International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

Predicting Student Dropouts in Schools Using Machine Learning Techniques

Ms. S. Pavithra, Student, Department of Information Technology, Dr.N.G.P. Arts and Science College

Mrs. N. Vanitha, Assistant Professor, Department of Information Technology, Dr.N.G.P. Arts and Science College

Dr. V. Vinodhini, Professor, Department of Information Technology, Dr.N.G.P. Arts and Science College

Dr. K. Santhi, Associate Professor, Department of Information Technology, Dr.N.G.P. Arts and Science College

Abstract

Dropping out of school has a serious consequence for students, their families. It's a difficulty because the coed who aren't inquisitive about studies or don't understand the relevant course, they'll show no interest in studies, and that they are going to be mostly likely to commit crimes, can start stoning up, there life are going to be spoiled, and will be useless for the country. Students have negative attribute towards school, risks factors might be for the family, individual, school or the community, they will earn less than educated people, there socioeconomic status could suffer, people might not respect them and cannot provide facilities to themselves and families. One personality could suffer plenty, as when he/she will become enthusiastic about the family or government. In past years the amount of student dropouts from the schools are raising rapidly. During a registered course, the coed dropout rate has been a threat for several schools.Reducing student dropout rate is one in all the challenges facing within theeducation sector globally. This problem has major concern within the field of education. Addressing this challenge requires an intensive understanding of the underlying issues and effective planning for intervention. Thus, this research work suggests a model which can predict whether the coed will continue his/her study or dropout his/her study using classification technique supported on machine learning algorithm.

1.Introduction

School dropout is absenteeism from school for no good reason for a never-ending number of days. Most of the students aren't ready to continue their education due to many

reasons. Dropout have to be cared as their performance could be better within the future, the dropout rates cost, our society plenty of cash, to save lots of them from bad activities is vital, if they are not cared they will commit various style of crimes, over 80% of the criminals are the school dropouts. For a few students, dropping out is that the culmination of years of educational hurdles, missteps, and wrong turns. For others, the choice to dropout could be a response to conflicting life pressures the necessity to assist support their family financially or the strain of caring for siblings or their own child. A student may dropout for various reasons like professional, health,academic, personal reasons and family and varies looking on the education system adopted by the school, in addition because the selected subject of studies. Machine learning approaches are one in all the well sought solutions to addressingdropout challenge.

2.Background Study

M. Solis, T. Moreira, R. Gonzalez, T. Fernandez and M. Hernandez in 2018, have used Random Forest, Neural Networks, Support Vector Machines and Logistic Regression techniques in perspectives to predict dropout in university students with machine learning and stated Random Forest algorithm is the best for predicting dropouts **[1].**

Sheikh Arif Ahmed, Shahidul Islam Khan in 2019, have used Support Vector Machine, random forest, neural network techniques to Predict the Engineering Students at risk of dropout and factors behind using a machine learning approach and stated Neural Networkis the best algorithm **[2].**

FisnikDalipi , Ali Shariq Imran , ZenunKastrati in 2018, have used Support Vector Machine (SVM) technique to predict MOOC dropout using machine learning **[3]**.

3.Existing Work

The existing method is incredibly time consuming and not very accurate and focuses on only specific factors. There's no early warning system to understand the potentialout student beforehand. While the facts are available in the market with us all the time, the administration were taking actions only after the dropout had happened.

4.Proposed Work

A system is proposed in a way that it can meaningfully understand the data from which it has to be trained upon and tried to develop behaviour from the data-sets. Data-sets are the backbone of this model

and hence it should be adequate and precise data for the student will dropout or continue their study for the model to be trained upon. The proposed method is also combined approach which takes several factors like course of study, school and also nationality which increases the accuracy and implements methods that reduces the time taken for prediction. To train a model an ensemble method random forest classifier is used.

5.Machine Learning

Machine Learning are often defined as how a machine imitate a typical creature and learns from user actions sort of a small kid. It's going to even be said that a machine is explicitly programmed inan exceedingly way that it understands from the user actions and execute in its applications.

Supervised machine learning algorithms

Supervised learning is where you've got input variables (x) and an output variable (Y) and you employ an algorithm to be toldthe mapping function from the input to the output.

$Y = f(X)$

The goal is to approximate the mapping function so well that after you have new computer file (x) that you simply can predict the output variables (Y) for that data.

6.Machine learning in Education

There have been significant advances within the field of machine learning, over the past twenty years. This field arouse because the method of choice for developing practical software for computer vision, robot control, speech recognition, tongue processing and other applications. There are some areas where machine learning can genuinely crash on education. On labeling student dropout issue, several predictive models were developed in developing countries to control complex data sets that include details about student performance, enrollment, gender and, school infrastructure, socio-economic demographics.

7.Significance of machine learning

Machine learning as technology helps analyze large chunks of information, easing the tasks of information scientists in an automatic process and is gaining a many of prominence and recognition. Machine learning has replaced the way of extracting data and interpreting works by involving automatic sets of generic methods that have replaced traditional statistical methods. The machine learning field is continuously evolving. There's one crucial reason why data scientists need machine learning, and which is: 'High-value prediction that will guide them to

make better decisions and take smart actions in real-time without human intervention'. Machine learning is useful because it automatically processes and saves time, so humans can focus their time and energy on more complex higher noesis.

8.Significance of Student Dropouts

The high dropout rates are costly for both students, educational institutions and society generally. We aren't that much aware of what causes the students to dropout from school. The foremost reliable knowledge available is based on register data, which tells us only little about the factors that are likely to possess the most effective significance for dropout rates.

9.Significance of Machine learning methods for student dropouts

Over the past twenty years, there has been a noteworthy approach in the field of machine learning, making an appearance as the approach of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications.Over the years machine learning has acquired the most consideration on labeling the issue of student dropout. This is because machine learning techniques can strongly ease resolution of at-risk students and timely planning for interventions.

10.Machine Learning Approach

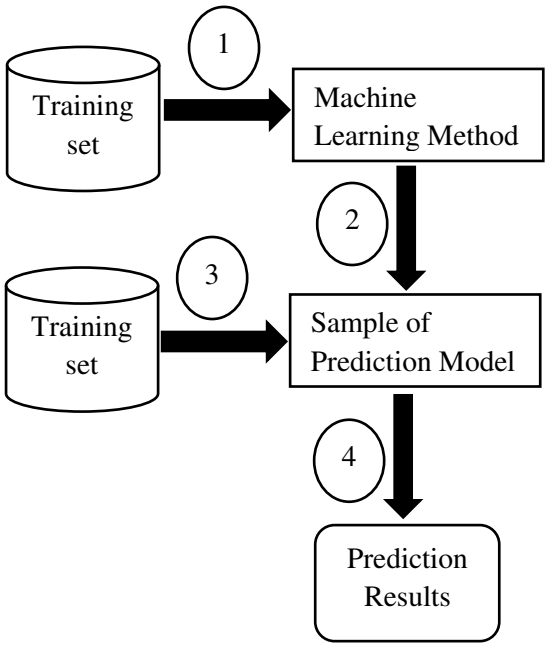
Random Forest Classifiers

Random Forest Classifier is another supervised machine learning algorithm which will be used for both classification and regression problems. It's generally classified into a special form of machine learning called an Ensemble learning during which a combination of weak classifiers is employed to create a robust classifier, which may be accustomed perform better predictive analysis. In case of the random forests, decision trees are the weak learners. Although decision trees perform rather well on some datasets, they're generally called weak learners as they have a tendency to own high variance once they are trained on different subsections of the identical data. Here, variance refers to the spread of the predictions and occurs when a model is sensitive to small changes within the training data, which ends up in overfitting. This happens because of the greedy approach of decision tree algorithm in selecting the simplest split to be told rules from the training data. Random forests on the opposite hand create a variety of decision trees during training by using different subsections of the training data.

Moreover, the method of finding the basis node and splitting the feature nodes occurs randomly in random forests. Once we've got a forest of trees, decisions from different trees are combined to form a final decision regarding the data. This way the random forests will generalize the predictions better since the combination of selections won't be sensitive to the trained data as each tree learns from different subsections of data. The more the number of trees within the random forest the better generalized the predictions.

11. Methodology

Figure 1. Overall Research Framework



12.Data Set

The dataset used for this research was taken from the UCI Machine Learning Repository. The data attributes include gender, nationality, course of study, etc.

13.Feature Extraction

A data in the dataset are imported as data frames, for that machine learning numpy package is used. The data can be understood by the pipeline. A pipeline consists of a chain of processing elements arranged so that the output of each element isthe input of the next; the name is by analogy to a physical pipeline.

14.Training Model

To train the model Stratified ShuffleSplit - cross validator that is imported using sklearn model from python sci-kit library. (From sklearn.model_selection import StratifiedShuffleSplit). It uses train data set for learning. After learning it prints score of trained model.

15.Testing Model

Pass the students information as inputs in to trained model. It provides whether the student will continue or dropout their studies and returns the output as Yes/No.

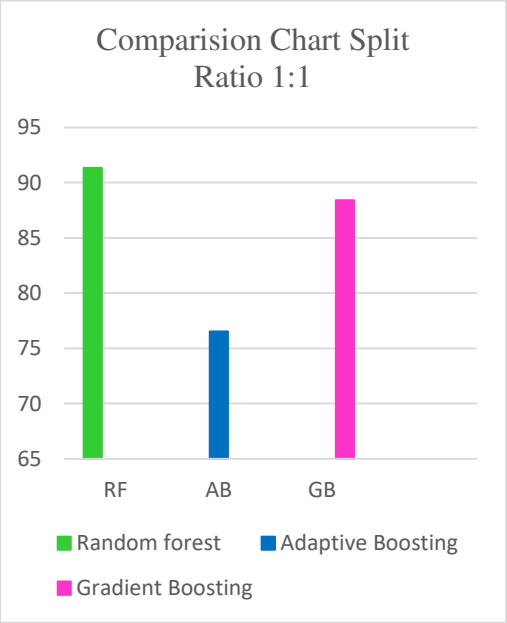
16.Experiment and result discussion

The model is to be sketched to predict the student will continue or dropout by using machine learning methods. It generally consists of two parts such as machine learning model and datasets. The data in the data sets are imported as data frames,

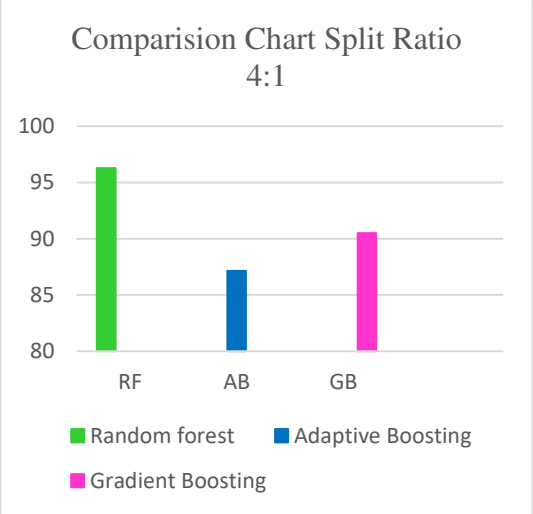
for that the machine learning numpy package is used. Then the data can be understood by the pipeline. To train and test the data ShuffleSplit from sklearn python model is imported.

The model is trained for multiple times with different data split ratio such as 1:1, 4:1, 10:1 and compared with previous models and the learning accuracy score iscompared. Comparison table is shown in comparison table and charts.

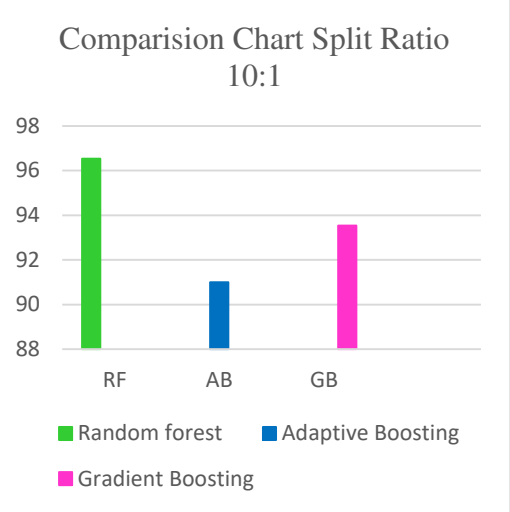
17.Comparision Charts



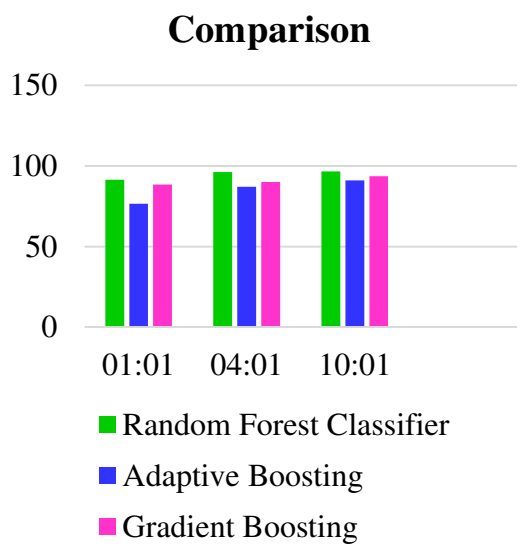
Random Forest Classifier gives the highest accuracy of 91.35 whereas, Adaptive Boosting gives 76.5 and Gradient Boosting gives 88.4 while comparing in the ratio 1:1



Random Forest Classifier gives the highest accuracy of 96.26 whereas, Adaptive Boosting gives 87.15 and Gradient Boosting gives 90.5 while comparing in the ratio 4:1



Random Forest Classifier gives the highest accuracy of 96.54 whereas, Adaptive Boosting gives 91 and Gradient Boosting gives 93.54 while comparing in the ratio 10:1



The comparison have been made by different attributes such as gender, nationality, elementary grades, assignment grades, presence an also absence of the student.

While comparing the three techniques Random Forest Classifier, Adaptive Boosting, and Gradient Boosting with different Split ratios Random Forest Classifier algorithm predicts the student dropout rate with highest accuracy.

18.Conclusion and future Enhancements

The power of machine learning can step in building better data to help authorities draw out crucial insights that changes outcomes. When student dropout of school instead of continuing with education, both students and communities lose out on skills, talent and innovation.

After analysing the results it was determined that the algorithm for classifying dropouts is the Random Forest. The predictive capacity of this algorithm was the best of the alternatives evaluated.

For Future enhancement, we need to include school level datasets on addressing this problem whereas, many research work focus on addressing student dropout using student level datasets. This will involve the use of new sources school level data, that will consider school needs related features and applying additional machine learning approaches to improve the predictive power of the proposed algorithm.

References

[1] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez and M. Hernandez, "Perspectives to Predict Dropout in University Students with Machine Learning," *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, San Carlos, 2018, pp. 1-6.

[2] S. A. Ahmed and S. I. Khan, "A machine learning approach to Predict the Engineering Students at risk of dropout and factors behind: Bangladesh Perspective," *2019 10th International Conference on Computing, Communication and Networking*

Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-6.

[3] F. Dalipi, A. S. Imran and Z. Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," *2018 IEEE Global Engineering Education Conference (EDUCON)*, Tenerife, 2018, pp. 1007-1014.

[4] K. Limsathitwong, K. Tiwatthanont and T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," *2018 5th International Conference on Business and Industrial Research (ICBIR)*, Bangkok, 2018, pp. 110-114.

[5] K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho and C. A. E. Montesco, "Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout," *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, Maceió, Brazil, 2019, pp. 207-208.

[6] G. Kostopoulos, S. Kotsiantis, O. Ragos and T. N. Grapsa, "Early dropout prediction in distance higher education using active learning," *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Larnaca, 2017, pp. 1-6.

[7] M. A. A. Dewan, F. Lin, D. Wen and Kinshuk, "Predicting Dropout-Prone Students in E-Learning Education System," *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, Beijing, 2015, pp. 1735-1740.

[8] A. Kashyap and A. Nayak, "Different Machine Learning Models to Predict Dropouts in MOOCs," *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Bangalore, 2018, pp. 80-85.

[9] P. M. da Silva, M. N. C. A. Lima, W. L. Soares, I. R. R. Silva, R. A. de A. Fagundes and F. F. de Souza, "Ensemble Regression Models Applied to Dropout in Higher Education," *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, Salvador, Brazil, 2019, pp. 120-125.

[10] B. Perez, C. Castellanos and D. Correal, "Applying Data Mining Techniques to Predict Student Dropout: A Case Study," *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, Medellin, 2018, pp. 1-6.