To what extent do mark expectations of first years align with their academic achievement within biology based degrees at a South African university?

Siyabonga Hlomuka
School of Computer Science and Applied Mathematics
The University of the Witwatersrand Johannesburg, South Africa
1384685@students.wits.ac.za

Shalini Dukhan
School of Animal, Plant and Environmental Sciences
The University of the Witwatersrand Johannesburg, South Africa
shalini.dukhan2@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*—The drop-out rate in South African universities is of concern because it is is costing the government substantial amounts of funding. There are historical reasons why there is a high drop-out rate, these reasons stem from the Apartheid era and how it had an impact on our education system. In this study we aim to shed light on whether the marks that students anticipate achieving are aligned to the marks that they obtain on assessments. Examining this link could shed light on whether first-year students have realistic views of the academic demands at university. This is done by training 6 different models namely Bagging, Random Forest, Decision table, Logistic Regression, Naive Bayes and Multi-Layer Perceptron. All the models are trained using a K-Fold cross validation K=5 to achieve a 80:20 split of training and test data. The best performing model is the Logistic Regression model with 0.692982 accuracy but with the problem in had it is found that the best model to use is the Decision table due to it's ability to classify students that failed the year with a recall of 0.719 for class "Fail".**

I. Introduction

STUDENT ENROLLMENT in South African universities has increased over the years, there has been a significant drop-out coupled with the increase of registrations, see[1]. Consequentially this has led to billions of rands in the form of grants and subsidies costing the National Treasury without a return in their investment, see[1].This high drop out rate resonates with the inequalities created by the Apartheid regime which impacted on the quality of education that was provided to the majority of the native population, see[2], thus it is even more important to help students at risk in there entrance year of university by assessing their academic performance throughout the first year, in doing so we might improve on the current drop-out rate.

This research is about identifying the accuracy of classification models in predicting first-year students' academic performance based on their mark expectation of that year. The factors that will be considered in the classification system include students' mark expectation taken twice throughout the first year, the first taken at the start of first semester and the second at the beginning of second semester, as well as the mark final year-end mark. Based on the accuracy of the classification models, it may be possible to predict student academic success.

The research question thus is can we identify classification models that will correctly classify students at risk based on their mark expectation. This leads on to the research aim to predict a student academic performance in the first year of university based on their mark expectation for that year within a first-year biology course.

The goal is to identify models that will correctly classify students who pass and fail, with more emphasis on the students that fail, this will reveal the students at risk which is our goal. The models will be assessed on the accuracy and how good it is in classifying students at risk.

II. Related Work

There has been much research done relating to the prediction of academic performance for students within institutions of higher education. Most of the research focuses on trying to predict academic performance using factors like socio-demographics as seen in [2], [3], [4] and previous marks attained from high school, see [5], [1]. From previous research, see [6], [7], [8], [9], [10], [11], [12], [5], [2], [3], [13], [14], [14], [1], [15], [4], it is noticeable that predicting students' performance from the start of a students' studies is very important. Applications if this is achieved can lead to early student interventions to try and help students at risk, thus providing academic support on time.

A. Data

Most of the data work that has been reported in literature was attained from High school and the universities the students attended. The high school data was the final results of the students, the graduation results, see [5], [6], [15]. The bulk of the data was attained from the universities only, the data from universities is comprised of academic records as seen in [16], [8], [9], [15], [7] and survey or questionnaires like [3], [2] whilst a minor of the data is synthetic data like [14], [1],

synthetic data is data that was made up to closely resemble the real data. The rest of the data is from public databases like the Kalboard 360 LMS(Learning Management System), see [12].

B. Features

Most of the dominant features include the high school grade (matric results in the case of South Africa). A few papers also included SAT scores like [16], [5], these papers were written based on Higher education institutions that are based abroad (e.g. America). SAT scores are used as predictors to predict academic performance in first year and staying the course for the first year, greater SAT score means higher probability to stay the whole first year course, see [16]. Whilst the SAT score in another paper was used to determine which students where likely to pass first year that major in computer science, see [5], In this paper the SAT score along with gender proved to be the best predictors. Another common feature was the initial academic record of the first years e.g. first semester marks. Papers like [5], [4], [9], [8] all utilized initial higher education academic results to predict future performance. The use of final results like degree outcome and year end first-year results are also used as the dependent variables to train models, see [6], [1]. Social demographics is a dominant feature in most papers related to my question, factors like location which the students come from, if it is a poor neighbourhood or wealthy one, the high school type, if it is a private and well rated high school or not etc. as seen in [2] and [3] place their focus on social economic backgrounds like family's state of finance and prior students' families involvement in the Higher Education, a close inspection of cultural and social capital as well as students' perception in their learning ways. Also worth noticeably both Dukhan papers [2], [3] and [1] focus on post apartheid attempts to correct the education systems in order to have more equality. Other papers also utilize background information, see [11], [15], [7], which focused more on location of high school, [12] which place more focus on students attitude and [4] is also another paper which utilized social demographic. Choice of academic course is another recurring feature used to predict future academic performance, see[14], [1]. Then we have some unique features like prior computer experience seen in [15] and involvement within the school systems e.g. online access to course material and forum participation, see[10].

C. Models

Naive Bayesian models are used quiet often in papers related to my research question [9] uses Bayesian Networks in Weka to predict performance, they compare the Bayesian network model against the Descision tree, Descision tree performed better. In articles [11], [14], [12], [1], [6], [15] all utilize Naive Bayes model, Naive Bayes model is one of the model used to classify the data, it does so through probalistic classifiers and an assumption of independence between features, it does not perform particularly well compared to other model, neither does it perform the worst. The same is seen of the Linear Logistic Regression but in one particular article the Logistic Regression performed the best, see[12]. Neural Networks is a popular model as well, in some article all the attention was dedicated to see the performance of the Neural Networks model, see [10], [8], [7],[8], all models were shown as accurate predictors. The setup of the Neural network seen in [10] is The Multilayer Perceptron trained using the gradient descent backwards propagation, it utilize the independent variables as input and the dependent variable as output, the output being 'success' or 'failure' of the course. In the article [8] they use regression model that predicts the actual mark the students will get, the network uses sigmoid in the input and hidden layers and utilizes a non linear activation on the output layer, the training of this model is done using Levenberg-Marquardt back propagation algorithm, similarly the neural network from [7] utilized the sigmoid function but on all levels and used Cuckoo Optimization Algorithm to train the model. The Decision Tree was another reoccurring model of classification used performing relatively well in some articles like[9], [12], [6]. The Random forest model was not used much but whenever used it mostly turned out to be the best model to use for classification of academic performance, see[11], [14], basically the random forest is a collection of decision trees put together to create the random forest, it is very good in avoiding over fitting of the model. Other least used algorithm that proved very powerful if used was the Bootstrap Aggregating(Bagging) as seen in[14] "Bootstrap Aggregating(Bagging) is a meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over fitting".

D. Accuracy

When it comes to accuracy of model classification the best performing model is the Neural Networks models with a 94% and above classification accuracy, see [10], [8], [7]. This model correctly provides the means to predict future academic performance. Another powerful model is the Random Forest model with an 82% accuracy, see[11], others perform well but none as well as the random forest and then Neural networks models, the Decision Tree and Bagging are the only other models that perform relatively well with 70% above accuracy, see[15] for decision tree and[14], [1] for Bagging.

E. Significance

This article [5] is one of the first papers to tackle predictions of students' academic performance, the models used are old and new models are available that can yeild better results. This paper [5] is a base for future studies. The paper[2] is very closely related to this research as it goes briefly into mark expectation vs actual marks, this research however purely focuses on mark expectation vs actual marks. More on mark expectation this paper [3] focuses on students' views and perception towards their academics which is closely linked with students mark expectation. A few of the papers focus on social demographic/biographical features which are

also a feature I aim to investigate in my Question, see[4], [15]. The article [16] is very similar to this research but this research assesses academic prediction every semester, the article [11] is important to this paper because it has the same goal to intervene on first-year students at risk based there predictive models. Worth also pointing out that all papers are connected to this research due to all the other papers investigating Higher Education students' performance.

Therefore the proposed research question 'To what extent do mark expectations of first years align with their academic achievement within biology based degrees at a South African university?' has not been adequately answered based on the research papers assessed, there has been instance where some papers eluded to the idea of mark expectation and how it will effect the performance of the student but no study has deeply hone in on this research question. Thus with that in mind the research question is one worth asking due to the inadequate answer to it.

III. Methodology

In this research we aim to classify students' academic performance by attributes like student mark expectation,the data used in this research will undergo pre-processing in order to allow for clear classification of students. Afterwards the data will be ran on machine learning classification models with K-fold cross validation. The model performances will be assessed through confusion matrix, accuracy and recall in order to determine the best model.

A. Data collection and pre-processing

The used data was previously collected by a lecturer in biology first-year course at Witwatersrand. The data was collected through means of survey at beginning of the year as well as the start of second semester.
The original data has 189 student participant, that is 189 entries. The 189 participating students were asked to give an expected year-end mark for their first year at university in a biology course. There was two instances where the student mark expectation was taken, the first students' mark expectation was taken at the begging the first semester, the second students' mark expectation was taken at the begging of the second semester. All the mark expectation taken down was is a percentage form from 0% to 100%. The data also has the actual final-year end marks obtained by the 189 participating students.

In the preprocessing phase the actual year-end final marks obtained are used to come up with a new feature called "Qualified". Qualified has two classes namely "pass" and "fail", class "pass" is year-end final marks greater than 49%, class "fail" is year-end final marks less than 50%. Class "pass" has 132 instances and class "fail" has 57 instances.
Since the classes of our new feature was unbalanced, the data was balanced by choosing at random 57 instances of class "pass", leaving us with 114 entries, 57 from class "pass" and 57 for class "fail".

TABLE I
Final Attributes

Qualified
Mark expectation 1
Mark expectation 2

As stated the data is from a previous study and thus requires an ethics clearance to be used. ethical procedures have been followed according to the University's Ethics board.

B. Attribute Selection

The data comes with the following attributes, namely students' mark expectation 1 taken at the beginning of semester 1 and students' mark expectation 2 taken at the beginning of semester 2 as well as the Final year-end mark and the derived feature "Qualified". The two mark expectations as well as feature "Qualified" are the selected attributes to be used for training our models.

The explanatory variables are the students' mark expectation(taken at the beginning of each semester of the first year) and the explained variable is "Qualified" which was derived from the final year-end mark. This means we have 3 Features namely: Mark expectation 1, Mark expectation 2 and Qualified.

C. Classification Models

- Bagging: An ensemble technique. The bagging works by running the same learning algorithm on different subsets from the same data, the aggregate prediction is taken as the true one.
- Logistic Regression: The logistic regression is the same as the linear regression but with the predicted values bounded between 0 and 1.
- Naive Bayes: Works by applying probability, particularly the Baye's theorem, to certain outcomes given an occurrence of another
- Random Forest: Uses an ensemble technique. It works by aggregating many decision trees, that was trained from subsets of the same data, the aggregate is taken as the true predictive value
- Decision table: Works by assessing a combination of conditions that will lead to a certain outcome.
- Multi-Layer Perceptron(MLP): Is a feed-forward artificial neural network with activation functions of perceptron in each layer(at least 3 layers in network)

D. Performance Metric

In order to evaluate the performance of models we will use evaluation metric like the confusion matrix, accuracy and precision.

- Confusion matrix: Contains true positive, true negative, false positive and false negative. The true means correctly classified values and false is incorrectly specified values.

- Accuracy: This tells you the percentage of correctly classified data from the test data.
- precision: A ratio telling you out of the positive classified values what proportion is correct
- Recall: A ratio telling you out of the actual positive values what proportion is correctly classified

E. Ethics Clearance

The study participants were learners who studied at a South African higher-education institution. The study ethics application has been approved by the university's Human Research Ethics Committee (Non-Medical). The ethics application addresses key ethical issues of protecting the identity of the learners involved in the study and ensuring security of data. The clearance certificate protocol number is CSAM-2021-03W.

IV. Results

The results are from 6 machine learning classification models. The attributes used are students' expected year-end results taken on two occasions, the begging of first and second semester. With the explained variable being 'Qualified', 'Qualified can take on two instances,'Pass' and 'Fail'.

The models used K-fold cross validation to train the models, we used 5-fold in order to achieve a 80:20 split of training and test data, this split is done on 114 instances of the whole dataset. The results are presented on Table 1 and table 2.

The best model in terms of accuracy is the **Logistic Regression(LR)** model with a **69.3%** accuracy. When looking at the confusion matrix we can see that **71.9%** of the students that passed where correctly classified, thus **71.9%** is the recall for class 'Pass'. When looking at the students who failed, **66.7%** of them where correctly classified, thus **66.7%** is the recall for class 'Fail'. This shows that the LR model is great at classifying students that actually passed correctly more so then those who failed just based on their mark expectation for the year.

The second best model is the **Decision Table(DT)** model with an accuracy of **67.54%**. When it comes to classifying the students that failed the DT performed the best with a recall of **71.9%**.

The **Naive Bayes(NB)** is the third best performing model with an accuracy of **65.79%**.When it comes to the classification of students that failed, recall for class 'Fail', the recall is **68.4%** which is the second best.

The **Bagging** model is the forth best based on accuracy then the **Multi-Layer Perceptron(MPL)** and **Random Forest(RF)** are the worst performing models.
All the detailed results are presented in Table 1 and Table 2.

TABLE II
Confusion Matrix

Model	Confusion Matrix				Accuracy
		Predicted outcome			
		Fail	Pass	Class	
LR	Actual outcome	38	19	Fail	69.2982 %
		16	41	Pass	
DT	Actual outcome	41	16	Fail	67.5439 %
		21	36	Pass	
NB	Actual outcome	39	18	Fail	65.7895 %
		21	36	Pass	
Bagging	Actual outcome	38	19	Fail	64.0351 %
		22	35	Pass	
MLP	Actual outcome	36	21	Fail	63.1579 %
		21	36	Pass	
RF	Actual outcome	36	21	Fail	63.1579 %
		21	36	Pass	

TABLE III

Model	Precision	Recall	Class
DT	0,661	0,719	Fail
	0,692	0,632	Pass
NB	0,650	0,684	Fail
	0,667	0,632	Pass
Bagging	0,633	0,667	Fail
	0,648	0,614	Pass
LR	0,704	0,667	Fail
	0,683	0,719	Pass
MPL	0,632	0,632	Fail
	0,632	0,632	Pass
RF	0,632	0,632	Fail
	0,632	0,632	Pass

V. Conclusion

From the resulting models we observe that the Logistic Regression model performed best based on accuracy and the given dataset. With the aim of our research, being able to detect students at risk of failing the year based on their mark expectation, the Logistic Regression model is not the best model since we are trying to detect students at risk from our models. Hence want a model that correctly classified students that failed, looking at the recall of class 'fail' the Logistic Regression model has recall **66.7%**, which is not the best from our 6 trained models. The best at classifying students that failed is the Decision Table model with a recall of **71.9%**, meaning that **71.9%** of the students that failed were correctly classified under fail. Thus the second best model, which is

the Decision table based on accuracy, is the best given what we want to achieve. It has an accuracy of **67.54%** but the highest recall for class 'fail'. The idea behind identifying a classification model is to come up with a system where we can utilize the classification model to identify students at risk of failing the year then intervening in time to try and help the student to pass the year, in doing so we aim to archive a better yield in the National Treasury money, in the form of bursaries and subsidies, as well as a decrease the drop out rate. Finally to address the question "To what extent do mark expectations of first years align with their academic achievement within biology based degrees at a South African university?", to quantify the extent to which mark expectations of first years align with their academic achievement, achievement meaning passing or failing the year, we would have to look at the accuracy of the best performing model which is the Logistic Regression with an accuracy of **69.2982**%.

The limitation in the research is that the dataset is only from one course, Biology in first year at Witwatersrand, it is thus hard to tell if this dataset is a true representation of all the courses, not to mention every University in South Africa. Therefore more data that is diversified across different courses at University of Witwatersrand might be a better sample representation of the Universities in South Africa. Also the research only focuses on mark expectation in determining students performance, it does not take into account allot of other factor that might results in students dropping out of university like psychological issues and money problems. This however serves as a good baseline for further study into this area.

ACKNOWLEDGEMENT

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

REFERENCES

[1] N. Philippou, R. Ajoodha, and A. Jadhav, "Using machine learning techniques and matric grades to predict the success of first year university students," in *The International Multidisciplinary Information Technology and Engineering Conference*, 2020, pp. 1–5.

[2] S. Dukhan, A. Cameron, and E. A. Brenner, "The influence of differences in social and cultural capital on students' expectations of achievement, on their performance, and on their learning practices in the first year at university," 2012.

[3] S. Dukhan, "Value for learning during this time of transformation: the first-year students' perspective," in *Higher Education Research and Development*, 2020, pp. 39–52.

[4] T. Thiele, A. Singleton, D. Pope, and D. Stanistreet, "Predicting students' academic performance based on school and socio-demographic characteristics," in *Studies in Higher Education*, 2016.

[5] P. F. Campbell and G. P. McCabe, "Predicting the success of freshmen in a computer science major," *Communication ACM*, vol. 27, no. 11, 1984.

[6] R. Asif, S. Hina, and S. Haque, "Predicting student academic performance using data mining methods," *IJCSNS International Journal of Computer Science and Network Security*, vol. 17, no. 5, 2017.

[7] J.-F. Chen, H.-N. Hsieh, and Q. Do, "Predicting student academic performance: A comparison of two meta-heuristic algorithms inspired by cuckoo birds for training neural networks," in *Creative Commons Attribution*, 2014.

[8] A. R. Iyanda, O. D. Ninan, A. O. Ajayi, and O. G. Anyabolu, "Predicting student academic performance in computer science courses: A comparison of neural network models," in *I.J. Modern Education and Computer Science*, 2018, pp. 1–9.

[9] N. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in *Computer Science and Information Management Program Asian Institute of Technology*, 2007.

[10] N. Z. Zacharis, "P redicting student academic performance in blended learning using artificial neural networks," *International Journal of Artificial Intelligence and Applications*, vol. 7, no. 5, 2016.

[11] R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *SAICSIT*, 2020.

[12] E. Buraimoh, R. Ajoodha, and K. Padayachee, "Application of machine learning techniques to the prediction of student success," in *National Research Foundation of South Africa*, 2021.

[13] D. Koller and N. Friedman, "The bayesian network representation: Bayesian networks," in *Probabilistic Graphical Models: Principles and Techniques*, 2009, pp. 51–68.

[14] J. Mabunda, A. Jadhav, and R. Ajoodha, "Predicting student success at various levels of their learning journey in a science programme," in *National Research Foundation of South Africa*, 2021.

[15] T. Ramaano, R. Ajoodha, and A. Jadhav, "Different models relating prior computer experience with performance in first year computer science," in *National Research Foundation of South Africa*, 2021.

[16] B. A. FRIEDMAN and R. G. MANDEL, "The prediction of college student academic performance and retention: Application of expectancy and goal setting theories," *J. COLLEGE STUDENT RETENTION*, vol. 11, 2009.