A Comparative Study of Machine Learning Algorithms for Student Academic Performance

[1]Cherukuri Durga Narayana Raju, [2]S Srinivas

[1]Student, [2]Assistant Professor

[1,2]Department of Computer Science and Engineering,

[1,2]KIET, East Godavari, India.

Abstract : Machine Learning Techniques find a myriad of applications in different fields. One such application is the use of these techniques in education. The research in the educational field that involves machine learning techniques is rapidly increasing. Applying machine learning techniques in an educational background aims to discover hidden knowledge and patterns about student's performance. This work aims to develop student's academic performance prediction model, among the various students from various departments using machine learning classification methods; K-Nearest Neighbor, Decision Tree, Support Vector Machines, Random Forest, and Gradient Descent Boost Algorithms. Parameters like living area, mother father relation, education and their employment, backlogs, attendance, Internet connection availability and smart phone usage are used. Resultant prediction model can be used to identify student's performance in the final examination and anticipate the final grade. Thereby, the college management or lecturers can classify students and take an early action to improve their performance. Due to early prediction, solutions can be sought for better results in the final exams.

IndexTerms - Educational Data Mining, Machine Learning, Classification, Student Academic Performance.

I. Introduction

Education is a very important issue regarding the development of a country [1]. The main objective of educational institutions is to present quality education to its students. One way to accomplish the higher level of quality in education scheme is by predicting student's academic performance and thereby taking early actions to improve student's performance and customize teaching techniques to improve the retention of information. Currently there is an increasing interest in Data mining, Machine learning and educational systems, making educational data mining a new growing research community. In the real world, predicting the performance of the students is a challenging task. Educational Data Mining (EDM) is the utilization of the extracting knowledge procedure from educational data. The objective of an EDM is to evaluate educational data in order to improve the teaching-learning process. EDM combines the computational theory, database management, machine learning for the betterment of learners. The quality of educational system is vital for any society and it has therefore emerged as a growing research area in recent years. The data that has been collected from various educational surveys, institution records can be processed through machine learning technique such as classification for the improvement of academic performance.

Traditionally educational institutions are collecting large volumes of data related to students, faculty members, the organization and management of the educational process, and other managerial issues. However, the extent to which the available and collected data is used is not significant. In general, the data is used for producing simple queries and traditional reports that are not highly significant in contributing to the decisions making process in the institutions. Moreover, the volume and complexity of the data is often very huge that it becomes difficult to the management of the educational institutions to handle the data and hence remains unused. The potentiality of the available volume of data can be exploited only if it transformed into useful information and in turn is used to generate knowledge to support decision making.

Machine Learning is the process of discovering meaningful patterns in large quantities of data. Machine Learning is emerging as a promising framework which provides wide variety of techniques, methods and tools that enable for thorough analysis of available data in various fields. Considering the potential applications of machine learning in educational sector, Educational Data Mining started out as a new stream in the machine learning research field. EDM concerns with new methods and techniques by inquiring into eccentric types of data from educational settings to understand student's learning ability. In the domain of education, machine learning techniques are very useful for enhancing the current educational standards and academic management. These techniques provide a route to a multiple levels of ranking, a finding which gives a new perception of how people can become proficient in the educational sectors. We know that wide range of data is stored in educational databases, so in order to get required data and to find the hidden relationships different machine learning classification techniques are developed and used.

In this work, we will analyze recent real-world data from various students. Two different sources were used: mark reports and questionnaires. Since the former contained scarce information (i.e. only the grades and number of absences were available), it was complemented with the latter, which allowed the collection of several demographic, social and institution related attributes (e.g. student's age, mother's education). The aim is to predict student achievement and if possible to identify the key variables that affect educational success/failure.

The paper is organized as follows: section 2 surveys machine learning classification techniques for clustering and evaluating student performance, section 3 describes dataset, section 4 represents our Proposed System, and section 5 shows our experimental results. Finally, in section 6 the conclusions are outlined.

II. RELATED WORK

Xiaofeng Ma et. al. have used decision tree based model. The data needed for the model construction and testing are taken out from the UCI Machine learning Repository. After collecting the data they are normalized and used for constructing the model. For the construction of a decision tree model, the attribute with the highest information gain is chosen as the root node for the initial partition of the data. The information gain is calculated using the concept of entropy. The partition continues until there are no data left the leaf node contains the label of the data. After the initial model construction a test data is used to test the model. [3]

Huda Al-Shehri et. al. presented two prediction models for the estimation of student's performance in final examination. The work made use of the popular dataset provided by the University of Minho in Portugal, which is related to the performance in the subject of mathematics and it consists of 395 data samples. Forecasting the performance of students can be useful in taking early precautions, instant actions, or selecting a student that is fit for a certain task. The need to explore better models to achieve better performance cannot be overemphasized. Most of earlier work on the same dataset used K-Nearest Neighbor algorithm and achieved low results, while Support Vector Machine algorithm was rarely used, which happens to be a very popular and powerful prediction technique. To ensure better comparison, they applied both Support Vector Machine algorithm and K-Nearest Neighbor algorithm on the dataset to predict the student's grade and then compared their accuracy. Empirical studies outcome indicated that Support Vector Machine achieved slightly better results with correlation coefficient of 0.96, while the K-Nearest Neighbor achieved correlation coefficient of 0.95[4].

Pauziah Mohd Arsad, et. al. utilized the method of Artificial Neural Network (ANN) for the prediction of academic performance of students. The cumulative grade points (CGPA) is used as the measuring criterion. The data needed for the project is collected from Electrical Department of Teknologi MARA University, Malaysia. The first semester result of students is taken as the input predictor variable (Independent variable) and eighth semester grade points are taken as the output variable (Dependent variable). The study was done for two different entry points namely Matriculation and Diploma intakes. Performances of the models were measured using the coefficient of Correlation R and Mean Square Error (MSE). The outcomes from the study showed that fundamental subjects at semester one and three have strong influence in the final CGPA [5].

Kayah, et. al. used two classifiers namely, Naïve Bayes and J48, for considering the data from the UCI Machine Learning Repository. Analysis for these algorithms are performed using WEKA tool and the accuracy of the models are increased using discretization of continuous features [6].

III. DATASET DESCRIPTION

This study will consider data collected from the various students. Although there has been a trend for an increase of Information Technology investment from the Government, the majority of the Government institutions, Colleges and Universities continue to rely mostly on manual maintenance of the data. Hence, the dataset was built from two sources: institution reports, based on papers and including few attributes (i.e. the final grades and number of institution absences); and questionnaires (Google Forms) are used to complement the previous information. We designed the latter with closed questions related to several demographic (e.g. mother's education, family income), and institution related (e.g. number of backlogs) variables that were expected to influence student performance. Technology variables (e.g. Internet, Smart Phone usage) are also considered to predict their academic performance. Social Factors and Technology factors are considered to understand the extent of impact of these parameters on their education. The data-set considered, contains some categorical data for the certain number of attributes. As the pre-processing step, all the categorical information is converted into binary data with "yes" being as "1" and "no" being as "0". After the pre-processing step, the data-set consists of totally 30 attributes for each student and only numerical data is present except for the final grade.

S.NO	FEATURES	DESCRIPTION
1	Gender	Student's Gender (binary: 'F' -female or 'M' - male)
2	Age	Student's age (numeric: from 15 to 22)
3	Address	Student's home address type(binary: 'U' - urban or 'R' - rural)
4	Fam_Size	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
5	P_Status	Parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
6	M_Edu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th class, 3 - secondary education or 4 - higher education)
7	F_Edu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th class, 3 - secondary education or 4 - higher education)
8	M_Job	Mother's job (nominal: 'teacher', 'health' care related, civil 'services'(e.g. administrative or police), 'at_home' or 'other')
9	F_Job	Father's job (nominal: 'teacher', 'health' care related, civil 'services', (e.g. administrative or police), 'at_home' or 'other')
10	Reason	Reason to choose this institution(nominal: close to 'home', institution'reputation', 'course' preference or 'other')
11	Guardian	Student's guardian (nominal: 'mother', 'father' or 'other')
12	Travel time	Home to institution travel time(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
13	Study time	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
14	Backlogs	Number of past class failures(numeric: n if 1<=n<3, else 4)
15	Institution support	Extra educational support (binary: yes or no)
16	Fam sup	Family educational support (binary: yes or no)
17	Tuitions	Extra paid classes within the course(binary: yes or no)
18	Extra Activities	Extra-curricular activities (binary: yes or no)
19	Primary_E	Attended nursery institution (binary: yes or no)

	ducation	
20	Higher_Studies	Wants to take higher education(binary: yes or no)
21	Internet	Internet access at home (binary: yes or no)
22	Fam_Rel	Quality of family relationships(numeric: from 1 - very bad to 5 - excellent)
23	Free_Time	Free time after institution (numeric: from 1 - very low to 5 - very high)
24	Go_out	Going out with friends (numeric: from 1 - very low to 5 - very high)
25	Health	Current health status (numeric: from1 - very bad to 5 - very good)
26	Absences	Number of college absences(numeric: from 0 to 93)
27	Having smart	Whether student having smart phone or not (binary: yes or no)
28	phone Usage time of smart phone	Usage time of Mobile Phones in Hrs(binary: 'LE4' – less than or equal to4 or 'GT4' - greater than 4)
29	Having Pc/ Laptop	Whether student having PC / Laptop or not (binary: yes or no)
30	Student Academic Performance	Final grade (numeric: from 0-3(Fail), 4-7(Average), 8-10(Good) output target)

. **Table 1: Dataset Description**

IV. PROPOSED SYSTEM

Prediction of student behavior is made by the classifier on the test data-set after the model is built in the training phase [12] [13]. Prediction is made by, 29 attributes given as an input and prediction is made for any one remaining attribute. The accuracy for the two different types of prediction is calculated separately. Steps involved are

1) Loading of the data-set

2) Calculating mean and standard deviation: The mean is the central tendency of the data, and it is used as the middle of our normal distribution when calculating the probabilities. The standard deviation is calculated for each attribute for a class value. The standard deviation describes the variation of spread of data which is used to characterize the expected spread of each attribute in our normal distribution when calculating probabilities.

3) Separation of data: In the proposed system, data-set is separated by class values.

4) Summarization of the data: The machine learning classification model consists the information about the summary of the data in the training dataset. This summary is used to make the predictions when test data-set is given as an input. The summary obtained from the classifier is the mean and standard deviation values for each attribute. These values will be used to calculate the probability of an attribute belonging to a particular class.

5) Making predictions: In this phase, prediction is made based on the summaries obtained from the training data. Predictions are made by calculating the normal probability density function which uses mean and standard deviation for the attribute from the training data.

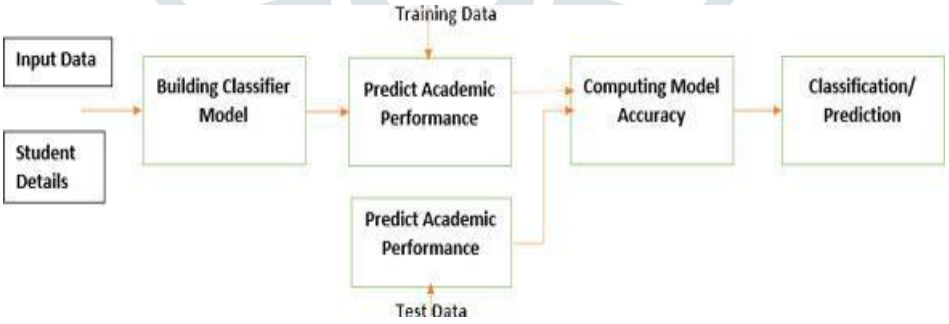
Training Data
Input Data
Student Details
Building Classifier Model
Predict Academic Performance
Computing Model Accuracy
Classification/ Prediction
Predict Academic Performance
Test Data

Fig 1. System Architecture

V. RESULTS

The main objective of the study is to explore if it is possible to predict the performance of the student (output) based on the various explanatory (input) variables which are retained in the model. The classification model was built using several different algorithms and each of them using different classification techniques. In order to approximately evaluate the efficiency and effectiveness of our experiment; we compared five selective classification algorithms on the data. Those algorithms are K-Nearest Neighbour, Decision Tree, Support Vector Machine, Random Forest and GDBoost. Decision Tree is Classification algorithm which is performed on our data with some specific parameters that is Gini impurity on 2 spilt tree. With max depth of 15, hyper parameter tuning is generated with 93.28% accuracy and with 6.82%error rate. Support Vector Machine predicted the academic grade with 98.60% and the error rate is 1.40%. K-NN Classifier Predicts with 78.86%, which has lowest classification rate. Random Forest is ensemble method to control the bias and variance in data, it is improves its accuracy when the depth of the tree is high and lessens when its depth is low. To avoid the over fitting of the model we put a threshold values to our model with the value 14. With that depth it generated 91.61% accuracy with 8.39% error rate. Gradient Boost is also an ensemble method to boost up the model performance, this method generated highest accuracy with the boosting technique at 99.66% and the error rate is 0.34%. **Accuracies table showed below.**

Classifier	Error rate	Accuracy
K-Nearest Neighbour	21.14%	78.86%
Support Vector Machine	1.40%	98.60%
Decision Tree	6.82%	93.28%
Random Forest	8.39%	91.61%
Gradient Boost	0.34%	99.66%

Table 2. Accuracies of Models

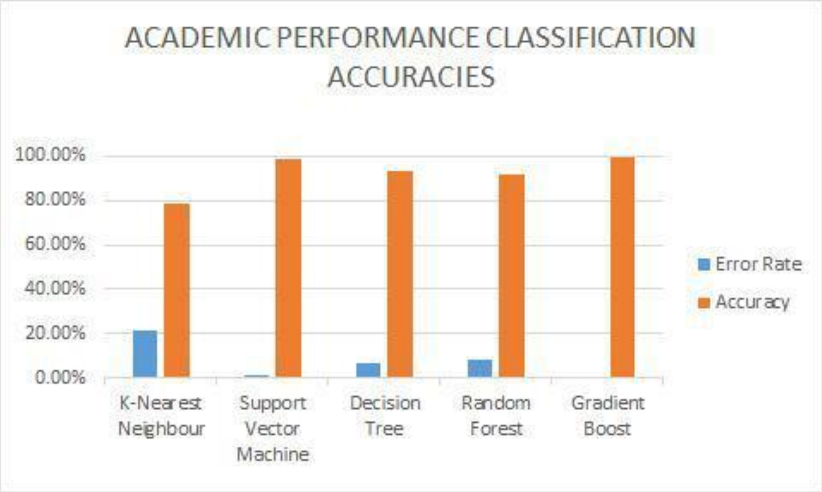

Fig 2: Accuracies of Models

VI. CONCLUSION AND FUTURE SCOPE

Education is very important in present generation and methods to analyze the educational system and learner's performance are essential for the advancement of institution and the student. The proposed automated system emphasizes on making predictions of student's academic performance, social behaviour and technology. The accuracy of the model is also calculated however, there remains a full exploration of the technological features of this model which will be the focus of the future work. Further scope is to build a deep learning classifier and analyze which classifier is more appropriate in carrying out the classification.

.

REFERENCES

[1] Hashmia Hamsa, Simi Indiradevi, Jubilant J. Kizhakkethottam, Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm, Procedia Technology, Volume 25, 2016, Pages 326-332.

[2] S. S. Athani, S. A. Kodli, M. N. Banavasi and P. G. S. Hiremath, "Student academic performance and social behavior predictor using data mining techniques," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, 2017, pp. 170-174.

[3] Xiaofeng Ma and Zhurong Zhou. "Student Pass Rates Prediction Using Optimized Support Vector Machine and Decision Tree", 978-1-5386-4649-6/18/$31.00 ©2018 IEEE.

[4] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi and Sunday O. Olatunji. "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor", 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE).

[5] Pauziah Mohd Arsad, Norlida Buniyamin and Jamalul-lail Ab Manan. "A Neural Network Students' Performance Prediction Model (NNSPPM)" IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), 26-27 November 2013.

[6] Kayah, F. "Discretizing Continuous Features for Naive Bayes and C4. Classifiers". University of Maryland publications: College Park, MD, USA.

[7] David, L. M. and Carlos E. G. Data Mining to Study Academic Performance of Students of a Tertiary Institute, American Journal of Educational Research,2(9), 2014, pp. 713-726. Doi: 10.12691/education-2-9-3.

[8] Romero, C. and Ventura, S., Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications, 33(1), 2007, pp. 135-146.

[9] Anuradha, C & T, Velmurugan. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. Indian Journal of Science and technology.974-6846. 10.17485/ijst/2015/v8i15/74555.

[10] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[11] https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[12] E.Chandra and K. Nandini, "Predicting Student Performance using Classification Techniques", Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India, p.no, 83-87.

[13] S. Huang, & N. Fang, Work in Progress - Prediction of Students' Academic Performance in an Introductory Engineering Course, In 41st ASEE/IEEE Frontiers in Education Conference, (2011), 11–13.