Linear Regression and Support Vector Machine for Classification of E-Learning Students' Engagement and Performance

C.S.Sasikumar, A.Kumaravel

Abstract: ***Dealing with continuous (frequently occurring) and huge volume data with conventional computing environment becomes challenging now days due to the memory and processing limitations present in the allocated resources though the network connectivity and site processing power are relatively high. However distributed processing approaches support this issue as a main theme through filtering data to our needs and analyzing the data to check the presence of required characteristics becomes solvable in specific contexts and IT industries thrives on this capability. In this paper the environment for e-Learning is selected as there are many inherent problems to be solved and researchers progress mainly due to technology advancement by using relevant tools. Here we address the problem of filtering or extracting the data not from any data warehouse but continuous data collection from connected nodes and isolated tables and generate the data for checking the relationships from inputs of student engagement activities to their performance. Since the inputs are not independent, it is applied with various filters through the queries and the patterns are detected by machine learning techniques like linear regression and support vector machine.***

Keywords **:** ***E-Learning, Data mining, Structured Query Language, Randomizer, Schema, Entity Relationship, R programming, Linear Regression, Support Vector Machine.***

I. INTRODUCTION

The environment for e-Learning is selected as there are many inherent problems to be solved in this domain and researchers progress mainly due to technology advancement. Here we address the problem of filtering or extracting the data not from any data warehouse but continuous data collection from connected nodes and separate tables and generate the data for checking the relationships from inputs of student engagement activities to their performance. Since the inputs are not independent it is applied with various filters through the queries and the patterns detected by machine learning techniques or data mining algorithms like linear regression and support vector machine.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

C.S. Sasikumar*, Research Scholar, Department of Computer Science and Engineering Bharath Institute of Higher Education and Research, Chennai, India.

A. Kumaravel, Professor, Dean, School of Computing, Bharath Institute of Higher Education and Research, Chennai, India.

1.1 Related Works

Educational activities in e-Learning context, determines the improvement in quality of learners' engagement and detecting the same is a key issue. The authors Wolff, et al, [1] tried to calculate probability in a spread of data of records in the order of millions for the event that the student submits the first assessment A1 and that the score of this assessment will be higher than 40.The detection of such improvement even from learners' facial expressions is the main research topic for long time [2, 3]. High drop-out rates especially in e-Learning courses made the concerns of research [4].

Interactive online learning facilities for providing personalized pedagogical platform try to detect learners' engagement and this becomes main stream in e-Learning research. Similar strategies are tried or well approached in case of conventional classrooms, intelligent learning packages and game based lessons [5]. The unique difference between dataset for e-Learning and that of conventional learning is the general availability as it is very rare in the former case. Moreover these two can overlap very well. The good organization of this data directly determines the quality of the predictions made over such data set which affects very easily the applications made out of them. Most of the time the redundancy in the data gives more time for processing and misses the chances of getting optimal accuracy. The problem of searching for redundant attributes demands either equivalence classes in which the equivalent attributes are put together and among which any one of them can be treated significant and the rest of the equivalent attributes as insignificant or by ranking them using criteria measures and removing the attributes in the tail end of rank list. Such elimination of attributes should not contribute to error or loss of information eventually. Learners' engagements can be obtained from the repository in the format of log file [6, 7] and this type data had been studied by Bosch [8-11], Anderson [12] elaborately by organizing them into cognitive, affective, behavioral and emotional. Affective type of engagement is based on the change of state in interest and joyful state in learning, whereas academic type of engagement depends on academic alignments with teachers and more involvement (e.g., completion time of tasks, not avoiding classes) in learning [12]. Behavioral type of engagement represents on the point of participation including in the classroom and external tasks, timely submission of assigned tasks, and strictness in following the teacher's instruction [13]. Cognitive type of engagement deals with the inclination and readiness to apply the efforts necessary to understand complex topics and learn difficult skills (e.g., concentration, and creative and critical thinking) [13]. Emotional type of engagement consists of bi polar reactions (either negative or positive) to monitors, peers, and people in academics. Psychological type of engagement based on the degree of belonging and interactions with mentors and peers [12, 13].

Different types of engagements are considered in some research studies emphasizing that measuring engagement through requires bringing together external data i.e, observational data with the data internal data belonging to the individual by 'self-reports'[2,3].The authors also published the feature selection procedures and their relative performance using rough set theory and data mining [14,15].

This paper presents answers to the questions involving the students' engagement detection techniques in the context of e-learning, specifically the identification of significant features for predicting the learners' performance and simpler mapping to model the relationship among various activities and their results obtained. Finally, we interpret the results by the plots generated for the selected algorithms.

The remainder of the article is sequenced as follows. In Section II, steps for data transformations and the corresponding structured queries are discussed. These queries make the original data set (used in [1] collected and collaborated for e-learning in an open university). and further detailed in Section III. Data scheme, tool selected along with some results are discussed. Final section concludes the paper with some critical discussions and future recommendations.

II. METHODS AND MATERIALS

In this section we describe the dataset derived from the ER diagram shown in Figure 1 containing the entities representing the components like student details, assessment details, activity details etc and it includes 7 entities. The data size happens to be 'Big' and it is not favoring the limitations of arbitrary computing node with even for configurations i7 processor with 16 GB RAM, 2 TB ROM. Hence in order to manage with our resources, we apply transformations to filter parts of the dataset with structured queries as seen in Figure 1. Next we deal with descriptions of the set of queries based on such entities /underlying tables derived from the dataset, as pointed out in Figure 2.

2.1 Data Set Descriptions

Open University data [29] yields students' data whom registered for e-learning courses, taking assessments under various activity types. Figure 1 shows the schema hidden in this data and we extract the final dataset using the queries as shown in Figure 2. Sample met data is given for three modules in Table 1.

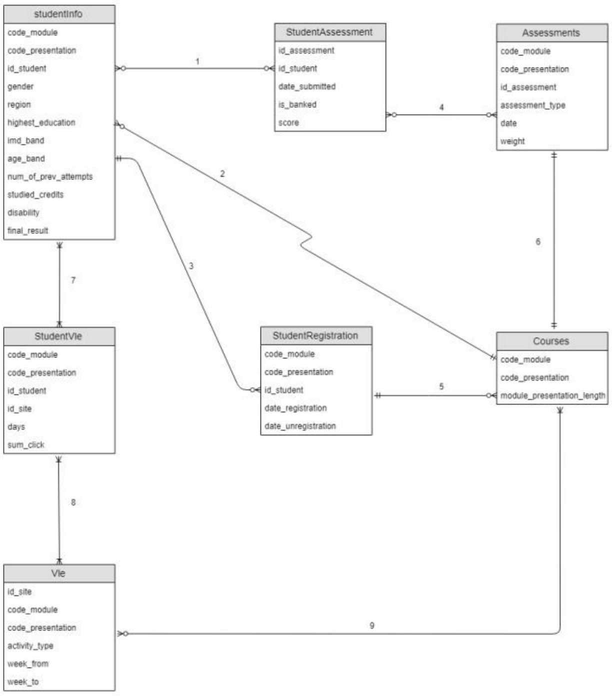

Fig. 1.Organization of dataset using seven entities in the schema for Open University [29]

Table 1: Three selected modules (AAA, BBB, CCC) for three assessments

Module ID	**# students**	**# VLE clicks**
AAA	4,397	1,570,402
BBB	1,292	2,750,432
CCC	2,012	1,218,327

VLE clicks were counted within a time period between assessments. Few weeks are given for this time period and the number of clicks decreases at the tail end of this time period. Among seven modules, only three modules selected to observe the sum of clicks made. In the assessment table assessment details can be either denoted according to the student obtained score or in terms of passing or failing limits with labels 'pass' or 'fail'.

The following bullets describe the links over entities as in Figure 1 for making possible joins and the output view tables are further filtered as shown Figure 4.

- The entity student_info related to student_assessment table in order to get the date submitted and score of the assessment.
- The entity student_info related to courses table in order to get the module presentation length of the courses.
- The entity student _info related to student registration table inorder to get the date registration and date unregistration of the student registration.
- The entity student _assessment related to assessment table inorder to get the assessment type, date and weight of the assessment.

- The entity student_registration related to courses table in order to get the module presentation length of the course.
- The entity assessment related to course table in order to get the module presentation length of the course.
- The entity student_info related to student vle table inorder to get the id site, date and sum click of the student_vle.
- The entity student_vle related to vle table inorder to get the activity type, week from and week to of the vle.
- The entity course related to vle table in order to get the activity type ,week from and week to of the vle.

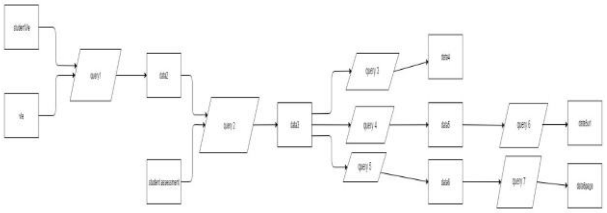

Fig. 2. Organization of data queries to filter for pre processing

There are seven queries are applied either sequentially or in parallel as organized in Figure 2. They are constructed based on following structured queries:

Query1: - SelectstudentVle.code_module,studentvle.id_site, studentVle.sum_click, vle.activity_type from vle inner join studentVle on studentvle.id_site=vle.id_site.

Query 2: - Selectdata2.code_module,data2.id_student,data2.id_site,studentAssessment.id_assessment,data2.activity_type, data2.sum_click,studentAssessment.score from data2 inner join studentAssessment ondata2.id_student=studentAssessment.id_student

Query 3: -Select * from data3 where code_module=’AAA’

Query 4: - Select * from data3 where code_module=’BBB’

Query 5: - Select * from data3 where code_module=’CCC’

Query 6: - Select * from data5 where activity_type =’url’

Query7: - Select * from data6 where activity_type=’page’

2.2 Selected Tool and its description

Our experiments for data mining based on uploaded dataset with students’ attributes as shown in the scheme, is iterated with R tool since it is capable of relatively high volume data and computing performance on a high end PC.

R is a tool based on enriched statistical operations for modeling graphics plotting and summarizing. This programming environment comes with an environment for integrated development referred as R studio with the facility for scripting. It is an open source package mainly with a work environment included for command lines facilities and the presentation of analysis output. This text oriented output summary and tables can be transferred to other tools or locations very easily.

2.3 Classifiers selected for student engagement and performance

Dividing the dataset for associating to the binary classes or clustering into two groups can easily be handled by linear regression and support vector machine algorithms. The model due to linear regression [32] employs simple line (or hyper plane, in case of multi-dimensional data space) based on least square principles applied in core statistics whereas instead of single linear boundary, the margin is identified either sides of the boundary supported by minimal distant data points (referred as ‘support vectors’) [30]. The non-linearity of the boundary can be tackled by the ‘kernel’ tricks as several applications implemented in literature [31]. The following Table 2 presents the R implementations of these two classifiers and mainly describes the commands with the input parameters and the intended output.

S.No	R-command	Brief description(input/output)
1.	lm(y ~ x, data)	The function used for building linear models is lm (). The lm() function has two main arguments, namely: Firstly, the expression in X and Y; Secondly the data Frame with data and the Expression are an object of class. The algorithm for linear regression is applied for predicting the value of an output variable Y based on one or more input variables X. The main objective is to identify a linear relationship (a numerical type formula) between the input variable(s) and the output variable, so that, we can use this regression expression to predict the value of the output Y, when only the inputs (Xs) values are known.
2.	svm (Y ~ X , data)	A Support Vector Machine (SVM) is a learning classifier properly defined by a dividing hyper plane and called by function svm(). data an non mandatory data frame containing the variables in the model. By default the variables are taken from the environment in which 'svm' is called from. x a matrix with data, or a row vector, or a matrix with lot of zeros (object of class Matrix provided by the Matrix package, or of class matrix's provided by the SparseM package, or of class simple_triplet_matrix provided by the slampackage). y an output vector with single label for every row/component of x. It can be either a nominal value (for classification tasks) or a numeral scalars constituting a vector (for regression).

Table 1: Commands description

III. EXPERIMENTAL SETUP

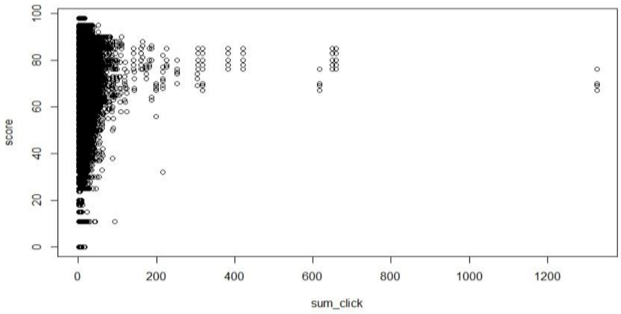

Fig. 3.Data distribution for (sum_click, score) for the dataset mean for code_module=’AAA’. (17,48,420 instances)

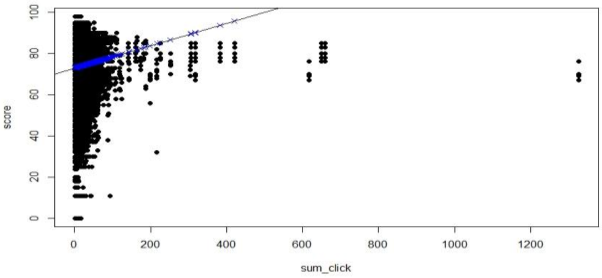

Fig. 4. Linear Regression and Support Vector Machine behaviors (almost close) for the dataset meant for code_module='AAA'.(17,48,420 instances)

Both figures partition data in a surprised pattern. This pattern may be varied if different datasets are selected, in specific the student with less number of clicks. In general, achieve higher score comparing to the student with higher number of clicks. Hence, the present of frequency clickers implies to present of slow learner. The moreover size of quick learner is small as seeing in the sparsity of the right side the figure.4 and figure.5. The above figure shows the relationship between sum click made by 17, 48,420 (data4) students using all activity types value to score obtained (0-100).

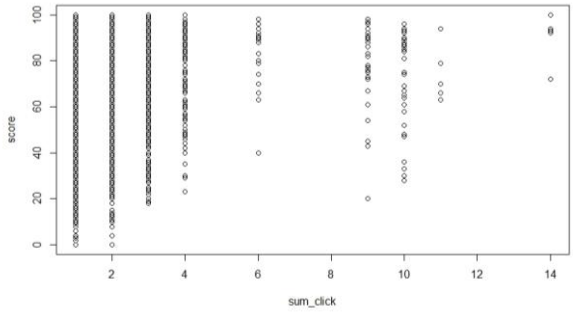

Fig. 5. Data distribution for (sum_click, score) dataset meant for (code_module='CCC') and (activity_type='page'). (39,266 instances)

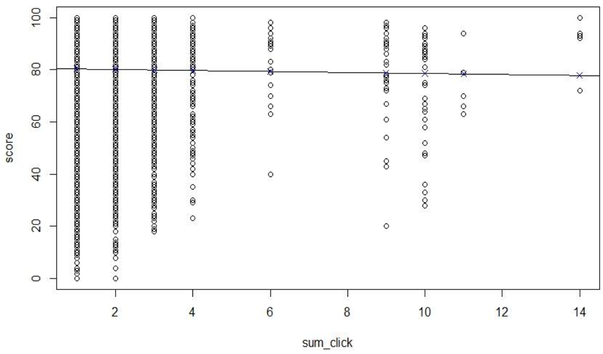

Fig. 6.Linear Regression and Support Vector Machine behaviors (almost close)for the dataset meant for (code_module='CCC') and (activity_type='page'). (39,266)

Both figures partition data in a surprised pattern. This pattern may be varied if different datasets is selected, in specific, student with less number of clicks. In general, achieve higher score comparing to the student with higher number of clicks. Hence, the present of frequency clickers implies to present of slow learner. The moreover size of quick learner is small as seeing in the sparsity of the right side the figure.5 and figure.6.The above figure shows the relationship between sum_click made by 39,266 (data6page) students using page number to score obtained (0-100).

IV. CONCLUSION

In this study we have handled big dataset in e-learning existing in Open University for observing the relationship between inputs namely clicks made the e-learners (derived from repositories as given by scheme in Figure 1) and their score. The results confirm the section of registered students whom have tendency of late pickup or slow learning in the selected module. The main classifiers we had selected namely the linear regression and support vector machine depict almost close behavior (Figure 4 & 6) We can extend the results for answering other characteristics of the e-learners with similar classifiers.

ACKNOWLEDGEMENT:

The authors feel grateful with Bharath University for their support and encouragement for this research study and also the Open University which made the dataset available public.

REFERENCES

1. MushtaqHussain , Wenhao Zhu , Wu Zhang , and Syed Muhammad RazaAbidi,(2018) "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores" , Computational Intelligence and Neuroscience , , Article ID 6347186
2. J. Whitehill, M. Bartlett, J. Movellan, Automatic Faacial Expression Recognition for Intelligent Tutoring Systems(IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, 2008)
3. J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, J. Movel, The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing 5(1), 86–98 (2014)
4. A.L. Rothkrantz, Dropout Rates of Regular Courses and MOOCs(International Conference on Computer Supported Education, Rome, 2016)
5. S. Karumbaiah, B. Woolf, R. Lizarralde, I. Arroyo, D. Allessio, N. Wixon, Addressing Student Behavior and Affect with Empathy and Growth Mindset (International Conference on Educational Data Mining, Wuhan, 2017)
6. M. Cocea, S. Weibelzahl, Log file analysis for disengagement detection in e-learning environment. User Model. User-Adap. Inter. 19, 341–385 (2009)
7. M. Cocea, S. Weibelzahl, Disengagement detection in online learning: Validation studies and perspectives. IEEE Trans. Learn. Technol. 4(2), 114–124 (2011)
8. N. Bosch, Detecting Student Engagement: Human Versus Machine (Conference on User Modeling Adaptation and Personalization, Halifax, 2016)
9. N. Bosch, Y. Chen, S. D'Mello, It's written on your face: Detecting affective states from facial expressions while learning computer programming (Intelligent Tutoring Systems, Honolulu, 2014)
10. N. Bosch, S.K. D'Mello, R.S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, W. Zhao, Automatic Detection of Learning-Centered Affective States in the Wild (International Conference on Intelligent User Interfaces, Atlanta, 2015)
11. N. Bosch, S.K. D'Mello, R.S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, W. Zhao, Detecting Student Emotions in Computer-Enabled Classrooms (International Joint Conference on Artificial Intelligence, New York, 2016)
12. S. Christenson, A. Reschly, C. Wylie, Handbook of Research on Student Engagement (Springer, New York, 2012)
13. S.Christenson, A.R. Anderson, The centrality of the learning context for students' academic enabler skills. School Psychological Review 31(3), 378–393 (2002)

14. C.S.Sasikumar, A.Kumaravel, Yet another Approach for Construction of Cost Sensitive Classifiers for E-Learning Datasets, International Journal of Engineering & Technology, 7 (4.39) (2018)1047-1052
15. C.S.Sasikumar, A.Kumaravel, Attribute Selection on Student Academic and Social Attributes Based on Randomized And Synthetic Dataset, International Journal of Engineering & Technology, 7 (4.39) (2018), 1069-1075
16. Wolff, A., Zdrahal, Z., Herrmannova, D., &Knoth, P. (2013). Predicting Student Performance from Combined Data Sources. Studies in Computational Intelligence, 175–202.
17. T. Aluja-Baneta, M.-R. Sanchob, I. Vukic, Measuring motivation from the virtual learning environment in secondary education. Journal of Computational Science, 1–7 (2017)
18. J.E. Beck, Engagement Tracing: Using Response Times To model Student Disengagement (Conference on Artificial Intelligence in Education, Amsterdam, 2005)
19. B.M. Booth, A.M. Ali, S.S. Narayanan, I. Bennett, A.A. Farag, Toward Active and Unobtrusive Engagement Assessment of Distance Learners (International Conference on Affective Computing and Intelligent Interaction, San Antonio, 2017)
20. G. Matthews, S. Campbell, S. Falconer, L. Joyner, J. Huggins, K. Gilliland, J. Warm, Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. Emotion 2(4), 315–340 (2002)
21. B. McDaniel, S. D’Mello, B. King, P. Chipman, K. Tapp, A. Graesser, Facial Features for Affective State Detection in Learning Environments (Proceedings of the Annual Meeting of the Cognitive Science Society, California, 2007)
22. H. Monkaresi, N. Bosch, R. Calvo, S. D'Mello, Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans. Affect. Comput. 8(1), 15–28 (2017)
23. H.L. O'Brien, E.G. Toms, The development and evaluation of a survey to measure user engagement. J. Am. Soc. Inf. Sci. Technol. 61(1), 50–69 (2010)
24. J. Parsons, L. Taylor, Student engagement: what do we know and what should we do? (University of Alberta, Technical Report, Edmonton, 2011)
25. S. Sathayanarayana, R.K. Satzoda, A. Carini, M. Lee, L. Salamanca, J. Reilly, G. Littlewort, Towards Automated Understanding of Student-Tutor Interactions Using Visual Deictic Gestures (IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, 2014)
26. M. Sathik, S.G. Jonathan, Effect of facial expressions on student’s comprehension recognition in virtual educational environments. SpringerPlus 2(455), 1–9 (2013)
27. P. Sundar, S. Kumar, Disengagement detection in online learning using log file analysis. International journal of computer technology and applications 9(27), 195–301 (2016)
28. A.K. Vail, J.B. Wiggins, J.F. Grafsgaard, K.E. Boyer, E.N. Wiebe, J.C. Lester, The Affective Impact of Tutor Questions: Predicting Frustration and Engagement (International Conference on Educational Data Mining, Raleigh, 2016b)
29. Data downloaded from: https://analyse.kmi.open.ac.uk/open_dataset
30. V. N. Vapnik, IEEE Trans. Neural Netw., 10, 988–999 (1999). An Overview of Statistical Learning Theory.
31. OvidiuIvanciuc, Applications of Support Vector Machines in Chemistry, *Rev. Comput. Chem.* 2007, 23, 291-400Gareth James, An Introduction to Statistical Learning: with Applications in R. 2017 Edition, Springer.

AUTHORS PROFILE

C.S.Sasikumar, having around 27 years of experience in computer application technology and also I am a Research Scholar in Bharath Institute of Higher education and Research, Chennai. My research interest includes Soft Computing, Cloud Computing, Machine Learning, Pervasive Computing, Mobile application development and Enterprise application development.

Dr Kumaravel is working as a Professor and Dean, School of Computing, Bharath Institute of Higher education and Research, Chennai. My research interest includes Soft Computing, Cloud Computing, Machine Learning, Pervasive Computing and Knowledge Engineering. I am a life Member of ISTE and IET.