The Influence of Clarity on Career Choice and Academic Success in First Year.
&
The Influence of Interest in a Career Field and Academic Success in First Year Biology Students.

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Xolani Mti - 1847445

Supervised by Dr R. Ajoodha Dr S. Dukhan

A project submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Bachelor of Science with Honours

Declaration

I Xolani MTi, (Student number: 1847445)
am a student registered for Research Project: Big Data Analytics in 2021.

This declaration applies to the Research Report document of
Research Project: Big Data Analytics.

I here declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that ALL the work submitted for assessment for the above course is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this in not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: Date: 27-Nov-2021

Contents

Chapter 1

Abstract

This report is a presentation of the research that seeks to use six machine learning algorithms to predict student success using the student's career choice, influences on their career choice, and the study approaches they use to learn as the features. This research is divided into 2 papers and both papers are still under consideration for publication in separate conferences. The first paper is on Chapter 2, in this paper we classified student success into two classes, pass or fail, using the student career choices, their interest factor that leads to the career choice, the learning style they use in university, and their results in first year. Using 10-fold cross-validation the support vector machine achieved, 74%, the highest accuracy against the other five models. The second paper is on Chapter 3, it is an extension of the first paper, where students' success is divided into three categories: low risk students, medium risk students, and high risk students, based on their biological science course results. Taking into account these features: career choices, the interest factor that led to the professional decision. Using 10-folds of cross-validation the simple logistic regression and the decision tree achieved, 82%, the highest accuracy against the other four models.

Chapter 2

The Influence of Clarity on Career Choice and Academic Success in First Year.

The Influence of Clarity on Career Choice and Academic Success in First Year.

Xolani Mti
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
1847445@students.wits.ac.za

Shalini Dukhan
School of Animal, Plant
and Environmental Sciences
The University of the Witwatersrand
Johannesburg, South Africa
shalini.dukhan@wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—In this study, we use various machine learning algorithms to understand the link between the grades that students achieve and their choice of career path within a first year biological sciences degree. This research will shed light on the students' academic success in relation to their motivation to pursue a career within the field which they are studying.**

Student success is classified into two classes, pass or fail, using the student career choices, their interest factor that leads to the career choice, the learning style they use in university, and their results in first year. We used six machine learning models to classify the students under two classes (Fail or Pass). The Support Vector Machine model had an accuracy of 74% which was the best model compared to the other 6 models, it had 10-fold cross-validation. The Simple Logistics model achieved the second-best accuracy of 73% also with 10-fold cross-validation.

***Index Terms*—Machine Learning, Career Choice, Student success, South Africa.**

1. Introduction

The South African educational system has become more equitable over the years after the institutionalised system of racial segregation, the apartheid system, came to an end in 1994. Disciplines such as science are now growing the pool of talent entering the domain. Despite the fact that the number of first-year students in higher education is increasing, data reveal that only a small percentage of those individuals qualified for a degree [1]. Challenges with decisions on career choice could be the underlining issues for the high rate of retention of students in science. We could gain more insight into why high attrition rates are still prevalent today by determining to what extent first-year students who are confident of their career path perform better than first-year students who are uncertain of their career path.

This research study will help in the prediction of a first-year student's success in the field of science based on their career path decisions and interests. The aim is to predict a student's academic achievement in biological science, depending on several findings of career path choices and desires. We will also determine the best machine learning model to predict student success using the student's career choice, influences on their career choice, and the study methods they use to pass their exams.

Using the collected information of students about their foremost career choices we classified what career path they would like to pursue, whether they wanted to pursue science or non-science career pathways and what are the factors that triggered interest in their chosen career pursuit, and their first year mark. We classified the student's success into two classes: Pass or Fail. We represented the results from the six machine learning models using accuracy and confusion matrices.

The contributions of this paper are: institutions can use an early detection tool to determine students who are at risk, contributes literature by demonstrating that one of the most important factors for academic success is student motivation and interest. The learning approaches students use helps them to learn more effectively. First year students have little guidance at university and they use the learning approaches they learnt in high school, and so they rely on their high school experience to assist them in putting together their notes. It is also possible that the learning approaches of students could be linked to their motivation to pursue a career within their field of study. Thus, this link could influence the grade that students achieve [2]. Universities are aware of the importance of student data to predict attrition rate as a result of scientific advancements in Machine Learning. Several studies apply a number of machine learning models, and from those models, the model with the best accuracy is used to identify students at risk [3], [4].

We use the students' career choice (science or non-science), factors that triggered their career path, the learning style(s) they use in university, and their first semester results to predict their success. The machine learning model with the best accuracy is the support vector machine, which achieved an accuracy of 74%. The second-best model is the simple logistics model, with an accuracy of 73%, which uses 10-fold cross-validation. The summary of the data collected is shown in tables 1, 2, and 3. Since these are first year students, it is not surprising that 36% prefers to use their own notes for studying. The quality of notes is related to the student academic achievement [2]. The data was collected from students who enrol into a general Biological Sciences degree 92% of them want to work in the field of science and 53% are self-interested in the field.

The structure of the document is as follows. The Related work section has an overview of previously published work, allowing us to find relevant theories, methods, and research gaps; the Methodology section highlights the data processing, features, and classification models used in this paper; the Results section explains the results; and the Conclusion section outlines the study, highlights our contributions, and makes suggestions for further research.

2. Related Work

This section reviews prior published work on the topic of student success and/or machine learning models. We will first look at the data used by several papers who had significant contributions towards predicting student success. Secondly, we will look at the list of prominent features which have been successful to predict success, and finally we will look at the machine learning models used to classify.

A. Data

A number of studies have been conducted to predict student success and, different papers defined student success in different ways and different, in terms of features, data sets have been used to predict student success [5]. Most departments within higher education institutions have access to demographic data for their student cohorts (e.g. gender, race, langauge) and some recent studies still use demographic data to predict student success [6], however surveys such questionnaires can be used to collect more nuanced data [7]. To predict student success, Zeineddine at el [3] analyzed information from admissions, registrar, and student service departments containing records of 1491 students.

The data collected by researchers can sometimes be biased, usually representing selection biased, where, for example, only a large percentage of academically achieving students are selected to predict student success. Zeineddine et al [3] collected 1491 records of student data and 1014 had a good academic standing. A data balancing technique had to be applied to ensure data balancing an unbiased prediction of the machine learning models. The use of SMOTE (Synthetic Minority Oversampling Technique) to balance the data has been used in recent student attrition studies [8].

B. Features

Student's career choice is very complicated because there are a number of factors that influence their choice, and background knowledge of a career is essential for fostering and developing interest in that career [9]. Family and friends have the most influence towards a student's career choice, followed by gender, media, financial difficulties and others. [10]. Research and consultations; training, exposure, and industrial attachment are ways to gain knowledge about a career. Mentorship, commitment, motivation, and socialising with people in the field of interest are other ways for people to gain foreknowledge for their careers.

Learning styles consist of different ways in which students gather and assimilate information. Age, achievement level, culture, and gender all affect learning styles [11]. There is a strong correlation between learning effectively and academic success and most students use the approach writing notes [12].

York, Gibson, and Rankin [5] explores the use of the term academic success in different educational research. Using the Inputs-Environments-Outcomes model academic success consists of academic achievement, knowledge building, skills, competence, persistence and retention [13]. Academic success is the measure of student's outcomes of academic work, such as course grades. Learning includes the student's cognitive development and effective skills.

The definition of academic success is complex and broad [5]. Academic attributes are mostly used in research for predicting and analyzing student success, other studies use external factors like demographic information, family, financial resources, and more. Most studies use easily accessible student features like age, gender, ethnicity, study program, course load, on/off-campus residency, academic probation, and school educational system [3]. While other studies take it further by considering the students' views on the methods they use to learn and the students' views of their learning experience, for example, see [14]. Using questionnaires, researchers are able to get more insight into students. Thus, more researchers rely on questionnaires (constructed using certain models) instead of simple biographical data accessible through the high institution's database. Recent studies consider using more detailed features and this proves and disproves models that explain and predict student attrition. McKenzie, Kirsten Schweitzer, and Robert [15] collected and annotated documents with textual citations during the literature review, with particular attention paid to how the author(s) described academic success, as well as what metrics were used in academic success in any empirical studies the data categorised into broad categories like "grades," "critical thinking," and "effective outcomes.".

Ajoodha, Jadhav, and Dukhan [4] use the Tintos framework [16] to analyse biographical and enrollment features, considering that the framework lists (1) Background or Family attributes, (2) Individual attributes, and (3) Schooling attributes as input factors to student attrition. Student institution integration is found to be negatively correlated with academic achievement, implying that students with high levels of integration into university are more likely to have low grade averages, contradicting Tinto's concept [15]. Academic success does not always imply retention, and low academic performance does not always imply attrition [15]. The explanation to this finding firstly is that there is a difference between academic success and attrition, and secondly, Tinto's model was created to explain student attrition, and scholars used Tinto's framework to explain academic performance.

C. Models

Machine learning is used to predict student attrition, it makes use of historical data to train its models. The models used, to predict student attrition, include classification models like decision trees, Naive Bayes, Logistic regression, etc. The prediction accuracy for each algorithm differs and usually, it depends on the features used. Determining which algorithm to use can be strenuous and other studies use Automated Machine Learning to predict student success to obtain the best accuracy [3]. The algorithms used are the Logistic regression, Artificial Neural Network, Naive Bayes, Decision Tree, Support Vector Machines, and K-Nearest Neighbour. Numerous machine learning classification models such as Naive Bayes, support vector machine, logistic regression, linear and logistic regression are used and the algorithm with the best accuracy is chosen to measures student success [4]. Most of the time five to six different machine learning models are used to predict student success [17].

3. Methodology

In this study, we attempt to predict student success using the student's career choices, their interest factor that leads to the career choice, the learning style they use in university, and their results in the first year. Using the student features we classify success in two classes: (1)*Fail* where the student has obtained less than 50% for their final mark in a course, (2)*Pass* where the student has obtained more or equal to 50% and above for their final course mark. We used various machine learning algorithms to classify the student into one of the two classes. The confusion matrix and the algorithm's accuracy to accurately classify will be used to evaluate each algorithm's performance.

A. Data Processing and Collection

In this section, we will describe the steps taken in data collection and pre-processing. The data was collected from a South African university from first year students who enrol in a general Biological science degree in 2019. The dataset is a sub-set of a larger dataset that was collected by one of my supervisors. Since the general Biological science degree has a variety of career options available the information about the student career choices, their interest factor that leads them to the career choice, the learning style they use in university, and their results in first year were collected. The dataset used contained 260 rows, and 17 columns (16 predictor variables and 1 target variable). The target variable had 2 classes, Pass and Fail, each with 130 instances. The predictor variables are explained below.

B. Features

First year students who enrol on a general Biological Sciences degree were asked questions about their learning style, career choice and factors that lead them to choose that career. For each student five learning approaches were collected and grouped in table 1, four career choices were collected and grouped in table 2, and four factors that lead them to choose a career were collected and grouped in table 3.

Table 1: Describe your learning approach that you use at university.

	Percentage
1. Use my own notes	36%
2. Focus on my understanding	18%
3. Prioritising tasks ahead	11%
4. Unsure	10%
5. Use resources on the Internet (e.g. YouTube)	6%
6. Used of lecture slides	6%
7. Use the prescribed textbook	6%
8. Practice questions (like Pearson online exercises)	5%
9. Peer assistance	1%

Table 2: What career field would you like pursuit.

	Percentage
Science	92%
Non-Science	8%

C. Classification and Evaluation

To analyse and compare the machine learning models we will use confusion matrices, the accuracy of the model and its classification error. An example of a confusion matrix is shown in table 4. The predicted and actual classification are shown in a confusion matrix of dimension 2x2 connected with a classifier [18]. The meaning of the entries in table 4 are:

TN is True Negative which means the number of correct negative predictions.
FP is False Positives which means the number of incorrect positive predictions.
FN is False Negatives which means the number of incorrect negative predictions.
TP is True Positives which means the number of correct positive predictions.

Table 4: Two-class confusion matrix [18]

	False Negative	False Positive
True Negative	TN	FP
True Positive	FN	TP

The accuracy and the classification error of the models will be obtained from the confusion matrix (Table 4) defined as:

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{1}$$

$$Error = \frac{FP + FN}{TN + FP + FN + TP} \tag{2}$$

Table 3: What factor/s triggered your interest in your chosen career pursuit.

	Percentage
1. Self-interest in the field	53%
2. To make an impact on the field	15%
3. Not sure	9%
4. Personal background	7%
5. To have a social status	6%
6. Job opportunities	6%
7. Consider the career achievable	4%

Six machine learning models will be used to predict student success. The following models will be used.

a) Naïve Bayes model: The probabilistic algorithm Naive Bayes is commonly used for classification problems. The probability can be expressed mathematically as follows:

$$P(A/B) = \frac{(B/A) * P(A)}{P(B)} \tag{3}$$

Naive Bayes is a simple, understandable algorithm that performs well in many situations. The approach is intended for use in supervised induction tasks in which the goal is to reliably predict the class of test instances, and the training instances contain class information, the implementation is adopted from [19]. It's a classification method based on Bayes' Theorem and the assumption of predictor independence.

b) Multi-layer Perceptron model: Multilayer perceptrons (MLP) are a type of deep artificial neural network that has a hidden layer with a nonlinear activation function. The MLP architecture is made up of three layers, each with nodes, including an input layer that defines the input value, one or more hidden levels that specify the mathematical function, and an output layer [20]. In the MLP used in this paper, all the nodes are sigmoid (Only if the class is numeric, in which case the output nodes are non-thresholded linear units.).

c) Simple Logistic Regression model: In a simple logistic regression model, a covariate X_1 is associated to a binary response variable Y in a model

$$log(\frac{P}{(1-P)}) = \beta_0 + \beta_1 X_1 \tag{4}$$

where $P = probability(Y = 1)$. The null hypothesis $H_0 : \beta_1 = 0$ is being tested against the alternative $H_1 : \beta_1 = \beta^*$, where $\beta^* \neq O$ denotes that the covariate is related to the binary answer variable [21]. The implementation is based on [22], [23].

d) Support Vector Machine model: The support vector machine (SVM) is a supervised machine learning model that uses classification techniques to handle two-group classification problems. It uses a multi-dimensional hyper-plane to partition the training data into classes. We used the popular SVM implementation (LibSVM) from [24].

e) Decision tree model: Ross Quinlan devised the $C4.5$ technique for generating decision trees. Quinlan's earlier Iterative Dichotomiser 3 (ID3) algorithm is extended in C4.5. The decision trees in C4.5 can be used for classification, which is why it's also known as a statistical classifier. In this paper we implemented the $c4.5$ from [25].

f) Random Forest model: During training, the model generates a large number of decision trees and outputs the class that is the mean/average prediction (regression) of the individual trees, or the mode of the classes (classification). The implementation is based on [26].

4. Ethics Clearance

The University's Human Research Ethics Committee has approved the study's ethics application. The ethics application handles important ethical issues such as safeguarding the identity of study participants and data protection. The protocol number for the clearance certificate is CSAM-2021-02W.

5. Results

Initially the data contained 22 variables (see Table 5) and for dimensional reduction we used the principle component analysis(PCA). Out of the 22 variables PCA returned 18 new combination features and the first feature had a standard deviation of 1.82, we divided the standard deviation by 2 which gave us 0.91. From the 18 new features we selected 11 features that had a standard deviation greater or equal to 0.91. Finally, the 11 features were used for machine learning classification.

Table 5: Student variables used in classification models.

Variable name
1. Student's gender
2. First career choice
3. Second career choice
4. Third career choice
5. Fourth career choice
6. What career field would you like pursuit.
7. First interest factor that lead to choosing a career
8. Second interest factor that lead to choosing a career
9. Third interest factor that lead to choosing a career
10. Fourth interest factor that lead to choosing a career
11. What factor/s triggered your interest in your chosen career pursuit.
12. Fist Learning style
13. Second Learning style
14. Third Learning style
15. Fourth Learning style
16. Fifth Learning style
17. Describe your learning approach that you use at university.
18. Block 1 exam mark
19. Block 1 final exam mark
20. Block 2 exam marks
21. Block 2 final exam mark
22. Semester final exam mark

We discuss results from the six machine learning algorithms used for classification. We used the Naïve Bayes, MLP, Simple logistics, SVM, Decision tree and Random forest. Fig. 1 shows the confusion matrices and accuracy, using 10-folds of cross validation, of each model.

The Confusion matrix (a) represent the results from the Naïve Bayes model which achieved an accuracy of 73%. It took the shortest amount of time to complete.

The Confusion matrix (b) represent the results from the Multi-layer perceptron model which achieved the worst accuracy of 71%. It took the longest amount of time to complete.

The Confusion matrix (c) represent the results from the Simple Logistic regression model which achieved the second best accuracy of 73%.

The Confusion matrix (d) represent the results from the Support Vector Machine model which achieved the best accuracy of 74%.

The Confusion matrix (e) represent the results from the Decision tree model which achieved an accuracy of 72%.

The Confusion matrix (f) represent the results from the Random Forest tree model which achieved an accuracy of

72%. The accuracy is the same as the accuracy achieved in the confusion matrix (e).

Using PCA to reduce the number of features, from 23 features to 11 features, improved the accuracy of our models. Fig. 2 shows an accuracy of the Support Vector Machine before the data dimension was reduced. The accuracy of SVM was 53% before PCA. Fig. 2 compared to Fig. 1 only incorrectly classified more students who passed. The classification for students who failed is the same on both confusion matrices. This means each feature has a contribution to student success. Thus considering the combination of all the features enables us to predict student success.

		Predicted Class	
		Pass	Fail
Actual Risk	Pass	32	98
	Fail	23	107

Figure 2: Confusion matrix of **Support Vector Machine** before PCA was used for dimensionality reduction. The model has an accuracy of 53% and an inaccuracy of 47% from the data set of 270 number of instances

6. Discussion and Conclusion

This study contributes to literature by examining whether student motivation and interest are the key factors for academic success, and It can be used to identify at-risk first-year students.

Using machine learning algorithms we predicted if a student will pass or fail and from the high accuracy results, we can see that these factors do influence if a student will succeed or not, whether they will pass or fail. The use of PCA improved the accuracy of the models, this shows that student success does not depend on one factor, but a combination of factors. The importance of providing a clear career direction as a target is related to the student's motivation to pursue a career within the field which they are studying, in this case science or non-science field. Therefore, students' need more exposure to different careers so that they can have a clear career direction.

In this study like most recent publications we used machine learning models to predict student success. But to make it a new research we used several machine learning models and we predicted student success using the student career choices, their interest factor that leads to the career choice, the learning style they use in university, and their results in first year.

It is also worth noting the study's limitations. The amount and quality of the collected data determine the predictive accuracy, the data used was collected from one institution and it contains only a small sample(270) of the first year students and so the results can not be generalised for all South African universities. Subsequent research should consider a large sample size and from different institutions.

Acknowledgment

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 121835 and 121960).

References

[1] S. Stats, "Education series volume v: Higher education and skills in south africa, 2017," 2019.

[2] S. Dukhan, "Note-making in biology: How the school experience influences note-making practice and approach at university," *African Journal of Research in Mathematics, Science and Technology Education*, vol. 22, no. 3, pp. 265–275, 2018.

[3] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," *Computers & Electrical Engineering*, vol. 89, p. 106903, 2021.

[4] 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.

[5] T. T. York, C. Gibson, and S. Rankin, "Defining and measuring academic success," *Practical assessment, research, and evaluation*, vol. 20, no. 1, p. 5, 2015.

[6] S. Batool, J. Rashid, M. W. Nisar, J. Kim, T. Mahmood, and A. Hussain, "A random forest students' performance prediction (rfspp) model based on students' demographic features," in *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. IEEE, 2021, pp. 1–4.

[7] P. Dass-Brailsford, "Exploring resiliency: Academic achievement among disadvantaged black youth in south africa," *South African Journal of Psychology*, vol. 35, no. 3, pp. 574–591, 2005.

[8] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and smote," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[9] J. Nyamwange, "Influence of student's interest on career choice among first year university students in public and private universities in kisii county, kenya." *Journal of Education and Practice*, vol. 7, no. 4, pp. 96–102, 2016.

[10] A. S. Kazi and A. Akhlaq, "Factors affecting students' career choice." *Journal of Research & Reflections in Education (JRRE)*, vol. 11, no. 2, 2017.

[11] E. Collinson, "A survey of elementary students' learning style preferences and academic success," *Contemporary Education*, vol. 71, no. 4, p. 42, 2000.

[12] M. Gokalp, "The effect of students learning styles to their academic success," *Educational Research and Reviews*, vol. 8, no. 17, pp. 1634–1641, 2013.

[13] A. Astin, "Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education. american council on education/macmillan series on higher education." 1991.

[14] S. Dukhan, "Value for learning during this time of transformation: the first-year students' perspective," *Higher Education Research & Development*, vol. 39, no. 1, pp. 39–52, 2020.

[15] K. McKenzie and R. Schweitzer, "Who succeeds at university? factors predicting academic performance in first year australian university students," *Higher education research & development*, vol. 20, no. 1, pp. 21–33, 2001.

[16] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research*, vol. 45, no. 1, pp. 89–125, 1975.

[17] J. Kuzilek, Z. Zdrahal, and V. Fuglik, "Student success prediction using student exam behaviour," *Future Generation Computer Systems*, vol. 125, pp. 661–671, 2021.

[18] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection." *MAICS*, vol. 710, pp. 120–127, 2011.

[19] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[20] G. Sahoo and Y. Kumar, "Analysis of parametric & non parametric classifiers for classification technique using weka," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 4, no. 7, p. 43, 2012.

[21] F. Y. Hsieh, D. A. Bloch, and M. D. Larsen, "A simple method of sample size calculation for linear and logistic regression," *Statistics in medicine*, vol. 17, no. 14, pp. 1623–1634, 1998.

[22] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer, 2005, pp. 675–683.

		Predicted Class	
		Pass	Fail
Actual Risk	Pass	76	54
	Fail	18	112

(a) Confusion matrix of **Naive bayes model**. The model has an accuracy of **72%** and an inaccuracy of 28% from the data set of 270 number of instances

		Predicted Class	
		Pass	Fail
Actual Risk	Pass	89	41
	Fail	35	95

(b) Confusion matrix of **Multilayer Perceptron model**. The model has an accuracy of **71%** and an inaccuracy of 29% from the data set of 270 number of instances

		Predicted Class	
		Pass	Fail
Actual Risk	Pass	76	54
	Fail	15	115

(c) Confusion matrix of **Simple Logistic model**. The model has an accuracy of **73%** and an inaccuracy of 27% from the data set of 270 number of instances. It has the same results as the Naive Bayes.

		Predicted Class	
		Pass	Fail
Actual Risk	Pass	86	44
	Fail	23	107

(d) Confusion matrix of **Support Vector Machine model**. The model has an accuracy of **74%** and an inaccuracy of 26% from the data set of 270 number of instances

		Predicted Class	
		Pass	Fail
Actual Risk	Pass	83	47
	Fail	26	104

(e) Confusion matrix of **Decision tree model**. The model has an accuracy of **72%** and an inaccuracy of 28% from the data set of 270 number of instances

		Predicted Class	
		Pass	Fail
Actual Risk	Pass	92	38
	Fail	35	95

(f) Confusion matrix of **Random Forest model**. The model has an accuracy of **72%** and an inaccuracy of 28% from the data set of 270 number of instances

Figure 1: Confusion matrices measuring the performance of the 6 classification models. The accuracy of each classification model is shown, as well as the occurrences that have been correctly and inaccurately classified.

[23] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," vol. 95, no. 1-2, pp. 161–205, 2005.

[24] Y. EL-Manzalawy and V. Honavar, *WLSVM: Integrating LibSVM into Weka Environment*, 2005, software available at http://www.cs.iastate.edu/ yasser/wlsvm.

[25] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

Chapter 3

The Influence of Interest in a Career Field and Academic Success in First Year Biology Students

Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET)
9-10 December 2021, Cape Town-South Africa

The Influence of Interest in a Career Field and Academic Success in First Year Biology Students

Xolani Mti
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
1847445@students.wits.ac.za

Shalini Dukhan
School of Animal, Plant
and Environmental Sciences
The University of the Witwatersrand
Johannesburg, South Africa
shalini.dukhan@wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—In this research, we utilize a variety of machine learning methods to investigate the relationship between students' career choice and grades during their first year of biological sciences. This study will examine students' academic achievement in connection to their desire to pursue a profession in the subject they are studying. Students' success is divided into three categories: low risk students, medium risk students, and high risk students, based on their biological science course results. Taking into account these features: career choices, the interest factor that led to the professional decision. Six machine learning models were employed to divide the Students' success into three groups. With 10-fold cross-validation, the simple logistic and the decision tree model scored the highest accuracy of 82%, while the multi-layer preceptron model earned the lowest accuracy of 76%.**

***Index Terms*—Machine Learning, Interests, Career Choice, Academic achievement, Classification.**

I. Introduction

There is an increase in the number of first year students entering the field of science but data reveal that only a small percentage of those individuals are eligible for a degree [1]. This research will help forecast a first-year student's success in the field of science based on their career goals. When students are interested, they will be eager to participate in learning [2].We want to find the best machine learning algorithm for classifying student performance based on the student's career choice and the factors that influence it.

Using the information collected from students who are studying biological science. We divided the student's achievements into three categories: low risk, medium risk, and high risk. The low risk students rank 65% and above for their first semester course mark. The medium risk students ranked between 65% and 50% above for their first semester course mark. The high risk students are ranked below 50% for their first semester course mark. A semester represents the firs 6 months of learning. We used accuracy and confusion matrices to depict the results of six machine learning models.

We use the student's interest in a field (science or non-science) and their marks in a biological science course to predict their success in the first six months of study. The results from the machine learning models ranged from 76%-82% using 10-fold cross validation, the simple logistic model and the decision tree model scored the highest accuracy of 82% and the Decision tree model scored the lowest accuracy of 76%. Table I, table II, and table III show the features used in the machine learning model. The following is a breakdown of the document's structure. The next section provides an overview of work that solves a different problem with the same proposed solution.

II. Related Work

This section summarizes previous research on student achievement and/or machine learning models. We'll start by review the features that have been successful in predicting student achievement, then go over the data utilized by numerous studies that have made substantial contributions to predicting student success. We will analyse the machine learning algorithms used for classifying student performance.

A. Feature

Due to funding opportunities, more people qualify to obtain a higher education qualification students' international mobility is increasing, and they're opting for more flexible study options like online, off-shore, and part-time classes [3]. Renninger and Hidi [4] defined interest as a psychological condition characterized by an effective response to and concentrated for a long-term attention for certain subject. Depending on a student's level of interest in a subject, [4] proposes that there are several forms of interest and achievement relationships, it also implies that pupils might be encouraged to develop an interest in and work with subject matter in which they initially had little interest. Eko, Ari and Yarmani observed students' activities in an online class to analyse their interest, motivation, and learning outcomes of sports sociology. They concluded that blended learning, the combination of online and face-to-face learning, and the jigsaw technique increases students' interest (70% of students had high interest), motivation, and learning outcomes in sports sociology [2].

B. Data

Several studies have been undertaken to predict student success, and different papers defined student success in different ways and used different data sets to predict student success in terms of attributes [5]. Most departments in higher education institutions have access to demographic data for their student cohorts (e.g., gender, race, and language), and some recent research [6] continue to use demographic data to predict student success. Synthetic data created by the machine learning model is used in studies like to predict student achievement [7].

Researchers' data can be skewed in some cases, mainly due to selection bias, such as when just a large number of academically successful students are chosen to predict student progress.Zeineddine, Braendle, and Farah [8] used student enrolment statistics from admission, registrar, and student service offices, as well as records of 1491 students to predict student performance. To achieve data balancing and an unbiased prediction of the machine learning models, a data balancing technique needs to be used. Recent studies like [9] have used SMOTE (Synthetic Minority Oversampling Technique) to balance the data.

C. Models

To predict student attrition, machine learning employs past data to train its models. Classification techniques such as decision trees, Naive Bayes, Logistic regression, and others were employed to predict student attrition. Using six machine learning algorithms [7] utilized students' grade 12 grades to predict whether they would perform well in each year of study until they received their degree.The prediction accuracy of each algorithm varies, and it is usually determined by the features employed. Choosing which algorithm to employ can be difficult, and previous related work found that Automated Machine Learning is the most accurate method for forecasting student achievement [9]. The methodology section describes the data processing, features, and classification models used in this paper.

III. Methodology

In this study, we use the student's profession choice, the interest factor that leads to the career choice, and their first-year results to try to predict student success. We divide students into three groups based on the final semester mark for biology: (1) low risk students - those who obtained more than or equal to 65% for their final semester grade in a course, (2) medium risk students - those who obtained less than 65% but more or equal to 50% for their final semester grade in a course, and (3) high risk students - those who obtained less than 50% for their final semester grade in a course. We will use six machine learning algorithms to classify the students and use the algorithm's accuracy in accurately classifying students and the confusion matrix to assess each algorithm's performance.

A. Data processing and collection

This section explains the data collection and pre-processing of the data. The information was gathered from first-year students enrolled in a general Biological Science degree at a South African institution in 2019. We have classed the professional choices available with a biological science degree as either science or non-science. The dataset is a subset of a larger dataset that one of my supervisors generated. The dataset contained a number of features, but for the purpose of this research we use the interest of the students as the main features. We use the gender feature and the fact that the biology students either want to work in science or non-science field, and their interests factors that led them to the career choice, and their first semester(6 months) results in a biology course.

B. Features

Students enrolled in a standard Biological Sciences degree were asked about their career choice and factors that influenced their decision. Table I shows the career choices of the students in the dataset and table II shows the four factors that led a student to choose a career, these factors are grouped as either self-interest or other interest.

Table I: What career path would you like pursuit.

	Percentage
1. Science	91%
2. Non-Science	9%

The two groups in the interest feature are composed as follows: Self-interest:

- To have a significant impact in the field.
- Self-interested in the field.

Other interests:

- To have a social standard.
- there are Job opportunities in the field.
- influenced by personal background.
- the career is considered attainable.
- not sure which career to choose.

Table II: What piqued your interest in your chosen profession?

	Percentage
1. Self-interested in the field	71%
2. Not self-interested in the field	29%

C. Classification and Evaluation

The confusion matrix and accuracy of each model will be used to analyze and compare the machine learning models. The equation for accuracy is show in equation 1

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{1}$$

To predict student achievement, six machine learning models will be employed. The following is a list of the models that were used.

a) Naïve Bayes model: The Naïve Bayes assumes that features of a class are independent, i.e P(X—C)= $\prod_{i=1}^{n} P(X_i|C)$, where $X = (X_1, X_2, ..., X_n)$ are features, and C is the class. Although the assumption of independence is often incorrect, Naïve Bayes often outperforms more complex classifiers in practice. The model has proven to be useful in a variety of fields, including medical diagnosis, text classification, and system performance management [10]. As in most paper, we use the NaBayes as a benchmark model. We used the Naïve Bayes model that was implemented in [11].

b) Multi-layer Perceptron model: Multi-layer perceptron(MLP) model are useful When you have minimal information of the shape of the relation between the independent and dependent variables. MLP allow for nonlinear mapping of input and output vectors, it uses non-linear activation function, such as logistic function. MLPs are commonly trained using a technique called the *generalized delta rule*, which computes derivatives using a simple *back-propagation* application of the chain rule [12] .

c) Simple Logistic Regression model: The definition of the word regression means the measure of the relation between the mean-value of the independent variable and the corresponding dependent variable. There are two regression models: logistic regression and linear regression. The process of fitting a simple model to the data in logistic regression is highly stable, resulting in low variance, but potentially greater bias [13], [14]. The simple logistic regression model simply relates the covariate X_1 to the binary target variable Y in a model $log(\frac{P}{(1-P)}) = \beta_0 + \beta_1 X_1$, where $P = probability(Y = 1)$. We test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 = \beta^*$, where $\beta^* \neq O$ indicates that the covariate is related to the binary answer variable [15]. The model used here is based on the implementation on related work; [16], [13].

d) Random forest tree model: The random forest tree is mostly applied in classification, regression, and unsupervised learning [17]. The algorithm contains tree-structured classifiers represented in Eq.2, where Θ_k represents independent identically distributed random vectors and each tree ranks the most occurring class at input X. Though, random features as input produce good results in classification that is not the case in regression [17]. [17] implemented the algorithm presented in this paper.

$$h(X, \Theta_k), k = 1, ...) \qquad (2)$$

e) Decision tree model: The type of decision tree used in this paper is C4.5 which is the extension of the Iterative Dichotomiser 3(ID3) algorithm. The algorithm is implemented by [18]. The training data consists of a collection of already categorized samples $S = s_1, s_2, ...s_n$. Each sample $s_i = x_l, x_2, ..., x_m$ is a vector, with $(x_l, x_2, ...x_m)$ denoting n sample features with m vectors. A vector $(C = c_l, c_2, ...)$ is added to the training data, where $(c_l, c_2, ...)$ represents the class to which each sample belongs [19].

f) Sequential Minimal Optimization: SMO (sequential minimum optimization) is a training method for Support Vector Machines that removes the requirement for additional matrix storage and a huge quadratic programming (QP) optimization problem. SMO makes use of the lowest possible QP problems, which are solved easily, allowing it to scale and compute faster [20]. This paper's method is based on [21], which normalizes attributes and converts nominal attributes to binary.

IV. Ethics Clearance

The University's Human Research Ethics Committee has approved the study's ethics application. The ethics application handles important ethical issues such as safeguarding the identity of study participants and data protection. The protocol number for the clearance certificate is CSAM-2021-02W..

Acknowledgment

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 121835 and 121960).

V. Results

The results of the machine learning algorithms will be discussed in this section. The features used in this shown on table III, the target variable which is the "Block 2 final exam mark" was categorised in 3 equal classes. Each class (Low risk students, medium risk students, and high risk students) had 80 instances.

Table III: Student features used in classification models.

Feature name
1. Student's gender.
2. What career field would you like pursuit.
3. First interest factor that lead to choosing a career.
4. Second interest factor that lead to choosing a career.
5. Third interest factor that lead to choosing a career.
6. What factor/s triggered your interest in your chosen career pursuit.
7. Block 1 exam mark.
8. Block 1 final exam mark.
9. Block 2 exam marks.
10. Block 2 final exam mark.

We used the Naïve Bayes, simple logistics, random forest tree, Multi-layer perceptron, decision tree, and sequential Minimal Optimization(SMO). Confusion matrices are graphs that show how well each of the six categorization models performs. For each algorithm we showed the accuracy and the confusion matrix.

Fig. 1 conveys the results from the six machine learning algorithms. We used the confusion matrices and accuracy. Each model runs on 10-folds of cross validation. The Fig. 1(a) shows the results from the Naïve Bayes model which had 78% accuracy. Fig. 1(b) shows the results from the multi-layer perceptron model which had 76% accuracy. Fig. 1(c) shows the results from the simple Logistic regression model which had the highest accuracy, 82% accuracy. Fig.

1(d) shows the results from the random forest tree model which had the second best accuracy of 81%. Fig. 1(e) shows the results from the decision tree model which also had the highest accuracy, 82% accuracy. Fig. 1(f) shows the results from the sequential minimal optimization model which also had the second best accuracy of 81%.

VI. Discussion and Conclusion

This study adds to the body of research by evaluating if self-interest in a profession is a determinant for academic performance and may be used to identify first-year students who are at risk. We employed machine learning models to predict student achievement in our study, as in other previous papers. However, in order to make it a different study, we utilized various machine learning models to predict student performance based on the student's profession choice, the interest factor that led to the chosen profession, and their first semester results.

It's also important to consider the study's limitations. The quantity and quality of the data gathered affect the prediction accuracy; however, because the data was collected from only one university and only a small sample of first-year students (240) was used, the results cannot be applied to all South African universities. Following that, future research should use a big sample size and come from a variety of universities.

The necessity of setting a clear career path as a goal is related to a student's desire to pursue a career in the subject in which they are studying, whether science or non-science. As a result, students require greater exposure to various professions in order to develop a clear professional path.

References

[1] S. Stats, "Education series volume v: Higher education and skills in south africa, 2017," 2019.

[2] Y. E. Nopiyanto, A. Sutisyana, S. Raibowo, and Y. Yarmani, "Blended learning with jigsaw in increasing interest, motivation, and learning outcomes in sports sociology learning," *Kinestetik: Jurnal Ilmiah Pendidikan Jasmani*, vol. 5, no. 1, pp. 26–34, 2021.

[3] S. Gamlath, "Peer learning and the undergraduate journey: a framework for student success," *Higher Education Research & Development*, pp. 1–15, 2021.

[4] K. Ann Renninger and S. Hidi, "Chapter 7 - student interest and achievement: Developmental issues raised by a case study," in *Development of Achievement Motivation*, ser. Educational Psychology, A. Wigfield and J. S. Eccles, Eds. San Diego: Academic Press, 2002, pp. 173–195. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780127500539500097

[5] T. T. York, C. Gibson, and S. Rankin, "Defining and measuring academic success," *Practical assessment, research, and evaluation*, vol. 20, no. 1, p. 5, 2015.

[6] S. Batool, J. Rashid, M. W. Nisar, J. Kim, T. Mahmood, and A. Hussain, "A random forest students' performance prediction (rfspp) model based on students' demographic features," in *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. IEEE, 2021, pp. 1–4.

[7] 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.

[8] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," *Computers & Electrical Engineering*, vol. 89, p. 106903, 2021.

[9] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and smote," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[10] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[11] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[12] W. S. Sarle, "Sas institute inc., cary, nc, usa,"," in *Neural Networks and Statistical Models", Proceedings of the Nineteenth Annual SAS Users Group International Conference*, 1994.

[13] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," vol. 95, no. 1-2, pp. 161–205, 2005.

[14] G. Sahoo and Y. Kumar, "Analysis of parametric & non parametric classifiers for classification technique using weka," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 4, no. 7, p. 43, 2012.

[15] F. Y. Hsieh, D. A. Bloch, and M. D. Larsen, "A simple method of sample size calculation for linear and logistic regression," *Statistics in medicine*, vol. 17, no. 14, pp. 1623–1634, 1998.

[16] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer, 2005, pp. 675–683.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[19] 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.

[20] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.

[21] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

		Predicted Class		
		Medium	Low	High
Actual Risk	Medium	56	9	15
	Low	7	70	3
	High	20	0	60

(a) Confusion matrix of **Naive bayes model**. The model achieved **78% accuracy**.

		Predicted Class		
		Medium	Low	High
Actual Risk	Medium	49	6	25
	Low	8	72	0
	High	17	1	13

(b) Confusion matrix of **Multilayer Perceptron model**. The model achieved **76% accuracy**.

		Predicted Class		
		Medium	Low	High
Actual Risk	Medium	59	2	19
	Low	10	69	1
	High	12	0	69

(c) Confusion matrix of **Simple Logistic model**. The model achieved **82% accuracy**.

		Predicted Class		
		Medium	Low	High
Actual Risk	Medium	59	4	17
	Low	8	72	0
	High	18	0	62

(d) Confusion matrix of **Random Forest Tree**. The model achieved **80% accuracy**.

		Predicted Class		
		Medium	Low	High
Actual Risk	Medium	53	4	23
	Low	7	73	0
	High	9	1	70

(e) Confusion matrix of **Decision tree model**. The model achieved **82% accuracy**.

		Predicted Class		
		Medium	Low	High
Actual Risk	Medium	56	1	23
	Low	12	67	1
	High	9	0	71

(f) Confusion matrix of **Sequential Minimal Optimization**. The model achieved **81% accuracy**.

Figure 1: Confusion matrix and accuracy of each of the six models used for classification. The dataset contained 240 number of instances