Regression Model for Predicting Engineering Students Academic Performance

R.R.Rajalaxmi , P.Natesan , N.Krishnamoorthy ,S.Ponni

Abstract: Prediction of academic performance of the students helps the educator to develop the good understanding of student's community and take the healthy measures to make their learning comfortable and understandable. Many data mining techniques can be used to predict the performance of the students in academics. But the main objective of the paper is to use linear regression techniques to build a model which predicts the performance of the students in Engineering Discipline. The predictor or independent variables of the model contain how many hours spent on the internet in some activities based on the data collected. The output or dependent variable is the prediction of end semester examination grades i.e. CGPA (Cumulative Grade Points). Multiple measures are used to calculate and corroborate the models that were predicted along with the percentage of good predictions. The results show that the predicted model gives the better accuracy in prediction.

Keywords: Data mining; Engineering students; Internet use; Regression.

I.INTRODUCTION

Internet has a wide door to an innovative style of gaining knowledge. The amount of information available therein exceeds that of any physical library. There are many uses of internet but academic purpose conquers the highest desirable position as far as students are concerned. Generally, the college-going students have an immense variety of subjects in the core area which is sometimes difficult for the students to understand. So they have chosen internet where there is a lot of information provided about the area of their study. Nowadays internet plays a major role in and around the people and makes them depend on it for each and everything. This is mainly affected by student's community where they use internet for various purpose like getting notes for studies, doing the assignment and other related activity and also for communications. Almost every engineering student is required to take the corresponding engineering course which is highly impacted over the area which they have chosen. The courses contain all the essential and basic subjects from all the engineering subjects. So there is a chance of engineering students using internet for the various purpose including the entertainment along with the studies. Prediction of student's academic performance has long been observed and considered to be an important research topic in many disciplines because it aids the educator and learners. Educators can use the predicted results to identify the number of students who will do well, averagely or poorly in a particular class to take measures accordingly. Many areas can be chosen to arrive for the prediction of student's academic performance. Data mining is a process of extraction of already existing previously unknown patterns and worthy information from the dataset. One can analyze the data for identifying the healthy information and patterns which can be used for further predictive analysis. It is a major such area where prediction can be done easily using mining algorithms and can infer the patterns [4]. There are several tasks that can be used to build a prediction model which includes classification, regression and categorization and also generate rules for prediction [3]. Generally the classification is used for the purpose of prediction. Some of the algorithms are Decision tree, Neural networks[5],SVM [8], Naive Bayes, KNN [9]. Our main idea is to present the model by using the regression techniques which can be used for prediction of student's performance in academics. A diverse of mathematical methods such as linear regression, logistic regression, multivariate linear regression have been used to predict the academic performance of the students. Multivariate linear regression is one such commonly used a mathematical model in the study of prediction. It is easy to use and understand since it doesn't require any mathematical skills for the researchers to learn. It also provides a model equation which allows the educators and researchers to know how the predicted values obtained. It also helps in interpreting the results which were experimented.

II. RELATED WORKS

In this section, the various techniques for prediction using different data mining techniques in the discussed below. There are many applications and area where prediction can be applied to predict some useful information. One such application/ area is Healthcare prediction. In the healthcare environment, there are many diseases which can be predicted before the analysis.

Revised Manuscript Received on March 25, 2019.

R.R.Rajalaxmi, Professor, Department of Computer Science &Engineering, Kongu Engineering College,Erode

P.Natesan , Associate Professor , Department of Computer Science &Engineering, Kongu Engineering College,Erode

N.Krishnamoorthy, Assistant Professor(Sr.Gr) , Department of Computer Science &Engineering, Kongu Engineering College,Erode

S.Ponni, Research Assistant , Department of Computer Science &Engineering, Kongu Engineering College,Erode

Some of the researchers focused on heart disease and the discovery of rules in medical data to predict the existence of disease [2] by using a data mining technique called association rule mining which maps the medical data to the transaction and forms association rules. Association rule mining can be used to extract the relationship between the item sets. Some authors presented a study which compares the data mining techniques used by different researchers for prediction of heart disease [1] and tabulated the accuracy of using different data mining techniques such as Decision tree, NaiveBayes, JRip, CART etc and finally concluded with the best data mining technique with number of correctly classified samples for further understanding.

Our present study is mainly focused on the predicting the student's academic performance by using data mining algorithms. Here are some researchers who conveyed on the usage of data mining algorithms for the student's academic performance. Oloruntoba et al(2017) proposed a study based on the prediction of student's academic performance using data mining techniques. The study identifies the relationship between the student's academic performance and their final scores. The model is built using the support vector machine technique and it was compared with other classification algorithms. The final result has shown that the accuracy obtained through SVM classification is much greater than the other algorithms. A research project done by Dorina Kabakchieva(2012)Bulgarian University mainly focused on the usage of data mining techniques for university management. The results achieved by selected data mining algorithms for classification doesn't reveal any worthy outcomes. Zlatko(2010) explored the student's demographic attributes along with their corresponding study environment which is used for the analysis of these factors affecting their success rates in their course of the study. The results show that the important factors to distinguish between successful and unsuccessful students and for predicting the category of students, the CART algorithm is used which produce an overall percentage of 60.5%. It does not contain adequate information for distinguishing between successful and unsuccessful students. Some authors explored the difference between data mining techniques [5] [6][7] and explored the comparison of the methods for educational learners and provided a better predictive model among all the data mining techniques. Some researchers focused on the use of classification algorithms and provided the comparison of all the classifiers in their paper [10]. In all these works, the authors concentrated on the students' performance prediction using different data mining techniques to carry out the analysis but the work mainly focuses on building and validating the regression technique of undergraduate students of Engineering stream who use the internet for various purposes. The rest of the paper is unfolded as follows, Section 3 discusses the objective and techniques in detail followed by the data collection and how research method is applied for the above-proposed problem and in last Section, the detailed explanations of the performance metrics are discussed. Finally, it ends with the conclusion along with the future work of the paper.

III.OBJECTIVE OF THE PRESENT STUDY

The main objective of the present study is to develop a set of multivariate linear regression models to predict the academic performance of the Engineering students based on the category of CGPA which they lie. The outcome of the regression models is the students end semester examination based on the inputs collected. The predictor of the regression model contains the independent variables as

x_1: Usage of Internet for the academic/education purpose

x_2: Usage of Internet for the Entertainment purpose

x_3: Usage of Internet for the Communication purpose

x_4: Active duration in social media networks

x_5: Usage of Internet before the end semester exam

y : Cumulative GPA

where the variable from x_1 to x_5 represents the student's behavior on the usage of internet for various activities and y represents the outcome of the model. The scope of the project is to observe the patterns of using the internet by various undergraduate engineering students to predict their cumulative grade points. Formulating the equations of the predictive models and analyzing the accuracy of the predicted models is some of the research questions included in the present study.

IV. OVERVIEW OF REGRESSION

Regression is a statistical method to identify the relationship between the variables present in the data. It mainly focuses on the relationship between the dependent variable and independent variables which is otherwise called as predictors. It helps to understand the changes occur in the value of dependent variable when anyone of the independent variables is changed. By using the value of the independent variable, an equation is formulated which contains the independent variables along with some coefficients and the slope value. There are lot many types of regression techniques. One such is linear regression technique which is mainly used for prediction. The linear regression is used to examine the relationship between one dependent variable and one or more independent variables. The simple linear regression with one dependent variable and one independent variable is given in Equation (1):

$$y = c + m*x \quad (1)$$

where y = predicted dependent variable value , c = constant, m = coefficient of the regression and x = the value of independent variable. There are several types of linear regression techniques: Simple linear regression, Multiple linear regression, Logistic regression, Ordinal regression, Multinominal regression and Discriminant analysis. For our study, the multiple linear regression is used. Multiple linear regression is used for one dependent variable and more than two independent variables containing dataset. The multiple linear regression equation is given in Equation (2):

$$y' = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \ldots\ldots\ldots b_kx_k \quad (2)$$

The multiple linear regression can be calculated by the following steps:

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \ldots\ldots\ldots\ldots x_{k1} \\ 1 & x_{12} & x_{22} \ldots\ldots\ldots\ldots x_{k2} \\ 1 & x_{13} & x_{23} \ldots\ldots\ldots\ldots x_{k3} \\ . & . & \ldots\ldots\ldots\ldots\ldots\ldots \\ 1 & x_{1n} & x_{2n} \ldots\ldots\ldots\ldots x_{kn} \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & . & 1 \\ x_{11} & x_{12} & x_{13} & . & x_{1n} \\ x_{21} & x_{22} & x_{23} & . & x_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{k1} & x_{k2} & x_{k3} & . & x_{kn} \end{bmatrix} Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_k \end{bmatrix}$$

$$B = (X^TX)^{-1}X^T.Y \quad (3)$$

$$B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \\ b_k \end{bmatrix} \quad (4)$$

where X contains the value of the independent variables and Y contains the value of dependent variables, X^T contains the transpose matrix of the matrix X. Equation (4) contains the predicted coefficients of the independent variables in a matrix form. Thus these variables are written in the following format,

$$y' = a + b_1x_1 + b_2x_2 + b_3x_3 \ldots\ldots\ldots + b_kx_k \quad (5)$$

The training dataset helps us to formulate the equation of a regression line and test dataset is used to predict the accuracy of the predicted equation. The test dataset is passed as values for the above Equation (5) to get the predicted scores.

V. DATA COLLECTION

Based on the proposed study, the questionnaire was prepared to gather information from the students. The input data on student performance in academics was collected from students of 150 undergraduate engineering disciplines. The questionnaire contains some demographic questions about the students and some questions related to the internet usage patterns and their end semester grades. The questionnaire was given to 150 students of different year of study in same branch in the institution. The data is analyzed to build a predictive model. Table 1 shows some demographic information about the undergraduate students.

Table 1. Depicts the demographic information about the students

Category	Attributes	Count
Gender	Male Female	50 100
CGPA wise split up	7.1-8.0 (Model 1) 8.1-9.0 (Model 2) 9.1-10 (Model 3)	70 56 24
Year wise	III IV	61 89
	Total number of students	150

VI. Research Method of the Present Study

A total of 150 undergraduate engineering students in their corresponding semesters were included in the present study to develop and validate the predicted regression model. The following steps describe the flow of the research methodology.

- Collecting the data from the undergraduate engineering students based on their performance in the academics up to the current semesters is named as S1, S2, S3, S4, S5, S6, and S7. Understanding the collected data and performing analysis over it
- Select the data collected and split it as training and testing dataset.
- After splitting the dataset, the multivariate linear regression technique is applied on the training dataset and from the regression Equation got it is used for predicting by passing the test dataset.
- In training dataset, there will be three categories of students based on the CGPA criteria. Based on the categories of training dataset, the models will be predicted. Figure (1) explains the flow of predictive model of the present study.

From the whole dataset, a sample category of students is considered for explanation. The values represent the regression coefficient of the variables (independent variables). The first value represents the intercept of the model.

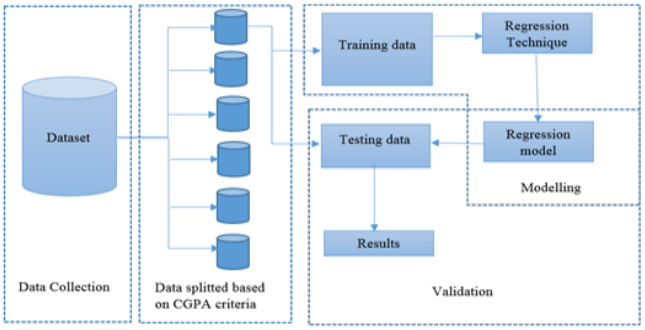

Fig 1. Flow of Predictive model implementation

So the prediction model can be written as the following equations for the category of students who lie in the cgpa range of Model 1 [M1 (7.1 -8.0)],

$$y'=7.47+0.26x_1-0.22x_2+0.48x_3-0.03x_4-0.66x_5 \quad (6)$$

Here are some sample predicted models given below to show the use of regression techniques for prediction of the students' academic scores of different category. Likewise, the models for each category of students can be calculated and shown below in the following equations.
and shown below in the following equations.

$$y'=7.32+0.055x_1+0.0633x_2+0.16x_3+0.007778x_4-0.19556 \quad (7)$$

$$y'=8.61-0.01x_1+0.01x_2+0.2x_3-0.07x_4+0.1x_5 \quad (8)$$

The presence of independent variables varies for each and every category of the students. From the projected models, the dependent variable can be calculated by passing the test data values.

VII. MODEL VALIDATION

The further model validation should be done to check how accurately the model is predicting and the test data is passed. In Equation (6), the value for the independent variable is substituted shown below. The sample test set values are given below for M1,

$$\textbf{M1:}x_1=4\ ;\ x_2=4\ ;\ x_3=3\ ;\ x_4=5\ ;\ x_5=3 \quad (9)$$

CGPA Criteria	Standard Error Rate
7.1 – 8.0	0.2473
8.1 – 9.0	0.304
9.1 – 10	0.1369

When the values in (9) substituted in Equation (6),

Table 2. Evaluation of Performance metrics

$$y'=7.47+0.26(4)-0.22(4)+0.48(3)-0.03(5)-0.66(3)$$

$$y'=6.94 \quad (10)$$

The value in (10) shows the predicted score value from the predicted model M1. The original score of the student given below in (11),

$$y=7.1 \quad (11)$$

Some of the sample test set values given below for Model 2, ie. in cgpa range (8.1 -9.0),

$$\textbf{M2:}x_1=11;\ x_2=8\ ;\ x_3=3\ ;\ x_4=2\ ;\ x_5=1 \quad (12)$$

When the values in (12) substituted in Equation (7),

$$y'=8.6968 \quad (13)$$

The value in (13) shows the predicted score value from the predicted model M2. The original score of the student given below in (14),

$$y=8.5 \quad (14)$$

Some sample test set values given below for Model 3, ie. in cgpa range (9.1-10),

$$\textbf{M3:}\ x_1=5;\ x_2=2;\ x_3=1;\ x_4=3;\ x_5=5 \quad (15)$$

When the values in (15) substituted in Equation (8),

$$y'=9.07 \quad (16)$$

The value in (16) shows the predicted score value from the predicted model M3. The original score of the student given below in (17),

$$y=9.16 \quad (17)$$

From the original scores and the predicted scores of the models, it reveals that the value of the predicted models M1, M2, M3 lies in the range ± 0.5 from the original scores. To prove this, the standard error is used to measure the accuracy of the regression predictions. It tells how far the values get deviated from the regression model. The equation is given by,

$$RMSE=\sqrt{\frac{(Y-Y')^2}{N}} \quad (18)$$

where Y represents the actual observation and Y' represents the predicted observation and N is the sample size. By using (18), the standard error rate was calculated and the results were tabulated in the table 2 which infers the error rate for each CGPA criterion students.

From the above table, it can be clearly noticed that the values are slightly deviated from the original regression line.

VIII. CONCLUSION

Generally various data mining methods can be applied for the analyzing the performance of the students. One such method is regression. The present study detailed the prediction of the students' performance in the end semester by the usage of internet of the students in their day to day life as the input. The model is built by using multivariate linear regression technique. From the model predicted, it shows how the value of the dependent variable differs based on the value of the independent variables. From the above, the educators can analyze the performance of the class and can also improvise the teaching techniques used based on the result of each category of the engineering students.

REFERENCES

1. Syed Immamul Ansarullah, Pradeep Kumar Sharma, Abdul Wahid, Mudasir M Kirmani, "Heart Disease Prediction System using Data Mining Techniques: A study", International Journal of Engineering Sciences & Research Technology, 1375-1381, (August 2016).
2. Carlos Ordonez, Edward Omiecinski, Levien de Braal, Cesar A. Santana, Norbert Ezquerra, Jose A. Taboada, David Cooke, Elizabeth Krawczynska, Ernest V. Garcia, "Mining Constrained Association Rules to Predict Heart Disease", IEEE Explorer, 433-440 (2001).
3. Sivagowry .S, Dr. Durairaj. M and Persia.A, "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease", IEEE Explorer (April 2013).
4. Pooja Thakar, Anil Mehta, Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", International Journal of Computer Applications, 60-68 (January 2015).
5. Norlida Buniyamin, Usamah bin Mat, Pauziah Mohd Arshad, "Educational Data Mining for Prediction and Classification of Engineering Students Achievement", IEEE Explorer, 49 -53 (2015).
6. Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashida, "A Review on Predicting Student's Performance using Data Mining Techniques", Elsevier, 414 – 422, (2015).
7. Geraldine Gray, Colm McGuinness, Philip Owende, "An application of classification models to predict learner progression in tertiary education", IEEE Explorer, 549 – 554 (March 2014).
8. S.A. Oloruntoba , J.L.Akinode, "Student Academic Performance Prediction Using Support Vector Machine", International Journal of Engineering Sciences & Research Technology, 588 – 598 (December 2017).
9. Dorina Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms", International Journal of Computer Science and Management Research, 686 – 690(November 2012).
10. Hilal Almarabeh, "Analysis of Students' Performance by Using Different Data Mining Classifiers", Modern Education and Computer Science, 9-15(2017).
11. Zlatko J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data", Proceedings of InSITE, 647- 665 (2010).