Linear Regression Technique for Student Academic Performance Prediction

[1] Kapil Saxena, [2] Shailesh Jaloree, [3]R.S.Thakur, [4] Sachin Kamley
[1] Research Scholar
[1,2,4] S.A.T.I., Vidisha, [3] M.A.N.I.T., Bhopal

Abstract: The amount of data in the education field is growing rapidly day by day. It is very complicated task to manage this data and making predictions of performance from it. In order to retrieve meaningful information and facts, data mining software's are used. Student performance prediction is an interesting task and gaining so much popularity due to competitiveness among education organizations. In this study, the well-known and efficient linear regression method is adopted to predict student performance. The main concern of this research study is to make use of this rich data set optimally to predict student performance and also help the academician, researchers and stack holders for improving the decision making quality. This research study also will be helpful to students as well as teachers to aware them and boosting the confidence so that they can gain better marks in the future.

Index Terms **- Education Data Mining, GGC, Linear Regression, Prediction.**

I. INTRODUCTION

Regression Analysis (RA) is one of the well-known statistical techniques that are used to draw relationships among dependent and independent variables [1] [2]. However, the applicability of these techniques is in almost every field such as economics, finance, education and engineering etc. On the other side, RA is most popularly known for studying the functional dependencies among variables. The functional dependency P->Q means Q is functionally dependent on P. for ex. There is a functional dependency between salary and experience because salary dependent on experience. If experience increases then salary automatically increases [3].

One of the simplest type of regression in regression family is linear regression i.e. regression with a single predictor variable. However, linear regression can be used if and only if relationship between and modeled with a straight line [4].
Finally, linear regression technique has much wide applicability for forecasting task. Presently, it is overlapped with machine learning field. There is a wide variety of data analysis methods available in regression family such as multiple regression, nonlinear regression, lasso regression and ridge regression etc. [5]. The main concern of this study is to design linear regression model for education dataset as well as to test and compare the accuracy of the model with past models.

II. LITERATURE REVIEW

Teir and Halees (2012) [6] have applied data mining techniques for extraction of knowledge from educational domain to improve the performance of graduate students i.e. low grades. In their exercise, considered the data of college of science and technology. Finally, experimental results stated that they could have been succeeded to affected problem of low grades of students.

Ibrahim and Rusli (2007) [7] have compared the performance of three different algorithms like linear regression, decision tree and artificial neural network. They have used CGPA and demographic variables for measuring the student performance. Finally, they have got the prediction accuracy over 80 and linear regression method performs outstanding for short data samples.

David de la Peña et al. (2017) [8] focuses on a logistic regression model. The collected data are preprocessed and data cleaning is performed after that a reference model is built for each of the course using logistic regression and are stored in the database which are then used for classification. Here the whole student details are being used to classify into two classes namely dropout or non-dropout

Rao and Nagraj (2014) [9] have applied liner regression analysis method for student performance prediction. In their study, they have considered student's data set of 49 students of a reputed institution consisting of the percentage of marks got by the students in their mid and final exams. Finally, error rates between original marks and predictive marks are very less.

Gadhavi and Patel (2017) [10] have applied linear regression approach for student final grade prediction on particular subject. They have considered the data set of 181 students to their study. Finally, limitation of their model is that it works only for one variable.

Kamber and Micheline (2006) [11] characterize data mining software that permit the users to examine data from different aspect, classify it and summarize the relationships which are recognized during the mining process.

Ayan and Garcia (2008) [12] have applied linear and logistic approaches to predict performance of university students. They have considered various academic, non-academic, social and demographic parameters for their study. finally, experimental results stated that logistic approach performs little bit better in terms of accuracy.

III. DATA PREPROCESSING

The Government Girls College (GGC), Vidisha is considered for prediction task [13]. However, dataset contain 250 samples. The data set consisting SSC marks, HSC marks, attendance, theory and practical marks, SGPA AND CGPA marks etc. [13]. During this study, mean (average) is calculated in order to fill missing values as well as data normalization formula is used to transform the value in [0, 1]. The data set contains so much attributes so describing all attributes is not possible. The sample of preprocessed data is given below by Table I [13].

Table I: Sample of Preprocessed Data

SSC	HSC	S1TH	SIPR	Attend.	Income
poor	verygood	good	good	good	verygood
poor	good	avg	avg	avg	good
avg	avg	avg	good	poor	good
poor	avg	avg	good	good	poor
verygood	poor	poor	good	good	poor
poor	good	poor	good	poor	good
avg	avg	avg	verygood	poor	Avg
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

IV. PROPOSED METHODOLOGY

A. Linear Regression

Regression analysis provides a way to identify the relation between two or more variables i.e. dependent and independent variable about which knowledge is available [11] [14]. In this way, linear regression method is used to make predictions based on two variables. The simplest form of linear regression with a single predictable variable and one response variable is given below by equation (1.1) [11] [15].

$$Y = u + wX \qquad (1.1)$$

Where and denotes regression coefficients as well as specifying the intercept and slope of the line respectively.
Equivalently, equation (1.1) can be written as

$$Y = u_0 + w_1 X \qquad (1.2)$$

Therefore, the least square method is used to solve the equation (1.1) which always minimizes the error between the actual and predicted data [11] [16].

Let predictable variable and response variable is denoted by training samples. In this study, SGPA is used as a predictable variable and CGPA is used as a response variable.
Now regression coefficients can be solved by following equations.

$$w_1 = \sum_{i=1}^{|n|} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \qquad (1.3)$$

$$u_0 = \bar{y} - w_1 \bar{x} \qquad (1.4)$$

Where, $\bar{y}$ = Average of CGPA marks,

$\bar{x}$ = Average of SGPA marks, and

n = Total no. of observations.

Here, the coefficient of regression equations (1.3) & (1.4) is calculated and used to solve the regression parameters. Therefore, the calculated value of variables w_1 and u_0 are given below:

$w_1 = 2.70$, $u_0 = -100.69$

The straight line trend in eq. (1.5) was used to get the result as shown below by equation no (1.2).

$$Y = -100.69 + 2.70X \qquad (1.5)$$

V. EXPERIMENTAL RESULTS

In this study, the last three years (2012-2015) Government Girls College (GGC) data is obtained [13]. Table II shows sample of student performance pattern.

Table II. Sample of Student Performance Pattern

S.NO.	SGPA Marks	CGPA Marks
1	52.67	55.17
2	61.33	58.00
3	56.67	65.42
4	64.50	62.29
5	53.83	60.08
6	59.78	58.06
7	52.67	59.83
8	50.50	57.33
9	56.00	57.88
10	55.67	59.17

Moreover, the least square method is used to predict student performance by putting the values of variables in equation (1.2). However, the dependent variable (CGPA) is calculated from independent variable (SGPA). Table III shows getting predictions (i.e. CGPA) from proposed system.

Table III. Sample of Forecasted Values

S.No.	Actual CGPA	Predicted CGPA
1	55.17	41.50
2	58.00	64.90
3	65.42	52.30
4	62.29	73.45
5	60.08	44.65
6	58.06	60.70
7	59.83	41.50
8	57.33	35.65
9	57.88	50.50
10	59.17	49.60
11	60.21	44.65
12	55.71	29.80
13	58.17	77.05
14	55.71	35.65
15	61.17	48.10
16	54.17	91.90
17	63.78	43.30
18	56.78	102.10
19	61.89	73.90
20	55.28	97.30
21	58.39	25.30
22	59.83	60.70
23	67.75	79.30
24	56.08	56.80
25	57.50	54.55

26	60.17	65.80
27	58.79	41.95
28	55.75	36.55
29	57.83	37.45
30	56.78	37.00
31	61.89	88.30
32	55.28	41.95

Figure 1 shows performance comparison against actual CGPA and predicted CGPA.

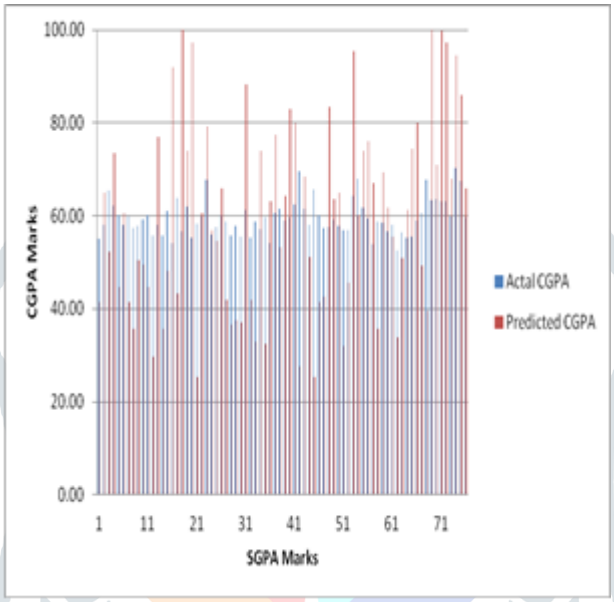

Figure 1: A Performance Comparison against Actual CGPA and Predicted CGPA using Bar Graph

In Figure 1, bar graph shows performance comparison against actual and predicted CGPA. Here, red line shows predicted CGPA and blue line shows actual CGPA. Next, Table IV shows residual error estimation. Therefore, correlation coefficient between the actual CGPA and predicted CGPA of the model is found to be implying a very poor dependency of the predictive model.

Table IV. Residual Error Estimation

CGPA (Y)	**Predicted CGPA** $(\bar{Y})$	**Residual** $(Y - \bar{Y})$	**Residual** $(Y - \bar{Y})^2$
55.17	41.50	13.67	186.7504
58.00	64.90	-6.90	47.6238
65.42	52.30	13.12	172.0207
62.29	73.45	-11.16	124.5307
60.08	44.65	15.43	238.1569
58.06	60.70	-2.65	6.998376
59.83	41.50	18.33	336.0744
57.33	35.65	21.68	470.1236
57.88	50.50	7.37	54.37588
59.17	49.60	9.57	91.50198
60.21	44.65	15.56	242.0306
55.71	29.80	25.91	671.1899
			$\Sigma(Y-\bar{Y})^2 = 31629.72$

Table IV states that residual (the sum of squared errors) of the prediction. The equation no. (1.6) shows residual error calculation formula [104].

$$\text{Residual Error Estimate (REE)} = \sqrt{\frac{\text{Residual Sum}(Y-\bar{Y})^2}{N}} \quad (1.6)$$

Where,

REE = Residual Error Estimation,
Y = Actual Value of dependent variable,

$\bar{Y}$ = Predicted value of dependent variable,
N = No. of observation or data points.

After applying formula, the value of REE is obtained. We have also calculated MAD that is 416.18 which is also high. Due to the large samples and presence of larger errors, prediction accuracy is over than 60% which is acceptable.

VI. CONCLUSION AND FUTURE SCOPES

In this study, we have presented a simple linear regression method for predicting the student academic performance, but limitation of this method is that it works only for two variables. For getting more accurate prediction, the consideration of more variable is needed. The Multiple Regression Analysis is more powerful tool for forecasting student performance using multiple parameters. In future, we will consider a multiple regression analysis approach to overcome the problem of proposed linear regression model.

REFERENCES

[1] C. Romero, S. Ventura, and, E. Garcia, "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, Issue (1), pp. 368–384, 2008.

[2] S. E., Sorour, T. Mine, K. Goda, and, S. Hirokawa, "A predictive model to evaluate student performance", Journal of Information Processing, Vol. 23, Issue (2), pp.192–201, 2015.

[3] T. Devasia, Vinushree T P, and V. Hegde, "Prediction of student's performance using educational data mining", In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), pp. 9195, March 2016.

[4] 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, pp. 568–573, Dec 2014.

[5] A. PENA˜ -AYALA, "Educational data mining: A survey and a data mining-based analysis of recent works", Expert Systems with Applications, Vol. 41, Issue (4), pp. 1432–1462, 2014.

[6] M. Abu Tair, Alaa M. ElHalees, "Mining educational data to Improve Students' performance", International Journal of Information and Communication Technology Research, pp. 140-146. 2012.

[7] Z. Ibrahim and D. Rusli, "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression", In 21st Annual SAS Malaysia Forum, 5th September 2007, Shangri-La Hotel, Kuala Lumpur, 2007.

[8] David de la Peña, Juan A. Lara, David Lizcano, María and Concepción Burgos, María L. Campanario. "Mining activity grades to model students' performance", ICEMIS2017, Monastir, Tunisia, 2017.

[9] G.N. Rao and S. Nagaraj, "A study on the prediction of student's performance by applying straight-line regression analysis using the method of least squares", International Journal of Computer Science Engineering (IJCSE), Vol. 3, Issue (1), pp. 43-45, 2014.

[10] M. Gadhavi and C. Patel, "Student final grade prediction based on linear regression", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 8, Issue (3), pp. 274-279, 2017.

[11] J.H. Kamber and Micheline, "Data Mining: Concepts and Techniques", second edition. San Francisco: Morgan Kaufmann, 2006.

[12] Ayan, M. N. R. and Garcia, M. T. C., "Prediction of University students' academic achievement by Linear and Logistic models," The Spanish Journal of Psychology, Vol. 11, pp. 275-288, 2008.

[13] Data obtained from Government Girls College (GGC), Vidisha.

[14] B.K. Pal and Bharadwaj, "Data mining: A prediction for performance", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, 2011.

[15] S. Borkar and K. Rajeswari, "Predicting students' academic performance using education data mining", International Journal of Computer Science and Mobile Computing, Vol. 4, pp. 273-279. 2013.

[16] E. Alfan, and M.N. Othman, "Undergraduate students' performance : The case of University of Malaya, quality assurance in education, Vol. 13, No. 4, pp. 329 – 343, 2005.