International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

Implementation of Student Performance Prediction and Course Material Recommendation System Using Data Mining

Niyati Lashkari
Department of Computer Science and Engineering, Sushila Devi Bansal College of Technology, Indore (M.P), India

Submitted: 01-07-2021 Revised: 10-07-2021 Accepted: 13-07-2021

ABSTRACT—Education is primary right of human that help to improve personality as well as help to improve their skills to deal with the real world problems. In educational domain a significant amount of data has been generated during different learning and teaching activity. In this context the mining and processing of the educational activity data may helpful for students, teachers and academic managers or administrators. In this work the student oriented data has been mined to recover the potential student learning growth patterns to predict the performance of student. Additionally the predicted performance of the student is used to recommend relevant study material to learn. The proposed concept includes the data mining techniques thus this domain of study is termed as the educational data mining. First the experimental database of student performance is collected from the online machine learning repository namely UCI. In next step the obtained dataset is pre-processed to refine dataset contents. Here, obtained dataset is unlabelled in nature thus we utilized k-means clustering algorithm for categorizing the data into three categories i.e. low, average and high performing students. Using these categories the dataset is labelled first and then used with the C4.5 decision tree algorithm. The trained decision tree algorithm is validated and their performance was measured. In addition, the outcome of student performance is used to recommend the suitable study material according to their learning behavior. The implemented model is evaluated and compared against the CART based algorithm. The experimental results reflect the superiority of the proposed model in terms of accurately prediction of student performance as well as recommendation..

Keywords—Educational data mining (EDM), Data mining, Supervised learning, unlabelled dataset, student performance prediction, study material recommendation system**.**

I. INTRODUCTION

The Educational Data Mining (EDM) is an application of data mining technique. In this application the educational data is mined for finding and discovering the patterns and trends from the data. These patterns are belongs to the two key objectives [1]:

1. Academic Objectives and
2. Administrative Objectives.

When the data is mined for Academic purpose the intention is to find information about the following criteria[2] [3]:

1. **Person Oriented:** to teaching and learning patterns such as Student learning, modeling, behavior, risk, performance analysis, predicting decision etc. for both in traditional and digital environment that also involve Faculty modeling-performance and satisfaction.
2. **Department/Institutions oriented:** that analysis is oriented to Particular department with respect to time, sequence and demand such as Redesign courses according to industry requirements, identify realistic problems.
3. **Domain Oriented:** This involves the Designing Methods-Tools, Techniques, Knowledge Discovery based Decision System (KDDS) etc.

In this presented work the person oriented EDM is the key point of attraction; more specifically the student performance prediction is the main aim of the proposed investigative work.The main aim of the proposed work is to predict the student performance additionally support the student by recommending the study material according to their reading behavior to improve their performance. In this context the following objectives are established for proposed investigation and implementation work:

1. **To investigate the EDM and the techniques of data mining:** the work includes the study of recently published articles related to EDM and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

data mining that help to improve the quality of education.

2. **To design an student performance prediction and study material recommendation model:** by using the different data mining techniques and method in this phase a EDM model is proposed that predict the student's performance and also recommend the study material to improve the student learning.
3. **To evaluate the performance of the proposed model with a relevant prediction model:** In this phase the proposed work is evaluated and compared with the traditional model to demonstrate the effectiveness of the model.

II. PROPOSED WORK

The proposed work is motivated to design a prediction model which allows the students to get feedback about the upcoming performance. Additionally this model also recommends the relevant study material according to their learning behavior. Therefore a student performance prediction and course material recommendation model is proposed for design and implement. This chapter provides the detailed understanding about the design of the proposed model.

A. System Overview

The data mining is a technique of analyzing large and complex data to providing the application centric patterns which can be used for grouping the data, predicting patterns, recognizing the patterns and mining the relationship among two patterns. These techniques are reducing the human efforts by automating the data analysis process.Therefore a wide range of applications are accepting the goodness of these data mining techniques, among them the educational domain is also one of the most essential area where these techniques are getting benefits of data mining. The application of data mining in educational data analysis is also known as the EDM (educational data mining). There are various prospective and applications are feasible for mining educational data. Therefore this work is including the data mining with the student performance prediction and recommendation of material to study.

The proposed EDM model includes two different data models and some manual efforts for performing the required task. Therefore first an online dataset for student performance prediction is obtained and then the clustering algorithm is employed for categorizing the data into three performances categorizes i.e. high, medium and low. After that a decision tree algorithm is applied for training with this categorized database. The trained decision tree algorithm is first validated with a test dataset and then by providing custom inputs by the model predicts the performance of student additionally during this outcome the system recommends the study material. The study material is collected and categorized manually to support the student learning. This section provides the overview of the proposed working model and the next section involve the detailed design of the proposed model.

B. Methodology

The proposed work is subdivided into two main parts first the performance prediction of students and the next is to recommend the study material. Both the modeling can be described using system architecture as demonstrated in figure 2.1.

Student dataset: the data mining models requires the initial training samples for learning or analysis. In this presented work the student performance data were used as the learning sample. This dataset is collected from the online machine learning dataset repository namely UCI repository. This dataset contains 650 student records (instances). Additionally the dataset not includes any target outcome or predefined classes. The dataset includes 33 features or attributes. This dataset is used for the experimentation in this work.

Data pre-processing: the datasets may contain the noise and unwanted information which might influence the performance and learning of the classifier. Therefore the pre-processing techniques are used for optimizing the quality of learning data. In this presented work the used dataset consist of text and numerical attribute values thus in pre-processing the dataset is transformed into numerical values. Therefore the categorical attributes are mapped into unique numerical numbers.

Dataset Splitting: after data pre-processing all the student performance dataset is converted into numerical values which are split into two parts i.e. training and testing datasets. The training set is further used for preparing the decision tree and test set is used for validation of trained data model.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

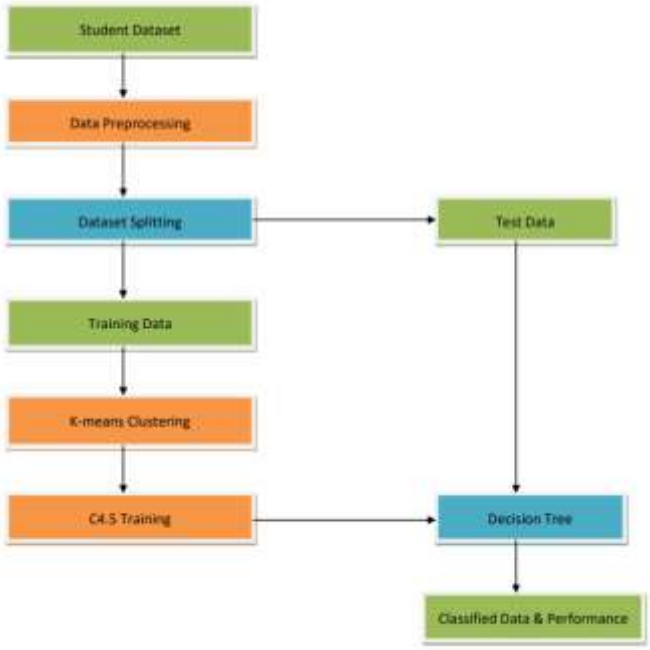

Figure 2.1 System Architecture

Training set: the training dataset contains the 70% of randomly selected data instances. This set of data is further used with the data mining algorithm to prepare the trained model for classifying or predicting the student's performance.

Test dataset: the test dataset is also created using the 30% of randomly selected data instances. That set of data is used for validating the trained decision tree classifier. After validation if the trained model performs well then we can use this classifier for predicting the student performance.

K-means clustering: That is an unsupervised learning algorithm. That is used to solve the clustering problems in data science. The K-Means Clustering algorithm groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that are needed to be created, for example if K=2 then two clusters will be created. It allows discovering the categories of the unlabeled dataset without any training. Each cluster of data is associated with a centroid. The aim is to minimize the sum of distances between the data point in corresponding clusters.

The algorithm takes unlabeled data as input, and divides the dataset into k-number of clusters. That iterates until it does not find the best clusters. The value of k should be provided as input in this algorithm. The k-means algorithm performs two tasks:

- Determines the best value for K centroids by iterative process.
- Assigns each data point to its closest centroids to create clusters

The working of the K-Means algorithm is explained as follows:

1. Select K number of clusters.
2. Select random K points as centroids.
3. Assign each data point to their closest centroid.
4. Calculate the variance and place a new centroid of each cluster.
5. Repeat the third steps, which reassign each data point to the new closest centroid.
6. If any reassignment occurs, then go to step 4 else FINISH

Here the input dataset is unlabelled thus in order to train the supervised classifier we need to assign some predefined class labels. Therefore the input training data is clustered into three categories namely high, low and medium. Thus we put k=3 for performing the clustering and dividing the entire dataset into three categories. The categorized dataset are further added to the training data samples.

C4.5 Decision Tree: The C4.5 algorithm is a Decision Tree Classifier which can be employed to generate a decision, based on a certain sample of data. Before understanding of C4.5, let's discuss a bit Decision Trees and how they can be used as classifiers. A Decision Tree is something like a flowchart. To be more concise if you know an event is very probable, it is no surprise when it happens, that is, it gives you information that it actually happened. From this statement we can

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

formulate that the amount of information gained is inversely proportional to the probability of an event happening. We can also say that the Entropy increases the information gain decreases. This is because Entropy refers to the probability of an event.

$$E(X) = -\sum_{i=1}^{n} P(x_i)\log_b P(x_i)$$

We say that the entropy is minimum because without any sort of trials we can predict the outcome of an event. In the case is Decision Trees, it is essential that the node are aligned as such that the entropy decreases with splitting downwards. This basically means that the more splitting is done appropriately, coming to a definite decision becomes easier. So, we check every node against every splitting possibility. Information Gain Ratio is the ratio of observations to the total number of observations $(m/N = p)$ and $(n/N = q)$ where $m + n = N$, $m + n = N$ and $p + q = 1, p + q = 1$. After splitting if the entropy of the next node is lesser than the entropy before splitting and if this value is the least as compared to all possible test-cases for splitting, then the node is split into its purest element. In the Decision Tree when the dataset is huge and there are more variables to take into consideration. This is where Pruning is required. Pruning refers to the removal of those branches in decision tree which we feel do not contribute significantly in decision process. The concept of Pruning enables us to avoid over fitting of the regression or classification model so that for a small sample of data, the errors in measurement are not included while generating the tree.

Algorithm steps:

1. Check for the base cases.
2. For each attribute a, find the normalized information gain from splitting on a.
3. Let a_{best} be the attribute with the highest information gain.
4. Create a decision node that splits on a_{best}.
5. Recur on the sub-lists obtained by splitting on a_{best}, and add those nodes as children of node.

Advantages of C4.5:

1. The algorithm inherently employs Single Pass Pruning to mitigate over-fitting.
2. It can work with both Discrete and Continuous Data
3. C4.5 can handle the issue of incomplete data very well

Decision Tree: After the assignment of initial classes to the data instances the dataset is modified which is being used with the decision tree C4.5. The decision tree algorithm generates a tree during the training process. Additionally after training the test dataset is used to validate the learned data model.

Classified Data and Performance: the input test dataset is consumed with the trained model to predict the class label of each test set instances. Using these predicted class labels and the actual class labels the performance of the trained model is measured.

C. Course Material Recommendation System

The proposed course material recommendation model is demonstrated in figure 2.2. In this model the previous stage trained C4.5 decision tree rules are used for predicting the individual student's performance. The student's properties are being used to traverse the decision tree to get the decision node. Based on the obtained decisions the student's performance is approximated. According to the decisions of student's performance if the student performance found higher than the system recommend the complex study material for those students. Further if the student performance found average than the average levels of study material has been recommended, and finally if the prediction is low than the basic level of study material has been recommended.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

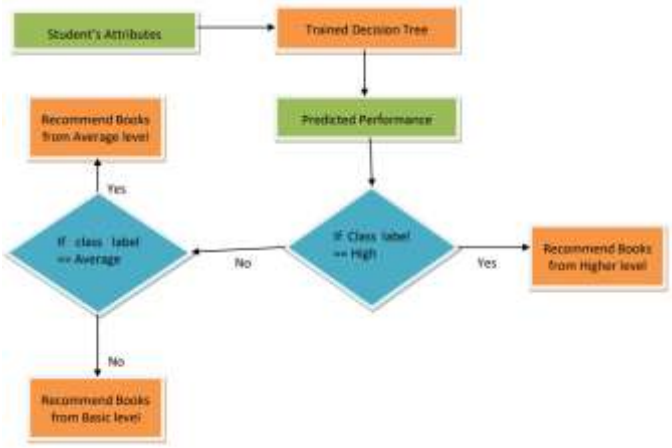

Figure 2.2 course material recommendation

D. Proposed Algorithm

The above given both the data models are combined in this section to predict the student's performance and the relevant study material. The summarized steps of both the process are explained in table 2.1.

Table 2.1 proposed algorithm

Input: student dataset D, Student's attributes A **Output:** student performance P, recommended study material S, Validation Performance V
Process:
1. $R_n = \text{readDataset}(D)$
2. $P_n = \text{preProcessData}(R_n)$
3. $[Cl, \text{centroid}] = \text{kMeans.CreateCluster}(k = 3, P_n)$
4. $[\text{Train}, \text{Test}] = \text{Cl.Split}(70\%, 30\%, \text{Random})$
5. $T_{model} = \text{J48.CreateTree}(\text{Train})$
6. $V = T_{model}.\text{Classify}(\text{Test})$
7. $P = T_{model}.\text{Classify}(A)$
8. if(P == High)
a. S = complex Material
9. else if(P == AVG)
a. S = Avg Material
10. else if(P == Low)
a. S = basic Material
11. End if
12. Return S, P, V

III. RESULTS ANALYSIS

This chapter provides the details about the experimental analysis carried out in order to predict the student performance using the decision tree algorithm. In this context the different performance parameters are measured and reported in this chapter.

A. Accuracy

The accuracy of a predictive model demonstrates the correctness of the model for predicting the measured values. That can be measured using the following formula:

$$\text{Accuracy}(\%) = \frac{\text{correctlypredictedinstances}}{\text{totalinstancestopredict}} X100$$

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

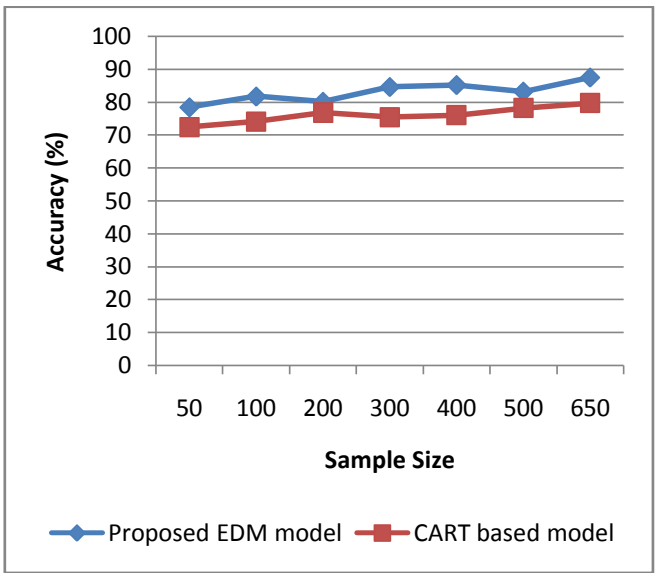

Figure 3.1 Accuracy (%)

The accuracy of the proposed EDM model is described in figure 3.1 and table 3.1. The table consists of the observations of experiments of both the models i.e. traditional and proposed EDM model for predicting the performance of student. Additionally the line graph visualizes the performance of experiments. The X axis of the line graph shows the samples size for experiment and Y axis shows the performance of the algorithm in terms of accuracy. The accuracy of the model has been measured in terms of percentage. Accuracy of the model demonstrates the superiority over the traditional technique of student performance prediction.

Table 3.1 Accuracy (%)

Sample size	Proposed EDM	CART based
50	78.5	72.5
100	81.8	74.2
200	80.2	76.9
300	84.7	75.5
400	85.3	76.1
500	83.3	78.3
650	87.6	79.8

B. Error Rate

The error rate of a data mining model describes the misclassification rate of a data model. In other words the error rate is the indicator of error possibility in prediction. That can be measured using the following equation:

$$\text{ErrorRate}(\%) = \frac{\text{MisclassifiedSamples}}{\text{totalsamplestoclassify}} \text{X}100$$

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

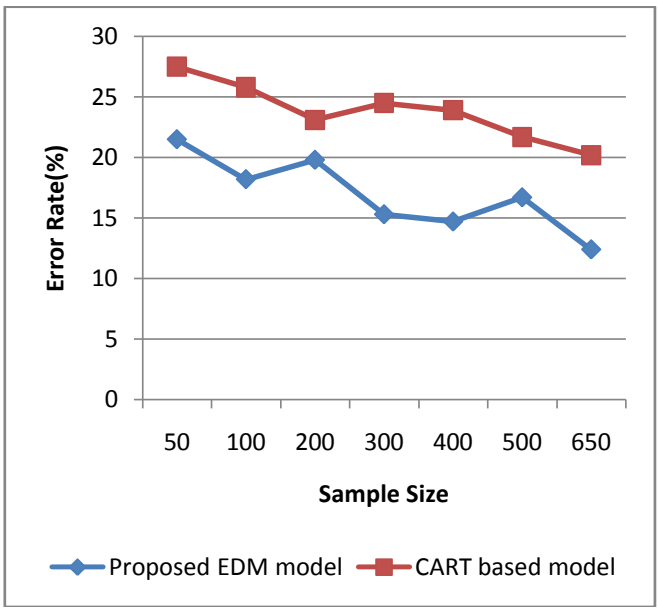

Figure 3.2 Error Rate (%)

The error rate of the proposed data model for predicting the performance of the student is reported in this section in terms of percentage. The observation of error rate percentage is given in table 3.2 and the line graph of this observation is given in figure 3.2. As similar to the previous line graph of accuracy, this line graph also contains sample size in X axis and Y axis contains the obtained error rate. According to the obtained results the proposed data model of student performance prediction results less error in prediction as compared to the traditional CART based performance prediction model. Therefore the proposed model is superior to the classical data model based on CART algorithm.

Table 3.2 Error Rate (%)

Sample size	Proposed EDM model	CART based model
50	21.5	27.5
100	18.2	25.8
200	19.8	23.1
300	15.3	24.5
400	14.7	23.9
500	16.7	21.7
650	12.4	20.2

C. Memory Usage

The memory usage is indicator of the space complexity of the executed algorithm. That demonstrates the amount of main memory utilized during the algorithm execution. In JAVA technology that is measured using the following formula:

$$\text{memoryusage} = \text{totalalloted} - \text{freememory}$$

Table 3.3 Memory Usage

Sample size	Proposed EDM model	CART based model
50	11022	10352
100	11271	10283
200	11524	10546
300	11722	10682
400	11996	10828
500	12173	10587
650	12489	10822

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

The memory usages of both the data models for predicting the student performance is described using figure 5.3 and table 5.3. The X axis of the diagram shows the sample size of experiment and the Y axis shows the corresponding memory usages of the algorithms. The results demonstrate the proposed model consumes higher memory as compared to treditional CART based model.

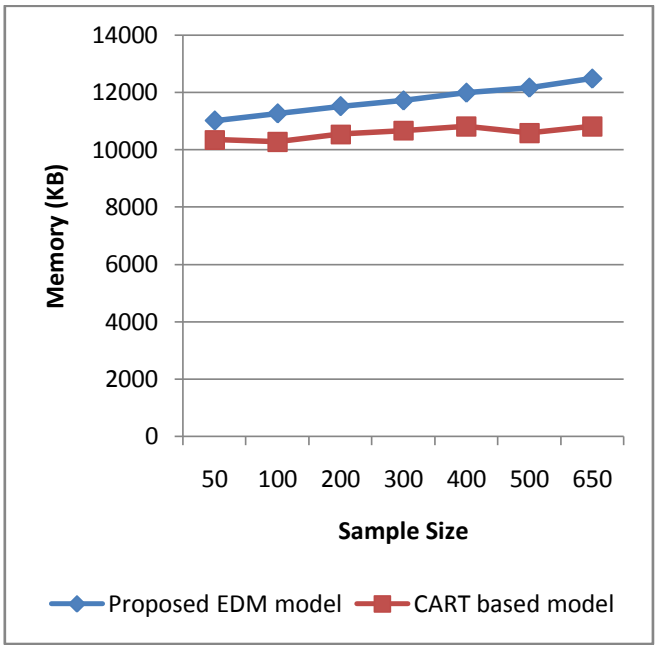

Figure 3.3 memory usages

D. Time consumed

The time consumption of an algorithm shows the required to complete the execution process. That can be difference between the algorithm initialization (T_i) and completing the execution (T_e). That can be represented using the following formula:

$$T = T_i - T_e$$

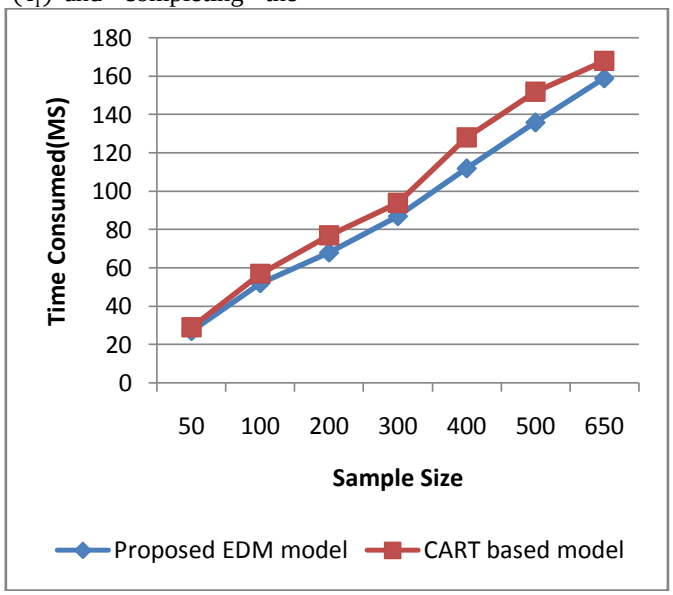

Figure 3.4 Time consumed

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1844-1854 www.ijaem.net ISSN: 2395-5252

The time consumption of the proposed model and the CART based student performance prediction model is demonstrated in figure 3.4 and table 3.4. The X axis of the diagram shows the experimental data sample size and Y axis of the model shows the time consumed. The Red line shows the performance of CART based model and Blue line shows the performance of proposed EDM model. The time consumption of the data models were measured in terms of milliseconds. According to the experimental results the performance of the proposed model demonstrates efficient results as compared to the traditional CART based model.

Table 3.4 Time consumed

Sample size	Proposed EDM model	CART based model
50	27	29
100	52	57
200	68	77
300	87	94
400	112	128
500	136	152
650	159	168

IV. CONCLUSIONS AND FUTURE WORK

The aim of the proposed work is to contribute in the domain of EDM by designing a data model for predicting the student's performance. Additionally based on the performance we have tried to recommend the relevant study material for student's use. This chapter provides the summary of the work carried out and also the future extension of the work is also suggested.

A. Conclusion

The EDM (Educational Data Mining) is a subject of data mining where the academic data is used with the data mining algorithms and techniques to support various educational activities. These activities not only include the learning and teaching, it may also involve various other management and sustainable educational activities for improving the quality of education. However there are a number of applications exist on educational domain using data mining techniques among them student performance prediction is one of the essential applications of EDM domain. The student performance prediction technique supports the students to improve their learning activity by providing feedback about their performance and possible future improvements.

In this context the proposed work is focused on designing a student performance prediction system using data mining techniques. This system also aimed to recommend the course material. In this context the proposed work is involve an data model which usages the student database as input for learning process. That base is developed with the help of historical student record. First the data is clustered using the k-means clustering algorithm, which categorize the data into three classes, i.e. low, medium and high performance of students. In next step the C4.5 or J48 decision tree algorithm is trained over the categorized dataset. The trained learning algorithm first tested using a validation dataset and then used for predicting the student's performance individually.

The implementation of the proposed EDM model has been carried out using JAVA technology. Additionally to implement the data mining algorithms the WEKA library has been used, and to store the performance observations the MySQL data base in used. The experimental analysis is conducted and the obtained observations are summarized in the table 4.1.

Table 4.1 performance summary

S. No.	Parameters	Proposed EDM	CART based
1	Accuracy	78.5-87.6	72.5-79.8
2	Error rate	12.4-21.5	20.2-27.5
3	Memory	11022-12489	10352-10822
4	Time	27-159	29-168

According to the performance summary as given in table 4.1 the proposed model found efficient and accurate for student performance prediction. Therefore that is a promising model

which can be providing extensive support for EDM domain for automation and new generation educational needs.

B. Future work

The established objectives of the proposed work have been accomplished successfully. The proposed data model is found promising and we can extend this model by incorporating the following features in the existing data model.

1. The proposed model is just identify the student's performance and based on this performance model suggest the study material in near future it is required to evaluate the course material complexity for suggesting the material automatically
2. The current approach include the single learning algorithm for predicting the student performance but the ensemble learning can increase the predictive performance significantly thus we need to include the ensemble learning approach for future model design

REFERENCES

[1] R. Jindal, M. D. Borah, "A Survey on Educational Data Mining and Research Trends", International Journal of Database Management Systems (IJDMS) Vol.5, No.3, June 2013

[2] R. Ferguson, "Learning analytics: drivers, developments and challenges", International Journal of Technology Enhanced Learning, 4(5/6) pp. 304–317.

[3] G. J. Hwang, H. C. Chu, C. Yin,"Objectives, methodologies and research issues of learning analytics, Interactive Learning Environments", 25:2, 143-146, DOI: 10.1080/10494820.2017.1287338

[4] R. Asif, A. Merceron, S. A. Ali, N. G. Haider, "Analyzing undergraduate students' performance using educational data mining", Computers & Education 113 (2017) 177-194

[5] J. Han, J. Pei, M. Kamber, "Data mining: concepts and techniques", Elsevier, 2011.

[6] B. M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering, Volume 1 Number 4, pp. 301-305

[7] M. H. Dunham, "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, New Delhi, 2003. Print. ISBN: 81-7758-785-4, 2006.

[8] K.Mehmed, "Preparing the Data." Data Mining: Concepts, Models, Methods, and Algorithms", Second Edition (2003): 26-52.

[9] N. Jain, V. Srivastava, "Data Mining Techniques: A Survey Paper", IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11, Nov-2013.

[10] S. Sumathi, S. N. Sivanandam, "Introduction to data mining and its applications", Volume 29, Springer, 2006

[11] R. S. Petre, "Data mining in cloud computing", Database Systems Journal 3.3 (2012): 67-71.

[12] F. Sebastiani, "Machine learning in automated text categorization", ACM Computing Surveys, Volume 34, Number 1, 2002, pp. 1–47

[13] S. B. Navathe, E. Ramez, "Data Warehousing and Data Mining", in "Fundamentals of Database Systems", Pearson Education pvt Inc., Singapore, 2002, 841-872.

[14] A. Algarni, "Data Mining in Education", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016

[15] U. K. Sen, "A Brief Review Status of Educational Data Mining", international Journal of Advanced Research in Computer Science & Technology (IJARCST 2015), Vol. 3, Issue 1 (Jan. - Mar. 2015)

[16] C. Romero, S. Ventura, "Educational Data Mining: A Review of the State-of-the-Art", IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, VOL. XX, NO. X, 200X

[17] L. C. Liñán, Á. A. J. Pérez, "Educational Data Mining and Learning Analytics: differences, similarities, and time evolution", RUSC. Universities and Knowledge Society Journal, Vol. 12, Núm. 3 (juliol 2015)

[18] V. Thanuja, B. Venkateswarlu, G. S. G. N. Anjaneyulu, "Applications of Data Mining in Customer Relationship Management", J. Comp. & Math Sci. Vol.2 (3), 423-433 (2011)

[19] Ms. T. K. Anusuya, P. Yasotha, "Estimation of Student Performance Using Artificial Intelligence with LMS Strategies", Journal of Information and Computational Science, Volume 10 Issue 3 – 2020

[20] Suhirman, J. M. Zain, H. Chiroma, T. Herawan, "Data Mining for Education Decision Support: A Review", iJET – Volume 9, Issue 6, 2014

[21] S. Angra, S. Ahuja, "Analysis of Student's Data using Rapid Miner", Journal on

Today's Ideas –Tomorrow's Technologies, Vol. 4, No. 2, December 2016 pp. 109-117

[22] J. Srivastava, Dr A. K. Srivastava, "Data Mining in Education Sector: A Review", Special Conference Issue: National Conference on Cloud Computing & Big Data

[23] E. Gomede, F. H. Gaffo, G. U. Briganó, R. M. d. Barros, L. d. S. Mendes, "Application of Computational Intelligence to Improve Education in Smart Cities", Sensors 2018, 18, 267; doi:10.3390/s18010267

[24] F. Martino, A. Spoto, "Social Network Analysis: A brief theoretical review and further perspectives in the study of Information Technology", PsychNology Journal, 2006, Volume 4, Number 1, pp. 53 – 86

[25] Meenakshi and Geetika, "Survey on Classification Methods using WEKA", International Journal of Computer Applications, Vol. 86, No.18, January 2014.

[26] S. Archana, Dr. K. Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Volume 2 Issue 2, February 2014.