are illustrated in Table 2.

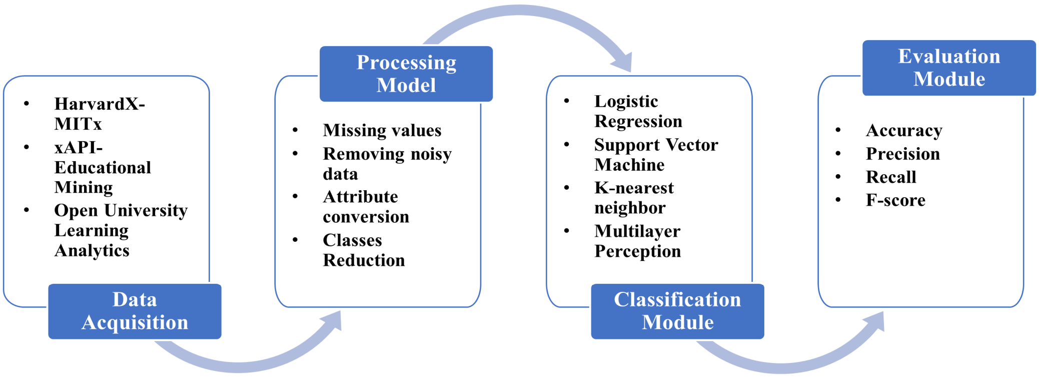


Fig. 1. Proposed methodology

Table 2. Dataset information

Dataset Name	No. of Instances	No. of Attributes	Classes	Dataset Size
HarvardX MITx	597,692	20	2 (Pass, Fail)	67.9MB
Open University Learning Analytics	32,593	15	2 (Pass, Fail)	10MB
xAPI-Educational Mining	480	17	3 (High, Low, Medium)	38KB

We shortly call the HarvardX-MITx dataset the big dataset(i.e. 597,692 instances), OUAL the medium dataset(i.e. 32,593 instances) and xAPI-EM a small dataset(i.e. 480 instances).

3.1.1 HarvardX MITx

The first 17 HarvardX and MITx courses were launched from Fall 2012 to Summer 2013 on the edX platform[21]. In the same year, 43,196 enrolled students earned certificates of degree completion and 35,937 registrants explored half or more of the course content without certification.

3.1.2 Open University Learning Analytics

The Open University Learning Analytics(OULA) dataset offers two elements in the framework: behavior and performance. It contains information about 22 courses, 32,593 students, their evaluation results, and student click logs(10,655,280 entries). This dataset contains demographic information about the students along with their results[22]. The features include "gender", "highest education", “age", "previous attempts", “studied credits" and "final results" representing a unique identification number for the student, the student’s gender, highest student education level, a band of the student’s age, the number of times the student has tried this module, the total number of credits the student takes, etc.

3.1.3 API-Educational Mining

This is an educational data set that was collected from the learning management system(LMS) called Kalboard 360[23][24]. Kalboard 360 is a multi-agent LMS, which aimed to promote learning by using of state-of-the-art technology. These systems provide users with simultaneous access to educational resources via an Internet connection.

The dataset consists of 480(i.e. 305 males and 175 females) student records and 16 features. The students came from twelve different origins. The features are classified into three major categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational stages and grades. (3) Behavioral features such as opening resources and parent parturition in the educational process[25][26].

3.2 Data preprocessing

The next step after data acquisition is data preprocessing. In this step, the collected data set is converted into a suitable format before applying the selected ML models. We have applied various pre-processing techniques for handling missing values, noisy data, categorical variables(i.e. one-hot encoding) and imbalanced data(i.e. class reduction) based on the types of dataset. Data preprocessing has been applied equally to all models instead of being applied to fit the models. The features in the original data were used as they are, and other feature extraction or feature selection methods were not used.

In the Harvardx-MITx dataset(HMD), which is a binary classification task, there are 2 target classes 0 and 1 representing fail and pass outcomes, respectively. The dataset contains a total of 641,139 instances. After pre-processing, 535,641 instances remained. The attributes such as ‘age’, ‘education’, ‘english_speaking’, ‘explored’, ‘genderCat’, ‘grade’, ‘incomplete flag’, ‘nchapters’, ‘ndays_act’, ‘nevents’, ‘nforum_posts’, ‘nplay_video’ and ‘viewed’ are used for classification purpose. The continuous data were scaled down using “Standard Scaler”. One-Hot encoding was applied to the “Data frame”. One hot encoding is a process of converting categorical data into binary vectors for ML algorithms to process categorical data. The “Class” column represents the target variable. The data was split into 70-30% for training and testing respectively.

The OULA dataset contained 32,593 instances classified into four classes, "withdraw", "pass", "fail" and "distinction". This dataset was converted into a format for binary classification. Withdraw and fail classes were converted into one class “fail”. Distinction and pass classes were merged into "pass". The aforementioned variable conversion techniques are used to convert the attribute values into categorical values. The training and testing set contains 70% and 30% instances respectively.

The xAPI-Educational Mining(xAPI-EM) data set has zero instances with missing values, therefore, it was used as a whole. We normalized the Class column data. There were three target classes such as High-level(H), Medium-level(M) and Low-level(L). The dataset has almost half of the samples of class M, while L and H have25% of dataset instances. This data set was assessed according to the belonging of students from different countries.

Encoding of categorical variables such as Binarization, LabelEncoding, and one-hot encoding was performed before conducting standardization, and normalization of numerical variables. The data was split into 70-30% for training and testing respectively.

3.3 Classification

In machine learning classification is a supervised learning task which uses a known dataset to train a model and use them to make predictions (i.e. “Pass, Fail” or “High, Low, Medium”). For classification purposes, Logistic Regression(LR), Support Vector Machines(SVM), K-nearest neighbor(KNN), Random Forests(RF) and Multilayer Perception(MLP) algorithms were employed. These five algorithms were often adopted in predicting students’ academic performance due to their high accuracy on imbalanced data[27].

LR is a generalized linear regression model(GLM) that is used in scenarios when a classification problem is binary. LR is a predictive analysis that determines “the relationship between a dependent binary variable and a set of independent variables”[28].

SVM is a widely used ML algorithm that separates the classes by forming a hyperplane. KNN classification algorithm is deemed as a non-parametric classifier. KNN has widely been harnessed as a baseline classifier in various pattern classification studies. KNN tackles the distance among the instances of training and testing data to determine the output class in classification problems[29].

Random forest(RF) is an ensemble method that assists in both classification and regression tasks. This process forms a forest having multiple decision trees, where each tree predicts a class. The final class is predicted based on majority voting. RF often results in high accuracy and reliability of the outcomes[28].

Multilayer Perception(MLP) is used for both classification and regression tasks. The learning process of the MLP network follows the data samples which are made up of the N-dimensional input vector and the M-dimensional required output vector d, called destination. MLP predicts the output single based on an input vector x[30].

3.4 Hyperparameter optimization

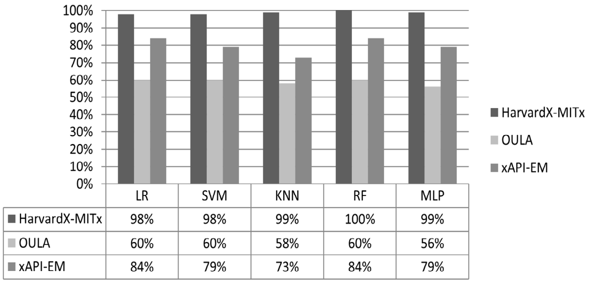
Hyperparameter optimization(or tuning) is the selection of the best set of hyperparameter values to maximize the performance of the model to produce better results. In[27], the authors used differenct academic and non-academic parameters to predict students' academic performance. Optimiazed parameters' settings for the trained model are given in Table 3.

Table 3. Hyperparameters' settings

Dataset	ML Model	Parameters
HarvardX-MITx, OULA, xAPI-EM	LR	random_state=42, C=0.1
	SVM	kernel='rbf'
	KNN	n_neighbor: 17, algorithm='auto', n_jobs=-1
	MLP	hidden_layer_sizes=(100,100,100), max_iter=300, alpha=0.0001,learning_rate='adaptive',activation='relu', solver='sgd', verbose=10, random_state=42,tol=0.0001
HarvardX-MITx	RF	n_estimators=1000, n_jobs=-1, max_depth=5, random_state=42
OULA, xAPI-EM	RF	n_estimators=2000, n_jobs=-1, max_depth=6, random_state=42

IV. Results

The overall accuracy of the ML classifiers is depicted in Figure 2.

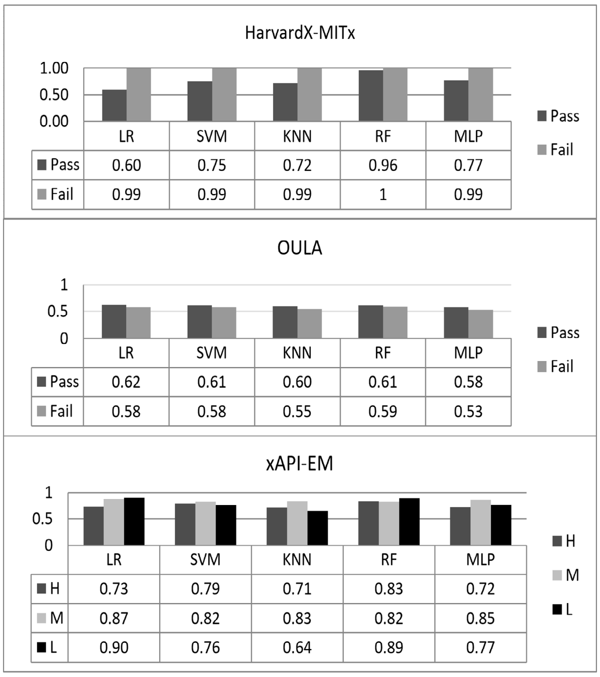


	LR	SVM	KNN	RF	MLP
HarvardX-MITx	98%	98%	99%	100%	99%
OULA	60%	60%	58%	60%	56%
xAPI-EM	84%	79%	73%	84%	79%

Fig. 2. Overall accuracy score achieved by the classifiers

The performance of classifiers in terms of accuracy varies depending upon the datasets. High classification accuracy was attained for the HarvardX-MITx dataset where an accuracy score of 98%, 97%, 99%, 99% and 98% was attained by LR, SVM, KNN, RF and MLP respectively. For the xAPI-EM dataset, less accuracy was recorded as compared to the HarvardX-MITx dataset. RF outperformed the other four classifiers by achieving an 84% score, followed by LR scoring 83% accuracy. KNN classifier achieved high accuracy for the HarvardX-MITx dataset, whereas for xAPI-EM and OULA datasets, the classifier attained 75% and 57% accuracy respectively.

Figure 3 illustrates the precision scores achieved by the classifiers using three education datasets. For the HarvadX-MITx dataset, all the classifiers attained a high precision score for the "Pass" class. RF and MLP classifiers performed well to predict fail class by attaining a precision of 0.96. Whereas, the SVM classifier achieved a low precision score for the fail class. For the xAPI-EM dataset, MLP outperformed all other algorithms for the M class(0.85) and RF attained a 0.89 precision score for the L class. For the H class, the LR classifier achieved a high score of 0.84. The LR algorithm performed well on the OULA dataset and achieved 0.63 and 0.59 precision scores for a pass and a fail class respectively.



HarvardX-MITx

	LR	SVM	KNN	RF	MLP
Pass	0.60	0.75	0.72	0.96	0.77
Fail	0.99	0.99	0.99	1	0.99

OULA

	LR	SVM	KNN	RF	MLP
Pass	0.62	0.61	0.60	0.61	0.58
Fail	0.58	0.58	0.55	0.59	0.53

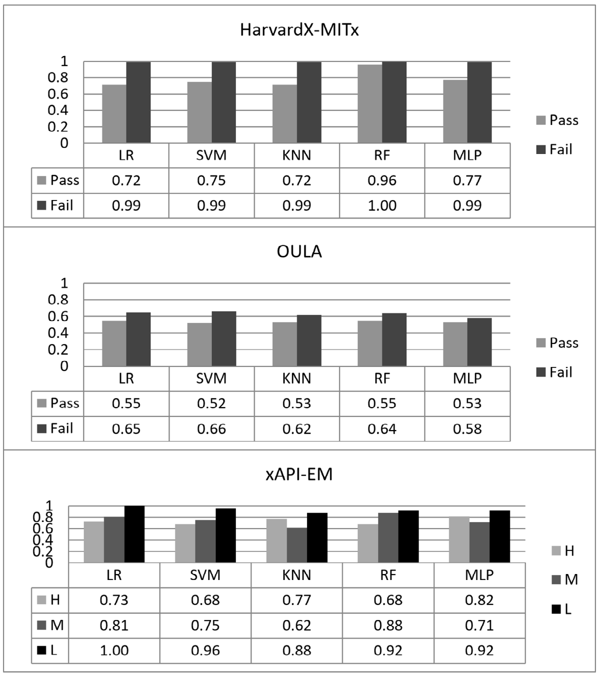
xAPI-EM

	LR	SVM	KNN	RF	MLP
H	0.73	0.79	0.71	0.83	0.72
M	0.87	0.82	0.83	0.82	0.85
L	0.90	0.76	0.64	0.89	0.77

Fig. 3. Precision values for individual classes by the classifiers

The recall values achieved by the classifiers are presented in Figure 4. All the classifiers attained a high recall score in predicting the pass class for the HarvardX-MITx dataset. Whereas for the fail class the LR and SVM achieved low recall values. For the H class in the xAPI-EM dataset, the MLP algorithm attained a high recall score of 0.82, the SVM recall value for the L class was 0.96 and RF attained a high score of 0.88 for the M class. MLP and LR performed well on the OULA dataset attaining 0.96 for the pass class and 0.57 for the fail class respectively.

Due to the imbalanced classes issue found in many datasets, the f-score is a suitable measure to compare the performance of classifiers. The f-score values achieved by the classifiers for the three datasets are shown in Figure 5. As it can be shown in the Figure, RF and MLP achieved high f-score values of 0.98 each for the HarvardX-MITx dataset. For the xAPI-EM dataset, RF attained a high f-score of 0.83. Finally, for the OULA dataset, LR attained an f-measure of 0.61.



HarvardX-MITx

	LR	SVM	KNN	RF	MLP
Pass	0.72	0.75	0.72	0.96	0.77
Fail	0.99	0.99	0.99	1.00	0.99

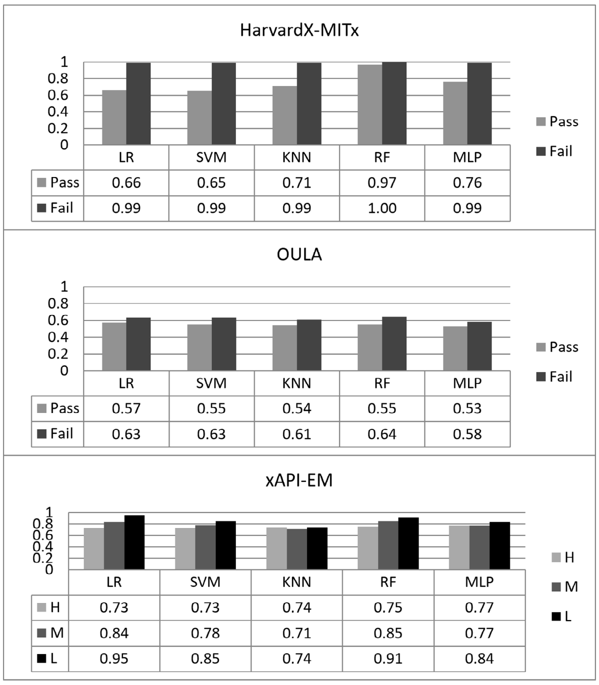
OULA

	LR	SVM	KNN	RF	MLP
Pass	0.55	0.52	0.53	0.55	0.53
Fail	0.65	0.66	0.62	0.64	0.58

xAPI-EM

	LR	SVM	KNN	RF	MLP
H	0.73	0.68	0.77	0.68	0.82
M	0.81	0.75	0.62	0.88	0.71
L	1.00	0.96	0.88	0.92	0.92

Fig. 4. Recall values for individual classes by the classifiers



HarvardX-MITx

	LR	SVM	KNN	RF	MLP
Pass	0.66	0.65	0.71	0.97	0.76
Fail	0.99	0.99	0.99	1.00	0.99

OULA

	LR	SVM	KNN	RF	MLP
Pass	0.57	0.55	0.54	0.55	0.53
Fail	0.63	0.63	0.61	0.64	0.58

xAPI-EM

	LR	SVM	KNN	RF	MLP
H	0.73	0.73	0.74	0.75	0.77
M	0.84	0.78	0.71	0.85	0.77
L	0.95	0.85	0.74	0.91	0.84

Fig. 5. F-score values for individual classes by the classifiers

V. Discussions and future work

Accurate prediction of student grades at the early stages of their degree program can play an important role in increasing retention rates at higher learning educational institutions. Researchers have investigated different approaches and proposed solutions to increase student retention. Different machine learning techniques have been proposed to predict the performance of the students. Successful prediction yields valuable insights into knowing the factors that impact the success of students. These insights inform stakeholders to surge the commitment of the students to increase their retention.

Machine learning has proved to be a powerful tool in educational data mining for predicting the future performance of students. This study has evaluated five machine learning classifiers, including LR, SVM, KNN, RF and MLP on three small, medium and large-sized datasets with different structures and classes. How do we determine the choices between each machine learning algorithm for different sizes and types of data sets? Accuracy measures the ratio of correctly predicted instances to that of total instances. If high accuracy is a priority, the best way is to test out a couple of different algorithms and try different parameters within each algorithm as well and then select the best one by cross-validation. Here is the result of our experiments.

Applying LR, SVM, KNN, RF and MLP classifiers, the outcomes revealed that RF achieved 81.3%, LR 80.7%, SVM 78.7%, MLP classifiers 79.0% and finally KNN achieved 76.7% accuracy. The highest accuracy was obtained from the HarvardX-MITx dataset, i.e., 98% to 100% from the 5 algorithms. The HarvardX-MITx dataset has the biggest dataset size (i.e. 597,692 instances versus OULA has 32,593 and xAPI-EM has 480). The result shows that if there is a huge dataset, then whichever classification algorithm is used does not make much difference. However due to the size of the dataset, the training time was extended, and if an approximation prediction is adequate, there will be a huge reduction in processing time. So, if there are a lot of instances, any algorithm can be chosen based on speed or ease of use instead of accuracy.

For the small and medium datasets(i.e. xAPI-EM has 480 and OULA has 32,593), LR and RF outperformed the others. KNN and MLP did not predict accurately(i.e. 58% and 56% respectively) with the OULA dataset. For the OULA dataset, 8 features(i.e. "gender", "region", "highest_education", "imd_band", "age_band", "num_of_prev_attempts", "studied_credits", "disability") were used for the inputs to develop a ML model. KNN did not handle a lot of irrelevant features so the performance of classifiers including accuracy, precision, recall and F-score was not achieved a good score. In all datasets, when RF and LR classifiers are in operation, the proposed model boosted their performance. The important features in Harvard’s data are ndays_act(i.e. the number of unique days students interacted with the course) and nchapters(i.e. the number of chapters in the courseware that students interact with) by evaluating the importance of features on the classification task[22]. The lowest accuracy was obtained from the OULA dataset, i.e., 56% to 60%. The features used in OULA are gender, region, highest education, age band, num of previous attempts, studied credits and disability. The OULA features are incapable of distinguishing between two students i.e. one high scorer and one low scorer because the features used in OULA are very much generalized. Certainly, adding features like how many times a student asked questions, what resources he visited, ndays_act and nchapters will add more distinctive value to the data and will increase the accuracy.

The objective of this study was to obtain the best prediction model so that in the following work, an individualized recommendation system will be developed based on prediction in students’ grades of the previous academic years of the subject. A comparison of ML model performance metrics revealed that random forest tended to score highly(relative to the other algorithms) across the tested data sets and metrics.

This finding suggests that the random forest algorithm represents a useful ML tool for predicting student academic performance. In the future, we plan to carry out the following research: (i) predicting student final grades at an early stage(e.g. before mid-term exam), (ii) developing a recommendation system for students based on their skill, or knowledge component and (iii) intelligent tutoring based on detecting student's motivation and engagement.

References

[1] UN, "Transforming our world: the 2030 Agenda for Sustainable Development", Sustainable Dev. Knowl. Platform. Sustain. un. org/post2015/ transformingourworld(consulté le 4 avril 2017), 2015.

[2] R. Conijn, A. Van den Beemt, and P. Cuijpers, "Predicting student performance in a blended MOOC", J. Comput. Assist. Learn., Vol. 34, pp. 615-628, May 2018. https://doi.org/10.1111/ jcal.12270.

[3] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A review and future research directions", IEEE Trans. Learn. Technol., Vol. 12, No. 3, pp. 384-401, Jul. 2018. https://doi.org/10.1109/ tlt.2018.2856808.

[4] T.-Y. Yang, C. G. Brinton, C. Joe-Wong, and M. Chiang, "Behavior-based grade prediction for MOOCs via time series neural networks", IEEE J. Sel. Top. Signal Proc., Vol. 11, pp. 716-728, May 2017. https://doi.org/10.1109/jstsp.2017.2700227.

[5] Z. Ren, H. Rangwala, and A. Johri, "Predicting performance on MOOC assessments using multi-regression models", arXiv Prepr. arXiv1605.02269, Jun. 2016.

[6] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "Predicting Student's Academic Performance in a MOOC Environment", Dec. 2017. https://doi.org/ 10.15242/dirpub.dir1217002.

[7] N. Harb and A. El, "Factors Affecting UAEU Students Performance", Res. Aff. Sect., pp. 146, Jul. 2006.

[8] A. Hellas et al., "Predicting academic performance: a systematic literature review", in Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education, Jul. 2018, pp. 175-199. https://doi.org/10.1145/3293881.3295783.

[9] B. Cotter, "Student retention: An issue, a discussion and a way forward, for higher education professionals", 2013.

[10] S. J. Hong and S. M. Weiss, "Advances in predictive models for data mining", Pattern Recognit. Lett., Vol. 22, No. 1, pp. 55-61, Jan. 2001. https://doi.org/10.1016/S0167-8655(00)00099-4.

[11] C. Lang, G. Siemens, A. Wise, and D. Gasevic, "Handbook of learning analytics", SOLAR, Society for Learning Analytics and Research, 2017. https://doi.org/10.18608/hla17.

[12] P. M. Moreno-Marcos, P. J. Munoz-Merino, J. Maldonado-Mahauad, M. Perez-Sanagustin, C. Alario-Hoyos, and C. D. Kloos, "Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs", Comput. Educ., Vol. 145, pp. 103728, Feb. 2020. https://doi.org/10.1016/j.compedu.2019.103728.

[13] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems", Computer(Long. Beach. Calif)., Vol. 42, No. 8, pp. 30-37, Aug. 2009. https://doi.org/10.1109/ mc.2009.263.

[14] H. Bydžovská, "Are collaborative filtering methods suitable for student performance prediction?", in Portuguese Conference on Artificial Intelligence, pp. 425-430, Jan. 2015. https://doi.org/10.1007/978-3-319-23485-4_42.

[15] L. Rechkoski, V. V. Ajanovski, and M. Mihova, "Evaluation of grade prediction using model-based

collaborative filtering methods", in 2018 IEEE Global Engineering Education Conference (EDUCON), Santa Cruz de Tenerife, Spain, pp. 1096-1103, Apr. 2018. https://doi.org/10.1109/EDUCON.2018.8363352.

[16] B. Khan, M. S. H. Khiyal, and M. D. Khattak, "Final grade prediction of secondary school student using decision tree", Int. J. Comput. Appl., Vol. 115, No. 21, pp. 32-36, Apr. 2015. https://doi.org/10.5120/20278-2712.

[17] J. Zimmermann, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance", J. Educ. Data Min., Vol. 7, No. 3, pp. 151-176, 2015.

[18] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics", Computer(Long. Beach. Calif)., Vol. 49, No. 4, pp. 61-69, Apr. 2016. https://doi.org/10.1109/mc.2016.119.

[19] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security, Vol. 9, No. 4, pp. 136-140, Apr. 2011. https://doi.org/10.48550/arXiv.1201.3418.

[20] S. Shahinfar, P. Meek, and G. Falzon, "'How many images do I need?' Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring", Ecol. Inform., Vol. 57, pp. 101085, May 2020. https://doi.org/10.1016/j.ecoinf.2020.101085.

[21] HarvardX, "HarvardX Person-Course Academic Year 2013 De-Identified dataset, version 3.0", Harvard Dataverse. https://doi.org/10.7910/DVN/26147.

[22] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset", Sci. data, Vol. 4, pp. 170171, Nov. 2017. https://doi.org/10.1038/sdata.2017.171.

[23] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance", in 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, pp. 1-5, Nov. 2015. https://doi.org/10.1109/aeect.2015.7360581.

[24] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods", Int. J. Database Theory Appl., Vol. 9, No. 8, pp. 119-136, 2016. https://doi.org/10.14257/ijdta.2016.9.8.13.

[25] I. Aljarah, "Student's Academic Performance Dataset-xAPI-Educational Mining Dataset", Vol. 2022, 2017. [Online]. Available: https://www.kaggle.com/aljarah/xAPI-Edu-Data/version/4.

[26] R. Russell, "Open University Learning Analytics Dataset", Kaggle, Vol. 2022, 2019. [Online]. Available: https://www.kaggle.com/rocki37/open-university-learning-analytics-dataset#assessments.csv.

[27] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of students' academic performance based on courses' grades using deep neural networks", IEEE Access, Vol. 9, pp. 140731-140746, Oct. 2021. https://doi.org/10.1109/ACCESS.2021.3119596.

[28] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study", Int. J. Adv. Comput. Sci. Appl., Vol. 9, No. 2, 2018. https://doi.org/10.14569/ijacsa.2018.090238.

[29] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets", Springerplus, Vol. 5, pp. 1304, Aug. 2016. https://doi.org/10.1186/s40064-016-2941-7.

[30] E. A. Zanaty, "Support vector machines(SVMs)

versus multilayer perception(MLP) in data classification", Egypt. Informatics J., Vol. 13, No. 3, pp. 177-183, Nov. 2012. https://doi.org/10.1016/j.eij.2012.08.002.

Authors

HeeJeong Jasmine Lee

1997 : Computer Science and Engignering, POSTECH

2003 : MSc degree in Compute Science, University of Edinburgh

2005 : MPhil degree in Technology Policy, Cambridge University

2013 ~ present : PhD degree in Information Technology, Monash University

Research interests : social network analaysis, artificial intelligence and survival analysis