National College of Ireland

Identifying At-Risk Students in Virtual Learning Environment using Clustering Techniques

MSc Research Project
MSc in Data Analytics

Kamalesh Palani

Student ID: x18180311

School of Computing

National College of Ireland

Supervisor:
Dr. Paul Stynes
Dr. Pramod Pathak

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Kamalesh Palani		
Student ID:	x18180311		
Programme:	MSc in Data Analytics	**Year:**	2019-2020
Module:	MSc in Research Project		
Supervisor:	Dr. Paul Stynes, Dr. Pramod Pathak		
Submission Due Date:	17th August 2020		
Project Title:	Identifying At-Risk Students in Virtual Learning Environment using Clustering Techniques		
Word Count:	**7612**	**Page Count**	**19**

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: Kamalesh Palani

Date: 17th August 2020

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):	

Identifying At-Risk Students in Virtual Learning Environment using Clustering Techniques

Kamalesh Palani
x18180311
MSc Research Project in Data Analytics

Abstract

Higher education institutions across the world have started using web-based learning platform to provide distance learning education online. These learning platforms provide opportunities for the eLearners and the working professional to learn on demand and gain knowledge remotely. This wide spread use of Virtual Learning Environments (VLEs) in universities has higher rate of students drop out percentage and difficulty in monitoring the student's online engagements during the courses. Therefore, goal of this research is to develop a data-driven clustering model which is aimed to identify the students who are at-risk during the course cycle in early stages. A publicly accessible Open University (OU) dataset which consist of more than 30,000 students for 7 different courses, is used to build clustering model based on individual student's behaviour in Virtual Learning Environment. This research was carried out using three unsupervised clustering algorithms, namely Gaussian Mixture, Hierarchical and K-prototypes. Models efficiency is measured using clustering evaluation metric to find the best fit model. Upon comparing the different models, the K-Prototype model clustered the at-risk students more accurately than the other proposed models and generated highly partitioned clusters. The outcome of the models can help the online instructors in distance learning universities to monitor the students based on the student online engagement in the VLEs and offer extra assistance to the students who are at-risk during the course cycle in early stages in order to minimize the drop-out rate.

1 Introduction

In recent years, web-based eLearning platforms in higher education institutions has been expanded worldwide to provide distance learning for the students and also for the working professionals. The rapid growth of eLearning platform in university leads to increase in abundance of educational data, which contains individual student login data for their respective courses in online.(Aljohani, Fayoumi and Hassan, 2019) describes that educational data from the VLEs provides opportunities to analyse the students learning pattern individually, and also helps to increase the performance of teaching and learning behaviour in the VLEs. Major challenge in the VLEs is the student drop out percentage, Students get low marks in the courses and they lose confidence which leads to withdraw from the courses. As a result, it impacts both the student's carrier and the reputation of the university. A study by (Hussain *et al.*, 2018) show

that due to the absence of face to face interaction in the web-based educational platform it is difficult to monitor the student online activity by the instructor. Therefore, student login data is the only source for the instructor to monitor the student's online engagement and provide high quality education in online courses. Because of the limited number of instructors in the online platform it is difficult to access all the individual student log in data to know about the student engagement level in their courses, resulting in absence of one to one supervision in the VLE. As a result, there is a lack of motivation for students to pursue the online courses and lose confidence resulting in withdrawn from the course. Hence the student drops out prediction is the ongoing challenge in the online learning platform which needs to get addressed so that both the student and the online educational institution will get benefited (Chui *et al.*, 2020).

Different Machine Learning (ML) algorithm has been used to build the dropout prediction model in the recent years. Most of which are not data driven and labelled data is used to train the model. Mainly most of the research work is based on traditional education in university and the student's interaction in various activity in the VLE has been overlooked. Moreover, in the previous research work the low engagements students in the VLE are predicted by using the students' performance and not compared with the interaction pattern of another students. This states that in the previous researches the predicting model was built under the assumption that all the students have same online engagement behavior. Moreover (Hassan *et al.*, 2019) said that to predict the at-risk students, individual student engagement pattern has to be identified from the VLEs along with academic performance to derive the valuable insights from the data. Since the educational data continues to increase, the diversity of the data changes based on the research question, hence there is not a standard way to monitor the individual student's online activities in the VLEs or tracking of individual student learning habit.

Therefore, the proposed research work uses the data-driven clustering algorithm on freely accessible OU dataset, to identify the at-risk students during the course cycle in early stages based on the individual student's online activities from the VLEs and academic performance and compared with Fuzzy C-means model (MacEdo *et al.*, 2019) to find the best fit model . Below is the research question aimed to address in this research work.

"To what extent the unsupervised clustering algorithm can be used to identify the at-risk students during the course cycle in early stages from the virtual learning environments?"

This study considered the following research objective to address the research question:

- Investigating the state-of-the-art in identifying the at-risk students in the VLEs using the clustering techniques.
- Design of clustering model using aggregated data which is derived from the raw dataset.
- Implementation of Fuzzy C-means and proposed model on the aggregated data to identify the at-risk students.
- Evaluation of the models based on the clustering evaluation metric and comparison of Fuzzy C-means and proposed models to find the best fit.

The key contribution of this paper is to help the online instructor to track the student's online activities and build students profile. And also, it helps to predict the future outcomes of the student performance which can be used to alter teaching content and also helps to optimize the learning environment in the VLEs. Additionally, by identifying the vulnerable students at

early stages during the course's instructor can motivate or provide additional support to those students which will increase the student performance in the online activities.

Below is the research section which is organized as follows: Section 2 precisely discuss the previous research related to the at-risk student prediction and clustering methods used in the VLEs. Section 3 describes the OU dataset and the methodology used in this research work. Section 4 present the implementation of the clustering algorithm. Section 5 provides the evaluation metric of the model used in the paper and in last section conclusion and the future outcome of this research is discussed.

2 Related Work

The Literature review for this research has been written from the peer reviewed paper published between 2010 to 2020 on the student dropout prediction in the VLEs. Below subsection will discuss about the uses and challenges of online education system in the school or university, challenges in predicting the student's dropout rate, student learning behaviors in VLEs, overview of data mining techniques used in dropout predictions and at last research gap will be discussed.

2.1 Study of Virtual Learning Educational Platform

Web based learning platform have shown a rapid growth in the higher educational institution in many forms like Virtual Learning Environment, E-Learning , Massive Open Online Courses (MOOCs) and Modular Object-Oriented Dynamic Learning Environment(Moodle).These online learning environments provides a new way of self-paced learning for the students at any time and also students can access the materials from any location. This leads to progression of educational data which contains the individual students learning behavior (Oztekin *et al.*, 2013). Therefore, in this section how the VLEs is used in the education institutions will be explained along with the challenges.

(Corsatea and Walker, 2015) has stated that most of the VLEs in the higher educational institutions are used as data container to upload the students study materials. And the teacher is not utilizing the full tools in the VLEs like blogs, chat forms, tracking of student's engagements in the VLEs because instructors lags the technical knowledge to use the system. Due to the absence of one to one interactions in the VLEs students lose motivation and finding difficulty in accessing the courses materials and as a result which affect their performance. (Hussain *et al.,* 2018) has used the VLEs log to overcome the above-mentioned challenges between the instructors and the learners. Educational logs of the individual students can be used to analyze the student's engagement behavior in the VLEs. Using the log stored in the VLE instructor can monitor the students in the online educational platform. However, due to the limited instructors in the higher educational institute it not possible to analyze the individual student logs for all the courses. Furthermore, author suggest that an automated intelligent system is required to process or extract the information from students logs which can be used by the instructor to profile the students and understand the student's engagement in the VLEs in the meaningful way. (Agnihotri *et al.*, 2015) did research on students log in data from an online assessment

platform tool called connect. Which contains the number of times the students logged in to the course for the entire course duration. Using these student log in details author used it for students profiling and monitoring. Limitations, in this research is choosing of limited factors when profiling the students.

From the above discussion, it is clear that students log in the VLEs can be used to predict the learner's behavior. And also, it can be used to profile or monitor the student's engagements in the VLEs courses. In the next topic reason for student drop out in the VLEs will be discussed which is the major challenge facing by the higher education institution.

2.2 Study of Student Drop Out in Virtual Learning Environment

One of the major challenges faced by the higher educational institution which incorporated VLEs is students drop out and failure rates. Student who are enrolled for the course at the beginning will not complete the course. Because lack of support from the institutions for the at-risk students. Therefore, in this section different researchers review of the student drop out in the VLEs is discussed.

(Dalipi, Imran and Kastrati, 2018) has reviewed the student dropout prediction from different research papers and their challenges. In his work he has stated two important factors which influence the student dropout in VLE, one is student related and VLEs related. Factors in students related are lack of motivation, lack of time and insufficient knowledge to the courses and for VLEs related factors are course design, hidden cost and lack of interactivity or monitoring in VLEs. These are the key factors that the author has stated for students drop out. However, in this work to overcome the challenges and to reduce the dropout rate institutions recommendations has been presented in this paper. In order to build the effective prediction model, students clickstreams data, Student academic performance and students' social engagement features or variables has to be considered. (Yi *et al.*, 2018) in his work the author had used non-cognitive skills of the students to predict the drop out students. The feature that are selected for the prediction contradicts from the above paper. Non-cognitive skills of the students taking the course like sleep hour, usage of smart phones, consumption of energy drink, number of visits to doctor and final semester grade. This data is collected by using the survey at the start of the semester taken from a total of 552 business students and model is build. Main limitation of this research is the data collected are course specific and not generalized to other courses and also the data used to train model is small. In (Liang *et al.*, 2016) data from Edx platform for 39 courses has been used to build the predictive model. Data were extracted from the Edx platform which contains the Enrolment, user and course feature and classification models was built to classify the students. In the user feature author has used the data from the student interaction from the video and the clicks the students has made for each course to build the model. But this, approach had not been carried out in VLEs and the students interacting with the video are not properly recorded. Therefore, in this research the model trained has data loss which is a major drawback.

Overall, from the above study to predict the students drop out in VLEs feature selection from the VLEs log and the size of the dataset are the important factors that has to considered in building the model. And also, as stated by (Hassan et al., 2019) due to the

growing educational data in the institution provides opportunities not only to improve the student performance but also it can be used to optimize the learning environment.

2.3 Understanding of Student Engagement in VLEs

Student engagement in the VLEs is the effort that student spends on interacting with the VLEs for the specific course. Student engagement metric in the prediction of students drop out is an important factor because lack of interaction in VLEs will affect the student grade. Due to the absence of face to face meetings in web-based system it is difficult to measure students' engagements in VLEs like attendance, interaction of the students in the courses and grades, Due to these challenges in measuring students' engagements there are no standard approach to find out the student behavioral in the VLEs.

(Waheed *et al.*, 2020) in this research work student's engagement is used as a key factor predict the student academic performance in the VLEs. Research agenda is to develop a prediction model on binary classification dataset whether a student will pass or fail at the end of the course is predicted using the deep learning models. From the VLEs log, clickstream from the VLEs is taken as an important factor in predicting the student performance. However, the model was built on the assumption that the student's behavior during the course is treated as equal. Absence of individual student's behavior pattern is not considered in this research. (Boroujeni and Dillenbourg, 2019) have tried different approaches to analyze the individual learning processes from the VLEs. In this research video, assessment details are extracted from the student's interaction logs on weekly basic and used Hypothesis-driven methods to analyze the individual student behavior. The outcome of this research shows that students who have taken the assessments watch video before the submissions and minimal students take the assessments without watching the videos. Limitation in this research is fixed study pattern which was used to train the model and the students who change the study pattern are given less importance.

Understating the individual students learning behavior in the VLEs is an important metric that has to be included while training the model so that the accuracy of predicting the low engagements students in the VLEs can be increased(Corrigan and Smeaton, 2017). In the next topic different Machine Learning (ML) and clustering techniques which was used to build the student drop out prediction model in the VLEs is discussed.

2.4 ML Technique used in Predicting the At-Risk Students

(Chui *et al.*, 2020) used support vector machine (RTV-SVM) to predict the at-risk students and marginal students in the VLEs. Author has transformed the dataset in to binary classification dataset. Which contains two class namely class 0 are the students who got pass mark and class 1 are the students who failed the course. However, in this work the students who are dropping out of the course cannot be identified in real time only after the completion of the course the drop out students are identified. (Macarini et al., 2019) has tried to predict the at-risk students at the early stages using the Moodle dataset. Four classification model was built namely AdaBoost, Decision Tree, Random Forest and Naive Bayes on the dataset which were transformed into weekly basis. And to evaluate the model Area Under Curve (AUC) is used.

Limitation in this research is dataset used to train the model is small therefore oversampling technique like SMOTE is used to balance the data and the performance of the model changes every time the model is trained. Drop out predicting system developed by (Hassan *et al.*, 2019) used the Deep learning models like Long Short-Term Memory (LSTM), Artificial Neural Network is used to build the model on the smart data which was transformed into week-wise clickstream data. And the author has made a statement that deep learning models perform better than the traditional machine learning models with better accuracy in predicting the at-risk students. However, in this research Students engagement pattern in their courses factors has to be considered in this work which is a major limitation in this project. Author also suggested that sequence to sequence approach on student's interaction pattern can be built in the model for better accuracy. (Corrigan and Smeaton, 2017) has used the students interaction pattern which was not used in the above research. Recurrent Neural Networks (RNN) along with student interaction pattern to predict how well the students will perform in their VLEs courses. However, in this work the major limitation is 2,879 students are used to train the model and if any new courses have been added to the curriculum then to include the new course in the model, one year data has to be collected and after that the model has to be trained.

2.5 An Understanding of Clustering in VLEs

(Agnihotri *et al.*, 2015) used data-driven clustering methods to identify the high and low achiever in the online courses. K-means clustering algorithm is used to group the students based on the Login behavior of the students and number of attempts to clear the course. Data aggregations used in this research is not properly processed there are lot of null values in training the model and less factors are used to build the model. (Preidys and Sakalauskas, 2010) extracted the huge data from BlackBoard Vista distance learning platform to analyze the learners study pattern. Three clusters were identified from the dataset namely Important, Unimportant and Average importance using K-means clustering. Limitation in this research is several outliers have been occurred in the dataset and the same is used to build the model. Above mentioned challenges have been resolved in (Navarro and Moreno-Ger, 2018). In this research huge dataset with no outlies has been used on education dataset to determine which clustering algorithm perform better in predicting the low learners in the VLEs. Seven clustering models has been used in this work and to benchmark the performance different evaluation metrics like Dunn Index, Silhouette score and Davies-Bouldin score have been compared to identify which algorithms performs better. However, specific limitation is this research is missing data in any instances in the factors are removed, which may contain the useful information and provide additional insights. 44 % of the data is cleaned from the original data.

2.6 Research Gaps

It is evident from the previous research that there is no specific way of predicting the at-risk students in VLEs. In most of the paper feature section and choosing the factors which contribute to predict the at-risk students are not properly considered. Also, the size of the dataset is the major limitation is the previous works most of the researches have used the student's data which is less than 1000 and performed the model. Most importantly to train

the model for early prediction of the students in the course all the weeks interaction data is used under the assumption that each student will have the same interaction behavior during the course. Therefore, aim of this research is to implement the clustering model on OU dataset and compare it with (MacEdo *et al.*, 2019) fuzzy C-means model to identify the at-risk students in the VLEs. Important attributes like interaction patterns of the individual students in the VLEs, academic success and student demographic information will be used and transformed into aggregated data to detect the at-risk students in the early stages. Clustering models like Gaussian mixture, K-prototype and Hierarchical clustering are used with different parameters and compared with the evaluation metric (Navarro and Moreno-Ger, 2018) to evaluate which model performance better. Overall, this research will be helpful to both the instructor and the eLearners in the online learning environments for profiling and tracking of students. And, also by knowing the student's behavior teaching content can be altered in VLEs.

3 Research Methodology

To extract the meaningful insights from the complex data Knowledge Discovery in Database (KDD) methodology is used in this research in order to achieve the goal of this paper.

3.1 Data Selection and Understanding

For this research dataset has been extracted from Open University platform [1]. This dataset is unique from the other educational data because it contains the student's demographic data along with the student's online activities in the VLEs which is clickstream. Totally there are 32,593 students in this dataset for 22 different courses for the period 2013 and 2014.Dataset used in this research is publicly available and contains students anonymized information and follow ethical and privacy requirement of Open University. There are 7 different CSV files which are used in this research work which contains different information related to demographic, assessments and interaction of the students in VLEs. Figure 1 shows the dataset structure that has been used in this work.

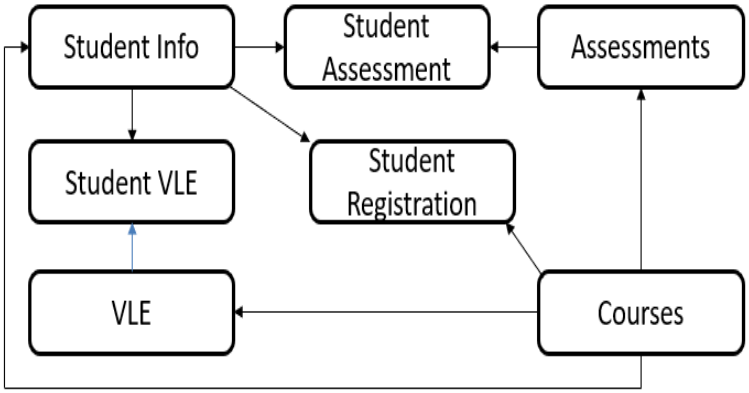

Figure 1. OU dataset structure (Kuzilek, Hlosta and Zdrahal, 2017)

[1] https://analyse.kmi.open.ac.uk/

In this research raw data is transformed to aggregated data with newly created attributes from different files of the data. Three different type of category are extracted from the dataset: learning behavior, student course performance and demographic details of students. Below tables shows the dataset description and the count.

Table :1 Dataset Description

Category	File Name	File Description	Record count
Demographic	Student info	Contains student information regarding age, studied credits, number of attempts and demographic details.	32,593
Demographic	Student Registration	Contains the date of registration and deregistration of the students for their courses.	32,593
Performance	Student assessment score	Assessment result for each student in each course.	1,73,912
Performance	Course details	Modules name, code and length of each modules are given in this file.	22
Performance	Assessment details	In this file whether students opted for social science or STEM (Science, Technology, Engineering and Mathematics) module is given.	206
Learning behavior	Student VLE clicks	Student interaction with the VLEs.	10,48,575
Learning behavior	VLE activity	Material available in the VLEs namely forums, homepages, URLs, accessing the study materials, forumng, etc.	6,364

3.2 Data Pre-Processing and Transformation

The Raw data is transformed into actionable aggregated data because it cannot be directly used as input to clustering model. All the pre-processing and transformation steps are performed in Python Jupyter Notebook using pandas library. First data exploration part is carried out to check the distribution of the data, finding out the missing values and checking the outliers in the data. Both univariate and bivariate analysis has been carried out and outlier and missing values are filtered from the dataset in the first step. In second step data transformation like encoding the categorical variables and standardization of the data is performed. In the third step new variables are created for each student namely the overall studied credits, total score, average clicks week wise and attempted weight for each course, as an additional success measure for students. To improve the clustering model performance one hot-encoding is done on the categorical column before giving as an input to train the models. At last columns which are not contributing for the at-risk student's prediction are dropped before implementing it in the model. A detailed description of aggregated data preparation and processing is explained is section 4.1.

3.3 Modelling

The aggregated and transformed data is given as an input to the clustering model. In this research, three clustering models will be implemented on the above transformed aggregated dataset namely K-Prototype, Gaussian Mixture and Hierarchical. Before running the model choosing the number of clusters for the dataset is done using the Gap Statistics (MacEdo *et al.*, 2019). K-Prototype model is the combination of K-means and K-mode clustering technique which can be used for the dataset having both numeric and categorical variable. The aggregated dataset used in this research contains both numeric and categorical therefore this specific type of clustering model is chosen(Wang *et al.*, 2016). As this analysis is based on finding the similar students interaction behavior in VLEs hierarchical clustering is used as it groups the students based on the similarity and also both top down and bottom up approach can be tested(De Morais, Araújo and Costa, 2015). Gaussian Mixture clustering model is chosen for this research because it is a probabilistic model and the approach will not get over until all the data points are converged in different clusters and also it uses soft clustering approach.

3.4 Evaluation

Multiple clustering evaluation metric are chosen based on the fact that the ground truth labels are not know in the dataset. These metrics will show the result whether the cluster are well separated and are not overlapped. Silhouette coefficient metric calculate the mean distance between the data points to find the better-defined clusters. And higher the Calinski-Harabasz index better the clusters are defined in the model. At last, Davies-Bouldin index is used to check the similarity between the clusters and lower the index value better the cluster partitioned(Navarro and Moreno-Ger, 2018).Therefore, in this research all the above-mentioned metric will be used to evaluate the model performance.

4 Implementation

In this section implementation of the Fuzzy C-means model(MacEdo *et al.*, 2019),proposed clustering models and preparation of aggregated data is discussed along with technical specifications.

4.1 Aggregated Data Preparation and Pre-processing

In this research to predict the at-risk students before the final exam in their modules for each student, raw OULA dataset is transformed to aggregated data by processing all the data from the 7 files to single table. The first step in data preparation is merging all the files together to a single table. Primary key column has been generated for all the files because students may take multiple courses, therefore by concatenating the student id, year of the course taken and module column, primary key is generated using Concat function in python pandas data frame. From these files, left outer joins has been performed with the Students info table using python panda's data frame and single data frame containing all the student's information has been formed. In the second step, column which are not contributing in identifying the at-risk students

are dropped like unregistration date, disability, module length, week_from and week_to, is_banked etc. In the third step data exploratory is done using the seaborn and matplotlib libraries in python jupyter notebook to find the distribution of the attributes in the merged data frame, missing values and check for outliers in the data. Null values, NAs and outliers which are present in the data dataset were removed. Also, in the VLE filed in the data frame contains date columns which have negative value which indicate the student's interaction before the starts of the modules which was also removed from the data.

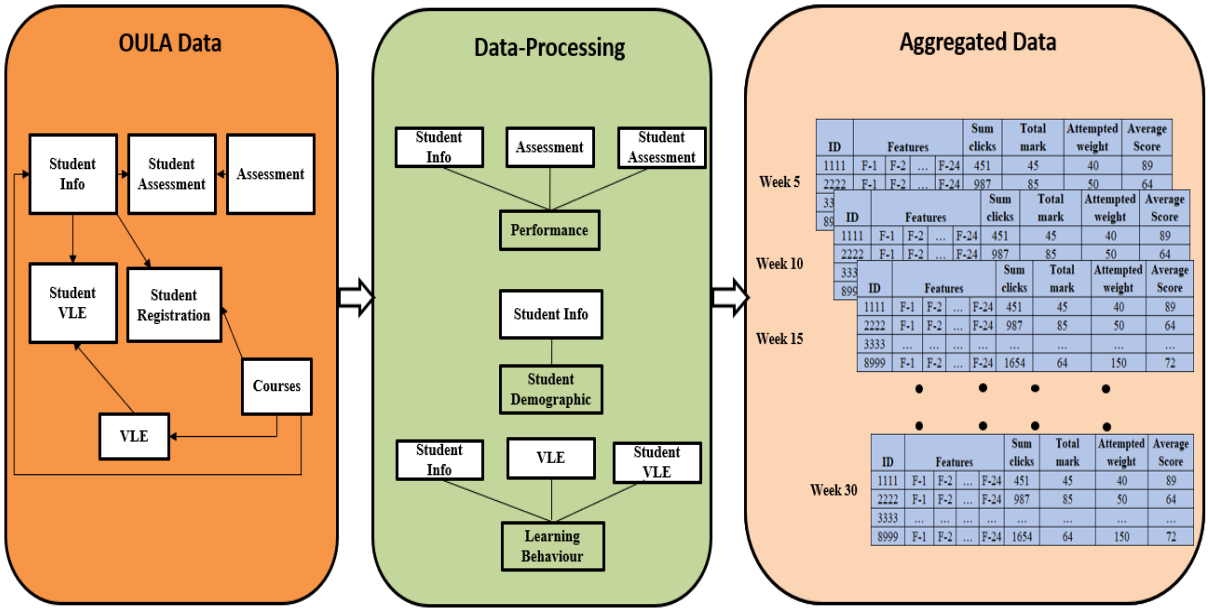

Figure 2. Aggregated Data Processing

Figure 2 shows the three steps process for the transformation of raw data to aggregated data. The aim of this research is to use the three important attributes – Learning Behavior, Performance and Demographic details of the students as an input to clustering models. Therefore, data transformation has been conducted in the cleaned dataset to derive the above-mentioned attributes. Firstly, to derive the learning behavior student's clickstream data has been processed to week wise for 20 different activity till week 30 from the VLEs namely URLs, Homepage, Forums, Quiz, Questionnaires, Folders, etc. Each week wise aggregation of clicks has been added to the previous week student click stream behavior this is achieved by using the aggregate, group by, melt and pivot function in python pandas data frame is used. Secondly, for student's performance average score the students has attained in all the assignments before the final exam has been added into a new column in the dataset. Also, total mark and attempted weights are calculated based on the assessments score and total credits they earned during the assignments. At Last, for demographic attributes one-hot encoding is done on the categorical columns like region, age_band, highest_education using get_dummies function in python using pandas. Before running the model, data were normalized and scaled down to fixed range between 0 to 1 with the help of MinMaxScaler function in python. This normalization of the data will improve the performance of the model, due to the fact that all the clustering models will use Euclidean distance to find the distance between the closest points to the near clusters.

Overall, after performing the above steps actionable aggregated data has been prepared and the same is given as an input to train the clustering models.

4.2 Implementation of Clustering Models

All the clustering model was implemented in Python 3.7 using Jupyter Notebook and Scikit-learn libraries. At First, to define the number of clusters for the clustering models is identified by using Gap Statistics (MacEdo *et al.*, 2019) on the aggregated data. Gap-stat library has been imported from python and used by the range of values from 0 to 11 for K by fitting the model and including all the indexes. The point of reflection of the curve was found at 3 which means for the dataset the number of clusters can be used is 3 to run the clustering models. Therefore, all the models were executed with 3 clusters to group the students based on the individual behaviors in the VLE.

4.2.1 Fuzzy C-means Clustering

First Fuzzy C-means clustering model has been implemented by defining three clusters which was found using the gap statistics. MAX_ITER parameter has been set to 20 to limit the model from running infinite loop. Also, m parameter value is given greater than 1 to avoid the model to run as K-nearest neighbors. After, passing the parameters model is fitted and cluster labels are stored in separate variable. Finally, to find the dispersion of the data and the clusters formed scatter plot is used to visualize the clusters.

4.2.2 Hierarchical Clustering

Agglomerative clustering has been imported from sklearn.cluster library in python to perform hierarchical clustering on the normalized data. This method follows the bottom-up approach, therefore at the start each data points will be treated as individual clusters and based on the similarity it will be successively merged together. Parameter n_clusters is set to 3 which was derived from the Gap Statistics and affinity parameter is set to Euclidean which will find the distance between the clusters. Finally, Linkage parameter is set to Ward which will minimize the variance between the groups of clusters. After defining all the parameters, the model is fitted using the method fit_predict which will return the labels or names for each data points in the dataset. The output of the method will be in one dimensional array for all the data points. This output result of clusters labels will be used to identify the students who are at-risk by plotting the scatter plot using matplotlib library in python and setting the parameter of x-axis to score attribute and y-axis to sum of clicks attribute in the dataset .Additionally, to check the performance of the model the clusters labels , metric and normalized data is used to find how well the clusters are separated between the datapoints using evaluation metric.

4.2.3 Gaussian mixture Clustering

Gaussian mixture is the second clustering model which is implemented on the normalized data. This clustering model is imported from sklearn.mixture library in python and created function to run the model using the defined parameters. n_components parameters has been set to 3 as

the number of mixture components in the model and init_params parameters as kmeans which is used to initialize the means and weights. After setting the parameters, model is built using the fit methods and output of the methods is the cluster labels. Using the clusters labels both the scatter plot and evaluation metric are performed to find the performance of the model.

4.2.4 K-Prototype Clustering

In this clustering model both categorical and numerical data has been given as input to the model. Because k-prototype algorithm works well with mixed numerical/categorical data. For numerical data this model uses the euclidean distance to cluster the data points and for categorical data it uses the similarity between the data points to group into clusters.10 Iteration has been carried out and for each iteration centroids and clusters are redefined and the best iteration is chosen based on the less variance between the clusters. After setting the parameters model were fitted into the fit method by defining the categorical variable separately. The output of the method showed that in the eighth iteration less variance has been achieved and the clusters labels are plotted in seaborn library in python to find whether the clusters are well separated.

5 Evaluation

This chapter discuss the results and performance of the clustering models which is implemented as a part of this research to identify the at-risk students. In the experiment 1 choosing the number of clusters for the aggerated data is discussed and in experiment 2,3,4 comparison of clustering model is discussed to identify the best model which has less overlap of data points between the clusters.

5.1 Experiment 1: Gap Statistics

Statistical testing methods was used to find the numbers of clusters in a dataset and in this research, it is found using Gap Statistics. Gap statistics is used in the (MacEdo *et al.*, 2019) has a statistical method to find the optimal clusters for the dataset. Figure 3 shows the results of gap statistics, that increase in the average sum of square within the clusters from 1 to 10 gets the elbow point at 3. Therefore, 3 optimal clusters can be used in clustering model to cluster the students using the aggregated dataset.

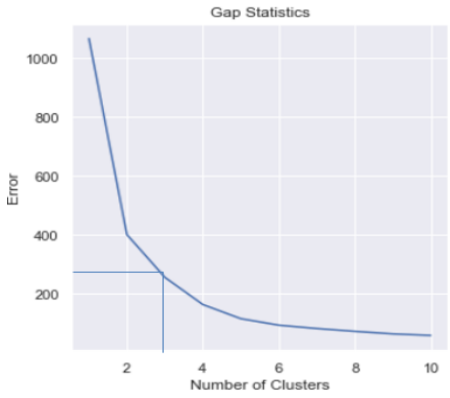

Figure 3. Result of GAP Statistics

Result of the Gap statistics was used in below clustering models to defines the clusters, so that result of the clustering model will be three different clusters for the input data.

5.2 Experiment 2: Fuzzy C-means vs Gaussian Mixture

In this experiment Fuzzy C-means and the Gaussian Mixture models were build and their result are compared. Model performance is compared using evaluation metric and scatter plot. Table 2 shows the metric result of the Fuzzy C-means and Gaussian Mixture clustering model.

Table 2: Fuzzy c-means vs Gaussian Mixture

Model	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
Fuzzy c-means	0.38	4731	0.94
Gaussian Mixture	0.51	3152	0.67

In this experiment of comparison between two clustering model, proposed model gaussian mixture outperformed fuzzy c-means model. Silhouette score of gaussian model show that 13% increase and Davis score of 27% when compared to the fuzzy c-means model. However, Calinski index metric which explains how well the data points are separated from other clusters shows less result for gaussian model. Figure 4 shows the scatter plot of gaussian mixture model which shows that the data are overlapped in cluster 1 and 2.

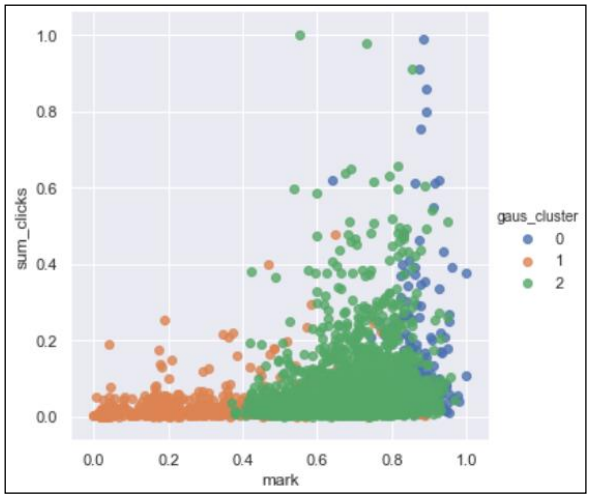

Figure 4. Gaussian model scatter plot

Therefore, to reduce the overlapping of the data points in the cluster the model outperformed in this experiment is compared with hierarchical clustering model in the experiment 2.

5.3 Experiment 3: Gaussian Mixture vs Hierarchical

In this stage of research, hierarchical clustering model was built and compared it with the model which outperformed in experiment 1, gaussian mixture. And, same evaluation metric was used

to compare the result of the model and result shows that hierarchical model performed better than the gaussian in calinski and davis index.

Table 3: Gaussian Mixture vs Hierarchical

Model	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
Gaussian Mixture	0.51	3152	0.67
Hierarchical	0.52	4552	0.52

Table 3 shows the performance comparison of the model conducted in this experiment. For hierarchical model the calinski score and davies index has been increased when compared to the Gaussian model and silhouette score shows 1 % decrease in hierarchical model. Figure 5 shows the clusters formed using the hierarchical model.

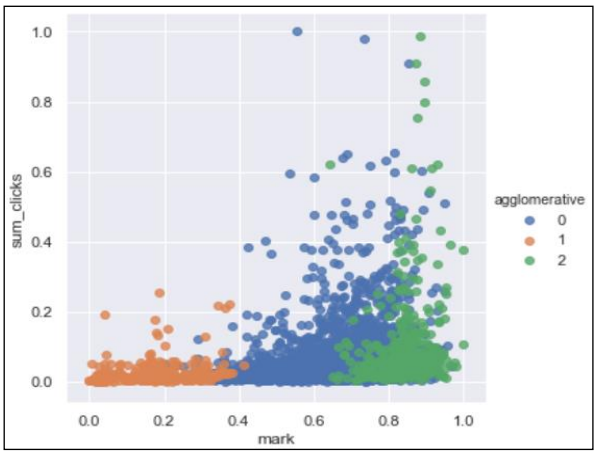

Figure 5. Hierarchical model scatter plot

From the hierarchical scatter plot, it is evident that hierarchical clustering model also have overlapping of datapoints between the clusters 1 and cluster 2. Therefore, to reduce the overlapping of datapoints between the clusters K-Prototype clustering model is built in next experiment and compared with the hierarchical model which outperformed in this experiment.

5.4 Experiment 4: Hierarchical vs K-Prototype

K-Prototype clustering model was implemented in this experiment and used different notion of distance to calculate the distance between the cluster. Totally 10 iterations were used to find the best separation of clusters and centroids. And the model produced the best result in the iteration 8. Table 4 shows the comparison of the performance of the model.

Table 4: Hierarchical vs K-Prototype

Model	Silhouette Coefficient	Calinski -Harabasz Index	Davies-Bouldin Index
Hierarchical	0.52	4552	0.52
K-Prototype	0.75	17847	0.28

K-Prototype clustering algorithm showed the better result when compared to all the experiment and the Davies Bouldin index is near to1 which indicates that the groupings of the students has been in better partition. And also silhouette and calinski value is higher when compared to the hierarchical clustering model which shows that the clusters are better defined in the k-prototype model. Figure 6 shows the clusters formed by the k-prototype. And the scatter plot shows that the clusters 1 and 2 are well partitioned and separated between the data points and overlapping of clusters is reduced in this model compared to the models implemented in the above experiments.

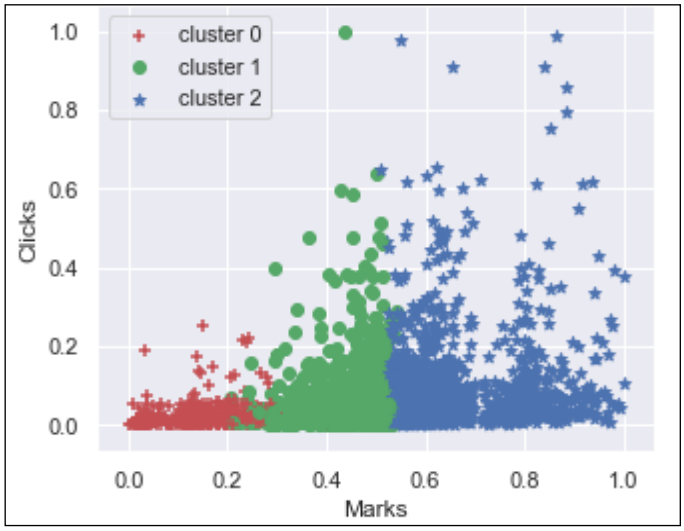

Figure 6. K-Prototype model scatter plot

5.5 Discussion

Result from the above conducted experiments shows that compared with all the clustering models used in this research, it is observed that k- prototype clustering model produced better partition of clusters compared to other models. The reason behind the performance improvement of k-prototype from the other models is, this model is designed to work on both categorical and numerical attributes in the dataset and also the distance between the data points to group the clusters is measured using two metrices, for numeric values euclidean distance is used and for categorical values similarity between the points are considered. In other model's categorical data is converted to numeric data before running the model through one-hot encoding which reduced the model's performance.

Figure 7 shows that, the silhouette coefficient score for k-prototype model is 0.75 which is 75 % and for fuzzy c-means which is 38 %. Therefore, the model showed 37 % increase in the separation of data points between the clusters in k-prototype model. Also, proposed hierarchical and gaussian mixture model also performed less compared to k-prototype.

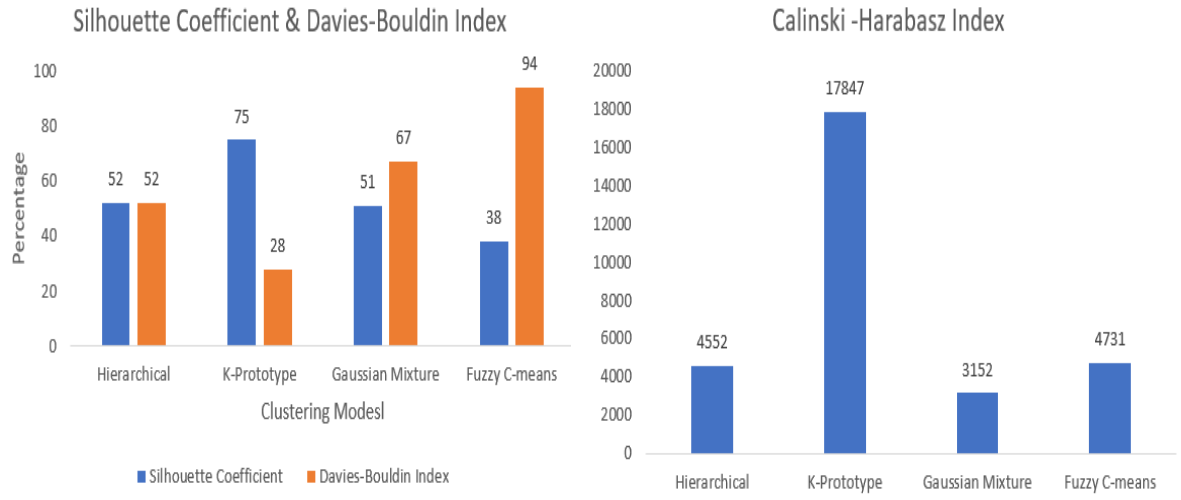

Figure 7. Comparison of evaluation metric for all the models

Another evaluation metric to find the variance of the data points between the cluster used is Calinski -Harabasz score. If the values of the score is higher than the cluster are dense and well separated. In this research calinski score for k-prototype is 17847 which is higher when compared to the other model. At last, Davies score is calculated for the scaled data, lesser the value of davies score better the separation of the clusters for k-prototype model the score is 28%. Table 5 shows the cluster labels that is observed in the clustering result of k-prototype model. Figure 7 shows that cluster 0, class represent the group of at-risk students with low interaction in VLE and low score in the modules. The cluster 1 are the marginal students who are also at risk are the group of students with medium engagement in VLE and attained low scores. Finally, cluster 3 class represent the distinction students who has high scores attained in assignments and high interaction in the VLEs.

Table 5. k-prototype clustering result

Cluster	Class
Cluster 0	At-risk students
Cluster 1	Marginal students
Cluster 2	Distinction students

Overall, the experiments conducted shows that the k-prototype model showed less overlapping of clusters compared to other model and identified the at-risk students with high performance.

6 Conclusion and Future Work

Identifying the at-risk students in the distance learning university is important because it allows the instructor to monitor the student's online activities for different courses. Therefore, in this research one of the biggest distance learning universities in UK, OU dataset is collected and formatted to actionable aggregated data in a form suitable for input to the clustering model. Then fuzzy c-means model and multiple clustering models has been applied on the data to

figure out the model which outperformed in identifying the at-risk students using evaluation metric.

The results of the experiment showed that K-Prototype clustering algorithm is the most appropriate in identifying the at-risk students in the VLEs compared to Fuzzy C-means and other proposed models showing the silhouette score of 75 % which indicates the clusters are better partitioned and davies score of 28 % which is near to zero which indicate the less variance between the cluster. Also, the result findings show that the clickstream behaviour of the students in VLE and academic success are the key factors which have an impact in identifying the at-risk students. overall, goal of this research project has been attained in identifying the at-risk students by using the k-prototype clustering model.

In the Future work, each student's day to day activity can be explored to get detailed understanding of student's behaviours in VLEs. Also, behavioural change of the students between the courses can also be analysed for examining student's behaviour. Mining the student's textual data from the feedback forms using the Natural Language processing from the VLEs can also be an important factor in identifying the student performance. Additionally, use of date attributes like assignments submission date and student's week wise interactivity in VLE can be used to build the model using time series which can result in monitoring the students daily or in weekly frequency.

Finally, this research work will be helpful for educational institution, learning analytics and future researcher in choosing the important attribute to identifying the at-risk students in the online learning environment and to figure-out how to pick the best performing clustering algorithm based on the clustering analysis in educational dataset.

References

Agnihotri, L. *et al.* (2015) 'Mining Login Data For Actionable Student Insight', *Proceedings of the 8th International Conference on Educational Data Mining (EDM)*, pp. 472–475.

Aljohani, N. R., Fayoumi, A. and Hassan, S. U. (2019) 'Predicting at-risk students using clickstream data in the virtual learning environment', *Sustainability (Switzerland)*, 11(24), pp. 1–12. doi: 10.3390/su11247238.

Boroujeni, M. S. and Dillenbourg, P. (2019) 'Discovery and temporal analysis of MOOC study patterns', *Journal of Learning Analytics*, 6(1), pp. 16–33. doi: 10.18608/jla.2019.61.2.

Chui, K. T. *et al.* (2020) 'Predicting at-risk university students in a virtual learning environment via a machine learning algorithm', *Computers in Human Behavior*. Elsevier, 107(December 2017), p. 105584. doi: 10.1016/j.chb.2018.06.032.

Corrigan, O. and Smeaton, A. F. (2017) 'A course agnostic approach to predicting student success from vle log data using recurrent neural networks', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10474 LNCS, pp. 545–548. doi: 10.1007/978-3-319-66610-5_59.

Corsatea, B. M. and Walker, S. (2015) *Opportunities for Moodle data and learning intelligence in Virtual Environments, 2015 IEEE International Conference on Evolving and Adaptive Intelligent Systems, EAIS 2015*. IEEE. doi: 10.1109/EAIS.2015.7368776.

Dalipi, F., Imran, A. S. and Kastrati, Z. (2018) 'MOOC dropout prediction using machine learning techniques: Review and research challenges', *IEEE Global Engineering Education Conference, EDUCON*. IEEE, 2018-April, pp. 1007–1014. doi: 10.1109/EDUCON.2018.8363340.

Hassan, S. U. *et al.* (2019) 'Virtual learning environment to predict withdrawal by leveraging deep learning', *International Journal of Intelligent Systems*, 34(8), pp. 1935–1952. doi: 10.1002/int.22129.

Hussain, M. *et al.* (2018) 'Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores', *Computational Intelligence and Neuroscience*, 2018. doi: 10.1155/2018/6347186.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) 'Data Descriptor: Open University Learning Analytics dataset', *Scientific Data*, 4, pp. 1–8. doi: 10.1038/sdata.2017.171.

Liang, J. *et al.* (2016) 'Big data application in education: Dropout prediction in edx MOOCs', *Proceedings - 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016*. IEEE, pp. 440–443. doi: 10.1109/BigMM.2016.70.

Macarini, L. A. B. *et al.* (2019) 'Predicting students success in blended learning-Evaluating different interactions inside learning management systems', *Applied Sciences (Switzerland)*, 9(24). doi: 10.3390/app9245523.

MacEdo, M. *et al.* (2019) 'Investigation of college dropout with the fuzzy c-means algorithm', *Proceedings - IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019*. IEEE, pp. 187–189. doi: 10.1109/ICALT.2019.00055.

De Morais, A. M., Araújo, J. M. F. R. and Costa, E. B. (2015) 'Monitoring student performance using data clustering and predictive modelling', *Proceedings - Frontiers in Education Conference, FIE*, 2015-Febru(February). doi: 10.1109/FIE.2014.7044401.

Navarro, Á. M. and Moreno-Ger, P. (2018) 'Comparison of Clustering Algorithms for Learning Analytics with Educational Datasets', *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(2), p. 9. doi: 10.9781/ijimai.2018.02.003.

Oztekin, A. *et al.* (2013) 'A machine learning-based usability evaluation method for eLearning systems', *Decision Support Systems*, 56(1), pp. 63–73. doi: 10.1016/j.dss.2013.05.003.

Preidys, S. and Sakalauskas, L. (2010) 'Analysis of students' study activities in virtual learning

environments using data mining methods', *Technological and Economic Development of Economy*, 16(1), pp. 94–108. doi: 10.3846/tede.2010.06.

Waheed, H. *et al.* (2020) 'Predicting academic performance of students from VLE big data using deep learning models', *Computers in Human Behavior*. Elsevier Ltd, 104(November 2018), p. 106189. doi: 10.1016/j.chb.2019.106189.

Wang, F. *et al.* (2016) 'Empirical comparative analysis of 1-of-K coding and K-prototypes in categorical clustering', *CEUR Workshop Proceedings*, 1751(c), pp. 248–259.

Yi, J. C. *et al.* (2018) 'Predictive analytics approach to improve and sustain college students' non-cognitive skills and their educational outcome', *Sustainability (Switzerland)*, 10(11). doi: 10.3390/su10114012.