Student's performace evaluation usinng Machine Learning Algorithms

Ritu Vinodbhai Kumbhani (c0865677)
Information Technology
Lambton College at Cestar college of
Business, Health and Technology

Toronto, Canada, Ontario

C0865677@mylambton.ca

Abstract

Prediction of the student performance is very important for a success of any educational institute. Make a use of the methods of machine learning algorithms to predict the student performance based on data available in school can explain their behavior and impact on their study. The prediction system has been proposed by using their grades. The study is evaluated using the machine learning algorithms Logistic regression, Gaussian Naïve bayes, Support Vector Classifier, Random Forest and Gradient Boost.

Keywords—Student Performance, ***Logistic regression, Gaussian Naïve bayes, Support Vector Classifier, Random Forest and Gradient Boost.***

Introduction

This report is focused on educational system. It is essential for every educational institute to provide the accurate grade to the students.

Due to the huge amount of data in educational databases, predicting the performance of students has become more difficult. The shortage of an established framework for evaluating and tracking the success of students also isn't currently being considered. There are two primary reasons why such kind of occurring. First, the research on existing methods of prediction is still insufficient to determine the most appropriate methods for predicting student performance in institutions. Second, is the absence of inquiry of the specific courses. [4-9]

The real goal is to have an overview of the systems of artificial intelligence that were used to predict academic learning. This research also focuses on how to classify the most relevant attributes in student data by using prediction algorithm. Using educational machine learning methods, we could potentially improve the performance and progress of students more efficiently in an efficient manner. Students, educator and academic institutions could benefit and also have an impact.

In this report we have used UCI Student performance dataset. We have used many classification machine learning algorithms to predict the student performance**.** Then we compared the performance of 5 algorithms ***Logistic regression, Gaussian Naïve bayes, Support Vector Classifier, Random Forest and Gradient Boost,*** to reach the best technique for prediction.

MACHINE LEARNING ALGORITHMS

A. LOGISTIC REGRESSION (LR)

Logistic regression is used for classification problems, it is predictive analysis algorithm based on the concept of probability.

B. Gaussian Naïve bayes (GNB)

Gaussian Naïve bayes is supervised machine learning classification algorithm and it is the variant of the Naïve bayes algorithm. It follows Gaussian normal distribution and supports continuous data

C. SUPPORT VACTOR CLASSIFIER (SVC)

Support vector machine uses the svc to create the hyperplane after placing the data into high-dimensional space to separate the data into different classes. It works well with the classification tasks.

D. RANDOM FOREST (RF)

In order to increase the dataset's predictive accuracy, a classifier called Random Forest uses many decision trees on different subsets of the input data. It is based on the idea of ensemble learning, which is the practise of integrating various classifiers to solve a challenging problem and enhance the model's performance

E. GRADIENT BOOST (GB)

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. Gradient boosting is one of the variants of ensemble methods where you create multiple weak models and combine them to get better performance as a whole.

II. DATASET INFORMATION

The data we used are student achievement data in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por)

Table 1. Dataset features discription

Features	Values	Description
School	GP, MS	Student's school name
Sex	F, M	Student's gender
Age	Numeric from 15 to 22	Student's age
Address	U – urban R- rural	Student's address
Famsize	LE3, GT3	Family size
Pstatus	T, A	Parent's cohabitation status
Medu	0,1,2,3,4	Mother's education
Fedu	0,1,2,3,4	Father's education
Mjob	Teacher, health, services, at_home, other	Mother's job
Fjob	Teacher, health, services, at_home, other	Father's job
Reason	Home, reputation, course, other	Reason to choose this school
Guardian	Mother, father, other	Student's guardian
Travel time	1,2,3,4	Home to school travel time
Study time	1,2,3,4	Weekly study time
Failures	1,2,3,4	Number of past class failures
Schoolsup	Yes, No	External educational support
Famsup	Yes, No	Family educational support
Paid	Yes, No	Extra paid classes within the course
Activities	Yes, No	Extra-curricular activities
Nursery	Yes, No	Attended nursery school
Higher	Yes, No	Wants to take higher education
Internet	Yes, No	Internet access to home
Romantic	Yes, No	Relationship
Famrel	1 to 5	Quality of family relationship
Free time	1 to 5	Free time after school
Go out	1 to 5	Going out with friends
Dalc	1 to 5	Workday alcohol consumption
Walc	1 to 5	Weekday alcohol consumption
Health	1 to 5	Current health status
Absences	0 to 93	Number of school absences.

A. METHODOLOGY

We have worked on various methodology to enhance the accuracy of Machine Learning classification methods of student performance. The performance of the classifiers has been tested on all attributes.

A classification technique is used to predict student performance as a commonly used tool in forecasting**.** The classifiers used in this report are based on algorithms often used in the literature.

B. DATA TRANSFORMATION

Data transformation is a critical step to remove inconsistencies in a dataset, making it more suitable for data mining. Converting String to Numeric Variables: Most data mining algorithms work only with numeric variables. Therefore, non-numeric data must be converted to numeric variables, the most common methods being to encode a string using a value in the range [0 to (N-1)], where N is the number of values. For example, gender attribute (F/M) is coded as 0 and 1.

C. DATA PARTITONING

The purpose of the dataset is to divide the dataset into two parts, training data and test data. The test data is the largest part of the dataset and is used to test the class. The test data is used to evaluate the output of the classifier (Han et al., 2012), in our experiments we use 80% for training and 20% for testing.

D. PERFORMANCE EVALUATION

In My experiments, four standard evaluation metrics are used to evaluate the performance of the classification algorithms, namely: accuracy.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (1)$$

E. TEST CLASSIFIERS

To identify the best classifier for predicting student performance in the dataset, I examine total 4 machine learning algorithms namely: ***Logistic regression, Support Vector Classifier, Random Forest and Gradient Boost***

Table 2 Classifier performance comparison

Classifier	**Accuracy**
Logistic Regression	90%
Support Vector classifier	88%
Random Forest	89%
Gradient Boost	90%

F. PARAMETER TUNING

Parameter tuning is important because default values cannot be suitable for all tasks and do not produce the best results(Smit & Eiben, 2009), so I applied parameter tuning in the classifier with the best performance from the previous experiment(i.e. Random Forest). Random Forest Hyperparameters

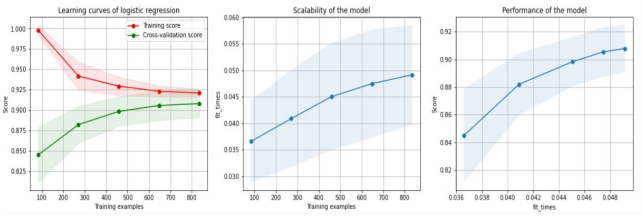
- N – estimators
- Criterion
- Max depth

I used a grid search technique that tests a set of hyperparameters to determine the best values for a given task based on validation accuracy. This method is computationally more complex than simply using the default values of the model parameters.

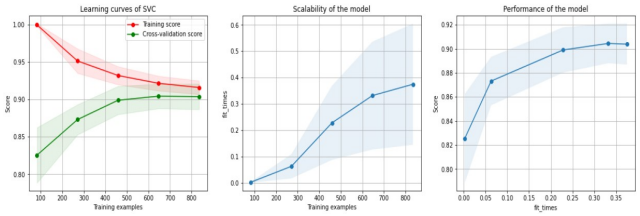
G. GRAPHS

I have plotted the graphs of all 4 machine learning algorithms for visual presentation.

Logistic Regression –

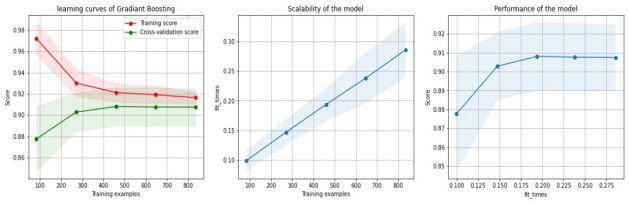


Support Vector Classifier –



Random Forest –

Gradient Boost –



CONCLUSION

Predicting student achievement can help an institution take timely action, such as designing appropriate training to increase student success rates. Analyzing training data can help achieve desired training goals. Predictive models can be created to improve student performance. In this report, we collect a dataset representing a sample of , students to investigate the possibility of predicting student performance. We applied machine learning techniques to datasets and tested 4 classification algorithms, and Logistic Regression gave the best results out of with an accuracy ratio of 90% . To get the best parameter values, I used the parameter tuning technique. So by using this models I have conclude that Logistic Regression is the best classifier to predict the student performance.

REFERENCES

[1] Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. In International Journal of Educational Technology in Higher Education (Vol. 17, Issue 1). https://doi.org/10.1186/s41239-020-0177-7 J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Amrieh, Elaf Abu and Hamtini, Thair and Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), 119–136.

[3] Athani, S. S., Kodli, S. A., Banavasi, M. N., & Hiremath, P. G. S. (2017). Student performance predictor using multiclass support vector classification algorithm. 2017 International Conference on Signal Processing and Communication (ICSPC), 341–346.

[4] Bajpai, P., Chaturvedi, R., & Singh, A. (2019). Conjecture of Scholars Academic Performance using Machine Learning Techniques. 2019 International Conference on Cutting-Edge Technologies in Engineering (ICon-CuTE), 141–146 Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on

[5] Breiman, L. (2001). Random Forests. Mach. Learn., 45, 5–32.

[6] Ghorbani, R., & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students Performance Using Machine Learning Techniques. IEEE Access, 8, 67899–67911.

[7] Hussain, M., Zhu, W., Zhang, W., Ni, J., Khan, Z. U., & Hussain, S. (2018). Identifying beneficial sessions in an e-learning system using machine learning techniques. 2018 IEEE Conference on Big Data and Analytics (ICBDA), 123–128.

[8] Kime, K., Hickey, T., & Torrey, R. (2019). Refining Skill Classification with Interactive Machine Learning. 2019 IEEE Frontiers in Education Conference (FIE), 1–8.

[9] Kiranmayee, A. H., Panchariya, P. C., Prasad, P. B., & Sharma, A. L. (2012). Biomimetic classification of juices. 2012 Sixth International Conference on Sensing Technology (ICST), 551–556.

[10] Ko, C.-Y., & Leu, F.-Y. (2020). Examining Successful Attributes for Undergraduate Students by Applying Machine Learning Techniques. IEEE Transactions on Education.

I