
**International Journal of Innovation in Education**

ISSN online: 1755-1528 - ISSN print: 1755-151X
https://www.inderscience.com/ijiie

# Using data mining techniques to predict university student's ability to graduate on schedule

Sampa Catherine Mwape, Douglas Kunda

**DOI:** 10.1504/IJIIE.2023.10053113

**Article History:**

| | |
|---|---|
| Received: | 01 July 2021 |
| Accepted: | 30 June 2022 |
| Published online: | 23 January 2023 |

# Using data mining techniques to predict university student's ability to graduate on schedule

Sampa Catherine Mwape*

Information Technology Department,
Zambia ICT College,
Ndola, 10101, Zambia
Email: mwapes19@gmail.com
*Corresponding author

Douglas Kunda

Department of Computer Science,
ZCAS University,
Lusaka, 10101, Zambia
Email: douglas.kunda@zcas.edu.zm

**Abstract:** Research in educational data mining to establish or predict the retention of students in higher education institutions, as well as predict graduation performance abounds. This research is a data mining based project aimed at generating a model that can be used for predicting student's ability to graduate on time. In this research we have examined various factors such as age, gender, continuous assessment results, and final exam results, determine how they influence a student's graduation schedule. We have demonstrated our application of classification as a data mining technique to identify interesting patterns, and subsequently use predictive techniques to predict the possible consequent outcome, and further have conducted a detailed examination of the J48, Bayes Net, PART and Random Forest predictive algorithms and compared to draw conclusions on the data mining prediction tools that give optimum results. The J48 stood out in terms of performance output.

**Keywords:** decision tree; comparison; clustering method; academic performance; modelling; higher education.

**Reference** to this paper should be made as follows: Mwape, S.C. and Kunda, D. (2023) 'Using data mining techniques to predict university student's ability to graduate on schedule', *Int. J. Innovation in Education*, Vol. 8, No. 1, pp.40–62.

**Biographical notes:** Sampa Catherine Mwape is a Lecturer in the Information Technology Faculty of Zambia ICT College. She holds a Master of Science degree in Information Technology from Mulungushi University, and undergraduate bachelor's degree in Information Technology from The Copperbelt University.

Douglas Kunda is an Associate Professor in Software Engineering and Vice Chancellor of ZCAS University. He is founding Dean of the School of Science, Engineering and Technology, Mulungushi University, Zambia. He has over 25 years of experience as in industry and academia. He has Ph.D. degree in Computer Science from the University of York (UK). He was Project

Director/Manager responsible for the development, management and implementation of the Integrated Financial Management Information Systems (IFMIS) Project for the Government of the Republic of Zambia. He is a member of ACM and Fellow of the ICT Association of Zambia. His research area and interest include: software engineering, artificial intelligence, component-based software engineering, ICT in Education, Internet of things, machine learning and data mining.

## 1 Introduction

The education sector has continued to record growth in the use of data mining tools and techniques, predictive data mining techniques being in predominant use, to predicting different academic outcomes. However, as the number of university entrants keep swelling, so does the data that needs to be handled and managed by universities. Data mining approaches are implemented to control huge volumes of data in order to uncover patters that lie therein and the subsequent revealing of relationships, all of which may consequently be helpful in making decisions (Bhardwaj and Pal, 2012)

Researchers have ventured into predicting graduation performance via the use of data mining methodologies (Hooshyar et al., 2019). Research pertaining to predicting university student's ability to graduate on time however, has not received much focus. And thus, as universities endeavour to provide remedial measures that can help support and improve student performance, minimal efforts if any are dispensed towards identifying factors that would influence student's ability to complete on schedule. Not to mention availing resources to mitigating these factors. Failure to determine if students will graduate on schedule, not only stifles educational administrator's mitigating efforts, but also blind sides sponsors of the learners, who might have to deal with unforeseen expenses due to delayed completion, not to mention failing to provide support that would otherwise avert such an eventuality.

Thus, this research's problem statement focuses how different factors can influence a student's ability to graduate on time, and how if these are not identified, delay to mitigate the situation can lead to institutions incurring avoidable expenses. If these factors left unchecked, the can limit educational technocrats potential engender procedures to mitigate these factors.

The main aim of the research to develop a model that predicts university student's expected graduation time, using data mining techniques. And all the efforts of the research work were dedicated to answering the research questions;

- What factors would affect higher education institutions and university student's ability to graduate on schedule?
- How can data mining algorithms can be applied as predictive tools of determining expected graduation dates.
- Which data mining algorithm provides accurate predictions of when a student is likely to graduate.

It is in this vain that this paper reveals the identified attributes, demonstrates how to apply different classification algorithms so as to compare their effectiveness and accuracy

levels of prediction. Among those analysed included Decision Trees, as it is evidenced to be one of the most popular classification techniques in data mining. They present several advantages of other techniques, that include easy understandability based on the simple presentation, and their versatility in working with varied attributes (Al-Barrak and Al-Razgan, 2016). They are not only easy to implement and use for classification and regression tasks, but also good predictive performance, computational efficiency (ElGamal, 2013). Further, the research applied Bayesian classifiers (Naïve Bayes) and Random Forest as comparison algorithms.

Findings from a study of this nature can help university and college administrators to not only identify learners falling behind in terms of graduation schedule, but it would also assist in revealing factors that would lead to learners failing to graduate on time and how they are related. Modelling this approach could help highlight remedial/mitigate measures for students that may show signs of falling behind and potentially fail to graduate on time. Once the influencing factors have been identified and necessary corrective measures put in place, it can help academic administration and sponsors to save on resources that would otherwise be expended towards learners that go beyond the anticipated graduation time.

The sections of the paper following the introduction, covers the related work, the methods, results and discussion, and ends with the conclusions and future works.

## 2 Related work

Varied data mining studies have been carried out and notably so in the education sector in relation to student performance. This section highlights the classification of data mining models and algorithms in Section 2.1, and then discusses the data mining in higher education Section 2.2, with Section 2.3 providing analysis of the application of data mining in predicting student performance and retention, and the section concludes with a look at challenges and open issues in Section 2.4

### *2.1 Classification of data mining models and algorithms*

The classification of the most common data mining models and algorithms in higher education is presented:

#### *2.1.1 Predictive modelling*

Predictive modelling has been described as referring to the act of constructing equations that use observed data in order to predict future instances with future unobserved data (Raju and Schumacker, 2015).

#### *2.1.2 Decision tree*

Decision Tree is a decision tree, Yadav et al. (2012) defined a decision tree as a flow-chart-like tree configuration, where each interior node is denoted by rectangles, and leaf nodes are denoted by ovals. All interior nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it.

### *2.1.3 Random forest*

As suggested by the name, this is an algorithm that forms a forest with multiple trees, and where the more trees in the forest, points a more robust forest, and thus linked to a high accuracy results. It is an easily applicable supervised algorithm of classification (Jalota and Agrawal, 2019).

### *2.1.4 K-Nearest neighbour*

A nearest neighbour classifier is a method for classifying elements centered on the classification of the elements in the training set that are most comparable to the test example. With the k-nearest neighbour technique, this is done by assessing the k number of closest neighbours (Kumar and Verma, 2012).

### *2.1.5 Support vector machines*

Known to be an effective method of regression, classification and general pattern recognition algorithm which recognises patterns without the need for background knowledge even when the dimension of input space is very high (Kumar and Verma, 2012).

### *2.1.6 Association rule learning*

Also referred to as market basket analysis or dependency modelling. It uncovers links and associations among variables (Bhargava et al., 2013).

### *2.1.7 Clustering*

Clustering is the fragmenting of dividing of a group of related records or similar items into a number of sets called clusters (Lekha and Prakasam, 2017).

### *2.1.8 Classification*

A frequently applied data mining technique, which works on a pre-classified samples of data to create a model that can categorise attributes based on their frequency of occurrence. This classification technique forms a link between a dependent variable and an independent variable by mapping the data points. And thus, classification is applied to uncover to which class a data occurrence is linked within a given dataset (Lekha and Prakasam, 2017).

### *2.1.9 Regression*

It tries to find a function that model the data with least errors (Bhargava et al., 2013).

### *2.1.10 Bayesian classifier*

Bayesian Classifier is a Bayes Theorem based probability algorithm designed to deal with classifications by computing a set of probabilities by adding up the occurrences of given value combinations from an available dataset. It is a classifier with a probability method

and statistics which forecasts opportunities in the future based on previous experiences (Peling et al., 2017).

Table 1 gives a summarised presentation of the data mining classifications and algorithms.

**Table 1** Classification of data mining models and algorithms

| *Data mining technique* | *Algorithm* | *References* |
|---|---|---|
| Predictive | Classification | • Jayaprakash (2018)<br>• Borges et al. (2013) |
| Classification | Various types | • Umadevi and Marseline (2017)<br>• Romero and Ventura (2010)<br>• Imran et al. (2019) |
| Classification | ID3, C4.5 and ADT | • Yadav et al. (2012)<br>• Yadav et al. (2012) |
| Classification | Naïve Bayes, Decision Tree, Neural Network, Multi-layer perception, K-Nearest Neighbour, Rule based Learners | • Asif (2015)<br>• Jishan (2015)<br>• Kabakchieva (2013)<br>• Ramesh (2013)<br>• Osmanbegovic and Suljic (2012)<br>• Al-Radaideh et al. (2006) |
| Classification and Clustering | Extreme learning machine, support vector machine and neural networks. Decision Tree, Neural Network, K-Nearest network, Naïve Bayes, Support Vector Machine and Logistic Regression | • Nhu (2020) |
| Clustering and Classification | K-Kmeans, Smooth support vector machine (SSVM) | Sembiring et al. (2011) |
| Classification | Linear Regression | Siguenza-Guzman et al. (2015) |
| Classification and Clustering | J48 and Random Tree | Moscoso-Zea et al. (2019) |
| Classification | CART | Yadav et al. (2012) |
| Classification | Decision Trees and Linear Models | Alyahyan and Düştegör (2020) |
| Classification | JRip | Anuradha and Velmurugan (2015) |
| Classification | Artificial Neural Networks and Decision Tree | Kuyoro'Shade et al. (2012) |

Research supporting descriptive models having been carried out by Braganca et al. (2019), who favoured regression and Umadevi (2017), favoured classification. Further Romero and Ventura (2010) in their research point out common use of classification techniques, specifically Bayesian networks, neural networks, and decision tree. Siguenza-Guzman et al. (2015) highlight linear regression and logistic regression as common regression techniques. Bhardwaj and Pal (2012) in their research used ID3, C4.5 and ADT as their classification tools, and noted how it can be used to determine the accuracy of the generated model.

Yadav (2012) to evaluate the accuracy of the resulting predictive model, used and compared three classification models: ID3, C4.5 and ADT with a 10-fold cross validation selected as their evaluation approach. The models helped them determine whether a new student would continue to enrol in the following year or not. Moscoso-Zea (2019)compared the percentage of correct and incorrect classification of two models J48 and Random tree, in determining graduation rate, and they concluded J48 provided the best result given the parameters applied. Borges et al. (2013) describe data classification as comprising two stages, one being the training and the other being the test stage, where the actual class of instance is compared with the predicted class. And they compared different algorithms based on this fact.

A study by Yadav (2012) used student attendance data, and assessment marks from various tasks to predict end of semester performance, by using three algorithms, ID3, C4.5 and CART with CART identified as the best algorithm for classifying data. Alyahyan and Düştegör (2020) Indicate that a choice of data mining model can be made between predictive or descriptive. They ranked 10 algorithms that can be used to build a model, and indicated that the choice from these 10 would be based on which is the most interpretable and understandable. And based on this they opted for Decision Trees and Linear Models and fitting the bill.

Anuradha and Velmurugan (2015) conducted a comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. The research carried out in three of the private colleges in Tamil Nadu state of India was aimed at using classification techniques to predict the performance of students in end semester university examinations. They compared Decision Tree, C45(J48), Bayesian classifier, K Nearest Neighbour and Two Rules Learner's Algorithms namely OneR and JRip, to determine their accuracy in predicting student performance. And the resulting observations where that overall accuracy of the tested classifiers was above 60%. The JRip produces highest classification accuracy for the Distinction.

Goga et al. (2015) conducted a research and observed the abundance of students' performance related studies, but with minimal focusing on applying machine learning algorithms to students, and thus focused theirs as such. They considered taking into account student background factors in predicting student performance. Their research was conducted on data gathered from 1500 students from three Nigerian tertiary institutions. The student's academic performance was to be measured based on the cumulative grade point average (CGPA) at eh end of the first year. They used WEKA to generated three decision tree models, Artificial Neural Networks and two rulesets. A comparison was later done based on the accuracy level and confusion matrices to determine the optimal mode. Jayaprakash (2018) points out the many machine learning tools that support predicative analysis and visualisation of datasets. They lean towards WEKA which has inbuilt tools for data preprocessing classification, association rules and visualisation among many other capabilities.

Although a research by Osmanbegovic and Suljic (2012), was aimed at using data mining to develop models that would derive the conclusions on students' academic performance, they focused it differently in that they also compared different data mining techniques during the predicting process. Data collection was based on a survey conducted during the semester at University of Tuzla for the academic year 2010–2011, from first year students. The success was evaluated with the passing grade at the examination. They investigated the impact of factors such as a student's socio-demographic variables, achieved results from high school and form the entrance exam, as well as attitude towards studying, on overall student success.

### *2.2 Data mining in higher education*

We point to the works of Moscoso-Zea (2019) who describe educational data mining as progressing discipline that is centered on creating models that aid improvement of learning practices and organisational competence. Because of its potential to transform the education sector, particularly the higher education, the datamining field has continued to gain popularity. It promises the potential to enhance the understanding of educational data as well as the learning process as it focuses on identifying and extracting and evaluating variables related to the learning process of students. (Abu Tair, 2012) Mining in education environment is called educational Data Mining. Han et al. (2012) further describe data mining as the tool that permits the users to examine the relationship which are identified during the mining process.

As higher institutions of learning collect and retain numerous types of student data, which could range from student academic data to their personal records. And studies that focus on data mining in higher education institutions abound, most seeking to monitor student performance. Veeramuthu (2014) carried out a study that was an attempt at using data mining processes, particularly predictive classification, with the objective to design a model that would aid higher institutions of learning in identifying factors that motivate new students to enrol in a given college or university would use to attract and retain an increased number of students.

For higher educational institutions, data mining techniques could also assist in giving more personalised learning, increase the learning systems efficiency and lessen the cost of the teaching and learning processes. It ultimately could lead to education administrators raising learner retention rate, increase educational improvement ration and raise the learner's performance outcome (Zhang et al., 2010). Moscoso-Zea (2019) conducted a research to review which algorithms of data mining could be used in the analysis of educational data. The aim was to discover trends and patterns of study in the graduation rate indicator. They compared the methods and algorithms and their findings highlighted random tress to have better precision although it had limitations and difficulty of interpretation while J48 algorithms had better possibilities of interpretation of results in the visualisation of the classification with slight inferior performance.

### *2.3 Data mining in predicting student performance and retention*

Research has gone into identifying predicators to student graduation, with most supporting academic ability as an outstandingly significant variable leading to student graduation (Raju and Schumacker, 2015).

In their study, Kovacic (2010) brought out ethnicity, program under study and course block as variables that influenced student A further study on student retention, revealed that a major factor that influenced student retention was their proximity to college or campus. Sivakumar (2016) conducted a study on how educational data mining can be used to reduce student dropout rate by using classification. In their study they used the classification technique to predict a student's performance. Their developed model used different Decision Tree and Bayes algorithms to evaluate student performance, which subsequently identified weak student having enrolment status at risk and identifying those that would need further help. Their research revealed Naïve Bayes as being among four algorithms having highest accuracy of above ninety percent. Bhardwaj and Pal (2012), in their research, recognising that academic performance is influenced by varied factors, deemed it cardinal to develop predictive data mining model that would identify between high learners and low learners.

Bhardwaj and Pal (2012) study on student performance revealed that factors like student grade in senior secondary exam, living location as well medium of teaching were highly correlated with the student's academic performance. In a related conducted by Pandey and Pal (2011) who used Bayes classification on category, language and background qualification, to determine whether new comer students will perform or not. In another related research for feature describe, this process as a dynamic and productive field and research area of machine learning and data mining (Zaffar et al., 2017).

A comparative study to predict student's retention was conducted by Yadav (2012). Their project sought to generate predictive models for student retention management. Their conclusions indicated some machine learning algorithms as being able to create effective predictive models from the existing student retention data. Other researchers highlight preadmission data as cardinal in selecting factors that influence student performance. Ahmad et al. (2015), Mesarić and Šebalj (2016) and Aluko et al. (2018) all in their studies indicate entrance test results as being the factors to predict student performance, whereas Ahmad et al. (2015), Almarabeh (2017), Hamoud et al. (2018) and Mueen et al. (2016); highlight university data such as grades already obtained by the learner while in college like GPA or CGPA and Mohamed and Waguih (2017) selected course marks. Other researchers identifying student assessment related data as factors that would influence student performance include (Al-Barrak and Al-Razgan, 2016) considered overall rating, whereas Hamoud et al. (2018) and Aluko et al. (2018) considered binary class problems like (pass/fail). Aluko et al. (2018) and Putpuek (2018) both pointed out university related data such as the faculty a student belonged to as another influencing factor to student performance. They also pointed out that aside demographic data, assessment grades in end of year examinations as other influencing factors to performance. Bhardwaj and Pal (2012) considered student data that included attendance, class tests, seminar and assignment marks as factors to for predicting student performance.

Daud et al. (2017) proposed a model for Predicting Student Performance using Advanced Learning Analytics. Their study applied learning analytics on data collected on scholarship holding student of different Pakistan universities, to predict whether a student will be able to complete his degree or not. The consequent experimental results indicated that the suggested method considerably outperformed the current methods due to exploitation of family expenditure and students' personal information feature sets. Narrowing their research, ElGamal (2013) focused on a model that predicts student performance in Programing Course. The study considered factors such as students'

mathematical background, programming aptitude, problem solving skills, gender prior experience would influence student performance.

### 2.4 Evaluation of data mining techniques

This section discusses research that has compared and evaluated varied data mining techniques and the performance of different data mining algorithms. Table 2 provides a summary of some of the comparative studies as investigated.

**Table 2** Evaluation of data mining techniques

| *Methodology* | *Key findings* | *References* |
|---|---|---|
| Compared performance metrics of probabilistic error, qualitative error and visual metrics | Compared to the Decision Tree, Naive Bayes, Bayes Network and CART, Random Forest provided the best results | Kumar et al. (2017) |
| Compared Precision | Reviewed data mining techniques Used in Educational Data Mining to Predict Student Amelioration | Anoopkumar and Rahman (2016) |
| Analysed and compared | KNN as well as classifiers related to Rule-based, Bayesian Decision Trees | Ranbaduge (2013) |
| Hybrid procedure | Analysed and compared Decision Tree methods and Data clustering of Data mining recommended K-Means and Decision Trees | Shovon et al. (2012) |
| Cross validation method and percentage split method | Compared J48, NBtree, Reptree and Simple cart, J48 was most outstanding as ideal for model construction to predict performance | Pandey and Pal (2011) |
| Classification algorithms and predictive analysis | Classification algorithms and predictive analysis | Shazmeen et al. (2013) |
| Clustering | Smooth Support Vector machine (SSVM) classification clustering technics like K-means | Sembiring et al. (2011) |
| 10-fold cross validation methods | Results showed that Nearest Cluster, ID3 and J48 technique has highest accuracy compared to other method | Jayakameswaraiah and Ramakrishna (2014) |
| Comparative analysis | Evaluated J48 decision tree algorithm which is an open source Java implementation of C4.5 algorithm Naive Bayes Classifiers, *k*-Nearest Neighbours algorithm (K-NN), OneR and JRip algorithm | Anuradhal and Velmurugan (2015) |
| Comparison of prediction precision | Naïve Bayes classifier outperformed other algorithms in prediction | Osmanbegović and Suljić (2012) |

Kumar et al. (2017) in their research evaluated algorithms based on performance metrics of probabilistic error, qualitative error and visual metrics, where they draw conclusions pointing to Random Forest algorithm for predictive modelling as giving them the best result as compared to the Decision Tree, Naive Bayes, Bayes Network and CART.

Anoopkumar and Rahman (2016) conducted a research that reviewed data mining techniques used in Educational Data Mining to Predict Student Amelioration. Data mining techniques such KNN as well as classifiers related to Rule-based, Bayesian Decision Trees and instance based learner classifiers were analysed and compared in a study conducted by Ranbaduge (2013) as they endeavoured to determine which would be ideal to examine a student performance.

Shovon et al. (2012) on evaluating Decision Tree methods and Data clustering of Data mining recommended K-Means and Decision Trees as sufficient to predict a student's performance. In the same light Bavisi et al. (2014) comparative study of four different decision tree algorithms, J48, NBtree, Reptree and Simple cart, used the cross validation method and percentage split method to determine which of the algorithms provided accurate results. They concluded that J48 was ideal for model construction to predict performance. In other related study, Shazmeen et al. (2013) performance evaluation of classification algorithms and predictive analysis was conducted and applied to varied dataset to establish the efficiency of the algorithms in feature selection and performance prediction. A further study to determine efficiency and performance of different algorithms led to Sembiring et al. (2011) proposing the use of kernel methods as data mining techniques. They analysed Smooth Support Vector machine (SSVM) classification clustering technics like K-means. Jayakameswaraiah and Ramakrishna (2014) in a study to predict student performance, evaluated some classification and clustering algorithms using the 10-fold cross validation methods. Further studies to determine prediction accuracies of classification algorithms was conducted through a comparative analysis (Anuradha and Velmurugan, 2015). They evaluated J48 decision tree algorithm which is an open source Java implementation of C4.5 algorithm Naive Bayes Classifiers, *k*-Nearest Neighbours algorithm (K-NN), OneR and JRip algorithm

## *2.5 Challenges and open issues*

Thus far, existing literature indicates research dominance in predicting student's academic success and retention, which begs the question on research specific to predicting student's ability to graduate on schedule. A legion of studies have been conducted in educational data mining, significantly those that predict student performance. However little literature reveals mining data to predict a student's expected graduation time. Much of the research has been carried out predominantly relating to performance to student, highlights on their learning capabilities, and factors leading to students dropping out. And in most of these studies reveal smaller datasets, as well as confinement to specific learning institutions mostly private colleges and universities. Table 3 highlights some observed challenges and open issues.

**Table 3** Challenges and open issues

| *Challenges and open issues* | *References* |
|---|---|
| Considered the overlooked factors that would influence student academic performance | Oskouei and Askari (2014) |
| Multi Agent Data mining, to predict students' performance | Almalaise (2013) |
| Detailed study of the precision of different data mining technics | Strecht et al. (2015) |
| Highlighted one of the challenges of educational data mining as being confidentiality issue related to mining personal data | Salenga and Villanueva (2018) |

The works of Oskouei and Askari (2014) involved a review of students from different countries to determine the various factors that influence performance of students. The study applied classification and prediction algorithms that accurately predict student performance results prior to examinations. They thus identified factors such as gender, family background and style of living among other things as determining factors in predicting student academic performance. Al-Malaise (2013) proposed a student's performance prediction system using Multi Agent Data mining, to predict students' performance based on their data with high precision of prognostication and provide an aid to the weaker student by optimisation rules. By using Adaboost. M1 and LogitBoost ensemble classifier methods and with the single classifier method C4.5, the implemented system was evaluated to determine which presented the most accurate results.

Further a comparative study by Strecht et al. (2015) evaluates the precision of Decision Tree and Bayesian Network algorithms for predicting the academic performance. The study reviewed a consistent precision result for the Decision Tree in comparison to the Bayesian Network which yielded lesser precision. The outcomes of the said study provided a basis for identifying the data mining algorithms that accurately predict performance of students, and evaluate these varied algorithms in terms of their precision.

As highlighted studies that predict student performance by considering factors such as class attendance, CA grades etc., however minimal investigations look into factors such as financial status, sponsorship terms, and as well as marital or family status of learner as influencing factors in graduation time. Though the identified factors largely point towards a predicting a student's academic success, minimal research applies these factors to determining when a student would likely graduate and this needs further attention and consideration.

Another issue that raises concerns is that although there is a range for creating comparative research that support in the valuation of the efficiency, accuracy and importance of the already existing techniques in educational data mining, it is rather a challenge to conclusively make comparisons and contrast the techniques as most researchers for confidentiality purposes, hide the raw data and only review the test results. They seek to assure the stakeholders that their data remains confidential and not open to the public. The abundance of educational data cannot be argued, what needs attention though is accessing datasets that are already structured, as such would ease the data mining process in educational institutions.

## 3 Methods

The steps followed in the research methodology process are depicted in Figure 1.

**Figure 1** Depiction of the research methodology process (see online version for colours)

### *3.1 Data collection*

The development process was preceded by the data collection; which data was obtained from the college's Student Management System that stores the enrolment data. The performance data was obtained from the Learning Management System, and as data was obtained from two different sources, it needed to be cleaned and normalised. And thus the collection method was followed by the preprocessing stage.

### *3.2 Data preparation and preprocessing*

This state involved the cleaning of the data, the author worked on an Automated Machine Learning Platform, by using TPOT, a Python Library that automates the whole machine learning pipeline. And thus the process of feature selection as well as model selection, data cleaning and evaluation was supported through the use of TPOT, in order to minimise errors in the prediction.

The preparation and preprocessing included the following stapes:

### *3.3 Data mining model development*

WEKA platform has been used as the development platform. WEKA toolkit, chosen as an option because it is widely used software for data mining that provides a broad array of varied data mining algorithms implemented in JAVA. It has evidenced wide use in educational data mining researches and instructional purposes (Whitley, 2018). It is freely available and is broadly used for research in the data mining realm (Anuradha and Velmurugan, 2015).

As a means to extracting useful knowledge from the data collected, 4 data mining models were used; Decision Tree Algorithm, that is C4.5 (J48), Bayesian Algorithms, Random Forest and PART. These so chosen due to their popular use in reviewed literature, which detail the advantages they present in predicting student performance.

### *3.4 Attribute selection*

Literature indicates that the outstanding factors/attributes contributing to student success have been the following:

| *Factors* | *References* |
|---|---|
| Prior academic achievement | Bhardwaj and Pal (2011) |
| Student demographics | Oshodi et al. (2018) |
| E-learning activity | Hussain et al. (2018) and Shayan and van Zaanen (2019) |
| Psychological attributes | Veeramuthu and Periasamy (2014) |
| Environments (living) | Thiele et al. (2016) |
| | Bhardwaj and Pal (2011) |
| Continuous assessment grades | Ahmad et al. (2015), Almarabeh (2017), Hamoud et al. (2018), Mueen et al. (2016) and Singh and Kaur (2016) |
| Examination Grades | Osmanbegovic and Suljic (2012) |

Thus, in this study, the researcher adopted some of the aforementioned attributes for consideration. However psychological attributes being subjective and being difficult to quantify was omitted from the list of attributes. The aforementioned attributes were used to populate the two (2) categories of datasets as stated in previous section.

As one of the objectives of the research is to establish the influential factors that influence student graduation time, no sampling techniques will be used, but rather the entire dataset will be analysed, that is the demographic data, the performance datasets respectively.

The resulting attributes are described in Table 4.

**Table 4** Table of the attributes and their descriptions

| *ID* | *ATTRIBUTE* | *TYPE* |
|---|---|---|
| 1 | Student No | Numeric |
| 2 | Gender | String |
| 3 | Course Code | Numeric |
| 4 | Continuous assessment | Numeric |
| 5 | Final Exam | Numeric |
| 6 | Total | Numeric |
| 7 | Grade | String |
| 8 | Result | String |

This table is a depiction of the selected attributes, and their associated descriptions.

## *3.5 Implementation using WEKA tool*

In this research, the following steps were performed to implementing the algorithms in WEKA tool:

*Step 1: Preprocessor*

The preprocessor imports the dataset into the tool and preprocess it. Output of preprocessor shown in Figures 2 and 3.

*Step 2: Classification*

This is the panel that allows for the user to choose to either use the classification or to use Regression methods to estimate the accuracy of the resulting model. The algorithms for classification used are J48, PART, Random Forest and Bayes Net.

## *3.6 Results*

Having created the classification using WEKA, results were analysed and presented as highlighted in this section:

*J48 ALGORITHM*

J48 is an algorithm that generates a pruned or unpruned C4.5 decision tree.

**Figure 2** Trained data as imported into WEKA (see online version for colours)

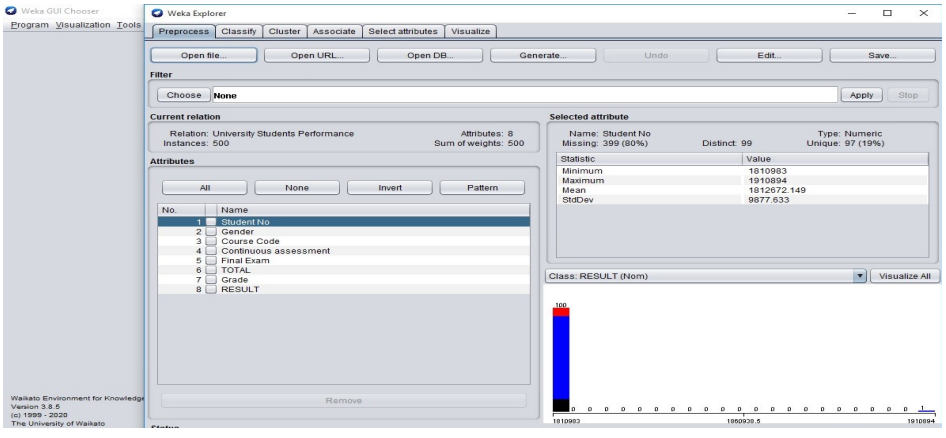



**Figure 3** Out of preprocessor visualisation of all attributes (see online version for colours)

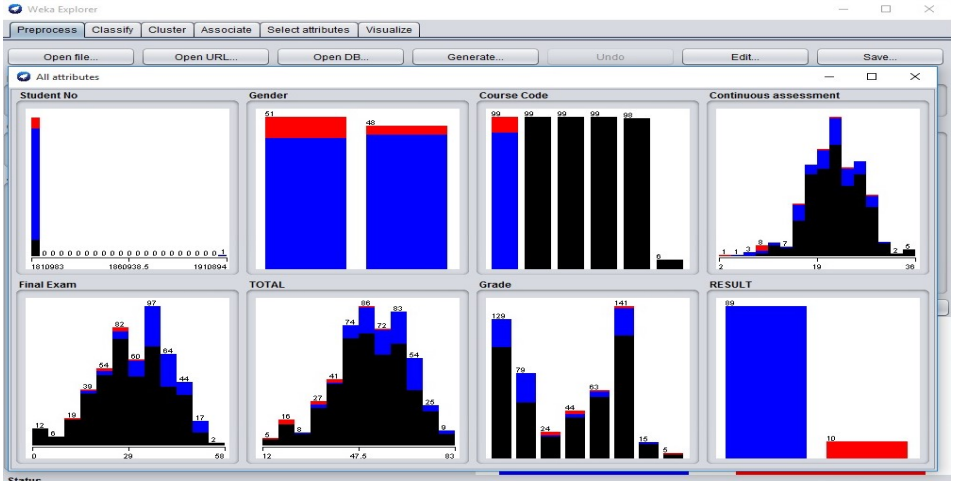



Figure 4 is a depiction of the performance output of the J48 algorithm which classifies all students with student 1811411 and below, with the final exam result of 40 or less as having a ration of (6.0) repeat likelihood and classified students with student number 1811411 and above, with a continuous assessment of 23.2 and above as having a (6.0) ration of graduating likelihood.

On visualisation, the J48 produced the depicted pruned tree, with node 'Total' producing the left branch of a total score of less or equal to 40 and classifying students with student number 1811411 and below as potentially repeating whereas those with student number 1811411 graduate. Whereas the right side braches indicate students with a total score above 40, with final exam score greater than 29.4 to graduate and those with final exam score, less or equal to 29.4 and continuous assessment less or equal to 23.2 to repeat their course.

**Figure 4** Output for J48 (see online version for colours)

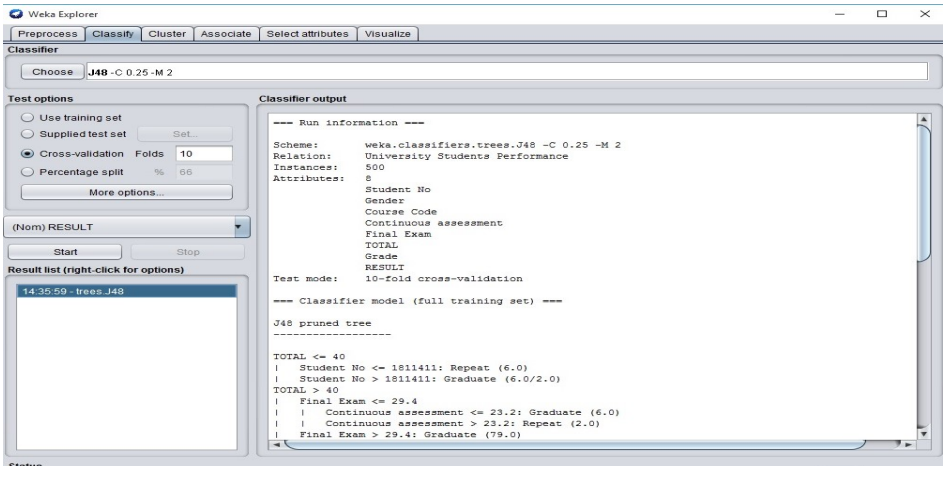



Figure 5 displays the Output of J48 visualization tree.

**Figure 5** Output for J48 visualised tree (see online version for colours)

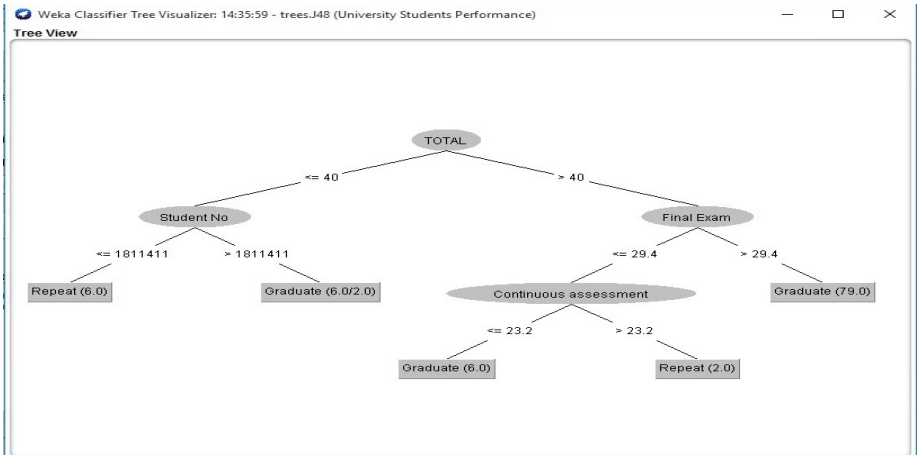



Figure 6 is the output for Bayes Net. The top left, 85, are things the model thinks are ‘a’ which really are ‘a’ ← these were correct

- bottom left, 3, are samples the model thinks are ‘a’ but which are really ‘b’ ← one kind of error
- top right, 4, are samples the model thinks are ‘b’ but which really are ‘a’ ← another kind of error
- bottom right, 7 are samples the model thinks are ‘b’ which really are ‘b’
- the top-left and bottom-right of the matrix is showing the samples the model gets right
- bottom-left and top-right of the matrix are showing where the model is confused.

**Figure 6** Output for Bayes Net (see online version for colours)

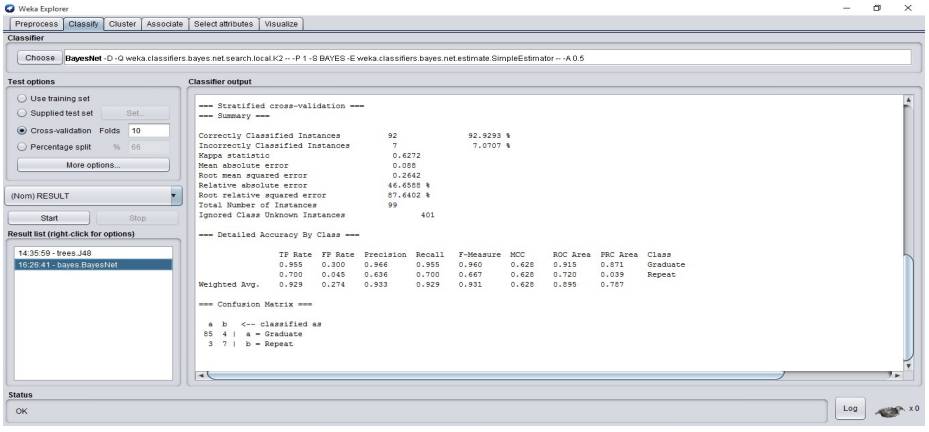
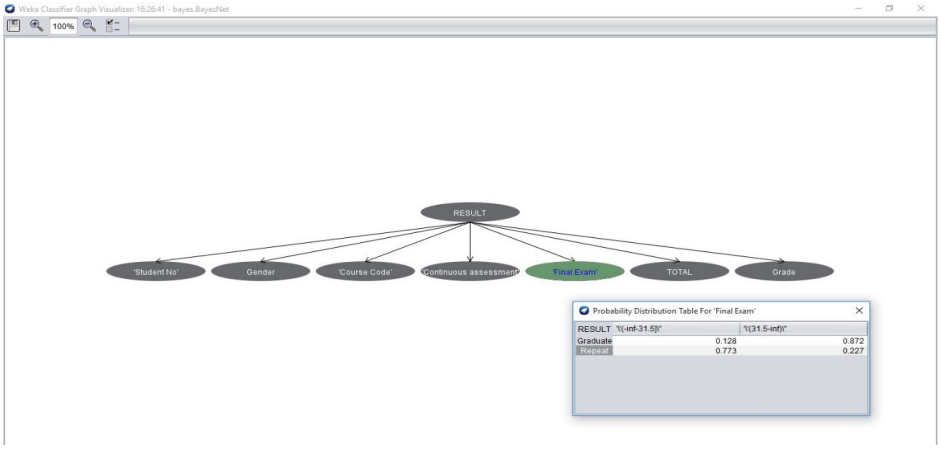



Figure 7 is a visualisation of the Bayes Net Visualisation, which can only present visual analysis of one node attribute at any given time, and the attribute displayed is the probabilistic table for the Final Exam score attribute.

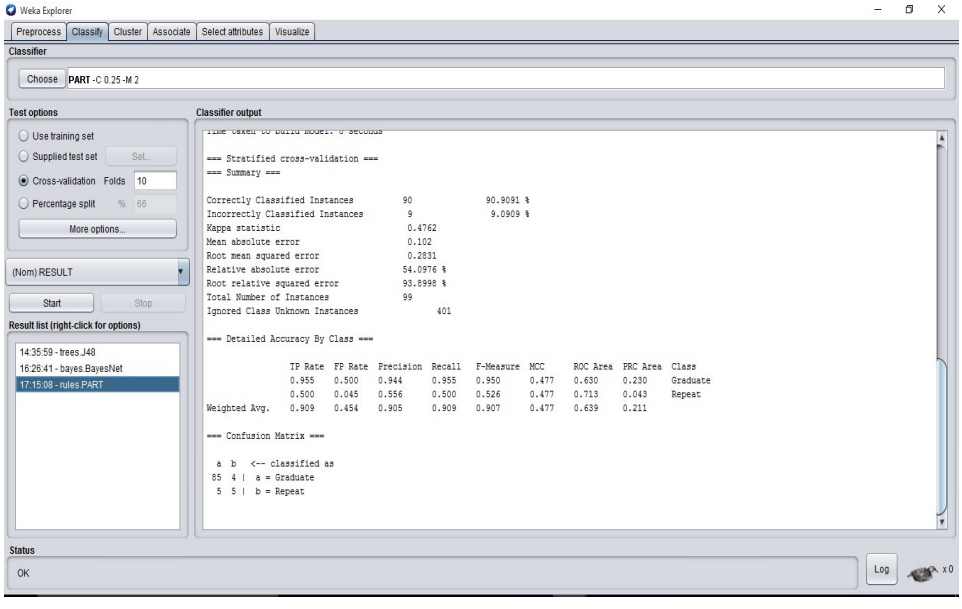
**Figure 7** Bayes net visualisation (see online version for colours)




The PART output is presented in Figure 8 and shows the following performance results of the algorithm. Given the two (2) variable s being considered, that is the Graduate and Repeat, the algorithm presents

- top left, 85, are things the model thinks are 'a' which really are 'a' ← these were correct
- bottom left, 5, are samples the model thinks are 'a' but which are really 'b' ← one kind of error
- top right, 4, are samples the model thinks are 'b' but which really are 'a' ← another kind of error
- bottom right, 5 are samples the model thinks are 'b' which really are "b".

**Figure 8** PART output (see online version for colours)




Random Forest in output in Figure 9 shows the top left, 85, are things the model thinks are ‘a’ which really are ‘a’ ← these were correct

- bottom left, 5, are samples the model thinks are ‘a’ but which are really ‘b’ ← one kind of error
- top right, 4, are samples the model thinks are ‘b’ but which really are ‘a’ ← another kind of error
- bottom right, 5 are samples the model thinks are ‘b’ which really are “b”
- the top-left and bottom-right of the matrix is showing the samples the model gets right
- bottom-left and top-right of the matrix are showing where the model is confused
- we can tell from the confusion matrix that it made 99 predictions. Out of the 99 predictions the Classifier predicted 90 to Graduate and 9 to Repeat when in reality 89 Graduate and 10 Repeat.

The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified.

## *3.7 Comparison of classifiers*

The tables depicted in this section provide comparisons of classifiers as done in WEKA with the results discussed below each given table.

**Figure 9** Random forest output (see online version for colours)

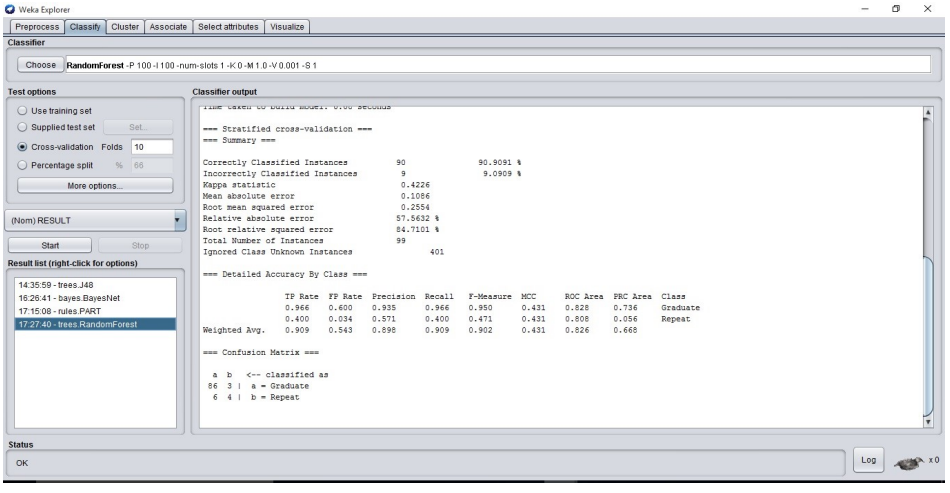



Table 5 is a depiction of the prediction performance of the algorithms detailing the correctly classified instance and the incorrectly classified instances. J48 produced a 92 correctly classified instances which represents a 92.9293% and 7 incorrectly classified instances which accounts for 7.0707%.

**Table 5** J48 output summary

| | | |
|---|---|---|
| Correctly classified instances | 92 | 92.9293% |
| Incorrectly classified instances | 7 | 7.0707% |
| Kappa statistic | 0.5509 | |
| Mean absolute error | 0.0824 | |
| Root mean squared error | 0.2575 | |
| Relative absolute error | 43.6634% | |
| Root relative squared error | 85.4202% | |
| *Total number of instances* | 99 | |

Table 6 is the Bayes Net output summary on the prediction performance of the algorithms detailing the correctly classified instance and the incorrectly classified instances. Bayes Net produced a 92 correctly classified instances which represents a 92.9293% and 7 incorrectly classified instances which accounts for 7.0707%.

The output summary for PART is given in Table 7. Its performance shows the correctly classified instance and the incorrectly classified instances. It produced a 90 correctly classified instances which represents a 90.9091% and 9 incorrectly classified instances which accounts for 9.0909%.

**Table 6** Bayes Net output summary

| | | |
|---|---|---|
| Correctly classified instances | 92 | 92.9293 |
| Incorrectly classified instances | 7 | 7.0707% |
| Kappa statistic | 0.6272 | |
| Mean absolute error | 0.088 | |
| Root mean squared error | 0.2642 | |
| Relative absolute error | 46.6588% | |
| Root relative squared error | 87.6402% | |
| *Total number of instances* | 99 | |

**Table 7** PART output summary

| | | |
|---|---|---|
| Correctly classified instances | 90 | 90.9091% |
| Incorrectly classified instances | 9 | 9.0909% |
| Kappa statistic | 0.4762 | |
| Mean absolute error | 0.012 | |
| Root mean squared error | 0.2831 | |
| Relative absolute error | 54.0976% | |
| Root relative squared error | 93.8998% | |
| *Total number of instances* | 99 | |

Table 8 gives the output summary of Random Forest and its performance output gave 90 instances that were correctly classified out of a total of 100, and 9 incorrectly classified instances. The Kappa statistic of 0.4226 is within the acceptable margins.

**Table 8** Random forest output summary

| | | |
|---|---|---|
| Correctly classified instances | 90 | 90.9091% |
| Incorrectly classified instances | 9 | 9.0909% |
| Kappa statistic | 0.4226 | |
| Mean absolute error | 0.0186 | |
| Root mean squared error | 0.2554 | |
| Relative absolute error | 57.5632% | |
| Root relative squared error | 84.7101% | |
| *Total number of instances* | 99 | |

The foregoing presentation highlights the results of the classifications, and the revelations are discussed as follows:

We can clearly see that the highest accuracy is 92.9293% and the lowest is 90.9091%. The other algorithm yields an average accuracy of around 90%. J48 is capable of generalising well. It exhibited the best performance with Correctly Classified Instances at 92.9293% and Relative absolute error of 43.6634 %. Whiles Bayes Net came in second with Correctly Classified Instances at 92.9293% and Relative absolute error of

46.6588%. Even with various experimentation with parameters in different models, it was not enough to beat J48.

Random forest came in last at around 90%. An average of 90 instances out of total 100 instances is found to be correctly classified with highest score of 92 instances compared to 90 instances, which is the lowest score. We are using the Kappa statistic to determine the accuracy of any particular measuring cases, it is the used to distinguish between the reliability of the data collected and their validity. The average Kappa score from the selected algorithm is around 0.4–0.6. Based on the Kappa Statistic criteria, the accuracy of this classification purpose is substantial.

And Thus, the classification discloses that the Decision Tree classifier (J48) performed very well with a percentage prediction of 92%. The other algorithm that compared was the Bayesian Classifier, BayesNet which also produced a 92% prediction rate. The other analysed classifiers that is PART and Random Forest even though they performed well at 90%, did not outperform their two counterparts.

## 4 Conclusions

The study sought to identify factors that would affect higher education institutions and university student's ability to graduate on schedule, and the selection process revealed, gender, assessments scores both in final examinations and continuous assessments, mode of study and sponsorship status as being dominant factors that would influence graduation times. Further the research applied J48, PART, Random Forest and Bayes Net, to demonstrate how data mining algorithms can be applied as predictive tools of determining expected graduation dates. The algorithm so chosen, due to literature supported popular use in field of education. The researcher also demonstrated a comparative analysis of the applied algorithms to determine which data mining algorithm provides accurate predictions of when a student is likely to graduate.

The findings in this study can be beneficial to higher learning institutions who would seek uncover factors that would lead to learners falling behind and subsequently identify learners falling behind in terms of graduation schedule and ultimately provide remedial measures. The study will also help academic and management team to provide additional consideration to improve learner's standing with regards the scheduled graduation time and aid them remain on track. The outcome of this study can help academic administration and sponsors to save on resources that would otherwise be expended towards learners that go beyond the anticipated graduation time.

For future studies, the researcher would be interested to study how additional attributes other than the ones considered in this study, would influence the performance of the classification algorithms.

## References

Abu Tair, M.M. and El-Halees, A. (2012) 'Mining educational data to improve students' *Performance: a Case Study. International Journal of Information*, Vol. 2, No. 2, pp.140–146.

Ahmad, F., Ismail, N.H. and Aziz, A.A. (2015) 'The prediction of students' academic performance using classification data mining techniques', *Applied Mathematical Sciences*, Vol. 9, No. 129, pp.6415–6426.

Al-Barrak, M.A. 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, p.528.

Al-Malaise, A. (2013) 'Implementation of apriori algorithm to analyze organization data: building decision support system', *International Journal of Computer Applications.*, Vol. 66, No. 9.

Almarabeh, H. (2017) 'Analysis of students' performance by using different data mining classifiers', *International Journal of Modern Education and Computer Science*, Vol. 9, No. 8, p.9.

Aluko, R.O., Daniel, E.I., Shamsideen Oshodi, O., Aigbavboa, C.O. and Abisuga, A.O. (2018) 'Towards reliable prediction of academic performance of architecture students using data mining techniques', *Journal of Engineering, Design and Technology*, Vol. 16, No. 3, pp.385–397.

Alyahyan, E. and Düştegör, D (2020) 'Predicting academic success in higher education: literature review and best practices', *International Journal of Educational Technology in Higher Education*, Vol. 17, No. 1, p.3.

Anoopkumar, M. and Rahman, A.M.Z. (2016) 'A review on data mining techniques and factors used in educational data mining to predict student amelioration', *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, March, IEEE, Ernakulam, India, pp.122–133.

Anuradha, C. and Velmurugan, T. (2015) 'A comparative analysis on the evaluation of classification algorithms in the prediction of students performance', *Indian Journal of Science and Technology*, Vol. 8, No. 15, pp.1–12.

Bavisi, S., Mehta, J. and Lopes, L. (2014) 'A comparative study of different data mining algorithms', *International Journal of Current Engineering and Technology*, Vol. 4, No. 5, pp.3248–3252.

Bhardwaj, B.K. and Pal, S. (2012) *Data Mining: A Prediction for Performance Improvement Using Classification*, arXiv preprint arXiv: 1201.3418.

Bhargava, N., Sharma, G., Bhargava, R. and Mathuria, M. (2013) 'Decision tree analysis on j48 algorithm for data mining', *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 6, pp.1114–1119.

Borges, L.C., Marques, V.M. and Bernardino, J. (2013) 'Comparison of data mining techniques and tools for data classification', *Proceedings of the International C* Conference on Computer Science and Software Engineering*, July, Porto, Portugal, pp.113–116.

Braganca, R., Portela, F. and Santos, M. (2019) 'A regression data mining approach in Lean Production', *Concurrency and Computation: Practice and Experience*, Vol. 31, No. 22, p.e4449.

Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F. and Alowibdi, J.S. (2017) 'Predicting student performance using advanced learning analytics', *Proceedings of the 26th International Conference on World Wide Web Companion*, April, Perth, Australia, pp.415–421.

ElGamal, A. (2013) 'An educational data mining model for predicting student performance in programming course', *International Journal of Computer Applications*, Vol. 70, No. 17, pp.22–28.

Goga, M., Kuyoro, S. and Goga, N. (2015) 'A recommender for improving the student academic performance', *Proceedings of the 6th International Conference Edu World 2014 "Education Facing Contemporary World Issues"*, pp.1481–1488.

Hamoud, A., Hashim, A.S. and Awadh, W.A. (2018) 'Predicting student performance in higher education institutions using decision tree analysis', *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 5, pp.26–31.

Han, J., Kamber, M. and Pei, J. (2012) 'Classification: advanced methods', *Data Mining Concepts and Techniques*, pp.393–443.

Hooshyar, D., Pedaste, M. and Yang, Y. (2019) 'Mining educational data to predict students' performance through procrastination behavior', *Entropy*, Vol. 22, No. 1, p.12.

Jalota, C. and Agrawal, R. (2019) 'Analysis of educational data mining using classification', *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing* (*COMITCon*), IEEE, February, pp.243–247.

Jayakameswaraiah, M. and Ramakrishna, S. (2014) 'A study on prediction performance of some data mining algorithms', *International Journal*, Vol. 2, No. 10, pp.141–144.

Kovacic, Z. (2010) 'Early prediction of student success: mining students' enrolment data', *Informing Science + Information Technology Education Joint Conference*, Cassino, Italy, pp.647–665.

Kumar, M., Singh, A.J. and Handa, D. (2017) 'Literature survey on student's performance prediction in education using data mining techniques', *International Journal of Education and Management Engineering*, Vol. 7, No. 6, pp.40–49.

Kumar, R. and Verma, R. (2012) 'Classification algorithms for data mining: a survey', *International Journal of Innovations in Engineering and Technology (IJIET)*, Vol. 1, No. 2, pp.7–14.

Lekha, K.C. and Prakasam, S. (2017) 'Data mining techniques in detecting and predicting cyber crimes in banking sector', *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, pp.1639–1643

Mesarić, J. and Šebalj, D. (2016) 'Decision trees for predicting the academic success of students', *Croatian Operational Research Review*, Vol. 7, No. 2, pp.367–388.

Mohamed, M. and Waguih, H.M. (2017) 'Early prediction of student success using a data mining classification technique', *International Journal of Science and Research*, Vol. 6, No. 10, pp.126–131.

Moscoso-Zea, O., Saa, P. and Luján-Mora, S. (2019) 'Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining', *Australasian Journal of Engineering Education*, Vol. 24, No. 1, pp.4–13.

Mueen, A., Zafar, B. and Manzoor, U. (2016) 'Modeling and predicting students' academic performance using data mining techniques', *International Journal of Modern Education & Computer Science*, Vol. 8, No. 11, pp.36–42.

Oskouei, R.J. and Askari, M. (2014) 'Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies)', *Computer Engineering and Applications Journal*, Vol. 3, No. 2, pp.79–88.

Osmanbegovic, E. and Suljic, M. (2012) 'Data mining approach for predicting student performance', *Economic Review: Journal of Economics and Business*, Vol. 10, No. 1, pp.3–12.

Pandey, U.K. and Pal, S. (2011) *Data Mining: A Prediction of Performer or Underperformer using Classification*, arXiv preprint arXiv:1104.4163.

Peling, I.B.A., Arnawan, I.N., Arthawan, I.P.A. and Janardana, I.G.N. (2017) 'Implementation of data mining to predict period of students study using naive Bayes algorithm', *Int. J. Eng. Emerg. Technol.*, Vol. 2, No. 1, p.53.

Putpuek, N., Rojanaprasert, N., Atchariyachanvanich, K. and Thamrongthanyawong, T. (2018) *Comparative Study of Prediction Models for Final GPA Score: A Case Study of Rajabhat Rajanagarindra University*, s.l., June.

Raju, D. and Schumacker, R. (2015) 'Exploring student characteristics of retention that lead to graduation in higher education using data mining models', *Journal of College Student Retention: Research, Theory & Practice*, Vol. 16, No. 4, pp.563–591.

Ranbaduge, T. (2013) 'Use of data mining methodologies in evaluating educational data', *International Journal of Scientific and Research Publications*, Vol. 3, No. 11.

Sembiring, R.W., Zain, J.M. and Embong, A. (2011) *A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course*, arXiv preprint arXiv:1101.4270.

Shazmeen, S.F., Baig, M.M.A. and Pawar, M.R. (2013) 'Performance evaluation of different data mining classification algorithm and predictive analysis', *Journal of Computer Engineering*, Vol. 10, No. 6, pp.01–06.

Shovon, M., Islam, H. and Haque, M. (2012) *An Approach of Improving Students Academic Performance by using k Means Clustering Algorithm and Decision Tree*, arXiv preprint arXiv:1211.6340.

Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J. and Cattrysse, D. (2015) 'Literature review of data mining applications in academic libraries', *The Journal of Academic Librarianship*, Vol. 41, No. 4, pp.499–510.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J. and Abreu, R. (2015) *A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance*, International Educational Data Mining Society.

Umadevi, S. and Marseline, K.J. (2017) 'A survey on data mining classification algorithms', *2017 International Conference on Signal Processing and Communication (ICSPC)*, July, IEEE, pp.264–268.

Veeramuthu, P. and Periasamy, R. (2014) 'Application of higher education system for predicting student using data mining techniques', *International Journal of Innovative Research in Advanced Engineering*, Vol. 1, No. 5, pp.31–36.

Whitley, L.A. (2018) *Educational Data Mining and Its Uses to Predict the Most Prosperous Learning Environment*, Master's Thesis, East Carolina University, April, Retrieved from the Scholarship.

Yadav, S.K., Bharadwaj, B. and Pal, S. (2012) *Data Mining Applications: A Comparative Study for Predicting Student's Performance*, arXiv preprint arXiv:1202.4815.

Zaffar, M., Hashmani, M.A. and Savita, K.S. (2017) 'Performance analysis of feature selection algorithm for educational data mining', *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, November, IEEE, pp.7–12.

Zhang, Y., Oussena, S., Clark, T. and Kim, H. (2010) 'Using data mining to improve student retention in higher education: a case study', *International Conference on Enterprise Information Systems*, Portugal.