Educational Data Mining to Investigate the Impact of Students' Online Learning Activities on Their Assessment Marks

Y.T.Mathotaarachchi

2021

Educational Data Mining to Investigate the Impact of Students' Online Learning Activities on Their Assessment Marks

A dissertation submitted for the Degree of Master of Computer Science

Y.T.Mathotaarachchi

University of Colombo School of Computing

2021

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. We have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Student Name: Y.T.Mathotaarachchi

Registration Number: 2017/MCS/052

Index Number: 17440526

____

(27-11-2021)

Signature of the Student & Date

This is to certify that this thesis is based on the work of ~~Mr.~~ /Ms. Y.T.Mathotaarachchi under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Dr. (Mrs.) T.A. Weerasinghe

____

(27-11-2021)

Signature of the Supervisor & Date

I would like to dedicate this thesis to my loving parents.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my supervisor Dr. (Mrs.) T. A. Weerasinghe, senior lecturer of the University Of Colombo School Of Computing - UCSC, for her unceasing guidance, constant supervision and providing valuable resources throughout my final year individual project.

Also, I would like to convey my utmost gratitude to all the other academic members of the University Of Colombo School Of Computing - UCSC for the knowledge they passed throughout all two years.

Finally, I would like to thank my family for their valuable support and encouragement given endlessly to make this project successful.

ABSTRACT

Discovering pedagogically useful information included in databases acquired from Web-based educational systems is made easier with the assistance of different data mining algorithms. Researchers are trying to continuously evolve virtual learning environments by collecting and analysing data related to all the aspects of online learning. The purpose of this study was to investigate whether student involvement in an online learning activities on virtual learning environment is a good predictor of final course grades and to determine whether the below suggested clustering method can achieve comparable accuracy to conventional clustering algorithms.

This study aims to determine the effect of a variety of online learning activities on students' learning progress. . We focused more on applying this data mining technique for virtual learning environment the better insights of students. The findings indicate that engagement in online learning activities had the most significant influence on students' final grades of the course module. We demonstrate how K-Means clustering can be used to understand online learning activities in a virtual learning environment and how data mining techniques may be used to aid in the discovery of student-related and educator-related information included in databases derived from a virtual learning environment. We identified four clusters out of which we have focused more on Cluster 2 because the students falling into this category have scored the maximum marks and performed the most important and important online learning activities. These results may be utilised to assist educators in managing their classes, comprehending their students' learning, reflecting on their own teaching, and encouraging learner reflection and constructive feedback. This study begins with data collection, pre-processing, data mining methods application, and display of the findings. We have conducted a survey for the Web-based educational data mining tool in which we have included ten educators and ten students, tried to reach on constructive feedback about the Web based educational data mining tool by asking certain number of questions. The aggregative outcome for the survey of the Web-based educational data mining tool, educators seems to be a bit more unsatisfied as we reached an average score of 77.2 per cent. Teachers require some extra and better features in the web-based tool, while the students are quite more comfortable as the average score was 86.0 per cent. Students do not require too many changes to the tool. It is identified from the study that there is huge impact for the final assessment marks from the online learning activities the students engaged in.

Key Words: Educational Data mining, Virtual Learning Environment, Online learning activities, clustering, final grade, K- Means Clustering algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CSS - Cascade Style Sheet

CMS- Course Management Systems

DM- Data Mining

EDM - Educational Data Mining

E-Learning - Electronic Learning

HTML - Hypertext Mark-up Language

KDD - Knowledge Discovery in Data

KPI - Key Performance Indicators

LMS- Learning Management Systems

PHP- Hypertext Pre Processor

RDBMS - Relational Database Management System

SQL - Structured Query Language

VLE- Virtual Learning Environment

WBEDMT - Web-Based Educational Data Mining Tool

WEKA - Waikato Environment for Knowledge Analysis

CHAPTER 1
1 INTRODUCTION

1.1 Overview

Education is critical for a country's growth, particularly in developing nations such as Sri Lanka, India, Pakistan, Bangladesh, and Nepal. However, it is critical to identify the reasons behind students' inability to enhance their educational development, as well as any gaps in this area. Educators used to anticipate student success based on their own experiences; they had a good understanding of the student's personality and temperament (Mehboob, 2017).

As our generation is moving forward with time, the e-learning platforms are evolving rapidly. Now, the most wanted quality of e-learning frameworks is that they must be customised, since they must be utilised by a wide assortment of understudies with various abilities (Brusilovsky and Peylo, n.d.). Various different kinds of learners and students learn in an unexpected way. Some of them observe and learn the information brilliantly, while others ineffectively. Some percentage of students lean toward theoretical materials while others favour substantial models. E-learning frameworks ought to think about every student's acquiring inclinations and abilities. A new learning style has different model groups for which the learners indicate where they fit on various scales. Students have a place with the manners by which they get and measure their capabilities. The utilisation of learning styles for trial research in Web-based instructive frameworks has shown that giving material as per learners' moulded and most soothed learning styles can improve learners' learning (Peña and Marzo, n.d.).

Virtual Learning Environments (VLEs) are extensively utilised by instructors, who upload educational materials to VLE, allowing for the exchange of educational content as well as debate among students. VLEs also enable for the online storage of learning materials, as well as the linking of resources and the creation of exams. They also enable the instructor to create online conversations and track whether or not their students participate in them. VLEs additionally allow for the creation of blogs, the evaluation of instructional materials, the use of email, and other similar activities (Black et al., 2008), as well as studying outside of the classroom, providing flexibility.

Instructive or Educational Data Mining (EDM) is the application of Data Mining (DM) procedures to extract educational information. Data Mining is the procedure which is associated with separating important information, interpreting the extracted data, and extracting information from data (Fayyad, n.d.). The different applications of data mining have been utilised for a long time by different organisations, researchers, and governments to filter through volumes of information like aircraft traveller records, enumeration information, and general store scanner information that produces statistical surveying reports (Ashby and Broughan, 2002). EDM has started to gain attention as an examination region lately for analysts all around the world in various regions to break down enormous informational collections to determine instructive examination issues. The EDM has concerns about creating strategies to investigate the extraordinary sorts of information for instructive settings and, utilising these techniques, all the students and the settings in which they learn should be focused on personalising according to the learners. The certain rise in both instrumental instructional programming as well as state data sets of understudy data has created enormous stores of information reflecting how students can learn(Romero et al., 2008). The maximum utilisation of the Internet in training has created another setting known as e-learning or online instruction, in which a lot of data about educating learning cooperation is interminably created and universally accessible (Mostow and Beck, 2006). This data gives a gold mine of instructive information.

The integration of adaptive framework that allows for the delivery of customised learning is the current issue for web-based learning systems. These solutions address the issues that arise from the conventional "one-size-fits-all" strategy, which distributes the same static learning content to everyone regardless of domain knowledge, information requirements, or preferences, which may differ significantly. These sophisticated e-Learning systems provide high-quality material, efficient structure, and complete support for all the user profiles involved in a typical distance learning situation(Markellou et al., n.d.).

1.2 Motivation

During the last few years, e-learning has turned into the important segment of the worldwide learning cycle and slowly changed the entire education system. There are numerous equivalent words for the expression "e-learning" that are utilised in various researches. The examination available will adjust the terms personal computers or systems based learning innovation improved learning or "organised learning" (McGill and Hobbs, 2007). Thusly, the expression "e-learning" will be utilised that could be perceived as consolidating different ones referenced. E-Learning is an exceptionally expansive term for web based learning overall. The expression "E-Learning" has made exposure in the course of the most recent few years and infers commercialisation perspectives. The Internet has permitted schools, corporate colleges and revenue driven organisations to start offering degrees and chief instruction by means of the web. This methodology on account of that the organisations offering Personal Computer based preparing set up supposed e-learning arrangements (E.Khedr and I. El Seddawy, 2015).

The reason for our motivation was the weaknesses and lacks of the e-learning stages in the early years, there were loads of annoying issues which influencing the two understudies and instructors execution. Most importantly, absence of coordinated effort and correspondence offices made understudies feel forlorn and unsupported. What's more, their instructive inclination to teacher drove learning caused them face new difficulties understanding self-guided learning materials. These issues propelled e-learning divisions to utilise web conferencing and virtual joint effort apparatuses to fulfil the understudies' requests (Falakmasir and Habibi, n.d.).

This thesis presents the significant discoveries that came about because of examining the e-learning exercises and their effects on understudies' final marks.

1.3 Problem Statement

In e-learning frameworks, people had limited courses a few years back and the educational materials were not powerful, descriptive, and guided enough or they introduced uneasy flow. In this way, the platforms couldn't react adequately to the needs and capabilities of the students, which left students helpless and demotivated if they couldn’t solve their problems and failed to move forward in their work. Through hyperlinked course material, permits students to follow any navigational path they pick and not really utilise the construction dictated by web specialists or content makers (who have a specific navigational example at the top of any student’s individual priority list). This opportunity might demonstrate a preventing factor, since much of the time, students don't have the fundamental development and ability to follow a viable path, and it is a normal situation that they wonder about subjects that are either excessively troublesome, excessively simple, or only superfluous to singular adapting needs. Another question arises about whether the created model is well fitted for the situation of many and how our approach can impact the efficiency of any student or any instructor. This approach can be utilised in any other field, like sports, politics, and co-curricular activities. If yes, then how can someone define these parameters without wasting resources?

1.4 Aim and Objectives

1.4.1 Aim

Aim of this research is to implement a Web-Based Educational Data Mining Tool (WBEDMT) for institutes, educators, and students through which they will provide personalised details of students' interactions with virtual learning environments (VLEs). With the help of this tool, educators can also give instructions to their students regarding the interaction with VLE and help them directly to learn things efficiently. A student can have the freedom to choose their interests. Our research plans to discover the components that impact last assessments of students and their final evaluations by tracking their online learning activities on VLE.

1.4.2 Objectives

The objectives of the research reported on in this thesis are to,

- find out the most appropriate data mining technique that is the most compatible with the data set in our research. Since there are a lot of data mining techniques, we will do a literature review to find the best data mining technique. This technique should support to discover any positive relationship between the online learning activities of students on VLE and their final grade and to use that technique to analyse the data in our research.
- identify the (if there is such a positive relationship), online learning activities can make the strongest impact on the best final grade.
- enabling the educator, relevant to his/her course modules, through the proposed web based educational data mining tool, to detect the online learning activities that have been attempted and not attempted by a particular student. This further enables the educator to send a personalised message to a student if adequate engagement on the part of the student has not been made. The educator also is able to find the overall performance of students' online learning activities and do norm-referenced assessments on their performance.

1.5 Scope

In this thesis, we have conducted a thorough Literature Review before moving towards our own work to find a suitable data mining technique and what difference we can implement in our research. According to the previous, we have decided to use the k-means clustering algorithm as a data mining technique. We will be creating a suitable number of clusters (more descriptive in the Methodology chapter) using WEKA. The clusters will be based on the grades the students achieved in their final exams. Many iterations will be performed by us with different values of K to achieve the best results and the proper segregation of students on a grade basis.

After performing the clustering algorithm, we will be able to find the relationship between the online learning activities which have weightage for learning purposes on the VLE platform and the clusters of the students we have made. After finding and visualising the relationship in the tabular format, we will be segregating the online learning activities based on their importance into three classes: most important activities, important activities, and average important activities. For the ease of understanding for students, instructors, and universities, we have proposed web-based educational data mining tool specially made for visualisation of the student interaction on VLE.

In the Web based tool, we have different specific visualisations for students and educators separately to understand the ongoing learning of students and trace their progress. We have provided an additional feature for educators through which they can send personalised mail to students who are not continuously interacting with the online activities or losing interest while learning.

1.6 De-limitations

The thing that is out of scope in our study is that if we could try to predict the final grades of the students before they attempt the final exam, based on the online learning activities they perform on the VLE to learn their course modules, it could be very helpful to the educators and the students mainly.

1.7 Structure of the thesis

The chapters of this thesis consists of specific details with charts and diagrams in order to give an overview of the individual project. After went through a fully detailed description and understanding about the quest domain and scope of the study and motivation in the Chapter 01, the next chapters of this thesis consists with vital descriptions of the study much deeper as given below.

The second chapter covers a detailed discussion about the literature review of the domain referred during the study. It provides an understanding of the background of the study along with the technological developments regarding the topic, which helped in understanding the existing research gaps. In order to do that, we will be discussing the comprehensive summaries of similar work done in the past. Along with that, we will be writing about the difference between our work and the previous work which has been done.

The third chapter explains the methodology selected to complete and achieve the targets of the research. The chapter outlines the proposed model, which elaborates the suggested solution to the existing gap in terms of the key inputs to the model and proposed development stages.

Chapter four unfolds the different evaluation strategies that were used in terms of surveys to obtain feedback from the lecturers from different subject areas and undergraduate students from different subject areas who would be practically using the developed solution. The survey evaluates the overall outcome of the project via structured interviews.

The final chapter, Chapter five presents the key findings and the overall conclusion of the research. It also elaborates on the prospective areas of future research related to the study by addressing the limitations of this study.

CHAPTER 2

2 LITERATURE REVIEW

2.1 Introduction

As mentioned in Chapter 1, this study correlates with educational data mining. In order to quest a better approach to the study, it is a must to do a literature survey about recent research carried out under these correlated fields of study, which would be a great advantage when it comes to the implementation phase. A vital overview of the researches done related to this study will be discussed in this chapter. In order to find an approach to the clustering K-Means algorithm, we referred to recent related research articles.

2.2 Overview

Learning and training with the help of digital resources are termed e-learning. E-learning system is a formalized system of teaching by using the resources in electronics or digital form. The means of the internet and the computer consider as the fundamental component in the e-learning system and the process of teaching can be room-based or out. With time, the new generation is rapidly migrating to e-learning platforms because e-learning is a method for delivering lessons quickly. E-learning is widely considered the most effective method for ensuring effective learning by using computer and communications tools. For example, Guo et al. (2014) investigated how focused students were when watching the video (Guo et al., 2014). The time being spent watching the video and the number of times the student answered to assessments were used to determine the study's input characteristics. Students were far more motivated by shorter videos than by scheduled lectures, as per the survey (Sibanda, 2014). Since it improves the quality of teaching and learning, e-learning has now become a core of success in higher education (Bhuasiri et al., 2012). The usage of learning technology and learning motivation, as well as training material, has a favourable link. Data mining is termed as the process used to determine patterns and other useful information from massive sets of data, also known as Knowledge Discovery in Data (KDD). One of the most useful applications of educational data mining is predicting student performance. Educational data mining is "a fast-growing method that focuses on developing effective strategies that can discover detailed information in the educational environment and provides its methods to obtain a greater understanding of students' learning performance and make objectives for them," as per the definition on the Educational Data Mining Web forum (Falakmasir and Habibi, n.d.). To collect information on students, you'll need to create a data warehouse using their activity logs. This method allows for more

in-depth research and monitoring of students' academic performances and related patterns. In education institutions and virtual universities, the utilisation of the Warehouse Data and On-Line Analytical Processing (OLAP) tools has become popular day by day.
There exist many methods developed for educational data mining, and particularly there exist different algorithms for various problems related to educational problems and are elaborated briefly below.

2.3 Conceptual Background

The fast advancement in technology makes e-learning a part of education offered in many institutions and they are offering many courses in the world and can be attended by any person in any country. From certifications to diplomas undergraduate studies to post graduate degrees, seeking languages and much more in between is not so convenient so far. Even nowadays, some top-ranked institutions are offering courses that are a dream for the students to get certified from them but this adds some convenience of seeking the experience of ones tailored schedules. Then, e-learning courses are available with a flexible schedule that can suit the lifestyle of almost every person. Students are mostly adopting the e-learning system as an alternative to physically studying in degree awarding institutions and campuses specifically in COVID 19 pandemic like situations. Rapid innovations in technology enable people to get their entire education online together with their fellow mates, watching the recorded lectures and then contributing in specific course modules discussions from home has its advantages. Support from institutions is an important factor in the e-learning system and care should be taken regarding support that students are receiving the same content, in simulation, related to courses and other supports just like they are physically in institutions.
To make the e-learning process successful, instructors must know about the tools and the pedagogy which are considered as the requirements of students. Many strategies could be found beneficial in the e-learning environment that are elaborated below
The first strategy in an e-learning system is to know about the capabilities related to the technology of the students. If the students are not coming to the institutions and cannot be relied on the institution resources. Then one might take for granted about their internet connection reliability, using the technology of institution, Virtual Private Network access in some cases and other important tools for guaranteeing the students privy for studying in foreign countries. Therefore it is the need of time to facilitate students to accommodate in the institution that is living too far away have no such facility for learning. Like in institution students knows about the professor or instructor place in the e-learning system students should be able to know about the coursework of the instructor and how to communicate with them. Many institutions have

their own Learning Management System but they must have a connection with the links of the important content and the guiding pace for their assessments and homework.

Communicating with consistency is another strategy that improves the e-learning system. Creating the same environment that instructor communicate in the institutions with student determine the student's attention in the e-learning system and leads to get better results. The learning and engagement process matters a lot in the online education system as it does in the physical education system. So the creation of an asynchronous learning environment is the best option to overcome this hurdle by using the platform of different video communication like zooms, Google meets etc. having same time zone then it is possible to create meetings in an online space but for variant time zone one can then adopt the strategy of the of asynchronous form. LMS forum of discussion permits the responses of the students and their dialogue for a specific period and knowing that all the students are not able to attend the session same time. In projects related to the group, students can use different tools of leverage like Google drive, for asynchronous collaboration. Supporting the connection of students is another strategy that enhances the performance of the students. Eric Hudson has a point of view about the e-learning system that students must be aware of using social media for their self-expression, exploration and personal connections. He has a view about social media that it is just like research reminds that majority of the students are independent and move freely between online and in spaces of persons and use of intentional social media for the communication in E-learning system having feelings like classroom (Islam et al., 2015).

In e-learning for instructors assessments, is considers as one of the most common challenges. So an effective strategy for online understanding is the use of assessment in an asynchronous format. Precordial test in an e-learning system is not impressive. Organized and concise designing plans for the students with specific tenure of submission and the elaboration of the completed coursework. Getting feedback from the student is another best approach to find the shortcomings in the e-learning system and to overcome it. Like in the classroom teachers are constantly paying attention to their students same in online education it is necessary to design a feedback channel for the students. It is necessary to reflect the role of an instructor in the e-learning space. It is necessary and helpful for the students to have the connection for expressing autonomy, purpose sense, and keen interests in the work of mastery in the e-learning system.

One can expect better outcomes from e-learning channels when there exist multiples personalized opportunities for e-learners. E-learning facilitates students with opportunities to seek from different places. Identification of student support is another important factor that results in good outcomes from the students' results. This can be ensured by the instructor that students are engaged and showing up else identify the reason and support them if they needed.

Informing the parents of the student regular by sending emails by this strategy helps in managing the students and encourage them to e-learning like students have physically.
Another aspect of the LMS system data can be mined conveniently. Data concern to resources and the test data cannot be stored the same way of example (Merceron and Yacef, 2005). Manipulation of complex data is required to obtain the whole consolidated inefficient and useful form Statics present by LMS but they have limitations. So LMS should be improved with a particular module having an excellent function for data exploring. Further, the tool of the LMS system used for data mining must have good rules for the association with functions for choosing various attributes for driving the rules of associations.
Thus, there are user-friendly virtual learning environments readily made for students to achieve the aforesaid learning conditions. Some of them are Moodle, Blackboard, and so on. Another advantage is that, educators can customise these virtual learning environments to suit the needs of the relevant learners appropriately. For example, Moodle is a learning management system related to the e-learning that helps the educators for an effective online learning environment and considers an alternative to proprietary commercial. In modular design, it is convenient to create courses and for the engaging of learner one can add content. This is particular designed for the learning style called social constructionist pedagogy which was based on believing that students can learn well when they have interaction with course material (Anon., n.d.). For example, students connect to the lecture through Moodle and participate in discussion forums, do quizzes, activities and assignments or share knowledge and attitudes through message passing options. In addition, they can check for any updates on the relevant course, their grades in the assignments and quizzes.
When the students participate in such educational activities, educators, on their view are able to check the number of participants for the relevant assignments and to find out which students either participated or not in the discussion forums and then decide and direct the students towards doing more important but not often touched topics.
Thus, it is clear that such interactions between the educator and the students through the virtual learning environment, help students' in e-learning. At the same time help the relevant higher education institute in making appropriate connections among the students and between the educator and the students. This makes the learning teaching experience, although virtual, feel closer to a physical learning environment.

2.4 Theoretical Background

This review looked at how data mining can help with e-learning and education system development by adopting various methodologies and algorithms. This part appears to be a review of previous data mining-based e-learning research. The results of the research study will be described in this section using data mining techniques like classification and clustering, as well as algorithms and methodologies such as fuzzy logic, neural networks, decision trees and evolutionary algorithms. The majority of extant research focuses on various classification and cluster techniques. Data mining is the method of analysing patterns in huge datasets, which gives decision-makers a range of possibilities. The logical decision may be made to enhance the learning management system and assessment process by examining those patterns. The research majority's interest in applying data mining in e-learning is growing all the time.

According to E.Kheder (E.Khedr and I. El Seddawy, 2015), e-learning is regarded as the optimal method for facilitating and enhancing learning and based on communication and information technologies. He proposed that educational data mining is focused with the development of techniques for analysing the specific kinds of data generated by academic institutions. Also, This research pin-pointed mainly on the application of data mining methods to the Masters and Ph.d education sector by using the most widely used techniques on the most widely used application, the Moodle way of education. Additionally, (E.Khedr and I. El Seddawy, 2015) suggested that data compiled from history and used data can be stored in educational institutions' databases. The suggested method assists in obtaining adequate findings, which in our case study include many stages beginning with data collection, pre-processing, using data mining techniques, and visualising the results. E.Kheder discussed several e-learning systems, including a Learning Management System, a Course Management System, a Learning Content Management System, a Managed Learning Environment, a Learning Support System, and a Learning Platform. These systems can provide a plethora of channels and workspaces for educators to distribute information to students, create content, prepare assignments and tests, engage in discussions, manage distance classes, and enable collaborative learning through forums, chats, file storage areas, and news services.

(E.Khedr and I. El Seddawy, 2015) has focused upon the data mining techniques and explored two ways of it to fetch the data and they are Association Rule and clustering, Authors have used the multiple RDBMS tools to create the dataset for applying the predictive modelling. Some of the attributes analysed are different from our attributes like total_time_assignment, n_read and distinguished features of quizzes in itself. Our focus was to analyse the data and give feedback to lecturers along with the learners which will help students to self-evaluate themselves while the lecturers can send motivational messages to students those who didn't interact with the online activities on Virtual Learning Environment.

Merceron showed that data mining techniques may aid in identifying pedagogically relevant information stored in databases derived from Web-based educational systems (Merceron and Yacef, 2005) . The results may be utilised to assist instructors in organizing their classes, comprehending their students' learning, and reflecting on their own teaching, as well as to encourage learner reflection and constructive feedback. They have used different tools like SQL queries and Excel for data preparation, clementine and Tada-ed for clustering, classification, association rules along with the visualizations. Also they have used clustering techniques to discover homogenous groupings inside data. Merceron included K-Means clustering with hierarchical clustering. Both approaches are based on the notion of individual distance. The authors referred to euclidian distance. Classification is used to forecast the values of a variable. (Merceron and Yacef, 2005) utilised the TADA-Ed C4.5 decision tree, which is based on the notion of entropy. A series of rules may be used to express the tree, such as: if $x=v_1$ and $y>v_2$, then $t=v_3$. Thus, one can anticipate an individual's value for t based on the values it assigns to variables such as x and y. The tree is constructed using a consumer involvement and is used to forecast new individual values. The authors focused their attention on individuals who tried but did not complete an activity satisfactorily. They clustered this subgroup using (i) TADA-k-means Ed's algorithm and (ii) Clementine's mix of k-means and hierarchical clustering. Because there is no predetermined quantity or set of activities against which students can be compared, establishing a gap between people was not apparent. The authors computed and used a new variable: the cumulative number of errors committed by each student throughout an activity.

Merceron began with data mining. SODAS partitions the population in sets referred as symbolic objects. The number of attempted exercises defined the symbolic objects, which were characterised by the values assigned to the obtained variables: the amount of actually completed question, the average amount of right steps for any attempted question, and the average number of errors per attempted exercise. They compiled a variety of tables to facilitate the comparison of all of these items. But in our study, we have more focused on the attributes which were collected by Online Learning platform like number of clicks on the online learning activities. Merceron have used tools like MSSQL and Excel for RDBMS instead we have used MYSQL and on the other hand k-means clustering is being used as data mining technique. Author's research was more of educator centric based while our study is educators' and student both centric.

Lopez Presented a classification method based on clustering to predict final grades in a university course using forum data (López et al., n.d.). Their goal was twofold: to evaluate if

student involvement in the course forum is a good, predictor of final course grades and to test whether the suggested classification through clustering method can achieve the same level of accuracy as conventional classification algorithms. The authors compared the suggested method to conventional different classifiers in predicting if student passed or failed a course based on their Moodle forum use statistics. The findings demonstrate that when just a subset of chosen characteristics is used, the Expectation-Maximisation (EM) clustering method produces results comparable to those of the top classification systems. The centroid of an EM groups are detailed in order to demonstrate the connection between various clusters and the two student classes. According to Lopez et al had a binary objective of predicting whether a student would pass or fail based on the data collected by Moodle forums statistically? Although Lopez et al have created two clusters by using multiple algorithms as mentioned above, instead we have used only one clustering algorithm, which is k-means, and we made 4 clusters. Lopez et al have applied multiple classification algorithms to attain the best accuracy and their other goal was to understand how well the students are at understanding information available in the Moodle Forum and what the important attributes are present in Forum.
Janesmanoharan have shown that predicting student's performance is a difficult job. They have utilised cluster model to investigate the outcomes of the students, and statistical methods have been used to categorise the student's grades according to their performance (Jamesmanoharan et al., 2014). However, the author believes it is less successful, therefore they have included the k-mean clustering method in conjunction with a deterministic model in order to evaluate and monitor the student's outcomes and overall performance in the course. It is possible to gain more efficiency in the monitoring of the development of academic performance of students in higher education institutions by using k-mean clustering, which allows us to give correct findings in a shorter period of time. He used the technique to figure out the different intriguing patterns by obtaining the student test results and presenting them in their paper. It was discovered by Jamesmanoharan et al that their model was based on an applied data set that was derived from one semester's academic results, which was the test score of Computer Science students. They have applied k-means clustering for different K values as 3, 4, and 5. After taking into account all the different clusters the results were analysed and put into a table.

Jamesmanoharan et al has segregated students having a major in computer science by using k-means clustering and trying to understand what percentage of students fall into which category while dividing them on the basis of their percentages attained in the exam. For example, out of a certain number of students, what percentage of students score above 80% and what percentage of students score between the ranges of 60 – 80% and so on. The number of intervals was directly decided by the number of clusters ‘k’ given. The students who are scoring less in their online activities have been tracked to understand why they have been underperforming and how we can increase their marks. Instead, in our project, we can track each and every student’s activities and try to analyse if those students are performing the most important activities or not, and how we can encourage both students to perform the most important activities.

Learning management system is an effective education method and have very little adaptivity. Felder-Silverman Learning Style Model (FSLSM) is a technique adopted by Sabine Graf and Kinshuk for the deduction data mining system in which Fielder and Silverman proposed the learner learning style in the suggested model and create specific four dimension differences like verbal-visual, global-sequential, reflective-active and intuitive-sensing (Graf and Kinshuk, 2006). Every learner has a specific choice among these dimensions and the other important thing in FSLSM is the tendency problem that suggests sometimes a specific behaviour seems to acts differently having a high preference.

Gartner makes a comparison of various classification methods is made to determine which method is the best for classifying the students depends on their final grade, whereas other researchers utilize neural network models to predict student grades (Rotondo and Quilligan, 2020). The above-mentioned classifiers are extremely valuable since they extract the data mining related to education that has linked with the topics of education and intelligently presents the result to the instructor. The available information, research goals, and desired results all determine the Data mining technique and its technically feasible algorithm. According to Khedr, the learning management system of the database, contains a huge amount of important data that may be used to enhance the e-learning approach. Focus on the fact that learning management systems collect a massive amount of data that is extremely beneficial for studying students' activity and might turn educational data into a gold mine (Khedr, n.d.). The experts underline the large amounts of data that can be generated daily; it is very difficult to evaluate this data manually, and a very promising approach to this analytical goal is the employment of Data mining techniques. Data mining methods can be used to find many different types of knowledge. This technique demands not only the selection of appropriate thresholds for the two traditional metrics of support and confidence but also the consideration of suitable procedures of reliability to keep relevant rules while filtering out those that are interesting (Merceron and Yacef, n.d.) . The results of data mining can be used to improve the teaching method. In Tada-Ed, they included a feature that allows teachers to discover patterns and integrate them into the ITS, where the information is stored. This feature is currently only accessible in the Association Rule module. That is, any association rule can be extracted by the

teacher. After that, the rules are recorded in an XML file and sent into the ITS's educational module.
Cristóbal Romero adopt a model of meta classifier based on the classification of a cluster and assume that every cluster corresponds to the specific class (Romero and Ventura, 2007). First of the useful and interactive data have been collected and then processed for selecting a specific group they applied one more optional attribute. The next step was the execution of cluster data. For predicting the unseen labelling of class mapping is used in test data instantly. Class attributes were not or clustering but for evaluation to get data as the classifier. For all the algorithms of the cluster, they make it necessary that the number of data set in the class label should be equal to the number of clusters produced yielding a better model for relating every class with the label.
To execute the case study, they utilized a variety of "Feature Selection" and "Attribute Evaluation" techniques, as well as real student usage data from the Moodle learning management system. The research found that attendance in classroom virtual sessions has the highest influence on learning effectiveness in the IUST e-learning centre's specific contexts. As a result, the instructors and managers were encouraged to pay significant attention to digital learning and particularly motivate the students for participation. According to A. Mercerone and K Yacef, the association rules are highly effective in data mining related to education because they extract associations between educational objects and provide the results in an intelligible form to the teachers (Francis and Raftery, n.d.). Association rules, according to the authors, need smaller amount of data mining knowledge than other methods. Francis and Raftery also suggested three digital learning models based on an improved education basic course management and assisting learners; hybrid learning improves the learning and the teaching process significantly; and the first two modes of guide personalization through many online modules and classes (Keržič et al., 2019) . Depends on the above three stages, it is suggested that the university's staff of academic management be adequately informed of each stage's tactics, structure, and support to improve higher education and learning environments. Keric, Tomaevic, Aristovnik, and Umek investigated the important components of collaborative learning for higher education students and found that e-learning was positively regarded when the teacher was participating in an e-course and that students' attitudes about the subject had a significant effect (Keržič et al., 2019).

Mehboob assessed the reasons for student failure based on prior data and predicted the probability of failure for the next course so that students may psychologically prepare for the given course as well as the course's dependence level (Mehboob, 2017). It is standard practice in engineering that if a student does not understand the fundamentals, he or she will struggle in advanced courses with similar scopes. With the assistance of a decision tree, the author was able to track down the source of the failure. Their study also aided in estimating risk in the early stages, which may aid instructors in developing effective plans for at-risk children. The 6 algorithms were utilized for prediction and risk analysis by (Mehboob, 2017). In comparison to the others, ID3 produced the best outcomes. They utilized the CEME and NUST data sets in their study. A total of four hundred and fifty records were retrieved from five degrees in their collection.

Mehboob have focused mainly on the reasons why students were failing in certain subjects, and they tried to analyse and improve the problems by doing. But instead we have worked upon creating a platform to improve study efficiency. Mehboob have tried to predict the risk of failing in the next course. They have tried to create a platform which will warn students to choose any vulnerable subject in which they might fail in the future. They have used 6 algorithms for risk (predictive) analysis, whereas we have used only one algorithm. Mehboob objective was to count the number of students who failed in their subjects while studying at any university by analysing their past behaviour towards the subject.

To overcome all these problems we research the implementation of Web-Based Educational Data Mining Tool (WBEDMT) for educational institutes so it can provide the facility of detailed personalisation while interacting with online learning activities on virtual learning environment. For data mining, we used the technique of K- Means clustering and performed multiple by taking different values of K to get the most improved result. Then we relate the cluster of a student with their online learning activities on VLE. Finally, with the tool of the web, we traced the students' individual interaction of online learning activities.

2.5 Summary

After looking at above-mentioned studies we can see that many approaches can be used to educational data mining. Among them, a similar approach that was used in educational data mining is clustering algorithms. The main difference between our work and other's work was providing web-based tool to interpret results from analysis, and educators can motivate their students those who are not interacting with online learning activities. The next chapter explains the steps followed in identifying and applying the best methods to address this study.

CHAPTER 3
3 METHODOLOGY

3.1 Introduction

Educational technologists and training administrators have no way of verifying how important each activity is to a student's performance. As previously stated, the goal of this study is to rate online learning activities according to their effect on students' final grades. As a result, certain factors have been designated as student key performance indicators (KPIs). The effect of each variable was then assessed based on its impact on student performance on final examinations. Data mining methods were used to examine the online learning activity logs of the virtual learning environment in order to derive certain rules regarding the significance of each activity in students' success.

In this chapter, we have explained the whole process that we will be performing in this research, from the dataset collection to the implementation of different algorithms on the dataset. The steps involved in the cleaning of dataset and the preparation of dataset have been explained in this chapter along with the explanation of the tools used in this research. We have explained the raw data transformation to the model implementation ready data. The numbers of clusters formed on this dataset after implementing the decided algorithm K- Means clustering, and then we have analysed the relationship between the clusters and the attributes we had on this platform.

3.2 Theoretical Framework

Currently, there is no method for pedagogical scientists or management teams for training to quantify the importance of individual activities to students' performance. Our study will rate online learning activities according to their effect on learners' final grades. Then, the effect of each variable such as the sum of clicks, time spent, etc., was determined by its impact on students' final test scores. Data mining methods were used to examine the web use statistics of the virtual learning environment to deduce certain principles regarding the relative relevance of each activity to students' success: Pre-processing, clustering, and visualisation of data.

3.3 Data Mining

Data mining is the method of examining data from various angles and distilling it into actionable knowledge (E.Khedr and I. El Seddawy, 2015). DM methods are the outcome of a lengthy research and development process (E.Khedr and I. El Seddawy, 2015). In other terms, it is the act of rapidly identifying non-obvious yet useful patterns within a huge collection of data (Romero et al., 2008).

The data mining stages in a virtual learning environment

a. Data extracting
b. Data preparing
c. Data transforming
d. Data mining applications i.e. K-Means Clustering
e. Analysing results.

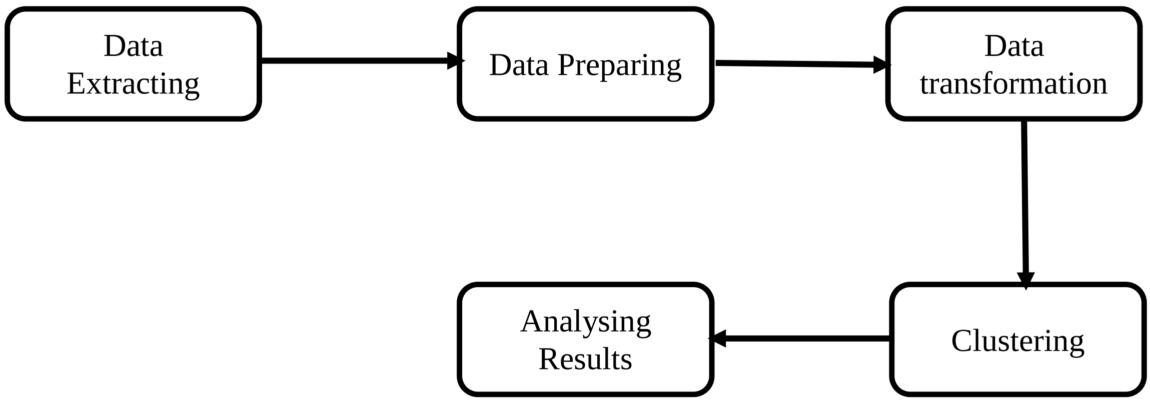

Figure 1: Data Mining Steps

3.3.1 Data Extraction

Extracting data is the most fundamental and critical phase of data mining. This task sometimes becomes difficult, particularly when the data has been fetched from databases such as MYSQL. The dataset for this work was retrieved from a UCI machine learning repository (Open University Learning Analytics dataset) (Anon., 2015) that contained numerous tables relating to students and the virtual learning environment. The website has fetched all tables in a single database. There were a varying number of tables in the database accessible via the website. Assessments, courses, student assessments, student information, student registration, student VLE, and VLE are listed in the tables (Romero and Ventura, 2010). After obtaining all of these tables, all pertinent information was gathered and aggregated into a single table that can be easily used for data mining using the tool, WEKA. All data pertaining to students' assessments, grades, and relationships with the VLE have been fetched and stored in a central location for further data mining steps.

3.3.2 Data Preparation

This step of data mining in a virtual learning environment involves aggregating multiple datasets into a single table and then applying some useful pre-processing filters such as replacing missing values in WEKA according to the dataset's requirements. After applying various filters, the dataset has been properly prepared and is ready to be transferred and loaded into the WEKA workspace for clustering. A flat table was created from the previously mentioned tables and then subjected to data mining techniques using the software. Simply connected to VLE data, was chosen to demonstrate the eccentricities of learner's exercises. Certain number of activities would go unnoticed due to their extremely uncommon application.

3.3.3 Data Transformation

After performing data preparation, in which all tables are figured out and aggregated into a single table. Then the single table can be fully utilised in data mining applications, a new transformed single table is created and fetched. This transformed table (Table 1) was created for the purpose of performing data mining operations.

Table 1: Sample dataset from aggregated table

codemodule	codepresentation	Idstudent	sum_click	idassessment	score	activitytype	finalresult
AAA	2013J	28400	4	1752	78	resource	Pass
AAA	2013J	28400	1	1752	70	oucontent	Pass
AAA	2013J	28400	1	1752	72	resource	Withdrawn
AAA	2013J	28400	11	1752	69	url	Pass
AAA	2013J	28400	1	1752	79	resource	Pass
AAA	2013J	28400	8	1752	70	homepage	Pass
AAA	2013J	28400	2	1752	72	url	Pass
AAA	2013J	28400	15	1752	72	oucontent	Pass
AAA	2013J	28400	17	1752	71	resource	Pass

This Table 1 is basically the finalised aggregated and cleaned sample table, which has been extracted from different tables in the dataset from the website (Anon., 2015). The data given in this table is as it was from the beginning. But this data first has been pre-processed using filtering techniques like replacing missing values, checking correlation, reducing hindrance, and then this table has been transformed into the workspace of the WEKA tool to perform clustering.

3.3.4 Clustering

Clustering is the process of separating a population or set of data points into several sets divided by the number in the same group that are more similar to each other and different to data points in other groups (Anon., n.d.). It is essentially a grouping of things based on their similarity and dissimilarity (Anon., n.d.).
In this thesis, the K-Mean clustering method was to evaluate the learning behaviour of the students. Students, tests, grades, final results, and activities are examples of things that may be utilised in remote learning. Data clustering was performed via the use of WEKA data mining software and quantifiable inspection software, respectively. K-Means clustering method, which is the most commonly used clustering algorithm, was utilised in this study. This clustering organises data according to the recipe. (Fayyad, n.d.).

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

Figure 2: K-means Clustering Algorithm

3.4 Tools and software used in research

In this research, we have used WEKA (Waikato Environment for Knowledge Analysis) software for data cleaning and clustering analysis. WEKA is a free application written in the Java programming language. WEKA is a collection of visualisation tools and algorithms for large-scale data processing, all accessible through graphical user interfaces (Menaka and Kesavaraj, 2019). It has a library of machine learning and data mining techniques for pre-processing, classification, regression, clustering, association rules, and visualisation of data. WEKA was used for cleaning the data as well. We have replaced all the null values with the average of the whole column. We have dropped the rows with special characters and wrong entries.

MS Excel spreadsheet software has also been included in my work to increase the visualisation power. This application provides a huge number of charts that anyone can include in their work after extracting their cleaned data in the form of CSV or xlsx format from the above (WEKA) software.

MYSQL was used to merge all the data, which was initially in the different tables. We have performed all the relational database management functions through this software. The server used for my work was XAMMP (Anon., n.d.) (To run the proposed Web-based educational data mining tool in local host environment- which is explained in the below section). It is an open-source server that was used for data storage, and the data retrieval was done with the help of MYSQL.

3.5 Dataset

We have a complete relational database management system (RDBMS) which can be processed with the help of any SQL variants like MYSQL, ORACLE SQL and MSSQL, Although We have used MYSQL for our study. The database consists of a total of 7 tables which are different in shapes, namely student registration, student assessment, StudentVle, StudentInfo, courses, assessments, Vle. All the mentioned tables provide information about the student information, the learning course modules and the online learning activity details. All these tables contain 41 columns in total. All these columns differ in shape as well. Each attribute (feature) has been analysed with the help of the MYSQL and EXCEL, and the essential columns have been extracted to use for modelling purposes (Kuzilek et al., 2017). The RDBMS platform solved the row shape mismatch and merged using different and appropriate joins like inner join, outer join, left join, and right join. The shape of the final data we achieved has 12 columns and 173913 rows, namely code module, code presentation, student id, site id, and date, the sum of clicks, assessment id, date submitted, is_banked, score, activity type and final result.

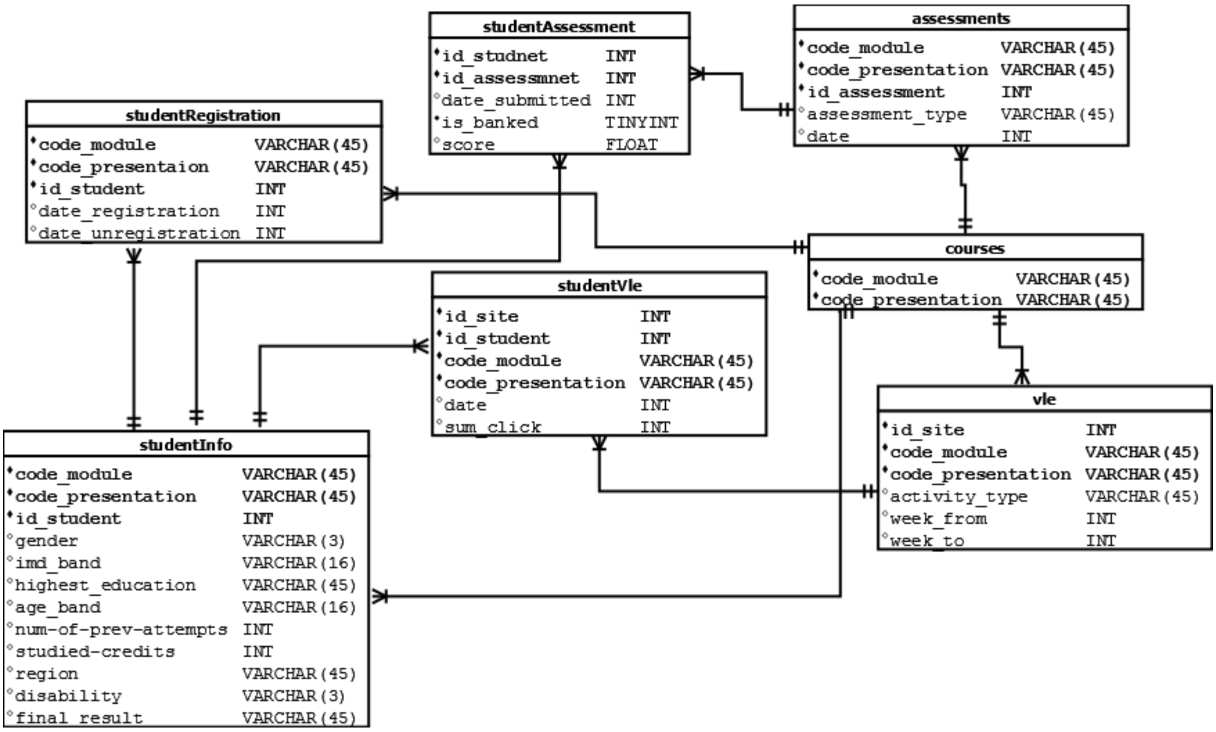

Figure 3: Table structure of Dataset

- Codemodule - The code name of the module, which acts as the module's identification, is shown below.
- codepresentation - This explains the presentation's code name. It consists of the year and the letter "B" for February presentations and "J" for October presentations.
- Idsite - a unique identifier for the material.
- Idstudent - a unique identifier for the student.
- Date - It contains information on the assessment's final submission date, which is computed as the number of days after the module's presentation began. The presentation's start date is 0 (zero).
- sum_click - It indicates how often a student engages with the content throughout that day.
- idassessment - It contains the assessment's identifying number.
- datesubmitted - It indicates the date of student submission, which is expressed as the number of days since the module presentation began.
- isbanked - It is a status indicator that indicates that the assessment result has been copied from a prior presentation.
- Score - This column indicates the student's performance on this examination. The value range is 0 to 100. A score of less than 40 is considered a failure. The scores range from 0 to 100.

- activitytype - This identifies the function of the module's content.
- finalresult - It serves as the final product of the student's module presentation.

3.6 Data Analysis

Clustered instances and cluster classes, i.e. activities, have been fetched for further data analysis. Clustering results are shown in Tables 2 and Figure 4.

Table 2: Model and evaluation on training set

Clustered Instances		
0	295011	28%
1	148941	14%
2	350298	33%
3	254325	24%

Cluster instances were obtained from a summary of the model's results after developing a clustering model in WEKA on the dataset. This indicates which clusters contain more and which clusters contain less information. On the far left, there is a number of clusters ranging from 0 to 3, which indicates that there are four clusters, and then cluster instances are listed in both percentage and exact values (Fayyad, n.d.). As a result of this, it has been determined that cluster number two contains more data than the other clusters.

As shown in Table 2, Cluster 2, which is ranked third, is the most valuable and weighted cluster in the entire model.

```
Class attribute: activitytype
Classes to Clusters:

      0      1      2      3  <-- assigned to cluster
 295011 148941 346594 254325 | resource
      0      0    996      0 | oucontent
      0      0    886      0 | url
      0      0     22      0 | homepage
      0      0   1055      0 | subpage
      0      0     21      0 | glossary
      0      0    194      0 | forumng
      0      0     82      0 | oucollaborate
      0      0     28      0 | dataplus
      0      0    127      0 | quiz
      0      0     21      0 | ouelluminate
      0      0      3      0 | sharedsubpage
      0      0     61      0 | questionnaire
      0      0    102      0 | page
      0      0     26      0 | externalquiz
      0      0     49      0 | ouwiki
      0      0     20      0 | dualpane
      0      0      5      0 | repeatactivity
      0      0      2      0 | folder
      0      0      4      0 | htmlactivity
```

Figure 4: cluster relationship with online activities

Figure 4 shows clusters with numbers ranging from 0 to 3. Each cluster acts as a column head in this instance, and each cluster contains all cluster instances associated with each student online learning activity. As indicated on the right, all student online learning activities have been included. Thus, the goal is to have all cluster instances based on each student's online learning activity in order to obtain the most important, important, and average important online learning activities based on their cluster instances or weight which will be effecting the final grades in the student's exams.

3.7 Web-Based Educational Data Mining Tool (WBEDMT)

The purpose of this tool is to interpret the results of the analysis. We have provided the option through which this platform can be accessed and used by educators as well as by students. Through this tool, educators can check and monitor the interaction of the students' online learning activities in the Virtual Learning Environment. The educators are provided with two options. One is the individual interaction of the students who are continuously following and studying the course. The other is the overall interaction, where the educators can have an entire statistical comparison of all the students' activities through different visualisations. The WBEDMT was implemented with the help of the following technologies

- Design: Hypertext Mark-up Language (HTML 5), Cascade Style Sheet (CSS3), Bootstrap
- Implementation Programing Language: Hypertext Pre Processor (PHP7)
- Relational Database Management System: MYSQL (Structured Query Language)
- Data Visualization: Google Charts.

The input of the system is the final table made from the combination of different tables, as explained above that we prepared for analysis purposes for our research.

The output of the tool is the set of charts that have been created by the tool which show the correlation of the student's final grade to the attributes.

3.7.1 High-Level Diagram of the WBEDMT

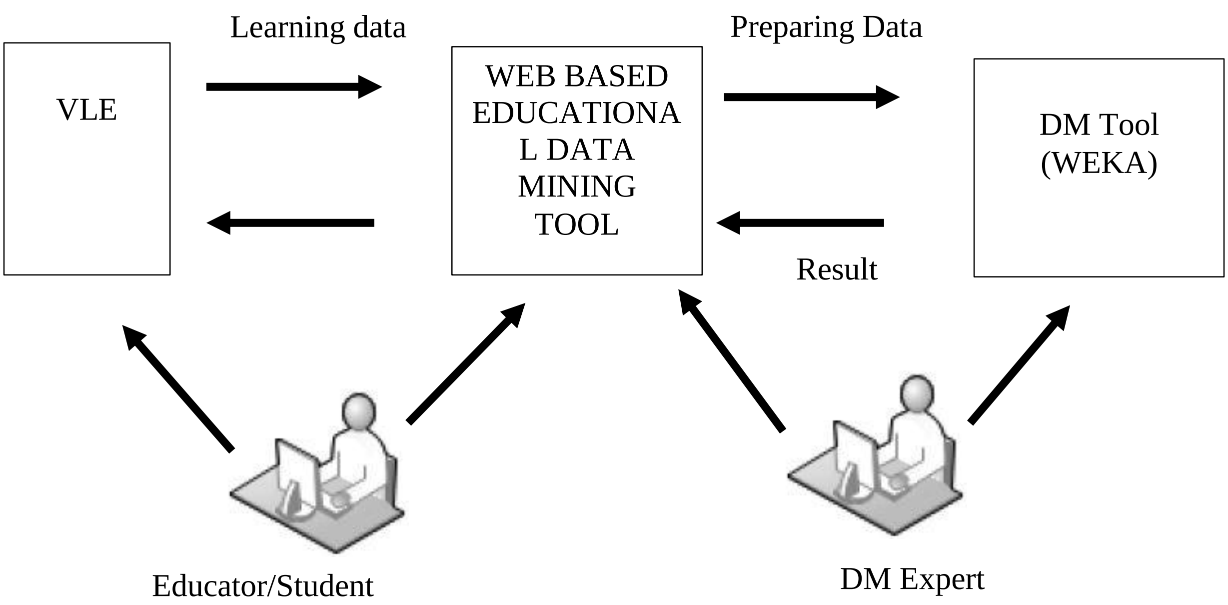

Figure 5: High-level diagram of the WBEDMT

3.7.2 Use Case diagram for Educator View

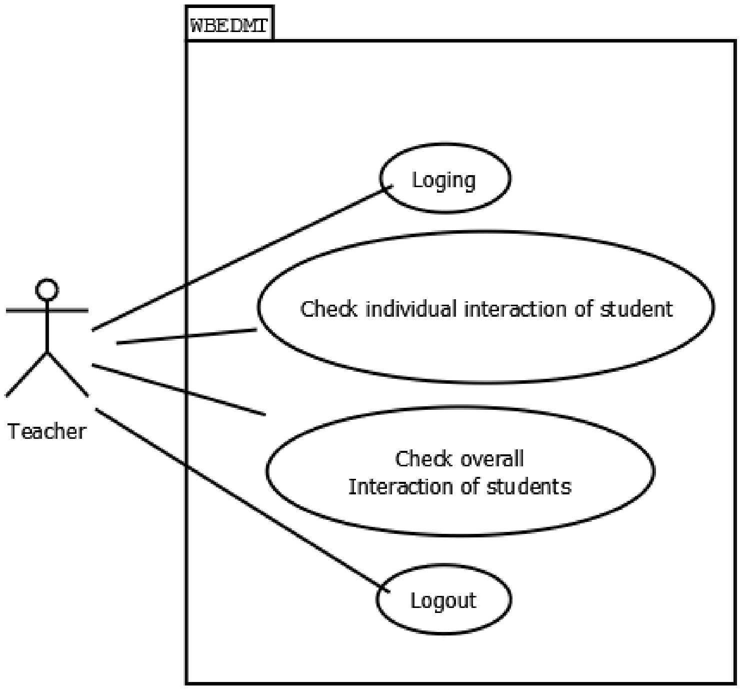

Figure 6: Use case diagram for teacher view

3.7.3 Use Case diagram for Student View

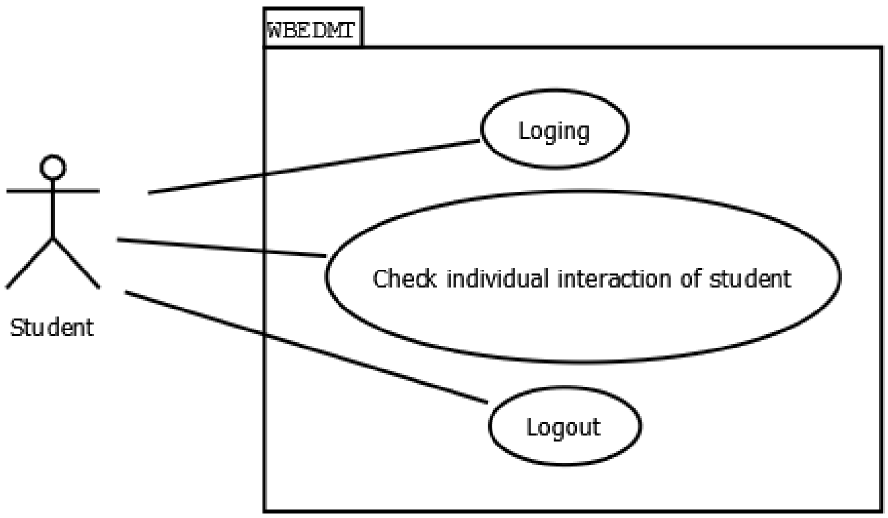

Figure 7: Use case diagram for student view

3.7.4 User Interface for WBEDMT

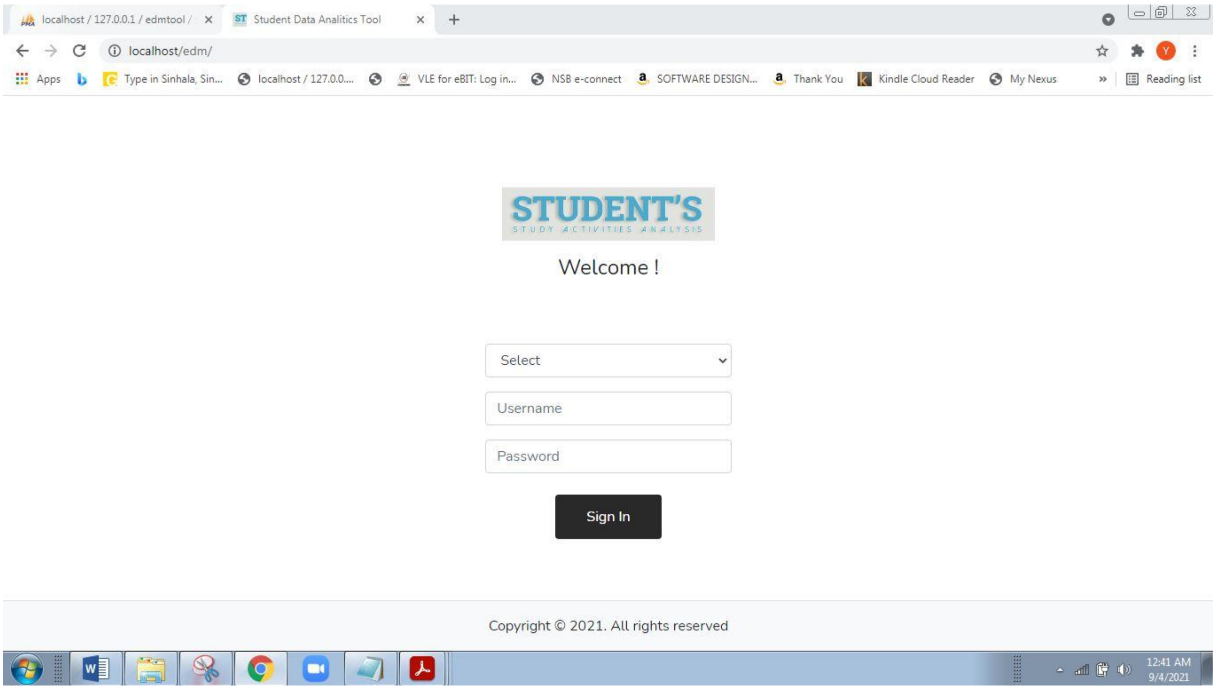

Figure 8: Login page of WBEDMT

3.7.5 Database for WBEDMT

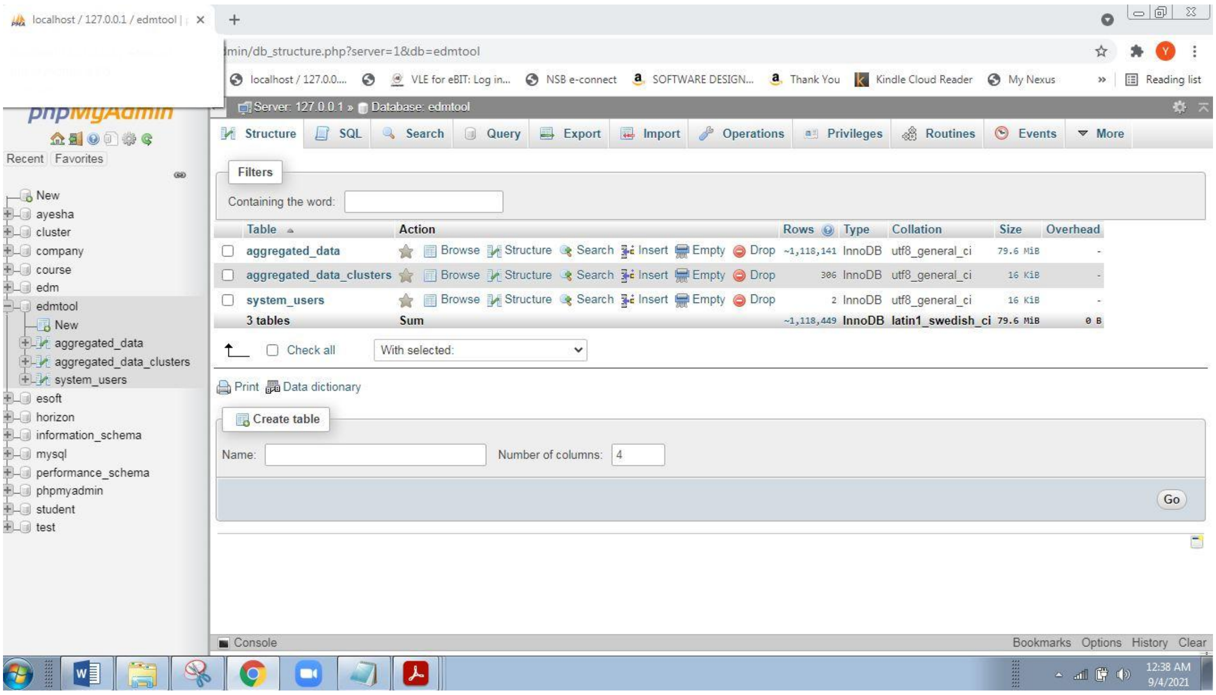

Figure 9: Database for WBEDMT

3.8 Summary

The overwhelming majority of the world's biggest schools and institutions, as well as in other nations, give WEKA priority. The majority of the students and actions are stored and fetched from MYSQL data sets that are accessible to framework administrators via the WEKA application. Data mining techniques were utilised to analyse the virtual learning environment's web use data in order to infer specific rules about the relative importance of each activity to students' performance. The dataset for this study was obtained from a UCI machine learning repository (Open University Learning Analytics dataset) that included many tables including information on students and virtual learning environments. All data related to students' tests, grades and interactions with the VLE have been collected and stored in a central place in preparation for future data mining processes. WEKA data mining software was used to cluster the data. To evaluate people learning behaviour, the K-Mean clustering method was employed. In next chapter explains the results of the analysis and evaluation of identified results.

CHAPTER 4

4 EVALUATION AND RESULTS

4.1 Introduction

To explain the results and illustrate them, we have discussed the K-Means clustering algorithm that will be applied to the collected datasets through the use of online activity logs on VLE. We have discussed how we collected the data, cleaned and prepared the data using different tools. In this research project, all analysis were performed using the tool WEKA, database was constructed using MYSQL, and visualisations and result evaluation were done using MS Excel. Here we will discuss the insights we have gained after performing the Exploratory Data Analysis (EDA) on the collected dataset. In the first section, the plots of the clustering and the insights will be drawn and plotted against the online activities. Then, moving forward, we have segregated the Cluster 2 students' activities into three categories according to their importance (discussed below) of the online activities on VLE. We have given a detailed explanation in the paragraphs of Cluster 0, Cluster 1 and Cluster 3. These three clusters did not carry the important information which we have explained below in detail.

In the 2nd part, we conducted a survey which tell us the performance of the web based educational data mining tool (WBEDMT) which we have created and gave two specific sections for academics (lecturers, demonstrators, instructors) and students and provided different features and functionalities. There was a set of five selected fixed formatted ranking questions which were asked by ten academics, and a set of four selected fixed formatted ranking questions were asked from ten students. An average was taken to come up with proper feedback for the web-based educational data mining tool (WBEDMT).

4.2 Results

If we have a look at Table 2 and observe, we can clearly state that Cluster 2 has more weight than the rest of the clusters. This certainly implies that Cluster 2 is the most essential cluster, and it contains more information about students than the rest of the clusters or Clusters 0, 1 and 3. So, Cluster 2 is considered the most important cluster in the entire model.

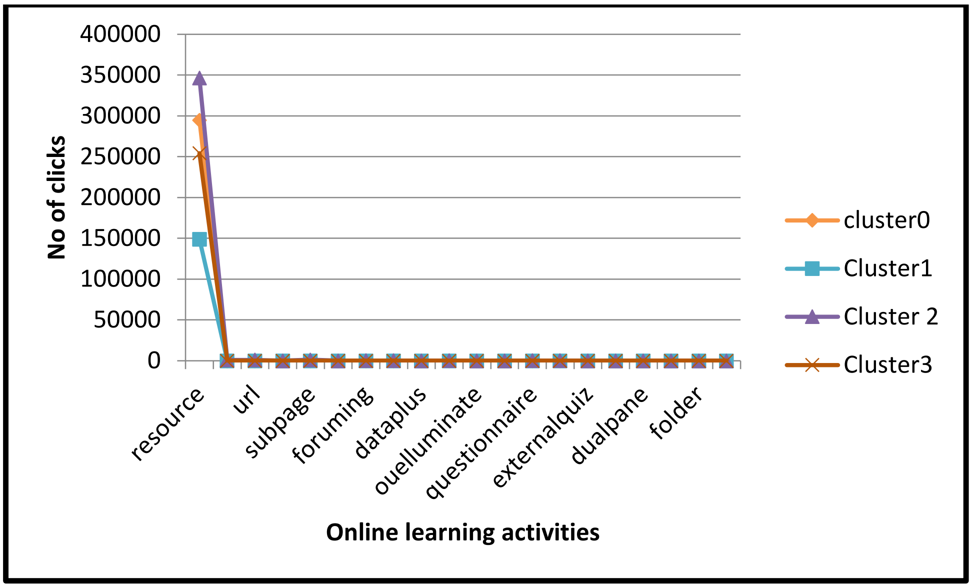

Figure 10: K-Means clustering of students 'online learning activities using the K-means algorithm

As shown in Figure 4, every single student from Cluster 2 is the most active learner, devoting the bulk of their time and effort to the activities on the Virtual Learning Environment. Incredibly, when compared to other groups, their final evaluations are the most minimum, despite the fact that they are not easily distinguishable. In order to determine the causes behind this, we must do a more in-depth analysis of the material from Cluster 2. Table 3 illustrates the results of categorising VLE online learning activities according to their impact on the type of research, as determined by the data collected.

Table 3: The influence of VLE online learning activities on the quality of studies

Most Important	Important	Average
• Resource • subpage • Oucontent • Url	• Foruming • Quiz • Page • Questionnaire	• Oucollaborate • Homepage • Glossary • Dataplus • Ouelluminate • Sharedsubpage • Externalquiz • Ouwiki • Dualpane • Repeatactivity • Older • Htmlactivity

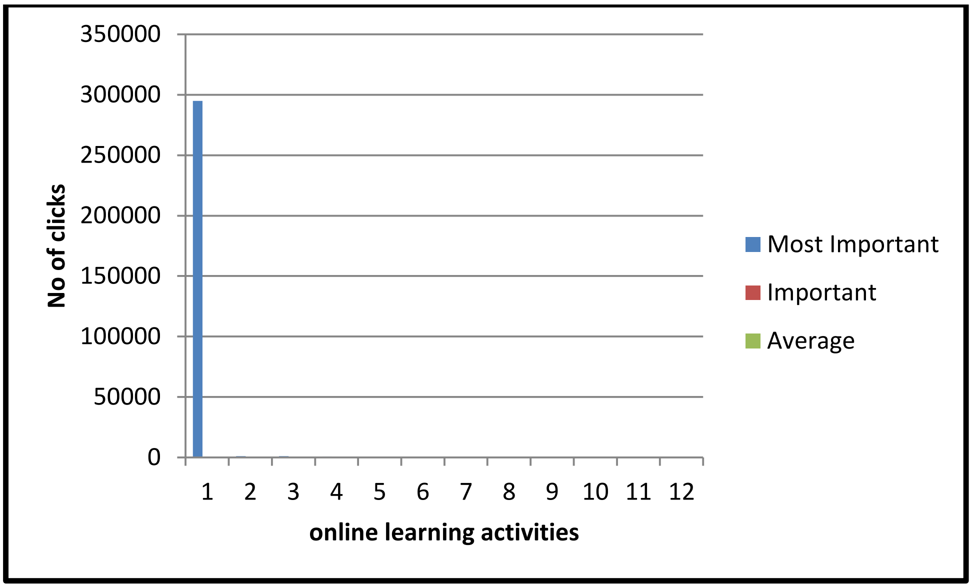

Figure 11: Analysis of students' online learning activities from Cluster 2 according to the effectiveness of their activities

Figure 11 shows clearly that all of the students' time spent on the VLE is aggregated into Cluster 2. From Figure 11, it can be seen that the vast majority of students from the same cluster have spent a substantial part of their time using important VLE platforms, such as reviewing materials, completing assessments, and so on. After that, students checked their own achievements, looked at who else was online and sitting on their system and performing learning activities, participating in conversations, and so on. During this period, there was no point to anything. As a result, it is possible to conclude that the high degree of movement associated with using VLE is not a guarantee of good report performance. Because the intentional, sensible use of significant apparatuses is a significant model for course engineers and educators as well, a course teacher who has information about this rule can upgrade the course route, to increase the amount of time students spend on the VLE, and redirect students to work in a purposeful manner, among other things. All things being equal, in certain instances, they use all of the available gadgets without first evaluating their potential for use and the overall rationale for doing so. The most popular student activity in Cluster 2 is the resources. Individuals from this cluster are the most latent VLE visitors, yet the average of their final evaluation is the most notable of the variables to consider. Once again, in order to properly examine this material, we must do a factual analysis of the data. Following the previous model, Figure 9 depicts students' activities as represented by various VLE items that are partitioned into three groups Most important, Important and average important, as in the previous model. Clearly, the students in this category make use of both important and trivial VLE devices, as shown in the chart. A few understudies have spent more than 70–80% of their study time on the major VLE gadgets. Only a small portion of their time was devoted to important and average important sections. It is possible to assert that the understudies in the second group make use of study resources that provide exact predetermined information. They didn't waste their time looking about in every direction. Unexpectedly, they were able to hunt down the necessary inspection supplies and equipment and put them to good use. As a result, we draw another conclusion: the students from this cluster do not make extensive use of the VLE's tools (conversations, talks, and so forth). They like researching different situations.

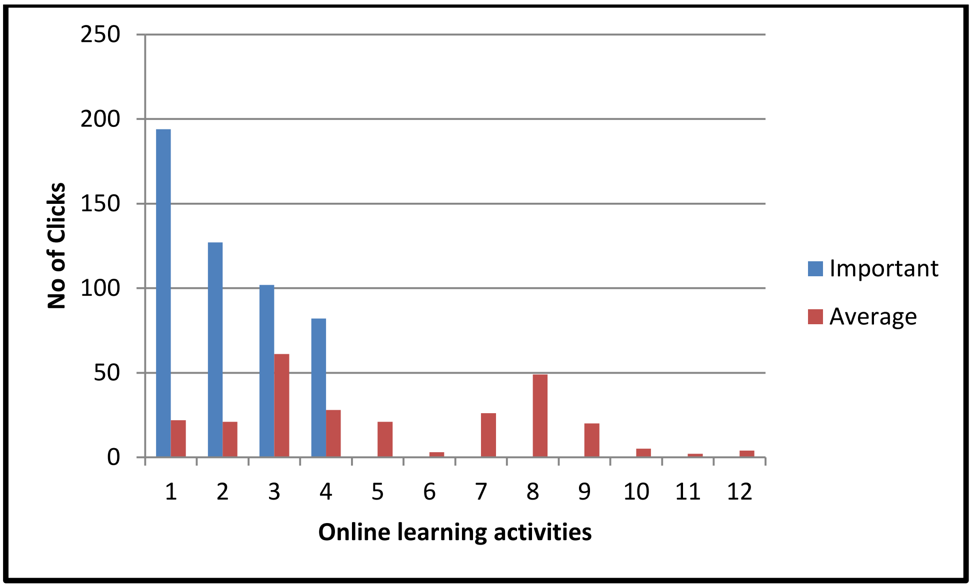

Figure 12: Analysis of students' important and average activities from Cluster 2 according to the effectiveness of their activities

It is now abundantly clear from Table 3 that only activity 'Resources' participates and dominates in Clusters 0, 1, and 3. On the other hand, it has also been noted from Table 3 that if the goal is to obtain the most important, important, and average important activities from clusters, then only Cluster 2 should be considered. For the reason that, in Cluster 2, all activities are included in a variety of cluster instances, a variety of activities can be organised into the most important, the important, and the average important activities on the basis of these cluster instances, which is one of the primary objectives of our research.

Another point that can be seen in Figure 10 is that the number of significant activities, particularly in the area of "resources," is so large that the rest of the activities are completely ignored. As a result, in order to take such activities into consideration as well, Figure 12 has been created, which contrasts important and average important activities.

After looking at the most important activities and their effects on the students who scored the maximum in their final grades, Cluster 2, we will be analysing the important and average important activities separately because we couldn't fix the index of important and average important activities with the most important activities due to the high difference in measuring values. As we can clearly see in Figure 12, there are fewer clicks on average important activities, and the performance of important activities is better than compared to the average important activities. The interpretation of the above graph was already expected, and the results are very convincing, which gives us proof that our research is moving in the right direction. The performance of important activities is much less compared to the most important, but the performance of average important activities is nearly negligible despite having a higher number of activities.

The activities of students from Cluster 0 and Cluster 3 represent 28 per cent and 24 per cent of the total students, respectively, and are very similar to one another. They lag behind the Cluster 2 by a small number of students, but they are slightly more active than the students from Cluster 1. Students in Cluster 1 are less active users of key online learning activities on VLE, whereas students in the Cluster 2 pay somewhat greater attention to less important tools. Despite the fact that the representatives of Clusters 0 and 3 are prospective learners, their abilities are not being used. Try as we may, we can get them closer to Cluster 2 students if we can get them back on track and urge them to participate in the necessary and more weighted activities. The students are able to concentrate on the most important course elements, and they are interested in the new online learning activities. The reason for this may be that the educators are not handling or encouraging the students, or it may be something else entirely, but they lose the opportunity to achieve better results in the course. If we identify such students and use active teaching techniques to help them, we may expect them to do better in the future.

4.3 Evaluation of Results

The web based educational data mining tool (WBEDMT) has helped us to understand different kinds of aspects of online learning. We have analysed almost all the related activities that can impact any student's grades and categorised those activities into three categories as discussed above. But as of now, we have not discussed the performance of our web based educational data mining tool, The academics as well as the students were granted permission to visit the WBEDMT and analyse their progress over time. Now we will be discussing the performance of our tool. We have selected ten different students, all from different streams and different graduation levels. So we must be able to cover the 360 degree view and be able to collect feedback from every corner. Similarly, while performing the survey among the educators, we asked questions of lecturers, demonstrators and instructors who belong to different streams. We have asked around five questions from ten educators and four questions from different ten students. At the end of the survey, we created a metric in excel and calculated the overall score both student and educator wise separately. In the questions we gave a statement which is a feature or the function of the web based educational data mining tool and we asked educators and students to give the feature a ranking ranging from 0 to 5.

Table 4: Web based educational data mining tool evaluation criteria

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

4.4 WBEDMT evaluation survey with educators

This survey consisted of five questions containing five features or functionality and was asked to the educators to rank it. All selected ten educators were from different streams. The formula for calculating the overall score is $\sum \frac{TotalScore}{Total\ teachers}$.

The first question asked was '**How much are you satisfied with the user interfaces of the web-based educational data mining tool?**' The overall score for this question was 3.9.

Table 5: Response for Question 01 - educator view

Teacher ID	q1
T001	5
T002	5
T003	4
T004	3
T005	4
T006	4
T007	5
T008	3
T009	4
T010	2
Average Satisfaction	3.9

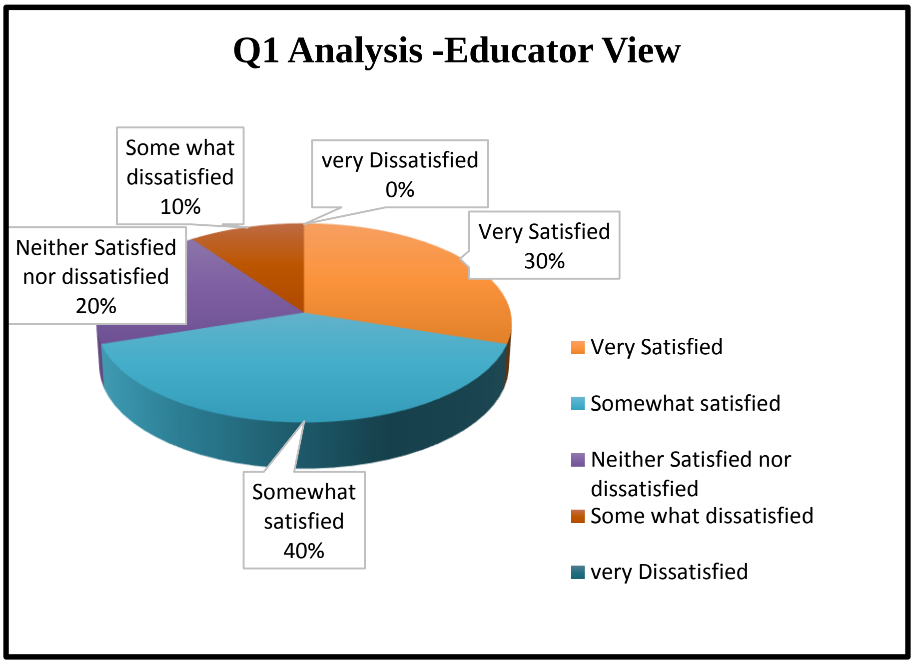

Figure 13: Question 01- educator view

The second question was, '**How much are you satisfied with the individual student's activity interaction feature of the web-based educational data mining system?**' The overall score achieved was 4.2.

Table 6: Response for Question 02- educator view

Teacher ID	q2
T001	5
T002	5
T003	3
T004	4
T005	5
T006	5
T007	4
T008	2
T009	4
T010	5
Average Satisfaction	4.2

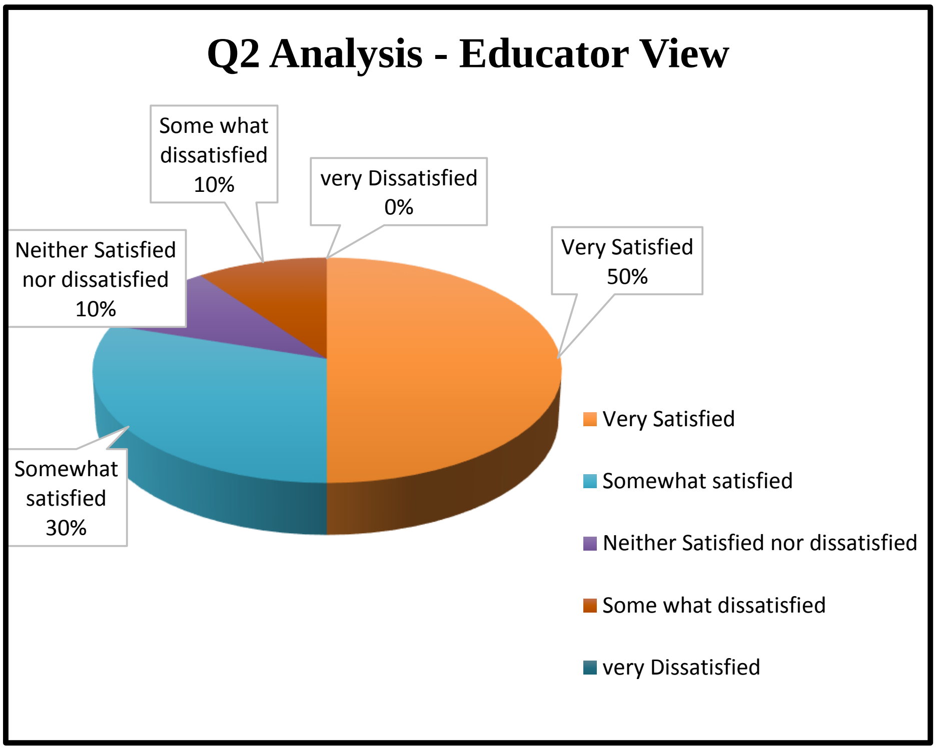

Figure 14: Question 02- educator view

The third question was '**how much are you satisfied with the overall module interaction feature of the web-based educational data mining system?**'. The overall score achieved was 3.9.

Table 7: Response for Question 03- educator view

Teacher ID	q3
T001	5
T002	5
T003	5
T004	3
T005	5
T006	4
T007	2
T008	1
T009	4
T010	5
Average Satisfaction	3.9

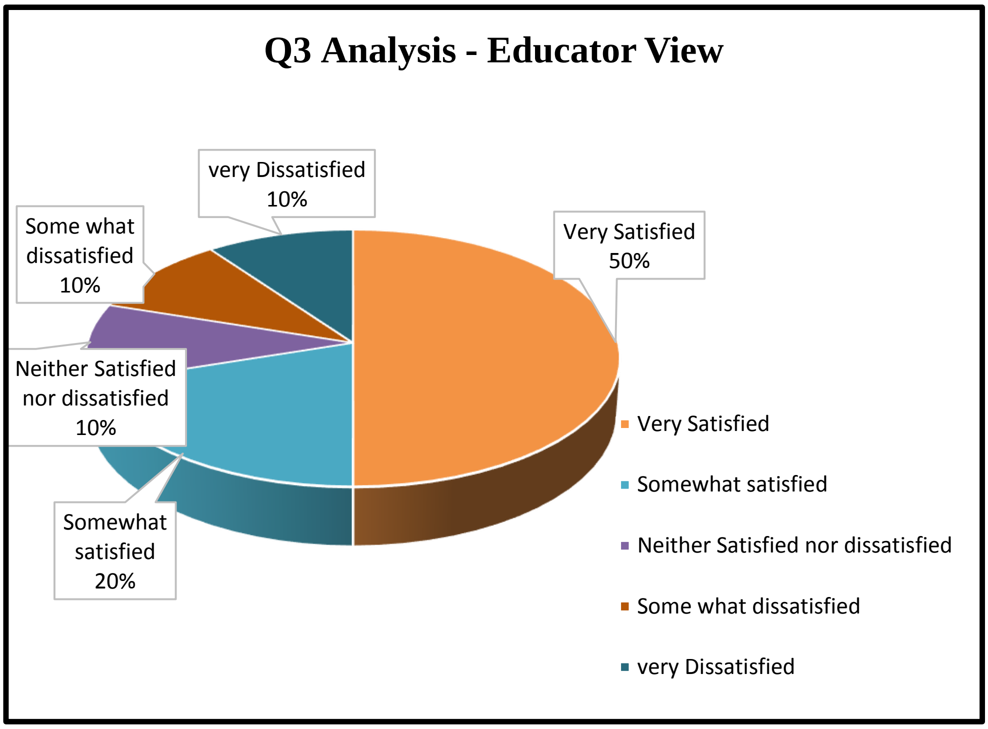

Figure 15: Question 03- educator view

Fourth question '**How much are you satisfied with the message sending functionality of the web-based educational data mining tool?**' The overall score achieved was 3.6.

Table 8: Response for Question 03- educator view

Teacher ID	q4
T001	5
T002	3
T003	4
T004	1
T005	5
T006	3
T007	5
T008	4
T009	2
T010	4
Average Satisfaction	3.6

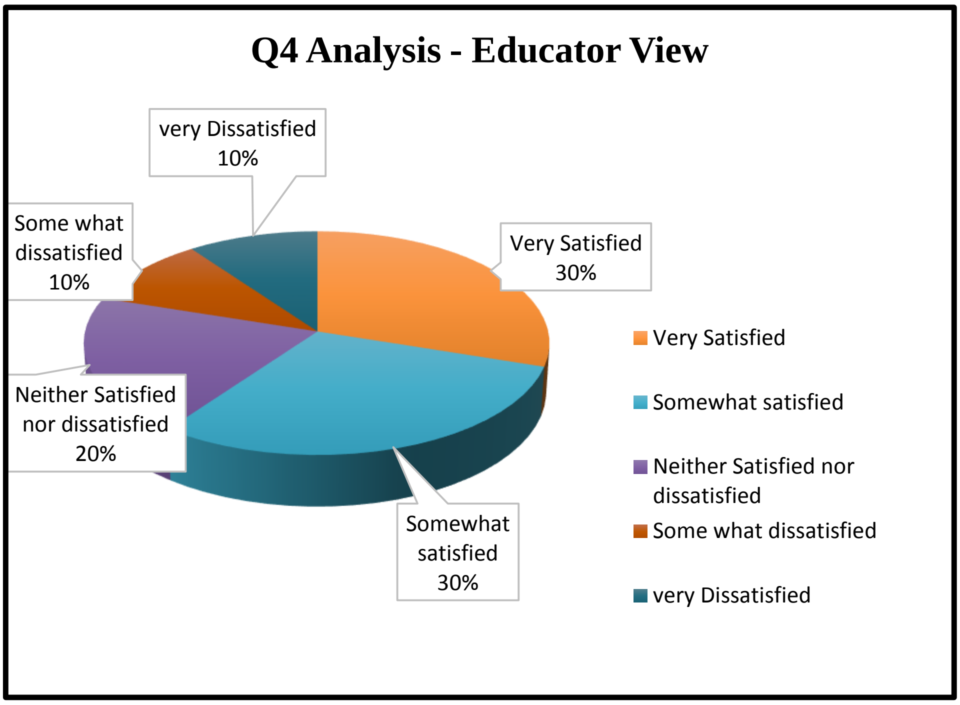

Figure 16: Question 04- educator view

The last question was '**How much are you satisfied with the activity categorisation according to the importance of the activities (as**

most important, important, and averagely important)?' The total overall score achieved was 3.7.

Table 9: Response for Question 05- educator view

Teacher ID	q5
T001	5
T002	5
T003	4
T004	3
T005	5
T006	4
T007	3
T008	2
T009	1
T010	5
Average Satisfaction	3.7

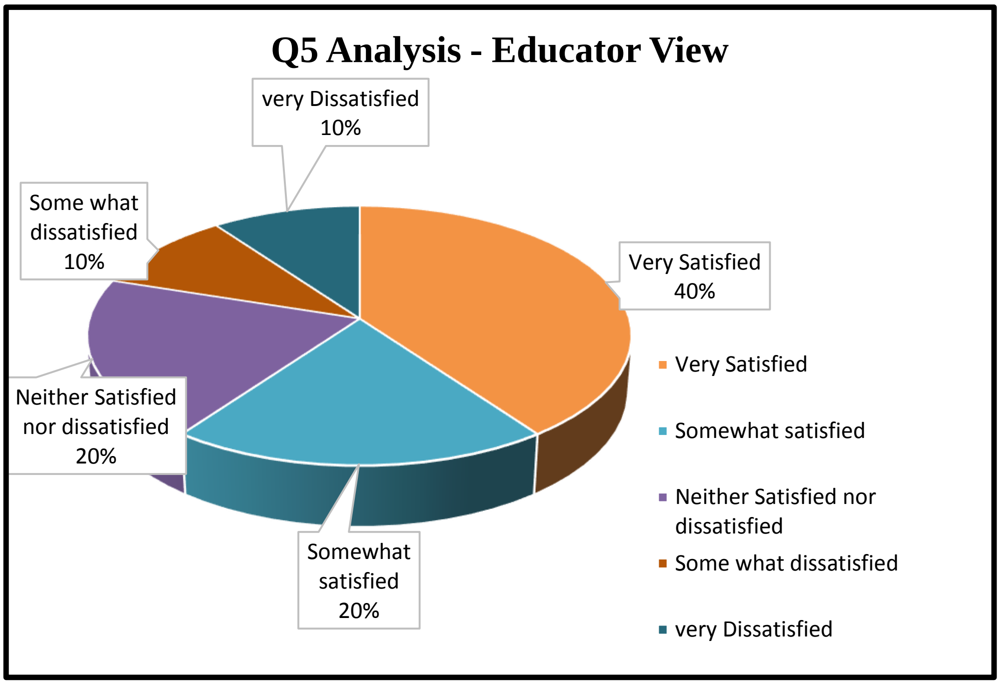

Figure 17: Question 05- educator view

After aggregating the scores of all the questions, we calculated the overall average performance score as 77.2 percent. This score implies that we need to work on the web-based tool to make it look better than the current version.

Table 10: Overall average performance - Educator View

Teacher ID	q1	q2	q3	q4	q5	Total marks	Overall Satisfaction (%)
T001	5	5	5	5	5	25	100.00
T002	5	5	5	3	5	23	92.00
T003	4	3	5	4	4	20	80.00
T004	3	4	3	1	3	14	56.00
T005	4	5	5	5	5	24	96.00
T006	4	5	4	3	4	20	80.00
T007	5	4	2	5	3	19	76.00
T008	3	2	1	4	2	12	48.00
T009	4	4	4	2	1	15	60.00
T010	2	5	5	4	5	21	84.00
Overall average performance							77.20

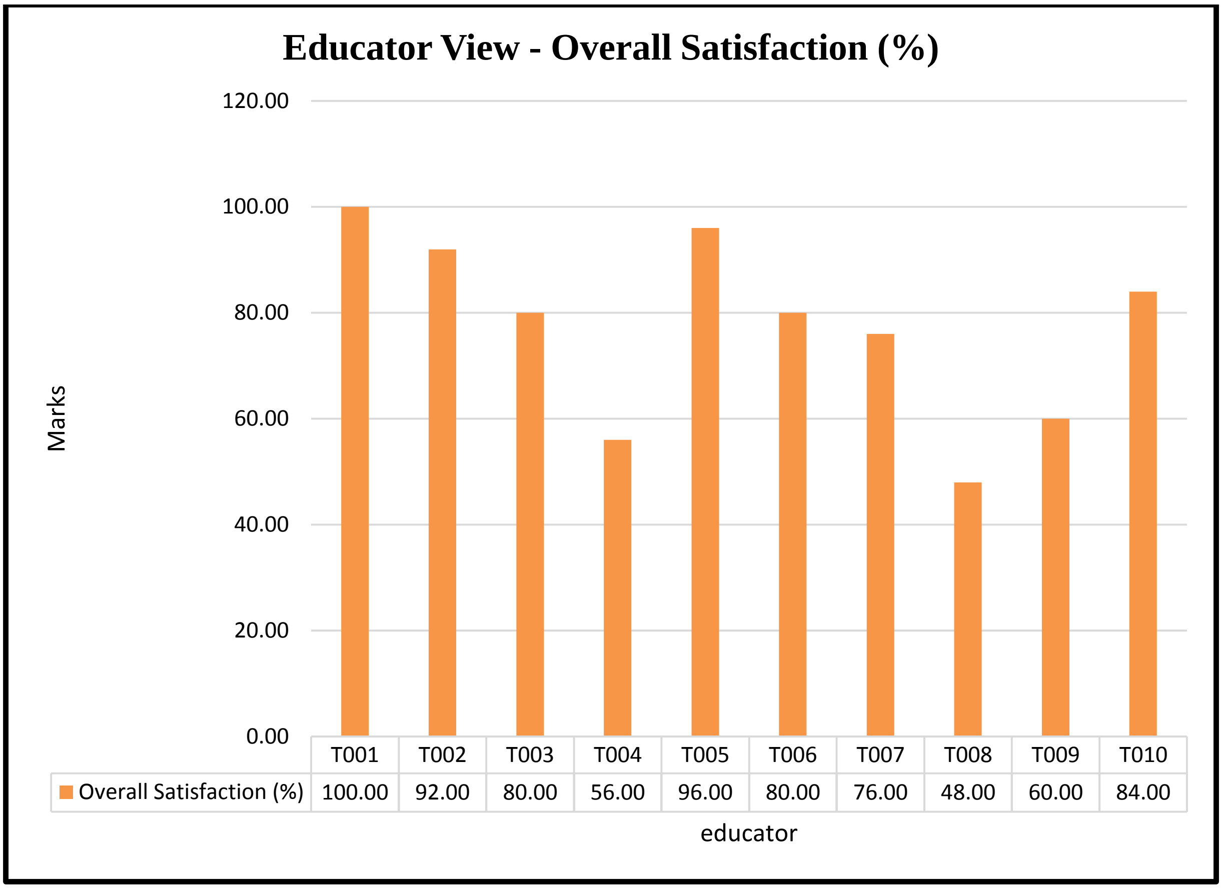

Figure 18: Overall average performance - Educator View

4.5 WBEDMT evaluation survey with students

The second part of the survey consists of four questions containing four features or functionality included in the web-based educational data mining tool focusing on students in the student section. The total number of students chosen was ten. Before conducting the survey, we made sure all the students were from different streams, and no biasing was done. The formula for the calculation of the overall score is $\sum \frac{TotalScore}{Totalstudents}$.

The first question is, '**How much are you satisfied with the user interfaces of the web-based educational data mining tool**?' The overall score achieved for this question is 4.3.

Table 11: Response for Question 01-student view

Student ID	q1
S001	5
S002	4
S003	4
S004	5
S005	4
S006	4
S007	3
S008	5
S009	4
S010	5
Average satisfaction	4.3

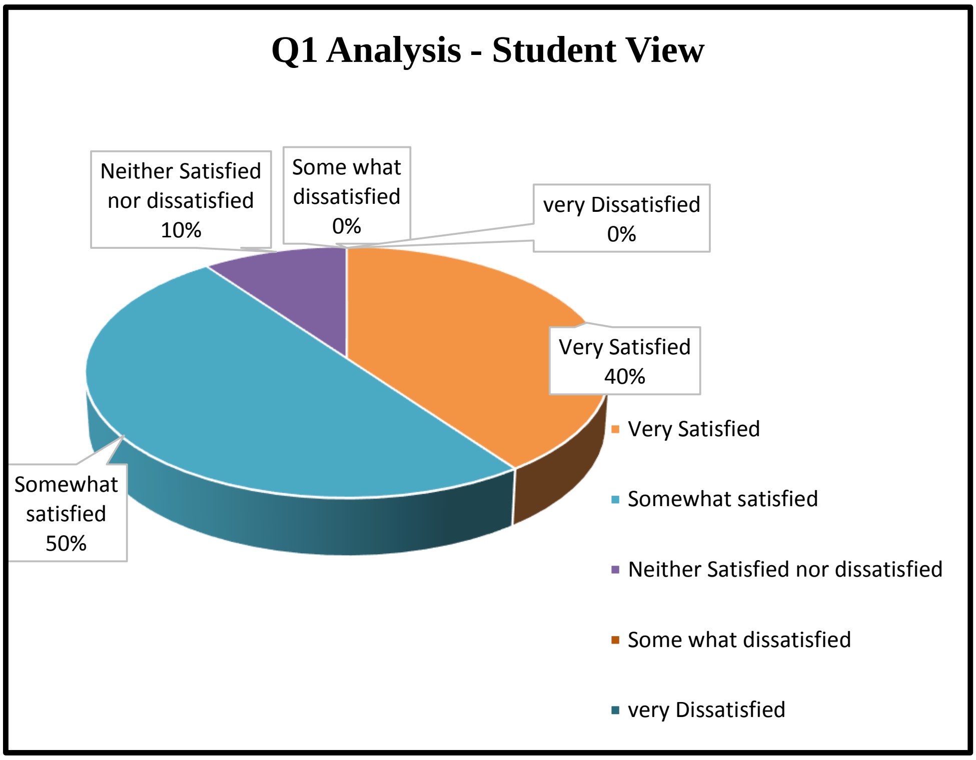

Figure 19: Question 01- student view

The second question is '**How much are you satisfied with the individual student's activity interaction feature of the web-based educational data mining system?**' The overall score achieved is 4.4.

Table 12: Response for Question 02-student view

Student ID	q2
S001	5
S002	4
S003	4
S004	5
S005	4
S006	4
S007	5
S008	4
S009	5
S010	4
Average satisfaction	4.4

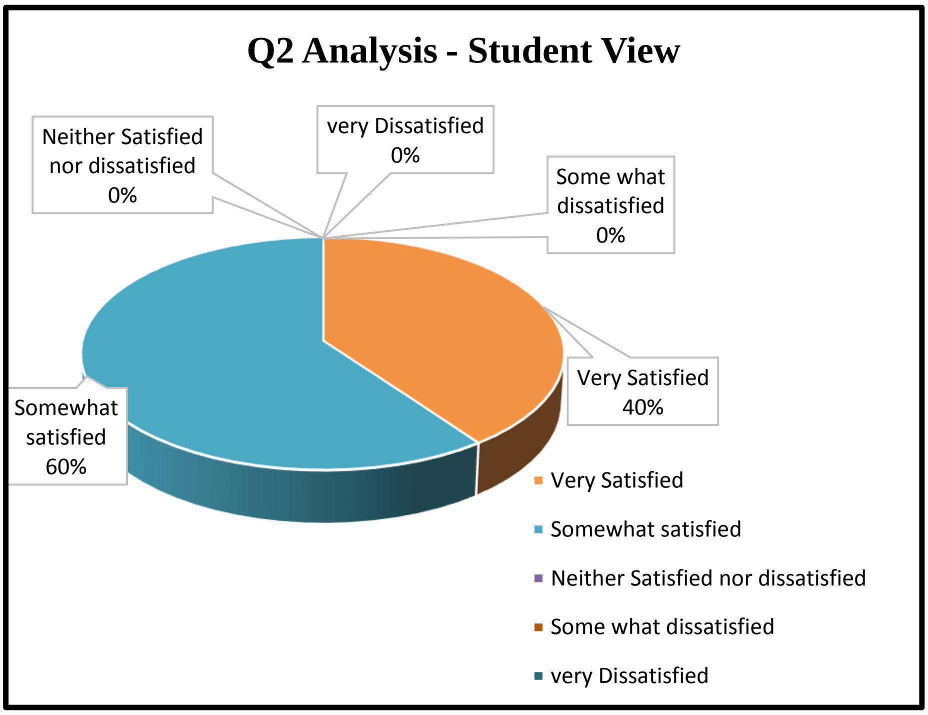

Figure 20: Question 02- student view

The third question, '**How much are you satisfied with the email receiving functionality of the web based educational data mining tool'**, The overall score achieved is 4.

Table 13: Response for Question 03- student view

Student ID	q3
S001	4
S002	4
S003	4
S004	4
S005	4
S006	5
S007	3
S008	2
S009	5
S010	5
Average satisfaction	4

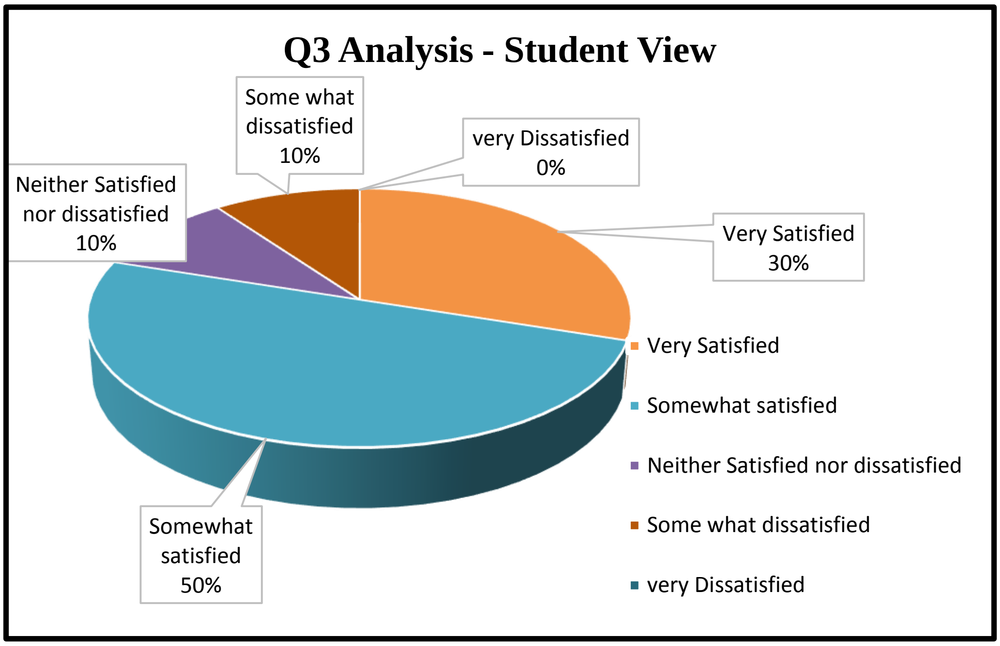

Figure 21: Question 03- student view

The fourth question was '**How much are you satisfied with the activity categorisation according to the importance of the activities (as most important, important, and averagely important)?**' The overall score achieved was 4.5.

Table 14: Response for Question 04- student view

Student ID	q4
S001	5
S002	4
S003	4
S004	5
S005	5
S006	4
S007	5
S008	3
S009	5
S010	5
Average satisfaction	4.5

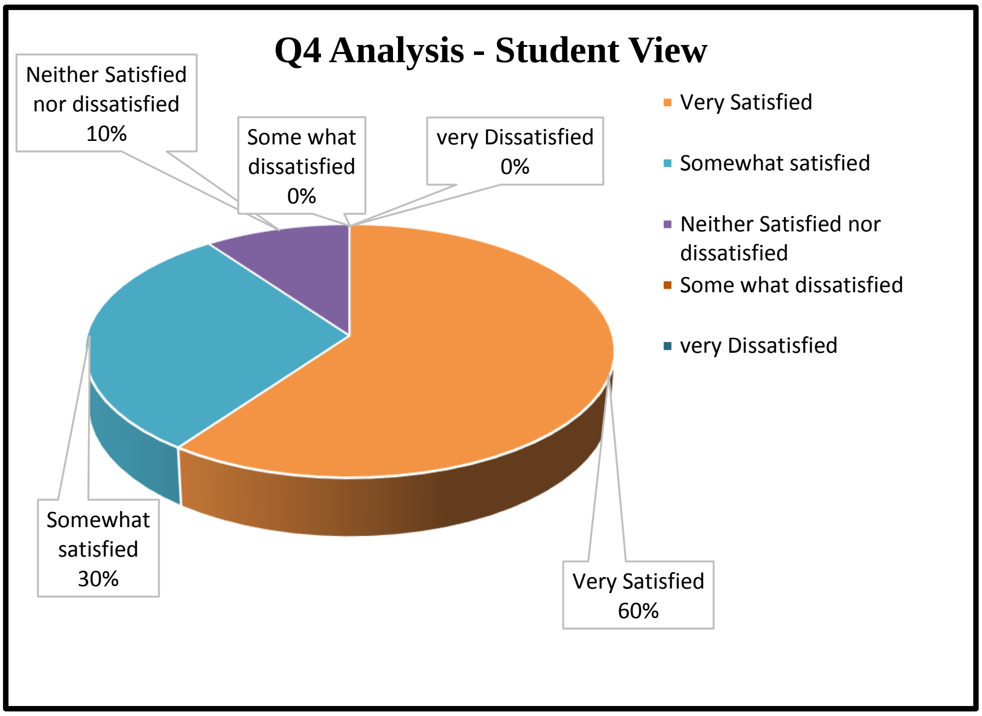

Figure 22: Question 04- student view

The overall average performance score achieved from the student survey is 86.0 percent, which is better than the teacher survey, which shows that I need to work on the teacher's section.

Table 15: Overall average performance - Student View

Student ID	q1	q2	q3	q4	Total marks	Overall Satisfaction (%)
S001	5	5	4	5	19	95.00
S002	4	4	4	4	16	80.00
S003	4	4	4	4	16	80.00
S004	5	5	4	5	19	95.00
S005	4	4	4	5	17	85.00
S006	4	4	5	4	17	85.00
S007	3	5	3	5	16	80.00
S008	5	4	2	3	14	70.00
S009	4	5	5	5	19	95.00
S010	5	4	5	5	19	95.00
Overall average performance						86.00%

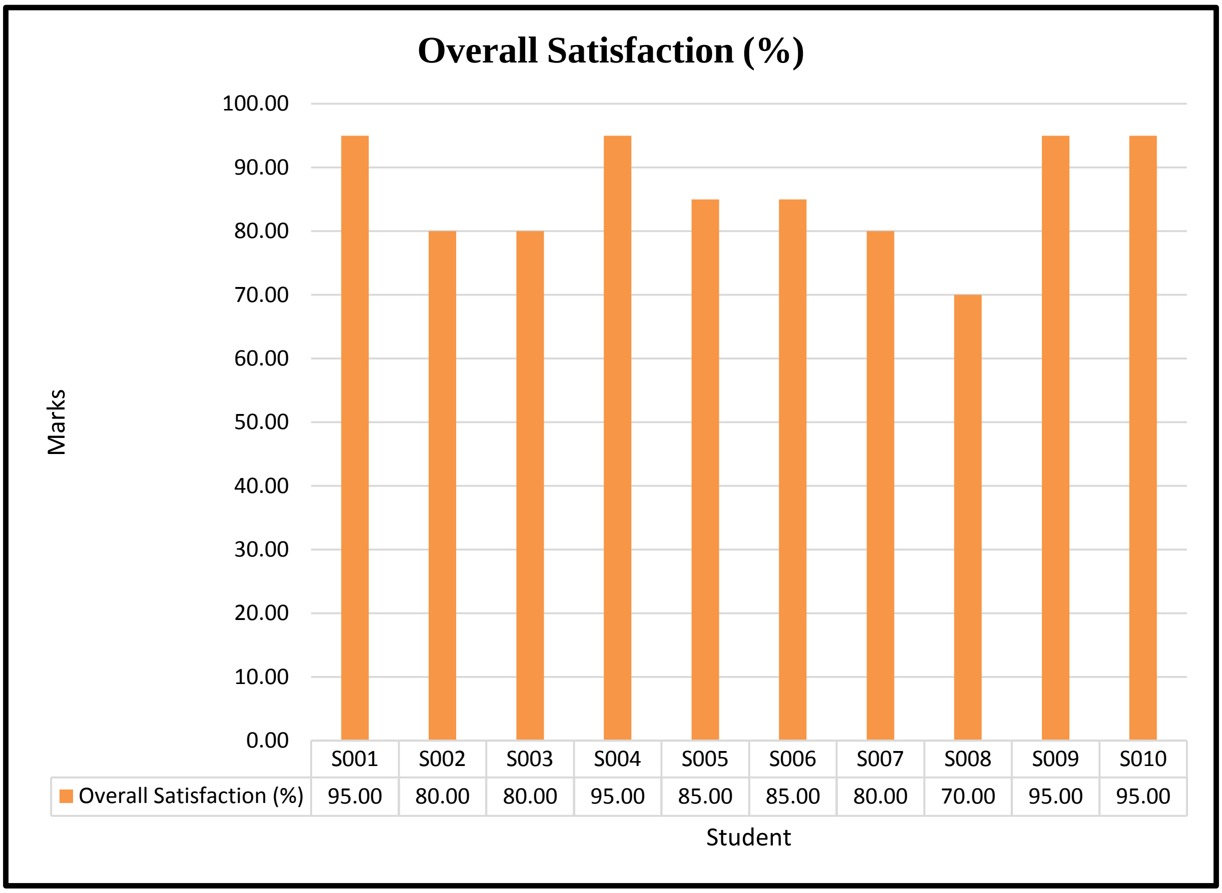

Figure 23: Overall average performance - Student View

We found out the most appropriate data-mining technique as k-means clustering after analysing the unlabelled data. Along with that, we dug out the relationship between the online learning activities and final grades by creating visualisations between the most important activities and the clicks, as the most important activities have the maximum number of clicks. Though we achieved objective number 2.

4.6 Summary

In this last section we will be summarising the above chapter, Cluster 2 has more weight than the rest of the clusters. Which certainly implies that Cluster 2 is the most important cluster and it contains more information about students than the Clusters 0, 1 and 3. So, cluster 2 is considered as the most important cluster in the entire model. Student from Cluster 2 is the most active learner, devoting the bulk of their time and effort to the virtual learning environment. Incredibly, when compared to other groups, their final evaluations are the most minimum, despite the fact that they are not easily distinguishable. In order to determine the causes behind this, we must do a more in-depth analysis of the material from Cluster 2 various VLE online learning activities that are partitioned into three groups most important, important and average important. Students have spent more than 70–80 per cent of their study time on the major VLE online learning activities. Only a small portion of their time was devoted to important and average important online learning activities on VLE.

Finally in the survey after aggregating the scores of all the questions, we calculated the overall average performance score as 77.2 per cent for educator view. This score implies that we need to work on the web-based tool to make it look better than the current version form the educators' perspective. The overall average performance score achieved from the student survey is 86.0 per cent, which is better than the educators'' survey, which shows that we need to work on the educators' view to provide better user experience. Next chapter based on overall conclusion of the study and future work.

CHAPTER 5

5 CONCLUSION AND FUTURE WORK

5.1 Introduction

This study is based on the potential of data mining in the educational context, along with the applications and implementations of k-means clustering, which was implemented on the number of clicks that were recorded by the WBEDMT for the online activities performed by the students. This chapter concludes the whole study that we have completed with the help of the tools WEKA, MYSQL, and Excel. Here we will be concluding all the results and evaluations found in the research in the below paragraphs. Along with the results, we will be giving attention to the limitations and the future work and its potential in very brief.

5.2 Conclusion

The performance of the web based educational data mining tool (WBEDMT) was very good and it gave quite convincing results as well. Still, our expectations were quite high regarding the performance of this tool as it was lacking in time-based activity measurement. However, the educators' and students both felt it was a major attribute of this tool. According to the educators', this tool is just 77.2 per cent workable, and according to the students, it was 86.0 per cent, although my expectation for this survey was to achieve more than 90%. It seems that these tools need some modifications as of now.

One of the objectives of the study was to find a suitable data mining technique for preparing the data for analysing with the help of different Literature Reviews. As per our review, the most appropriate technique we found is K-means clustering for our research project.

The secondary objective was to show the relationship between online learning activities on the VLE platform, and the student's final grades achieved based on the online activities performed on the platform. Still, this objective cannot be achieved completely due to the lack of information in the dataset, as the dataset contains only the number of clicks on the online learning activities, not the time spent on them. Hence, the objective achieved was partial. If we can include time spent on the activities, I think we can have better results.

5.3 Limitations

The limitation is that the dataset we used for this project for analysis is the number of clicks on the online learning activities of VLE, which the VLE traced. But the main problem is that we cannot estimate any student’s or an individual’s performance by only the number of clicks. We must consider the time spent on a single online learning activity. We could have done that by including a digital clock to measure any student's time spent on each activity.

Though our work gives a bigger picture of any student’s visiting pattern on any online learning platform. Further, the dataset didn’t provide activity logs about educators, and it was a major problem when implementing the web-based educational data mining tool for educators’ functional requirements.

5.4 Contribution to Computer Science field

We have successfully accumulated the raw data and cleaned it, giving it a proper story, and applying the most suitable algorithm to obtain the desired results, which tells us how all the students perform their activities while preparing for their final exams. We have made it easier for future generations to use the platforms for online learning with great efficiency, and research students can find and add some additional features and directions to our research to get to a better conclusion. Data mining experts can use our tools and track how the students perform while learning their subjects.

5.5 Future work

We can plan to implement our proposed web based educational data mining tool with the VLE, which can generate the data with hours or minutes spent on a particular online learning activity on a daily basis. For example, students and educators can have their interaction on the platform and, at the end of the day, they can measure the time they have spent on the platform. Via the above technique, we can predict the students' marks with the help of different machine learning techniques. The generated dataset can help students predict their marks before attempting their final exams. Through this, any student can analyse and improve their performance in the final exam.

5.6 Summary

After completing this evaluation and handling its results, we may infer that educational data mining is a wide data mining application area. In order to make reference to students' online learning activities on the VLE, the k-Means clustering method was used, and students were divided into four groups. The author conducted an in-depth analysis and depiction of the student’s online learning activities in each cluster. Because the students are unique people, all of the desired objectives may be achieved via the use of customised teaching and learning methods. In this case, data mining methods may help, and the investigation prompted authors to develop basics for computer experts, presenting unique resources for each and every student. Also, during this investigation, it was discovered that the actions of educators’ in the VLE had a significant impact on the online learning activities of students.

APPENDICES

APPENDIX A: SURVEY FOR EDUCATORS

- This survey is a part of research material for my master thesis. It would be great pleasure if you could fill out this questionnaires.
- Target audience: University academic staff members (lecturers, demonstrators, instructors)
- Description:

  The Virtual Learning environment is an online platform which can be used to handle online learning. This proposed web based educational data mining tool can be used as track student online learning activities on VLE and based on their interaction educators can send motivational messages.

Email Address: ______________________________________

1. **How much are you satisfied with the user interfaces of the web based educational data mining tool?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

2. **How much are you satisfied with the individual student's activity interaction feature of the web based educational data mining system?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

3. **How much are you satisfied with the Overall module interaction feature of the web based educational data mining system?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

4. **How much are you satisfied with the message sending functionality of the web based educational data mining tool?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

5. **How much are you satisfied with the activity categorization according to importance of the activities (as most important, important and averagely important)?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

6. **Any suggestion about the web based educational data mining tool.**

APPENDIX B: SURVEY FOR STUDENTS

- This survey is a part of research material for my master thesis. It would be great pleasure if you could fill out this questionnaires.
- Target audience: students from different graduate levels
- Description:

  The Virtual Learning environment is an online platform which can be used to handle online learning. This proposed web based educational data mining tool can be used as track student online learning activities on VLE and based on their interaction educators can send motivational messages.

Email Address: ____

1. **How much are you satisfied with the user interfaces of the web based educational data mining tool?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

2. **How much are you satisfied with the individual student's activity interaction feature of the web based educational data mining system?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

3. **How much are you satisfied with the message receiving functionality of the web based educational data mining tool?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

4. **How much are you satisfied with the activity categorisation according to importance of the activities (as most important, important and averagely important)?**

Answer	mark
Very Satisfied	5
Somewhat satisfied	4
Neither Satisfied nor dissatisfied	3
Somewhat dissatisfied	2
very Dissatisfied	1

5. **Any suggestion about the web based educational data mining tool.**

REFERENCES

Agarwal, S., 2012. Data Mining in Education: Data Classification and Decision Tree Approach. IJEEEE. https://doi.org/10.7763/IJEEEE.2012.V2.97

Ashby, R., Broughan, C., 2002. Factors Affecting Students' Usage of Virtual Learning Environments. Psychology Learning & Teaching 2, 140–141. https://doi.org/10.2304/plat.2002.2.2.140

Bhuasiri, W., Xaymoungkhoun, O., Zo, H., Rho, J.J., Ciganek, A.P., 2012. Critical success factors for e-learning in developing countries: A comparative analysis between ICT experts and faculty. Computers & Education 58, 843–855. https://doi.org/10.1016/j.compedu.2011.10.010

Black, E.W., Dawson, K., Priem, J., 2008. Data for free: Using LMS activity logs to measure community in online courses. The Internet and Higher Education 11, 65–70. https://doi.org/10.1016/j.iheduc.2008.03.002

Brusilovsky, P., Peylo, C., n.d. Adaptive and Intelligent Web-based Educational Systems 14.

E.Khedr, A., I. El Seddawy, A., 2015. A Proposed Data Mining Framework for Higher Education System. IJCA 113, 24–31. https://doi.org/10.5120/19839-1693

Falakmasir, M.H., Habibi, J., n.d. Using Educational Data Mining Methods to Study the Impact of Virtual Classroom in E-Learning 9.

Fayyad, U., n.d. From Data Mining to Knowledge Discovery in Databases 18.

Francis, R., Raftery, J., n.d. Blended Learning Landscapes 3, 6.

Graf, S., Kinshuk, P., 2006. An Approach for Detecting Learning Styles in Learning Management Systems, in: Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06). Presented at the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06), IEEE, Kerkrade, The Netherlands, pp. 161–163. https://doi.org/10.1109/ICALT.2006.1652395

Guo, P.J., Kim, J., Rubin, R., 2014. How video production affects student engagement: an empirical study of MOOC videos, in: Proceedings of the First ACM Conference on Learning @ Scale Conference. Presented at the L@S 2014: First (2014) ACM Conference on Learning @ Scale, ACM, Atlanta Georgia USA, pp. 41–50. https://doi.org/10.1145/2556325.2566239

Islam, N., Beer, M., Slack, F., 2015. E-Learning Challenges Faced by Academics in Higher Education: A Literature Review. Journal of Education and Training Studies 3, 102–112. https://doi.org/10.11114/jets.v3i5.947

Jamesmanoharan, J., Ganesh, S.H., Felciah, M.L.P., Shafreenbanu, A.K., 2014. Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm, in: 2014 World Congress on Computing and Communication Technologies. Presented at the 2014 World Congress on Computing and Communication Technologies (WCCCT), IEEE, Trichirappalli, India, pp. 200–202. https://doi.org/10.1109/WCCCT.2014.75

Keržič, D., Tomaževič, N., Aristovnik, A., Umek, L., 2019. Exploring critical factors of the perceived usefulness of blended learning for higher education students. PLoS ONE 14, e0223767. https://doi.org/10.1371/journal.pone.0223767

Khedr, A.E., n.d. A Proposed Data Mining Framework for Higher Education System. International Journal of Computer Applications 113, 9.

Kuzilek, J., Hlosta, M., Zdrahal, Z., 2017. Open University Learning Analytics dataset. Sci Data 4, 170171. https://doi.org/10.1038/sdata.2017.171

López, M.I., Luna, J.M., Romero, C., Ventura, S., n.d. Classification via clustering for predicting final marks based on student participation in forums 4.

Mamcenko, J., Sileikiene, I., n.d. Intelligent Data Analysis of E-Learning System Based on Data Warehouse, OLAP and Data Mining Technologies 5.

Markellou, P., Mousourouli, I., Spiros, S., Tsakalidis, A., n.d. Using semantic web mining technologies for personalized e-learning experiences 7.

McGill, T.J., Hobbs, V.J., 2007. How students and instructors using a virtual learning environment perceive the fit between technology and task: VLEs and task-technology fit. Journal of Computer Assisted Learning 24, 191–202. https://doi.org/10.1111/j.1365-2729.2007.00253.x

Mehboob, B., 2017. Student Performance Prediction and Risk Analysis by Using Data Mining Approach 8, 9.

Menaka, S., Kesavaraj, G., 2019. Predicting Student Performance Using Data Mining Techniques: A Survey Of The Last 5 Years 4, 5.

Merceron, A., Yacef, K., 2005. Educational Data Mining: a Case Study. pp. 467–474.

Mostow, J., Beck, J., 2006. Some useful tactics to modify, map and mine data from intelligent tutors. Nat. Lang. Eng. 12, 195–208. https://doi.org/10.1017/S1351324906004153

Peña, C.I., Marzo, J.-L., n.d. UN SISTEMA DE TUTORÍA INTELIGENTE ADAPTATIVO CONSIDERANDO ESTILOS DE APRENDIZAJE 14.

Romero, C., Ventura, S., 2010. Educational Data Mining: A Review of the State of the Art. IEEE Trans. Syst., Man, Cybern. C 40, 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Rotondo, A., Quilligan, F., 2020. Evolution Paths for Knowledge Discovery and Data Mining Process Models. SN COMPUT. SCI. 1, 109. https://doi.org/10.1007/s42979-020-0117-6

Sibanda, M., 2014. The Impact of E-Learning on Student Performance: A Case Study of an Entry-Level Module at a South African University. MJSS. https://doi.org/10.5901/mjss.2014.v5n9p478