National College of Ireland

Technology-enhanced Learning Impact Analysis through Categorization and Pattern Recognition by using Clustering Machine Learning Algorithm

MSc Research Project
Data Analytics

Neeraj Choudhary
x17156611

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Neeraj Choudhary
Student ID:	x17156611
Programme:	Data Analytics **Year** : 2020
Module:	Research Project
Supervisor:	Dr. Cristina Muntean
Submission Due Date:	17/08/2020
Project Title:	Technology-enhanced learning impact analysis through categorisation and pattern recognition by using clustering machine learning algorithm
Word Count:	7014 **Page Count** 24

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:	Neeraj Choudhary
Date:	17/08/2020

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):	

Technology-enhanced Learning Impact Analysis through Categorization and Pattern Recognition by using Clustering Machine Learning Algorithm

Neeraj Choudhary
x17156611

Abstract

STEM related learning and critical thinking is plays a pivot role in a student knowledge development. These skills build the foundation for a student to learn mathematics and programming concepts but in the recent time education institution have observed that student shows less interest in learning math and programming concept with traditional teaching approach which have forced the institution to explore new teaching approach and modernized there education structure. One efficient approach which most of the education institution had accepted is game based learning which allows younger students to learn concepts while playing interesting games or puzzles and based on the student feedback after they have played the game they can evaluate their development on programming skills and education institution can understand the impact of the game on the student learning. This project is to identify the efficient methodology to understand the impact and behaviour pattern of the students by using machine learning clustering algorithm. The main aim of this research to highlight the limitation of the current approach of evaluating the student's feedback and provide the solution to it with the help of clustering algorithm. The three game (variable, function, loop) data has been imported from the Newton platform and process with the clustering algorithms and evaluated based on how many clusters and efficient grouping of students has been created. The results show the Kmean is more efficient than agglomerative hierarchical clustering in grouping and defining the potential number of the cluster to be used. The insight from this research could be used by National College of Ireland and Dublin City University for future analysis on game-based students' feedback.

1 Introduction

In the recent years, many discussions have been made on the modernizing the status of curriculum to keep students interested and make sure proper development in their STEM related learning and critical thinking skills. These skills are very important for their future success and development. Most of the educational institutions are looking for changing their current way of teaching by introducing game-based learning. Games are one of the important ways of teaching for centuries and the most important aspect of the game-based learning is learning and teaching from failure and repetition of learning goal. This concept has been applied for example in a video game-based learning where students are provided with video games and they were asked to play and learn programming skills by solving tasks from different

level of games. The level is designed based on the difficulties level related to each learning concept. Textbook learning has always been preferred and educational institutions have their concern to use technology-based learning over textbook approach. The change is slow, and it is good for some cases where it is important the new approach should be adopted only after being tested. On the other hand, it may also lead to the situation of sluggish.

As game-based learning (Huizenga, Admiraal, Akkerman and Dam, 2009) is designed in such a way that students enjoy the playing activity at the same time they solve different puzzles or questions while learning programming concepts and skills. Before designing a game, a proper process is followed where curriculum and concepts need to be taught to the student which leads the type of games are designed. This approach easily addresses the primary, secondary, and tertiary learning style where a game can include multiple concepts together and make it easier for a student to learn while playing it.
Feedback process where student is asked to provide their feedback related to the game and learning activity is a very important to understand how effective game-based learning is against the traditional learning styles. Analysis has been done based on the pre- and post- tests or questionnaires approach that assesses average score as well as based on, statistical models that group students into categories. Each category consists of students that are identical in terms of their behaviour related to the game. The current approaches used to analyse the feedback data to understand the student behaviour are not based on the quantitative approach. Most of the time institutions use statistical approach that involves pre and post assessments to measure the student's acceptance towards the game to group the students into identical behaviour group. However, most of the time this approach is based on the human assumption rather than based on the methodology. That is why it is very important to group the students into identical behaviour which should be done with the help of the quantitative approach.
A quantitative approach involves processing of the collecting dataset such as revenue, market share during analysis related to the financial dataset and evaluate the result to make the recommendation. In this research processing of the game-based dataset which includes questionnaires related to the game, student's info and student answers towards the game-based questionnaires follow by evaluating with machine learning clustering algorithms to recommend the group or cluster of students.

This research project aims to provide the answer regarding the following question:
Is machine learning clustering algorithms such as Kmean and Agglomerative hierarchical clustering for categorization groups of students based on their feedback for an educational game is efficient compared to the statistical categorization methodology?

The research objectives are: -

1. Use machine learning clustering algorithms to group the students based on their response/feedback for a questionnaire.
2. Develop a methodology to be used in the analysis of the feedback data based on the machine learning to categorize the students into groups or sub-groups based on their response.

This research document is structured as follows: Section 2 discuss the literature review based on the current approaches used in evaluating feedback, methodologies, and limitations of the current approaches. Section 3 presents the methodology used in this research, Section 4 presents the design specification, Section 5provides implementation information of the approach and technique used, Section 6 presents results analysis while Section 7 concludes the research and further research avenues are presented.

2 Related Work

Recently, research work is performed on e-learning methods and platforms to be used to improve student's knowledge and problem-solving skills. To understand - learning various methods has been used and they are discussed and summarised below. E-learning has been introduced in education to improve the problem-solving skills and programming skills of the students and evaluate feedback collected by the students. Research done by (Kazimoglu, Kiernan, Bacon and Mackinnon, 2012) talked about game-based learning to improve the computation thinking and concepts of programming. This research explains how to program a robot to help the students to learn programming construct to improve computation thinking skills. There are six levels in the game to teach programming concepts such as computational thinking, designing algorithm to design your solution, condition logic to achieve high score in the game, simulation and debugging. This research was evaluated by collecting feedback from 25 students. They all agreed that the game helped them to improve computer programming skills and to develop computer programming solving skills. However, there is no critical methodology has been used to categories the students feedback.
The research reported by (Burgos, Nimwegen, Oostendorp and Koper, 2007) shows e-games importance for learning and improvement based on the feedback from the users. In this research planning educational task approach has been applied and students were asked to use the game with two versions: with feedback and with no feedback. The feedback version of the game provided the users with information related to the actions, moves and hints to solve the problem whereas in the no feedback game version, users received no information regarding the moves or actions they should take to solve the puzzle of the game. This research was evaluated by analyzing the feedback and no feedback data using ANOVA statistical methodology to report results with p value < 0.05 significant level. The evaluation results shows that ANOVA statistical methodology had limitation in processing the feedback data which shows that it's was not very efficient to understand the student behavior based on the feedback.

A learning platform called Scratch was introduced to students to improve the concepts of mathematics and computation programming (Ferrer-Mico, Prats-Fernàndez and Redo-Sanchez, 2012). The aim was to evaluate the impact of the Scratch learning platform on the students by collecting feedback from two groups consists of 19 and 22 students. Total 5 different questions related to the platform such as is platform able to tech , student experience , how useful is the platform , time importance and feedback usefulness were provided to the students and collected there response in terms of agree , disagree , neutral , agree and strongly agree .The result had been evaluated which shows positive impact of the platform on the

student and their concepts' related to learning has been improved but no quantitative analysis had been performed to categorized the students behavior towards the game.
The motivation and computation skills were improved by introducing digital game in the research (Kazimoglu, Kiernan, Bacon and MacKinnon, 2012) where students had been asked to complete the following tasks such as Problem identification & decomposition , creating efficient and repeatable patterns , practice debug mode , practicing run-time mode and brainstorming to developed problem solving, building algorithm , debugging, simulation and socializing skills. 25 students have provided there feedback on the experience of the game and improvement in their knowledge regarding programming concepts and also recommend to add new features such as introducing achievement section after every task to rewards the players after they demonstrate good practice in programming as this will increase their motivation level to continue playing the game but no statistical analysis has been performed to evaluate the impact of the approach and categorization of the students behavior towards the game has been done.
The research (Chi and Jain, 2011) focused on improving teaching computations skills by using visualization computations tool. In this research a visualization tool had been used with computation project to provide students a better understanding of the concepts. Hands on labs were designed to provide support to students to understand the problem and try to find the solutions by using the tools. Students feedback had been collected on the pre and post survey at beginning of the semester and end of the semester by asking students 10 questions related to the computation's concepts. 200 students have took a part in the survey and 85% of student think that visualization tool is important for learning, but no statistical or quantitative analysis had been performed to measure the impact and categorized the student into different group based on their behavior towards the game.
This research (Makri, Choudhary and Muntean, 2019) introduced Loop game for teaching loop programming concept.. The paper reports results of a pilot that was run at National College of Ireland and used the Loop game and it was run at National College of Ireland and evaluated student's improvement in learning OOP concepts through pre and post assessment. 23 students took part in the pilot. Statistical approach pre and post T test analysis had being applied to analyze the student feedback which shown improvement in learning the loop concepts after students have played the Loop game but no quantitative analysis had been performed to categorized the student into different groups based on their response towards the questionnaires which could have helped to understand their behavior towards the game.
A recent research reported in (Sharma et al., 2019) analyzed students learning improvement in a programming module by using multimedia assistant learning and problem-based learning methodologies . This research shows that in the recent year students' interest in the science filed has experienced decline across Europe which became a concern among educational institutions. One of the main factors are due to not enough technical and scientific staff's availability. In this research multimedia assistant learning and problem-based learning teaching approach has been introduced to the student of Dublin city university as part of the Newton (NEWTON EU HORIZON 2020 Project, 2020) large scale programming pilot which shows improvement of programming concepts among students as overall 84.38% of total students scored above 80 out of the 100 related to the programming projects and also assessment based on the students response in terms of feedback towards the questionnaires shows satisfaction among students related to the approach but no quantitative evaluation had been done to measure the impact of the approach and categorization of students into different

group based on their response towards the questionnaires which is important to understand the student behavior.
(Zhao et al., 2019) research based on the Newton loop game to improve the students programming concepts related to the JAVA programming. In the recent times among European countries has observed low interest among the students regarding STEM subjects which is due to no improvement has been made over the year in the traditional teaching methods. In this research Newton loop game has been introduced at national college of Ireland as part of the case study to assess the game effectiveness in terms of knowledge improvement related to programming concepts among 31 students. Assessment were made based on the pre and post average score range of 1 to 5 which shows average score has been 2.12 out of 5 compared to 1.29 out of 5 once they have played the game. The categorization of student into multiple groups had been done based on the statistical approach result but categorization efficiency could be improved with introducing quantitative analysis as statistical pre and post approach has limitation and it could lead to biased categorization of the student.
Research done by (Hussein, Ow, Cheong and Thong, 2019) to improve critical thinking skills by using digital game based DGBL approach. A total 127 students have participated in this game-based learning activity where two group has been formed where one group exposed with experimental group of developed game whereas other group of control group with conventional method of learning. This research has been evaluated by pre and post statistical T test score which shows there is a significant improvement of the critical thinking skills among the students but efficiency of the analysis could be improved with grouping of the student on pre and post by using quantitative analysis to understand the impact of the game on their behavior which is useful for future improvement of the game.
(Hussein, Ow, Cheong and Thong, 2019) research based on the game-based learning impact on the motivation and efficiency of the student learning. In this research, with the help of the game they were trying to improve the science concept among the students by providing the card game called Conveyance Go. The feedback has been taken from the students by asking them to provide their response for each questionnaire related to game feedback which based on hypothesis, induction, explanation, and evaluation and calculate the total score against 100. The evaluation had been based on pre and post statistical T-test which shows the majority of the student had accepted there learning had been improved related to science concepts and willing to keep using this game card but impact and categorization of the students to could have helped in order to understand the pre and post behavior change which had not been implemented in this research.

In this literature review we have observed the limitation in the analysis approach towards the feedback data which had been collected to understand the impact and improvement of the student learning. Most of the analysis concentrated on the statistical approach or human based assumption to understand the impact of the game-based learning on the student problem solving skills and critical thinking which could lead to the biased results. Therefore, in this research we will try to introduce the quantitative approach such as machine learning clustering methodology to form unbiased group of the students more efficiently based on the their identical response towards the game in order to understand their behavior patterns.

3 Methodology

This main aim of this research work is to apply clustering algorithms which is efficient to find out the group of feature which are linked or similar to each other and each group are different from the other groups. In this research we have applied two clustering algorithm such as Kmean (Li, 2020) and Agglomerative hierarchical clustering (Naeem, Rehman, Anjum and Asif, 2019) to categorize the students into groups based on their responses towards the Newton game such as function, loop, and variable. These two methods are very useful and efficient in processing the large sample size dataset and having different varieties of feature compared to other clustering approach such as Dbscan which has limitations in processing dataset with different feature. The clustering algorithm is very useful approach to categorise the data into different groups based on the similar behaviour or qualities. In this research it is important to understand the student behaviour pre and post scenario. The data mining technique Knowledge Discovery in database (KDD) methodology (Wang, 2014) is been used in this research and follow the following steps:

1. Business Understanding
2. Data Understanding
3. Data Pre parathion or Data Pre processing
4. Data Mining

The following flow chart in figure 1 shows the steps we will follow :

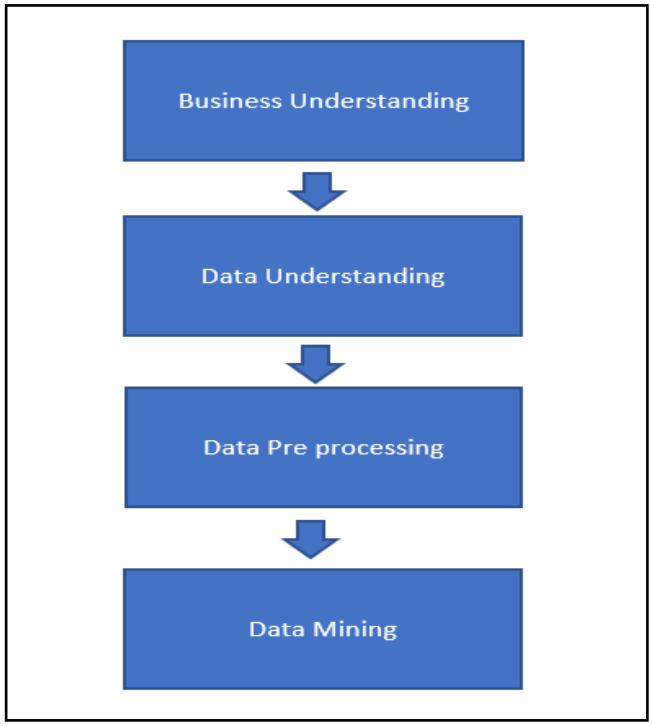


Figure 1: Flowchart to showcase the steps to be follow in KDD

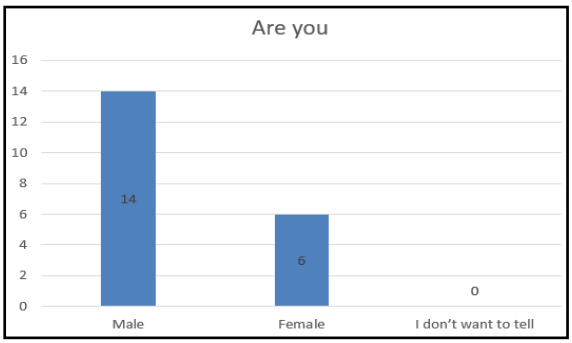
3.1 Business Understanding

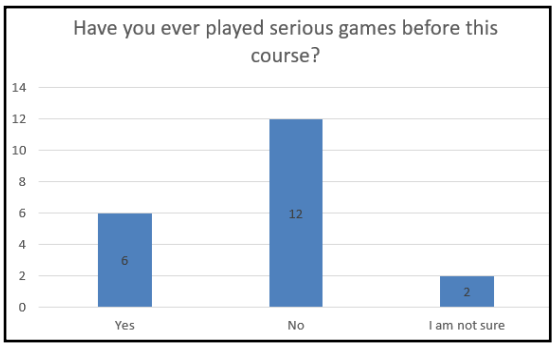
Improvement of the Programming and problem-solving skills of the students by introducing game-based learning gave good results in the recent times. However, application of the machine learning is important for the assessment of the results related to the student's

improvement once they have started using e-based learning platform. Lots of evaluation methods has been used in e learning research to understand the improvement of the student's skills but there is still no clear single and effective methodology to assess the improvement of knowledge with the application of the E-learning. As discussed in the related work summary, machine learning clustering methodology is used in this research. Clustering methods are very efficient to assess the behaviour with the students in terms of the improvement and feedbacks related to the game-based learning. In recent year due to no single way to assess the improvement on computation skills of the students due to game leads to the situations where each way of the assessment can only apply to the certain area of the feedback. to overcome this, we are going to use clustering methodology which is efficient and can be applied to all different evaluation dataset.

3.2 Data Understanding

In this research we have used programming large scale pilot dataset where 133 students' feedback had been taken for game-based learning which is part of the Newton research project (NEWTON EU HORIZON 2020 Project, 2020). The pilot was deployed for two educational institutions: National College of Ireland and Dublin City University.
This Newton game based learning data set consists of various demographic data related to the students which includes information such as gender , age , gaming experience , learning outcome experience in math, science and technology and also students responses related questionnaires on learning impact related to three different game such as Loop, Function and Variable games which includes programming concepts related to C and java programming. These timeline of the datasets is 12- week programming large scale pilot that was run during one semester in both National College of Ireland and Dublin City University .
Game based data set consists of 15 questionnaires based on game feature and learning such as how useful the game in order to understand the programming concepts, is any further improvement student think should be made in the game? or is game based learning is better than textbooks based learning and student response such as agree, strongly agree , neutral , disagree and strongly disagree. Considering all these categories of the Newton dataset, all data preparations steps are discussed in the following section
We have analysed the survey data of National College of Ireland and Dublin City University to understand about the student background. Figure 2 shows the survey data descriptive analysis for National College of Ireland in which out of total 20 students 14 are male and 6 are female and majority of them were playing the serious game for first time and they showed the interest in learning science, technology and maths.





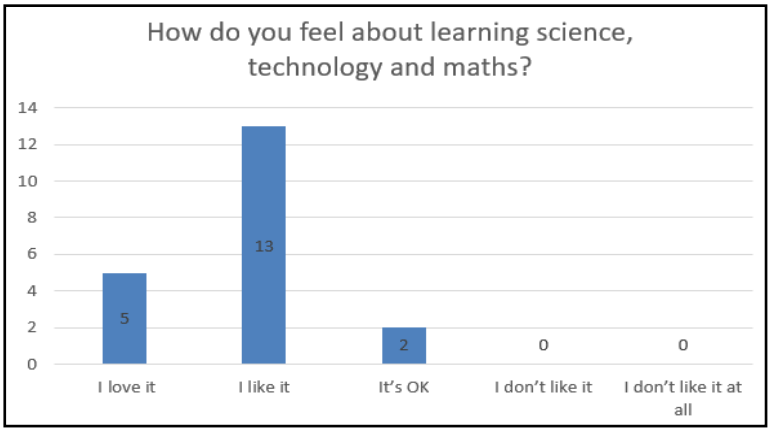
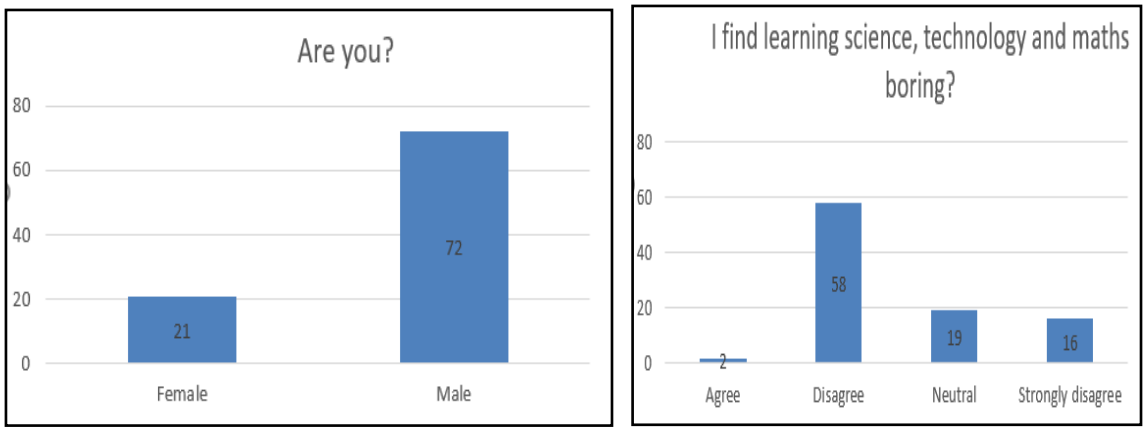


Figure 2: Structure of the survey dataset of National College of Ireland

Similarly ,Figure 3 shows the survey descriptive analysis for Dublin City University in which majority of student who have participated are male (72 out of 93) and 21 (out of 93) are female and majority of them have shown there interest in learning science, technology and maths and also gave positive response towards learning science , technology an maths from game based learning.



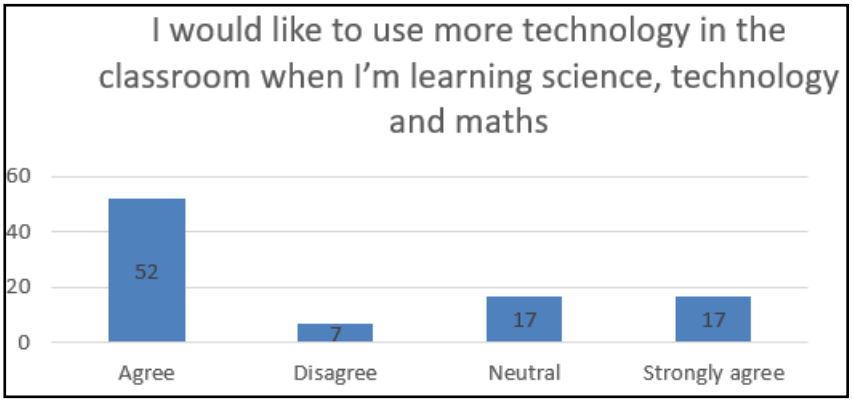


Figure 3: Structure of the survey dataset of Dublin City University

3.3 Data Preparation

Newton platform game-based dataset consists of student's feedback towards the questionnaires related to game (Loop, Function, and Variable) and demographic survey where information such as gender , whether student have any experience with the serious game and will they like learning science , maths and technology from game has been collected. Newton dataset has been pre-processed before processing it with the clustering algorithm and survey data has been analysed to understand the structure of the dataset.

3.4 Data Cleaning

Newton game-based dataset which include game such as Variable, Loop and Function consists of students id, game-based questionnaires, and student responses such as agree, strongly agree, neutral, disagree and strongly disagree. The current format of the dataset need to preprocessing to overcome the format issue related to the structure of the dataset before using it with the clustering algorithm. The student response needs to convert from the categorical format to the numerical range such as 1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree as shown in the table 1. As we need numerical format response dataset to process it by clustering models Kmean and Agglomerative Hierarchical to create a cluster or group based on the identical response behavior.

Strongly disagree	1
Disagree	2
Neutral	3
Agree	4
Strongly agree	5

Table 1: Formatting of the response with the numerical range

Null value present in the dataset has been formatted with mean value and scaled the dataset to make the range of the dataset in between -1 to 1 which is necessary to process it with the clustering algorithm. This scaling help in increasing the efficiency of the clustering modeling.

3.5 Modelling

This section explained the implementation of the clustering modelling and architecture of implementation of Kmean and Agglomerative Hierarchical clustering. Clustering modelling is used to understand the structure of the dataset and its main aim to identified and defined the group, sub group based on the feature and structure of the datasets by measuring it against the Euclidean distance or correlation based distance. Clustering is used for unsupervised learning where structure of the dataset is unsupervised.

3.6 Kmean Clustering

Kmean is called as an iterative algorithm which defined the dataset into kpre -defined unique subgroup in which data points are unique and linked to only one group. Its main aim is to create group or cluster on the similar structure of the data sets and other cluster or group different from the other clusters. The data points assigned to each cluster is based on the sum of the square distance between the data points and centroid of the cluster is to be minimum. If

we have less variation within cluster then we have more homogenous data points are related to the same cluster.

3.7 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical clustering also called as an agglomerative nesting is a method which is used to cluster the objects based on the common behaviours or data points. The implementation is based on the considering data points or also called an object as a one cluster follow by keep merging all the data points or object until we will have one cluster which includes all data points or objects. This approach creates a tress-based diagram which we called as a dendrogram. There are two types of approach in the agglomerative clustering are bottom up or top down. In the bottom up approach, during each iteration similar behaviour cluster combined and formed a node which we called as a big cluster and in top down approach also called as a devise analysis where each iteration most common or heterogenous node or cluster is break down into two and keep iteration until they object get into bigger cluster.

4 Design Specification

In this research we will be using design called 2-tier which is represent in the Figure 4 in which we are using the game based data set which we have to took from the E learning game based platform called Newton and its present in the unsupervised format. So, we will start with the pre-processing of the dataset to format the data as per the clustering algorithm so to obtain the results. We have used programming language python in this research for implementing clustering algorithm such as Kmean and Agglomerative Hierarchical clustering. Descriptive analysis has been done by using google spreadsheet and python programming to format and understand the structure of the dataset. We have used libraries such as pandas and NumPy for data import and data manipulation and for visualization we have used matplotlib and seaborn libraries.

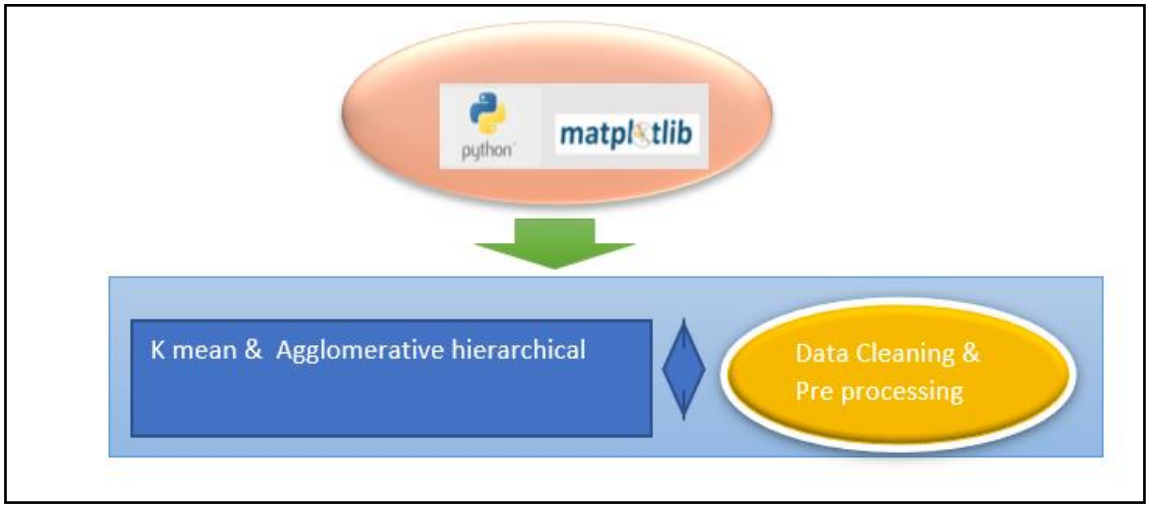


Figure 4: Design of 2- tier architecture

4.1 Architecture of the Kmean model

The implementation of the Kmean follows the following steps:

1. First step is to define the number of cluster K.
2. Arrange the data first to initialize the centroids and select k points randomly related to centroids.
3. We must keep iterating until there is no change to centroids.
4. Squared distance between data points and centroids must summed up.
5. Clusters are been assigned with the data points which are closest.
6. Take the average of all data points as per assigned clusters to calculate the centroid.

The equation 1 gives us the mathematical representation of the Kmean.

$$distance(x, y) = \sum_{i}^{n} (x_i - y_i)^2$$

Where,

n is the number of the cluster centre.
x- y is the Euclidean diatnce between x and y
x is the data point
y is the centres

4.2 Architecture of the Agglomerative Hierarchical Clustering

The implementation of the Agglomerative Hierarchical clustering follows the following steps:

1. Import the data and pre-process the data.
2. For each pair of the data or objects, compute the similarity.
3. Linkage function is being applied to combined object or common group into the tree or also called as a hierarchical tree where common objects are linked with the help of linkage function,
4. In this step we decide whether to cut the tree into cluster or not.

5 Evaluation

In this section we will discussed about the evaluation of our research results based on the objective of this research work. We have decided the number of clusters in Kmean based on the cluster error for each game such as loop, variable and function whereas in Agglomerative Hierarchical clustering its deicide the number of clusters based on the cut off. Once we have identified the number of clusters, we have created cluster or group and assigned each student to specific cluster based on the similarities related to response for each questionnaire. We have created cluster distribution visualization for each game games (Loop, Variable and Function) is discussed below.

5.1 Kmean

Kmean clustering method has been applied to the variable, loop, and function game. The table 2 represents the number of clusters can be selected based on the cluster error where x axis represents the number of the cluster and y axis represent cluster error.

The figure 5 represents pre stage of the clustering where each question represents a specific cluster and number of the student grouped to each question. This shows how to categorise the student into the group as per each questionnaire for three games (variable, loop, and function). this is very useful to understand the behaviour pattern of the student response across each questionnaire.
The Figure 6 represents elbow method to understand number of clusters to be used in the implementation based on the cluster error. In the case of variable game based on the cluster error (810 .288) after that cluster error is constant, number of clusters to be formatted is decided which is 6.

	num_clusters	cluster_errors
0	1	1980.000000
1	2	1446.800691
2	3	1154.310269
3	4	996.235164
4	5	893.031436
5	6	810.288487
6	7	721.667042
7	8	621.064089
8	9	570.618883
9	10	504.950221
10	11	455.091632
11	12	431.678442
12	13	358.462494
13	14	329.157103

Table 2: Cluster error per cluster for variable game

	num_clusters	cluster_errors
0	1	1980.000000
1	2	1446.800691
2	3	1154.310269
3	4	996.235164
4	5	896.233555
5	6	792.596788
6	7	698.678376
7	8	634.851896
8	9	565.180009
9	10	514.218799
10	11	455.091632
11	12	402.072707
12	13	378.731955
13	14	323.641721

Table 3:Cluster error per cluster for function game

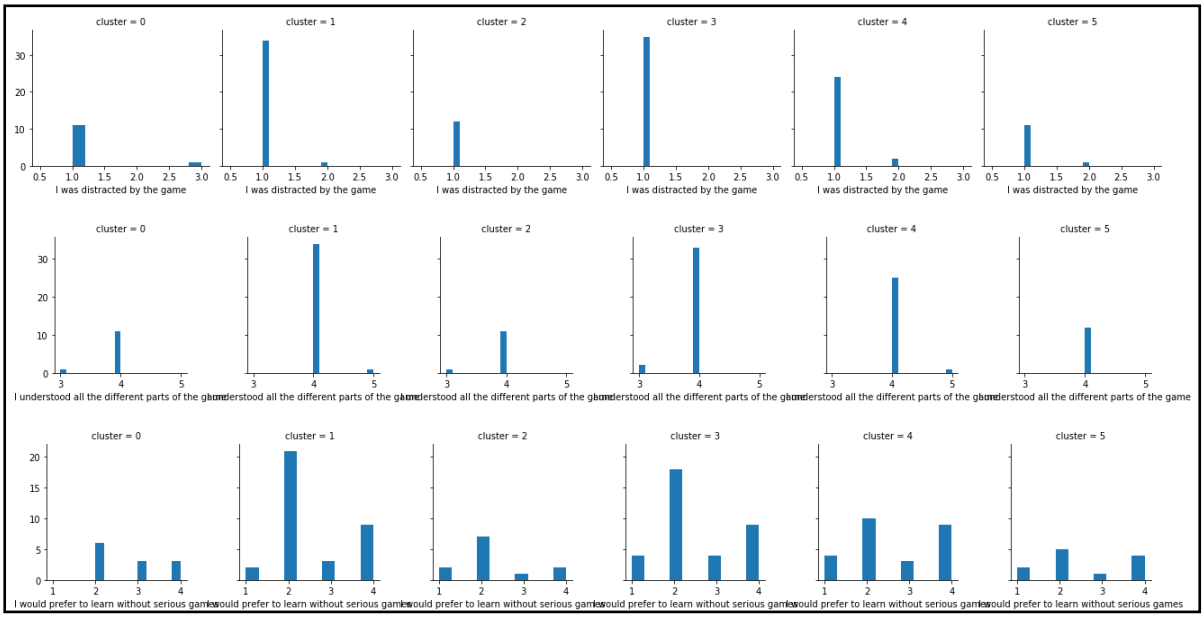


Figure 5: Grouping of the student to each questionnaires and clusters

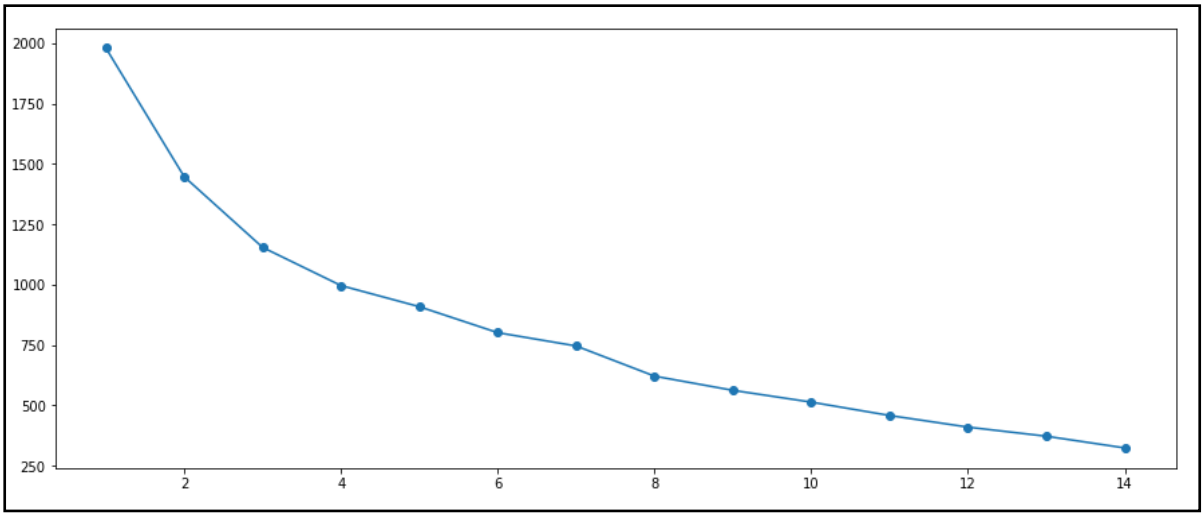


Figure 6: Cluster error per number of clusters for variable game

Figure 7 shows the results of the categorisation of the student into different clusters based on the patterns of their response towards the questions related to the variable game.

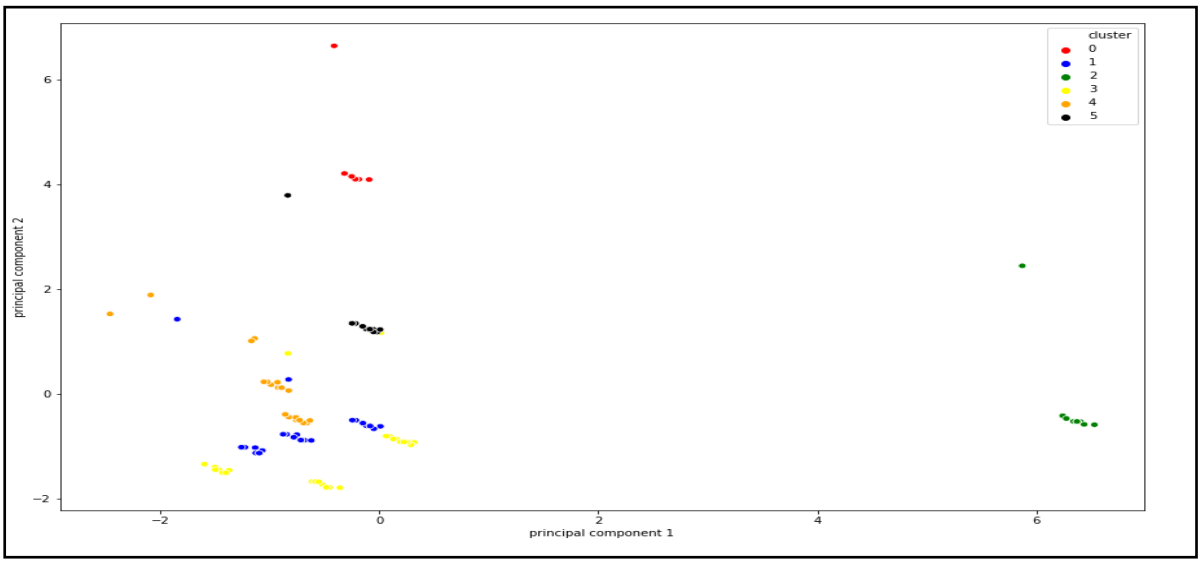


Figure 7: Categorization of the students belongs to the variable game into clusters

The table 3 represents the number of clusters can be selected based on the cluster error where x axis represents the number of the cluster and y axis represent cluster error for function game.
The Figure 8 represents elbow method to understand number of clusters to be used in the implementation based on the cluster error. In the case of function game based on the cluster error (698.67) after that cluster error is constant, number of clusters to be formatted is decided which is 7. Figure 9 represent group of the student's categories into different cluster clusters based on the patterns of their response towards the questions related to the function game.

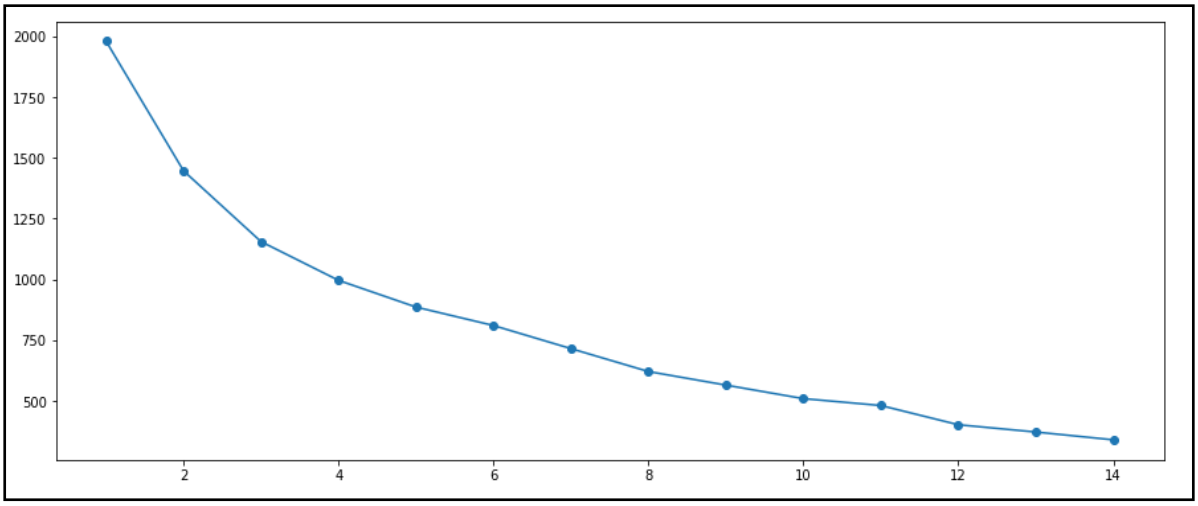


Figure 8: Cluster error per number of clusters for function game

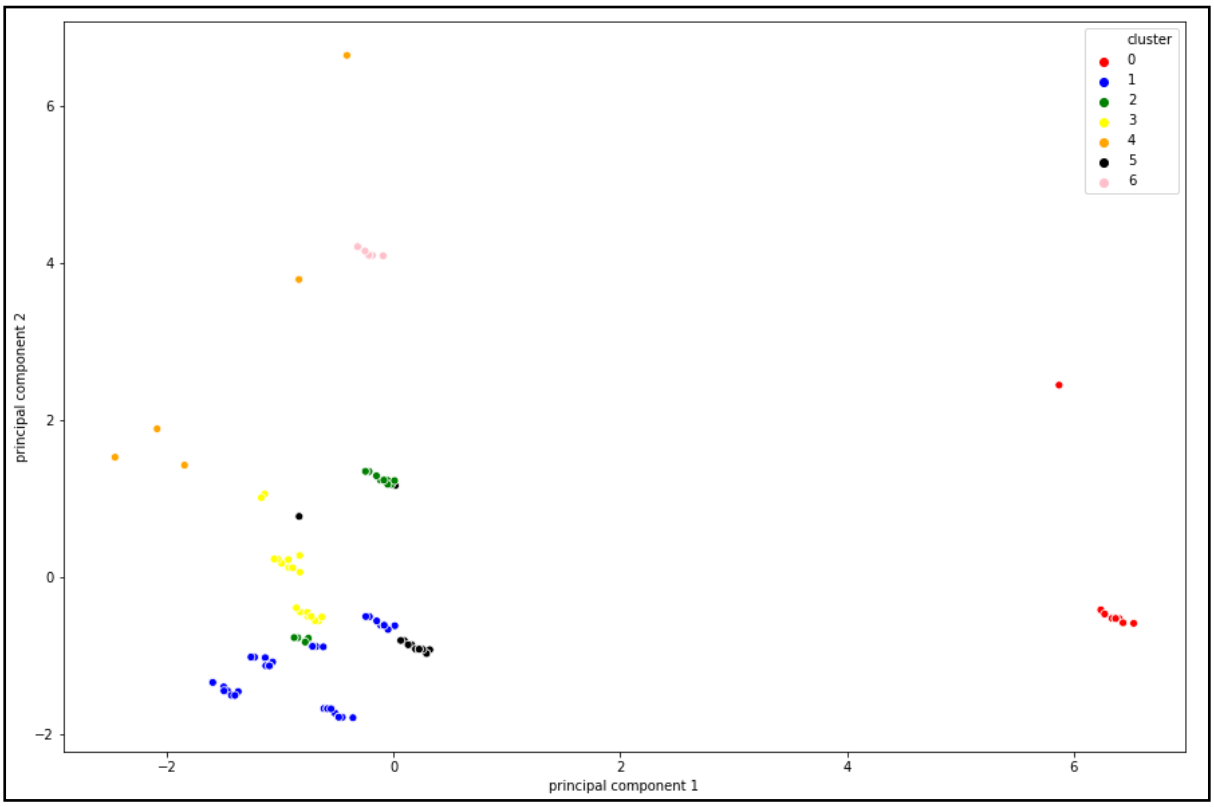


Figure 9: Categorization of the students belongs to the function game into clusters

	num_clusters	cluster_errors
0	1	1980.000000
1	2	1446.800691
2	3	1154.310269
3	4	996.235164
4	5	906.592778
5	6	792.596788
6	7	709.179922
7	8	640.984223
8	9	569.573130
9	10	506.582592
10	11	458.532477
11	12	415.749891
12	13	378.979007
13	14	337.946543

Table 4: Cluster error per cluster for loop game

The table 4 represents the number of clusters can be selected based on the cluster error where x axis represents the number of the cluster and y axis represent cluster error for loop game. The Figure 10 represents elbow method to understand number of clusters to be used in the implementation based on the cluster error. In the case of loop game based on the cluster error (709.17) after that cluster error is constant, number of clusters to be formatted is decided which is 7. Figure 11 represent group of the student's categories into different cluster clusters based on the patterns of their response towards the questions related to the function game.

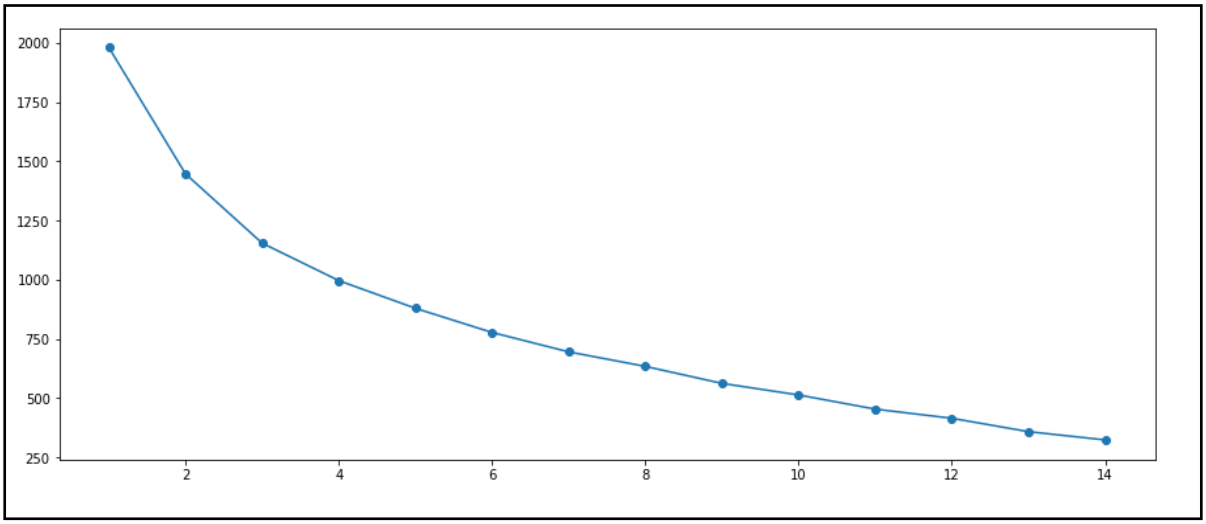


Figure 10: Cluster error per number of clusters for loop game

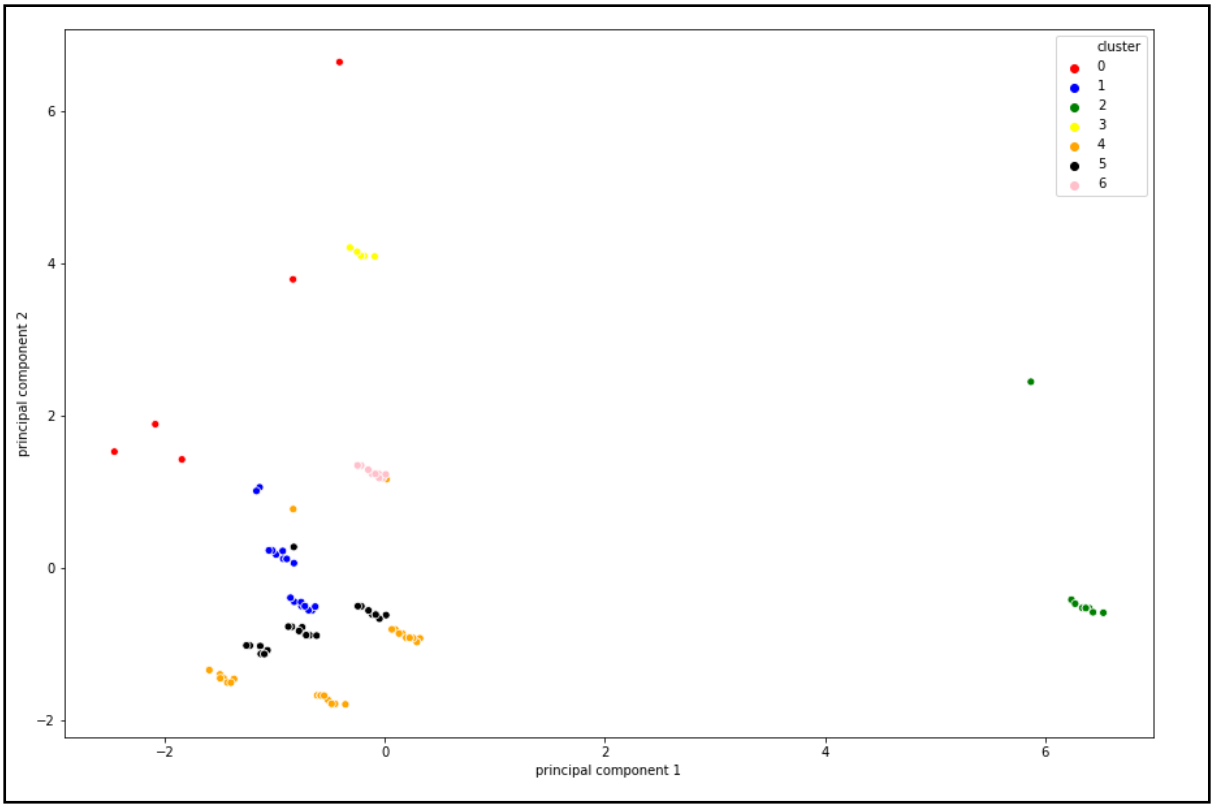


Figure 11: Categorization of the students belongs to the loop game into clusters

The following table 5,6,7 represent the result on number of the cluster and size of each cluster which represent the number of the student for variable, loop, and function game. This clustering of the students into different group based on the output of the Kmean shows the behaviour pattern of the students towards the game.

Number of the Cluster (K)	Sample Size (No of the student to each cluster)
0	23
1	13
2	24
3	12
4	48
5	12

Table 5: Represents the results of the variable games

Number of the Cluster (K)	Sample Size (No of the student to each cluster)
0	24
1	12
2	22
3	12
4	25
5	36
6	1

Table 6: Represents the results of the loop games

Number of the Cluster (K)	Sample Size (No of the student to each cluster)
0	12
1	50
2	16
3	25
4	5
5	13
6	11

Table 7: Represents the results of the function games

5.2 Agglomerative Hierarchical Clustering

The figure 12 represents result of tree-based representation on the variable game where x axis represents the number of the students and y axis represent the Euclidean distance. In node represent the distance shortest distance between two clusters which become node. In figure 12 also shows leaf which represent single data point and root which calculate the distance between new cluster and another member cluster outside the structure until data points get assigned to one cluster and based on the node cut off we have identified number of number for variable game is 4.

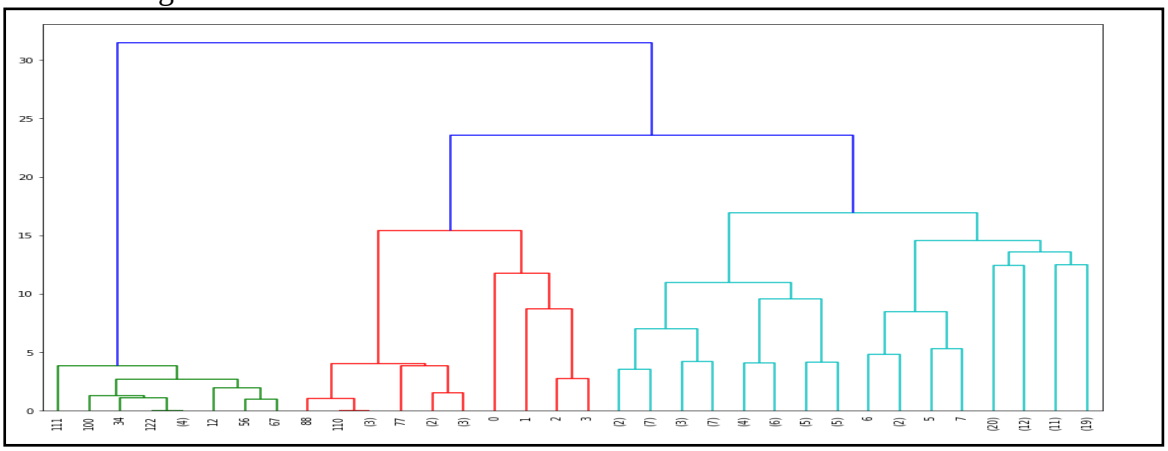


Figure 12: Dendrogram output of the variable game for Agglomerative Clustering

Figure 13 display the distribution of the students into different cluster where axis represent student numerical range response.

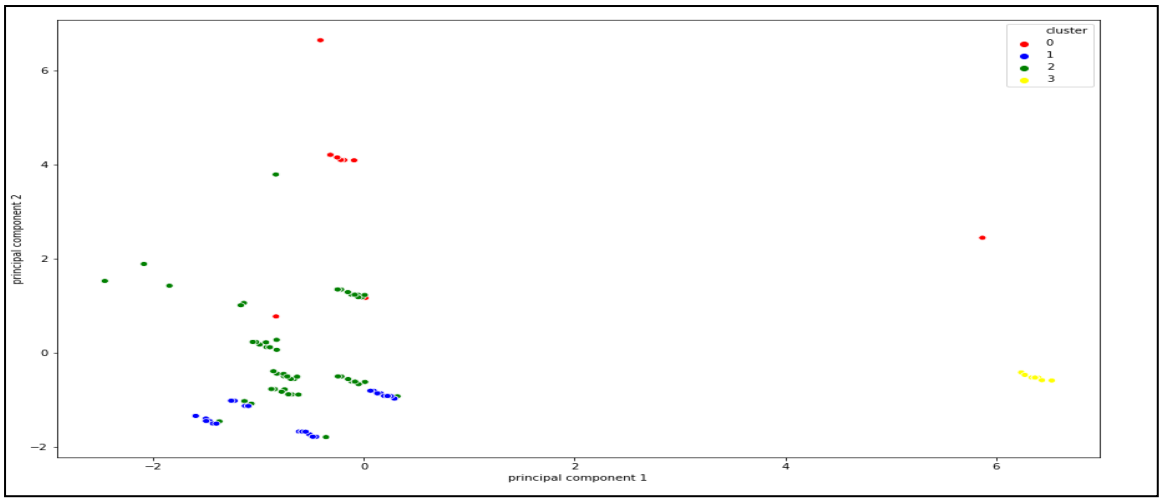


Figure 13: Cluster of the students belongs to the variable game into clusters

The figure 14 represents result of tree-based representation on the loop game where x axis represents the number of the students and y axis represent the Euclidean distance. In node represent the distance shortest distance between two clusters which become node. In figure 14 also shows leaf which represent single data point and root which calculate the distance between new cluster and another member cluster outside the structure until data points get assigned to one cluster and based on the node cut off we have identified number of number for loop game is 7.

Figure 15 display the distribution of the students into different cluster where axis represent student numerical range response.

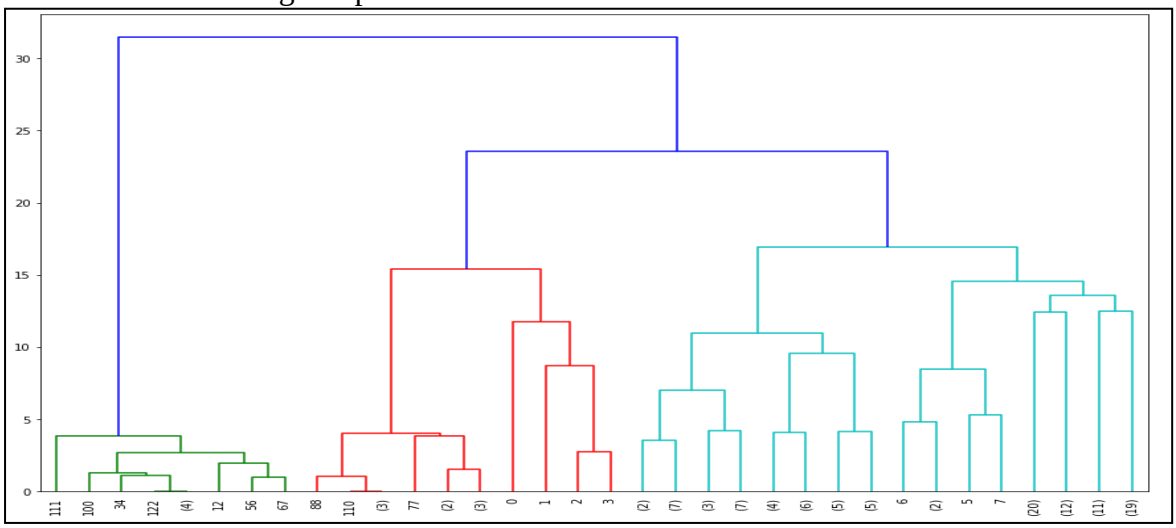


Figure 14: Dendrogram output of the loop game for Agglomerative Clustering

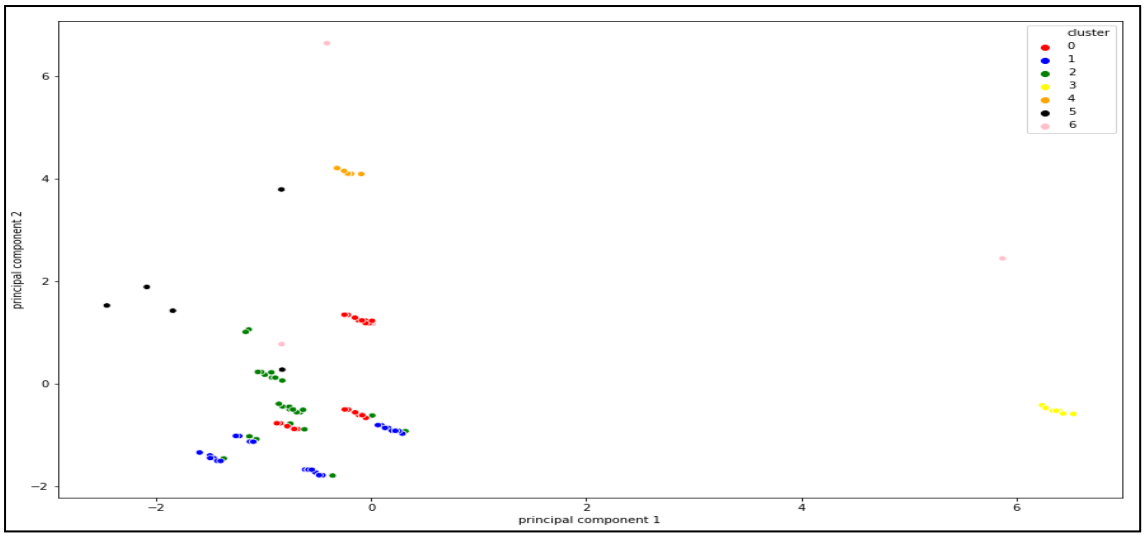


Figure 15: Categorization of the students belongs to the loop game into clusters

The figure 16 represents the result of tree-based representation on the function game where x axis represents the number of the students and y axis represent the Euclidean distance. In node represent the distance shortest distance between two clusters which become node. In figure 16 also shows leaf which represent single data point and root which calculate the distance between new cluster and another member cluster outside the structure until data points get assigned to one cluster and based on the node cut off we have identified number of number for function game is 5.
Figure 17 display the distribution of the students into different cluster where axis represent student numerical range response.

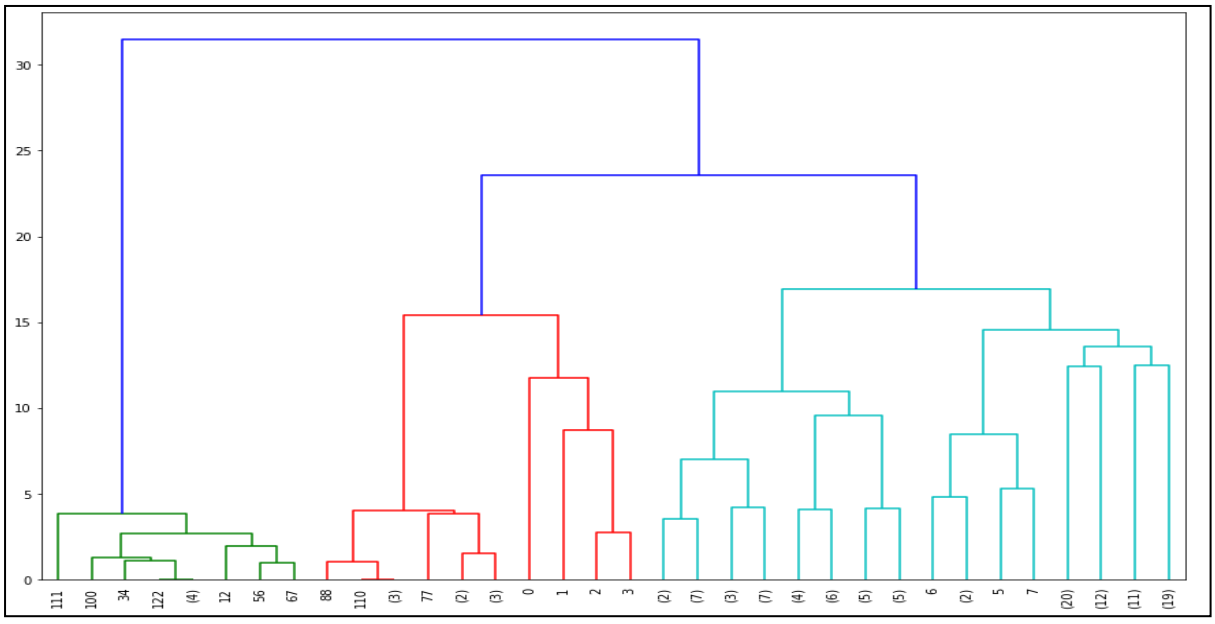


Figure 16: Dendrogram output of the function game for Agglomerative Clustering

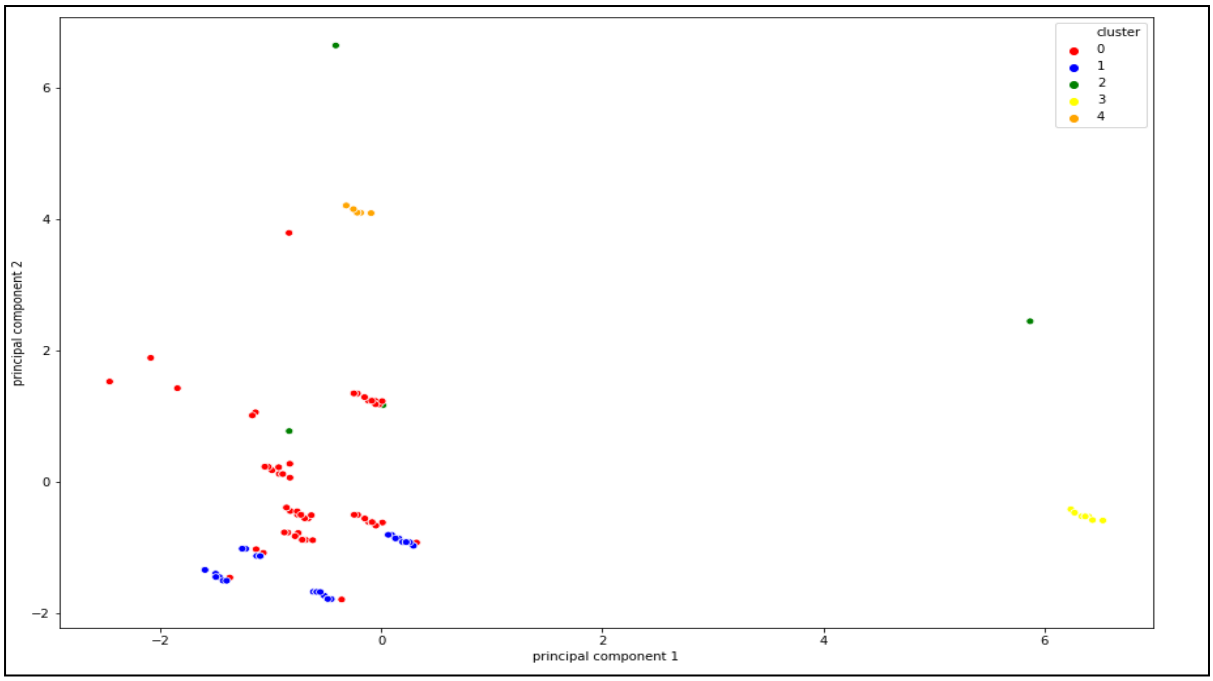


Figure 17: Categorization of the students belongs to the function game into clusters

The following table 8,9,10 represent the number of the cluster and size of each cluster which represent the number of the student for each game loop, function, and variable. This clustering of the students into different group based on the output of the Agglomerative Clustering shows the behaviour pattern of the students towards the game

Number of the Cluster (K)	Sample Size (No of the student to each cluster)
0	15
1	39
2	67
3	11

Table 8: Represents the results of the variable game

Number of the Cluster (K)	Sample Size (No of the student to each cluster)
0	30
1	39
2	32
3	11
4	11
5	5
6	4

Table 9: Represents the results of the loop games

Number of the Cluster (K)	Sample Size (No of the student to each cluster)
0	67
1	39
2	4
3	11
4	10

Table 10: Represents the results of the function games

5.3 Discussion

This research main objective is to implement and use machine learning clustering algorithm such as kmean and Agglomerative Hierarchical clustering to categorize the student into similarities group based on their response towards the game questionnaires more efficiently than traditional based approach where no quantitative approach had been used to analyse the student feedback and we have seen it during the literature review where we have highlighted the limitation of the methodology had been used to process the feedback. Clustering methodology such as kmean and Agglomerative Hierarchical clustering are used in this research after data pre-processing.

Game	Kmean clustering	Agglomerative clustering
Variable	6	4
Function	7	5
Loop	7	7

Table 11: Represents the number of the cluster for two methods

Kmean clustering algorithm has an advantage over Agglomerative clustering as we can identified the cut off of number of cluster to be created based on the clustering error whereas it is difficult the identify the cut off of number of suitable cluster in Agglomerative clustering and also in this research kmean is more suitable and efficient as sample size of the dataset is not too large. For each game bigger number of cluster to do categorization of the students is better than smaller number of cluster because it help us to understand the student behavior more in depth with capturing all important information and provide us more insights related to their pattern towards the game. In this research we recommendation bigger number of clusters so that education institutions can have into multiple groups of students which give more useful insight to understand their behavior and learning improvement more efficiently. The number of the cluster in Kmean has been selected through elbow method based on the cluster error before implement the clustering algorithm and results has been evaluate in terms of grouping of number of the student per clusters. In Agglomerative Hierarchical clustering number of the cluster to be used for the implementation is based on the cut off number which algorithm decides by itself. The results have been evaluated in terms of grouping of the number of the students per clusters. This approach helps us to get more insight and recommendation towards the grouping in the form of the cluster to understand the student's reaction towards the games and impact of the game towards there problem solving and critical skills. Table 11 represent the number of clusters for two methods.

Number of Cluster K	Sample Size (No of the student)
0	23
1	13
2	24
3	12
4	48
5	12

Agglomerative clustering table result for variable game:

Number of Cluster K	Sample Size (No of the student)
0	15
1	39
2	67
3	11

Table 12: Represents the number of the variable game comparison for two methods

Table 12 represent the Kmean has created 6 cluster to categorize the student based on their similarities. This will help us to understand the student behavior more clearly. Agglomerative clustering algorithm defined 4 cluster to categorize the students into 4 cluster, but this will restrict us to get all the information and not covered all minor information which is important to understand the student behaviors.

One limitation was the availability of limited feature such as missing of platform, UX and game design specific feedback data which makes research more area specific towards the student response game knowledge specific questions and other limitation was in Agglomerative Hierarchical clustering where number of the cluster to be used in the modelling is not as efficient as compared to the Kmean where we can decide the number of the cluster as per the cluster error and Kmean results are based on the multiple iteration of the results where as in Agglomerative Hierarchical clustering results are reproducible which makes Kmeans more efficient.

6 Conclusion and Future Work

In this research as discussed in the section 6 shows that the clustering models had been implemented on the game feedback data and shows the efficient way of the categorization of the students compared to the previous methodology has been used in the analysing the feedback dataset. The kmean and Agglomerative Hierarchical clustering has performed well in terms of the group the students based on their response which is acceptable.

In future scope of the research and analysis of student feedback towards there game based learning platform by including more questionnaires related to the feature and UX design and also increasing the same size of the feedback, this further improvement will definitely improve the analysis. the result is helpful for the educational institution to understand how to form the group the students which help us to understand their behaviours towards the game.

7 Acknowledgement

I would like to thank you my supervisor Dr.Cristina Muntean for her all support and guidance during my research work. she has provided me recommendation which was very useful to shape my research work and approach. I would also like to thank you National College of Ireland for providing me the opportunity for research.

I would also like thanks my family for all their support and encouragement during my research work.

References

Hussein, M., Ow, S., Cheong, L. and Thong, M., 2019. A Digital Game-Based Learning Method to Improve Students' Critical Thinking Skills in Elementary Science. *IEEE Access*, 7, pp.96309-96318.

Liu, E. and Chen, P., 2013. The Effect of Game-Based Learning on Students' Learning Performance in Science Learning – A Case of "Conveyance Go". *Procedia - Social and Behavioral Sciences*, 103, pp.1044-1051.

Zhao, D., Chis, A., Choudhary, N., Makri1, E., Muntean, G. and Muntean, C., 2019. IMPROVING LEARNING OUTCOME USING THE NEWTON LOOP GAME: A SERIOUS GAME TARGETING ITERATION IN JAVA PROGRAMMING COURSE.

Sharma, S., Choudhary, N., Zhao, D., Muntean, G. and Muntean, C., 2019. IMPROVING STUDENTS' LEARNING EXPERIENCE IN A PROGRAMMING MODULE WITH MULTIMEDIA ASSISTED MATERIALS AND PROBLEM-BASED LEARNING.

Makri, E., Choudhary, N. and Muntean, C., 2019. Game-based Learning in Computer Programming: Preliminary Results of a Large Scale Pilot Study Run at National College of Ireland.

Chi, H. and Jain, H., 2011. Teaching Computing to STEM Students via Visualization Tools. *Procedia Computer Science*, 4, pp.1937-1943.

Kazimoglu, C., Kiernan, M., Bacon, L. and MacKinnon, L., 2012. Learning Programming at the Computational Thinking Level via Digital Game-Play. *Procedia Computer Science*, 9, pp.522-531.

Ferrer-Mico, T., Prats-Fernàndez, M. and Redo-Sanchez, A., 2012. Impact of Scratch Programming on Students' Understanding of Their Own Learning Process. *Procedia - Social and Behavioral Sciences*, 46, pp.1219-1223.

Kazimoglu, C., Kiernan, M., Bacon, L. and Mackinnon, L., 2012. A Serious Game for Developing Computational Thinking and Learning Introductory Computer Programming. *Procedia - Social and Behavioral Sciences*, 47, pp.1991-1999.
Burgos, D., Nimwegen, C., Oostendorp, H. and Koper, R., 2007. GAME-BASED LEARNING AND THE ROLE OF FEEDBACK: A CASE STUDY. *Advanced Technology for Learning*, 4(4).

Wang, L., 2014. Performance Evaluation System of Software Enterprise Knowledge Management Based on Fuzzy Evaluation. *Applied Mechanics and Materials*, 687-691, pp.4990-4995.

Li, H., 2020. Application of K-means clustering algorithm in the analysis of college students' online entertainment consumption. *Journal of Physics: Conference Series*, 1570, p.012018.

Naeem, A., Rehman, M., Anjum, M. and Asif, M., 2019. Development of an Efficient Hierarchical Clustering Analysis using an Agglomerative Clustering Algorithm. *Current Science*, 117(6), p.1045.

Huizenga, J., Admiraal, W., Akkerman, S. and Dam, G., 2009. Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. *Journal of Computer Assisted Learning*, 25(4), pp.332-344.

NEWTON. 2020. *Publications*. [online] Available at: <http://www.newtonproject.eu/publications/> [Accessed 17 August 2020].