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 Decision tree 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 Naíve 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.

$$h(X, \Theta_k), k = 1, ...) \tag{2}$$

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