Factors Affecting Programming Skill of the Students – An Exploratory Analysis

Sherna Mohan(✉) and E. R. Vimina(✉)

Department of Computer Science and IT, Amrita Vishwa Vidyapeetham, Kochi Campus, Kochi, India

Abstract. One of the issue encountered in computer programming courses are the high failure rate among the students. This is a serious concern for educators and the students. So there is a dire need to diagnose the factors affecting the same. Hence the objective of this study is to analyse the programming skill of the students by considering the factors like educational background, program debugging skill etc. Methods like correlation and regression are adopted for analysing these factors. It is observed that the debugging skill of the student has an upper hand in determining the programming skill compared to the marks secured in the examinations.

Keywords: Programming skill · Performance · Behavior · Correlation · Debugging · Regression

1 Introduction

There is a high failure rate and drop outs occurring in the computer programming courses. Students are often finding difficulty in programming. So educators have to take extra effort in identifying the weak students and device proper strategy to improve their programming performance. In this study certain factors are investigated which are obligatory for the programming skills of the students. Hitherto major studies focused on the demographics [2], educational background [4], prior programming experience [6], mathematical ability of the students in the examination [3, 7] etc. are used to predict the programming skill.

In computer programming papers, students are given various programming assignments which are expected to be completed within a stipulated time period. The evaluation of these assignments helps the educators to identify the programming skills of the students. Hence the objective of the proposed work is to explore the factors that can be used to determine the programming skills of the students. In order to predict the programming behavior, some methods are used for recognizing the hidden relationship among the attributes of the dataset. The following factors are considered for the analysis.

- Marks obtained in the qualifying examination, especially the marks scored in mathematics [3, 7].
- Programming behavior of the student for the given assignment which is analysed through debugging capability.

P. K. Behera and P. C. Sethi (Eds.): CSI 2020, CCIS 1372, pp. 12–18, 2021.
https://doi.org/10.1007/978-981-16-2723-1_2

This paper is organized as follows. The background work is described in Sect. 2 and proposed approach in Sect. 3. In Sect. 4, results and discussions are analysed and finally, in Sect. 5 the conclusions and future scope are explained.

2 Related Works

Over the past decades, several researchers have been attempting to study the academic performance of the students by analysing the factors which affect the programming behavior of the students. So there are various data mining techniques can be used to predict students programming performance. In [1] relationship between the gender and the marks obtained in the final examination by the students are analysed using chi square test and it is observed that there is no significant relationship between the same. They used deep learning methodology to predict the grade of the students. Another study [2] deals with prediction of students' performance in final examination using the linear regression and multilayer perceptron in WEKA tool and compared the greatness between the mean absolute error value differences. Based on the 58 participants from Taiwan University [3], the study predict the academic performance using students final grades to improve learning performance with the help of multiple linear regression and principal component analysis.

Sujatha [4] predict student performance with the help of regression algorithms and found risky students who need more attention in the programming based on the features of higher secondary school background details, the medium of study, syllabus covered, marks scored in mathematics and English etc. Based on the data points collected from different undergraduate courses [5], a new set of multivariate linear regression model is used to predict the final exam score in the Engineering dynamics course. The data mining technology aids to assess the learner's performance and help us to implement various new trends and technologies to analyse the data [6]. Another paper [7] statistically found that there is no correlation between the performance of computer subjects in high school and the performance in the first year programming course.

3 Proposed Approach

The objective of the investigation is to analyse the programming skills of the students by considering the factors like marks obtained in the qualifying examination, debugging skill etc. So the weaker students can be identified in advance and the educator can help them to improve their programming skills. In the present scenario, while analysing the factors affecting the programming skills of the student, we consider:

- The marks secured in the qualifying examination especially the marks obtained in the mathematics subject [3, 7].
- The programming behavior of the student by analysing the debugging capability of a student.

The dataset for this study is prepared by collecting the details of first year computer science students. The details of 108 students were collected, out of which 35 were males

and 73 were females. The chosen students were from different educational backgrounds. Programming questions and the difficulty levels (Easy (E), Medium (M) and Hard (H)) were prepared by the educators. For the dataset preparation, the programming assignments were allotted to the students based on the first year curriculum programming language. The teacher indistinguishably distributed programming assignments to each students and finally collect the debugging outputs of each student.

Programming outputs were analysed based on the number of errors in each compilation, number of errors occurred in the penultimate compilation, number of compilation attempts etc. Various types of errors like syntax errors, semantic errors were also identified during the analysis process. While debugging a program, students may attempt multiple compilations to arrive at the final output. The maximum number of errors occurred and penultimate compilation error values are recorded during the compilation process. This way we reach the debugging capability of the students in three levels (E, M and H) and the numerical values should be normalized within the range from 0 to 1. The compilation attempts and errors occurred during the debugging stage are used to quantify the debugging capability. The formula for calculating debugging capability is given as

$$\mathrm{DC}_{std\,1} = 1 - \left(\frac{PCE_{\max} - PCE_{min}}{CA} \right) \tag{1}$$

Where

- $\mathrm{DC}_{\mathrm{std1}}$ = Debugging capability of individual student
- $\mathrm{PCE}_{\max}$ = maximum number of errors occurred while debugging a single program in the compilation
- $\mathrm{PCE}_{\min}$ = number of errors occurred in the penultimate compilation while debugging a single program in the compilation
- CA = Total number of compilation attempts made by the student to reach the programming output.

3.1 Methodology

As discussed in the introduction, the primary goal of this study is to investigate the factors which influence the programming skills of the students by considering the marks obtained in the qualifying examinations and programming behavior of the students. Better programming practices not only solves the programming problem but also expand their coding creativity. Pearson correlation and Regression methods are adopted for analysing the factors. 70% data were considered as training set and remaining set as testing data.

Factors Influencing the Programming Skill

Here we find the correlation analysis of marks obtained by the students in various examinations and analysing the programming behavior using linear regression methodology.

In order to figure out the dependence between various potential factors we first correlate the final marks and mathematics marks secured in qualifying examination with

final marks obtained in the examination of the current programming course. The criterion variable used for this research was the marks obtained in the final examination of the programming course. Pearson correlation coefficient is the statistical correlation method used in this study and it was calculated as

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{2}$$

In the above correlation method, n is the number of students, x and y determines the marks obtained in qualifying examination and in the final examination. By definition, the coefficient of correlation assumes any values in the interval between -1 and $+1$. The statistical correlation values between the above attributes shown in Table 1 and the Fig. 1 shows the correlation between the marks obtained in the qualifying examination and the marks secured in the final examination. It is pointed out that programming behavior of a student is weakly correlated with marks secured in the qualifying examination.

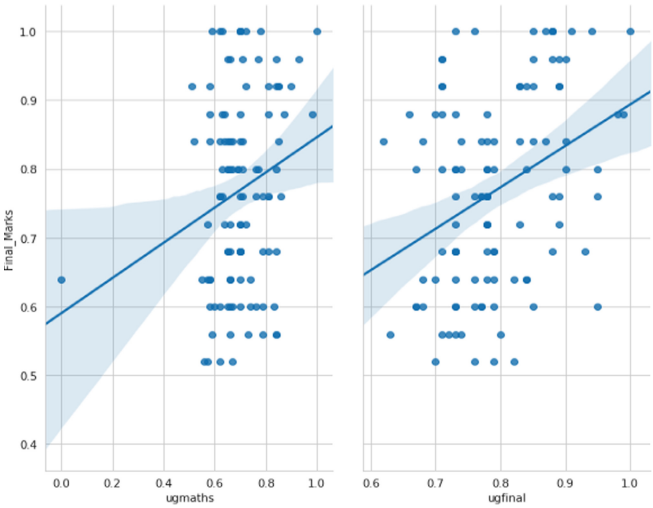

Fig. 1. Correlation between the qualifying marks and final marks

Next, for analysing the programming behavior of the students, we correlate debugging skill and the final marks obtained in the examination of the current programming course. The debugging capability is used for predicting the programming behavior of the students and the quantified value of debugging capability can be computed using the Eq. (1) which is prescribed in Sect. 3. Linear Regression method is used for analysing the programming behavior which indicates the strength of the impact between multiple independent variables and a dependent variable. The linear regression prediction result was displayed as below:

$$Y = 0.405 - 0.011 * DC_E + 0.465 * DC_M + 0.11 * DC_H \tag{3}$$

Where DC_E, DC_M *and* DC_H deals with debugging capability of each student in different levels (Easy, Medium, Hard) and the computation is shown in Eq. (1). Here the response variable (dependent variable) used in this study is final marks secured in the examination of the current programming course. Figure 2 shows the strong positive correlation between the marks obtained in the final examination and the debugging capability of students in different levels (Easy, Medium and Hard). From this, it is pointed out that debugging skill of the student has an upper hand in determining the programming skill compared to the marks secured in the examination.

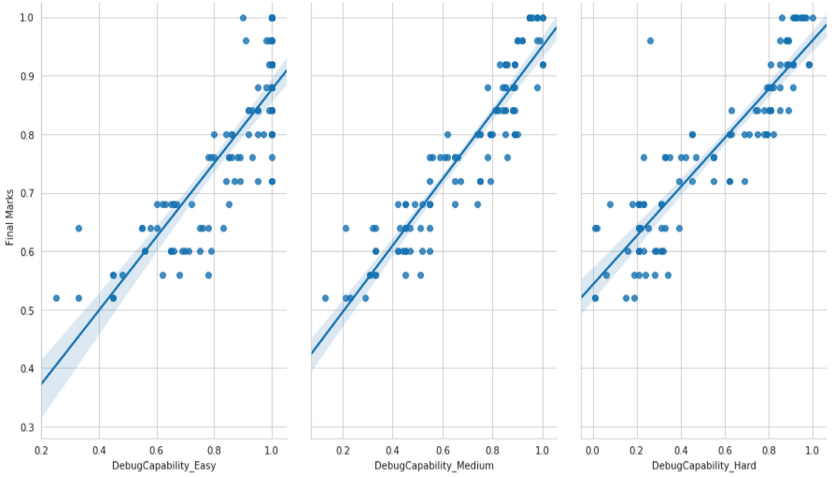

Fig. 2. Correlation between the debugging capability of each levels and the final marks

Furthermore, the study needs an evaluation metric in order to compare the predictions with the actual values. So, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are the two common prediction error measurement methods used for finding the score of the continuous variables. They are used to measure the difference between values predicted by the regression model and the values actual observed. MAE and RMSE are calculated using the following formulas:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum_{j=1}^{n} \left| \mathrm{y_j} - \mathrm{y'_j} \right| \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}} \sum_{j=1}^{n} \left(\mathrm{y_j} - \mathrm{y'_j} \right)^2} \tag{5}$$

Here $\mathrm{y_j}$ and $\mathrm{y'_j}$ are the attributes specified in the above discussion. R^2 and MSE measures can evaluate the goodness of fit of a regression model. The accuracy of prediction is measured by the standard error of the estimate. By comparing Fig. 2 with Fig. 1, the values are very closer to the regression line which minimizes the sum of squares error. Therefore by minimizing the sum of squared deviations of prediction, the predictions in Fig. 2 is more accurate than Fig. 1. The R^2, mean absolute error, mean squared error and root mean squared error values were shown in Table 2.

Table 1. Results obtained by the correlation analysis

Performance Criteria	Marks obtained for Mathematics in qualifying examination	Final Marks secured in qualifying examination	Debugging capability (Easy Level)	Debugging capability (Medium Level)	Debugging capability (Hard Level)
Final marks secured in the examination	0.21	0.35	0.83	0.92	0.89

4 Results and Discussion

In this segment, we will discuss the results of factors influencing the programming skill of the students using correlation and regression. From Table 1, it is observed that the marks obtained in qualifying examination (especially mathematics marks and final marks 0.21 and 0.35) are positively correlated with final marks obtained in the current programming course. Also it is observed that there is a strong positive correlation between the debugging capability of the student (0.83, 0.92, and 0.89) with the final marks. By considering the regression model, R^2 value is 0.81 and evaluation metric values MAE and RMSE have very less residual values (from Table 2) which produces better prediction. From Fig. 2, it is analysed that regression line minimizes the sum of squared errors. So more accurate the prediction. From the above analysis it can be predicted that programming behavior is strongly correlated with final marks in the examination than marks obtained in the qualifying examination. This will aid the educators to learn more about the weak students and identify better techniques to impart programming skills to those students.

Table 2. The results of linear regression model

Feature affecting the programming skill	Coefficient of Determination	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Debugging Capability	0.81	0.047	0.0033	0.218

5 Conclusion and Future Study

Apart from predicting the programming skills, the presented work represents the exploratory analysis of the factors which affect the programming behavior of the students. It is observed that debugging skill of a student is highly correlated with final marks obtained in the examination of the current programming course. So it is concluded that debugging skill of the student has an upper hand in determining the programming skill

compared to the marks secured in the examinations. Therefore as a future work, by considering the programming behavior of the current situation, we plan to carry out a similar study covering more number of potential success factors to improve the performance of the students in programming.

References

1. Pal, V.K., Bhatt, K.K.V.: Performance prediction for post graduate students using artificial neural network. Proc. Int. J. Innov. Technol. Explor. Eng. **8** (2019)
2. Widyahastuti, F., Tjhin, V.U.: Predicting students performance in final examination using linear regression and multilayer perceptron. In: 2017 10th International Conference on Human System Interactions (HSI), pp. 188–192. IEEE (2017). https://doi.org/10.1109/HSI.2017.8005026
3. Yang, S.J., Lu, O.H., Huang, A.Y., Huang, J.C., Ogata, H., Lin, A.J.: Predicting students' academic performance using multiple linear regression and principal component analysis. J. Inf. Process. 170–176 (2018). https://doi.org/10.2197/ipsjjip.26.170
4. Sujatha, G., Sindhu, S., Savaridassan, P.: Predicting students performance using personalized analytics. Int. J. Pure Appl. Math. 229–238 (2018). https://www.ijpam.eu
5. Huang, S., Fang, N.: AC 2010-190: regression models of predicting student academic performance in an engineering dynamics course. In: American Society for Engineering Education (2010)
6. Sagar, M., Gupta, A., Kaushal, R.: Performance prediction and behavioral analysis of student programming ability. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1039–1045. IEEE (2016). https://doi.org/10.1109/ICACCI.2016.7732181
7. Ayalew, Y., Tshukudu, E., Lefoane, M.: Factors affecting programming performance of first year students at a university in Botswana. Afr. J. Res. Math. Sci. Technol. Educ. **22**(3), 363–373 (2018). https://doi.org/10.1080/18117295.2018.1540169