National College of Ireland

Predicting knowledge level of learners using Machine Learning Algorithms

MSc Research Project
MSc Data Analytics

Sachin Belagur Ramesh
Student ID: x17170834

School of Computing
National College of Ireland

Supervisor: Dr Anu Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sachin Belagur Ramesh

Student ID: x17170834

Programme: Msc Data Analytics **Year:** 2019

Module: Research Project

Supervisor: Dr Anu Sahni
Submission Due Date: 12/08/2019

Project Title: Predicting knowledge level of users using Machine Learning Algorithms
Word Count: 7646 **Page Count** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting knowledge level of learners using Machine Learning Algorithms

Sachin Belagur Ramesh
X17170834

Abstract

The evolution of Technology in the last decade has witnessed ground-breaking innovations and one of the major innovations being in the e-learning landscape. Over the years we have witnessed how online learning has morphed into an easily accessible learning platform. Currently with extensive research on areas concentrated on offering accurate and tailored information to the user this area has seen plenty of new developments and has promising returns on providing quality and targeted learning. This research aims at understanding the user of an e-learning platform which is essentially vital to deliver accurate and quality learning in this age where the attention span of humans has seen a drastic drop and the online world is driven by distractions of a million types. At the stage of discovering insights this research uses a set of machine learning algorithms pre-dominantly being classification algorithms, which predict the users' level of knowledge and the resulting outcome. Linear discriminant Analysis, classification and Regression trees, Support vector machine, Random forests and K-Nearest Neighbor have been used to mine for insights. On a high level this research found that Random Forest algorithm was more efficient and accurate in comparison to other algorithms, achieved a 62% accuracy in predicting the knowledge level of the user.

Keywords: E-Learning; Data correlation Analysis; Machine Learning Classification Algorithms

1 Introduction

Integrating technology for gaining knowledge purpose has changed learning process of people. There are multiple e-learning environment available for people to choose courses and understand the concepts in their own pace. With millions of people enrolling for these platforms producing massive data. The data is produced from electronic devices like Cell phones, Tabloid, Laptops, Computers and sensors. Advantage of these online platforms is easy to connect and all they require is internet. These massive online open course platforms recommend course for the user based on his information to given to the platforms and it also provides opportunity for users select course based on their requirement (Moubayed, Injadat, Nassif, et.al, 2018). The information produced from these environments is massive quantity is due to several reasons people need to create account for the accessing these courses and they need to provide information mail and qualification of the user (Huang, Chen, Tzeng, et. al, 2018). The key feature of these e-learning systems is assessing the knowledge of the learner because on e-learning systems there is no instant communication between the user and tutor it creates hurdle for user to not progress further since user needs to clarify about his or her doubts before jumping into next chapter. Furthermore, these platforms require to moderate the course syllabus with the perspective of user rather than dumping with unnecessary information. Series of tests taken by user does not gives us proper insight about his level of knowledge. There is need for upgrading the way assessing the knowledge level of user rather than just evaluating his excellence in exams e-learning systems should also consider amount of duration spent on the course by users.

The importance of assessing the e-learning environment is necessary for the increasing the competence of these platforms. For assessing these platforms, we need to consider key components such as users' requirement, feedback of users about course structure and necessity of collaboration on social media sites. The progression towards e-learning systems should be encouraged since it is helping many users even though there are hurdles in e-learning environment. There are multiple researches suggesting that e-learning system will be moving to mobile learning which will be more successful and affordable for users providing the services at reasonable price.

In going ahead, we need depth analysis of student's activities on these e-learning systems rather than just analysing their performance of users in tasks completion. We need to make use of AI and ML methods for analysing the student's activities on these e-learning platforms. The information accumulated regarding the actions of users are collected from various devices and examined using classification methods. In this research paper will use various ML methods for analysing the data and compare these algorithms. With comparison of these algorithms will help us identifying which is better suitable for classification. For classification we consider more input variables and increase in levels of output variable.

2 Research question

Which machine learning algorithm play a vital role of examining the knowledge level of user based on his online activities?

To solve the research question, A critical review of literature was carried out, in the next stage implementation and evaluation of the following classification algorithms Linear Discriminant Analysis, Classification and Regression Trees, K-Nearest Neighbours, Support Vector Machines and Random Forest. After implementing these algorithms, the models will be compared with their accuracy.

3 Related Work

In this section will discuss about the issues that are faced by the researchers and how they addressed the problems.

Researchers in this paper discuss about the rise in use of the e-learning environment for learning various across this platform (Zafar and Ahmad, 2006). The paper mainly concentrates on important issues faced by the instructors on these platforms are Content Authoring System and Assessment System. The main challenge for the instructors regarding Content Authoring System is that they have to develop the course content which will be relevant to the course learner. The Assessment System will provide opportunity for instructors to check the performance of learners based on continuous assessment of learners by conducting exams periodically. One more challenge discussed in this paper by researchers is about Content Delivery System. The instructors can be connected to learner using this system for constantly improving the content on the e-learning platform. Researchers have provided information about enhancing the content on the e-learning platform but they have not concluded how these technologies can be implemented.

To boost evaluation technique and providing better quality content on the e-learning system a new proposal was given by the researchers (Daradoumis, Bassi, Xhafa, et. al, 2013). The researchers in this paper mainly focus on how a software agent can help the instructors to evaluate the learner in a more efficient way and which will help in providing the better quality of education to the user. The authors also discuss about the issues that are faced by users. The users tend to lose interest in the

course if the course has many exercises or if the requires lot of time to finish the course. Authors points out that the learners are more interested in courses which provide information which are more relevant to their academic rather than courses which provide lot of information which is of no use for them. Researchers talks about how these software agents can help the instructors in keeping progress of the learners. The limitations of this research paper these software agents need huge data for finding the learners patterns.

Researchers introduces new method which is concept based learning (Nair, Archana, Chatterjee, et. al, 2015). The researchers discuss about the concept-based learning method which would increase the efficiency of instructors teaching the course on e-learning platforms. The concept-based method uses the feedback provided to the instructors on these platforms for enhancing their teaching methods in turn the quality of the content on these platforms will be improved. Researchers talks how this method can help in finding where student in stuck in the course. The findings can help the instructors to provide required feedback to students. These feedbacks can help the student to perform well in the upcoming exams. In this research we see that performance of learners was improved based on constant feedbacks. However, the limitations in this research was remarks was uploaded to the system manually.

Authors in this research paper talks about how e-learning platforms such as Coursera, edx and udemy update course content on a periodic basis and how the content is relevant to industry standards (Pang, Wang and Wang, 2014). The authors also discuss about these e-learning platforms by comparing the same course provided on these platforms and how they are beneficial to the learner. Outcome of this research paper indicates that how the learners who have enrolled for the courses on these platforms are benefitted and data from anonymous users is also considered for the experiment purpose. The major outcome of this paper is that how these e-learning platforms are able to enrol new students to the course. The drawback in this paper is that the authors more focused on comparing the e-learning platforms rather than how the content can be enhanced.

Researchers in this paper introduces the framework developed for the university which will help in enhancing the course syllabus of the modules like Big Data, Data analytics and other modules (Demchenko, Gruengard and Klous, 2014). The authors also suggest how integrating these modules with cloud computing modules will be beneficial for students in learning the modules above mentioned. The main drawback of this paper is that the proposed method is still developing stage not yet fully implemented.

(Gauthier, 2013) Researcher in this paper discusses about how teaching analytics visualizing frame work can improve the performance of the instructor evaluating techniques. The researcher elaborates about what are the hypothesis is considered in this experiment. Author also talks about strengthening the teaching on e-learning platforms and how learners can engage on these platforms using visual data analysis. The outcome of this research paper indicates that how visual data analysis will increase the standards of examining the learners and also enhance the teaching method on these mooc environment.

(Zafra and Ventura, 2009) The researchers in this paper discuss about prediction of user's performance who are enrolled for courses in traditional class room teaching environment. The researchers in this paper use multiple techniques for predicting whether the students will clear the final exam before only. Researchers have used G3P-MI technique for prediction. The factors which are influencing the learners to improve their performance will be identified using this method. The results from this experiment indicates that it is performing better than machine learning methods when compared with 74% accuracy. This will help the tutor to identify where the students are lagging and give the students more attention towards.The constraint in this experiment that it is

applied to classroom teaching environment and its yet to be applied on e-learning environments for checking the performance of the users who are enrolled on these environments.

The researchers in this paper investigated how the tutors of the e-learning environment are able to simplify the concepts of subjects so that users can easily grasp the content of the subjects (Mestadi, Nafil and Touahni, 2015). The researchers focused on identifying the challenges of users such as how users could not effectively apply the knowledge gained over these platforms and also addresses what are the reasons for decrease in students learning capacity. The authors introduce new concept called 3 levels of knowledge structure and 2 methods which would eventually increase the learning capacity of students and evaluation capacity of tutors of e-learning environment. Conclusion from this experiment reveals that knowledge structure levels effectively increases the standards of content provided on e-learning environment. The setback in this research is that knowledge accumulated by learners is directly proportional to knowledge level of tutors on these platforms.

The authors in this research paper mainly focused on assessing the performance of the users on the e-learning environment (Eremin, 2014). The assessment of the users provides the tutors analysing how much students have accumulated the knowledge. For assessing the students, the researchers have used interrelation's method which includes remarks given by the tutors to students and quantitative benchmark. Outcomes of this paper suggests that interrelation's method is instrumental in analysing the students who have finished the course exams. This method also suggests better than existing evaluating methods for checking student performance and it can be applied to any courses.

(Tuparov, Kosradinova and Raykova, 2014) The major attribute on any e-learning environment is efficiency of providing relevant course structure to the user and it's also one of the toughest challenges for e-learning system providers. This research paper involves examining the efficiency of the e-learning systems and authors focuses on components of these systems. The authors talk about the benefits of these components and play major role in evaluating efficiency of the different e-learning systems. The major drawback in this paper is that they don't highlight how can they improve efficiency of evaluating techniques.

The author in this paper majorly concentrated on understanding the behaviour of users and recommending relevant courses of his/her area of interest (Peng, 2008). To provide relevant courses it depends on the ability of e-learning system providers. The effectiveness of e-learning environment can be measured on the basis of whether the these can provide students friendly user interface, high quality of videos and course content which needs to be engaging for the users. The author introduces two agents which are Learner Interface and Content Pushes agent which would eventually help in raising standards of course syllabus and teaching methods on mooc environment. Results of this paper advocates that course syllabus will be provided on individual preferences and this technology will regulates recommending unnecessary content to the users.

The authors in this paper explore how the traditional class room colleges are integrating with e-learning environment and how these online learning systems have changed the way of accumulating knowledge (Ping, Yanni, Jinping, et. al, 2009). The authors also mention importance of these environments and hurdles that are faced by e-learning course providers. The most challenging thing on these platforms is providing course content is solely depends on nature of the user learning interest. This paper particularly concentrates on strengthening the personality learning process and for this purpose the authors make use of net learning method. The outcomes of this paper show that when technology meets traditional classroom teaching will help personalized service to the students and the limitation in this paper does not indicates whether integrating will help the students to excel in the exam.

In this research paper introduces us to the key features that are necessary for boosting the content on the mooc environment (Hariri and Amoudi, 2013). The course that are available on these mooc environment should give overview about course syllabus and instructions about who can take up these courses. The content of the course should be structured in a such a way that it should be helpful in achieving the objective of the course. The course content should be authenticated and researchers provides insight how it is authenticated. The researchers in this paper developed the courses for the KFUPM university and content of the course was subjected multiple validations before it was loaded on the university online platform. The quality of the content was regulated on regular basis and met the global standards of other e-learning platforms. The drawback in this paper is that they have not talk about obstacles they faced once it was developed.

(Zahra,2013) The author in this paper proposes in developing a Decision Support System which would provide correct condition of the system. The author elaborates about with introduction of latest technology how the consumption of knowledge has changed over the years from classroom to online. The authors also talk about reporting mechanism provided by mooc environment for understanding position of the students to online tutors. They have used data mining techniques for increasing the standards of course syllabus. Results from this paper concludes that the evaluation techniques used in this experiment provides the online tutors to give users necessary comments for improving their performance in early stages. So, in this paper the author is more focused analysing the comments of the online tutors which would eventually affect the users.

The authors of this papers predominantly focus in raising the standards of delivering styles on these e-learning systems (Lu, He, Qin, Jiang, et. al, 2006). For raising the standards, the researchers have proposed a method called multimedia course ware. The method was introduced for delivering course biomedicals materials where they have used more than 2000 slides and flash movies were optimized. This course was introduced in foe university for 2 years where the content of this course developed with cautious including only needed information. Outcome from this paper has shown good sign that it helps in raising the standards of the online tutors delivering style. The setback of this research paper it did good job with university e-learning system but they haven't discussed how could it can be applied for global e-learning systems.

The main objective in this paper the authors discusses about using Intelligent Agents techniques for developing the course structure of the e-learning systems (Elghibari, Elouahbi, Elkhoukhi, et. al, 2015). The paper also talks about how e-learning environment used for developing new technical and non-technical skills. The authors mention about tackling challenges like needs of regularly updating course structure. It is very hectic job for e-learning environment providers to do this if they do not update the users are at the verge of missing relevant knowledge and e-learning environment providers also at the risk of losing valuable customers from business point of view. The outcome of this paper indicates that intelligent agents provide necessary step for automatically updating relevant content to the course.

The content on the e-learning needs to be supervised on a regular basis so that users don't miss out on latest information regarding those subjects which are very essential (Kasim and Gunawan, 2012). In this paper the technique which is used for delivering the course to the students is problem-based learning technique. For tackling this issue, the authors make use of VLCMS which primarily concentrates finding correlation between students' performance and structure of the course and they also focus on cutting direct relationship between students and tutor by upgrading content of the course well in advance. The outcome of using VLCMS method showing positive sign in regulating the course content. In this research the everything depends on the user perspective not from the online tutor perspective.

(Santur, Karakose and Akin, 2006) This paper discusses about the progress of e-learning systems from the beginning to now and how e-learning systems have increased their efficiency in providing content to the user. The paper involves examining the user's requirement by making use of ML and Big data techniques. The paper considers attributes of users for analysing his activities on e-learning system. Outcomes from this paper indicates for analysing the information provided by e-learning systems we need to use ML and Big data techniques and it would not be possible for analysing using existing traditional techniques.

(Hassine, Marinca, Minet and et. al, 2016) In this paper the researchers mainly focused on enhancing information provided through videos. For enhancing purpose, the authors use ML methods and ML method used for experiment is Delivery Network technique. The Delivery Network method will predict which videos will get more views. The outcomes of this paper are compared with multiple forecasting methods for evaluation. The paper did not provide information will prediction accuracy be same when cached stored near end user is considered.

(Ghatasheh, 2015) Researcher in this paper focuses on requirement of assessing the knowledge level of users and providing courses which are relevant to his knowledge level. For assessing the knowledge level of users, the author makes use of multiple classification methods for predicting the user level of knowledge. The outcomes in this paper tell us that SVM is capable of providing better accuracy then other algorithms. The major drawback in this paper is that features considered for the experiment is limited to 6.

4 Research Methodology

There are two methodologies available for data mining research which are KDD and CRISP DM. For my research purpose I have considered the KDD process as this is more appropriate for my research which can be seen at Fig 1 (Fayyad et. al, 1996).

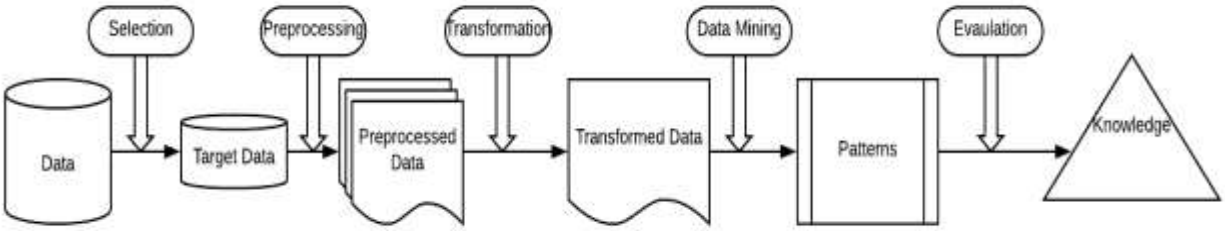

Fig 1. Overview of KDD Process

- **Selection:** In this step we will be concentrated on subset of attributes or will be involved in generating target data upon which we are to perform the exploratory analysis.

- **Pre-Processing:** In this step, in order to get accurate data, we clean the data and pre-process that dataset.

- **Transformation:** In this step using transformation/dimension reduction techniques we will be transform the data based on the requirement.

- **Data Mining**: In this step to obtain the patterns we make use of data mining techniques.

- **Evaluation:** In this step after finding the patterns we will be evaluating or interpreting them.

5 Design and Specification

In this research will mainly concentrates on evaluating knowledge level of users. For assessing the user knowledge, we consider his activities on these e-learning systems number of days active, number of times videos played, grade, number of events participated etc and we will be considering more input variables for checking the accuracy of algorithms. A document will be produced which gives overview of the features which represent the input variables and output variables.

The flowchart is developed which will help in meeting the targets of these research. The flow chart is mainly divided into five parts which can be seen from Fig 2.

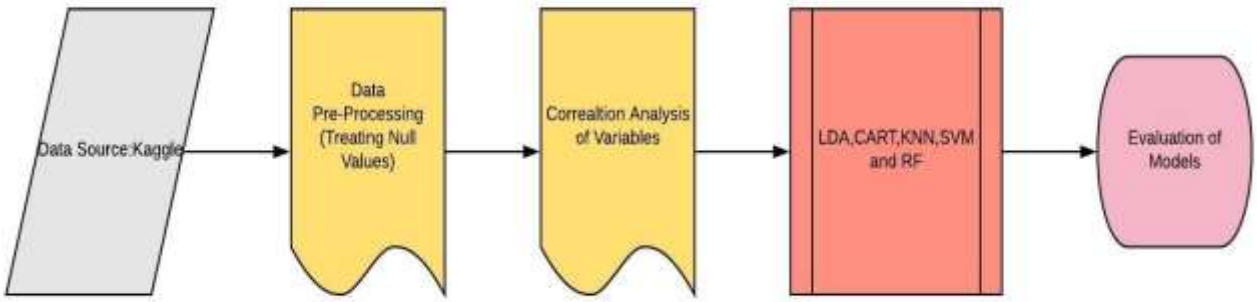

Fig. 2 Work Flow chart

6 Implementation

Data Source: The data for this research project has been downloaded from Kaggle.com website which is an open source platform.

Data Pre-Processing: For pre-processing data was loaded on R studio. I have used R programming language where library 'dplyr' was installed and loaded for data manipulation which was used for dropping columns and rearrange the required columns. The null values were treated using na.omit function. The data was divided into user attributes and user qualification attributes. The user attributes were related his activities on the e-learning system and user qualification attributes were his education background. The input variables are nevents, ndays, nplay video, nchapters, nforum posts, incomplete flag and age whereas output variable is divided into five classes as Bachelors, Doctorate, Less than secondary, Masters and Secondary.

Attributes	Information
nevents	Number of events participated by user
ndays	Number of days active on course
nplay video	Number times of videos viewed by user
nchapters	Number of chapters completed by user
nforum posts	Number of posts by user
Incomplete flag	Course completed by user or not
age	Age of the users
LoE DI	Level of education (output variable) 1.Bachelors 2.Doctorate 3.Less than Secondary 4.Masters 5.Secondary

Fig 3. Input and Output Variables

Data Correlation Analysis of Variables:

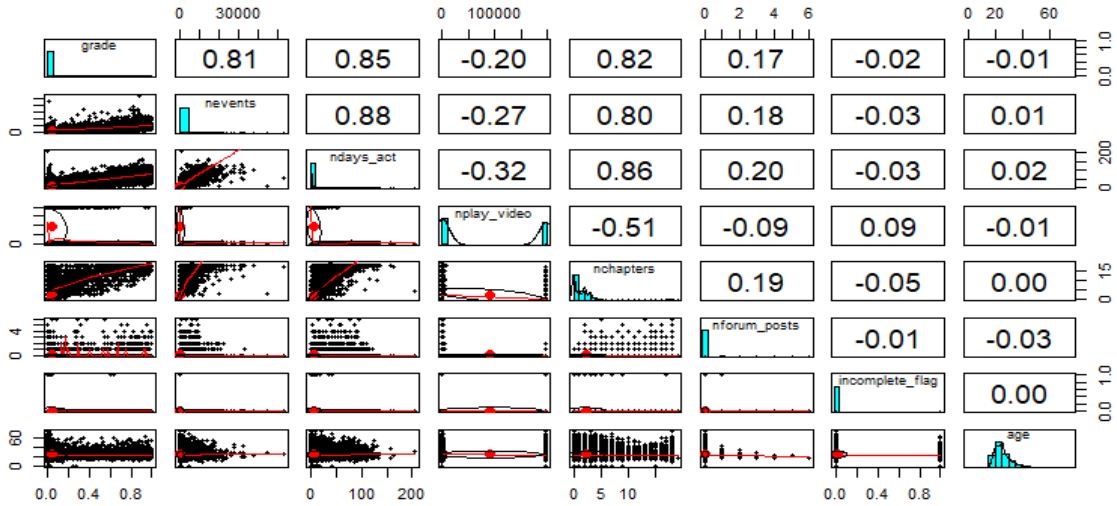

Fig 4 Correlation between the variables

To find the correlation among the input variables I have installed and loaded 'pysch' library which is available in R studio. This library helps us to determine the correlation between more than two variables since we have eight variables it was very appropriate use this library. This library helps in finding the personality of the e-learning system users. From the Fig 4 we can see the correlation between the variables shown using histogram, scatter plot in the same graph.

After finding the correlation between the variables the dataset was divided into 75% and 25%. The 75% of the dataset which will be used for training and testing the models and 25% of the data set will be used for validation purpose. The division of the dataset will be helpful in finding which algorithm gives the highest accuracy since the it will be evaluated on the masked data. In this research we will be using 75% of instance for classifying using the developed methods. For

examining the performance of the models will be using confusion matrix which depends on the various measures. The measures which are considered for evaluating performance of algorithms are Total Accuracy, Sensitivity, Specificity.

- Total Accuracy = TP + TN / TP + TN + FP + FN
- Sensitivity = TP / TP + FN
- Precision = TP / TP + FP

	Predicted Class	
ACTUAL CLASS	TRUE	FALSE
TRUE	TP	FN
FALSE	FP	TN

Fig 5 Confusion Matrix

Implementing machine learning algorithms:

For running machine learning algorithms, I have installed and loaded 'caret' package. This library provides numerous machine learning algorithms using which we can tune the parameters based on requirement, we can use for visualization purpose and data manipulation. We will be creating the models and predict their accuracy on masked data. Our data will be divided into 10 slices where 9 will be used for training and remaining one will be used for testing. In order to get better accuracy this process will be performed three times with various splits among the 10 batches.

Implementation of Linear Discriminant Analysis (LDA)

Linear Discriminant analysis is a statistical machine learning method to solve numerous classification issues and it has been for human behaviour recognition. The LDA has been very effective getting discriminant features. In LDA each instance will be represented by its corresponding mean and differentiate among the instances which is based distribution of the instances with respect mean of total data. Linear Discriminant analysis formula is given by (Xu, L, et. Al, 2018).

$$J(W) = \max W (tr (WT SBW) / tr (WT SWW))$$

SB = within the classes and SW = Between class scatter matix

Linear Discriminant Analysis is executed by dividing the dataset into 75 % and 25% where 75% of data is used for training and 25 % for testing purpose. Th library 'caret' has been used for implementing linear discriminant analysis algorithm. The model will be trained on the model which is developed using numerous features such as ndays, nevents, nplay video and other features as well. In order get better accuracy we are using 10-fold cross validation where data will be sliced into 9 slices for training 1 slice for testing and for evaluation of model metric accuracy has been used. The output of the LDA has been shown in fig 6.

```
Linear Discriminant Analysis

43467 samples
    8 predictor
    5 classes: 'Bachelor's', 'Doctorate', 'Less than Secondary', 'Master's', 'Secondary'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 39120, 39121, 39121, 39119, 39120, 39122, ...
Resampling results:

  Accuracy   Kappa
  0.5364763  0.1881424
```

Fig 6 Output of Linear Discriminant Analysis.

The accuracy produced by Linear Discriminant Analysis is with all the attributes is 53%.

Implementation of Classification and Regression Trees (CART)

Classification and Regression Tree is basically classified as binary tree and the main principal of this algorithm is Gini index which is important from selection of attributes at nodes since it acts as criterion. Dirtiness of the selected attributes will be assessed by Gini coefficient between 0 and 1. If the value is less then dirtiness of the selected attributes will be less. The Gini coefficient formula is given by (Xiao, B, et. al, 2018).

$$Gini = 1 - \Sigma (Pk)^2 \text{ where } k = 1$$

K = number of categories and Pk = Probability of the Kth Category

Classification and Regression Trees is executed by dividing the dataset into 75% and 25% where 75% of data is used for training and 25% of data is used for testing. The library 'caret' has been used for implementing the CART algorithm. The model will be trained on the model which is developed using several attributes. From the point of getting good accuracy 10-fold cross validation where data will be divided into 9 slices which is used for training and 1 slice of data for testing. For evaluation of developed model metric accuracy is considered. The output of Classification and Regression Trees shown in the fig 7.

```
CART

43467 samples
    8 predictor
    5 classes: 'Bachelor's', 'Doctorate', 'Less than Secondary', 'Master's', 'Secondary'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 39120, 39121, 39121, 39119, 39120, 39122, ...
Resampling results across tuning parameters:

  cp           Accuracy   Kappa
  0.001729655  0.5729181  0.2597403
  0.002680965  0.5708473  0.2529561
  0.192207905  0.5170740  0.1217930

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.001729655.
```

Fig 7 Output of Classification and Regression Trees.

The accuracy produced by the Classification with all the features considers is 57%. The accuracy of the model has seen varying with respect to the value of cp. Higher the cp value accuracy is reduced. When cp is 0.19 the accuracy of the model is 51% but when the cp value is lowest 0.00172 the model produced the highest accuracy of 57%.

Implementation of K- Nearest Neighbour (KNN)

K-Nearest Neighbour algorithm works on basis of nearest neighbour. KNN algorithm is widely used for classification purpose and it does not make any presumption about the dataset. It is very powerful and training of data will be will be fast in KNN. The distance between the attributes calculated by Euclidean distance as shown

$$Dist(p,q) = \sqrt{(p1 - q1)^2 + (p2-q2)^2 + \ldots + (pn-qn)^2}$$

n = attributes, p1 = first attribute instance of p value, q1 =first attribute instance of q value

K-Nearest Neighbour algorithm is run by dividing the dataset into 75% and 25% where 75% of data is used for training and 25% of data is used for testing. The library 'caret' has been used for implementing the CART algorithm. The model will be trained on the model which is developed using several attributes. The KNN algorithm implemented using several k values (5, 7, 9), method is knn, metric is metric and training control is control. After tuning of the parameters would give the better accuracy. The output of K-Nearest Neighbour algorithm is shown in the fig 8.

```
k-Nearest Neighbors

43467 samples
    8 predictor
    5 classes: 'Bachelor's', 'Doctorate', 'Less than Secondary', 'Master's', 'Secondary'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 39120, 39121, 39121, 39119, 39120, 39122, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.5062699  0.1844218
  7  0.5114228  0.1859402
  9  0.5163692  0.1892891

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
```

The accuracy of the algorithm varies depending the k value as seen from the output when k values is 5 it gives low accuracy of 50 % when the value of k is 9 it gives highest accuracy of 51.63 %.

Implementation of Support Vector Machines (SVM)

Support Vector Machine is a machine learning algorithm which is used for examining the data used for classification purpose. In SVM for multiple classification it uses numerous binary classifiers and it works on detecting the planes which distinguish the class accurately. SVM Linear formula is given by

$$K(Xi, Xj) = Xi * Xj$$

Support Vector Machines is run by dividing the dataset into 75% and 25% where 75% of data is used for training and 25% of data is used for testing. The library 'caret' has been used for implementing the CART algorithm. The model will be trained on the model which is developed using several attributes. The SVM algorithm is implemented using combination of attributes where method is Linear, metric for accuracy, train control is control, where c is center and scale and tune length is 10. The combination of these parameters will provide better accuracy. The output of support vector machines is shown by fig 9.

```
Support Vector Machines with Linear Kernel

43467 samples
    8 predictor
    5 classes: 'Bachelor's', 'Doctorate', 'Less than Secondary', 'Master's', 'Secondary'

Pre-processing: centered (8), scaled (8)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 39120, 39120, 39119, 39121, 39120, 39120, ...
Resampling results:

  Accuracy   Kappa
  0.4692526  0.00366275

Tuning parameter 'C' was held constant at a value of 1
```

Fig 9 Output of Support Vector Machine

The accuracy generated for support vector machine is 46% when the value of c was at constant and it was 1. The accuracy was produced for SVM when tune length was 10 and method was linear.

Implementation of Random Forest (RF)

Random Forest is an ensemble classifier which is used for classification and regression. In Random Forest setting the parameters is easy and it is not much affected by outliers. Support Vector Machines is run by dividing the dataset into 75% and 25% where 75% of data is used for training and 25% of data is used for testing. The library 'caret', 'randomforest' was installed and loaded for implementing the random forest algorithm. The model will be trained on the model which is developed using several attributes. The random forest algorithm is implemented using combination of several features where method is rf, training control is control and metric is used for measuring accuracy. The output of random forest algorithm is shown in the fig 10.

```
Random Forest

43467 samples
    8 predictor
    5 classes: 'Bachelor's', 'Doctorate', 'Less than Secondary', 'Master's', 'Secondary'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 39120, 39121, 39121, 39119, 39120, 39122, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  2     0.5738151  0.2594641
  5     0.5446668  0.2465393
  8     0.5352343  0.2371440

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

Fig 10 Output of Random Forest

The accuracy produced by random forest algorithm when the value of mtry = 2 is 57%, when mtry =5 is 54%, when mtry = 8 is 53% and mtry is a number of predictor variables. When optimising purpose model selected largest value which produced overall accuracy of 57% using metric for accuracy.

7 Evaluation and Results Discussion

To achieve the objectives of this research paper multiple experiments to examine the performance of the classification algorithms which is used for predicting level of learner knowledge. The classification algorithms which are considered for the experiments are Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), K-Nearest Neighbour (KNN), Support Vector Machines and Random Forest (RF).

The models were created using training data set which is of 75% and which subsequently tested using 25% of data set. For Linear Discriminant Analysis parameters lda method and metric for accuracy. The accuracy of the results shown in the Fig 10 are when validated on the test data set.

```
Confusion Matrix and Statistics

                    Reference
Prediction           Bachelor's Doctorate Less than Secondary Master's Secondary
  Bachelor's               6055       131                  42      2368      2986
  Doctorate                  42        17                   0        37        15
  Less than Secondary         7         0                   1         4         4
  Master's                  310        31                   0       207       149
  Secondary                 366         2                 260        16      1438

Overall Statistics

               Accuracy : 0.5327
                 95% CI : (0.5246, 0.5409)
    No Information Rate : 0.468
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1823

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bachelor's Class: Doctorate Class: Less than Secondary Class: Master's Class: Secondary
Sensitivity                     0.8931         0.093923                 3.300e-03          0.07865          0.31315
Specificity                     0.2830         0.993430                 9.989e-01          0.95867          0.93492
Pos Pred Value                  0.5228         0.153153                 6.250e-02          0.29699          0.69068
Neg Pred Value                  0.7505         0.988593                 9.791e-01          0.82416          0.74577
Prevalence                      0.4680         0.012493                 2.091e-02          0.18167          0.31695
Detection Rate                  0.4179         0.001173                 6.902e-05          0.01429          0.09925
Detection Prevalence            0.7994         0.007662                 1.104e-03          0.04811          0.14371
Balanced Accuracy               0.5880         0.543676                 5.011e-01          0.51866          0.62404
```

Fig 10 Results of Linear Discriminant Analysis

The overall accuracy produced for Linear Discriminant Analysis is 53%. For evaluation we are using total accuracy, sensitivity, specificity. The sensitivity has highest for bachelor's class which has identified 89% of positive instances. The Balanced accuracy has been highest for secondary class. The overall accuracy produced is 53% when it had five classes.

The models were created using training data set which is of 75% and which subsequently tested using 25% of data set. For Classification and Regression Trees parameters are cart method and metric for accuracy. The accuracy of the results shown in the Fig 11 are when validated on the test data set.

The overall accuracy produced for Classification and Regression Trees is 57%. For evaluation we are using total accuracy, sensitivity, specificity. The sensitivity has highest for bachelor's class which has identified 83% of positive instances and least was doctorate class with less than 10%. The Balanced accuracy has been highest for secondary class and has been same for doctorate class and master's class which 50% The overall accuracy produced is 53% when it had five classes.

```
Confusion Matrix and Statistics

                    Reference
Prediction           Bachelor's Doctorate Less than Secondary Master's Secondary
  Bachelor's               5685       173                  21      2578      2037
  Doctorate                   0         0                   0         0         0
  Less than Secondary        17         0                  69         4        32
  Master's                    0         0                   0         0         0
  Secondary                1078         8                 213        50      2523

Overall Statistics

               Accuracy : 0.5713
                 95% CI : (0.5632, 0.5794)
    No Information Rate : 0.468
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2559

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bachelor's Class: Doctorate Class: Less than Secondary Class: Master's Class: Secondary
Sensitivity                     0.8385          0.00000                   0.227723          0.0000          0.5494
Specificity                     0.3761          1.00000                   0.996264          1.0000          0.8637
Pos Pred Value                  0.5417              NaN                   0.565574             NaN          0.6516
Neg Pred Value                  0.7258          0.98751                   0.983712          0.8183          0.8051
Prevalence                      0.4680          0.01249                   0.020914          0.1817          0.3170
Detection Rate                  0.3924          0.00000                   0.004763          0.0000          0.1741
Detection Prevalence            0.7243          0.00000                   0.008421          0.0000          0.2673
Balanced Accuracy               0.6073          0.50000                   0.611993          0.5000          0.7066
>
```

Fig 11 Results of Classification and Regression Trees

The models were created using training data set which is of 75% and which subsequently tested using 25% of data set. For K-Nearest Neighbour parameters are knn method and metric for accuracy. The accuracy of the results shown in the Fig 12 are when validated on the test data set.

```
> predictions <- predict(fit.knn, validation)
> confusionMatrix(predictions, validation$LOE_DI)
Confusion Matrix and Statistics

                    Reference
Prediction           Bachelor's Doctorate Less than Secondary Master's Secondary
  Bachelor's               4964       125                  71      1950      2189
  Doctorate                   2         0                   0         2         0
  Less than Secondary         3         0                  21         3        15
  Master's                  564        40                   8       381       265
  Secondary                1247        16                 203       296      2123

Overall Statistics

               Accuracy : 0.5169
                 95% CI : (0.5087, 0.5251)
    No Information Rate : 0.468
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1932

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bachelor's Class: Doctorate Class: Less than Secondary Class: Master's Class: Secondary
Sensitivity                     0.7322          0.0000000                 0.069307          0.14476          0.4623
Specificity                     0.4376          0.9997204                 0.998520          0.92603          0.8219
Pos Pred Value                  0.5338          0.0000000                 0.500000          0.30286          0.5465
Neg Pred Value                  0.6500          0.9875035                 0.980479          0.82986          0.7671
Prevalence                      0.4680          0.0124931                 0.020914          0.18167          0.3170
Detection Rate                  0.3426          0.0000000                 0.001449          0.02630          0.1465
Detection Prevalence            0.6418          0.0002761                 0.002899          0.08683          0.2682
Balanced Accuracy               0.5849          0.4998602                 0.533913          0.53539          0.6421
>
```

Fig 12 Results of K-Nearest Neighbour

The overall accuracy produced for Classification and Regression Trees is 51%. For evaluation we are using total accuracy, sensitivity, specificity. The sensitivity has highest for bachelor's class which has identified 73% of positive instances and least was doctorate class with less than 5%. The Balanced accuracy has been highest for secondary class and least for doctorate class and master's class which 49%. The overall accuracy produced is 53% when it had five classes.

The models were created using training data set which is of 75% and which subsequently tested using 25% of data set. For Support Vector Machines parameters are linear method, c is centre and scaled at constant value of 1, tune length is 10 and metric for accuracy. The accuracy of the results shown in the Fig 13 are when validated on the test data set.

```
> confusionMatrix(predictions, validation$LoE_DI)
Confusion Matrix and Statistics

                    Reference
Prediction           Bachelor's Doctorate Less than Secondary Master's Secondary
  Bachelor's               6769       181                 303      2628      4570
  Doctorate                   0         0                   0         0         0
  Less than Secondary         0         0                   0         0         0
  Master's                    0         0                   0         0         0
  Secondary                  11         0                   0         4        22

Overall Statistics

               Accuracy : 0.4687
                 95% CI : (0.4606, 0.4769)
    No Information Rate : 0.468
    P-Value [Acc > NIR] : 0.4305

                  Kappa : 0.0022

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bachelor's Class: Doctorate Class: Less than Secondary Class: Master's Class: Secondary
Sensitivity                  0.998378         0.00000                   0.00000          0.0000          0.004791
Specificity                  0.003373         1.00000                   1.00000          1.0000          0.998484
Pos Pred Value               0.468410             NaN                       NaN             NaN          0.594595
Neg Pred Value               0.702703         0.98751                   0.97909          0.8183          0.683759
Prevalence                   0.467973         0.01249                   0.02091          0.1817          0.316952
Detection Rate               0.467214         0.00000                   0.00000          0.0000          0.001518
Detection Prevalence         0.997446         0.00000                   0.00000          0.0000          0.002554
Balanced Accuracy            0.500875         0.50000                   0.50000          0.5000          0.501638
>
```

Fig 13 Results of Support Vector Machines

The overall accuracy produced for Support Vector Machines is 46%. For evaluation we are using total accuracy, sensitivity, specificity. The sensitivity has highest for bachelor's class which has identified 99% of positive instances and same has been for doctorate class and secondary class with less than 5%. The Balanced accuracy has been highest for secondary class and has been same for doctorate class and master's class which 50% The overall accuracy produced is 53% when it had five classes.

The models were created using training data set which is of 75% and which subsequently tested using 25% of data set. For Support Vector Machines parameters are linear method, c is centre and scaled at constant value of 1, tune length is 10 and metric for accuracy. The accuracy of the results shown in the Fig 13 are when validated on the test data set.

```
Confusion Matrix and Statistics

                    Reference
Prediction           Bachelor's Doctorate Less than Secondary Master's Secondary
  Bachelor's               5932       158                  24      2267      2002
  Doctorate                   0        15                   0         0         0
  Less than Secondary         3         0                 115         0        10
  Master's                   17         0                   0       328         2
  Secondary                 828         8                 164        37      2578

Overall Statistics

               Accuracy : 0.619
                 95% CI : (0.611, 0.6269)
    No Information Rate : 0.468
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3442

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bachelor's Class: Doctorate Class: Less than Secondary Class: Master's Class: Secondary
Sensitivity                     0.8749         0.082873                  0.379538          0.12462          0.5614
Specificity                     0.4225         1.000000                  0.999084          0.99840          0.8952
Pos Pred Value                  0.5713         1.000000                  0.898438          0.94524          0.7131
Neg Pred Value                  0.7934         0.988530                  0.986908          0.83707          0.8148
Prevalence                      0.4680         0.012493                  0.020914          0.18167          0.3170
Detection Rate                  0.4094         0.001035                  0.007938          0.02264          0.1779
Detection Prevalence            0.7167         0.001035                  0.008835          0.02395          0.2495
Balanced Accuracy               0.6487         0.541436                  0.689311          0.56151          0.7283
```

Fig 14 Results of Random Forest

The overall accuracy produced for Random Forest is 62%. For evaluation we are using total accuracy, sensitivity, specificity. The sensitivity has highest for bachelor's class which has identified 87% of positive instances and least has been for master's class with around 12%. The Balanced accuracy has been highest for secondary class, second highest has been for Less than secondary class which is around 68%. The overall accuracy produced is 62% when it had five classes. By checking the performance measure table, the Random forest algorithm is surely performing better than other algorithms. Random forest identifying the negative instances up to 40% and identifying the positive instances up to 42%. Random forest also identifying correctly classified instances up to 87%. However, the support vector machine has been correctly classifying the positive instances up to 99% but overall accuracy of the support machine less than random forest and support vector machine also showing only 35 of positive instance. Random forest-based classification has been more accurate in providing the accuracy.

Algorithms	Sensitivity	Detection Rate	Specificity
RF	0.87	0.40	0.42
SVM	0.99	0.46	0.03
KNN	0.73	0.34	0.43
CART	0.83	0.39	0.37
LDA	0.89	0.41	0.28

Table 1 Performance Measures

By looking at table we can see that Random Forest is giving the highest overall accuracy when compared to other algorithms. This overall accuracy is produced with increase in the number of input variables. For this experiment we considered around 8 input variables and increased levels in the output variables up to 5. It shows that with increase in the input accuracy random forest gives us better accuracy than the other classification algorithms.

Algorithms	Accuracy
RF	62%
SVM	46%
KNN	51%
CART	57%
LDA	53%

Table 2 Accuracy of Each Model

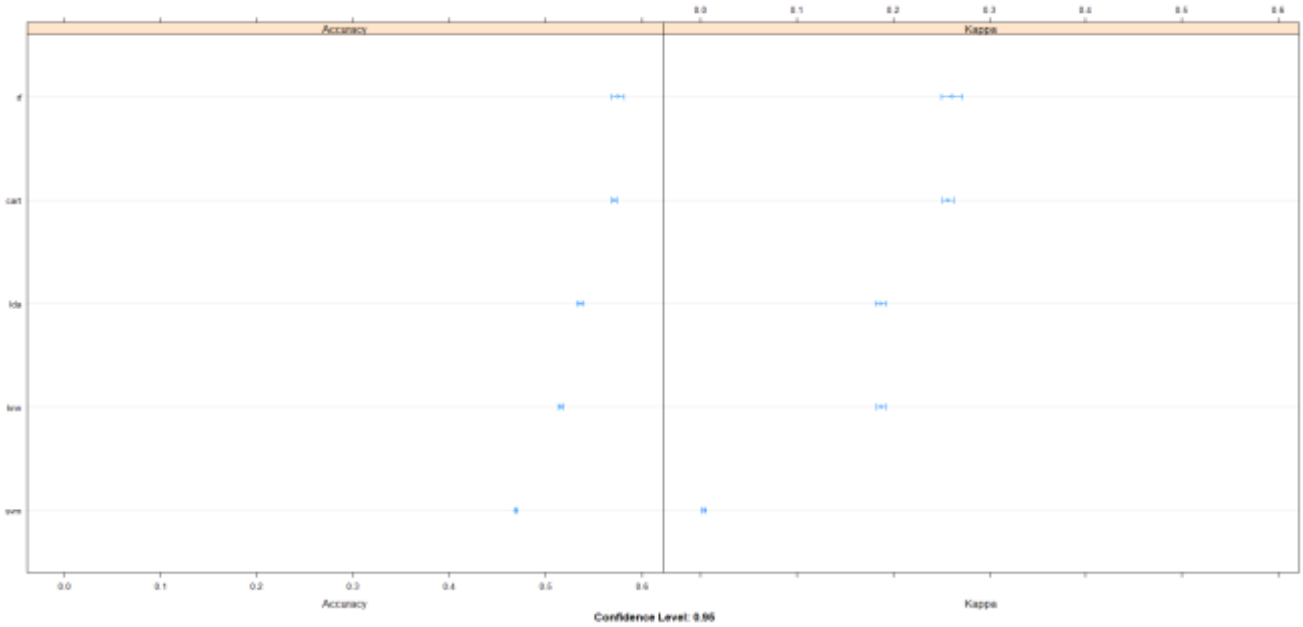

Fig 15 Comparison of Model Accuracy

From the fig 15 we can see that we comparing the accuracy of the models and the plot shows that Random First giving highest accuracy. We can also observe that classification algorithms which have high kappa value they giving better accuracy than the algorithms with less kappa value.

8 Conclusions and Future Work

The finding of this research will play a key role in identifying the classification algorithm for a better accuracy to achieving an increase in efficiency of the evaluation system in a massive open online courses environment. As part of the research a design work flow chart was introduced for examining the classification algorithms. An analysis for the correlation between the input variables which are very fundamental for the classification algorithms was done, these input variables are significant in enhancing the accuracy of the classification algorithms. In the literature review several methods have been used in examining level of user knowledge and to enhance the evaluation system on the e-learning systems. Through this research the model built with Random Forest algorithm performs comparatively better than other classification algorithms in terms of accuracy and specificity. Thus we propose based on our findings that with Random Forest classification for achieving better accuracy in aspects of evaluation of user on mooc environment. Followed by Random forest, models built with Classification and Regression Tree algorithms give better performance, however models with Support Vector Machine gave least accuracy.

For future work based on this research work where we have focused on examining the importance of input variables and this can be further followed up by introducing dimension reality. E-learning systems can also be integrated in corporate world for training their employees and evaluating their performance.

9 Acknowledgement

I would like to thank especially my supervisor Dr Anu Sahni for her constant support throughout this research with her guidance, technical inputs and motivation. I would like to thank School of Computing and National College of Ireland for providing the resource materials which were required for completing this research project. I would like to thank my family members and friends for providing constant support.

References

Fayyad, U. and Uthurusamy, R. (1996). Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11), pp.24-26.

Xu, L., Iosifidis, A., & Gabbouj, M. (2018). Weighted Linear Discriminant Analysis Based on Class Saliency Information. *2018 25th IEEE International Conference on Image Processing (ICIP)*. doi:10.1109/icip.2018.8451614

Xiao, B., Liang, M., & Ma, J. (2018). The Application of CART Algorithm in Analyzing Relationship of MOOC Learning Behavior and Grades. *2018 International Conference on Sensor Networks and Signal Processing (SNSP)*. doi:10.1109/snsp.2018.00055

Zafar, A., Ahmad, N., 2006. Towards adaptive e-learning: Technological challenges and enabling technologies, in: 2006 Annual India Conference, INDICON. doi:10.1109/INDCON.2006.302838

Daradoumis, T., Bassi, R., ... Caballé, S., 2013. A review on massive e-learning (MOOC) design, delivery and assessment, in: Proceedings - 2013 8th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2013. IEEE Computer Society, pp. 208–213. doi:10.1109/3PGCIC.2013.37

Nair, N.C., Archana, J.S., ... Bijlani, K., 2015. Knowledge representation and assessment using concept-based learning, in: 2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015. Institute of Electrical and Electronics Engineers Inc., pp. 848–854. doi:10.1109/ICACCI.2015.7275716

Pang, Y., Wang, T., Wang, N., 2014. MOOC data from providers, in: Proceedings - 2nd International Conference on Enterprise Systems, ES 2014. Institute of Electrical and Electronics Engineers Inc., pp. 87–90. doi:10.1109/ES.2014.45

Demchenko, Y., Gruengard, E., Klous, S., 2015. Instructional model for building effective big data curricula for online and campus education, in: Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom. IEEE Computer Society, pp. 935–941. doi:10.1109/CloudCom.2014.162

Gauthier, G., 2013. Using teaching analytics to inform assessment practices in technology mediated problem solving tasks, in: CEUR Workshop Proceedings. CEUR-WS.

Zafra, A., Ventura, S., 2009. Predicting Student Grades in Learning Management Systems with Multiple Instance Learning Genetic Programming, in: Proceedings of the 2nd International Conference on Educational Data Mining. pp. 309–318.

Mestadi, W., Nafil, K., & Touahni, R. (2015). Knowledge structuring for learning by level. *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*. http://doi.org/10.1109/sita.2015.7358392

Yeen-Ju, H.T., Mai, N., ... Haw, L.C., 2013. Authentic learning strategies to engage student's creative and critical thinking, in: Proceedings - 2013 International Conference on Informatics and Creative Multimedia, ICICM 2013. IEEE Computer Society, pp. 57–62. doi:10.1109/ICICM.2013.19

Eremin, E.A., 2014. New proposals about evaluation of complete e-learning course digestion, in: 8th IEEE International Conference on Application of Information and Communication Technologies, AICT 2014 - Conference Proceedings. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ICAICT.2014.7036003

Tuparov, G., Kostadinova, H., ... Raykova, M., 2014. Approaches for competencies assessment in open source e-learning environments, in: IEEE Global Engineering Education Conference, EDUCON. IEEE Computer Society, pp. 529–532. doi:10.1109/EDUCON.2014.6826143

Peng, Y. (2008). Intelligent Content Push for SCORM-based E-Learning Systems. *2008 International Symposium on Intelligent Information Technology Application Workshops*. http://doi.org/10.1109/iita.workshops.2008.81

Ping, Y., Yanni, W., Jinping, L., & Bo, K. (2009). Study on Personality Learning in E-Learning. *2009 International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government*. http://doi.org/10.1109/eeee.2009.39

Hariri, M.M., Amoudi, S.M.A., 2013. Standards and process applied in development of comprehensive online courses at KFUPM, Saudi Arabia, in: Proceedings - 2013 4th International Conference on e-Learning Best Practices in Management, Design and Development of e-Courses: Standards of Excellence and Creativity, ECONF 2013. pp. 413–416. doi:10.1109/ECONF.2013.29

Zahra, K.D., 2013. New approach to the design of decision support system to improve e-learning environments, in: 4th International Conference on E-Learning and e-Teaching, ICELET 2013. pp. 26–29. doi:10.1109/ICELET.2013.6681640

Lu Xiaoying, He Jian, Qin Tian, Jiang Dongxu, & Chen Wei. (2006). Construction of Multimedia Courseware and Web-based E-Learning Courses of "Biomedical Materials" (pp. 2886–2889). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/iembs.2005.1617077

Elghibari, F., Elouahbi, R., Elkhoukhi, F., Chehbi, S., & Kamsa, I. (2015). Intelligent e-learning system model for maintenance of updates courses. *2015 International Conference on Information Technology Based Higher Education and Training (ITHET)*. http://doi.org/10.1109/ithet.2015.7218028

Kasim, N. A. A., & Gunawan, T. S. (2012). Virtual-learning content management system for problem-based learning (PBL) courses. *2012 International Conference on Computer and Communication Engineering (ICCCE)*. http://doi.org/10.1109/iccce.2012.6271356

Santur, Y., Karakose, M., Akin, E., 2016. Improving of personal educational content using big data approach for MOOC in higher education, in: 2016 15th International Conference on Information Technology Based Higher Education and Training, ITHET 2016. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ITHET.2016.7760728

Hassine, N. B., Marinca, D., Minet, P., & Barth, D. (2016). Expert-based on-line learning and prediction in Content Delivery Networks. *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*. http://doi.org/10.1109/iwcmc.2016.7577054

Ghatasheh, N., 2015. Knowledge Level Assessment in e-Learning Systems Using Machine Learning and User Activity Analysis. International Journal of Advanced Computer Science and Applications 6. doi:10.14569/ijacsa.2015.060415

Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.2381263

Huang, N.-F., Chen, C.-C., Tzeng, J.-W., Fang, T.- T., & Lee, C.-A. (2018). Concept Assessment System Integrated with a Knowledge Map Using Deep Learning. *2018 Learning With MOOCS (LWMOOCS)*. http://doi.org/10.1109/lwMOOCs.2018.8534674

Liu, Y. (2010). An Automatic Grading Model for Learning Assessment. *2010 International Conference on e-Education, e-Business, e-Management and e-Learning*. http://doi.org/10.1109/ic4e.2010.32

Moubayed, A., Injadat, M., Nassif, A. B., Lutfiyya, H., & Shami, A. (2018). E-Learning: Challenges and Research Opportunities Using Machine Learning & Data Analytics. *IEEE Access*, *6*, 39117–39138. http://doi.org/10.1109/access.2018.2851790