STUDENT FINAL GRADE PREDICTION BASED ON LINEAR REGRESSION

Mahesh Gadhavi

Smt. Chandaben Mohanbhai Patel
Institute of Computer Applications
CHARUSAT
Changa-388421, Gujarat, India
Email: maheshgadhavi.mca@charusat.ac.in

Dr. Chirag Patel

Smt. Chandaben Mohanbhai Patel
Institute of Computer Applications
CHARUSAT
Changa-388421, Gujarat, India
Email: chiragpatel.mca@charusat.ac.in

Abstract

In the world of Massive Open Online Course (MOOC) and open education systems, students have flexibility to learn anything with ease as the learning content is easily available. But this facility can make student complacent. Therefore, it becomes difficult to predict student performance in advance. In this research an attempt is made to help the student to know his/her performance in advance by using univariate linear regression model. We collected the marks of internal exam components of one subject to predict the final grade in that subject. The internal marks are normalized to 100 (percentage) to have accurate results. The model provides predicted grade of final examination in particular subject. It also helps students to know how many marks in the internal examination are required to get particular grade.

***Keywords*:** Grade, Marks, Linear regression, Decision tree

1. INTRODUCTION

In present educational systems, student performance prediction is getting worsen day by day. Predicting student performance in advance can help students and their teacher to keep track of progress of a student. Many institutes have adopted continuous evaluation system today. Such systems are beneficial to the students in improving performance of a student. The purpose of continuous evaluation system is to help regular students.
In continuous evaluation system, unit tests or class tests are conducted at regular period. To have consistence performance in the final grad of the subject it is required to appear in all the unit tests or class test. In this paper we have considered the examination pattern of our institute Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA). As per the internal examination pattern of CMPICA, at least three unit tests of any subject should be conducted. Then after, one sessional examination is conducted. The weightage of internal examination is 30% and external examination in 70%.

2. LITERATURE REVIEW

During the last few decades the educational pedagogy has been improved a lot. With the advent of Learning Management Systems (LMS), it becomes easier to deliver the subject contents to the students. But still it is difficult to predict the student's performance in advance. To predict student's achievement a genetic algorithm based algorithm is proposed by Amelia Zafra et al. [17]. They applied G3P-MI algorithm to predict whether a student will pass or fail in certain subject. The experiments were carried out on 10-fold stratified cross validation to judge the validity of their algorithm and to achieve 74.29% of accuracy. A decision tree based algorithm is proposed in [13] to predict marks of a student. The authors applied c4.5 decision tree to forecast to predict student marks. The authors mention that decision three is the core element of data mining process which is divided in the stages. In the first stage decision tree is generated by using the training data. In the second stage the decision tree is validated by using test data to check whether the decision tree provides accurate results or

not. In the third stage forecasting of unknown data and generate the data which can be useful for the decision makers.

Today huge amount of student data is available so majority of the researchers adopted data mining technique to predict student's performance. A decision table and One-R based method is proposed in [12] to predict student's performance in higher education. The authors mention that decision table is easy way to represent complex logic. The logic can be transformed into simple if-else or flow chart. They also mention that One-R method is used to generate one level decision tree. The also mention that One-R algorithm is better to have good predictive accuracy among large class of data. Besides that it is also good to test new machine learning algorithm. Different mathematical and machine learning techniques are proposed by [9]. The authors applied multilayer perceptron neural networks, radial basis function neural networks, support vector machines and multivariate linear regression to predict student performance. They did comparative study of all these methods by applying these methods on 323 student data. They found that there is no such difference between these methods. Their model was able to get more than 80% of accuracy.

The authors in [3] found determinant for student performance in advanced programming subject. The scope of this paper was to provide different hypothesis to find the relationship between different variables. In [16] authors recorded programming behavior of students, and then examined the recorded information. Based on this information they predicted their capabilities in basic programming subjects. They developed one dynamic algorithm called Watwin to predict potential of a student in basic programming subjects. This model reduced efforts in evaluating student using different methods to predict the performance of a student in programming. Predicting potential and achievements of students using various institute analytic tools is presented in [2]. Author have proposed model to increase students capabilities in different categories such as percentage/grade, attendance in class, knowledge of prerequisite subjects. They also proposed the use of Artificial Neural Network and Decision Tree for prognosticative modeling. In [15], to predict performance of engineering students using different classification algorithm author have proposed method to predict students grade / percentage. They initially developed classifier using different classification methods and then using Bootstrap method they increased the reliability of the individual classifier. In [7], authors have collected students data generated through LMS, online web based assessments at the time of traditional daily teaching in classroom. The data provides students behavior and grade / percentage during the course is still in progress. Then using latest classification algorithms and genetic algorithm they predicted the performance of students in final examination. An improvement in early warning system is proposed by [8] to predict future marks of a student. They applied Naive Bayes method to predict student performance from data warehouse of the student data. The collected records of 220 students out of which 175 records are considered as training data, 45 records are considered as test data to achieve 86.66% accuracy. An artificial neural network based system is proposed by [14] to predict grades of the students. The authors invited comments from students after each lesson. Then they applied Latent Semantic Analysis (LSA) to extract semantic data from comments made by students. Then they applied Artificial Neural Network (ANN) to predict the grades of student. The authors claimed to achieve 82.6% of accuracy. To predict the performance of the students in future academic examinations, authors have developed web based application in [5]. They applied Nave Bayesian mining technique as it provides more accuracy compare to other techniques. They collected various historical information related to students and then using Nave Bayesian mining technique, they predicted future academic results of the students. In [4], author have used log data of different activities which are performed by the students in the Learning Management System (LMS). Based on this information, using classification techniques they predicted performance of learner in examination. They applied the decision tree algorithm based on J48 and Multiple linear regression to predict student performance. To help the students in achieving their targeted goal, based on students education data author in [11] have developed system which can generate weak and strong area of the students. They formed rules from various algorithms such as J48, Logistic Model Tree(LMT), Random Tree and REPTree, and predicted semester wise results of the students.

Apart from this, several researchers tried to measure the student performance using different methods such as self-assessment [18], analysis of difficulty level of course using student performance prediction [10], software matrices to predict difficulty of code writing questions [6] and student behavior analysis in virtual learning environment [1]. In this section detailed literature review has been conducted. We have observed many methods such as linear regression, ANN and decision tree. It is observed that linear regression based model is well suited for our proposed work as it predicts the future value rather than a class label. We also studied that for predefined class prediction ANN or decision tree or any other supervised learning method are suitable but in our research we want to predict future grade of a student. Therefore, linear regression is applied in our work which is mentioned in the next section.

3. PROPOSED WORK

In this research, we applied linear regression based model to predict the final grade for student in particular subject. The model is trained using marks of existing student in one subject. In this research the X variable is considered as average of unit test and sessional examination marks. Furthermore, the marks are converted into percentages to gain proper accuracy of the system. In the present system the grade are converted in the non-numeric forms which are converted into numerical form as depicted in Table 1.

Table 1 Percentage, Grade point and Grade

Grade	Grade point range	Equivalent Grade point
AA	>=80	80
AB	>=75 and <80	75
BB	>=70 and <75	70
BC	>=65 and <70	65
CC	>=60 and <65	60
CD	>=55 and <60	55
DD	>=50 and <55	50

As illustrated in above table, the students are provided grade 80 marks, the student gets AA grades. Likewise, this rule is applicable to other grade and marks range too. To check the co-relational between the variable X and Y the co-relation coefficient r is calculated based on the following equations:

$$Z_x = \frac{x - \mu_x}{\sigma_x} \quad (1).$$

$$Z_y = \frac{y - \mu_y}{\sigma_y} \quad (2)$$

$$r = \frac{Z_x Z_y}{n - 1} \quad (3)$$

The Z-scores of X and Y are calculated using equation 1 and equation 2 respectively. Then the co-relation coefficient r is calculated using the equation 3. The value of r is 0.64 which indicates that there is positive co-relation between X and Y. This means that increasing the value of X increases value of Y. The value of r indicates that the grade marks can be predicted using linear regression as there is positive correlation between internal exam marks and final grade of the subject. As per the equation 4, the linear regression model is trained with records of 181 of students in one subject.

The model is represented in figure 1 The function of linear regression is mentioned in equation 4.

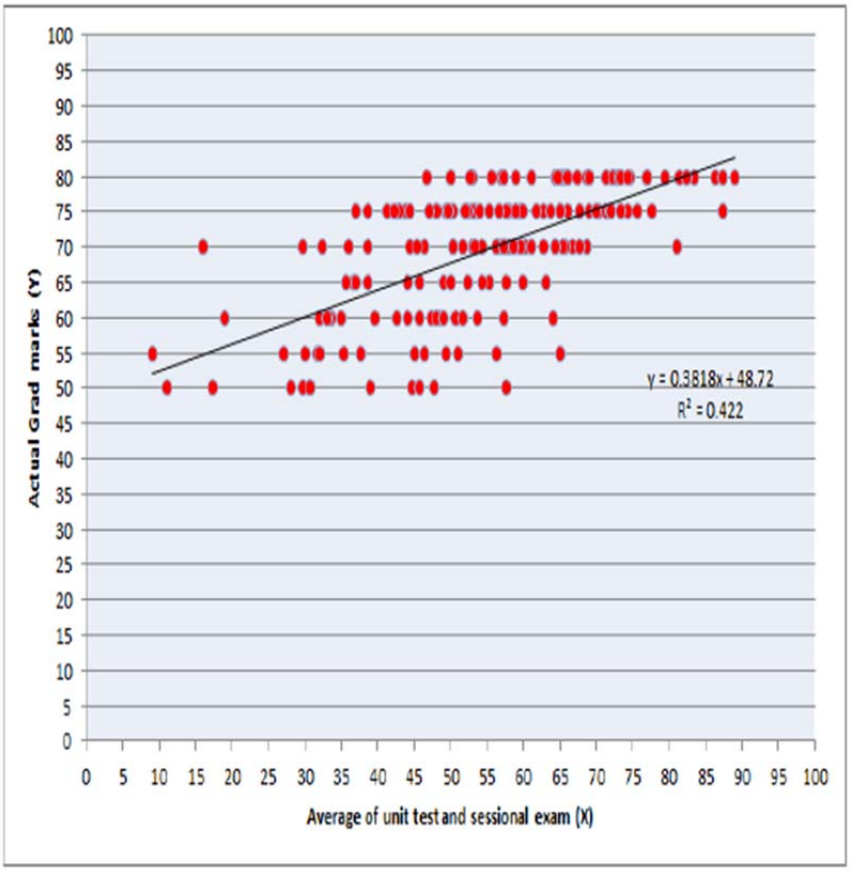

Fig 1. Linear regression model

$$y = \theta_0 + \theta_1 x \tag{4}$$

The main goal is to find the optimum values $of\ \theta_0$ and θ_1 which minimize the cost function of the model. The cost function of the model is defined as per equation 5:

$$C(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h(x) - y)^2 \tag{5}$$

The optimized values of θ_0 and θ_1 can be calculated by using the iterative process which minimized the cost function C. The cost function can be minimized by using the well-known gradient descent function. As per gradient descent, the value of theta can be updated until the local minimum value of cost function is found. The update of θ_0 and θ_1 should be done simultaneously as depicted in equation through equation6 to equation 9. Then provide these values to calculate cost function as per the equation 5.

$$temp_0 = \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h - y) \tag{6}$$

$$temp_1 = \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h - y)x \tag{7}$$

$$\theta_0 = temp_0 \tag{8}$$

$$\theta_1 = temp_1 \tag{9}$$

The process continues until the minimum cost function is calculated. The variable α is a learning rate used in equation 7 and equation 8. It remains constant through all the iterations. The value of α is taken as 0.001 for convergence of cost function. The higher value of α does not allow the cost function to converge at local minimum which calculates wrong values of θ_0 and θ_1. Therefore, the value of α should not be too large or two small.

4. Results and Discussion

The model is tested on same data set of 181 students. The minimum value of cost function is ~23.32. The optimum value of θ_0 is 49.10 and θ_1 is 0.37 which fits the model best. The model with training and testing data is shown in Figure 2.

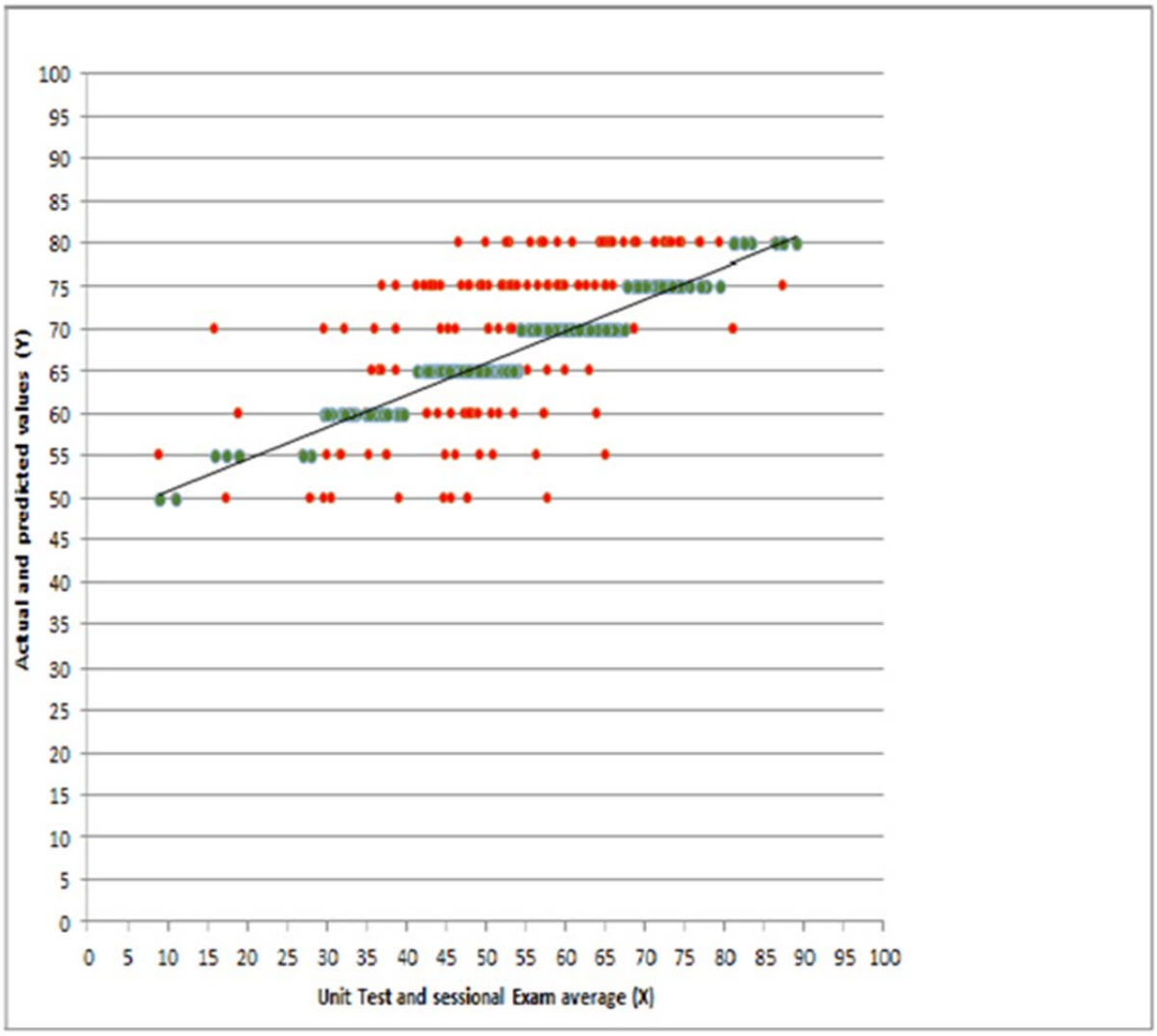

Fig 2 Final Model

5. Conclusion

The prediction of student performance is getting difficult day by day. In this research we have developed a linear regression based model which will help students in knowing final grade in particular subject. To accomplish this research, internal exam marks out of 30 are taken into consideration. Then the marks are converted into 100 (percentage) to have uniformity benchmark. These data is used to train the linear regression model to calculate the appropriate value of θ_0 and θ_1. This model is a univariate i.e. it takes only one variable but it can be extended as multivariate model by adding more parameters to get more accurate results.

6. Acknowledgement

The authors would like to thank Charotar University of Science and Technology (CHARUSAT) for providing necessary resources to accomplish the research.

References

[1] L. Benko, J. Reichel, and M. Munk. Analysis of student behavior in virtual learning environment depending on student assessments. In 2015 13th International Conference on Emerging eLearning Technologies and Applications (ICETA), page 16, Nov 2015.

[2] U. bin Mat, N. Buniyamin, P. M. Arsad, and R. Kassim. An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In 2013 IEEE 5th Conference on Engineering Education (ICEED), page 126130, Dec 2013.

[3] Y. Y. Chen, S. Mohd Taib, and C. S. Che Nordin. Determinants of student performance in advanced programming course. In 2012 International Conference for Internet Technology and Secured Transactions,page 304307, Dec 2012.

[4] B. E. V. Comendador, L. W. Rabago, and B. T. Tanguilig. An educational model based on knowledge discovery in databases (kdd) to predict learner's behavior using classification techniques. In 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), page 16, Aug 2016.

[5] T. Devasia, Vinushree T P, and V. Hegde. Prediction of students performance using educational data mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), page 9195, March 2016.

[6] S. Elnaffar. Using software metrics to predict the difficulty of code writing questions. In 2016 IEEE Global Engineering Education Conference (EDUCON), pages 513–518, April 2016.

[7] J. Gamulin, O. Gamulin, and D. Kermek. Comparing classification models in the final exam performance prediction. In 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pages 663–668, May 2014.

[8] H. Goker and H. I. Bulbul. Improving an early warning system to prediction of student examination achievement. In 2014 13th International Conference on Machine Learning and Applications, pages 568–573, Dec 2014.

[9] S. Huang and N. Fang. Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques. In 2012 Frontiers in Education Conference Proceedings, pages 1–2, Oct 2012.

[10] K. Kaur and K. Kaur. Analyzing the effect of difficulty level of a course on students performance prediction using data mining. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT), page 756761, Sept 2015.

[11] P. Kaur and W. Singh. Implementation of student sgpa prediction system (ssps) using optimal selection of classification algorithm. In 2016 International Conference on Inventive Computation Technologies (ICICT), volume 2, page 18, Aug 2016.

[12] S. Anupama Kumar and M. N. Vijayalakshmi. Mining of student academic evaluation records in higher education. In 2012 International Conference on Recent Advances in Computing and Software Systems, page 6770, April 2012.

[13] Z. Liu and X. Zhang. Prediction and analysis for students' marks based on decision tree algorithm. In 2010 Third International Conference on Intelligent Networks and Intelligent Systems, page 338341, Nov 2010.

[14] S. E. Sorour, T. Mine, K. Goda, and S. Hirokawa. Predicting students' grades based on free style comments data by artificial neural network. In 2014 IEEE Frontiers in Education Conference (FIE) Proceedings, page 19, Oct 2014.

[15] S. Taruna and M. Pandey. An empirical analysis of classification techniques for predicting academic performance. In 2014 IEEE International Advance Computing Conference (IACC), page 523528, Feb 2014.

[16] C. Watson, F. W. B. Li, and J. L. Godwin. Predicting performance in an introductory programming course by logging and analyzing student programming behavior. In 2013 IEEE 13th International Conference on Advanced Learning Technologies, page 319323, July 2013.

[17] A. Zafra, C. Romero, and S. Ventura. Predicting academic achievement using multiple instance genetic programming. In 2009 Ninth International Conference on Intelligent Systems Design and Applications, page 11201125, Nov 2009.

[18] S. M. Zvacek, M. de Ftima Chouzal, and M. T. Restivo. Accuracy of self-assessment among graduate students in mechanical engineering. In 2015 International Conference on Interactive Collaborative Learning (ICL), page 11301133, Sept 2015.