A Study on the Prediction of Student's Performance by applying straight-line regression analysis using the method of least squares

G.Narasinga Rao[1], Srinivasan Nagaraj[2]
[1]Assistant Professor, Department of Computer Science & Engineering, GMRIT, Rajam.
[2]Assistant Professor, Department of Computer Science & Engineering, GMRIT, Rajam.
Email Id: [1]narasingarao.g@gmrit.org, [2]sri.mtech04@gmail.com

***Abstract*— Now a days, the challenge for many educational institutions is to make their students do better in their exams and secure good percentage of marks so that many of them get placed in reputed multinational companies. If the educational institutions are able to predict the percentage of marks that the students are going to secure in their final exams, it will help them to focus only on those students who secure less percentage of marks. Thereby, they can make the weaker students improve in their respective subjects to secure good percentage in their final exams.**

Keywords-Data Mining, Data Sets, Classification, Linear Regression

I. INTRODUCTION[2]

Data Mining is the process of analyzing the data that resides in large data repositories. That is extracting the useful information from the large data sets available, thereby converting the useful information into knowledge. This is called as Knowledge Discovery in Databases (KDD). The knowledge that is extracted can be useful for variety of purposes. Predictive tasks are used to predict the value of a particular attribute based on the values of other attributes that are known. Predictive modeling refers to the task of building a model for the target variable as a function of the explanatory variables. There are two types of predictive modeling tasks: Classification, which is used for discrete target variables and Regression, which is used for continuous target variables. The objective of the paper is to predict the percentage of marks secured by the students in their final exams based on the percentage of marks secured by the students in their mid exam marks by applying straight-line regression analysis using the method of least squares.

II. METHODOLOGY[1]

Straight-line regression analysis involves a response variable, y, and a single predictor variable, x. It is the simplest form of regression, and models y as a linear function of x.

That is, $y=b+wx$, (1)

where the variance of y is assumed to be constant, and b and w are regression coefficients specifying the Y-intercept and slope of the line, respectively. The regression coefficients, w and b, can also be thought of as weights, so that we can equivalently write

$y=w0+w1x.$ (2)

These coefficients can be solved by the method of least squares, which estimates the best-fitting straight-line as the one that minimizes the error between the actual data and the estimate of the line. Let D be a training set consisting of values of predictor variable, x, for some population and their associated values for response variable, y.

The regression coefficients can be estimated using this method with the following equations:

$$W_1 = \frac{\sum_{i=1}^{|D|}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{|D|}(x_i-\bar{x})^2} \quad (3)$$

$$W_0 = \bar{y} - W_1\bar{x} \quad (4)$$

III. RESULTS

We have taken a student data set consisting of 49 student's information of a reputed institution considering the percentage of marks secured by the students in their mid exams and final exams of a semester. Out of the 49 student's data set, we have considered only 22 student's data as the training data set and applied the method of

least squares to predict the percentage of marks secured by the students in their final exams based on the percentage of marks secured by the students in their mid exams.

TABL 1 The following is the training data set that we considered to apply the method of least squares

Serial Number	Percentage of marks secured in mid examinations(x_i)	Percentage of marks secured in final examinations(y_i)
1	78	67
2	82	60
3	68	66
4	88	75
5	87	65
6	63	61
7	75	75
8	47	60
9	43	63
10	80	71
11	75	66
12	92	73
13	75	63
14	57	63
15	90	69
16	68	75
17	68	60
18	50	65
19	52	69
20	82	74
21	57	68
22	85	64
Total	1562	1472

Now by applying the method of least squares,

we get $\bar{x}$ =$\Sigma x_i/n$=1562/22=71 and $\bar{y}$ = $\Sigma y_i/n$=1472/22=66.9=67

Therefore

$$W_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}$$

= 656/4608 = 0.142361

That is w_1=0.142361

Also $W_0 = \bar{y} - W_1\bar{x}$

Therefore w_0=56.892369

Now based on the values of w_0 and w_1, if we give x value we can predict the value of y i.e., if the percentage of marks secured by the students in the mid exams is given, then we can predict the percentage of marks secured by the student in the final exams.

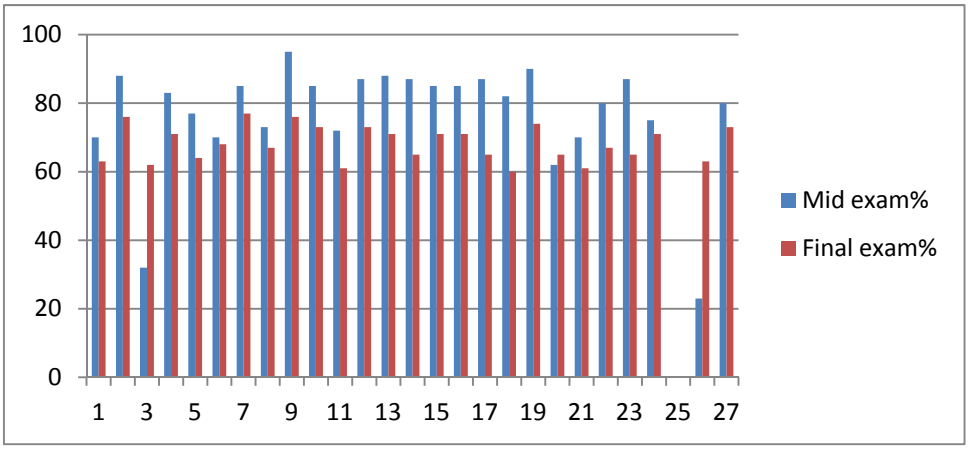

Figure 1: This graph represents the original data of the mid exam% and final exam% of the remaining 27 students out of 49 students

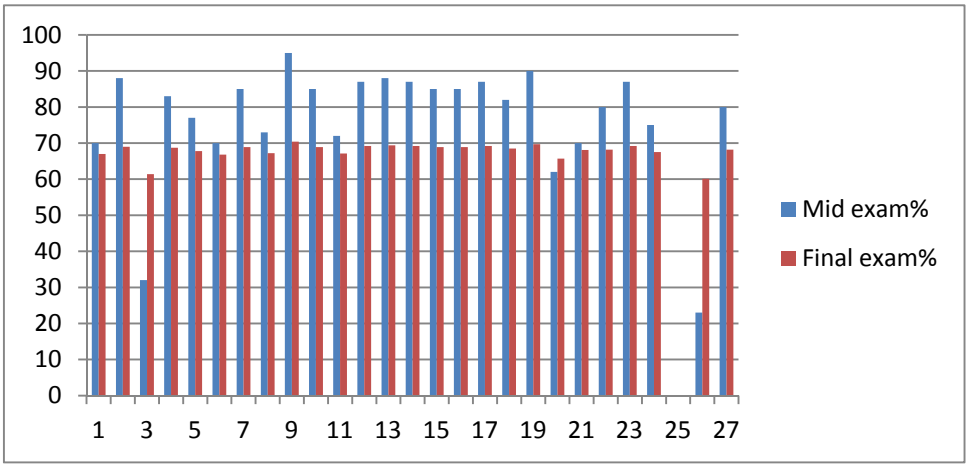

Figure 2: This graph represents the predicted data of the mid exam% and final exam% of the remaining 27 students (test data) by applying the method of least squares

IV. CONCLUSION

In this paper, we have taken student's data set consisting of 49 student's information of a department of a reputed institution consisting of the percentage of marks secured by the students in their mid exams and final exams. Out of 49 students, we have taken 22 students data and applied the straight line regression analysis using the method of least squares. Then we made a comparison between the original data and predicted data by the straight line regression analysis using the method of least squares of the remaining 27 students and obtained the results in the above graphs. The graphical results show that the error rate between the original data and the predicted data is very less. The error rate could also be due to the several factors like the student's attendance, attitude and the interest that he had in attending the course. Therefore we can apply the straight line regression analysis using the method of least squares for a college data set and predict the percentage of marks secured by the students of a college.

V. REFERRENCES

[1] Data Mining, Concepts and Techniques, 3/e, Jiawei Han, Micheline Kamber, Elsevier.
[2] Introduction to Data Mining: Pang-Ning tan, Micheline Steinbach, Vipin Kumar, Pearson
[3] Larose, D. T., "Discovering Knowledge in Data : An Introduction to Data Mining", *Hoboken, NJ, USA: Wiley*, 2005.
[4] S.Sujit Sansgiry, M.Bhosle and K. Sail, "Factors that affect academic performance among pharmacy students," American Journal of Pharmaceutical Education, 2006
[5] N. V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique,"
[6] Garcia, E. P. I., & Mora, P. M., "Model Prediction of Academic Performance for First Year Students", *Paper presented at the Artificial Intelligence (MICAI)*, 10th Mexican International Conference on, Nov. 26 2011-Dec. 4 2011.
[7] Sachin, R. B., & Vijay, M. S, "A Survey and Future Vision of Data Mining in Educational Field", Paper presented at the Advanced Computing & Communication Technologies *(ACCT)*, Second International Conference on 7-8 Jan. 2012.