Analysis of Data Mining, Big Data and Machine Learning Techniques to Predict Academic Performance

Sharath Sriram
B.Tech Computer Science and Engineering
School of Computing
SASTRA Deemed University
Thanjavur, India
E-mail: sharathsriram@sastra.ac.in

Vishal R
B.Tech Computer Science and Engineering,
School of Computing,
SASTRA Deemed University
Thanjavur, India
E-mail: vishalr@sastra.ac.in

J. Naren
Assistant Professor
School of Computing,
SASTRA Deemed University
Thanjavur, India
E-mail: naren@cse.sastra.edu

***Abstract*- Education is the key for solving majority of world problems. It is important that the education imparted is assimilated by every student who aspires to learn. Classifying slow learners in the early stages so that necessary help can be given to improve their performance is one of the crucial responsibilities for educational institutions. Technology plays an important role in aiding in the process of classifying students. Various techniques have been developed to predict the performance of students accurately for addressing their learning needs and styles. In the present paper, a bird’s eye view of the data mining and machine learning techniques used for predicting student performance is presented.**

Keywords - Machine Learning, Big data, Data Mining

I. INTRODUCTION

Education is a fundamental right given to every human being. Every person with a passion to learn must be given the resources and opportunity to get educated. In educational institutions, there is a possibility of some students being slow learners in comparison to others. It is highly crucial for the institution to identify these students so that necessary support can be given to help them finish the course. It is even more essential that this process is carried out in the early stages of the course so that students have enough time to get back on track. A report by the UNESCO in 2016 states that 47 million youth in India drop out of schools before they reach Grade 10. While dropouts are due to various reasons, one among them is the student’s inability to match the standards of education. Identifying such inability in the early stages of their education can help in motivating and guiding students in the right track and can prevent a considerable number of people from dropping out of school.

The marriage between technology and education brings about a myriad of possibilities. Both technology and education at their core, have a common goal- empowering and transforming the world. Technology can help in classifying and identifying students to get them back on track. With the recent advent of data mining and machine learning, the advancements in the process of classification have taken a giant leap. In the present paper, various techniques in data mining, machine learning have been explored for their usefulness in predicting academic performance of students and assisting them in their education.

Data mining is the process of discovering patterns in data sets using statistical methods and algorithms. Educational data mining has become a buzzword in recent years and there has been an abundance of data mining based research for predicting student performance. However, most of the research work tends to overlook important parameters and the feature sets used, often end up being prior academic records. Many times, other factors like cultural background, character of the individual etc. are not taken into consideration.

Data Mining and Machine Learning are growing fields in recent years. Big data simply refers to large volumes of unstructured data that is difficult to process using traditional techniques. Machine Learning is defined as an art that gives computers the ability to learn without being explicitly programmed. Of late, the techniques of machine learning and data mining have also been used in academic performance prediction. Some other use cases in educational domain include prediction of career choices, identifying learning capabilities of students etc. Early detection is very important in education as it can help in providing vital help to necessary students.

II. Data Mining for Student Performance Prediction

Muslihah Wook et.al compared the techniques of artificial neural networks and combination of clustering and decision tree classification. Student data of the Computer Science Department of National Defence University of Malaysia was used. Features like demographics (age, gender, religion), education background and personality were taken into consideration and 60% of data was used for training with the rest of the data allotted for validation. The patterns influencing academic performance were identified by them to improve efficiency in the future. [1]

Hashmia Hamsa et.al proposed the usage of decision tree and fuzzy genetic algorithm for predicting student performance. Early classification was done to selectively focus and improve performance of slow learners. Internal marks, admission scores, high school scores were taken as parameters in predicting performance. The results from the algorithm were analysed and used by teachers to guide slow learners and the top students were identified to prepare them better for placements in top companies. [2]

In the paper by Norlida Buniyamin et.al, some common classifier techniques were explored to predict and classify students' achievements in a Malaysian Public University. The study was performed to judge single/group performance, so that assistance could be given to guide students. [3]

Ishwank Singh et.al proposed the use of data mining technique for understanding the performance of students to group them into categories. A clustering analysis was done to understand student behaviour. The analysis was done to help during admission and placement process and hence parameters like high school marks, college grades were taken into consideration. [4]

Yohannes Kurniawan et.al proposed the use of data warehouse and data mining techniques for predicting student performance in schools. A discussion has been made on how the inference from the data mining model can be used to assist students. Parameters like test scores, attendance and assignments were taken into consideration with the view that inappropriate data in these parameters could result in false predictions. Students with higher probability of failing were identified by the algorithm for motivating them to pass the course so that they are not forced to repeat the year. [5]

A novel method of decision tree induction approach called Multivariate Regression Prediction Model M5P was proposed by S. Chaitanya Kumar et.al for predicting performance of students based on online-learning skills, problem solving efficiency, time management, adaptable nature, sports participation, versatility, practical knowledge etc. Data collected from 307 students doing their third year in Computer Science and Engineering in India was used. The outcome from the model achieved an accuracy of 97.17%. It is proposed to use the model for early prediction of performance so that remedial measures can be taken. [6]

Tismy Devasia et.al used a web based application based on Naive Bayesian mining technique for prediction of academic performance. A data set of 700 students from Amrita Vishwa Vidyapeetham, Mysuru was used with 19 attributes taken into consideration. The corresponding results and comparison proved that Naive Bayesian algorithm provided better accuracy than other methods like Regression, Decision Tree, Neural Networks etc. Future work with additional attributes and newer data sets have been considered. [7]

Zhenpeng Li et.al illustrated a novel approach for predicting students final grade using attributes related to students past academic records and attributes of normal study behaviour. A fuzzy clustering and multi-variate regression approach were used and comparative experimental investigations were carried out. [8]

III. Big Data and Machine Learning for Student Performance Prediction

Muhammad Fahim Uddin et.al explored the potential of using personality data from social networking such as Facebook, LinkedIn and correlating it with academic and career data to improve prediction and classify good and bad fit students. The techniques of Stochastic Probability Distribution, Bayesian Networks were used. Future measures of understanding career changes, re-admission to colleges, study performances, real world job progression have been discussed. [9]

Roshani Ade and P.R. Deshmukh proposed the use of a pair of classifiers to predict student career choices. Students' marks in psychometric test were used as training set for the study. Four pairs of classifiers were used in the study. The incremental learning algorithm was found to be efficient in predicting the career choices on a dataset created by conducting a psychometric test on 1333 students of the age group of 16 to 20. 14 attributes were used. Usage of web log and multimedia data set has been proposed as future work. [10]

Carlos J. Villagra-Arnedo et.al proposed the use of black box techniques for predicting student performance and to output rich and meaningful data for easy understanding by teachers. A Support Vector Machine Model was used in the study and enhanced data visualization in the form of graphs and heat maps were presented. A high accuracy in the prediction was obtained by the end of the ten weeks of course. A set of tips for better prediction were proposed. [11]

S. Rajeswari et.al developed a model using big data for predicting student grades and final year campus placements. Predictive Analytics(EDM) were used and it was concluded that EDM was one of the best models for predicting student performance. The identification of potential weak students was done to take appropriate action for improving their performance. [12]

Dr. Manju Jose et.al proposed a predictive model to forecast student performance based on contextual factors. A data set of 245 students was taken under the study. A huge data set for more effective and accurate prediction has been proposed as future work. [13]

Jie Xu et.al proposed a novel machine learning method that can be used to predict student performance in degree programs. A data set collected over three years at University of California, Los Angeles was used. A data driven approach based on latent factor models and probabilistic matrix factorization was employed to perform the study. Extending the performance prediction to elective courses has been proposed as future work. [14]

Radhika R Halde et.al proposed two machine learning algorithms to test the impact of students' psychology on academic performance. Neural Networks and Decision Trees were the techniques used in the study. The psychological state of the student was deeply considered as a parameter in the study, adding novelty in comparison to other research on student performance. The impact of the mental state of students in academic performance was observed through the study. [15]

Name/Title	Data Set	Method
a) *Data mining for modelling students' performance: A tutoring action plan to prevent academic dropout* *Concepción Burgos, María L. Campanario, David de la Peña, Juan A. Lara, David Lizcano, María A. Martínez*	Data obtained from students of five different BS in Computer Engineering Programmes taught during 2013/2014 academic year.	Logistic regression technique is used in order to build a reference model for the grades of all the students.
b) *Towards the integration of multiple classifier pertaining to the Student's performance prediction* c) **Mrinal Pandey, S. Taruna**	Data obtained from engineering college in India, consisting of demographic information, behavioural characteristics etc.	Three complementing classifier namely DT (J48), KNN (IBK) and AODE are integrated and proposed a single composite model (KNNAD) based on voting strategy.
Improving the expressiveness of Black Box Models for predicting Student Performance *Carlos J. Villagra- Arnedo, Francisco J. Gallego-Duran, Faraon Llorens-Largo, Patricia Compan-Rosique, Rosana Satorre-Cuerda, Rafael Molia-Carmona*	Data collected for the subject Computational Logic in which students were provided with an interactive web application where they had to program a character in a game to overcome obstacles.	The prediction system is based on a standard C-parameterized margin, SVM classifier with Radial Basis Function kernel, adding probability estimates using Pairwise Coupling
On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees *Keeley Crockett, Annabel Latham, Nicola Whitton*	Data collected from questionnaire completed by 75 students with knowledge in SQL	A module is used in the OSCAR architecture known as Fuzzy Learning Styles Predictor(FLSP). OSCAR is the name of a Conversational Intelligent Tutoring System that models a person's learning style using natural language dialogue during tutoring in order to dynamically predict, and personalise, their tutoring session.

Table1: Comparison of various Machine learning and big data techniques to predict academic performance

IV. CONCLUSION

Advancements in technology have made possible a system of education where prediction and identification can help students overcome hurdles and ace their courses. While there is abundance in the presence of data mining techniques used to predict student performance, machine learning and big data enhance and scale up this process. It helps in dynamically helping students to pursue their education without any roadblocks. The current trend has been to relate academic output with prior academic performance. Newer research suggests that personality features play an equally important role in student education.

In the present paper, the techniques of data mining, machine learning and their usefulness in predicting student performance were explored. There is a lot of scope for future work in this growing field of applying meaningful technology to empower education. Early prediction of learning styles and academic performance can be used to create a personalised learning experience for each student. Knowing the academic performance and behavioural habits can be used to predict the employability of an individual. These theories are yet to be tested but if successfully implemented, have a high probability of succeeding in creating a more efficient academic experience.

V. REFERENCES

[1]. MuslihahWook, Yuhanim Hani Yahaya, NorshahriahWahab, Mohd Rizal Mohd Isa, Nor Fatimah Awang, Hoo Yann Seong,(2009),"Predicting NDUM Student’s Academic Performance Using Data Mining Techniques", 2009 Second International Conference on Computer and Electrical Engineering

[2]. HashmiaHamsa, Simi Indiradevi, Jubilant J.Kizhakkethottam,(2016), "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm", Procedia Technology, Volume 25, 2016, Pages 326-332

[3]. NorlidaBuniyamin, Usamah bin Mat, PauziahMohd Arshad, (2015), "Educational Data Mining for Prediction and Classification of Engineering Students Achievement", Engineering Education (ICEED), 2015 IEEE 7th International Conference on 17-18 Nov. 2015

[4]. Ishwank Singh, A Sai Sabitha, Abhay Bansal, (2016), "Student performance analysis using clustering algorithm", Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference.

[5]. YohannesKurniawan, Erwin Halim, (2013), "Use data warehouse and data mining to predict student academic performance in schools: A case study (perspective application and benefits)", Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference on 26-29 Aug. 2013

[6]. S Chaitanya Kumar, Deepak Chowdary, VenkatramaPhani Kumar S, Krishna Kishore, (2016), "M5P model tree in predicting student performance: A case study", 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)

[7]. TismyDevasia, Vinushree T, Vinayak Hegde,(2016), "Prediction of Students Performance using Educational Data Mining", 2016 International Conference onData Miningand Advanced Computing**(**SAPIENCE2016)

[8]. Zhenpeng Li, Changjing Shang and Qiang Shen, (2016), "Fuzzy-clustering embedded regression for predicting student academic performance", Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on 24-29 July 2016

[9]. Muhammad Fahim Uddin, Jeongkyu Lee, (2016), "Utilizing Relevant Academic and Personality Features from Big Unstructured Data to Identify Good and Bad Fit Students", Procedia Computer Science, Volume 95, 2016, Pages 383-391

[10]. Roshani Ade, P. R. Deshmukha, (2015), "Efficient Knowledge Transformation System Using Pair of Classifiers for Prediction of Students Career Choice", Procedia Computer Science, Volume 46, 2015, Pages 176 – 183

[11]. Carlos J.Villagrá-Arnedo, Francisco J.Gallego-Durán, FaraónLlorens-Largo, Patricia Compañ-Rosique, Rosana Satorre-Cuerda, Rafael Molina-Carmona, (2017), "Improving the expressiveness of black-box models for predicting student performance", Computers in Human Behavior, Volume 72, July 2017, Pages 621-631

[12]. S. Rajeswari, R. Lawrance, (2016), "Classification model to predict the learners' academic performance using big data", Computing Technologies and Intelligent Data Engineering (ICCTIDE), International Conference on 7-9 Jan. 2016

[13]. Manju Jose, PreethySinu Kurian, Biju V. (2016), "Progression analysis of students in a higher education institution using big data open source predictive modeling tool", Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference on 15-16 March 2016.

[14]. Jie Xu, KyeongHo Moon, Mihaela van der Schaar, (2017), "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs", IEEE Journal of Selected Topics in Signal Processing, Volume: 11,Issue: 5, Aug. 2017,Pages 742 - 753.

[15]. Radhika R Halde, Arti Deshpande, Anjali Mahajan, (2016), "Psychology assisted Prediction of Academic Performance using Machine Learning", 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)