International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 03, March-2015

Student Performance Prediction using Machine Learning

Havan Agrawal, Harshil Mavani
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, India

***Abstract* - In this paper, a model is proposed to predict the performance of students in an academic organization. The algorithm employed is a machine learning technique called Neural Networks. Further, the importance of several different attributes, or "features" is considered, in order to determine which of these are correlated with student performance. Finally, the results of an experiment follow, showcasing the power of machine learning in such an application.**

Keywords— Artificial intelligence, machine learning, student performance, neural networks

I. INTRODUCTION

There are many studies in the learning field that investigated the ways of applying machine learning techniques for various educational purposes. One of the focuses of these studies is to identify high-risk students, as well as to identify features which affect the performance of students.

The study conducted by Kotsiantis et al [1] is one of the initial studies which investigated application of machine learning techniques in distance learning for dropout prediction. The most significant contribution by this study was that it was a pioneer and carved the path for several such studies. While machine learning algorithms had been previously implemented in several settings, this was perhaps the first time that these techniques were applied to an academic environment.

Bhardwaj and Pal [2] conducted a study in India, Faizabad to determine factors that most heavily affected student performance. They used Bayesian Classification for their study.

The study by Erkan Er [3] was based upon Kotsiantis' as well as other similar studies. It concluded that Naive Bayes indeed performed better than any other machine learning algorithm. However, the crucial contribution of this study was that time-invariant features may be detrimental to the machine learning process, and hence are better left out of the study entirely. He also concluded that "Instead of demographic characteristics of students, using initial attendance and homework grades produces better prediction rate at earlier stages."

II. BACKGROUND AND RELATED WORK

A. Algorithm

Classification is one of the most frequently studied problems by data mining and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). There are different classification methods.

Bayesian classification is an algorithm that is based on Bayes rule of conditional probability. Bayes rule is a technique to estimate the likelihood of a property given the set of data as evidence or input. Bayes rule or Bayes theorem is-

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

A more recent development in classification is that of artificial neural networks. These networks are modeled after the human neural system (hence the name), and have proven to be as powerful, if not more, as any other algorithm. While implementations may be complex, these networks are capable of understanding non-linear patterns in data.

A detailed description of the algorithms can be found in [4].

Kotsiantis et al [1] compared five algorithms, viz. Decision Trees (C4.5), Naive Bayes algorithm (Bayesian networks), 3-NN (kNN), RIPPER (Rule Learning) and WINNOW (Perceptron based neural networks). This study was composed of two experimental stages, training and testing. During these stages, number of attributes was increased step-by-step. For example, while only demographic data was included in the first step, performance attributes were added in the next step. Five algorithms were tested for each these subsequent steps and then they were compared. This comparative study helped in narrowing down candidates for our own application.

However, classification of data into binary groups seems insufficient. The primary goal of this study was only detecting at-risk students instead of determining performance levels of students. Classifying students according to their performances in different levels (e.g. poor, average, good, excellent, etc.) might be more useful. In this way, instructors can provide more adaptive feedback for each student.

B. *Features*

Bhardwaj and Pal [2] conducted a study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application) course of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayesian classification method on 17 attributes, it was found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, students other habit, family annual income and student's family status were highly correlated with the student academic performance.

In the present study, those variables whose probability values were greater than 0.70 were given due considerations and the highly influencing variables with high probability values have been shown in Table 1. These features were used for prediction model construction. For both variable selection and prediction model construction, the publishers have used MATLAB.

From the table, it is found that the second high potential variable for students' performance is their living location, and the third high potential variable for students' performance is medium of teaching. In Uttar Pradesh the mother tongue language of students is Hindi. Hence, students tend to be more comfortable in Hindi and other languages, than in the English language.

C. *Uniqueness*

The study conducted by Erkan Er [3] proved valuable in confirming the uniqueness of the proposed application. His work concluded that all current applications of machine learning in an academic setting were to predict dropout rates in a distance learning program. There is perhaps no application that attempts to predict the absolute performance of the student. If one does exist, it has not been published yet.

D. *Inference*

We analyzed the experiments and results of the aforementioned studies, and two prominent inferences were drawn. The first is that Naive Bayes Classification proves to be an excellent algorithm for the application of predicting student performance in an academic setting. Further, a worthy contender for the same is neural networks. Secondly, several factors contribute to a student's performance, apart from previous academic performance.

Table 1: Study Results

Variable	Description	Probability
GSS	Student's Grade in Secondary Education	0.8642
LLoc	Living Location	0.7862
Med	Medium of Teaching	0.7225

III. EXPERIMENTAL WORK

Initially, the existence of a linear relationship between a student's previous academic performance levels was considered. This relationship can be expressed accurately using Multivariate Linear Regression. Multivariate Linear Regression uses past semester marks of a student and marks scored by this student's senior batches to predict future marks of this student. Octave was used for test purposes. The marks of 80 B.E. I.T (Bachelor of Engineering, Information Technology) students from semester 3 to semester 6 were used. The algorithm is trained on a training set of 60 students, and tested on a cross-validation set of 10 students, to predict marks in 6 subjects. This is done 7 times, varying the training and test sets each time (k-fold cross validation). An error of plus or minus 8 marks was considered as accurate. The error statistics were as follows:

Average error = 6
Accurate = 296
Erroneous = 124
Accuracy Rate = 70.48%

Once it was confirmed that the data conforms well to a machine learning algorithm, we conducted a comparative study of neural networks and Bayesian classification, on the basis of varying training and test sets. The results were fairly surprising. In general, the neural networks tend to outperform Bayesian classification. This is somewhat justified once one realizes that the input provided to the algorithm was on a continuous range, and Bayesian classification traditionally requires discrete data.

Finally, an application was made that employed neural networks (Figure 2). The application provides to and fro access of data from .csv (Comma Separated Values) files. When a prediction is required, it dynamically trains a network of 3 layers, and provided prediction of marks in discrete classes of 20 marks.

The training dataset size was increased in increments of 10, starting from 40, for 17 subjects. The test set was of 10 students, to predict a single subject. The accuracy results are summarized in Table 2.

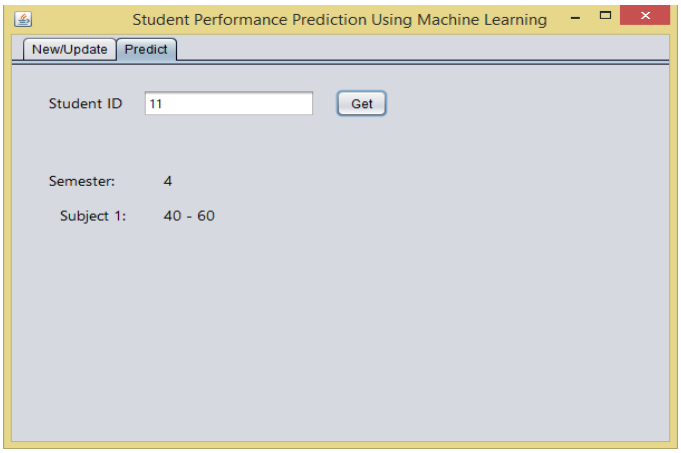

Figure 1: Screenshot of Application

Table 2: Neural Network Accuracy

Training Dataset Size	Accuracy
40	50 %
50	50 %
60	60 %
70	70 %

IV. CONCLUSION

Present studies shows that academic performances of the students are primarily dependent on their past performances. Our investigation confirms that past performances have indeed got a significant influence over students' performance. Further, we confirmed that the performance of neural networks increases with increase in dataset size.

Machine learning has come far from its nascent stages, and can prove to be a powerful tool in academia. In the future, applications similar to the one developed, as well as any improvements thereof may become an integrated part of every academic institution.

ACKNOWLEDGMENTS

We thank Prof. Yogita Borse. Without her guidance, this paper could never have been accomplished.

REFERENCES

[1] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, “Preventing student dropout in distance learning systems using machine learning techniques,” AI Techniques in Web-Based Educational Systems at Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, pp. 3-5, September 2003.

[2] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[3] Erkan Er. "Identifying At-Risk Students Using Machine Learning Techniques", International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp. August 2012.

[4] S. Kotsiantis, I.D. Zaharakis, and P. Pintelas, "Assessing Supervised Machine Learning Techniques for Predicting Student Learning Preferences"