Improvement of Online Course Content using MapReduce Big Data Analytics

Shehnaz Anjar Siddique[1]

[1]*Department of Computer Science and Engineering, V.Y.W.S Prof Ram Meghe Institute of Technology and Research Badnera, Maharashtra, India*

---***---

Abstract - *The rapidly growing online learning resources such as e- learning, distance learning and the most recent and popular MOOC (Massive Open Online Course) are revolutionizing online education. This learning platform generates enormous amount of unused data which gets wasted as traditional learning analytics are not capable of processing them. This has motivated researchers to put forward solutions by utilizing big data techniques to process the large amount of data involved related to the educational field. In this paper, a machine learning and big data based approach has been presented for the online education systems. The analysis is done on the Open University Learning Analytics dataset (OULAD), a large complex dataset that requires significant pre-processing and feature extraction. With the proposed approach, it is aimed to perform an analysis based on various criteria such as age, gender, highest education etc and to help improve the performance of students in online courses by providing informed guidance.*

Keywords*:* MOOC, Big Data, Machine Learning

1. INTRODUCTION

The changing nature of technology and access to internet have brought more learning opportunities for the individuals in the field of education with various types of teaching, learning and assessment methods that can be achieved on/off campus, inside the classroom or virtualized environment. Online and distance learning are becoming more popular as opposed to traditional learning methods because of its accessibility and low cost for everyone regardless of age, gender or employability status. At present there are number of engaging online environments and platforms such as e- learning, distance learning, virtual classrooms etc of which MOOCs have become quite popular. The expansion of MOOCs in the era of online learning is supported by the fact that millions of students are enrolling themselves in these courses.

1.1 Massive Open Online Courses (MOOC)

MOOC are online educational courses available to anyone with a computer or any electronic device and an internet connection. It provides students an environment similar to a classroom or an online class setting and it can be accessed from anywhere across the globe. There is no limitation of paying tuition fees or committing to an academic course. Opting for MOOCs depends on individual interest/s and enhancement of their personal or professional goals. At the end of the courses, students can choose to take part in the examinations to finish the course and get the course-completion certificate. Universities in some countries have even set up the credit transferring system, i.e., the scores in MOOCs can be converted to the credits in universities. After finishing the required courses, students can get the graduation certificate or vocational certificate in a relevant field. It is different from the classic e-learning systems by the following characteristics [11].

1) Massive: Refers to huge in scale, amount or degrees. Since it aims at large audiences from all over the world they are termed ‘Massive’.

2) Open: No prerequisite for participation in courses. They are open to all, mostly free of charge or only certain modules are paid such as certification.

3) Online: All the course content and educational activities such as assignments and exams are online.

4) Course: The course is a collection of learning material developed by the teachers for a particular program. The course content can either be instructional, certification course, or courses based on learning activities.

MOOCs exceptional quality is that it brings together people who are interested in learning and an expert who seeks to impart this learning to students. Furthermore, they generally do not require any prerequisites, fees, or formal accreditation etc. MOOCs scope is quite wide and has the potential to further education in many different fields and subjects. A report by Online Course Report (2016) concluded that Computer Science and Programming offers largest percentage of MOOCs. There has also been a substantial growth of MOOCs in Science, Technology, Engineering, and Mathematics (STEM) fields. In addition, the benefits of MOOCs include the improvement of educational outcomes and accessible to all thereby eliminating barriers in the learning process, providing equality of opportunity in education and, most importantly ensure the liberalization of knowledge.

MOOCS have developed partnerships with 62 world class universities and developed platforms such as Coursera (www.coursera.org) led by Stanford University, edX (www.edx.org) which includes the Massachusetts Institute of Technology, ÉcolePolytechniqueFédérale de Lausanne etc. Besides these there are other platforms as well such as Udemy, Udacity, Futurelearn (the UK Open University's MOOC platform). In fact, the number of available MOOCs has expanded with over 4,500 courses provided by renowned universities across the USA and Europe.

1.2 Big Data

With the advent of technology, data is being created every second in an exponential manner. This digital data is available in every sector, educational institutions being no exception. Since the data is available and it is possible to share it over digital network, it has led to a massive increase in data volumes. Social networking, online course content etc are few means through which this huge data is being created. This huge data is termed as "Big Data" as these datasets are beyond the ability of traditional database tool to capture, search, store, transfer, manage, share, query, analyse, or update and provide information privacy. The definition of Big Data can vary from sector to sector depending on the size of the dataset associated with each sector. Big Data has been defined by International Data Corporation (IDC) as "New generation of technologies and architecture designed to extract value for large datasets of wide variety by high velocity capture, discovery and analysis". Specifically, big data can be divided into data science and big data technologies. Data science is "the study of techniques covering the acquisition, conditioning, evaluation and exploitation of data", while big data technologies are "systems, software libraries, tools, frameworks with their algorithms associates that allow distributed processing and analysis of big data problems between clusters of machines" [6].

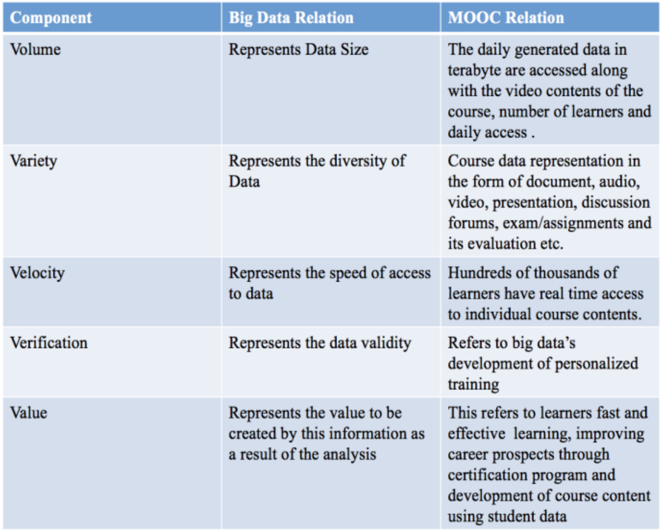
Big Data Analytics(BDA) combines enormous volumes of diversely outsourced data, analyses them, using complex algorithms to discover patterns and create informed decisions. BDA automated reports facilitate increased efficiency, better insights, and improved awareness which qualifies education services to be suited to individuals and institutes requirements. Big data offers, in addition to massively parallel computational powers, algorithms dedicated to machine learning to process and extract knowledge from the various types of data produced by e-learning systems, including learner profile information, activities, preferences, results, etc [1].

1.2.1 MOOC and Big Data Analytics

Big Data Analytics(BDA) combines enormous volumes of diversely outsourced data, analyses them, using complex algorithms to discover patterns and create informed decisions. BDA produces automated reports enable increased efficiency, better insights, and improved awareness which modifies education services to be suited to individuals and institutes requirements. Big data offers, in addition to massively parallel computational powers and distributed storage capabilities, sophisticated methods, and algorithms dedicated to machine learning to process and extract knowledge from the various types of data produced by e-learning systems, including learner profile information, activities, preferences, results, etc [1].

The concept of Big Data is expressed by five main components which are briefly known as 5V in the literature, these are Volume, Variety, Velocity, Value and Verification. The subject of big data can be associated as in Table I in accordance with the MOOC and 5V components [8, 9].

Table-1: Big Data and MOOC relation



Component	Big Data Relation	MOOC Relation
Volume	Represents Data Size	The daily generated data in terabyte are accessed along with the video contents of the course, number of learners and daily access .
Variety	Represents the diversity of Data	Course data representation in the form of document, audio, video, presentation, discussion forums, exam/assignments and its evaluation etc.
Velocity	Represents the speed of access to data	Hundreds of thousands of learners have real time access to individual course contents.
Verification	Represents the data validity	Refers to big data's development of personalized training
Value	Represents the value to be created by this information as a result of the analysis	This refers to learners fast and effective learning, improving career prospects through certification program and development of course content using student data

Big data technologies will allow an online learning system to automatically evolve and adapt to different situations depending on the learner's profile and interactions. An online learning system now has the ability to make decisions and make predictions automatically, without the intervention of a human being, through very advanced models and algorithms of machine learning, which is an integral part of big data[2].

1.3 MOOC and Machine Learning

Machine learning can be defined as performing analysis and producing result on large and/or complex data which cannot be analyzed by human intervention but by using intelligent and learning algorithms. They are further subdivided into following three classes:

1. Supervised Learning

This algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning algorithms are: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc [7].

1. Unsupervised Learning

In this algorithm, there is no target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means [7].

2. Reinforcement Learning:

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process [7].

Here the following classification algorithms of supervised learning were considered for predicting the final result:

Decision Tree Algorithm

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It is used for both categorical and continuous input and output variables. In this technique, the population or sample is split into two or more homogeneous sets (or sub-populations) based on most significant differentiator in input variables. It makes use of a decision tree as a prediction model to go from observation of an item (mentioned in the branches) to conclusions of the items target value (represented in leaves) [8].

Random Forest

Random forest is a tree based algorithm which involves building several trees and combining with the output to improve generalization ability of the model. This method of combining trees is known as an ensemble method. Ensemble is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical [9].

K Nearest Neighborhood

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) calculating the distance between points on a graph [10].

1.4 MapReduce

It is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. The basic unit of information, used in MapReduce is a (Key, value) pair. All types of structured and unstructured data are required to be translated to this basic unit, before supplying the data to MapReduce model. A MapReduce model consist of two separate routines, namely Map-function and Reduce-function. The computation on an input (i.e. on a set of pairs) in MapReduce model occurs in three stages: 1. Map Stage, 2. Shuffle Stage and 3. Reduce Stage

The map and shuffle phases distribute the data, and the reduce phase performs the computation. MapReduce logic, unlike other data frameworks, is not restricted to just structured datasets but it has an extensive capability to handle unstructured data as well. Map stage is the critical step which makes this possible. Mapper brings a structure to unstructured data.

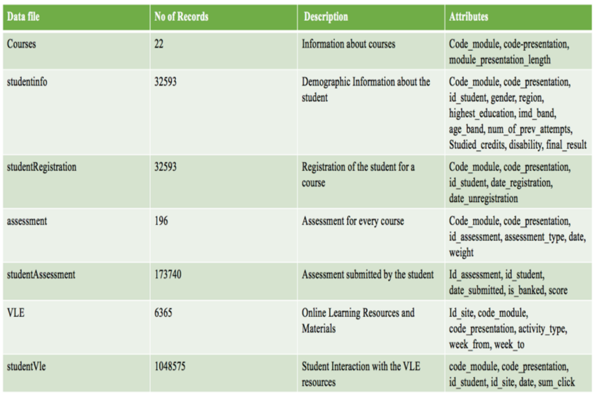
In this experiment, a big-data based analytical model is presented and based on the it, a framework has been defined to perform the intended analyses using machine learning to predict which algorithm gives the best possible result. This experiment investigates big data analytics and machine learning methods in MOOCs on large open data set which requires extensive pre-processing, evaluating the learner's performance and predicting the possible outcomes to improve the performance of the system using machine learning algorithms such as decision trees, K-Nearest Neighborhood(KNN) and Random Forest.

2. PROPOSED METHODOLOGY

The field of Education is rich of data, but putting them into data science has a short history. Analysis of educational content is an emerging field of data science. They aim at utilizing big data methods to give institutions and individuals the power to enhance student performance. Advances in modern software technology allow us to capture a complex number of specific user actions to get closer to students' academic performance. Based on these results, instructors can also adjust their instructional contents, sequences and activities. This could include early identification of at-risk students so that timely interventions can be designed for these students to support them in improving their academic performance and overall institute retention rate. Another application is prediction of student's drop out, as increased dropout rate is a potential. This paper investigates prediction performance using machine learnings classification algorithms models based on different types of attributes such as demographic info, assessment info and interaction with the Virtual Learning Environment (VLE). An analysis on the OULAD dataset is done here.

2.1 Dataset Selection and Description
Following the review of dataset during the data selection step, the decision was made to use the OULAD dataset from Open University UK [12]. It includes the result of the assessments submitted by the students, detailed VLE data which can be used by scholars to engineer various features and build various models for predicting students' performance during the course.

Table 2: Summary of OULAD dataset



Data file	No of Records	Description	Attributes
Courses	22	Information about courses	Code_module, code-presentation, module_presentation_length
studentinfo	32593	Demographic Information about the student	Code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, Studied_credits, disability, final_result
studentRegistration	32593	Registration of the student for a course	Code_module, code_presentation, id_student, date_registration, date_unregistration
assessment	196	Assessment for every course	Code_module, code_presentation, id_assessment, assessment_type, date, weight
studentAssessment	173740	Assessment submitted by the student	Id_assessment, id_student, date_submitted, is_banked, score
VLE	6365	Online Learning Resources and Materials	Id_site, code_module, code_presentation, activity_type, week_from, week_to
studentVle	1048575	Student Interaction with the VLE resources	code_module, code_presentation, id_student, id_site, date, sum_click

Table 2 provides a summary of the OULAD dataset. The dataset included 7 separate files converted to csv files using processing and transformation to extract features before building prediction models. A student demographic information is linked with other information segments of assessments and VLE interactions.

2.2 Pre-processing and Transformation

This section mentions the data pre-processing and transformation carried out before building predictive models. Features from different files were extracted based on three main keys that identify a student uniquely in the entire database: id_student, course_module, module_presentation.

2.3 Data Mining and Evaluation

This work aims to investigate the performance of the student in the course using analytics. The exploration question was then: to predict which algorithm give the best result and to predict whether a student will pass or fail the course.

Figure 1 demonstrates the methodology from an implementation perspective. The dataset was retrieved as a compressed zip from OULAD. OULAD dataset contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules) mentioned earlier. The dataset consists of tables connected using unique identifiers. All tables were converted using MapReduce and stored in the csv format.

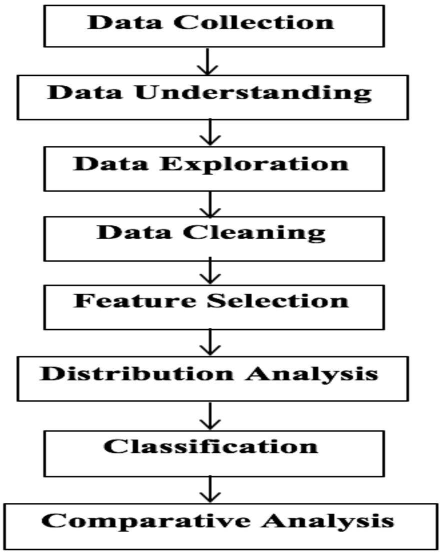


Fig-1: Implementation Architecture

The experiment is developed on the programming language Python. Python is the most ideal language for data science and machine learning. Python language provide hundreds of libraries and packages for developing application in the fields of data science. The details about the technologies used are as below:

1. Python programming language
2. Flask Web Framework
   A. Werkzeug
   B. Jinja2
   C. WSGI
3. Data science libraries

There are various packages or libraries of Python used for data mining, machine learning, data analysis, data visualization like numpy, pandas, matplotlib, seaborn, bokeh, scipy, scikit, keras, tensorflow, theano, glueviz, orange and many more.
The current work uses:

a)NumPy : It is the fundamental package needed for scientific and mathematical computing with Python.

b) Pandas: It is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.

c)Matplotlib: It is the most imp data visualization library to plot graphs in the program. Matplotlib (Mathematical Plotting Library) is a plotting library for the Python language and it is a numerical mathematics extension of NumPy.

After applying the algorithms of machine learning the following results were obtained:

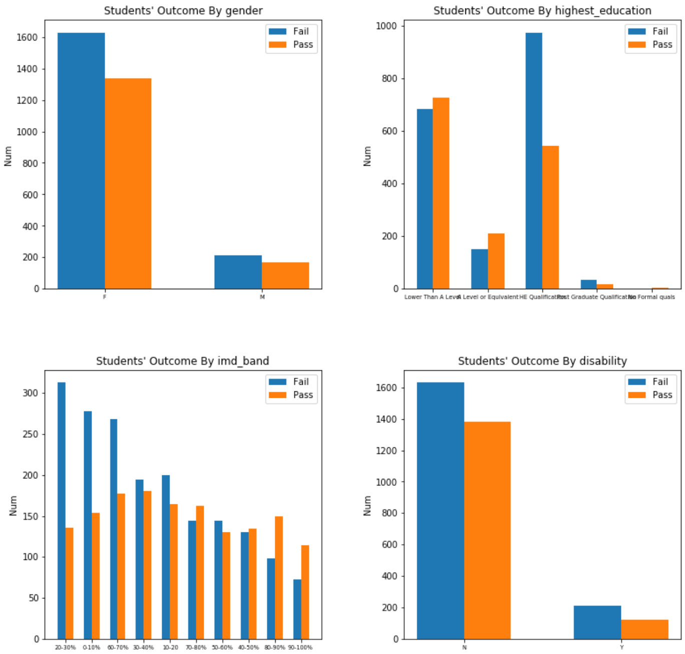


Fig-2: Student Outcome based on varied parameters

Here the output displayed in figure 2 indicates students learning outcome based on parameters such as gender, level of highest education, imd band and disability.

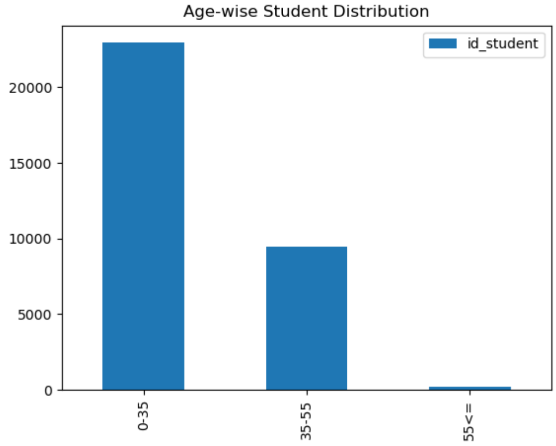


Fig-3: Age wise student distribution

Figure 3 displays the age-wise distribution of student who were active on the MOOC platform. As seen by the representation of the graph, the proportion of the age group 0-35 outnumbers the rest of the age group indicating that most learners fall in the age group of 0-35.

Another important parameter in this analysis was the gender-wise distribution of the students participating in the MOOCs which concludes that the count of the males surpassed that of the females as shown in figure 4.

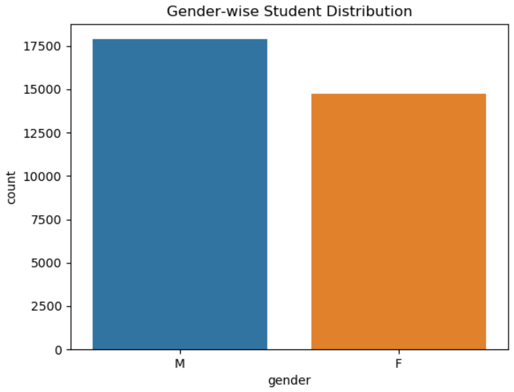


Fig-4: Gender-wise student distribution

Figure 5 and Figure 6 displays the comparative analysis done on the dataset using various machine learning algorithms on the number of students who have passed/failed their assessments. Random Forest classification provides the highest accuracy result for the given dataset.

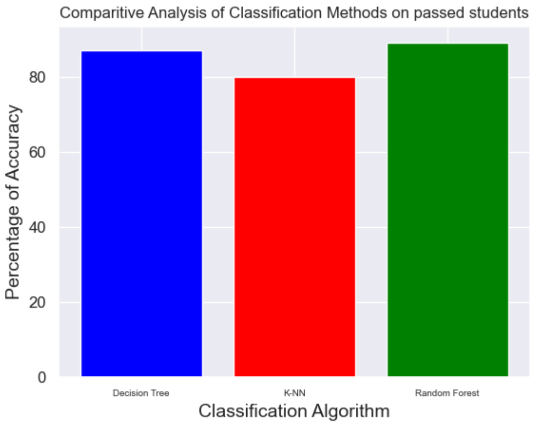


Fig-5: Classification analysis on passed students

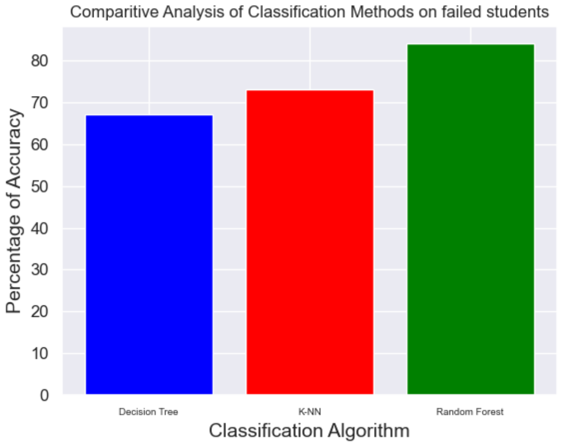
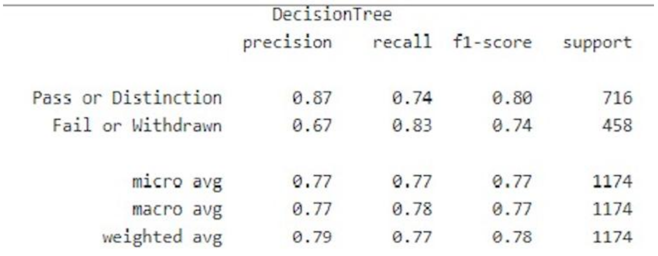


Fig-6: Classification analysis on failed students

3. RESULT

This study investigated the performance of various machine learning algorithm to draw out conclusion as to which classification model gives the best result and also whether the student will pass or fail. For this purpose, I used the OULAD dataset which included 7 separate files converted to csv files using MapReduce and features were extracted to build prediction model based on criteria such as age, gender, higher education etc and predict performance of machine learning algorithms based on different categories of predictors such as demographics, assessment scores, VLE interaction etc. The result was assessed based on the classification algorithms and it was found that Random Forest provided the most accurate results based on multiple attributes. The results of various classification are shown as follows:

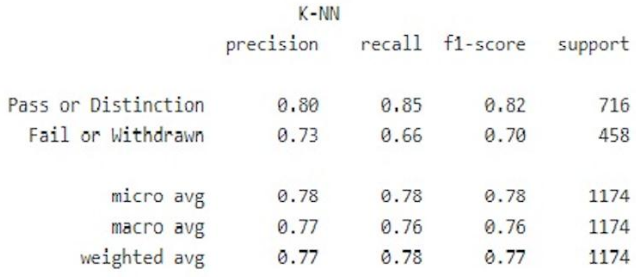


```
                       DecisionTree
                    precision    recall  f1-score   support

Pass or Distinction      0.87      0.74      0.80       716
  Fail or Withdrawn      0.67      0.83      0.74       458

          micro avg      0.77      0.77      0.77      1174
          macro avg      0.77      0.78      0.77      1174
       weighted avg      0.79      0.77      0.78      1174
```

Fig-7: Classification result on Decision Tree

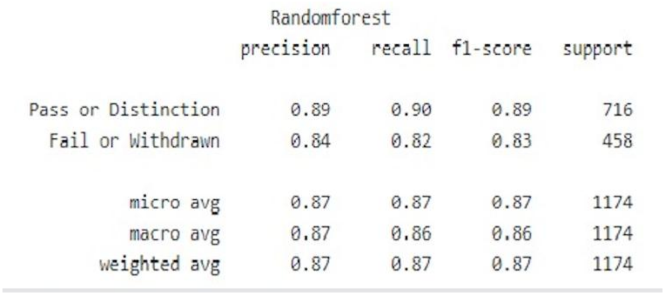


```
                       K-NN
                    precision    recall  f1-score   support

Pass or Distinction      0.80      0.85      0.82       716
  Fail or Withdrawn      0.73      0.66      0.70       458

          micro avg      0.78      0.78      0.78      1174
          macro avg      0.77      0.76      0.76      1174
       weighted avg      0.77      0.78      0.77      1174
```

Fig-8: Classification result on K-NN



```
                       Randomforest
                    precision    recall  f1-score   support

Pass or Distinction      0.89      0.90      0.89       716
  Fail or Withdrawn      0.84      0.82      0.83       458

          micro avg      0.87      0.87      0.87      1174
          macro avg      0.87      0.86      0.86      1174
       weighted avg      0.87      0.87      0.87      1174
```

Fig-9: Classification result on Random Forest

Since Random Forest algorithm offers good feature selection, it can be used to produce good prediction result based on various attributes mentioned before. From the result, it is evident that Random Forest classification provides the best possible outcome on all parameters with a precision value of 0.89 which is the highest among the used classification algorithms.

4. CONCLUSION

Open online education should be open with respect to people, places, and methods. Online curriculum and course development, instructional design, quality assurance, student and faculty support, technological platforms, and infrastructure among other things are important issues to consider, not only in the context of MOOCs, but in online, open and distance learning in general. Therefore, it is imperative to build upon the theory, research, and practice in the field of education to reinvent a flexible learning method for both students and educators as technology changes. In this experiment, it was found that after applying the various machine learning algorithms such as random forest, decision tree and KNN on the given current dataset, Random Forest algorithm provides the most accurate result and using that result, predictions for future learning outcomes can be done.

REFERENCES

[1] Improving Online Education Using Big Data Technologies DOI: http://dx.doi.org/10.5772/intechopen.88463

[2] Mervat A. Bamiah, Sarfaraz N. Brohi, Babak Bashari Rad, Journal of Engineering Science and Technology Special Issue on ICCSIT 2018, July (2018) 229 - 241: BIG DATA TECHNOLOGY IN EDUCATION: ADVANTAGES, IMPLEMENTATIONS, AND CHALLENGES

[3] Siemens, G. Massive Open Online Courses: Innovation in Education?. Commonwealth of learning, perspectives on open and distance learning. Open Educational Resources: Innovation, Research and Practice. 2013, 5-16.

[4] Sinclair, Jane, Boyatt, Russell, Rocks, Claire and Joy, Mike. Massive open online courses (MOOCs): A review of usage and evaluation. International Journal of Learning Technology 2015, 10 (1): 1-23.

[5] Mohammad Khalil PhD dissertation titled 'Learning Analytics in Massive Open Online'

[6] Chakhari A. La digitalisation est une guerre mondiale armez-vous. 2015. p. 70

[7] https://stepupanalytics.com/introduction-to-machine-learning/

[8] https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4

[9] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[10] https://www.saedsayad.com/k_nearest_neighbors.htm

[11] Yunus Santur, Mehmet Karaköse, Erhan Akın ,"Improving of Personal Educational Content Using Big Data Approach for Mooc in Higher Education" , Fırat University, Computer Engineering Department, 23119 Elazig, Turkey.

[12] https://analyse.kmi.open.ac.uk/open_dataset

[13] Cormier, D. 2010, what is a MOOC? [Video file]. Retrieved from https://www.youtube.com/watch?v=eW3gMGqcZQ

[14] S. I. De Freitas, J. Morgan, and D. Gibson, Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision, British Journal of Educational Technology, vol. 46, issue 3, pp. 455-471, 2015

[15] H. Fournier, and R. Kop, MOOC learning experience design: Issues and challenges, International Journal on E-Learning, vol. 14, issue 3, pp. 289-304, 2015.

[16] Hollands, F. M., &Tirthali, D. (2014). MOOCs: Expectations and reality. Full report. New York, NY: Columbia University. Retrieved from https://files.eric.ed.gov/fulltext/ED547237.pdf

[17] Mohammad Khalil PhD dissertation titled "Learning Analytics in Massive Open Online".

[18] Siemens, G. Massive Open Online Courses: Innovation in Education. Commonwealth of learning, perspectives on open and distance learning. Open Educational Resources: Innovation, Research and Practice. 2013, 5-16.