International Journal of Engineering & Technology, *7 (3.4) (2018)* 197-199

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Evaluating the Effectiveness of Machine Learning Algorithms in Predictive Modelling

Deepali Vora[1]*, Kamatchi Iyer[2]

[1]*Research Scholar, Amity University, Mumbai*
[2]*Department of Computer Science and Engineering, Amity University, Mumbai*
**Corresponding author E-mail: deepali_as@yahoo.com*

Abstract

Predictive modelling is a statistical technique to predict future behaviour. Machine learning is one of the most popular methods for predicting the future behaviour. From the plethora of algorithms available it is always interesting to find out which algorithm or technique is most suitable for data under consideration. Educational Data Mining is the area of research where predictive modelling is most useful. Predicting the grades of the undergraduate students accurately can help students as well as educators in many ways. Early prediction can help motivating students in better ways to select their future endeavour. This paper presents the results of various machine learning algorithms applied to the data collected from undergraduate studies. It evaluates the effectiveness of various machine learning algorithms when applied to data collected from undergraduate studies. Two major challenges are addressed as: choosing the right features and choosing the right algorithm for prediction.

Keywords:*Machine Learning, Predictive Analytics, SVM*

1. Introduction

India being the developing country; education plays a vital role in development. Specifically undergraduate studies require a special attention if India has to grow as a developed country. Students' retention is an important challenge in graduate studies. Now a day with higher availability of seats for undergraduate studies; the universities and colleges are facing problem of retention of students. As well it's a challenge to ensure that students graduate in timely fashion. So there is a critical need to develop innovative approaches that ensure students graduate in a timely fashion and are well trained and workforce ready in their field of study. Students' training can be planned based on predictive analytics which can help colleges to make student ready for higher studies or placements. [1]

Educational Data Mining (EDM) is a new field of research in the data mining and Knowledge Discovery in Databases (KDD) field. It mainly focuses in mining useful patterns and discovering useful knowledge from the educational information systems from schools, to colleges anduniversities. EDM is very useful in formation of predictive analytics about the student.

This paper talks about two major challenges in above process: (a) Choosing the features for predictive analytics and (b) Choosing the algorithms for performing the analytics. The paper presents comparative study on various machine learning algorithms and tries to identify the prominent factors in prediction of performance. Section 2 talks about approach followed in our experimentation. In Section 3 we present the results from experiments and discussion of the same. Section 4 presents the conclusion and future work.

2. Approach

a. Purpose of Study

In any education system predicting students' progress is very important. Early tracking will help boosting the students' weak area to improve the overall performance. Various EDM methods can be used effectively to do the same. The objective of this experimentation is to find out which EDM algorithms are best suitable for predictive analytics of students' performance. The higher education data is collected and various machine learning algorithms are applied. The performance prediction is a classification task, so is there need of new classification model or current algorithms are sufficient for the prediction is another goal. By comparing the performance of algorithms on various performance parameters we will be able to decide on the same.

b. Dataset and Performance Factors considered

The dataset under consideration is collected by surveying undergraduate students from various engineering colleges. The data is collected when students are studying in third year and final year of undergraduate studies. The dataset contains the attributes which can be classified as cognitive or non-cognitive factors. Cognitive factors refer to characteristics of the person that affect performance and learning. Non cognitive factors are the mental constructs which indirectly contribute to the success of the students' performance. Non cognitive factors are also considered as they play an important role in determining the performance of the student [2]. Figure 1 depicts the summary of cognitive and non-cognitive factors under consideration.

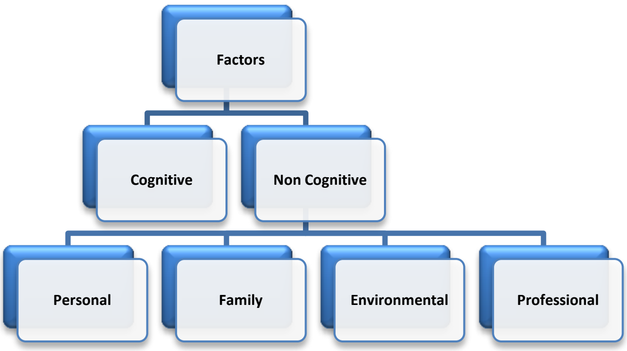


Figure 1: Cognitive and Non Cognitive Factors

Cognitive factors considered are 10th Exam and 12th Exam Marks, Marks secured in semesters and score of entrance exam. 4 to 5 factors from each non cognitive domain are also considered; for example Age and Gender from personal, Parents Education in Family etc. [3]

c. Pre-processing

The data collected was pre-processed to remove the noisy data. Records with major missing values are removed. Depending on number of subjects in which students' have failed in different semesters; KTScore is assigned from 0 to 3. Based on the education taken in 10th and 12th Class the student is evaluated as Home University student or Other than Home University student (OHU). Other attributes are assigned default values if found missing data.

d. Algorithms considered

The general goal of this study is to compare the effectiveness of existing EDM techniques for early identification of students likely to fail with increased accuracy and precision. There are plethora of machine learning algorithms available out of which we have chosen few by looking at the results required and data available.

Students' data is collected to analyze the effectiveness of algorithms from various machine learning domains as: Regression - Logistic Regression (LR), Dimensionality Reduction -Linear Discrimination Analysis(LDA), Instance Based Algorithms - K-Nearest Neighbourhood(KNN), Bayesian Algorithms - Gaussian Naïve Bays(NB), Support Vector Machine (SVM) [4] and Neural Network(NN) [5].

Following diagram depicts the experimental setup to evaluate the algorithms under consideration.

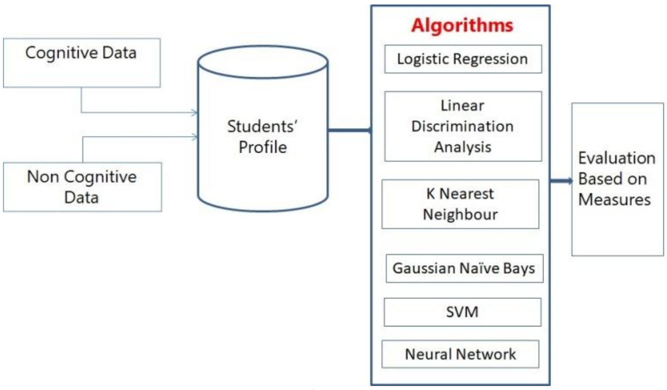


Figure 2: Setup for Experimentation

e. Evaluation Measures

To evaluate the effectiveness of the Machine Learning algorithms applied in this experiment, we decided to adopt the Accuracy, Precision, Recall and F-Measure [6] which is widely used in domains such as information retrieval, machine learning and other domains that involve binary classification [7]. Confusion matrix is the base for the determination of Precision and Recall as follows:

Precision = TP/(FP+TP)
Recall = TP/(FN+TP)
Where -
True Positive (TP) = Number of positive instances correctly classified as positive.
False Positive (FP) = Number of negative instances correctly classified as positive.
True Positive (TP) = Number of positive instances incorrectly classified as negative.

Precision is the measure of exactness and Recall is the measure of completeness. F-Measure is the harmonic mean between Precision and Recall as described below:
F-Measure= 2 * (Precision * Recall) / (Precision +Recall)

3. Results and Discussion

a. Scores of Algorithms Dataset

The algorithms discussed in Section 2 are evaluated on the dataset collected. The dataset was divided into training dataset and testing dataset. Following table depicts the results of various measures on the dataset.

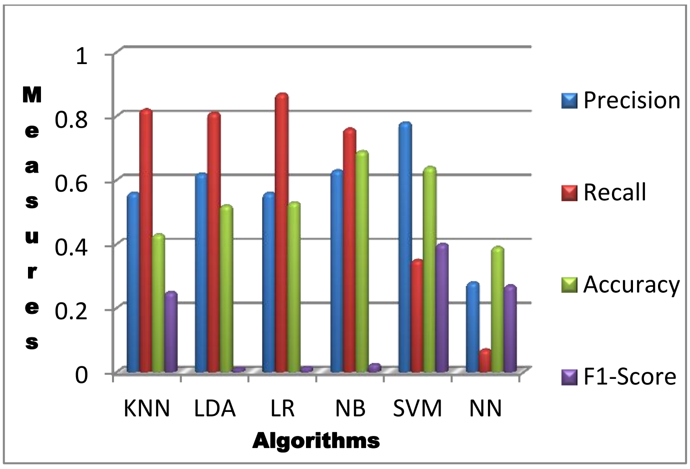
Table1: Evaluation Measures of different algorithms on dataset

Algorithm → Measures↓	LR	LDA	KNN	NB	SVM	NN
Precision	0.56	0.62	0.56	0.63	0.78	0.28
Recall	0.87	0.81	0.82	0.76	0.35	0.07
Accuracy(%)	53	52	43	69	64	39
F1 Measure	0.014	0.011	0.25	0.024	0.4	0.27

The accuracy of NB and SVM are between 60% to 69%. But the F1-measure of SVM is better than NB. This indicates SVM may be used to get accurate predications. Also the precision of SVM is better than other algorithms.

b. Discussion

Following Graph shows the scores of all algorithms on dataset



. **Figure 3:** Effectiveness of Algorithms

The F1 Measure clearly indicates that SVM is better choice than other algorithms under consideration. But still the scores of all the measures can be improved by using better model.

The features were evaluated using ANOVA test and top features were selected. The features like 12th Marks, Semester Exam Marks, KT Score, and Entrance Exam Score were identified as key features. The performance of algorithms was found sensitive to these features.

4. Conclusion and Future Work

We have presented the results of evaluation of various algorithms for predictive analytics in educational data. The algorithms were evaluated for their effectiveness in predicting the students' performance. It is evident that SVM is prominent in prediction on the collected dataset but not the best. There is a need of new clasiification model which can give improved results for the data under consideration.

There are two challenges for which these algorithms may be evaluated as (a) Scalability of algorithms when data size grows (b) Accurate prediction of performance when number of features increase. Educational domain can be considered as Big Data Domain because of drastic increase in digitization of data. Basic machine learning algorithms may not be so effective when the data increases.

We were able to evaluate the effectiveness of most popular machine learning algorithms on the collected data of Students. Total 6 algorithms were evaluated on dataset having 35 features. It will be interesting to evaluate more advanced machine learning algorithms for predictive analytics. Specifically in Educational Domain their effectiveness will help in early failure prediction of students.

References

[1] AsmaaElbadrawy, AgoritsaPolyzou, Zhiyun Ren, Mackenzie Sweeney, George Karypis, HuzefaRangwala, "Predicting Student Performance Using Personalized Analytics," IEEE , April 2016

[2] "The important role of Non Cognitive Factors in School Performance," http://singteach.nie.edu.sg/issue25-hottopic/", Accessed on 20/08/2017

[3] DeepaliVora, Dr. KamatchiIyer, "EDM - Survey of Performance Factors And Algorithms Applied", International Journal of Engineering and Technology, Vol:7 ,No.2.6 (2018), Pages: 93-97

[4] Vapnik, V. N.," The nature of statistical learning theory", New York, NY, USA: Springer-Verlag New York, Inc.,1995

[5] Nürnberger, A., Pedrycz, W., & Kruse R., "Handbook of data mining and knowledge discovery", Oxford University Press, Inc., Pages 304-317

[6] Han, J., Kamber, M., & Pei, J.," Data Mining: Concepts and techniques", 3rd ed., 2011

[7] Olson, D. L., &Delen, D.," Advanced data mining techniques", 1st ed., Springer Publishing Company, Incorporated, 2008