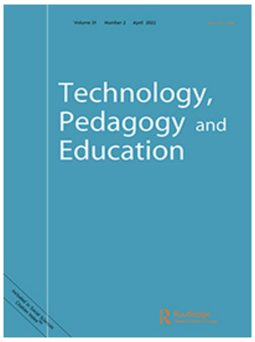

Routledge
Taylor & Francis Group

Technology, Pedagogy and Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rtpe20

# Using online student interactions to predict performance in a first-year computing science course

Sam Goundar, Arpana Deb, Goel Lal & Mohammed Naseem

**To cite this article:** Sam Goundar, Arpana Deb, Goel Lal & Mohammed Naseem (2022): Using online student interactions to predict performance in a first-year computing science course, Technology, Pedagogy and Education, DOI: 10.1080/1475939X.2021.2021977

**To link to this article:** https://doi.org/10.1080/1475939X.2021.2021977

TECHNOLOGY, PEDAGOGY AND EDUCATION
https://doi.org/10.1080/1475939X.2021.2021977

Routledge
Taylor & Francis Group

# Using online student interactions to predict performance in a first-year computing science course

Sam Goundar, Arpana Deb, Goel Lal and Mohammed Naseem

School of Computing and Information Sciences, The University of the South Pacific, Suva, Fiji

**ABSTRACT**

Student performance is a critical factor in determining a university's reputation because it has a negative effect on student retention. Students who do not perform well in a course are more likely to drop out from their programmes before graduating. Many students who enrol in Computing Science programmes struggle to find success because it is considered a difficult discipline. In this study, a sample of 918 observations were selected containing demographic and academic information about students enrolled in a first-year undergraduate Computing Science course at a university. Classification algorithms such as Decision Tree, Random Forest, Naïve Bayes and Support Vector Machine were used to build predictive models to determine whether a student will pass or fail the course. The results showed the Random Forest algorithms are capable of producing better predictive performance compared with traditional Decision Tree algorithms.

**ARTICLE HISTORY**
Received 14 March 2020
Accepted 10 September 2021

**KEYWORDS**
Predictive analytics; data mining; classification algorithms; random forest algorithm; decision trees

## Introduction

In recent years, technology has influenced the education sector (Abad-Segura et al., 2020; Alexander et al., 2019) and challenged the higher education universities to operate in a highly competitive environment (Subhash & Cudney, 2018). New technologies have been incorporated in the learning and teaching processes (Deng et al., 2019), and the demands on educational institutes to provide quality education (Latif et al., 2019) have increased. One of the most important factors when deciding on the quality of education a higher education institute provides is student performance (Suljic & Osmanbegovic, 2012). The success of the students at a higher education institute has become a concern owing to the fact that many students who tend to enrol are not able to complete their studies, i.e. the success rate is low, and the attrition rate is higher (Shaleena & Paul, 2015; Sukhija et al., 2015). The attrition and success rates are a vital measure of student performance, and, in order to improve these rates, performance measures need to be captured before students sit for their final examinations so that those showing to be at risk of low performance are identified earlier and remedial sessions can take place, enabling them to improve on their performance (Yassein et al., 2017). Students who do not actively participate, do not perform well and this eventually leads them to fail and drop out of the courses. At the university where this study was carried out, a grade point average (GPA) of less than 2.0 meant the student would not be allowed to continue their studies because of unsatisfactory academic performance.

**CONTACT** Sam Goundar sam.goundar@gmail.com School of Computing and Information Sciences, The University of the South Pacific, Suva, Fiji.

Supplemental data for this article can be accessed here

The use of educational data-mining techniques is the key to analysing these large volumes of educational data to find the different trends and patterns to predict student performance (Soni et al., 2018). One such prediction provided by data mining is to be able to understand if the learning outcomes of a course are being attained and to create a model that helps predict student performance measures such as their continuous assessments and even their grades. Such a learning model acts as an enabler to improve on teaching processes and gives educators data to help them prepare assessments (Shaleena & Paul, 2015; Suljic & Osmanbegovic, 2012). Higher education institutes are moving more into learning management systems, which is the source of most of the data needed for predictions; however, students' academic performance is affected by many factors such as personal, socio-economic and other environmental factors. The use of correct data-mining techniques can find the relationship and correlation between these factors and student performance (Yassien et al., 2017).

The application of data-mining techniques to large educational datasets yields many benefits to educators. According to Dutt et al. (2017), it enables educators and other interested stakeholders to analyse student motivation, attitude and behaviour, understand their students' learning styles, customise e-learning and promote collaborative learning. Research conducted so far has proved that students' performance can be predicted using a dataset consisting of students' gender, parental education, their financial background (Tie et al., 2010), attendance, performance in class tests and assignments in their studies (Chi et al., 2008). Nasiri et al. (2012) used regression analysis and classification (CS5.0 algorithm, which is a type of decision tree) to predict the academic dismissal of students and the GPA of graduated students in an e-learning centre. As universities compete to attract more students, educational data-mining tools are used to predict students' results to prevent drop-out and focus on both academically good and poor performers.

Self-efficacy is the belief about the personal capabilities to perform a task and attain the established goals. Over the years, self-efficacy has become one of the most important variables not only in research on motivation, but also in research on self-regulated learning (Schunk & Usher, 2011), and self-efficacy has therefore been incorporated into self-regulated learning models (Panadero, 2017). Academic self-efficacy is also an important factor when predicting academic success in college. Dewberry and Jackson (2018), when applying the Theory of Planned Behaviour to student retention, also found self-efficacy and academic integration to be critical factors in student retention. While looking at student retention models in higher education, Burke (2019) identified students' engagement and how well they integrated into the higher education environment (transition from secondary school to tertiary) as extremely important to student retention. Self-efficacy and academic integration also feature in studies by Drysdale and McBeath (2018), Jüttler (2020) and Tudor (2018).

This article comprises a study that was undertaken amongst first-year undergraduate Computing Science students at a university in the South Pacific region. Data-mining classification algorithms were used to predict the student performance for these students. Correlation analysis was used to find the association between the factors and the students' performance, and decision trees were used to build the predictive models to evaluate student success and failure rates. This research will look at how to use data extracted from the university's Learning Management System (LMS), namely Moodle, to ascertain how to best utilise the vast tools available on the platform and how they can be used to predict students' performance through the LMS. The analytics model that we will be proposing will provide us with trends and predictive behaviour of learners.

The following sections will first look at the background to provide a grounding for the research and explain why the research was undertaken. The background also looks at similar research carried out at other universities. After the background, a comprehensive literature review is provided to reinforce the theoretical foundations of this study on educational data mining. The methodology then follows, where this article defines its proposed model and data-collection techniques. The methodology will also explain data-pre-processing and data-processing methods. After the methodology, the article discusses its experimental set-up and explains its full dataset as used in the

R studio program. The results and findings will follow to explain the results against the four algorithms used to build the classification model. The discussion section shows data when compared with original results and explains what the authors thought to be a plausible solution and model. The article ends with a conclusion explaining what knowledge was gained and then puts forward recommendations for future research.

## Background

Educational data mining is the process of analysing data from different sources and datasets found within a Learning Management System and studying the patterns and relationships to derive meaningful information. Educational data mining provides information, knowledge and intelligence to the management team of any educational institution to analyse its educational system and students' performance. This analysis enables the educational institution to take stock of its current educational practices and institute changes for improvement and better student performance. Mining educational data reveals a number of cause-and-effect relationships that help determine the causes of student performance or the lack of it. Data can also be mined to search for associative, predictive or descriptive relationships. Educational data mining has become a trending topic of research within the field of educational technology.

Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimising learning and the environments in which it occurs. According to Viberg et al. (2018), learning analytics can improve learning practice by transforming the ways we support learning processes. They stated that 'overall there is little evidence that shows improvements in learner practice, however, the identified potential for improving learning support and teaching is high' (p. 98) and learning analytics might be the alternative. While delivering a keynote address on 'Learning Analytics: Utopia or Dystopia', Kirschner (2016) spoke about the myopic vision of what learning is under the learning analytics model. The dystopia is that learning analytics depends on purpose and thus might be used selectively, and the results might be biased. On the other hand, being able to predict what will happen, when and why is the utopia. Learning sciences theory is the missing link. As interest in learning analytics among higher education institutions grows, Tsai and Gasevic (2017) evaluated learning analytics policies and recommended adopting pedagogy-based approaches to learning analytics.

This study was conducted at a university within the South Pacific region. The university is a major provider of higher education for all the countries in the region. It has contributed towards many measures in improving the quality of degree-level programmes offered to the multiply culturally diverse students to enhance their learning ability and in producing graduates with excellent academic performance to meet current and future industry requirements in their field of study. In order to achieve this, the university has incorporated a number of technological developments and initiatives to improve the teaching and learning process of their students. One key source of data for the university is the LMS. The university has invested greatly in this area of technology-enhanced learning and has continuously introduced more learning tools to help engage students in an online environment. This study is based on the first-year undergraduate Computing Science students at the university.

The LMS, namely Moodle, has existed in this university for more than a decade. It provides flexible learning opportunities for students who are not able to attend face-to-face lectures on the main campus. Most of the courses are offered on the main campus in face-to-face mode. Students from the region and other campuses can enrol on the same course via online mode. The course coordinator will put all the course materials online organised into weeks and specify when assessments are due. Students will study the course materials and will have some face-to-face support from their local tutors. The LMS also provides a supplementary system for accessing course materials, assessments and participation and submitting assignments for students on the main campus who are enrolled to study in face-to-face mode. Recently, the university has started recording face-to-face lectures and providing them for students to view in their own time.

Other LMSs that have compared algorithms and provide tools for analysis are available on the market, but they are not in use at this university. For example, according to LinkedIn,

> Civitas Learning is a student success company that helps colleges and universities harness the power of their student data to improve outcomes. Our connected infrastructure and software ensure that higher education can better coordinate student success strategies, deliver proactive, collaborative care, power holistic advising, and quickly measure what's working for whom.[1]

EAB,[2] another LMS with learning analytics uses research, technology, data analytics and other services to universities to support students from enrolment to graduation and beyond. Rogers (2021) wrote that GoSkills, Moodle, TalentLMS and Thinkific are the four best free and open-source LMS tools. As the university where this study was conducted uses Moodle, the analytics were extracted from Moodle.

A number of similar studies has been carried out to mine educational data collected from online logs of students' interaction with their LMSs. Most of these studies have been done outside the region. A comprehensive literature review of all these research studies is provided in the next section. This university has been collecting its LMS's student interaction data for a number of years, but to date, no proper or formal research study has been done to mine this interaction data for predictive analytics and publish its findings. There have been a few internal reports coming from the team that administer the LMS and minor student projects within courses, but nothing as comprehensive as this.

It is expected that this research will assist in identifying students who are at risk of failing the courses in which they are enrolled, based on their interactions with the LMS. Having identified such students within the first few weeks from the beginning of the course will enable the course coordinators to intervene early and identify what is causing the lack of interaction, if any. Appropriate support and assistance can then be provided to these students.

## Literature review

Educational data mining (EDM) is a developing field based on data-mining techniques. EDM emerged as a combination of areas such as machine learning, statistics, computer science, education, cognitive science and psychometry. EDM focuses on learner characteristics, behaviours, academic achievements, the process of learning, educational functionalities, domain knowledge content, assessments and applications. EDM is defined by Baker (2010) as 'an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in'. EDM is concerned with improving the learning process and environment.

There has been increasing interest in data mining and educational systems, which turns educational data mining into a new and growing research area. Romero and Ventura (2007) explored the application of data mining to traditional educational systems, well-known learning content management systems, some web-based courses, and adaptive and intelligent web-based educational systems. They proposed that after pre-processing the available data in each case, the following data-mining techniques can be applied: statistics and visualisation, clustering, classification, outlier detection, association rule mining, pattern mining and text mining. The authors concluded that the success of the educational data mining needs much more specialised techniques.

The paper by Baker and Yacef (2009) reviewed the history and current trends in the field of EDM. The methodological profile of research in the early years of EDM is compared to that in 2008 and 2009, and the status explored. The trends and shifts in the research conducted by this educational data-mining community are discussed. The increased emphasis on prediction, the emergence of work using existing models to make scientific discoveries ('discovery with models') and the reduction in the frequency of relationship mining within the EDM community are also discussed. Two ways that researchers have attempted to categorise the diversity of research in educational data-mining research are explored, and the types of research problems that these methods have been used to address are reviewed. The most cited papers in EDM between 1995 and 2005 are listed, and their influence on the EDM community (and beyond) is discussed.

Educational data mining as an emerging interdisciplinary research area was explored by Romero and Ventura (2010). They specifically focused on areas that deal with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyse educational data in order to study educational questions. This paper surveys the most relevant studies carried out in this field to date. Firstly, it introduces EDM and describes the different groups of users, types of educational environments and the data they provide. It then goes on to list the most typical/common tasks in the educational environment that have been resolved through data-mining techniques, and finally some of the most promising future lines of research are discussed.

In an attempt to preserve and enhance the literature on recent EDM, Peña-Ayala (2014) looked into several constructs of educational data mining including development in advances, and to organise, analyse and discuss the content of the review based on the outcomes produced by a data-mining approach. Thus, as result of the selection and analysis of 240 EDM works, an EDM work profile was compiled to describe 222 EDM approaches and 18 tools. A profile of the EDM works was organised as a raw database, which was transformed into an ad-hoc database suitable to be mined. As result of the execution of statistical and clustering processes, a set of educational functionalities was found, a realistic pattern of EDM approaches was discovered, and two patterns of value-instances to depict EDM approaches based on descriptive and predictive models were identified. One key finding is that most of the EDM approaches are grounded on a basic set each composed of three kinds of educational systems, disciplines, tasks, methods and algorithms. The review concludes with a snapshot of the surveyed EDM studies, and provides an analysis of the EDM strengths, weakness, opportunities and threats, whose factors represent, in a sense, future work to be fulfilled.

The research paper by Papamitsiou and Economides (2014) aimed to provide the reader with a comprehensive background for understanding current knowledge on learning analytics (LA) and EDM and its impact on adaptive learning. The paper constitutes an overview of empirical evidence behind key objectives of the potential adoption of LA/EDM in generic educational strategic planning. The literature on experimental case studies conducted in the domain during the previous six years is examined (2008–13). Search terms identified 209 mature pieces of research work, but inclusion criteria limited the key studies to 40. The authors collated and categorised the research questions, methodology and findings of these published papers and organised them accordingly for further exploration by others. Non-statistical methods to evaluate and interpret findings of the collected studies are studied. The results highlighted four distinct directions of the LA/EDM empirical research. The emerged added value of LA/EDM research and the significance of further implications are explored.

According to Sachin and Vijay (2012), educational data mining is an emerging trend, concerned with developing techniques for exploring and analysing the huge amount of data that come from the educational context. EDM is poised to leverage an enormous amount of research from the data-mining community and apply that research to educational problems in learning, cognition and assessment. In recent years, EDM has proven to be more successful at many of the problems relating to educational statistics owing to enormous computing power and data-mining algorithms. This paper surveys the history and applications of data-mining techniques in the educational field. The objective is to introduce data mining to traditional educational systems, web-based educational systems, intelligent tutoring systems and e-learning. The paper describes how to apply the main data-mining techniques such as prediction, classification, relationship mining, clustering and social area networking to educational data.

The paper by Huebner (2013) discussed the emerging discipline of EDM that focuses on applying data-mining tools and techniques to educationally related data. The discipline focuses on analysing educational data to develop models for improving learning experiences and institutional effectiveness. A literature review on educational data-mining topics covers topics such as student retention and attrition, personal recommender systems within education, and how data mining can be used to analyse course management system data. Gaps in the current literature and opportunities for further research are presented.

Newly developed web-based educational technologies offer researchers unique opportunities to study how students learn and what approaches to learning lead to success (Minaei-Bidgoli et al., 2003). Web-based systems routinely collect vast quantities of data on user patterns, and data-mining methods can be applied to these databases. This paper presents an approach to classifying students in order to predict their final grades based on features extracted from logged data in an education web-based system. We design, implement and evaluate a series of pattern classifiers and compare their performance on an online course dataset. A combination of multiple classifiers leads to a significant improvement in classification performance. Furthermore, by learning an appropriate weighting of the features used via a genetic algorithm (GA), we further improve prediction accuracy. The GA is demonstrated to successfully improve the accuracy of combined classifier performance by about 10 to 12% compared with a non-GA classifier. This method may be of considerable usefulness in the early identifying of students at risk, especially in very large classes, and in allowing the instructor to provide appropriate and timely advice.

A paper by Ahmed and Elaraby (2014) stated that there is a huge amount of data stored in educational databases and these databases contain useful information for the prediction of students' performance. The most useful data-mining technique in educational databases is classification. In this paper, the classification task is used to predict the final grade of students, and, as there are many approaches used for data classification, the decision tree (ID3) method is used.

The main objective of higher education institutions is to provide quality education to their students. One way to achieve the highest level of quality in higher education is by discovering knowledge in order to make predictions regarding the enrolment of students in a particular course, alienation from the traditional classroom teaching model, detection of unfair means used in online examinations, detection of abnormal values in students' result sheets, prediction of students' performance and so on. The knowledge is hidden in the educational dataset and it is extractable through data-mining techniques. A research paper by Baradwaj and Pal (2012) justified the capabilities of data-mining techniques in the context of higher education by offering a data-mining model for the higher education system in the university. In their research, a classification task was used to evaluate students' performance, and, as there were many approaches that were used for data classification, the decision tree method was used there.

Web-based educational systems collect large amounts of student data, from web logs to much more semantically rich data contained in student models. Whilst a large focus of *AIED* research is to provide adaptation to a learner using the data stored in his/her student model, Merceron and Yacef (2005) explored ways to mine data in a more collective way: just as a human teacher can adapt to an individual student, the same teacher can also learn more about how students learn, reflect and improve their practice by studying a group of students. The field of data mining is concerned with finding new patterns in large amounts of data. Widely used in business, it has scarce applications to education. Of course, data mining can be applied to the business of education, for example, to find out which alumni are likely to make larger donations. The authors were interested in mining student models in a pedagogical perspective. The goal of their project was to define how to make data possible to mine, to identify which data-mining techniques are useful and understand how to discover and present patterns that are pedagogically interesting both for learners and for teachers.

Educational data mining is concerned with developing methods for discovering knowledge from data that come from the educational domain. Abu Tair and El-Halees (2012) used educational data mining to improve graduate students' performance and overcome the problem of low grades. In their case study they extracted useful knowledge from graduate students' data collected from the college of Science and Technology – Khanyounis. The data included a period of 15 years (1993–2007). After pre-processing the data, they applied data-mining techniques to discover association, classification, clustering and outlier detection rules. In each of the four tasks, they presented the extracted knowledge and described its importance in the educational domain.

The extant literature also shows that there were collections of reviewed papers that cover the important aspects of data in educational research that come from educational sources (systems and processes). EDM is an emerging data-mining field that is concerned with developing methods for exploring the educational data (Yahya, 2017). The data-mining field plays a fundamental role in EDM, for example, education teachers and instructors are classifying students according to their behaviour and level of knowledge and motivation.

According to Depren et al. (2017), using data-mining techniques in the field of education provides educators and education planners with a better understanding of huge datasets that include hidden useful patterns. In recent years, several studies falling under the EDM concept have been done to investigate students' academic performances at the national level. It was discovered that using educational data mining minimises the issue of inaccuracy in generating reports and saves a lot of time. Aldowah et al. (2019) found that specific EDM and LA could offer the best means of solving certain learning problems. Applying EDM and LA in higher education can be useful in developing a student-focused strategy and providing the required tools that institutions will be able to use for the purposes of continuous improvement.

Data-mining methods are often implemented at advanced universities today for analysing available data and extracting information and knowledge to support decision making (Kabakchieva, 2013). However, universities today are operating in a very complex and highly competitive environment. The main challenge for modern universities is to deeply analyse their performance, to identify their uniqueness and to build a strategy for further development and future actions. Further to this, university management should focus more on the profile of the students admitted, becoming aware of the different types and specific students' characteristics based on the data received. They should also consider if they have all the data needed to analyse students at the entry point of the university or they need other data to help managers support their decisions such as how to organise the marketing campaign and approach promising potential students (Kabakchieva, 2013).

A study was conducted by Ramaswami and Bhaskaran (2009) in India on the essence of data-mining concepts used in the educational field for the purpose of extracting useful information on the behaviours of students in the learning process. The results of the present study effectively support the well-known fact of the increase in predictive accuracy with the existence of a minimum number of features. The expected outcomes show a reduction in computational time and constructional cost in both training and classification phases of the student performance model. Educational data mining has provided useful benefits towards students' performances.

Much research has been done on EDM. Earlier research has tended to concentrate on different aspects of EDM. It has branched out into different aspects and is considered as one of the most appropriate technologies in providing knowledge. Research by Chalaris et al. (2014) was conducted in different timelines to find out different types of educational systems and how data mining can be applied in each system.

Data mining is useful in the field of education especially in student learning platforms in online environments. According to Mohamad and Tasir (2013), EDM often stresses improvement in motivation, metacognition and attitudes. The purpose of this literature review has been to outline why educational data mining plays a vital role in the educational sector and how it benefits instructors and students. The intention in this study is a rather broad review to show the necessity of EDM and uncover how EDM benefits the LMS.

EDM extends the benefits to LMS and provides insight to users of LMS as stated by Hämäläinen and Vinni (2010). EDM has become a very new and common tool in education systems that is used for describing the research disciplines which use data from educational settings. Sukhija et al. (2015)

discussed various methods and techniques for exploring the data collected from different educational sources, and they categorised the data in two forms, which are offline data which can be collected from the traditional databases of the educational institutions and online data which can be collected from the online media.

An investigation by Prabha and Shanavas (2014) discovered that EDM research uses technical methods such as prediction, clustering, relationship mining and modelling, thus working from student data can help educators both track academic progress and understand which instructional practices are effective. Most of the previous studies on educational data mining in various fields use a variety of data types ranging from text to images and which are stored in a variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety of databases.

## Research methodology

Based on a search of Google Scholar, Scopus and Web of Science databases, we have found that data-mining techniques for educational data mining have so far used the following methods: mostly different techniques converge, including clustering, decision trees, classification, neural networks, sequential patterns and association rules. A research paper by Manjarres et al. (2018) looked at more than 100 published papers on educational data mining between 1993 and 2016 and reported that data-mining techniques were used to analyse, understand or solve a particular situation in an educational environment. This indicates solid support for using data-mining techniques for educational data mining. The multivariate analysis technique examines the relationship between several categorical independent variables and two or more metric dependent variables. There are a few publications where techniques in multivariate analysis have been used for educational data mining, for example (Abdi et al., 2019; Alzahrani & Stojanovski, 2018; Hu & Rangwala, 2020). However, none of them were used to predict performance. We believe the two techniques should complement each other, as both of them have something to offer.

We used the Knowledge Discovery in Databases (KDD) steps in this educational data-mining project. It is a widely used (Guleria & Sood, 2014; Shruthi & Chaitra, 2016) iterative process for knowledge extraction from a collection of data (Noah et al., 2013). Priyadharsini and Thanamani (2014) defined KDD as 'the broad process of finding knowledge in data & Rangwala, the "high-level" application of particular data mining methods' (p. 1571).

KDD has gained the interest of researchers in the field of machine learning, pattern recognition, predictive modelling and data visualisation, to name a few. In the work of Kaur et al. (2015), the authors used the steps of the KDD process to predict slow learners so that teachers can provide individual assistance. In similar research by Pratiwi (2013), KDD steps were applied to discover important knowledge from data about high school students and to apply classifier models to predict student placement.

The KDD process is simple to use, which was the main reason for its selection in this research. The use of a standardised process such as KDD ensures that there is a systematic and easy step to follow to find knowledge in data in the context of large databases. It provides the opportunity to refine the extraction of knowledge through its iterative process. Additionally, it allows the enhancement of the evaluation measures and the integration of new data to obtain more results.

The KDD framework compromises of four main steps:

1. Dataset selection– in this step a dataset is created from which knowledge will be extracted.

In this research, we used students' institutional data stored in a university's LMS database. Five years of data (2014–2018) were gathered about students on a first-year computer science programming course. Students' enrolment data were extracted from the Students Online Services (SOLS), while course interaction and assessment data were extracted from Moodle, which is the LMS used by the university. The target dataset compromised of student demographic variables, course interaction variables and course assessment variables (Table 1).

2. Pre-processing – in this step, the data is cleaned.

Some variables contained missing values. The cleaning method is described below:

- Instances containing missing values for high school mathematics grades were completely removed from the dataset.
- There were also missing values for assignment 1 and assignment 2, which were replaced by the respective average values.

The final dataset was compromised of 918 observations and 18 variables. The decision variable (GRADE) was a binary attribute with two levels **F** (fail grade) and **P** (pass grade).

Table 1. Description of variables used in prediction.

| Variables | Data Type | Description |
|---|---|---|
| CAMPUS | Factor | Campus where student is taking the course. 10 levels 'Alafua','Kiribati'. |
| PROGRAMME | Factor | 7 levels 'BA','BAGCED'. |
| AGE | Factor | Student age at enrolment. 5 levels '<25', '>40', '26–30'... |
| STUDY_TYPE | Factor | 2 levels 'FULL TIME','PART TIME' |
| SEX | Factor | 2 levels 'F','M' |
| MARITAL_STATUS | Factor | 2 levels 'M','S' |
| DISABILITY | Factor | 2 levels 'N','Y' |
| NATIONALITY | Factor | Country of residence. 10 levels 'China', 'Fiji'... |
| SEC_SCH | Factor | Highest secondary school level attended. 5 levels 'Foundation', 'Other' |
| SEC_MATH | Factor | Secondary school Maths grade attained. 9 levels 'A', 'A+', 'B', 'B+'... |
| SPONSOR | Factor | 2 levels 'Private', 'Sponsored' |
| ASSIGNS_SUB | Integer | Number of assignments submitted CS111 in Moodle dropboxes. |
| QUIZ_ATTEMPT | Factor | Number of Moodle quizzes attempted in CS111. 5 levels '11 to 15', '16 to 20'... |
| FORUM_POSTS | Factor | Number of posts made in CS111 Moodle discussion forums. 6 levels '1 to 5', '11 to 15'... |
| FREQ_CP_ACCESS | Factor | Number of times the CS111 Moodle course page accessed. 7 levels '1 to 100', '101 to 200'... |
| A1_MARK | Factor | Assignment 1 mark attained in CS111. 9 levels 'A', 'A+', 'B', 'B+'... |
| A2_MARK | Factor | Assignment 2 mark attained in CS111. 9 levels 'A', 'A+', 'B', 'B+'... |
| CW | Factor | Final Course work attained in CS111. 8 levels 'A', 'B', 'B+', 'C'... |
| GRADE | Factor | Class variable with 2 levels 'F' for fail grade, 'P' for passing grade |

3. Data mining – this step requires choosing the appropriate data-mining tasks.

This is a classification problem, so we decided to use some popular data-mining classification algorithms from the literature. The classifiers used to build the predictive models included a Decision Tree classifier, a Naïve Bayes classifier and a Support Vector Machine classifier. We also used an ensemble algorithm called Random Forest.

Decision Tree is a classification algorithm which has a tree-like structure providing a graphical depiction of all possible solutions to a decision. The nodes in a decision tree represent a test on an attribute; the branches represent the outcomes of the test while the class label is represented by the leaf node (Sharma & Kumar, 2016). Decision Tree algorithms are popular in the literature for predictive modelling because domain knowledge is not required to build a decision tree classifier and due to its ability to handle data with large numbers of attributes (Kotsiantis, 2013).

The Naïve Bayes algorithm is based on Bayes' Theorem, which uses conditional probabilities and the maximum likelihood occurrence. The Naïve Bayes algorithm assumes that the attributes are not correlated to each other, implying that they should be independent (Xu, 2018). It is a simple classification algorithm which works quite well with real-world situations, and it is widely used for classification and prediction (Wei et al., 2011).

Support Vector Machine (SVM) is a popular machine-learning algorithm for classification and regression problems. It is a distance-based algorithm which is well known for its robust mathematical theory. The output of an SVM is an optimal hyperplane (a separating line that maximises the margin between training data and the class boundary). The input data are transformed to a higher dimension using a kernel function in which the SVM attempts to find an optimal hyperplane. The decision function of SVM is then based on the support vectors that are closest to the decision boundary (Parikh & Shah, 2016). The choice of this algorithm in this work was influenced by the fact that it has been found to produce better predictive performance compared with other classification techniques (Huang et al., 2017).

Random Forest is an ensemble classifier technique based on the decision tree algorithm. The Random Forest classifier uses bagging and attributes randomness to generate multiple decision trees where each individual tree has its own class prediction. The final class of the test object is determined by aggregating the outcomes of individual trees (Breiman, 2001). It is considered an effective algorithm due to the notion of 'the wisdom of crowds', where the class of an unknown data is determined by the behaviour of a group of uncorrelated trees rather than those that are based on individual trees (Yiu, 2019).

The tool used in this work to carry out the data-mining tasks was **R**, which is a free software environment widely used by many researchers for statistical analysis and for building data-mining predictive models.

4. Interpretation and evaluation – this step focuses on the clarity and usefulness of the model.

Evaluation techniques such as accuracy, specificity, sensitivity and kappa were employed in this step to verify the performance accuracies of the various classification models built during experimentation.

***Accuracy*** is the measure of the model's ability to differentiate between the pass and fail classes correctly, while the ***specificity*** of a model is its ability to determine the pass cases correctly. The ***sensitivity*** of a model is its ability to determine the failure cases correctly.

The calculation for the model's accuracy, specificity and sensitivity is represented by equation 1, 2 and 3, respectively:

## Results and findings

The objective of the article is to see the possibility of predicting data from Moodle logs and demographic data provided by the Student Academic Services. Four algorithms are used to build the classification model using the **R** studio application as the data-mining tool. The algorithms used are Decision Tree, Random Forest, Naïve Bayes and Support Vector Machine.

Our results indicate that student interactions have a direct impact on student retention. Students that interact more perform better and they continue with their studies, while those that do not interact much, fail and drop out. At the university where this study was carried out, a GPA of less than 2.0 meant the student will not be allowed to continue their studies because of unsatisfactory academic performance. According to Caruth (2018), the variables used in the prediction are not the only incentive to measure student performance because there are other reasons such as interactions with the course coordinators, their interest in learning and their expectations which also determine student performance. From the pedagogical point of view, the link between student engagement and student retention also depends on the resources available online for students and the curriculum design of the course (Kahu & Nelson, 2018; Pechenkina et al., 2017; Tight, 2020).

Table 2 depicts the number of instances correctly and incorrectly classified together with its prediction accuracy of the four models used for analysis in the article. From a total of 174 instances, Random Forest model had the highest number of instances correctly identified, followed by Support Vector Machine with 161 out of 174 instances correctly identified. Naïve Bayes had the lowest number predicted correctly with a total of 147 instances over 174 instances.

**Table 2.** Predictive performance of the classifiers.

| EVALUATION CRITERIA | DECISION TREE | RANDOM FOREST | NAÏVE BAYES | SVM |
|---|---|---|---|---|
| Correctly Classified | 151 | 163 | 147 | 161 |
| Incorrectly Classified | 23 | 11 | 27 | 13 |
| Prediction Accuracy | 86.78% | 93.67% | 84.48% | 92.53% |

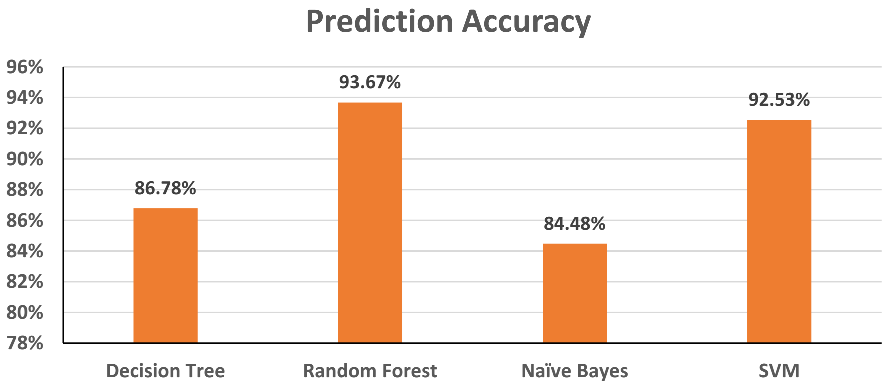

Figure 1. Prediction accuracy.

Relationship mining can identify relationships among variables and encode them in rules for later application. Through relationship mining, with the 918 datasets, we found the following pattern and trend: student success was related to the number of times students accessed their course Moodle page, number of times they posted on their course discussion forum, number of quizzes they attempted, number of assignments submitted, their secondary school mathsgrade and their highest secondary school level attended. Ensemble techniques create multiple models and then combine them to produce improved results. These techniques usually produce more accurate solutions than a single model would. This has been the case in several EDM scenarios, where the best results were derived from ensemble techniques (Ajibade et al., 2019; Ashraf et al., 2020; Kumari et al., 2018). The graph in Figure 1 shows that Random Forest has the highest prediction accuracy of 94% followed by Support Vector Machine with 93%. The Decision Tree model gave an accuracy of 87%. The lowest prediction accuracy was retrieved from Naïve Bayes with 84%.

Table 3 shows the confusion matrix of data accumulated using the Decision Tree model. This model correctly classified 31 out of 37 students who passed the course with a true positive rate of 65% and a true negative rate of 95%.

Table 3. Confusion matrix: Decision Tree.

| CLASS | FAIL | | PASS | | PRECISION |
|---|---|---|---|---|---|
| Fail | True Positive | 31 | False Positive | 6 | 83.78% |
| Pass | False Negative | 17 | True Negative | 120 | 87.59% |
| | SENSITIVITY (TPR) | 64.58% | SPECIFICITY (TNR) | 95.24% | |

Table 4 shows the confusion matrix of data accumulated using the Random Forest model. This model correctly classified 40 out of 43 students who passed the course with a true positive rate of 83% and a true negative rate of 98%.

Table 4. Confusion matrix: Random Forest.

| CLASS | FAIL | | PASS | | PRECISION |
|---|---|---|---|---|---|
| Fail | True Positive | 40 | False Positive | 3 | 93.02% |
| Pass | False Negative | 8 | True Negative | 123 | 93.89% |
| | SENSITIVITY (TPR) | 83.33% | SPECIFICITY (TNR) | 97.62% | |

Results for Naïve Bayes in Table 5 were similar to the results for the Decision Tree with low true positive rate of 67% and high true negative rate of 91%. True positive rates for the Random Forest model (Table 3) and the SVM model (Table 6) were significantly high with 83% and 81% respectively compared with the other two models. The true negative rates were also relatively high for these two models.

**Table 5.** Confusion matrix: Naïve Bayes.

| CLASS | FAIL | | PASS | | PRECISION |
|---|---|---|---|---|---|
| Fail | True Positive | 32 | False Positive | 11 | 74.42% |
| Pass | False Negative | 16 | True Negative | 115 | 87.79% |
| | SENSITIVITY (TPR) | 66.67% | SPECIFICITY (TNR) | 91.27% | |

Table 6 shows the confusion matrix of data accumulated using the Support Vector Machine model. This model correctly classified 39 out of 43 students who passed the course with a true positive rate of 81% and a true negative rate of 97%.

**Table 6.** Confusion matrix: Support Vector Machine.

| CLASS | FAIL | | PASS | | PRECISION |
|---|---|---|---|---|---|
| Fail | True Positive | 39 | False Positive | 4 | 90.70% |
| Pass | False Negative | 4 | True Negative | 122 | 93.13% |
| | SENSITIVITY (TPR) | 81.25% | SPECIFICITY (TNR) | 96.83% | |

Table 7 compares the TP (True Positive) Rate and the Precision of the four models against the classes Pass and Fail.

**Table 7.** Comparison of evaluation measure by TP rate and precision.

| | DECISION TREE | | RANDOM FOREST | | NAÏVE BAYES | | SVM | |
|---|---|---|---|---|---|---|---|---|
| CLASS | TP Rate | Precision | TP Rate | Precision | TP Rate | Precision | TP Rate | Precision |
| Fail | 0.646 | 0.838 | 0.833 | 0.930 | 0.667 | 0.744 | 0.813 | 0.907 |
| Pass | 0.952 | 0.876 | 0.976 | 0.939 | 0.913 | 0.878 | 0.968 | 0.931 |

True positive rate (TPR) is also called the sensitivity which measures the proportion of the true positives which have been correctly identified.

$$TP\ Rate = \frac{True\ Postive\ (TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

Precision also known as positive predictive value is the proportion of positive and negative values in the statistics.

$$Precision = \frac{True\ Postive\ (TP)}{True\ Positive(TP) + False\ Positive(FN)}$$

According to Table 7, the Random Forest model has the highest TP rate and precision for both classes, Fail and Pass, 83% and 98% respectively. The lowest TP rate for the Fail class is denoted in the Decision Tree model with 65% TP rate, and the lowest TP rate for the Pass class is denoted in the Naïve Bayes model with 91% TP rate.

This result indicates high TP rates for the Random Forest and SVM models in the Pass category with 98% and 97% respectively. It is also noted that the TP rate of the Fail category is significantly low when compared with the Pass category.

The comparison of the results of the four models in Table 8 and Figure 2 shows that the Random Forest and SVN models predicted the pass and fail rate of students with a high accuracy; 94% and 92% respectively.

Table 8. Comparison of evaluation measure by accuracy, Kappa, sensitivity and specificity.

| CLASSIFICATION MODEL | ACCURACY | KAPPA | SENSITIVITY | SPECIFICITY |
|---|---|---|---|---|
| Decision Tree | 0.8678 | 0.6439 | 0.6458 | 0.9524 |
| Random Forest | 0.9368 | 0.8365 | 0.8333 | 0.9762 |
| Naïve Bayes | 0.8448 | 0.5987 | 0.6667 | 0.9127 |
| SVM | 0.9253 | 0.8068 | 0.8125 | 0.9683 |

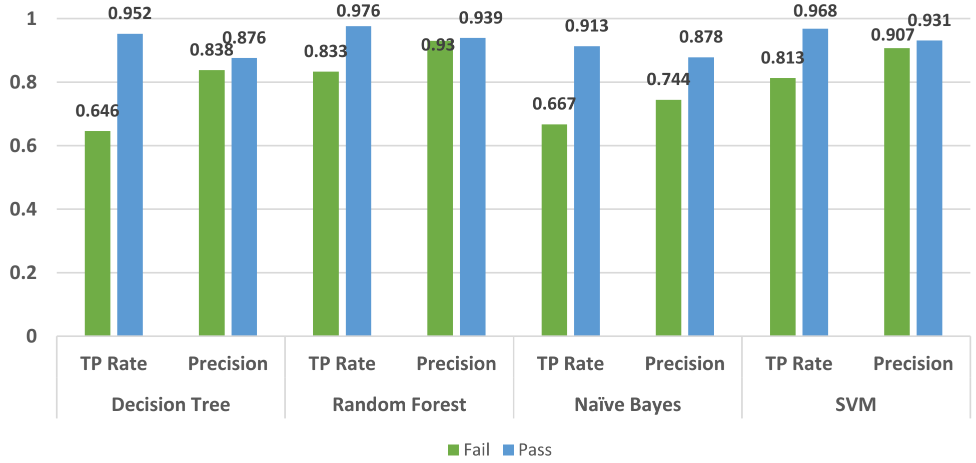

Figure 2. TP rate and precision for the four models.

The best model for educational data mining after running through the analysis is the Random Forest Model as it has the highest prediction accuracy compared with the other models.

Four-classification model algorithms (Decision Tree, Random Forest, Naïve Bayes and Support Vector Machine) were used to mine 918 student interaction logs data with 18 variables using R Studio. The prediction accuracy with the respective classification algorithm was as follows: Random Forest (93.67%), Support Vector Machine (92.53%), Decision Tree (86.78%) and Naïve Bayes (84.48%). Random Forest seems to be the best classification algorithm for our research, given our dataset and variables to predict student performance. Random Forest has the highest TP rate and precision for both classes, Fail and Pass, at 83% and 98% (Figure 3).

## Discussion

Competition exists amongst educational institutions in the country of study, and to survive in a competitive environment, educational institutions need to analyse their data, gain knowledge and use the intelligence from that knowledge to gain a competitive advantage. Students would like to study in institutions that are able to support them in successfully completing/passing the course. Successful students result in student retention and more student enrolments, which in turn means more revenue for the educational institution. Since the introduction of technological tools such as learning management systems (Moodle), the educational institutions have experienced positive paradigm shifts reflected through an enhanced delivery of services. These systems collect huge

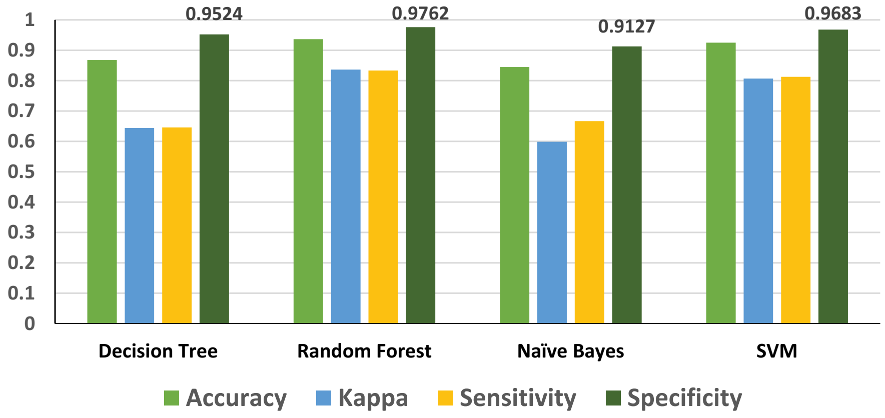

**Figure 3.** Evaluation measure.

amounts of data in regard to the students' interactions with the system. Technologies such as big data analytics, data mining and artificial intelligence can assist in analysing data and providing intelligence. This research study has mined student interaction data to predict students' success.

A sample of 918 student interaction logs with their course Moodle page was analysed. The analysis included collecting raw interaction data from logs, cleaning this data and then mining the data with R.[3] The research methodology followed the Knowledge Discovery in Databases (KDD) steps: 1). Dataset selection; 2). Pre-processing; 3). Data mining; 4). Interpretation and evaluation. Similar steps have been used for educational data mining by Guleria and Sood (2014), Kaur et al. (2015), Noah et al. (2013), Pratiwi (2013), and Shruthi and Chaitra (2016). Eighteen variables were scrutinised for making predictions. The decision variable was the grade of students, while some of the other variables were course work marks, assignment marks, number of times students accessed their course Moodle page, number of times they posted on their course discussion forum, number quizzes they attempted, number of assignments submitted, secondary school maths grade and highest secondary school level attended.

The student interaction logs, data-mined along 18 variables using R Studio with Decision Tree, Random Forest, Naïve Bayes and Support Vector Machine algorithms, resulted in a number of predictions. The prediction accuracy with the respective classification algorithm was as follows: Random Forest (93.67%), Support Vector Machine (92.53%), Decision Tree (86.78%) and Naïve Bayes (84.48%). The True Positive (TP) and Precision computation were as follows: Random Forest model has the highest TP rate and precision for both classes, Fail and Pass, 83% and 98% respectively. The lowest TP rate for class Fail is denoted in Decision Tree with 65% TP rate and the lowest TP rate for class Pass is denoted in the Naïve Bayes model with 91% TP rate.

Random Forest seems to be the best classification algorithm for our research, given our dataset and variables to predict student performance. It is evident from the results that students who interact more with their course Moodle page perform better than students who do not. Students who are active on their course Moodle page and contribute to forum discussions are more likely to pass than students who do not. Educators and administrators can now look at student Moodle interaction data logs early into the semester and identify students that are likely to be at risk of failing and take corrective action with intervention programmes. It is also suggested that course coordinators make quizzes, discussion forums and online participation count towards the course and move tests and assignments online as well. Learning management systems such as Moodle should not be seen as a supplementary course materials dissemination and catch-up web pages, but rather an integral part of the learning experience. Researchers such as Hussain et al. (2018), Mishra et al. (2014), Namdeo and Jayakumar (2014) and Peña-Ayala (2014) have all found the Random Forest classification algorithm is the best prediction model for education data mining.

## Conclusion

Data-mining algorithms were used to predict students' performance in an undergraduate course (first-year computing science course) at a university in the South Pacific. The dataset that was mined included 918 student interaction logs with their Learning Management System (Moodle) and 18 variables as described in the methodology section. R Studio was used with four classification algorithms (Decision Tree, Random Forest, Naïve Bayes and Support Vector Machine). The Random Forest classifier was the most accurate with 93.67% of prediction accuracy. For all the four models, the calculation for accuracy and precision was computed using the True Positive (TP) and Precision formula as in the results section. Based on the predictive performance of the classifiers, confusion matrix and comparison of evaluation measure by TP Rate and Precision, the Random Forest algorithm was the best for education data mining in our research study.

The prediction accuracy with the respective classification algorithm was as follows: Random Forest (93.67%), Support Vector Machine (92.53%), Decision Tree (86.78%) and Naïve Bayes (84.48%). The True Positive (TP) and Precision computation were as follows: Random Forest model has the highest TP rate and precision for both classes, Fail and Pass, 83% and 98% respectively. The lowest TP rate for class Fail is denoted in Decision Tree with 65% TP rate and the lowest TP rate for class Pass is denoted in the Naïve Bayes model with 91% TP rate. Random Forest seems to be the best classification algorithm for our research, given our dataset and variables to predict student performance. It is evident from the results that students that interact more with their course Moodle page perform better than students that do not. Students that are active on their course Moodle page and contribute to forum discussions are more likely to pass than students that do not. Educators and administrators can now look at student Moodle interaction data logs early into the semester and identify students that are likely to be at risk of failing and take corrective action with intervention programmes.

## Recommendations and future research

It is recommended that the learning analytics and variables consider and incorporate learning that takes place outside of the LMS environment. The course design needs to integrate and assess this learning so that a holistic view of student learning is considered for predicting student performance. In the future, in regard to extending this research, access and interaction of students with their course Moodle page need to be considered on a weekly basis. Variables such as number of clicks per assessment items and number of minutes spent within the learning environment need to be part of the dataset. The tools utilised by the students within their LMS environment and whether the student is a face-to-face student, blended or online student need to be included in the dataset as well.

Different data-mining methods, approaches and algorithms are also recommended to be used for future research on educational data mining for comparison of results and to come up with a best-practice algorithm. However, the caveat is that the results and the resulting process depend more on dataset, variables and data formats than on algorithms. With a different dataset and variables, Random Forest might not have been the best prediction algorithm. Future researchers can additionally consider clustering and classifications along the lines of pattern generation and pattern analysis when they are considering analytics inside a learning management system such as Moodle. Future research could also look at different learning management systems other than Moodle and different courses and even different levels of students within the computing science course. This research can be extended with the same set of students as they move into year two.

## Notes

1. https://www.linkedin.com/company/civitas-learning/
2. https://eab.com/
3. R is a free software environment for statistical computing and graphics. It is widely used in academia and research, as well as industrial applications. http://www.rdatamining.com/r

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Notes on contributors

***Sam Goundar*** is an International Academic having taught at 12 different universities in 10 different countries. He is the Editor-in-Chief of the *International Journal of Blockchains and Cryptocurrencies*, the Editor-in-Chief of the *International Journal of Fog Computing*, the Section Editor of the *Journal of Education and Information Technologies* and the Editor-in-Chief (Emeritus) of the *International Journal of Cloud Applications and Computing*. He is also on the Editorial Review Board of more than 20 high-impact factor journals.

***Arpana Deb*** is an Educational Technologist at The University of the South Pacific in Suva, Fiji. She has attained a Master of Science in Information Systems. Her research interests and areas of expertise include predictive analytics, big data and deep learning.

***Goel Lal*** is a Teaching Assistant at The University of the South Pacific in Suva, Fiji. His research interests are in data science, data mining, machine learning, artificial intelligence and knowledge discovery in databases. He teaches a first-year compulsory programme UU100 which involves lab sessions, assisting students in online discussion forums, marking assignments and updating student' marksheets on Moodle.

***Mohammed Naseem*** is a Teaching Assistant at The University of the South Pacific in Suva, Fiji. His research interests are in data science, data mining, machine learning, artificial intelligence and knowledge discovery in databases. He teaches a first-year compulsory programme UU100 which involves lab sessions, assisting students in online discussion forums, marking assignments and updating students marksheets on Moodle.

## ORCID

Sam Goundar http://orcid.org/0000-0001-6465-1097
Mohammed Naseem http://orcid.org/0000-0001-9740-6450

## References

Abad-Segura, E., González-Zamar, M. D., Infante-Moro, J. C., & Ruipérez García, G. (2020). Sustainable management of digital transformation in higher education: Global research trends. *Sustainability*, *12*(5), 2107. https://doi.org/10.3390/su12052107

Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. *Information Processing & Management*, *56*(4), 1245–1259.

Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: A case study. *Computer Science*, *2*(2), 140–146. http://hdl.handle.net/20.500.12358/25066

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, *2*(2), 43–47. https://doi.org/10.13189/wjcat.2014.020203

Ajibade, S. S. M., Ahmad, N. B. B., & Shamsuddin, S. M. (2019, August). Educational data mining: Enhancement of student performance model using ensemble methods. In *IOP Conference Series: Materials Science and Engineering*, IOP Publishing (Vol. 551, No. 1, p. 012061).

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 13–49. https://doi.org/10.1016/j.tele.2019.01.007

Alexander, B., Ashford-Rowe, K., Barajas-Murph, N., Dobbin, G., Knott, J., McCormack, M., . . . Weber, N. (2019). *Horizon report 2019 higher education edition*. EDU19.

Alzahrani, A. R., & Stojanovski, E. (2018). Multivariate assessment of mathematics test scores of students in Qatar. *International Journal of Educational and Pedagogical Sciences*, *12*(12), 1668–1671.

Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Computer Science*, *176*(2020), 1420–1429. https://doi.org/10.1016/j.procs.2020.03.358

Baker, R. S. J. D. (2010). Data mining for education. *International Encyclopaedia of Education*, *7*(3), 112–118. https://www.igi-global.com/dictionary/learning-analytics/59252

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM. Journal of Educational Data Mining*, *1*(1), 3–17.

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance . *arXiv preprint arXiv:1201.3417*.The Science and Information (SAI) Organization Limited.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burke, A. (2019). Student retention models in higher education: A literature review. *College and University*, *94*(2), 12–21. ProQuest document ID2232610556.

Caruth, G. D. (2018). Student engagement, retention, and motivation: Assessing academic success in today's college students. *Participatory Educational Research*, *5*(1), 17–30. https://doi.org/10.17275/per.18.4.5.1

Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., & Tsolakidis, A. (2014). Improving quality of educational processes providing new knowledge using data mining techniques. *Procedia-Social and Behavioral Sciences*, *147* (2014), 390–397. https://doi.org/10.1016/j.sbspro.2014.07.117

Chi, C. C., Kuo, C. H., Lu, M. Y., & Tsao, N. L. (2008, July). Concept-based pages recommendation by using cluster algorithm. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, IEEE, (pp. 298–300).

Deng, R., Benckendorff, P., & Gannaway, D. (2019). Progress and new directions for teaching and learning in MOOCs. *Computers & Education*, *129*, 48–60. https://doi.org/10.1016/j.compedu.2018.10.019

Depren, S. K., Aşkın, Ö. E., & Öz, E. (2017). Identifying the classification performances of educational data mining methods: A case study for TIMSS. *Educ Sci Theory Pract*, *17*(5), 1605–1623. doi:10.12738/estp.2017.5.0634.

Dewberry, C., & Jackson, D. J. (2018). An application of the theory of planned behavior to student retention. *Journal of Vocational Behavior*, *107*(1), 100–110. https://doi.org/10.1016/j.jvb.2018.03.005

Drysdale, M. T., & McBeath, M. (2018). Motivation, self-efficacy and learning strategies of university students participating in work-integrated learning. *Journal of Education and Work*, *31*(5–6), 478–488. https://doi.org/10.1080/13639080.2018.1533240

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247

Guleria, P., & Sood, M. (2014). Data mining in education: A review on the knowledge discovery perspective. *International Journal of Data Mining & Knowledge Management Process*, *4*(5), 47. https://doi.org/10.5121/ijdkp.2014.4504

Hämäläinen, W., & Vinni, M. (2010). Classifiers for educational technology. In *Handbook on educational data mining* (pp. 54–72). https://doi.org/10.1201/b10274

Hu, Q., & Rangwala, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. International Educational Data Mining Society.

Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. *PloS One*, *12*(1), e0161501. https://doi.org/10.1371/journal.pone.0161501

Huebner, R. A. (2013). A survey of educational data-mining research. *Research in Higher Education Journal*, *19*.

Jüttler, M. (2020). Predicting economics student retention in higher education: The effects of students' economic competencies at the end of upper secondary school on their intention to leave their studies in economics. *PloS One*, *15*(2), e0228505. https://doi.org/10.1371/journal.pone.0228505

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61–72. https://doi.org/10.2478/cait-2013-0006

Kahu, E. R., & Nelson, K. (2018). Student engagement in the educational interface: Understanding the mechanisms of student success. *Higher Education Research & Development*, *37*(1), 58–71. https://doi.org/10.1080/07294360.2017.1344197

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, *57*, 500–508. https://doi.org/10.1016/j.procs.2015.07.372

Kirschner, P. A. (2016, April). Learning analytics: Utopia or dystopia. Keynote presentation at 6th International Conference on Learning Analytics and Knowledge (LAK16), Edinburgh, UK.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kumari, P., Jain, P. K., & Pamula, R. (2018, March). An efficient use of ensemble methods to predict students' academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*, IEEE (pp. 1–6).

Latif, K. F., Latif, I., Farooq Sahibzada, U., & Ullah, M. (2019). In search of quality: Measuring higher education service quality (HiEduQual). *Total Quality Management & Business Excellence*, *30*(7–8), 768–791. https://doi.org/10.1080/14783363.2017.1338133

Manjarres, A. V., Sandoval, L. G. M., & Suarez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education Review*, *33*, 235–266.

Merceron, A., & Yacef, K. (2005). Educational data mining: A case study. *AIED*, 467–474.

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: An application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education, 2003. FIE 2003*, IEEE. (Vol. 1, pp. T2A–13).

Mishra, T., Kumar, D., & Gupta, S. (2014, February). Mining students' data for prediction performance. In 2014 Fourth International Conference on Advanced Computing & Communication Technologies (pp. 255–262). IEEE.

Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. *Procedia-Social and Behavioral Sciences*, *97*, 320–324. https://doi.org/10.1016/j.sbspro.2013.10.240

Namdeo, J., & Jayakumar, N. (2014). Predicting students performance using data mining technique with rough set theory concepts. *International Journal*, *2*(2).

Nasiri, M., Minaei, B., & Vafaei, F. (2012, February). Predicting GPA and academic dismissal in LMS using educational data mining: A case mining. In *6th National and 3rd International conference of e-Learning and e-Teaching*, IEEE, (pp. 53–58).

Noah, O. F., Barida, B. A. A. H., & Egerton, T. O. (2013). Evaluation of student performance using data mining over a given data space. *International Journal of Recent Technology and Engineering (IJRTE)*, *2*(4).

Noroozi, O., Alikhani, I., Järvelä, S., Kirschner, P. A., Juuso, I., & Seppänen, T. (2019). Multimodal data to design visual learning analytics for understanding regulation of learning. *Computers in Human Behavior*, *100*, 298–304. https://doi.org/10.1016/j.chb.2018.12.019

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*(422). https://doi.org/10.3389/fpsyg.2017.00422

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64.

Parikh, K. S., & Shah, T. P. (2016). Support vector machine – A large margin classifier to diagnose skin illnesses. *Procedia Technology*, *23*, 369–375. https://doi.org/10.1016/j.protcy.2016.03.039

Pechenkina, E., Laurence, D., Oates, G., Eldridge, D., & Hunter, D. (2017). Using a gamified mobile app to increase student engagement, retention and academic achievement. *International Journal of Educational Technology in Higher Education*, *14*(1), 1–12. https://doi.org/10.1186/s41239-017-0069-7

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Penteado, B. E., Paiva, P. M. P., Morettin-Zupelari, M., Isotani, S., & Ferrari, D. V. (2018). Toward better outcomes in audiology distance education: An educational data mining approach. *American Journal of Audiology*, *27*(3S), 513–525. https://doi.org/10.1044/2018_AJA-IMIA3-18-0020

Prabha, S. L., & Shanavas, A. M. (2014). Educational data mining applications. *Operations Research and Applications: An International Journal (ORAJ)*, *1*(1), 23–29.

Pratiwi, O. N. (2013). Predicting student placement class using data mining. *In Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, IEEE, (pp. 618–621).

Priyadharsini, C., & Thanamani, A. S. (2014). An overview of knowledge discovery database and data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(1), 1571–1578.

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924*.

Rogers, S. (2021, April 13). *4 best free and open source LMS tools* [blog post]. https://blog.capterra.com/top-8-freeopen-source-lmss/

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. https://doi.org/10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Sachin, R. B., & Vijay, M. S. (2012). A survey and future vision of data mining in educational field. In *2012 Second International Conference on Advanced Computing & Communication Technologies*, IEEE, (pp. 96–100).

Schunk, D. H., & Usher, E. L. (2011). Assessing self-efficacy for self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of selfregulation of learning and performance* (pp. 282–297). Routledge.

Shaleena, K. P., & Paul, S. (2015, March). Data mining techniques for predicting student performance. In 2015 IEEE international conference on engineering and technology (ICETECH) (pp. 1–3). IEEE.

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, *5*(4), 2094–2097.

Shruthi, P., & Chaitra, B. P. (2016). Student performance prediction in education sector using data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(3), 2012–2018.

Simsek, Ö., & Yazar, T. (2017). Investigation of teachers’ educational technology standards self-efficacy. *Pegem Egitim Ve Ogretim Dergisi= Pegem Journal of Education and Instruction*, *7*(1), 23.

Soni, A., Kumar, V., Kaur, R., & Hemavathi, D. (2018). Predicting student performance using data mining techniques. *International Journal of Pure and Applied Mathematics*, *119*(12), 221–227.

Subhash, S., & Cudney, E. A. (2018). Gamified learning in higher education: A systematic review of the literature. *Computers in Human Behavior*, *87*, 192–206. https://doi.org/10.1016/j.chb.2018.05.028

Sukhija, K., Jindal, M., & Aggarwal, N. (2015). The recent state of educational data mining: A survey and future visions. In *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*, IEEE, (pp. 354–359).

Suljic, M., & Osmanbegovic, E. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3–12.

Tie, Z., Jin, R., Zhuang, H., & Wang, Z. (2010, June). The research on teaching method of basics course of computer based on cluster analysis. In *2010 10th IEEE International Conference on Computer and Information Technology*, IEEE, (pp. 2001–2004).

Tight, M. (2020). Student retention and engagement in higher education. *Journal of Further and Higher Education*, *44*(5), 689–704. https://doi.org/10.1080/0309877X.2019.1576860

Tsai, Y. S., & Gasevic, D. (2017). Learning analytics in higher education – Challenges and policies: A review of eight learning analytics policies. In *Proceedings of the seventh international learning analytics & knowledge conference*, IEEE, (pp. 233–242).

Tudor, T. R. (2018). Fully integrating academic advising with career coaching to increase student retention, graduation rates and future job satisfaction: An industry approach. *Industry and Higher Education*, *32*(2), 73–79. https://doi.org/10.1177/0950422218759928

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110. https://doi.org/10.1016/j.chb.2018.07.027

Wei, W., Visweswaran, S., & Cooper, G. F. (2011). The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*, *18*(4), 370–375. https://doi.org/10.1136/amiajnl-2011-000101

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48–59. https://doi.org/10.1177/0165551516677946

Yahya, A. A. (2017). Swarm intelligence-based approach for educational data classification. *Journal of King Saud University–Computer and Information Sciences*.

Yassein, N. A., Helali, R. G. M., & Mohomad, S. B. (2017). Predicting student academic performance in KSA using data mining techniques. *Journal of Information Technology & Software Engineering*, *7*(05).

Yiu, T. (2019, June 13). Understanding random forest. *Medium. Towards Data Science*. https://towardsdatascience.com/understanding-random-forest-58381e0602d2