International Journal of Modern Computation, Information, and Communication Technology
2022;5(2):40-46.
ISSN: 2581-5954

A Statistical Feature Data Mining Framework for Students' Progress over Gender Variations

R. Dhanalakshmi[1], A.V. Kalpana[2], J. Umamageswaran[3], B. Praveen Kumar[4*]

[1]*Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India.*
[2]*Department of Artificial Intelligence, Shri Vishnu Engineering College for Women, Bhimavaram, Andra Pradesh, India.*
[3]*Department of Information Technology, R.M.K Engineering College, Chennai, Tamil Nadu, India.*
[4]*Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Tamil Nadu, India.*

[*]*Corresponding author: bprawinkumar@gmail.com*

Abstract. Gender differences in academic progression are examined in this article using two novel indicators. Two indicators are used to show how long students have studied above the minimal amount of time needed to graduate. We examine these characteristics using Kaplan–Meier techniques, Expedited Fault Timeframe, Regression trees, Multiple Regression analysis tree branches, and a redesigned Gender Parity Index. It is possible to analyse data that has been censored as well as add other factors into the study using this unique mix of statistical methodologies. Gender disparities in university education may be studied using a combination of survival analysis and the Accelerated Failure Time (AFT) model, according to data from a Greek as well as Italian institution. It is possible to quantify the relevance of these influencing factors using data mining approaches such as survival trees and multivariate regression trees. The proposed indicators concurrently use the Multivariate Regression Tree technique, which considers more than one consistent outcome variable. Gender plays an essential part in this research, as women outperform males in terms of these new metrics by completing their studies in a shorter amount of time and with a greater level of performance, regardless of other student characteristics. The findings were improved across the board when using a gender parity index that was tweaked.

Keywords: Data Mining, Regression Tree, Statistical Analysis, Student Data, Learning Management System (LMS).

INTRODUCTION

Basically, Gender is widely accepted to have a significant impact on academic research. When discussing the gender gap, "frequently regards all dimensions of education as acting against women," says Jacobs. In the first place, there were stark and apparent inequities that harmed women in terms of education opportunities and, as a result, degree awards. However, this is no longer the case [1]. In most nations, women currently outnumber males in higher education, and this trend is expected to continue in the foreseeable future. The percentage of women in higher education is expected to climb from 48 percent in 2015 to 59 percent in 2025, assuming the current trends continue, according to the Organization for Economic Co-operation & Development (OECD).

Women's greater proclivity for higher education appears to inspire a higher level of degree achievement. However, these patterns appear to be consistent at least in OECD member nations, where the gender disparity in degree awarding is rising. It is projected that by 2025, 70 per cent of all degrees would be given to women, an increase of 54 percent since 1998. In the above-mentioned reality, women enrolled in higher education at 51 percent and 57 percent, respectively, in Greece and Italy, the two nations at the centre of this study in 2005, and these numbers are predicted to rise to 53 percent and 57 percent, respectively, by 2025. The ratio of female graduates in Italy and the United States was 61 percent in 2005, respectively, but is anticipated to rise to 70 percent in 2025 [2].

The Gender Parity Index (GPI)2 for the gross enrolment ratio (the ratio of female to males registered at common organizations) for Greece and Italy from 1971 to 2016. Women's access to university education is no longer a problem in Greece and Italy, where the GPI value is larger than 1 since 1985 and 1992 respectively; this shows that women's access to tertiary education is no longer an issue in these two countries. From a position

Received: 17.05.2022 Revised: 06.07.2022 Accepted: 24.07.2022

in which males had an advantage of 0.54 to a situation where women had an advantage of 0.76 by the mid-1990s, the indicator consistently increased for both nations [3].

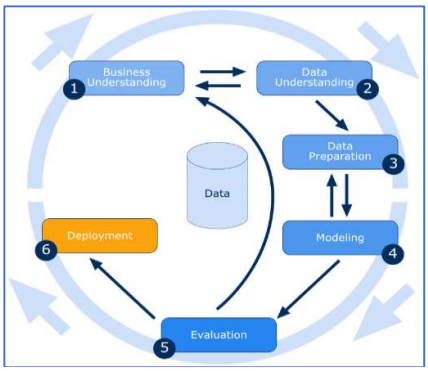

FIGURE 1. General Steps of Data Mining

For this study, we will examine the gender imbalance at an interim stage by investigating the disparities in academic achievements between male and female students in higher education, given that women have already surpassed men in terms of enrolment and graduation rates [4]. It is expected that the more sophisticated variables, such as the time of study and the final grade earned by students, will be used in the analysis, as these are two of the most important indicators currently used in assessing colleges. Data mining for educational purposes has recently piqued the attention of researchers (EDM).

EDM is a new area that employs data mining (DM) techniques to examine and extract the hidden knowledge from educational data contexts. It's essential to note that EDM has a diverse range of users, all of whom have distinct goals and visions for utilizing EDM. For example, instructors may utilise the concealed knowledge to enhance their teaching methods, students can use it to understand better their own learning styles and the learning process itself can benefit from it. The administrator can make better selections with the aid of this tool [5].

Web-based learning, instructional archives and conventional surveys may all be used to gather educational data. Several DM strategies can be employed in EDM to address various educational issues. Classification is the most used method for predicting educational models, for example. Decision trees, neural networks, and Bayesian networks are just a few examples of the many methods that fall under the categorization of classification [6].

EXISTING WORK

Predicting the success of students in online learning settings is a critical challenge. Classification, regression, and grouping are some of the DM techniques used to develop a predictive model. Classification is the most widely used method for predicting student achievement. Artificial Neural Networks (ANN), DT as well as Naive Bayes are some of the classification algorithms that may be used (NB). A decision tree is a hierarchical arrangement of circumstances. The ease with which it may be turned into a collection of categorization criteria is what drew many researchers to this method. C4.5 and CART are two well-known DT algorithms [7].

The DT method was used by Romero et al to estimate students' final grades based on their Moodle usage statistics. Moodle is a well-known and widely adopted Learning Management System (LMS). Real-world data from eight Moodle classes at Cordoba University has been used by the author to identify students as either passing or failing. Classifying students with identical final grades into separate categories depending on the behaviours they engage in throughout an online course is the goal of this study [8]. Another common data mining tool is the neural network, which has been employed in educational research. Neural networks are intelligent systems inspired by biology and composed of neurons, which are small, interconnected units that help to complete a certain task.

To determine the educational outcomes of engineering students, Arsad et al. employed an ANN model. Students' Grade Point (GP) in core subjects is used as an input, but no information about their demographics is considered when calculating their cumulative Grade Point Average (CGPA). Students pursuing an engineering degree use NNs to learn how to get certain results. This study indicated that the final CGPA after graduation is strongly influenced by foundational topics [9]. A student's CGPA was predicted by use of Bayesian networks based on his or her prior academic performance at the time of application for admission. In today's world, educational institutions are looking for a way to identify the best students graduating from different universities [10].

Research presented in this paper introduces a fresh method to the prediction model that incorporates a case-based component. Using a case-based approach, the prior student most comparable to the candidate being evaluated is found [11]. Defining similarity between instances (candidates can apply) in a way that is compatible with the forecasting model is a difficult task. Any university with a robust repository of student and application information can use this strategy. Learning management systems (also referred simply web-based learning or web-based schooling) are the result of a rise in educational internet use [12].

As a digital foundation for online learning, the LMS is a must-have. It is the primary function of the LMS to manage students, evaluate students' engagement, and keep track of students' progress throughout the LMS. Resource allocation and management are handled via the Learning Management System [13]. From a learning management system, educational data is gathered in this study. Kalboard 360 seems to be a multi-agent learning management system built with cutting-edge technologies to help students' study more effectively. Users can access instructional content simultaneously from any Internet-connected device with such a system. In addition, parents and school administrators should be involved in the educational process [14].

In this way, it becomes a comprehensive process that involves and links all stakeholders. Using a technology called experience API, learner activity data is gathered. Learner behaviors, such as reading an article or viewing an instructional video, can be tracked using the xAPI, a component of TLA's Training and Learning Architecture. Learning activity providers can utilize the Experience API to identify the learner, activity, and objects that comprise a learning experience. X-API is used in this study to track students' behavior throughout the educational process and identify characteristics that may impact their academic success [15].

Only 151 students' records, with 11 characteristics, were included in the educational data set used in the prior study. The present study includes 500 pupils and 16 characteristics. Three primary categories are used to categorise the features: Gender and nationality are examples of demographic characteristics. The educational stage, grade level, and section are all examples of academic background information. Behavioural characteristics such as raised hands in class, visits to resources and parent answers to surveys. This tool tracks the LMS progress of students and their guardians. These elements show the involvement of both the student and the parent.

PROPOSED SYSTEM

If a student dropped out of school before earning their degree, it was a failure in their educational endeavour. Attrition is a term that is still used today, but often, it is referred to as pupil abrasion, which is clear now as the number of pupils those fail to get gradation, too estimated as a proportion of those who enrol in University [16] Different statistical measures used for different classification in [17-18]. It's important to note that attrition and retention are synonyms, and that the percentage of students who successfully manage their degree is sometimes referred to as a graduation rate as opposed to a completion rate.

As per literature, the phrase "student attrition" has been widely used in the United States from the early 20th century, if not before. Those with a degree are no longer as well-rewarded in terms of monetary compensation and, by extension, social and private benefits. Student development may be stalled or demotivated due to uncertainty about their future career path due to the shifting landscape of graduate jobs. Alternatively, they may be compelled to change their field of study entirely. Students may also drop out of a course if they believe it is never relevant since their job routes are changing so quickly.

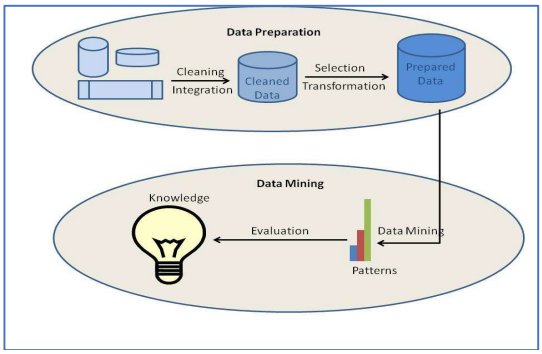

FIGURE 2. Proposed System Architecture

Due of the large amounts of money that people, companies, and societies have invested in recent graduates, this is a topic that has recently piqued public interest. Some statistical approaches have been employed and implemented to the available data to support the hypothesis outlined above. Kaplan–Meier estimators, accelerating failure time (AFT) models, and regression models (survivability vegetation and multinomial logit trees) are some of the methods used in this study. These include In a survival study, the term "T" generally refers to the amount of time that individual has left until an event occurs.

$$SSM_K = \sum_{i=1}^{n}\sum_{j=1}^{r}(y_{ij} - \overline{y_j})2$$

We define T as a student's length of studies as the period from the date of first registration to the date of commencement or the end of the follow-up period, whichever comes first. Survival analysis has the capacity to handle properly censored data, which is a crucial characteristic of the approach. In cases whenever the components have not been exposed to the event by a certain follow-up period, right filtering takes place. A student who has not yet earned a degree and is still enrolled in school at the conclusion of the follow-up period is represented by the right-censored data in the study of university advancement.

If T is supposed to be continuous or discrete, then it is a non-negative fuzzy number. The density function, typically abbreviated as f (t), and the gamma distribution are analysed in a survivability analysis technique for a given period T. It is known as an unqualified failure rate because it depicts the likelihood of an event occurring at any given moment t without regard to any factors. If an individual's life span exceeds or falls short of a certain value, F(t) represents the likelihood of this happening. This paper's research focuses on enrolling university students who are tracked for a certain amount of time until they are expected to graduate. The likelihood of a student earning a degree in t months or less is thus provided by the function F(t).

CART approaches for censored data have been extended to include survival trees. Each successive node may be judged based on the log-rank statistic in this situation. Predictor values may be used to characterise the output of the algorithm when a learning sample, which includes time, status, and predictor values, is supplied into it. With the Kaplan–Maier estimate of the group unit's cumulative response variable, the overall and suppressed data points falling into each final category are also included in every end node. Consequently, we have subsets of the underlying survival data, each with its own set of unique properties. The survival trees have mostly been used in medical data thus far, according to the available literature.

Algorithms such as Boosting can transform weak students into strong students. It is easy to boost a series of learners consecutively and aggregate them for prediction; but, by adjusting the weights of the weaker learner, it is possible to focus more on errors made by prior learners. In the case of binary classification, boosting has a special restriction. The AdaBoost algorithm removes this restriction. AdaBoost is an illustration of an adaptive boosting algorithm. This algorithm's main goal is to focus on patterns that are difficult to categorise. Each training set has a weight given to it, and this weight is used to determine how much attention is being paid to each subset. Equal weights are given to each subgroup. The weights of incorrectly categorised occurrences rise with each repetition, while the weighting of correctly classified examples fall. Through this voting mechanism, the AdaBoost ensemble produces an effective learner from poor classifiers.

RESULTS AND DISCUSSION

Graduated women (82 percent) outnumber their male counterparts (72 percent) by a statistically significant margin over the study period. Supporting the point presented is this finding, which is in line with earlier relevant findings outlined by Vincent-Lancrin. There are roughly 16 and 39 months between the average TGaT S(t) = 1.7 and the upper quartile S(t)= 1.15, indicating that a fourth of the graduates nearly cut-off date. While comparing the graduation rates of men and women, the male rate is consistently lower than the female rate.

TABLE 1. Median Probability Distribution

	Median	Probability
Male	22	0.514
Female	34	0.489
Total	20	0.568

As a result, the log-rank test shows that this difference is statistically significant (z = 12.822, p-value 0.0002), demonstrating that females and males behave differently with regard. It has been backed by data on graduation dates, the risk (probability) that a student will not graduate in 1, 2, or 3 years after the graduation threshold, and the anticipated risk (probability) that a student would not graduate after 4 years. Women take an average of 14 months to complete their TGaT, whereas men take an average of 24 months, and the odds are consistently lesser for females than for men.

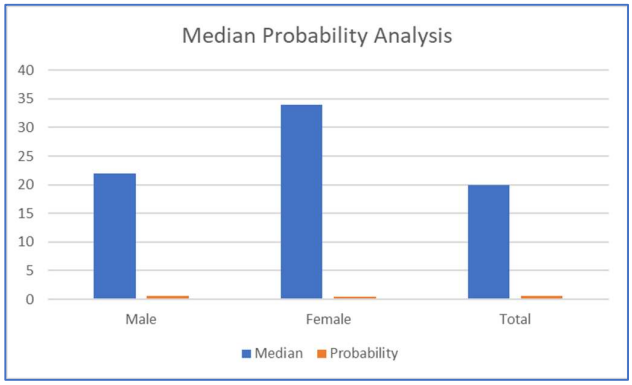

FIGURE 3. Comparative Study

The performance is superior for women based on these findings. Once at value compared with fewer than 1 on the GPI, the proportion of "survived" women is lower than it is for males at that point in time. If GPI is bigger than 1, then the reverse is true. In any case where GP It is less than 1, it shows that there is a gender gap. After the initial 20 months of observation, the GP It curve begins to flatten out and eventually settles at 0.6 after another 20 months of steady decline.

TABLE 2. Failure Model Analysis

	Exponential	Welbull	Log
Log	45.356	54.879	41.387
ATC	48.365	50.368	38.547
BTC	55.36	50.247	56.321

To put it another way, the number of female students graduating from college is always higher than the number of males. It turns out, however, that throughout the course of the first 29 months of the study, the disparity between the sexes steadily widens in favour of women before stabilising. This graduation behaviour might be attributed to the fact that females are more determined to complete their educations and earn a gradation, especially at this early stage of the evaluation period.

With the inclusion of previously discussed factors, GPI t findings are usually consistent with the results provided before. This difference is more evident for women in Behavioral sciences than for other departments, according to GP It data. According to Mullen and Backer, traditionally female-dominated fields like psychology and sociology have a higher percentage of PhDs than traditionally male-dominated fields like business and economics.

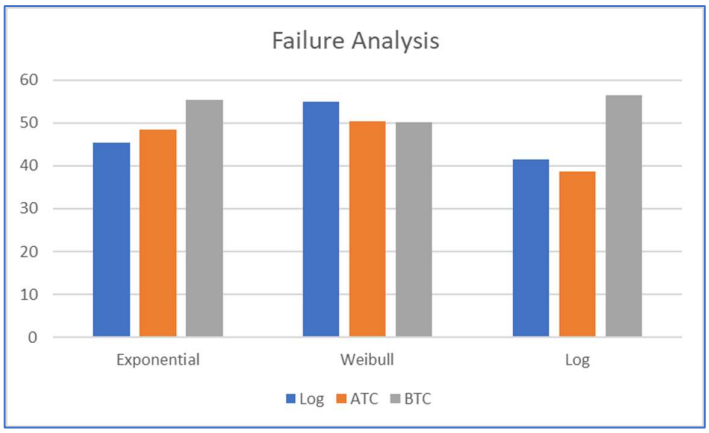

FIGURE 4. Failure Analysis of Proposed Model

Native students have a bigger discrepancy for women than foreign scholars, and steady pupil stake a better discrepancy than non-regular students when it comes to gender inequality. GP It numbers confirm this. For the first 24 months, both Italian and Greek undergraduates have a gender difference in favour of women, according to GP data, but the disparity expands beyond that period.

CONCLUSION

An in-depth examination of intermediate-stage indicators has been carried out because women have surpassed males when it comes to enrolment and graduation rates in higher education. The proposed indicators concurrently use the Multivariate Regression Tree technique, which considers more than one consistent outcome variable. Students' graduation marks and time of study are two of the most essential indications for companies nowadays and for assessing colleges. Web-based learning, instructional archives and conventional surveys may all be used to gather educational data. This analysis focuses on these two more refined factors. Examining the academic achievement of students in social science-oriented courses at an Italian and Greek institution was the inspiration for this article. Students at both institutions have the same duty to study for a certain amount of time, but there is no cap on how long they can study for. However, our suggested technique might be used to any level of schooling to study similar factors.

REFERENCES

[1]. E. A. Amrieh, T. Hamtini, I. Aljarah, 2015, “Preprocessing and analyzing educational data set using X-API for improving student's performance,” *2015 IEEE Jordan Conf. on Applied Electrical Eng. and Computing Technologies (AEECT)*, pp. 1-5. IEEE.

[2]. M. Hanna, 2004, “Data mining in the e-learning domain”, *Campus-wide information systems*, **21(1)**, pp. 29-34.

[3]. F. González-Gómez, J. Guardiola, ÓM. Rodríguez, MÁ Alonso, 2012, “Gender differences in e-learning satisfaction,” *Computers & Education*, **58(1)**, pp. 283-90.

[4]. G. Kakasevski, M. Mihajlov, S. Arsenovski, S. Chungurski, 2008, “Evaluating usability in learning management system Moodle,” *Iti 2008-30th Int. Conf. on Information Tech. Interfaces*, pp. 613-618. IEEE.

[5]. V. Moisa, 2013, “Adaptive learning management system,” *J. of Mobile, Embedded and Distributed Systems*.; **5(2)**, pp. 70-7.

[6]. C. S. Ong, J. Y. Lai, 2006, “Gender differences in perceptions and relationships among dominants of e-learning acceptance,” *Computers in Human Behaviour*, **22(5)**, pp. 816-29.

[7]. S. Putrevu, 2001, "Exploring the origins and information processing differences between men and women: Implications for advertisers," *Academy of Marketing Science Review*, **10(1)**, pp. 1-4.

[8]. S. Rapuano, F. Zoino, 2006, "A learning management system including laboratory experiments on measurement instrumentation," *IEEE Trans. on Instrumentation and Measurement*, **55(5)**, pp. 1757-66.

[9]. C. Romero, S. Ventura, 2010, "Educational data mining: a review of the state of the art," *IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **40(6)**, pp. 601-18.

[10]. A. M. Shahiri, W. Husain, 2015, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*. **72**, pp. 414-22.

[11]. M. E. Zorrilla, E. Menasalvas, D. Marin, E. Mora, J. Segovia, 2005, "Web usage mining project for improving web-based learning sites," *Int. Conf. on Computer Aided Systems Theory*, pp. 205-210. *Springer, Berlin, Heidelberg.*

[12]. J. P. Salazar-Fernandez, M. Sepúlveda, J. Munoz-Gama, 2019, "Influence of student diversity on educational trajectories in engineering high-failure rate courses that lead to late dropout," *In 2019 IEEE Global Eng. Education Conf. (EDUCON)* pp. 607-616. IEEE.

[13]. O. Viberg, M. Khalil, M. Baars, 2020, "Self-regulated learning and learning analytics in online learning environments: A review of empirical research," *Proc. of the Tenth Int. Conf. on Learning Analytics & Knowledge*, pp. 524-533.

[14]. P. Mukala, J. Buijs, M. Leemans, W. van der Aalst, 2015, "Exploring Students' Learning Behaviour in MOOCs Using Process Mining Techniques," *Computing Conf.*, pp. 1–12.

[15]. V. Naderifar, S. Sahran, Z. Shukur, 2019, "A Review on Conformance Checking Technique for the Evaluation of Process Mining Algorithm," *TEM J.* **8**, pp. 1232–1241.

[16]. S. Murugan, S. Mohan Kumar, and T.R. Ganesh Babu, 2020, "CNN model Channel Separation for glaucoma Color Spectral Detection," *Int. J. of MC Square Scientific Res.* **12(2),** pp. 1-10.

[17]. E. A. Kolog, S. N. O. Devine, 2019, "Texture Image Classification by Statistical Features of Wavelet," *Int. J. Adv. Sig. Img. Sci*, **5(1)**, pp. 1–7.

[18]. M. Alagirisamy, 2021, "Micro Statistical Descriptors for Glaucoma Diagnosis Using Neural Networks," *Int. J. Adv. Sig. Img. Sci,* **7(1)**, pp. 1–10.