Different Models Relating Prior Computer Experience with Performance in First Year Computer Science

Thabo Ramaano
School of Computer Science and Applied Mathematics
The University of the Witwatersrand,
Johannesburg
South Africa
1134539@students.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

Ashwini Jadhav
Faculty of Science
The University of the Witwatersrand,
Johannesburg
South Africa
ashwini.jadhav@wits.ac.za

***Abstract*—According to a South African university, the current minimum admission requirements to study a degree in computer science is 70 percent or above in pure mathematics and 60 percent and above in English. This shows that students are not evaluated based on their computer skills. The evaluation of a student's computer skills could provide one with information on whether or not the student is comfortable with using a computer. This may have an impact on whether a student passes or fails first year computer science. In order to investigate this further, two groups of students were considered. The first group had no prior computer experience while the second group had prior computer experience. Based on the accuracies of three classification models, we were able to determine group 2 as a better predictor of performance in first year computer science with accuracies of more than 60 percent for each model. Furthermore, a hypothesis test and confidence interval test was conducted. This was used to establish whether or not the final computer science results of the second group were greater than or equal to the first group. The hypothesis test and confidence interval test resulted in students in the second group performing better as compared to students in the first group at a significant level of 5 percent (95 percent level of confidence).**

KEY WORDS

***Index Terms*—Prior computer experience, Pearson correlation coefficient, Linear regression, Logistic regression, Naïve Bayes model, Decision tree classifier, Performance in first year computer science, Secondary school results**

INTRODUCTION

To enter the computer science programme in South African universities, potential students are required to have performed well in both pure mathematics and English as confirmed by [10] and [12]. The aim of this investigation is to determine whether or not the evaluation of a student's past computer experience along with their results in both mathematics and English is a better predictor of performance in first year computer science as compared to only evaluating their mathematics and English results.

This study will compare the performance of first year computer science students with prior computer experience to students without prior computer experience. Four machine learning models will be used. The linear regression model will allow us to predict the actual results of students. The outcome produced from the model is a real number [9]. The classification models which include: the logistic regression model; the decision tree classifier and the naïve Bayes model will allow us to predict a student's results as either a *PASS* or *FAIL*.

Similar studies have been done in finding effective predictors for performance in first year computer science. Language performance as a predictor was studied by [10]. The study used the Pearson product-moment correlation to establish language performance as a better predictor as compared to mathematics.

Similar to [10], [12] used the Pearson product-moment correlation to establish the strength between mathematics results and performance in computer science. [5] also investigated mathematics results as a predictor and was confirmed as a good predictor for performance in computer science.

A past study that investigated past computer experience as a predictor was by [7]. The study could not establish any significant relationship between computer-related subjects and performance in computer science [10]. For the purpose of our study, we consider past computer experience prior to entering first year as our predictor and not first year computer-related subjects.

The predictors of concern in this investigation are grade 12 results in pure mathematics; English first language and computer studies. We will be applying our different models to our chosen predictors. We do this to evaluate the strength of past computer experience as a predictor in comparison to the known predictors, namely, mathematics and English. We will evaluate the accuracy of each chosen model on the predictors. More on the methodology will be discussed in section 3.

Students are evaluated on the basis of their results in both grade 12 pure mathematics and English, however, they are not evaluated based on their computer skills. The contribution made by this study will be to show whether or not the

combination of a students' results in mathematics; English and prior computer experience is an effective way to evaluate a potential student to study first year computer science as compared to only evaluating their results in mathematics and English.

In section 2, we will be elaborating on the past research studies that have done similar work in using past results as a predictor of performance in computer science. We will evaluate the different features and models used as well as results from the respective past research studies. In section 3, we will provide a detailed outline of the methodology used in conducting our investigation. Section 4 will provide results from our investigation. Finally, we end with a conclusion and recommendations for future work based on this investigation.

Related Work

There has been several past research studies that have studied potential predictors of performance in first year computer science. Some studies even went as far to study the perception that students have towards studying computer science [13]. For the purpose of our investigation, we will only be studying grade 12 results in pure mathematics, English first language and computer-related subjects as potential predictors. For each related work that we discuss, we will be concerned with the features; the models used as well as the accuracy of the models from each investigation.

Table 1 is displaying information pertaining to the features, models and accuracies for previous research studies that have investigated at least one of the features: mathematics; English and past computer experience as predictors of performance in first year computer science. The logistic regression model, used by [2], produced the highest accuracy (80 percent) out of all the research studies listed in *Table 1*.

TABLE I
Table displaying a summary of previous research studies that have investigated past computer experience; mathematics and English as predictors of performance in computer science

Authors	Features	Models	Accuracy
[2]	MATH12	LogR	80 %
[4]	MATH12	LinR; PPMC	< 0.3
[4]	COMS12	LinR; PPMC	> 0.04
[4]	ENGFL12	LinR; PPMC	< 0.3
[5]	MATH12; ENGFL12	PPMC; DA; WL	68.4 %
[8]	COMS12	Regression	Between 45% and 65 %
[10]	MATH12; ENGFL12	PPMC	Refer to Table II
[1]	MATH12; COMS12; ENGFL12	LogR: SVM: NB: KC: MP:	59 % 59 % 58 % 57 % 62 %

Key for *Table I*
MATH12: *Grade 12 pure mathematics results*
ENGFL12: *Grade 12 English first language results*
COMS12: *Prior Computer experience*
PPMC: *Pearson product-moment correlation*
LogR: *Logistic Regression model*
LinR: *Linear Regression model*
SVM: *Supprort Vector Machine model*
NB: *Naïve Bayes model*
KC: *K* Classification*
MP: *Multilayer Perceptron*
DA: *Discriminant Analysis*
WL: *Wilks' Lambda*

Mathematics as a predictor

The first predictor that we concern ourselves with is mathematics. This is a known predictor of performance for computer science as confirmed by [12]. This investigation studied several features that may have an impact on results for computer science [12]. One of the features was the final secondary school results in mathematics [12]. The Pearson product-moment correlation was used to establish a relationship between results in mathematics and computer science. Results showed that a significant relationship existed between performance in mathematics and computer science [12].

Results from [5] were similar to that of [12]. Amongst the features studied included: background in high school mathematics and science; SAT scores as well as gender [5]. The study used the discriminant analysis; Multivariate t-test as well as the Wilk's Lambda and correlation coefficient to conduct the investigation [5]. Results from this investigation confirmed the findings of [12].

Past computer experience as a predictor

One would expect past computer experience to be the best predictor of performance in computer science. A study by [4] produced results contradictory to what many would expect. The study investigated several predictors including mathematics and prior computer experience [4]. Mathematics did not disappoint as a predictor. However, prior computer experience was found to have no impact to the performance in computer science [4]. The study tested using the Pearson product-moment correlation; ANOVA and regression. The Statistical Analysis System (SAS) was used to perform all the statistical analysis [4].

Other studies which investigated prior computer experience as a predictor was by [7] and [8]. [7] used the multiple regression analysis to confirm performance in computer related modules as a weak predictor of performance in computer science. However, [8] found prior computer experience to have a significant relationship with overall performance in computer science when using a regression model.

Language performance as a predictor

Another useful predictor is language performance. English is known as one of the admission requirements to study

computer science and this was verified by [10]. The results produced from this investigation were really interesting. It showed language performance as a better predictor of performance in computer science as compared to mathematics [10]. Similarly to the study by [12], [10] used the Pearson product-moment correlation to establish a relationship between its chosen predictors and performance in computer science modules. It was successful in realising language performance as a better predictor as compared to mathematics as shown in *Table II*.

TABLE II
Table sourced from [10] that shows the results from using Pearson's correlation. The independent variables are: mathematics; English first language; English second language and all other first languages. The dependent variables are: BCO and FAC

Category	BCO to FAC	Category to FAC	Category to BCO	n
Mathematics	0.7607	0.2664	0.2782	90
English first language	0.7667	0.4571	0.4063	48
English second language	0.7755	0.2343	0.1232	46
All first language	0.7651	0.3213	0.2440	96

Key for *Table II*
BCO: *Basic Computer Organisation*
FAC: *Fundamental Algorithmic Concepts*

Moving on to section 3, we will be discussing the methodology used to conduct our investigation. This will include the information about the data collection process; the features to be used; the models that will be employed as well as the measures of accuracy for each of our models.

METHODOLOGY

In this section, we will be discussing the methodology used to conduct our investigation. We will be including the data collection process; the features used; the models used as well as the measures of accuracy. Four different machine learning models were used in this investigation to make our predictions. These included: the linear regression model; the logistic regression model; the decision tree classifier and the the naïve Bayes model. We used confusion matrices; F1 score; recall; precision and the Pearson correlation coefficient to measure the accuracy of our models. We also used a hypothesis test and confidence interval test to confirm our findings.

We divided the features into two groups *(Table III)*. Group 1 consisted of students with mathematics and English first language but with no prior computer experience. Group 2 consisted of students with mathematics, English first language and prior computer experience.

We used the hypothesis test and confidence interval test to test whether or not the students in group 2 outperformed the students in group 1.

TABLE III
Table displaying the groups considered for this investigation. We have divided the data into two groups namely: Group 1 and Group 2

Group number	Group Description
Group 1	Group consists of students with no prior computer experience. (We consider only the results for pure mathematics and English first language)
Group 2	Group consists of students with prior computer experience. (We consider the results for pure mathematics; computer studies and English First language)

The structure of the methodology process is as follows: firstly, we introduce our data collection process. Secondly, we discuss the features that we used in the investigation. Thirdly, we introduce the machine learning models that we used along with their respective measures of accuracy. Next, we discuss the hypothesis test and confidence interval test that we conducted. Lastly, we include information about the ethics clearance certificate that was obtained for this research.

Data collection and pre-processing

The data consisted of high school results; biographical information; university results and registration information of students from a South African university. The data had 14326 samples.

The focus of this study was on computer science majors, therefore, data pertaining to computer science majors were extracted from the original 14326 samples. This decreased the number of samples to 428. The new sample consisted of students who either had prior computer experience or did not have prior computer experience.

Features and target value

Table IV shows the features used in the investigation. We used the final aggregate first year computer science results as our target value (the variable we are aiming to predict).

TABLE IV
Table displaying the features used in this investigation

Features
English first language grade 12 results
Pure mathematics grade 12 results
Computer studies grade 12 results

Past studies by [12], [5] and [10], suggested mathematics and English were good predictors so we used our own data to confirm this.

Linear Regression Model

$$Y = B_0 + B_1 X \tag{1}$$

We used this model to predict the aggregate of a student's final results using testing data. *Equation 1* shows the linear

equation used in the linear regression model to predict the final aggregate results (Y) of students given their feature results (X). B_0 represents the y intercept while B_1 represents the gradient. *Equation 1* resulted in predicted values (final aggregate results) given a feature in *Table IV*. The accuracy of our model was calculated using the Pearson correlation coefficient similar to what was done in past studies, [10], [12] and [4] to analyse the strength of the correlation between each feature in *Table IV* and final aggregate results in first year computer science.

Logistic Regression Model

This classification model was used to predict whether a student had passed or failed based on the results of the chosen predictors as shown in *Table IV*. For the purpose of this investigation, an aggregate score of 50 percent and above was considered a *PASS*. Furthermore, an aggregate score of less than 50 percent was considered a *FAIL*. A confusion matrix was constructed and displayed for the group (*Table III*) which produced the highest accuracy. From the confusion matrix, we were able to evaluate the F1 score; recall; accuracy and precision of the logistic regression model.

Naïve Bayes Model

This classification model was used to classify a student as having passed or failed based on the results of each chosen feature. To use this model, the features needed to be independent of each other and the data needed to be normally distributed [11]. A confusion matrix was constructed and displayed for the group (*Table III*) which produced the highest accuracy. From the confusion matrix, we were able to evaluate the F1 score; recall; accuracy and precision of the naïve Bayes model.

Decision Tree Classifier

The decision tree classifier was used to classify a student as having passed or failed based on the results of each chosen feature. We constructed a decision tree using the gini index as well as using entropy. Based on the results of the students from the chosen features (depending on the group), a decision was made on whether a student had passed or failed. A confusion matrix was constructed and displayed for the group (*Table III*) which produced the highest accuracy. From the confusion matrix, we were able to evaluate the F1 score; recall; accuracy and precision of the decision tree classifier model.

Hypothesis Test and Confidence Interval Test

We constructed a hypothesis test at a significant level of 5 percent to test whether or not the students in group 2 outperformed the students in group 1. Our null hypothesis was that the aggregate results of the two groups were equal. The alternative hypothesis was that the aggregate results of students in group 2 were greater as compared to students in group 1. A 95 percent confidence interval was constructed alongside the hypothesis test to validate our results.

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 the security of data. The clearance certificate protocol number is H19=03=02.

RESULTS AND DISCUSSION

We have used four different machine learning models in this investigation. We also constructed a hypothesis test and confidence interval test which gave us interesting results. Results from the hypothesis test showed that students with prior computer experience performed better as compared to students without prior computer experience using a significant level of 5 percent. The confidence interval test produced similar results to the hypothesis test at a 95 percent level of confidence.

Linear Regression Model

The results in *Table V* shows the correlation between each feature with the final first year computer science results as was done in studies by [12], [10], [5] and [4]. Grade 12 pure mathematics produced the greatest correlation with a value of 0.35. It was followed by computer studies with 0.32 and finally English first language with 0.16. Mathematics and computer studies proved to be the best predictors as compared to English first language.

TABLE V
Table displaying the Pearson correlation coefficient for pure mathematics, English first language and computer studies with the final first year computer science results

Modules (features)	**Pearson Correlation Coefficient**	**n (sample size)**
Pure mathematics grade 12 results	0.346117	214
Computer studies grade 12 results	0.320650	119
English First Language grade 12 results	0.164604	214

Logistic Regression Model

We constructed a logistic regression model and applied it to the two groups in *Table III*. Results in *Table VI* indicate that group 2 produced the highest accuracy by correctly classifying 62.5 percent of students' results as either a *PASS* or *FAIL*. A confusion matrix was constructed for group 2, as seen in *Figure 1*, as it was the group which produced the highest accuracy when using the logistic regression model.

TABLE VI
Precision, recall, F1 score and accuracy from the logistic regression model

Group	Precision	Recall	F1 score	Accuracy
Group 1	58.71 %	66.42 %	62.33 %	59.85 %
Group 2	62.33 %	63.21 %	62.76 %	62.5%

		Predicted Outcome	
		Pass	Fail
Actual Outcome	Pass	134	81
	Fail	78	131

Fig. 1. Confusion matrix for the logistic regression model classifying students' results in group 2 as either a *PASS* or *FAIL*.

Naïve Bayes Model

We constructed a naïve Bayes model and applied it to the two groups in *Table III*. Results in *Table VII* indicate that group 2 produced the highest accuracy by correctly classifying 61.32 percent of students' results as either a *PASS* or *FAIL*. *Figure 2* shows a confusion matrix generated for group 2.

TABLE VII
Precision, recall, F1 score and accuracy from the naïve Bayes model

Group	Precision	Recall	F1 score	Accuracy
Group 1	68.29 %	40.88 %	51.14 %	60.95 %
Group 2	61.54 %	60.38 %	60.95 %	61.32 %

		Predicted Outcome	
		Pass	Fail
Actual Outcome	Pass	128	80
	Fail	84	132

Fig. 2. Confusion matrix for the naïve Bayes model classifying students' results in group 2 as either a *PASS* or *FAIL*.

Decision Tree Classifier

Table VIII is displaying the precision; recall; F1 score and accuracy for each group when using the decision tree classifier. We produced results using both the gini index and entropy. Group 2 produced the highest accuracy by correctly classifying 68.75 percent of students' results as either a *PASS* or *FAIL* when using the gini index and 73.44 percent of students' results as either a *PASS* or *FAIL* when using entropy. Confusion matrices for both the gini index and entropy are shown in *Figure 3* and *Figure 4* respectively for group 2.

TABLE VIII
Precision, recall, F1 score and accuracy from the decision tree classifier model

Group	Precision	Recall	F1 score	Accuracy
Group 1 using the gini index	62 %	60 %	56 %	60.24 %
Group 1 using entropy	62 %	60 %	56 %	60.24 %
Group 2 using the gini index	81 %	69 %	66 %	68.75 %
Group 2 using entropy	74 %	73 %	73 %	73.44 %

		Predicted Outcome	
		Pass	Fail
Actual Outcome	Pass	26	0
	Fail	40	62

Fig. 3. Confusion matrix for the decision tree classifier classifying students' results in group 2 as either a *PASS* or *FAIL* using the gini index.

		Predicted Outcome	
		Pass	Fail
Actual Outcome	Pass	44	12
	Fail	22	50

Fig. 4. Confusion matrix for the decision tree classifier classifying students' results in group 2 as either a *PASS* or *FAIL* using entropy.

Hypothesis Test and Confidence Interval Test

A hypothesis test and confidence interval test was conducted to test whether or not there was a difference in average aggregate final first year computer science results between students with prior computer experience (group 2) and students without prior computer experience (group 1).

Table IX is displaying results from the hypothesis test and confidence interval test. We represented the average aggregate final first year computer science results for group 1 as μ_1 whereas μ_2 represented the average aggregate final first year computer science results for group 2.

Based on the results from *Table IX*, the null hypothesis was rejected at a 5 percent level of confidence. The confidence interval test results showed that the first year computer science results of students with prior computer experience was greater as compared to students without prior computer experience at a 95 percent level of confidence.

Based on the hypothesis test and the confidence interval test, we can conclude that Students with prior computer experience outperformed students without prior computer experience in first year computer science.

TABLE IX
Table displaying information and results from the hypothesis test and confidence interval test.

Section	Content / Value
Null hypothesis (H_0)	μ_2 - $\mu_1 = 0$
Alternative hypothesis (H_A)	μ_2 - $\mu_1 > 0$
Z score	1.645
The calculated Z statistic	6.034
Confidence interval test lower bound	6.137
Confidence interval test upper bound	10.737
Hypothesis test decision	Reject null hypothesis at a 5 percent level of significance
Confidence interval test result	First year computer science results of students in group 2 was greater as compared to students in group 1 at a 95 percent level of confidence.

CONCLUSION AND RECOMMENDATIONS

We have investigated the effect that prior computer experience had on the final first year computer science results. Two groups were considered. Group 1 consisted of students who had no prior computer experience. Group 2 consisted of students who had prior computer experience. The final grade 12 results for pure mathematics; English first language and computer studies were used to predict the final aggregate first year computer science results. We used three classification models, namely: the logistic regression model; the naïve Bayes model and the decision tree classifier model. Another model used was the linear regression model.

The three classification models resulted in group 2 with the highest accuracy. However, the difference in accuracies between group 1 and group 2 was not great when using the logistic regression model and naïve Bayes model. The difference in accuracies observed in the decision tree classifier model was greater between the two groups as compared to using the logistic regression model and naïve Bayes model. The difference in accuracies between the two groups using the logistic regression model was 2.65 percent. Furthermore, The difference in accuracy between the two groups when using the naïve Bayes model was 0.37 percent. The difference in accuracy between the two groups when using The decision tree classifier was 8.51 percent when using the gini index and 13.2 percent when using entropy.

Results from the hypothesis test and confidence interval test showed that the students in group 2 outperformed the students in group 1. Based on these results, it is worth considering past computer experience as an additional criterion to studying computer science. However, we should not ignore students without prior computer experience considering the fact that the accuracies produced from both group 1 and group 2 differed by a small percentage when using both the logistic regression model (difference of 2.65 percent) and naïve Bayes model (difference of 0.37 percent).

More work can still be done in this investigation. There is the possibility of investigating the students' self-efficacy towards programming which is not based on the students' marks. It is based on the mind set of the students towards programming as was done in the study by [13]. This additional work would give us an indication as to the effect that the students' perception towards programming would have on their final first year computer science results. Further additional work that can be done is investigating the first year performance in computer science as a predictor of performance on the final year results in computer science. First year computer science modules will be considered as the features while performance in final year computer science is what we will be predicting. With this additional work; we will be able to get an idea as to which modules in first year computer science are good predictors of performance for final year computer science.

REFERENCES

[1] Ajoodha, R., Jadhav, A., Dukhan, S. (2020). Forecasting Learner Attrition for Student Success at a South African University. Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (SAICSIT . Cape Town: Association for Computing Machinery

[2] Bergin, S., Reilly, R. (2006). Predicting introductory programming performance: A multi-institutional multivariate study. Computer Science Education (pp. 303-323). Routledge.

[3] Brownlee, J. (2016, April 1). Logistic Regression for Machine Learning. Retrieved August 2020, from https://machinelearningmastery.com/logistic-regression-for-machine-learning/

[4] Butcher, D. F., MUTH, W. A. (1985). Predicting Performance In An Introductory Computer Science Course. Communications of the ACM. West Virginia: Communications of the ACM. [5]Campbell, P. F., McCabe, G. P. (1984). Predicting The Success Of Freshman In a Computer Scuence Major. Communications of the ACM.

[5] Campbell, P. F., McCabe, G. P. (1984). Predicting The Success Of Freshman In a Computer Scuence Major. Communications of the ACM.

[6] Chakure, A. (2019, July 5). Decision Tree Classification An introduction to Decision Tree Classifier. Retrieved August 2020, from https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac

[7] Golding, P., McNamarah, S. (2005). Predicting Academic Performance in the School of Computing Information Technology (SCIT). Proceedings Frontiers in Education 35th Annual Conference, S2H-S2H.

[8] Goold, A., Rimmer, R. (2000). Factors Affecting Performance in First-year Computing. Communications of the ACM

[9] Martin, R. (2018, May 16). Using Linear regression for predictive Modelling in R. Retrieved August 2020, from https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/

[10] Rauchas, S., Rosman, B., Konidaris, G., Sanders, I. (2006). Language Performance at High School and Success in First Year Computer Science. SIGCSE Bull (pp. 398–402). New York: Association for Computing Machinery.

[11] Ray, S. (2017, September 11). 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. Retrieved August 2020, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[12] Werth, L. H. (1986). redicting Student Performance in a Beginning Computer Science Class. SIGCSE Bull (pp. 138–143). New York: Association for Computing Machinery.

[13] Wiedenbeck, S. (2005). Factors Affecting the Success of Non-Majors in Learning to Program. Proceedings of the First International Workshop on Computing Education Research (pp. 13–24). Seattle: Association for Computing Machinery.

[14] Wits. (n.d.). Entry Requirements. Retrieved May 29, 2020, from https: https://www.wits.ac.za/course-finder/undergraduate/science/computer-science/

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
School of Computer Science and Applied Mathematics

Declaration
2020

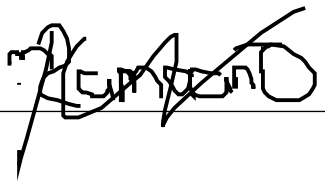
I, Thabo Ramaano, (Student number: 1134539)

am a student registered for Introduction to Research Methods in 2020.

This declaration applies to the Research Report (RR) document of Introduction to Research Methods.

I hereby 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 the work submitted for assessment 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, Johannesburg may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  Date: 1 December 2020