Classifying Student's Academic Performance using SVM

Jabeen Sultana

Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Riyadh, Kingdom of Saudi Arabia , j.sultana@mu.edu.sa

Kishwar Sadaf

Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Riyadh, Kingdom of Saudi Arabia, k.sadaf@mu.edu.sa

Abdul Khader Jilani

Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Riyadh, Kingdom of Saudi Arabia a.jilani@mu.edu.sa

Abstract

Learning management systems are mainly concerned to educational sectors play a dominant role in leading the nation across the globe. These days as everything goes online in terms of data storage across the globe. Lots of data emerges from learning systems and makes very promising to predict and classify learner's performances. In order to classify students' performance, various techniques are available. One of the most popular techniques to classify students' performance is Machine Learning (ML) and is widely used in learning systems to process Informative facts about learners. Processing Educational data involves usage of several data processing methods like forecasting, clustering and finding out associations in order to extract the valuable information of the learners, their mood changes in shifting of subjects and accordingly their performances by extracting the hidden knowledge. Subsequently the obtained useful information and patterns can be used in predicting student's performance. This research work suggests the effective technique in order to process and classify learner's performance. Data is gathered from a middle east university concerning to graduate course. ML techniques like Support Vector Machine-SVM, Multi-Layer Perceptron-MLP, Random Forest-RF, Decision Tree-DT, Naïve Bayes-NB and K-Nearest Neighbor-KNN are applied after preprocessing the data. The outcomes attained are assessed on few metrics like Accuracy, TPR, TNR, Kappa Statistics and ROC Curve. SVM outperforms in classifying learners' part linked to other methods by yielding optimal classification results like high accuracy and Sensitivity followed by MLP, RF, DT, NB and KNN.

Keywords:

Educational Data; Support vector Machine (SVM); Multi-Layer Perceptron (MLP); Random Forest (RF); Decision Tree (DT); Naïve Bayes (NB); K-Nearest Neighbor (KNN).

1.Introduction

Currently, Educational data mining (EDM) is demanded and attaining additional response due to increase in the data generation from learning management systems. EDM is an evolving area, startled with progressing methods to distinguish different categories of data emerging from learning systems. Simultaneously, valuable patterns are recognized so that improvements can be suggested at learner's front [1]. Accordingly, conventional data search can offer solutions to identify the learner's performances and suggestions can be made to improve their performance by finding out where the learners lack in understanding the concepts or fail to perform well in the exams etc. for the complete course duration in particular learning systems. The task is make this data wisely in improving the educational process.

Imminent learning systems advanced the learner prototypes in order to boost the performance of learners. Prediction and classification of learner's performance plays a key role in learning systems and is the hot area among the researchers as nations target to improve the learning experience and attract the learners for new courses. Consequently, lots of academicians and scholars are in full swing to explore several methods using machine learning so that course instructors can get benefitted in to assisting the learner's performance towards a particular course according to their interest [2]. Middle East student's data was classified using different machine learning classifiers and it was observed that neural networks attained good classification results compared to rest of the classifiers in predicting and classifying students' academic performance [3].

In this paper, Section II covers the detailed related work and Section III presents the proposed methodology and methods. Section-IV consists of classification results and discussions followed by Section-V describing conclusion and future work.

2. Related Work

Learner's interactive attributes in learning the course, difficulties faced and absent rate are given priority along with some other attributes and a method was suggested using machine learning techniques which yielded an improved classification accuracy of more than 22% by taking out interactive attributes. Moreover, classifiers were ensemble to improve the classification accuracy and it was observed that by using ensemble process, more than 25% accuracy was obtained [4]. Sentiments were analyzed to interpret the learner's way of dealing with the course right from start of the course till course completion, in this whole process how learners plan their schedule was analyzed so that a teacher can change the way of his or her approach towards the learner's. A brief comparative analysis was performed using other methods of ml and it was found that MLP attained the highest classification results in classifying learner's data [5]. Also, sentiment analysis of Telugu language tweets was performed to classify user sentiments of learning in Telugu. ML methods were applied and models were built accordingly and it was observed that Passive Aggressive Classifier attained utmost accurateness near to 80% with high precision rate-0.77, recall rate-0.78 and F1_score of 0.77 [6].

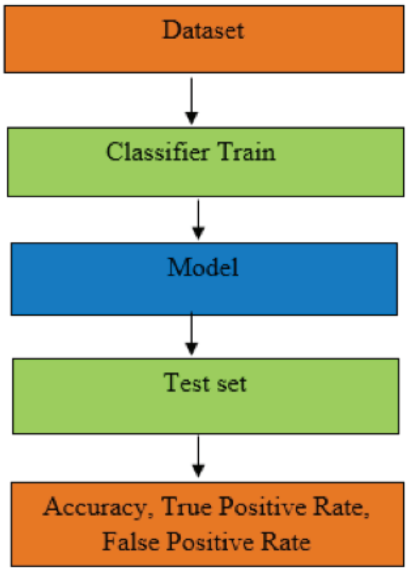
ML techniques are being used in processing the natural language of any medium and it was observed that they yielded promising results in classifying NLP data [7]. Discovering facts and gaining useful insights from big data pertaining to health sector [8,9] to link goods using association rules, one of the mining technique used in market basket analysis [10] and educational sectors [11], necessitate and slender in the direction of knowledge discovery methods to understand the ML methods expectancy. Traditional classifiers of deep learning were used to understand the models way of classifying learner's data gathered from a middle east university. Similarly, educational data was collected from Twitter to understand the public demanded courses and their way of approach in learning the new courses. Tweets were well processed and classified using ML and DL classifiers and it was observed DL classifiers yielded supreme results compared to ML classifiers [12, 13]. To analyze the academic potential of students, classification was performed on learner's data using ML and DL methods and found that both the methods yield superior results in classifying the students into groups based on their performances[14]. Basing the above studies, we suggest a method to classify student's academic performance on real time data collected from a middle east university database.

3.Proposed Methodology

Data was gathered from a LMS System of a middle east university database in order to predict the learners' performances. The dataset consists of 1300 instances with 11 different features namely Student id, section id, age group, grades, student absent count, raised hands, resources visited, participating discussions, viewing announcements, student answering, parent's feedback. Data had nominal and numerical features and therefore data was normalized. Further data was categorized into three classes based on the learner's total marks or internal values. The students who scored more than 85% of marks are been classified as high level and the students who scored between the range of 70% to 85% were classified as Middle level and furthermore the students who scored below 70% were classified as low level.

The below figure 1 demonstrates the suggested methodology in a frame work for classifying real time data using ML classifiers.

Fig.1: A Frame Work to Classify Real time data



3.1 Methods Used

SVM: Support vector machines are mainly utilized to perform supervised classification tasks. Linear and non-linear classifications can be performed using SVM by employing a kernel function. Also, used for regression tasks and it performs classification by constructing ideal hyperplane on the trained data. Two parallel hyperplanes are constructed on the sides of the separating hyperplane. The distance between the two parallel hyperplanes gets maximized by separating the hyperplane. Test data is imparted on the model built after training grounded on this hyperplane using SMO classifier. SVMs can be used for the classification of complex datasets in sophisticated applications such as handwritten digit recognition, object recognition, and text classification. [15].

MLP: Back propagation algorithm is used to train perceptron's; a model is built. Functions like logistic and hyperbolic tangent sigmoid functions are used to activate the perceptron's. There are some layers involved in this classifier where inputs are trained and are associated with proper weights and summed up to generate the activation function and is passed to other layers until results are attained in MLP [16].

Naive-Bayes: It is a classifier which constructs a model on trained data grounded on probabilities using Bayes concept. Cluster of attributes categorized by NB is free of each other. It is a technique to model the future possibilities in a class based on previous experiences. It predicts new classes on the imparted test data by utilizing numerical values on the model built. Naive Bayes classifier can be applied to binary and multi-class classification problems [17].

Random Forest: It is used for classification tasks with known class and is quite powerful in classifying the data. Unlike decision tree, random forest constructs many decision trees and considers the decision tree which yields the best results. RF's are constructed in a random fashion and in vast manner, gives us the impression like a forest. [18].

Decision Trees: DT's are constructed beginning from the root and continues till it stretches to its leaf nodes using if-then rules. The branches in the tree represent non similar attributes and the nodes at leaves on each branch represent a class. The training data is used to make the system learn the rules of classification. The decision tree is simple, doesn't require complex data representation, and gives exceptional performance for categorical and numerical data features. The decision tree construction is a two-step process, namely tree building, and tree pruning. Tree building refers to generating a decision tree on the training data employing a recursive breadth-first search algorithm. The tree pruning uses the remaining data to test the tree and correct the errors. The decision tree algorithm features simplicity, easy analysis, high accuracy in classification, and efficiency in execution [19].

KNN: K-Nearest Neighbor, is a basic classifier widely used in ML in classifying data with unknown class labels. Samples which resemble some similarity are grouped by calculating distance measures from a certain sample. It performs classification by grouping similar samples and calculates Euclidean distance measure to group them accordingly [20].

3.2 Metrics

Accuracy: Accuracy is a metric which describes or classifies data into proper class label or class. The extent to which classification of data is done correctly is observed by calculating accuracy.

Accuracy = (TP+TN)/(TP+FP+FN+TN)

TPR: True Positive Rate is known as TPR. It is also called as sensitivity. It is the possibility of classifying actual data belonging to positive class into positive class instead of other class. It is calculated as

TPR = TP/(TP+FN)

TNR: True Negative Rate is known as TNR. It is also called as specificity. It is the possibility of classifying actual data belonging to negative class into negative class instead of other class. It is calculated as TNR = TN/(TN+FP)

ROC: It is significantly used to measure the performance of a classifier in classifying data. It signifies how well the model created by the classifier is classifying the data. The ROC curve signifies the better performance of the model. It is plotted with respect to TPR against FPR.

4.Classification Results on Educational Dataset

Student's academic data was classified using ML classifiers, training and testing the data was carried on with the help of Rapid Miner tool. Classification results were analyzed in terms of accuracy, TP rate, FP rate, ROC curve area and Kappa statistics.

Table 1. Shows the Classification Results by ML Models

Methods	Accuracy	TPR	FPR	ROC	Kappa Statistics
SVM	93.90	1.00	0.10	0.94	0.89
MLP	86.72	1.00	0.00	1.00	0.75
Random Forest	78.33	0.34	0.78	0.89	0.66
DT	76.66	0.33	0.76	0.89	0.64
Naive Bayes	75.83	0.36	0.75	0.83	0.62
KNN	67.70	0.39	0.67	0.85	0.51

It was observed that SVM outstands by yielding 93% classification accuracy compared to other techniques like MLP, RF, NB. Whereas MLP yields the promising classification results of 86%, followed by RF-78%, DT-76%, NB-75% and KNN-67% attains the finest outcomes next to SVM. Also, SVM tops in achieving highest TP rate of 1.00, FP rate of 0.10, ROC curve of 0.94 followed by RF, DT, NB and KNN.

The below Fig.2 shows that SVM attains highest accuracy of 93.90 in classifying students into low level, medium level based on various attributes. MLP yields better results with an accuracy of 86.72% followed by Random Tree and decision tree. KNN failed to yield accurate results and is prone to incorrect classification by yielding low accuracy of 67% among the rest of the classifiers.

The below graphs show the comparison of all ML Classifiers used.

Fig.2: Shows Accuracy obtained by ML Classifiers

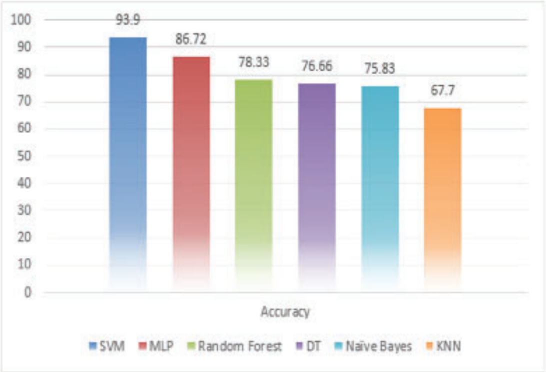


Fig.3: Shows TP rate obtained by ML Classifiers.

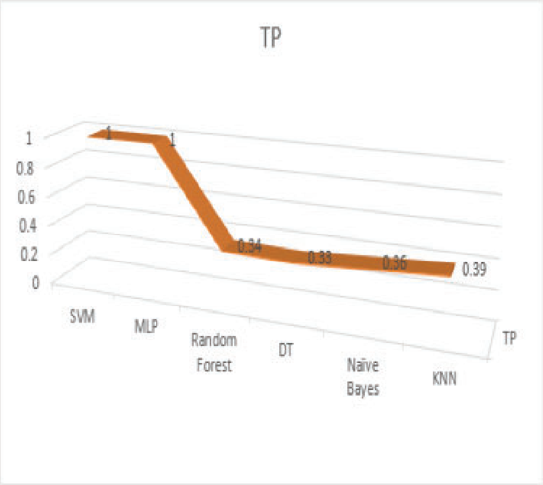


Fig.4: Shows FP rate obtained by ML Classifiers.

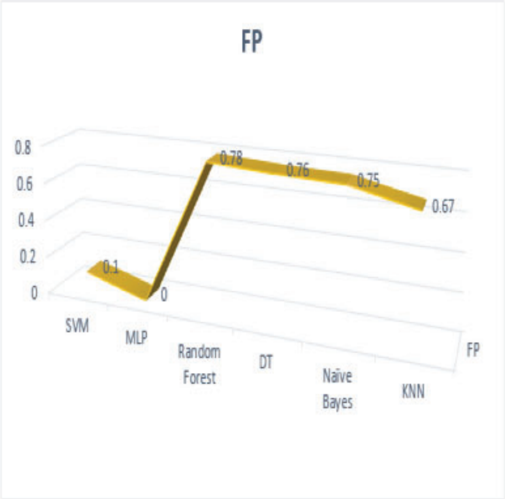


Fig.5: Shows ROC obtained by ML Classifiers.

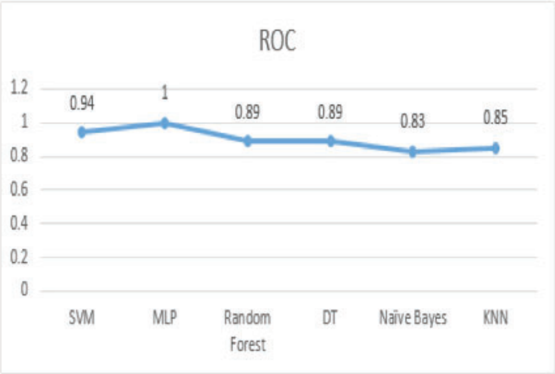


Fig.6: Shows Kappa Statistics obtained by ML Classifiers

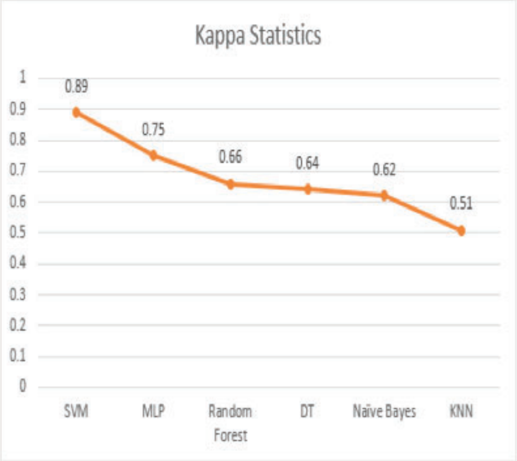
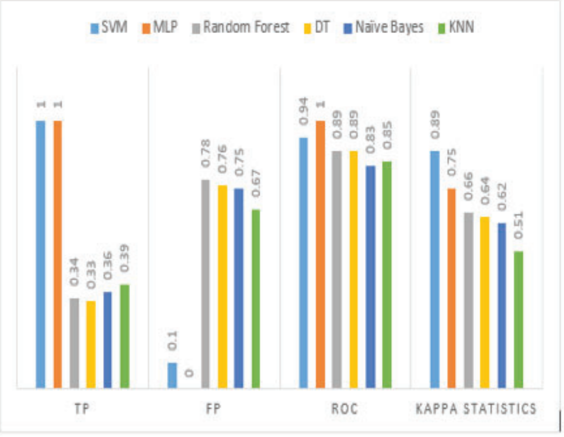


Fig.7: Shows TP, FP rate, ROC and Kappa Statistics obtained by ML Classifiers.



5.Conclusion and Future Work

In this research, we discuss about the proposed framework for classifying students' performance over Real time data using EDM. ML classifiers are used to classify learner's data into 3 classes namely slow level learners, middle level learners and top level learners, which was collected from a middle east university. in different class categories like High, medium and low. The results of all the machine Learning techniques are assessed based on few metrics like accuracy, sensitivity, specificity and ROC curve area. Data was well preprocessed and classified into proper class labels. SVM outstands by yielding 93% classification accuracy compared to other techniques like MLP, RF, NB. Whereas MLP yields the promising classification results of 86%, followed by RF-78%, DT-76%, NB-75% and KNN-67% attains the finest outcomes next to SVM. Even though MLP and DT yield good classification results but here on this data they failed to yield appropriate results. Also, Random Forest usually yields better performance on different data, here it fails to yield accurate results. KNN attains poor performance and is prone to incorrect class prediction. In future, we try to boost the performance of MLP and DT by using feature selection techniques. Also, Random Forest and KNN performance has to be boosted in the future work by using hybrid classifiers.

Conflict of Interest

None declared

References

[1]Scheuer et al., “Educational data mining”, In Encyclopedia of the sciences of learning pp 1075–1079, Springer, 2012.

[2]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.

[3]Sultana, J., Rani, U., Farquad, M. “An Efficient Deep Learning Method to Predict Student’s Performance”, 2019. https://www.researchgate.net/publication

[4]Amrieh et al. “Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods”, International Journal of Database Theory and Application, vol 9, issue 8, pp119-136, 2016.

[5]Sultana, J., Sultana, N., Yadav, K., & AlFayez, F., “Prediction of sentiment analysis on educational data based on deep learning approach”, 21st Saudi Computer Society National Computer Conference (NCC) pp. 1-5, 2018. IEEE.

[6]Priya, G.B.K., Sultana, J., Rani, “Telugu News Data Classification Using Machine learning Approach”, Handbook of Research on Advances in Data Analytics and Complex Communication Networks, IGI Global, pp 181-194, 2021.

[7]Sultana, J., Rani, M. U., Farquad, M. A. H., “An Extensive Survey on Some Deep-Learning Applications”, In Emerging Research in Data Engineering Systems and Computer Communications, pp. 511-519, Springer, Singapore, 2020.

[8]Sultana, J., Jilani, A. K., “Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers”, International Journal of Engineering & Technology, 7(4.20), 22-26, 2018.

[9]Sultana, J., Sadaf, K., Jilani, A. K., Alabdan, R. “Diagnosing Breast Cancer using Support Vector Machine and Multi-Classifiers”, International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), pp. 449-451, IEEE, 2019.

[10] Sultana, J., Nagalaxmi, G., “How Efficient is Apriori: A Comparative Analy-

sis", International Journal of Current Engineering and Scientific Research, ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, 2(8), pp 91-99, 2015.

[11]Sultana, J., Rani, M. U., Farquad, M. A. H., "Discovery from Recommender Systems using Deep Learning", International Conference on Smart Systems and Inventive Technology (ICSSIT) pp. 1074-1078, IEEE, 2019.

[12]Sultana, J., Rani, M. U., Farquad, M. A. H., "Deep Learning Based Recommender System Using Sentiment Analysis to Reform Indian Education", International Conference On Computational and Bio Engineering, pp. 143-150, Springer, Cham, 2019.

[13] Sultana, M. J., Rani, M. U., Farquad, M. A. H., "Sentiment Analysis based Recommender System for Reforming Indian Education using Multi-Classifiers", TEST Tets Engineering and Management Journal, 2020.

[14]Sultana, J., Rani, M. U., Farquad M. A. H., "Student's Performance Prediction using Deep Learning and Data Mining Methods" International Journal of Recent Technology and Engineering (IJRTE), vol. 8, iss. 1S4, 2019.

[15] Martens, D., Baesens, B., Gestel, T. V., "Decompositional Rule Extraction from Support Vector Machines by Active Learning", IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 2, pp 352- 358, 2009.

[16]Delashmit, W. H., Manry, M. T. "Recent developments in multilayer perceptron neural networks", In Proceedings of the seventh Annual Memphis Area Engineering and Science Conference, MAESC, 2005.

[17]Kohavi, R., "Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision Tree Hybrid", In Proceedings of KDD-96, Portland, USA, pp 202-207, 1996.

[18]Breiman L., "Random forests", Machine Learning:45(1)-5-32, 2001

[19]Quinlan, R., "Induction of decision trees", Machine Learning, vol 1, pp 81-106, 1986.

[20]Aha, D., Kibler D., "Instance-based learning algorithms", Machine Learning. vol 6, pp.37-66, 1991.