Using Predictive Analytics to Identify First-Year Engineering Students at Risk of Failing Engineering Physics

Low Beng Yew, Temasek Polytechnic, Singapore
Cha Cher Liang, Temasek Polytechnic, Singapore
Teoh Cheng Yong, Temasek Polytechnic, Singapore

The Southeast Asian Conference on Education 2021
Official Conference Proceedings

Abstract

Due to a lack of continual assessment or grade related data, identifying first-year engineering students in a polytechnic education at risk of failing Engineering Physics is challenging. Our experience over the years tells us that there is no strong correlation between having good entry grades in Mathematics and the Sciences and excelling in hard-core engineering subjects. Hence, identifying students at risk of failing cannot be on the basis of entry grades in Mathematics and the Sciences alone. These factors compound the difficulty of early identification and intervention. In this paper, we describe the development of a predictive analytics model in early detection of students at risk of failing and evaluates its effectiveness. Data from continual assessments conducted in term one, supplemented by data of student psychological profiles such as interests and study habits, were used. Three classification techniques, namely Logistic Regression, K Nearest Neighbour, and Random Forest, were used in our predictive model. Based on our findings, Random Forest was determined to be the strongest predictor with an Area Under the Curve (AUC) value of 0.994. Correspondingly, its Accuracy, Precision, Recall, and F-Score were also highest among these three classifiers. Using this Random Forest Classification technique, students at risk of failing could be identified at the end of term one. They could then be assigned to a Learning Support Programme at the beginning term two. This paper gathers the results of our findings. It also proposes further improvements that can be made to the model.

Keywords: Predictive Analytics, Random Forest, Students at risk, Early Intervention, Student Psychological Profile, Continual Assessment

Introduction

Engineering Physics is a first year engineering subject that is taught in the second semester in Temasek Polytechnic School of Engineering (TP ENG). It is a pre-requisite to most engineering studies. Poor grounding in this subject can compound learning difficulties in many related and subsequent subjects. Because of its importance, TP ENG has been offering a remedial service called Learning Support Programme (LSP) since 2010.

The main target audience of the LSP program is students from vocational schools instead of mainstream secondary schools in Singapore. This is because students from vocational schools are not taught Mathematics and the Sciences, and as such, have knowledge gaps in these subjects compare to students from mainstream secondary schools. The LSP program can also be extended to students from mainstream secondary schools who have weak foundation in these subjects, but there is always a delicate balance between teaching resources available and class size. As such, these precious limited vacancies should be allotted to mainstream students who are at risk of failing.

The current practice is to allocate these vacancies to mainstream students who have weaker entry grades in Mathematics and the Sciences. The assumption is that students with weaker entry grades need more teaching support. Surprisingly, our experience and observations showed that, for mainstream students, there is no strong correlation between having poor entry grades in Mathematics and the Sciences and failing Engineering Physics. A quick look up into the data of students who have failed their semester examination or declined in performance often show that they started off with fairly decent entry grades in Mathematics and the Sciences. Some students may have fallen by the wayside as early identification of such students and support for these students were not in place.

Hence, early intervention is imperative and in this paper, we examine the viability of using predictive analytics (Martin et al, 2019) as an early intervention device. Early intervention does improve academic success (Zhang et al, 2014). In our context, early intervention means being able to sieve out students are at risk of failing Engineering Physics in term one of the semester and then enrolling them in the LSP program in term two of the semester. In this way, at risk students would receive one semester term of additional help.

Since identifying mainstream students to be sent to the LSP classes on the basis of entry grades in Mathematics and the Sciences is not a good determinant, other factors have be considered. In our analysis, we gathered and examine factors related to student aptitude and psychological profiles such as interests and study habits.

Literature Review

Predictive analytics in higher education has evolved fairly recently as a result of the availability of more data set. Indeed, the development of educational learning tools and educational management system have created large databases which has enabled data mining (Calvet Liñán & Juan Pérez, 2015). Data mining has evolved from the classical regression analysis to present day machine learning. While data mining relies on human intervention and decision making, machine learning trains a computer using a set of existing data to predict future outcomes and hence the term, predictive modelling. Today, predictive modelling is used in discovering patterns of knowledge about educational phenomena and the learning process (Anoopkumar & Rahman, 2016). Predictive modelling has been also used in predicting educational outcomes, such as student performance (Hamoud et al, 2018), academic success (Martins et al, 2019; Richard-Eaglin, 2017), and dropout rate (Pérez et al, 2018).

These literatures reaffirm that using predictive analytics to sieve out student at risk is a viable approach. However, at the point of investigation, little or no empirical studies using predictive analytics in the context of Polytechnics in Singapore was done. In addition, most of the studies done were on students in their sophomore or senior years. As such, data related to continual assessments, grade point average and cumulative grade point average are readily available in their investigations. Our challenge is the lack of such data as our investigation was into first year students.

In determining students' academic success in higher education, prior academic achievements and student demographics were the top two factors quoted in 69% of the research papers (Alyahyan and Dustegor, 2020). However, our experience pointed away from a direct correlation between prior (entrance) academic achievements and student success. Besides, due to policies on data privacy, data of prior academic achievements, that is, student performance in their secondary schools, were not readily available and would consume much time to collect. We also do not suspect that demographic factors, such as financial background, play a key influence in academic success since the Ministry of Education in Singapore has a "No Child Left Behind" program instituted since 2012. As such, we were more interested in investigating other factors.

Instead, in our analysis, we considered students' aptitude, interests and study habits in the determination of their success in Polytechnic as more applicable. Key questions that we ask in these respects are the percentage of tutorial questions students are able to do unassisted, the amount of time a student spends self-studying each week, and the student's level of interest in Engineering Physics.

Methodology

A common technique that is generally used in the literatures to build a supervised learning predictive model is shown in Figure 1. The main stages are 1) factors, 2) data collection, 3) data pre-processing, 4) data mining, and 5) result evaluation.

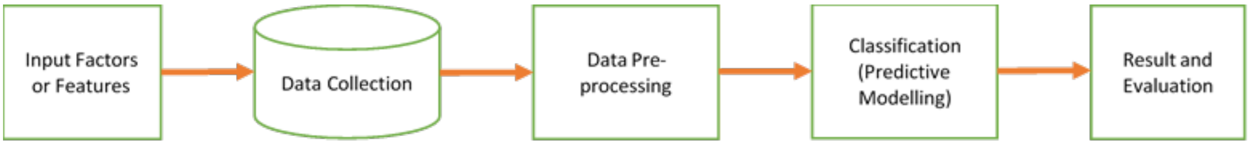

Input Factors or Features
Data Collection
Data Pre-processing
Classification (Predictive Modelling)
Result and Evaluation

Figure 1: Method Used in Prediction of Student at Risk

Stage 1: Input Factors or Features

Input factors, or features that were used are summarised in Table 1. These features fall broadly into two category. In the first category, the features were *tutorial attempt*, *like Physics*, *weekly self-study* and *prior knowledge*. These features are aptitude, attitudinal and psychological profile of student that focus on learning abilities, interests and study habits. The inclusion of *prior knowledge* as a feature is not to gather input of student's past academic achievements but to gather input of the relevancy of a past subject taken at the secondary school level. The second category of data are continual assessments marks that students have taken in the first term of the semester.

Input Factors / Features	Category	Explanation
Tutorial attempt	First	Captures input of percentage of questions student could answer unassisted
Like Physics	First	Captures whether student likes or dislikes Engineering Physics
Weekly Self-study	First	Captures the amount of time spent self-studying Engineering Physics each week
Prior Knowledge	First	Identifies if students has taken pure physics at "O" levels against other combinations
Assignment 1, Class Participation 1, Online Test 1, Term Test	Second	Continual assessment components from semester term 1
End-semester Examination	Target	Target which was converted into a derived column with either an "NR" or an "R" to indicate not-at-risk and at-risk

Table 1: Input Factors and Target Used in Our Predictive Analytics

Stage 2: Data Collection

The data for these first category features were information that needed to be collected directly from the students and were done through a student survey. The features from the second category were the outcome of assessment components and were collected from various marks and online assessment systems.

Stage 3: Data Pre-processing

A total of 200 students were surveyed, but due to incompleteness of data, some data were removed. Imputation of missing values was used where possible, without over making assumptions (Aleryani et al, 2018). The target used in training the model was the outcome of end-semester examination. End-semester examination marks were collected and a derived column indicating an "R" or an "NR" was introduced as the machine learning target. "R" indicates at-risk and "NR" indicates a not-at-risk. Students who failed the end-semester examination or had borderline passes were categorized as "R". By assigning students with C-grade and below into the "R" category, we could resolve the issue of data imbalance (Maheshwari, 2017). Indeed, students with borderline passing marks can be considered as students at risk.

We also made use of the linear projection and feature statistics widgets (see Figure 2) to ensure that data are not subjected to outliers that would affect the machine learning. Categorical data were converted to numerical data. After the data pre-processing and transformation, only 166 data points were used in this study.

Stage 4: Classification

Commonly used classification techniques are neural networks, K-nearest neighbor (kNN) and decision trees (Romero & Ventura, 2010). Their advantages and suitability were well discussed (Brohi et al, 2019). Our choice of classifiers used were random forest (a type of decision tree), logistic regression and kNN. We made use of a freeware called Orange to perform our analysis. The Orange tool provides built-in algorithms that simplify analysis. A diagram of our classification analysis workflow is shown in Figure 2.

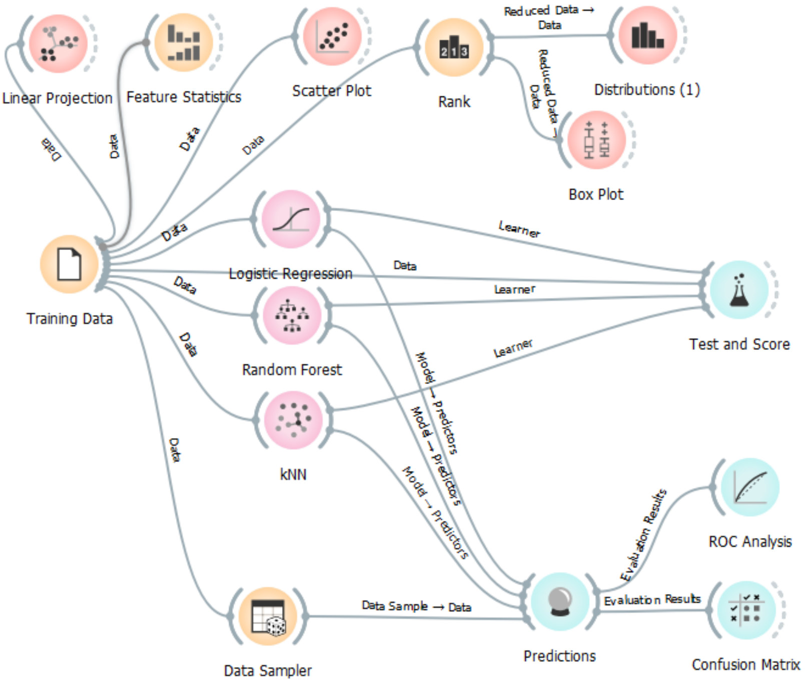


Figure 2: Workflow of Classification Analysis Using Orange

With reference to Figure 2, our collected data is used as the training data and is fed into the three classifiers to train these classifiers to identify the target outcome. To test the classifiers, we make use of the data sampler widget from the Orange tool. 35% of the data from the training data were sampled in a deterministic manner. Based on the sampled data input, the classifiers predict an “R” or an “NR”. The prediction outcome of each classifier is then evaluated by the predictions widget. Further evaluations are made using the receiver operating characteristic (ROC) analysis widget and confusion matrix widget.

The linear projection and feature statistics widgets are used to provide insights of the training data as part of data pre-processing. The scatter plot, rank, distributions, and box plot widgets were used in feature selection and evaluating the significance of the features.

Stage 5: Results and Evaluation

Standard performance evaluation parameters (Alyahyan and Dustegor, 2020) such as confusion matrix, classification accuracy (CA), precision, recall, F1 score, and area under the curve (AUC) were used to evaluate the performance of the classifier.

The confusion matrix is a table that displays the number of actual and predicted values. If the predicated outcome is the same as the actual, then we have a true positive (TP) or a true negative (TF). Otherwise, we have a false positive (FP) or a false negative (FN). This is illustrated in Table 2 below.

		Predicated	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Table 2: Confusion Matrix

The other performance evaluation parameters are defined as follows.

CA measures the proportion of predictions that are correct and is calculated as

$$CA = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the proportion of positive cases and is calculated as

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

Recall measures the proportion of positive cases that is correctly identified and is calculated as

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

F1 score conveys the balance between precision and recall and is calculated as

$$F1\ score = 2\times \frac{Precision\times Recall}{Precision + Recall}$$

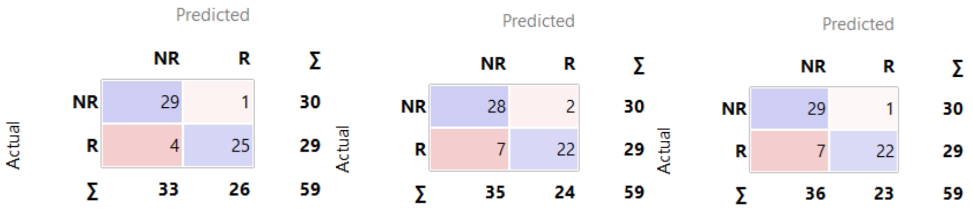
AUC is the area under the ROC curve and represents the probability of making a correct prediction.

All these parameters have values between 0 and 1 and are generally better when the values are closer or equal to 1.

Besides determining the classifier with the best performance, we also want to differentiate the features that are more significant in training the model. Scoring and ranking of features are performed using the rank widget. The performance indicators that we used were information gain and Gini gain. The information gain ratio determines the "pureness" of the information that is rendered by the feature towards identifying target. Information gain measures the amount of entropy or disorderliness that is removed. The higher the information gain, the more the entropy is removed. Likewise, the Gini gain determines the quality of the split between classes. The higher the Gini gain, the better the split.

Empirical Results and Evaluation

These performance evaluations are readily produced by the confusion matrix, predictions and ROC analysis widgets. The performance results of the classifiers are summarized in Tables 3 and 4 and Figure 3. These results show that random forest is the best performing classifier.



Predicted

Actual	NR	R	Σ
NR	29	1	30
R	4	25	29
Σ	33	26	59

(a) Random forest

Predicted

Actual	NR	R	Σ
NR	28	2	30
R	7	22	29
Σ	35	24	59

(b) Logistic regression

Predicted

Actual	NR	R	Σ
NR	29	1	30
R	7	22	29
Σ	36	23	59

(c) KNN

Table 3: Confusion Matrices of Classifiers

Classifiers	CA	Precision	Recall	F1	AUC
Random Forest	0.915	0.919	0.915	0.915	0.994
Logistic Regression	0.847	0.857	0.847	0.846	0.883
KNN	0.864	0.880	0.864	0.863	0.909

Table 4: Performance Evaluation of Classifiers (Generated by the Predication Widget)

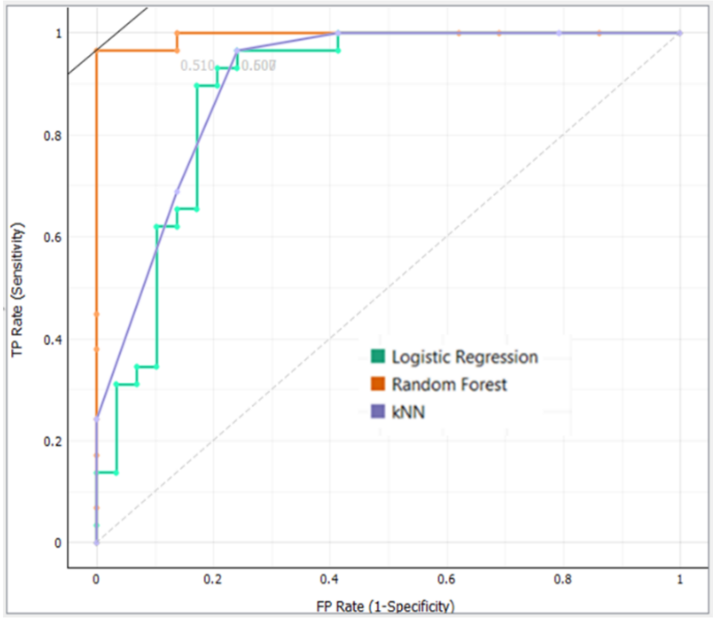
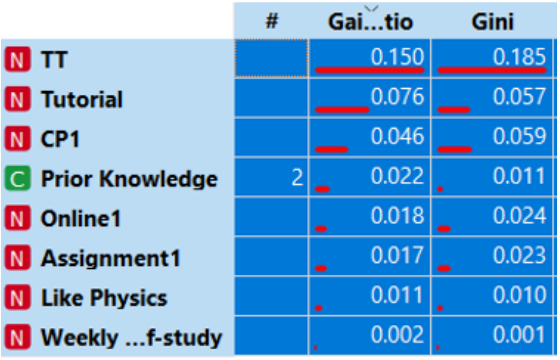


Figure 3: ROC Curves of Classifiers (AUC Values Can Be Found in Table 4)

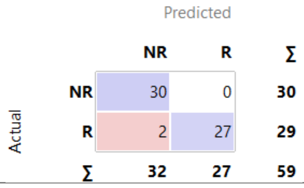
Using the rank widget, feature selection can be made and the significance of the features were ranked in descending orders as shown in Table 5.



	#	Gai...tio	Gini
N TT		0.150	0.185
N Tutorial		0.076	0.057
N CP1		0.046	0.059
C Prior Knowledge	2	0.022	0.011
N Online1		0.018	0.024
N Assignment1		0.017	0.023
N Like Physics		0.011	0.010
N Weekly ...f-study		0.002	0.001

Table 5: Scoring and Ranking of Features

The next stage of our evaluation was to use the outcome of feature selection to re-evaluate our model. The information gain ratio and Gini gain of Table 5 suggest that *like Physics* and *weekly self-study* are among the least significant features. Since information gain ratio and Gini gain are also part of the random forest algorithm, these features were turned off to evaluate the performance of the random forest classifier. The result is summarized in Table 6. It does show improvements in all performance categories. In particular, the new AUC is 0.999 compared to the previous values of 0.994 and the CA also improved from 0.915 to 0.966.



Predicted

Actual	NR	R	Σ
NR	30	0	30
R	2	27	29
Σ	32	27	59

Classifiers	CA	Precision	Recall	F1	AUC
Random Forest	0.966	0.968	0.966	0.966	0.999

Table 6: Performance of Random Forest After Feature Selection

Conclusions and Future Improvements

A predictive model was built to help identify first-year student at risk of failing Engineering Physics. This predictive model uses the limited data from continual assignment components from term one of the semester, and data collected from a student survey. The data from the student survey focuses on student psychological profiles such as aptitude, interests and study habits instead of student demographics. The end-semester examination was used as the target. A sample size of 166 data were used and three classifiers, namely, logistic regression, random forest and kNN were evaluated.

Random Forest was the best classifier and gave an AUC of 0.994 and a CA of 0.915. Feature scoring was used to rank the significance of the features. Two non-continual assessment related features stood out. *Prior knowledge* and *tutorial attempt* were significant features that help to train the classifiers. Two other non-continual assessment related features, namely *like Physics* and *weekly self-study,* were found to have low information gain ratio and Gini gain and were thus less significant. When *like Physics* and *weekly self-study* were turned off, the Random Forest showed an improved performance, with AUC of 0.999 and CA of 0.966. Using our predictive model, we could identify students at risk and take an early intervention actions such as assigning them to our LSP to receive additional tutoring help.

For future work, we can feed this trained model with data of the next student batch so that at-risk students of the next batch may be identified. We can then evaluate the outcome of our intervention by tracking students' performances in the end-semester examination and by comparing year-on-year end-semester examination performances. We were unable to perform this work in the October 2020 semester due to the Covid-19 situation, which resulted in a shift towards home-based learning and a change in assessment components.

In our next study, three additional features could also be incorporated into our model. The first is data of students' usage of Blackboard's Learning Management System. The second is tutor recommendation, which would be tutor's rating of students based on their attitude towards studies, and level of active engagement in class. The third is the non-medical attendance ratio, which disregards medical excuses and considers such cases as absent. These features, based on our teaching experience, apply very well in Singapore’s Polytechnic context and could enhance the performance of the next classifier.

Acknowledgements

The authors wish to express their gratitude to the Bioinformatics Lab at University of Ljubljana, Slovenia, for the use of Orange.

References

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied mathematical sciences, 9*(129), 6415–6426. doi:10.12988/ams.2015.53289

Aleryani, A., Wang, W., De, B., & Iglesia, L. (2018). Dealing with missing data and uncertainty in the context of data mining. *Hybrid artificial intelligent systems*, *10870*(24), 289-301. doi:10.1007/978-3-319-92639-1_24

Alyahyan, E., & Dustegor, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International journal of educational technology in higher education*, *17*(3). doi:10.1186/s41239-020-0177-7

Anoopkumar, M., & Rahman, A. M. J. M. Z. (2016). A review on data mining techniques and factors used in educational data mining to predict student amelioration. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 122–133.

Brohi, S.N., Pillai, T.R., Kaur S., Kaur, H., Sukumaran, S., & Asirvatham, D. (2019). Accuracy comparison of machine learning algorithms for predictive analytics in higher education. In Miraz M., Excell P., Ware A., Soomro S., & Ali M. (Eds), *Emerging technologies in computing*: *Proceedings of iCETiC 2019, 285*. doi:10.1007/978-3-030-23943-5_19

Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Educational data mining and learning analytics: Differences, similarities, and time evolution. *International journal of educational technology in higher education, 12*(3), 98-112. doi:10.7238/rusc.v12i3.2515

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., & Zupan, B. (2013) Orange: Data mining toolbox in python, *Journal of Machine Learning Research, 14*, 2349−2353.

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International journal of interactive multimedia and artificial intelligence, 5*(2), 26-31. doi:10.9781/ijimai.2018.02.004

Maheshwari, S., Jain, R. C., & Jadon, R. S. (2017). A review on class imbalance problem: Analysis and potential solutions. *International journal of computer science issues (IJCSI), 14*(6), 43-51.

Martins, M., Miguéis, V.L., Fonseca, D., & Alves, A. (2019). A data mining approach for predicting academic success – A case study. In Rocha Á., Ferrás C., & Paredes M. (Eds), *Information technology and systems: Proceedings of ICITS 2019, 918,* 45-56. doi:10.1007/978-3-030-11890-7_5

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International journal of modern education and computer science, 8*(11), 36–42.

Pérez, B., Castellanos, C., & Correal, D. (2018). Predicting student drop-out rates using data mining techniques: A case study. *2018 IEEE 1st Colombian conference on applications in computational intelligence (ColCACI)*, 1-6.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE transactions on systems, man, and cybernetics, part c (applications and reviews), 40*(6), 601–618.

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.

Zhang, Y., Fei, Q., Quddus, M., & Davis, C. (2014). An examination of the impact of early intervention on learning outcomes of at-risk students. *Research in higher education journal, 26*, 57-70.

Contact email: Low_Beng_Yew@tp.edu.sg
Cha_Cher_Liang@tp.edu.sg
Max_Teoh@tp.edu.sg