

A SURVEY ON STUDENT PERFORMANCE PREDICTION

¹Mudit Panwar, ²Pratiti Hazarika, ³R. Vidhya³Assistant Professor^{1,2,3}Department of Computer Science and Engineering

SRM Institute of Technology

¹muditpanwar113@gmail.com, ²hazarikapratiti364@gmail.com³vidhya.r@ktr.srmuniv.ac.in

Abstract: In today's world, the academic performance of students is necessary for the development of students. Prediction of students' grades helps in finding out about the various academic opportunities that are in store for them. It helps in giving better future prospects for them and opening a plethora of opportunities for them. It will also help in choosing the better career option for the students. And thus, it helps a lot if the performance of the students can be predicted before time so that the educational institutions have a clear idea of which set of students might need assistance and which set of students will perform well. In this model, it will mostly be focused on predicting the performance of the students in school based on their subject marks. Taking all these parameters into account, a model will be designed to predict the performance of each student and if he needs further assistance or not. And in order to achieve this, we will be making use of multiple regression for the prediction of student performance.

Keywords: Student Performance, Prediction, Multiple Regression, Descent Gradient

1. Introduction

With the advancements in educational systems all over the world, the learning environment is becoming more and more competitive every day. Thus, prediction of students' grades have become a very crucial part of education. Various educational institutions all over the world are adopting different types of student prediction techniques. Prediction of student grades helps the student as well as the teachers in various educational institutions in finding out about the performance of the students before time. This helps in having a clear idea about the group of students who are strong in certain subjects and some who are weak in certain areas and providing them with assistance accordingly and this prediction of student performance can also help them to prepare for the higher

studies that is this method can provide them basic idea about their anemic points and capable points, so that they can plan out the preparations accordingly.

In most of the educational institutions all over the world, the Grade Point Average (GPA) system is used for the final assessment of the student. The GPA and the performance of the particular student can be predicted taking a certain number of attributes. A varied number of attributes can be used to predict a student's performance and interest in a particular subject. The basic attributes that can be used are the students' basic information like gender, age, study time, failures and grades. There are some further characteristic that can be used relating to the students' family like mother's education, father's education, mother's job, father's job and familial support. The other additional attributes that can be used are the courses that the particular student has selected and the school that the particular student is from. These characteristics aid in predicting the end grades and achievement of the student.

In light of the review, it was discovered that a fundamental issue was looked by everybody. The procedure of expectation isn't a one-time process however a nonstop procedure implies understudy can take diverse course in his school life and can get distinctive outcome each time that is we should refresh the information in like manner. . Due to new information being included each day, the test information must be checked every day to get precise outcomes. In this way, checking the information regular made the procedure of expectation extremely dull.

Multiple Regression can be used to know about the connection between the assorted reliant and self reliant variables. We will solve the above mentioned problem using the regression in which one variable i.e GPA will depend on various factors such as demographics, scores of students and we will find the cost function which will be reduced using the Gradient Descent method. After which the value will be predicted.

2. Literature Survey

There have been many methods which have been used for the prediction of student performance. The various methods that have been used can be classified into many categories.

A. Clustering

Clustering is the method of combining real or unreal collection into similar collection. It is the process by which the data or objects are converted into meaningful sub category known as clusters. Clusters are a batch of akin objects where the gap among any two body in the bundle is less than the gap between an body in the bundle and an body which is not located inside the bundle. Clustering appear under unsupervised learning as it has no pre categorized division.

Since clustering is used to group the similar data together that is why it is use to categorise the students in different sections. In this the data of course taken which was finished in 130 days. From this the starting nine days were removed as they were of no use. The data is fed to the clustering algorithm through which we came to the conclusion that there are total of five clusters which are formed after the process which the students are divided into these five categorise namely achievers , regular, half-hearted , underachievers and average. This cluster was formed for the II year students. The same methodology was used for the I year students which provide the same five clusters.

After getting the clusters for two different year students, both the cluster are analysed using different techniques that is the differences are checked between the two clusters using:-

(1) Kruskal-Wallis Test:- This test is very important as it tell the user whether the sample is originated from the same group of cluster or different group. It is non-parametric method of finding the difference between the clusters. In this method the data is rated from 1 to N ignoring combined association

$$T = (N - 1) \frac{\sum_{i=1}^k n_i (\bar{r}_i - \bar{r})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2}$$

Where, n is illustration in the student dataset r is the rank N is the no. of illustration among all the clusters r is the average rank.

The significance which is observed after the cluster analysis in the study was 0.8% using which the grouping

is done using condition $p < 0.008$. This criteria helped to do the comparative study between the different student group.

B. Correlation

Correlation is often used to catch the relation between variables. In more unequivocal terms, relationship is utilized to discover the continous communiacion between two factors. It shows how skilled the connection between two factors is.

From the study, it is found that Pearson correlation analyses is used to find the predictor variables of all the courses combined. And form the study it was found that the predictor variables of all the courses combined had a statistically significant correlation with the final exam grades. The effect size for the in-between assessment grades were large and for the number of wiki views were moderate. The effect size of all the other variables were significantly less. The correlations between the predictor variables and the end exam grades showed different results. The midterm grades were the only ones which correlated significantly with the end exam grade. The other factors which represented significant correlation for at least a significant percentage of all the courses are the number of online working, the number of basic viewed, the number of subjects started and the number of basics passed.

The correlation test that was used was:

$$\text{corr}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2 \right]^{1/2}}$$

Where

x_i is the first variable of the first set of observations

y_i is the first variable of the second set of observations

n is the no. of illustration

$\bar{x}$ is the average of the first set of illustration

$\bar{y}$ is the average of the second set of illustration

It is seen that the predictor variables did not correlate significantly with the final exam grades and the effect sizes were also varied and the direction of correlation sizes were also very different. This shows that the effect of the variables as predictors will differ across various courses in a multivariate environment.

The main drawback of correlation is its continous accord bound. Even if the result is almost null or close to null, then the variables may be non-linearly correlated.

C. Regression

Regression is another data mining approach which has been used in contemporary years for prediction desire. A regression task usually begins with a dataset in which the target values are known. While building the regression model, an algorithm for regression is used for finding the value of the objective as functions of the predictors in each case and then using the values of this model for another dataset where the values are unknown. The regression models are designed with the help of various statistics which help the user to find the variation between the anticipated values and the expected values. The conclusive objective of regression analysis is to resolve the values of the guidelines of an activity that cause the activity to best fit the set of data observations that we cater.

In the study, we found that the predictor variables were having mixed results across a number of courses, using correlation. And in order to determine the effects that the predictors had across different courses, an ordinary least squares regression was run on all courses.

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

where

$$\bar{x} = \frac{\sum x}{n} = \text{mean of the } x \text{ data}$$

$$\bar{y} = \frac{\sum y}{n} = \text{mean of the } y \text{ data}$$

a is the intercept

b is the slope

n is the number of observations

—
xy is the product of the scores

x is the sum of x scores

y is the sum of y scores

x^2 is the sum of squares of x scores

In the study, the regression is used to get the prediction about the pass and the fail probabilities of the student. In the study, the GPA of the students has been predicted with students having GPA more than 5.5 declared as pass and students having GPA less than 5.5 declared as fail. This gave them an accuracy of 60%. The study also further checked whether LMS data can be used to predict the final grades and student performance. After adding the LMS data, it was found that the amount of variance level dropped from 48% to 38% and the course level raised from 8% to 18%. This tells the importance and capability of LMS data for the grades. In order to check the differences between the prediction models which are there per course, separate multiple linear regression models were run per course and the final exam grade was taken as the outcome and all basic and pattern variables as

predictors as they are available in all courses. The results which persisted shows that the LMS data explains some of the variance in the final exam grade in each course, but the amount of variance which is there differs strongly from 8% for course 9 to 37% in course 7. It was also observed that the irregularity in the study time per session is the least present with 6 out of 17 courses whereas the total time online had the most often regression coefficient with 12 out of 17. It was also observed that the coefficient of time spent online was different for different courses. In some courses, the time spent online had a negative impact on the final exam grade and in some other courses, it had a positive impact on the final exam grades. Henceforth, many such factors influence the final grades of the student and accordingly have a negative or positive impact on them. But ultimately, the number of online sessions are necessary for the effect they have on the final grades of the student. Therefore, the huge variety in the various prediction models reveals low portability of the prediction models. These models give insight into the factors affecting the student performance in each course. With the average mean residual of 1.56, where almost 70% grades are between 4.0 to 9.0, the models are very far away from an accurate prediction.

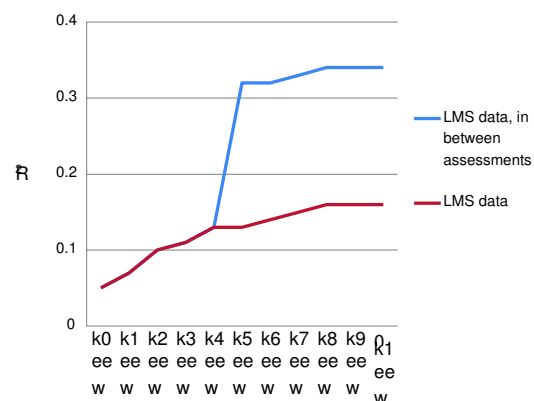


Figure 1. R^2 of multiple linear regressions on final exam grade over weeks, using LMS data with and without in-between assessment grade.

Thus the formula for multiple regression analysis is as follows:

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Where

x_1 is one independent variable

x_2 is another independent variable

y is the intercept

Multiple regression allows us to calculate which factors will be more likely dependant linearly on the dependant variable and as such identifying such dependency. In the study, a series of assumption tests were run before the multiple regression analysis to test whether the data met the necessary requirements. While operating a standard residuals test it was established that no outliers were found and tests for multi-colinearity showed that the values of tolerance and VIF were within the adequate restrictions. The value of the Durbin-Watson which is 2.20, also showed the assumption of independence of errors were met. Two models were prepared in the study to predict the student performance.

In the native model it was initaite that test nervousness and negative self-controlled tutoring anticipated the understudy scholarly execution with a little impact where the two factors chargeable for only 7% of the variance.

In the second model, an additional 6% of the variance in academic performance is explained. Here, there were three variables, namely test anxiety, resource and MCQ which contributed greatly to student performance. It was observed that interaction with dashboard and exercises were not so significant.

D. Other Methods

Neural Networks method is use to predict the change in grade of the students according to their behaviour. In the study we found that each student video watching click stream data is taken plus test grades are taken which are then divided into testing and training data. In the study two layers of neural network are used which are define as

$$z_i(n) = f_i(y_i(n-1), y_i(n-2), \dots, y_i(n-d), x_i(n-1), x_i(n-2), \dots, x_i(n-d)),$$

Where,

Z is predicted average grade

D is feedback delay

N is number of quizzes

3. Conclusion

In the examination done we have watched that different information mining procedures have been utilized for the expectation of understudy execution. While some of these strategies gave more prominent precision than whatever is left of them, just few of them were extremely exact. There are numerous elements which influence the last grade expectation and also understudy execution. Some of these elements influence the last outcome while some of them don't influence the outcome to such an extent. Later on, more factors can be utilized which can be influence the last outcome and give more noteworthy

exactness with the goal that the expectation is more precise.

4. References

- [1] Javier García-Martín, Jorge Enrique Pérez-Martínez, and Almudena Sierra-Alonso "Teamwork, Motivational Profiles, and Academic Performance in Computer Science Engineering" IEEE REVISTA IBEROAMERICANA DE TECNOLOGIAS DEL APRENDIZAJE, VOL. 10, NO. 2, MAY 2015
- [2] Guillermo Rodríguez, Álvaro Soria, and Marcelo Campo "Measuring the Impact of Agile Coaching on Students' Performance"
- [3] Jie Xu, Member, IEEE, Kyeong Ho Moon, Student Member, IEEE, and Mihaela van der Schaar, Fellow, IEEE "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs" 2017
- [4] S. Ikbāl A. Tamhane B. Sengupta M. Chetlur S. Ghosh J. Appleton "On early prediction of risks in academic performance for students" VOL. 59 NO. 6 PAPER 5 NOVEMBER/DECEMBER 2015
- [5] AMIN ZOLLANVARI , (Member, IEEE), REFIK CAGLAR KIZILIRMAK , (Member, IEEE), YAU HEE KHO , (Senior Member, IEEE), AND DANIEL HERNÁNDEZ-TORRANO "Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors" VOLUME 5, 2017
- [6] Gabriel Barata, Sandra Gama, Joaquim Jorge, Senior Member, IEEE, and Daniel Gonçalves "Early Prediction of Student Profiles based on Performance and Gaming Preferences" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES
- [7] Rianne Conijn, Chris Snijders, Ad Kleingeld, Uwe Matzat "Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES
- [8] Jose-Luis Poza-Lujan, Member, IEEE, Carlos T. Calafate, Juan-Luis Posadas-Yagüe, and Juan-Carlos Cano "Assessing the Impact of Continuous Evaluation Strategies: Tradeoff Between Student Performance and Instructor Effort" IEEE TRANSACTIONS ON EDUCATION

[9] Abelardo Pardo, Feifei Han, and Robert A. Ellis
“Combining University Student Self-regulated Learning
Indicators and Engagement with Online Learning Events
to Predict Academic Performance” IEEE
TRANSACTIONS ON LEARNING TECHNOLOGIES

[10] Tsung-Yen Yang, Christopher G. Brinton,
Member, IEEE, Carlee Joe-Wong, Member, IEEE and
Mung Chiang, Fellow, IEEE “Behavior-Based Grade
Prediction for MOOCs via Time Series Neural
Networks” [http://www.ieee.org/publications_standards/pub
lications/rights/index.html](http://www.ieee.org/publications_standards/publications/rights/index.html)

