Using numerous biographical and enrolling observations to predict student performance

Mpho Mendy Nefale, Ritesh Ajoodha

University of the Witwatersrand, Computer Science and Applied Mathematics, Johannesburg, South Africa

Abstract. One of the most worrying aspects for students, when they enroll in universities, is how they will perform and whether they will finish their degrees on time if they will even finish the degree at all. Students may, however, fail to complete their studies on time or at all for a variety of reasons, including but not limited to biographical background, academic obstacles, and enrolment status. Another massive assignment is to discover the possibilities of whether or not the student will be successful or not earlier than the beginning of the educational year. The goal of this study is to forecast student performance by utilizing biographical history and enrollment observations so that the likelihood of a learner's success can be determined early. Identifying the likelihood of a learner's success or failure earlier will allow us to provide students with relevant support to boost their chances of success. It will also assist in recommending a course in which the student is more likely to succeed. In this study, predictive machine learning approaches such as SVMs, logistic regression, decision trees, KNN, Naive Bayes, Decision trees, and random forest were applied. SVMs outperformed all other classifiers with an accuracy of 95.56%. By considering the students' biographical background and enrolment statuses, this study argues for a more nuanced view of forecasting vulnerable learners.

Keywords: Classifications, Machine Learning, students performance.

1 INTRODUCTION

The research of forecasting student success using biographical variables, precollege observations, and enrolling data is introduced in this chapter. The problem statement, related work, purpose statement, overview of techniques, contribution, results in summary, and overall report structure will all be discussed in this chapter.

1.1 Problem statement

For many students, student performance has been a major concern. For most students, getting a spot at a university is a life-changing experience, with many

students picking a course only because they heard it pays better. The sad reality is that the majority of students who are accepted into university programs do not complete their studies owing to a lack of academic ability or performance in their field of study [1].

1.2 Purpose Statement

The major goal of this research is to create a system that is capable of accurately predicting a student's likelihood of success in a given set of disciplines based on various background and enrollment data. Throughout their studies, students are anxious about the success of their degrees. Having a system that predicts student performance would help students determine how much they need to invest in order to ensure their success [1].

1.3 Related work

Several studies have been conducted to predict student performance. A study conducted by Osmanbegovic [2], Kabakchieva [3] and Minaei-Bidgoli [4], data mining technologies were used to predict student performance. A study conducted by Elbadrawy [5] used personal data to predict students performance by the application of machine learning algorithm logistic regression. A study conducted by Ajoodha[1] used features of the learner's background, personality, and school to predict student performance using several machine learning algorithms.

1.4 Methodology

Several machine learning models were utilized to predict student performance in this study. The synthetic data included biographical information, pre-college information, and enrollment information. Highest risk, high risk, lowest risk, and medium risk are the four risk categories for the response variables. SVMs (Support Vector Machines), Random Forest, Decision Tree, Naive Bayes, Logistic Regression, and KNN (The k-nearest neighbors) were among the machine learning models utilized. Accuracy, confusion matrix, precision, recall, and f1-score were used to evaluate the models.

1.5 Results

After applying all of the models for this investigation, SVMs outperformed all other models with an accuracy of 95.54%, precision of 96%, recall of 96%, and f1-score of 96%.

1.6 Contribution

This project will add to current knowledge by developing a model that can predict learner risk status before completing a certain course of study and comparing aspects that are more successful in predicting student performance.

1.7 Overview

In chapter 2 we will be discussing literature review in details, Chapter 3 will focus on the methodology of this study including evaluations,in chapter 4 we will discus the findings of the research and chapter 5 will be the conclusion of the report and some future work recommendations.

2 Related work

Students that participate in a students success course have a higher chance of succeeding than students who do not participate in a students success course [6]. Students success courses assist students in making a smooth transition from their previous level of education or experience. This may be a group of students ranging from high school to post-secondary education. Students enrolled in student success courses accounted for 68% of students who graduated in record time,indicating that students enrolled in students success courses have a better chance of succeeding. However, some students succeed despite not being enrolled in students success courses [7].
By 2000, just over 20 percent of women had earned a bachelor's degree in engineering, but that number has since dropped dramatically. According to the results, students who drop out of engineering while maintaining reasonable grades had high grades in high school and showed less interest in the subject [8]. These students may have been encouraged to study engineering by family members, but they ultimately chose to pursue other interests, according to the findings [9]. Students who switched majors after a poor performance in engineering, on the other hand, appeared to have high hopes for the course, which could have been affected by the possibility of financial benefits in post-graduate studies or jobs [10].
First entering students are at the highest the possibility of dropping out during the first semester of study or failing to complete their program/degree on time in all institutions of higher learning [8]. The majority of students are between the ages of 30 and 40, with more than 68% of students being over 30. This age category is also associated with a higher likelihood of failing the course, with a rate of students failing the course of 37.7%, which is higher than the overall percentage of students failing the course in the student population (38%) [11].

Adelman looked at data from the NCEs', the sophomores cohort study was a 13-year study that followed a group of sophomores from high school to college [12]. In high school, choosing advanced math and science courses was related to obtaining an engineering bachelor's degree, according to the report. Engineering majors were more likely to be pursued by students that have a high average academic performance and quantitative score's test than their lower performing peers. Students with superior overall academic performance and standardized students with higher exam scores were more likely to pursue engineering majors than their less-achieved counterparts; students who studied engineering for the love of it were more likely to excel than those who pursued it for the sake of

higher pay after graduation [10].

The random forest algorithm is one of the well known supervised learning system. It creates a "forest" out of a collection of decision trees that are commonly trained using the "bagging" method [13]. The bagging method's primary idea is that it combines many learning models enhances total output. In a random forest, the process of dividing a node analyzes only a subset of the features at random [6]. A study conducted by Ndiatenda Ndou [14] used random forest to predict the student performance and obtained 94.04% accuracy.
According to the statistics, the majority of information students (63 percent) are female; nevertheless,the percentage of female students who successfully complete the course (65%) is higher than that of male students, indicating that female students are more likely to finish the course. Most students in the information system who are disadvantaged have been demonstrated to be at a disadvantage due to their disability [15]. Students with disabilities have a higher chance of failing than those who do not. Depending on their ethnic heritage, the percentage of students that successfully finished the course varies dramatically.33.33% of the students on these courses were enrolled in bachelor's degree programs in applied sciences [15]. In comparison to students enrolled in bachelor's degree programs in business, they have a higher chance of failing the course. Finally, students who enroll in this course during the summer semester are more likely to fail than those who enroll during the fall semester who take it during the fall or spring semesters [16]. Year of commencement, plancode, plan description, stremline, age in first year, school quintile, mathematics matric major, home province, rural or urban, life orientation, physics Chem, English first language, home country, additional mathematics, mathematics matric literacy, computer studies, and English first additional were all included in a study that yielded an accuracy of 80% when employing a Naive Bayes model [7].
Nghe [17] and Ajoodha Jadhav [1] used BNT Models to predict students performance and achieved 61.54% and 70% accuracy respectively

Author(s)	**Models**	**Accuracy**
Ndiatenda Ndou	Random Forest	94.04%
Ossama E	Decision tree	96.5%
Madhuri T	CART	79.48%
Marin AJ	MIMIC	AGFI = 0.95
Noluthando Mngadi	Random forest	82% with AUC = 0.95
Macdaline RM	Bayesian networks	82%
Ajoodha	BNT	70%
Sangodiah	Linear SVM	89.95&

Table 1. Literature Review Summary Table

3 Research methodology

This chapter will provide a fast overview of the data collecting and the methods that has been used to analyze and model the data.

3.1 Data collection and sampling methods

The study used second-hand data; it is synthetic data based on the learned Bayesian network structure modeling. The value of the parent node is taken from their unconditional distribution, and the value of the child node is taken from the parent set. Repeat the sampling process until all node values have been generated. Use Gaussian distribution to model continuous variables. The Gaussian distribution, usually called the normal distribution, is a continuous function with a mean and standard deviation, provided that the data is normal. Negative values such as cumulative estimates and probabilities have negative values. Remove negative values from the data set; they cannot change them,because it would change the network's distribution. The level of the factor is represented by tabular conditional probability density (CPD), which is used to model discrete variables. The data was not balanced, however, SMOKE was used to balance the data.

3.2 Features

The data includes biographical information, pre-college observations, and university enrollment observations. We have utilized this information to perform our research. Gender, race, first-year age, home language, home province, home country, and place of origin, whether rural or urban, are all biographical characteristics that has been employed in the study. In the pre-university observations, we utilized the school quintile to indicate school courage, with quintile 1 being the poorest and quintile 5 being the least deficit in basic math, English as a second language, finally, for the observation of university enrollment, we used the year of commencement of studies, the description of the plan, professional history, the possibility of success in different branches of science (mathematics, physics, earth sciences and life sciences), the totality of course grades and the number of years that the Completion of studies was spent.

The table below summarises the features used in this study.

3.3 Methods

Random forest Random forest is a simple machine learning technique that in most cases gives excellent results even without hyperparameter tuning and is widely used due to its simplicity and versatility [18]. The random forest produces

Biographical Characteristics	Pre-College Observations
Race	School Quantile
Gender	English FAL
Home Language	Mathematics major
Age at first year	Computers
Home province	Additional Maths
Rural or Urban	
Home country	

Table 2. List of Biographical and pre-college features to be used in the study

Enrollment Observation
Year started
Probability of streams
Plan Description
Aggregate
Number of yrs in degree

Table 3. List of Enrolment features to be used in the study

understandable predictions and can handle large datasets effectively. The random forest has been discovered to be more accurate in predicting outcomes [18]. By expanding the number of trees, this strategy improves accuracy [19]. Random forest has proven to be useful in different kinds of problems [20]. Random forest has been successfully applied in remote sensing, bioinformatics, predicting students' performance, analyzing customers' behavior, and many other predicting problems. [21–25].
Random forest generates and blends numerous decision trees to produce a more exact and dependable prediction. The random forest has the benefit of being able to be used for both classification and regression issues, both of which are common in modern machine learning systems. Random forest has nearly identical hyperparameters to decision trees and bagging classifiers. The precise information is which you do not want to mix a bagging classifier decision tree due to the fact that you may use random forest classifier-class [26]. Random forest also can manage regression duties. Another great advantage of the random forest approach is that determining the proportional value of each feature on the prediction is a breeze.

We use the Gini index when performing random forests on categorization data. To determine how nodes on a decision tree branch, we apply the formula below:

$$Gini = 1 - \sum_{i=1}^{c} (pi)^2$$

The Gini of each branch on a node is calculated using the class and probability formula above.
Entrophy can also be used to figure out how a decision tree's nodes branch.

$$Entrophy = \sum_{i=1}^{c} -pi * log_2 pi$$

The relative frequency of the class you're looking at in the dataset is represented by pi, and the number of classes is represented by c.

3.4 Decision tree

One of the supervised machine learning algorithms is decision tree. It can be used to tackle problems related to regression and classification [27]. The decision tree method is often the method of choice for predictive modeling since it is both simple to learn and highly effective [28]. The decision to make strategic splits has a significant impact on the correctness of a tree [29]. Classification and regression trees have different decision criteria. Decision tree been successfully applied in a range problems including business, medicine, computer science and many more [30–32] and it was found to give highly accurate results. Because the method clearly spell out the problem and allow all choices to be altered, decision tree is an effective technique of decision-making.

3.5 Linear logistic regression

Linear regression is a well-known simple version of modeling the linear relationship between fixed and undistorted variables [33]. "Logistics" refers to categorical responses; for two categories, (binomial / binary logistic regression) is binary or dichotomous. (polynomial logistic regression) Our risk status is the dependent variable Bi, the categories are minimal risk, medium risk, high risk, and very high risk; the independent variables Ai are biographical features, observations prior to university, and very high risk and enrollment according to Table 2.1

$$\gamma_i = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + ... + \beta_n A_n$$

$$\gamma_i = log \frac{P_r(lowest)}{P_r(highest)}$$

When comparing the lowest and greatest risk categories

$$\gamma_i = log \frac{P_r(medium)}{P_r(highest)}$$

When comparing the medium risk group to the highest risk category

$$\gamma_i = log \frac{P_r(high)}{P_r(highest)}$$

When comparing the high risk group to the highest risk category,**The reference category is the one with the highest risk.**
Linear logistic regression was used for predicting corporate financial distress and the method was found to be efficient [34].

3.6 K-Neighbour

K-Nearest Neighbor, or KNN, is a supervised learning technique that may be applied to both regression and classification problems. The method is commonly used in machine learning to solve categorization difficulties. The KNN was used for the prediction of stock price, the method was found to be efficient and gives highly accurate results [35]. KNN was also used to predict students' performance the method was found to be accurate [36].
When utilizing KNN, the first thing you should do is convert your data into mathematical values. The method will calculate the distance between these points' mathematical values [37].
This distance is calculated using the Euclidean distance formula, as illustrated below:

$$d(p,q) = d(p,q) = \sqrt{(q_1 - p_1)^2 + ... + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

3.7 Naive Bayes

Naive Bayes is a machine learning model that classifies data points using Bayes' Theorem. Naive Bayes was used for predicting the system for heart diseases and decision support in heart disease prediciton system, the method was found to be effective [38][39]. The method has been used in many prediction problems, including criminal prediction, software defect prediction, and complex networks [40][41][42].
With only two occurrences — event A and event B — the formula below is the simplest version of Naive Bayes.

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

When a data collection has more than two possible events, the formula will assume this form:

$$P(x|c) = \prod_{k=1}^{d} P(x_k|c_j)$$

Simply said, Naive Bayes is a set of methods that can be used to classify huge data sets using probability.

3.8 SVMs

SVM is a type of supervised machine learning model that can be used to solve classification problems [43]. We depict each data item as a point in n-dimensional space in this approach, where n is the number of features you have. The SVM approach is beneficial because it recognizes non-linearity in the data and produces

an accurate prediction model [44, 45]. When the data is linearly or non-linearly separable, SVMs perform well in terms of accuracy. When the data is linearly separable, the SVMs produce a separating hyperplane that optimizes the margin of separation between classes along a line perpendicular to the hyperplane. The SVM's primary role is to look for a hyperplane that can distinguish between the two classes [46]. The SVM was applied to predict membrane protein types and virulent protein in bacterial pathogens [47]. The method was found to be accurate. Based on the related work, we can say that SVM is a good method for predicition problems [47][48].

3.9 Evaluations

The goal of evaluation is to put a model through its paces on data that differs from what it was trained on. This gives an unbiased assessment of learning performance.
In this study, we employed the evaluation functions listed below:

- Accuracy
- Confusion matrix
- Precision, recall and f1-score

4 Results and Discussion

The outcomes of our experiment are presented and discussed in this section.

4.1 Data Analysis

Python 3 was used to evaluate the bogus data. The findings were predicted using the machine learning models mentioned in the preceding chapter. When a student begins a program, the data displays four risk statuses: highest risk (students at this risk status end up dropping out), high risk(students at this risk status fails to finish the degree in more than 3+ years), medium risk (students at this risk status finish their degree in more 3+ years), and low risk (Students at this risk status gets to finish their degree on time or in less than 3+ years).

Feature Information Gain. This section investigates the contribution of each characteristic to classifying the risk status class variable. Using the IGR, the most contributing features were found as illustrated in the table below by order from the top one. The table below will show only 7 most important features.

Features Statistics In this brief section, we will investigate how the significant features connect to other variables and the target variables.

Rank	Feature
1	English First Lang
2	English First Aditional
3	Plan Description
4	Number of years for degree
5	Race Description
6	Progress outcome
7	Quintile

Table 4. Top 7 most important features

4.2 Classification

The outcomes of the six fitted machine learning classification algorithms are examined in this section : SVMs, Decision tree, Random forest, Logistic regression, K-neighbour, and Naive Bayes.

Confusion Matrix and Accuracies : **Accuracy** is the first metric we used to and is the most basic. It provides an answer to the question:
How often does the classifier get it right?
It is easily obtained by applying the following formulas:

$$Accuracy = \frac{\text{number of correctly classified items}}{\text{number of all classified items}}$$

Confusion Matrix is yet another indicator commonly used to assess the success of a classification algorithm. This metric was used in this study. If we were to utilize a confused matrix to forecast if an email is spam or not, we would have the following matrix:

	Predicted:RE	Predicted: SE
Actual:RE	TN	FP
Actual:SE	FN	TP

Where:

- RE = Real Email
- SE = Spam Email
- TN = True Negatives
- TP = True Positives
- FN = False Negatives
- TF = False Positive

The anticipated classes in the matrix's columns, are represented. The actual classes, on the other hand, are represented in the matrix's rows. There are four cases:

- TP : when the classifier expected spam and The emails were unquestionably spam.
- TN : where the classifier expected "not spam" and therefore the emails were real
- FP : where the classifier expected "spam," though The emails were authentic.
- FN : where the classifier expected "not spam," however the emails were spam.

True or false suggests if the classifier appropriately expected the class, while positive or negative shows whether or not the classifier efficaciously expected the goal class. The accuracy is then predicted by the use of following formula:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

The method of 10-fold cross validation was used to assess accuracy.
Figure 1 shows the results the accuracy obtained for the 6 different algorithms:

According to figure 1, SVMs outperformed all other algorithms with an accuracy of 95.56%, which means that the classifier got it right 95.56% times out of the entire testing data, followed by Random Forest with an accuracy of 94.23%, which means that the classifier got it right 94.23% of the time, and Decision Tree with an accuracy of 93%, which means that the classifier got it right 93% of the time,K-neighbor came in 4th with an accuracy of 87%, indicating that the classifier got it right 87% of the total testing data, followed by Naive Bayes with a 77% accuracy, indicating that the classifier was correct. 77% of the total testing data, and finally, Logistic Regression with an accuracy of 76%, indicating that the classifier was correct. 76% of the total testing data

Confusion Matrix for the Outperformed Model : SVMs

High Risk : The actual high risk is 1706. predicted high risk in actual high is 1637, out of 1706 actual high risk, 1637 are predicted correctly as high risk wheres 29, 21 and 19 were incorrectly predicted as highest, lowest and medium risk respectively.
Highest Risk : The actual highest risk is 1681. Predicted highest risk in actual higher is 1670, 1670 are predicted correctly as highest risk wheres 9 and 2 and incorrectly predicted as high risk and medium risk respectively.
Lowest Risk : The actual lowest risk is 1706. Predicted lowest risk in actual lowest risk is 1635, out of 1706 actual lowest risk, 1635 are correctly predicted

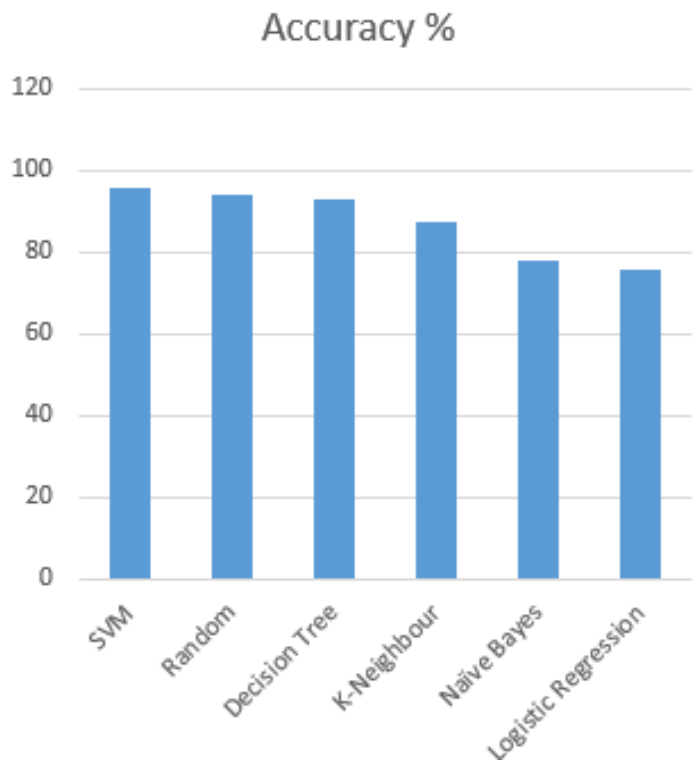

Fig. 1. Accuracy Bar Graph

col_0 RiskStatus	High Risk	Highest Risk	Lowest Risk	Medium Risk
High Risk	1637	29	21	19
Highest Risk	9	1670	0	2
Lowest Risk	28	0	1635	43
Medium Risk	69	14	71	1593

Fig. 2. Confusion Matrix for SVMs

wheres 28 and 43 are incorrectly predicted as high risk and medium respectively.
Medium Risk : The actual medium risk is 1747. Predicted medium risk in actual medium risk is 1593, out of 1747 actual medium risk, 1593 are correctly predicted wheres 69, 14 and 17 are incorrectly predicted as high, highest and lowest respectively.

Precision, recall and f1-score Aside from accuracy, the confusion matrix was used to construct a number of additional performance indicators.
Precision
Precision provides an answer to the question:
How often does it get it right when it forecasts the outcome?

This is accomplished through the application of the formula:

$$Precision = \frac{TP}{TP + FP}$$

When the goal is to reduce the quantity of FP, precision is typically used.

Recall
Precision provides an answer to the question:
How often does it anticipate properly when the outcome is positive?

This is accomplished through the application of the formula:

$$Recall = \frac{TP}{TP + FN}$$

When the purpose is to restrict the amount of FN, recall is frequently used.

F1-Score
Simply said, this is the arithmetic mean of precision and recall:

$$F1score = \frac{\text{Precision x recall}}{Precision + recall}$$

When both precision and recall are considered, the result is it is beneficial. If you only want to improve recall, your algorithm will predict that the majority of occurrences will fall into the positive category, however, this will lead to having a lot of false positives and there isn't a lot of precision. If you strive to maximize precision, on the other hand, Only a few positive cases will be predicted by your model. yet have a very poor recall.

The figure below shows the Precision, recall and f1-score Results for the 6 model employed for this study.

Fig. 3. Precision, Recall and F1-score

SVM has outperfomed all other model with an Accuracy of 95.54%, 96% Precision, 96% Recall and 95% f1-score.The precision and recall are significant metric scores, but it is difficult to maximize both of them, so one has to trade off one metric.

4.3 Interpretation

SVMs performs better mostly reason being that it offers superb consequences in phrases of accuracy whilst the information are linearly or non-linearly separable. When the information are linearly separable, the SVMs end result is a isolating hyperplane, which maximizes the margin of separation among classes, measured alongside a line perpendicular to the hyperplane.
Random Forest is one the models that gave good results in this study mostly reason being that a large number of relatively independent models (trees) operating in committee will prevail over the individual constitutive models. The key is the low correlation between the models.
Decision Tree had a high accuracy in this study, decision tree is one of the most used models due to the fact they wreck down complicated records into extra practicable parts.
due to the fact they wreck down complicated records into extra practicable parts.
KNN : If all of the data has the same scale, KNN performs substantially better. SMOTE was used to balance the data so that the performance for KNN can be maximized. KNN was performing poorly when that data was imbalanced
Naive Bayes : Naive Bayes gave better results however they were not good given that some models are giving way better results.The overall performance of Naive Bayes can degrade if the statistics includes fantastically correlated capabilities. This is due to the fact the fantastically correlated capabilities are voted for two times within side the model, over inflating their importance.

Logistic Regression : This model did not perform well compared to the rest of the models used in this study. The predominant predicament of Logistic Regression is the idea of linearity among the structured variable and the unbiased variables.

5 Conclusion

Our research into classifying students into the appropriate risk profiles using biographical (background), individual, and schooling variables revealed that biographical characteristics, followed by individual traits, have the greatest impact on student attrition or risk classification. Pre-college factors have little or no impact on determining student risk profiles. Similarly, the eight most significant (contributing) attributes are biographical and individual characteristics, according to. They serve a significant (essential) role in classifying students into the appropriate risk profiles, according to the conceptual model.When compared to models fitted with a controlled balanced class data set using the SMOTE technique, the results show that the fitted models performed well on an imbalanced class data set. This could be related to the fact that the smote method does not take into account neighboring samples from different classes when generating synthetic samples. It can then lead to class overlap and the introduction of additional noise. The fitted machine learning algorithms were able to recognize (deduce) the various risk profiles successfully. However, the positive class identification rate varies depending on the quantity of each class [7].

The findings show that the machine learning models used were able to estimate learner susceptibility based on the attributes provided [1] [14]. The SVMs outperformed all other models with an accuracy of 95.56% followed by random forest with 94.23%, decision tree with 93%, KNN with 87%, Naive Bayes with 77% and Logistic Regression with 76% . The least performing model on this study was found to be logistic regression.
The observe concludes that scholar attrition is stricken by biographical and individual attributes, and consequently those elements have to be considered in the better schooling enrollment system.
Significance : This study was important in assisting students in determining which level of risk the course they are taking has based on their biographical and pre-college,because many students do not consider their biographical and pre-college when choosing a course of study, and the majority of them do not complete their degree, which is unfortunate.
Future Work : This research can be expanded in a number of ways, including developing a model that will recommend the most appropriate course for a student's success, thereby reducing the number of students who drop out.

References

1. R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *Conference of the South African Institute*

of Computer Scientists and Information Technologists 2020, 2020, pp. 19–28.
2. E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3–12, 2012.
3. D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and information technologies*, vol. 13, no. 1, pp. 61–72, 2013.
4. B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system," in *33rd Annual Frontiers in Education, 2003. FIE 2003.*, vol. 1. IEEE, 2003, pp. T2A–13.
5. A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics," *Computer*, vol. 49, no. 4, pp. 61–69, 2016.
6. I. Tien and D. K. Armen, *Reliability Engineering System Safety.*
7. S.-W. Cho and M. Karp, "Student success courses in the community college," *Community College Review*, vol. 41, pp. 86–103, 02 2013.
8. J. Bossart and N. Bharti, "Women in engineering: Insight into why some engineering departments have more success in recruiting and graduating women." *American Journal of Engineering Education*, vol. 8, no. 2, pp. 127–140, 2017.
9. D. Kember, *Open learning courses for adults: A model of student progress.* Educational Technology, 1995.
10. G. M. Nicholls, H. Wolfe, M. Besterfield-Sacre, L. J. Shuman, and S. Larpkiattaworn, "A method for identifying variables for predicting stem enrollment," *Journal of Engineering Education*, vol. 96, no. 1, pp. 33–44.
11. A. S. Arnold, J. S. Wilson, M. G. Boshier, and J. Smith, "A simple extended-cavity diode laser," *Review of Scientific Instruments*, vol. 69, no. 3, pp. 1236–1239, 3 1998. [Online]. Available: http://link.aip.org/link/?RSI/69/1236/1
12. A. Berger, N. Adelman, and S. Cole, "The early college high school initiative: An overview of five evaluation years," *Peabody Journal of Education*, vol. 85, no. 3, pp. 333–347, 2010.
13. T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *International workshop on machine learning and data mining in pattern recognition.* Springer, 2012, pp. 154–168.
14. N. Ndou, R. Ajoodha, and A. Jadhav, "Educational data-mining to determine student success at higher education institutions," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC).* IEEE, 2020, pp. 1–8.
15. J. B. Caruso and G. Salaway, "The ecar study of undergraduate students and information technology, 2007," *Retrieved December*, vol. 8, p. 2007, 2007.
16. O. L. Herrera, "Investigation of the role of pre- and post-admission variables in undergraduate institutional persistence, using a markov student flow model," *Conference Paper*, vol. 1, no. 16, 06 2006.
17. T.-O. Tran, H.-T. Dang, V.-T. Dinh, X.-H. Phan *et al.*, "Performance prediction for students: A multi-strategy approach," *Cybernetics and Information Technologies*, vol. 17, no. 2, pp. 164–182, 2017.
18. G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
19. D. Yates and M. Z. Islam, "Fastforest: Increasing random forest processing speed while maintaining accuracy," *Information Sciences*, vol. 557, pp. 130–152, 2021.

20. T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006.
21. M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24–31, 2016.
22. Y. Qi, "Random forest for bioinformatics," in *Ensemble machine learning*. Springer, 2012, pp. 307–323.
23. A. M. Shahiri, W. Husain *et al.*, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.
24. T. Abed, R. Ajoodha, and A. Jadhav, "A prediction model to improve student placement at a south african higher education institution," in *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 2020, pp. 1–6.
25. N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? predicting customers' purchase behavior using deep learning," *Decision Support Systems*, vol. 149, p. 113622, 2021.
26. A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved random forest for classification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012–4024, 2018.
27. J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 71–72, 1996.
28. I. S. Damanik, A. P. Windarto, A. Wanto, S. R. Andani, W. Saputra *et al.*, "Decision tree optimization in c4. 5 algorithm using genetic algorithm," in *Journal of Physics: Conference Series*, vol. 1255, no. 1. IOP Publishing, 2019, p. 012012.
29. Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.
30. H. Lipyanina, A. Sachenko, T. Lendyuk, S. Nadvynychny, and S. Grodskyi, "Decision tree based targeting model of customer interaction with business page." in *CMIS*, 2020, pp. 1001–1012.
31. V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: an overview and their use in medicine," *Journal of medical systems*, vol. 26, no. 5, pp. 445–463, 2002.
32. C. Jin, L. De-Lin, and M. Fen-Xiang, "An improved id3 decision tree algorithm," in *2009 4th International Conference on Computer Science & Education*. IEEE, 2009, pp. 127–130.
33. R. Christensen, *Log-linear models and logistic regression*. Springer Science & Business Media, 2006.
34. Z. Hua, Y. Wang, X. Xu, B. Zhang, and L. Liang, "Predicting corporate financial distress based on integration of support vector machine and logistic regression," *Expert Systems with Applications*, vol. 33, no. 2, pp. 434–440, 2007.
35. K. Alkhatib, H. Najadat, I. Hmeidi, and M. K. A. Shatnawi, "Stock price prediction using k-nearest neighbor (knn) algorithm," *International Journal of Business, Humanities and Technology*, vol. 3, no. 3, pp. 32–44, 2013.
36. I. A. A. Amra and A. Y. Maghari, "Students performance prediction using knn and naïve bayesian," in *2017 8th International Conference on Information Technology (ICIT)*. IEEE, 2017, pp. 909–913.
37. S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

38. S. A. Pattekari and A. Parveen, "Prediction system for heart disease using naïve bayes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.
39. G. Subbalakshmi, K. Ramesh, and M. C. Rao, "Decision support in heart disease prediction system using naive bayes," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 2, pp. 170–176, 2011.
40. M. S. Vural and M. Gök, "Criminal prediction using naive bayes theory," *Neural Computing and Applications*, vol. 28, no. 9, pp. 2581–2592, 2017.
41. T. Wang and W.-h. Li, "Naive bayes software defect prediction model," in *2010 International Conference on Computational Intelligence and Software Engineering*. Ieee, 2010, pp. 1–4.
42. Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: A local naïve bayes model," *EPL (Europhysics Letters)*, vol. 96, no. 4, p. 48007, 2011.
43. W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
44. D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*. Elsevier, 2020, pp. 101–121.
45. A. Ukil, "Support vector machine," in *Intelligent Systems and Signal Processing in Power Engineering*. Springer, 2007, pp. 161–226.
46. F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.
47. Y.-D. Cai, P.-W. Ricardo, C.-H. Jen, and K.-C. Chou, "Application of svm to predict membrane protein types," *Journal of theoretical biology*, vol. 226, no. 4, pp. 373–376, 2004.
48. A. Garg and D. Gupta, "Virulentpred: a svm based prediction method for virulent proteins in bacterial pathogens," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–12, 2008.