Multiple Regression: A Data Mining Technique for Predicting the Student Performance

[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: In these days, predicting student future performance based on their past records is an important concern. The education institutes and universities collect huge amount of data yearly. However, making accurate prediction from that data is very challenging issue due to the involvement of various factors such as academic, non-academic, demographic and social etc.. Due to the competiveness among the institutes and remain in the lime light, this area has gained so much popularity among scientists, researchers, academicians etc.. This study presents well known and efficient Multiple Regression approach for predicting student performance. The main objective of this research study is to find out student performance so that it will be helpful for teachers and parents to know actual conditions of students before going to final exams.

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

I. INTRODUCTION

In statistics Community, regression techniques are well suited for drawing relationships among dependent and independent variables [1] [2]. However, these techniques cover almost all areas such as agriculture, weather forecast, economics, finance, education and engineering etc. [3].

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 age and experience because experience dependent on age. If age increases then experience automatically increases [3] [4].

In regression family, there are many methods available such as linear regression, multiple regression, nonlinear regression, lasso regression and ridge regression etc.

Multiple Regression (MR) analysis is more powerful tool for forecasting student performance using multiple parameters. The term MR is firstly coined by Pearson in 1908 [4]. However, the term is used to study about the several dependent and independent parameters in the social and natural sciences context [2] [4].

Finally, regression technique is most widely used for forecasting task. Presently, it is overlapped with machine learning field [5]. The main objective of this research study is to design and test multiple regression model student dataset. Moreover, asses the model quality with different error measures.

II. LITERATURE REVIEW

This section describes the brief literature of significant researchers. Table I describes brief description of literature review.

Table I: Brief Description of Literature Review

S. No.	Authors	Input/Output Parameters	Technique Used	Accuracy/ Outcome
1	Teir and Halees (2012) [6]	Semester Marks and CGPA	Regression and Classification	Outstanding Performance
2	Gadhavi and Patel (2017) [7]	Final Grade	Linear Regression	Low Accuracy
3	Huang and Fang (2010) [8]	Previous Sem Marks Vs CGPA	Multiple Regression	86%
4	Potgieter AND Coetzee (2013) [9]	Personal Characteristics and Academic attributes	Multiple Regression	Significant relationships found between the Participants' personality preferences and their employability attributes

5	Gokuladas (2011) [10]	Previus Sem Marks	Correlation And Multiple Regressions	CGPA and proficiency in English language are important predictors of employability and female students are better performers.
6	Gokuladas (2010) [11]	Academic Factors	Correlation and e Regression Analysis	Graduates need to possess specific skills beyond general academic education to be employable.
7	Saxena et al. (2018) [12]	Academic and Non-Academic	Linear Regression	Over 60%

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]. Table II shows sample of some descriptive statistics on student data set. [13].

Table II: Descriptive Statistics for Student Dataset

Attributes	N	Min	Max	Mean	Std. Dev.
SSC	248	222	545	371.44	77.55
HSC	248	200	434	319.16	58.44
Theory4 Marks	248	204	434	313.41	53.95
Practical4 Marks	248	11	130	44.18	25.89
Attendance	248	45	79	63.25	68.95
Income	248	20000	60000	38545.45	9535.23
SGPA	248	45.33	77.56	59.86	56.05

Table II clearly states that min, max, mean and std. dev. for all attributes. The high value for standard deviation shows more variability in the student dataset and low values show less variability in the student dataset. So therefore, SSC, attendance and income attributes have high standard deviation values.

IV. PROPOSED METHODOLOGY

A. Multiple Regressions

One way of identifying a relation between two or more variables i.e. dependent and independent variable is known as regression analysis [2] [14].

Multiple Regression is one of the most effective and usable data mining techniques for predicting the dependent variable based on other variables (independent) [3] [15]. However, the value of dependent variable Y can be predicted by putting the values of independent variables $X_1, X_2, ..., X_n$ into the regression equation. However, the term linear is used because the model $Y = a + C_1X_1 + C_2X_2 + C_3X_3 + C_4X_4 + ...C_nX_n$, is a linear function of the unknown parameters $C_1, C_2, ..., C_n$, where the response variable Y is related to n regressor variables [4] [16].

The equation (1.1) [] is basic multiple regression equation used to obtain the value of Y (dependent variable) corresponding to independent variables $X_1, X_2, ..., X_n$. In this study, Y denotes the CGPA, and X_1, X_2 denotes the attendance, SGPA

respectably.

$$\bar{Y} = a + C_1X_1 + C_2X_2 + C_3X_3 + C_4X_4 + ...C_nX_n \qquad (1.1)$$

Where,

Y = predicted value of the response variable,
a = Intercept,
C_i = Slope of the "estimated" regression equation associated with X_i .

The least-squares criterion estimates the parameters in such a way that it should be minimize the total error. This process also maximizes the correlation between actual and predicted variables (Y and Y bar). We define some formulas that describe the procedure for multiple regression analysis.

$$C_1 = \frac{\left(\Sigma X_2^2\right)\left(\Sigma X_1 Y\right) - \left(\Sigma X_1 X_2\right)\left(\Sigma X_2 Y\right)}{\left(\Sigma X_1^2\right)\left(\Sigma X_2^2\right) - \left(\Sigma X_1 X_2\right)^2} \qquad (1.2)$$

$$C_2 = \frac{\left(\Sigma X_1^2\right)\left(\Sigma X_2 Y\right) - \left(\Sigma X_1 X_2\right)\left(\Sigma X_1 Y\right)}{\left(\Sigma X_1^2\right)\left(\Sigma X_2^2\right) - \left(\Sigma X_1 X_2\right)^2} \qquad (1.3)$$

$$a = \bar{Y} - C_1X_1 - C_2X_2 \qquad (1.4)$$

Therefore, the coefficients of multiple regression models are calculated using equation no. (1.2), equation no. (1.3) & equation no. (1.4) [2] [3] [17]. However, the respective value of multiple regression variables C_1, C_2, and a are given below:

$C_1 = 0.00255$, $C_2 = 0.02187$, $a = -466.37$

The equation (4.6) gives resultant multiple regression equation.

$$\bar{Y} = -466.37 + 0.00255X_1 + 0.02187X_2 \qquad (1.5)$$

V. EXPERIMENTAL RESULTS

In this study, the last three years (2012-2015) Government Girls College (GGC) data is obtained [13]. The performance of proposed system is predicted by using multiple regression equations no (1.5). The Table III shows sample of forecasted values i.e. dependent variable (CGPA) based on independent variables and (i.e. SGPA and attendance) respectively.

Table III. Sample of Student Forecasted Performance

S.No.	SGPA (X_1)	Attendance (X_2)	CGPA (Y)	$\bar{Y}$
1	49.83	58%	52.58	48.22
2	56.17	74%	56.33	52.32
3	60.00	63%	55.29	61.28
4	64.83	75%	55.54	58.09
5	66.89	51%	59.06	57.97
6	55.56	72%	60.61	55.58
7	52.00	59%	67.78	66.42
8	77.78	58%	63.39	60.09
9	63.56	78%	63.50	66.44
10	74.22	57%	63.17	62.93
11	73.33	65%	63.17	61.39
12	62.50	77%	60.00	58.43
13	72.33	71%	70.25	71.53
14	69.17	66%	67.58	62.48

A bar graph showing the predicted patterns of student performance using Figure 1.

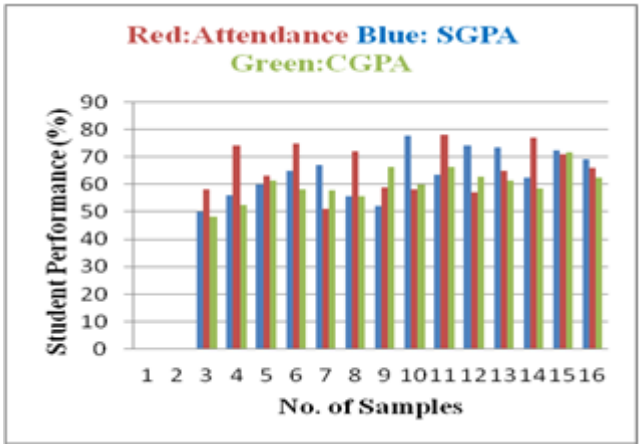

Figure 1: Bar Graph for Predicted Patterns of Student Performances against the SGPA and Attendance

Figure 1 clearly states that student performance patterns against SGPA and attendance. In bar graph green color shows CGPA performance against SGPA and attendance.

At last, to measure the performance, if degree of dispersion around the regression gets smaller then results came to be more accurate. To measure this variation, we shall use the measure called the Residual Error Estimate (REE) which is shown by equation (1.6).

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

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.

The Table IV shows residual (error) between actual CGPA and predicted CGPA.

Table IV. Residual Error Estimation

CGPA (Y)	Predicted CGPA $(\bar{Y})$	Residual $(Y - \bar{Y})$	Residual $(Y - \bar{Y})^2$
55.17	51.45	3.72	13.8384
58.00	57.43	0.57	0.3249
65.42	66.32	-0.9	0.81
62.29	61.38	0.91	0.8281
60.08	56.56	3.52	12.3904
58.06	58.65	-0.59	0.3481
59.83	57.27	2.56	6.5536
57.33	60.33	-3	9
57.88	56.61	1.27	1.6129
59.17	54.77	4.4	19.36
60.21	61.89	-1.68	2.8224
55.71	49.43	6.28	39.4384
58.17	60.21	-2.04	4.1616
55.71	51.43	4.28	18.3184
61.17	62.93	-1.76	3.0976
54.17	50.25	3.92	15.3664
63.78	67.12	-3.34	11.1556
56.78	54.54	2.24	5.0176
61.89	57.43	4.46	19.8916
55.28	53.85	1.43	2.0449
58.39	59.65	-1.26	1.5876
59.83	61.23	-1.4	1.96
67.75	75.54	-7.79	60.6841
56.08	49.92	6.16	37.9456
57.50	52.42	5.08	25.8064
60.17	65.25	-5.08	25.8064
58.79	57.45	1.34	1.7956
.	.	.	
.	.	.	
59.58	59.12	0.46	0.2126

			$\sum(Y - \bar{Y})^2 = 35451.17$

In Table IV, residual or sum of the squared errors of the prediction. We got the residual or error for multiple regression for three variables (formula is used above) is 13.61 which is higher than 10. Finally, we have got the prediction accuracy is 74.38% which is acceptable. Based on the results we can say that prediction results are good than linear regression approach.

V. CONCLUSION AND FUTURE SCOPES

Today, due to the competitive environment among educational institutes the prediction plays a key role to predict student academic performance and the process is being very complicated and challenging. As per the discussed work above this proposed system, predicts the student academic performances based on multiple regression based approach using three variables and found the accuracy of system is 74.38% respectively, which is more accurate than previous linear regression model. Today, education dataset is growing at very rapid speed and it needs the intelligence of human for effective prediction. In this regard, multiple regression approach seems to be more complicated and challenging. So in the future to overcome the drawbacks of these approach, the Neural Network (NN) based approach is proposed, which can simultaneously works on multiple variables as well as on large dataset.

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] B. Richard A., "Regression analysis: A constructive critique", Sage Publications, 2004.

[3] Armstrong Scott J., "Illusions in regression analysis", Journal of Forecasting (forthcoming), 28 (3):689. Doi:10.1016/j.ijforecast, 2012.

[4] World wide web multiple regression Types available at "http: www.enwikipedia.org/Multiple Regression Types".

[5] 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.

[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] 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.

[8] S. Huang and N. Fang, "Regression Models for Predicting Student Academic Performance in an Engineering Dynamics Course", American Society for Engineering Education, PP. 1-15, 2010.

[9] I. Potgieter, and M. Coetzee, "Employability attributes and personality preferences of postgraduate business management students." SA Journal of Industrial Psychology, Vol 39, Issue (1), pp.01-10, 2013.

[10] V. K. Gokuladas,. "Technical and non- technical education and the employability of engineering graduates: an Indian case study." International Journal of Training and Development 14.2 (2010): 130-143.

[11] V. K. Gokuladas, "Predictors of Employability of Engineering Graduates in Campus Recruitment Drives of Indian Software Services Companies." International Journal of Selection and Assessment 19.3 (2011): 313- 319.

[12] K. Saxena, S. Jaloree, R.S. Thakur and S. Kamley, "Linear regression Technique for student academic performance prediction", Accepted for Presentation in International Conference on Recent Trends in Computer Science and Electronics (ICRTCSE-18), Indore, MP, 7th July, 2018.

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

[14] 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.

[15] A. PENAˇ -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.

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

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