EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance in Academics

Mukesh Kumar, Gaurav Mehta, Nandini Nayar, Abha Sharma

Department of Computer Science and Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Himachal Pradesh, India
mukesh.kumar@chitkarauniversity.edu.in, gaurav.mehta@chitkarauniversity.edu.in, nandini.nayar@chitkarauniversity.edu.in, abha.sharma@chitkarauniversity.edu.in

Abstract: Data Mining is a field in which hidden information is extracted from a large database by using some algorithms implementation. These algorithms are further divided into some categories like classification, clustering, association rule mining etc according to information we want to extract. Data mining is a field which is widely spread over different areas like telecommunication, marketing, operation, hospitals, hotel industry, education etc. Predicting the academic's performance and progress of the students has revealed the attention of the young researchers. To facilitate the task of building an academic prediction model, historical student academic dataset is used. In this paper, the contributions are exhibited in two different folds. In the first fold, the main aim is to build the prediction model by different families of the Machine Learning Techniques on the selected dataset for consideration. In the second fold, implementations of different ensemble meta-based model are presented by combining with different classification algorithms of Machine Learning Techniques. Different ensemble meta-based model taken into consideration for implementation are Bagging, AdaBoostM1, RandomSubSpace. The implementation results demonstrate that the ensemble meta-based technique (AdaBoostM1) gained a superior accuracy performance with MultilayerPerceptron Machine Learning technique reaching up to 80.33%.

Index Terms: Educational Data Mining, NaiveBayes, MultilayerPerceptron, Locally Weighted Learning, DecisionTable, J48

1. Introduction

Data Mining is a field in which hidden information is extracted from a large database by using some algorithms implementation. These algorithms are further divided into some categories like classification, clustering, association rule mining etc according to information we want to extract. Data mining is a field which is widely spread over different areas like telecommunication, marketing, operation, hospitals, hotel industry, education etc. Data mining in the field of education is known as Educational data mining (EDM). In EDM, data related to student's are analysed and some useful information is extracted from that which help us to improve the education system and hence the overall educational development of the student.

Due to the increased use of digital data storage system throughout the world, a lot of data is generated in the education system. The student's data may be in the form of academic, social-economic, family, personal, institutional attributes. Today's young researchers are working on the areas of educational data mining and try to develop a new model which successfully analyse student's data more accurately and implement it to improve the education system. The row data which are stored in the educational system using regular process do not have any importance in future if that data is not analysed properly to get some useful information from that. By means of Educational Data Mining tools, we are able to understand the student's data in a better way which is not readable by a human, which may have a big impact on the educational system.
EDM is used in many areas of education like predicting student academic performance, educational dropout, student behaviour, student's placement, student academic progress etc. In this paper, the main focus is on predicting student academic performance in education. By predicting the academic performance of the student's, timely help will be provided to the needy student's. At present, lots of technical institutions in India are shorts of student admission and if they will not taking care of their educational process then there will be a danger of shutdown. So it's critical for every institution to properly analyse the student academic data and find out some good information from that which further used to enhance the academic process in future.

By applying some data mining techniques on student dataset, we can find out some good information. These data mining techniques are classification, clustering and association rule mining which are implemented on the student dataset. The student's data analyse enhance the progress of the student's and hence overall performance and learning environment of the university improves. These classifiers model provide different result on a different dataset. A good classifier may provide a bad result on some dataset. So it's the responsibility of the data analyst that, the select a classification algorithm which is suitable for their dataset.

2. Literature Survey:

In the past, numerous prediction models have emerged that enhance the students’ overall academic performance. These emerged models are either focus on the effect of different student's attributes on academic performance, predict the academic future of the students, predict the placement, predict the educational dropout etc. Therefore, this below-mentioned section discusses the two aspects of research in EDM:

**The important student's factor which affects or predict the academic performance:** In the literature review, we can across so many students attribute which helps in predicting student's academic performance. These student's attributes are from an academic background, family-economic attributes, social attributes, institutional attributes and personal attributes. All these listed attributes are extracted from the student databases which are maintained by the institutional authority at a different level. So before implementing data mining algorithms, a well-prepared dataset is created according to what types of result we want to predict. Different researcher in the domain of EDM investigated different student's attributes which influence the academic performance of the student's at a different level. The student's attributes that are collected during the academic period of the students are internal and external marks of the students, attendance of the students, demographics attributes, cumulative grade point average, extra-curricular activities, school background, institutional type, external assessments which includes the behaviour of student outside the class etc. Table-1 given below explains about different attributes which are taken into consideration by the different researcher during their research work in the domain of educational data mining:

| Student’s Attribute Types | Description of the Student’s Attributes                                                      |
|--------------------------|------------------------------------------------------------------------------------------------|
| Student’s Percentage/CGPA| Cumulative grade point average/ Student Marks / Percentage                                    |
| Student’s Attributes     | Demographic attributes such as age, sex, family status, education level, income, occupation, and race. |
| Pre-admission Attributes | Student’s attributes such as high school grades, General Aptitude Test score, Scholastic Achievement Admission Test scores, high school background |
| Internal-Assessment Attributes | Student's attributes such as class attendance, internal marks, internal lab tests, class tests etc. |
Prediction Methods which predict students’ academic performance: An imperative task for predicting students’ academic performance is to develop a superior classifier model by using different classification algorithms. In data mining, there are lots of families of classification algorithms build. At the time literature review, we can across such different algorithms which gave different types of accuracy according to the selected datasets for predicting students’ academic performance. Prominent classification algorithms for predicting student’s academic performance by different researchers are summarized in Table 2.

Table 2: Common Classifiers used in predicting Student's Academic Performance

| S.N | Author’s | Paper Title | Techniques used |
|-----|----------|-------------|-----------------|
| 1   | Ammar Almasri, et. al. | “EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance” | EMT based classification Model: Bagging, Boosting, Adaboost, PART, A2DE, J48, Multilayer perceptron, LocalKNN |
| 2   | Mudasir Ashraf, Majid Zaman et. al. | “Using Predictive Modeling System and Ensemble Method to Ameliorate Classification Accuracy in EDM” | Boost with J48, Boost with Random tree, Boost with NaiveBayes, Boost with KNN |
| 3   | A. Daud, N. R. Aljohani | “Predicting student performance using advanced learning analytics” | BN, c4.5, NB, SVM, and CART |
| 4   | C. Romero, M.-I. L’opez, et. al. | “Predicting students’ final performance from participation in online discussion forums” | BayesNet, SMO and NaiveBayes with the highest accuracy |
| 5   | S. Natek and M. Zwilling | “Student data mining solution knowledge management system related to higher education institutions” | REPTree, J48 |
| 6   | S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman | “Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique” | Neural network, Naive Bayes classification with SMOTE technique yielded maximum accuracy as compared to the Decision tree. |
| 7   | K. Bunkar, U. K. Singh, B. Pandya, and R. Bunkar | “Data mining: prediction for performance improvement of graduate students using classification” | C4.5, ID3 and CART were implemented. The result demonstrates that C4.5 gave the superior accuracy |
8. G. Gray, C. McGuinness, and P. Owende
   “An application of classification models to predict learner progression in tertiary education”
   Naive Bayes, Logistic Regression, Decision Tree, Support Vector Machine, k-nearest neighbour, Neural Network. Result shows that KNN as compared to other algorithms implemented.

9. T. Mishra, D. Kumar, and S. Gupta
   “Mining student’s data for prediction performance”
   J48 and Random Tree

10. T. M. Christian, M. Ayub
    “Exploration of classification using NBTree for predicting students’ performance”
    NBTree classification algorithm gave the maximum accuracy as compared to other implemented algorithms.

11. M. Mayilvaganan and D. Kalpana Devi
    “Comparison of classification techniques for predicting the performance of students academic environment”
    C4.5, AODE, Naive Bayesian, Multi-Label K-Nearest Neighbor algorithms. Multi-Label k-Nearest Neighbor gave maximum accuracy among other algorithms.

12. E. Osmanbegovi´c and M. Sulji´c
    “Data mining approach for predicting student performance”
    Multilayer Perceptron, C4.5 and Naive Bayes algorithms were implemented. Results demonstrate that Naive Bayes gave the maximum prediction accuracy.

13. A. M. Ahmed, A. Rizaner, and A. H. Ulusoy
    “Using data mining to predict instructor performance”
    J48 Decision Tree, Naive Bayes, SMO, and Multilayer Perceptron were implemented. The result shows that the J48 Decision Tree algorithm achieves the maximum accuracy as compared to other algorithms implemented.

14. S. Kalaivani, B. Priyadharshini, and B. S. Nalini
    “Analyzing student’s academic performance based on data mining approach”
    Decision stump and J48 classification algorithms were implemented and the result shows that J48 provides maximum accuracy on the taken dataset.

15. G. Ayyappan, K. SivaKumar
    "A Noval approach of Ensemble Models by using EDM" 
    BayesNet, Compliment NaiveBayes, NaiveBayes, Adaboost M1, Bagging, Dagging, ConjuctiveRule, Decision Table, JRip, BFTree, Decision Stump, J48 were the classification algorithms used and the result shows that the maximum accuracy 72.18% had AdaBoostM1 with BayesNet.

In EDM, to improve the overall accuracy of any classification algorithms we have some ensemble techniques. These ensemble techniques are bagging, boosting, Random Forest and voting. We further proceed by taking the two most important questions in our mind which are mentioned below:
How different classification algorithms of data mining are valuable for predicting students’ academic performance and explain how the implemented algorithms improve the academic success of the new scholars enrolled in the institution. How the implemented classification algorithms performance are improved by using different EMT classification algorithms.

Environment setup and Ensemble Meta-Based Tree Model:
Dataset Description: The considered dataset for this study has been taken from Kalboard-360, which is a Multi-agent Learning Management System. In this technological era, such an online learning platform provides the user with unlimited access to educational resources from several places and on any device which uses an internet connection. The dataset consists of 16 features and 480 records of students, in which 305 (Male) and 175 (Female) students. These features can be classified into three major categories namely academic background, demographic as well as behavioral features. All 16 attributes are considered for the analysis on the above-mentioned. More information about dataset is taken from https://www.kaggle.com/aljarah/xAPI-Edu-Data link.

Data Mining Tool and Environment Setup: WEKA tool kit is a collection of different Machine Learning algorithms which are further used to solve different types of problems like classification, clustering, association rule mining of real-world. This tool is written in Java programming and almost run on any platform. On this platform, you may directly apply algorithms on your dataset or you may directly call dataset from your java code. It contains tools for numerous tasks like pre-processing, features selection, classification, visualization, regression as well as clustering. It is also pertinent for developing new Machine Learning algorithms. WEKA is an “Open-source” Machine Learning software which is freely available to use for analysis and can be easily used via graphical user interface, standard terminal applications /Java API. This tool is extensively used in a different area of research, teaching and industrial application and gives transparent access to toolboxes like sci-kit-learn R, and Deep learning. WEKA was primarily designed for analyzing data from agricultural domains. However, the recent version is primarily used for education as well as research. Numerous benefits of Weka include: It’s freely available under the “GNU General Public License”. Moreover, it is entirely implemented in Java programming language and is able to run on almost every computing platform. It is easily used due to its graphical user interfaces. Below figure. 1 shows that the dataset is uploaded on the WEKA interface through explorer.

![Figure 1: Dataset uploaded on the WEKA tool kit](image-url)

Below figure. 2 show that how different classification algorithms is used with bagging, AdaBoostM1 and RandomSubSpace (ensembling techniques).
3. Results and Discussion:

The implementation result conducted is used to answer the research questions, which ensures that the proposed algorithm is more effectual for predicting academic performance of the students. As we know that is main issues concerning to any country’s development is the education system over there. The key goal of this paper is to present an organized analysis of numerous classification algorithms for predicting the students’ academic performance. This analysis will be beneficial for the institutional / college authorities, for identifying the weak students that can further help in enhancing their performance and improve the learning, and students’ growth that also reflects on the students’ background.

**Prediction Model Analysis:** Most of the researcher in recent research papers found the most appropriate attributes which have made a significant influence on the student’s academic performance. The main question here is to find effect of these attributes on the classification algorithms. So, in this section, five classification algorithms were implemented using the selected dataset. The implemented classification algorithms are NaiveBayes, MultilayerPerceptron, Locally Weighted Learning, DecisionTable and J48.

| Classification Algorithms       | Accuracy  | Mean Absolute Error | Root Mean Squared Error | Relative Absolute Error | Root Relative Squared Error |
|--------------------------------|-----------|---------------------|-------------------------|-------------------------|-----------------------------|
| NaiveBayes                     | 67.70%    | 0.2251              | 0.397                   | 52.00 %                 | 85.33 %                     |
| MultilayerPerceptron           | 78.33%    | 0.151               | 0.349                   | 34.88 %                 | 75.01 %                     |
| Locally Weighted Learning      | 66.04%    | 0.3265              | 0.3968                  | 75.41 %                 | 85.29 %                     |
For table 3, it is very much clear that MultilayerPerceptron algorithm is performed well as compared to other algorithms implemented. The classification accuracy of MultilayerPerceptron algorithm is approximately equal to 78.33% as compared to J48 (75.83%), DecisionTable (72.70%), NaiveBayes (67.70%) and Locally Weighted Learning (66.04%). As we can see those other performance parameters like Mean Absolute Error, Root Relative Squared Error, Root Mean Squared Error and Relative Absolute Error gave the best result with MultilayerPerceptron algorithm. So, overall according to our implementation result, we can say that the above mention algorithm (MultilayerPerceptron algorithm) perform exceptionally well as compared to other classification algorithms. Figure 3 demonstrates the Graphical representation of Classification algorithms with performance parameters taken into consideration.

| Algorithm     | Accuracy | Mean Absolute Error | Root Relative Squared Error | Relative Absolute Error | Root Mean Squared Error | Mean Absolute Error | Accuracy |
|---------------|----------|---------------------|-----------------------------|-------------------------|------------------------|--------------------|----------|
| DecisionTable | 72.70 %  | 0.2953              | 0.3718                      | 68.19 %                 | 79.92 %                | 72.70 %            |          |
| J48           | 75.83 %  | 0.2155              | 0.3632                      | 49.77 %                 | 78.07 %                | 75.83 %            |          |

**Figure 3: Graphical representation of Classification algorithms with performance parameters**

**Ensemble Meta-Based Tree Model:** The academic performance monitoring of the student’s in any institution indicate to student’s parents that the institution is taken the student performance as a major concern. The importance of implementing Machine Learning algorithms, exploring the chronological data of the students to envisage their future academic performance, has gained massive attention that motivated us to develop a model for predicting the “unknown labels” of future instances. Proposed algorithms are ensemble algorithms which further combines the selected classification algorithms to build a final prediction model. The proposed methodology is used to demonstrate how ensembling algorithms improved the predicting performance in terms of different performance parameters taken into consideration. The different ensembling algorithms implemented were bagging, AdaBoostM1 and RandomSubSpace with selected classification algorithms. In Table 5, it is very much clear that the ensembling algorithms improve the envisaging performance of the classification algorithms. The prediction accuracy is improved from 67.70% to 72.29% (AdaBoostM1 + NaiveBayes), 66.04% to 72.91% (AdaBoostM1 + Locally Weighted Learning), 72.70 % to 76.25% (AdaBoostM1 + DecisionTable, Bagging + DecisionTable), 75.83 % to 77.91% (AdaBoostM1 + J48).
Table 4: Implementation results of Classification with Ensembling algorithms

| Classification Algorithms     | Accuracy without Ensembling | Accuracy with Bagging | Accuracy with AdaBoostM1 | Accuracy with RandomSubSpace |
|------------------------------|-----------------------------|-----------------------|--------------------------|-----------------------------|
| NaiveBayes                   | 67.70%                      | 67.70%                | 72.29%                   | 69.1667 %                   |
| MultilayerPerceptron         | 78.33%                      | 78.12%                | 80.33%                   | 76.4583 %                   |
| Locally Weighted Learning    | 66.04%                      | 70.20%                | 72.91%                   | 71.6667 %                   |
| DecisionTable                | 72.70%                      | 76.25%                | 76.25%                   | 73.9583 %                   |
| J48                          | 75.83%                      | 74.37%                | 77.91%                   | 76.875 %                    |

For table 4, it is very much clear that MultilayerPerceptron algorithm is performed well as compared to other algorithms implemented. The classification accuracy of MultilayerPerceptron algorithm is approximately equal to 80.33% as compared to j48 (75.83%), DecisionTable (72.70%), NaiveBayes (67.70%) and Locally Weighted Learning (66.04%). So, overall according to our implementation result, we can say that the above mention algorithm (MultilayerPerceptron algorithm) perform exceptionally well as compared to other classification algorithms. Figure 4 demonstrates the Graphical representation of Classification with Ensembling algorithms.

![Graphical representation of Classification with Ensembling algorithms](image)

Figure 4: Graphical representation of Classification with Ensembling algorithms

From figure 4, the prediction accuracy of multilayer perceptron (78.33%) is superior in comparison to other classification algorithms.

4. Conclusion and Future Work:

In Educational data mining, we use a data analytical tool that exploited and explore the new information from the historical academic dataset. The data analytical tools convert that academic dataset which is not in human-readable form into the interpretable form which may have a huge impact on the educational research purpose. On the basis of 10-fold cross-validation method, a
comparison among various “Machine Learning” techniques is conducted. The set of evaluation metrics are considered for evaluating these techniques comprise of: Accuracy, Mean Absolute Error, Root Relative Squared Error. Root Mean Squared Error and Relative Absolute Error. The results demonstrate that the maximum accuracy given by different are MultilayerPerceptron (78.33%), NaiveBayes (67.70%), Locally Weighted Learning (66.04%), DecisionTable (72.70%), J48 (75.83%), which ensure the validity of the first research question. Moreover, the implementation results tend to increase the performance of the selected classifier by applying the ensemble techniques (Bagging, AdaBoostM1 and RandomSubSpace). The outcomes demonstrate a noteworthy augmentation utilizing the proposed EMT model algorithm. It has been demonstrated that the proposed approach can yield superior accuracy (up to 80.33%).

References:

[1] Ammar Almasri, Erbug Celebi, Rami S. Alkhawaldeh, 2019 EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance, Hindawi Scientific Programming Volume 2019, Article ID 3610248, 13 pages https://doi.org/10.1155/2019/3610248
[2] Mudasir Ashraf, Majid Zaman and Muheet Ahmed, 2018 Using Predictive Modeling System and Ensemble Method to Ameliorate Classification Accuracy in EDM”, Asian Journal of Computer Science and Technology, ISSN: 2249-0701 Vol.7 No.2, pp.44-47©, The Research Publication, www.trp.org.in
[3] A. Daud, N. R. Aljohani, R. A. Abbas, M. D. Lytras, F. Abbas and J. S. Alowibdi, 2017 Predicting student performance using advanced learning analytics, in Proceedings of the 26th International Conference on World Wide Web, Companion, Perth, Australia, April.
[4] C. Romero, M.-I. L’opez, J.-M. Luna, and S. Ventura, 2013 Predicting students’ final performance from participation in online discussion forums,” Computers and Education, vol. 68, pp. 458–472.
[5] S. Natek and M. Zwilling, 2014 Student data mining solution knowledge management system related to higher education institutions, Expert Systems with Applications, vol. 41, no. 14, pp. 6400–6407.
[6] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, 2015 Improving accuracy of student’s final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique, Decision Analytics, vol. 2, no. 1, p. 1.
[7] K. Bunkar, U. K. Singh, B. Pandya, and R. Bunkar, 2012 Data mining: prediction for performance improvement of graduate students using classification, in Proceedings of 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN), Indore, India.
[8] G. Gray, C. McGuinness, and P. Owende, 2014 An application of classification models to predict learner progression in tertiary education in Proceedings of 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, India.
[9] T. Mishra, D. Kumar, and S. Gupta, 2014 “Mining students’ data for prediction performance,” in Proceedings of 2014 Fourth International Conference on Advanced Computing and Communication Technologies (ACCT), Rohtak, India.
[10] T. M. Christian and M. Ayub, 2014 Exploration of classification using NBTree for predicting students’ performance,” in 12 Scientific Programming Proceedings of 2014 International Conference on Data and Software Engineering (ICODSE), Bandung, Indonesia.
[11] M. Mayilvaganan and D. Kalpana Devi, 2014 Comparison of classification techniques for predicting the performance of students academic environment,” in Proceedings of 2014 International Conference on Communication and Network Technologies (ICCNT), Sivakasi, India.
[12] E. Osmanbegovi´c and M. Sulji´c, 2012 Data mining approach for predicting student performance,” Economic Review, vol. 10, no. 1, pp. 3–12.
[13] A. M. Ahmed, A. Rizaner, and A. H. Ulusoy, 2016 Using data mining to predict instructor performance,” Procedia Computer Science, vol. 102, pp. 137–142.
[14] S. Kalaivani, B. Priyadharshini, and B. S. Nalini, 2017 Analyzing student’s academic performance based on data mining approach,” International Journal of Innovative Research in Computer Science and Technology, vol. 5, no. 1, pp. 194–197.
[15] G.Ayyappan, K.SivaKumar, 2018 A Noval approach of ensemble models by using EDM*, in Indian Journal of Computer Science and Engineering (IJCSE), e-ISSN: 0976-5166, p-ISSN: 2231-3850, Vol. 8 No. 6.

[16] Kumar M., Singh A. J., Handa D. 2017 Literature Survey on Educational Dropout Prediction*, International Journal of Education and Management Engineering (IJEME), Vol.7, No.2, pp.8-19.DOI: 10.5815/ijeme.2017.02.02.

[17] Kumar M., Singh A.J., Handa D., 2017 Literature Survey on Student’s Performance Prediction in Education using Data Mining Techniques, International Journal of Education and Management Engineering (IJEME), Vol.7, No.6, pp.40-49, DOI: 10.5815/ijeme.2017.06.05