Learning Analytics: A Study of the Dynamics impacting Students Performance in a VLE

Tolanitoluwa Rita Awoliyi

Abstract

Virtual Learning environments have directly reflected that there are newer and more advanced methods in which higher educations are seeking to embrace learning because of the affordances it offers to its users, the students, instructors and administrators. It is from the various activities that are involved in this environment that massive amount of data is generated and accumulated into learning analytics. Some scholars suggest that the pandemic sporadically increased the participation of students in online learning activities.

The study sought to understand and explore the various factors such as age, gender, location, course module, course presentation, education qualification that would affect their academic performance from their engagements in the VLE using the Open University Learning Analytics Dataset. These are further indicated as independent variables and the performance outcome as the dependent variable.

Different studies have looked at how they can use ML algorithms to predict a better performance outcomes and reducing the rate of failures and withdrawals especially for OULAD while other researches were conducted to explain different factors affect students' academic performances in a VLE. A python programming software was used to conduct a multilayered statistical analysis from an EDA to a Chi square to linear regression analysis was used to obtain appropriate result outputs.

The findings revealed that there is a high level of significant relationship between students who had a high number of sum of clicks and their performance. Also course module had a statistically significant relationship with academic performance.

Keywords: Virtual Learning Environment (VLE); Exploratory Data Analysis (EDA); Open University Learning Analytics Dataset(OULAD); Learning Analytics, Performance; sum of clicks

1.0 Introduction

1.1. Introduction Background of the study

The world of education has changed dramatically in the last few decades. With the advent of the internet, people can now access information and education from anywhere worldwide (Hashim, Tlemsani and Matthews 2022). This advent led to the development that sporadically changed the delivery and administration of knowledge. It brought about an increasing type of education known as online education. Online learning environments have been called different names, including virtual learning environment (VLE), learning management systems (LMS), collaborative learning software (Flavin and Bhandari 2021), educational technology, distance Learning, and online education (Da Silva et al. 2021).

Over the years, learning has improved through fundamental impacts and shifts in technological advancements, which is evident in the transformation of its environment (Thornbury 2020). The advent of COVID-19 brought about an accelerated technological conversion in the learning process, causing more educational facilities to opt for e-learning over conventional face-to-face learning (Dhawan 2020; Chiodini 2020; Adedoyin and Soykan 2020). Institutions sought new teaching platforms such as Massive Open Online Courses (MOOCs), VLEs, and LMS (Khan 2021; Hasan and Khan 2020; Al-

Maroof et al. 2021). However, some scholars acknowledged that due to the sudden pandemic, many educational institutions were unprepared for the transition from face-to-face to open and distance learning. (Chiu, Lin and Lonka 2021; Ghani et al. 2021).

Student's engagement in various online learning platforms has generated massive amounts of data. The use and interpretation of this enormous dataset have been linked to the emergence of learning analytics (Agudo-Peregrina et al. 2012). Adejo and Connoly (2017 p.2) explained learning analytics (LA) as "the collection, processing, analysis and reporting of data and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs." LA is a field of educational research whose usage generates voluminous information on learners' data which can be explored and harnessed through the learning analytics process (Banihashem et al. 2018; Hidalgo and Evans 2020). It employs a technological form of acquiring relevant students' information in computer language to improve their learning and performance options (Adejo and Connolly 2017). In other words, LA seeks to capture data from various VLEs powered by the interactions in a data-driven platform. This enables the VLEs to thrive with accessibility, rich information, and manipulation of resources for both learners and instructors (Raj et al. 2021). These VLEs comprises course materials, tests, and assessments, which may also have communicative tools (like chat box, forums, etc.) for learners and instructors to interact (Abuhlfaia 2020; Hamid et al. 2018). With a VLE, one may create and administer a whole online course or use it to supplement conventional teaching.

The immersion of virtual learning into learning analytics as described by Hamid et al. (2018), is the learning environment mediated by computers, digital technologies, and augmented reality. VLEs are the future of education because they rapidly surpass the experience students have in traditional classrooms (Rashid et al. 2021). The implementation of LA produced a more beneficial and convenient learning environment that provides the framework for the educational sector to make the required modifications to enhance the learning opportunities and performance of students (Alves, Miranda and Morais 2017). According to Aljohani et al. (2019), educational institutions today concentrate on enhancing the calibre of teaching and learning while simultaneously improving students' achievement. This has made learning analytics a viable method to accomplish these objectives (Raj et al. 2021; Banihashem et al. 2018).

Past research has shown that the experience students have in a VLE differs from the physical atmosphere; hence, many students struggle to excel in online learning settings, and consequently, they either quit or are unable to receive passing grades (Namoun and Alshanqiti, 2020). These experiences often lead to different interactions in a VLE, and these interactions become a vital and pertinent part of learning. Sheung Au (2019) identified different types of interactions in online learning. These interactions are generated from the engagement of multiple channels, which are beyond physical proximity. Such interactions are among complex agents and have allowed users to manipulate different applications. These interactions are identified as peer and device; peer, device and tutor; peer, device and course content; peer, device, tutor and administrator. These interactivities produce a significant quantity of information, including user details, user behaviours, test scores, assessments, and the count of engagements with course materials (Kuzilek et al. 2015).

In a VLE, there are several internal and external factors why students pass, fail, or drop a course. According to Eriksson, Adawi and Stöhr (2017), Onah, Sinclair and Boyatt (2014), and Christensen and Spackman (2017), the proportion of students dropping an online course is significantly higher than that of a conventional course. This is partly attributed to health problems, inadequate preparation, lack of academic vigour, financial reasons, poor time management, and personal issues (Ahmad et al. 2021). Past studies explained that students' performance could be better understood through the examination of students' behavioural and demographic characteristics and learning conditions (Alves, Miranda and Morais 2017; Wakelam et al. 2020). Kuzilek et al. (2019) and Omona (2022) agree that

characteristics such as students' time, duration and location of study can be used in the analysis and prediction of students' behaviour in a VLE. Early identification of performance-influencing factors may help to improve substandard academic outcomes (Adejo 2017).

Some academics have referred to the VLE as the future of education because it rapidly surpasses the experience students had in traditional classrooms. These studies acknowledge that data generated from VLEs could be used to improve user interfaces, reduce the academic discouragements or challenges experienced by students during their educational journey and strengthen the interaction of teaching and learning between instructors and students (Da Silva et al. 2022; Kuzilek, Hlosta and Zdrahal 2017). Davies (2020) acknowledged that the availability of VLE brings about a variety of learning options among learners, instructors, and administrators. It also eliminates the constraints of typical face-to-face interaction.

Some drawbacks have been documented, such as poor connection, unfavourable learning conditions and a lack of minimum learning requirements (Mseleku 2020). Adejo (2017) believes that developing reliable and valid metrics that can be used to assess learning outcomes can be a challenge. He acknowledged that more research is required in analytics to improve student success rates, for instance, identifying the best LA tools and services that can achieve specific educational objectives, such as improving students' performance (Adejo 2017).

1.2. Rationale

According to a study by Al-Azawei and Al-Masoudy (2020) and Lotsari et al. (2014), there is still a dearth of studies looking into the variables that affect students' performance. This sets the premise of the need to investigate the factors and characteristics of students that impact learners' academic performance in a VLE.

Ahmad et al. (2021) highlighted more benefits of learning analytics, such as predicting students' performance. The study emphasised the significance of determining the variables that can be used to track students' performance and accurately predict their future performance, such as the date of graduation, estimated final grade point averages (GPAs), and the likelihood of succeeding, failing, and withdrawing.

According to Jha et al. (2019), there are nine publicly available learning analytics datasets, including Open University Learning Analytics Dataset (OULAD), available online. Some datasets, including KDD Cup 2015, Khan Academy, Coursera, and others, have explored students' interactions in a VLE. However, this research used the OULAD, which includes students' demographic, behavioural and assessment information. This research seeks to establish the impact of the variables in the OULAD on students' academic performance using statistical techniques such as exploratory data analysis (EDA) and associational statistical analysis.

1.3. Aim

This study aims to explore the variables in the OULAD to understand the factors that impact students' academic performance.

1.4. Objectives

The following objectives guide this research:

1. To critically review the literature on factors that impact students' performance in a VLE.

2. To explore the OULAD by applying Exploratory Data Analysis (EDA) techniques to understand the variable(s) that impact students' performance.

3. To test each selected variable against the performance outcome and establish their impacts on performance in a VLE.

4. To critically assess the statistical significance of the variables influencing students' performance in a VLE.

5. To propose features and strategies that administrators and instructors can use to improve academic performance.

Hypotheses

Hypothesis one:
$H0$: There is no significant relationship between the code module and performance
$H1$: There is a significant relationship between the code module and performance
Hypothesis two:
$H0$: There is no significant relationship between the code presentation and performance
$H1$: There is a significant relationship between the code presentation and performance
Hypothesis three:
$H0$: There is no significant relationship between the gender and performance
$H1$: There is a significant relationship between the gender and performance
Hypothesis four:
$H0$: There is no significant relationship between the region and performance
$H1$: There is a significant relationship between the region and performance
Hypothesis five:
$H0$: There is no significant relationship between the highest education and performance
$H1$: There is a significant relationship between the highest education and performance
Hypothesis six:
$H0$: There is no significant relationship between the age band and performance
$H1$: There is a significant relationship between the age band and performance
Hypothesis seven:
$H0$: There is no significant relationship between the disability and performance
$H1$: There is a significant relationship between the disability and performance
Hypothesis eight:
$H0$: There is no significant relationship between the assessment type and performance
$H1$: There is a significant relationship between the assessment type and performance

1.5. Justification and significance of the study

The pandemic outbreak in 2019 brought about a rapid change in the learning, and teaching approaches used not only in the education sector but in vast communication methods across businesses and organisations locally and globally (Khan 2021; Hasan and Khan 2020; Al-Maroof et al. 2021). Containment of the outbreak led to the abrupt interruption of many activities by limiting physical contact hence, the need to transition into virtual communication. Universities adopted virtual and digital tactics inside their selected Virtual Learning Environments (VLEs), such as Blackboard, Canvas, and Moodle (Albreiki et al., 2021).

Based on the scope of the research, which is limited to the Open University, UK's VLE, this paper will identify the variables in the dataset that can influence students' academic performance. Identifying these factors will guide educational administrators to pay attention to the learner and their behavioural characteristics. It will help students understand the results of their learning patterns. Overall, this research will benefit academic institutions in planning, improving, and facilitating students' academic performance, learning culture and teaching platforms. It would also help the

institutions in using the vast data generated by students in making insightful and value-adding decisions.

2.0 Review of Literature

2.1 Introduction

This chapter critically examines a diverse range of published literature on the factors that affect students' performance in a VLE. Considering the aim and objectives of this study, a variety of studies relating to the topic are presented below. First, the nature of VLEs will be examined by looking at the definitions, benefits and the challenges encountered in its environment. Additionally, the connection between the VLE and the application of learning analytics to the data generated from a VLE will be explored. Also, the factors influencing students' behaviour and academic performance will be discussed. Furthermore, ethical considerations around these factors will be considered. Finally, a review will be done to acknowledge the studies that have analysed the Open University Learning Analytics Dataset (OULAD) to reveal the goal of this research.

2.2 Conceptual understanding of a VLE

In several studies, VLEs have been identified to mean any form of learning that involves merging internet-related connectivity or the application of computer-generated software to enhance learning (Omona 2022; Alhakbani and Alnassar 2022). Ketelhut and Nelson (2021) and Casalino, Castellano and Vessio (2020) identified it as a more common and acceptable form of engaging learners in a virtual environment that universities have come to adopt their academic structure. VLEs represent a comprehensive platform that acts as an extended form of distance learning while merging it with online activities (Da Silva, Lidia Martins et al. 2021; Casalino, Castellano and Vessio 2020).

Nazif, Sedky and Badawy (2020) and Jha, Ghergulescu and Moldovan (2019) define VLEs in terms of the volume of students it accommodates, their online accessibility and course pattern. Studies in Nazif, Sedky and Badawy (2020) differentiate distance/online learning from MOOCs because the former mainly involves a specific institution and registered students. However, the study highlights that both environments harness technologically aided devices and applications regardless of whether they are an extended version of the institution or part of an unaccredited programme. These studies have explained the spectacular factors that help determine the components attributable to a VLE.

2.2.1 Benefits of a VLE

Several scholars have pointed out some benefits of a VLE. According to Da Silva et al. (2021) and Flavin and Bhandari (2021), VLEs encompass a structure that gives students the advantage of learning at their convenience, subsidised prices, high quality, and the opportunity to make education a system devoid of physical limitations. Dhawan (2020) notes that students gain a lot from using A VLE because of its flexibility and advantages, causing universities to invest in the technological requirement of the virtual environment (Omona 2022). Adnan and Anwar (2020); Dhawan (2020); Marinoni, Van't Land and Jensen (2020); Coman et al. (2020) proposed that VLEs provide the student with simultaneous experiences between virtual reality and real-life experiences because it uses applications such as zoom, google fora (classrooms), chats, video conferences and others. This ability to utilise the VLE to satisfy the students' purpose has led to an impactful development to infuse the physical experiences into the virtual counterpart. These scholars explain that the advent of technological development helps to facilitate the dissemination of content to a vast number of students in real-time. Through this, learners can adapt to their learning pace and study objectives at a reduced cost.

According to the United Nations Educational, Scientific, and Cultural Organisation (UNESCO), COVID-19 contributed to the strengthening of one of the sustainable development goals (SDGs), notably digital mediation and its application in education for enhancing learning, inclusion, and equity

(Adedoyin and Soykan 2020; Williamson, Eynon and Potter 2020; Pedro et al. 2019). In addition to the technological development occurring periodically, Adnan et al. (2022) and Da Silva et al. (2021) explained that the advent of the pandemic extended the methods of tutoring into distance learning (DL) or e-learning following World Health Organisation (WHO) guidelines to combat the viral infection. Digital intervention in education allowed educational advancement in countries that had not yet assembled their pedagogical system, especially in third-world countries (UNESCO 2021). This led to the need to understand further and study the difficulties encountered while using the VLEs to ameliorate the effects of social distance safety measures.

2.2.2 Challenges of a VLE

Several researchers have highlighted several weaknesses of the VLE. Omona (2022), Pardo and Siemens (2014) and Flavin and Bhandari (2021) identified some challenges within a VLE which hinder the students' highest academic expectations. Some of these are delayed communication from lecturers due to poor connectivity and not being online simultaneously. Additionally, feeling isolated compromises opportunities for some learners.

Additionally, several studies, including those by Fonseca, Trimmel and Bachmann (cited in Weaver et al. 2021) asserted that substantial progress in student motivation, involvement, sustained knowledge acquisition, and excellence have been linked to the utilisation of VLE for web-based self-directed study.

Further difficulties students experience in a VLE are summarised in table 1 below.

Table 1: Indicating the challenges students and institutions encounter in a VLE.

	VLE Challenges	Description	Study
1	Difficulty in merging the curriculum with Computer Assisted Language learning.	Curriculums are usually designed with clear links, formats and expectations where they use interactive devices and multimedia for physical classes but this has posed a challenge especially for teachers/instructors teaching students with special needs. They find it difficult to use CALL to immerse their students in learning activities.	Guangul et al. (2020) and Dung (2020)
2	Interpersonal communication is reduced in a VLE when compared to a face-to-face class.	It involves peer-to-peer, peer-to-tutor, and peer-to-multimedia devices where they engage in interactions. More than a third percentile of students in 800 colleges expressed their displeasure in having virtual classes.	Klawitter (2022)
3	Learning expectations are not met.	Based on the design of the VLE, it is challenging to monitor and achieve maximum level of concentration from all learners. Often, not all participants are carried along equally. Tutors are not able to adequately monitor students who are not participating.	Omona (2022) and Alhakbani and Alnassar (2022)

4	Low motivation, loneliness and high level of distractions.	Without physical meetings and communications, learners have less drive to learn, many feel isolated and some students are unable to concentrate/focus in the online sessions.	Omona (2022) and Klawitter (2022)
5	Technical issues of devices, gadgets, multimedia equipment and connectivity.	Poor internet connection can cause a disruption in the learning experience of the student. Also, students' classes can be affected due to faulty gadgets. These technical glitches in the environment causes dissatisfaction among learners and tutors.	Klawitter (2022); Da Silva et al. (2022) and Alhakbani and Alnassar (2022)

Regardless of these challenges of the VLE, studies reveal that universities have made corrective methods for both students and lecturers through student retention, learner's behaviour and maintaining a balanced environment by providing financial support in their quota distribution for the institution and designing an analytical system which utilises an environment of sharing meaning and understanding (Bart et al. 2020; Hassan et al. 2019; Hamid et al. 2018; Adejo and Connolly 2017a; Adejo and Connolly 2017b).

2.3 The use of VLEs and the development of learning analytics (LA)

According to Alves, Miranda and Morais (2017) and Agudo, Hernandez and Iglesias (2012), VLEs currently exist within the pedagogical systems in higher education. As a result of these current integration, LA has been utilised to redefine knowledge transmission and assessment by incorporating academic analytics. Agudo, Hernandez and Iglesias (2012) recognized that LA could be applied to enhancing VLE by using the big, massive data generated from VLE platforms. Researchers have acknowledged that these large quantities of data can be problematic to comprehend. (Lidia M. et al. 2022; Da Silva et al. 2021; Alonso and Casalino 2019). Nevertheless, Yousafzai et al. (2021), Ahmad et al. (2021), Dhawan (2020) and Adejo (2017) understood that the availability of big data allows university institutions to apply LA in making informed decisions to support students. Hasan et al. (2021) and Agudo, Hernandez and Iglesias (2012) further acknowledged that the data in the educational sphere is snowballing and explains that the component of these big data includes admission data, academic information and data showing the microscope, and moment-by-moment activities during the learning process.

Big data is a numerous collection of different information on a set of defined populations with the aim of understanding patterns, outliers, associations, and relationships within the population (Da Silva et al. 2022; Banihashem et al. 2018; Rienties et al. 2017; Holmes et al. 2019). Shaffer and Ruis (2017), Liu et al. (2022) and Adam et al. (2020) admitted that big data could not speak for itself; as such, Educational Data Mining (EDM) and LA can be used in predicting and classifying students' behavioural and demographic characteristics.

Some scholars have identified several LA tools that can be applied to big data and their challenges. According to Lester et al. (2019), poorly designed data displays, and communications may hamper the usage of LA. Also, teachers' use of LA tools may be limited by time restrictions, motivation, and practical training. These LA tools and methods include "visualisation methods, social network analysis, semantic, and educational data processing such as prediction, association mining, model development, data separation for human analysis. Gapminder, IBM many eyes, are some visualization tools used to uncover hidden patterns and trends in large, unstructured information. Other tools such as Rapidminer, SPSS, Weka use programming languages like Python, R and Java to perform complex statistical analyses.

Several techniques such as Exploratory data analysis, statistical techniques and Machine Learning can be used to view, explore and analysis big data. Other uses of these approach include include outlier detection, relationship mining and creation of predictive models (Lester et al. 2019; Avella et al. 2016; Romero and Ventura 2020).

2.3.1 EDA, Statistical techniques and ML

Exploratory Data Analysis has been described as the foremost stage in data analysis to understand a dataset. This involves the application of statistical and visualization methods to identify patterns and relationships in the dataset. Several academic materials have identified univariate, bivariate and multivariate analysis as the three ways of performing EDA (Sayad 2022 and Magdum 2022).

Univariate analysis is applied if the dataset has a single variable, bivariate methods are used for datasets with two variables while multivariate methods are used when the dataset exceeds two variables. Figure 1 below shows a breakdown of some analytical and graphical approaches of conducting a data exploration as identified by Sayad (2022). The quantity, distribution, spread, percentage and relationship of the values in the study variables can be determined using various statistical methods, test, frequency tables, plots and charts.

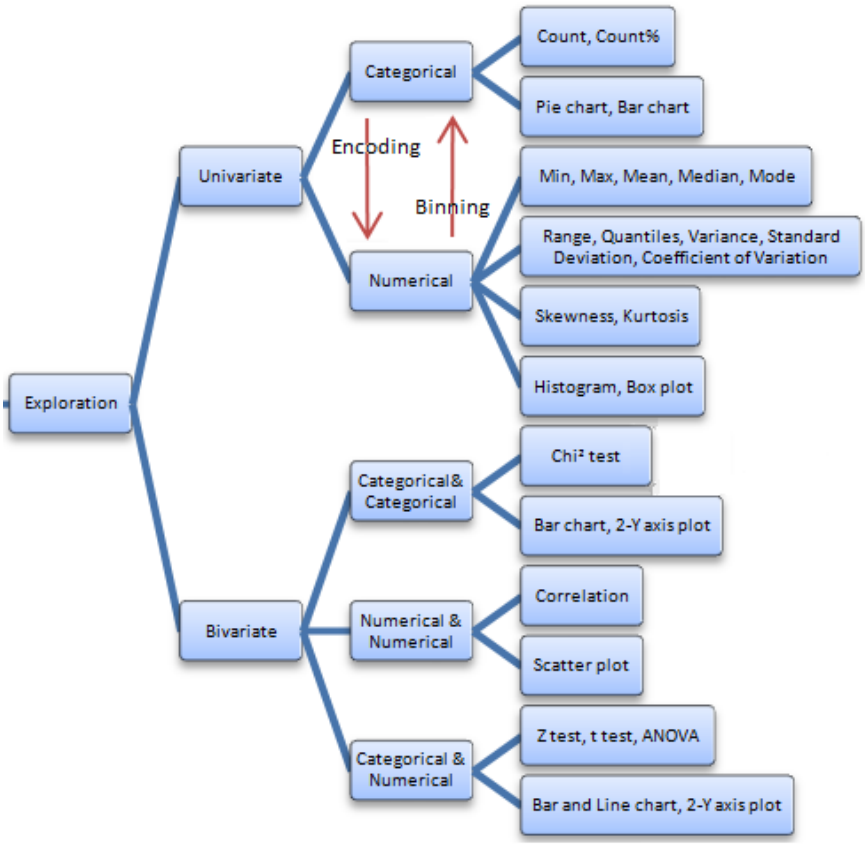

Figure 1: Approaches of Data Exploration (Sayad, 2022)

The use of Machine Learning Algorithms (MLA) in understanding, discovering insights from, and making decisions from massive datasets has been widely advocated. Both supervised and unstructured datasets can be used with ML to do predictive analysis. The prediction model is known as a classification model if the target variable is categorical. However, if the outcome is numerical, it is referred to as a regression model. Similar groupings are created via clustering, while association rules can be used to identify relationships between observations. Figure 2 shows the main groups and sub-methods of the classification, regression, clustering and association algorithms. This includes simple classification methods like ZeroR, which only considers the target variable, Naive Bayesian,

which assumes that one feature in a class has nothing to do with the presence of any other features and K nearest neighbors, which categorises new cases based on a similarity metric (Sayad, 2022; Alnassar et al. 2021; Rahmany, Zin and Sundararajan 2020; Xia 2020).

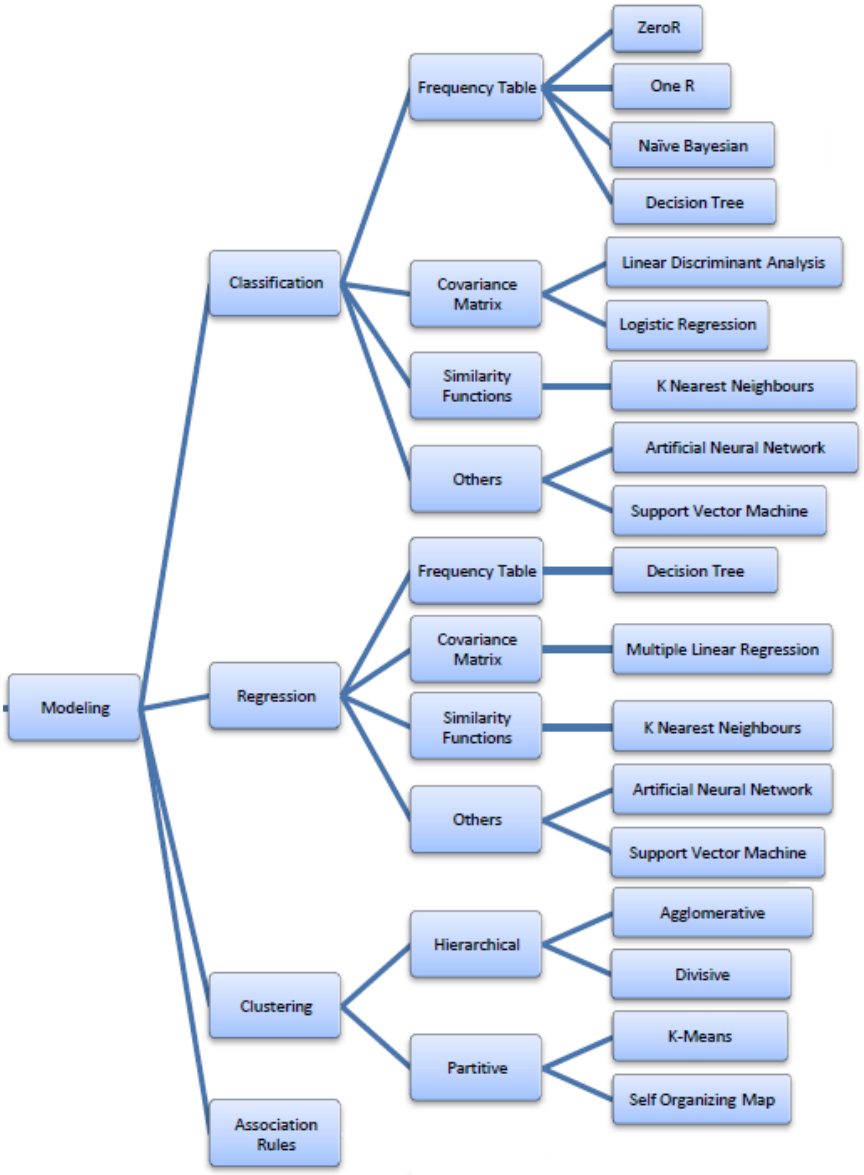

Figure 2: Approaches of Modelling (Sayad, 2022)

2.4 Factors that impact performance outcome in a VLE

Several pieces of literature have discussed students' performance and the factors influencing it in a VLE. According to Qiu et al. (2022), factors affecting academic achievement in a VLE can be classified under tendency indicators or behaviour indicators. Tendency indicators are classified under static data and these are accumulated prior the commencement of a programme. Some of these are location, economic status, gender and past academic results. Chen, Wang and Zhou (2022) and Bilal et al. (2022) defined behaviour indicators using consumption laws, living habits, learning and internet access to measure the behaviour.

On the other hand, Coldwell et al. (2008) noted that demographic factors such as gender and nationality influence students' performance. Additionally, class interactions and involvement were found to have a substantial and positive impact on performance. However, there was no correlation between the age of students and their academic performance. Regarding gender, female students had more engagement in the VLE activities than their male counterparts. The study also claimed that nationality affected student performance. The research showed that students from western countries fared better than those from Asian countries, despite the latter's greater online learning engagement. Dhawan (2020) had similar findings. He noted that students from minority and black heritage had lower chances of passing a course when compared to their Asian and Black skinned course mates. Overall, the white-skinned learners performed better.

Students' engagement in a VLE was analysed using 38 courses at an anonymous on-campus university in the United Kingdom. Findings revealed that students' performance was directly related to engagement with the VLE. Nonetheless, the drawback of the study was determining the VLE usage alone because of the students' blended learning. In the same survey, individual modules were analysed against science-based and non-science-based courses, and it was discovered that science-based subjects have a higher dependency on VLE activity (Boulton, Kent and Williams 2018).

Other factors, for instance, geographical location, economy, and technological infrastructure, can influence the number of withdrawals from online courses. Al-Azawei and Al-Masoudy (2020) discovered that there are more withdrawals from students in developing nations from online learning courses than traditional learning. This is attributed to the lack of a learning environment and prompted researchers to explore behavioural and demographic factors to comprehend learners' online accomplishments.

In a study to clarify whether engagement with VLE self-directed study can enhance the learning process and augment academic achievement, Weaver et al. (2021), in their research of the learning process and outcomes of VLE self-directed study, concluded that digital technology is not enough to influence performance. Instead, its effective incorporation and alignment to planned learning goals and summative evaluations are essential.

Yunita, Santoso and Hasibuan (2021) believe that student's success is determined mainly by tutors' grading procedures and policies. In addition, some scholars investigated the impact of VLE on learning outcomes and identified that software design plays a role of importance (Wessa, De Rycker and Holliday 2011). Furthermore, the study suggested that rather than adopting a general-purpose VLE design, a course or content-focused method is more effective, giving allowance for learning materials and activities tailored to specific topics.

Students' levels of resilience and involvement, mainly their capacity to complete their studies relative to their final performance was investigated by (Ahmad et al. 2021). The study found that in addition to resilience and student engagement, gender plays a role in performance when combined with other variables. Other knowledge, attitudes, and practices (KAP) factors such as physical activity, aerobic fitness, motor skills, motivational beliefs and learning strategy, family background, culture and region were identified to impact academic performance.

2.5 Ethical considerations

Data is accumulated through students' use and interaction in VLEs. These include their information and activities. (Slade and Prinsloo 2013) identified three specific areas that encapsulate ethical issues in a VLE. Firstly, the environment of data collection and its interpretation. Secondly, the permission, privacy, and removal of data identification. Thirdly, the classifying, management, and storage of data. Regarding user privacy, there have been concerns about ownership of data and the access of third

parties to the data (Prinsloo and Slade 2017). Also, Qiu et al. (2022) acknowledge that students' behavioural characteristics can be linked directly to their performance, and therefore data that contains such information has ethical and privacy implications. According to Qiu et al. (2022), there are certain information that does not raise ethical concerns, such as frequency of logins or clicks, accessibilities, and the number of students in a group. Corrin et al. (2019) propose that curbing misuse of such submitted data for educational purposes on VLEs is tedious and has various complexities. It becomes essential for policymakers to observe these complexities and challenges to minimise ethical concerns. Spencer and Patel (2019) suggest providing a guideline to cover a multistage approach to address ethical issues. According to Kuzilek, Hlosta and Zdrahal (2017), the students of Open University UK are usually informed that their data will be collected and shared with third parties for research purposes. However, the OULAD dataset was anonymised so that all personal identification information was deleted, for example, date of birth and social security numbers.

2.6 Open University Learning Analytics Dataset (OULAD)

Enough data sources are available on the internet to analyse and anticipate students' and learners' data patterns in a learning environment. The Society for Learning Analytics Research (SOLAR) asserted that the Open University, UK is the global leader in using big data for gathering and analysing voluminous details of students and applying the findings to improve students' academic performance (Khan 2021; Hidalgo and Evans 2020). The table below summarises the aim, objectives, methods, and findings of some past studies that used the OULAD. Please refer to the bottom of the table for the meaning of abbreviations used in table 2.

Table 2: Table of Summary Presenting the Past Research Using Open University Learning Analytics Dataset. summary of past research using OULAD.

Author	Methodology	Research Aims	Results	Dominant factors	Featured used
Jha et al. (2019)	ML Algorithms such as DRF, GBM, DL and GLM were used.	To use regression and deep learning methods to forecast dropout rates using attributes such as demographic information, assessment score and observations from the VLE interaction.	The models created based on demographics information achieved a minimal value between 0.62 and 0.65 validation data. The models based on all attributes achieved close to 0.01 higher AUC than the models based on the VLE interactions alone. The machine learning models based on student's interaction with the VLE achieved high performance in terms of AUC.	No dominant factors were identified. The focus was on the overall accuracy of the model.	All features were used
Poudyal, Mohammadi-Aragh and Ball (2022)	2D CNN . The data obtained was converted into 2D format using zero padding to increase the features from 37 to 40 and then a 40-length array was reshaped to be suitable for 2D CNN.	CNNs was used to test the hybrid model and compare it with baseline models and different learning rate was used to determine the performance model.	After the conversion from 1D to 2D, the scholars were able to predict academic performance with a high accuracy of 88% using their model. The hybrid CNN can be applied to numerical 1D educational datasets to predict student academic performance.	No dominant factors were identified. The focus was on the overall accuracy of the model.	All features were used
Verma, Singh and Srivastava (2021)	Experimental analysis using the k-NN, SVM and ANN approach.	To prediction of the performance of the dataset using data mining and ML methods.	The K-NN algorithm generally performed better than the ANN, SVM, Naïve Bayes and Random Forests for feature variations.	No dominant factors were identified	Demographic, Engagement and past performance features.
Alnassar et al. (2021)	SVC, K-NN, and ANN ML methods were used.	To predict students' performance using three selected ML algorithms	The K-NN approach was the most appropriate for OULAD.	No dominant factors were identified	Demographic, Engagement and past performance features.
Casalino, Castellano and Vessio (2020)	Two ML algorithms: RF and ARF were used.	To predict students' outcome and obtain key features by focusing only on 25,819	Higher accuracy was achieved for the prediction of semester 3 and semester 4.	Study credits, average score and clicks	All features were used

Author	Methodology	Research Aims	Results	Dominant factors	Featured used
		students. Also, to assess the effect time has in the prediction of students' performance.	While semester 2 had the lowest accuracy using both the RF and ARF techniques.		
Nazif, Sedky and Badawy (2020)	PNN was the primary ML techniques used. Also, MATLAB programming languages, Microsoft SQL and Visual Studio packages were used.	To examine the trends in students outcomes based on the feature selection algorithms.	An accuracy of 93.4% was attained using PNN in addition to the FSCNCA feature selection Algorithm. RF, DT and KNN performed well with 91.7%, 91.2% and 89.8% respectively.	Study credits and sum clicks	16 features were selected from the whole dataset.
Casalino, Castellano and Mencar (2019)	The utilization of the DISSFCM algorithm in processing the dataset as a data stream.	To show the effectiveness of the adaptive fuzzy clustering algorithm in educational data analysis and the prediction of students performance .	The DISSFCM algorithm can accurately detect underlying insights in educational data, even if certain students observations were missing.	N/A	Behavioural and demographic observations.
Poudyal et al. (2020)	KNN, DT and LR ML model was used.	To validate whether hidden insights and trends in academic data can be uncovered by prediction and dimensional reduction algorithm.	The dimensional reduction algorithm and the prediction algorithm reached high accuracy for forecasting students final result.	N/A	18 features were used
Casalino et al. 2020	ARF algorithm was utilised.	To develop a model to project the academic performance of students using a comparison of the different batches. To examine the most independent variables affecting the dependent variables.	The algorithm successfully adapted and evolved its internal dimensions to incoming data.	Interaction with learning materials - quiz and outcollaborate	Behavioural and demographic characteristics.

Author	Methodology	Research Aims	Results	Dominant factors	Featured used
Qiu et al. (2022)	BCEP prediction framework	To propose a behaviour classification based on e-learning performance BCEP prediction framework	The BCEP framework had performed well at predicting students performance in comparison to the conventional classification techniques.	Interaction with learning materials	N/A
Kuzilek et al. (2019)	The approach of expectation maximisation clustering was used to split students into VLE intensity categories and groups.	To examine and analyze the presence or absence and the depth of any connection between recorded students VLE and study outcomes.	In as much as the data did not contain all the necessary details about students academic performance, they were able to effectively analyse and note the groups of students who are already having a low performance outcome at the beginning of a course.	Students' final performance can be predicted based on their (attitude to assessments) whether they submit their first, second, and fifth assessment	Assessments
Raj et al. (2021)	Deep Learning Approach using the CNN Algorithms. The GridSearchCV hyperparameter tuning technique.	Ascertaining the relative closeness or students that are lagging have had towards online courses which decides if they will take or drop the course. To explore the fundamental reasons that determine students that were on the verge of withdrawal from their programme and then to develop a prediction model to make adjustments against future occurrences.	The algorithm was able to identify leaarners who were unlikely to complete their academic programme.	Deep learning is effective in forecasting early withdrawals of students.	Sum_clicks, activity_type, highest education, count and score
Kuzilek et al. (2015)	Time-series sequential method and LSTM	To develop a system for classifying students' academic achievement levels.	The LSTM model surpassed the previous model which was able to detect 90% performance higher than the pass/fail method.	Student activities	Courses, demographic features

Author	Methodology	Research Aims	Results	Dominant factors	Featured used
Aljohani, Fayoumi and Hassan (2019)	Bayesian approach to select the most germane activity in the VLE.	To identify and predict vulnerable students who are susceptible to failure early in their courses by using demographic and students' interactions regarding their course selection and performance.	The probability of failure changed when augmented with VLE attributes.	Students' assessment score and student activities	Behavioural and academic factors

Definition of abbreviations: RF = Random Forest; **DRF** = Distributed Random Forest; **GBM** = Gradient Boosting Machine; **DL** = Deep Learning; **AUC** = Area under the ROC curve; **ROC** = Receiver Operating Characteristic Curve; **CNN** = Convolutional neural network; **1D** = Kernel moves in 1 direction; **2D** = Kernel moves in 2 directions; **KNN** = K-Nearest Neighbour; **ANN** = Artificial Neural Network; **SVM** = Support-Vector Machines; **SVC** = Sector Vector Classifier; **ARF** = Adaptive Random Forest; **DT** = Decision Tree; **LR** =Logistic Regression; **FSCNCA** = Feature Selection for Classification using Neighbourhood component analysis; **DISSFCM** = Dynamic Incremental Semi-Supervised Fuzzy C-Mean Algorithm; **BCEP** = Behaviour Classification Performance Framework; **LSTM** = Long Short-term memory.

From table 2 above, it can be deduced that many scholars applied various ML techniques to the OULAD to predict students' performance. Additionally, some studies compared different models to know which one performed better.

In addition to the findings in the table, Tomasevic, Gvozdenovic and Vranes (2020) stated that students' demographic data did not have a substantial impact on the prediction of performance. Al-Azawei and Al-Masoudy (2020) studied three features: behavioural, demographic, and performance. These features predict students' academic achievements based on multi-time periods in a VLE. The research indicated that demographic and behavioural features significantly predicted students' performance. The study also confirmed that the most important demographic factors are the level of the students' educational attainment before enrolling in the course and the level of financial and service stability based on students' area (Al-Azawei and Al-Masoudy 2020). Furthermore, Adam et al. (2020) claimed that demographic characteristics, behavioural factors, and past course scores might be used to predict students' academic success.

Some publications discussed the drawbacks and difficulties associated with these forecasts. Yousafzai et al. (2021) and Kaliisa, Mørch and Kluges (2019) established that some major limitations and challenges associated with the prediction of students' academic performance are the irrelevant information that comes with this dataset as well as the poor variables for prediction and inability to ascertain real-time learning curve. However, researchers suggest that machine learning should be used with learning analytics to address these challenges. This can focus on feature representation techniques, finding a balance between learning theory and computational measurement followed by a classifier and combining the epistemology and educational studies (Alshabandar et al. 2020; Ellis, Han and Pardo 2017). With appropriate ability to use, it would help them formulate policies for maintaining optimum teaching methods.

According to Bilal et al. (2022), when using demographic, personality traits, economic status, and location data, there is an overload of factors that interfere with one another when used to predict students' academic performance. As a result, it can be challenging to create a predictive model suitable for choosing the best factor(s) to consider. Prediction is receiving greater focus; thus, this study will apply EDA methodologies and statistical approaches to pinpoint the variables that influence student performance and the potency of these variables. This is the identified gap in the literature that this study will fill. The results of this study can assist stakeholders in understanding the elements that call for action to enhance academic performance. Additionally, knowing the factors that are more important to use in the ML algorithm can be helpful for data analysts, particularly in the feature selection process.

3.0 Methodology

3.1 Introduction

This chapter described the design and systematic techniques for addressing the aim and objectives of this research. The research objectives served as a guide for selecting the proper statistical methods for the study. In addition, the chapter justified the chosen approach. The study used EDA on the OULAD to understand the variable(s) and determine their relationships using frequency tables, plots and charts. Furthermore, this study established the influence of the factors on performance results and critically examined the statistical significance of the variables that influence students' performance outcomes in a VLE. Chi-square and regression analysis were used to examine the associations and the strength of the variable(s) that influence students' performance. Applying these methods offered insight into how learning analytics can enable academics and stakeholders to understand better students' behaviour and success rate in a VLE.

3.2 Research Design

The study will examine the design approaches and research techniques used in this study by looking at various design quirks and errors, as well as strengths and weaknesses. Research methodology offers a framework for gathering reliable information on the issue being investigated to identify ways to solve the research problems. It explains the systematic strategy that is utilised in comprehensively solving research difficulties. According to Mishra and Alok (2017), "research methodology" refers to the scientific technique utilised in a research project. It describes the way data is collected and analysed.

According to Osuagwu (2020), research methods generally involve planning, collecting or gathering data, and analysing essential relationships or differences among variables. This research design will help provide insightful responses to the research objectives, and recommendations will be made based on the research outcome. The quantitative method used in this study is deductive and uses probabilistic techniques such as correlation analysis and regression models to assess theories and earlier empirical research findings (Wildemuth 2016).

3.3 Proposed Method

The proposed model aims to identify the different determinants of students' performance. The Knowledge Database Discovery (KDD) methodology, as suggested by Fayyad and his colleagues in 1996 will be adopted to provide a clear perspective for the research methodology. The first stage is the data selection, followed by the data pre-processing, data transformation stage, and data mining stage. The final stage is the interpretation and the evaluation stage (Foroughi and Luksch 2018). Please refer to figure 2 for the five stages of the KDD process. However, the OULAD dataset will be discussed under steps 1,2 and 3 because they are relevant to this chapter. Finding patterns and trends in the fourth step, which is covered in detail in chapter 4. Step five entails evaluating and interpreting the findings, and this will be discussed in chapters four and five, respectively.

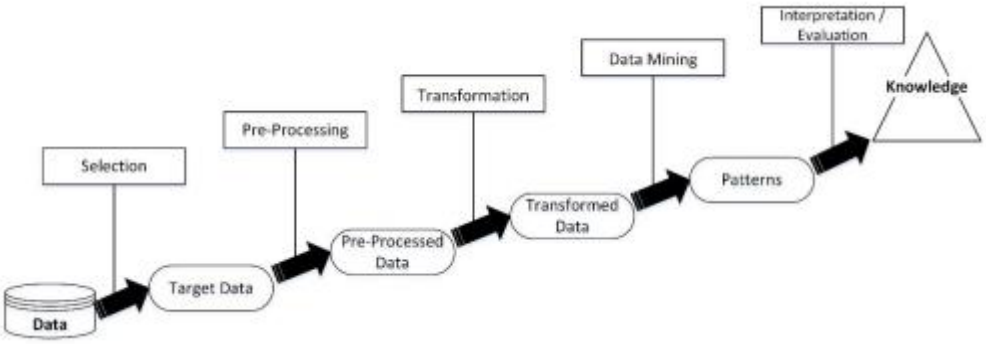

Figure 3: The KDD lifecycle (Foroughi and Luksch 2018 p.5)

According to Embarak (2018), Python is a must-have tool for data analysis. Python is a widely used high-level programming language. It provides many libraries for data accessing, pre-processing, and manipulating complex data. It is an excellent tool for exploratory data analysis, statistical analysis and visualisation of massive data (Nagpal and Gabrani 2019; Nelli 2018; Londhe and Rao 2017). Figure 3 shows the python libraries that were used for this research. For example, seaborn is a python library used for data visuals, while panda is used for data transformation and analysis. However, some scholars pointed out some disadvantages, such as speed limitation and a lot of errors at runtime, such as parse errors and indentation errors (Londhe and Rao 2017). Nevertheless, Python was selected as the primary tool for this research because of its ease of use, shallow learning curve and readability. In addition, Python has a large community of users on websites, notably Github and StackOverflow where experts collaborate, interact, share problems and answer questions (Vadlamani and Baysal 2020).

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.formula.api as sm
import statsmodels.api as sm
from scipy.stats import ttest_ind
from scipy.stats import f_oneway
from scipy.stats import chi2_contingency
import researchpy
%matplotlib inline
import seaborn as sns
import warnings
```

Figure 4: Image showing the Python libraries used in the methodology.

The methods applied in this research will be discussed under the following headings.

1. The data selection, data pre-processing and data transformation stages as introduced in 3.3 above. This stage involves data cleaning, dealing with missing values, dropping duplicate observations, and identifying possible errors and irregularities in the dataset.
2. Exploratory data analysis to obtain valid information relating to the variable and visualise the variable.
3. Statistical modelling of the data set.

3.3.1 Step one: Data selection

This stage is where the data is chosen in line with the research objectives. This stage will discuss the description of the data set, and the data set selected and the criteria and rationale behind the choice. The Open University in the UK offers courses online at both undergraduate and postgraduate levels. The OULAD was obtained online from Open Learning Analytics | OU Analyse | Knowledge Media Institute | The Open University. The OULAD contains a selection of the data from the students who took courses between 2013-2014. The dataset is made up of 7 CSV files which include information related to learners' demographics, registrations, assessments, and VLE interactions. Figure 4 below shows the seven tables, the names of the variables in the data and how they are linked.

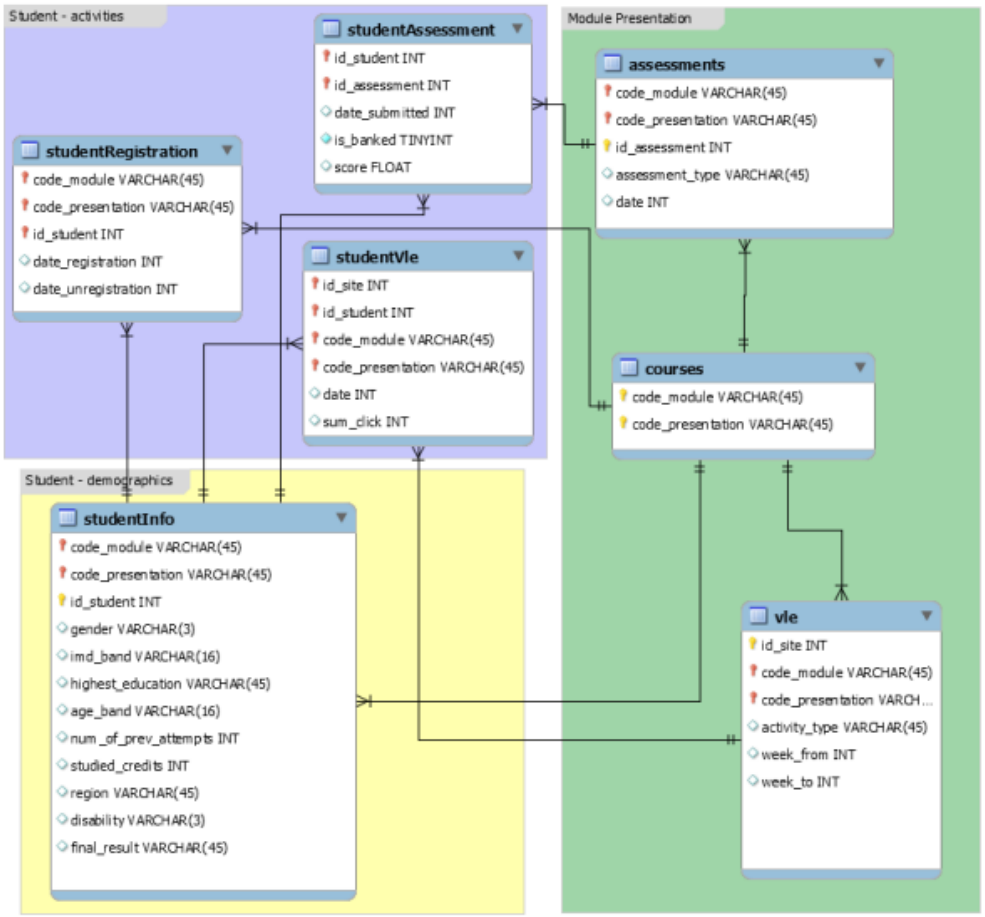

Figure 5: Database schema of the OULAD (Poudyal et al. p6 2020).

Generally, the studentInfo table contains mainly the students' demographic attributes. The studentRegistration, studentVle, and the studentAssessment tables contain information related to the student's activities. While, the assessment, courses and vle table include information related to the module presentation. Table 3 below shows the summary of OULAD dataset. The summary captures the total number of rows, the dataset description and their attributes. Please refer to Appendix 1 to 7 for a more detailed description of each attribute of the dataset, for instance, the description of each feature and the data type.

Table 3: Summary of the OULAD (Jha et al. p.159 2019).

Data File	No of Records	Description	Attributes
courses	22	Information about the courses	code_module, code_presentation, module_presentation_length
studentInfo	32593	Contains demographic information about the student	code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_result
studentRegistration	32593	Registration of the student for a course presentation	code_module, code_presentation, id_student, date_registration, date_unregistration
assessments	196	Assessments for every course presentation	code_module, code_presentation, id_assessment, assessment_type, date, weight
studentAssessments	173740	Assessments submitted by the students	id_assessment, id_student, date_submitted, is_banked, score
vle	6365	Online learning resources and materials	id_site, code_module, code_presentation, activity_type, week_from, week_to
studentVle	1048575	Student interaction with the VLE resources	code_module, code_presentation, id_student, id_site, date, sum_click

Initially, all the tables appeared to have essential variables and were selected to answer the research objectives. The studentinfo table captured student demographic, academic information, and final result. The studentAssessment table contained the students' scores. Also, the courses table had information on the course modules, while the studentVle showed students' engagement through the number of clicks.

3.3.2 Step two: Data pre-processing

In this stage, the data set is examined to find anomalies, missing numbers, noisy data and inconsistencies. This stage is crucial because it makes the dataset accurate and usable. Each table selected in the previous stage was scanned for missing values. Refer to appendix 8 for details. The fields with up to 50% of missing values were dropped, for instance, week_to, week_from and date unregistered.

Some of the students' IDs were duplicated in the studentInfo table. This was because some students had attempted the course previously. The student details that captured the last attempt were used to handle this duplication. In other words, all the information containing the previous attempts at the course was dropped.

3.3.3 Step three: Data Transformation

This stage involves modification to the variables that will aid the objectives. This includes data reduction and aggregation. From table 3, it is evident that the stuVle file had the highest number of records. The file initially captured the daily clicks of students. The table was compressed for ease of understanding by changing the daily clicks to reflect each student's total number of clicks during each module.

The target variable has four attributes: pass, fail, distinction and withdraw. We dropped the withdrawal field. The research will focus only on the students who completed the programme and received a grade. In other words, students who withdrew were dropped from the selection. It is worth noting that the dataset does not contain the reasons for withdrawal.

The selected tables (studentInfo, course, studentAssessment, stuVle and assessment) were merged to have a consolidated table with the necessary variables. To achieve this, we referred to figure 4, which shows the unique identifiers and how each table is linked. All observations that did not meet that condition were dropped.

Finally, we arrived at the final merged dataset that contained students' demographic factors, average scores and the total number of clicks for each module.

3.4 Educational Data Analysis (EDA)

According to Sahoo et al. (2019), EDA allows data to be represented in rows and columns. It is a form of analysis that present complex data in a more understandable format. It facilitates data analysis. EDA emphasises the graphical and tabular representation of statistical measures such as measures of central tendency, spread, the shape of the data and outliers. Furthermore, these researchers highlighted how certain data types and structures are better suited for specific representations. Histograms, for example, can represent continuous data, whereas box plots help identify outliers. A bivariate graphical EDA (GEDA) can show associations between two variables, whereas a multivariate GEDA can show associations between more than two variables (Patel et al. 2022; Milo and Somech 2020; Sahoo et al. 2019). Sahoo et al. (2019), emphasised that EDA can detect mistakes, provide insights and test assumptions. This research will use various charts and plots to reveal anomalies, patterns and trends in the dataset.

3.5 Statistical Techniques

A multi-stage method was carried out because of the scale of measurements of the variables; the dependent variable which is the final result for the categorical case and the scores for the continuous case since the final results were based on the score. The independent variables are both categorical and continuous is indicated by the student's VLE activities. The research objectives required a combination of different stages which are explained below.

1. Chi-square
2. Cramers' V
3. Correlation Analysis
4. Regression Analysis

In figure 1 in chapter 2, the nature of variables in the OULAD dataset are categorical and continuous which has one dependent variable as performance (final grade or score) against multiple independent variables (module, presentation, gender, region, highest education, age band, disability and assessment type).

In addition to using statistical analysis, according to Mertler and Reinhart (2017), there are assumptions which must be adhered to justify statistical method can be used. Some of the assumptions are:

1. Dependent variable is to be measured on a continuous scale using interval or ratio scale of measurement. The OULAD dependent variable is a 3-point ordinal and categorical scale of measurement variable indicated by pass, fail and distinction dataset, hence it applies to this assumption rule.
2. There must be two or more independent variables which may either be continuous or categorical. There are nine independent variables being checked against the performance outcome in the OULAD. Some of them are gender, region, highest education, age band etc.
3. The researcher must ensure that there is no significant outlier. Since the variables are categorical, there was no significant outlier observed in the pre analysis step; data cleaning stage.
4. It is essential to have independence of observations. The dataset has been anonymized and the data was obtained on different aspects of the independent variables such as course modules, regions, education level etc.

It is of utmost importance that there should be a linear relationship between the dependent and the independent variables it the performance and the following independent variables.

3.4.1 Chi-square

This is a parametric and inferential test used in statistical analysis. It was used because it allowed the researcher to evaluate two different possibilities using relationship between two or more categorical variables (Pagano 2013). Also, it allows the researcher the opportunity to compare the observed and expected frequency outcomes. This is represented as χ^2. A null hypotheses was written to express the notion that there no difference between the factors that affect student's performance in a VLE. To determine if the statement made (the null hypothesis), some statistical test were done to either accept or reject the statement made about the population (students in the VLE). The chi-square statistic is given by:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

Where,

O_i is the observed value

E_i is the expected value
χ^2 is chi-square

3.4.2 Cramer V

This will be used to determine the strengths of the association between two nominal variables after testing with the chi-square (Akoglu 2018). There is a minimal difference with correlation because Cramer V gives the exact level of strength or weakness between associated variables (Magdum 2022). Cramer V was chosen because the variables have two or more unique values in each of the category. It is also calculated from the chi-square results

$$.V = \sqrt{\frac{\chi^2}{n(q-1)}} \qquad (2)$$

Where q is the smaller number of row or column.

Table 4: Showing the degree of freedom (Zach, 2022).

Degrees of freedom	Small	Medium	Large
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29
4	0.05	0.15	0.25
5	0.04	0.13	0.22

$V \leq 0.2$ The strength of the relationship is weak
$0.2 < V \leq 0.6$. The strength of the relationship is moderate
$V > 0.6$ The strength of the relationship is strong
The Cramer V statistics and the minimum degree of freedom which is (c—1, r—1) is used to determine the strength of this relationship given then table above.

Where C= the number of columns
r= number of rows

3.4.3 Correlation Analysis

This is used to determine the existence of a relationship between units or multiple independent or dependent variables whose relationship are not static (Holmes and Illowsky Dean 2017). This is indicated as 'r' and its determinant varies between -1 and +1. The negative and positive connotes the direction which is either positive or negative relationship.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Where $n\sum xy - \sum x\sum y$ is the covariance of x and$y$, $(n\sum y^2 - (\sum y)^2)$ is the variance of y and $(n\sum y^2 - (\sum y)^2)$ is the variance of x

3.4.4 Regression

This was used to predict the value of a dependent variable based on the value of the independent variable (Runker 2020). The multiple regression will be adopted because of the number of independent variable being measured against the dependent variables.

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where ϵ is the error term, β_0 is the intercept and β_1 is the slope and

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

n is the number of observation, y is the dependent variable and x is the independent variable.

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

Where $\overline{x}$ is the mean of x and $\overline{y}$ is the mean of y.

3.4.5 Summary

A conventional data analysis project includes gathering data, pre-processing, analysis, and visualisation. Afterwards, the findings are interpreted. (Runkler, 2016; Magdum, 2022). This research follows this specified procedure. In addition, statistical analysis was conducted to investigate factors and the relationships that determined students' performance in a VLE.

The methodology answered the research questions. Objective two aims to explore the OULAD dataset using EDA techniques. This was done using cross-tabulation followed by a visual presentation using charts and plots. Objective three was achieved using chi-square test to test the variables against performance. The analysis used the Pearson chi-squared value, the p-value, and Cramer's V value. However, the p-value was used to test the significance level of relationships between variables. The fourth objective was achieved using the ordinary least square regression models. The findings in the study will be the basis for proposing strategies administrators can use to improve outcomes.

4.0 Analysis of result

4.1 Introduction

This chapter discusses the results and observations from the findings. The results were presented according to the research objectives, and conclusions were made from applying EDA and statistical methods.

4.2 Results

The results were deduced from the final dataset, which contained information on students' learning behavior, performance, and demographics. The merged dataset contained 10308 rows and 25 columns. Table 4 below captures all the variables used for this research and their frequencies and percentages.

Table 5: Showing the count and percentage of the variable.

Variables	Frequencies	Percentage
Final result/ performance		
Distinction	1860	(18.0%)
Fail	875	(8.5%)
Pass	7573	(73.5%)
Module		
AAA	476	(4.6%)
BBB	2412	(23.4%)
CCC	1027	(10.0%)
DDD	2322	(22.5%)
EEE	651	(6.3%)
FFF	2620	(25.4%)
GGG	800	(7.8%)
Presentation		
2013B	1424	(13.8%)
2013J	2820	(27.4%)
2014B	2077	(20.1%)
2014J	3987	(38.7%)
Gender		

Female	4897	(47.5%)
Male	5411	(52.5%)
Region		
East Anglian Region	1097	(10.6%)
East Midlands Region	717	(7.0%)
Ireland	457	(4.4%)
London Region	946	(9.2%)
North Region	592	(5.7%)
North Western Region	814	(7.9%)
Scotland	1047	(10.2%)
South East Region	725	(7.0%)
South Region	1063	(10.3%)
South West Region	778	(7.5%)
Wales	674	(6.5%)
West Midlands Region	764	(7.4%)
Yorkshire Region	634	(6.2%)
Highest Education		
A Level or Equivalent	4866	(47.2%)
HE Qualification	1719	(16.7%)
Lower Than A Level	3538	(34.3%)
No Formal qualification	70	(0.7%)
Post Graduate Qualification	115	(1.1%)
Age Band		
Less than 35 years	7009	(68.0%)
55 years and above	84	(0.8%)
Between 35 years and 55 years	3215	(31.2%)
Disability		
No	9487	(92.0%)
Yes	821	(8.0%)
Assessment Type		
CMA	5156	(50.0%)
Exam	2732	(26.5%)
TMA	2420	(23.5%)

4.2.1 Exploratory Data Analysis

Pie chart distribution of Final Grade

The pie chart in figure 5 shows the distribution of the final result, which is the dependent variable in the dataset. From the chart, it is evident that over 90% of the learners scaled through their course modules.

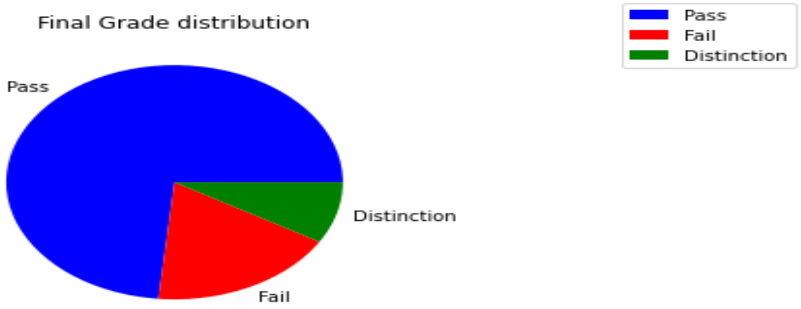

Figure 6: A pie chart showing the distribution of final grade.

Code Module of Courses

Figure 6 shows that more students took FFF, GGG and BBB over AAA, EEE and GGG. This depicts that more students took more social science courses. Please refer to appendix 9 for the description of the course modules.

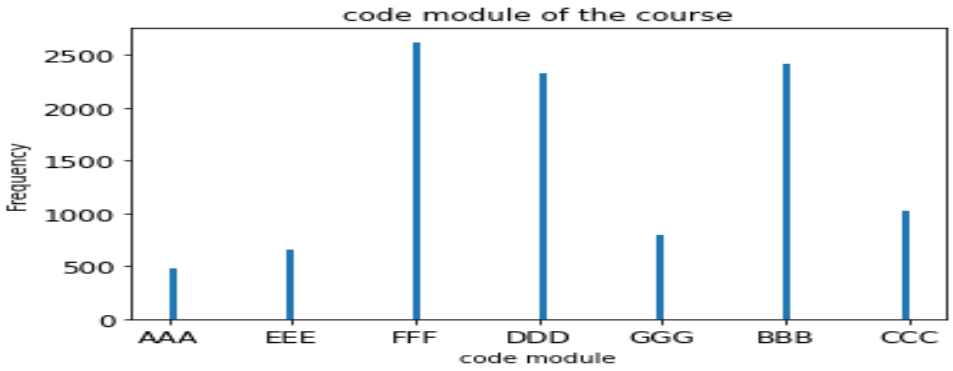

Figure 7: A chart showing the frequency of the code module of the course.

Assessment Type for Students

The assessment type comprises Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Exam. The pie chart in figure 7 shows that about half of the assessments in the Open University were Computer Marked Assessments. This number might be attributed to the nature of learning (virtual learning) at the university.

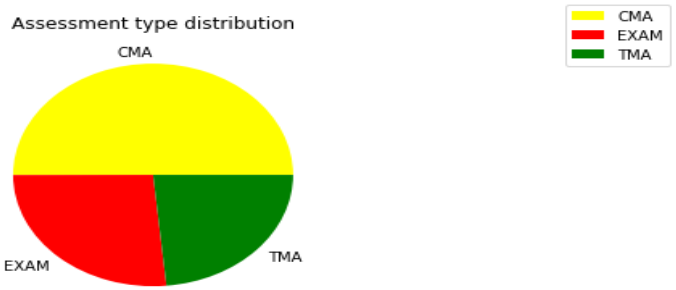

Figure 8: A pie chart showing the distribution of assessment type.

Code Presentation Distribution

The combination of 2013J and 2014J represents the code presentation of over half of the students. This means that in 2013 and 2014, more students started their course modules in November compared to those who began in February. However, more students enrolled in 2014 compared to 2013, as seen in figure 8.

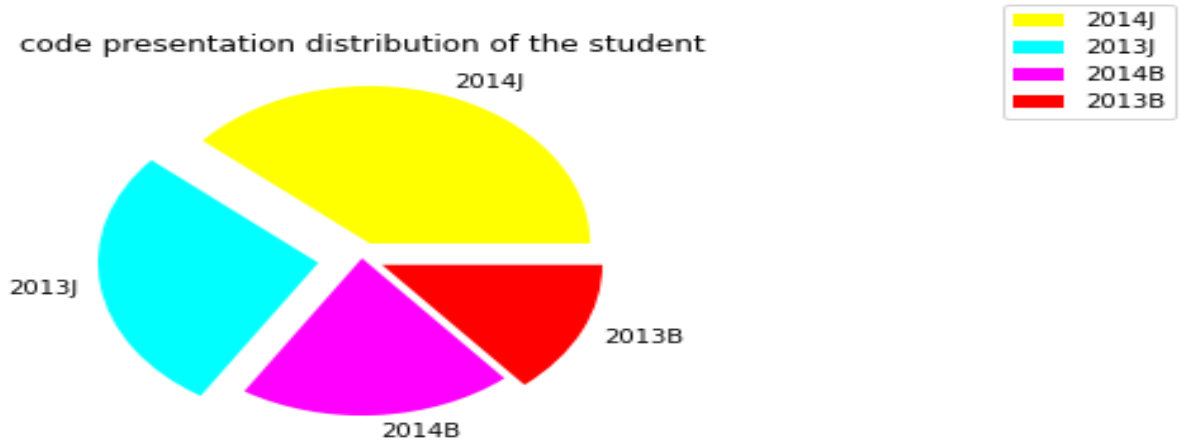

Figure 9: Code Presentation of Students.

Gender Distribution of Students

As depicted in figure 9, the difference between male students and female students that enrolled in 2013 and 2014 was not a considerable one. The difference was 5%.

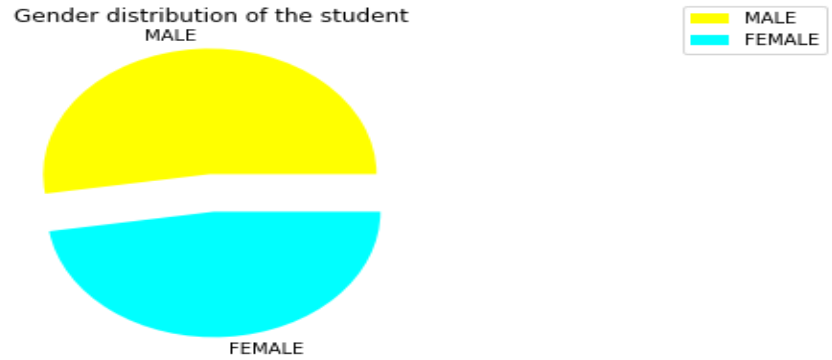

Figure 10: Pie chart showing the distribution of the gender of the students.

Region Distribution of the Students

East Anglian, South, Scotland and London region accounted for the highest number of learners. The remaining 60% of learners are connected from the other nine regions. This is shown in figure 10.
sam

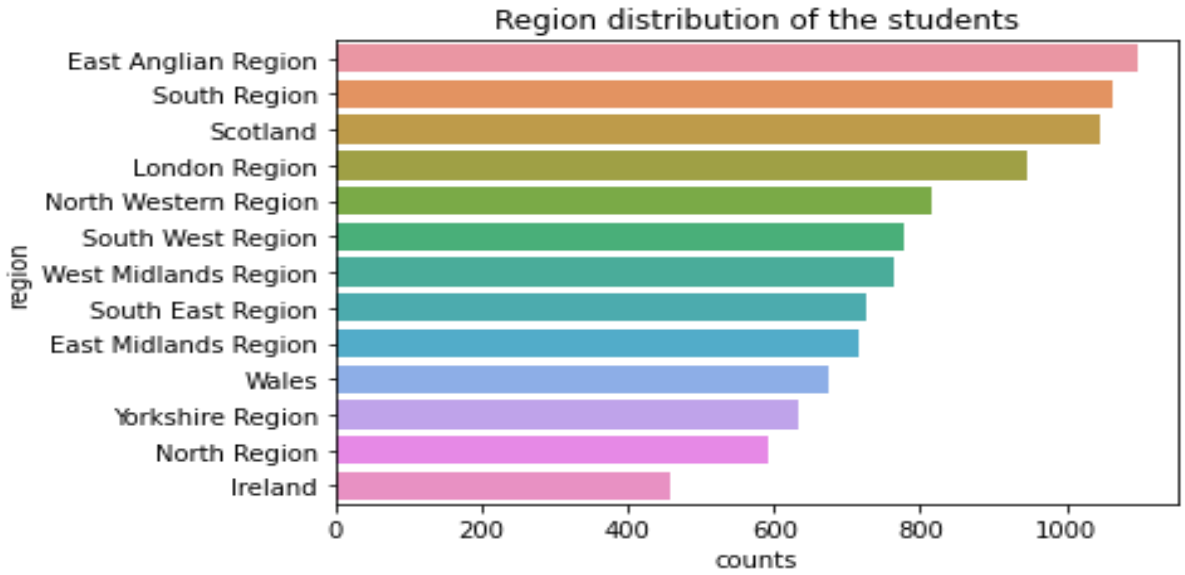

Figure 11: Bar chart showing the distribution of students across regions.

Age Distribution of the Students

Over half of the class were younger than 35 years. This shows that the population of OU students are composed of young accounts. Only 0.8% of the learners were older than 55. The older population are not inclined to distance learning. The result is presented in Figure 11 below.

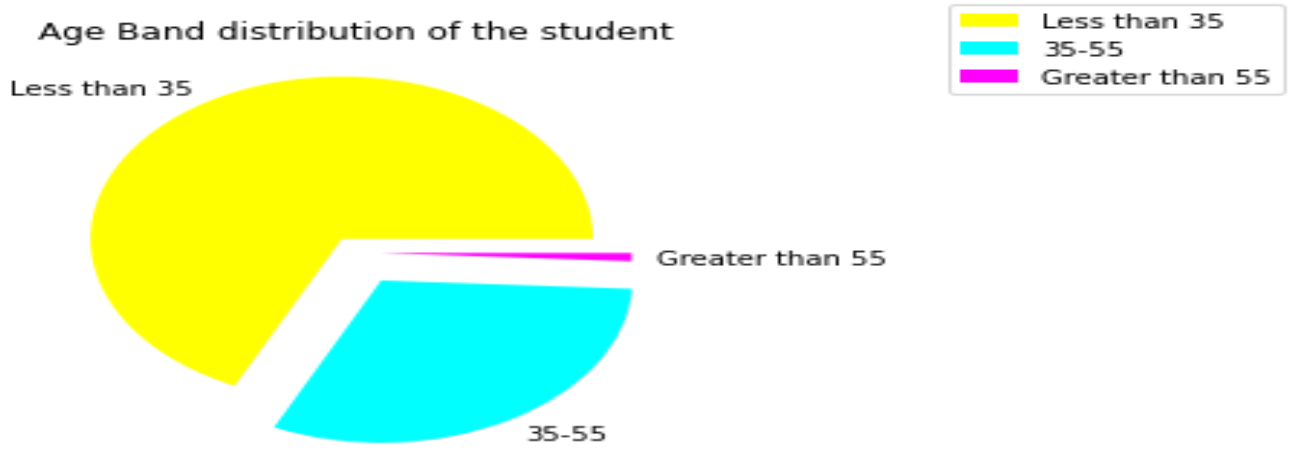

Figure 12: A pie chart showing the age distribution of students.

Disability Status of Students

As shown in figure 12, about 92% of students who took courses at the OU had no disability.

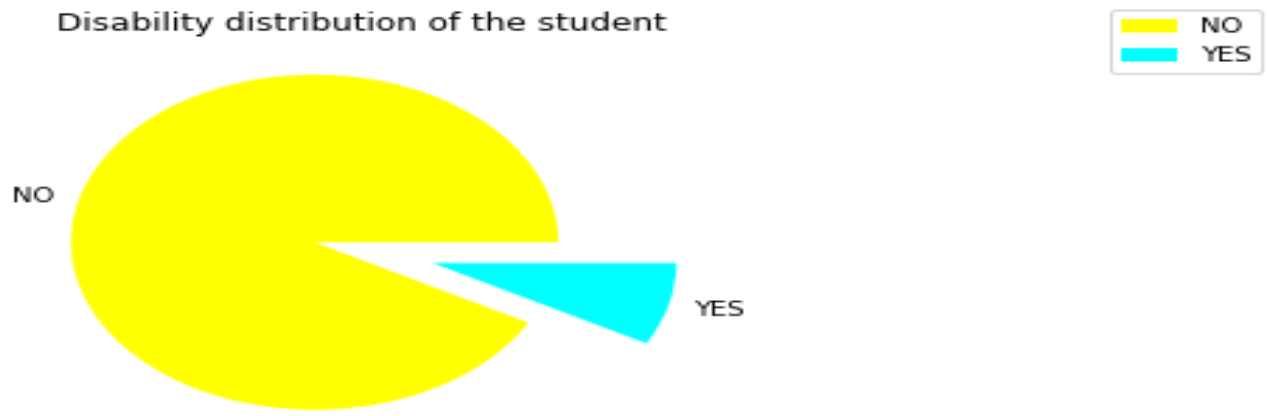

Figure 13: A pie chart showing the distribution of students with disabilities.

Level of Highest Education Distribution

Figure 13 shows that almost half of the students had a minimum qualification of A Level or equivalent. However, there was only a difference of 0.4% between learners with no formal qualification and learners with the highest possible qualification (Post Graduate Qualification).

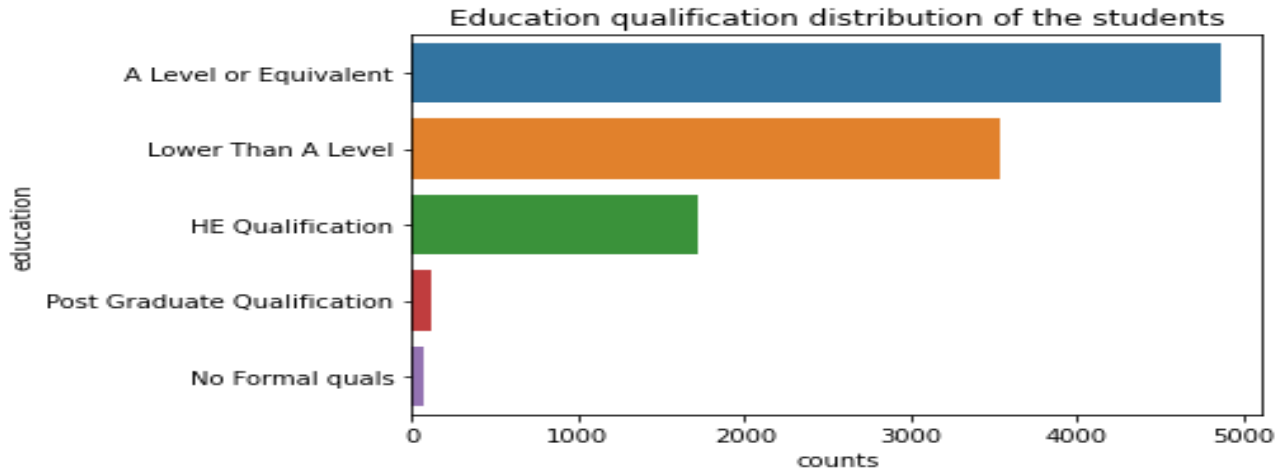

Figure 14: : A bar chart showing the highest educational qualification of students.

Sum of Clicks and Performance

The average clicks of those that had distinction are higher than those that had a pass. Those that failed recorded the least number of sum clicks. Based on the distribution, it can be said that the more students click, the better their performance in a VLE, as seen in figure 14.

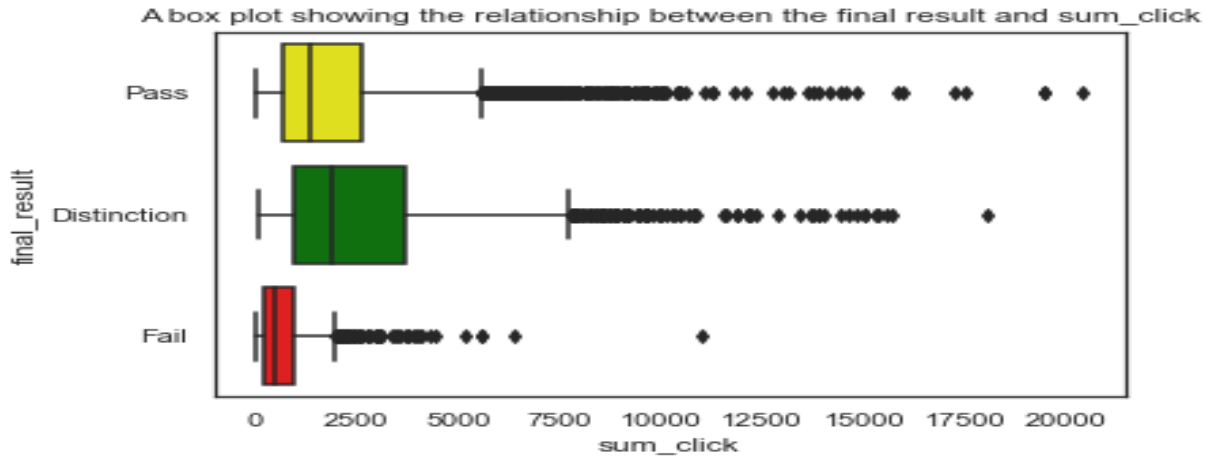

Figure 15: A box plot showing the relationship between sum clicks and final result.

Gender and Performance

The boxplot below shows the summary of the scores of students by gender. The is a slight difference of 1.31 when comparing their average scores, as shown in see figure 15.

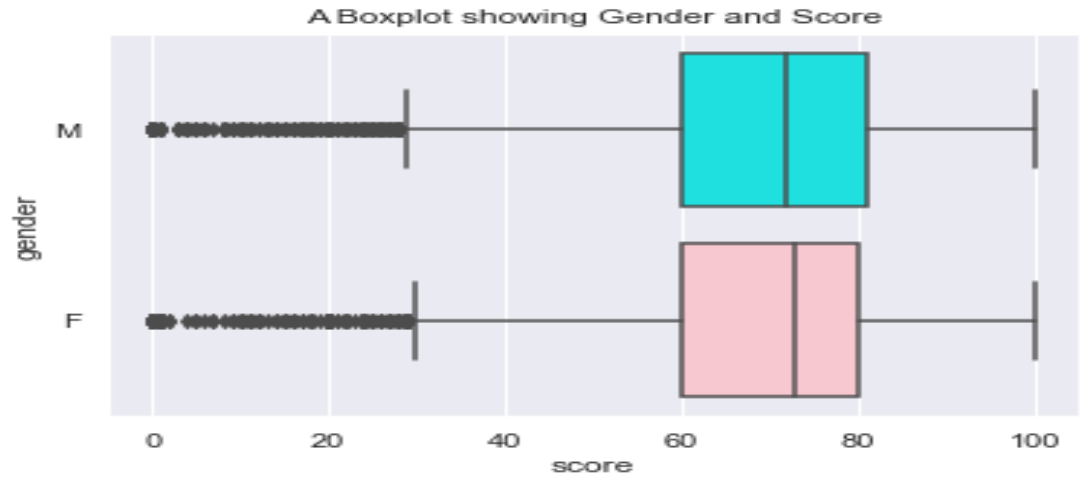

Figure 16: A box plot showing the relationship between gender and score.

4.2.2 Statistical Techniques

Table 16 below shows the variables and some statistical measures such as count, percentages, chi-square and crammer's V statistic. These will be explained further in the sub-sections. Performance and final results are used interchangeably and mean the same thing.

Table 6: Table showing the comparison of variables to performance.

Variable	**Performance**			**Chi square**	**Crammer's V**
	Distinction	Fail	Pass		
Module				922.99 (0.000)	0.21
AAA	27 (0.3%)	14 (0.1%)	435 (4.2%)		
BBB	476 (4.6%)	99 (1.0%)	1837 (17.8%)		
CCC	306 (3.0%)	172 (1.7%)	549 (5.3%)		
DDD	238 (2.3%)	415 (4.0%)	1669 (16.2%)		
EEE	176 (1.7%)	16 (0.2%)	459 (4.5%)		
FFF	382 (3.7%)	133 (1.3%)	2105 (20.4%)		

GGG	255 (2.5%)	26 (0.3%)	519 (5.0%)		
Presentation				58.6 (0.000)	0.05
2013B	211 (2.0%)	132 (1.3%)	1081 (10.5%)		
2013J	447 (4.3%)	207 (2.0%)	2166 (21.0%)		
2014B	460 (4.5%)	202 (2.0%)	1415 (13.7%)		
2014J	742 (7.2%)	334 (3.2%)	2911 (28.2%)		
Gender				30.8 (0.000)	0.06
Female	938 (9.1%)	343 (3.3%)	3616 (35.1%)		
Male	922 (8.9%)	532 (5.2%)	3957 (38.4%)		
Region				87.3 (0.000)	0.07
East Anglian Region	194 (1.9%)	80 (0.8%)	823 (8.0%)		
East Midlands Region	116 (1.1%)	52 (0.5%)	549 (5.3%)		
Ireland	58 (0.6%)	28 (0.3%)	371 (3.6%)		
London Region	163 (1.6%)	102 (1.0%)	681 (6.6%)		
North Region	143 (1.4%)	43 (0.4%)	406 (3.9%)		
North Western Region	125 (1.2%)	83 (0.8%)	606 (5.9%)		
Scotland	214 (2.1%)	77 (0.7%)	756 (7.3%)		
South East Region	163 (1.6%)	60 (0.6%)	502 (4.9%)		
South Region	210 (2.0%)	67 (0.6%)	786 (7.6%)		
South West Region	147 (1.4%)	77 (0.7%)	554 (5.4%)		
Wales	104 (1.0%)	70 (0.7%)	500 (4.9%)		
West Midlands Region	112 (1.1%)	80 (0.8%)	572 (5.5%)		
Yorkshire Region	111 (1.1%)	56 (0.5%)	467 (4.5%)		
Highest Education				269.7 (0.000)	0.11
A Level or Equivalent	937 (9.1%)	360 (3.5%)	3569 (34.6%)		
HE Qualification	438 (4.2%)	110 (1.1%)	1171 (11.4%)		
Lower Than A Level	420 (4.1%)	391 (3.8%)	2727 (26.5%)		
No Formal qualification	9 (0.1%)	12 (0.1%)	49 (0.5%)		
Post Graduate Qualification	56 (0.5%)	2 (0.0%)	57 (0.6%)		
Age Band				106.1 (0.000)	0.07
Less than 35 years	1095 (10.6%)	663 (6.4%)	5251 (50.9%)		
55 years and above	26 (0.3%)	7 (0.1%)	51 (0.5%)		
Between 35 years and 55 years	739 (7.2%)	205 (2.0%)	2271 (22.0%)		
Disability				18.7 (0.000)	0.04
No	1741 (16.9%)	777 (7.5%)	6969 (67.6%)		
Yes	119 (1.2%)	98 (1.0%)	604 (5.9%)		
Assessment Type				393.1 (0.000)	0.14
CMA	1036 (10.1%)	189 (1.8%)	3931 (38.1%)		
Exam	514 (5.0%)	290 (2.8%)	1928 (18.7%)		

TMA	310 (3.0%)	396 (3.8%)	1714 (16.6%)		

Relationship between Code Module and Performance

Based on the result in table 6, there is a significant relationship between the code module and performance. ($\chi^2 = 923, p < 0.05$). The Cramer's V statistic is 0.21, and the minimum degrees of freedom = 2, indicating a moderate relationship between code module and performance.

The relationship between the code module and final result is depicted in figure 16. BBB recorded a high number of passes and a relatively lower number of failures compared to module DDD which had the second highest number of pass and a relatively high failure rate. While the difference in the number of passes for AAA and EEE was low (0.3%), the difference in the number of distinctions was higher.

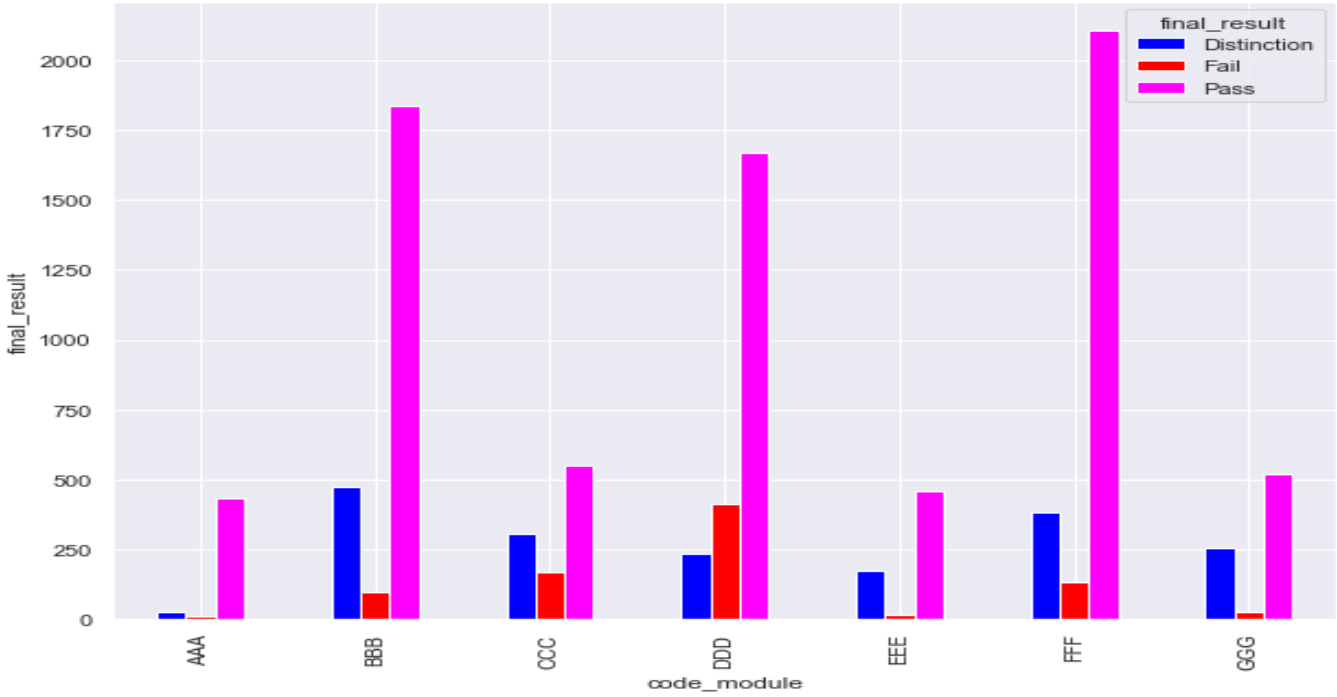

Figure 17: Bar chart showing the relationship between code module and performance.

Relationship between Code Presentation and Performance

According to table 6, there is a strong link between code presentation and performance ($\chi^2 = 58.6, p < 0.05$). The Cramer's V statistic is 0.05, with the minimum degrees of freedom equal to 2, indicating a weak relationship between the code presentation and performance.

Across all the modes of code presentation, more students had a pass in their courses, while failure was the least recorded result across all the code presentations, as seen in figure 17.

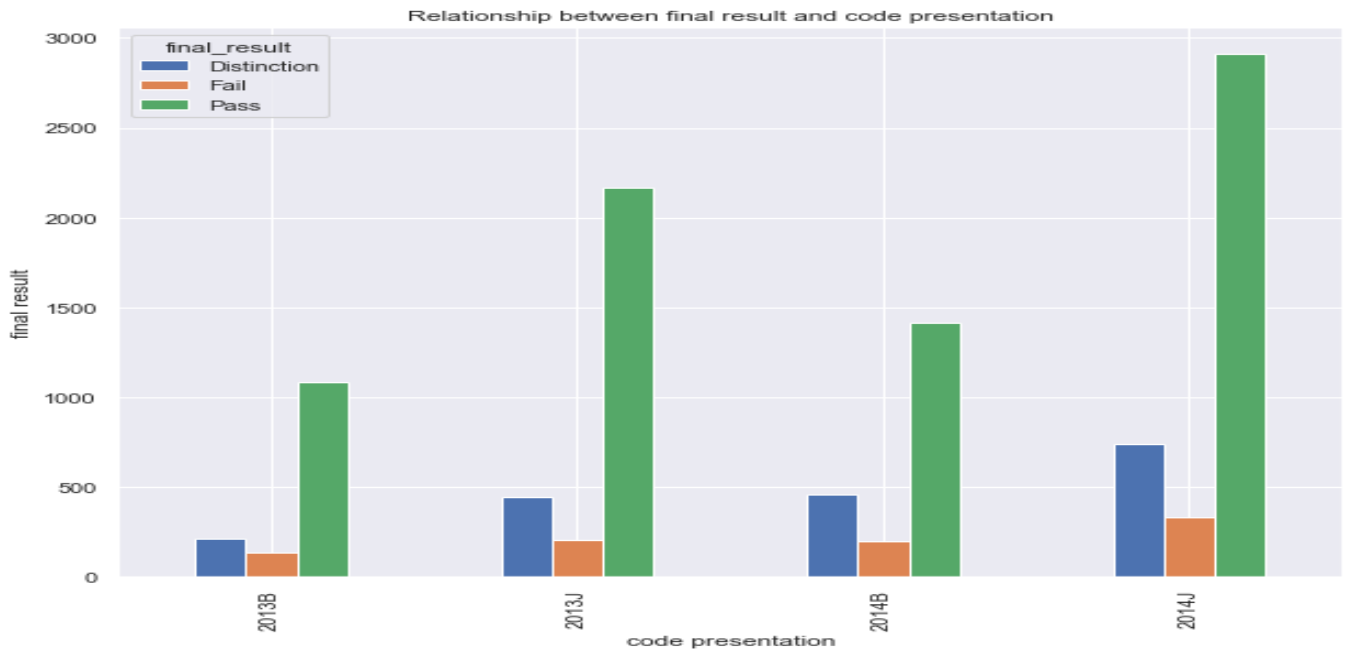

Figure 18: Bar chart showing the relationship between code presentation and performance.

Relationship between Gender and Performance

The findings in table 6 indicate a strong relationship between gender and performance ($\chi^2 = 30.8, p < 0.05$). Cramer's V statistic is 0.06, with a minimum degree of freedom of 2, indicating a weak relationship between gender and performance.

About 3.3% of males had more passes than their female counterparts. However, the likelihood of both genders having a distinction is almost equal, as seen in figure 18.

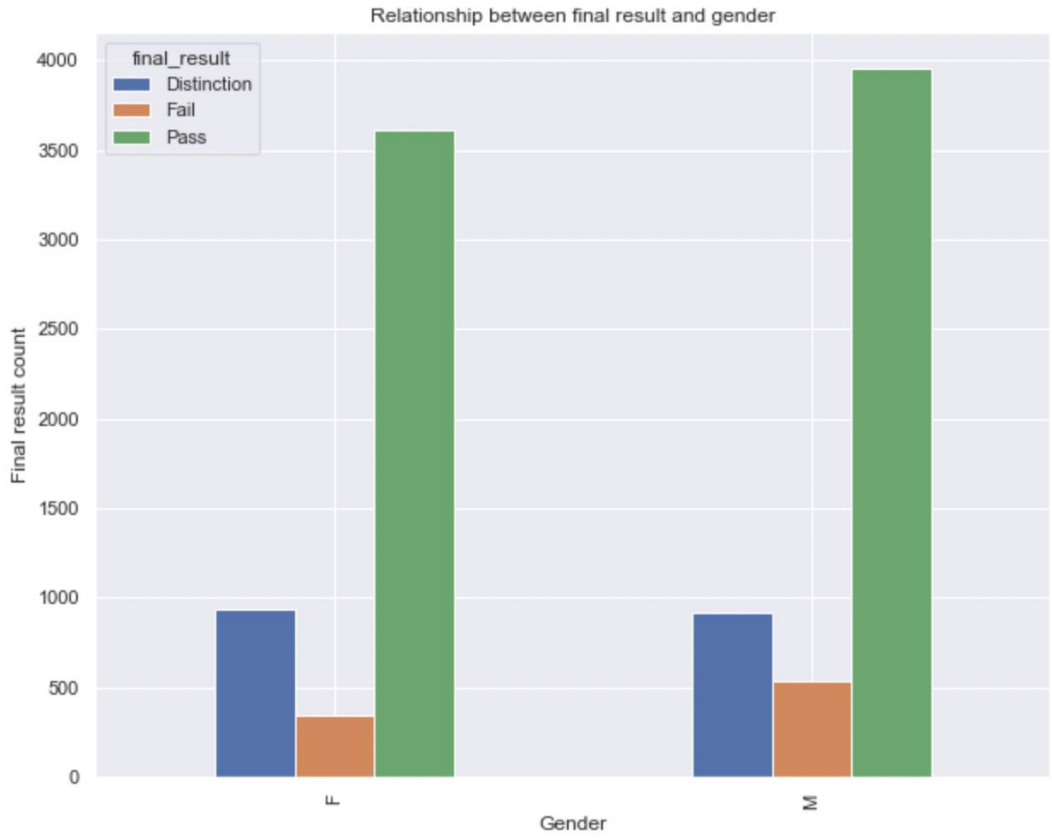

Figure 19: Bar chart showing the relationship between gender and performance.

Relationship between Region and Performance

The results in table 6 show a strong correlation between region and performance ($\chi^2 = 87.3, p < 0.05$). Cramer's V statistic is 0.07, with a degree of freedom of 2 as the minimum, indicating weak relationship between region and performance.

Across all the regions, the pattern of performance was the same. In all the regions, there was a higher number of passes, followed by distinction. Even though the East Anglian region produced the highest number of passes compared to Scotland, Scotland had more students with distinction, as seen in figure 19.

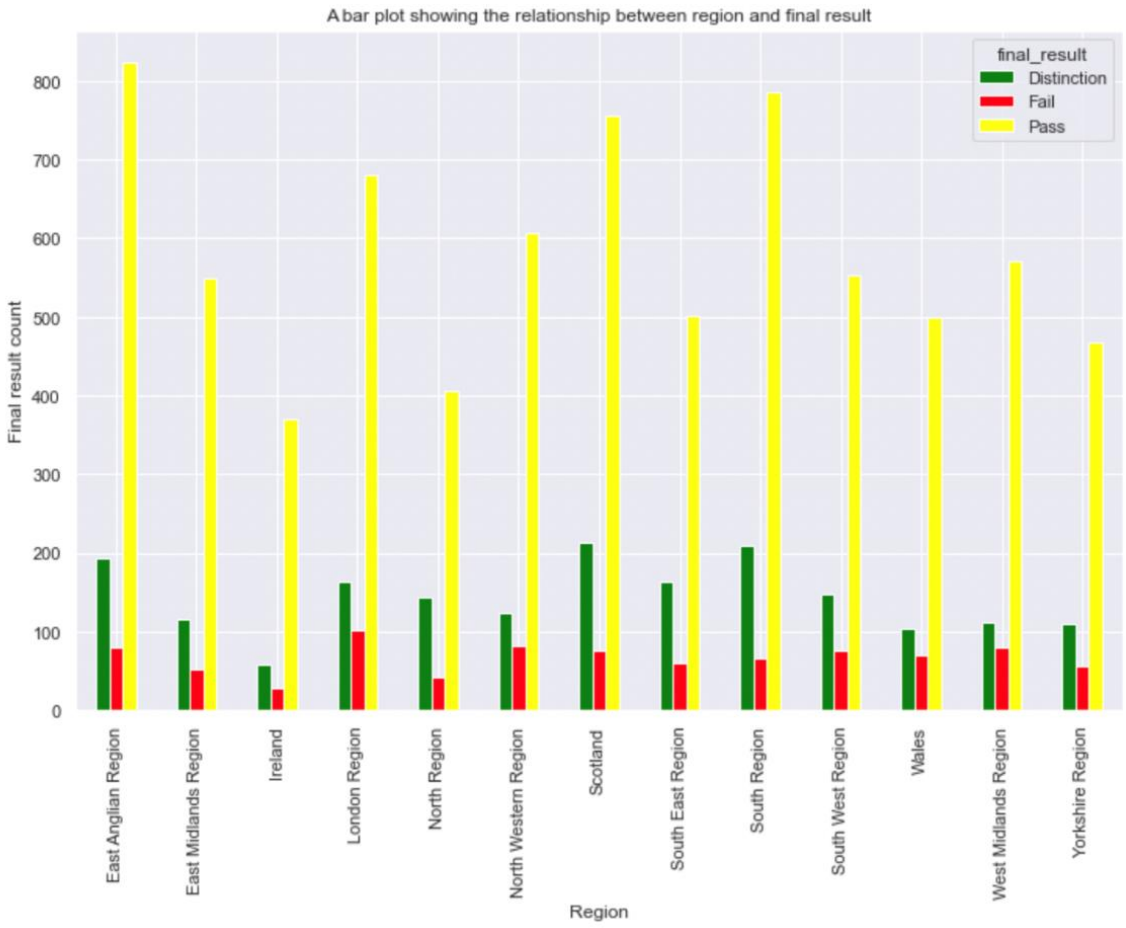

Figure 20: Bar chart showing the relationship between region and performance.

Relationship between Highest Education and Performance

Based on the result in table 6, there is a significant relationship between the highest education and performance ($\chi^2 = 269.7, p < 0.05$). The Cramer's V statistic is 0.11, and the minimum degrees of freedom = 2, indicating a weak relationship between the highest education module and performance. There are differences in the proportion of learners who pass, fail or have distinctions in their courses across different levels of education. While those with A-levels or equivalent and HE qualifications recorded a comparatively higher number of distinctions, those with lower than A-level have an almost equal number of distinctions to failure rate, as seen in figure 20.

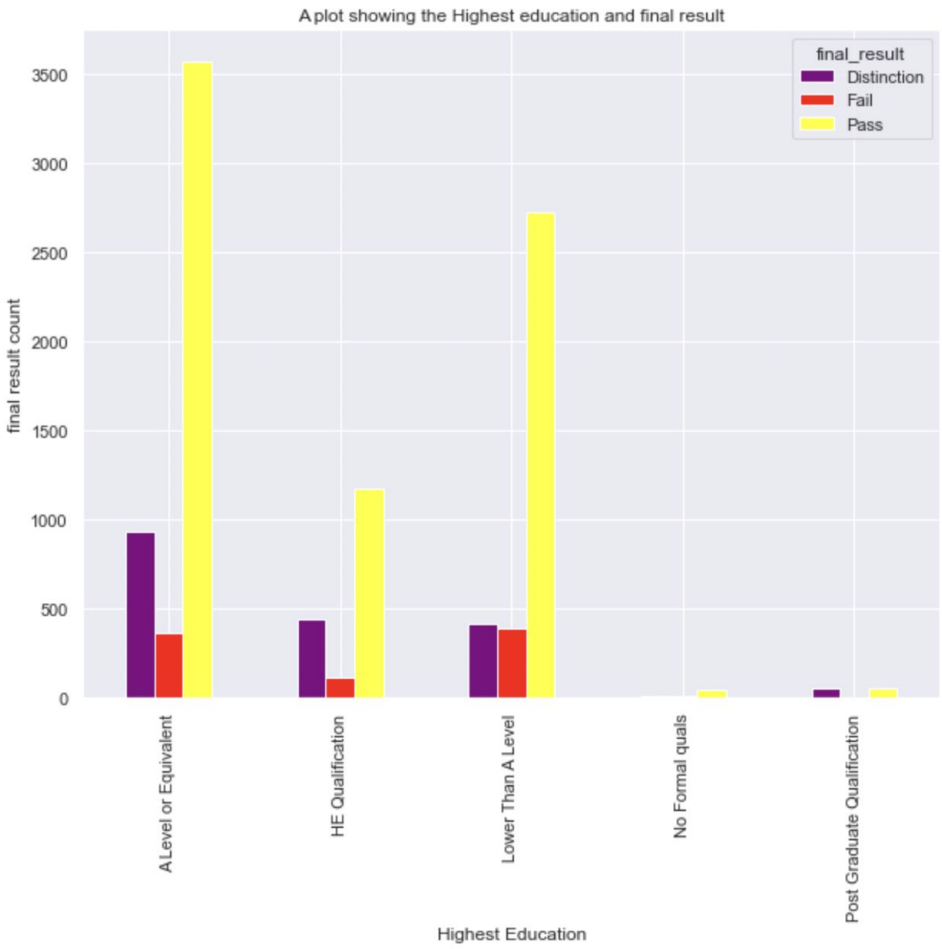

Figure 21: Bar chart showing the relationship between highest education and performance.

Disability and Performance

The findings in table 6 indicate a strong relationship between disability and performance ($\chi^2 = 18.7, p < 0.05$). Cramer's V statistic is 0.04, with a minimum degree of freedom of 1, indicating a weak relationship between disability and performance.

For every student with no disability that fails a course, there are almost twice as many students that get a distinction. However, for students with disabilities, there is an equal ratio for students having a distinction or a failure, as seen in figure 21.

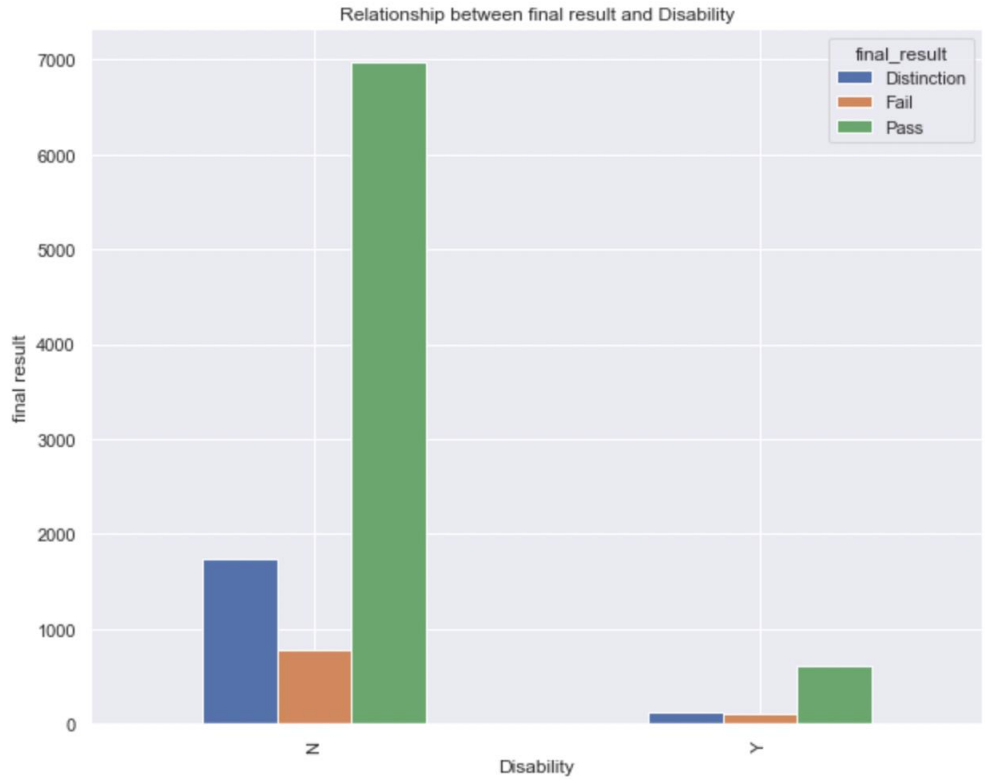

Figure 22: Bar chart showing the relationship between disability and performance.

Age and Performance

According to table 6, there is a strong link between age and performance ($\chi^2 = 106.1, p < 0.05$). The Cramer's V statistic is 0.07, with the minimum degrees of freedom equal to 2, indicating a weak relationship between age and performance.

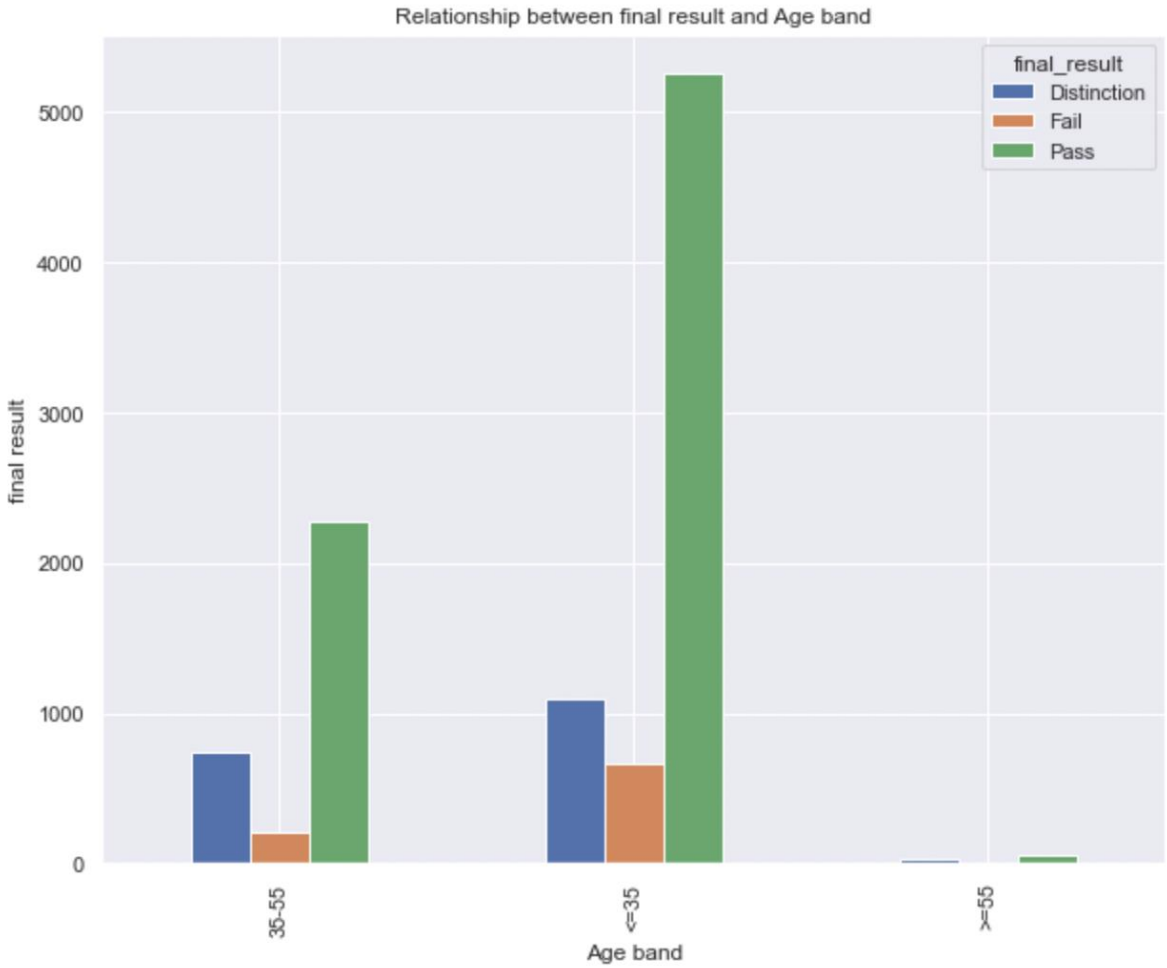

Figure 23: Bar chart showing the relationship between age and performance.

Assessment Type and Performance

The results in table 6 show a strong correlation between assessment type and performance ($\chi^2 = 393.1, p < 0.05$). Cramer's V statistic is 0.14, with a degree of freedom of 2 as the minimum, indicating a weak relationship between assessment type and performance.

As seen in figure 23, the highest pass rate and the lowest failure rate was recorded for the Computer-Marked Exam. In contrast, the lowest pass rate was recorded for the Tutor-Based Exam, regardless of a relatively high failure rate.

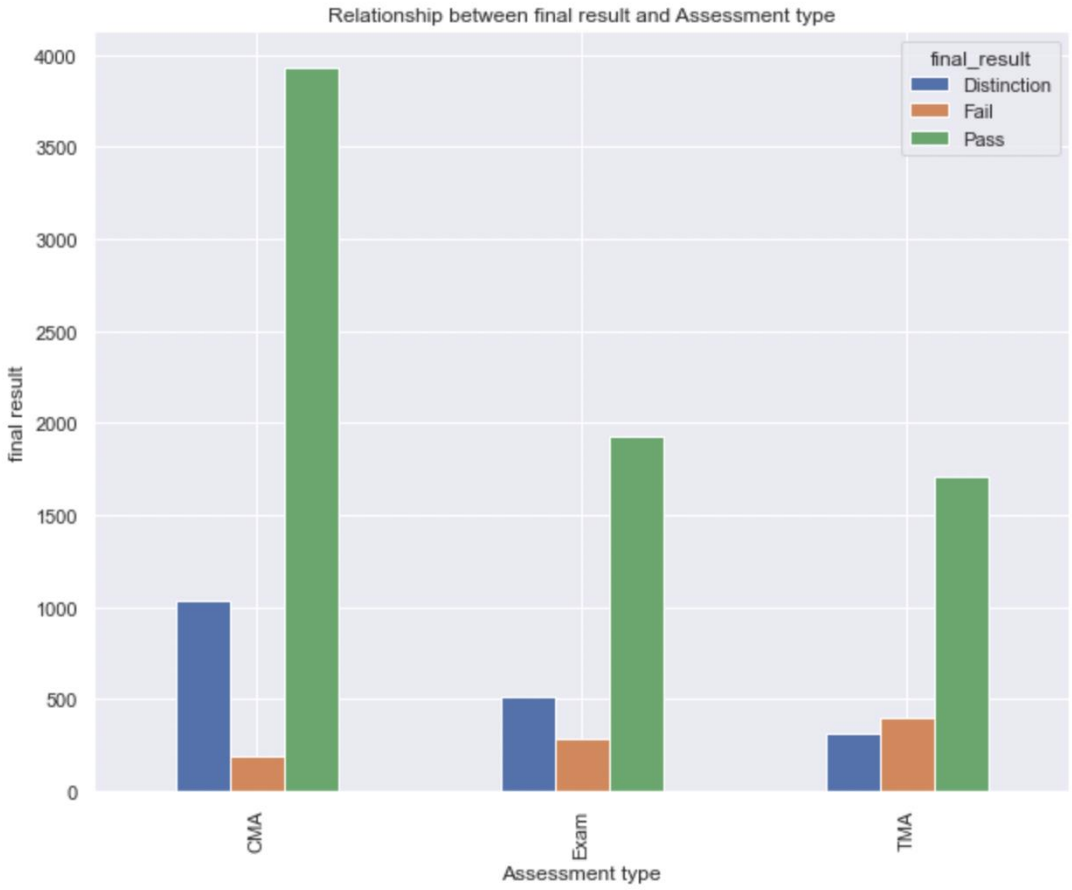

Figure 24: Bar chart showing the relationship between assessment type and performance.

4.2.3 Correlation between Variables

Table 7: Showing the correlation of scores and other numeric variables.

	Score(Depemdent Variable)	sum_click	studied_credits	num_of_prev_attempts
score	1	.253	-.050	-.110
sum_click	.253	1	.057	-.036
studied_credits	-.050	.057	1	.165
num_of_prev_attempts	-.110	-.036	.165	1

Table 7 shows the relationship between the response variable (score) and other numeric variables (sum_click, studied_credits and num_of_prev_attempts) in the table. The result showed a weak positive correlation between score and sum_clicks (0.25) which implies that as the number of clicks increases, the student scores also increase. It also shows a weak negative correlation between the scores and the number of previous attempts (-0.11); this implies that as the number of previous attempts increases, the scores of the students decrease. There is a very weak negative relationship between scores and studied credits (-0.05).

The regression model (See appendix 11,12 and 13) was used to model the relationship between the response variable (score) and the independent variables (num_of_prev_attempts , studied_credits, sum_click).

4.2.4 The Regression Model

OLS Regression between Score and Number of Previous Attempts

R-squared is the amount of variation in the score that can be accounted for by the number of previous attempts. OLS was used to test the relationship between the score and the number of previous

attempts. The results show that 12% of the variation in the score is accounted for by the number of previous attempts. The Prob (F-statistic), 0.000 shows a statistically significant influence of the number of previous attempts on scores. (See Appendix 11)

OLS Regression between Score and Studied Credit

The regression analysis showed that 3% of the variation in the scores is accounted for by studied credit. There was a statistically significant relationship between the score and the studied credits. It was also concluded that the model is a good fit since the prob (F=34.09) is less than 0.05 (See Appendix 12)

OLS Regression between Sum Click and Scores

From the regression summary, R^2 implies that 6.4% of the total variation in scores can be accounted for by sum click. It can also be seen that the model is a good fit since the p-value (p= 0.00) is less than the α value = 0.05. Hence, the test of the relationship between scores and the number of clicks is statistically significant. (See Appendix 13)

From the result of the regression model, all the independent variables have a significant relationship with performance (scores).

4.3 Discussion

This study set the objectives to examine the various factors that impact students' performance outcome in a VLE using the OU learning analytics dataset. Results obtained indicate the students with the highest number of clickstream, the higher its impact on academic performance. Other essential variables that had impact on the performance are the type of course module because the results reveal that there are more students offering course modules in the social sciences. From the OULAD, STEM based courses represented with CCC, DDD, EEE and FFF while non-STEM courses were represented with AAA, BBB, and GGG. Students passed the courses, FFF, DDD with distinction and also passed the non-STEM course BBB with distinctions. Hence, non-STEM courses AAA, and GGG had poor performance with STEM course CCC and EEE.

This finding is unreflective of Gasevic, et al. (2016) and Wessa, De Rycker, Holliday (2011), and Williams (2018) who assert that STEM based courses have higher VLE activities and cannot be adequately used to adjudge the performance outcome. The reason for the contradiction can be explained from the opinion of Gasevic et al (2016) that teaching conditions such as length of time spent in the module, assessment types, module presentations are taken at different and alternate times. This notion opposes Alves et al (2017) opinion. Their study revealed that there is correlation between courses related variables and performance.

Interestingly, results obtained revealed that factors such as location, region, gender and age had statistically significance and weak relationships on performance except for code module which had a moderate strength using the Cramers' V statistical analysis.

Codewell et al (2008) noted that demographic factors such as age, gender, nationality had impacts on performance. The study noted further that there is no correlation between students' age and their academic performance, however, analysis from this research indicated that students below 35 years had distinction while those older than 55 had the least performance in all three categories. A keen understanding of the factors in the OULA datasets helped the researcher understand that despite equal opportunities provided for the students by the university, their behavioural activities unequally affected their performance. Students with higher age and levels of qualification differed with their respective counterpart. The students with lower grades were younger than the students older and

similarly, the students with higher qualifications performed higher than the students without prior qualifications and this is in contrast agreement with the study carried out by Ahmad et al. (2021).

Performance impact obtained from the analysis by region which is indicated by the imd band, gives a detailed outcome that Scotland had most distinction result and London with failed students. East Anglian region had the most number of students with a pass. Comparing the outcome of this results with that of Dhawan (2020) who studied the learning and VLE engagement of minority groups, there is inadequate information to determine the direction and impact of minority regional groups in the OULA Datasets.

Analysing the learning outcomes of the learners in VLE gave adequate knowledge to give better advise regarding course modules as it was observed from the results that the type of course students offered and the credits ascribed had a correlative effect on their performance. Boulton, Kent and Williams (2018) notes that STEM based courses have higher dependency on VLE activity and from the results gathered during the analysis. Yunita, Santoso and Hasibuan (2021) express their opinion that students' success is determined mainly by tutor's grading procedures and policies however, the results from this analysis indicate that students had distinctions with computer assessment type than tutor marked and exam assessment.

In regards to using statistical methods to understand the various factors that affect students' performance in a VLE, it is worthy of note that past OULAD studies focused more on predictive models to improve performance outcomes and also detecting at-risk students. In the case of Aljohani, Fayoumi and Hassan (2019), the sum of clicks of students' in a VLE using a time-series was used in predicting students that had more possibilities of withdrawing from the VLE. This is an example of one among many of such prediction using ML algorithms. The results from this study pinpoints a contrary opinion to the studies in the past seeking to predict students' performance outcomes from their VLE activities. Such studies should have at-hand an in-depth understanding of these mediating demographic factors of students despite their different socio-economic and background information which could negatively impart their academic performance. This study goes all way to explain that gender, educational qualification, age band, disability status, number of previous attempts, total sum of clicks had impact on performance and learning outcomes. It means that improving student clicks, encouraging, and engaging in more clicks are avenues towards improving students' performance outcomes.

5.0 Limitation, Recommendations and Conclusion

5.1 Limitation

The limitations of the study include.

1. The availability of data was for only two years. This was a limiting factor because access to data from other years would have provided a broad scope for the research.
2. The dataset did not capture the reasons for students withdrawing from the programme. This information would have made more insightful findings on why students withdrew from the programme. For example, it might have been due to financial challenges, academic misconduct, or health issues.
3. The Open University is based in the UK. However, they have students from all over the globe. The sample dataset was limited to only learners in the UK. This limited more detailed research on how geographic locations affect students' performance.
4. Due to time constraints, this research could not apply machine learning algorithms to build predictive models' to predict students performance. Therefore, the research could not validate the claims other researchers made about some ML algorithms and compare the performance and accuracy of these models.
5. This study relied on a secondary data set that was publicly available. The validity of the data cannot be ascertained. This may result in the findings being biased.
6. Another limitation finding a current learning analytics dataset. Many academic institutions are unwilling to share information with third parties due to data protection and privacy issues.
7. The dataset captured students that took only a course module. It did not show students taking a combination of courses. Having a more robust dataset that contains course combinations of students can give insights into how course combinations affect performance.

5.2 Recommendation for further studies

There are various recommendations for further studies that can enable stakeholders to understand the learning outcomes in a virtual learning environment, particularly how the VLE can thrive in this digital age as well as provide a strong and quality alternative to a physical learning environment. These will be discussed below.

- ML experts should continue to conduct their research without overfitting the models. Research should be done by adopting algorithms that can perform more in-depth analysis based on clustering, combination, and combined impacts of variables can be carried out to expand understanding of the research work further.
- It is recommended that the Open University should collate a post-COVID dataset with the same objective. This will help to discover the effect of the digital transformation brought about by the COVID-19 pandemic. This research should bring to light strategies for ensuring student performance does not decline due to the transition to digital learning.
- Data from different geographical zones could be merged and explored to ascertain if the impact of these variables aligns across the board and establish if zones and geographic location impact learning outcomes.
- Past research focused on dataset that can be extracted from the VLE and MOOC platform. Other external factors affect students' performance. Research should be conducted to know these factors (which may be unstructured) and find how to integrate them into the data from the VLE. This will help to provide more hidden insight on factors that impact performance.
- Some studies conducted studies to compare online learning and face-to-face learning. However, this was done using different populations in different locations. Similar research can be done in an academic environment that runs both face-to-face and online learning to compare students' performance under the same conditions.

5.3 Conclusion

This study's results will help decision-makers develop policies to help students improve their success based on various behavioural and virtual learning adjustments. It was also concluded that the study will assist education communities and open universities in developing interventions that improve student performance outcomes based on teaching and learning factors in the OULAD. This will provide adequate pedagogical support in terms of virtual learning strategies. This will also assist in developing early intervention mechanisms to improve Performance.

Aljohani et.al. (2019) explained that there has been a sharp rise in people's interest in online and virtual learning. Virtual learning has become increasingly popular and accepted despite concerns about effectiveness and quality of learning. There is therefore the need to pay more attention to this mode of teaching and learning. Administrators must seek to improve quality of instructors that are involved in the learning process. Support technology that helps to improve the delivery mode of the different modules as well as the presentation of the modules which affects the learning behaviour of the student should be improved. The modules delivered should be learner-centered involving a feedback mechanism between the instructors and the student. This method improves the performance outcome as the student are largely involved and their ideas can be improved upon.

Since one major factor that encourages the student to engage in VLE is flexibility and convenience, the modules design approach should be centered around the learners such that they are encouraged to finish a module with incentives to boost their participation. Also, the VLE should be easy to navigate through and students who are struggling shouldn't be relegated. Encouraging and incentivising the finishing of modules and participation in the final exams would go a long way to improve the performance of the students.

References

ABUHLFAIA, K.M.O., 2020. Assessing the usability of virtual learning environments in higher education.

ADAM, I.Y., et al., 2020. Using Formative and Summative Assessments in Data Mining to Predict Students' Final Grades. *International Research Journal of Innovations in Engineering and Technology,* 4(11), pp. 43.

ADEDOYIN, O.B. and SOYKAN, E., 2020. Covid-19 pandemic and online learning: the challenges and opportunities. *Interactive learning environments,* , pp. 1-13.

ADEJO, O.W., 2017. Data mining and learning analytics: a multi-model approach to predicting student performance using aggregated data sources. *[online] Ethos.bl.uk. Available at: <https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.810776> [Accessed 10 August 2022].*

ADEJO, O. and CONNOLLY, T., 2017a. Learning analytics in a shared-network educational environment: Ethical issues and countermeasures. *Learning,* 8(4), pp. 156-163.

ADEJO, O. and CONNOLLY, T., 2017b. Learning Analytics in Higher Education Development: A Roadmap. *Journal of Education and Practice,* 8(15), pp. 156-163.

ADNAN, M. and ANWAR, K., 2020. Online Learning amid the COVID-19 Pandemic: Students' Perspectives. *Online Submission,* 2(1), pp. 45-51.

ADNAN, M., et al., 2022. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Computer Science,* 8, pp. E803.

ADU-MANU, K. and ARTHUR, J., 2013. Analysis of Data Cleansing Approaches regarding Dirty Data A Comparative Study. International Journal of Computer Applications. 76. 14-18. 10.5120/13258-0736.

AGUDO, A., HERNANDEZ, A. and IGLESIAS, S., 2012. Predicting academic performance with learning analytics in virtual learning environments: a comparative study of three interaction classifications. *IEEE Xplore, digital library,*

AGUDO-PEREGRINA, ÁF., et al., 2014. Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior,* 31, pp. 542-550.

AHMAD, N., et al. , 2021. Students' performance prediction using artificial neural network. *IOP conference series: Materials science and engineering.* 2021. IOP Publishing, pp. 012020.

AKOGLU, H. 2018. *User's guide to correlation coefficients,*Turkish Journal of Emergency Medicine, Vol 18, Issue 3 Pages 91-93, ISSN 2452-2473,

AL-AZAWEI, A. and AL-MASOUDY, M., 2020. Predicting Learners' Performance in Virtual Learning Environment (VLE) based on Demographic, Behavioral and Engagement Antecedents. *International Journal of Emerging Technologies in Learning (IJET),* 15(9), pp. 60-75.

ALHAKBANI, H.A. and ALNASSAR, F.M., 2022. Open learning analytics: A systematic review of benchmark studies using open university learning analytics dataset (OULAD). *2022 7th international conference on machine learning technologies (ICMLT).* 2022. , pp. 81-86.

ALJOHANI, N.R., FAYOUMI, A. and HASSAN, S., 2019. Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability,* 11(24), pp. 7238.

AL-MAROOF, R.S. et al., 2021. Factors that affect e-learning platforms after the spread of COVID-19: post acceptance study. *Data,* 6(5), pp. 49.

ALNASSAR, F., BLACKWELL, T., HOMAYOUNVALA, E. and YEE-KING, M., 2021. How Well a Student Performed? A Machine Learning Approach to Classify Students' Performance on Virtual Learning Environment. *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM).* IEEE. pp. 1-6.

ALONSO, J.M. and CASALINO, G., 2019. Explainable artificial intelligence for human-centric data analysis in virtual learning environments. *International workshop on higher education learning methodologies and technologies online.* 2019. Springer, pp. 125-138.

ALSHABANDAR, R., et al. , 2020. Students performance prediction in online courses using machine learning algorithms. *2020 international joint conference on neural networks (IJCNN).* 2020. IEEE, pp. 1-7.

ALVES, P., MIRANDA, L. and MORAIS, C., 2017. The influence of virtual learning environments in students' performance. *Universal Journal of Educational Research,* 5(3), pp. 517-527.

AVELLA, J.T. et al., 2016. Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning,* 20(2), pp. 13-29.

BANIHASHEM, S.K., et al., 2018. Learning analytics: A systematic literature review. *Interdisciplinary Journal of Virtual Learning in Medical Sciences,* 9(2),

BART, R., et al., 2020. Effective usage of learning analytics: what do practitioners want and where should distance learning institutions be going? *Open Learning: The Journal of Open, Distance and e-Learning,* 35(2), pp. 178-195.

BERTHOLD, M.R et al., 2020. *Guide to Intelligent data science:How to Intelligently make use of real data.* 2nd ed.: Switzerland AG-Springer Nature.

BOULTON, C.A., KENT, C. and WILLIAMS, H.T., 2018. Virtual learning environment engagement and learning outcomes at a 'bricks-and-mortar'university. *Computers & Education,* 126, pp. 129-142.

BROWNLEE, J., 2020. How to Calculate Feature Importance With Python. https://machinelearningmastery.com/calculate-feature-importance-with-python/ . Accessed [24th August 2022].

CASALINO, G., CASTELLANO, G. and MENCAR, C., 2019. Incremental and adaptive fuzzy clustering for virtual learning environments data analysis. *2019 23rd international conference information visualisation (IV).* 2019. IEEE, pp. 382-387.

CASALINO, G., CASTELLANO, G. and VESSIO, G., 2020. Exploiting time in adaptive learning from educational data. *International workshop on higher education learning methodologies and technologies online.* 2020. Springer, pp. 3-16.

CHIODINI, J., 2020. Online learning in the time of COVID-19. *Travel Med.Infect.Dis.,* 34, pp. 101669.

CHIU, T.K., LIN, T. and LONKA, K., 2021. Motivating online learning: The challenges of COVID-19 and beyond. *The asia-pacific education researcher,* 30(3), pp. 187-190.

CHRISTENSEN, S.S. and SPACKMAN, J.S., 2017. Dropout rates, student momentum, and course walls: A new tool for distance education designers. *Journal of Educators Online,* 14(2), pp. n2.

COMAN, C. et al., 2020. Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective. *Sustainability,* 12(24), pp. 10367.

DA SILVA, L.M., et al., 2022. Learning analytics and collaborative groups of learners in distance education: A systematic mapping study. *Informatics in Education,* 21(1), pp. 113-146.

DA SILVA, L.M., et al., 2021. A literature review on intelligent services applied to distance learning. *Education Sciences,* 11(11), pp. 666.

DAVIES, C.P., 2020. Are VLEs still worthwhile? Journal of Learning Development in Higher Education, (18),

DHAWAN, S., 2020. Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems,* 49(1), pp. 5-22.

GHANI, T. et al., 2021. Development and analysis of a machine learning based software for assisting online classes during COVID-19. *Journal of Software Engineering and Applications,* 14(3), pp. 83-94.

GUANGUL, F.M. et al., 2020. Challenges of remote assessment in higher education in the context of COVID-19: a case study of Middle East College. *Educational assessment, evaluation and accountability,* 32(4), pp. 519-535.

EMBARAK, D.O., 2018. *Data analysis and visualization using python.* New York. Springer.

FOROUGHI, F. and LUKSCH, P., 2018. Data science methodology for cybersecurity projects. *arXiv preprint arXiv:1803.04219.*

ELLIS, R.A., HAN, F. and PARDO, A., 2017. Improving learning analytics–Combining observational and self-report data on student learning. *Journal of Educational Technology & Society,* 20(3), pp. 158-169.

ERIKSSON, T., ADAWI, T. and STÖHR, C., 2017. "Time is the bottleneck": a qualitative study exploring why learners drop out of MOOCs. *Journal of Computing in Higher Education,* 29(1), pp. 133-146.

FLAVIN, M. and BHANDARI, A., 2021. What we talk about when we talk about virtual learning environments. *International Review of Research in Open and Distributed Learning,* 22(4), pp. 164-193.

HAMID, Z., et al., 2018. The concept and use of the virtual learning environment in teaching: a literature review. *International journal of academic research in business and social sciences,* 8(6), pp. 1293-1301.

HASAN, R., et al., 2021. Dataset of Students' Performance Using Student Information System, Moodle and the Mobile Application "eDify". *Data,* 6(11), pp. 110.

HASHIM, M.A., TLEMSANI, I. and MATTHEWS, R., 2022. Higher education strategy in digital transformation. *Education and Information Technologies,* 27(3), pp. 3171-3195.

HASSAN, S., et al., 2019. Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems,* 34(8), pp. 1935-1952.

HIDALGO, R. and EVANS, G., 2020. Analytics for Action: Assessing effectiveness and impact of data informed interventions on online modules. *RIED.Revista iberoamericana de educación a distancia,*

HOLMES, A., ILLOWSKY, B., and DEAN, S. 2017. Introductory business statistics. Houston Texax:OpenStax.

INFANTE-BLANCO, L., ROUSSANALY, A. and BOYER, A., 2018. METALRS: Towards effective learning analytics through a hybrid data collection approach for the french lower secondary education system. *2nd annual learning & student analytics conference.* 2018.

JHA, N.I., GHERGULESCU, I. and MOLDOVAN, A., 2019. OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques. *Csedu (2).* 2019. , pp. 154-164.

JOKSIMOVIĆ, S., KOVANOVIĆ, V. and DAWSON, S., 2019. The journey of learning analytics. *HERDSA Review of Higher Education,* 6, pp. 27-63.

KALIISA, R., MØRCH, A.I. and KLUGE, A., 2019. Exploring social learning analytics to support teaching and learning decisions in online learning environments. *European conference on technology enhanced learning.* 2019. Springer, pp. 187-198.

KEARNEY, M. 2017. Cramer's V. In M. R. (ed), The Sage Encyclopedia of Communication Research Methods. Thousand Oaks, CA: Sage Publications.

KHAN, M.A., 2021. The impact of COVID-19 on UK higher education students: experiences, observations and suggestions for the way forward. *Corporate Governance: The International Journal of Business in Society,*

KIM, D., PARK, Y., YOON, M. and JO, I., 2016. Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments, The Internet and Higher Education. Volume 30. Pages 30-43. ISSN 1096-7516,https://doi.org/10.1016/j.iheduc.2016.03.002.

KUZILEK, J. et al., 2015. OU Analyse: analysing at-risk students at The Open University. *Learning Analytics Review,* , pp. 1-16.

KUZILEK, J., HLOSTA, M. and ZDRAHAL, Z., 2017. Open university learning analytics dataset. *Scientific data,* 4(1), pp. 1-8.

KUZILEK, J., VACLAVEK, J., ZDRAHAL, Z. and FUGLIK, V., 2019. Analysing student vle behaviour intensity and performance. *European Conference on Technology Enhanced Learning.* Springer. pp. 587-590.

KLAWITTER, A. 2022. 5 Challenges students face with online learning in 2022. *Meratas In*c.[online]. Available from: https://meratas.com/blog/5-challenges-students-face-with-remote-learning/ [Accessed 12 August 2022].

LESTER, J. et al., 2019. Learning analytics in higher education : current innovations, future potential, and practical applications. New York: Routledge.

LIU, T., et al., 2022. Predicting High-Risk Students Using Learning Behavior. *Mathematics,* 10(14), pp. 2483.

LONDHE, A. and RAO, P.P., 2017. Platforms for big data analytics: Trend towards hybrid era. *2017 international conference on energy, communication, data analytics and soft computing (ICECDS).* 2017. IEEE, pp. 3235-3238.

LOTSARI, E., et al. , 2014. A learning analytics methodology for student profiling. *Hellenic conference on artificial intelligence.* 2014. Springer, pp. 300-312.

MARINONI, G., VAN'T LAND, H. and JENSEN, T., 2020. The impact of Covid-19 on higher education around the world. *IAU global survey report,* 23.

MERTLER, C. A. and REINHART, R. V., 2017. *Advanced Multivariate Statistical Methods.* 6th ed. New York and London: Routledge Taylor and Francis Group

MILO, T. and SOMECH, A., 2020. Automating exploratory data analysis via machine learning: An overview. *Proceedings of the 2020 ACM SIGMOD international conference on management of data.* 2020. , pp. 2617-2622.

MSELEKU, Z., 2020. A literature review of E-learning and E-teaching in the era of Covid-19 pandemic. *SAGE,* 57(52), pp. 588-597.

NAGPAL, A. and GABRANI, G., 2019. Python for data analytics, scientific and technical applications. *2019 amity international conference on artificial intelligence (AICAI).* 2019. IEEE, pp. 140-145.

NAZIF, A.M., SEDKY, A.A.H. and BADAWY, O.M., 2020. MOOC's student results classification by comparing PNN and other classifiers with features selection. *2020 21st international arab conference on information technology (ACIT).* 2020. IEEE, pp. 1-9.

NELLI, F., 2018. Python data analytics. *Apress Media, California.*

OMONA, K., 2022. Addressing virtual learning challenges in higher institutions of learning: a systematic review and meta-analysis. *Journal of STEAM Education,* 5(2), pp. 100-112.

ONAH, D.F., SINCLAIR, J. and BOYATT, R., 2014. Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 proceedings,* 1, pp. 5825-5834.

PARDO, A. and SIEMENS, G., 2014. Ethical and privacy principles for learning analytics. *British journal of educational technology,* 45(3), pp. 438-450.

PATEL, H., et al. , 2022. Advances in exploratory data analysis, visualisation and quality for data centric AI systems. *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining.* 2022. , pp. 4814-4815.

PEDRO, F. et al., 2019. Artificial intelligence in education: Challenges and opportunities for sustainable development.

POUDYAL, S., MOHAMMADI-ARAGH, M.J. and BALL, J.E., 2022. Prediction of Student Academic Performance Using a Hybrid 2D CNN Model. *Electronics,* 11(7), pp. 1005.

POUDYAL, S., NAGAHI, M., NAGAHISARCHOGHAEI, M. and GHANBARI, G., 2020. Machine Learning Techniques for Determining Students' Academic Performance: A Sustainable Development Case for Engineering Education. *2020 International Conference on Decision Aid Sciences and Application (DASA).* IEEE. pp. 920-924.

RAHMANY, M., ZIN, A.M. and SUNDARARAJAN, E.A., 2020. Comparing Tools Provided By Python And R For Exploratory Data Analysis. *IJISCS Int.J.Inf.Syst.Comput.Sci,* 4(3).

RUNKER, T., 2020. *Data analysis:Models and Algorithms for intelligent data analysis.* 3rd ed. Munchen Germany: Springer View.

SAHOO, K., et al., 2019. Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* 8(12), pp. 2019.

QIU, F. et al., 2022. Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports,* 12(1), pp. 1-15.

RAJ, N.S., PRASAD, S., HARISH, P., BOBAN, M. and CHERIYEDATH, N., 2021. Early prediction of at-risk students in a virtual learning environment using deep learning techniques. *International Conference on Human-Computer Interaction.* Springer. pp. 110-120.

RASHID, A.H.A. et al., 2021. Teachers' Perceptions and Readiness toward the Implementation of Virtual Learning Environment. *International journal of evaluation and research in education,* 10(1), pp. 209-214.

ROMERO, C. and VENTURA, S., 2020. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 10(3), pp. e1355.

RIENTIES, B., et al., 2016. Analytics4Action Evaluation Framework: A Review of Evidence-Based Learning Analytics Interventions at the Open University UK. *Journal of Interactive Media in Education,* 2016(1),

RIENTIES, B., et al., 2017. Making sense of learner and learning Big Data: reviewing five years of Data Wrangling at the Open University UK. *Open Learning: The Journal of Open, Distance and e-Learning,* 32(3), pp. 279-293.

SAYAD, S., 2022. *Data Mining Map*. [online] Saedsayad.com. Available at: <https://www.saedsayad.com/> [Accessed 1 October 2022].

SCHEFFEL, M., et al. , 2019. Policy matters: Expert recommendations for learning analytics policy. *European conference on technology enhanced learning.* 2019. Springer, pp. 510-524.

SHAFFER, D. and RUIS, A., 2017. Epistemic network analysis: A worked example of theory-based learning analytics. *Handbook of learning analytics,*

THORNBURY, E.E., 2020. The relationship between instructor course participation, student participation, and student performance in online courses.

TLADI, L.L. and SERETSE, T.E., 2019. Learning Analytics: Analysing Trends in Online Learning Activities for Masters' Students at Botswana Open University (BOU).

VADLAMANI, S.L. and BAYSAL, O., 2020. Studying software developer expertise and contributions in stack overflow and GitHub. *2020 IEEE international conference on software maintenance and evolution (ICSME).* 2020. IEEE, pp. 312-323.

VERMA, B.K., SINGH, H.K. and SRIVASTAVA, N., 2021. Prediction of Students' Performance in e-Learning Environment using Data Mining/Machine Learning Techniques. *vol,* 23, pp. 586-593.

WEAVER, K., et al., 2021. How far does VLE self-directed study facilitate improvements in written, practical and overall assessment results? Sports therapy case study. *Innovations in Education and Teaching International,* 58(2), pp. 219-229.

WESSA, P., DE RYCKER, A. and HOLLIDAY, I.E., 2011. Content-based VLE designs improve learning efficiency in constructivist statistics education. *PloS one,* 6(10), pp. E25363.

WILLIAMSON, B., EYNON, R. and POTTER, J., 2020. Pandemic politics, pedagogies and practices: digital technologies and distance education during the coronavirus emergency. *Learning, Media and Technology,* 45(2), pp. 107-114.

XIAO, J., HOEL, T. and LI, X., 2019. Constructing an open learning analytics architecture for an open university. *European conference on technology enhanced learning.* 2019. Springer, pp. 609-612.

YOUSAFZAI, B.K., et al., 2021. Student-Performulator: Student Academic Performance Using Hybrid Deep Neural Network. *Sustainability*, 13(17), p.9775.

YUNITA, A., SANTOSO, H.B. and HASIBUAN, Z.A., 2021. Research review on big data usage for learning analytics and educational data mining: A way forward to develop an intelligent automation system. *Journal of physics: Conference series.* 2021. IOP Publishing, pp. 012044.