UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Interpreting Student Performance through Predictive Learning Analytics

Fei Zhi Tan[1] & Weng Howe Chan[2]*
Faculty of Computing,
Universiti Teknologi Malaysia
81310, UTM Johor Bahru, Johor, Malaysia
Email: fztan2@graduate.utm.my[1]; cwenghowe@utm.my[2]

Submitted: 21/10/2023. Revised edition: 2/6/2024. Accepted: 29/7/2024. Published online: 25/11/2024
DOI: https://doi.org/10.11113/ijic.v14n2.434

***Abstract*—In today's information-rich world, accurately predicting student performance is crucial for institutions seeking to support at-risk students and ensure their success, but this task can be challenging. Learning analytics (LA) can help identify students who are struggling and provide them with the tools and opportunities they need to succeed, benefiting both students and institutions. However, data integration from various sources can be challenging in learning analytics, causing educators to struggle with managing and keeping track of students' progress and dropouts. The goal of this project is to generate insights into student performance through the application of machine learning methods, including Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Machine (SVM). These methods were used to predict students' future results and the likelihood of students' dropout based on predictive learning analytics. RF, ANN, and SVM predictive models were constructed to predict students' future final results and dropout. The dataset from The Open University is used in this study, which consists of information from multiple aspects including data about courses, student registration, results, and their interactions with virtual learning environment. RF, ANN, and SVM models were constructed to predict students' future final results based on their learning behaviour. The performance of the models is evaluated based on accuracy, precision, recall, and time taken for training. In this study, the RF model demonstrated the best performance among the three predictive models for predicting final results and dropout with the shortest training time and achieved the highest accuracy. The RF model achieved an accuracy of 87.8% in predicting final results and 82.3% in predicting dropout while maintaining an average training time of 3.6 seconds. At the end of the study, the dashboards visually presented the results, offering valuable insights into students' learning outcomes. This enables educators to effectively support their students by utilizing predictive analytics, which includes identifying potential dropouts and tailoring assistance based on these predictions.**

***Keywords*—Learning Analytics (LA), data integration, machine learning, Virtual Learning Environment (VLE), predictive models**

I. Introduction

Accurately predicting student performance is crucial but challenging for institutions seeking to support at-risk students and ensure their success. Learning Analytics (LA) can help identify students who are struggling and provide them with the tools and opportunities they need to succeed. Recent studies have shown that students experience increased stress due to online learning and express a preference for face-to-face instruction [1], [2]. Consequently, this has had a detrimental impact on students' academic performance, resulting in a decline in their grades. Hence, LA benefits both students and institutions. Although learning analytics is a relatively new field, LA has quickly become advanced, particularly in LA applications for higher education [3]. Generally, learning analytics involves measuring, collecting, analysing, and reporting learners' data in the learning context [4]. A rising number of LA models have been created for application in a variety of industries, including education. These models include the academic resistance model [5], the online teaching-learning model [6], and the technology acceptance model. Through a variety of instrumental research, the LA model is being used to enhance learning and obtain a better understanding of how students learn [7]. The utilization of LA has simplified the process of making well-informed decisions regarding students' performance.

In the literature, there are many ways to define student performance. Student performance is defined as the outcome of measuring learning assessment and co-curriculum activities. However, many studies have focused on graduation as a measure of student success. In Malaysia, many higher learning institutions evaluate student performance using final results, which are determined by a combination of factors such as course structure, assessment marks, final exam scores, and extracurricular activities. Assessing student achievements is essential for maintaining the quality of the education. By analyzing student performance, institutions can develop strategic programs that help students succeed during their studies. Utilizing artificial intelligence (AI) and data from multiple public and private databases and applications, Malaysia has invested a lot in predictive data analytics associated with student movement and job prospects [8]. Prediction based on learning analytics has become a trend in the education sector.

Machine learning (ML) has been widely used in predicting student performance in education [9] such as the tendency of failure and dropout from studies. Meaningful insights from massive student academic datasets can be generated through the application of machine learning. Many different machine learning approaches have been utilized in previous research work such as Naïve Bayes (NB), and Artificial Neural Networks (ANN) to predict the student's performance based on their learning behaviour. In this project, there are three ML techniques including Random Forest (RF), Support Vector Machine (SVM), and ANN were applied to predict students' future academic outcomes. This research will provide valuable insights, knowledge, and intelligence for policymakers for educators to assist the students and for students to improve themselves based on their learning behaviours.

In recent years, learning platforms have not only focused on in-person classes but also online platforms such as Moodle, Virtual Learning Environment (VLE), and Learning Management System (LMS) [10]. Data integration solutions should be able to combine data from several sources, which may be in various forms and have variable degrees of organization, to obtain deeper insights into LA and education. The scalability of LA solutions, which is a big difficulty, depends on data integration as well [11]. Integrating and working with data from various sources can be challenging. Data integration from various sources and formats is a challenging task to provide effective insights for LA. This study involved the integration of data from VLE to make the study more meaningful.

Moreover, student dropout is a great concern [12] and an extraordinarily challenging issue. Student dropout refers to the phenomenon of students leaving or dropping out of school, college, or university. This can be due to a variety of reasons, including academic struggles, pressure from peers, financial difficulties, work commitments, or other personal factors [13]. Students dropout is usually linked to crucial factors, including academic progress [14]. Finding the cause of a dropout is challenging. Research on the variables affecting students' success in different courses in Malaysia is lacking [15].

Moreover, the COVID-19 pandemic has led to a substantial rise in the demand for online learning platforms [16]. The extensive embrace of online education has resulted in a notable surge in the quantity of data generated by the online activities of students worldwide [17]. An educator can tacitly identify a student's learning progress and performance in a face-to-face classroom. However, the conversion from physical classes to online classes was a big challenge for educators. The data generated in online learning platforms such as Blackboard and Moodle are complex, large, and heterogeneous. The data might be difficult for the lecturer to interpret meaningfully. In this case, learning analytics is important in analyzing the data of the learners including tracking their learning progress and predicting their performance in the future examination.

Furthermore, traditional approaches to education are no longer effective for engaging modern students [18]. Conventional methods of teaching often fail to hold students' attention and high enrolment rates might be difficult for schools to manage their students. At the same time, university students are in a class from different backgrounds. When students from diverse backgrounds are enrolled in the same course, this situation can create imbalanced learning progress, posing a challenge for educators in terms of providing support and tracking individual student progress. Some students may struggle to follow lectures and other course materials. Learning analytics can help educators monitor and evaluate students' progress and behaviour.

This study aims to predict students' performance, including their academic results and dropout probability, based on various factors such as demographics, courses, assessments, and VLE behaviour. Previous studies have often overlooked VLE behaviour in predicting student performance due to the dominance of traditional learning methods. However, with the increasing popularity of VLEs and online learning platforms, it has become essential to integrate these behaviours into the analysis. By incorporating students' VLE activities along with other relevant information, this study seeks to provide a more comprehensive dataset for model training. This integration is expected to enhance the accuracy of predictive models. Consequently, more precise predictions can be made regarding student performance and dropout risks.

As a result, accurate predictions enable educators to offer tailored support to students, addressing their specific needs and improving their chances of success. This personalized approach can significantly contribute to better educational outcomes and reduced dropout rates.

This study could inspire future researchers to delve deeper into various aspects of student behaviour, such as their involvement in co-curricular activities or other extracurricular engagements. By broadening the scope of research to include these additional behaviours, future studies can become more intriguing and comprehensive. Understanding the impact of a wider range of student activities on learning outcomes can provide educators with valuable insights into how different behaviours may enhance students' learning abilities. This expanded knowledge can lead to the development of more effective and holistic educational strategies, ultimately supporting student success in a more detailed and thorough manner.

II. METHODOLOGY AND DATASET

This research proposed the predictive model to predict student results and dropout. The research methodology for this study is displayed in Fig. 1.

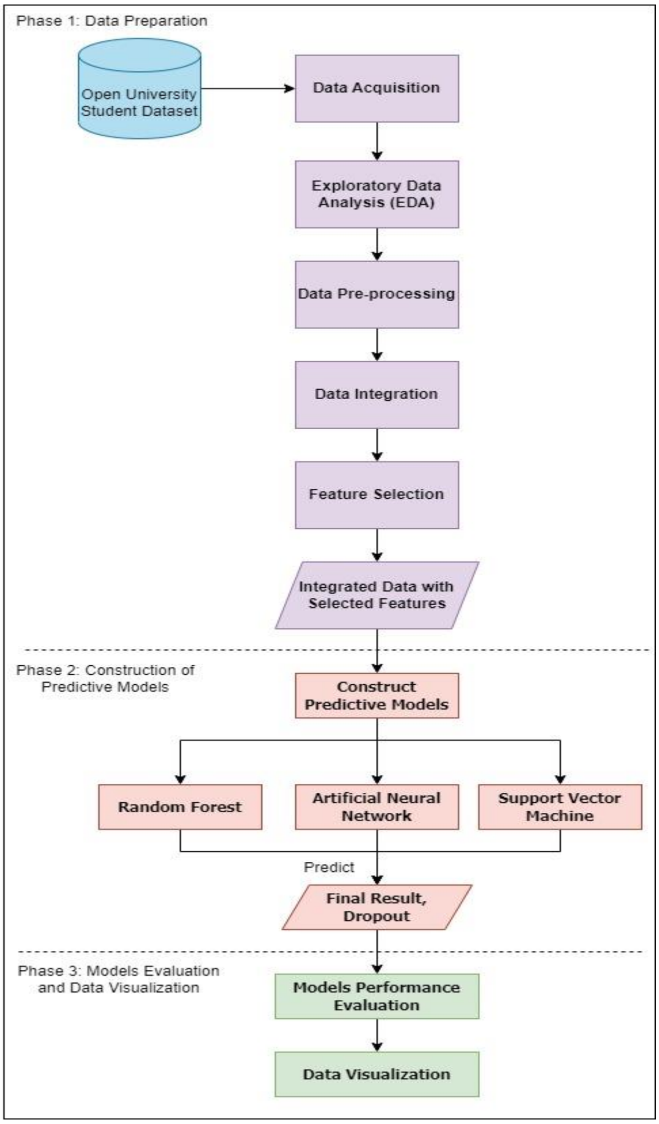


Fig. 1. Research methodology

A. Dataset

The dataset used in this project is the Open University Learning Analytics Dataset (OULAD) which was obtained from [19]. The dataset comprises 7 datasets with a total of 41 attributes. The 7 datasets are named assessments, *studentAssessment*, *vle*, *studentVle*, *courses*, *studentInfo*, and *studentRegistration*. These datasets consist of various attributes which can be meaningfully interpreted in this study to discover more about student's performance. These datasets were integrated into a final dataset.

B. Exploratory Data Analysis (EDA)

During the data exploration phase, analysis was carried out on the structure and specifics of the datasets which involved examining the shape of the data, data types, and the information contained. Additionally, thorough checks were performed to identify any null values present in each dataset. While the null values were found, appropriate preprocessing steps were taken to handle null columns and rows in the subsequent phase.

The *studentInfo* dataset consists of 4 semesters and 2 years of data from 2013 to 2014. The starting month of the semesters is February and October in both years. From data exploration, the dataset was found that the February semester is approximately 20 days shorter than the October semester.

The weight of assignments in some of the modules was found to be imbalanced. The total weight of exams was supposed to be 100%, and the total weight of other assessments (TMA, CMA) was supposed to be 100%. However, some modules have imbalanced weights. The imbalanced weight issue was addressed in the data pre-processing phase. Fig. 2 shows the imbalanced weights found during data exploration.

GGG	CMA	2013J	229.0	229.0	0.0	0.0
		2014B	222.0	222.0	0.0	0.0
		2014J	229.0	229.0	0.0	0.0
	Exam	2013J	229.0	229.0	100.0	100.0
		2014B	222.0	222.0	100.0	100.0
		2014J	229.0	229.0	100.0	100.0
	TMA	2013J	61.0	173.0	0.0	0.0
		2014B	61.0	166.0	0.0	0.0
		2014J	61.0	173.0	0.0	0.0

Fig. 2. Imbalanced assessments weight

Besides that, some graphs were plotted to visualize the relationships and overview of the datasets. For instance, VLE significantly influences student results by offering an array of resources and tools that support learning and facilitate communication between students and educators. Students employ VLE to access learning materials, submit coursework, and engage with fellow students and educators. We visualized the most accessed activity types by students based on the total number of clicks, as illustrated in Fig. 3.

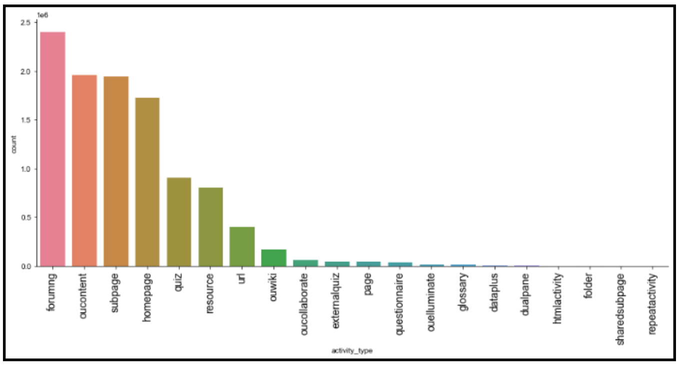


Fig. 3. The most accessed activity types by students in VLE

The distribution of students based on their final results is presented in Fig. 4. The pass grade had the highest percentage, accounting for 37.93% of the students, while the distinction grade had the lowest percentage at 9.28%. This indicates that the data is balanced, as the total number of students who passed and received a distinction is close to the number of students who failed or withdrew. Therefore, up or down sampling is not needed in this study.

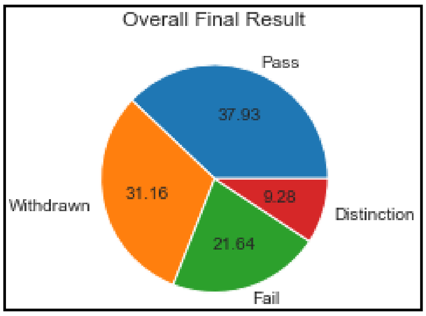


Fig. 4. Distribution of students based on final results

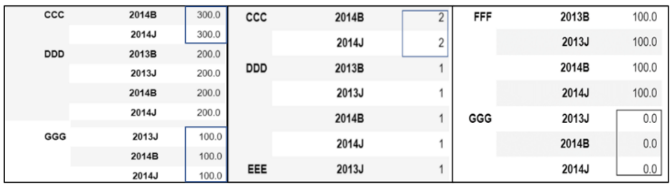
C. Feature Selection

In data pre-processing, a few steps were conducted to preprocess the data, such as converting the datatypes of the attributes, assigning the weight of imbalanced data, removing null columns, adding new columns, and replacing null values. Some datatypes of the attributes were incorrect. Hence, the attributes were converted into correct datatypes.

There was an imbalance in the assessment weight between modules CCC and GGG which was found in data exploration. The intended total weight for exams and other assessments is 200. However, module CCC weighs 300, while module GGG weighs 100. To rectify this issue, the assessment weights were adjusted to restore balance across the assessments.

Additionally, the number of exams was examined for each module. Module CCC had 2 exams, which contributed to the imbalance in the assessment weight. Conversely, the module GGG was found that did not include any non-exam type assessments. Fig. 5 illustrates the imbalanced number of exams and the discrepancy in assessment weight. Since the CMA tends to be 0 in module GGG, hence assign 100% to TMA to ensure that module GGG has balanced data. Next, the same method was applied for reassigning the weight for module CCC, ensuring that every module attains a balanced weight distribution.

Exam Type | **Number of Exams** | **Non-Exam Type**



CCC	2014B	300.0
	2014J	300.0
DDD	2013B	200.0
	2013J	200.0
	2014B	200.0
	2014J	200.0
GGG	2013J	100.0
	2014B	100.0
	2014J	100.0

CCC	2014B	2
	2014J	2
DDD	2013B	1
	2013J	1
	2014B	1
	2014J	1
EEE	2013J	1

FFF	2013B	100.0
	2013J	100.0
	2014B	100.0
	2014J	100.0
GGG	2013J	0.0
	2014B	0.0
	2014J	0.0

Fig. 5. Imbalanced number of exams and weight

There were 173 students' scores found missing in the StudentAssessment dataset, indicating that the students did not receive any results for their courses. As a result, the scores with null values had been dropped from the results data frame. In the studentInfo dataset, there were 1111 rows with missing values in the "imb_band" column. These null values were filled with NaN. In the registration dataset, some rows have null values in the "unregistration_date" column, as most of the students have completed their studies. Therefore, the null columns are filled with 0.

D. Data Integration

Data integration allows the combining of data from multiple sources which can provide a more comprehensive view of data. OULAD comprises 7 datasets and some of the datasets have mutual variables. Hence, some datasets were combined to provide a more complete overview of this study. By consolidating the datasets around mutual variables, "student_Info" serves as a robust foundation for subsequent feature selection and machine learning model training, ultimately enhancing the study's analytical capabilities. The integration of the dataset is presented in Table I.

TABLE I. DATA INTEGRATION

Dataset 1	Dataset 2	Integrated Dataset
studentVle	vle	Studentvle_merge_vle
Student-Registration	courses	reg_merge_courses
assessment	studentAssessment	results_merge_ass
studentInfo	studentvle_merge_vle	studentInfo
studentInfo	reg_merge_courses	student_Info

E. Feature Selection

Feature selection is crucial in developing a machine learning model, as the features chosen can significantly affect the model's performance. The goal of feature selection is to select features that improve the model's prediction part. Fig. 6 shows the overall flow of feature selection.

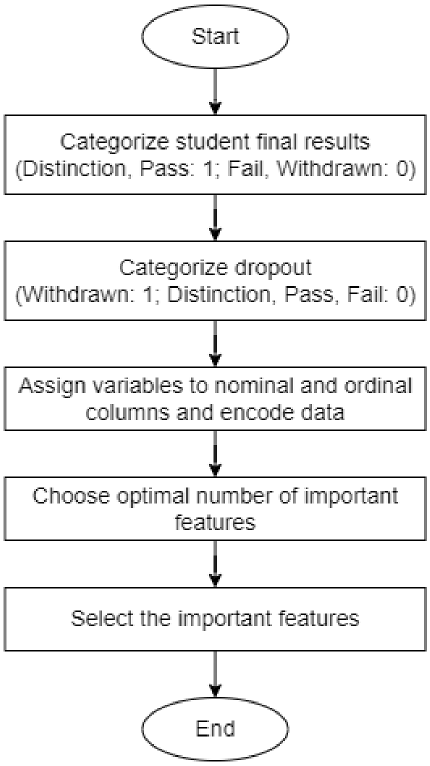


Fig. 6. Flowchart of feature selection

The feature selection method used in this study was Recursive Feature Elimination with Cross-Validation (RFECV). RFECV works by recursively eliminating less important features from a dataset while assessing the model's performance using cross-validation. The technique starts by training a model on the entire feature set and assigns importance scores to each feature. RFECV then eliminates the least important feature and repeats the process, iteratively reducing the feature set until a desired number of features is reached. This iterative process is guided by cross-validation, which aids in evaluating the model's potential performance on unseen data.

RFECV helps identify the most relevant features for a given problem, improving model performance by focusing on the most informative attributes. The inclusion of cross-validation within RFECV enhances the reliability in estimating model performance since it assesses the model on various data subsets.

The categorical data were assigned to nominal and ordinal columns. Nominal data is categorical data without any order, while ordinal data is categorical data with a meaningful order or ranking. After that, the attributes were encoded using a defined encoding function before the feature selection process was conducted.

The process of choosing the optimal number of features with RFECV involves iteratively eliminating less important features while evaluating the model's performance using cross-validation. The Random Forest Classifier was selected as the classifier to run RFECV due to the built-in feature importance measure in this study. By utilizing Random Forest as the classifier, RFECV can leverage the feature importance information provided by Random Forest to rank and eliminate features based on their importance scores. This simplifies the feature selection process. Additionally, RF, known for the robustness against overfitting and noise in the data, proves to be a suitable choice for this task. Random Forest can handle various data types, including numerical and categorical features, without requiring extensive pre-processing.

Following the feature selection procedure utilizing the RFECV technique, different sets of features were chosen for predicting final results and dropout. The features selected for predicting final results can be found in Table II.

TABLE II. SELECTED FEATURES FOR PREDICTING THE FINAL RESULTS

No	Features
1.	imd_band
2.	num_of_prev_attempts
3.	age_band
4.	final_result
5.	studied_credits
6.	sum_click
7.	After_Clicks
8.	Before_Clicks
9.	date_registration
10.	module_representation_length
11.	dropout
12.	highest_education_HE Qualification
13.	gender_F
14.	code_module_AAA
15.	code_module_EEE
16.	code_module_FFF
17.	id_student

The selected features for predicting the dropout were shown in Table III.

TABLE III. SELECTED FEATURES FOR PREDICTING THE DROPOUT

No	Features
1.	imd_band
2.	studied_credits
3.	final_result
4.	sum_click
5.	After_Clicks
6.	Before_Clicks
7.	date_registration
8.	module_representation_length
9.	id_student
10.	dropout

F. Model Training

The data with the selected features were employed in three different algorithms to assess the performance of the models based on this feature set. This approach ensures a fair evaluation of the model's performance. For this study, three methods were employed: RF, ANN, and SVM algorithms. These algorithms were used to construct predictive models capable of predicting students' final results and likelihood of dropout based on their learning behaviour. The construction of predictive models was conducted in several steps, as depicted in Fig. 7.

Since the final results and dropout of students were predicted in this study, the target for predicting the final results was assigned as "final_result," and the other selected features, excluding "final_result," were assigned as features. These features were labelled as "train," while the target was labelled as "test." Similarly, for predicting the dropout of students, the target was assigned as "dropout," and the other features, excluding "dropout," were assigned as features. Once again, the features were labelled as "train," while the target was labelled as "test."

In this research, the dataset was divided into three parts: training, validation, and testing sets. This data partitioning strategy is employed to evaluate the machine learning model's performance [20]. The reason for dividing the data into these sets is to accomplish different objectives within the model training process and is mainly used for hyperparameter optimization purposes [21]. In general, the training set is employed for model training, the validation set is utilized to fine-tune the hyperparameters, and the final model's performance is assessed using the test dataset. [22].

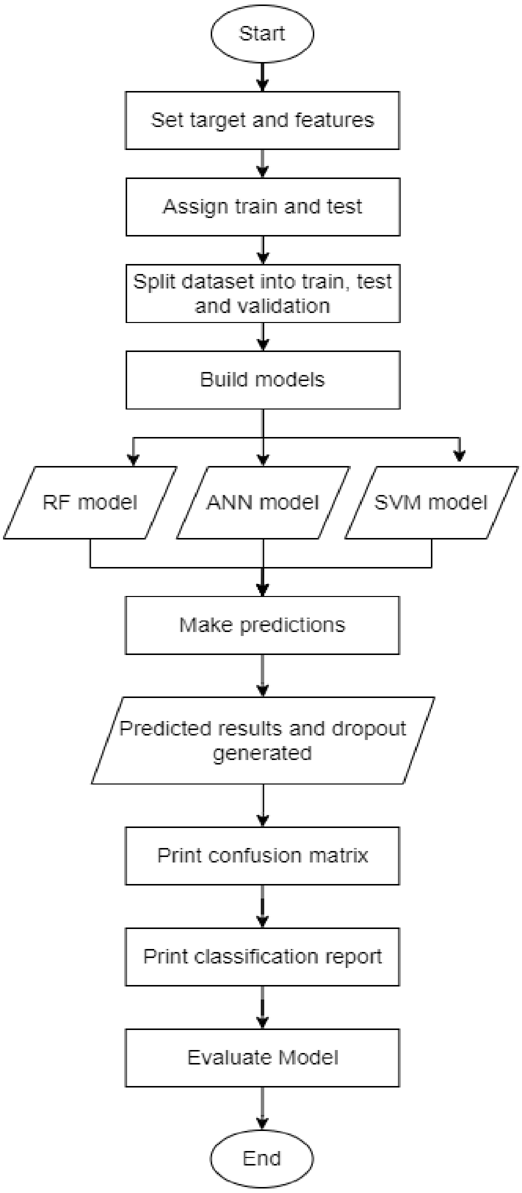


Fig. 7. Flowchart of model training

Train data is used for learning, which fits the parameters of the machine learning model [20]. The effectiveness of a machine learning model is assessed using test data while the effectiveness of the model during training is assessed using validation data.

The dataset is split into 60% for train data, 20% for test data, and 20% for validation data as shown in Fig. 8. The dataset consists of data from 32,593 students. The data is divided into three sets: 19,555 (60%) training data, 6,519 (20%) validation data, and 6,519 (20%) testing data.

For the model training process, the model was trained using 60% of the training data. Once the training was completed, a trained model was obtained. To validate the model, 20% of the validation data was used, and the performance of the model was accessed. If the model did not perform well, the parameters were adjusted, and the model was retrained. Otherwise, the testing process was initiated, using the remaining 20% of the test data to assess the model. Lastly, the model was run on the 20% test data, which represented data the model had never encountered before. In simple terms, the test set offers a fair assessment of the final model's performance whereas the validation set was utilized to refine the model parameters [23].

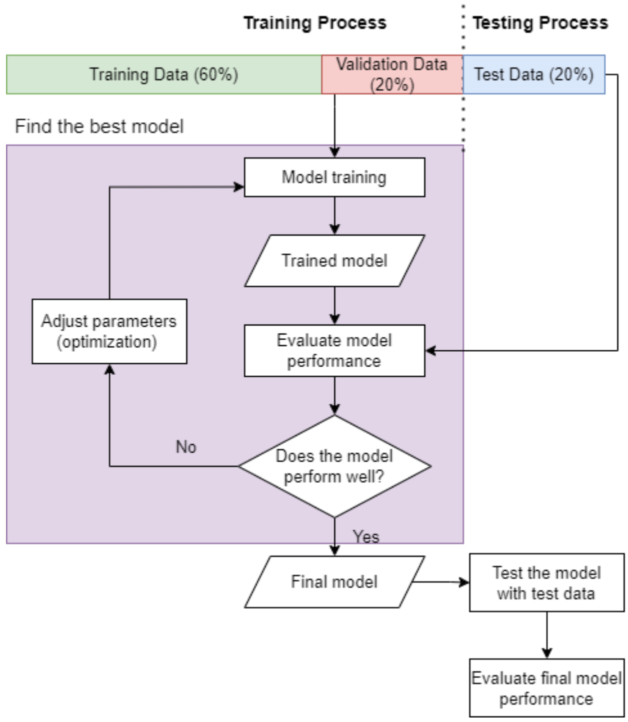


Fig. 8. Train validation data splitting

Three predictive models with RF, ANN, and SVM algorithms, have been developed based on algorithms to facilitate predictions on the given dataset. These models were trained using a dataset and can effectively forecast students' final results and the likelihood of dropout, utilizing their past learning behaviour and academic performance as key indicators. The same three predictive models were used for both final results prediction and dropout prediction, ensuring a fair and consistent evaluation process. The primary objective of these predictive models was to anticipate students' future final results and dropout rates, empowering educators to identify and comprehend which students, from which courses and semesters, are at risk of academic failure and likely to drop out. The models calculated the probability of dropout, allowing for the ranking of students most susceptible to this outcome. By leveraging the insights provided by these predictive models, educators can proactively intervene and provide targeted support to students who require additional assistance, thereby improving overall educational outcomes and minimizing dropout rates.

G. Performance Evaluation

The RF, ANN, and SVM models are evaluated based on a few metrics which are accuracy, precision, recall, and time taken for training.

Accuracy is a measure of how well a model or algorithm can predict the correct outcome or label for a given data point. The accuracy is calculated from the equation:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision is defined as the ratio of true positive predictions (correctly classified positive samples) to the total number of positive predictions made by the model (true positives and false positives). The precision is derived from:

$$precision = \frac{TP}{TP + FP} \quad (2)$$

Recall is a measure of how well a model is able to identify positive examples in a dataset. Recall is calculated by dividing the number of true positive predictions made by the model (correctly classified positive samples) by the total number of positive examples in the data. Recall is calculated based on the following equation.

$$recall = \frac{TP}{TP + FN} \quad (3)$$

H. Data Visualization

For the data visualization, the predictions for students' final results and dropout were visualized using Power BI Desktop, allowing for the creation of interactive dashboards. These dashboards covered three types of analytics: descriptive, diagnostic, and predictive.

A total of four dashboards have been created to showcase the predicted results obtained from the previous phase, which is the model training. The first dashboard is the Predictive Learning Analytics (PLA) Dashboard, which provides descriptive analysis by offering an overview of the results obtained from the prediction. The Dropout Analysis Dashboard focuses on the variables that impact dropout. The Final Results Analysis Dashboard displays the relationships between variables and predicted final results. Additionally, the Student Analysis Dashboard provides a detailed analysis at an individual student level, offering insights into each student's performance. This level of analysis assists educators in understanding their students better. The wireframes of four dashboards were visualized in Fig. 9, Fig. 10, Fig. 11, and Fig. 12.

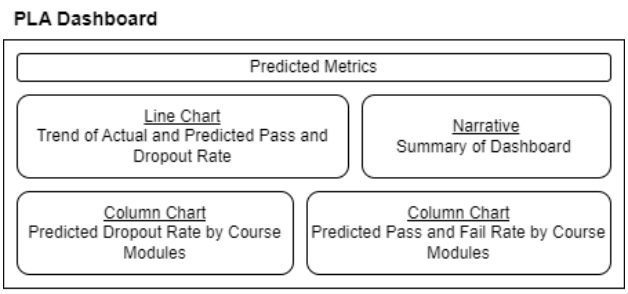


Fig. 9. Wireframe of PLA dashboard

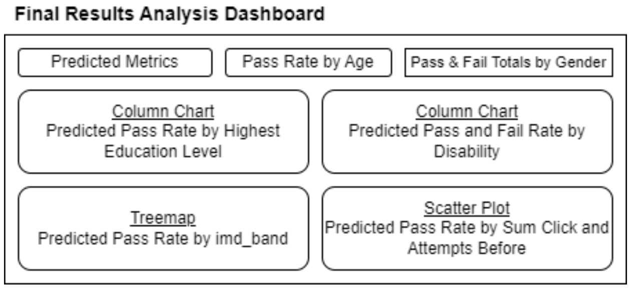


Fig. 10. Wireframe of Final Result Analysis Dashboard

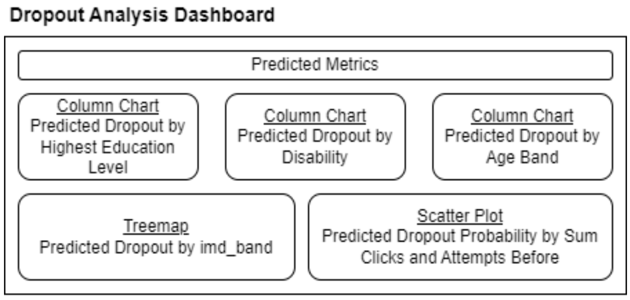


Fig. 11. Wireframe of Dropout Analysis Dashboard

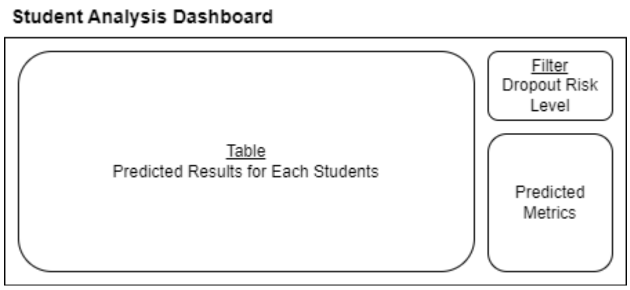


Fig. 12. Wireframe of Student Analysis Dashboard

III. RESULTS AND DISCUSSION

This section discusses the results gained from the research work.

A. Predictive Models Performance Evaluation

For evaluating the performance of three models, the confusion matrix serves as a useful tool for assessing the performance of a predictive model when predicting students' actual pass, actual fail, and the corresponding predicted pass and predicted fail outcomes, as well as predicting students' actual dropout, actual non-dropout, predicted dropout, and predicted non- dropout. The confusion matrix provides a condensed overview of the model's predictions and how they correspond to the actual labels.

Fig. 13 presents the confusion matrix of the models that predicted the students' results. Based on the prediction results, the RF model achieved a prediction accuracy of 5724 labels that matched the actual labels. Similarly, the ANN model correctly predicted 5628 labels, while the SVM model achieved 5432 correct predictions.

	Predicted Final Results		
Random Forest	2866 TN	602 FP	Actual Final Results
	193 FN	2858 TP	
Artificial Neural Network	2769 TN	699 FP	
	192 FN	2859 TP	
Support Vector Machine	2929 TN	539 FP	
	548 FN	2503 TP	

Fig. 13. Confusion matrix analysis based on final results prediction

Additionally, the predictions were conducted for dropouts as well. As a result, the RF model achieved a prediction accuracy of 5367 labels that matched the actual labels. Similarly, the ANN model correctly predicted 5087 labels, while the SVM model achieved 5122 correct predictions. Based on the performance, the RF model demonstrated the highest in predicting final results labels correctly compared to the other two models. However, in contrast to the final result prediction, the SVM model predicted more correct labels compared to the ANN model. These prediction performances for final results are summarized in Fig. 14.

	Predicted Dropout		
Random Forest	3916 TN	630 FP	Actual Dropout
	522 FN	1451 TP	
Artificial Neural Network	3167 TN	1379 FP	
	53 FN	1920 TP	
Support Vector Machine	3772 TN	774 FP	
	623 FN	1350 TP	

Fig. 14. Confusion matrix analysis based on dropout prediction

Apart from the confusion matrix, the predictive models were evaluated based on crucial metrics including accuracy, precision, recall, and the time required for training. The higher the accuracy, the more correct predictions the model made compared to the actual results. Interestingly, both prediction outcomes demonstrated remarkable consistency, highlighting the reliability and robustness of the created predictive models. These evaluations provide valuable insights into the models' effectiveness and their potential for practical applications. The performance of these models in predicting final results and dropout has been summarized in Table IV.

While predicting the final results, the Random Forest model showed exceptional performance with an accuracy of 0.88. This indicated a high level of overall correctness in predictions. RF model demonstrated a precision of 0.83, implying a relatively low rate of false positives. The RF model achieved the recall at 0.94. The RF model achieved a good performance while keeping the training time reasonably low at 3.98 seconds.

TABLE IV. PREDICTION REPORT OF EACH PREDICTIVE MODEL

	Final Result	**Dropout**	**Average**
Random Forest			
Accuracy	0.88	0.82	0.85
Precision	0.83	0.70	0.76
Recall	0.94	0.74	0.84
Time taken for training (s)	3.98	3.24	3.61
Artificial Neural Network			
Accuracy	0.86	0.78	0.82
Precision	0.80	0.58	0.69
Recall	0.94	0.97	0.96
Time taken for training (s)	79.18	228.78	153.98
Support Vector Machine			
Accuracy	0.83	0.79	0.81
Precision	0.82	0.64	0.73
Recall	0.82	0.68	0.75
Time taken for training (s)	2477.09	1776.97	2127.03

On the other hand, the ANN model achieved an accuracy of 0.86, slightly lower than the RF model. ANN model exhibited a precision of 0.80, indicating a moderate rate of false positives. The ANN model matched the RF model in terms of recall, also achieving a value of 0.94. However, the ANN model required a significantly longer training time of 79.18 seconds. The SVM model, while falling slightly behind in terms of accuracy at 0.83, still delivered commendable results. The SVM model demonstrated a precision of 0.82 and a recall of 0.82. Notably, the SVM model required a significantly longer training time of 2477.09 seconds.

Shifting the focus to the prediction of dropout, the RF model achieved an accuracy of 0.82, indicating a relatively high level of overall correctness in the predictions. The training time for dropout prediction was 3.24 seconds. In contrast, the ANN model exhibited an accuracy of 0.78, a precision of 0.58, and a remarkably high recall of 0.97. This implied that while the ANN model had a lower precision, the ANN model excelled in capturing the majority of positive instances accurately. The training time for dropout prediction using the ANN model was 228.78 seconds. The SVM model recorded an accuracy of 0.79, a precision of 0.64, and a recall of 0.68, and demonstrated relatively better precision compared to the ANN model but had a lower recall. The training time for dropout prediction using the SVM model was 1776.97 seconds.

Considering both the prediction of final results and dropout, the Random Forest model consistently delivered robust performance. Besides that, the RF model achieved outstanding performance among the three predictive models. On average, the RF model achieved an accuracy of 0.85, a precision of 0.76, and a recall of 0.84. The training time for the RF model was relatively low, averaging 3.61 seconds.

In a nutshell, the RF model proved to be the most reliable performer for predicting both final results and dropout, demonstrating a well-balanced combination of accuracy, precision, and recall. While the SVM model achieved slightly higher accuracy in predicting dropout compared to the ANN model, the SVM model took significantly longer to make predictions for both final results and dropout. Therefore, the predicted results based on the RF model were utilized to plot visualizations and enhance the accuracy of the data analysis process.

B. Insights Interpretation for Data Visualization

The goal of data visualization is to convey insightful information and analysis, encompassing descriptive, diagnostic, and predictive analytics. These three types of analytics are important in showcasing the anticipated performance of students and aiding educators in decision-making processes. The dashboards utilized the predicted results generated by the RF model. The RF model's exceptional accuracy enhances the decision-making process, making the decision-making process more effective. A total of 4 dashboards were created based on the wireframe design outlined in (Fig. 9-Fig. 13).

In the dataset, there are four semesters available: 2013B, 2013J, 2014B, and 2014J. Assuming the current semester is 2014J, three previous semesters have already passed. Consequently, there is a total of 2525 students enrolled in the current 2014J semester. The performance of these 2525 students has been predicted in this research to predict the student's outcomes and help in the decision-making process.

The Predictive Learning Analytics Dashboard consists of both descriptive and predictive analytics. This dashboard provides an overview of the anticipated student performance, as illustrated in Fig. 15. The Predictive Learning Analytics Dashboard offers valuable insights into anticipated student performance, empowering educational institutions to identify trends, take proactive measures, and support struggling students. By analysing key metrics such as dropout rates, pass/fail rates, and module-specific rankings, the dashboard provides a comprehensive understanding of student outcomes. The observed decline in pass rates and increase in dropout rates over four semesters highlight the need for preventive measures and targeted interventions. The educators may take some actions to enhance the academic performance of students. For instance, educators can try to view from the course module perspective and identify any difficulties the students may encounter.

The Dropout Analysis Dashboard as shown in Fig. 16 focuses on diagnostic analytics, aiming to provide a comprehensive understanding of the factors influencing predicted dropout rates and the correlation between student characteristics or behaviours and dropout. The Dropout Analysis Dashboard provides a deep understanding of the factors influencing predicted dropout rates and highlights the correlation between student characteristics and behaviours and dropout. This valuable information empowers educators and institutions to take proactive measures to support struggling students and mitigate dropout rates. By analysing the visualizations and insights offered by this dashboard, educational institutions can identify at-risk students based on various factors such as education level, disability, age range, environmental conditions, and vle behaviour. Armed with this knowledge, institutions can implement targeted interventions, provide tailored support, and create a supportive learning environment to improve student retention and success. The Dropout Analysis Dashboard serves as a powerful tool in shaping educational strategies, fostering student engagement, and enhancing the overall learning experience.

The Final Results Analysis Dashboard focuses on diagnostic analytics, aiming to provide a comprehensive understanding of the factors influencing predicted final results and the correlation between student characteristics and final results. This valuable information is visually represented in Fig. 17. In short, the Final Results Analysis Dashboard provides valuable insights into the factors influencing predicted final results and their correlation with student characteristics. Through visualizations such as pie charts, column charts, bar graphs, tree map, scatterplot, and a line graph, this dashboard offers a comprehensive understanding of the relationships between age, gender, highest education level, disability, environmental conditions, vle behaviours, and sum clicks. By analysing these factors, educators and institutions can identify areas of improvement and implement targeted interventions to support students' academic success. The dashboard enables educators and institutions to make data-driven decisions and allocate resources effectively to enhance student outcomes. Overall, this dashboard serves as a powerful tool for educational institutions, providing a holistic view of student performance and highlighting areas that require attention, ultimately leading to improved educational experiences and increased student success rates. By understanding the factors that affect students' final results, educators can take action to help them. For instance, educators can provide extra classes for students who do not have the formal qualifications for their education level.

The Students Analysis Dashboard as shown in Fig. 18 is a descriptive dashboard that offers a range of benefits for educators in tracking and supporting individual students. By providing various filtering and search options, such as disability, imd_band, code_module, highest education level, dropout risk level, predicted results, and predicted dropout, educators can easily locate specific students and tailor their interventions to address their unique needs. This dashboard allows educators to gain insights into students' progress, performance, and potential risk factors, enabling them to proactively identify struggling students and provide timely support. Additionally, the dashboard provides a comprehensive overview of each student's predicted results and dropout risk, empowering educators to prioritize their attention and allocate resources effectively.

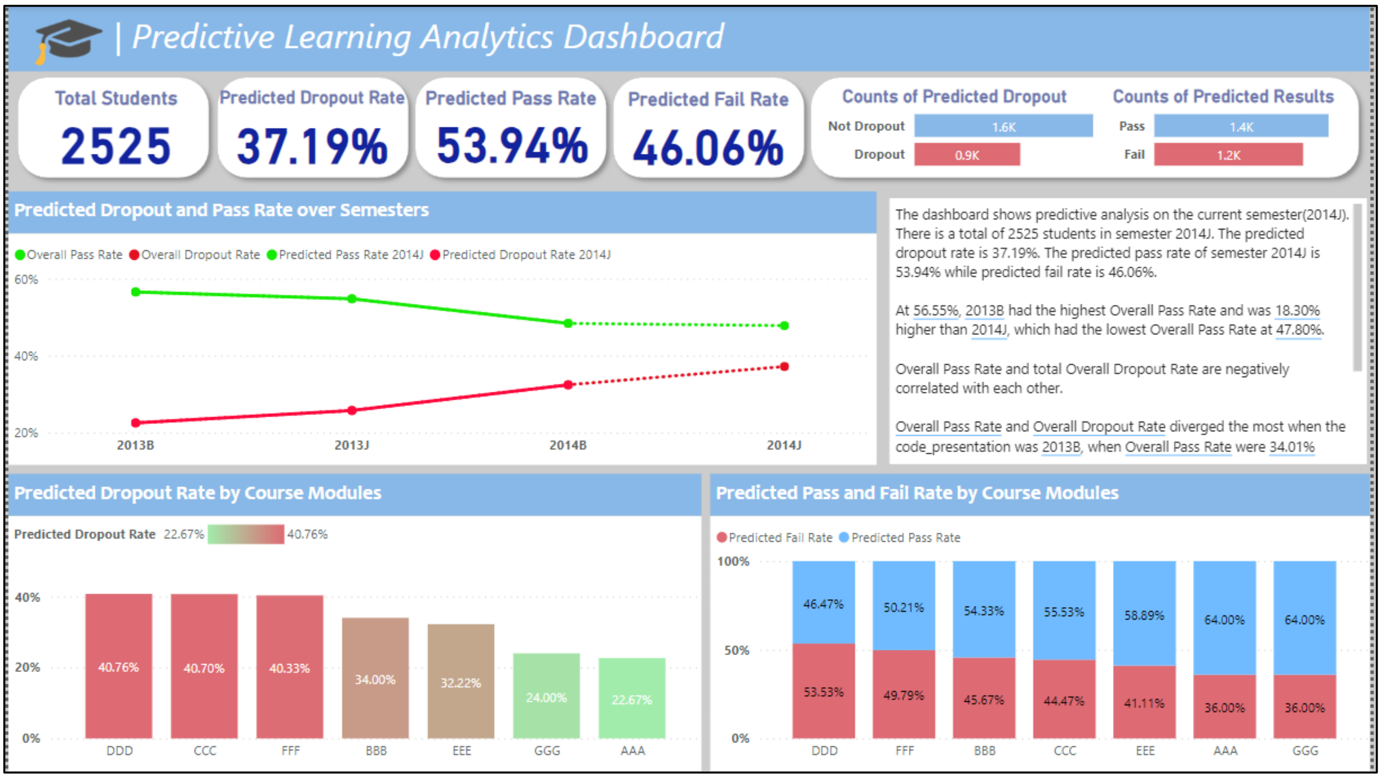


Fig. 15. Predictive Learning Analytics Dashboard

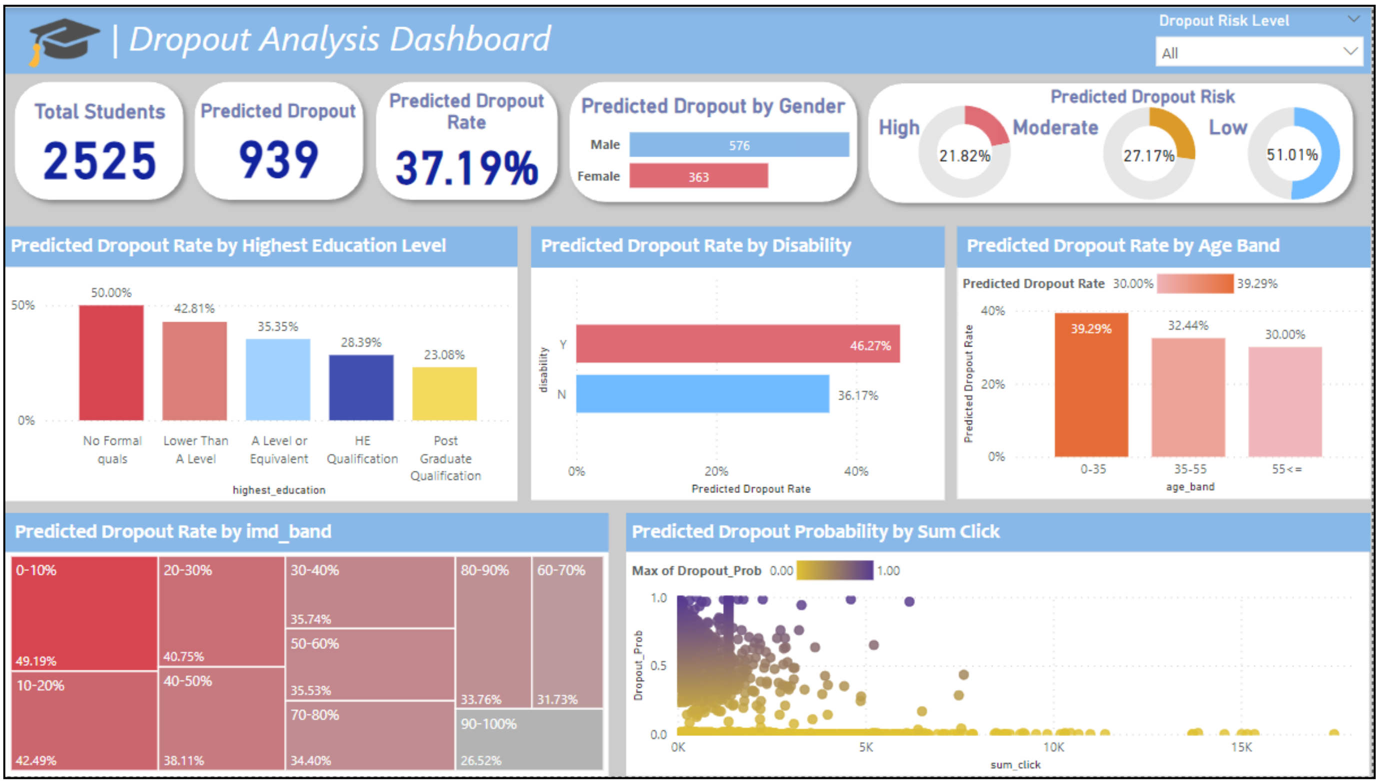


Fig. 16. Dropout Analysis Dashboard

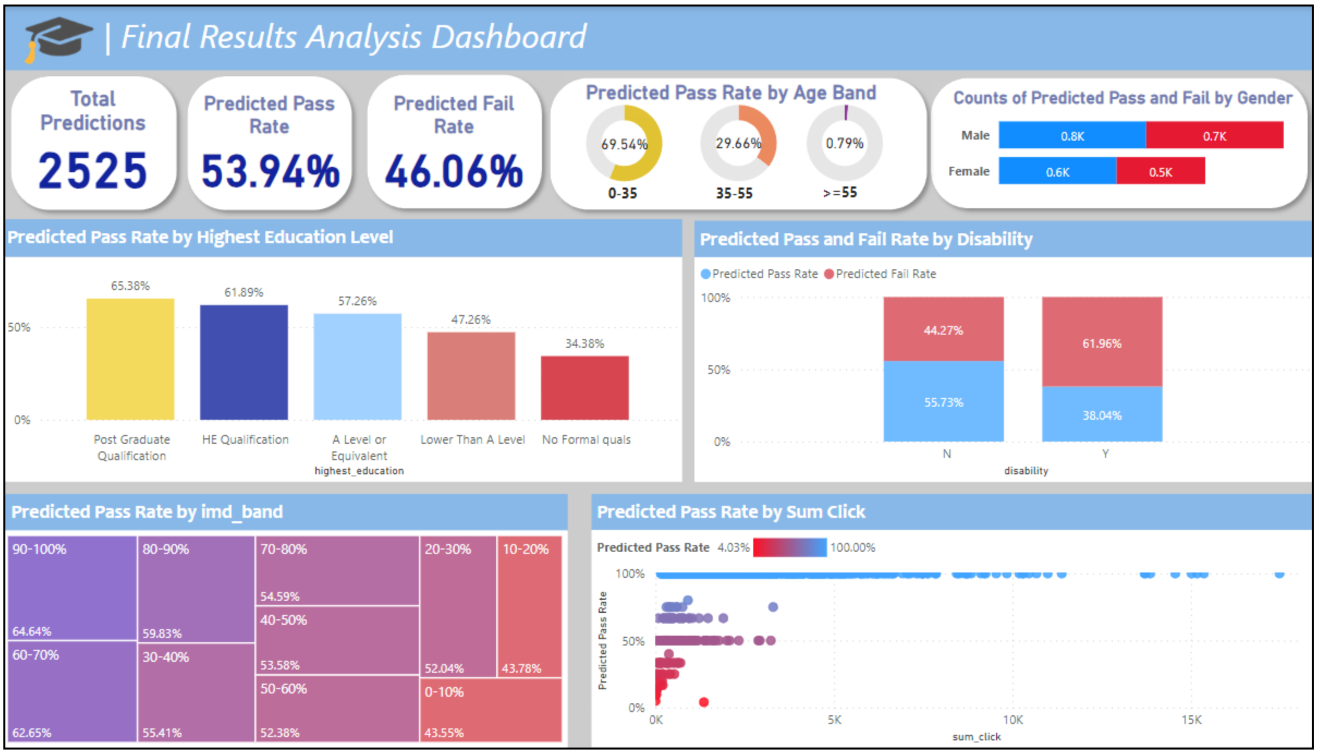


Fig. 17. Final Results Analysis Dashboard

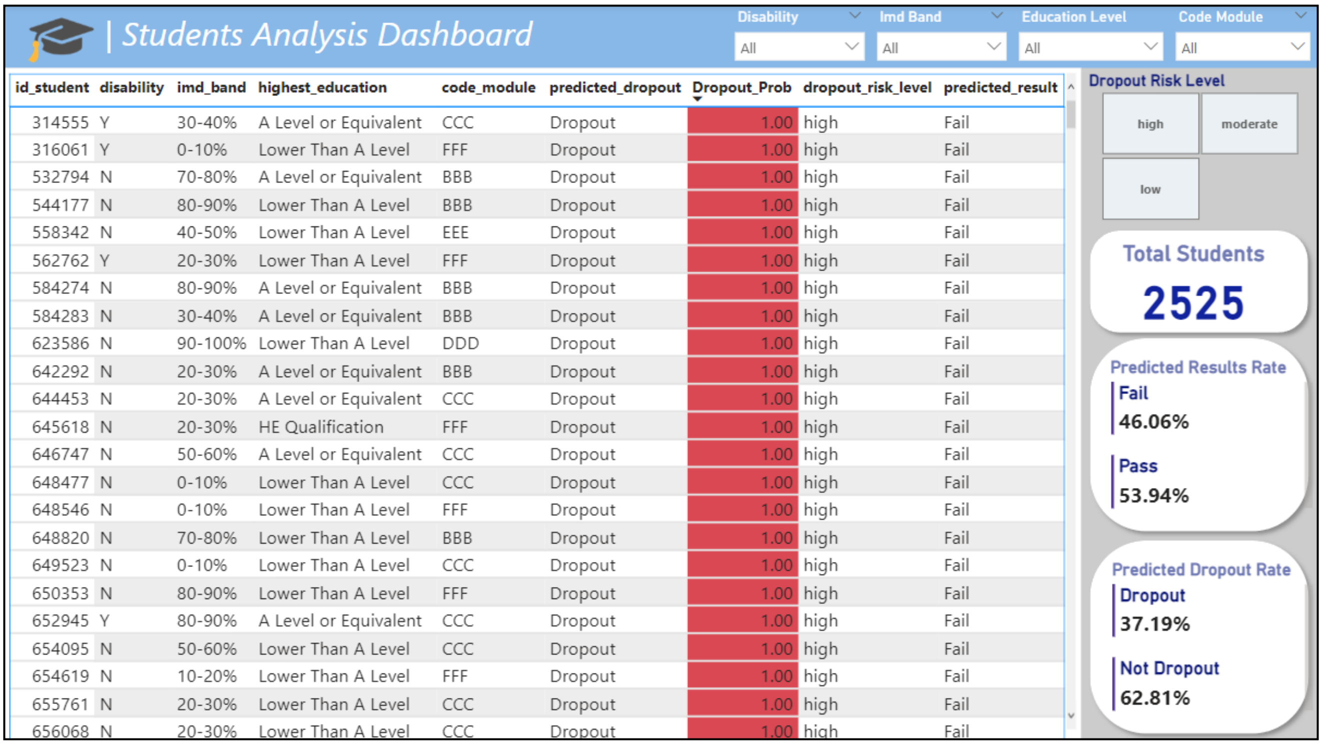


Fig. 18. Students Analysis Dashboard

The performance of three predictive models was evaluated based on various performance metrics and discussed in brief. In a nutshell, the RF model showed the best performance among the three predictive models for both cases of final result prediction and dropout prediction. The trained RF model is then implemented into our learning analytics dashboards as the main prediction model used for predicting student final results and student dropout. The predicted results were visualized in several dashboards as shown in Fig. 15-18, which highlighted the factors such as students' highest education level, learning environment, and Virtual Learning Environment behaviour significantly impacted their academic results and dropout rates. Overall, our designed dashboards with predictive capability intend to demonstrate the potentials that can be leveraged by educators in monitoring the student performance and getting more insights from the learning environment. These insights aim to assist educators in the decision-making process, empowering them with valuable information to support their actions.

IV. Conclusions

In a nutshell, three predictive models have been developed to predict students' performance including the final results and dropout. The data used for model training consisted of various kinds of students' data such as VLE behaviour and the student's results. The datasets were integrated into a single integrated dataset and the final dataset was used for the features selection process. The feature selection was also conducted using the RFECV technique to select the important features to enhance the prediction part. For the model training, the performance of the models was evaluated by accuracy, precision, recall, and time taken for training.

The Random Forest model demonstrated outstanding performance among the models, with an average accuracy of 0.85 and a training time of 3.61 seconds. In contrast, the SVM model recorded the lowest accuracy at 0.81 and took the longest training time, totalling 2127.03 seconds. Consequently, the results obtained from the RF model's predictions were used to create dashboards, which assist educators in customizing learning materials.

For future work, the datasets used for predicting students' performance could be enhanced by incorporating a more comprehensive range of features, such as students' e-learning activities, behaviours, and co-curricular involvement. Previous research has predominantly concentrated on academic outcomes.

Acknowledgment

Authors wish to thank Universiti Teknologi Malaysia and UTM Space for supporting this work through Quick Win Research Grant (R.J130000.7751.4J548)

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

References

[1] Patricia. (2020). Impact of online learning on students' stress levels: A Case Study. *International Journal of Online and Interactive Learning, 11*(30), 28-34.

[2] Fawaz, M. M., & Samaha, M. (2021). Student's perception of online learning during the COVID-19 pandemic: A case study of Lebanese Higher Education. *Frontiers in Education*, *6*, 615733.

[3] De Sousa, E. B. G., Alexandre, B., Mello, R. F., Falção, T. P., Vesin, B. and Gašević, D. (2021). Applications of learning analytics in high schools: A systematic literature review. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2021.737891.

[4] Long, P. D., Siemens, G., Conole, G., & Gašević, D. (2011). International Conference Proceeding Series (ICPS). *Proceedings of the 1st International Conference on Learning Analytics and Knowledge (LAK'11)*.

[5] Herodotou, C., Rienties, B., Boroowa, A., Zdráhal, Z. and Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: The teachers' role and perspective. *Educational Technology Research and Development, 67*(5), 1273-1306. https://doi.org/10.1007/s11423-019-09685-0.

[6] Torras-virgili, M., & Bellot-urbano, A. (2018). Learning analytics: Online higher education in management. *Proceedings of the 4th Annual International Conference on Management, Economics and Social Development (ICMESD 2018)*. https://doi.org/10.2991/icmesd-18.2018.46.

[7] Sengupta, S., Banerjee, A., & Chakrabarti, S. (2020). In-detail analysis on custom teaching and learning framework. *International Journal of Computer Applications*. *176*(33), 10-15. https://doi.org/10.5120/ijca2020920390.

[8] Maszlee. (2020). Malaysia to invest in data analytics to pursue spirit of mobility in education. https://www.thesundaily.my/local/malaysia-to-invest-in-data-analytics-to-pursue-spirit-of-mobility-in-education-maszlee-KX1606815.

[9] Liu, L. T., Wang, S., Britton, T. and Abebe, R. (2023). Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(9). https://doi.org/10.1073/pnas.2204781120.

[10] Pellas, N. (2021). The impact of learning platforms on education: A review of the literature. *Journal of Educational Technology Systems, 49*(3), 356-381.

[11] Samuelsen, J., Chen, W. and Wasson, B. (2019). Integrating multiple data sources for learning analytics—review of literature. *Research and Practice in Technology Enhanced Learning, 14*(1). https://doi.org/10.1186/s41039-019-0105-4.

[12] Guzmán, A., Barragán, S. and Vitery, F. C. (2021). Dropout in rural higher education: A systematic review. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.727833.

[13] Wintre, M. G., Bowers, C., Gordner, N. and Lange, L. (2006). Re-evaluating the University Attrition Statistic. *Journal of Adolescent Research, 21*(2), 111-132. https://doi.org/10.1177/0743558405285658.

[14] Beer, C. and Lawson, C. (2016). The problem of student attrition in higher education: An alternative perspective. *Journal of Further and Higher Education, 41*(6), 773-784. https://doi.org/10.1080/0309877x.2016.1177171.

[15] Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*. *72*, 414-422. https://doi.org/10.1016/j.procs.2015.12.157.

[16] (2020). The COVID-19 pandemic has changed education forever. This is how. *World Economic Forum*. https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/.

[17] Cao, J., & Qiao, X. (2020). Educational data mining and learning analytics in student online engagement. *Journal of Educational Computing Research*, *58*(3), 553-573.

[18] Huang, Y. M., Liang, T. H., & Su, Y. N. (2020). Predicting student learning in a flipped classroom using learning analytics. *Computers & Education, 145*, 103725.

[19] Open Learning Analytics. (2017). OU analyse. Knowledge Media Institute. The Open University. https://analyse.kmi.open.ac.uk/open_dataset.

[20] Agrawal, S. (2022). How to split data into three sets (train, validation, and test) and why? *Medium*. https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c#:~:text=Definition%20of%20Train%2DValid%2DTest,of%20these%20datasets%20is%20below.

[21] (2023). Split data: train, validate, test. https://apmonitor.com/pds/index.php/Main/SplitData.

[22] Or, B. (2023). On common split for training, validation, and test sets in machine learning. *Medium*. https://pub.towardsai.net/breaking-the-mold-challenging-the-common-split-for-training-validation-and-test-sets-in-machine-271fd405493d.

[23] Zach. (2021). Validation set vs. test set: What's the difference? *Statology*. https://www.statology.org/validation-set-vs-test-set/.