IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

A Survey on Student Performance using Data Mining Techniques

Zainab Fatema[1*], Geeta Pattun[2]

[1,2]Dept. of Computer Science and Information Technology, Maulana Azad National Urdu University, Hyderabad, India

[*]*Corresponding Author: f.zain.2016@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i3.707710 | Available online at: www.ijcseonline.org

Accepted: 13/Mar/2019, Published: 31/Mar/2019

Abstract— Students are the main stakeholders of institutions and their performance plays a significant role in country development. The aim of institutions is to give excellence educations to their students. It has been observed in the previous works, students of slow performance and dropouts are the most vital issues. Due to early detection of slow performers and dropouts of students, help the teachers, administrator, and management to take appropriate actions at the right time for improving the overall performance of the students. The purpose of this study is to analyze different data mining and machine learning techniques on student data and find which technique gives better accuracy. And, we also find different factors like socio-demographic, psychological factors, attendance of students, understanding level of students, previous grades, study time, parent's status, internet usage, travel time, extracurricular activities, and also health factors affect the performance of students. Data mining is a process of analyzing data and turns it into useful information.

Keywords— Students Academic Performance, Data Mining, Machine Learning.

I. INTRODUCTION

Educational Data Mining doesn't generate exact results for future but it concentrates on prediction. EDM is used for resolving educational research issues [5]. Data mining techniques are used to discover meaningful information from a huge amount of data [9]. Different data mining techniques which are repeatedly used in papers are, Naïve Bayes, Multilayer Perceptron, REP tree, J48, SMO, Random Forest, Logistic Regression, Support Vector Machine, optimized Support Vector Machines, K-Nearest Neighbor, C4.5 and so on. WEKA, Rapid Miner, DB Miner, Intelligent Miner tools are used for mining.

Different researchers give their own definition of education but the meaning of all is the same. Performance of students based on different factors such as Socio-demographic variables, Psychological factors, understanding levels, the capacity to learn, ability to perform well in exams, Student attendance, extracurricular activities, Psychological and Health-related factors, Home environment, Teaching-Learning methods [13]. To predict the performance of students is a challenging task for researchers. For this, we need to analyze the previous performance of students, so that by using that information, the future prediction can be done. For this researchers take data from the University database or by questionnaire based on identified attributes. And, by applying different machine learning and data mining techniques this prediction can be done. By predicting the student performance in future may be useful for management or administrator to improve education quality for students.
Rest of the paper is organized as follows, Section II contains the related work of student performance, Section III concludes the research work.

II. RELATED WORK

(Ms. Ashna Sethi, Mr. Charanjit Singh 2017) presented a study on 300 engineering students of Gulzar Group of Institutes and develop a predictive data mining model using Improved Naïve Bayes technique. This study helps to identify those students who have fewer marks and poor performance and to take action at the right time and also reduce the fail ratio. They use Naïve Bayes and Improved Naïve Bayes algorithms. Improved Naïve Bayes give better performance. The Accuracy of Improved Naïve Bayes was 86.66% [1].

(Edin Osmanbegovic, Mirza Suljic 2012) predicted student's success using data mining algorithms. The algorithms used in this study are Naïve Bayes, Multilayer Perceptron, C4.5. In this study, Naïve Bayes predicts better than other algorithms. The data is collected from the University of Tuzla, the academic year 2010-2011 by the surveys in the summer semester [2].

(Brijesh Kumar Baradwaj, Saurabh Pal 2011) present a study on 50 students from VBS Purvanchal University,

Janpur(Uttar Pradesh) of course MCA from 2007 to 2010. In this study classification task is used to evaluate student performance by using a decision tree method. For predicting student performance, attendance, class test, seminar and assignment marks data are collected in this study [3].

(Farshid Marbouti, Heidi A. Diefes-Dux, Krishna Madhavan 2016) develop prediction models for identifying at-risk students as early as possible in a course using standards-based grading. The data is collected from the course conducted in the second semester of first-year engineering students from Midwestern U.S university. The author uses six different prediction methods to identify at-risk students are logistic regression, support vector machine, decision tree, multilayer perceptron, Naïve Bayes classifier, and K-nearest neighbor. The author compares all the six different prediction methods in which logistic regression overall accuracy is 92.6%, k-nearest overall accuracy is 94.9% and multilayer perceptron accuracy is 93.1%, decision tree accuracy is 92.3%, SVM accuracy is 87.2% and Naïve Bayes accuracy is 86.9% [4].

(Parneet Kaur, Manpreet singh, Gurpreet Singh Josan 2015) focus on identifying slow learners by using classification algorithms. For this study, they used a high school record of 152 students. Naïve Bayes, J48, Multilayer Perceptron, SMO algorithms are used in this study for prediction. WEKA an open source tool is used. They tested and analyze all the algorithms in which multilayer perceptron produce the best accuracy of 75% focus on identifying slow learners by using classification algorithms. For this study, they used a high school record of 152 students. Naïve Bayes, J48, Multilayer Perceptron, SMO algorithms are used in this study for prediction. WEKA an open source tool is used. They tested and analyze all the algorithms in which multilayer perceptron produce the best accuracy of 75% [5].

(M. RamaSwami and R. Bhaskaran 2010) conducted a study on the higher secondary school for predicting students' performance. The data is collected from a survey. From five different schools of Tamilnadu total of 1000 datasets were collected. The CHAID algorithms are used in this study for prediction. The accuracy of this model was 44.9% [6].

(Mrs. VarshaPDesai 2018) uses C# algorithm on a set of attributes to predict slow learner, average learner and fast learner students. In this naïve Bayes algorithm is used 14 attributes are taken for this prediction. this prediction is helpful for teachers for providing and improving the quality of education to the students [7].

(R. Sumitha, E.S. VinothKumar 2016) uses senior data for predicting a student's future outcome. The dataset is collected from B.E CSE of KLN college of about 350 students. A questionnaire is used for collecting real data. The dataset attributes were 24 initially. For this work, WEKA explorer is used. The algorithms are used are naïve Bayes, multilayer perceptron, REPtree, and J48. For creating the model java coding is used in Net Beans and J48 algorithm is used for designing [8].

(Ahmed Mueen, Bassam Zafar, Umar Manzoor 2016) predict students' academic performance by applying data mining techniques. They classification algorithms used in this study are a neural network, naïve Bayes and decision tree. They collected data of undergraduate students, these students take courses of Programming Fundamental and Advanced Operating System from August 2014 to May 2015. They built three classification models for predicting student academic performance. From all models, naïve Bayes classification produced the best accuracy of 86% [9].

(Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D.lytras, Farhat Abbas, Jalal S.Alowibdi 2017) predicted student performance using generative and discriminative classification models. They use two generative model- Bayes network(BN) and naïve Bayes(NB) and three discriminative models-support vector machine(SVM), C4.5 and classification and regression tree(CART). They collected data from different universities of Pakistan about 3000 student records from 2004 to 2011. After applying to preprocess, they considered 776 student instances. The aim of this research is to predict which student complete his/her degree or which drop. They found SVM classifier is effective for family expenditure and student personal information categories [10].

(Humera Shaziya, Raniah Zaheer, G.Kavitha 2015) presents a study to predict the performance of students in a semester exam based on Naïve Bayes Classifier. The author takes the data set of MCA students. All the implementations are done in WEKA 3.7.10. in this naïve Bayes model is used and seven attributes were selected to the prediction of final results [11].

(Mohammadreza Zahedifard, Iman Attarzadeh, Hadi Pazhokhzadeh, Javad Malakzadeh 2015) conducted a study on school students to predict the students' performance. By using this study manager of high school find important factors which affect a student's success and can improve the quality of the school. This study uses 386 student's data set of high school in Bushehr province. The data were collected by using questionnaire from five different fields physics, mathematics, humanities, natural science, technical and occupational field. Authors use different methods, naïve Bayes, nearest neighbor, Bayes nets, support vector machines, artificial neural networks, decision tree, logistic regression [12].

(K. Prasada Rao, M.V.P. Chandra Sekhara, B.Ramesh 2016) conducted a study on students which give a better perspective for student performance in the future by using classification techniques. The algorithms used in this study was J48, Naïve Bayesian Classifier, and Random Forest Algorithm. The data was collected from rural-based computer science and engineering students about 200 records. For implementation WEKA software package was used. By comparing all the algorithms, the conclusion was when the data size increases Random Forest Algorithm shows better accuracy [13].

(Swati and Rajinder Kaur 2016) conducted a study on predicting slow learner using factor classification. The data was collected from high school students from the diverse region in Punjab. 2400 data was used for a test using the proposed model. Naive Bayes classification model is used. The proposed model work in online and offline both [14].

(Mukesh Kumar, Shankar Shambhu, Punam Aggarwal 2016) conducted study on high school students to identify slow learner. The algorithms used in this study was J48, REPTree, Naïve Bayes, SMO, Multilayer Perceptron. The data was collected from two different high schools help of a survey. For implementation, WEKA tool was used. Multilayer Perceptron gives higher accuracy of 87.43% [15].

(Farshid Marbouti, Heidi A. Diefes-Dux, Krishna Madhavan 2016) focuses on identifying at-risk students early using standards-based grading. In this study, seven different models are compared using academic factors. The aim of this study is to find the best prediction method. The data used in this study was from first-year engineering students of Midwestern U.S. University. The algorithms used in this study are logistic regression, Support Vector Machine, Decision Tree, Multilayer Perceptron, Naïve Bayes Classifier, K-Nearest Neighbor. By comparing all the models, found K-Nearest Neighbor was the best model for overall accuracy. Which identifies 94.9% of failed and passed students correctly [16].

(Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, Mustafa I. Al-Najjar) conducted a study on the improvement of the quality of education in higher education students. The data was collected from the students of computer science of Yarmouk university through the questionnaire. The classification algorithms used in this study are ID3, C4.5, Naïve Bayes. Implementation was done in WEKA tool. Because of small dataset classification, accuracy was very low. So, by using large sample size classification accuracy can be improved [17].

(V. Ramesh, P. Parkavi, K. Ramar 2013) conducted a study to find out different factors which influence the students' performance in the final exam. The data was collected from higher secondary students of Kancheepuram district. Collected data was from both primary and secondary source. The primary source was a questionnaire and secondary source was school database. For this study WEKA tool is used. The algorithms used in this study are Naïve Bayes, Multilayer Perception, SMO, J48, REP Tree. It was concluded that Multilayer Perception Classifier gives higher accuracy of 72.38% [18].

III. CONCLUSION

Student's performance plays an important role in a country's social and economic growth by producing graduates, innovators, and entrepreneurs. This paper surveys the various methods and techniques used to predict the performance of students. The data mining and machine learning techniques used in previous researches are – decision tree, support vector machine, naïve Bayes, neural network, k-NN, regression, random forest algorithm etc. These techniques are predicting the future behavior of students by analyzing the student's data. The prediction results are helpful for institution, students or teachers for improving student performance.

REFERENCES

[1] A. Sethi, "Data Mining for Prediction and Classification of Engineering Students achievements using Improved Naïve Bayes", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. **6**, Issue **7**, pp. **966-971, 2017**.

[2] E. Osmanbegović, "Data Mining Approach for Predicting Student Performance", Economic Review – Journal of Economics and Business, Vol. **X**, Issue **1**, pp.**3-12**, May **2012.**

[3] B. K. Baradwaj, "Mining Educational Data to Analyze Students" Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. **2**, No. **6**, pp. **63-69**, **2011**.

[4] F. Marbouti, "Models for early prediction of at-risk students in a course using standards-based grading", ELSEVIER, Vol. **103,** pp.**1-15, 2016.**

[5] P. Kaur, "Classification and prediction based data mining algorithms to predict slow learners in education sector", 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015), Ghaziabad, **India**, pp.**500-508, 2015**.

[6] M. Ramaswami, "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues, Vol.**7**, Issue.**1, pp.10-18, 2010.**

[7] V. P. Desai, "Classification Technique for Predicting Learning Behavior of Student in Higher Education", International Conference on Digital Economy and its Impact on Business and Industry, Sangli, **India,** pp.**163-166, 2018.**

[8] R. Sumitha, "Prediction of Students Outcome using Data Mining Techniques", International Journal of Scientific Engineering and Applied Science ,Vol.**2**, Issue.**6**, pp.**132-139, 2016.**

[9] A. Mueen, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques", I.J. Modern Education and Computer Science, Vol.**8**, No.**11**, pp.**36-42, 2016**.

[10] A. Daud, "Predicting Student Performance using Advanced Learning Analytics", International World Wide Web Conference Committee, Perth, **Australia**, pp.**415-421, 2017.**

[11] H. Shaziya, "Prediction of Students Performance in Semester Exams using a Naïve Bayes Classifier", International Journal of Innovative Research in Science, Engineering and Technology, Vol.**4**, Issue.**10**, pp.**9823-9829, 2015**.
[12] M. Zahedifard, "Prediction of Students Performance in High School by Data Mining Classification Techniques", International Academic Journal of Science and Engineering Vol. **2**, No. **7**, pp.**25-33**, **2015.**
[13] K. P. Rao, "Predicting Learning Behavior of Students using Classification Techniques", International Journal of Computer Applications, Vol.**139**, No**7**, pp.**15-19, 2016**.
[14] Swati, "Using Factor Classification for the Slow Learner Prediction over Various Class of Student Dataset", Indian Journal of Science and Technology, Vol.**9**, Issue.**48,** pp.**1-5, 2016.**
[15] M. Kumar, "Recognition of Slow Learners using Classification Data Mining Techniques", Imperial Journal of Interdisciplinary Research, Vol.**2**, Issue.**12,** pp.**741-747, 2016.**
[16] F. Marbouti, "Models for early prediction of at-risk students in a course using standards-based grading", ELSEVIER, Vol.**103**, pp.**1-15, 2016.**
[17] Q. A. Al-Radaideh, "Mining Student Data using Decision Trees", International Arab Conference on Information Technology, Jordan, **Arab, 2006.**
[18] V. Ramesh, "Predicting Student Performance: A statistical and Data Mining Approach", International Journal of Computer Applications, Vol.**63**, No.**8**, **2013.**

Authors Profile

Ms. Zainab Fatema completed Bachelor of Technology(B.Tech) from Maulana Azad National Urdu University (MANUU), Hyderabad, India, in 2017 and currently pursuing Master of Technology(M.Tech) form the same university that is MANUU. Her research interest is in machine learning and big data.

Mrs. Geeta Pattun obtained her Masters of technology degree from Osmania University, Hyderabad, India. At present working as an Assistant Professor in department of Computer Science and Information Technology, Maulana Azad National Urdu University(MANUU), Hyderabad, India.

