Student’s Performance Prediction using Deep Learning and Data Mining Methods

Jabeen Sultana, M. Usha Rani, M.A.H. Farquad

ABSTRACT--- Educational organizations are unique and play utmost significant role for the development of any country.As Education transforms the lives of individuals, families, communities, societies, countries and ultimately the world! This is why we live comfortable lives today. Now a day’s education is not limited to only the classroom teaching but it goes beyond that like Online Education System, Web-based Education System, Seminars, Workshops, MOOC course. becomes It’s more challenging to Predict student’s performance because of the huge bulks of data stored in the environments of Educational databases, Learning Management databases.Students’ performance can be evaluated with the help of various available techniques.Data Mining is the most prevalent techniques to evaluate students’ performance and is extensively used in Educational sector known as Educational data mining. It is evolving area of study that emphases on various techniques of data mining like classification, prediction, feature selection. It is employed on learning recordsor data related to education to predict the students’ performance and learning behavior by extracting the hidden knowledge. EDM is a methodology or like a procedure which is used to mine valuable information and patterns or forms from a massive educational database. Subsequently, the student’s performance is predicted from the obtained useful information and patterns. The prime motto of our study is to discover the performance of students using some classification techniques and discovering the best one which yields optimal results. Educational Dataset is collected from a Saudi University database. The dataset is pre-processed to filter duplicate records; missing fields are identified and filled with the destined data. Deep Learning techniques like Deep Neural Net and Data Mining techniques like Random Forest, SVM, Decision Tree and Naïve Bayes are employed on the data set using Weka and Rapid Miner tools. Results achieved are evaluated on few metrics. Deep Neural Network and Decision Tree outstands in predicting students’ performance compared to other techniques by producing deep predictions and obtains the best results like high accuracy, kappa-statistic, Sensitivity and Specificity are also determined.

Keywords—Educational Data Mining (EDM), Deep Learning, Random Forest, Decision Tree, Naïve Bayes and SVM.

I. INTRODUCTION

Discovering Knowledge from huge Databases is known as Data mining. It discovers hidden information from diverse data sources pertaining to diverse fields. Several techniques can be employed in different fields of data mining together with weather forecasting, oil research, business, medical, marketing and EDM etc. [2].To extract and analyze the knowledge present in educational data sources, a subdomain of data mining has also been developed termed as Educational data mining (EDM). Data mining, statistics and machine learning are applied on EDM data to derive knowledge from educational environments.Currently, it is in demand and gaining more attention because of increase in the educational data of e-learning systems, and even progressing traditional education. Alarmed with evolving techniques for discovering the distinctive types of data present in scholastic environments,It seeks to extract meaningful information in order to advance and appreciate learning processes from vast amounts of raw data [6]. Probing traditional records of database can offer answer to Problems such as “find the students who failed the examinations”, whereas EDM offers answers to additional problems like “predicting the students who are more likely to pass”.

Coming to educational institutions, advancing the student models so that student characteristics or performances can be predicted well in advance is distinct key areas of EDM applications. Therefore, many researchers started exploring various data mining techniques in order to assist educationalists or instructors to assess and progress their respective course organization [7].

Student performance prediction is going worst in our current educational systems. If the performances of students are predicted well in advance, then it can upkeep or improve the quality of education by predicting student' subject interests, student level activities, and assists in improving their performances in school’s universities and educational institutes. classifying dropout points can also be done by this [4].By means of machine learning techniques along with EDM, continuous evaluation system is practiced by several institutions today. These schemes are helpful in improving the student’s performance. Benefiting the regular students is the prime motto of continuous evaluation systems. Preprocessing pipelines and data transformations is the by-product of the effort in strategic spreading out of machine learning algorithms. They contribute in data representation to provide for active machine learning, and focus the drawbacks of current learning algorithms [1]. Survey findings on some Deep learning applications that can be applied in different fields like Image Processing, Natural Language Processing, and Object detection were found [8].

In order to predict students’ performance, knowledge discovery is suggested here to mine rules from the dataset of Systems of Learning Managements. Deep learning and data mining techniques are employed here. Deep Learning

Revised Manuscript Received on June 10, 2019.

Jabeen Sultana, Research Scholar,Department of Computer Science, SPMVV UniversityTirupati, A.P, India. (E-mail: jabeens02@gmail.com)

M. Usha Rani, Department of Computer Science, SPMVV University,Tirupati, A.P, India. (E-mail: musha_rohan@yahoo.com)

M.A.H. Farquad, Faculty of Computer and Information Systems, Islamic University of Madinah,Madinah Al-Munawwara, Kingdom of Saudi Arabia. (E-mail: farquadonline@gmail.com)

classifier-MLP and other classifiers like KNN, Naïve Bayes are used in our research work. A model is constructed by applying the classifiers on our data. 10-fold Cross Validation is accomplished. Parameters like Accuracy, Specificity, Sensitivity, Kappa-statistic and ROC curve are considered for evaluating the Classifiers.

Next section deals with brief review donein section II and the suggested techniques and approaches are explained in the section III. In Section IV, detailed Experimental results are shown with proper discussion and followed by Conclusion and expected work in the future is presented in Section-V.

II. LITERATURE REVIEW

Educational data mining in short EDM is widespread nowadays due to increase in e-resources, usage of online tools for education and Internet. Lots of research is taking place to make best of education tools and technologies [3]. The usage of EDM techniques to predict or analyze the students' performance and improve the students who are falling below satisfactory grades, an Artificial neural network classifier model was built which can be beneficial for both students and teachers to discover knowledge from huge data present in educational sector [12]. Sentiment analysis was carried on to understand the different way of students learning and their plan of study in order to improve teaching [14].

Student's behavioral features were considered with other features and a model was proposed based on data mining techniques which yielded 22.1% high accuracy after removing behavioral features.Further, by employing ensemble methods there was found 25.8%increase in accuracy [5].

Academic data set consisting of 473 instances, and found that 70% accuracy was yielded by Bayesian classifier. The naive Bayes classifier, ANN, KNN and j48 were used to categorize student's dropouts. 87% and 79.7% accuracy was yielded by K-nearest neighbors and decision trees applying 10-fold cross-validation [10].

SVM's separates the classes in high-dimension space by constructing hyper plane [15]. Data mining techniques like Logistic regression and Multi-classifiers produced outstanding results on data related to health [16]. Optimal approaches are used by the Decision Trees in order to identify or to reach a definite target in the practical world [13]. Combining this concept and speeding up the time of training [9] explains their widespread use in EDM. Decision Tree was used to predict classes pass or fail on a dataset of 15150 instances, 85.92% accuracy was attained [11].

By analyzing the above studies, we propose efficient techniques to predict students' performance. Deep Neural Network Classifiers namely MLP is used for Deep Learning approach, and Classifiers of Data mining namely Support Vector Machine, Naïve-Bayes, Multi-class Classifiers, Random Forests and Decision tree-j48 are considered. Later, the proposed classifiers were supplied with data set to undergo training and a Model was obtained. The obtained Model was supplied with test data. The dataset was collected from Learning Management System from a Saudi University database.In this Paper, Best classifiers of data mining and deep learning are compared. Later, assessed on few factors like Accuracy of the model built, Kappa-statistics, TP, TN and ROC curve.

DATASET AND DATA PREPROCESSING

Source of data for building the proposed techniques to envisage the performance of students is collected from learning management system called D2L. The dataset consists of 1100 student records. It has 11 different features.

Students Dataset		
Name	Type of Data	Range/Values
ID	Nominal	4
Raisedhands	Numeric	0-100
VisITedResources	Numeric	0-100
AnnouncementsView	Numeric	0-100
Discussion	Numeric	0-100
ParentAnsweringSurvey	Nominal	2
ParentschoolSatisfaction	Nominal	2
StudentAbsenceDays	Nominal	2
Internal assessment	Nominal	0-60
External assessment	Nominal	0-40
Total Marks	Nominal	0-100

Table-1

The dataset has three classes based on their numerical interval values.

Classes	
Internal-values	Class Label
0-69	Low
70-84	Medium
85-100	High

Table-2

WEKA tool is used in our research work. It is a standard Machine Learning tool, offers data preprocessing, cleaning and removing anomalies. Classification, regression, clustering, market-basket analysis and feature selection [19].

III. SUGGESTED METHODOLOGIES

A. Outline about the Proposed Methodology to predict performance of the students'

Data mining and Deep learning classifiers are given to the data set collected from educational environments. Data is preprocessed and check for missing values. Classifiers are applied on the data set to build the models. Models are tested with test data to predict the students' performance and the best models yielding high accuracy are considered.

The Methodology Proposed in this research work has the important phases described below.

1. Firstly, MLP-a Deep Neural network classifier and classifiers of data mining namely Bayes Net, SVM, Random Forest, Decision tree and Multi-class Classifiers. Training was performed on Educational dataset and a Model was obtained.
2. The Obtained Model from the first phase is supplied with test data set and the results are obtained. 10-fold cross validation is done here in the second phase which involves both training and testing in 10-folds.
3. The obtained results are assessed on parameters like Accuracy, Specificity, Sensitivity, Kappa-statistic, and ROC curve and the best model yielding high results is selected in the third phase.

The Framework for predicting student's performance is summarized below.

Proposed Frame Work

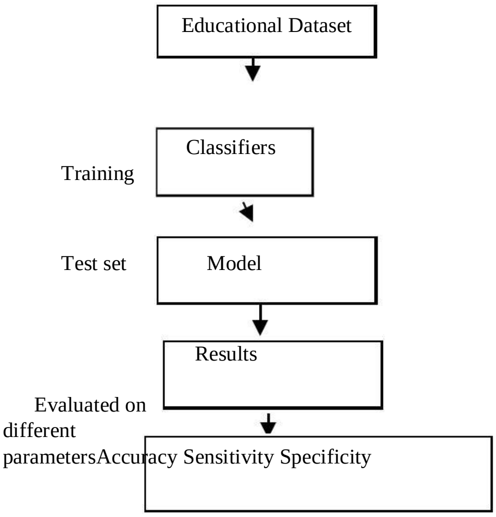
Educational Dataset
Training
Classifiers
Test set
Model
Results
Evaluated on different parameters
Accuracy Sensitivity Specificity

Figure-1

B. Brief discussion on different classification techniques used:

Dataset was collected from an educational institute of Saudi University. Experimentations were done using open source tool named as WEKA. WEKA can proficiently work with limited data and used in data mining, Data is pre-processed by using filtering techniques and Classification is done. Data set is taken in .csvformat, 10-fold cross-validation was done. The results were deeply analyzed on different parameters like Accuracy, Kappa-statistic Specificity, Sensitivity, and ROC curve area.

Weka is widely used software to perform classification and regression tasks [18].

C. Description of Techniques:

MLP:MLP, kind of neural network follows the standard way to train multilayer perceptron's and uses back propagation algorithm [18].

SVM: Classification and regression analysis is observed using SVM. It constructs optimal hyperplane on the training data, further classifies new instances based on this hyperplane. SMO classifier is used [21].

Naïve-Bayes: A Classifier, grounded on the concept of probabilities, estimates classes by considering numeric precision values, Group of Features classified by this classifier is independent of each other [18].

Multi-class Classifier: It handles datasets comprising of Multi-classes by combining 2-class classifiers [18].

Lazy-Star: It is an instance-based classifier [18].

Random Forest: Randomly trees are constructed leading to a forest [20].

Decision Trees: Constructed, starting from the root and continues until it reaches to its leaf nodes. J48 algorithm is employed to implement. [17].

The knowledge discovery is done in this research paper to predict student's performance by applying Deep learning and Data mining techniques. The obtained results are assessed using few parameters like Accuracy; Classifiers accuracy is calculated by seeing the correctly classified instances to the total test dataKappa-statistic, Specificity and ROC Curve are being calculated from the Confusion Matrix using Weka tool.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Different classifiers were chosen in this research work and comparative analysis of their performance was done using WEKA tool. Educational Dataset was pre-processed and later fed to Neural Network-MLP, SVM, K-NN classifiers, Decision Tree, Naïve-Bayes, Random Forest and Multi-class Classifier. training and testing was performed in ten different folds resulting in accurate Model. The obtained results from the model built were measured in different terms like Accuracy, Kappa-statistic and ROC curve area.

End results of Classifiers used

Methods	**Accuracy**	**TP**	**FP**	**KappaStatistic**	**ROC**
MLP	99.45	1.00	0.00	0.99	1.00
Multi-Class Classifier	99.81	1.00	0.00	0.99	1.00
SVM	93.90	1.00	0.10	0.89	0.94
Naïve-Bayes	97.45	0.98	0.00	0.95	0.99
IBK	79.81	0.91	0.16	0.63	0.87
Lazy LWL	86.72	1.00	0.00	0.75	1.00
Random Forest	100	1.00	0.00	1	1.00
Decision Tree	100	1.00	0.00	1	1.00

Table 3

It is experimented that MLP-Deep Learning technique, Data mining technique-Random Forest, performed well in predicting student's performance. Techniques that gave optimal results are MLP, Decision trees and Random Forest with maximum accuracies of99.45%, 99.81% and 100%.

V. CONCLUSION AND FUTURE WORK

Performance of student's using EDM is carried out in this research work. Classification is done in order to predict students in different class categories like High, medium and low. Classifiers used are Support Vector Machine (SVM), Multiple-Layer Perceptron approach (MLP), decision tree and Other Multi-classifiers for classifying students whether they belong to either High, medium or low classes. The results of both Data mining techniques and deep learning were compared on the basis of accuracy and precision. It was found and detected that classification implemented by MLP Multi-class Classifier, Decision trees and Random forest technique in this paper is more efficient compare to other classifiers as seen in the accuracy and precision. Based on the results, MLP technique is more efficient compared to other technique in prediction of students’ performance. Rules can be mined and accuracy needs to be improved in SVM, K-NN as part of the future work.

REFERENCES

1. J. Schmidhuber, Deep learning in neural networks: An overview. Vol.61, pp.85–117, 2015.
2. Han, pei and Kamber, Data Mining Concepts and Techniques, The Morgan Kaufmann series in data management systems, 3rd edition, 2011.
3. Pena A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications,vol.41 (4), 1432–1462, 2014.
4. Oyerinde et al. Predicting Students’ Academic Performances- A Learning Analytics Approach using Multiple Linear Regression, International Journal of Computer Applications, Vol.157, No.4,pp.37-44, 2017.
5. Amrieh et al. Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods, International Journal of Database Theory and Application, Volume 9, issue 8, pp.119-136, 2016.
6. Scheuer et al.,Educational data mining. In Encyclopedia of the sciences of learningpp. 1075–1079, Springer, 2012.
7. Romero & Ventura, Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics, vol. 0 (6), pp. 601–618, 2010.
8. Sultana et al. An Extensive Survey on Some Deep Learning Applications, SSRNe-journal, 2018.
9. Rokach&Maimon, Data mining with decision trees: theory and applications. World scientific. 2014.
10. Yukselturku et al. Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and eLearning, vol.17(1),pp.118- 133, 2014.
11. Jayaprakash et al. Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics, vol.1(1), pp. 6–47,2014.
12. Livieris, et al. Predicting students performance using artificial neural networks, 8th PanHellenic conference with International participation Information and communication technologies, pp.321-328, 2012.
13. Baker et al. Educational data mining and learning analytics. In Learning analytics pp. 61–75, Springer. 2014.
14. Sultana et al. Prediction of Sentiment Analysis on Educational Data based on Deep Learning Approach.*21st Saudi Computer Society National Computer Conference (SCSNCC)*. IEEE, 2018.
15. Meyer, D. Support vector machines the interface to libsvm in package e1071. 2014.
16. Sultana, J. Jilani, A.K. Predicting Breast Cancer using Support vector Machine and Multi-classifiers. International Journal of Engineering and Technology, Special Issue Vol. 7. No.4 20, pp.22-26, 2018.
17. Quinlan. R. Induction of decision trees. Machine Learning, vol.1, 81-106, 1986.
18. WEKA http://www.cs.waikato.ac.nz/_ml/weka.
19. Hall et al.. The WEKA data mining software: an update. SIGKDD Explorations, 11(1). 2009.
20. Leo Breiman. Random forests, Machine Learning:45(1)-5-32, 2001.
21. J. Platt: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C.Burges and A. Smola, editors, Advances in Kernel Methods-Support Vector Learning, 1998.