Pak. j. life soc. Sci. (2024), 22(2): 19224-19238 E-ISSN: 2221-7630;P-ISSN: 1727-4915

Pakistan Journal of Life and Social Sciences
www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.001410

RESEARCH ARTICLE

# Prediction of Student Academic Status in Higher Education Through Machine Learning

Ting Tin Tin[1*], Loh Eng Sin[2], Koong Jie Lum[3], Tham Hiu Huen[4], Ayodeji Olalekan Salau[5,6]

[1,5]Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Malaysia
[2-4]Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia
[5]Department of Electrical/Electronics and Computer Engineering, Afe Babalola University, Ado-Ekiti, Nigeria
[6]Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India

***ARTICLE INFO***

Received: Sep 12, 2024

Accepted: Nov 3, 2024

***Keywords***

Higher education institutions
Student dropout rates
Academic success
Artificial intelligent models
Random Forest
Logistic regression
Decision Tree Support Vector Machines
K-Nearest Neighbors

***Corresponding Author**

tintin.ting@newinti.edu.my

**ABSTRACT**

Academic achievement and student dropout rates are issues that higher education institutions worldwide are dealing with. The rapid growth of educational data makes it difficult for institutions to use them to improve the educational system. To predict student dropout and academic success, this study developed and compared several artificial intelligence (AI) models, including Decision Tree (DT), Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and K Nearest Neighbor (KNN). This study uses a data set that contains educational, socioeconomic, and macroeconomic data which helps in gaining further insight into how these factors influence student dropout rates or academic success outcomes. Predictive model validations were achieved by using multiple assessment criteria such as the ROC Curve, cross-validation, and evaluation metrics, which are Accuracy, Recall, Precision, and F1 score. The results of this study proved that the LR model performed the best, with the highest area under the curve and evaluation metrics of 0.92011, 0.93281, 0.85199, and 0.89057. This machine learning-based study would be helpful to higher education institutions in predicting students in the risk of dropping out and providing interventions to reduce dropout rates.

## 1. INTRODUCTION

Academic success is something that every student in the world. This is because good academic success will most likely ensure a successful future because they will be recruited by high-profile companies that pay a generous amount of salary to their employees to motivate and encourage them to give their all to the company. However, there are many levels to academic success, and different students may view academic success in different lights. Some may think that graduating with a passing score is considered academic success whereas others will only consider graduating with first-class honors an academic success.

Dropping out of higher education programs before obtaining a degree is a widespread issue that affects students around the world. According to recent data provided by the OECD, approximately 33% of students, on average, who enroll in universities fail to complete their studies and obtain a degree (Aina et al., 2022; Ting et al., 2022). This statistic is concerning, as it highlights the challenges that students face in higher education institutions. Research has shown that the first academic year is the most vulnerable period for students to drop out of higher education programs (Willcoxson et al., 2011; Chen, 2012; Wild and Schulze Heuling, 2020; Wild et al., 2023). Therefore, providing adequate support to students during their first year of higher education is crucial to ensure their

success and reducing the risk of dropping out. Education institutions should be able to predict a student's chance of dropping out as early as possible to address this problem (Singh and Alhulail, 2022). To reduce the issue and enable focused interventions to support these students, it is essential to classify students who are at risk of early dropouts (Niyogisubizo et al., 2022). Machine learning is used to develop a prediction model for student dropouts that can give educational institutions an early warning and allow them to take different actions to prevent students from leaving their programs (Del Bonifro et al., 2020; Niyogisubizo et al., 2022).

This study makes important contributions to theory, concept, application, and society. Understanding how machine learning algorithms were used to demographic, socioeconomic, macroeconomic, and educational data helps to find characteristics that influence student dropout and academic success is one of the theoretical implications of this research. The concept developed can be applied to other academic institutions and student populations to address educational challenges. From a practical perspective, academic institutions can use the predictive tool developed to identify at-risk students and provide targeted interventions and support systems to improve their academic success and reduce dropout rates. In terms of society, this research has the potential to improve individual students' lives and contribute to the creation of a more skilled and educated workforce, leading to broader social and economic benefits.

The primary goal of this research is to preprocess the students' dropout and academic success dataset from Kaggle, including data cleaning, data integration, data transformation, feature selection, and data splitting to prepare for analysis. Then use supervised machine learning techniques, specifically Decision Tree (DT), Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and K Nearest Neighbor (KNN), to develop statistical models that estimate the likelihood of dropout and academic success. The model will be fed with the designed train data set. Finally, test the statistical model using a designated test data set to evaluate its performance on new, unseen data and to assess its ability to generalize to new situations. This allows us to identify any potential problems before deploying the model in a real-world application.

The research structure is as follows: Section 2 states the related approaches, while Section 3 describes the methodology. This section includes details of the description of the dataset used, the preprocessing techniques applied to it, the machine learning methods used in the analysis and the methods used for model evaluation. Section 4 presents the results of the analysis, which includes an assessment of the performance of various models using four measures: Accuracy, Recall, Precision, and F1 score. Finally, in Section 5, conclusions were drawn based on the findings, and we discuss potential applications and extensions of this study.

### 1.1 Problem Statement

Higher education institutions around the world face a significant challenge with student dropout and academic success rates. According to UNESCO findings, COVID-19 has brought disruption to education that causes about 24 million students in 2020 to be at risk of being unable to continue their studies and eventually most of them dropout. Among the affected students ranging from pre-primary to university level, university students make up a significant portion, estimated to be around 7.9 million students with a decline rate of 3.5% in enrollment due to costs of studies (Song *et al.*, 2023). Financial constraints are not the only factors that lead students to abandon their studies or hinder them from achieving academic success, other factors such as family issues and learning difficulties can also contribute to student dropout. In today's competitive job market, where there is a growing demand for highly educated and skilled workers, academic success has become more critical than ever for students. Premature entry into society for these dropout students will likely subject them to significant social pressure, resulting in missed employment opportunities, reduced earning potential, poorer quality of life, and adverse health outcomes. High dropout rates can be detrimental to academic institutions such as wasted resources, reduced revenue, and reputation which affect their ability to attract students and result in staff reductions. Therefore, academic institutions are under pressure to retain their students to ensure their academic success, as this can have a significant impact on funding and rankings. Society as a whole may face a shortage of skilled workers, leading to a decrease in economic growth and productivity, while simultaneously causing an increase in crime and poverty rates.

Another challenge higher education institutions are facing today is the explosive growth in educational data, as well as the difficulty of using this information to make informed decisions and improve the quality of the education system (Bhutto et al., 2020). However, recent years have seen a rise in the application of educational data mining (EDM) techniques, which can provide a solution to this problem. The meaning of EDM is to apply data mining techniques to the enormous amount of data generated within the education sector. By analyzing the dataset, EDM can identify patterns and relationships that can be used to develop appropriate corrective strategies to refine teaching methods (Casquero et al., 2013; Fidalgo-Blanco *et al.*, 2015; Yağcı, 2022). For example, educational institutions can use EDM to identify successful students and those at risk of failure and determine the most important factors that contribute to their academic performance. Once these students have been identified, institutions can develop targeted interventions that address their specific needs such as providing financial aid or scholarships to students who are at risk of dropping out due to financial constraints, develop more effective curricula that are better aligned with students' needs and interests, provide teacher training, and assessment methods to enhance student learning outcomes (Yi et al., 2022).

It is important to identify prospective dropouts among the students and take proactive steps before dropout behavior occurs to prevent all the negative consequences for students, education institutions, and society. The aim of this research is to make use of the machine learning methods to develop a predictive model and to implement EDM by utilizing demographic, socioeconomic, macroeconomic, and educational data of students for model training. After training, the prediction model can accurately identify students at risk of dropping out and identify the most significant factors that contribute to academic success and dropout behavior. This enables educational institutions to identify at-risk students early and provide targeted interventions to support their academic success.

## 2.0 LITERATURE REVIEW

### Logistic regression

Statistically, the connection between a binary or categorical dependent variable with one or more independent variables may be modeled using logistic regression. In machine learning, it is a well-liked and often used classification technique, especially for binary or dichotomous dependent variables. The main objective of logistic regression is to forecast the likelihood that an event or result will occur based on the values of one or more independent variables. In other words, given a set of predictor variables, logistic regression can be used to predict the probability of an event or result occurring. When the dependent variable is dichotomous, or only has one of two potential values, logistic regression is especially helpful (Kleinbaum, 1994).

### Support Vector Machines

SVMs can be useful because they are able to handle nonlinear decision boundaries, which can be important when the classes are not easily separable by a linear decision boundary. SVMs are also robust to overfitting and can handle high-dimensional datasets, which is useful when dealing with datasets that have many features (Kecman, 2005).

### Random Forest

Using diverse subsets of the training data and various random selections of the characteristics, the random forest ensemble learning technique constructs several decision trees. The outcome is then determined by combining all of the decision trees' projections. The number of decision trees that should be constructed in the parameter of forest is determined by the n estimators. In general, increasing the number of trees in the forest can improve the performance of the random forest algorithm, but doing so also adds to the computational expense of creating and utilizing the model. Finding the ideal balance between model performance and computational expense is crucial (Breiman, 2001).

### K-Nearest Neighbors

Instances are supposed to be classified according to their nearest neighbor's class in nearest neighbor classification. The nearest neighbors K are employed in the (k-NN) classification method to determine the class. It is also known as Memory-based classification since the training samples must be available at runtime, i.e. in memory. The fact that induction is postponed until run-time qualifies it as a lazy learning strategy. Since it simply uses the training examples, this can be identified as case-based classification, or example-based classification and this distance metric is used to pick the K-Nearest Neighbors. (Cunningham and Delany, 2021).

**Decision Tree**

Decision trees are popular tools due to their ability to extract classification rules that are easier to understand from feature-based examples. Recent developments in decision tree induction techniques have been effectively used for a wide range of classification challenges. On small to medium-sized data sets, several of the decision tree methods that have been presented so far work well. Due to various memory constraints, time constraints, and data complexity bottlenecks, it is challenging to train a decision tree from huge data sets (Fletcher and Islam, 2019).

**Related Work in Student Dropout Prediction**

Machine learning (ML) techniques can be used in our analysis and research on the prediction of students' dropout rate according to the papers conducted by Balqis Albreiki (2021), which uses Educational Data Mining (EDM) method and uses ML as an essential role in similar prediction of student performance. Machine learning algorithms accompanied by statistical methods are used to analyze educational data to explore meaningful insights that help improve their academic performance with the criteria of massive data needed to perform accurately. The overview of this research focuses on two critical aspects, which are accurate students' risk prediction and student dropout, while 3 features are used in the studies: demographics, academic and e-learning interactions. Conclusions have been made such as most of the studies use traditional algorithms such as Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) compared to the investigation of deep learning potential and ML has potential to provide great improvements in education system through transformation of teaching, learning and research (Albreiki et al., 2021).

According to research on the students' academic performance of students through educational data mining approaches by Mustafa Yagci, he proposed and suggested that Nearest-Neighbor, Random Forest and Logistic Regression are algorithms with higher classification accuracy of 75% to be used to predict the grades of undergraduate students. 2 parameters are focused, which is the prediction of student academic grades based on their previous results and the comparison of performance between the chosen machine learning algorithms, while 3 types of parameters are used in prediction which are midterm grades, department data, and faculty data. Artificial neural network (ANN), Logistic Regression (LR) and random forest (RF) algorithms is used to identify the risk of students' academic failure of students depending on their demographic characteristics and logistic regression has the highest degree of precision after comparison. It is said that education staff should benefit from these results where it helps in early recognition of students who have below average academic motivation and therefore machine learning would be a great assistance in establishing the learning analytics framework and contributing to the decision-making process in the education world (Yac, 2022).

From the review by Rastrollo-Guerrero (2020) of student performance through applications of machine learning, approximately 70% of the articles analyzed in this review had chosen to explore student status or performance during tertiary education. The primary goal of this paper is to reveal a more in-depth overview of the algorithms proposed in this field of study. More than 70 articles were included to determine common techniques that has been widely used in students' performance prediction, and these techniques are mainly related to Artificial Intelligence (AI), Collaborative Filtering, Recommender Systems and Machine Learning (ML). Few objectives have been emphasized as weight indications of interest for research, which is student dropout, performance, knowledge and recommended activities or resources. Supervised learning is the most widely used in this research to provide more reliable results, where it can reason from instances to produce general hypotheses and predictions about future instances. The SVM algorithm is ranked highest among all followed by Naive Bayes and Random Forest but, however, these classification algorithms should be applied together

with preprocessing techniques to improve prediction results. Unsupervised learning is not preferred to be used because of its low accuracy but it is still usable to be applied as one of the approaches. Neural networks are less used, but this study also obtains great precision in predicting students' performance (Rastrollo-Guerrero et al., 2020).

The research conducted by Engr. Sana Bhutto and few University Departments of International Conference (2020) regarding students' academic performance prediction through supervised machine learning states that the introduction of the Support Vector Machine (SVM) and Logistic Regression (LR) is one of the few powerful algorithms. The knowledge found in this paper is able to assist in predicting the future behavior of students to categorize their performance and also provides the best technique to find impactful features to work on that will decrease the dropout ratio. It is stated that one of the main challenges education institutes face is the exponential growth in educational data to conduct proper analysis helping to improve overall quality of the system. Data cleaning, attribute extraction, and data transformation is performed as data preprocessing to assist in providing clean data to increase efficiency of applying data mining techniques. Three evaluation measures, which are recall, precision and f1 score are calculated using a confusion matrix to choose the most suitable predictive model, and the accuracy of both stated models is at least 70% accuracy (Bhutto et al., 2020).

According to the research overview on quitting studies and changing majors between university students by Lisa Baulke (2021), determinants of significant correlations between selected variables leading to dropout intentions have been found. This study was conducted first by providing online surveys and questionnaires to all students of one of the German universities through emails for further analysis. Few phases leading to student dropout have been determined through model assumptions, which are non-fit perceptions, dropout intentions, deliberation, information enquiries, and towards final decisions. Structural Equation Modeling (SEM) approach is used to check the acceptance of hypotheses created and the factors that play a major role in this phenomenon or intentions are academic self-concept, procrastination, anxiety, subjective task value, and learned helplessness (Bäulke et al., 2022).

From the conference paper on Student Dropout Prediction released by Francesca Del Bonifro (2020), machine learning approaches are used to analyze insightful information behind the datasets through design principles such as early risk prediction stage of quitting during their first year of study. The time frame is determined as one of the most important factors for dropout and personal information associated with their previous high school academic results is chosen to train selected models because this research is conducted between freshmen. The selected models in this research are Random Forest (RF), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) and found out that the performance of Random Forest is the highest compared to Support Vector Machine and Linear Discrimination Analysis according to their results of different combination of feature sets (Del Bonifro et al., 2020).

## 3.0 RESEARCH METHODOLOGY

### Data Description

This study uses the Kaggle dataset which includes demographic data, socioeconomic data, macroeconomic data and academic performance information to understand and predict student dropouts and academic success. Demographic, socioeconomic, and the previous academic performance data were collected during student enrollment, while academic performance data was collected at the end of the first two semesters. The macroeconomic data were collected to gain more insight on how economic factors influence student dropout rates or academic success outcomes. The data set comprised a total of 4,424 student records, with 2,209 graduates, 1,421 dropouts, and the remaining 794 students currently enrolled. The data set consists of 4424 records, each of which has 35 features and without missing values. It was obtained from undergraduates in 17 programs spanning various fields during a decade of academic years, from 2008/2009 to 2018/2019.

To develop a model that predicts student dropout or academic success, the rows with the target value of "Enrolled" were removed from the dataset. This decision was made because the "Enrolled" status only indicates that a student is currently studying at the university and does not provide information

on whether the student will drop out or graduate in the future. This study predicts academic status after their successful enrollment into their degree studies. In this study, we also decided not to remove any features and used all the features available in the dataset. This decision was made because the available features are not as much as expected and, by reference from literature review, it is well known that many features of different categories of information from individual demographic, family backgrounds, institution information to even surrounding social affections have potentials on affecting the students' academic status of students. Details of attributes are summarized in Table 1.

**Table 1. Attributes used grouped by class of attributes.**

| Category | Attribute | Type | Values |
|---|---|---|---|
| Demographic data | Marital status | Numeric/discrete | 1—6 |
| | Nationality | Numeric/discrete | 1—21 |
| | Displaced | Numeric/binary | 1—yes, 0—no |
| | Gender | Numeric/binary | 1—male, 0—female |
| | Age at enrollment | Numeric/discrete | 17—70 |
| | International | Numeric/binary | 1—yes, 0—no |
| Socioeconomic data | Mother's qualification | Numeric/discrete | 1—34 |
| | Father's qualification | Numeric/discrete | 1—34 |
| | Mother's occupation | Numeric/discrete | 1—46 |
| | Father's occupation | Numeric/discrete | 1—46 |
| | Educational special needs | Numeric/binary | 1—yes, 0—no |
| | Debtor | Numeric/binary | 1—yes, 0—no |
| | Tuition fees up to date | Numeric/binary | 1—yes, 0—no |
| | Scholarship holder | Numeric/binary | 1—yes, 0—no |
| Macroeconomic data | Unemployment rate | Numeric/continuous | 7.6—16.2 |
| | Inflation rate | Numeric/continuous | -0.8—3.7 |
| | GDP | Numeric/continuous | -4.06—3.51 |
| Academic data at enrollment | Application mode | Numeric/discrete | 1—18 |
| | Application order | Numeric/ordinal | 0—9 |
| | Course | Numeric/discrete | 1—17 |
| | Daytime/evening attendance | Numeric/binary | 1—daytime,0—evening |
| | Previous qualification | Numeric/discrete | 1—17 |
| Academic data at the end of 1st semester | Curricular units 1st sem (credited) | Numeric/discrete | 0—20 |
| | Curricular units 1st sem (enrolled) | Numeric/discrete | 0—26 |
| | Curricular units 1st sem (evaluations) | Numeric/discrete | 0—45 |
| | | Numeric/discrete | 0—26 |
| | Curricular units 1st sem (approved) | Numeric/continuous | 0—18.9 |
| | Curricular units 1st sem (grade) | Numeric/discrete | 0—12 |

| | | | |
|---|---|---|---|
| | Curricular units 1st sem (without evaluations) | | |
| Academic data at the end of 2nd semester | Curricular units 1st sem (credited) | Numeric/discrete | 0—19 |
| | Curricular units 1st sem (enrolled) | Numeric/discrete | 0—23 |
| | Curricular units 1st sem (evaluations) | Numeric/discrete | 0—33 |
| | Curricular units 1st sem (approved) | Numeric/discrete | 0—20 |
| | Curricular units 1st sem (grade) | Numeric/continuous | 0—18.6 |
| | Curricular units 1st sem (without evaluations) | Numeric/discret | 0—12 |
| Target | Target | Numeric/discrete | 1—dropout<br>2—graduate |

**Data Preprocessing**

Data preprocessing is a collection of methods used to clean, organize, and arrange raw data so that it can be utilized quickly and efficiently for analysis or machine learning. Although it may significantly affect the quality and dependability of the results, it is a crucial phase in the statistical modeling and machine learning process. The pre-processing steps that were performed in this study:

1. Data cleaning: This required carefully inspecting the data set for errors, discrepancies, and missing numbers. The data set was found to be error-free and devoid of contradictions or missing values, which meant that no entries had to be deleted.
2. Data Encoding: To use the data in machine learning, category variables must be converted to numerical variables. This study represented the categorical variables "dropout" and "graduate" numerically as 1 and 2, respectively.
3. Data Splitting: To assess the effectiveness of the machine learning model, this includes splitting the data into training and testing sets. This study separated the data set into the following sections to make sure the model developed is reliable: The training set received 80% of the data, and the testing set was given 20%.

**Model Selection**

This study uses kernel-based, tree-based, and linear models to predict whether a university student would drop out or graduate, which is a binary classification problem. If-then-else rules are used in tree-based models to address issues. Both classification (which predicts categorical values) and regression (which predicts numerical values) can be used with all tree-based models. To resolve the issue, kernel-based models convert nonlinear problems into linear problems in feature space. It is possible to think of linear models in terms of functions that generate predictions based on linear combinations of features. In this paper, 5 classification models are chosen: Random Forest, Logistic regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor.

These five models have been selected to be used in this investigation due to the references from a few researches included in the literature review. These five models are commonly used in the predictions of similar datasets that are closely related to students' performance and their significantly higher performances such as accuracy and precision scores after comparison with other models.

**Model Evaluation**

The ROC curve is used to ensure the overall performance of all models is good to be used. Cross-validation is used to verify that the fitted model is not overfitted or underfitted. After confirming that the three models selected are good to use, performance measures are performed to decide which is the best model.

### Test for over- and underfitting

A model is overfit if it has been overtrained on the data to the extent that it has even learned the noise from it. Because it learns every sample with such precision, an overfit model incorrectly classifies a previously unidentified or novel case. If the train accuracy is significantly higher than the test accuracy, it is said to be overfitting. The model is considered an underfit if the model is not able to recognize the data patterns correctly. Every example in the dataset is not fully learned by an underfit model and both the training and testing set will display a lower score.

### ROC curve

An illustration of a binary classifier system's performance on a graph is called a receiver operating characteristic (ROC) curve. It is a helpful instrument for assessing the efficacy of a model that generates probabilities or scores for binary classification issues. For each threshold value, the True Positive Rate (TPR) and False Positive Rate (FPR) will be calculated. The TPR measures the proportion of real positive cases that the model correctly identifies as positive, while the FPR measures the proportion of real negative cases that the model misidentified as positive. For each threshold value, this study plots the TPR against the FPR to obtain the ROC curve (Fan et al., 2006). The area under the ROC curve (AUC) is calculated to assess the overall performance of the model. The AUC runs from 0 to 1, with the closer the ROC value to 1, the better the overall performance of the model and the closer it is to 0, the worse the overall performance of the model.

### Cross-validation

One score array is returned by the cross_val_score method for each fold of cross-validation. These ratings are used to determine whether a model is over- or underfitting. If the scores are consistent across all folds, the model is most likely well-fit and not over- or under-fitting. If the training and validation scores are considerably different, the model may be overfitting. If the model's training and validation scores are poor, it may be underfitting. If the score range is smaller than 0.05, it is considered consistent across folds. This indicates that the model is a good fit for data if the difference in accuracy scores between the highest and lowest folds is less than 5% (Bro et al., 2008).

### Performance measures

Performance metrics (F1 score, recall, accuracy, precision) for machine learning classification models are used to assess how well they perform under specific conditions. This study compares the value between these four measures to determine the best model among the three valid models.

Accuracy is the proportion of the entire sample that the model correctly predicted and is determined below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Recall (or true positive rate), which is defined as follows, measures the model's ability to recognize positive samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision (or positive predictive value) is the ratio of samples that are correctly categorized as positive to all samples that are correctly or mistakenly classified as positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The definition of F1 score provides a thorough assessment of the classifier performance throught the uses of recall and precision.

$$\text{F1 Score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

Figure 1 shows the process and flow of our research from the determination of the dataset to model evaluation:

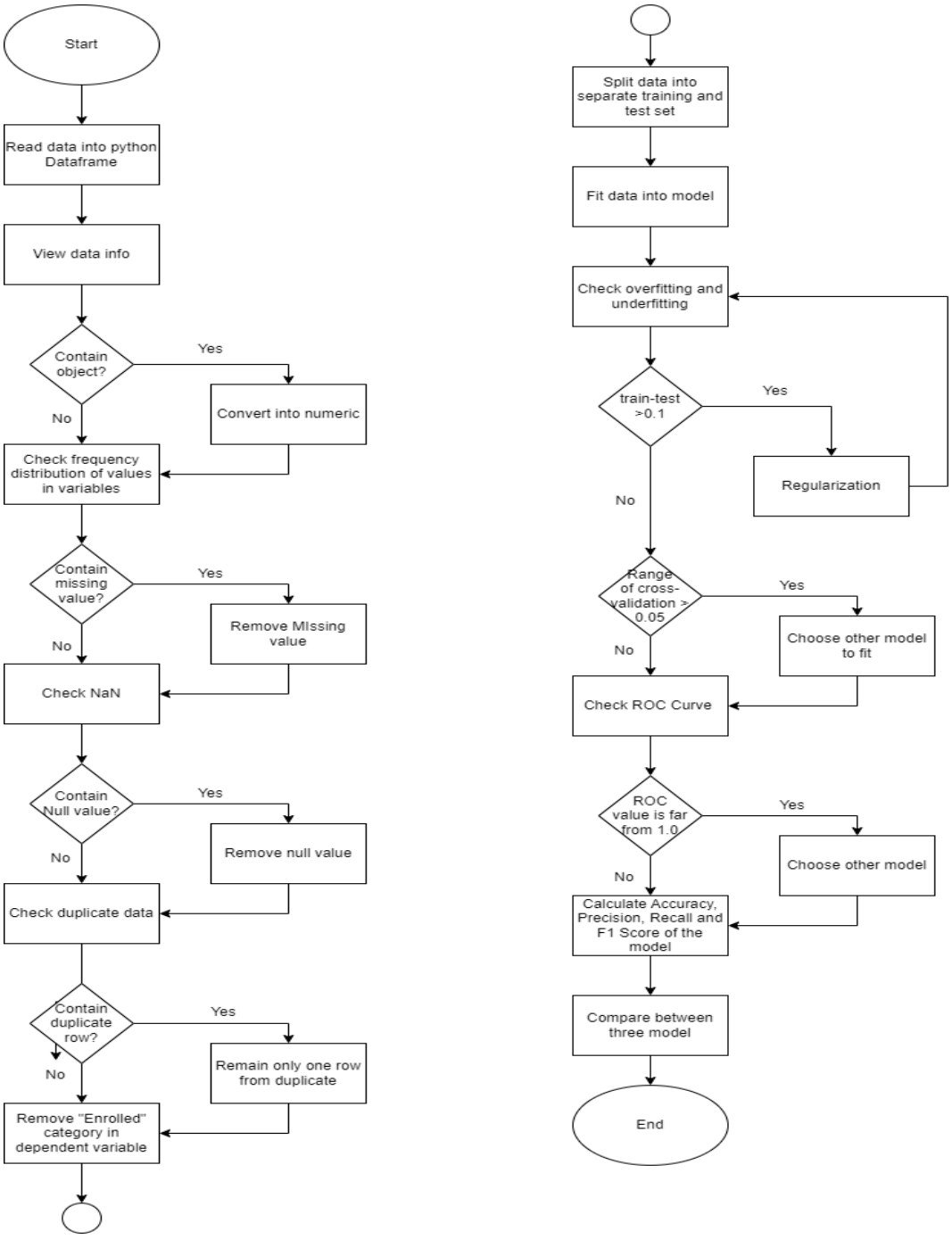


**Figure 1: Research Path of this research study.**

## 4.0 RESULTS AND DISCUSSIONS

The selected models below are listed in ascending order: Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, K-Nearest Neighbor. These are the results of this research after the implementation of our dataset into the five selected models to find the most suitable predictive model to perform evaluation on student academic status.

**Performance metrics**

**Table 2: Comparison of performance metrics among selected models**

| | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| **LR** | 0.92011 | 0.93281 | 0.85199 | 0.89057 |
| **RF** | 0.89669 | 0.89764 | 0.82310 | 0.85876 |
| **RF(regularized)** | 0.89118 | 0.91949 | 0.78339 | 0.84600 |
| **SVM** | 0.90771 | 0.93750 | 0.81227 | 0.87041 |
| **DT** | 0.84298 | 0.84277 | 0.84298 | 0.84287 |
| **DT(regularized)** | 0.88567 | 0.89190 | 0.88567 | 0.88253 |
| **kNN** | 0.91667 | 0.91667 | 0.59567 | 0.72210 |
| **Highest (1,2,3)** | LR<br>SVM<br>KNN | SVM<br>LR<br>RF(regularized) | DT(regularized)<br>LR<br>DT | LR<br>DT(regularized)<br>SVM |

Table 2 above shows the overall performance metrics of all the selected models to show a clearer vision of the comparisons to determine the best-performing predictive model. The logistic regression model is the best performance model among all the models because of the evaluation metrics: Accuracy: 0.92011, Precision: 0.93281, Recall: 0.85199, F1-Score: 0.89057 are always the top three among all the models.

**Overfitting/Underfitting Test**

Regularization of Random Forest is done by:

(max_depth=5, min_samples_split=5, min_samples_leaf=5, random_state=42)

Default model score: 0.8966942148760331 Regularized model score: 0.8911845730027548

Regularization of Decision Tree is done by:

(max_depth=3, min_samples_split=5, min_samples_leaf=2)

Default model score: 0.8429752066115702 Regularized model score: 0.8856749311294766

The Overfitting/Underfitting test is performed and the results are shown in Table 3 above to determine whether regularization of the model is needed. The test shows that logistic regression, support vector machine, and nearest neighbor K-Nearest Neighbor do not have to perform any model regularization because the difference between their training set and testing set accuracy are not significant. Random Forest and Decision Tree are required to perform regularization to further reduce the difference between their dataset accuracy because the difference is quite significant which are both higher than 0.1 and are considered overfitting as training set accuracy is higher than testing set accuracy. Logistic regression has the lowest accuracy difference after comparison between all the models and regularized models.

**Table 3: Results and Difference Calculated in Overfitting/Underfitting Test**

| | Training Set Accuracy | Testing Set Accuracy | Difference |
|---|---|---|---|
| **LR** | 0.91667 | 0.92011 | -0.00344 |
| **RF** | 1.00000 | 0.89669 | 0.10331 |
| **RF(Regularized)** | 0.90978 | 0.89118 | 0.01860 |
| **SVM** | 0.91563 | 0.90771 | 0.00792 |
| **DT** | 1.00000 | 0.84298 | 0.15702 |
| **DT(Regularized)** | 0.89360 | 0.88567 | 0.00792 |
| **KNN** | 0.84780 | 0.82507 | 0.02273 |

**Cross-Validation Test**

The Cross-validation test is conducted, and the results are shown in Table 4 above to determine whether the trained model is a good fit model for prediction and performance. The above results show that the score range of all the models and the regularized models to be used are smaller than 0.05 which means the accuracy scores between folds are consistent where all the models are good fit for this dataset. The regularized random forest model is shown to be the lowest range score compared to other selected models, followed by logistic regression, regularized decision tree, support vector machine and K-Nearest Neighbors.

**Table 4: Results and Range Difference Calculated in the Cross-Validation Test**

| | **Mean** | **Maximum** | **Minimum** | **Range** |
|---|---|---|---|---|
| **LR** | 0.91925 | 0.92424 | 0.02066 | 0.02066 |
| **RF (Regularized)** | 0.89587 | 0.89807 | 0.89532 | 0.00275 |
| **SVM** | 0.91157 | 0.92700 | 0.89945 | 0.02755 |
| **DT (Regularized)** | 0.89174 | 0.90358 | 0.87879 | 0.02479 |
| **KNN** | 0.82920 | 0.84573 | 0.81405 | 0.03168 |

**Receiver Operating Characteristic (ROC) Curve Test**

The ROC curve test is conducted, and the results are shown in Figure 2 above to check the overall performance of selected models. The results above show that Logistic Regression and Regularized Random Forest have higher Area Under Curve (AUC) which are both very close to 1 (0.96), followed by SVM, Regularized Decision Tree and K-Nearest Neighbors. This indicates that the overall performance of regularized random forest and Logistic Regression is slightly better than other selected models.

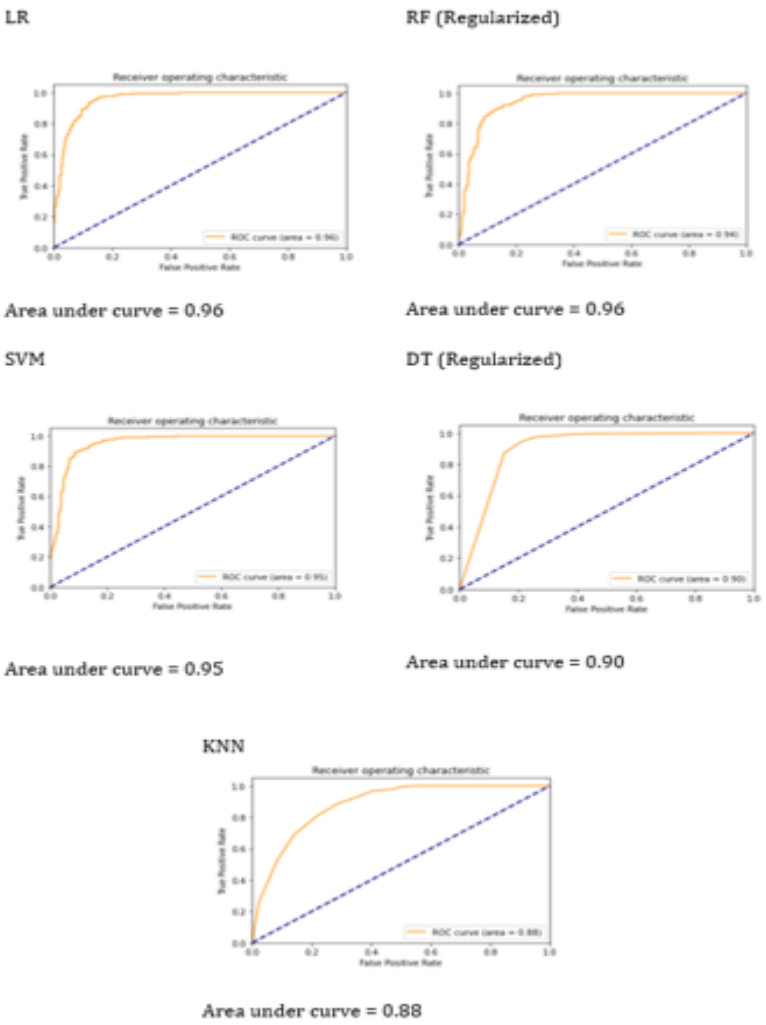


**Figure 2: Results and graph of ROC Curve**

Based on all the evaluations that have been done, the best model to predict the student academic status in university is logistic regression. Z

## 5.0 CONCLUSION

In conclusion, after a rigorous evaluation of various machine learning models, this study has determined that logistic regression is the optimal model for the research. The conclusion is based on a thorough analysis of performance metrics, which consistently demonstrate superior performance compared to other models. Furthermore, the AUC of the ROC curve is the highest, suggesting that the model is less susceptible to overfitting. Moreover, the minimal difference between the training and testing accuracy for logistic regression indicates that the model is well suited for the research and does not suffer from overfitting or underfitting.

Moreover, logistic regression is a probabilistic modeling approach that models the logarithmic odds of the dependent variable taking on one value versus the other, making it well-suited for dichotomous outcomes and allowing for nonlinear relationships between predictor variables and the outcome, ensuring that the predicted probabilities are always between 0 and 1. Logistic regression is easy to interpret and provides meaningful information on the relative importance of various predictor variables in predicting the outcome. Hence, these factors conclude that logistic regression is the most suitable model in this research.

During the whole process of this research. First, the number of features is comparatively lower than expected, which could affect the accuracy of the training and testing results. Additionally, strong connections between some of these features have been revealed which further affects the issue of accuracy. To resolve these constraints in future research, this study suggests further enlargement of

the dataset to include a wider range of universities and academic disciplines. Through this consideration, in addition to increasing the number of features, the diversity of the data will also be improved following the capability of the model to predict outcomes accurately in different scenarios.

## REFERENCE


Aina, C. *et al.* (2022) 'The determinants of university dropout: A review of the socio-economic literature', *Socio-Economic Planning Sciences*, 79, p. 101102. https://doi.org/10.1016/J.SEPS.2021.101102.

Albreiki, B., Zaki, N. and Alashwal, H. (2021) 'A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques', *Education Sciences 2021, Vol. 11, Page 552*, 11(9), p. 552. https://doi.org/10.3390/EDUCSCI11090552.

Bäulke, L., Grunschel, C. and Dresel, M. (2022) 'Student dropout at university: a phase-orientated view on quitting studies and changing majors', *European Journal of Psychology of Education*, 37(3), pp. 853–876. https://doi.org/10.1007/S10212-021-00557-X/TABLES/4.

Bhutto, S. *et al.* (2020) 'Predicting Students' Academic Performance Through Supervised Machine Learning', *ICISCT 2020 - 2nd International Conference on Information Science and Communication Technology* [Preprint]. https://doi.org/10.1109/ICISCT49550.2020.9080033.

Del Bonifro, F. *et al.* (2020) 'Student dropout prediction', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12163 LNAI, pp. 129–140. https://doi.org/10.1007/978-3-030-52237-7_11/TABLES/3.

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

Bro, R. *et al.* (2008) 'Cross-validation of component models: A critical look at current methods', *Analytical and Bioanalytical Chemistry*, 390(5), pp. 1241–1251. https://doi.org/10.1007/S00216-007-1790-1/FIGURES/8.

Casquero, O. *et al.* (2013) 'Students' personal networks in virtual and personal learning environments: a case study in higher education using learning analytics approach', 24(1), pp. 49–67. https://doi.org/10.1080/10494820.2013.817441.

Chen, R. (2012) 'Institutional Characteristics and College Student Dropout Risks: A Multilevel Event History Analysis', *Research in Higher Education*, 53(5), pp. 487–505. https://doi.org/10.1007/S11162-011-9241-4/TABLES/4.

Cunningham, P. and Delany, S.J. (2021) 'k-Nearest Neighbour Classifiers - A Tutorial', *ACM Computing Surveys (CSUR)*, 54(6). https://doi.org/10.1145/3459665.

Fan, J., Upadhye, S. and Worster, A. (2006) 'ORIGINAL RESEARCH • RECHERCHE ORIGINALE Understanding receiver operating characteristic (ROC) curves', *Can J Emerg Med*, 8(1), pp. 19–20. https://doi.org/10.1017/S1481803500013336.

Fidalgo-Blanco, Á. *et al.* (2015) 'Using Learning Analytics to improve teamwork assessment', *Computers in Human Behavior*, 47, pp. 149–156. https://doi.org/10.1016/J.CHB.2014.11.050.

Fletcher, S. and Islam, M.Z. (2019) 'Decision Tree Classification with Differential Privacy', *ACM Computing Surveys (CSUR)*, 52(4). https://doi.org/10.1145/3337064.

Kecman, V. (2005) *Support Vector Machines: Theory and Applications - Google Books*, *Springer*. https://books.google.com.my/books?hl=en&lr=&id=uTzMPJjVjsMC&oi=fnd&pg=PA1&dq=support+vector+machine&ots=GFHDas1Fo8&sig=edkKuDNVpe6RpMLEWaMqW2wP_HE&redir_esc=y#v=onepage&q=support vector machine&f=false (Accessed: 20 April 2023).

Kleinbaum, D.G. (1994) 'Logistic regression : a self-learning text', p. 282.

Niyogisubizo, J. *et al.* (2022) 'Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization', *Computers and Education: Artificial Intelligence*, 3, p. 100066. https://doi.org/10.1016/J.CAEAI.2022.100066.

Rastrollo-Guerrero, J.L., Gómez-Pulido, J.A. and Durán-Domínguez, A. (2020) 'Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review', *Applied Sciences 2020, Vol. 10, Page 1042*, 10(3), p. 1042. https://doi.org/10.3390/APP10031042.

Singh, H.P. and Alhulail, H.N. (2022) 'Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach', *IEEE Access*, 10, pp. 6470–6482. https://doi.org/10.1109/ACCESS.2022.3141992.

Song, Z. *et al.* (2023) 'All-Year Dropout Prediction Modeling and Analysis for University Students', *Applied Sciences 2023, Vol. 13, Page 1143*, 13(2), p. 1143. https://doi.org/10.3390/APP13021143.

Ting, T. T., Teh, S. L., & Wee, M. C. (2022). Ascertaining the Online Learning Behaviors and Formative Assessments Affecting Students' Academic Performance during the COVID-19 Pandemic: A Case Study of a Computer Science Course. *Education Sciences*, 12(12), 935. https://doi.org/10.3390/educsci12120935

Wild, S., Rahn, S. and Meyer, T. (2023) 'The relevance of basic psychological needs and subject interest as explanatory variables for student dropout in higher education — a German case study using the example of a cooperative education program', *European Journal of Psychology of Education*, pp. 1–18. https://doi.org/10.1007/S10212-022-00671-4/FIGURES/3.

Wild, S. and Schulze Heuling, L. (2020) 'Student dropout and retention: An event history analysis among students in cooperative higher education', *International Journal of Educational Research*, 104, p. 101687. https://doi.org/10.1016/J.IJER.2020.101687.

Willcoxson, L., Cotter, J. and Joy, S. (2011) 'Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities', 36(3), pp. 331–352. https://doi.org/10.1080/03075070903581533.

Yağcı, M. (2022) 'Educational data mining: prediction of students' academic performance using machine learning algorithms', *Smart Learning Environments*, 9(1), pp. 1–19. https://doi.org/10.1186/S40561-022-00192-Z/TABLES/14.

Aina, C. *et al.* (2022) 'The determinants of university dropout: A review of the socio-economic literature', *Socio-Economic Planning Sciences*, 79, p. 101102. https://doi.org/10.1016/J.SEPS.2021.101102.

Albreiki, B., Zaki, N. and Alashwal, H. (2021) 'A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques', *Education Sciences 2021, Vol. 11, Page 552*, 11(9), p. 552. https://doi.org/10.3390/EDUCSCI11090552.

Bäulke, L., Grunschel, C. and Dresel, M. (2022) 'Student dropout at university: a phase-orientated view on quitting studies and changing majors', *European Journal of Psychology of Education*, 37(3), pp. 853–876. https://doi.org/10.1007/S10212-021-00557-X/TABLES/4.

Bhutto, S. *et al.* (2020) 'Predicting Students' Academic Performance Through Supervised Machine Learning', *ICISCT 2020 - 2nd International Conference on Information Science and Communication Technology* [Preprint]. https://doi.org/10.1109/ICISCT49550.2020.9080033.

Del Bonifro, F. *et al.* (2020) 'Student dropout prediction', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12163 LNAI, pp. 129–140. https://doi.org/10.1007/978-3-030-52237-7_11/TABLES/3.

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

Bro, R. *et al.* (2008) 'Cross-validation of component models: A critical look at current methods', *Analytical and Bioanalytical Chemistry*, 390(5), pp. 1241–1251. https://doi.org/10.1007/S00216-007-1790-1/FIGURES/8.

Casquero, O. *et al.* (2013) 'Students' personal networks in virtual and personal learning environments: a case study in higher education using learning analytics approach', 24(1), pp. 49–67. https://doi.org/10.1080/10494820.2013.817441.

Chen, R. (2012) 'Institutional Characteristics and College Student Dropout Risks: A Multilevel Event History Analysis', *Research in Higher Education*, 53(5), pp. 487–505. https://doi.org/10.1007/S11162-011-9241-4/TABLES/4.

Cunningham, P. and Delany, S.J. (2021) 'k-Nearest Neighbour Classifiers - A Tutorial', *ACM Computing Surveys (CSUR)*, 54(6). https://doi.org/10.1145/3459665.

Fan, J., Upadhye, S. and Worster, A. (2006) 'ORIGINAL RESEARCH • RECHERCHE ORIGINALE Understanding receiver operating characteristic (ROC) curves', *Can J Emerg Med*, 8(1), pp. 19–20. https://doi.org/10.1017/S1481803500013336.

Fidalgo-Blanco, Á. *et al.* (2015) 'Using Learning Analytics to improve teamwork assessment', *Computers in Human Behavior*, 47, pp. 149–156. https://doi.org/10.1016/J.CHB.2014.11.050.

Fletcher, S. and Islam, M.Z. (2019) 'Decision Tree Classification with Differential Privacy', *ACM Computing Surveys (CSUR)*, 52(4). https://doi.org/10.1145/3337064.

Kecman, V. (2005) *Support Vector Machines: Theory and Applications - Google Books*, *Springer*. https://books.google.com.my/books?hl=en&lr=&id=uTzMPJjVjsMC&oi=fnd&pg=PA1&dq=support+vector+machine&ots=GFHDas1Fo8&sig=edkKuDNVpe6RpMLEWaMqW2wP_HE&redir_esc=y#v=onepage&q=support vector machine&f=false (Accessed: 20 April 2023).

Kleinbaum, D.G. (1994) 'Logistic regression : a self-learning text', p. 282.

Niyogisubizo, J. *et al.* (2022) 'Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization', *Computers and Education: Artificial Intelligence*, 3, p. 100066. https://doi.org/10.1016/J.CAEAI.2022.100066.

Rastrollo-Guerrero, J.L., Gómez-Pulido, J.A. and Durán-Domínguez, A. (2020) 'Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review', *Applied Sciences 2020, Vol. 10, Page 1042*, 10(3), p. 1042. https://doi.org/10.3390/APP10031042.

Singh, H.P. and Alhulail, H.N. (2022) 'Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach', *IEEE Access*, 10, pp. 6470–6482. https://doi.org/10.1109/ACCESS.2022.3141992.

Song, Z. *et al.* (2023) 'All-Year Dropout Prediction Modeling and Analysis for University Students', *Applied Sciences 2023, Vol. 13, Page 1143*, 13(2), p. 1143. https://doi.org/10.3390/APP13021143.

Wild, S., Rahn, S. and Meyer, T. (2023) 'The relevance of basic psychological needs and subject interest as explanatory variables for student dropout in higher education — a German case study using the example of a cooperative education program', *European Journal of Psychology of Education*, pp. 1–18. https://doi.org/10.1007/S10212-022-00671-4/FIGURES/3.

Wild, S. and Schulze Heuling, L. (2020) 'Student dropout and retention: An event history analysis among students in cooperative higher education', *International Journal of Educational Research*, 104, p. 101687. https://doi.org/10.1016/J.IJER.2020.101687.

Willcoxson, L., Cotter, J. and Joy, S. (2011) 'Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities', 36(3), pp. 331–352. https://doi.org/10.1080/03075070903581533.

Yi, A. C. Y., Ying, T. K., Yee, S. J., Chin, W. M., & Tin, T. T. (2022). 'InPath Forum: A Real-Time Learning Analytics and Performance Ranking Forum System.' *IEEE Access*, 10, pp. 128536-128542. https://doi.org/10.1109/ACCESS.2022.3227430

Yağcı, M. (2022) 'Educational data mining: prediction of students' academic performance using machine learning algorithms', *Smart Learning Environments*, 9(1), pp. 1–19. https://doi.org/10.1186/S40561-022-00192-Z/TABLES/14.