International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

Student Intervention System using Machine Learning Techniques

Shubhangi Urkude, Kshitij Gupta

Abstract: ***Now a days, the educational institutes are adopting technologies for betterment of student's quality, in respect to teaching methodologies etc. For which the huge information available with educational institutes can be used to predict student's future in academics. The main objective of this paper is to predict the student performance in the examination and also to predict the student will graduate or not. Hence forth we are using statistical analytical method which is F1 score. F1 score or F measure is used to test the prediction accuracy by considering precision and recall to compute the score. To fulfill this requirement in machine learning, classification technique is used. The dataset used in this analysis contains 395 student records, having attributes, such as age, health, internet, school, father job, mother job etc. Using support vector machines (SVM), Decision Tree and Naïve Bayes (NB) classification algorithms F1 score is calculated for each algorithm. Based on the analysis done the F1 score of support vector machine is giving the better prediction compared to rest of the two algorithms.***

Keywords: Support Vector Machines (SVM), Decision Tree, Naïve Bayes (NB), Classification Algorithm, Prediction.

I.INTRODUCTION

In educational institutes like colleges and many universities huge data is collected on daily basis. This data contains examination data, student's internal marks, their personal data, health data, various evaluation components data and logs of student's activity.

The data is caught through learning administration frameworks like Canvas and Edmodo. Canvas and Edmodo are particularly intended for online classrooms, which are getting to be well known in the highereducation.

This huge collection of data can be used to analyze the student's performance in the future to increase the graduation rate. Graduation rates are frequently the criteria of decision for this. Graduation rate is defined as the time when student enters the institution and completed the degree with in four years, it is usually expressed as % graduation rate. To finalize the graduationrate the data collected from different sources should be converted into single form so that this data can be retrieved and analyzed to make specific decision. To take any decision, data is analyzed in such a way that it should able to answer certain questions such as Who are the students likely to fail?

What is the courses student may fail?

Manuscript published on 30 September 2019.

***** Correspondence Author (s)

Shubhangi Urkud, Department of Computer Science and Engineering Faculty of Science and Technology The ICFAI Foundation for Higher Education, Hyderabad.

Kshitij Gupta, Department of Computer Science and Engineering Faculty of Science and Technology The ICFAI Foundation for Higher Education, Hyderabad

What is the quality of the student'scontribution to the marks he obtained?

Which courses should be focused more to help the student?

What are the easy subjects to get good score?

Based on the answers collected in the analysis phase, the perfect decision can be made by using different machine learning techniques and students can be guided in a particular direction without diluting educational standards.

Student's academic performance influenced by numerous components, like personal, financial and other environmental variables. Learning about these factors and their impact on student performance can help dealing with their impact. Recently, much aid has been paid to educational mining research. Educational Data Mining (EDM) concerns techniques, tools, and research planned for automatically extracting meaning from big repositories of data generated by or related to people learning activities in educational surrounding. Anticipating student's performance turns out to be all the more challenging because of the expansive volume of information in educational databases. Forecasting of academic performance is widely explored. The ability to foresee student performance is important in educational institutions. Expanding student's achievement is a long-haul objective in every single educational institute.Predicting student'sacademicperformance ahead of schedule before their last examination is really a great achievement. At this point additional measures can be taken by educational institutes to orchestrate appropriate help for the low performing students to enhance their studies and help them to progress. On the other hand, distinguishing characteristics that influence course achievement rate can aid courses enhancement. Recently created web-based educational technologies and the use of quality standard offer researchers' novel chances to contemplate how students realize and what ways to deal with learning lead to progress. The main objective of this paper is to distinguish the two factors that influence student's performance, courses achievement rate and student's achievement rate. By considering these components as early indicator analysis can be done and necessary action can be taken to improve.

II. LITERAYURE SURVEY

Many researchers have used statistics and machine learning algorithms for predicting the student's performances in educational institutes.

Edin Osmanbegovic etal. [1]uses three supervised learning algorithms, Bayesian, Decision trees and Neural Networks on the assessment data to predict number of students got success in a course and the performance of the teaching,learning methods were evaluated based on their predictive precision and ease of learning and user-friendly features.

It is proved that this method can be used to benefit students and teachers to progress student's performanceand decrease failing ratio by taking proper steps at right time to improve the quality of education.

MladenDragicevic, et al. [2] used decision tree classification to predict student'sperformance based on Grade Point Average (GPA) criteria. Two parameters are considered for analysis as time-in-degree and GPA, in that GPA is giving better result in predicting the student's performance.

CH.M.H.Sai Baba, et al. [3]used decision tree and multivariant regression analysis to predict the number of students getting job.

Behrouz MinaeiBidgoli Iet al. [4] compared four different classifiers and united the results into a multiple classifier. Their research separated the data into three diverse classes as high, middle and low. Genetic algorithm is used to improve the accuracy of prediction in all the classes. For the data set having less features, feature weighting mechanism is better as compare to feature selection. Using LON-CAPA, e-learning platform results are validated.

Furthermore, In the research workRahel Bekele et al. [5], applied Bayesian network in the education field, toforecaststudent's performance. This model is also tested on real world data where students are assigned t fill the data and analysis is done on the actual data to predict their performance. It shows that results obtained are very useful for the teacher to assist the student to improve the academic performance.

Paulo Cortez et al. [6] addressed the prediction of secondary school student's performance in two core subjects as mathematics and Portuguese by using their past score in the earlier session and other demographic factors. Using Business Intelligence (BI) and Data mining techniques such as Decision trees, Random Forests,Support Vector machines method and Neural network real world data is analyzed.The real data may contain information related to student's grade, social and school related features. The model is tested with and without previous semester grades.Model is showing good predictive accuracy

Surjeet Kumar Yadav et al. [7], concluded that Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods. Frequently used decision tree classifiers are studied and the experiments are conducted to find the best classifier for Student data to predict the student's performance in the end semester examination. The experimental results show that Classification and Regression Tree (CART). CART is the best algorithm for classification of data. It is giving the accuracy of 56.25% as compare to other algorithms like ID3 and C4.5. This study is useful for the students who are going to fail,and also used to identify the poor student which requires more attention.

From the study conducted by Zlatko J. Kovacic[8], proposed a case study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. Two algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying, successful and unsuccessful students. The result obtained showed that the accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

Abeer Badr El Din Ahmed et al. [9] used classification technique to predict the final grade of students. It was done by the use of ID3 decision tree method. To determine best attribute for particular node in the tree Information Gain measure is used over the collected sample S. Midtrem attribute is having highest gain, so it is selected as root node for decision tree and the same process is followed to classify entire attributes.

Sembiring S et al. [10] showed that Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. The study further showed how useful data mining can be in higher education particularly to predict the final performance of student. The researchers gathered data from student by using questionnaire to find the relationships between behavioral attitude of student and their academic performance. Data mining techniques were then applied. They obtained the prediction rule model using decision tree as well as applying the rules into Support Vector Machine (SVM) algorithm to predict the student's final grade. Also, the students were clustered into groups using kernel k-means clustering. The study expressed the strong correlation between mental condition of student and their final academic performance. Ogunde A. O, et al. [11] used Iterative Dichotomiser 3 (ID3) decision tree algorithm to predict the university student's grade in Nigeria. A prediction accuracy of 79.556% was obtained from the model. They recommendeddifferent decision-tree model to perform similar analysis with extended data set to get better results. DorinaKabakchieva et al. [12] proposed classification method to predict student performance. This paper compares different data mining algorithms using WEKA tool and results are likely to vary between 52%-67%. For the analysis university sample data related to student is collected containing admission score, number of failures at the first year etc. HashmiaHamsa, et al. [13] proposed academic performance model using decision tree and fuzzy genetic algorithm. In this research work different parameters like internal marks, attendance, average marks etc. are considered to identify the student's performance in degree and master's degree students. Degree student's performance is evaluated by using decision tree algorithm which gave more students at risk. Fuzzy genetic algorithm giving more pass students by considering students in between risk and safe state.

III. MOTIVATION

However, with constrained assets and budgets, the leading body of administrators needs us to locate the best model with minimal measure of calculation costs. With the end goal to build the intervention software, we first should break down the dataset on student's execution. Our goal is to choose and build up a model that will foresee the probability that a given student will pass, along diagnose whether or not an intervention is necessary. Our model is developed based on a subset of the information that we give, and it will be tested against a subset of the information that is kept hidden from the learning algorithm, with the end goal to test the model's effectiveness on information outside the training set.

IV. PROBLEM STATEMENT

The college administration a has an objective to achieve a 95% graduation rate before the decade's over by distinguishing students who require intervention before they drop out of school. To recognize whether the student will graduate or not. We will probably display the variables that foresee how likely a student is to pass their secondary school final exam. We being a sharp engineer choose to implement a student intervention framework utilizing ideas we gained from supervised machine learning. Rather than purchasing costly servers or actualizing new information models starting from the earliest stage, it can be connected to outsider organization who will give the essential programming libraries and servers to run our software.

V. MACHINE LEARNING ALGORITHM

Unsupervised learning

At the point when the dataset isn't characterized or hard for interpretation, it is called unsupervised learning. The labels for the information are not characterized. There no correct method to partition information collected except performingiterations. As per this research problem, unsupervised learning isn't advisable for prediction.

Supervised learning

Supervised learning can be said as function approximation, in whichtraining examples lead to function generation. If the learning is finished with right training set, a well-behaved function can be expected. Supervised learning develops reliably with the data. It is a sort of induction learning, and it causes one-sided supervised learning.

E.g.: The function generated with supervised learning will be $X2$, if X is simply the information esteem and the output is self-duplicated.

Classification algorithm

In machine learning and measurements, classification is the issue of distinguishing in different labels, based on a training set of information containing observations (or occurrences) whose category is known.Classification istechnique where details arranged into a given number of classes. Classifier: A algorithm that maps the input data to a particular classification. A classification model tries to make some inference from the input values given for training.

VI. DATA COLLECTION AND PREPARATION PHASE

The dataset was collected from the website Kaggle.com. This dataset consists of 395 student records; each record consists of 30 attributes with their domain values. The attributes are listed in Table1. The dataset was divided two parts, training dataset (75%) and testing dataset (25%).

Attribute Name	Description
School	General public
Sex	Male or female
Age	Between 15-18
Address	Urban or rural
Family size	Numeric value(1-5)
Parental status	Yes or no
Mother Education	Numeric value(1-5)
Father education	Numeric value(1-5)
Mother job	At_home, services, other
Father job	Teacher , services and other
Reason	Course or home
Gaurdian	Mother or father
Travel time	Numeric value(1-5)
Study time	Numeric value(1-5)
Faliures	Numeric value(1-5)
Paid	Yes or no
Activities	Yes or no
Nursery	Yes or no
Higher	Yes or no
Internet	Yes or no
Romantic	Yes or no
freetime	Numeric value(1-5)
Go out	Numeric value(1-5)
Health	Numeric value(1-5)
Absences	Numeric value(1-5)
Passed	Yes or no
Schools up	Yes or no
Families up	Yes or no
femeral	Numeric value(1-5)
Dalc	Numeric value(1-5)
Walc	Numeric value(1-5)

Table 1: List of attributes

VII. CLASSIFICATION MODELS

The classification algorithms used in this research is Support vector Machine (SVM)

SVM are supervised learning models that analyze data and distinguish patterns and used for classification analysis.Some of its advantages are SVM is very efficient in high-dimensional spaces, and in situations when we have a non-linear separation problem. With SVM we have the probability to apply new kernels that allows tractability for our decision boundaries, contributing to a better classification performance. One major disadvantage of the SVM is the deciding the kernel and also it is slowwhen compared to other models like Decision trees or Naive Bayes.

Decision Trees

The advantages of a decision trees are that nonlinear relationships between parameters do not influence our performance metrics and they give us quicker prediction as compared to other models like SVMs. Decision Trees do not function well if we have smooth boundaries. i.e they work best when we have discontinuous piece wise constant model. If we really have a linear target function decision trees are not the best.

Naive Bayes

Naive Bayes is one of the quickest learning algorithms. It is mainly used for classic problems such as spam detection, recognizing letters from handwritten texts and facial analysis. The advantages of naive Bayes is that it is faster than other discriminative models like logistic regression, so we need less training data.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The same conditional independence assumption can be a drawback when we have no occurrences of a class label and a feature value untidily, that will give us a zero frequency-based value probability that affects any posterior probability estimate. These are the Classification models that were assessed and all had promising outcomes. Analyzed outcomes can be utilized in enhancing student's academic performance. It is pointed out that the corrective action must be the joint endeavors of educators, guardians/parents, the advanced education organization and the students. Some corrective measures that are distinguished incorporate support and inspirations, nearly checking the students and extra or extramural instructional exercises.

VIII. RESULT

This segment will examine the outcomes analysis of the ongoing works in predicting student's performance. This Meta investigation depends on the highest accuracy of prediction methods and furthermore the primary vital components that may impact the student's performance. F1 score measure is taken to find student's performance. F1 score achieves its best value at 1 and worst at 0.For finding the best model among the selected models first sample data of 100 students was tested and F1 score is calculated, then sample size 200 students were taken and finally same model is tested with sample size of 300 records. For each sample size F1 score is calculated.Figure 1 shows the F1 score for decision tree. In this for a sample size of 100 records F1 score is 0.6552, for 200 records F1 score is increase to 0.75 and for sample of 300 records it is decreases to 0.6613. From this plot it is understood that performance of decision tree is increase up to certain input limit and then it suddenly decreases to initial score.

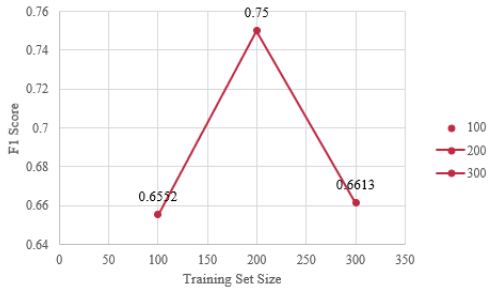

Figure 1: F1 Score for Decision Tree

Figure 2 shows the F1 score for Naïve Bays technique. In Naïve Bays a sample size of 100 records giving F1 score as 0.8029, for 200 records F1 score is decrease to 0.7244 and for sample of 300 records it is increase to 0.7634. From this plot it is understood that performance of decision tree is excellent for a smaller number of inputs.

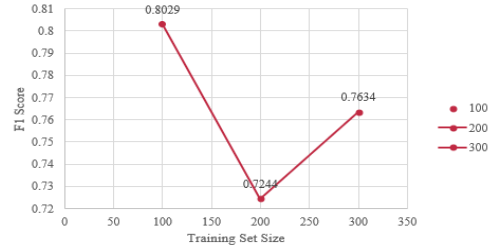

Figure 2: F1 Score of Naïve Bayes

Figure 3 shows the F1 score for Support Vector Machine technique. In this a sample size of 100 records giving F1 score as 0.7746, for 200 records F1 score is increase to 0.7815 and for sample of 300 records it is increase to 0.7638. From this plot it is understood that performance of support vector machine is increasing with data set size. So, for analyzing the student's performance this is best model as number of student's intake is increases prediction accuracy also increase.

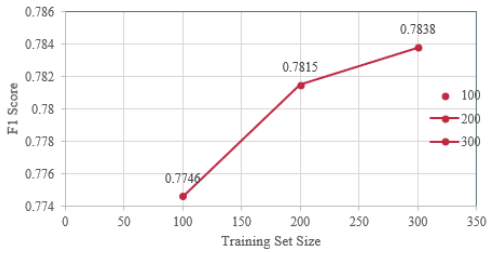

Figure 3: F1 Score for Support Vector Machine

Finally, all machine learning algorithms are compared with same input records and comparison table is prepared. Table 2 shows the F1 score for all algorithms with their accuracy for training data set and testing data set.

Algorithms	Decision Tree	Naïve Bayes	Support Vector Machine
Training set size	300	300	300
F1 score (Training set)	1.000	0.8038	0.8761
F1 score (Testing set)	0.6613	0.7634	0.7838

Table2:Final F1 score for all algorithms

Figure 4 shows the graphical representation for all algorithms with F1 score value. According to the analysis support vector machine is giving the best result.

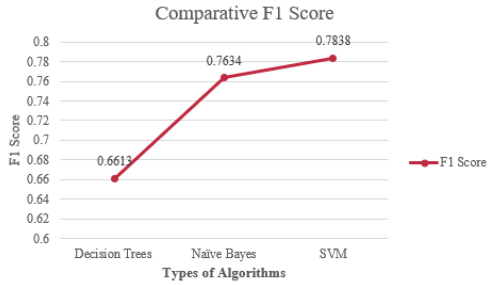

Figure 4: Final F1 Score Comparison for student intervention system

IX. CONCLUSION AND FUTURE SCOPE

This study ofresearch paper thought inthe beginning endeavor to utilize machine learning techniques to break down and discover student academic performance and to enhance the nature of the education framework. The administrations can utilize these systems to enhance the course results to enhance the student performance. Such information can be utilized to give a decent comprehension of student's enrollment pattern in the courses under study.

The faculty and managerial decision maker can utilize this research to improve performance. Other hand, such sort of learning encourages the administration to upgrade their policies, enhance their strategies and enhance the quality of the system.

In this research Support vector machines, Decision Trees, Naïve Bayes algorithm techniques are utilized to predict student's learning activities. We trust that the information created might be useful for educator and also for students. This investigation enhances student's performance; reduce failing ratio by making suitable strides at correct time to enhance the quality of education. In Future this kind of analysis can be implemented with a bigger dataset and more background factors of student at the first-year academic level. This study concentrated on deciding to what degree family background factors and past scholastic accomplishments influence student'sfirst year academic performance, recognizing the particular machine learning algorithm that best model to predict student academic performance.

REFERENCES

1. Osmanbegovic E. and Suljic M., "Data mining approach for predicting student performance",Journal of Economics and Business,Vol. X, Issue 1, 2012.
2. MladenDragicevic, Mirjana Pejic Bach, and VanjaSimicevic, "Improving University Operations with Data Mining: Predicting Student Performance", International Journal of Social, Behavioral, Educational, Economic and Management Engineering Vol. 8, Issue 4, 2014.
3. CH.M.H.Sai Baba, AkhilaGovindu,Mani Krishna Sai Raavi, and VenkataPraneethSomisetty, "Student Performance Analysis Using Classification Techniques", International Journal of Pure and Applied Mathematics, Vol. 115, No. 6, pp. 1-7, 2017.
4. Behrouz M, Karshy D, Korlemeyer G and Punch W., "Predicting student performance: an application of data Mining methods with the educational web-based system", IEEE Frontiers in Education Conference,2003.
5. BekeleR. andMenzelW. "A bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students", Journal of Information Science 2016.
6. Cortez P and Silva A., "Using data mining to predict Secondary school student performance", Journal of information science, Vol. 2, issue 6, 2013.
7. Surjeet K, Yadav, Bharadwaj B and Pal B., "Data Mining Applications: A comparative Study for Predicting Student's performance", International journal of innovative technology & creative engineering, Vol. 1, issue 12,2012.
8. Kovacic Z., "Early prediction of student success: Mining student enrollment data", Informing Science & IT Education Conference,pp. 647-665, 2010.
9. Ahmed A. B. E and Ibrahim S. E.,"Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology, Vol. 2, issue 2, pp. 43-47,2014.
10. Sembiring S, Zarlis M, Hartama D., Ramliana S andElvi W., "Prediction of student academic performance by an application of data mining techniques.", International Conference on Management and Artificial Intelligence, Vol.6, pp. 110-114,2011.
11. Ogunde A.O. andAjibade D.A., "A data Mining System for Predicting University Students Graduation Grade Using ID3 Decision Tree approach", Journal of Computer Science and Information Technology, Vol. 2, No 1, pp. 01-26,2014.
12. DorinaKabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernetics and Information Technologies, Vol. 13, No 1, pp. 61-72, 2013.
13. HashmiaHamsa, SimiIndiradevi and Jubilant J. K., "Academic Performance Model Using Decision Tree and Fuzzy Genetic Algorithm",International Conference on Recent Advancement and Effectual Researches in Engineering, Science and Technology, pp. 326-332, 2016.