Identifying Data Mining Techniques and Tools for Improving Student's Academic Performance

Ancy A[1]
IInd M.Sc. IT.,
Department of Computer Science
Stella Maris College, Chennai, India
deborah.ancy@gmail.com

Ms. Birunda Antoinette Mary J[2]
Asst. Professor.,
Department of Computer Science
Stella Maris College, Chennai, India
birunda78@gmail.com

Abstract

Data mining is the evolving process of identifying and extracting the hidden information from a data warehouse. Data Mining is widely used in business, medical, engineering and educational areas for analyzing existing data, identifying measures for improvement and also forecasting the future prospects. This study covers the application of data mining in education for predicting the academic performance of the students. Educational Data Mining(EDM) plays a dominant role in the data mining era. There is an essential need to identify effective algorithms for predicting the student's performance. With the help of EDM we can predict the academic performance of the students using different data mining techniques like Decision tree, Random Forest, Gradient Boosted tree, Naive Bayesian and Multilayer Perceptron. Each technique has its own advantages and disadvantages. This paper discusses about the different types of EDM techniques and what are the different tools used for implementation. The use of effective EDM techniques and tools would enable educators to improve the process by identifying any existing lacunae. EDM helps in developing a warning system for identifying weak student's prior and give adequate training to improve the academic performance of the students.

Keywords:
Student Performance Prediction, Educational Data Mining, Data Mining Technique, Academic Performance, Decision tree, Random Forest, Gradient Boosted tree, WEKA.

INTRODUCTION

Data Mining is process of analyzing the important information from a large set of data and come up with the prediction model. Data Mining is also called as Knowledge Discovery in databases(KDD). The data mining plays an important role in all the fields like medical, airline, banking sector, movies, scientific information and numerous new data types. Data mining can be used to solve real time problems. Currently data mining technique has been applied in educational system. Educational Data Mining(EDM) is the emerging technique for developing the prediction model with the help of available dataset, and extract the prediction of students' academic performance using machine learning technique. The prediction model acts like a warning system which is used to identify the weak students. EDM is a new field of research in data mining. The recent increase in online learning by the students have led to the progression in development of EDM. The data set for EDM is collected from different educational institutions like school, colleges and universities which has enormous data available in the database for each year. The highly reputed educational institution mainly focuses on improving the performance of the students in order to retain the standard rank of the institution, hence they train the students in such way that they perform well in academics and extra-curricular activities. The data mining is classified into:

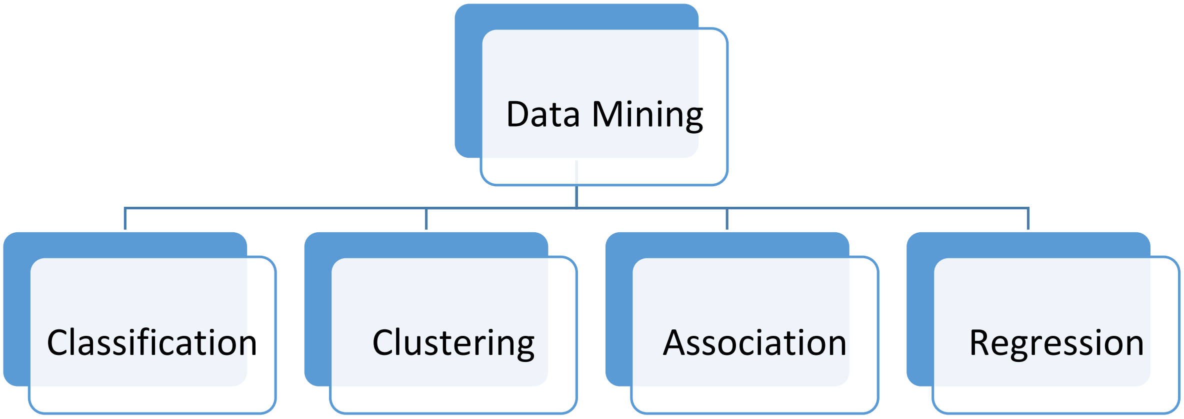

Figure 1: Data Mining Techniques

By using Educational data mining technique, the educational institutions can predict the performance of the students and identify low performing student early enough to overcome their difficulties in learning and improve their learning outcomes. Day by day the volume of the data is increased, hence there are different data mining algorithms which are used for predicting the performance of the students like supervised and unsupervised techniques to get the maximum accuracy. The supervised method is divided into Classification or Categorization and Regression. The unsupervised method is classified into Clustering and Association. Each technique has its own advantage and disadvantage more the accuracy rate the more specific the prediction is. Some of the algorithms which are popularly used in prediction are Decision tree, KNN, Naïve Bayes, Random forest, Gradient boosted trees, ID3 and J48. This paper comprises of what are the different techniques which are used in the educational data mining for predicting the performance of the students and a study is made on different types of students attributes and factors needed for developing the prediction model. Major data mining techniques which are used for predicting the student's performance are shown below:

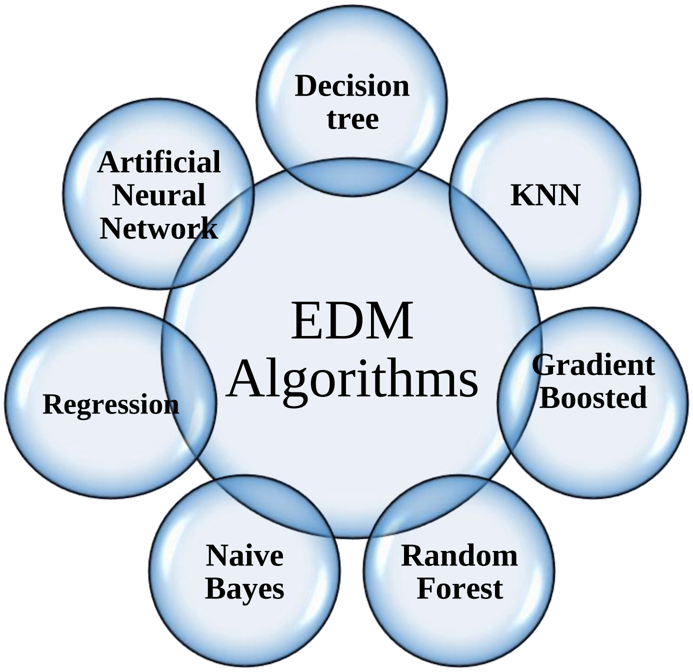

Figure 2: Data Mining Algorithm

LITERATURE SURVEY

Students' social activities and background details were used by Ching-Chieh Kiu [1] in the year 2018 for predicting the student's academic performance. This study used several data mining technique has been applied on predicting the accuracy of each dataset, of which decision tree J48 has given the best accuracy rate. The decision tree model has achieved 95% accuracy rate compared to other data mining technique. The dataset was divided into three subsets, students background details, student's social activities and student course work. The algorithms were implemented in WEKA tool. A comparative study on decision tree and random forest was done by Fergie Joanda and Reymon in the year 2018. The dataset was collected using questionaries' from the computer science students. It had 249 instances out of which 3 different classes of data are used. The study proves that decision tree gives more accurate result of 66.9% compared to random forest which gave 61.44%. The comparison was implemented using WEKA tool [2].

A comparative study was done on Naïve Bayes, Decision Tree, K-Nearest Neighbor and Discriminant Analysis by Samuel, Nor Bahiah and Siti Mariyam in the year 2019. The study was done to identify the best data mining technique for predicting the student's academic performance. It used 10 datasets from the University of California Irvine Repository. The decision tree out performed with the accuracy of 81.94% compared to other data mining techniques. The accuracy of other techniques was Naïve Bayes-73.61 %, KNN-80.56 % and Discriminant Analysis -77.78% accuracy rate. The tool used in the study was WEKA. In future hybrid metaheuristics algorithms will be for feature selection on the student data [3]. According to the study of Nongnuch Ketui, Warawut Wisomka and Kanitha Homjun in year 2019, Gradient boosted trees has given the best accuracy compared to other classification data mining techniques like Decision Tree, Weighted Decision Tree, Iterative Dichotomiser 3 (ID3) and Random Tree. WEKA is the data mining tool which is used for implementing the data mining techniques. A raw dataset was collected from the Rajamangala University of Technology Lanna Nan. The gradient boosted tree and decision tree gave good accuracy rate of 92.31% and 91.03% compared to other techniques. The Weighted Decision Tree 84.14%, ID3-89.66% and Random Tree-84.14% accuracy rate. The classification technique is widely used in predicting the student's performance [4].

According to Romero, Cristóbal, et al [5] SMO gives more accurate result compared to other techniques like BayesNet, Naïve Bayes Simple and EM. The paper used four different datasets and each dataset produced its own accuracy. The algorithms were implemented using WEKA tool. The

SMO produced 82.4% accuracy result and BayesNet produced accuracy result of 81.5% accuracy. The Naïve Bayes Simple produced an accuracy of 82.4% and EM has 80.7% accuracy rate. All the algorithms performed equally good. In the year 2014 [6], Hu, Ya-Han, Chia-Lun Lo, and Sheng-Pao Shih made a study on predict students' online learning using C4.5, CART and LGR. The WEKA tool was used for implementation. The dataset used was learning portfolio data and C4.5 produced more accurate result of 93.4%. Yu, Liang-Chih, et al [7] used Sentiment analysis in order to predict the student's academic performance in the year 2018. The study used unstructured dataset. The WEKA tool was used for implementing the sentiment analysis which produced an 76% accuracy rate. According to Deepika, K., and N. Sathvanaravana Support Vector Machine produces more accurate result compare to Linear Regression and Random forest. The study used student dataset of various academic disciplines of higher educational institutions in Kerala, India. The Linear Regression produced 89.96% accuracy and Random forest produced 89.98% and Support Vector Machine produced 91.43% accuracy result [8].

A comparative study was done in the year 2018 by Uzel and Vahide Nida on Multilayer Perceptron, Random Forest, Naïve Bayes, Decision Tree and Voting classifiers. The study used an educational dataset (xAPI) which is generated from an e-learning system includes 480 instances and 16 attributes. The Voting classifiers has highest accuracy of 80.6% [9]. The Artificial Neutral Network outperformed the decision tree and Naïve Bayes. The study used dataset from Kalboard 360 e–learning system with 500 instances and 17 attributes. The Decision Tree has 71.1% accuracy, Naïve Bayes has 67.5% accuracy and ANN has highest accuracy of 78.1% [10]. According to Rawat, Keshav Singh, and I. V. Malhan a hybrid classification gives more accurate result for predicting the student's academic performance. The study used data set of Department of Computer Science with 27 instances and 11 attributes. The Decision Tree produced 86.7% and KNN produced 87.5%, ANN produced 81.3% and NB produced 87.5%, Hybrid produced highest accuracy rate of 93.3% [11].

PAPER	EDM TECHNIQUE	TOOLS	DATASET	RESULT ACCURACY
[1] Kiu, Ching-Chieh. "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities." *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, 2018.	1.Naïve Bayesian(NB) 2.Multilayer Perceptron 3. Decision Tree(DT) 4.J48 5. Random Forest	WEKA	Used 395 instances with 33 attributes that described performance in Mathematics subjects	DT-95% NB-76%
[2] Kaunang, Fergie Joanda, and Reymon Rotikan. "Students' Academic Performance Prediction using Data Mining." *2018 Third International Conference on Informatics and Computing (ICIC)*. IEEE, 2018.	1.Decision Tree(DT) 2. Random Forest(RF)	WEKA	Used 249 records with 3 different classes	DT -66.9% RF-61.14%
[3] Ajibade, Samuel-Soma M., Nor Bahiah Ahmad, and Siti Mariyam Shamsuddin. "An Heuristic Feature Selection Algorithm to Evaluate Academic Performance of Students." *2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)*. IEEE, 2019.	1.Naïve Bayes (NB) 2.Decision Tree (DT) 3. K-Nearest Neighbor (KNN) 4. Discriminant Analysis (DISC)	WEKA	Used 10 different datasets that are gotten from the University of California Irvine (UCI) Repository.	NB-73.61% DT-81.94 % KNN-80.56 % DISC-77.78%
[4] Ketui, Nongnuch, Warawut Wisomka, and Kanitha Homjun. "Using Classification Data Mining Techniques for Students Performance Prediction." *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*. IEEE.	1.Decision Tree 2.Weighted Decision Tree(WDT) 3.Iterative Dichotomiser 3 (ID3) 4.Random Tree 5.Gradient Boosted Trees	WEKA	Education Division of Rajamangala University of Technology Lanna Nan (RMUTL Nan) for gave the raw dataset	DT-91.03% WDT-84.14% ID3-89.66% RT-84.14% GBT-92.31%

PAPER	EDM TECHNIQUE	TOOLS	DATASET	RESULT ACCURACY
[5] Romero, Cristóbal, et al. "Predicting students' final performance from participation in on-line discussion forums." *Computers & Education* 68 (2013): 458-472.	1.SMO 2.BayesNet 3.NaiveBayesSimple 4.EM	Meerkat ED SNAPP	Used four different student dataset	SMO -82.4% BayesNet - 81.5% NaiveBayes82.4% EM -80.7%
[6] Hu, Ya-Han, Chia-Lun Lo, and Sheng-Pao Shih. "Developing early warning systems to predict students' online learning performance." *Computers in Human Behavior* 36 (2014): 469478.	1.C4.5 2. CART 3. LGR.	WEKA	Used learning portfolio data	C4.5-93.4% CART-76.9% LGR.-95%
[7] Yu, Liang-Chih, et al. "Improving early prediction of academic failure using sentiment analysis on self-evaluated comments." *Journal of Computer Assisted Learning* 34.4 (2018): 358-365.	1.Sentiment Analysis	WEKA	Used unstructured data	Sentiment Analysis-76%
[8] Deepika, K., and N. Sathvanaravana. "Analyze and Predicting the Student Academic Performance Using Data Mining Tools." *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2018.	1.Linear Regression 2.Random forest 3.SVM	WEKA	Used student dataset of various academic disciplines of higher educational institutions in Kerala, India	LR-89.96% Random forest - 89.98% SVM-91.43%
[9] Uzel, Vahide Nida, Sultan Sevgi Turgut, and Selma Ayşe Özel. "Prediction of Students' Academic Success Using Data Mining Methods." *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2018.	1.Multilayer Perceptron(MP) 2. Random Forest 3.NaïveBayes 4.Decision Tree 5.Voting classifiers(VC)	WEKA	Used an educational dataset which includes 480 instances and 16 attributes	MP -78.3 % RF-76.6% NB -67.7% DT -75.8% VC-80.6%

PAPER	EDM TECHNIQUE	TOOLS	DATASET	RESULT ACCURACY
[10] Siddiqui, Isma Farrah, and Qasim Ali Arain. "ANALYZING STUDENTS'ACADEMIC PERFORMANCE THROUGH EDUCATIONAL DATA MINING." *3C Tecnologia* (2019).	1.Decision Tree 2.Naïve Bayes 3.Artificial Neural Network	WEKA	Used dataset from Kalboard 360 e–learning system with 500 instances and 17 attributes	DT-71.1% NB -67.5% ANN -78.1%
[11] Rawat, Keshav Singh, and I. V. Malhan. "A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining." *Proceedings of 2nd International Conference on Communication, Computing and Networking*. Springer, Singapore, 2019.	1.Decision tree 2.KNN 3.Artificial neural network 4.Naïve Bayes 5.Hybird	WEKA	Used data set of Department of Computer Science with 27 instances and 11 attributes	DT-86.7% KNN-87.5% ANN-81.3% NB-87.5% Hybird-93.3%

Decision tree has accuracy level from 66.9% to 95% and Random Forest has accuracy level from 61.14% to 89.98%. Naïve Bayes has accuracy level from 67.5% to 82.4% and KNN has accuracy level from 80.56% to 87.5%. ANN has accuracy level from 78.1% to 81.3% and Discriminant Analysis has accuracy level from77.78%. ID3 has accuracy level from 89.66% Weighted Decision Tree has accuracy from 84.14%. Gradient Boosted Trees has accuracy level of 92.31% and Sentiment Analysis has accuracy level 76%. Linear Regression has accuracy level 89.96% and Support Vector Machine has accuracy level 91.43%. Multilayer perceptron has 78.3% accuracy and Voting classifier has 80.6% accuracy. CART has 76.9% accuracy and LGR has 95% accuracy.

FINDINGS FROM SURVEY

Following observations are noted based on the survey:

a) Algorithms

Identification of right algorithms is the key for successfully prediction of the student's performance. From the figure2 shows the algorithms which are used for Educational data mining. Among the mentioned algorithms decision tree is the widely used algorithms, which also produces the more accurate result compared to other algorithms. Gradient boosted trees are also emerging in the EDM, which also give more accurate result like decision trees.

b) Feature Selection

Feature selection is one of the major task for predicting the performance of the student's. Identification of right features from the dataset is the challenging task of EDM. The feature selection can also be done using the different types of Educational data mining which is listed in the figure2.

c) Tools

The most popular data mining tool which is used for implementing the EDM algorithms is WEKA. The WEKA tool is an open source software which has built-in algorithms. The algorithms can either be applied directly to a dataset or called from Java code.

d) Data Mining Technique

The figure1 explains the different data mining techniques which are used for predicting the student's performance. The classification technique is one of the most widely used technique which implemented in many EDM

CONCLUSION

Educational data mining is the interesting field of research for educationalist. With the help of EDM the educational institutions can be benefitted by identifying the week student's and give adequate training for improving the performance of the student. This paper identified the most common data mining approaches, tools, techniques and algorithms which are used for predicting the student's performance. WEKA is the most common and widely used data mining tool for implementing the data mining algorithms. The data mining technique which are used for student's performance prediction are classification, clustering and regression. The classification technique is the popular data mining technique used in Educational Data Mining. The algorithms which are used for EDM are Decision tree, Random Forest, Gradient Boosted trees, Support Vector Machine, Naïve Bayes, Artificial Neutral Network and KNN. The algorithms which gave more accurate results are Decision tree and Gradient boosted trees. The widely used algorithm in EDM is Decision tree with high accuracy rate. Hybrid technique can also be used to get more accuracy result set. Identifying the correct features which affect their behavior or performance is an important task.

REFERENCES

[1] Kiu, Ching-Chieh. "Data Mining Analysis on Student’s Academic Performance through Exploration of Student’s Background and Social Activities." *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, 2018.

[2] Kaunang, Fergie Joanda, and Reymon Rotikan. "Students' Academic Performance Prediction using Data Mining." *2018 Third International Co-nference on Informatics and Computing (ICIC)*. IEEE, 2018.

[3] Ajibade, Samuel-Soma M., Nor Bahiah Ahmad, and Siti Mariyam Shamsuddin. "An Heuristic Feature Selection Algorithm to Evaluate Academic Performance of Students." *2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)*. IEEE, 2019.

[4] Ketui, Nongnuch, Warawut Wisomka, and Kanitha Homjun. "Using Classification Data Mining Techniques for Students Performance Prediction." *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*. IEEE.

[5] Romero, Cristóbal, et al. "Predicting students' final performance from participation in on-line discussion forums." *Computers & Education* 68 (2013): 458-472.

[6] Hu, Ya-Han, Chia-Lun Lo, and Sheng-Pao Shih. "Developing early warning systems to predict students’ online learning performance." *Computers in Human Behavior* 36 (2014): 469-478.

[7] Yu, Liang-Chih, et al. "Improving early prediction of academic failure using sentiment analysis on self_evaluated comments." *Journal of Computer Assisted Learning* 34.4 (2018): 358-365.

[8] Deepika, K., and N. Sathvanaravana. "Analyze and Predicting the Student Academic Performance Using Data Mining Tools." *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2018.

[9] Uzel, Vahide Nida, Sultan Sevgi Turgut, and Selma Ayşe Özel. "Prediction of Students' Academic Success Using Data Mining Methods." *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2018.

[10] Siddiqui, Isma Farrah, and Qasim Ali Arain. "ANALYZING STUDENTS’ACADEMIC PERFORMANCE THROUGH EDUCATIONAL DATA MINING." 3C Tecnologia (2019).

[11] Rawat, Keshav Singh, and I. V. Malhan. "A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining." Proceedings of 2nd International Conference on Communication, Computing and Networking. Springer, Singapore, 2019.