INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2020.14.51 Volume 14, 2020

APTITUDE Framework for Learning Data Classification Based on Machine Learning

Adelina Aleksieva-Petrova[1], Veska Gancheva[1], Milen Petrov[2]
[1] Faculty of Computer Systems and Technologies, Technical University of Sofia,
8 Kliment Ohridski Blv., Sofia
[2] Faculty of Mathematics and Informatics, Sofia University
5 J. Boucher Blv., Sofia
Bulgaria

Received: April 30, 2020. Revised: June 10, 2020. Accepted: June 30, 2020. Published: July 21, 2020.

Abstract—Learning analytics refers to the machine learning to provide predictions of learner success and prescriptions to learners and teachers. The main goal of paper is to proposed APTITUDE framework for learning data classification in order to achieve an adaptation and recommendations a course content or flow of course activities. This framework has applied model for student learning prediction based on machine learning. The five machine learning algorithms are used to provide learning data classification: random forest, Naïve Bayes, k-nearest neighbors, logistic regression and support vector machines.

Keywords—averaged perceptron, classification, learning analytic, machine learning.

I. Introduction

LEARNING analytic (LA) is a research field which gives new challenge in order to analyzing a large amount of learning data produced by students as effectively and meaningfully as possible. In recent years, many research works have been discuss LA and proposed different approaches in learning processes. It has strong potential for learning dispositions to be used in combination with learning analytics trace data to provide better predictions and intervention handles for students at risk of failure in both the short and long term [1].

Some authors address the relationship between LA and research areas of technology enhanced learning, such as academic analytics, action research, educational data mining, recommender systems, and personalized adaptive learning. They reviewed literature related to this field and defined that centralized web-based learning systems represent the most widely used data source for LA and the most frequently used LA techniques are classification and prediction [2].

A. Aleksieva-Petrova is with the Computer Systems Department, Faculty of Computer Systems and Technologies, Technical University of Sofia, Bulgaria (e-mail: aaleksieva@tu-sofia.bg).

V. Gancheva, is with the Programming and Computer Technologies Department, Faculty of Computer Systems and Technologies, Technical University of Sofia, Bulgaria (e-mail: vgan@tu-sofia.bg).

M. Petrov is with the Department of Software Engineering, Faculty of Mathematics and Informatics, Sofia University Bulgaria (e-mail: milenp@fmi.uni-sofia.bg).

At the other side the study indicate that students perceive value of learning analytics features in terms of learning but they were also concerned that learning analytics might be invasive or reduce their autonomy in terms of how to learn [3].

One area of impact of learning analytics is the usage into the recommendation systems. Knight et al. give recommendations for implementing learning analytics for learning impact, such as a focus on impact on learning through augmentation of existing practice; the centrality of tasks in implementing learning analytics; the commensurate centrality of learning in evaluating learning analytics; inclusion of co-design approaches in implementing learning analytics across sites; and an attention to both social and technical infrastructure [4].

In every education organization there is some Learning Management System (LMS) which support the main learning process. LMS produce and generate data from different activities and resources, as learners' activities logs, courses' activities logs, course contents, or student assessments results. There is an urgent need for analysis of these data in order to achieve better student results. The synergy between Big Data, learning analytics, and knowledge management plays growing roles to through the adaptive and personalized learning, educational data mining, data visualization, visual analytics and give better inform higher education officials and teachers [5].

In this paper we addressed to create a model of most used student and teaching activity in order to help the teacher build adaptation and recommendations a course content or flow of course activities. The paper proposed a multilayer framework based on that model which used different machine learning algorithms for LA.

The paper is structured as follows. Section II reviews the related literature. Section III presents APTITUDE framework for adaptation and recommendation based on learning analytics which is validate the learning data classification using machine learning (ML). Section IV is focused on design of algorithm for student learning prediction based on machine learning. The next section describe learning data classification using different machine learning algorithms. The experimental results and analyses are explained in section VI and finally the conclusion.

II. Related Works

In general, most data mining techniques are well suited for LA. A survey shows that the major data mining techniques of clustering, association rule, visual data mining, statistics, and regression are commonly used and some techniques, such as sequential pattern mining, text mining, correlation mining, outlier detection, causal mining, and density estimation, are not commonly used due to the complexity in obtaining the attributes necessary to regulate or adapt to individual needs [6]. The paper covered the most relevant studies (402 articles) describing applications of educational data mining and LA in higher education in order to provide opportunities and solutions to various learning problems related to four main dimensions: computer-supported learning analytics, computer-supported predictive analytics, computer-supported behavioral analytics, and computer-supported visualization analytics from 2000 till 2017 [6].

LA addresses a number of challenges, such as handling increasing data volume, finding meaningful metrics and appropriate information visualization [2]. The research work has shown how useful the application of data mining techniques in course management systems can be for online instructors according a visualization techniques to obtain a general view of the student's usage data. These techniques can use separately or applied together in order to obtain interesting information in a more efficient and faster way, such as apply clustering techniques in order to obtain the exact groups students can be divided into; use of classifier which shows what the main characteristics of the students in each group are; and apply association rule mining to discover if there is any relationship between these characteristics and other attributes [7].

The number of publications study the learning analytics and how it is apply in recommendation systems. Some authors propose the approach uses the combination of two clustering technique: Simple K-means and association rule algorithm – Apriori and finds the result. That method is applied in Course Recommendation System which helps the students to select proper course combination according to their interests [8]. The other algorithm is an iterated local search-based algorithm to solve the personalized itinerary recommendation problem, which has been beforehand formulated as a variant of the Orienteering Problem contains user interests and visit durations [9]. The new multi-constraint learning path recommendation algorithm, based on knowledge map, is generated by combination of the domain knowledge structure and cognitive structure of the learners [10]. The study has proposed and verified an approach to overcome the different learning path preferences of e-learners in four different learning scenarios according to the e-learning process.

The most used similarity metrics according to data nature and type for developing recommendation systems are: search, rating, reactive recommendations, cloud tag, proactive recommendations, crawled web pages, external domain events, domain heuristics and navigation history [11]. The authors have used for automated tag recommendation for large software information sites four new deep learning-based methods TagCNN, TagRNN, TagHAN and TagRCNN. Two of approaches TagCNN and TagRCNN achieved significant performance improvement. Training of recommendation models can be done off-line and only needs to be done once. TagCNN and TagRCNN require longer training time than TagMulRec, the overhead is acceptable for real-world practice [12].

The authors introduced a learning analytics dashboard for supporting advisers in decision-making through comparative and predictive analysis which enabled advisers to evaluate a greater number of scenarios in a similar amount of time before making a final decision, particularly for difficult cases and to increase the number of potential avenues that are evaluated with regards to the student's future development [13].

The recommendation systems usually apply in e-commerce field in order to user's satisfaction. Such example is proposed RESYGEN (REcommendation SYstem GENerator) tool which generate recommendation systems of different products such as books, video, music, clothes, and food, among others [11]. The system is operated with different data types over which a similarity metric for generating recommendation will be applied.

Moodle is one of the most wide used LMS and LA is integrated as using machine learning backend that goes beyond simple descriptive analytics and has two built-in models: students at risk of dropping out and no teaching activity [14]. That is the reason that some researchers developed a specific Moodle data mining tool and applied clustering techniques in order to obtain the exact groups students in order to classify students [15]. The classifier shows what the main characteristics of the students in each group are, and it allows new online students to be classified.

In the paper [16], authors are clustering students by mining Moodle log data. A main objectives are to define relevant clustering features and to determine if our students show different learning behaviors. The experiment executed the following clustering algorithms provided by Weka: Expectation Maximization, Hierarchical Clustering, Simple K-Means, and X-Means. The results show the method is sufficient to help identify students in trouble groups since is obtained by clustering were always neatly segregated by their activity level, and this activity level was also correlated to the grades they obtained in graded activities.

The paper proposes a hybrid recommender system implemented as a plug-in of the Moodle LMS which has implemented the approach to retrieve learning objects [17]. The recommendation strategies operate on two levels: a ranked list of Learning Objects is created, which are ordered by their correspondence to the query, and by their quality, and social generated features to show the teacher how they have been exploited in other courses.

Other researches present a learning environment formulated as a social network, including the interactions between users as

well as their relationships with the provided learning resources. They develops two ontologies: relationship between users versus resources and recommendations, and ontology to categorize the different resources [18].

III. APTITUDE SOFTWARE ARCHITECTURE

Some previous research works investigate into a software architecture for production and delivery learning resources with audio elements in university programming courses [19]. Other study create Moodle's Web Service API, which wraps up the retrieving of course details and resources, in order to deliver all available resources of a given course keeping the internal structure of course organization. The proposed solution is suitable for real-time searching of resources and learning analytics. All of them working with data from different source of information [20].

Getting knowledge to generate recommendations is usually based on data mining techniques. The main purpose of these approaches is to identify patterns by analyzing learner and teacher behaviors. The APTITUDE project is to create a solution of the gap between the lacks of simple but effective open software platform for learning and gaming analytics of big data extracted from learning modules of given course in both LMS and educational games (EGs), and applied for a learner-centric adaptation of technology enhanced learnings.

The main layers and elements in software architecture for adaptation and recommendation of course content and activities based on learning analytics is shown in Fig. 1 [21].

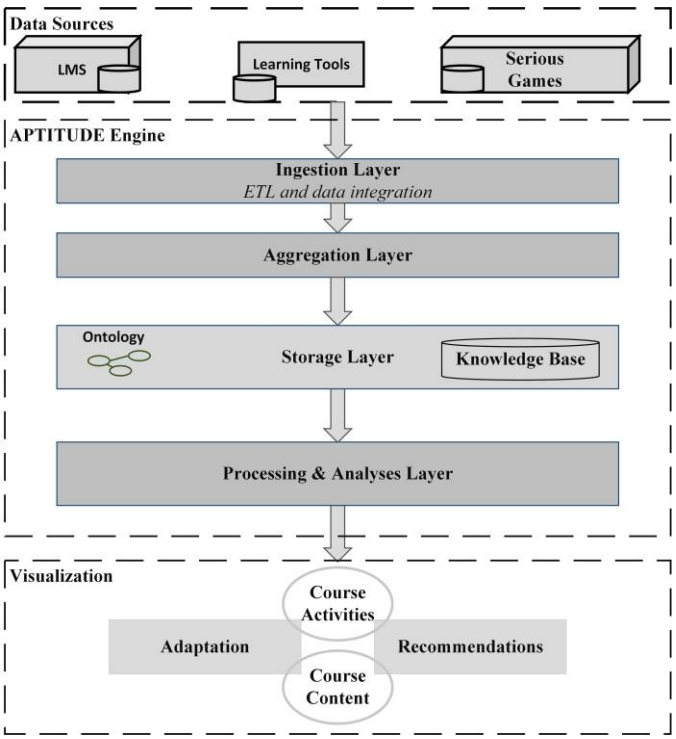

Fig. 1 APTITUDE Framework

The Ingestion Layer, Aggregation Layer and Storage Layer apply methods for retrieving, structuring and storing large volumes of data collected by LMS and educational games. The first layer (*Ingestion Layer*) is used for extract, transform, load (ETL) and data integration from different LMSs and EGs. These processes depend on the data source, for example it is a query from database or log file. The extract data should be transform in defined data format, which will help the load and data integration processes.

The second layer (*Aggregation Layer*) is responsible for compiled into data summaries according data anonymization process which is defined in [22]. The summaries data is stored into Knowledge Database in *Storage Layer*. In this layer are defined some ontology in order to extract information and support effective and timely solutions for adapting the content of the course, the workflow of its activities, and generating adaptive learning games.

The core of the software architecture is an Analytics Engine from big data *Processing & Analyses Layer* which performs the processing and analysis to make predictive modeling based on the data collected in the repository and the defined predicates in ontology. Statistical analysis, semantic analysis and text analysis is used as basic methods of data analysis.

In order for the analysis engine to provide guidance, it will use the available information base for each user as well as aggregate statistics on the behavior of a learner group. The information will be retrieved from the repository, and the machine will analyze and generate personal recommendations.

Therefore, it is important to validate different algorithms for student learning prediction based on machine learning.

IV. ALGORITHM FOR STUDENT LEARNING PREDICTION BASED ON MACHINE LEARNING

The first step in the presence of data is to examine them. During this activity it is imperative to determine how the data could be used to build the desired model and solve the specific problem. The data is also examined for problems such as those mentioned earlier in the text. There are several basic types of issues in *Ingestion Layer* (fig. 1) that will be described in the following lines.

A. Issue 1: Missing and / or incomplete data, as well as correcting them

Quite common in real data is the presence of incomplete records [23]. One or more of these fields are of no value. This lack of data completeness could lead to delay in training and inaccuracy in results. Due to the frequent presence of this problem, various ways have been developed to overcome it, depending on the data and the model being built.

B. Issue 2: Data of different scale and range and their normalization

In training neural networks, if the data is of different scales, this would lead to instability and inaccuracy in the model [24]. In order to avoid the dominance of one independent variable over another, the data themselves need to be scaled, otherwise normalized. This process is also called normalization.

C. Issue 3: Corrupted, conflicting and misleading data

Data collection is often caused by human or machine errors. As a result, the data cannot be misinterpreted by computers and this makes them unusable. Mathematical regression or clustering is often used to deal with this problem.

D. Issue 4: Categorical data

This is data whose value is in text format, not numeric format. Most machine learning algorithms use only numbers as input and output. This is due to the fact that they basically use mathematical functions and perform complex mathematical calculations.

Once the input data has been processed, it should be decided which variables to use in the model. Choosing the right input variables is at the heart of forecasting models. In recent years, with the accumulation of vast amounts of digital data, hundreds or thousands of input sets may be available to solve a problem and create a model. By filtering and selecting the most appropriate input variables, many benefits can be gained such as better understanding of the data, easier visualization, reducing the time and resources required to train and use the model [25]. Last but not least, the input variables need to be good at describing the data and carrying enough information.

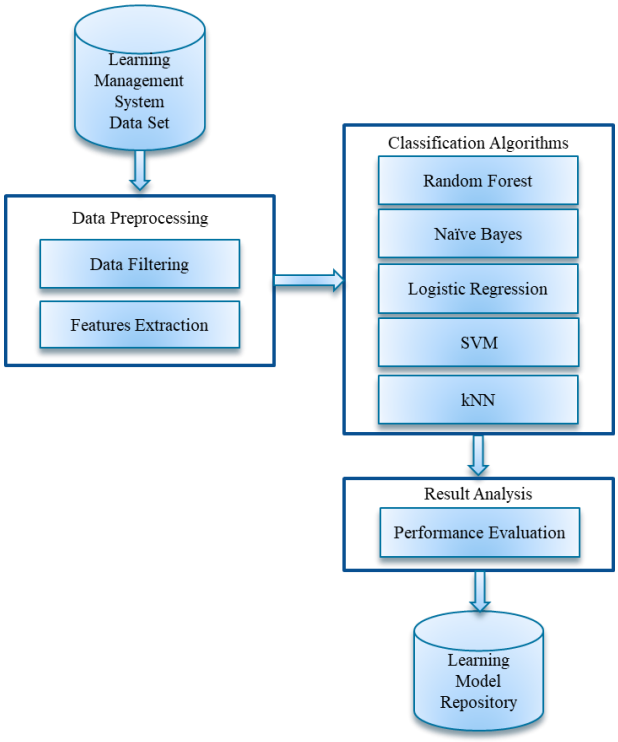

Fig. 2 Algorithm for student learning prediction based on machine learning.

The student learning prediction algorithm based on machine learning is presented in Fig. 2. That algorithm follow the processing pipeline for discovering useful knowledge from a collection of data and cover the following steps: data preparation, cleansing and selection; knowledge discovery and decision making, and comprising results and interpreting accurate solutions from the observed results. The responsibilities of the first phase (Data preprocessing) is preprocessing of the training and validating data sets and includes: data clearing in terms of accuracy; selection of functions in terms of relevance and features extraction. The aim of this phase is to establish model repository using training and validation data sets.

Analytical model is created after execution of the feature extraction and dataset reduction process and validated utilizing the validation data set. The second phase (Classification algorithms) is applying various machine learning algorithms such as classification and clustering As a result, various classification models are created and are used to build analytics workflows.

V. Learning Data Classification Based on Machine Learning

A. Data Set Selection and Preprocessing

The experimental data is obtained from learning management system. The data set contains of 63774 instances or samples characterized by 8 attributes as following: Date, Time, Event context, Component, Event name, Description, Origin, IP address. Essential of the activities logs is the Event name and how it is classified according two main groups of user (learner and teacher). The attribute Event name contains 32 different values. It is an interesting task to classify the dataset with respect to this attribute.

Essential of the activities logs are the name of event and how it is classified according two main groups of user (learner and teacher). That is the reason we select for learner point of view the following elements: course module viewed, course viewed, discussion viewed, grade user report viewed

According the teacher significant elements are: item created, course module created.

The example view of original data set is shown in Fig. 3, where each instance is represented by a row associated with the attributes value.

2/11/19,11:15,Course: Programming Languages,System,Course viewed,The user with id '7160' viewed the course with id '49'.,web,193.57.20.13

3/11/19,11:32,File: Lection,File,Course module viewed,The user with id '2' viewed the 'resource' activity with course module id '708'.,web,212.5.158.162

3/11/19,11:32,Course: Programming Languages,Activity report,Activity report viewed,The user with id '2' viewed the outline activity report for the course with id '49'.,web,212.5.158.162

Fig. 3 Example of original data set

B. Data Set Processing

The training phase is aimed to establish models repository using training data set, and applying classification machine learning algorithm. The training data set has prepared from

original data and has size 70% of them. The test are repeated 10 times. Analytical models are created after execution of the feature extraction and data set reduction process.

The attribute Event name is selected as target for the classification (Fig. 4). The attributes Time, Event context, Component and Origin are selected as features.

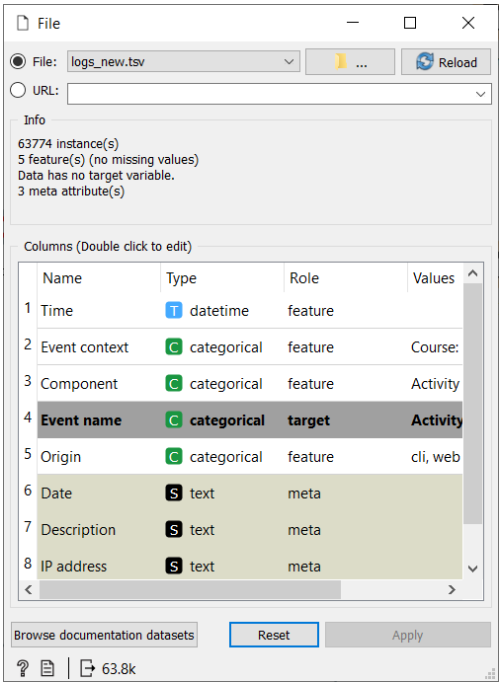

Fig. 4 Learning data set description

The proposed framework will be used different machine learning algorithms for LA. In order to validate this approach we select and use five machine learning algorithms for classification. The experiment is implemented as workflow using Orange Data Mining tool (Fig. 5) [26].

The first ML algorithm is Random forest which consists of a large number of individual decision trees and is "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [27].

The second one, the Naïve Bayes classifier, is a simple probability classifier that is based on the Bayes theorem for strong (native) independence between the individual features of the objects and is used in probability theory to calculate the probability of occurrence of an event once some of the information about it is known. Classification is performed as a result, the class with the highest probability in the presence of certain characteristics is accepted [28].

The third machine learning algorithm is k-nearest neighbors (KNN) and it was selected which "finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood" [29]. Logistic Regression is fourth ML algorithm and used to assign observations to a discrete set of classes and it based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

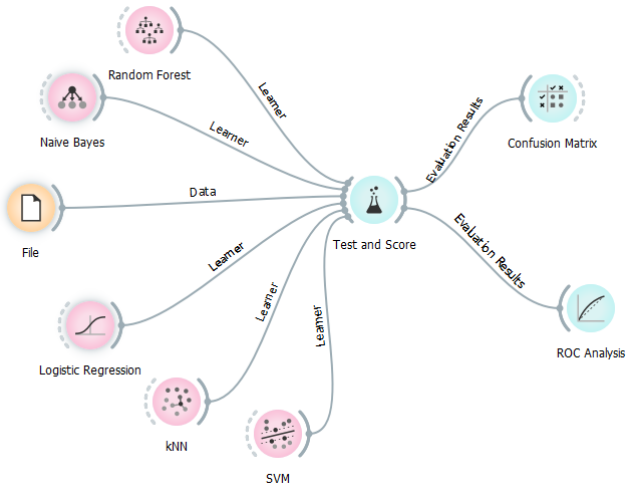

Fig. 5 Workflow for learning data classification

The last machine learning algorithm is support vector machines (SVM) which requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training. SVM find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function corresponds to a separating hyperplane that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance can be classified and belongs to the positive class [29].

VI. Experimental Results and Analysis

Precision is one of evaluation metrics of the model performance and is calculated as a ratio of true positive classified items divided by sum of true positive and false positive items in the test set. The precision range is from 0 (least precision) to 1 (most precision). The measured results for precision obtained from the selected classification algorithms are shown in Fig. 6. Best result in terms of precision is achieved through classification algorithm Random Forest: 0.934.

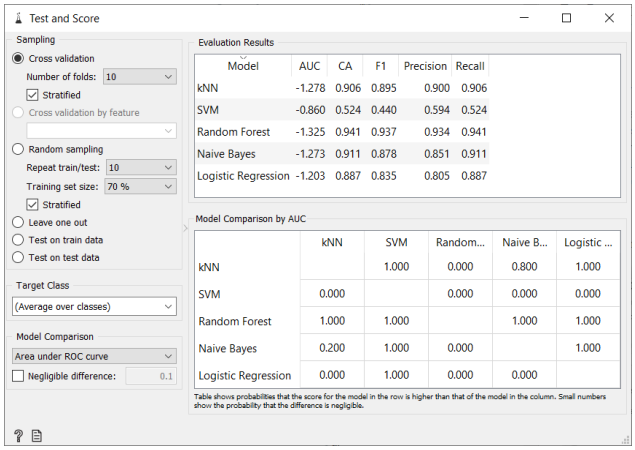

Fig. 6 Test and score of selected classification algorithms

The receiver operating characteristic (ROC) curves are used to observe the classifiers and comparison between classification models. A decision on whether to include a weak classifier in an ensemble may depend on whether its ROC curve is above or below the diagonal line of no-discrimination near 0 or 1. The ROC curves for the tested models and results of testing classification algorithms are shown in Fig. 7.

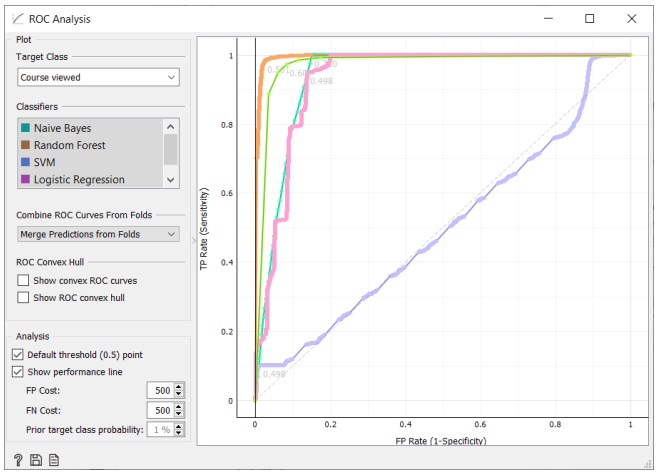

Fig. 7 ROC Analysis of the classification models

ROC curve demonstrates several things: It shows a trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left border and then the upper border of the ROC space, the more accurate the test is. The curve plots a false positive rate on an x-axis against a true positive rate on a y-axis. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the classifier.

Given the costs of false positives and false negatives, it can be also determined the optimal classifier, in this case that is Random Forest (0.934) follow by KNN (0.900). Logistic Regression (0.805) and Naïve Bayes (0.851) classifier have similar values and are considered acceptable. The SVM (0.594) is near to diagonal line and indicating not suitability for inclusion in ensembles of weak classifiers operating at thresholds close to 1.

The confusion matrix in Table I gives the number of instances between the actual and the predicted class. The matrix is useful for monitoring which specific cases are misclassified.

TABLE I. CONFUSION MATRIX FOR CASE OF RANDOM FOREST ALGORITHM

No	Value	Actual	Predicted
1	Activity report viewed	1	1
2	Calendar event created	6	0
3	Calendar event updated	1	0
4	Course module created	13	8
5	Course module deleted	11	13
6	Course module instance list viewed	7	9
7	Course module updated	95	98
8	Course module viewed	26223	26245
9	Course searched	6	4
10	Course section updated	2	0
11	Course user report viewed	44	0
12	Course viewed	30255	30653
13	Discussion created	1	0
14	Discussion viewed	88	67
15	Enrolment instance created	7	0
16	Grade overview report viewed	131	131
17	Grade user report viewed	1362	1362
18	Grader report viewed	2	2
19	Grouping deleted	6	0
20	Item created	8	8
21	Item deleted	8	8
22	Live log report viewed	1	0
23	Log report viewed	5	6
24	Recent activity viewed	163	3
25	Role assigned	1048	1146
26	Role unassigned	1032	984
27	Some content has been posted	1	1
28	User enrolled in course	1048	1015
29	User list viewed	277	406
30	User profile viewed	380	9
31	User report viewed	510	510
32	User unenrolled from course	1032	1085

Table 1 shows which values of selected learning activities are correct classified. It is obvious that this is mostly the case in the section that applies to grade reports activities. The other values must be review and evaluate in order to define which ML algorithms for student learning prediction are adequate and effective. Using results in table 1 we can also define which kind of error is admissible which depends of specific value of learning activity and role to achieve the desired goal.

VII. CONCLUSION

The real time learning and gaming analytics of big data produced by modern e-learning platforms and educational games, for a learner-centric adaptation of technology enhanced learning is one of main challenge. The paper proposed software architecture for adaptation and recommendation of course content and activities based on learning analytics. It helps to structure and storage of big data from heterogeneous sources as both LMS and educational game; identify patterns by analyzing learners' behavior and allowing data analyses with descriptive, predictive, and prescriptive results.

The student learning prediction algorithm based on machine learning for processing and analysis of data and knowledge discovery with respect to main learner and teacher activities is designed. Experimental results are presented and discussed. The experimental data set is obtained from learning management system and contains of 63774 instances characterized by 7 attributes.

As some studies [14, 15, 16, 17] we also use log files in Moodle system for analyses. Compared to some of them [14, 15] we have not classified the students rather have made the student activities prediction. For analysis [16] are used clustering algorithms provided by Weka: Expectation Maximization, Hierarchical Clustering, Simple K-Means, and

X-Means to find correlation in graded activities. The paper is implemented other five ML algorithms in order to validate their applicability for student learning prediction. As future work we can try to find correlation to grades based on learning activities.

Future work also includes designing and implementing an experiment prototype of proposed architecture. Based on different analytical models which created after execution of the feature extraction and data set reduction process, the prototype will be validate and verification the usability of the proposed architecture.

As main advantage of proposed framework is used not only data from LMSs but extract data from EGs too. For future work we will try to find intersection between different data sources (LMS and EG). Another advantage of proposed model is that data extraction from different sources will give opportunity for design and development of different interfaces (API) which will be aggregate data.

Main limitation is fact that the new data specification will be defined which will required extra data transformation according the data sources.

ACKNOWLEDGMENT

The research reported here was funded by the project "An innovative software platform for big data learning and gaming analytics for a user-centric adaptation of technology enhanced learning (APTITUDE)" - research projects on the societal challenges – 2018 by Bulgarian National Science Fund with contract №: KP-06OPR/1 from 13.12.2018.

REFERENCES

[1] Tempelaar, Dirk, et al. "Student profiling in a dispositional learning analytics application using formative assessment." *Computers in Human Behavior* 78 (2018): 408-420.

[2] Chatti, Mohamed Amine, et al. "A reference model for learning analytics." *International Journal of Technology Enhanced Learning* 4.5-6 (2012): 318-331.

[3] Schumacher, Clara, and Dirk Ifenthaler. "Features students really expect from learning analytics." *Computers in Human Behavior* 78 (2018): 397-407.

[4] Knight, Simon, Andrew Gibson, and Antonette Shibani. "Implementing learning analytics for learning impact: Taking tools to task." *The Internet and Higher Education* 45 (2020): 100729.

[5] J. Liebowitz, "Thoughts on recent trends and future research perspectives in big data and analytics in higher education," in *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*, Springer International Publishing, 2016, pp. 7–17.

[6] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis", *Telematics and Informatics*, vol. 37. Elsevier Ltd, pp. 13–49, 01-Apr-2019, doi: 10.1016/j.tele.2019.01.007.

[7] Romero, Cristóbal, Sebastián Ventura, and Enrique García. "Data mining in course management systems: Moodle case study and tutorial." *Computers & Education* 51.1 (2008): 368-384.

[8] Aher, Sunita B., and L. M. R. J. Lobo. "Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data." *Knowledge-Based Systems* 51 (2013): 1-14.

[9] Chen, Lei, et al. "Personalized itinerary recommendation: Deep and collaborative learning with textual information." *Expert Systems with Applications* 144 (2020): 113070.

[10] Zhu, Haiping, et al. "A multi-constraint learning path recommendation algorithm based on knowledge map." *Knowledge-Based Systems* 143 (2018): 102-114.

[11] Monfil-Contreras, Erick Ulisses, et al. "RESYGEN: A Recommendation System Generator using domain-based heuristics." *Expert systems with applications* 40.1 (2013): 242-256.

[12] Zhou, Pingyi, et al. "Is deep learning better than traditional approaches in tag recommendation for software information sites?" *Information and software technology* 109 (2019): 1-13.

[13] Gutiérrez, Francisco, et al. "LADA: A learning analytics dashboard for academic advising." *Computers in Human Behavior* (2018): 105826.

[14] Moodle, "Analytics - MoodleDocs," 2019. [Online]. Available: https://docs.moodle.org/36/en/Analytics. [Accessed: 13-Mar-2020].

[15] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Comput. Educ.*, vol. 51, no. 1, pp. 368–384, Aug. 2008, doi: 10.1016/j.compedu.2007.05.016.

[16] A. Bovo, S. Sanchez, O. Heguy, and Y. Duthen, "Clustering moodle data as a tool for profiling students," in *2013 2nd International Conference on E-Learning and E-Technologies in Education*, ICEEE 2013, 2013, pp. 121–126, doi: 10.1109/ICeLeTE.2013.6644359.

[17] De Medio, Carlo, et al. "MoodleREC: A recommendation system for creating courses using the moodle e-learning platform." *Computers in Human Behavior* 104 (2020): 106168.

[18] Khaled, Abdelaziz, Samir Ouchani, and Chemseddine Chohra. "Recommendations-based on semantic analysis of social networks in learning environments." *Computers in Human Behavior* 101 (2019): 435-449.

[19] Milen Petrov, Asen Asenov, Adelina Aleksieva-Petrova, "A as in Audio: Facilitating the Automatic Generation of Audio Lectures ", *Proceedings of the International Conference on E-Learning* in the Workplace New York, NY, USA June 15-17, 2016 (ICELW), editor: David Guralnick, Ph.D., 2016

[20] Petrov M., Aleksieva-Petrova A., Design Of Rest Client Architecture For Course Resources Download And Package, *10th International Technology, Education and Development Conference*, editors: L. Gómez Chova, A. López Martínez, I. Candel Torres, IATED Academy, 2016, pp.6513-6521, doi:doi:10.21125/inted.2016.0535

[21] Adelina Aleksieva-Petrova, Veska Gancheva and Milen Petrov "Software Architecture for Adaptation and Recommendation of Course Content and Activities Based on Learning Analytics" in *Proc of Int. Conf. on Applied Mathematics & Computational Science,* Venice, Italy, March 21-23, 2020, to be published.

[22] Aleksieva-Petrova, A., I. Chenchev, and M. Petrov. "LMS Data-Collection, Processing and Compliance with EU GDPR.", EDULEARN19 Proceedings, IATED, ISBN: 978-84-09-12031-4 / ISSN: 2340-1117, 2019

[23] Lean Yu, Shouyang Wang, and K. K. Lai, "An integrated data preparation scheme for neural network data analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 217–230, Feb. 2006, doi: 10.1109/TKDE.2006.22.

[24] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003, doi: 10.1162/153244303322753616.

[25] S. I. Gallant, "Perceptron-Based Learning Algorithms," *IEEE Trans. Neural Networks*, vol. 1, no. 2, pp. 179–191, 1990, doi: 10.1109/72.80230.

[26] Orange Data Mining, [Online]. Available: https://orange.biolab.si/

[27] L. Breiman, "Random Forests." *Machine Learning*. 45 (1): 5–32. doi:10.1023/A:1010933404324.

[28] A. Aleksieva-Petrova, M. Petrov, P. Georgonikos, "Web application for document classification with Naïve Bayes Algorithm." in *Proceedings of the International Scientific Conference Computer Science'2018*, Kavala, Greece, 2018.

[29] Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and information systems* 14.1 (2008): 1-37.