ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse591022021.pdf
https://doi.org/10.30534/ijatcse/2021/591022021

Predictive Model of Postgraduate Student's Dropout and Delay Using Machine Learning Algorithms

Muhammad Nadeem[1], Sellapan Palaniappan[2]
[1]Department of Information Technology, School of Science and Engineering, Malaysia University of Science and Technology, Petaling Jaya 47810, Selangor, Malaysia,nadeem.muhammad@phd.must.edu.my
[2]Department of Information Technology, School of Science and Engineering, Malaysia University of Science and Technology, Petaling Jaya 47810, Selangor, Malaysia,sell@must.edu.my

ABSTRACT

Neural networks and Logical Regression algorithm provide the best ways to classify data, but they are outperformed continuously by the Decision Tree in analyzing student performance. Therefore, many scholars have used the Decision Tree to predict student performance with greater success. This research analyzed postgraduate student degree outcomes using socioeconomic data to develop a prediction model, where Decision Tree recorded the highest accuracy of 92.79%, better than Logical Regression and Neural Network. For brevity, the Decision Tree was used to produce the prediction model. Based on the study findings, postgraduate students who delay or drop out at the university mostly lack sponsors or had decreased income. Besides, male students delay or drop out if they had financial issues more than their female counterparts. Age, money management skills, number of children, and health expenses are the other factors that contribute to higher dropout or delay at the university. Therefore, this study provides a reliable prediction model for degree outcomes, allowing personalized follow-up to improve graduation rates.

Key words: Algorithm, Decision Tree, Degree outcome, Delay, Dropout, Prediction model, Postgraduate students

I. INTRODUCTION

Decisions trees are reliable types of supervised learning algorithms that build regression or classification models in a tree's shape. They are applicable for both continuous and categorical data, but this study uses only categorical data. The paper presents a data mining case study to create a model that would help learning institutions identify the key factors of Postgraduate degree students' dropout and delay at university. A decision tree algorithm is a feasible alternative to accomplish this task since it has achieved the best outcomes, reaching 100% recall, 99.34% accuracy, and 98.69 precision [1]. Other algorithms, such as Neural Network and Logistic Regression, provide better means of classifying data, although they are outperformed continuously by decision tree [2].

Therefore, it provides a cost-sensitive learning and prediction improvement model without collecting additional data from students. This data analysis shows that the Decision Tree Algorithm performs better with 92.79% (see figure 1) accuracy that is hard to achieve with other sophisticated models. According to [3], [4], and [5], many scholars have used the Decision Tree to predict student performance with greater success. The developed decision tree presents a viable model for universities to predict students likely to drop out or overstay. Understandably, this study seeks to evaluate factors influencing dropout and delay at university by analyzing socioeconomic data to develop a model that will help the university predict postgraduate students' degree outcomes.

II. LITERATURE REVIEW

Educational institutions are fundamental assets in any society because they help advance and nurture academic intellectuals. These institutions continuously strive to improve the quality of education through various technological interventions to improve learning outcomes. Machine learning is an emerging trend in education, where it is applied in learning historical data and using it to predict learners' future behavior [6]. Therefore, many researchers have used the machine learning technique to predict students' future outcomes by classification, a popular machine learning technique [7], [8], [9]. Notably, classification offers a simple means to classify a large population and generate a robust prediction model to predict student performance [10], [11]. The study examines the performance of Decision Tree, Neural Network, and Logistic Regression to generate an accurate model for predicting postgraduate students' degree outcomes.

Researchers who have used various algorithms find the Decision tree more intuitive, powerful, and easy to predict [3], [11], [5]. For instance, [12] studied Portuguese student performance in Math using Naive Bayes and Decision Tree, achieving the highest accuracy (93.0%) with the Decision Tree method against Naive Bayes (91.9%). Another study by [13]combined a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) to predict student drop. The result showed a high level of compatibility with other methods that require manual data mining by experts.

Another study developed by [14]using Deep Neural Network to predict weekly temporal dropout indicated it performed better while considering personalized and prioritized intervention for each student. Also, [15] developed a study to predict student dropout in the future from students' data on attendance, discipline, grades, and course, using a logistic regression framework. The model predicting dropout achieved 76% accuracy, indicating it can precisely distinguish students who will drop out or not at 76%. Most of these research focused mainly on dropout prediction, but only a few focused on socioeconomic factors for in-person dropout or delay at the university. Therefore, this research fills the gap by applying Neural Network, Decision Tree, and Logical Regression to develop a dropout or delay model at the university using socioeconomic data.

III. METHODOLOGY

A. Analysis Tool: Rapidminer Studio

Rapidminer studio is an open-source data mining software solution that runs on Linux, Windows and Mac OS. It provides advanced predictive analytics features for building decision trees with no coding required. Besides, it holds various data mining functions, such as cleansing, clustering, data pre-processing and visualization that makes it suitable for this project. Therefore, the Rapidminer Studio, 64-bit version installed on Windows, was used to analyze the data. The software support function for classifying data using Decision Tree, Neural Network and Logistic Regression algorithms.

B. Data Classification

The study considered data collected about postgraduate students in science and engineering programs. The selected datasets included 742 students enrolled in the various postgraduate programs (i.e. PhD Information Technology, PhD Computer Science, MS Bio-Medical Engineering, MS Civil Engineering, MS Computer Engineering, MS Electronic Engineering, MS Computer Science, MS Software Engineering and MS Electrical Engineering). However, the final datasets included 741 students after data cleaning to remove null values and other errors. The factors are classified as shown in table 1 to obtain labels for predicting the model's supervised learning.

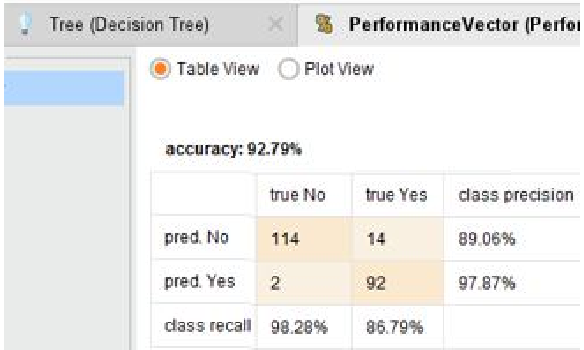
The final data was subjected to the Rapidminer studio to produce results using Decision Tree, Neural Network, and Logical Regression algorithms.

Table 1. Description of Attributes Used To Create Labels

S.No.	Attribute	Value	Description
1	Program Description	Science programme, Engineering programme	Programme taken by the students
2	Study Mode	Full Time, Part Time	The mode of study
3	Sponsor	Yes, No	Loan or scholarship taken by the student
4	Disability	Yes, No	Student's disability
5	Gender	Male, Female	Student's gender
6	Martial Status	Married, UnMarried, Divorced	Student's marital status
7	No. of Childern	0, 1, 2, 3, 4, 5, 6	Student's No. of Childern
8	Age	25-40	Student's age
9	FP_Decrease in income	Yes, No	Financial Problems - decrease in income
10	FP_Marital_family issues	Yes, No	Financial Problems - Marital and family issues
11	FP_Health_Medical expense	Yes, No	Financial Problems - Health and Medical expense
12	FP_Lack of budgeting _money management skills	Yes, No	Financial Problems - Lack of budgeting and money management skills
13	Degree_Status	Yes, No	Student's Degree Status

IV. ANALYSIS AND FINDINGS

Figure 1 and 2 shows the accuracy of the Decision Tree (92.79%) and Logical Regression (83.33%) algorithms, respectively. The Neural Network algorithm does now give sufficient result because it does not work correctly with polynomial attributes, which was used in the socioeconomic data. For brevity, only the model generated by the best performing algorithm is discussed.

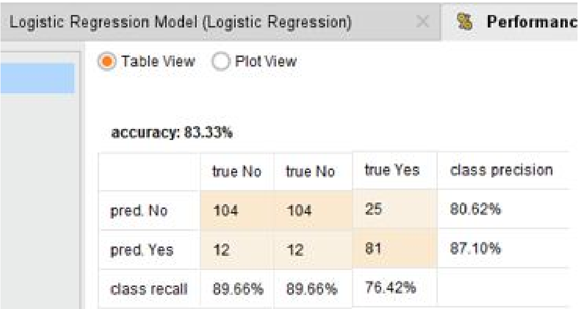


● Table View ○ Plot View

accuracy: 92.79%

	true No	true Yes	class precision
pred. No	114	14	89.06%
pred. Yes	2	92	97.87%
class recall	98.28%	86.79%	

Figure1 The accuracy of the Decision Tree Algorithm used to analyze the data



● Table View ○ Plot View

accuracy: 83.33%

	true No	true No	true Yes	class precision
pred. No	104	104	25	80.62%
pred. Yes	12	12	81	87.10%
class recall	89.66%	89.66%	76.42%	

Figure2The accuracy of the Logical Regression Algorithm used to analyze the data

The tree shown in figure 3 comprises a series of branches and nodes (labels). At the root node (labeled FP_Health_Medical expense), a student either incur medical expenses or none. Each of the branches represents an alternative decision or factor determining a student's degree status in the end- whether or not the student will dropout/delay. Top management of the universities has different views towards risks; therefore, they will draw varying conclusions in the circumstances described by the decision tree in figure 3.

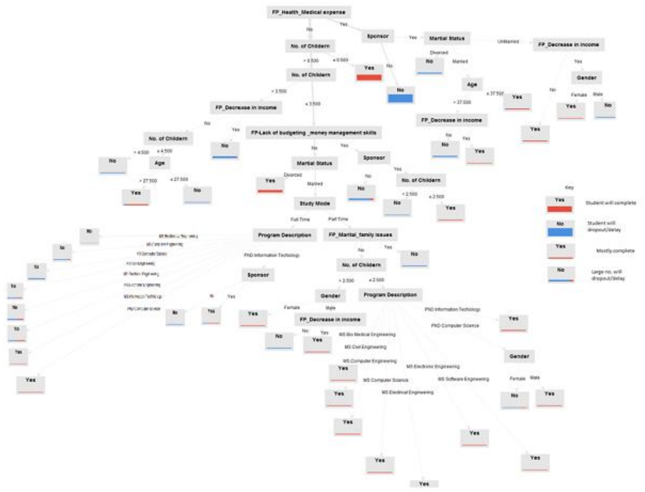


Figure3 Decision tree for various factors that determines postgraduate student dropout or delay at the university

A. Analysis of the Left Side of the Decision Tree

Towards the left of the root of the tree shows the outcome for students who do not have health medical expenses. The factor alone does not determine the degree status; hence, the number of children determines the next outcome. Notably, students with no kids and no health expenses will certainly graduate on time, unlike those with kids. If they have more than three children and recently had a decrease in income (FP_Decrease in income), then all except a few will drop out or delay at the university. Nevertheless, if they did not have decreased income, the outcome will depend on the number of children. Those with more than four children despite having a stable income will drop out or delay, whereas age will determine the fate of students having less than or precisely four kids. At this level, students aged 27 years and below will delay or drop out of the university while the rest will complete on time with few exceptions.

If the student has no health expense and has between one and three children, their success depends on a lack of budgeting money management skills (i.e. labeled FP_Lack of budgeting_money management skills). Students with better money management skills, but no sponsors are more likely to drop out or delay university. However, the number of children is the determining factor for students with sponsors and good money management skills. If they have more than two kids, they will drop out or delay, unlike those with two or less. The success of students with no money management skills further depends on the marital status. Those who are divorced are certain to graduate on time despite poor money management skills, whereas the married success depends on the study mode- whether part-time or full time.

Among full-time students who are married and have poor money management skills, those who enroll in PhD Computer science and MS information technology program graduate successfully. Those taking MS Biomedical Engineering has the highest guarantee of dropout or delay at the university. However, some students enrolled in MS computer Engineering, MS civil engineering, MS Electrical engineering, and MS electronic engineering succeed to graduate on time. Still, the majority drop out or delay their studies. For students enrolled in PhD Information technology, those having sponsors have high graduation success rate while those without sponsors drop out or delay in university.

Part-time students with marital family issues (i.e. labeled FP_Marital_family issues) has higher chances of dropping out or delaying. Nevertheless, those with no marital family issues and have two or fewer kids complete their studies regardless of the program. If the student has no marital issues and more than two kids, then females succeed to graduate while the male student's success depends on whether their income decreased. A decrease in income automatically reduces a male student's chance of graduating on time.

B. Analysis of the Right Side of the Decision Tree

Students who incur health medical expenses, but no sponsorship ultimately drop out or delay their studies. Nevertheless, those who have medical expenses and sponsors depend on marital status to determine their studies' completion. If they are divorced, such students will likely drop out or delay their studies. The married with 37 years or less have a higher completion rate excluding a few who fail to graduate or delay their studies. The married and aged above 37 years tend to complete their studies when income decreases compared to those with a stable income. Besides, sponsored and unmarried students who have stable income complete school within the stipulated timeline. Nevertheless, unmarried female students tend to complete university than their male counterparts who have high dropout or overstay rates.

C. Findings

The decision tree does not give the management the answer to university completion problem; instead, it helps institutions determine which alternative factors will lead to dropout or delay. Notably, it shows the administration of the university what decision today will contribute to the institution's long-term goal. The goal is to retain more students by identifying those at risks of not completing and applying various interventions on time. Here are the findings where a series of factors lead to dropout or delays:

- Students who have medical expenses tend to complete education on time unless they have no sponsors or experience decreased income.
- Male students are at high risk of dropping or delaying than female counterparts, particularly if they experience a decrease in income.
- Income fluctuation (decreased income) and sponsorship are the major factors influencing whether postgraduate students dropout or delay at the university.
- Part-time students tend to drop out or delay their studies than part-time students regardless of the study program.
- The number of children (3 or more) also affects how long postgraduate students stay in school or drop out altogether.
- Age has little impact on the dropout or delay but can have a greater impact if the student experienced reduced income.

A substantial subset of these students chose not to continue or delay their studies, although all indicators show successful continuation were present. The experimental results show that a simple classifier can give a more accurate prediction (92.79%) that is hard to achieve with other sophisticated models. Therefore, this study provides a reliable model for predicting possible dropout for new students entering a high-level course, allowing personalize follow-ups to reverse possible adverse educational outcomes. One interesting aspect of this model regards divorced students. For instance, students with no money management skills and medical expenses but divorced are successful than those divorced but have sponsors and medical expenses. It shows medical expense is imposing a heavy burden on the divorced. Another observation is that students aged 27 years or below tend to drop out or delay even if they do not experience decreased income or incur medical expenses.

V. CONCLUSION AND RECOMMENDATION

The number of successful graduates indirectly affects the operation cost of a university per student. If the dropout or relay is high, the cost can increases per student aside from reputation damage. The decision tree algorithm provides a trivial solution for classifying data to identify students’ weaknesses, interests and abilities. Accordingly, this study developed a predictive model that can predict students’ graduation status based on socioeconomic data. Performance evaluation of this algorithm shows it has 92.79%, which provides a reliable prediction model. Various socioeconomic factors affect student graduation, such as health expenses, marital status, family size, sponsorship, age, but only a series of these factors can determine the final degree outcome. Notably, students who experience decreased income and have no sponsorship have the highest dropout or delay rates, particularly male students. Medical expense is the other factor that significantly influences the final degree outcome besides study mode, domestic issues, marital status, and the number of children. Hopefully, this predictive model will benefit university management, and academicians devise strategies to improve weak students and ensure they graduate on time.

A. Recommendations

Most students are hard hit by lack of sponsorship and health expenses; therefore, institutions should devise intervention measures on these factors to curb dropout and delays in universities. In the future, researchers can improve this study by including data from different programs, particularly from fields unrelated to sciences and technology. Besides, there is a need to expand the system, developing it to include other socioeconomic data to create a robust model [16], [17]. A similar approach should be applied to a more extensive database from several universities to generate a reliable and inclusive model.

REFERENCES

[1] F. Freitas, F. Vasconcelos, S. Peixoto, M. M. Hassan, A. A. Dewan, V. H. Albuquerque and P. R. Filho, "IoT System for School Dropout Prediction Using Machine Learning Techniques Based on Socioeconomic Data," *MDPI*, pp. 2-11, 2020.

[2] G. W. Dekker, M. Pechenizkiy and J. M. Vleeshouwers, "Predicting Students Drop Out: A Case Study," in *Educational Data Mining 2009*, 2009.

[3] S. A. Kumar and M. D. Vijayalakshmi, "Efficiency of Decision Trees in Predicting Student' Academic Performance," *Cs & It-Cscp*, p. 335–343, 2011.

[4] S. K. Yadav and S. Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," *World of Computer Science and Information*, p. Vol 2(2): 51–56, 2012.

[5] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and Information Technologies*, pp. 61-72, 2013.

[6] M. Solís, T. Moreira , R. González, T. Fernández and M. Hernández, "1Perspectives to Predict Dropout in UniversityStudents with Machine Learning," *ResearchGate*, pp. 1-5, 2018.

[7] N. Gilbert, "Predicting Success: An Application of Data Mining Techniques to Student Outcomes," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, pp. Vol 7(2): 1-19, 2017.

[8] C. Romero, S. Ventura, P. G. Espejo and C. Hervás, "Data Mining Algorithms to Classify Students," *EDM*, pp. 8-17, 2008.

[9] M. López and J. M. Luna, "Classification via clustering for predicting final marks based on student participation in forums," in *Proceedings of the 5th International Conference on*, 2012.

[10] A. B. Ahmed and I. S. Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method," *World Journal of Computer Application and Technology*, pp. Vol 2(2): 43-47, 2014.

[11] S. Yadav, B. Bharadwaj and S. Pal, "Data mining applications: A comparative study for predicting student's performance," *International Journal of Innovative Technology & Creative Engineering*, pp. Vol 1(12): 13-19, 2012.

[12] P. Cortez and A. Silva, "Using Data Mining to Predict

Secondary School Student Performance. In A. Brito and J. Teixeira Eds.," in *Proceedings of 5th FUture BUsiness TEChnology Conference*, Porto, Portugal, 2008.

[13] W. Wang, H. Yu and C. Miao, "Deep model for dropout prediction in MOOCs," in *Proceedings of the 2nd International Conference on Crowd Science and Engineering*, New York, 2017.

[14] W. Xing and D. Du, "Dropout prediction in MOOCs," *Using deep learning for personalized intervention.J. Educ.Comput.Res*, p. 547–570, 2019.

[15] R. S. Bakera, A. W. Berning, S. M. Gowda, S. Zhang and A. Hawn, "redicting K-12 Dropout,," *Journal of Education for Students Placed at Risk (JESPAR),* pp. 1-15, 2019.

[16] N. M. Suhaii, S. Abdul-Rahman, S. Mutalib, N. Hamid and A. Malik, "Predictive Model of Graduate-On-TimeUsing Machine Learning Algorithms," *ResearchGate,* pp. 131-140, 2019.

[17] S. Herzog, "Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression," *New Directions for Institutional Research,* pp. 17-33, 2006.