Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

The Development of Student Academic Performance Prediction System for UPTM using RepTree Algorithm

Noornajwa Md Amin[1,*], Siti Robaya Jantan[1], Ramlan Mahmod[1], Nor Hafiza Haron[1], Airuddin Ahmad[1], Omar Al Tarawneh[2]

[1] Faculty of Computing and Multimedia, Universiti Poly-Tech Malaysia, Taman Shamelin Perkasa, 56100 Kuala Lumpur, Malaysia
[2] Department of Software Engineering, Amman Arab University, Amman, Jordan

ABSTRACT

Accurate prediction of students' performance is critical in this digital age for educational institutions to identify at-risk pupils and provide timely interventions. Numerous models have been proposed under different educational contexts to address it, but there is a lack of sophisticated models causing difficulty for the user in giving guidance to the stakeholders to take appropriate measures to counter student's problems. The RepTree algorithm is a decision tree-based machine learning approach to modelling and forecasting student academic achievement. This study adopted it to develop a Student Academic Performance Prediction System (SAPPS) for the University Poly-Tech Malaysia (UPTM). This study is using a mixed-method research design, based on historical student data to train the predictive model, including demographics, prior academic performance, and sponsorship attributes. The evaluation result shows how well the established system predicts students' academic success and how reliable it is. The system's predictive capabilities give educational institutions the ability to spot students who are likely to have academic difficulties and take proactive measures to improve their results. The use of such a predictive system could improve the learning environment and boost students' achievement at UPTM. By having this information, the target users like administrators, lecturers, and academic advisers can access and decipher the predictions produced by the RepTree algorithm using the suggested system's user-friendly interface.

Keywords:
Prediction model; Artificial intelligence; Student's performance; RepTree algorithm; Education

1. Introduction

Educational institutions primarily rely on their students as their main product. Therefore, students' academic performance is of utmost importance to these institutions. The ability of an educational institution to produce high-quality graduates largely depends on its capabilities. Consequently, most educational institutions aim to uphold high standards of teaching and learning to attract prospective students and maintain their reputation. However, some institutions prioritize their prestige over the quality of education they provide [1]. As mentioned by Chi *et al.,* [42]

* *Corresponding author.*
E-mail address: n_najwa@uptm.edu.my

https://doi.org/10.37934/araset.60.1.5981

integrating various technological approaches from different disciplines into higher education is becoming increasingly popular. Many universities are now adopting this practice to improve student learning, cultivate practical skills, and maintain high educational standards. Most governments have established accreditation agencies to ensure that educational institutes adhere to strict procedures and uphold high standards of quality in their learning environment. The ultimate objective of these institutions is to produce graduates of exceptional quality. To achieve this objective, they must closely monitor students' academic performance by gathering as much information as possible and providing recommendations to maintain good performance and prevent academic issues [2].

Project-based learning (PBL) proves to be an effective method for cultivating leadership abilities in student teachers [43]. By incorporating technology and various instruments to monitor academic performance, educational establishments can guarantee the optimization of the advantages of PBL. At the beginning of a course, instructors may not be able to distinguish students' levels of knowledge. However, once struggling students are identified, instructors can design preventive measures to help them. Therefore, it is essential to develop sophisticated prediction models that can forecast students' performance and enable instructors to provide adequate support to struggling students. Previous studies have shown that machine learning algorithms such as Naïve Bayes, decision tree, neural networks, detections by outliers, and sophisticated statistics are efficient and productive tools for developing models to monitor and predict students' academic performance [3].

However, existing models are primarily tailored to specific local institutions and are only effective for a single course. Therefore, this research aims to develop an AI-based approach to monitor and predict the academic performance of an entire program taught at UPTM. This research's primary goal is to monitor and interpret students' ongoing performance and develop a prediction model that can guide stakeholders in taking appropriate measures to address students' academic issues.

The key research questions addressed in this study are:

i. Which variables or attributes have a strong significant in predicting the student academic performance risk status?
ii. Which model or algorithm had the highest accuracy in predicting the student's academic performance?
iii. What are the most suitable tools and programming languages for the construction and implementation of the system?

Corresponding to these questions, the research objectives are:

i. To identify the attributes for predicting student's academic performance
ii. To propose a decision tree model and RepTree algorithm to discover the essential features which influence student's academic performance
iii. To develop a Student Academic Performance Prediction System (SAPPS) for UPTM using RepTree algorithm.

Based on the past literature, there are some difficulties in identifying the attributes that influence student academic performance, a lack of sophisticated models that can effectively guide users in monitoring student academic performance, and the absence of tools that can assist institutions in monitoring student academic performance.

Providing clearer guidance and support is crucial for enhancing students' academic performance and ensuring they are well-equipped to thrive in the context of the Fourth Industrial Revolution (IR4.0) [44]. Based on the research conducted by Abdalla *et al.,* [4], identifying the most relevant

attributes for predicting student performance may be complex. The researcher discussed using different attributes, including academic and uncertainty, to predict student performance. As stated by Abdallah *et al.,* [5], identifying the dominant factors impacting student outcomes, especially attributes that influence student academic performance, can be complex. However, the paper also highlighted that student online learning activities, term assessment grades, and academic emotions were the most evident predictors of learning outcomes. Jie Xu *et al.,* [6] mentioned the challenges of predicting student performance in completing college programs due to the diversity of courses selected by students and the requirement of continuous tracking and incorporation of students' evolving progress. The paper proposed a novel algorithm that utilizes education-specific domain knowledge to enable progressive prediction of students' performance.

Besides that, the lack of sophisticated prediction models caused the user difficulty in guiding the stakeholders to take appropriate measures to counter student's problems. According to Shouq *et al.,* [7], predicting students' academic performance is a challenging task that educational systems face every semester, and the proposed hybrid method aims to build robust strategies to enhance educational systems. Jastini *et al.,* [8] stated that the paper mentions a need for more investigations on exploring patterns of students' behaviour that affect their academic performance within the Malaysian context. It also suggests that there is a need for more sophisticated models to effectively guide users in monitoring student's academic performance.

Research conducted by Samantha *et al.,* [9] mentioned that there is a need for practical and easy-to-implement tools to monitor student progress. Implementing Student Academic Performance Systems (SAPS), which are analytical instruments to monitor student advancement, can improve academic development and learning outcomes. Nevertheless, the effectiveness of these monitoring systems is contingent upon their ease of implementation, clarity of interpretation, and flexibility of framework to suit various educational levels and course styles. Deevika *et al.,* [10] adopted data mining techniques, and various institutions used it to locate hidden data of the student and identify functional information from massive data. The research paper also focuses on developing methods to locate weak students and improve their cognitive levels and academic performance. Therefore, there is a need for practical tools to monitor and improve student's academic performance.

2. Literature Review

2.1 Educational Data Mining

Educational data mining (EDM) is a well-established research field focused on extracting and analysing data from various sources in educational institutions according to user-defined patterns [11]. Its primary goal is to uncover new insights and hidden patterns within student data [3]. Many models and techniques have been developed in different educational contexts to address student performance prediction. To build a predictive model, several tasks such as classification, regression, and clustering have been employed, along with algorithms such as Bayesian Network (BN), Decision Tree (DT), Artificial Neural Networks (ANN), Naive Bayes (NB), K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) [11].

In a recent study by Kausar *et al.,* [12], ensemble techniques were utilized to investigate the relationship between students' semester courses and final results. The experimental results revealed that Random Forest and Stacking Classifiers achieved the highest accuracy. Similarly, Chen and Ward [13] developed models using decision tree and linear regression, employing features extracted from the institution's auto-grading system. This research helps the institution identify struggling students and allocate teaching hours efficiently.

Hamsa *et al*., [14] proposed a decision tree model and fuzzy logic algorithm to identify the essential features that affect students' academic performance. The data on students' demographic, academic and social behaviour was collected via a survey. Latrellis *et al.,* [15] proposed a machine learning approach that uses the K-Means algorithm to generate a set of coherent clusters, followed by supervised machine learning algorithms to train prediction models for predicting students' performance. Additionally, Yaacob *et al.,* [16] developed predictive models using classification algorithms to predict students' performance as either excellent or non-excellent based on their academic performance results via educational data mining.

2.2 Attributes for Predicting Student's Performance

Vijayalakshmi *et al.,* [17] assert that attributes and methods are significant factors in determining student performance. Figure 1 presents the essential attributes utilized in forecasting student performance. Ten (10) attributes are formed by consolidating standard classes. This figure is crafted after a comprehensive literature review of various articles about utilizing data mining techniques in student performance models.

The conducted research chose four common attributes for analysis, namely, Cumulative Grade Point Average (CGPA), Grade Point Average (GPA), demographic attribute (specifically gender), and sponsorship. The systematic review revealed that CGPA is frequently used in predicting student performance [18-25]. On the other hand, GPA is a widely used indicator of academic success, with many universities requiring a minimum GPA to be met. Therefore, GPA remains a typical element academic planners consider when assessing academic development. Several factors can hinder students from achieving and maintaining a high GPA during their college years, which measures their overall academic performance [24]. This research also chose GPA as one of the attributes to monitor and predict student performance [18,24].

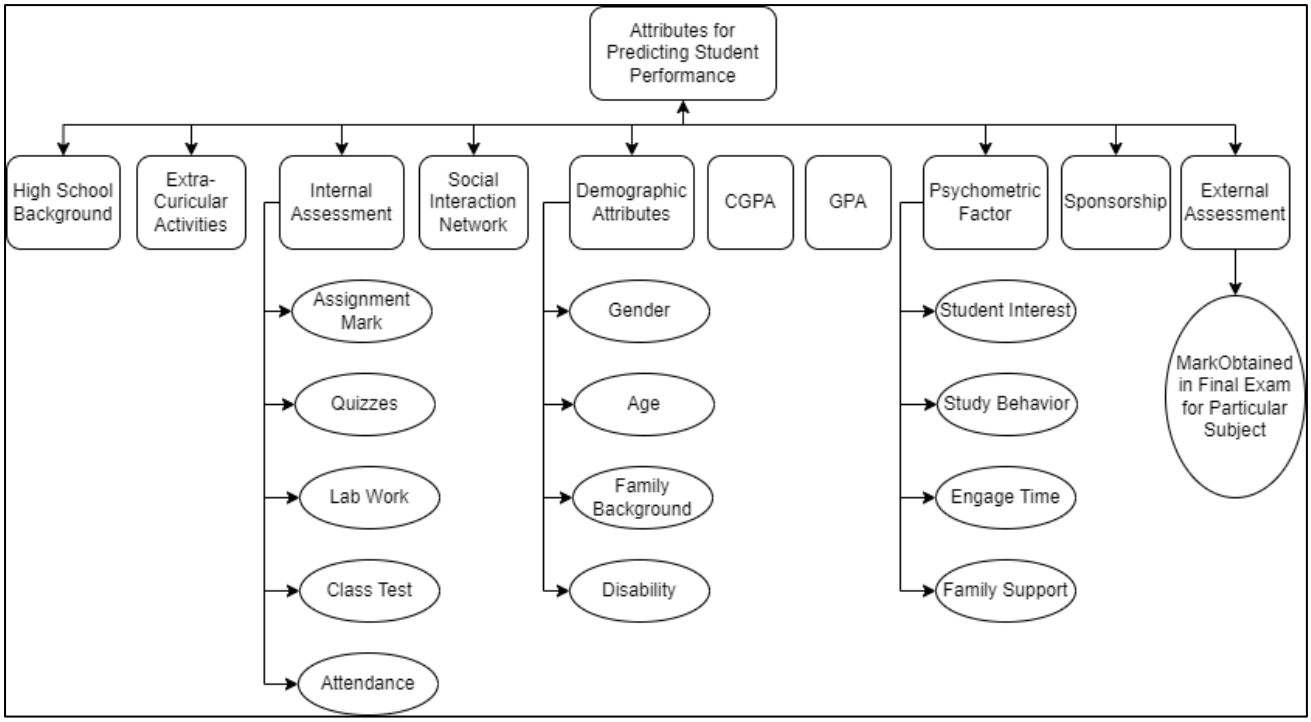


Fig. 1. Attributes for predicting student performance

Another attribute that significantly influences student performance is sponsorship. This factor, often overlooked, plays a crucial role in shaping a student's academic journey [2,26]. The research, based on the raw data collected from University Polytech Malaysia (UPTM) in the Faculty of Computing and Multimedia (FCOM), has identified sponsorship as a key attribute in predicting student performance. The selection of these attributes was based on their validity, novelty, potential usefulness, and understandability, providing a comprehensive understanding of the research.

2.3 Algorithm for Predicting Student's Performance

Many classification algorithms have been employed in educational data mining to predict student achievement, such as Decision Tree, Naive Bayes, K-Nearest Neighbor, Neural Network and Support Vector Machine. In light of this, the current study, "An Artificial Intelligence Approach to Monitor and Predict Student Academic Performance," conducted a systematic literature review on student performance and opted to utilize only three distinct methods or algorithms: Decision Tree, Naïve Bayes, and K-Nearest Neighbor.

2.3.1 Decision tree

Decision Trees have gained popularity as a favoured method of prediction. The simplicity and readability of this technique have prompted many academics to employ it in identifying data structures of varying sizes and forecasting values [23,26,27]. The Decision Tree models are easy to comprehend owing to their straightforward logic and ability to translate into a set of IF-THEN rules [28]. A strategy that made use of a Decision Tree was utilized. The Decision Tree method has been previously used in research to predict academic dropout characteristics in student data [27], identify the best career for a student based on their behavioural patterns [29], and forecast MCA students' performance in the third semester. Moreover, decision trees utilize a top-down, recursive divide-and-conquer approach. Their functioning properties and growing use as a prediction method in data science have made these algorithms useful in identifying valuable models. Researchers have chosen this method due to its simplicity and adaptability to all sizes of data structures for value estimation. Decision trees are easily understood as they take on a tree-like form, which applies classification principles to actual human reasoning [30].

2.3.2 Naïve-Bayes

According to Viet *et al.,* the Naive Bayes algorithm is utilized for problems involving multiple classes. The nomenclature "Naive Bayes" is assigned due to the direct calculation of probabilities for each class. Bayesian classification techniques serve as the fundamental basis for Naive Bayes classifiers. These techniques rely on Bayes' theorem, which is an equation that defines the connection between conditional probabilities of statistical quantities [31].

2.3.3 IBK (K-Nearest Neighbor)

IBK is a classifier which belongs to the K-Nearest Neighbor family. The development of this classifier necessitates minimal effort and is typically completed concurrently with the classification process. This approach has earned its moniker as the "lazy learning" technique [32]. To simplify the process of locating the closest neighbours, different search algorithm combinations may be employed. While linear search is the most widely utilized method, KD-trees, ball trees, and cover

trees are viable alternatives. According to Vijayarani and Muthulakshmi, the distance from the test instance can be utilized to weight predictions made by considering several neighbours. To translate the distance into the weight, two separate formulas are used [33].

2.3.4 Neural networks

Gray *et al.,* [34] have noted that neural networks are a popular method for data mining in education. The primary advantage of neural networks is their ability to identify all possible interactions between predictor variables, even in complex nonlinear relationships between dependent and independent variables, resulting in comprehensive detection. A primary neural network comprises three layers: the input layer, the hidden layer, and the output layer, each containing varying numbers of neurons. The number of hidden layers required for a network of this nature has yet to be fundamentally researched, but one or two layers may be used to constitute the hidden layer. The hidden layers determine the network size, and a more extensive network requires more extended training. In hidden layers, the output of one layer serves as the input for the subsequent layer. The transfer function of a neuron transforms the input into the neuron's output. Compared to single-layer neural networks, multilayer neural networks are much more potent tools to solve various problems [35].

2.3.5 Support vector machine

Janan *et al.,* [36] assert that support vector machines are supervised machine learning techniques that utilize associated learning methodologies for regression and classification tasks. These methods can be classified as discriminative classifiers as they can determine separating hyperplanes. The critical characteristic of SVM is its ability to minimize the empirical classification error and maximize the geometric margin. This property has led to SVM being referred to as a maximum margin classifier. The foundation of SVM is the Structural Risk Minimization (SRM) technique. The input vectors from SVM are mapped to a higher-dimensional space where a maximal separation hyperplane is constructed. Two parallel hyperplanes are built on either side of the hyperplane, which divides the data. The hyperplane that maximizes the separation between the two parallel hyperplanes is called the separating hyperplane [37].

3. Methodology

This section describes the research methodology used in the project, which covers all the processes involved in developing the Student Academic Performance Prediction System for UPTM using the RepTree Algorithm, including phases, activities, methods, and deliverables. The mixed method research paradigm using both qualitative and quantitative research approaches was adopted in this study in order to collect all data related to the project. This research methodology will show the flow of the project and ensure that the activities done in each phase are mapped with the objectives stated in the previous section.

3.1 Mixed Method Approach

Four phases in this mixed method approach should be followed to complete the research entitled The Development of Student Academic Performance Prediction System for UPTM using the RepTree Algorithm. It comprises user requirements and specifications, system architecture design, coding and

implementation, and software testing phases. This mixed-method approach entails the combination of both qualitative and quantitative techniques. The utilization of the qualitative approach was observed throughout all phases. In contrast, in the second phase, the application of the quantitative approach was evident in the data preparation and preprocessing technique. This research covers coding and implementation only, where the last software testing process will be covered in a future enhancement. Figure 2 shows the phases flow to conduct this research.

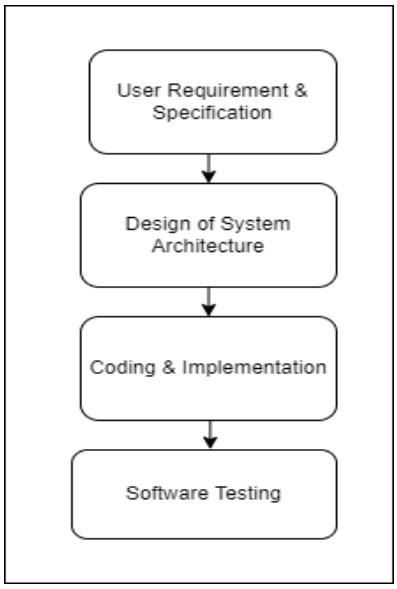


Fig. 2. Mixed method approach

3.1.1 Phase 1: User requirement and specification

Three main activities had been identified in phase 1 shown in Table 1 below that involved purposing to ensure that the objective, scope, and research gap had been defined. Table 1 below summarizes all the activities, objectives, methods, and deliverables that already being stated in the introduction section followed by the literature review section. This first phase adopted the qualitative technique to fulfil all the objectives stated in the table below.

Table 1
Research operational framework phase 1

Phase 1: Research Operational Framework Phase 1			
Activities	Objectives	Methods	Deliverables
Activity 1: Identify and understand the research problem	To identify the objectives, scope, and significance of the research.	Literature Analysis (Search Method)	Defined objectives, and scope of the study.
Activity 2: Review the literature in the domain of numerous models that have been proposed under different educational contexts to address student performance prediction.	To identify the prediction model in the educational context.	Literature Analysis (Search Method) Comparative Study	A defined research gap based on previous case studies related to the prediction model is selected.
Activity 3: Identifying system specifications and requirements	To identify the tools and programming language for the construction and	Literature Analysis	Defined the suitable tools and programming language.

	implementation of the system.	(Search Method) Comparative Study

Based on all the requirements already documented in sections 1 (introduction) and 2 (literature review), the data collection has been summarised in the Figure 3 flowchart below. The flowchart shows the system flows that consist of uploading raw data, previewing data sets, training or testing predictions, and displaying sample data sets. After testing the data set, the rule-based approach identified based on the RepTree classification and algorithm result will be set in the system to ease the prediction of the risk to students' academic performance. Users need to input all the student information such as name, ID, semester, gender, sponsorship, GPASem1, GPASem2, GPASem3, GPASem4 and CGPA, and the prediction class will be displayed either low, medium or high risk based on the rules that had been set in the code. The mitigation also will be displayed after the prediction process.

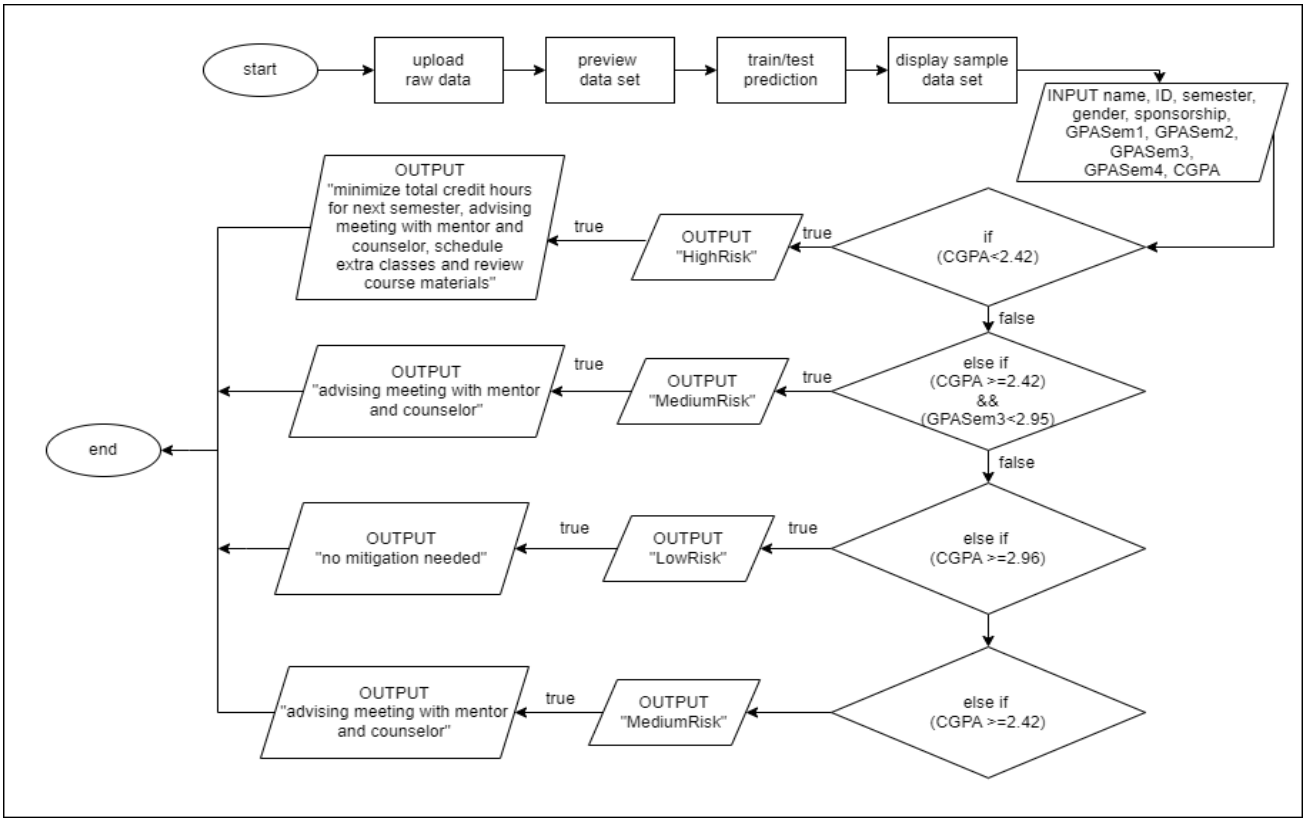


Fig. 3. Flowchart

3.1.2 Phase 2: Design of system architecture

Phase 2 consists of three activities, as shown in Table 2 below, that involve identifying the population, designing general architecture, and data preparation. The purpose of the first activity is to identify the population for this project to get the sample of a dataset based on the attributes of students identified in Phase 1. By using the literature analysis method, the population frame was decided, and data was collected from the Admission and Student Record Unit of University Poly-Tech Malaysia (UPTM) related to students' academic performance. The training dataset focuses on

students from the Faculty of Computing and Multimedia, covering four semesters with more than 150 instances.

Table 2
Research operational framework phase

Phase 2: Research Operational Framework Phase 2			
Activities	Objectives	Methods	Deliverables
Activity 1: To identify the population	To get the sample	Literature Analysis	Population frame
Activity 2: Designing general architecture of the application which involves the stakeholders and computing components.	To design the system architecture	Literature Analysis	System Architecture
Activity 3: Data Preparation	To get the clean dataset	Data Description Data Cleansing Data Pre-processing	Clean dataset

The second activity of the project phase 2 pertains to the design of the application's general architecture. The objective of this phase has been accomplished, as evidenced by Figure 4 below. The pertinent data regarding the students has been sourced from the Campus Management System (CMS) of UPTM. This overarching architecture encompasses all relevant stakeholders and computing components, including the students, university, parents, and student sponsors. The flow of the process is straightforward and streamlined. The system maintains a secure server for the student's academic records. Additionally, it is equipped with an internal AI-based data-defining and evaluating module and a Data Analytic engine that can generate reports, predictions, and suggestions on demand by the stakeholders. Users may access the system from static or mobile devices at any time and location to review the student's performance.

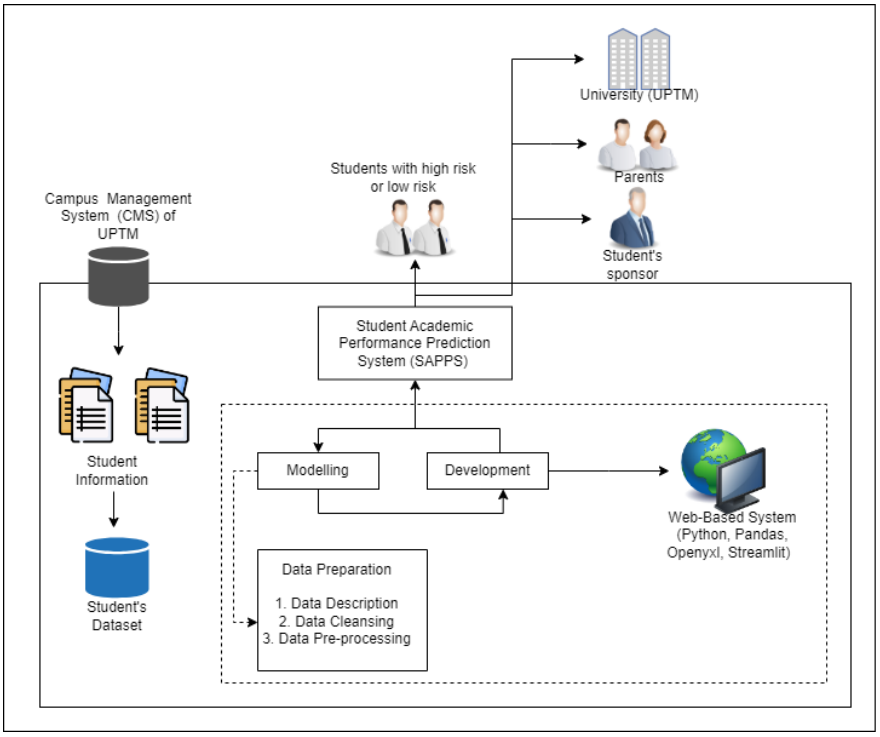


Fig. 4. System architecture

Meanwhile, the quantitative approach was employed, utilizing the data preparation and processing technique. This stage is integral to system design, with emphasis placed on preparing data for the system. As such, the data will undergo three stages, including data description, data cleansing, and data pre-processing. Table 3 below shows the data description, which consists of each data set's features, descriptions, and values.

Table 3
Data description of UPTM students

Features	Description	Value
Gender	Student's gender either male (M) or female (F)	Nominal (M or F)
Sponsorship	Whether sponsored for study by MARA or others either yes (Y) or no (N)	Nominal (Y or N)
GPASem1	Grade Point Average of Student for semester 1	Numeric (0-4)
GPASem2	Grade Point Average of Student for semester 2	Numeric (0-4)
GPASem3	Grade Point Average of Student for semester 3	Numeric (0-4)
GPASem4	Grade Point Average of Student for semester 4	Numeric (0-4)
CGPA	Cumulative Grade Point Average of Student	Numeric (0-4)
Class	Prediction either low, medium or high risk	Nominal (LowRisk, MediumRisk, HighRisk)

Data cleansing is an essential process that involves the elimination of anomalies, such as irrelevant features, from existing data to ensure the accuracy and uniqueness of the data collection. In dealing with inaccurate, inconsistent, and incomplete data, as illustrated in Table 4, the information about UPTM students that comprises irrelevant features will be eliminated.

Table 4
Examples of noisy instances in the dataset

Gender	Sponsorship	GPASem1	GPASem2	GPASem3	GPASem4	CGPA	Class
M	Y	3.78	3.13	3.26	3.06	3.31	LowRisk
M	Y	2.78	1.89	?	2.13	2.57	MediumRisk
M	Y	3.72	3.82	3.15	?	3.46	LowRisk
M	Y	3.83	3.83	3.89	?	3.75	?

The final stage of the research involves the pre-processing of data. Data pre-processing will be done using the Waikato Environment for Knowledge Analysis (WEKA), which provides a "filter" option. WEKA has two types of filters, supervised and unsupervised, which can filter attributes and instances separately. WEKA is an open-source software that offers a wide range of algorithms for data pre-processing, classification, clustering, regression, and association rules. For the purpose of this study, four commonly used machine learning algorithms were selected. The k-Nearest Neighbors (k-NN) algorithm, which is implemented as IBk in WEKA, was chosen from the lazy algorithms. The RepTree algorithm, which implements a decision tree in WEKA, was also selected. In addition, the Multilayer Perceptron (MLP) algorithm, a type of artificial neural network (ANN), and the Naïve Bayes algorithm were chosen. Tenfold cross-validation was employed to evaluate the performance of the algorithms. According to Hastie *et al.,* [38], this technique involves splitting the training dataset into ten identical-length intervals. During each iteration, nine intervals are used for learning purposes, while the tenth is used for testing the algorithm's performance. This is an iterative process, and a new interval is selected for the testing part in each iteration. The successful execution of the process involved the application of appropriate data types, specifically numeric or nominal, in order to effectively filter and clean the data. Before filtering, the unsupervised attribute filter was applied, with the attribute indices being set to 3. The research utilized three distinct classification processes,

namely RepTree, k-NN, and Naïve Bayes. The accuracy of the instance, as summarized using Rep Tree, is presented in Figure 5 below.

```
=== Summary ===

Correctly Classified Instances          85               95.5056 %
Incorrectly Classified Instances         4                4.4944 %
Kappa statistic                          0.8419
Mean absolute error                      0.0308
Root mean squared error                  0.1712
Relative absolute error                 15.8525 %
Root relative squared error             56.1444 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.987    0.000    1.000      0.987    0.993      0.960    0.992     0.998     LowRisk
                 0.889    0.038    0.727      0.889    0.800      0.780    0.929     0.697     MediumRisk
                 0.600    0.012    0.750      0.600    0.667      0.654    0.794     0.472     HighRisk
Weighted Avg.    0.955    0.004    0.958      0.955    0.955      0.924    0.975     0.938

=== Confusion Matrix ===

  a  b  c   <-- classified as
 74  1  0 |  a = LowRisk
  0  8  1 |  b = MediumRisk
  0  2  3 |  c = HighRisk
```

Fig. 5. Summarization of RepTree classification result

The decision tree depicted in Figure 5 utilizes the RepTree algorithm with a 10-fold cross-validation approach. It has demonstrated a prediction accuracy of 95.50%. This suggests that the data set has been classified as possessing a positive value. A notably strong and statistically significant relationship exists between CGPA and GPASem3. Furthermore, the findings reveal that variables such as Gender, Sponsorship, GPASem1, and GPASem2 do not correlate with the dependent variable. The insignificance of variables such as Gender and Sponsorship can be attributed to their exclusion from the CGPA calculation. At the same time, GPASem1 and GPASem2 do not exert a considerable influence on the CGPA of students. Conversely, GPASem3 demonstrates a highly significant association with the CGPA, as it forms an integral part of its calculation.

The primary objective of prediction accuracy is to ascertain that the employed methodology and approach have yielded the appropriate quantification. Table 5 compares RepTree, k-NN, and Naive Bayes, revealing their respective predictive capacities. Specifically, k-NN exhibits a prediction accuracy of 88.76%, whereas Naïve Bayes demonstrates a higher accuracy of 94.38%. The findings indicate that RepTree surpasses both k-NN and Naïve Bayes in accuracy.

Table 5
Prediction Accuracy

RepTree Decision Tree	k-NN Lazy Algorithm	Naïve Bayes
95.50%	88.76%	94.38%

3.1.3 Phase 3: Coding and implementation

Table 6 below shows the coding and implementation phase involves the activity of analysing the data after the process of filtering and data cleaning. The data utilized in this study was procured from the academic archives of UPTM, with a specific focus on students from the Faculty of Computing and

Multimedia (FCOM) who are enrolled in the Diploma in Computing Science program (CC101). Following a rigorous process of filtering and data cleaning, a total of 89 data sets were deemed suitable for analysis.

Table 6
Research operational framework phase 3

Phase 3: Research Operational Framework Phase 3			
Activities	Objectives	Methods	Deliverables
Activity 1: Analyse the data after the process of filtering and data cleaning.	To find the most accurate method for student's academic performance prediction.	Classification process using WEKA	Algorithm and attributes that influence the students' performance had been identified.
Activity 2: Design the prototype.	To design web-based system based on findings	Sketch the user interface	Graphical User Interface
Activity 3: Develop web-based system for student's academic prediction.	To develop a web-based system using IF-THEN rules	Decision tree-based model IF-THEN rules	Academic Performance Prediction Web-Based System

The employment of the RepTree algorithm with 10-fold cross validation approach in Figure 6's decision tree has yielded a prediction accuracy of 95.50%. This observation suggests that the data set has been categorized as a positive value. A robust and meaningful relationship has been established between CGPA and GPASem3. Conversely, the outcomes indicate that the dependent variable exhibits no correlation with Gender, Sponsorship, GPASem1, and GPASem2. The insubstantiality of variables such as Gender and Sponsorship can be attributed to their exclusion from CGPA, while GPASem1 and GPASem2 have a negligible impact on student CGPA. On the other hand, GPASem3 displays a highly significant correlation with CGPA, as it constitutes a fundamental component of CGPA.

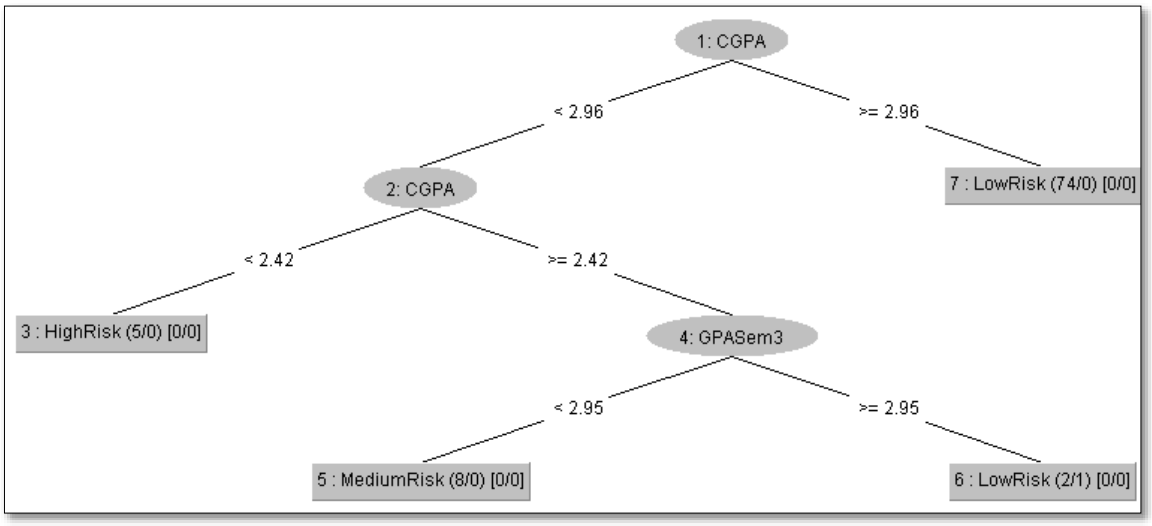


Fig. 6. Visualization of RepTree algorithm result

The deliverables for the second activity in phase 3 which design the prototype will be shown and be explain further in the next section of result and analysis. For the third activity, the development of a web-based system to monitor and predict student performance at UPTM entails the successful implementation of IF-THEN rules as shown in Figure 7 below. This model had been executed through a simple and comprehensible procedure, accompanied by a preventive approach for students who

exhibit poor performance. As a measure of support, the output model highlights the significance of GPASem3 as a key factor, thereby emphasizing the need for faculty members, including program coordinators, mentors, and subject lecturers, to consistently underscore its importance. Such measures have the potential to significantly reduce the number of inefficient students.

```
REPTree
============

CGPA < 2.96
|   CGPA < 2.42 : HighRisk (5/0) [0/0]
|   CGPA >= 2.42
|   |   GPASem3 < 2.95 : MediumRisk (8/0) [0/0]
|   |   GPASem3 >= 2.95 : LowRisk (2/1) [0/0]
CGPA >= 2.96 : LowRisk (74/0) [0/0]

Size of the tree : 7

Time taken to build model: 0 seconds
```

Fig. 7. Run information of RepTree

4. Results and Analysis

4.1 The Web-Based Prediction System

As a result, the Student Academic Performance Prediction System (SAPPS) for UPTM has been developed using the RepTree Algorithm by implementing the IF-THEN rules. Figure 8 below shows the SAPPS Homepage, consisting of three main menus that predict risk status, generate graphs and prediction form. This SAPPS system was developed using Python with packages that are commonly used for data analysis, visualization, and web application development, such as plotly version 4.14.3, pandas version 1.2.0, openpyxl version 3.0.6, and streamlit version 0.74.1. These packages, when combined, can form a powerful toolkit for data analysis and visualization and for building web applications with interactive visualizations.

Plotly is a powerful library for creating interactive and visually appealing visualizations in Python. It supports various chart types and is often used for data exploration and presentation. Pandas is a widely used library in Python for data manipulation and analysis. It provides data structures like DataFrames and Series and functions to handle and transform data efficiently. Openpyxl is a library that allows people to use Python to work with Excel files (both .xlsx and .xlsm formats). It enables reading, writing, and modifying Excel spreadsheets. Streamlit is a popular library that allows the creation of web applications in Python with minimal effort. It is designed for data scientists and developers to turn data scripts into shareable web apps quickly.

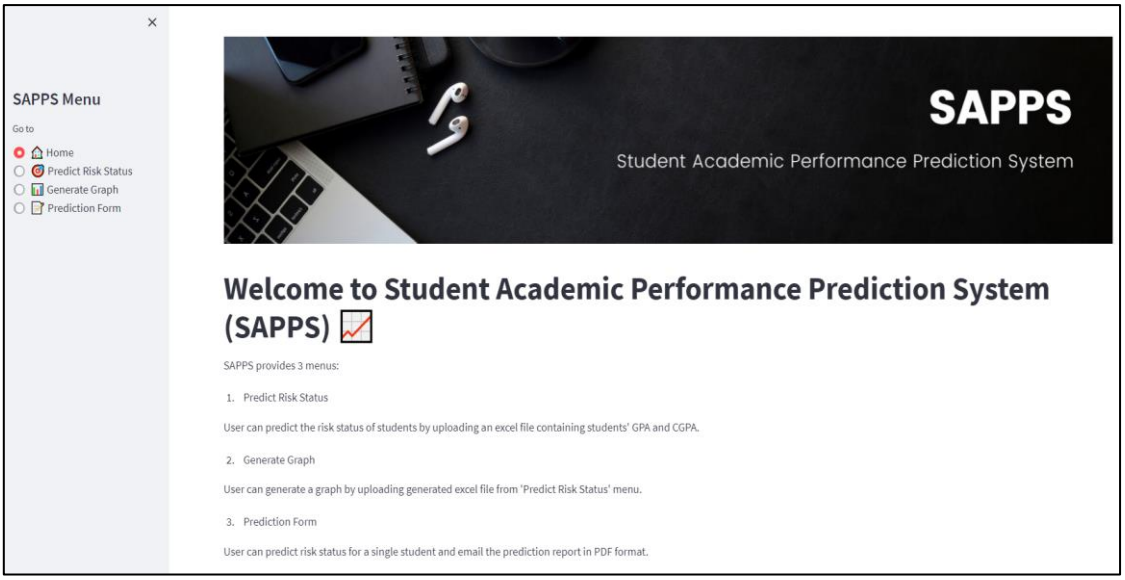


Fig. 8. Student Academic Performance Prediction System (SAPPS) homepage

The first menu predicts risk status, as shown in Figure 9 below. This menu displays the instructions on predicting the risk status by uploading the Excel file first. Before uploading the Excel file, the users need to ensure that they follow the rules which are in terms of file format (.xls and .xlsx only), file limit size (maximum 200 MB) and uploading the proper data train test that already goes through the process of filtering and data cleaning using WEKA. The user will be assisted with a user-friendly interface where the Excel file template and the data set example will be provided in this menu.

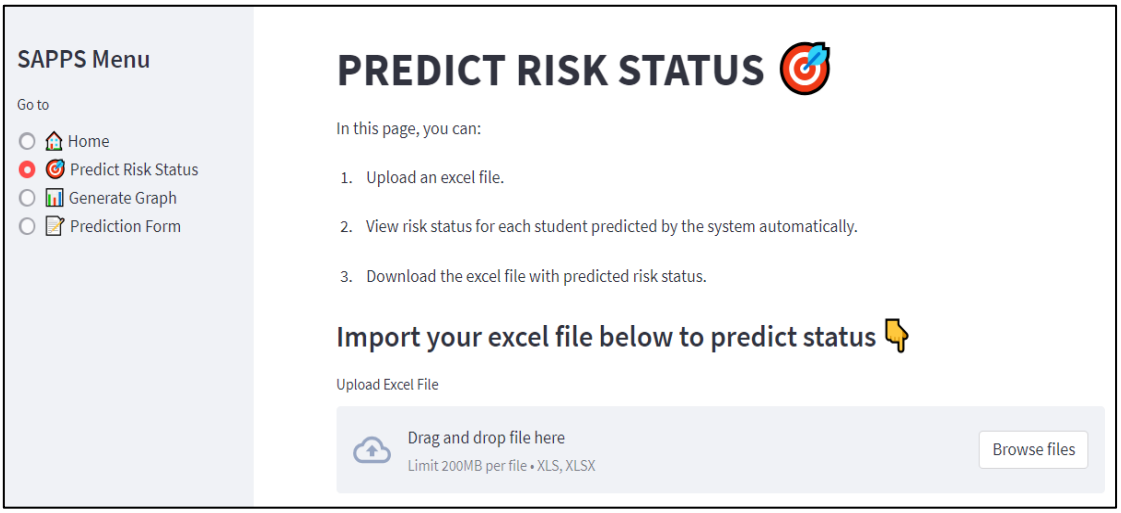


Fig. 9. Menu 1.1 - predict risk status

After the dataset or the original Excel file that consists of all attributes such as ID, semester, gender, sponsorship, GPASem1, GPASem2, GPASem3, GPASem4 and CGPA uploaded, all the data will display on the screen first, followed by the data with the student risk that already being identified using the IF-THEN rules that had been set in the code as shown in Figure 7. This menu is purposely for training the dataset, and the rules will be identified so that the system can directly use the same rule to predict the risk in the prediction form menu in Figure 10.

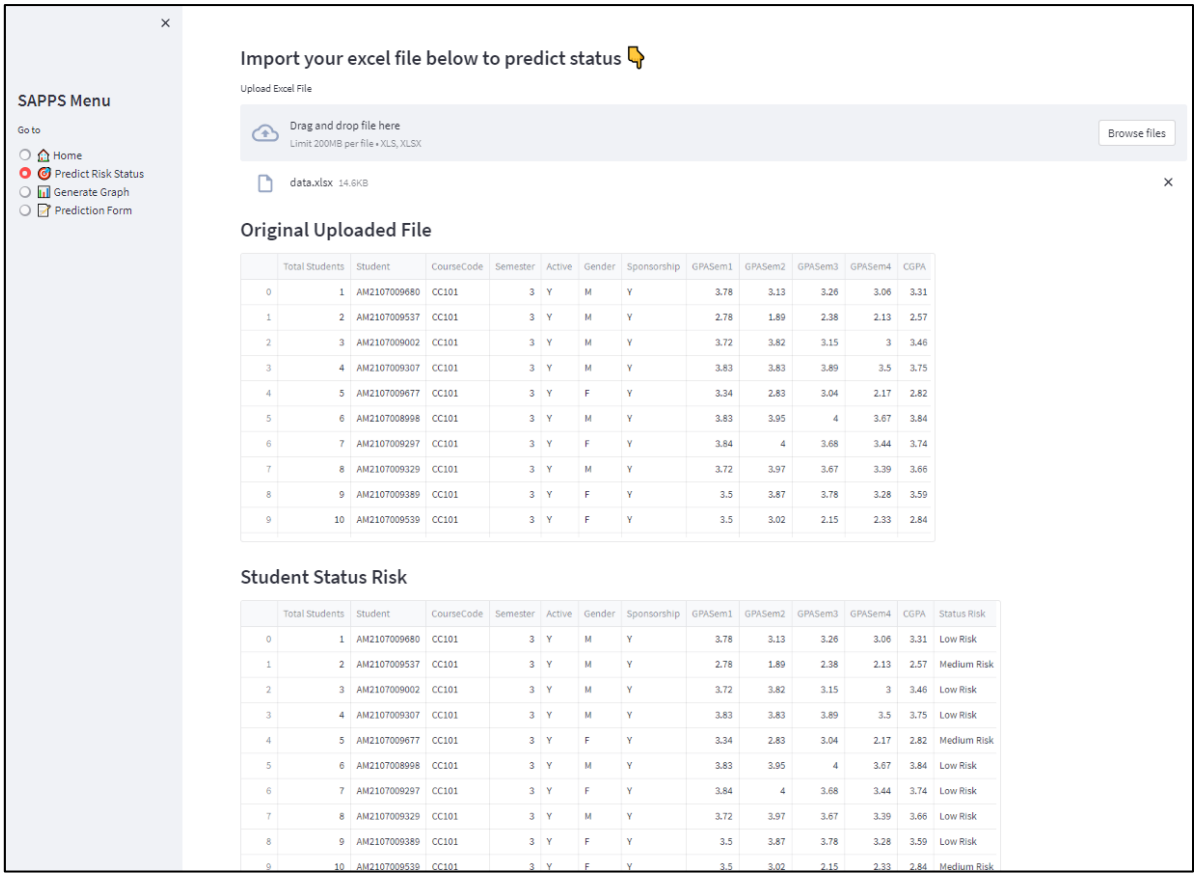
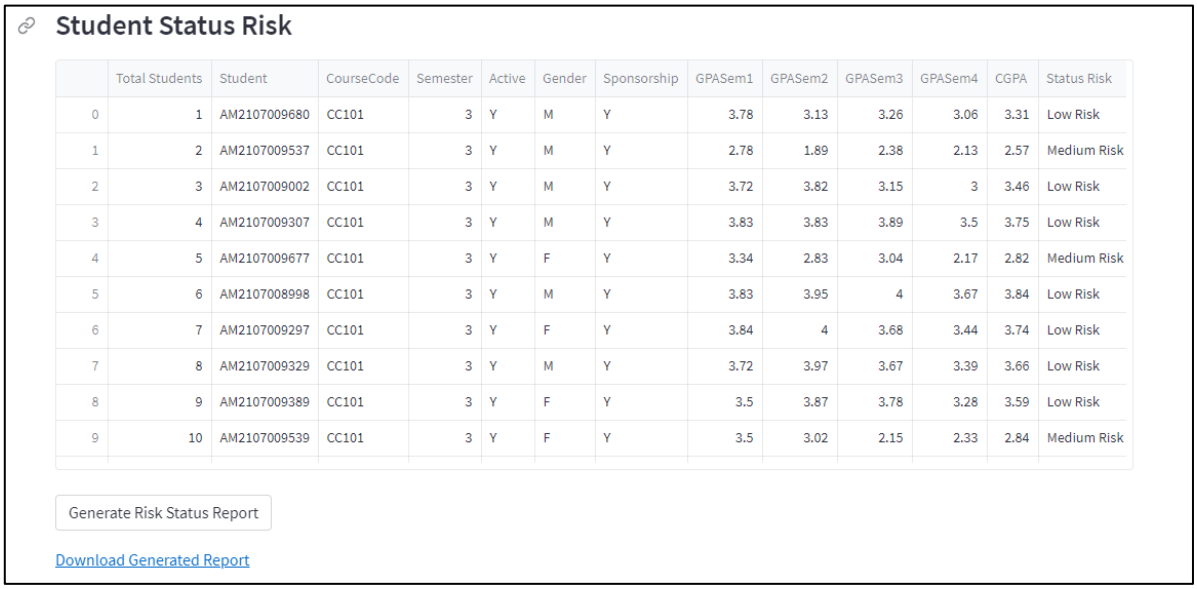


Fig. 10. Menu 1.2 - upload excel file

Figure 11 below shows the data of students with high, low, or medium-risk statuses. This output also includes a generate risk status report button that provides a link for the user to download the generated report, which will be used in the following menu (Figure 8) to display the graph.



	Total Students	Student	CourseCode	Semester	Active	Gender	Sponsorship	GPASem1	GPASem2	GPASem3	GPASem4	CGPA	Status Risk
0	1	AM2107009680	CC101	3	Y	M	Y	3.78	3.13	3.26	3.06	3.31	Low Risk
1	2	AM2107009537	CC101	3	Y	M	Y	2.78	1.89	2.38	2.13	2.57	Medium Risk
2	3	AM2107009002	CC101	3	Y	M	Y	3.72	3.82	3.15	3	3.46	Low Risk
3	4	AM2107009307	CC101	3	Y	M	Y	3.83	3.83	3.89	3.5	3.75	Low Risk
4	5	AM2107009677	CC101	3	Y	F	Y	3.34	2.83	3.04	2.17	2.82	Medium Risk
5	6	AM2107008998	CC101	3	Y	M	Y	3.83	3.95	4	3.67	3.84	Low Risk
6	7	AM2107009297	CC101	3	Y	F	Y	3.84	4	3.68	3.44	3.74	Low Risk
7	8	AM2107009329	CC101	3	Y	M	Y	3.72	3.97	3.67	3.39	3.66	Low Risk
8	9	AM2107009389	CC101	3	Y	F	Y	3.5	3.87	3.78	3.28	3.59	Low Risk
9	10	AM2107009539	CC101	3	Y	F	Y	3.5	3.02	2.15	2.33	2.84	Medium Risk

Fig. 11 Menu 1.3 – generate risk status report

Figure 12 below shows the second main menu, a generated graph. This menu consists of instructions on generating the graph by uploading an Excel file that the user downloads in menu 1.3 (Figure 11) first.

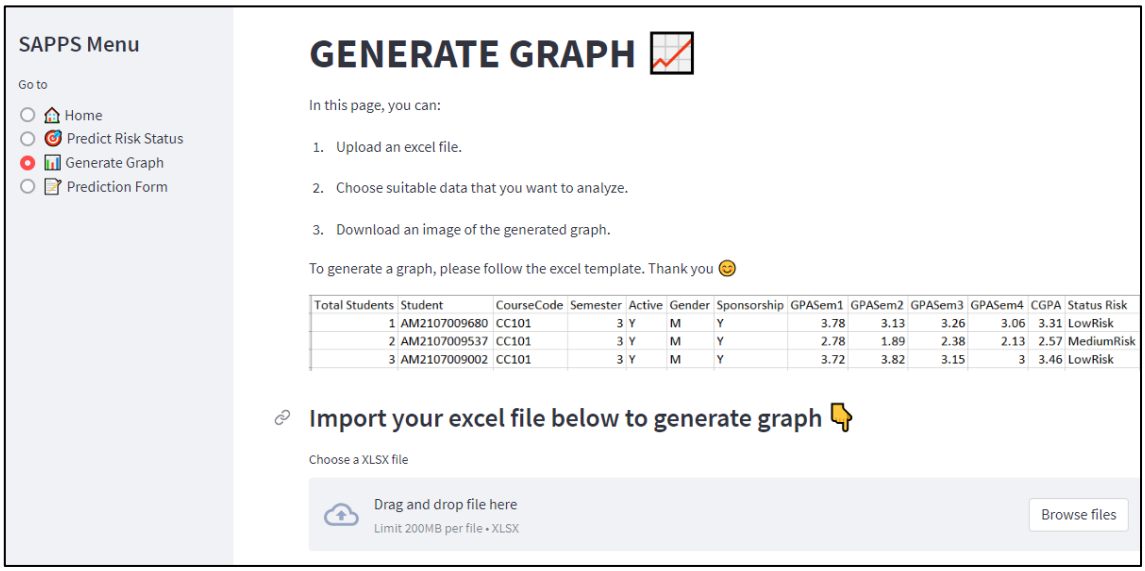


Fig. 12. Menu 2- generate graph

The list of students' information will be displayed in a table, and at the same time, users can choose suitable data based on their preferences, as shown in Figure 13 below. Students can choose to display the graph based on all attributes or specify the specific attribute only.

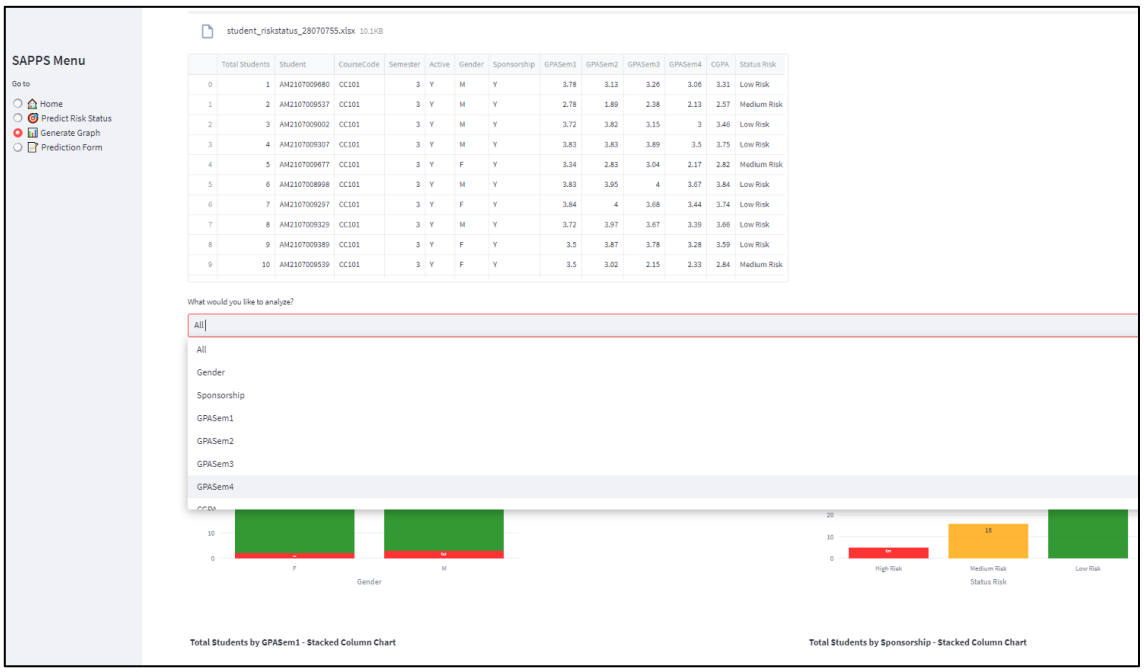


Fig. 13 Menu 2.1- graph analysis preferences

The sample output of the graph can be referred to in Figure 14 and Figure 15 on the next page.

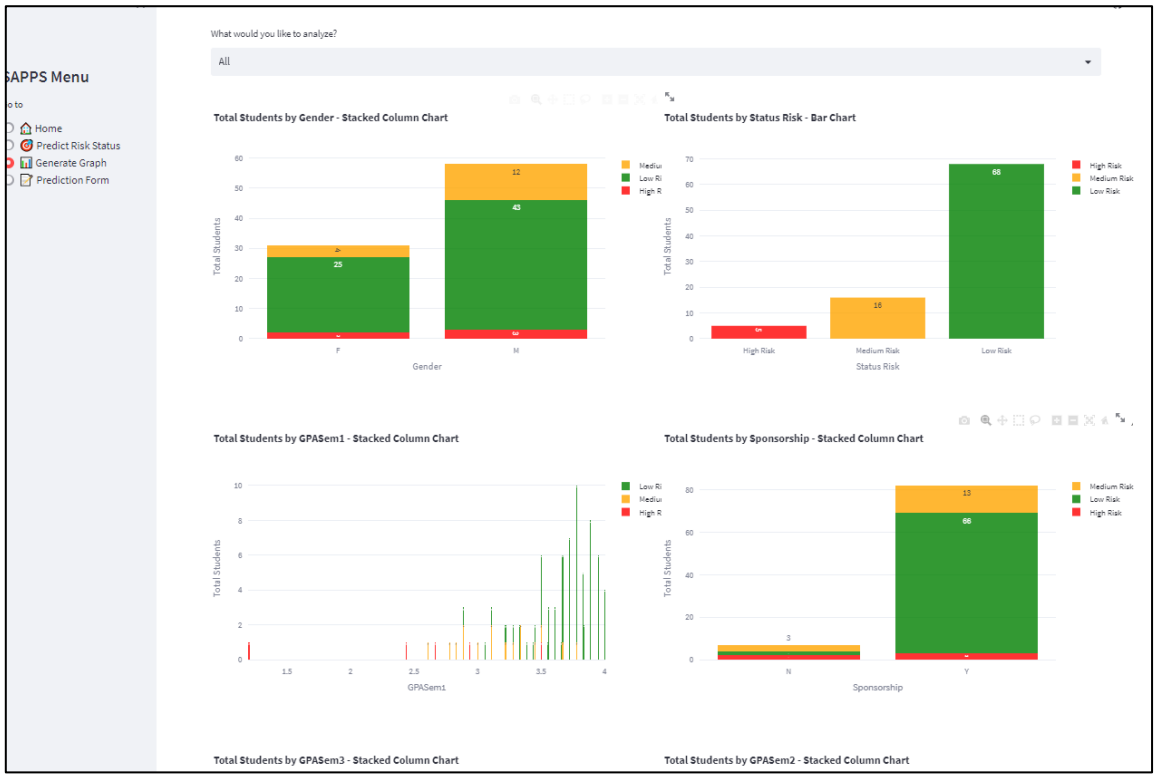


Fig. 14. Graph dashboard for all attributes

Fig. 15. Graph based on total students by gender

For Figure 16, the prediction form consists of all the data that the user needs to fill out first and will be displayed so that the user can predict the risk status for each student individually. At the same time, the system also provides different mitigations for each risk status, whether high, medium, or low. The user can also provide their email so that the report can be sent to the user.

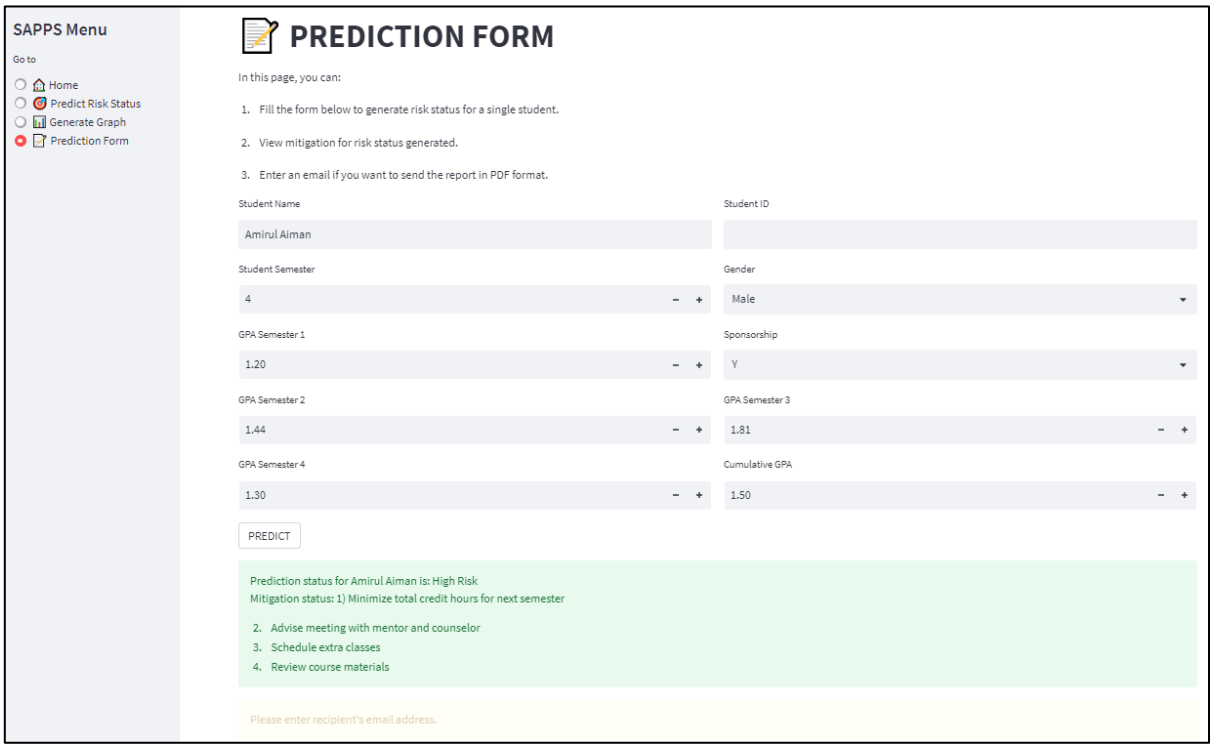


Fig. 16 Prediction form

4.2 Comparison with Existing System

Table 7 compares three existing student academic performance prediction systems to SAPPS based on five features: attributes, prediction class, system features, algorithm or model used, Prediction's Mean Absolute Error and programming language. Compared to the SAPPS, the other three existing systems have similar system features, which can train data and predict risk status. Each system was trained using different data sets and attributes, which led to different prediction results.

From this analysis, it becomes clear that the development of student academic performance prediction systems is not limited to a single approach. The potential for innovation is vast, with various algorithms and prediction classes being viable options depending on the data sets or attributes collected. This should inspire and motivate researchers and educators alike.

Table 7
Student Academic Performance Prediction Existing System

System Title / Features	Predict Student Performance by Iman Abdulrahman bin Faisal University (Aboaneen *et al.,*) [39]	Academic Performance Prediction based on Multisource, Multifeatured behavioural data by JP Infotech (Zhao, L. *et al.,* 2020) [40]	Decision Support System for Predicting Students' Performance by Technical Educational Institute of Peloponnese (Livieris *et al.,*) [41]	Student Academic Performance Prediction System (SAPPS) by UPTM
Attributes	Course, course level, midterm exam score, lab score, city, marital status, children, health	Sex, age, mother's job, father's job, reason, guardian, study time, failures, school support, family support, paid classes, activities, internet, free time, go out, health, absences, period 1 score, period 2 score, final score	Student personal data, 1st semester's grade, 2nd semester's grade	Gender, Sponsorship, GPASem1, GPASem2, GPASem3, GPASem4, CGPA
Prediction Class	Total score for each course out of 100	Final grade (poor, fair, good)	Grade academic semester (fail, good, very good, excellent)	Risks (low, high, medium)
System Features	Train data, predict total score status, generate report	Train data, predict final grade status, generate graph	Train data, predict grade academic semester status	Predict risk status, generate graph, generate report and mitigation
Algorithm/Model Used	Random Forest	Random Forest	Simple voting methodology using RIPPER, 3NN, BP and SMO	RepTree Decision Tree
Prediction's Mean Absolute Error	4.80%	Not stated	Not stated	3.08%
Programming Language	Python	Python	Java	Python

5. Conclusion

The present study applied the principles of data mining to the domain of classification. Applying Classification Algorithms like Decision Tree, Naïve Bayes and Support Vector Machine is useful in predicting student performance. The research employed RepTree, k-NN and Naïve Bayes as the prediction models, with k-NN registering a prediction accuracy of 88.76% and Naïve Bayes displaying an accuracy of 94.38%. The findings indicate that RepTree has exhibited the highest degree of accuracy, surpassing the techniques of k-NN and Naïve Bayes. Thus, it can be inferred that the RepTree algorithm is the most suitable prediction model, which will aid in developing a system for monitoring students' academic performance at UPTM. The system employed an artificial intelligence approach that ensured a good model of the system for users, particularly those belonging to the educational institution. This prediction model shall aid parents, teachers, or lecturers in tracking students' performance and provide the necessary guidance and counselling. The study also provides a detailed analysis that can be utilised to provide scholarships and other essential training programs for the students. Therefore, the educational institution shall benefit from the proposed system implementation, as it shall ensure the organisation's success by facilitating the monitoring of the student's performance and predictive models.

In subsequent research, additional attributes, such as the student's attendance rate per semester, familial background, student enthusiasm, and involvement in extracurricular activities, could be incorporated and assessed through various algorithms. Given that the data collection in question is focused solely on the Diploma in Computer Science program at UPTM, divergent outcomes would likely emerge if a diverse array of data were included, encompassing students pursuing degrees in various disciplines. This is because dissimilar levels of academic pursuits may yield distinct patterns. Besides, additional features can be added when comparing existing systems in terms of scalability and ease of use of the system in the future.

Acknowledgement
The authors thank Prof. Dr. Ramlan Mahmod for motivating the writing of this paper and the research, and also to the University Poly-Tech Malaysia for funding the research.

References

[1] Norris, Donald, Linda Baer, Joan Leonard, Louis Pugliese, and Paul Lefrere. "Action analytics: Measuring and improving performance that matters in higher education." *EDUCAUSE review* 43, no. 1 (2008): 42.
[2] Khan, Ijaz, Abdul Rahim Ahmad, Nafaa Jabeur, and Mohammed Najah Mahdi. "An artificial intelligence approach to monitor student performance and devise preventive measures." *Smart Learning Environments* 8 (2021): 1-18. https://doi.org/10.1186/s40561-021-00161-y
[3] Dixit, Prashant, Harish Nagar, and Sarvottam Dixit. "Decision Support System Model for Student Performance Detection using Machine Learning." *vol* 10: 25-31.
[4] Ali, Abdalla M., J. Joshua Thomas, and Gomesh Nair. "Academic and uncertainty attributes in predicting student performance." In *Intelligent Computing and Optimization: Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020)*, pp. 838-847. Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-68154-8_72
[5] Namoun, Abdallah, and Abdullah Alshanqiti. "Predicting student performance using data mining and learning analytics techniques: A systematic literature review." *Applied Sciences* 11, no. 1 (2020): 237. https://doi.org/10.3390/app11010237
[6] Xu, Jie, Yuli Han, Daniel Marcu, and Mihaela Van Der Schaar. "Progressive prediction of student performance in college programs." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1. 2017. https://doi.org/10.1609/aaai.v31i1.10713
[7] Alraddadi, Shouq, Shurug Alseady, and Sultan Almotiri. "Prediction of students academic performance utilizing hybrid teaching-learning based feature selection and machine learning models." In *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, pp. 1-6. IEEE, 2021. https://doi.org/10.1109/WiDSTaif52235.2021.9430248
[8] Jamil, Jastini Mohd, Nurul Farahin Mohd Pauzi, and Izwan Nizal Mohd Shahara Nee. "An analysis on student academic performance by using decision tree models." *The Journal of Social Sciences Research* (2018): 615-620. https://doi.org/10.32861/jssr.spi6.615.620
[9] Robinson, Samantha E., and Joon Jin Song. "Student academic performance system: quantitative approaches to evaluating and monitoring student progress." *International Journal of Quantitative Research in Education* 4, no. 4 (2019): 332-353. https://doi.org/10.1504/IJQRE.2019.100170
[10] Deepika, K., and N. Sathvanaravana. "Analyze and predicting the student academic performance using data mining tools." In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 76-81. IEEE, 2018. https://doi.org/10.1109/ICCONS.2018.8663197
[11] Asiah, Mat, Khidzir Nik Zulkarnaen, Deris Safaai, Mat Yaacob Nik Nurul Hafzan, Mohamad Mohd Saberi, and Safaai Siti Syuhaida. "A review on predictive modeling technique for student academic performance monitoring." In *MATEC Web of Conferences*, vol. 255, p. 03004. EDP Sciences, 2019. https://doi.org/10.1051/matecconf/201925503004
[12] Kausar, Samina, Solomon Oyelere, Yass Salal, Sadiq Hussain, Mehmet Cifci, Slavoljub Hilcenko, Muhammad Iqbal, Zhu Wenhao, and Xu Huahu. "Mining smart learning analytics data using ensemble classifiers." *International Journal of Emerging Technologies in Learning (iJET)* 15, no. 12 (2020): 81-102. https://doi.org/10.3991/ijet.v15i12.13455
[13] Chen, Huanyi, and Paul AS Ward. "Predicting student performance using data from an auto-grading system." *arXiv preprint arXiv:2102.01270* (2021).

[14] Hamsa, Hashmia, Simi Indiradevi, and Jubilant J. Kizhakkethottam. "Student academic performance prediction model using decision tree and fuzzy genetic algorithm." *Procedia Technology* 25 (2016): 326-332. https://doi.org/10.1016/j.protcy.2016.08.114
[15] Iatrellis, Omiros, Ilias K. Savvas, Panos Fitsilis, and Vassilis C. Gerogiannis. "A two-phase machine learning approach for predicting student outcomes." *Education and Information Technologies* 26 (2021): 69-88. https://doi.org/10.1007/s10639-020-10260-x
[16] Yaacob, Wan Fairos Wan, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, and Norafefah Mohd Sobri. "Supervised data mining approach for predicting student performance." *Indones. J. Electr. Eng. Comput. Sci* 16, no. 3 (2019): 1584-1592. https://doi.org/10.11591/ijeecs.v16.i3.pp1584-1592
[17] Venkatachalapathy, K. "A Comparison of Classification Techniques on Prediction of Student Performance in Educational Data Mining."
[18] Razak, Rohaila Abdul, Mazni Omar, Mazida Ahmad, and P. Mara. "A student performance prediction model using data mining technique." *International Journal of Engineering & Technology* 7, no. 2.15 (2018): 61-63. https://doi.org/10.14419/ijet.v7i2.15.11214
[19] Buniyamin, Norlida, Usamah bin Mat, and Pauziah Mohd Arshad. "Educational data mining for prediction and classification of engineering students achievement." In *2015 IEEE 7th International Conference on Engineering Education (ICEED)*, pp. 49-53. IEEE, 2015. https://doi.org/10.1109/ICEED.2015.7451491
[20] Arsad, Pauziah Mohd, and Norlida Buniyamin. "A neural network students' performance prediction model (NNSPPM)." In *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, pp. 1-5. IEEE, 2013. https://doi.org/10.1109/ICSIMA.2013.6717966
[21] Jishan, Syed Tanveer, Raisul Islam Rashu, Naheena Haque, and Rashedur M. Rahman. "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique." *Decision Analytics* 2 (2015): 1-25. https://doi.org/10.1186/s40165-014-0010-2
[22] Mayilvaganan, M., and D. Kalpanadevi. "Comparison of classification techniques for predicting the performance of students academic environment." In *2014 International Conference on Communication and Network Technologies*, pp. 113-118. IEEE, 2014. https://doi.org/10.1109/CNT.2014.7062736
[23] Natek, Srečko, and Moti Zwilling. "Student data mining solution–knowledge management system related to higher education institutions." *Expert systems with applications* 41, no. 14 (2014): 6400-6407. https://doi.org/10.1016/j.eswa.2014.04.024
[24] Sharma, Mamta, and Monali Mavani. "Accuracy comparison of predictive algorithms of data mining: Application in education sector." In *International Conference on Advances in Computing, Communication and Control*, pp. 189-194. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. https://doi.org/10.1007/978-3-642-18440-6_23
[25] Al Shibli, Kdhaiya Sulaiman, Amal Sulaiman Sayed Al Abri, Linitha Sunny, Nandakishore Ishwar, and Sherimon Puliprathu Cherian. "Model for Prediction of Student Grades using Data Mining Algorithms." *European Journal of Information Technologies and Computer Science* 2, no. 2 (2022): 1-6. https://doi.org/10.24018/compute.2022.2.2.47
[26] Osmanbegovic, Edin, and Mirza Suljic. "Data mining approach for predicting student performance." *Economic Review: Journal of Economics and Business* 10, no. 1 (2012): 3-12.
[27] Quadri, M. M., and N. V. Kalyankar. "Drop out feature of student data for academic performance using decision tree techniques." *Global Journal of Computer Science and Technology* 10, no. 2 (2010): 2-5. [28] C. Romero, S. Ventura, P. G. Espejo, C. Herv´as, Data mining algorithms to classify students, in: Educational Data Mining 2008
[28] Romero, Cristóbal, Sebastián Ventura, Pedro G. Espejo, and César Hervás. "Data mining algorithms to classify students." In *Educational data mining 2008*. 2008.
[29] Elakia, Gayathri, and Naren J. Aarthi. "Application of data mining in educational database for predicting behavioural patterns of the students." *Elakia et al.,/(IJCSIT) International Journal of Computer Science and Information Technologies* 5, no. 3 (2014): 4649-4652.
[30] Lynn, N. D., and A. W. R. Emanuel. "Using data mining techniques to predict students' performance. a review." In *IOP Conference series: materials science and engineering*, vol. 1096, no. 1, p. 012083. IOP Publishing, 2021. https://doi.org/10.1088/1757-899X/1096/1/012083
[31] Viet, Tran Ngoc, Hoang Le Minh, Le Cong Hieu, and Tong Hung Anh. "The Naïve Bayes algorithm for learning data analytics." *Indian Journal of Computer Science and Engineering* 12, no. 4 (2021): 1038-1043. https://doi.org/10.21817/indjcse/2021/v12i4/211204191
[32] Khoshgoftaar, Taghi M., Moiz Golawala, and Jason Van Hulse. "An empirical study of learning from imbalanced data using random forest." In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, vol. 2, pp. 310-317. IEEE, 2007. https://doi.org/10.1109/ICTAI.2007.46

[33] Vijayarani, S., and M. Muthulakshmi. "Comparative analysis of bayes and lazy classification algorithms." *International Journal of Advanced Research in Computer and Communication Engineering* 2, no. 8 (2013): 3118-3124.
[34] Gray, Geraldine, Colm McGuinness, and Philip Owende. "An application of classification models to predict learner progression in tertiary education." In *2014 IEEE international advance computing conference (IACC)*, pp. 549-554. IEEE, 2014. https://doi.org/10.1109/IAdCC.2014.6779384
[35] Arsad, P. M., N. Buniyamin, and J. A. Manan. "Profiling the performance of electrical engineering bachelor degree students based on different entry levels." *International Journal of Education and Information Technologies* 5, no. 2 (2011): 267-274.
[36] Janan, Farhatul, and Sourav Kumar Ghosh. "Prediction of student's performance using support vector machine classifier." In *Proc. Int. Conf. Ind. Eng. Oper. Manag*, vol. 11, no. 1, pp. 7078-7088. 2021. https://doi.org/10.46254/AN11.20211237
[37] Oloruntoba, S. A., and J. L. Akinode. "Student academic performance prediction using support vector machine." *International Journal of Engineering Sciences and Research Technology* 6, no. 12 (2017): 588-597.
[38] Franklin, James. "The elements of statistical learning: data mining, inference and prediction." *The Mathematical Intelligencer* 27, no. 2 (2005): 83-85. https://doi.org/10.1007/BF02985802
[39] Alboaneen, Dabiah, Modhe Almelihi, Rawan Alsubaie, Raneem Alghamdi, Lama Alshehri, and Renad Alharthi. "Development of a web-based prediction system for students' academic performance." *Data* 7, no. 2 (2022): 21. https://doi.org/10.3390/data7020021
[40] Zhao, Liang, Kun Chen, Jie Song, Xiaoliang Zhu, Jianwen Sun, Brian Caulfield, and Brian Mac Namee. "Academic performance prediction based on multisource, multifeature behavioral data." *IEEE Access* 9 (2020): 5453-5465. https://doi.org/10.1109/ACCESS.2020.3002791
[41] Livieris, Ioannis, Tassos Mikropoulos, and Panagiotis Pintelas. "A decision support system for predicting students' performance." *Themes in Science and Technology Education* 9, no. 1 (2016): 43-57.
[42] Chi, Cai, Melor Md Yunus, Karmila Rafiqah M. Rafiq, Hamidah Hameed, and Ediyanto Ediyanto. "A Systematic Review on Multidisciplinary Technological Approaches in Higher Education." *International Journal of Advanced Research in Future Ready Learning and Education* 36, no. 1 (2024): 1-10. https://doi.org/10.37934/frle.36.1.110
[43] Hishamuddin, Fatimah, Khalidah Ahmad, Halina Kasmani, Nur Bahiyah Abdul Wahab, Mohd Zulfahmi Bahaudin, and Elme Alias. "Empowering Leaders: A Work in Progress on Promoting Leadership Roles in Online Learning through Project-Based Learning (PBL)." *Semarak International Journal of Innovation in Learning and Education* 2, no. 1 (2024): 65-73. https://doi.org/10.37934/sijile.2.1.6573
[44] Sidhu, Pramita, Fazlin Shasha Abdullah, and Mohamad Sirajuddin Jalil. "Awareness and Readiness of Malaysian Generation Z Students towards the Fourth Industrial Revolution (IR4. 0)." *Semarak International Journal of STEM Education* 1, no. 1 (2024): 20-27. https://doi.org/10.37934/sijste.1.1.2027