ISSN 2278-3091
Volume 10, No.6, November - December 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse041062021.pdf
https://doi.org/10.30534/ijatcse/2021/041062021

A Machine Learning Approach to Detect Student Dropout at University

Shiful Islam Shohag[1], Masum Bakaul[2]
[1]Department of Computer Science & Engineering, Britannia University, Cumilla, Bangladesh, shifulshohagbd@gmail.com
[2]Lecturer, Dept. of Computer Science & Engineering, Britannia University, Cumilla, Bangladesh, masumbakaul.cse@gmail.com

Received Date : October 08, 2021 Accepted Date : November 08, 2021 Published Date : December 06, 2021

ABSTRACT

In universities, student dropout is a major concern that reflects the university's quality. Some characteristics cause students to drop out of university. A high dropout rate of students affects the university's reputation and the student's careers in the future. Therefore, there's a requirement for student dropout analysis to enhance academic plan and management to scale back student's drop out from the university also on enhancing the standard of the upper education system. The machine learning technique provides powerful methods for the analysis and therefore the prediction of the dropout. This study uses a dataset from a university representative to develop a model for predicting student dropout. In this work, machine-learning models were used to detect dropout rates. Machine learning is being more widely used in the field of knowledge mining diagnostics. Following an examination of certain studies, we observed that dropout detection may be done using several methods. We've even used five dropout detection models. These models are Decision tree, Naïve bayes, Random Forest Classifier, SVM and KNN. We used machine-learning technology to analyze the data, and we discovered that the Random Forest classifier is highly promising for predicting dropout rates, with a training accuracy of 94% and a testing accuracy of 86%.

Key words: Data Mining, Dropout assessment (detection, classification), University Student Dropout, Dropout Prediction.

1. INTRODUCTION

The Dropout at university is the most concern in the world. It is considered to be one of the most educational problems in the world. However, there is not an effective way to find dropouts yet, the key to reducing the rate of dropouts in the early detection and analysis of dropouts. Accurate detection of dropout normally requires analysis of university student data of different modalities. An automated method is desperately needed.

This issue is not only affecting the academic field but also influences the image of the country. Student retention, particularly in higher education, is a difficult endeavor that indicates the institution's efficiency and dependability. Finding hidden patterns or prediction trends in a vast database helps to improve the quality of management decision-making which can allocate resources appropriately with a better understanding of the student learning environment. The ability to predict student dropout with high accuracy is advantageous since it aids in identifying students who are at risk of poor academic performance. Data mining has been shown a successful benefit in the business domain and it can be a suitable tool to benefit in the educational domain for finding useful information hidden in the huge dataset. The classification method constructs a model based on the training set of known class labels data to classify unknown objects [1].

The primary goal of this study is to analyze student dropouts using longitudinal data. An advantage of our data in comparison to previous international longitudinal studies is that it covers dropout rates across the study period [2]. CS students, like students in other disciplines, drop out before graduating. So, it is necessary to track down the original reasons behind the dropout of CS graduates. From various studies, it is clear that mainly, two types of factors are responsible for these dropouts. One is personal another is institutional [3].Machine learning is a viable approach for developing a predictive model for an early warning system for dropouts. However, one of the possible challenges in developing a dropout early warning system based on machine learning is class imbalance[4].The purpose of this research was to enhance the performance of a dropout early warning system and to forecast dropout. There is still no agreement in the literature on the reasons of university dropout Despite, the availability of several studies on university dropout, there is a little study in the field of computer science. Furthermore, the great majority of research focuses on static factors, ignoring the dynamic component of the grades obtained by the student during his studies. It is critical to study the factors that drive students in the Systems Engineering program to drop out of the course before it is completed[5].

However, considering the factors that influence the dropout in universities the present study aims to answer the following questions. (a) Which attributes are the key determiners of student dropout and (b) Which machine-learning model is more suitable to find these key determinants?

We have three separate data modules in this study: Student Progress Data, Student Financial Data, and Personal Information Data. At first, we mapping the Data according to StudentId. After merge the three datasets we have 3562 records with 28 attributes. The application of machine learning in the field of data mining diagnosis is increasing gradually. After reviewed some papers, we found that several techniques were used for dropout detection. After reviewing these papers, we have selected five models for detecting dropouts. These models are Decision tree, Naïve bayes, Random Forest Classifier, SVM and KNN. We applied machine learning technology to the dataset and we have seen that is very beneficial for dropout rate with training accuracy 94% and testing accuracies 86% of Dropout prediction.

This paper is structured into five sections, which are as follows: Section.2 Review of the literature. Section. 3 Methodology. Section. 4 Result and Discussion. Finally, Section. 5 Conclusion and Future Work.

2. LITERATURE REVIEW

In the recent past, various studies have been carried out in connection to study the students' academic achievement, dropout, and organizational performance using the application of data mining methods by the number of authors in the area of traditional and professional education particularly in information science, management and social sciences [6].

In 2020 Del Bonifro et al (2020), published the paper "Student dropout prediction" where the author explains running a series of exams depending on a student's credits after a certain period. As one might anticipate, this contributes significantly to the model's overall performance. This fact can be used by the institution to decide whether to act as early as possible, based on the information available at enrolment time, or to wait for More data will be collected in the first year, allowing for more precise projections. In any event, the results suggest that, starting with data that has no educational or didactic value, our technology may practically aid in the effort to reduce dropout rates[7].There is strong evidence that university dropout prediction is of considerable interest among academic investigators, and that high precision algorithms are being developed to solve this critical issue[8].

Educational research has taken advantage of data mining. The present rate of use of data mining approaches in this sector has accelerated for a number of goals, including assessing student needs, forecasting dropout rates, analyzing and enhancing student academic performance. Student dropout prediction is an important and challenging task.

Iam-On et al (2017) says as suggested by many kinds of research works on the subject of student dropout, family background, financial support, and university-event participation may provide a complementary interpretation of student achievement. Academic aptitude is a key obstacle to success for some pupils, while social and socioeconomic considerations might be critical for others. Unfortunately, these characteristics may not be accurately recorded, making the related study difficult. However, with the aforementioned variables incorporated, a better knowledge of non-academic incentives for student performance may be obtained through collaboration with relevant divisions [9]. The scientific community is interested in university dropout prediction, as indicated by the vast amount of publications on the issue and its socioeconomic consequences. To address the issue of dropout, very accurate approaches are being created; nevertheless, we cannot identify a superior methodology since prediction accuracy is mostly determined by context, data, and technique features; any prospective alternative must take these aspects into consideration[10].

Mardolkar, Mahesh et al (2020), states the problem we aim to tackle is predicting student performance and which students will drop out with reasonable accuracy. The various features of the students are defined and categorized into two as student welfare feature and student performance feature Student welfare feature defines parent involvement, medium of instructions, qualified parent, earning members in a family, annual income of parents, time spent with friends, and playing, working in the family shop, doing a part-time job, liberty is given and financial assistance. Student performance feature defines previous examination performance, everyday reading and writing activity, academic pressure, need for extra classes, and re-examination performance [11].

Tsai, Shuo Chang et al (2020) were recommended to include other characteristics linked to students' involvement, family, and learning behavior as variables to concurrently enhance the accuracy and sensitivity of prediction models to 80% or higher, which is also a target of our future study. There were several drawbacks to their investigation, as mentioned by Lee, Sunbok et al (2019). First, in our investigation, we have restricted access to the NEIS database. Although we included the major risk factors for dropout prediction in our study, we were unable to access many other characteristics in the NEIS database at the time of our analysis, such as instructors' evaluations of students.

In this research, we try to find the key determiners of dropout related to the above limitations with many restrictions. We selected a dataset that included different student characteristics such as financial information, personal information, academic progress information, transfer intent, or not information. There are 28 different aspects of a student's information.

3. METHODOLOGY

3.1 Data preparation

This study analyzed student behavior data for the 2015–2016 school years to extract from a reputed university's institutional research database, these data included student financial information (e.g., ParentGrossIncome, GrossIncome, Father and Mother Education, Scholarship, Marital status Housing, Loan, Grant status, etc.), student progress at school in the first year (e.g., major course complete or not, different course score,

Term GPA, Cum GPA records, etc.), and study status (e.g., whether a student has dropped or not corresponding to student ID). In this research, we have three different data modules: Student Progress Data, Student Financial, and Personal Information data. At first, we mapping the Data according to StudentId. After marge, cleaning, and handling the missing value of three datasets we have finally 1896 records without null values with 28 Attributes these are the following 1. Student Financial & personal data (StudentID, MaritalStatus, AdjustedGrossIncome, ParentAdjustedGrossIncome, FathersHighestEducation, Mothers HighestEducation, Housing, Loan, Scholarship, PertimeJob/study, Grant), 2. Student Progress Data (Cohort, CohortTerm, Term, Academic Year, CompleteDevMath, CompleteDevEnglish, Major1, Major2, Complete1, Complete2, CompleteCIP1, CompleteCIP2, TransferIntent, DegreeTypeSought, TermGPA, CumGPA), 3. Label Data (Dropout (0 or 1)).

3.2 Encoding Method

In this dataset, we have some Categorical values which are MaritalStatus, FathersHighestEducation, MotherHighestEducation, and Housing. MaritalStatus contains four categories married, single, divorced, and not available. FathersHighestEducation contains School college and not available, MotherHighestEducation also contains this. The housing feature contains On-Campus Housing, Off-Campus, and WithParent. Machine learning algorithm Works with numbers not string. The above four attributes contain the object. So, before we selecting a machine learning model, we need to transform this attribute to numeric. To transform object to numeric we used an encoding technique from the scikit learn library. There is various technique to transform object to numeric we used Label Encoder based on our data.

3.3 Feature Selection

The data attributes we utilize to train our machine learning models have a significant impact on the results we can accomplish. Model performance can be harmed by irrelevant or partially relevant features. Feature selection is a procedure in which we automatically choose those characteristics in our data that contribute the most to the prediction variable or output of interest. Having irrelevant features in our data can decrease the accuracy of many models, especially linear algorithms. Here we use Univariate Selection in our research. Statistical tests can be used to identify the attributes with the strongest relation to the output variable. The scikit-learn package includes the SelectKBest class, which may be used in conjunction with a variety of statistical tests to choose a given number of features. Many different statistical test scans are used with this selection method. This can be used via the f_classif () function [12]. We select the 15 best features using this method. Table **Error! Reference source not found.**.

Table 1: Feature Score of best 15 feature

Selecting index	Feature index	Feature Name	Score
0	17	TermGPA	241.52
1	18	CumGPA	206.68
2	10	ChortTerm	59.56
3	7	Scholarship	31.66
4	0	MaritalStatus	22.99
5	13	Major1	19.56
6	2	ParentAdjustGrossIncome	17.61
7	4	MotherHightEducation	13.00
8	3	FatherHightEducation	11.92
9	9	Grant	11.81
10	6	Loan	10.81
11	15	Complete1	6.56
12	16	CompleteCIP1	6.56
13	12	CompleteDevEnglish	4.50
14	11	CompleteDevMath	4.30

3.4 Cross Validation

Validation via Cross-Checking We couldn't fit the model on the training data in machine learning, thus we can't assure the model will operate properly on real data. To do so, we must ensure that our model extracted the relevant patterns from the data and that it is not generating too much noise. We employ the cross-validation approach for this purpose. Cross-validation is a method that allows us to train our model using a portion of the data set and subsequently assess it using the complementary portion of the data set.The generated models were validated using cross-validation with ten folds [13]. In this method, we perform training on 75% of our data-set, and the rest 25% is used for testing purposes. The primary downside to this approach is that we only train on 25% of the dataset; it is likely that the remaining 50% of the data contains critical information that we are missing while training our model, resulting in larger bias.We divide the data set into k subsets (known as folds), then train on all of them while leaving one (k-10) subset for evaluating the trained model.In this approach, we iterate k times, each time with a distinct subset designated for testing.

3.5 Method

Using machine learning models, we've been able to detect dropout rates accurately, the use of machine learning in data mining diagnostics is growing. An examination of certain studies revealed that dropout detection relies on several different ways It took us some time to sift through all of these papers and select five models as an example, there's the Decision Tree model, Random Forest, KNN, and the Naive Bayes classifier as well as SVM. Data cleaning and preparation processes were described in machine learning basics. For example, the scikit learn library's Label Encoder is used for data preparation, while pandas are used for data cleaning.

4. RESULT AND DISCUSSION

The five distinct machine learning models that we utilized in this work were: As a result, we evaluate the model's accuracy, precision, recall, and efficiency. A detailed comparison is

shown below. Table 2 Training and testing Accuracy shows that Decision Tree training accuracy is 80%, Nave Bayes training accuracy is 71%, the SVM training accuracy is 83 %, KNN training accuracy is 85%, and Random Forest training accuracy is 86%. From these comparisons, we may conclude that Random Forest performs better in training. Also, the accuracy of testing is the best in the Random Forest classification Model.

Table 2: Training and testing Accuracy

Model Name	Training Accuracy	Model Name	Testing Accuracy
Decision Tree	80%	Decision Tree	80%
Naïve Bayes	74%	Naïve Bayes	71%
SVM	83%	SVM	84%
KNN	85%	KNN	84%
Random Forest	94%	Random Forest	86%

In this research, we had gone through the specifics of five models. 75% of data in our study assignment is for training, and 25 % of data is for testing. Even though training has 1422 records and relaxing has 474 records, it's clear that both are important. From table (**Error! Reference source not found.**) For Decision Tree model number of TP is 336, FN is 53 and FP is 44, TN is 41. For the Naïve Bayes model number of TP is 301, FN is 88 and FP is 48, TN is 37. For the SVM model number of TP is 380, FN is 9 and FP is 67, TN is 18. For the KNN model number of TP is 379, FN is 10 and FP is 63, TN is 22. And for the Random Forest model number of TP is 372, FN is 17 and FP is 48, TN is 37. Above the compares we can see that Random Forest is the best model for predicting dropout. See the table below (**Error! Reference source not found.**).

Table 3: Testing details five models

Decision Tree	**TP**	336	**FN**	53
	FP	44	**TN**	41
Naïve Bayes	**TP**	301	**FN**	88
	FP	48	**TN**	37
SVM	**TP**	380	**FN**	9
	FP	67	**TN**	18
KNN	**TP**	379	**FN**	10
	FP	63	**TN**	22
Random Forest	**TP**	372	**FN**	17
	FP	48	**TN**	37

Accuracy	ACC = (TP + TN) / (P + N)
Sensitivity or Recall	TPR = TP / (TP + FN)
Precision	PPV = TP / (TP + FP)
F1 Score	F1 = 2TP / (2TP + FP + FN)

To measure the efficiency that each technique demonstrates incorrectly identifying dropout students, the sensitivity criterion is used. This criterion measures the proportion of students that were correctly identified by a technique as dropouts, versus the total number of actual dropout students [14].

Decision Tree precision, recall, and f-1 score of NOT Dropout are 88%, 46%, and 87%, respectively, and precision, recall, and f-1 score of Dropout are 44%, 48%, and 46%. For Naïve Bayes, precision, recall, the f-1 score of NOT Dropout respectively 86%, 77%, and 82% and precision, recall, the f-1 score of Dropout is respectively 30%, 44%, and 35%. For SVM, precision, recall, the f-1 score of NOT Dropout respectively 85%, 98%, and 91% and precision, recall, the f-1 score of Dropout is respectively 67%, 21%, and 32%. For KNN, precision, recall, the f-1 score of NOT Dropout respectively 86%, 97%, and 91% and precision, recall, the f-1 score of Dropout is respectively 69%, 26%, and 38%. For Random Forest, precision, recall, the f-1 score of NOT Dropout respectively 89%,95% and 92% and precision, recall, the f-1 score of Dropout is respectively 66%, 45%, and 53%. For best estimation table-4 shows the comparison.

Table 4: Comparison of all classifier

Classifier		**Precision**	**recall**	**F-1 score**	**support**
Decision Tree	Not Dropout	88%	86%	87%	389
	Dropout	44%	48%	46%	85
	Accuracy			80%	474
	Macro Average	66%	67%	67%	474
	Weighted Average	80%	80%	80%	474
Naive Bayes	Not Dropout	86%	77%	82%	389
	Dropout	30%	44%	35%	85
	Accuracy			80%	474
	Macro Average	58%	60%	58%	474
	Weighted Average	76%	71%	73%	474
SVM	Not Dropout	85%	98%	91%	389
	Dropout	67%	21%	32%	85
	Accuracy			84%	474
	Macro Average	76%	59%	62%	474

	Weighted Average	82%	84%	80%	474
KNN	Not Dropout	86%	97%	91%	389
	Dropout	69%	26%	38%	85
	Accuracy			85%	474
	Macro Average	77%	62%	64%	474
	Weighted Average	83%	85%	82%	474
Random Forest	Not Dropout	89%	95%	92%	389
	Dropout	66%	45%	53%	85
	Accuracy			86%	474
	Macro Average	77%	70%	72%	474
	Weighted Average	85%	86%	85%	474

In the above Testing accuracy of the Decision Tree is 80%, Naïve Bayes is 71%, SVM is 84%, Random Forest is 86% and KKN is 85%. Accuracy of testing is also highest for the Random Forest classification Model. Now we see the ROC accuracy Score of these five models in (**Error! Reference source not found.**).

Table 5: ROC Accuracy Score

Model Name	ROC Accuracy score
Decision Tree	0.516
Naïve Bayes	0.736
SVM	0.788
KNN	0.746
Random Forest	0.782

Now see the ROC accuracy curve in (**Error! Reference source not found.**) for best estimation.

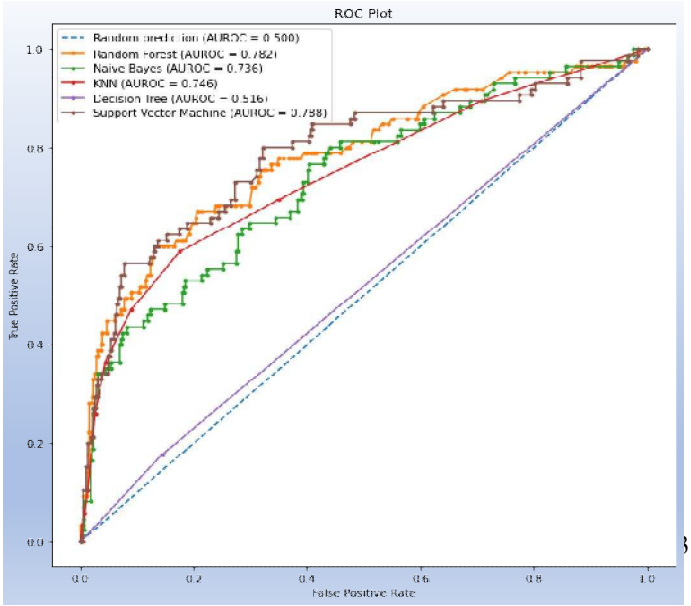
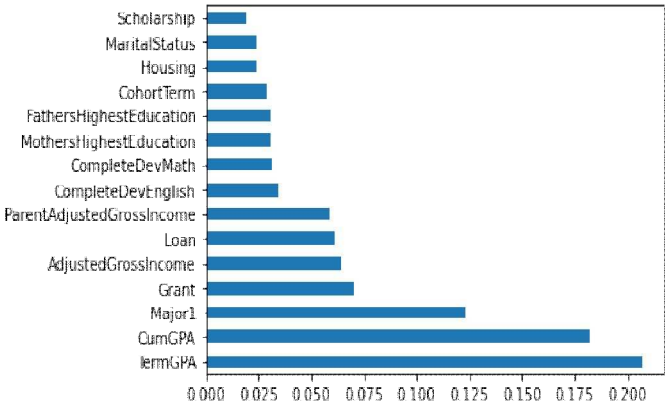


Figure 1: ROC Accuracy Curve

Random Forest Classifier is the model which we recommended. There are methods for estimating the significance of characteristics such as Random Forest and Extra Trees Given our dataset, we create an ExtraTreeClassifier. Prediction dropout characteristics are important as may be shown in the (Feature Selection for Machine Learning in Python, n.d.). As you can see, each attribute is assigned a value based on its relevance. The higher the score, the more essential as a result of the scores, we may determine the relative significance of different aspects Then, you can see (Figure 2 Feature Importance for Random Forest). In figure 2 we see that TermGPA and CumGPA are most significant for dropout.

Now we see the Precision-Recall curve for our recommended model. Precision and recall can be calculated in scikit-learn.



The precision and recall can be calculated for thresholds using

Figure 2: Feature Importance for Random Forest

the precision_recall_curve () function that takes the true output values and the probabilities for the positive class as input and returns the precision, recall, and threshold values [15]. Here AP is 0.58. See the (**Error! Reference source not**

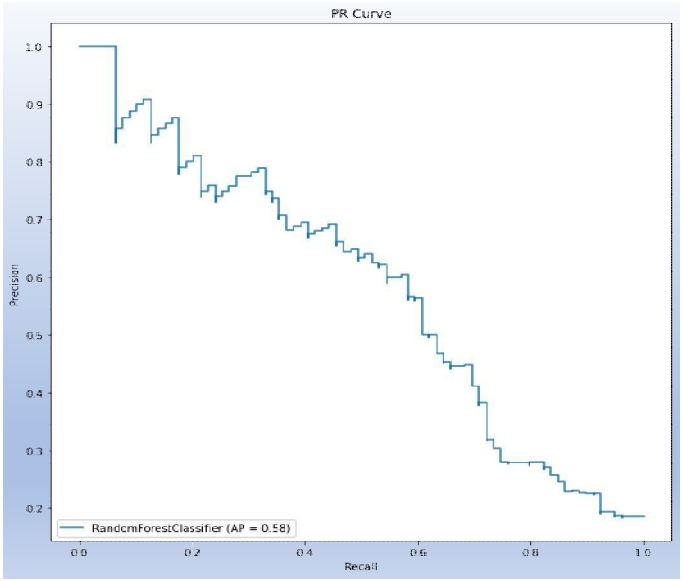


Figure 3: PR Curve

found.).

The accuracy of our recommended model is seen. Our selected model correctly predicts dropout 86 % of the time. We identify the top 15 Attributes for our recommended model based on our hypothesis Q-(a). The best predictors of dropout, in this case, are GPA characteristics, personality traits, and economical attributes. Sufficient experiments are developed and implemented from diverse viewpoints [16]. Students go away from their families all over the world to begin university studies, and it is a period of major transition and adjustment in which students must cope with new situations, loneliness, the creation of new friendships, and involvement in independent learning. Helping these students to access family support and strengthening the social support available to compensate for the absence of family support should help increase retention and decrease dropout during this critical period [2]. These factors were classified into five dimensions personal, academic, economic, social, and institutional; the most commonly studied was the personal dimension, which considers factors such as age, ethnicity, and gender [10]. From our investigation, we found that the main attributes influencing student dropout are mostly concerning the academic aspect, specifically the term they are studying in and a low average grade. In addition to a low average grade, the method of entering the university and the major of the students also affect the student dropout. These findings are consistent with previous findings by Pereira and Zambrano, who found that the most common causes of student dropout from university include academic characteristics such as a low-grade point average, the semester of the program, and the faculty to which the student belongs. Moreover, the location of the high school that students graduated from also related to the students' dropout. Similar was stated by Rahman and Dash who found that the location of student's residence either rural or urban is related with the discipline they chose to study [1]. The experimental findings validated our algorithm's effectiveness and universality [16]. We are also looking for technological enhancements to our prediction models [17].

5. CONCLUSION

Our study is primarily focused on reducing the dropout rate. As a result, there is an 86 % accuracy rate for machine learning in this case as a result of this model's efficient utilization, there will be as opposed to taking a long time for an institution of higher learning to find out, a machine learning model can predict in a matter.

Firstly, machines can work much faster than humans. A computer can do thousands of operations in a matter of seconds. Machines can accomplish things that people aren't very good at. They are capable of repeating themselves hundreds of times without tiring. The machine repeats the procedure after each iteration in order to improve it. Another benefit is the high precision of machinery. With the introduction of Internet of Things technology, there is now so much data in the globe that people cannot possibly sift through it all. That is when machines come in handy. They can complete tasks quicker than we can, do precise computations, and identify patterns in data. The accuracy of the machine is higher than an analyst. Secondly, this paper also reduces the cost of detection of dropout which is very needed in developing countries. This Project is highly performed at dropout detection and reduces the dropout of the student. And it also finds out the probability of a student dropout using various parameters. Experimental results on a large-scale public dataset show that the proposed model can achieve comparable performance to approaches relying on feature engineering performed by experts [18].This study will also work to identify those students which needed special attention to reducing the drop-out rate [6]. The best machine learning tool in the university student management system is like a professional analyst. This system may be a substitute or assistant to an analyst. This prediction by machine learning is using the large scale in university education. This system is used to predict dropouts. Assist university student dropout analysts with high-level accuracy analysis. Fully automated dropout detection. finds out the probability of a student dropout using the various parameter.

This is an important model for predict Dropout by machine learning. There needs far better technology to predict dropout. We want to build a model that can predict dropout accurately from various features of a student which paper will contribute to reducing dropout of a student in university. We also want to extend our research beyond the first term's transcript data and take a more in-depth look at attrition mechanisms [17].

REFERENCES

1. J. Pattanaphanchai, K. Leelertpanyakul, and N. Theppalak, "The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining: A Case Study of Faculty of Science, Prince of Songkla Uni-versity," *J. Univ. BABYLON Pure Appl. Sci.*, vol. 27, no. 1, pp. 356–367, 2019, doi: 10.29196/jubpas.v27i1.2191.
2. E. M. Sosu and P. Pheunpha, "Trajectory of University Dropout: Investigating the Cumulative Effect of Academic Vulnerability and Proximity to Family Support," *Front. Educ.*, vol. 4, no. February, pp. 1–10, 2019, doi: 10.3389/feduc.2019.00006.
3. S. A. Ahmed, M. A. Billah, and S. I. Khan, "A Machine Learning Approach to Performance and Dropout prediction in Computer Science: Bangladesh Perspective," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, no. October, 2020, doi: 10.1109/ICCCNT49239.2020.9225356.
4. S. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Appl. Sci. Switz.*, vol. 9, no. 15, 2019, doi: 10.3390/app9153093.
5. B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," *Commun. Comput. Inf. Sci.*, vol. 833, no.

April 2019, pp. 111–125, 2018, doi: 10.1007/978-3-030-03023-0_10.

6. S. Rai and A. Kumar Jain, “Students' Dropout Risk Assessment in Undergraduate Courses of ICT at Residential University A Case study,” *Int. J. Comput. Appl.*, vol. 84, no. 14, pp. 31–36, 2013, doi: 10.5120/14645-2965.
7. F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, *Student dropout prediction*, vol. 12163 LNAI. Springer International Publishing, 2020. doi: 10.1007/978-3-030-52237-7_11.
8. F. Agrusti, G. Bonavolontà, and M. Mezzini, “University dropout prediction through educational data mining techniques: A systematic review,” *J. E-Learn. Knowl. Soc.*, vol. 15, no. 3, pp. 161–182, 2019, doi: 10.20368/1971-8829/1135017.
9. N. Iam-On and T. Boongoen, “Generating descriptive model for student dropout: a review of clustering approach,” *Hum.-Centric Comput. Inf. Sci.*, vol. 7, no. 1, pp. 1–24, 2017, doi: 10.1186/s13673-016-0083-0.
10. M. Alban and D. Mauricio, “Predicting University Dropout trough Data Mining: A systematic Literature,” *Indian J. Sci. Technol.*, vol. 12, no. 4, pp. 1–12, 2019, doi: 10.17485/ijst/2019/v12i4/139729.
11. M. Mardolkar and N. Kumaran, “Forecasting and Avoiding Student Dropout Using the K-Nearest Neighbor Approach,” *SN Comput. Sci.*, vol. 1, no. 2, pp. 1–8, 2020, doi: 10.1007/s42979-020-0102-0.
12. “Feature Selection For Machine Learning in Python.” https://machinelearningmastery.com/feature-selection-machine-learning-python/ (accessed Jun. 21, 2021).
13. N. E. Rodríguez-Maya, C. Lara-Álvarez, O. May-Tzuc, and B. A. Suárez-Carranza, “Modeling Students’ Dropout in Mexican Universities,” *Res. Comput. Sci.*, vol. 139, no. 1, pp. 163–175, 2017, doi: 10.13053/rcs-139-1-13.
14. I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, “Dropout prediction in e-learning courses through the combination of machine learning techniques,” *Comput. Educ.*, vol. 53, no. 3, pp. 950–965, 2009, doi: 10.1016/j.compedu.2009.05.010.
15. T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS ONE*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/journal.pone.0118432.
16. J. Chen, J. Feng, X. Sun, N. Wu, Z. Yang, and S. Chen, “MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine,” *Math. Probl. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/8404653.
17. L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, “Predicting Student Dropout in Higher Education,” 2016.
18. W. Wang, H. Yu, and C. Miao, “Deep model for dropout prediction in MOOCs,” *ACM Int. Conf. Proceeding Ser.*, vol. Part F1306, pp. 26–32, 2017, doi: 10.1145/3126973.3126990.