Research in Engineering Education Symposium &
Australasian Association for Engineering Education Conference
5 – 8 December, 2021 - Perth, WA

PREDICTING STUDENT PERFORMANCE IN ENGINEERING COURSES: A RISK MODEL ANALYSIS

Veronica Abuchar, Jose De La Hoz, Camilo Vieira, and Carlos Arteta
Universidad del Norte
Corresponding Author's Email: vabuchar@uninorte.edu.co

Abstract

CONTEXT

Improving student academic success in higher education courses is a central objective for educational institutions. Hence, student academic failure and dropout rates are of significant concern. Recent studies link academic success to student self-efficacy, academic performance, social environment, demographics, and performance expectations of students. One of the strategies to evaluate academic success is through risk analysis: a set of methods to analyze, understand, and predict student outcomes before enrolling in specific majors or challenging college courses.

PURPOSE OR GOAL

Contributing to the goal of academic prediction, the purpose of this research is to develop a simple methodology to estimate fragility curves for students entering an engineering course. A fragility function describes the probability of succeeding in a course, given the students' GPA. The implementation of the proposed methodology facilitates the generation of models and decision-making according to the estimation of the probability of a student surpassing or not a specific grade for a course.

APPROACH OR METHODOLOGY/METHODS

The data used to generate fragility functions comes from a database of engineering courses collected over several years at a particular university. The data includes Course Grade of interest (CG) after taking a class, and the Grade Point Average (GPA) of the students before taking it. The methodology estimates the probability of surpassing a specific performance level in a course implementing the idea of fragility functions used in the earthquake engineering field but adapted to engineering education. For example, the data can be organized to developed cumulative distribution functions to represent the probability of surpassing or failing a specific course given the students' GPA.

ACTUAL OR ANTICIPATED OUTCOMES

The resulting fragility curves have the potential to achieve two goals: (i) assessing the population at risk for a course to take actions for improving student success rates, and (ii) assessing a course difficulty based on the fragility function parameters. A practical case in which fragility curves are helpful is to compare the difficulty of two or more engineering courses, detecting subjects in which students tend to have more challenges to succeed.

CONCLUSIONS/RECOMMENDATIONS/SUMMARY

In the literature, there are research studies that have focused on predicting student failure or dropping out in the first academic year or models to predict academic performance in the last semester of the program; however, this research focused on predicting academic success in any course of the program, provided that the GPA information is available. The procedure used to generate fragility curves used in seismic engineering is applicable to generate risk curves that estimate the probability of academic success in engineering courses.

Keywords

Fragility functions, academic success, engineering education, risk assessment, retention.

Introduction

Improving student academic success in higher education has been an important objective for academic institutions over the years. Student academic failure and dropout rates in engineering are a significant concern in several countries, including Colombia (Casillas, Robbins, Allen, Kuo, Hanson, & Schmeiser, 2012; Lucio, Hunt, & Bornovalova, 2012; Vieira, Aguas, Goldstein, Purzer & Magana, 2016). In Colombia, engineering dropout rates are more than 50%. Students drop engineering programs for several reasons, but academic performance is one of the main predictors at all educational levels (Casillas et al., 2012). Past academic performance and student demographics are some of the main predictors of academic success (Shahiri, Husain, Rashid, 2015; Alyahyan & Düştegör, 2020). Predicting student failure becomes relevant for institutions to develop procedures to support engineering students and avoid student dropout (Knight, Carlson & Sullivan, 2007).

Several approaches have been used to predict student success/failure rates. For instance, Lucio and colleagues (2012) used the receiver operating characteristic (ROC) curves to identify the optimum number of risk factors. Vandamme and colleagues (2007) implemented mathematical techniques (decision tree; neural networks and linear discriminant analysis) to predict the probability of failing or dropping out in their first academic year. Educational data mining (EDM) methods have also been used to predict students' performance. EDM methods extract relevant information from a large educational database to predict or analyze students' performance (Angeline, 2013; Shahiri et al., 2015). Risk analysis is another important process that has been used to analyze, understand, or predict students' outcomes before enrolling in specific majors or particularly difficult college courses (Bernacki et al. 2020; Alipio, 2020; Esmat & Pitts, 2020; Wilson & low, 2014; Dekker et al., 2009; Ohland et al., 2011; Marbouti et al., 2016; Belfield & Crosta, 2012). The importance of predicting student risk failure lies in the possibility of improving the teaching-learning process (Shahiri et al., 2015; Alyahyan & Düştegör, 2020), allowing teachers to make informed instructional decisions. This process may also minimize student repeating attempts at courses and improve completion rates through timely actions (Esmat & Pitts, 2020).

While all these different methods may help predict student failure or academic success in undergraduate programs, our approach will focus on predicting student success in individual courses. We argue that institutions may benefit from lower student dropout rates by improving the course-specific success rate at the program level. This study proposes a model to predict student success in specific undergraduate courses using their past grade point average (GPA). The model is based on fragility functions used in the earthquake engineering field to estimate the chance of structural damage given the ground-motion intensity. This approach also allows comparing two different courses and may help higher education institutions to make informed decisions to support student learning.

Theoretical Framework

In earthquake engineering, fragility functions are useful to describe the effect of earthquakes in a building. Given a particular building, a fragility function helps to estimate the probability of exceeding a specific limit state of an engineering demand parameter (EDP) as a function of ground motion intensity measure (IM). For example, the limit state of an EDP could be an acceleration threshold at the roof of a building which can vary according to different values of IM. Note, this is only a statistical data organization procedure that may be expanded to other fields. In this sense, this paper adapts this organization procedure to engineering courses when generating fragility functions to estimate the chance of obtaining a certain course grade

(CG) as a function of the grade point average (GPA) of the students before taking such course (**Figure 1**).

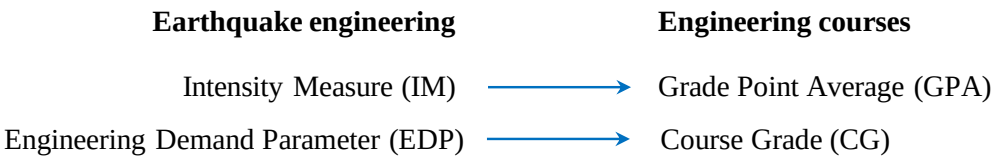

Figure 1: Equivalence of concepts from earthquake engineering to engineering courses

Baker (2015) presents two methods to obtain the data for estimating fragility curves, both fulfilling the need of finding correlating pairs of a cause and a consequence. Fragility curves are defined as a cumulative distribution function (CDF), which depends on the statistical distribution of the data treated. Typically, the lognormal distribution is used to elaborate these functions, as is shown in Equation (1)

$$\boldsymbol{P(CG > cg|GPA = x) = \Phi\left(\frac{\ln\left(\frac{x}{\theta}\right)}{\beta}\right)} \tag{1}$$

where $P(CG > cg|GPA = x)$ is the probability of obtaining a course grade greater than *cg*, given a test value of $GPA = x$; and Φ() is the standard normal cumulative distribution function. According to Baker (2015), logistic regression is also used to describe fragility functions. These are special cases of generalized linear models (GLMs) and will be the preferred option used in this paper. All GLMs have three components: the random component, the systematic component, and the link function. According to Agresti (2012):

- **Random component:** identifies the response variable *Y* (i.e., a consequence) and chooses a probability distribution for it. When the *Y* observations are binary, as is the case of success or failure, then a binomial distribution must be assumed for *Y*.
- **System component:** specifies the independent (predictor or explanatory) variable(s) (i.e., the cause). These variables get in as predictors and the linear combination of them is known as a linear predictor.
$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$
- **Link function:** Specifies a function of the expected value of *Y*, this is, $E(Y) = \mu$. When μ takes values between 0 and 1, then is appropriate to use a logit link function, this is, $g(\mu) = \log[\mu/(1-\mu)\,]$. When a GLM has a logit link function, then is called a logistic regression model, which is the case for this study.

The distribution of *Y* is represented by the probability $P(Y = 1) = \pi$ of success, $P(Y = 0) = 1 - \pi$, and $E(Y) = \pi$. The binomial distribution of *Y* follows Equation (2).

$$\boldsymbol{P(y) = \binom{n}{y} \pi(x)^y (1 - \pi(x))^{n-y}} \tag{2}$$

where *n = 1* when we work with binary observations, and $\pi(x)$ represents the conditional mean of *Y* given the independent variable *x* according to Equation (3)**.** The corresponding logistic regression function is presented in Equation (4), which implies that $\pi(x)$ increase or decease as an *s*-shaped function of the independent variable *x*.

$$\boldsymbol{\pi(x) = \frac{e^{\beta_o + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}} \tag{3}$$

$$\boldsymbol{logit[\pi(x)] = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x} \tag{4}$$

In this Logistic regression, or logit model, the parameter β_1 indicates if the curve increase ($\beta_1 > 0$) or decrease ($\beta_1 < 0$), and its magnitude defines how fast increase or decrease, that is, the slope. When $\pi(x) = 0.5$, *x* corresponds to the median effective level (EL_{50}) which represents the probability for success equals to 50% and can be calculated as $x = -\beta_0/\beta_1$.

According to Hosmer (2013), there are two significant reasons for selecting the logistic distribution. The first one is that logistic regression is an extremely flexible and easily used function, mathematical speaking. The second one is that model parameters provide "*the basis for clinically meaningful estimates of effect*".

The maximum likelihood method is used to estimate the parameters of the function for this model (Equation (5)):

$$\boldsymbol{l(\beta)} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\pi(x_i)^{y_i} * (1 - \pi(x_i))^{1-y_i}} \tag{5}$$

where $\beta = (\beta_0, \beta_1)$. Taking advantage of the logarithm's properties, then Equation (5) can be transformed to Equation (6).

$$\boldsymbol{L(\beta)} = \boldsymbol{\ln[l(\beta)]} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}} \tag{6}$$

Procedures for Estimating the Fragility Curves

In this section, we present the steps to estimate the risk of failure given the GPA of the student before taking a specific course. To explain the procedure, we use the data from a mid-sized private university in Colombia. The sample course is Calculus II which has 6,709 data points collected between 2008 and 2017. In the next section, the courses Physics I and Statistics are included to compare the three courses.

1. **Collect GPA versus CG pairs for the concerned course**
   Collect (GPA, CG) pairs, where the GPA is that of the students before taking the course of interest. Additional metadata may be included depending on the purpose of the fragility curve. For example, if the idea is to compare the evolution of a course, a third parameter can be the period in which the course was taken (e.g., semester, year). On the other hand, if the purpose is to compare the success in different educational institutions, it will be important to separate the information according to its origin. Note that the use of only one input variable (i.e., GPA) is a limitation of this methodology.
   The scatter plot in **Figure 2** helps visualize the data distribution. For the case study presented here, good-standing students at the college of engineering must have a GPA ≥ 3.3; hence the X-axis range starts there. The course grade scale goes from 0 to 5, and the minimum approving course grade is 3.0.

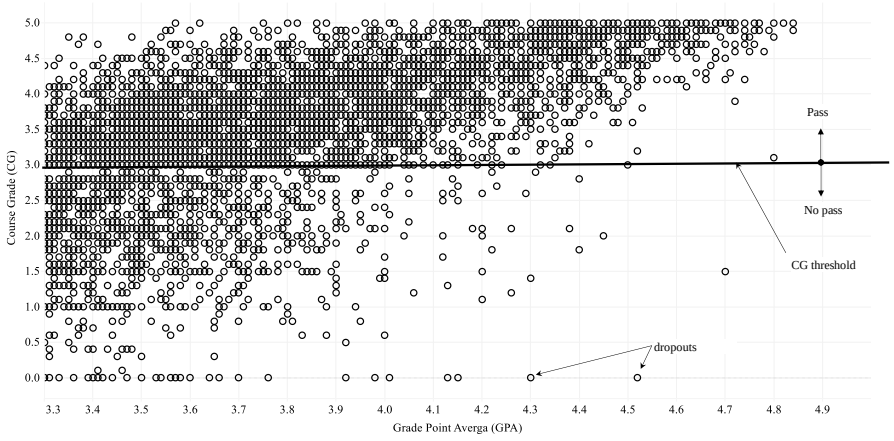

Figure 2: Scatter plot of GPA and CG of the course of Calculus II

2. **Select GPA level of interest and bin the data**

Define GPA bins from the minimum applicable GPA to the maximum GPA, depending on the institution's standards. Here, we use the range $3.3 \leq \text{GPA} \leq 5.0$, and the bins increments of 0.1. When defining the bin size, one must consider the amount of data available. Fewer data points require larger bins. **Figure 3** shows a bubble plot of CG versus GPA bins. Note, the size of each bubble indicates the concentrations of data around specific pairs of (GPA, CG).
In this stage, also define the CG threshold, which depends on the purpose of the fragility curve. For the case study, CG = 3.0 is selected as a threshold because this is the grade from which a student approves or not a course in the institution under study. However, any other threshold can be selected. For example, in the case study, the so-called *distinguished students* have a GPA ≥ 3.8, so a CG = 3.8 could be another possible threshold to analyze.

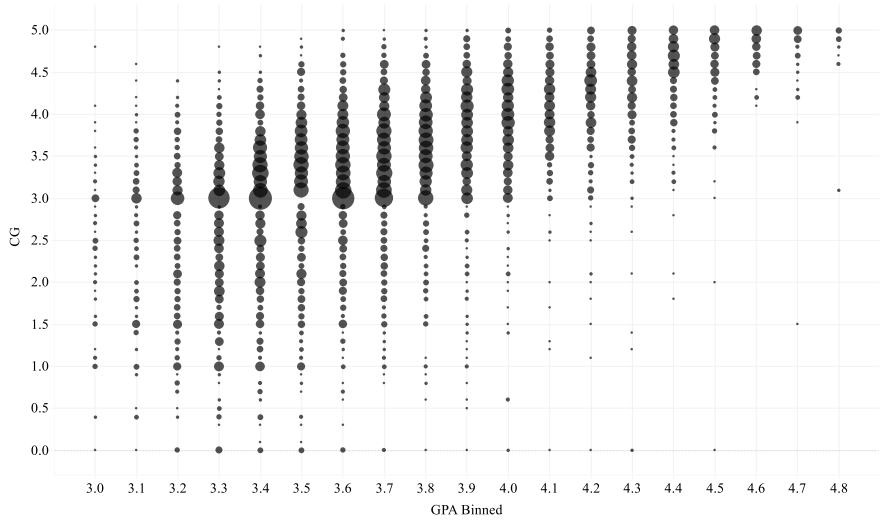

Figure 3: Bubble plot of binned GPA and CG of the course of Calculus II

3. Estimate logit coefficients and standard deviation

Once a CG threshold is defined, it is necessary to create a binary vector with the same size as the amount of data (i.e., of students evaluated). For each student, this vector has values of 1 when the $\text{CG} \geq \text{CG}_{\text{threshold}}$, and 0 otherwise. The fragility curves are estimated by a generalized linear model (GLM) using binomial probability distribution and logit as the link function in MATLAB (see code in Appendix). The inputs of the function are a vector collecting the GPA of the students, and the corresponding binary vector explained above. The code estimates the logit coefficient of the function.

4. Computes predicted values of GLM and plots fragility curves

Knowing the parameters β_0 and β_1, we can use Equation (3) to estimates the probability of surpassing the $CG_{threshold}$ for each GPA; hence, the fragility curve is estimated as $1 - \pi(GPA = x)$. **Figure 4** shows two fragility curves: the first one evaluates the probability of failing the course of Calculus II, while the second one evaluates the probability of obtaining CG < 3.8 for the same course. These fragility curves must be interpreted in this way: a student with a GPA = 3.6 has a probability of 22% of not passing the course, while the same student has a probability of 75% of obtaining a CG < 3.8. The complement to these probabilities offers another perspective from the same data. **Figure 4** also shows the binary vector plotted against GPA. It is worth mentioning that observations showed in this figure are not binned GPA, so they overlap.

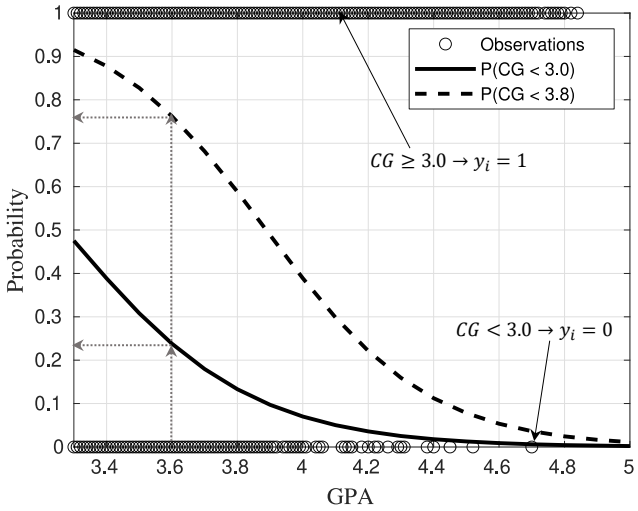

Figure 4: Fragility function of the course of Calculus II for a probability of CG < 3.0

Application case

An application case of these fragility curves compares the estimated academic performance that a student with a specific GPA would obtain in each course of interest. **Figure 5** presents the fragility curves of three courses: Calculus II, Physics I, and Statistics. **Figure 5a** shows the probability of failing each course given the student's GPA. This figure shows that Statistics is the most difficult subject among these three, and for Calculus II the students show a better performance. For example, a student with a GPA = 3.4 has a 40% chance of failing the course of Calculus II, while for Physics I and Statistics, this student has a 50% chance, approximately. **Figure 5b** presents a CG threshold of 4.0 and depicts a different behavior in comparison with **Figure 5a**. Note that both, Physics I and Statistics cross at an about GPA = 4.3, which also coincides with the 50th percentile. This indicates that, in an average sense, for both courses, a GPA of at least 4.3 is required to surpass the 4.0 grading.

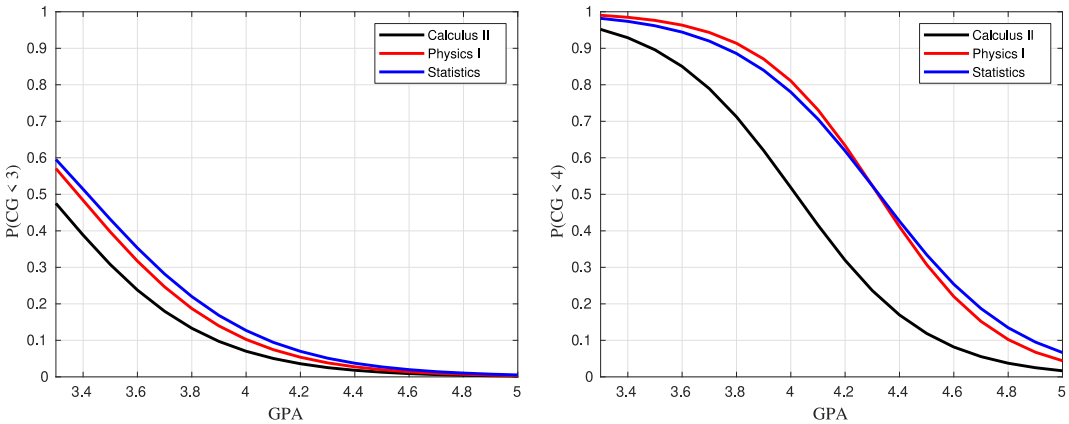

Figure 5. Fragility curves of the course of Calculus II, Physics I, and Statistics comparing: (a) the probability of not passing each course; (b) the probability of obtaining a CG < 4.0

Two important parameters for each curve are shown in **Table 1**. The first parameter is β_1 and its magnitude shows the rate at which the curve is decreasing, that is, the slope of the curve. For instance, **Figure 5b** shows that the curve of Physics I is stepper than Statistics and Calculus II as confirmed by the values of β_1 in **Table 1**. Note that **Figure 5** shows plots for $1 - \pi$, hence, the slopes are negative. A flatter slope indicates the data is more scattered. A

second more important parameter is EL_{50} which indicates the 50th percentile of the *GPA* data. As commented previously, one can use the EL_{50} to directly compare the difficulty of each course on an average sense, as it defines the overall horizontal position of the curves along the X-axis. For example, from the $CG_{threshold} = 3.0$ data in **Table 1**, Statistics with the larger EL_{50} value indicates that there is at least a 50% chance of failing the course for students of GPA equal to or less than 3.42. This GPA threshold is smaller for the other two courses; hence, students with lower GPAs are more likely to pass it.

Table 1. Parameters of the fragility curves for different CG of the course of Calculus II, Physics I, and Statistics

Parameter	$CG_{threshold}$ = 3.0			$CG_{threshold}$ = 4.0		
	Calculus II	Physics I	Statistics	Calculus II	Physics I	Statistics
β_0	-11.61	-11.85	-11.30	-16.70	-19.59	-16.91
β_1	3.55	3.51	3.30	4.15	4.53	3.91
$EL_{50} = -\beta_o/\beta_1$	3.27	3.38	3.42	4.02	4.32	4.32

As was mentioned before, this model may be used for other application cases. Students' academic performance in course offerings may be useful to identify how different strategies have contributed (or not) to student success. Likewise, this model may also be helpful to compare the same courses at various institutions, or over the years.

Conclusions

A significant concern in higher education is to enhance academic success in engineering programs. This paper contributes towards this goal by describing a methodology that enables instructors and decision-makers to predict students' future performance in a specific course from historical past performance in an objective manner. The proposed methodology uses fragility functions with historical course grades and corresponding grade point average (GPA) before taking the course. Once the fragility curves are created, it is possible to predict the probability of exceeding a specific CG given the GPA for a particular student.

Fragility functions were elaborated using a generalized linear model (GLM) with the binomial logistic method. Once fragility functions are created for the courses of interest, it becomes a functional tool to assess the population of risk according to their GPA. When this population is detected, it is possible to create mitigation actions to improve their academic performance.

One application case was presented, which consisted of comparing three courses: Calculus II, Physics I, and Statistics. Knowing the fragility curves parameters of each course is possible to compare the difficulty between one and others depending on the GPA of students and the CG threshold selected.

While we believe that this model can be helpful to inform instructional decisions, we recognize that other factors beyond the GPA may influence student success in a given course. We argue against providing students themselves with the outcomes of this model, as this may affect their self-efficacy towards the course and the program, and may misinform their future decisions. This model may be useful to inform teaching practices and to assess the consistency of the course difficulty.

Appendix

MATLAB code

```
%% LOGIT - FRAGILITY CURVE

% b: file with 4 columns: 1) ID of the observation, 2) the course grade of each
observation, 3) GPA of each observation, 4) GPA binned each observation

b = importdata('Calculus_II.txt');
values_b = b.data;
GOI = 3.0; % CG Threshold

GPA = values_b(:,3);
GPA_binned = values_b(:,4);
CG = values_b(:,2);
cond = zeros(length(values_b),1);

for i = 1:length(values_b)
    if CG(i)>= GOI
        cond(i) = 1;
    end
end

[logitCoef] = glmfit(GPA_binned, [cond],'binomial','logit');

beta_0 = logitCoef(1);
beta_1 = logitCoef(2);
EL_50 = -beta_0/beta_1;

GPA_x = 3.3:0.1:5;
for i=1:length(GPA_x)
    logitFit_plot(i)=exp(beta_0+beta_1*GPA_x(i))/(1+exp(beta_0+beta_1*GPA_x(i)));
end

%% Graphics
plot(GPA, cond, 'ok')
hold on
plot(GPA_x,1-logitFit_plot,'-','lineWidth',2,'Color', [0 0 0]);
hx = xlabel('GPA');
hy = ylabel('P(CG < 3.0)’);
ylim([0 1]);
axis([3.3 5 0 1])
grid on
```

References

Agresti, A. (2012). Categorical Data Analysis. Vol. 45 Wiley Series in Probability and Statistics.

Alipio, M. (2020). Predicting Academic Performance of College Freshmen in the Philippines using Psychological Variables and Expectancy-Value Beliefs to Outcomes-Based Education: A Path Analysis. Education and Administration, 1-15. DOI: 10.35542/osf.io/pra6z.

Alyahyan, E., Düştegör, D. Predicting academic success in higher education: literature review and best practices. Int J Educ Technol High Educ 17, 3 (2020). https://doi.org/10.1186/s41239-020-0177-7

Angeline, D. (2013). Association Rule Generation for Student Performance Analysis using Apriori Algorithm.

Baker, J. W. (2015). Efficient analytical fragility function fitting using dynamic structural analysis. Earthquake Spectra, 31(1), 579-599.

Bernacki, M., Chavez, M., Merlin,P. (2020). Predicting achievement and providing support before STEM majors begin to fail. Computers & Education, 158,103999. https://doi.org/10.1016/j.compedu.2020.103999.

Casillas, A., Robbins, S., Allen, J., Kuo, Y. L., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. Journal of Educational Psychology, 104(2), 407.

Dekker, G., Pechenizkiy, M., Vleeshouwers, J. (2009). Predicting Students Drop Out: A Case Study. Computers, Environment and Urban Systems. 41-50.

Esmat, T., Pitts, J. (2020). Predicting success in an undergraduate exercise science program using science-based admission courses. Adv Physiol Educ, 44(2):138-144. doi: 10.1152/advan.00130.2019. PMID: 32108508.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Knight, D., Carlson, L.,Sullivan, J. (2007). Improving engineering student retention through hands-on, team based, first-year design projects. In Proceedings of the International Conference on Research in Engineering Education. Honolulu, HI.

Lucio, R., Hunt, E., & Bornovalova, M. (2012). Identifying the necessary and sufficient number of risk factors for predicting academic failure. Developmental psychology, 48(2), 422.

Ohland, M., Brawner, C., Camacho, M., Layton, R., Long, R., Lord, S., Wasburn, M. (2011). Race, Gender, and Measures of Success in Engineering Education. Journal of Engineering Education, 100, 225-252. https://doi.org/10.1002/j.2168-9830.2011.tb00012.x

Shahiri, A., Husain, W., Rashid, N. (2015). A review on predicting Student's performance using data mining techniques. Procedia Computer Science, 72, 414–422.

The MathWorks Inc. (2021). *MATLAB* (2021a) [Computer Software]. Retrieved from https://la.mathworks.com/

Vandamme, J., Meskens, N., Superby, J., (2007). Predicting academic performance by data mining methods. Education Economics, 15(4), 405–419.

Vieira, C., Aguas, R., Goldstein, M. H., Purzer, S., & Magana, A. J. (2016). Assessing the Impact of an Engineering Design Workshop on Colombian Engineering Undergraduate Students. International Journal of Engineering Education, 32(5), 1972-1983.

Acknowledgments

The author would like to thank Beatriz Sanjuanelo and Rubiel Velasquez for helping in collecting the data.