THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-41-December-2019

Student Grade Prediction Using Data Analytics

Lahari Pampati, B. Varshini, V. Tanmaya and P. Archana
Department of Computer Science and Engineering, Sreyas Institute of Engineering and Technology, Hyderabad

ABSTRACT

In the world of open education systems, students have flexibility to learn anything with ease as the learning content is easily available. But this facility can make student complacent. Therefore, it becomes difficult to predict the student's performance in advance. In this project, an attempt is made to help the student to know his performance in advance. This is done by using univariate linear regression model. This would help students to improve their performance based on predicted grades and would enable teachers to identify those individuals who need assistance.

The Main Objective of "Student Grade Prediction Application" is to implement a simple algorithmic model that predicts the score of an individual student at he /she end of the year. "G3" or the final grade is our label (output) and the rest of the columns will be our features (inputs).

INTRODUCTION

In the present educational systems, student performance is getting worsen everyday gradually . Predicting student performance in advance can help students as well as their teachers to keep track of progress of the student. Many educational institutes have adopted continuous evaluation system today. Such systems are favourable to the students in improving their performance . The purpose of the continuous evaluation system is to help the regular students in their academics. In continuous evaluation system, unit tests or class tests are conducted at regular period. To have consistent performance in the final grade it is required to appear in all the unit tests or class test.

The core function of Student Grade Prediction is to help the student to know his/her performance in advance by using univariance Linear Regression Model. Such techniques would help the students to improve their performance based on the predicted grade and would enable teachers to identify those individuals who might need assistance.

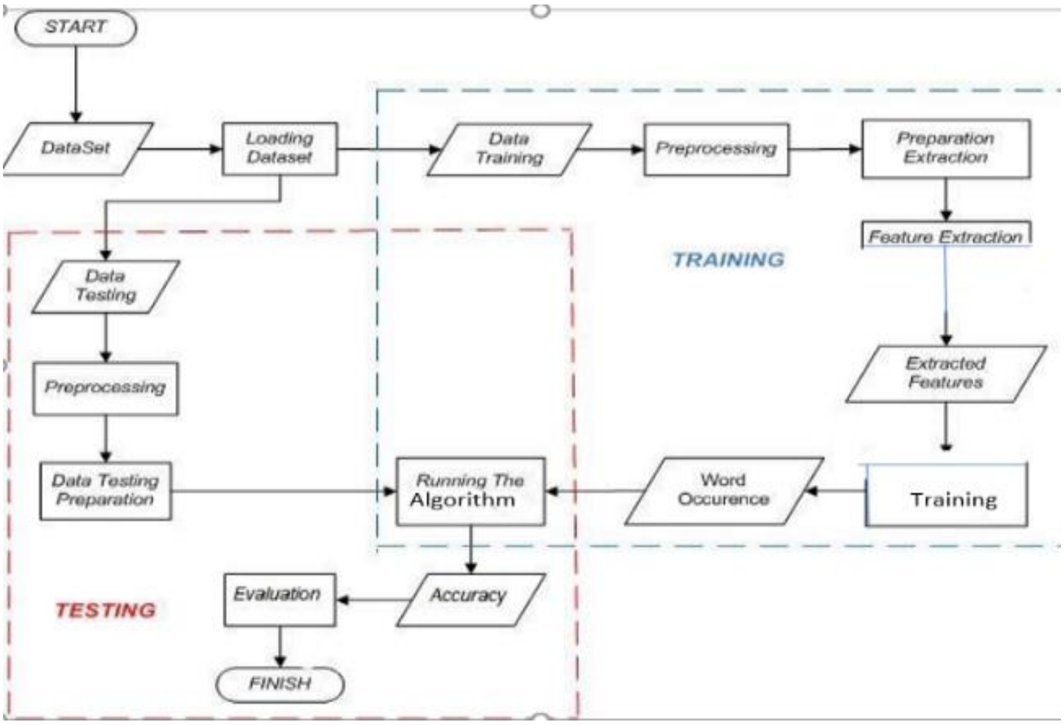
IMPLEMENTATION

Let's consider a school and data of its students. We collect every single minute information about the students and we put in an excel file. This gives a shape to the data. Suppose that we have around thirty three different attributes or features of data that we collected for every student. The first step in our process is identifying the attributes. We should now consider only those attributes that depend on the Grade of the student. To find out these attribute we should find correlations. Correlations give the dependency of a dependent variable on an independent variable. We then consider only the attributes that are mostly related and discard the rest. We then change the data types of the file as the system cannot compute multiple data types at once. So we convert the data type of the file and send it for training and testing. In training and testing, the data is partitioned randomly for training and testing. The partitions are sent for training and testing respectively. Then, it sent for a fit function using Linear Regression algorithm. The result is then predicted and the result obtained is the final grade (G3). For portraying accuracy it is shown using Boxplot.

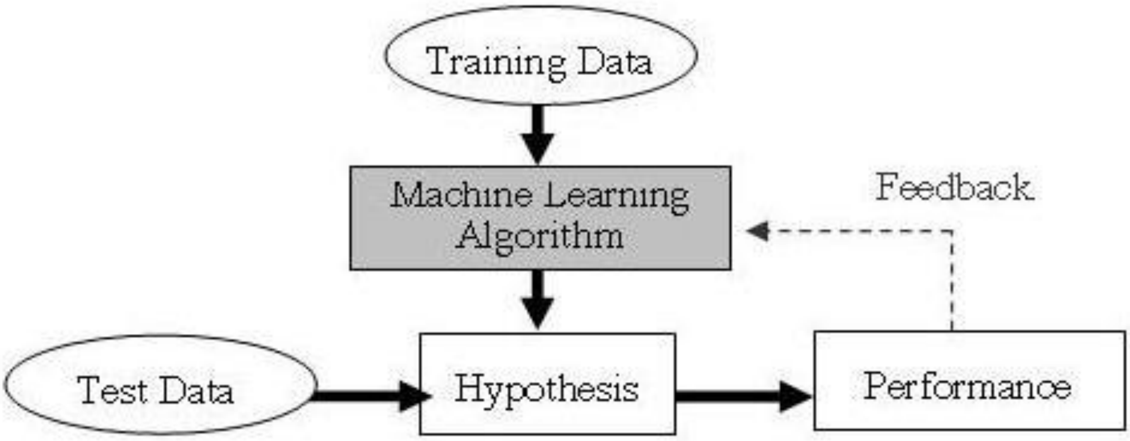
ISSN:0971-1260

Vol-22-Issue-41-December-2019

CONTENT DIAGRAM



SYSTEM ARCHITECTURE



METHODOLOGY

1. Reading the dataset

We first read the dataset from the client/user. Here the dataset is the students' data in the form of a file (CSV/EXCEL).We read this csv file by importing pandas module.

2. Dependency of various features/attributes on the final grade(G3)

Secondly, we are finding the Correlations. The dependency of all the attributes with the final grade is found. The correlation values range from -1 to 1.If the value is negative, the attribute is inversely related .If it is positive, it is directly related. If the value is near to the extreme value or the domain value, then it is highly correlated.

3. Removing the least Correlated attributes

The next step is to discard the least correlated attributes. This is done for more accuracy. For better computational results the least correlated attributes are removed from the file.

4. Converting the datatypes

The next step is to convert the data types of the dataset. Since the computations cannot be performed on multiple datatypes simultaneously, we are converting the data types of the values into binary i.e. 0s and 1s.This is done by using a function called get_dummies().

5. Splitting the data for training and testing

The most crucial part of the project is explained here. Here we import sklearn module for training and testing the data as well as performing our prediction algorithm. We perform the training and testing the data by send X and Y labels. The X and Y labels are the parameters that on which the training and testing of the data is computed.

Here, the X parameter is the dataset and the Y parameter is the required output i.e. the final grade(G3). Now that the data is sent for training and testing using train_test_split() function. It also includes test size or train size, which means the fraction of data that must be sent for testing or training respectively, Random state is an optional parameter.

6. Prediction

After the data is trained and tested, now it is fitted using the Linear Regression algorithm. The fit() function is used to set the accuracy of the computed values. The fit() function gives the desired results/values. The predict() is used for prediction which too uses the Linear Regression model.

7. Graphical Representation of the Result

The final step is the graphical representation of the predicted results. The graph that is being used is the "Box-plot". Here, we are comparing the given final grade with the predicted results which also the final grade(G3). The reason for using boxplot is that is shows the representation accurately in terms of statistics i.e. mean, median, quartiles etc. Through this we can say if the predicted values are accurate or not.

TOOLS USED

Python

Python's popularity may be due to the increased development of deep learning frameworks available. As Python has readable syntax and the ability to be used as a scripting language, Python proves to be powerful both for preprocessing data and working with data directly.

Pandas

Pandas is an open source library that allows to you perform data manipulation in Python. Pandas provide an easy way to create, manipulate and wrangle the data. *Pandas* is the most popular python library that is used for data analysis. It provides highly optimized performance.

Sklearn

Scikit-learn (also known as **sklearn**) is a free software machine learning library in Python. Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. The functionality that scikit-learn provides include:

- **Regression**, including Linear and Logistic Regression
- **Classification**, including K-Nearest Neighbours
- **Clustering**, including K-Means and K-Means++
- **Model selection**
- **Preprocessing**, including Min-Max Normalization.

LINEAR REGRESSION

Linear Regression is an approach to model the relationship between a dependent variable and one or more independent variable .In the case of a single independent or explanatory variable, it is called Simple Linear Regression or Linear Regression with single variable .For more than one dependent or explanatory variable, it is known as Multiple Linear Regression or Linear Regression with multi variable.

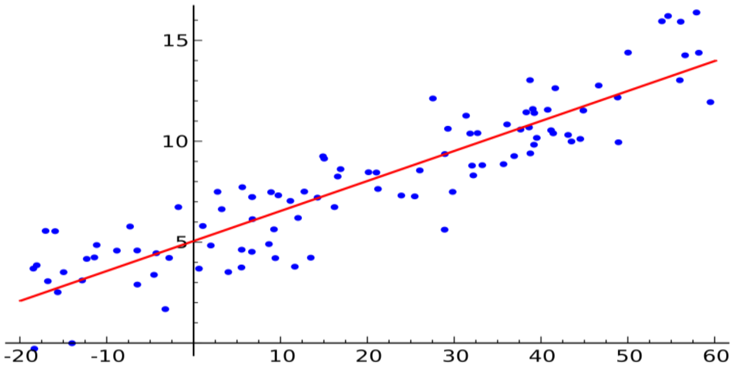
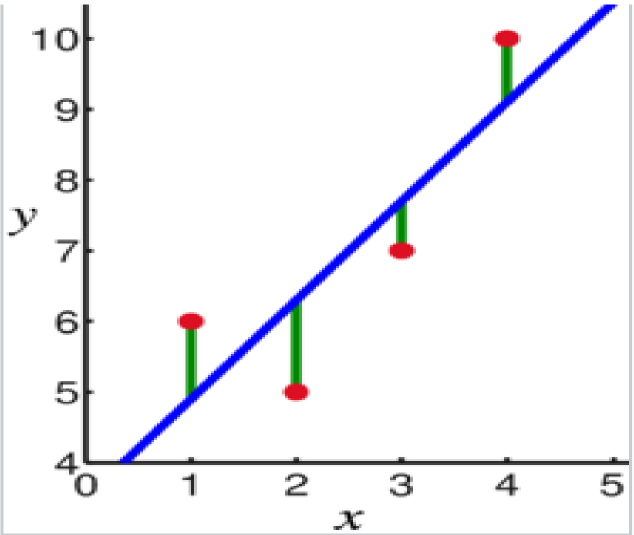


Fig: Example of Linear Regression that has one independent variable.

Linear Regression was the first type of regression models for the study of analysis. It is also widely used in practical applications. This is because the models which are linearly depend on unknown parameters are easier to fit than those models which depend non-linearly on their parameters.

Most applications of Linear Regression fall into one of the following two categories:

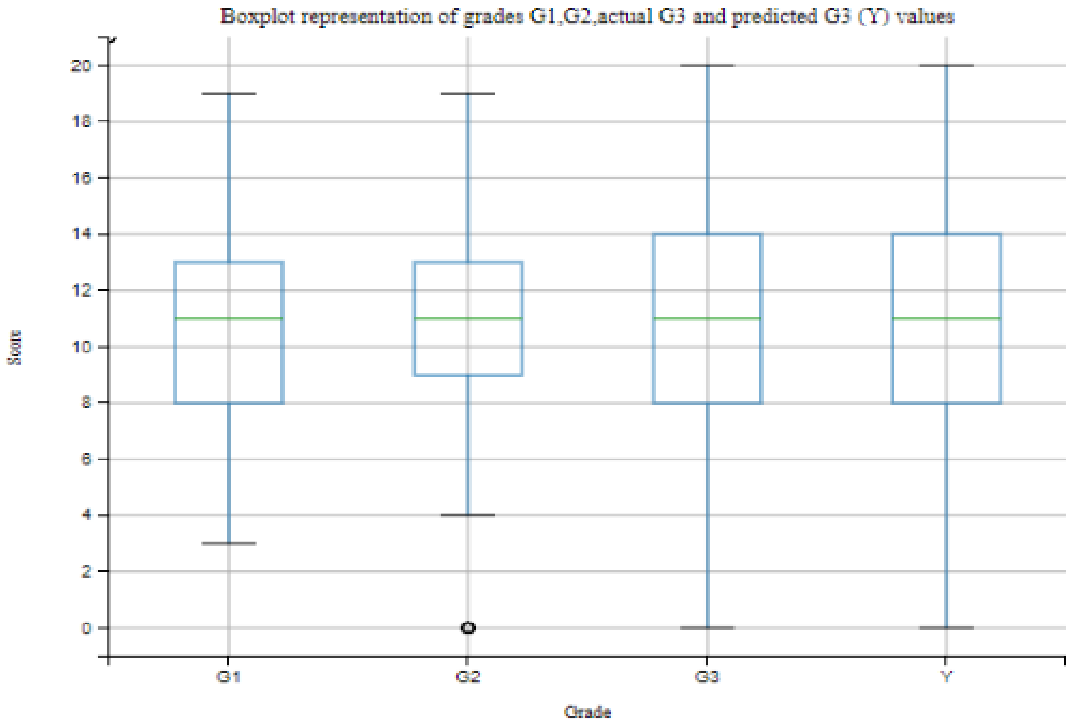
- if the aim is prediction ,or forecasting, or error reduction, linear regression is used to fit the predictive model to an observed set of data values of the independent variable. After developing the model, if additional values of the independent variables are gathered without any accompanying response values, the fitted model is used to make prediction of the response.
- if the aim is to explain the variation of a dependent variable attributed to an explanatory variable, linear regression can be applied to measure the strength of the relationship between the response and the explanatory variables.



In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable (*y*) and an independent variable (*x*).

Like most forms of regression, Linear Regression focuses on conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

RESULTS



CONCLUSION

In this project we did a deep analysis of what could be possible factor on whether a student is likely to get a high score or a low score. The data does not contain that much information but still we were able to predict a pretty precise Linear Regression algorithm that predicts what score a student will get in the foreseen feature by analysing the features.

It is to my understanding that the linear regression model is used to predict values with a given number of features. Here we learned how to tuned Hyper parameters in a more automatic way in order for our model to have better predictions when new features of students will come in and will let us know as data comes by if a student is most likely to pass the class or not.

FUTURE SCOPE

Data cleaning and analysis can be better done and other machine learning algorithms can be applied on the model to improve the accuracy. Increased dataset will give out more accurate predictions. To improve the results, a dataset with sufficient features and increase in quantity must be obtained. Further research must be conducted in enhancing the existing machine learning techniques to work in real time and develop an efficient model. Also, the models developed must be tested on data with different volumes to test its scalability and performance.

In future work, the result of regression on balanced dataset can be studied by changing the data distribution. This can be done by selecting a sample of dataset or removing certain records to balance the type of data.

REFERENCE

[1] Sumati Pathak, Rohit Raja, Vaibhav Sharma, Srinivas Ambala, ICT Utilization and Improving Student Performance in Higher Education, International Journal of Recent Technology and Engineering (IJRTE) at Volume-8 Issue-2, pp. 5120-5124, July 2019.

[2] Laxmikant Tiwari, Rohit Raja, Vaibhav Sharma, Rohit Miri, Adaptive Neuro Fuzzy Inference System Based Fusion of Medical Image, International Journal of Research In Electronics And Computer

Engineering, Vol 7, Iss. 2, pp. 2086-2091, ISSN: 2393-9028 (PRINT) |ISSN: 2348-2281 (ONLINE).

[3] Sumati Pathak, Rohit Raja, Vaibhav Sharma, and K. Ramya Laxmi, A Framework of ICT Implementation on Higher Educational Institution with Data Mining Approach, European Journal of Engineering Research and Science, ISSN (Online): 2506-8016,

[4] Sumati Pathak, Rohit Raja, Vaibhav Sharma The Impact of ICT in Higher Education. Published in IJRECE Vol. 7 Issue 1 January-March, 2019. ISSN: 2393-9028 (PRINT) ISSN: 2348-2281 (ONLINE) ISSN: 2393-9028 (PRINT). Vol 7, Issue 1, pp 1650-1656.

[5] Rakesh Kumar Lenka, Amiya Kumar Rath, Zhiyuan Tan, Suraj Sharma, Deepak Puthal, N V R Simha, Rohit Raja, Shankar Sharan Tripathi, and Mukesh Prasad Building Scalable Cyber-Physical-Social Networking Infrastructure Using IoT and Low Power Sensors, , Vol. 6, Iss. 1, pp. 30162-30173, Print ISSN: 2169-3536, Online ISSN: 2169-3536, Digital Object Identifier: 10.1109/ACCESS.2018.2842760. (SCI Index)

[6] Rohit Raja, Tilendra Shishir Sinha, Raj Kumar Patra and Shrikant Tiwari(2018), Physiological Trait Based Biometrical Authentication of Human-Face Using LGXP and ANN Techniques, Int. J. of Information and Computer Security, Vol. 10, Nos. 2/3, pp. 303- 320. (Scopus Index)

[7] Rohit Raja, Tilendra Shishir Sinha, Ravi Prakash Dubey (2016), Soft Computing and LGXP Techniques for Ear Authentication using Progressive Switching Pattern, Published in International Journal of Engineering and Future Technology, Vol. 2, Iss. 2, pp.66-86, ISSN: 2455-6432.

[8] Rohit Raja, Tilendra Shishir Sinha, Ravi Prakash Dubey (2016), Orientation Calculation of human Face Using Symbolic techniques and ANFIS, Published in International Journal of Engineering and Future Technology, Vol. 7, Iss.7, pp. 37-50, ISSN: 2455-6432.

[9] Rehmat Khan, Rohit Raja (2016) Introducing L1- Sparse Representation Classification for facial expression, Published in Imperial Journal of Interdisciplinary Research (IJIR), Vol. 2, Iss. 4, pp. 115-122, ISSN: 2454-1362.

[10] Nikita Rawat, Rohit Raja (2016), A Survey on Vehicle Tracking with Various Techniques", International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 5 Iss. 2, pp. 374-377, ISSN: 2278-1323.

[11] Nikita Rawat, Rohit Raja (2016), Moving Vehicle Detection and Tracking using Modified Mean Shift Method and Kalman Filter and Research, International Journal of New Technology and Research (IJNTR), Vol. 2, Iss. 5, pp. 96-100, ISSN: 2454-4116.

[12] Rohit Raja, Tilendra Shishir Sinha, Ravi Prakash Dubey (2015), Recognition of human-face from side-view using progressive switching pattern and soft-computing technique, Association for the Advancement of Modelling and Simulation Techniques in Enterprises, Advance B, Vol. 58, N 1, pp. 14-34, ISSN:-1240-4543. (Scopus Index)

[13] Rohit Raja, Tilendra Shishir Sinha, Ravi Prakash Dubey (2015), Biometrical Authentication of Twins from Side-View using Hybrid Approach, (BJSTH) Bharat Journal of Science Technology and Humanities, , ISSN: 2454-6151.

[14] Rohit Raja, Tilendra Shishir Sinha, Ravi Prakash Dubey (2015), An Empirical Analysis for Detection of Occlusion for face image parallel to the surface plain using Soft-Computing technique, Mats Journal of Engineering & Vol. I (1), pp. 1-6Technology, Vol. 1, Iss. 2, pp. 95-102, ISSN 2394-0549.

[15] Rehmat Khan, Rohit Raja (2015) Neural Network Allied With Recognition of Facial Expressions of Basic Emotions, International Journal of Emerging Trends in Science and Technology, Vol. 2, Iss. 11, pp. 3311-3315, ISSN: 2348-9480. A. C. Bhensle and Rohit Raja (2014), An efficient face recognition using PCA and Euclidean Distance classification, International Journal of Computer Science and Mobile Computing, Vol. 3 Issue.6, pp. 407-413. ISSN: 2320–088X.

[16] A. C. Bhensle and Rohit Raja (2014), A survey on side-view based face recognition, International Journal for Scientific Research and Developement (IJSRD), Vol. 2, Iss. 4, pp.574- 577, ISSN: 2321–0613.

[17] Keshika Jangde, and Rohit Raja (2013), Study of An Image Compression Based on Adaptive Direction Lifting Wavelet Transform Technique", International Journal of Advanced and Innovative Research (IJAIR), Vol. 2, Iss. 8, pp. ISSN: 2278–7844.

[18] Keshika Jangde and Rohit Raja (2014), Image Compression Based on Discrete Wavelet and Lifting Wavelet Transform Technique", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 3, pp. 394-399, ISSN: 2278–7798.

[19] Sandeep Kumar & Hemlata Dalmia, "A Study on Internet of Things Applications and Related Issues", International Journal of Applied and Advanced Scientific Research, Vol. 2, No. 2, pp. 273-277, 2017 with ISSN: 2456-3080.

[20] Tilendra Shishir Sinha, Raj kumar Patra, and Rohit Raja (2011) A Comprehensive analysis of human gait for abnormal foot recognition using Neuro-Genetic approach, International Journal of Tomography and Statistics (IJTS), Vol. 16, No. W11, pp. 56-73, ISSN: 2319-3339, http://ceser.res.in/ceserp/index.php/ijts. (Scopus Index)