Towards Predicting Student's Dropout in Higher Education Using Supervised Machine Learning Techniques

Raven Miguel M. Capuno, Chester Juliane M. Ferrer, Brian Troy L. Manaloto, Sean Redd Villafria
College of Computer Studies
Angeles University Foundation
Angeles City, Philippines
capuno.ravenmiguel@auf.edu.ph, ferrer.chesterjuliane@auf.edu.ph, ,
manaloto.briantroy@auf.edu.ph, villafria.seanredd@auf.edu.ph

James A. Esquivel
College of Computer Studies, Faculty
Angeles University Foundation
Angeles City, Philippines
esquivel.james@auf.edu.ph

Abstract

This paper is conducted to develop a system that can aid universities in tracking students' academic progress by accurately determining dropout rates using machine learning techniques. This has been done by collecting and organizing data into machine readable forms, using predictive methods that determines significant patterns and trends, constructing different models involving machine learning techniques: Random Forest, Logistic Regression, and Neural Network, for forecasting and predictions, and finally interpreting the statistical results of the model. After the evaluation, it has been concluded that the Random Forest achieved the highest accuracy of 95.00%, precision of 97.78%, and F1 score of 96.70%. Despite these results, the researchers recommend developing a more interesting and engaging prototype GUI and utilizing a more balanced and diverse dataset that will enable higher accuracy and deeper understanding of the results.

Keywords
Student Dropout, Machine Learning Technique, Algorithms, and Predicting Student Dropout

1. Introduction

Education is one of the most important things that all or any students of the coming generation must have. The education of students is crucial to the country's overall progress. However, numerous individuals are considered capable of dropping out of school for a variety of reasons. Several of the possible factors include family issues, financial difficulties, and numerous more. There is a desire to predict the number of students who will drop out of school due to a variety of factors. This prediction would be beneficial in assisting students in comprehending the precise reasons for their call in, as well as assisting other institutions to produce options in determining what could be done to eradicate that. (Selvan et al. 2019).

For other students, dropping out is the culmination of years of academic obstacles, blunders, and missed opportunities. For others, dropping out is a result of conflicting life circumstances, such as the necessity to financially support their family or the responsibilities of caring for siblings or their own child. Sometimes, students drop out because they are bored and find no link between school and the "real world." It's about the youth feeling detached from their friends, instructors, and other people. Insufficient resources in schools and communities to satisfy the diverse emotional and mental requirements of their child. Given the differences of reasons for dropping out, the outcomes and repercussions are remarkably consistent.

University dropout is one of the most difficult and unfavorable occurrences, both for students and the institution, among the many distinct observable phenomena in students' careers. A dropout is a potentially devastating event in a student's life, and it also has a severe economic impact on the college or rather to the University. Furthermore, it might be an indication of potential problems affecting course quality and the economy.

Machine learning approaches are one of the most sought-after options for solving the issue of school dropout. Numerous studies on building student prediction algorithms have been undertaken in industrialized countries. Additionally, there is a sizable body of literature on approaches to dropout prevention that are based on machine learning (Sales et al., 2016; Lakkaraju et al. 2015; Ameri et al. 2016).

1.1 Objectives

The goal of this research is to develop a program for universities that can help in monitoring academic performances of students and identify the probability of dropping out.
The specific objectives of the research are as follows:

- To present an analysis of different machine learning techniques applied to the task of dropout occurrences prediction for university students.
- To obtain accurate prediction using a specific machine learning technique in determining dropout rates.

2. Literature Review

According to Guerrero, Pulido & Domínguez (2020), in order to make improvements for the curriculum design and plan interventions for the academic support and guidance on the system that is being offered to the students, predicting future students' behavior will play a huge role. With regards to this, as the author of this study stated, this is where Data Mining (DM) will be needed for this matter. This system technique will be able to analyze certain datasets and may extract information in order to become an understandable structure for future use. Thus, the main computation techniques that allow the use of certain information in order to predict the students' performance are the following: Machine Learning (ML), Collaborative Filtering (CF), Recommender Systems (RS), and Artificial Neural Networks (ANN). In addition, these computation techniques will also determine the grades of the students and the risk of dropping out.

As stated by Tamada et al. (2019), distance learning has been significant to the gradual progress in the field of information technology. These changes led the stakeholders to have a new challenge in managing the learning process via a virtual platform. Thus, for the past 10 years, Education Data Mining (EDM) has emerged as a new area in the field of DM, Machine Learning, the statistics of information from an educational setting, and educational technique. Hence, the Machine Learning technique contributes to predicting the rates of dropout students in the Virtual Learning Environment (VLE) — a virtual classroom that allows teachers and students to communicate with each other online.

Based on the study conducted by Donggeun and Seoyong (2018) entitled Sustainable Education: Analyzing the Determinants of University Student Dropout by Nonlinear Panel Data Models. The high participation rate in higher education does not convey the whole picture. Not all freshmen graduate with a bachelor's degree from college. There are several factors besides the inability to catch up that hinder students from graduating; these may be divided into student-related and university-related causes. The former involves a lack of complete knowledge and dissatisfaction with university education, as well as a lack of optimism for the future. Students are often dissatisfied with institutions that have low-quality instructors, inadequate facilities, and a precarious financial situation. No of the origin, a reason might cause children to drop out of school.

The study of Berens et al. (2018) stated that the high rates of student wearing away in higher education are a significant source of concern for institutions and public policy since dropout is not only expensive for students but also a waste of public expenditures. To effectively minimize student attrition, it is critical to determine which students are at risk of dropping out and the underlying factors that contribute to dropping out. The high rates of school dropout globally and their significance show the need for conducting an in-depth examination of its origins and repercussions. The literature suggests that school dropout may be explained by a variety of factors operating at various levels (individual, family, school, and neighborhood). The study also examined the relationship between individual characteristics (defiant attitude, irresponsibility, alcohol abuse, and illegal drug use), family characteristics (educational figure absence and parental monitoring), school characteristics (truancy and school conflict), and school dropout.

Dropout is a significant problem when it comes to online course continuance. As a result, educators and academics have demonstrated a strong interest in developing several models and techniques for reducing online course student dropout through assessment of student's behavior and academic and personal details (Mubarak, Cao, & Zhang, 2020).

3. Methods

3.1 Description of Methodology Used

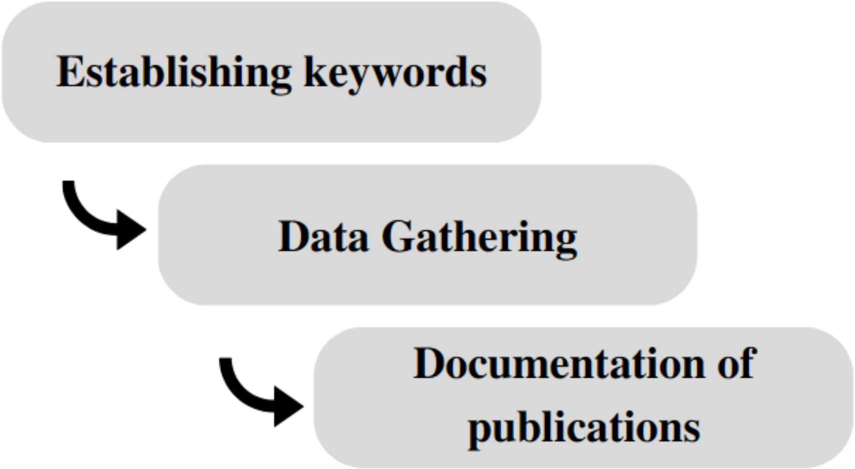

Figure 1. To locate a Machine Learning-related article or journal

To locate a Machine Learning-related article or journal: (i) established a keyword, (ii) gathered related studies, articles, and journals by searching via Multidisciplinary Digital Publishing Institute (MDPI) and other searching as Github, (iii) documented the publications gathered, (iv) worked and tested our code, (v) designed the interface of the program, and (vi) execution of the code and the program. The initial step was to gather keywords that would restrict the topic of this literature review by using "Student Dropout" "Machine Learning Technique" ″Algorithms″ and "Predicting Student Dropout" that will contribute to the information included in this paper. The next step was to submit the terms to Github. The following process was to sort out related articles or journals from 2018 to the year 2021 to Mendeley Reference Manager.

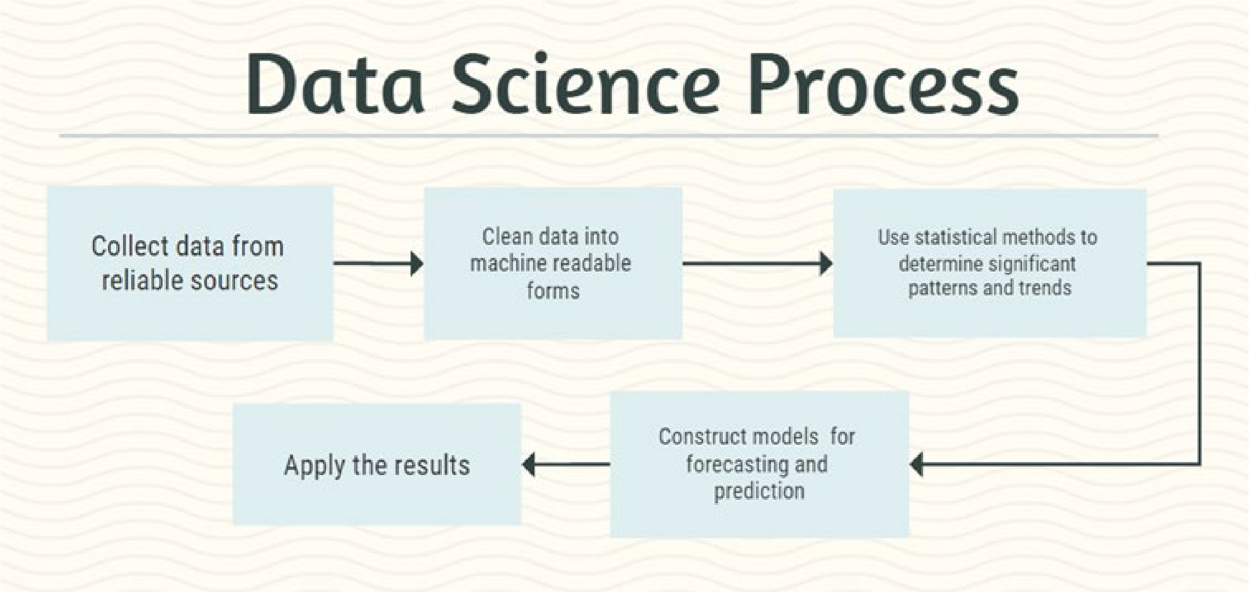
Data Science Process
Collect data from reliable sources
Clean data into machine readable forms
Use statistical methods to determine significant patterns and trends
Construct models for forecasting and prediction
Apply the results

Figure 2. Collected data into machine readable form.

The figure interprets that the researchers then modify the collected data into machine-readable forms. Furthermore, the data collected will be used to construct Neutral Network, Logistic Regression, and Random Forest, and to further understand the study, these models will be used to enhance the forecasting and prediction process. The researchers use HTML, CSS, and Bootstrap for designing the interface of the program. Moreover, we use Python to generate the code for the program. Finally, in order to examine the performance of the program we stored the code in our library in Azure App Service (Microsoft Azure Cloud Platform) and then a website will open and will direct you to a page where the program will ask the user if he/she wants to evaluate a dataset or predict an observation.

3.2 Data Collection

3.3 Data Preprocessing

3.4 Data Exploration

3.5 Data Modeling

3.6 Model Evaluation

4. Results and Discussion

The main objective of the study was to predict whether the student will drop out or not using supervised learning techniques. The program will receive manually entered data or a CSV file and the algorithm will calculate it. Additionally, three supervised learning approaches displayed various evaluation scores that indicated whether or not the student would graduate. The maximum accuracy of 95.00%, the precision of 97.78%, and the F1 score 96.70% were achieved by Random Forest. As a result, the program was successful in achieving both the primary and secondary objectives of the study by differentiating performance metrics of the supervised learning techniques and was able to predict accurately.

In the prototype, the researchers observed that the algorithm Random Forest has the highest percentage of the students passing their grades. The researchers also noticed that the algorithm Logistic Regression has the highest percentage of students failing their grades.

Table 2 shows the final results of the model evaluation where the researchers show if the algorithm used is accurate enough to know what percentage students pass or fail. Based on the results, the highest percentage is the Random Forest, it is more visible and more accurate to analyze to know how many are failing and passing.

5. Conclusion

In the study among the three machine learning techniques the Random Forest Classifier had the highest accuracy, precision, and F1 scores, as well as the second highest recall value. Upon analyzing the data it showed that the researchers had implemented normalization on the data before feeding it to the model, both in training and evaluation. Moreover, the researchers also used the same, fitted normalizer from the training process in the evaluation process to ensure consistency. Unlike in the mother journal, normalization seems to only be applied in the data preparation during the training process. Given this scenario, it would be expected that the model may perceive significant differences in evaluation if input data is not normalized as to how the training data was normalized during the training process.

References

Berens *et al*., Early Detection of Students at Risk - Predicting Student,2018.

Dropouts Using Administrative Student Data and Machine Learning Methods. Retrieved from https://www.econstor.eu/handle/10419/181293

Borrella, I., Caballero, S.C., & Cueto, E.P. (2019). Predict and Intervene: Addressing the Dropout Problem in a MOOC-based Program. Retrieved from https://mitili.mit.edu/sites/default/files/project-documents/a24 Borrella_Caballero_Ponce_2019.pdf

Computer Systems Technology. (n.d.). Retrieved from https://tec.illinoisstate.edu/computer-systems-technology/ CSU Global. (2021, December. Is Studying Computer Science Really That Difficult? | CSU Global. https://csuglobal.edu/blog/is-computer-science-really-difficult

Dalipi, F., Imran, A., & Kastrati, Z. (2018). *MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges*. 2018 IEEE Global Engineering Education Conference (EDUCON), (p1007). DOI: 10.1109/EDUCON.2018.8363340

Faas, C., Benson, M. J., Kaestle, E. C., & Savla, J., Socioeconomic success and mental health profiles of young adults who dropout of college. J. Youth Stud. 21, 669–686,2018. DOI: 10.1080/13676261.2017.1406598

Gausel & Bourguignon, Dropping Out of School: Explaining How Concerns for the Family's Social-Image and Self-Image Predict Anger,2020. Retrieved from https://pubmed.ncbi.nlm.nih.gov/32849096/

International Student. (n.d.). What is Computer Science? in the US. https:// www.internationalstudent.com/study-computer-science/what-is-computer-science/

Lee, Z. & Lee, C. , *A Parallel Intelligent Algorithm Applied to Predict Students Dropping Out of University*. The Journal of Supercomputing, p76,2020. DOI: 10.1007/s11227-019-03093-0

Machine Learning. (n.d.). Retrieved from https://www.mathworks.com/discovery/machine-learning.html

Mduma, N., Kalegele, K., & Machuve, D., A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. Data Science Journal, 18(1), 14,2019. DOI: http://doi.org/10.5334/dsj-2019-014

Mubarak A.A., Cao H., & Zhang W. (2020). Prediction of student's early dropout based on their interaction logs in online learning environment, Interactive Learning Environments, DOI: 10.1080/10494820.2020.1727529

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. Applied Sciences, 10(3), 1042. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app10031042

Sage Dictionary of Social Research Methods. (n.d.). Retrieve from https://methods.sagepub.com/reference/the-sage-dictionary-of-social-research-methods/n155

Selvan, M.P., & Prasanna, N.L. (2019). An Efficient Model for Predicting Student Dropout using Data Mining and Machine Learning Techniques. International Journal of Innovative Technology and Exploring Engineering.

Sosu, E., & Pheunpha, P. (2019). Trajectory of University Dropout: Investigating the Cumulative Effect of Academic Vulnerability and Proximity to Family Support. Frontiers in Education, (4). DOI: https://doi.org/10.3389/feduc.2019.00006

Tamada, M.M., Netto, J.M., & Lima, D. (2019). Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review. Frontiers in Education Conference. DOI: 10.1109/FIE43999.2019.9028545

Tamm, S. (2020). What is the Definition of E-Learning? Retrieved from https://e-student.org/what-is-e-learning/ What is a Dropout? (2015). Retrieved from https://www.purdue.edu/hhs/hdfs/fii/wp-content/uploads/2015/07/s_ncfis04c03.pdf

What is Computer Science? (n.d.). Retrieved from https://undergrad.cs.umd.edu/what-computer-science

What is Computer Programming and How to Become a Computer Programmer. (2021). Retrieved from https://www.snhu.edu/about-us/newsroom/stem/what-is-computer-programming

Voelkle, M. C., & Sander, N. (n.d.). A structural equation approach to discrete-time survival analysis. J. Individid. Dif.
29, 134–147. DOI: 10.1027/1614-0001.29.3.134

Lehr, C. A., Johnson, D. R., Bremer, C. D., Cosio, A., & Thompson, M. (2004). *Essential tools: Increasing rates of school completion: Moving from policy and research to practice.* Minneapolis, MN: University of Minnesota, Institute on Community Integration, National Center on Secondary Education and Transition.

Sanchez, C. (2011). *Why Dropout Data Can Be So Unreliable?* Retrieved from https://www.npr.org/2011/07/28/138750527/why-dropout-data-can-be-so-unreliable

Sustainable Education: Analyzing the Determinants of University Student Dropout by Nonlinear Panel Data Models (2018). Retrieved from https://www.mdpi.com/2071-1050/10/4/954/htm

Kabathova, J., & Drlik, M. (2021). Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. Retrieved from https://www.mdpi.com/2076-3417/11/7/3130/htm

Biographies

Raven Miguel M. Capuno is an undergraduate student from Angeles University Foundation, currently undertaking the B.S. in Computer Science program. His research interest

Chester Juliane M. Ferrer is an undergraduate student from Angeles University Foundation, currently undertaking the B.S. in Computer Science program. His research interest

Brian Troy L. Manaloto is an undergraduate student from Angeles University Foundation, currently undertaking the B.S. in Computer Science program. His research interest

Sean Redd Villafria is an undergraduate student from Angeles University Foundation, currently undertaking the B.S. in Computer Science program. His research interest

James A. Esquivel is an instructor at the College of Computer Studies at Angeles University Foundation, Angeles City.