IADIS International Journal on Computer Science and Information Systems
Vol. 18, No. 2, pp. 1-21
ISSN: 1646-3692

# HIGH ACADEMIC PERFORMANCE AND ENGAGEMENT BEHAVIOURS IN ONLINE LEARNING ENVIRONMENTS

Rhodora Abadia, Sisi Liu and Qiaoling Sun
*University of South Australia, Australia*

**ABSTRACT**

Most of the studies in identifying the factors contributing to high academic performance of students focus on both traditional and blended learning environments. This study explores the online engagement behaviours of high-performing online students. Three years of data derived from students' online activities and academic performance across six computing courses was analysed. Our findings reveal a continuity of high-performance among students who excel in online introductory programming courses, extending their success to subsequent online computing courses. Furthermore, we establish a positive association between programming experience and improved performance. High-performing students exhibit engagement levels that exceed their peers by at least 100% in online formative learning activities, with particular emphasis on forum participation and quiz contribution. This study also looked into the predictive potential of past academic performances in anticipating future achievements. It demonstrated a predictive accuracy of 75% for programming courses and 91% for non-programming ones. When combined with student engagement data, the predictive accuracy for programming courses increases to 81%. This approach requires substantial data until at least week 7 or 8 of a 10-week course, potentially delaying intervention until the latter stage of the course. To address this, exploring models exclusively reliant on engagement data is recommended. Our research offers actionable insights for educators, enabling them to identify critical content design elements that enhance academic performance across diverse student profiles, fostering improved online engagement and narrowing performance gaps. Moreover, this study calls for a re-evaluation of learning analytics tools within educational institutions to better understand and enhance student behaviours and patters, ultimately improving the overall online learning experience.

**KEYWORDS**

Descriptive Learning Analytics, Predictive Learning Analytics, Online Learning, Online Engagement Behaviors, Online Academic Performance

# 1. INTRODUCTION

Self-motivation and regulation are important for students to achieve high academic performance in a fully online learning and teaching environment. Numerous educational research and studies have focused on exploring the factors that differentiate academically high-performing students from low-performing students. To date, studies on understanding students' high academic performances have concentrated on the traditional or blended classroom learning using self-report instruments (Guo et al., 2019; Kaplan, 2018; Wang & Liou, 2018). However, self-report instruments have been criticised for being limited due to subjectivity. Inaccuracies in self-assessment can be resulted from ingenuine assessments, such as participants altering answers to make them more socially acceptable (Duckworth & Yeager, 2015), being unable to assess themselves accurately (Araujo et al., 2017), exhibiting response bias and having difficulty mapping their answers to the rating scales and being unable to fully retrieve information from memory (Rosen et al., 2017). Students tend to retrieve distinctive information that is time-bound, but self-reported questions are seldom distinct. With the popularity of online learning and the advancement of data mining techniques, Learning Management Systems (LMS) have been developed and implemented by various online education providers in online courses to automatically record students' engagement and performance data. It is reported that using LMS data to support learning analytics and educational data mining provides a more objective picture of students' learning through data-driven approaches (Liu et al., 2017).

Understanding the online engagement behaviors that contribute to online students achieving high performance will assist online facilitators in continually supporting high-performing students and identifying critical factors to help other students improve their academic performance. Online facilitators can set a guide to online engagement to uplift the performance of non-high performing students. This reasoning is consistent with Dweck (2006)'s work on growth mindset, which is now synonymous with high expectation. It is hypothesised that students' achievement is strongly affected by what the teacher expects of them, and this has been justified by many education researchers (Campbell et al., 2020; Johnston et al., 2019; Robinson, 2017).

## 1.1 The Study

The engagement data used in this study were collected from first- and second-year students who were enrolled in a 100% asynchronous online IT degree. Demographics of online students include those who are already working, parents who are unable to go to campus, have disabilities or live or work in a remote area. For them, online learning is more flexible and accessible. Identifying online engagement behaviors of high-performing students using the engagement and performance data recorded through the LMS not only provides a more objective view, but also gives the instructors an insight into how these students' online behaviors differ from the rest of the online students. Identifying these online behaviors can also help instructors highlight these behaviors and encourage changes in practice. Specifically, instructors can suggest specific strategies and provide guidance to non-high performing students on how to adopt these online engagement behaviors. Moreover, personalised learning interventions can be motivated from these behaviors and designed such that students are guided in the development of these positive online behaviors. Results can also help students reflect on their behaviors and learn to self-regulate. Finally, this study will further help educational institutions identify what needs to be improved in the current learning analytics tool to better understand student behaviors and patterns.

This paper presents a longitudinal study investigating online student engagement and academic performance from LMS data of high-performing students enrolled in fully online computing courses. It aims to answer the following research questions:

1. Do high-performing students in introductory online computing courses (programming and non-programming) continue to have high performance in their succeeding computing courses?
2. What are the associations between students' academic performance and engagement behaviors in these courses? What student academic and engagement data predict high performance in these courses?

# 2. RELATED STUDIES

This section discusses related literature on high-performing students, self-regulated learning, and student engagement data from LMS used in learning analytics.

## 2.1 High-Performing Students

Studies have shown a positive relationship between student motivation and academic achievement (Pintrich & De Groot, 1990). It has also been indicated that high-achieving students show a refined ability to select and modify study behaviors (Mai et al., 2021), manage their time and use more effective study strategies (Dunlosky et al., 2013) and routinely engage in and adapt skills to pursue these behaviors (Zimmerman & Schunk, 2001). Despite evidence of the association between high performance and effective study behaviors as a primary study strategy, Persky (2018) found in his longitudinal study that high-achieving students relied on re-reading texts and re-watching videos. All these studies revolved around the idea that students perform better if they can fully understand concepts when they self-regulate their learning behaviors. This idea is especially true in an online learning environment where self-regulated learning is critical to students' academic success due to them having limited interactions with peer learners and assistance from the instructors.

A study undertaken by Alqurashi (2022) compared seven aspects of student engagement (i.e., higher-order learning, reflective and integrative learning, learning strategies, quantitative reasoning, and collaborative learning, student–faculty interaction, and effective teaching practices) using survey data collected from senior-level undergraduate students who studied online courses. Their findings showed that low achieving students had significantly higher student-faculty interactions than high-performing students. Although this study looked at students' engagement in an online learning environment, the data used were from student survey, which may include inaccuracies in self-assessment of their engagement. Also, the study was not for students who studied fully asynchronous online courses and did not explore the engagement behaviors when interacting with online course activities. In this paper, the focus is the monitoring of students' learning process manifested through their online engagement behaviors using engagement and academic performance data captured in the LMS.

## 2.2 Student Engagement and Performance through LMS Data

In an online learning environment, data is often collected from an LMS. An LMS is a critical tool in a fully online learning environment to facilitate the teaching and learning process. In addition to distributing and managing course materials, LMS also captures students' engagement and performance data that can be further utilised to support the learning analytics and educational data mining. Recent LMS-related research has shifted the focus from exploring interactive and creative functionalities of an LMS to analysing the LMS data, such as the log data (Henrie et al., 2018) and activity data (Simanullang & Rajagukguk, 2020), to discover patterns of student engagement, evaluate students' academic performance and improve instructors' teaching pedagogical practices. Recent studies revealed that learning analytics (Conole et al., 2011) can be applied to LMS data to visualise student engagement patterns, derive insights from student engagement data for better learning design and improve student learning experience (Henrie et al., 2018; Toro-Troconis et al., 2019). However, empirical data was limited in these studies due to an LMS implemented as a supplement to the traditional classroom teaching or very few fully online courses available for research purposes.

To date, research on understanding student high academic performance has focused on studying students in a traditional classroom or blended learning environment. Nevertheless, few studies have attempted to look at the longitudinal data in online learning and identify the online engagement behaviors of students achieving high academic performance in a fully online learning environment.

## 2.3 Education Data Mining and Predictive Analytics Models

In recent educational research, a significant focus has been on students' academic performance prediction, with the aim of improving teaching methods and supporting student success. This area often involves using engagement data to predict dropout rates and academic performance.

Several studies used regression analysis to examine dropout rates, with Mubarak, Cao, and Zhang (2020) revealing that efficient model performance typically requires seven weeks of data mining. Burgos and colleagues (2018) found that the first assessment must be graded before predictive models yield results. Rovira, Puertas and Igual (2017) explored logistic regression, naive Bayes, random forest, and AdaBoost, with naive Bayes and logistic regression proving effective for smaller datasets, while random forest and AdaBoost perform better with larger datasets.

Predicting student course outcomes or identifying at-risk students employs various models, including k-nearest neighbors (k-NN) (Aluko et al., 2016), J48 decision trees (Gotardo, 2019), CART (Hu, Lo & Shih, 2014), logistic regression (Soffer & Cohen, 2019; Hu, Lo, & Shih, 2014), random forest, and bootstrap (Beaulac & Rosenthal, 2019; Kotsiantis, Pierrakeas & Pintelas, 2004; Golino & Gomes, 2014). Tree-based models generally outperform logistic regression, except in cases like Aluko et al. (2016) and Soffer & Cohen (2019).

Pardo, Han, and Ellis (2017) combined engagement and self-report data using multiple regression to predict course grades. Other studies leverage machine learning feature selection and support vector machine (SVM) (Liu & Cheng, 2016), J48 algorithms (Al-Barrak & Al-Razgan, 2016), and random forest models (Beaulac and Rosenthal, 2019) to predict cumulative grade point averages upon graduation.

While previous research has primarily centered on predicting students' academic outcomes, such as passing or failing a course or dropping out, utilising enrollment data, engagement metrics, or course grades, our study takes a unique approach. We shift the focus towards comprehending the engagement behaviours of high-performing students and harnessing this engagement data to pinpoint crucial factors for predicting high-performance among students.

In this paper, three-year data from six courses were collected from an LMS that stores various online courses having frequent engagement data of different online activities to conduct a longitudinal study to further analyse the online students' engagement behaviors and their association with students' academic performance.

## 3. WORKFLOW MODEL

The analysis used students' academic performance and online student engagement data. Online engagement behaviors were explored through the following formative(non-graded) online data activities: forum, quiz, URL, lesson, file, and folder. For empirical experiments, six courses were analysed, among which three were a series of programming courses and three were a series of non-programming IT courses. The data was divided into two case scenarios, including student academic performance alone, and student academic performance with online student engagement. These scenarios were subsequently used to identify the high-performing students in each course, i.e., if they continue to perform well in the succeeding courses, and the relationship of their engagement with the online course to their high performance. For example, Programming 1 is a prerequisite for the Programming 2 course, thus in the experiment, the number of high-performing students in the pre-requisite course that continue to be high-performing in the subsequent courses were identified. The workflow model used in this study comprises of four main phases: data collection, processing and transformation, data exploration and visualisation, experiments and evaluations (Figure 1).

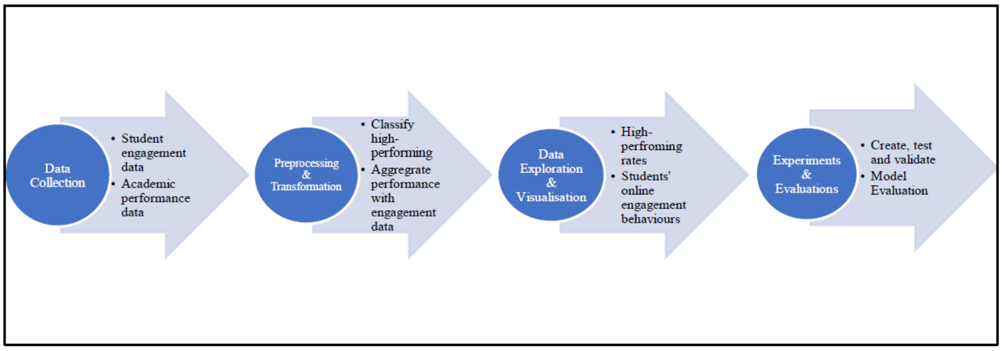


Figure 1. The workflow of the model used for analysing online engagement behaviours of high-performing students

## 3.1 Data Collection

The data were collected from student engagement and academic performance reports from the LMS. A total of 804 student academic and engagement data were used in the study, from which 612 distinct students whose data were collected and analysed. These reports were taken from a series of three online programming courses and a series of three online non-programming courses. For programming courses, three years (2018-2020) of reports were taken from two first year courses and one second year course. For non-programming courses, three years (2019-2021) of reports were taken from two first year courses and one second year course. The data contained a list of students who were enrolled in the courses with de-identified student IDs, their online engagement data, including the number and percentage of forum views and contributions, quizzes attempted and scores (if recorded), files, URLs, lessons, folders accessed, and students' academic performance involving summative(graded) assessment grades and final course grades. The forums are discussion forums where students can post questions and contribute to answering the questions; quizzes are formative assessments for students to self-check what they've learned; files are content related materials (instructions, additional materials); URLs are links to content related materials; and folders contain a group of content material files usually containing instructions and files associated to learning activities or assessments. Demographic data of the students were not gathered as a result of previous study (Bretana et al., 2020), which discovered that the available demographic characteristics (e.g., age, location, grade point average, gender, degree, full-time/part time study) were not effective indicators of achievement in online programming courses.

## 3.2 Preprocessing and Transformation

Two distinct datasets collected were student engagement data and academic performance data. Preliminary exploration was carried out to determine necessary data processing. The data was used to create two datasets for the two scenarios. The first scenario used only the student's academic performance data while the second scenario used both student engagement and academic performance data. These datasets were used later for the identification of factors that predict a student's high performance.

### 3.2.1 Scenario 1: Student's Academic Performance

The number of columns in the academic performance dataset varied across different courses within a given semester. This variation is caused by some course's dataset reports having only major summative assessments and a final examination, while others breaking down their major summative assessments into several continuous evaluations like practical exercises. Additionally, some courses have a "hurdle" requirement, such as a minimum score required for an oral assessment or a final exam. Students who do not meet this hurdle will fail the course despite achieving an overall passing grade. Due to these inconsistencies, it became essential to create a more generalised version of the dataset. A generalised data structure was employed to address these disparities, streamlining the data while retaining critical information.

For uniformity, the number of major summative assessments across all courses was standardised to three, along with one final assessment. Each of these assessments carries a specific weight corresponding to its importance as written in the course outline. If an assignment's weight within a course is marked as 0, it indicates the absence of that assignment.

The course grade is determined by aggregating the total score and accounting for any hurdle requirements where applicable.

The outcome variable was named as "IsHighPerformance", derived from the Grade column. A student is considered high-performing if they attain a final raw score of 75 or above. Table 1 presents the generalised data structure used to evaluate students' academic performance.

Table 1. The generalised data structure of students' academic performance

| Column Name | Data Type | Description |
|---|---|---|
| StudentID | Nominal | Unique ID of a student |
| Assignment1_Weight | Continuous | Weight of assignment 1 in the course outline |
| Assignment2_Weight | Continuous | Weight of assignment 2 in the course outline |
| Assignment3_Weight | Continuous | Weight of assignment 3 in the course outline |
| FinalExam_Weight | Continuous | Weight of final exam in the course outline |
| Assignment1 | Continuous | Assignment 1 score |
| Assignment2 | Continuous | Assignment 2 score |
| Assignment3 | Continuous | Assignment 3 score |
| FinalExam | Continuous | Final Exam score |
| CourseTotal | Continuous | Total score in this course |
| Grade | Ordinal | Grade in the grade system |
| SP | Nominal | Study period |
| Year | Nominal | Year |
| Course | Nominal | Course name |
| IsHighPerformance | Binary | High-performing or non-high performing grade |

In the academic performance dataset, each student's record consists of grades from summative assessments. These datasets employ two grading formats, a 100-point scale and a grading system based on the assignment's percentage contribution to the total course grade. To standardise the data, we converted grades given in percentage format to the 100-point scale. For instance, if an assignment contributes to 40% of the overall course grade, its grading scale ranges from 0 to 40. This score can be converted to the 100-point scale by dividing it by its percentage weight (in this case, 40) and then multiplying the result by 100.

### 3.2.2 Scenario 2: Student's Academic Performance Combined with Online Engagement Data

Students can engage in seven distinct activities within a course, including forums, quizzes, assignments, URLs, lessons, files, and folders. Among them, only the assignments activity is considered a summative activity, while the others are formative. Our analysis focused on the formative activities, particularly since they are optional for students. Activities such as URLs, lessons, files, and folders primarily present content materials without necessitating active participation from students. Conversely, forums and quizzes encourage and require student interaction. For each of these activities, three types of data are recorded: the total number of engagements by each student, the number of views by a student, and the number of contributions made by a student. This breakdown is detailed further in Table 2.

Table 2. The generalised data structure of each activity engagement

| **Column Name** | **Data Type** | **Description** |
|---|---|---|
| SP , Year | Nominal | Study Period and Year |
| Course | Nominal | Course name |
| Sum<activity>Views | Continuous | Total view count of a student in an activity |
| SumPercent<activity>Views | Continuous | Total percentage of views of a student in an activity |
| Sum<activity>Contributions | Continuous | Total contributions of a student inan activity. |
| SumPercent<activity >Contributions | Continuous | Total percentage of contributions ofa student in an activity |
| IsHighPerformance | Binary | Target variable |

The student engagement dataset captured multiple entries for each student, as there are nine online activities with which students can engage. Every interaction by a student resulted in a new system-generated record. For our data analysis, it’s essential to consolidate these multiple engagement records for each student into a singular entry. This consolidation was achieved by aggregating all related records.

### 3.2.3 Feature Selection and Transformation

Before building the model for each dataset, to select the attributes for training the model, variables that had at least 0.1 correlation with the target variable *isHighPerformance* were selected. Additionally, columns *Year*, *CourseTotal*, *Course*, *CourseThisYear*, *GradeInThisYear*, *StudentID*, *SP*, and *SPInThisYear* were removed for all data. Variables such as *SumForumViews* also have an equivalence *SumPercentForumViews* and are highly correlated with each other. Hence, only the one with total views (i.e., only *SumForumViews*) was used for modelling.

Min-max normalisation was applied to the continuous variables relating to *assignment scores* and *FinalExam*. The 100-point scale grades were transformed to values between 0 and 1 using the transformation shown in the following equation:

$$x^{y}\,\frac{x - min_{old}}{max_{old} - min_{old}}\,(max_{new} - min_{new}) + min_{new}$$

In addition, the final grade (*Grade)*, which is a categorical variable, was transformed to a binary variable using one-hot encoding. The values of the *Grade* column are spread into multiple columns and are assigned binary values 0 or 1.

## 3.3 Data Exploration and Visualizations

The datasets for all courses were initially explored by examining students’ academic performance, followed by exploring the details of the students’ online engagement combined with their academic performance. This approach helped pinpoint patterns and provided a descriptive analysis of the data. The academic performance of students who continued to enroll in the series of programming courses from 2018 to 2020 and those who enrolled for the non-programming courses from 2019 to 2021 were examined. This examination also encompassed the views and contributions these students made to formative activities.

Given our emphasis on formative activities, we excluded summative actions from our engagement behaviour analysis. To clarify, metrics such as views of assignments and file submissions weren't factored into our count. This exclusion stems from the fact that these activities are mandatory, and high-achieving students naturally engage in summative activities to get high grades. Our primary intent is to decipher the behavioral trends associated with formative activities.

We sought patterns to determine if students excelling in initial programming courses maintain their high academic performance in subsequent programming courses, and to model their engagement behaviours across these courses. We observed the same patterns for non-programming courses. Results from this data exploration not only helped our understanding of the dataset but also revealed the patterns whether high-performing students consistently outperformed in successive courses. The forum views and quiz contributions showed the most number of course engagements.

It is important to note that "Programming 1" serves as a pre-requisite course for both "Programming 2" and "Programming 3". Ideally, students should progress from "Programming 1" to "Programming 2", and finally to "Programming 3" over separate study periods. But in the face of scheduling conflicts, students might opt to tackle "Programming 1" followed by both "Programming 2" and "Programming 3" simultaneously. Additionally, "IT Fundamentals" is a pre-requisite for the "System Analysis" course, which further leads to the "System Design" course. The results and analysis from this exploration are presented in the subsequent section.

## 3.4 Experiments and Evaluations

Six supervised-learning models involving Logistic regression, Support Vector Machines (SVM), k Nearest Neighbor(k-NN), Naive Bayes, decision tree, and random forest were employed for the prediction of students' high-performance outcomes. Logistic regression, initially chosen as the baseline model, served a dual purpose. Firstly, it was used to assess the linearity of the data and to gauge prediction accuracy when no preprocessing, transformation, or dimensionality reduction techniques were applied. Secondly, logistic regression was selected due to its proven efficacy when dealing with binary classification tasks. In this study, the target variable is binary, representing whether a student is high-performing or not (i.e., "*isHighPerformance*"). Logistic regression computes the probability of the occurrence of the variable y (*isHighPerformance*) based on the predictor *x* (e.g., formative activities).

The Sequential Minimal Optimisation classifier algorithm was employed for SVM, with hyperparameter tuning conducted using the radial basis function kernel. Additionally, the Naïve Bayes model was also benchmarked. The models were trained twice:

- The first training iteration utilised the "Grade" variable, employing the nominal-to-binary filter and setting the attribute indices to correspond with the index of the "Grade" variable.
- The second training iteration employed a binary variable derived from the "Grade" variable. In this binary variable, the value was set to TRUE if the grade was either "D" or "HD," while all other grades were marked as FALSE.

For the development of the kNN model, a consistent k value of 5 was chosen across all case scenarios as it consistently yielded optimal results. In the case of the decision tree model, the J48 algorithm was selected due to its suitability for small datasets, as supported by previous literature (Priyam, Abhijeet, Gupta, Rathee, & Srivasta, 2013). This algorithm is known for predicting student performance while reducing overfitting. The minimum number of instances per leaf was set to 2, facilitating post-build pruning to eliminate unnecessary sub-trees. In pursuit of the highest accuracy, no specific minimum tree depth was enforced. In the training phase, both kNN and decision tree models were iteratively trained twice: first using the "*Grade*" variable, and subsequently by creating a new attribute, "*IsHighPerformance*" derived from the "*Grade*" variable.

The random forest classification model was used with the *bagSizePercent* parameter set to 100%, indicating that the entire dataset served as the training set. For optimal accuracy, an unlimited maximum tree depth was selected, with a value of 0 allowing the tree to grow unrestrictedly. This random forest ensemble consisted of 100 individual trees. The model was trained twice. The first iteration utilised the "Grade" variable, implementing one-hot encoding through the nominal-to-binary filter and configuring the *attributeIndices* to be aligned with the index of the "Grade" variable. In contrast, the second iteration employed a binary variable derived from the "*Grade*" variable, where the value was set to TRUE for grades "D" or "HD" and FALSE for all other grades.

All models underwent consistent performance evaluation, which included the generation of a confusion matrix and a Receiver Operating Characteristic (ROC) curve. The ROC curve serves as a graphical representation by plotting the true positive rate, often referred to as precision, against the false positive rate. This visualised the diagnostic prowess of a binary classifier system. The Area Under the Curve (AUC) in the ROC curve quantifies the model's ability to distinguish between the positive class (i.e., high-performing students) and the negative class. In essence, a higher AUC value indicates superior model performance, approaching ideal values of 1 or 0, signifying flawless classification. Conversely, an AUC of 0.5 suggests no predictive power, implying random class assignment for each record.

The primary objective of our model evaluation was to identify online student engagement behaviors that could predict high performance in online programming courses. Subsequently, we selected the model demonstrating the best overall performance. To facilitate a robust comparison between models, we leveraged performance metrics, particularly accuracy and the F1 score, based on test sets. To mitigate the risk of overfitting, we employed the k-fold cross-validation method with 10 folds. This approach allowed us to assess model performance, conduct an independence test, and gain valuable insights into its generalisation capabilities when applied to new data.

# 4. RESULTS AND ANALYSIS

The results and analysis in the following subsections explored and visualised the consistency of high-performing students in their series of introductory computing courses, and the associations between students' high academic performance and engagement behaviors in a fully asynchronous online degree.

## 4.1 Exploring Consistent High Performance of Students in Programming and Non-programming Courses

Table 3 suggests that 52.00% of the high-performing students in Programming 1 continue to obtain high grades in the succeeding programming course. 12.5% of the students who were non-high performing in Programming 1 and enrolled in Programming 2 had a better grade outcome and become high performing in Programming 2. The significant difference between the two percentages (52.00% - 12.50% = 39.50%) indicates that students are more likely to perform high in the successive courses (e.g., Programming 2) if they performed high in prerequisite courses (e.g., Programming 1). Similar observations have existed for students who are high performing in Programming 1 that enrolled in Programming 3 (71.43% continued to be high performing), and Programming 2 and then enrolled in Programming 3 (75% continued to be high performing). Note that 53.33% students who were in the non-high performing category in Programming 2 have performed well in Programming 3. This result is consistent with the experience factor that has been widely studied (Wilcox & Lionel, 2018). Although these studies looked programming experience prior to taking the introductory course, the experience factor can be applied in the study, that is programming experience influence better performance.

Table 3. Academic performances of students in the series of courses

| | Prog 1 to Prog 2 | Prog1 to Prog 3 | Prog 2 to Prog3 | IT Fund to Sys Analysis | Sys Analysis to Sys Design | IT Fund to Sys Design |
|---|---|---|---|---|---|---|
| Continued to be High performing | 52% | 71.43% | 75% | 41.38% | 68.75% | 25% |
| Non-high-performing to high-performing | 12.50% | 10% | 53.33% | 19.15% | 22.73% | 44.44% |

The academic performances of students in the non-programming courses (Table 3) show the percentages of students who have high-performance in pre-requisite that continue to have high performance in succeeding non-programming courses. However, compared to the programming courses, the percentages of high-performing students are lower. For high-performing students in IT Fundamentals that enrolled in System Analysis, 41.38% continued to perform well. For IT Fundamentals to System Design, only 25% continued to perform well. However, from Systems Analysis to System Design, the percentage of students who continually performed high was significant (68.75%). The above results indicated that out of six pairs of prerequisites and succeeding courses, five courses were observed that high-performing students in prerequisite courses had a much higher chance of being a high-performing student in the succeeding courses. Only one pair of courses (IT Fundamentals to System Design) showed the opposite trend. Compare to System Analysis and System Design where the contents are more related in the context of software development, IT Fundamentals covers a broad range of IT topics and may not be directly connected to System Design.

Students who were high-performing in the introductory programming courses had a higher chance of getting a high-performing grade in the succeeding programming courses, whereas in non-programming courses, it was observed that there was a lower percentage of students continuing to have high-performance if the pre-requisite course contents are not significantly related to the contents of the succeeding courses.

## 4.2 Online Engagement

Online activities are designed to help students progress their learning and prepare them for assessment tasks. The total views of formative activities in programming courses were 108,668 and non-programming courses was 126107 (see Figure 2).

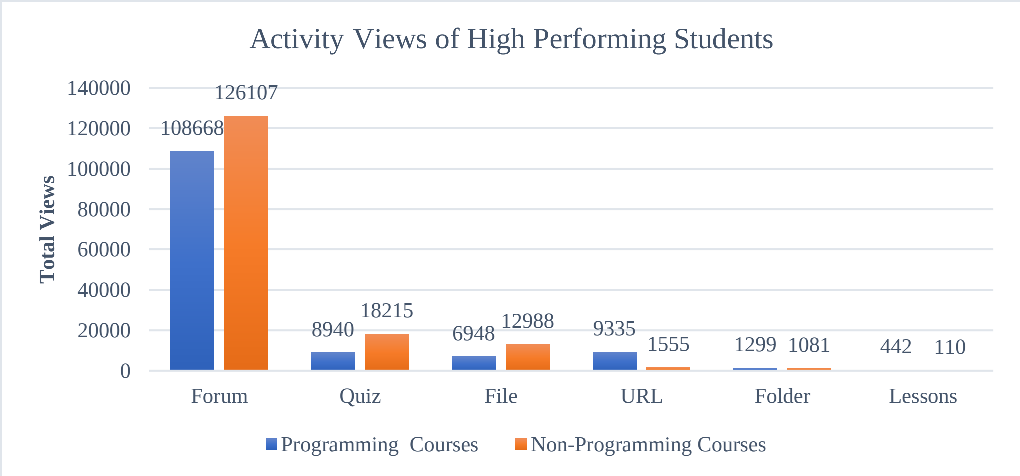


Figure 2. Total activity view for programming and non-programming courses

The next highest views were the URL and Quiz. The total views in forum posts were 91% higher than URL and 92% higher than quizzes. For non-programming courses, the total views in forum posts were 99% higher than URL and 86% higher than quizzes. This showed that high performing students significantly viewed the forum posts. Comparing to non-high performing students, the online engagement data of high-performing students (Figure 3) suggested that the distribution of forum views where the average views of high-performing students for both programming and non-programming courses was more than two-fold (112 % higher) than that of non-high performing students. The active use of forums by high-performing students is consistent with studies from Moström (2011) and Pedrosa et al. (2016) of successful programming students' behaviour in traditional classroom learning where they apply different strategies when they get stuck in programming. One common strategy is through social interaction with peers and teachers. In online learning, one equivalent form of interaction through peers and teachers is through engagement in forums. High-performing students' behaviour of having almost twice activity views than non-high performing is consistent with the self-regulated learner's behaviour of reflection. Viewing helps students to reflect and think about what they are seeing, which helps develop their skills and knowledge to analyse what they have just viewed. It is also consistent with the studies that indicate that students' viewing activities have direct positive influence on their completing learning tasks (Ma et al., 2015).

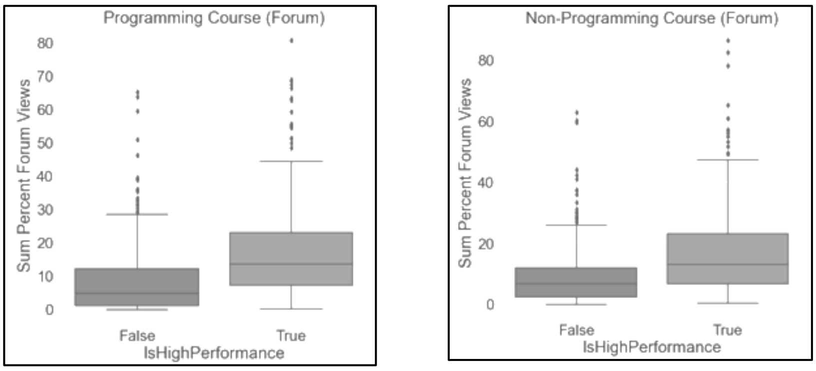


Figure 3. Boxplot of percentage forum views

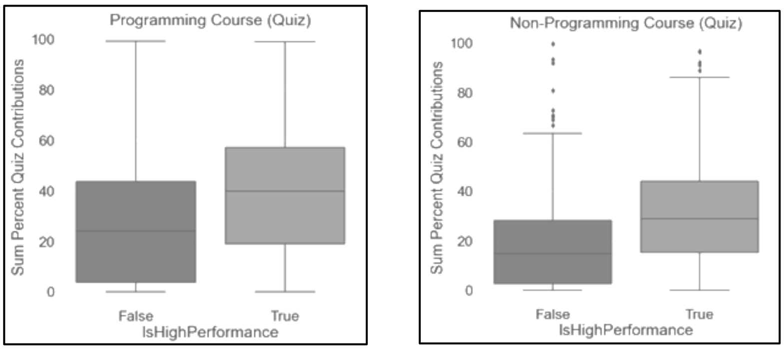


Figure 4. Boxplot of percentage quiz contributions

The quiz had significantly higher contributions for both programming and non-programming courses (57,393 total contributions for programming courses and 76,211 for non-programming). In contrast, forum contributions were significantly lower (90% lower for programming and 92% lower in non-programming courses). Figure 5 shows the comparison between forum and quiz contributions.

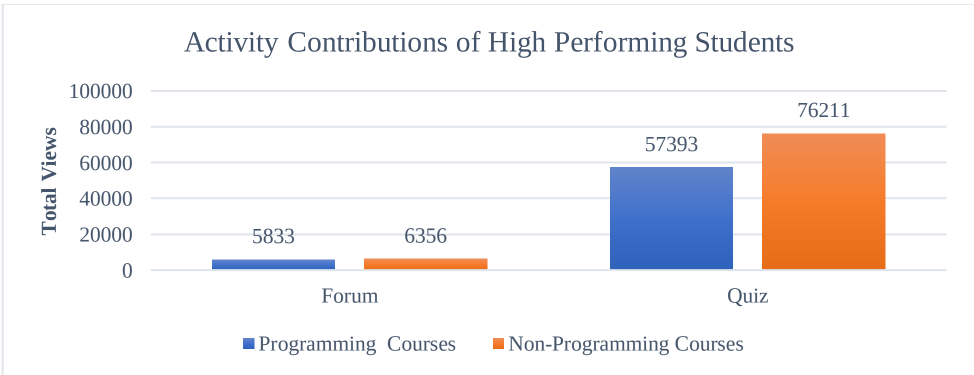


Figure 5. Total activity contributions for programming and non-programming courses

Other activities had zero contribution because these formative activities provide contents to the online students but were not designed for interaction with them. For quiz contributions (Figure 4), the percentage difference in the average quiz contributions in programming courses of high-performing students was 44% higher compared to non-high performing students, while for non-programming courses, the percentage difference was 112%. We observed that students who were high performing for all programming courses (Programming 1, Programming 2, and Programming 3) had a 159% difference in contributions and 198% difference in views. For non-programing courses (IT Fundamentals, System Analysis and System Design), students who were consistently high performing in these three courses had a 170% difference in views and 105% in contributions.

The quiz formative activity is the online activity where high-performing students contributed the most in programming courses and non-programing courses. This behaviour of having high contributions to formative activities such as quizzes in coding is consistent with the studies that high-performing students in programming education have been consistently active in practice as their study progresses (Hassinen & Mäyrä, 2006; Mai et al., 2021; Pedrosa et al., 2016).

### 4.2.1 Statistical Significance

Student's t-test has been performed to evaluate statistically if the mean values of online engagement between high performance and non-high-performance students are equal. Table 4 shows that all p-values of two online engagement data sets of both programming (8.99e-11 for forum views and 2.69e-9 for quiz contributions) and non-programming courses (9.83e-16 for forum views and 5.33e-15 for quiz contributions) are much lower than 0.05, indicating a rejection of the null hypothesis, i.e., the online engagement is statistically significantly different between high-performing and non-high performing students.

Table 4. P-values of the students' t-test on two different online engagement

| | Programming Courses | Non-Programming Courses |
|---|---|---|
| Sum Percentage Forum Views | 8.99e-11 | 9.83e-16 |
| Sum Percentage Quiz Contributions | 2.69e-9 | 5.33e-15 |

## 4.3 Academic and Engagement Data that Predicts a Student's High Performance

Six machine learning models were compared on accuracy, precision, recall, F1 score, and AUC. The comparison tables for each of the scenarios of all the cases are shown in the following subsections.

### 4.3.1 Using Students' Previous Academic Performance Data Only

In Scenario 1 where we aimed to predict future academic performance in programming courses using students' previous academic records, two standout models emerged: SVM and Logistic Regression. However, when it came to predicting performance in non-programming courses, a slightly different landscape emerged, with the top models being: Logistic Regression and kNN. Table 5 presents a comparison of all predictive models specifically for programming courses, while Table 6 provides a parallel analysis focused on non-programming courses.

Table 5. Comparing predict models' performances for programming courses

| | **Accuracy** | | | **Precision** | | | **Recall** | | | **F1 Score** | | | **AUC** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Models** | **C1*** | **C2*** | **C3*** | **C1** | **C2** | **C3** | **C1** | **C2** | **C3** | **C1** | **C2** | **C3** | **C1** | **C2** | **C3** |
| Logistic Regression | 67.2 | 71.7 | **71.2** | 66.7 | 72.5 | **70.6** | 94.7 | 93.6 | **68.6** | 78.3 | 81.7 | **69.6** | 73.0 | 74.0 | **76.0** |
| SVM | **75.4** | **78.3** | 71.2 | **82.5** | **78.4** | 66.7 | **86.8** | **93.5** | 74 | **80.5** | **84.8** | 68.7 | **70.9** | **75.0** | 78.5 |
| naïve Bayes | 68.9 | 73.91 | 68.5 | 72.1 | 78.8 | 63.6 | 81.6 | 83.9 | 80.0 | 76.5 | 81.3 | 70.9 | 66.9 | 69.7 | 74.7 |
| k-NN | 69.6 | 69.9 | 63.9 | 54.6 | 71.0 | 52.9 | 40.0 | 62.9 | 39.1 | 46.2 | 66.7 | 45.0 | 64.0 | 66.0 | 74.0 |
| Decision tree | 62.3 | 80.43 | 68.5 | 70.3 | 77.5 | 67.7 | 68.4 | 100 | 65.7 | 69.3 | 87.3 | 66.7 | 61.0 | 61.9 | 74.4 |
| Random forest | 63.9 | 73.9 | 68.5 | 72.2 | 80.7 | 67.7 | 68.4 | 80.7 | 65.7 | 70.3 | 80.7 | 66.7 | 72.0 | 71.0 | 68.0 |

*Case 1 –C1) - using Programming 1 to predict Programming 3; Case 2 (C2) – using Programming 2 to predict Programming 3; Case 3 (C3) – using Programming 1 to predict Programming 2

For Case 1, as shown in Table 5 (C1 columns), SVM outperformed other metrics except AUC. With the confusion matrix, 31 values were accurately classified as TRUE, 15 as FALSE, yielding a commendable AUC of 70.9% compared to alternative models.

In Case 2, which involved predicting Programming 3 based on Programming 2 (as depicted in Table 5 – C2 columns), the decision tree model displayed the highest values in three of the five metrics, but its low AUC score signaled suboptimal class distinction. Therefore, SVM was chosen for their leading AUC performance at 75% and strong rankings in three other metrics.

Moving on to Case 3, which focused on predicting Programming 2 using Programming 1, logistic regression was favored as the baseline model despite slightly lower values in some metrics when compared to other models (see Table 5 – C3 columns). Logistic regression's confusion matrix revealed accurate predictions of 24 high-performing students and 28 non-high-performing students, with a commendable AUC of 76%.

Table 6 illustrates the performance comparisons of predictive models for non-programming courses. Similar to the programming courses, there were three cases, namely Cases 4, 5, and 6.

Table 6. Comparations of predictive model performances for non-programming courses.

| | **Accuracy** | | | **Precision** | | | **Recall** | | | **F1 Score** | | | **AUC** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Models** | **C4*** | **C5*** | **C6*** | **C4** | **C5** | **C6** | **C4** | **C5** | **C6** | **C4** | **C5** | **C6** | **C4** | **C5** | **C6** |
| Logistic Regression | 91.7 | **97.3** | 83.3 | 93.3 | **97.4** | 88.0 | 91.7 | **97.3** | **83.3** | 91.9 | **97.3** | 83.8 | 96.9 | **100** | **100** |
| SVM | 89.7 | 81.1 | 83.3 | 91.3 | 81.0 | 86.7 | 89.7 | 81.1 | 83.3 | 89.4 | 81.0 | 81.5 | 88.5 | 80.4 | 75.0 |
| Naive Bayes | 82.8 | 81.1 | 83.3 | 86.9 | 81.0 | 88.9 | 82.8 | 81.1 | 83.3 | 81.9 | 81.0 | 83.8 | 92.2 | 90.0 | 87.5 |
| k-NN | **93.1** | 81.1 | 83.3 | **93.1** | 81.3 | 86.7 | **93.1** | 81.1 | 83.3 | **93.1** | 81.1 | 81.5 | 97.8 | 93.6 | **100** |
| Decision Tree | 89.7 | 81.1 | **100** | 91.3 | 81.0 | **100** | 89.7 | 81.1 | **100** | 89.4 | 81.0 | **100** | 98.6 | 80.4 | **100** |
| Random Forest | 89.7 | 81.1 | **100** | 91.3 | 81.5 | **100** | 89.7 | 81.1 | **100** | 89.4 | 80.7 | **100** | 98.6 | 93.6 | **100** |

*C4 - using System Analysis to predict System Design; C5 – using IT Fundamentals to predict System Analysis; C6 – using IT Fundamentals to predict System Design

In Case 4 where System Analysis served as the predictor for System Design (as presented in Table 6 - C4 columns), the kNN model performed the best across all metrics. With a notable accuracy of 93.1%, alongside impressive precision, recall, and F1 scores, it stood as an excellent model overall. The AUC further corroborated its performance, registering an impressive 97.8%.

Case 5 involved IT Fundamentals as the predictor for System Analysis, Table 6 (C5 columns) reveals that logistic regression leads across all metrics. A perfect AUC score of 100 lent further credence to its excellent predictive performance.

Lastly, in Case 6 where IT Fundamentals predicted System Design, logistic regression again excelled across all metrics (Table 6 - C6 columns). The confusion matrix confirmed its precision, correctly classifying 3 high-performing students and 2 non-high-performing students. Yet again, the AUC score attaining a perfect 100 underscored its exceptional predictive capabilities.

### 4.3.2 Using Previous Student Academic Performance and Online Student Engagement Data

The best models for scenario 2 were SVM and random forest for programming courses where both student academic performance and student engagement data were combined to predict high performing students. For non-programming courses, random forest and logistic regression were the best models.

Table 7 shows the comparison of the performances of the predictive models for programming courses while Table 8 shows the results for non-programming courses.

Table 7. Comparations of predictive model performances for programming courses

| | **Accuracy** | | | **Precision** | | | **Recall** | | | **F1 Score** | | | **AUC** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Models** | **C1*** | **C2*** | **C3*** | **C1** | **C2** | **C3** | **C1** | **C2** | **C3** | **C1** | **C2** | **C3** | **C1** | **C2** | **C3** |
| Logistic Regression | 77.4 | 75.6 | 67.7 | 76.0 | 75.6 | 69.4 | 100 | 100 | 71.4 | 86.4 | 86.1 | 70.4 | 76.0 | 50.0 | 77.0 |
| SVM | **81.1** | 80.5 | 81.5 | **85.7** | 78.9 | **86.7** | **94.7** | **96.9** | **74** | **88.1** | **88.2** | **81.8** | **75.4** | 80.0 | **86.4** |
| Naive Bayes | 69.8 | 70.7 | 70.8 | 84.3 | 88 | 75.0 | 71.0 | 71.0 | 68.6 | 77.1 | 78.6 | 71.6 | 75.1 | 68.1 | 59.0 |
| k-NN | 75.6 | 56.9 | 73.6 | 50.0 | 58.1 | 55.6 | 20.0 | 71.4 | 33.3 | 28.6 | 64.1 | 41.6 | 81.0 | 81.0 | 59.0 |
| Decision tree | 73.6 | 63.4 | 66.1 | 80.0 | 80.6 | 67.6 | 84.2 | 73.5 | 71.4 | 82.0 | 76.9 | 69.4 | 71.0 | 47.6 | 67.6 |
| Random Forest | 77.4 | **80.5** | 73.9 | 80.1 | **82.9** | 75.0 | **89.5** | 93.5 | 77.1 | 85.0 | 87.8 | 76.1 | 72.0 | **83.0** | 84.0 |

*C1 - using Programming 1 to predict Programming 3; C2 – using Programming 2 to predict Programming 3; C3 – using Programming 1 to predict Programming 2

In Case 1, the focus is on predicting Programming 3 using Programming 1 as the predictor, the standout model was SVM, as highlighted in Table 7 (C1 columns). This choice was driven by its impressive F1 score and AUC, which reached 75.4%—the second highest among the models evaluated. SVM exhibited robust predictive capabilities, with the confusion matrix accurately classifying 37 values as TRUE and 6 values as FALSE.

In Case 2, the aim was to predict Programming 3 based on Programming 2, the random forest model emerged as the most suitable choice. As depicted in Table 7 (C2 columns), this model showcased its performance compared to other models, boasting the highest AUC, the second highest F1 score, with a negligible gap between these two metrics. The confusion matrix further attested to its effectiveness, correctly classifying 29 values as TRUE and 4 values as FALSE, resulting in an AUC of 83%.

In Case 3, which involved predicting Programming 2 using Programming 1, the SVM model again displayed superior metrics compared to its counterparts (Table 7 - C3 columns). The confusion matrix reinforced its predictive prowess, accurately classifying 27 values as TRUE and 26 values as FALSE, with a notable AUC of 86.4%.

Table 8. Comparations of predictive model performances for non- programming courses

| | Accuracy | | | Precision | | | Recall | | | F1 Score | | | AUC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Models** | **C4*** | **C5*** | **C6*** | **C4** | **C5** | **C6** | **C4** | **C5** | **C6** | **C4** | **C5** | **C6** | **C4** | **C5** | **C6** |
| Logistic Regression | 69.2 | 74.1 | **75.0** | 73.5 | 73.2 | **82.1** | 69.2 | 74.1 | **75.0** | 66.8 | 73.5 | **70.8** | 81.0 | 75.0 | **100** |
| SVM | 53.8 | 56.8 | 62.5 | 53.8 | 56.8 | 62.5 | 53.8 | 56.8 | 62.5 | 70.0 | 56.8 | 76.9 | 50.0 | 50.0 | 50.0 |
| Naive Bayes | 61.5 | 64.9 | 62.5 | 77.6 | 65.3 | 62.5 | 61.5 | 64.9 | 62.5 | 52.9 | 80.7 | 76.9 | 66.1 | 76.2 | 68.8 |
| k-NN | 73.1 | 81.1 | 62.5 | 82.1 | 81.5 | 62.5 | 73.1 | 81.1 | 62.5 | 70.2 | 80.7 | 76.9 | 81.0 | 93.5 | 50.0 |
| Decision tree | 76.9 | 81.1 | 62.5 | 79.3 | 81.0 | 62,5 | 76.9 | 81.1 | 62.5 | 76.1 | 81.1 | 77.0 | 75.6 | 80.4 | 50.0 |
| Random forest | **76.9** | **83.8** | 50.0 | **79.3** | **84.0** | 35.7 | **76.9** | **83.8** | 50.0 | **76.1** | **83.6** | 41.7 | **78.9** | **93.9** | 53.3 |

*C4 - using System Analysis to predict System Design; C5 – using IT Fundamentals to predict System Analysis; C6 – using IT Fundamentals to predict System Design

The objective for Case 4 was to predict System Design based on System Analysis, the random forest model consistently demonstrated the highest performance across most metrics (Table 7 and 8 – C4 columns). The corresponding confusion matrix underscored its effectiveness, accurately classifying 13 out of 18 students as high-performing and 7 out of 8 as non-high-performing, ultimately yielding an impressive AUC of 78.9%.

In Case 5, which involved using IT Fundamentals as a predictor for System Analysis, random forest once again took the lead across all metrics (Table 8 - C5 columns), and resulting in a remarkable AUC of 93.9%.

Lastly, in Case 6 where IT Fundamentals was employed to predict System Design, logistic regression emerged as the top performer across all metrics (Table 8 - C6 columns) and achieving a perfect AUC score.

### 4.3.3 Summary of the Predictive Analytics Results

The analysis and experiments conducted in this study underscore the significance of leveraging students' previous academic performance as a robust predictor, with an average model accuracy of 75%. Specifically, it was found that academic performance in Programming 1 reliably forecasts high academic achievement in subsequent programming courses. Notably, among these programming courses, it is the academic performance in Programming 2, particularly in assessment 1, that emerges as the most effective predictor for high performance in Programming 3.

In contrast, for non-programming courses, the predictive power is even more compelling, showing an average accuracy of 91%. The academic performance in System Analysis is the top indicator for high academic performance in System Design, attaining a remarkable accuracy rate of 93.1%. In contrast, using academic performance in IT Fundamentals as a predictor yields a still respectable, yet comparatively lower, accuracy of 83%. It's important to note that relying solely on previous performance necessitates careful course selection as the predictor, given the subject matter correlation between the courses.

To further enhance the predictive accuracy, combining academic performance with student engagement data proves particularly effective for programming courses, resulting in 81%. accuracy However, this approach exhibits somewhat lower accuracy for non-programming courses. A notable constraint is the requirement for substantial engagement data, ideally spanning up to week 7 or 8, to empower the model's predictive capabilities. Consequently, intervention strategies, if necessary, can only be implemented in the final weeks of the study period. In light of this, it is advisable to explore a sequential model that relies solely on engagement data as a potential solution to address this temporal limitation.

# 5. CONCLUSION AND RECOMMENDATION

The study presented in this paper analysed three-year engagement and academic performance data of students enrolled in a 100% asynchronous online IT courses. Students' academic performance and online engagement data were collected from the LMS to provide a more objective view of data and explore the relationship between online engagement behaviors and academic performance. The data were further analysed to discover the patterns of online engagement behaviors of high-performing students. The study revealed that an average of 54.5% of high-performing students in the introductory programming courses continued to have high performance in the succeeding programming courses. High-performing students in the first programming courses have a higher chance (61.43%) of maintaining the high performance in Programming 3 than Programming 2. The study also showed that as students get more programming experience, that is after they have completed Programming 2, more than fifty percent of the non-high-performing students became high-performing in Programming 3. For non-programming courses, high-performing students in their first course (IT Fundamentals) has a lower percentage of achieving a high performance in the succeeding courses. Students who are high performance in System Analysis has a higher chance (46.025) of having consistent high performance in System Design. As the results indicated a continuity of high-performing students in an introductory course and the succeeding courses especially for programming courses, their online engagement behaviors were further explored.

The online engagement data of high-performing students is consistent with self-regulated learner behaviors. The average forum views of high-performing students for both programming and non-programming courses were 112 % higher than that of non-high performing students. In the quiz activity contributions, the average contributions of high-performing students for programming courses were 44% higher while 112% higher for non-programming courses comparing to non-high performing students. Online students use the online equivalent medium or tool to manifest the face-to-face student behaviors (e.g., social interaction with peers and teachers in face-to-face and the equivalent discussion forum tool in online learning as a form of social interaction).

This study also looked at the value of students' past academic performance as a predictor of their future achievements, demonstrating an accuracy of 75% for programming courses and 91% for non-programming ones. Specifically, Programming 1 results forecast success in later programming courses, with Programming 2's first assessment being the most predictive of high performance in Programming 3. For non-programming subjects, System Analysis stands out as the strongest predictor for success in System Design, with an accuracy of 93.1%. While the predictive accuracy was enhanced to 81% for programming courses when combined with student engagement data, this approach requires extensive data until at least week 7 or 8, delaying potential intervention strategies. To overcome this, exploring a model relying solely on engagement data is recommended.

Though this study utilised a data-driven approach based on learning analytics to understand the online engagement behaviors of high-performing students studying at a 100% asynchronous online IT courses, the study was limited to the data available through the LMS. Viewing of content videos is an activity that has been excluded due to the transition from one video tool to another which resulted in the inability to access the older data. Other learning analytics data points that are crucial for gaining deeper insights of students online learning behaviours are self-assessment data where students reflect on their progress, intervention and support data, and peer interaction data including collaborative activities. For future directions, we aim at further analysing the data collected in this study to identify the predictive factors that contribute to students' high performance and the continuity of their high performance and extending this study by understanding the mindset of these high-performing students when engaging and learning online.

## REFERENCES

Al-Barrak, M. and Al-Razgan, M. (2016). Predicting students final GPA using decision trees: A case study. *International Journal of Information and Education Technology*, Vol. 6, No. 7, pp. 528–533. https://doi.org/10.7763/IJIET.2016.V6.745

Aluko, R. et al. (2016). Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques. *Construction Economics and Building*, Vol. 16, No 4, pp. 86-98. https://doi.org/10.5130/AJCEB.v16i4.5184

Alqurashi, E. (2022). What do students engage with the most? A comparative study between high and low achieving students within online learning environments. *Open Learning: The Journal of Open, Distance and e-Learning*, Vol. 37, No. 3, pp. 219-234.

Araujo, T., Wonneberger, A., Neijens, P., and de Vreese, C. (2017). How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, Vol. 11, No. 3, pp. 173-190.

Beaulac, C. and Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, Vol. 60, No. 7, pp. 1048-1064. https://doi.org/10.1007/s11162-019-09546-y

Bretana, N. A. et al. (2020). Predicting Student Success for Programming Courses in a Fully Online Learning Environment. In H. So et al. (eds.) *Proceedings of the 28th International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education.

Burgos, C. et al. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, Vol. 66, pp. 541-556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Campbell, A., Craig, T. and Collier-Reed, B. (2020). A framework for using learning theories to inform 'growth mindset'activities. *International Journal of Mathematical Education in Science and Technology*, Vol. 51, No. 1, pp. 26-43.

Conole, G., Gašević, D., Long, P. and Siemens, G. (2011). Message from the LAK 2011 general & program chairs. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. Banff, Canada.

Duckworth, A. L. and Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational researcher*, Vol. 44, No. 4, pp. 237-251.

Dunlosky, J. et al. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, Vol. 14, No. 1, pp. 4-58.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random house.

Golino, H. and Gomes, C. (2014). Four machine learning methods to predict academic achievement of college students: a comparison study. *Revista Electrónica de Psicologia, Educação e Saúde*, Ano 4, Vol. 1, pp. 68-101.

Gotardo. M. A. (2019). Using decision tree algorithm to predict student performance. *Indian Journal of Science and Technology*, Vol. 12, No. 8, pp. 1-8. https://doi.org/10.17485/ijst/2019/v12i5/140987

Guo, W., Lau, K. L. and Wei, J. (2019). Teacher feedback and students' self-regulated learning in mathematics: A comparison between a high-achieving and a low-achieving secondary schools. *Studies in Educational Evaluation*, Vol. 63, pp. 48-58.

Hassinen, M. and Mäyrä, H. (2006). Learning programming by programming: A case study. *Proceedings of the 6th Baltic Sea Conference on Computing education research: Koli Calling 2006*, pp. 117-119.

Henrie, C. R., Bodily, R., Larsen, R., and Graham, C. R. (2018). Exploring the potential of LMS log data as a proxy measure of student engagement. *Journal of Computing in Higher Education*, Vol. 30, pp. 344-362.

Hu, Y., Lo, C.-L. and Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior, Vol. 36*, pp. 469-478. https://doi.org/10.1016/j.chb.2014.04.002

Johnston, O., Wildy, H., and Shand, J. (2019). A decade of teacher expectations research 2008–2018: Historical foundations, new developments, and future pathways. *Australian Journal of Education*, Vol. 63, No. 1, pp. 44-73.

Kaplan, H. (2018). Teachers' autonomy support, autonomy suppression and conditional negative regard as predictors of optimal learning experience among high-achieving Bedouin students. *Social Psychology of Education*, Vol. 21, pp. 223-255.

Kotsiantis, S., Pierrakeas, C. and Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, Vol. 18, No. 5, pp. 411-426. https://doi.org/10.1080/08839510490442058

Liu, D. Y.-T., Bartimote-Aufflick, K., Pardo, A. and Bridgeman, A. J. (2017). Data-driven personalization of student learning support in higher education. In A. Peña-Ayala (ed.) *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-learning*, Studies in Systems, Decision and Control, Vol. 94, pp. 143-169. Springer: Cham.

Liu, W. and Cheng, C. (2016). A hybrid method based on MLFS approach to analyze students' academic achievement. *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1625-1630. https://doi.org/10.1109/FSKD.2016.7603420

Ma, J., Han, X., Yang, J. and Cheng, J. (2015). Examining the necessary condition for engagement in an online learning environment based on learning analytics approach: The role of the instructor. *The internet and higher education*, Vol. 24, pp. 26-34.

Mai, T. T., Crane, M. and Bezbradica, M. (2021). Student behaviours in using learning resources in higher education: How do behaviours reflect success in programming education? *7th International Conference on Higher Education Advances (HEAd'21)*. Editorial Universitat Politecnica de Valencia, pp. 47-55.

Moström, J. E. (2011). *A study of student problems in learning to program*. Umeå University, Department of Computing Science, Sweden.

Mubarak, A. A., Cao, H. and Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, Vol. 30, Issue 8, pp. 1414-1433. https://doi.org/10.1080/10494820.2020.1727529

Pardo, A., Han, F. and Ellis, R. A. (2017). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies*, Vol. 10, No. 1, pp. 82-92. https://doi.org/10.1109/TLT.2016.2639508

Pedrosa, D., Cravino, J., Morgado, L. and Barreira, C. (2016). Self-regulated learning in computer programming: strategies students adopted during an assignment. In C. Allison et al. (eds.) *Proceedings of Immersive Learning Research Network: Second International Conference, iLRN 2016*. Springer: Santa Barbara, CA, USA.

Persky, A. M. (2018). A four year longitudinal study of student learning strategies. *Currents in Pharmacy Teaching and Learning*, Vol. 10, No. 11, pp. 1496-1500.

Pintrich, P. R. and De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, Vol. 82, No. 1, pp. 33-40.

Robinson, C. (2017). Growth mindset in the classroom. *Science Scope*, Vol. 41, Ed. 2, pp. 18-21.

Rosen, J. A., Porter, S. R. and Rogers, J. (2017). Understanding student self-reports of academic performance and course-taking behavior. *AERA Open*, Vol. 3, Issue 2. https://doi.org/10.1177/2332858417711427

Rovira, S. Puertas, E. and Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PloS One*, Vol. 12, No. 2, e0171207. https://doi.org/10.1371/journal.pone.0171207

Simanullang, N. and Rajagukguk, J. (2020). Learning Management System (LMS) based on Moodle to improve students learning activity. *Journal of Physics: Conference Series*, Vol. 1462, No. 1, 012067.

Soffer, T. and Cohen, A. (2019). Students' engagement characteristics predict success and completion of online courses. *Journal of Computer Assisted Learning,* Vol. 35, No. 3, pp. 378-389. https://doi.org/10.1111/jcal.12340

Toro-Troconis, M., Alexander, J., and Frutos-Perez, M. (2019). Assessing student engagement in online programmes: Using learning design and learning analytics. *International Journal of Higher Education*, Vol. 8, No. 6. https://doi.org/10.5430/ijhe.v8n6p171

Wang, C.-L. and Liou, P.-Y. (2018). Patterns of motivational beliefs in the science learning of total, high-, and low-achieving students: Evidence of Taiwanese TIMSS 2011 data. *International Journal of Science and Mathematics Education*, Vol. 16, pp. 603-618.

Zimmerman, B. J. and Schunk, D. H. (2001). *Self-regulated learning and academic achievement: Theoretical perspectives*. Routledge.