International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-3, February 2019

Performance Analysis of Students Using Machine Learning & Data Mining Approach

Mukesh Kumar, A. J. Singh

Abstract: ***Performance evaluation of students is essential to check the feasibility of improvement. Regular evaluation not only improves the performance of the student but also it helps in understanding where the student is lacking. It takes a lot of manual effort to complete the evaluation process as even one college may contain thousands of students. This paper proposed an automated solution for the performance evaluation of the students using machine learning. A threshold-based segmentation is employed to complete the evaluation procedure over MATLAB simulation tool. The performance of machine learning is evaluated by accuracy and mean square error.***

Index Terms*: *Performance Evaluation, Machine Learning, Performance Improvement, MATLAB, Mean Square Error, Estimation Effort

I. INTRODUCTION

An education system is one of the most important parts for the development of any country. So it should be taken very seriously from its start. Most of the developed countries have their own education system and evaluation criteria. Now a day's education is not limited to only the classroom teaching but it goes beyond that like Online Education System, MOOC course, Intelligent tutorial system, Web-based education system, Project based learning, Seminar, workshops etc. But all these systems are not successful if they are not evaluated with accuracy. So for making any education system to success, a well-defined evaluation system is maintained. Every educational institution generate lots of data related to the registered student and if that data is not analysis properly then all afford is going to be wasted and no future use of data happen. This institutional data is related to the student admission, student family data, student result etc. Every educational institution applies some assessment criteria to evaluate their students. For decades, the students are analyzed through various numbers of processes. Student's performance evaluation is as important as that of the study process of the student. The regular and traditional procedure is to take examinations or class assessments etc [1]. This process takes a lot of manual effort to complete the evaluation and it is very time consuming as well. This paper proposed the usage of machine learning in the evaluation of the performance of the students. It does not only evaluate the performance but also helps in improving various aspects of the students [2, 3, 4]. A lot of evaluation process may put the student into a lot of stress. Keeping the student in stress is not going to help the student perform well.

Manuscript published on 28 February 2019.
* Correspondence Author (s)
Mukesh Kumar*, Assistant Professor, Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India.
Prof. A. J. Singh, Department of Computer Science, Himachal Pradesh University, Summer-hill, Shimla. Himachal Pradesh.

II. DATA PRE-PROCESSING

Now a Dataset used for implementation is taken from **http://invertory.data.gov/dataset** named as "**Results for State Assessments in Reading/Language, Arts and Mathematics"** respective year of school result are (2008-09, 2009-10, and 2010-11) which was collected by EDFacts. EDFacts is an organization in United State which is under department of Education project to gather student academic information, analyze the gathered data, generate a report on that data, and then endorse the use of high-quality data in educational planning, making new policy and management of budget for decision-making to improve the educational outcomes for students. Below description gives an overview of the dataset used with different variables used with their description. The dataset used contain 34 different attributes which contain some school attributes and some subject marks related attributes. The different attribute names used within the dataset are structured with different abbreviations. Below is the list of all abbreviations used to make variable structure.

Syntax for give a variable name:
[Subgroup]_ [Subject][Grade][Metric]_ SchoolYear

For example:

1. The attribute name (ALL_MTH00numvalid_1011) contains the information of every student who is participating in evaluation in mathematics and scored valid marks, in all grades in academic session year 2010-11.
2. The attribute name (MHI_RLA08pctprof_0809) contains the information of Hispanic students scoring at or above proficient in Reading/Language Arts in eighth grade in Session Year 2008-09.

III. PROPOSED FRAMEWORK & METHODOLOGY USED

The proposed framework of estimation of the performance of the students in academic is divided into two different parts:
a) Data reconstruction and layout formation using threshold segmentation architecture (DRLFT). b) Analysis of the segmented data for precise prediction.

DRLFT: The dataset utilized for implementation is taken from **http://invertory.data.gov/dataset.** The dataset contains the records of different schools/colleges and different subjects. But the problem is that the data in the dataset is not in the structured format which is used for the MATLAB implementation [11].

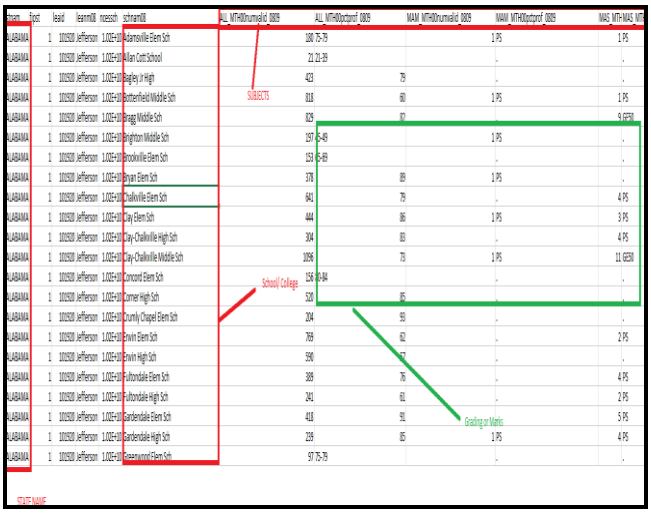

Fig 1: The unstructured data set

This data in the dataset need a reconstruction and the reconstruction is done using the following algorithmic architecture.

Algorithm 1: Function Reconstruct_Data (un_st)

1. // un_st is the unstructured data.
2. sub_codes = find (un_st = = sub_code); // the data is getting arranged on the basis of the subject codes.
3. data_st = [ ]; // It is the structured set of the data.
4. data_st[count, sub_data] = sub_code.find(data (un_st)); // Finding the details of the subject out of the given series of data
5. count = count+1;
6. end for
7. end function

Algorithm 1 arranges the data in the dataset into a structured format on the base of the subject code data. The same structure of above algorithm can be employed to arrange the data based on the school/college name.

Algorithm 2: Function reconstruct_data_college (un_st)

1. coll_code = find(un_st == coll_code);// coll_code is the school code
2. col_count = 1;
3. For each col in coll_code
4. data_st [col_count. coll_code.st] = col_data; // Filling the school information into the school structure
5. col_count = col_count + 1;
6. end for
7. end function

Algorithm 2 arranges the data from the dataset into a structured format on the base of the school/ college code used. By performing above two algorithms on our dataset, we put our data into a structured data format which is much need for our implementation purpose.

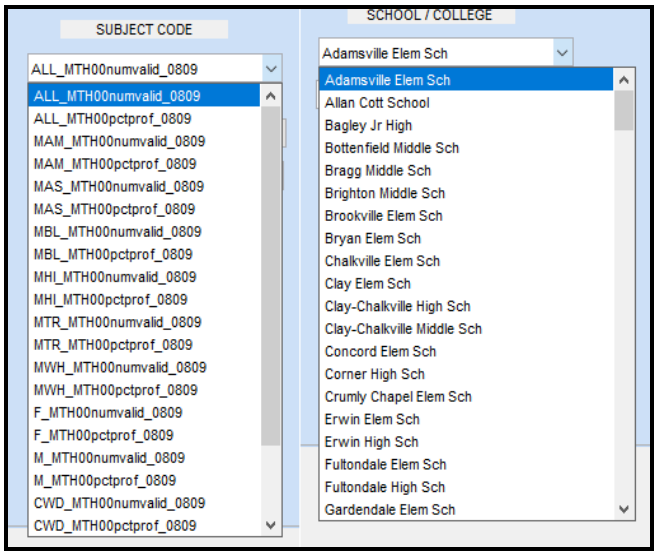

Fig 2: Segregated Schools / Colleges and Subject Data

In order to club Subject Code along with the School name, we are implementing K-means clustering algorithm based on the threshold segmentation is applied [7, 8, 9].

Algorithm 3: Apply Threshold_Segmentation_Kmeans (st.data):

1. For each class label in label properties
2. X (1) = class.label.name;// Collecting the school/college and class related data
3. X (2) = class.label.value
4. end for
5. end Function

After implementing Algorithm 3 on the above structured dataset, we find the following output values.

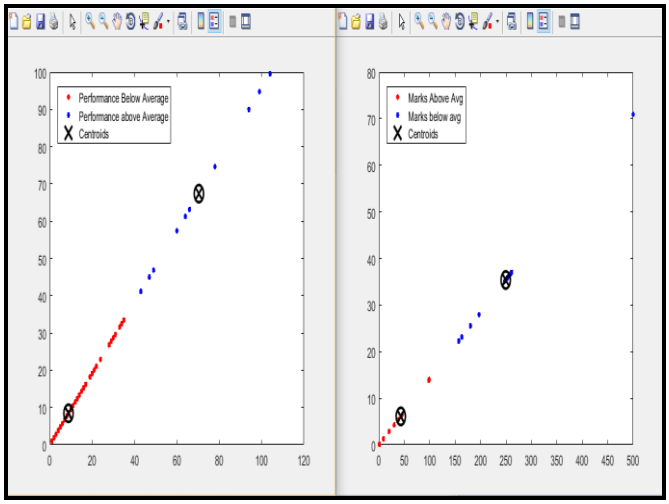

Fig3: a. Subject Code K-means b. School Code K-means

Figure 3, gave us the visualization of the K Means clustering algorithm after implement on the taken dataset. Figure 3(a) gave us the clustered data according to subject code and Figure 3(b) gave us the data clustered with respect to school code used in the dataset.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Analysis of the segmented data for precise prediction:
K-Means clustering algorithm not only provides the grouped elements together but also marks their region centre head. The data is clustered into two different segments namely "Performance below average students" and "Performance above average students". The clustered elements are validated through Artificial Neural Network (ANN) and Support Vector Machine (SVM). The validation structure is as follows:

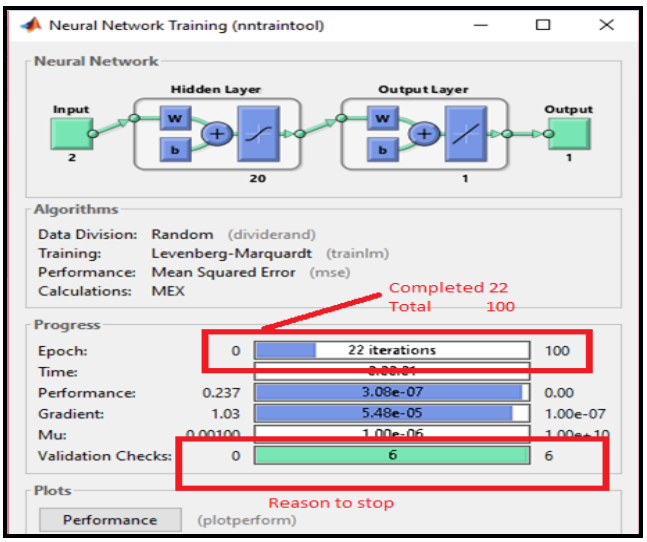

Fig 4: Training Architecture of Artificial Neural Network

Artificial Neural Network classification algorithm trains on the training data taken from given dataset based on the segmented values of K-Means. This particular classification algorithm provides multiple validation factors likes Mean Square Error as shown in Figure.4. So, our evaluation criteria for this dataset are based on *Mean Square Error* for Artificial Neural Network.

In Figure 5, the same type of training dataset as taken as input to implement the *Support Vector Machine* algorithm and the following results are obtained.

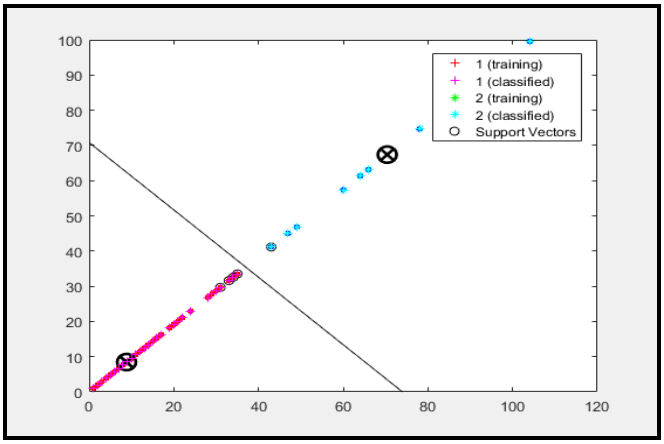

Figure 5: Training and Classification of SVM

The training and classification structure of *Artificial Neural Network* and *Support Vector Machine* ensures that the performances of the students are evaluated on the true pattern.

IV. RESULT ANALYSIS

The results of the above implementation are evaluated on the basis of Mean Square Error and the Effectiveness of the effort required for the evaluation. A total of 10,000 records are trained and the following results are evaluated.

Mean Square Error: Mean Square Error is the difference of the prediction and actual analysis. Table 1 illustrates the values for 10,000 records.

Table 2: Mean Square Error (MSE) result using K-means clustering for ANN and SVM

Record Count	Mean Square Error K Means - ANN	Mean Square Error K Means – SVM
100	12.24	17.96
200	14.14	18.35
500	15.23	22.21
700	18.21	24.53
800	18.78	26.45
1000	19.02	28.69
5000	20.03	32.45
10000	23.5	35.36

Table 2 shows that, the Mean Square Error value increases as the value of the record count increase. So, it is obvious from the above implementation that as the number of record count increase then the error rate will be high. As compared to Support Vector Machine (SVM), Artificial Neural Network (ANN) shows better efficiency in terms of producing error rate. The overall difference in producing the error rate is about 20% between ANN and SVM. Figure 6 shows the graphical representation of the data shown in Table 2.

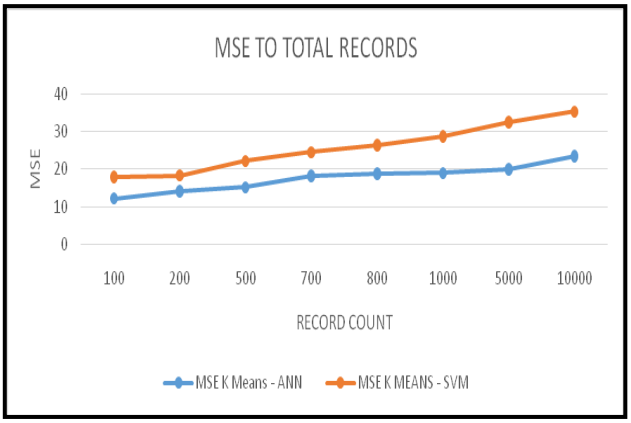

Fig 6: Mean Square Error vs. Record Count

Form Figure 6, it is very much clear that the Mean Square Error rate of Support Vector Machine algorithm is always higher then Artificial Neural Network as the record count increase every time. The difference in the error rate is higher as be reach at the 10,000 or maximum record count. Effort Estimation: The second judgment parameter is effort required for the evaluation of the student result. Below is the formula to calculate the effort estimation: Total Effort = ((1-overallmse) * 100) – (overallmse * group_element_count);

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Total Effort Estimation helps in the evaluation of the correctness and effort required
to group the elements for the evaluation of the students. Table 3 illustrates the values of the total effort applied for both the algorithms.

Table 3: Estimation Effort required by using K-means clustering for ANN and SVM

Record Count	Effort Estimation K Means - ANN	Effort Estimation K Means - SVM
100	32	47
200	38	48
500	39	49.36
700	39.5	49.87
800	40	53.56
1000	41	55
5000	41.3	58
10000	42	69

Table 3 shows that, the Effort Estimation value increases as the value of the record count increase. So, it is obvious from the above implementation that as the number of record count increase then the Effort Estimation value will be high. As compared to Support Vector Machine (SVM), Artificial Neural Network (ANN) shows better efficiency in terms of Effort Estimation value. The overall difference in producing the Effort Estimation value is about 27% between ANN and SVM. Figure 7 shows the graphical representation of the data shown in Table 3.

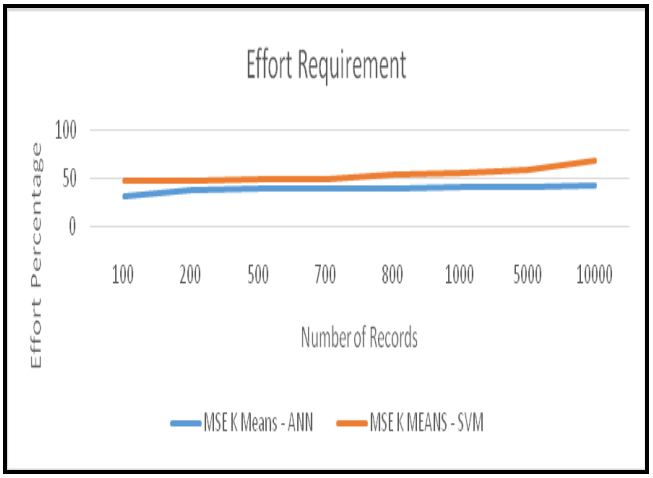

Fig 7: Illustrate the Graphical representation of Table 3

Form Figure 7, it is very much clear that the Effort Estimation value of Support Vector Machine algorithm is always higher then Artificial Neural Network as the record count increase every time. The difference in the error rate is higher as be reach at the 10,000 or maximum record count.

V. CONCLUSION

The paper proposed a combination of K-Means clustering algorithm with Artificial Neural Network and Support Vector Machine classification algorithm to evaluate the student performance in order to reduce the human effort. The evaluation is done on the basis of Mean Square Error and Effort Estimation. The results of our implementation show that the performance of Artificial Neural Network in comparison to Support Vector Machine is better. The Mean Square Error is 5-20% better where as the Effort Estimation is around 15-27% better. The current research work opens a lot of future aspects for the researchers. The future possibilities include changing the total number of neurons or varying the satisfying parameters.

REFERENCES

1. Ramanathan L., Parthasarathy G., Vijayakumar, K., Lakshmanan, L., & Ramani, S. (2018). Cluster-based distributed architecture for prediction of student's performance in higher education. Cluster Computing, 1-16.
2. Thomas, C. L., Cassady, J. C., & Heller, M. L. (2017). The influence of emotional intelligence, cognitive test anxiety, and coping strategies on undergraduate academic performance. Learning and Individual Differences, 55, 40-48.
3. Keyes K., & Dworak E. (2017). Staffing Chat Reference with Undergraduate Student Assistants at an Academic Library: A Standards-Based Assessment. The Journal of Academic Librarianship, 43(6), 469-478.
4. Yahya A. A. (2017). Swarm intelligence-based approach for educational data classification. Journal of King Saud University-Computer and Information Sciences.
5. Bharara S., Sabitha S., & Bansal A. (2017). Application of learning analytics using clustering data Mining for Students' disposition analysis. Education and Information Technologies, 1-28.
6. Brown S., Bowmar A., White, S., & Power, N. (2017). Evaluation of an instrument to measure undergraduate nursing student engagement in an introductory Human anatomy and physiology course. Collegian, 24(5), 491-497.
7. Ojha A. K. (2017). Management education in India: avoiding the simulacra effect. In Management Education in India (pp. 55-77). Springer, Singapore.
8. Asif R., Merceron A., & Pathan M. K. (2015, March). Investigating performance of students: a longitudinal study. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 108-112). ACM.
9. Pandey M., & Taruna S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. Perspectives in Science, 8, 364-366.
10. Almutairi F. M., Sidiropoulos N. D., & Karypis G. (2017). Context-aware recommendation-based learning analytics using tensor and coupled matrix factorization. IEEE Journal of Selected Topics in Signal Processing, 11(5), 729-741.
11. Oskouei R. J., & Askari M. (2014). Predicting academic performance with applying data mining techniques (Generalizing the results of two different case studies). Computer Engineering and Applications Journal, 3(2), 79.
12. Wook M., Yusof Z. M., & Nazri M. Z. A. (2017). Educational data mining acceptance among undergraduate students.
13. Education and Information Technologies, 22(3), 1195-1216.
14. Hussain M., Al-Mourad M., Mathew S., & Hussein A. (2017). Mining educational data for academic accreditation: Aligning assessment with outcomes. Global Journal of Flexible Systems Management, 18(1), 51-60.
15. Costa E. B., Fonseca B., Santana M. A., de Araújo F. F., & Rego J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 73, 247-256.
16. Polyzou A., & Karypis G. (2016, April). Grade prediction with course and student specific models. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 89-101). Springer, Cham
17. Wisneski J. E., Ozogul G., & Bichelmeyer B. A. (2017). Investigating the impact of learning environments on undergraduate students' academic performance in a prerequisite and post-requisite course sequence. The Internet and Higher Education, 32, 1-10.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

18. Ghani A. A., & Mohamed R. (2017). The Effect of Entry Requirement for Civil Engineering Student Performance. Journal of Science and Technology, 9(4).
19. Chakraborty T., Chattopadhyay S., & Chakraborty, A. K. A novel hybridization of classification trees and artificial neural networks for selection of students in a business school. OPSEARCH, 1-13.
20. Kumar M., & Singh A. J. (2017). Evaluation of Data Mining Techniques for Predicting Student's Performance. International Journal of Modern Education and Computer Science, 9(8), 25.
21. Meedech P., Iam-On N., & Boongoen T. (2016). Prediction of student dropout using personal profile and data mining approach. In Intelligent and Evolutionary Systems (pp. 143-155). Springer, Cham.
22. Yehuala M. A. (2015). Application of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University). International Journal of Scientific & Technology Research, 4(4), 91-94.
23. Asif R., Haider N. G., & Ali S. A. (2016). Prediction of Undergraduate Student's Performance using Data Mining Methods. International Journal of Computer Science and Information Security, 14(5), 374.
24. Tran T. O., Dang H. T., Dinh, V. T., & Phan, X. H. (2017). Performance Prediction for Students: A Multi-Strategy Approach. Cybernetics and Information Technologies, 17(2), 164-182.
25. Kim K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. Pattern Recognition, 60, 157-163.
26. Kumar M., Singh A.J. , Handa D.,"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, 2017.DOI: 10.5815/ijeme.2017.06.05
27. Guarín C. E. L., Guzmán E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de tecnologias del Aprendizaje, 10(3), 119-125.
28. Kumar M., Singh A. J., Handa D."Literature Survey on Educational Dropout Prediction", International Journal of Education and Management Engineering (IJEME), Vol.7, No.2, pp.8-19, 2017.DOI: 10.5815/ijeme.2017.02.02.

AUTHORS PROFILE

Mukesh Kumar is currently working as Assistant Professor in Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India. He completed his M.Tech in Computer Science and Engineering from HPU Shimla in 2008. He is currently pursuing PhD degree in the department of Computer Science, Himachal Pradesh University, Summer hill, Shimla, India and His research interest includes Machine learning, Artificial intelligence, Information security, Educational Data Mining. He has 10 years of teaching experience and published 15 research papers in different international journals.

Prof A.J. Singh is a Professor in the Department of Computer Science in Himachal Pradesh University, Shimla. He has been in this Department since 1992. He has obtained his B.Tech in Computer Technology from National Institute of Technology (MANIT) Bhopal, Master of Science in Distributed Information Systems from University of East Landon (UK) and PhD degree from Himachal Pradesh University Shimla. He has published more than 50 research papers, supervised PhD and M.Tech students. His area of interests are Distributed System (Networks and DBMS), and ICT for Development.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)