Student Performance Prediction by means of Multiple Regression

Anand Tamrakar[1], Dr. J P Patra[2], Deepak Khadatkar[3]

[1]Asst. Professor, Shri Shankaracharya Institute of Professional Management & Technology, Raipur, a.tamrakar@ssipmt.com
[2]Professor, Shri Shankaracharya Institute of Professional Management & Technology, Raipur, patra.jyotiprakash@gmail.com
[3]Asst. Professor, Shri Shankaracharya Institute of Professional Management & Technology, Raipur, d.khadatkar@ssipmt.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 May 2021

***Abstract*:** Education organizations incorporate and store enormous volumes of information, for example, student involvement, enrolment records, and academic records. Mining such information yields fortifying data that serves to institutional administrators. Hefty data yields ambiguity for Educational Data Mining (EDM) which reduces the correctness of prediction model applied over EDM. Subsequently going through numerous literature we came into conclusion that sorting out data into category is a characteristic decision. Essentially, arranging data into category is prevalent in numerous logical fields as per institutional administrator's requirement. In this research we have projected a prediction model as Multiple Regression and by means of curve fitting we found the relation of independent and dependent variable with lesser MSE which will predict student performance parameters.

Keywords: *ML, EDM, ID3, F-SCORE, Regression*

1. Introduction

Education organizations commonly experience the ill effects of restricted funding, work issues, and poor thoughtfulness regarding real training. Nowadays, most of the educational administration are facing issue of how to enhance the performance of students, although immense effort put for the same, henceforth a mechanism required which will provide an analytic information which will help education policymakers to take decisions to improve performance of students.

EDM (Education Data Mining) develops methods to that will help education system to cherish their performance. Education system having massive amount of data like student bio details, student test results, student interest, student's day by day activity etc., these data very useful information. EDM technology [1] can provide assistances for enlightening the excellence of education, skilled development, and ability policy maker can improve through examinations of key areas.

In this paper we have proposed use of machine learning classifier for furcating of student performance i.e. proposed system will predict the student result based on their earlier test results. Machine leaning classifier applied over data, depend on nature of the data that we want to predict if the data is heterogeneous in nature then, we need to apply unsupervised machine learning classifier for example we want to forecast rainfall of any state here we can group data that need to predict, and if data is homogenous in nature in which response can be grouped in confined class. Fig.-1 depicts the categories of machine learning classifiers and algorithms.

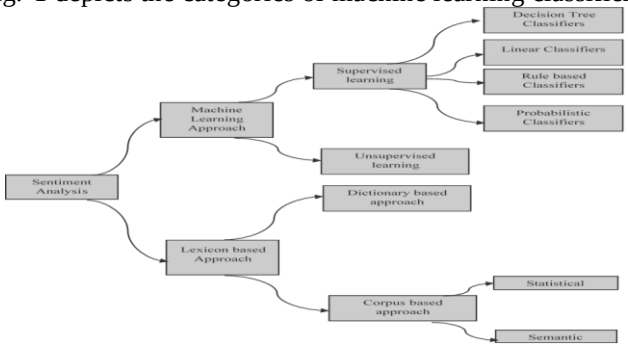

Fig.1. Classification techniques

"With the increasing capabilities of machine learning, there is a unique opportunity to personalize learning to individual students," says Erik Choi, Principal Researcher at Brainly.

With the cumulative competences of machine learning (ML), there is an exceptional chance to personalize learning to each and every student. Even we can predict the performance of student so as to he/she can know, in which section they need improvement and can perform well.

In this research we have applied proposed algorithm over Engineering graduate class test data. We took two class test marks as independent variable. We have applied curve fitting to find the relation among dependent and independent variable.

Let us understand machine learning model in the form of mathematical equation, if we read input data, there will be a dependent variable and an independent variable. Recognizing the dependent and independent variables in a mathematical equation will help us to classify what you are answering for in the equation. The independent variable is a variable whose value regulates the value of the dependent variables. Independent variable is plotted on the X axis, and the dependent variable is plotted on the Y-axis. Other variables may also be present in equations. These may be constants or other variables. They may be given to you or you may be required to get them by performing curve fitting. The example below illustrates this point.

$$y = mx + C \text{.....(1)}$$

Where y = dependent (Meant for Prediction) variable

x = independent (Other than response variable) variable

m and C = constants

Example Dataset

Table 1. Example dataset

X- independent variable	Y- dependent variable
0	80
30	100
40	120
50	140
60	160
70	180

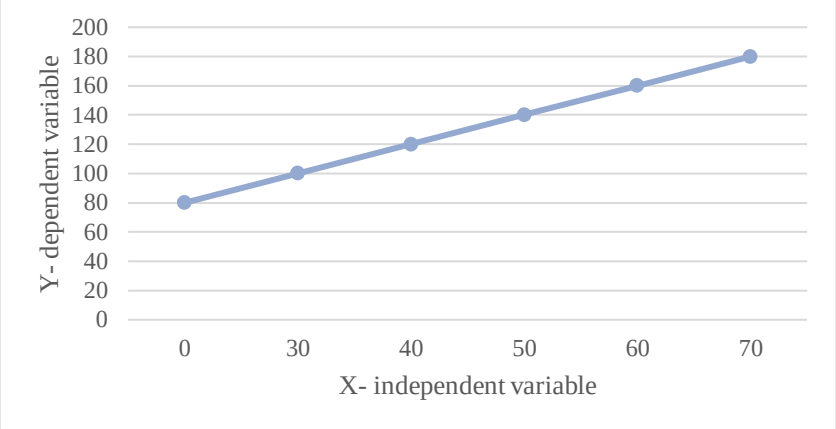

Fig.2. Plot between X and Y depend in Table 1

Curve fitting, also recognised as regression analysis, which will identify the "best fit" line or curve for a sequence of data points. Curve fit produces the equation that can be used to identify points along the curve.

In certain cases, we will be finding problem in finding equation. In its place, we can use curve fit to smooth the data and improve the look of plot.

There are several different models available for curve fitting:

- Straight Line
- Logarithmic

- Exponential
- Power
- Logistic Regression
- Non-logarithmic X-values
- Log10-transformed X-values
- Polynomial
- Gaussian
- Holt-Winters Forecast

2. Background Study

Sunita Jacob Radhai S. et al. this paper is grounded on a survey directed for the students and teachers to study and determine whether our higher education system and administrator of institute are prepared for introducing digital e-learning. Paper also deliberates whether the necessity of the hour is conventional learning method, e- learning method or an amalgamated method to both.

The researchers Magdalene D et. al. proposed Association rule mining to extracts expedient information from a hefty set of data. Similarly, this method is functional to student's data, their methods stated are cast-off for matching the organization with the students. This process is very difficult and includes a number of stages. Raheela, A et. al. presented a case study on forecasting performance of students [2]. The data of four academic associates comprising 347 undergraduate students have been mined with different classifier. The results obtained showed a reasonable accuracy. Recently, Abdulsalam et. al. investigated a method based on Decision tree algorithms BfTree, J48 and CART to predict student performance [3].

S. No.	Author/Title/Publication	Conclusion
1.	Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang et. al./ Data Mining for Internet of Things: A Survey/IEEE 2014	In this paper, author examination studies on smearing data mining technologies to the IoT, which entail of clustering, classification, and frequent patterns mining technologies, from the perspective of infrastructures and from the standpoint of services.
2.	Pooja Kumari et. al./An Efficient use of Ensemble Methods to Predict Students Academic Performance/IEEE 2018	The accuracy of the projected model is attained by using Ensemble Methods. Author have applied Bagging, Boosting, and Voting method that are the common ensemble methods.
3.	Y. K. Saheed Et.Al./Student Performance Prediction Based On Data Mining Classification Techniques/ Nijotech 2018	The investigational results presented that an ID3 accuracy of 95.9% , specificity of 95.9%, precision of 95.9%, recall of 95.9%, f-measure of 95.9% and incorrectly classified instance of 3.83. The C4.5 gave an accuracy of 98.3%, specificity of 98.3%, precision of 98.4%, recall of 98.3%, f-measure of 98.3% and incorrectly classified instance of 1.70.

3. Problem Identification

Students' academic performance is the major issue in front of most of the education policy makers. Forecasting of student performance index may help administration to take measures to improve student's academic performance. After going through several literature we came across following bottleneck of existing tutoring system as follows:

- Existing machine learning classifier needs improvement to increase the accuracy. Relation among predictor and response variable plays major role to improve the machine learning model accuracy.
- Value of MSE (Mean Square Error) is high while modelling between impendent and dependent parameters, hence accuracy decreases Scope of smart tutoring.

4. Solution Methodology

The institution goal is to develop a platform that provides real-time feedback and assistances online tutors become better at tutoring. For example, the system will perceive if a student's response to a concept follows a pattern of misinterpretation. By giving premature warning to teachers, the platform can help preclude problems

further in the teaching learning.

Figure 3 shows the layout of proposed scheme, Devices will be gather the student performance index i.e. their class test marks, gathered data will be stored in storage device. Upon gathered data (Knowledge base) we will apply our proposed algorithm, which will train the data and predict the student performance.

Predicted student performance provides the positive feedback to administrator of institute so that they can identify slow learner and fast leaner students, accordingly they can do some assignment to improve the performance of slow learner students. Matlab tool delivers curve fits that can be cast-off in both of directly above situations.

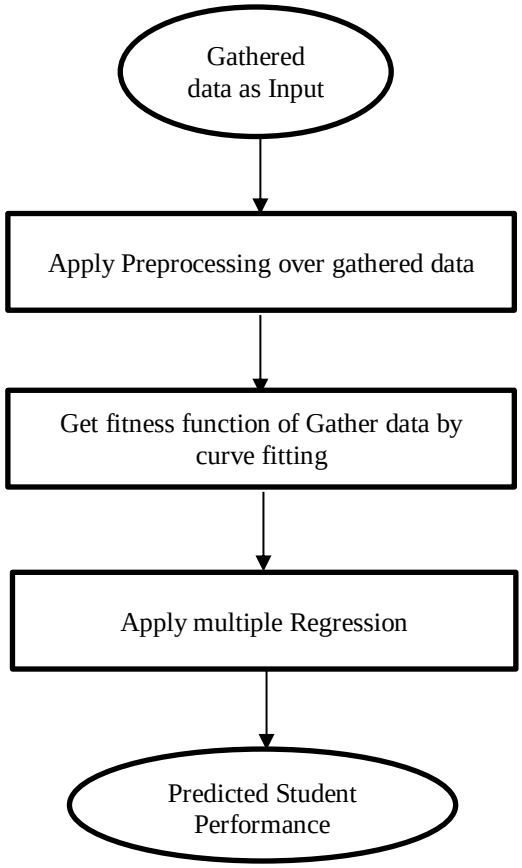

Fig.3. Proposed ML Model

Proposed Algorithm

Step-1. Input training dataset.
Step-2. Input test dataset.
Step-3. Process training dataset to find fitness function between dependent and independent variables of training dataset.
Step-4. Apply polynomial curve fitting.
Step-5. Find optimal fitness function as MSE (Mean square error) will be lesser.
Step-6. Apply regression statistical modelling as per optimal fitness function.
Step-7. Predicted data as output.
Step-8. Apply classifier to predicted output.

Multiple Regression (KNN)

Regression (featTrain classTrain, featTest, classTest, featName, classifier)

/*featTrain- A NUMERIC matrix of training features (N x M)
classTrain- A NUMERIC vector representing the values of the dependent variable of the training data (N x 1)
featTest- A NUMERIC matrix of testing features (Nts x M)

classTest- A NUMERIC vector representing the values of the dependent variable of the testing data (Nts x 1)
featName- The CELL vector of string representing the label of each features, (1 x M) cell*/
//classifier as KNN Regression
Step-1 NNBestFeat = floor(Datapoints()/10) //nearest neighbor
Step-2 trainModel=KNN Regression model
Step-3 NNSearch=Initialize earch function for KNNReg as linearsearch
//Set the distance measure for NNSearch
Step-4 distFunc = Euclidean distance (or similarity) function
Step-5 trainModel.setNearestNeighbourSearchAlgorithm (NNSearch)
Step-6 trainModel.setKNN(NNBestFeat)

Curve fitting is the procedure of assembling a curve, or mathematical function that has the best fit to a series of data points, conceivably subject to constraints. Curve fitting can comprise either interpolation, where a strict fit to the data is essential, or smoothing, in which a "smooth" function is constructed that just about fits the data.

Regression analysis, which emphases more on questions of statistical inference such as how much indecisiveness is existing in a curve that is fit to data experiential with random errors. Fitted curves can be cast-off as an aid for data visualization, to infer values of a function where no data are attainable, and to encapsulate the relationships among two or more variables. Extrapolation mentions to the use of a fitted curve beyond the range of the observed data, and is subject to a degree of doubt since it may reproduce the method used to build the curve as much as it echoes the experiential data.

If we apply curve fitting to any dataset then there are some terms which we need to know:

Dependent and Independent Variables: The 'independent' variable in machine learning is what we want control. The 'dependent' variable is what we want to extent, i.e., it depends on the independent variable. The 'coefficients' are the parameters that the fitting algorithm estimates.

For example, if we have census data, then the year is the independent variable since it does not be contingent on anything. Population is the dependent variable, because its value depends on the year in which the census is taken. If a parameter like growth rate is part of the model, so the fitting algorithm approximations it, then the parameter is one of the 'coefficients'.

Table 2. Gathered Dataset Fragment

Roll No.	Class Test-1 Marks	Class Test-2 Marks	End Semester Result
3422214001	81.66667	83.33333	1
3422214002	80.83333	80.83333	1
3422214003	85.83333	81.66667	1
3422214004	79.16667	72.5	1
3422214006	81.66667	70.83333	1
3422214007	80	81.66667	1
3422214008	70.83333	74.16667	1
3422214010	82.5	87.5	1
3422214011	90.83333	90	1
3422214012	91.66667	89.16667	1
Independent Variables			Dependent Variable

5. Result and Discussion

For implementation of our proposed smart tutoring we have used Matalb 2015b, dataset used in our experiment is data gathered from engineering college student dataset. We have obtain relation among dependent and independent variable using Matlab curve fitting tool.

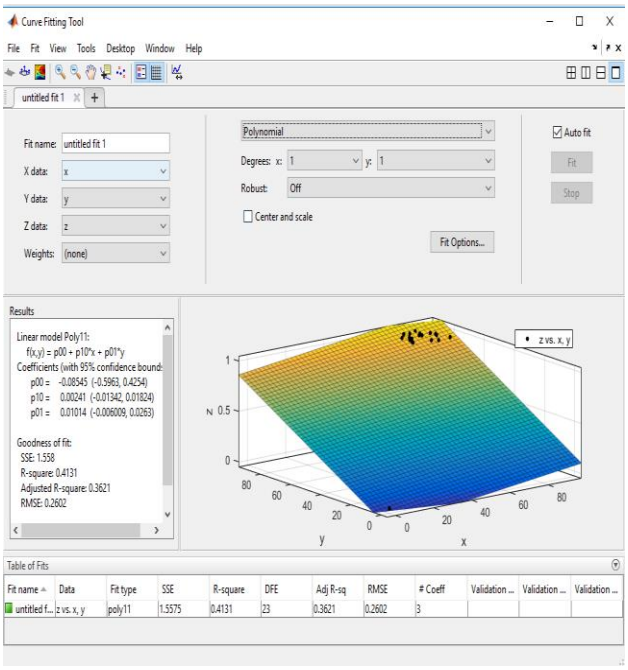

Fig.4. Curve Fitting

```
Linear model Poly11:
fitresult(x,y) = p00 + p10*x + p01*y
Coefficients (with 95% confidence bounds):
  p00 =  -4.579e+08  (-1.496e+09, 5.802e+08)
  p10 =      0.1338  (-0.1695, 0.4371)
  p01 =      0.9114  (0.7448, 1.078)


gof =

           sse: 1.0719e+03
       rsquare: 0.8487
           dfe: 23
     adjrsquare: 0.8355
          rmse: 6.8266
```

Fig.5. Matlab Curve Fitting best fit

Figure 5 snippet of matlab command window which show best fitness result, R-square, SSE, RMSE values.

```
3422214001,81.66666667,83.33333333,0,1
3422214002,80.83333333,80.83333333,0,1
3422214003,85.83333333,81.66666667,0,1
3422214004,79.16666667,72.5,0,1
3422214006,81.66666667,70.83333333,0,1
3422214007,80,81.66666667,0,1
3422214008,70.83333333,74.16666667,0,1
3422214010,82.5,87.5,0,1
3422214011,90.83333333,90,0,1
3422214012,91.66666667,89.16666667,0,1
3422214013,74.16666667,80.83333333,0,1
3422214014,86.66666667,79.16666667,0,1
3422214017,78.33333333,85.83333333,0,1
3422214018,70,70,0,5.000000e-01
```

Fig.6. Predicted Value of Student End semester result

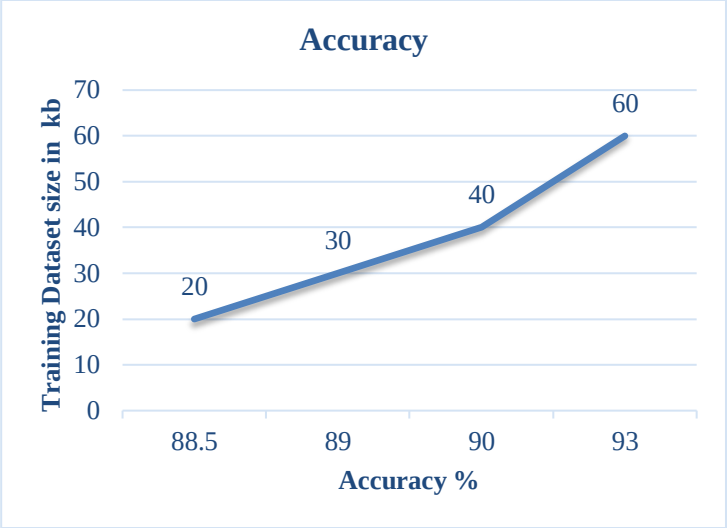

Fig.7. Performance Plot

Figure 7 shows the performance plot, from there we can say that as the training dataset size increases, accuracy of proposed algorithm also increases.

6. Conclusion

The institution goal is to develop a platform that provides real-time feedback and assistances better at tutoring. Nowadays machine learning provides the proactive information from historical data which helps decision makers. Proposed method achieved lesser MSE (Mean Square Error), we have achieved average prediction accuracy as 90%.

References

1. Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang et. al. Data Mining for Internet of Things: A Survey IEEE 2014
2. Pooja Kumari et. al./An Efficient use of Ensemble Methods to Predict Students Academic Performance/IEEE 2018
3. Y. K. Saheed Et.Al./Student Performance Prediction Based On Data Mining Classification Techniques/ Nijotech 2018