Using Advanced Analytics to Assist Stakeholders in Higher Education Institutions

Preethi Dsouza
ped9186@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Master's Project is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
Using Advanced Analytics to Assist Stakeholders in Higher Education Institutions

by

Preethi Dsouza

A Capstone Submitted in Partial Fulfilment of the Requirements for the

Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

Spring 2021
Student Name: Preethi Dsouza
Graduate Capstone Title: Using Advanced Analytics to Assist Stakeholders in Higher Education Institutions

Graduate Capstone Committee:

Name: Dr. Sanjay Modak                   Date:
Chair of committee

Name: Dr. Ehsan Warriach                Date:
Member of committee/Mentor
Acknowledgements

I would like to thank Dr. Ehsan for his guidance and valuable advice during the capstone project. I would also like to thank the RIT Dubai higher management for granting me this opportunity to study in this program.

A special word of gratitude to my husband and my son who have encouraged and supported me during this journey to achieve my dream of graduating with a Master’s Degree. And above all, my praise and thanks to Almighty God for providing me the strength and wisdom to complete this program and project.
# Table of Contents

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>5</td>
</tr>
<tr>
<td>Problem Statement</td>
<td>6</td>
</tr>
<tr>
<td>Background of the Problem</td>
<td>6</td>
</tr>
<tr>
<td>Business Understanding of the problem</td>
<td>7</td>
</tr>
<tr>
<td>Project Definition and Goals</td>
<td>8</td>
</tr>
<tr>
<td>Goals of this Project</td>
<td>8</td>
</tr>
<tr>
<td>Limitations of the Project</td>
<td>9</td>
</tr>
<tr>
<td>Literature Review</td>
<td>10</td>
</tr>
<tr>
<td>Methodology</td>
<td>13</td>
</tr>
<tr>
<td>Sources of Data</td>
<td>15</td>
</tr>
<tr>
<td>Analysis</td>
<td>17</td>
</tr>
<tr>
<td>A. Data Description</td>
<td>17</td>
</tr>
<tr>
<td>B. Data Cleaning and Reduction</td>
<td>21</td>
</tr>
<tr>
<td>C. Data Exploration</td>
<td>32</td>
</tr>
<tr>
<td>Hypotheses</td>
<td>62</td>
</tr>
<tr>
<td>D. Data Mining</td>
<td>63</td>
</tr>
<tr>
<td>A. Data Transformation</td>
<td>64</td>
</tr>
<tr>
<td>• Outlier Detection</td>
<td>64</td>
</tr>
<tr>
<td>• Encoding</td>
<td>67</td>
</tr>
<tr>
<td>• NA values</td>
<td>70</td>
</tr>
<tr>
<td>B. Feature Importance</td>
<td>70</td>
</tr>
<tr>
<td>Correlation matrix</td>
<td>71</td>
</tr>
<tr>
<td>C. Sampling</td>
<td>72</td>
</tr>
<tr>
<td>D. Confusion matrix</td>
<td>75</td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>80</td>
</tr>
<tr>
<td>Hyperparameter Tuning</td>
<td>83</td>
</tr>
<tr>
<td>Results</td>
<td>85</td>
</tr>
<tr>
<td>Conclusions and Future Work</td>
<td>87</td>
</tr>
<tr>
<td>Bibliography</td>
<td>88</td>
</tr>
</tbody>
</table>
Table of Figures

Figure 1- Ages of students per code module .................................................................35
Figure 2- Ages of student per code module ...............................................................36
Figure 3- Counts of previous attempts ........................................................................37
Figure 4- Correlation of Final results with previous attempts ....................................38
Figure 5- Percentage Distribution of Final results ......................................................39
Figure 6- Distribution of students per IMD Band .......................................................40
Figure 7- Correlation of Final results and imd_band ..................................................41
Figure 8- Correlation of Final results with Age of students .........................................42
Figure 9- Correlation of Final results with Gender .....................................................43
Figure 10- Correlation of Final results and Highest Qualification of Students ..........44
Figure 11- Distribution of Code_presentations ..........................................................45
Figure 12- Distribution of Code Modules ....................................................................46
Figure 13- Distribution of Final results per course modules ......................................47
Figure 14- Frequency of clicks per VLE page ............................................................48
Figure 15- Number of Student Clicks per VLE resource ............................................49
Figure 16- Average number of clicks per VLE resource ............................................50
Figure 17- Average number of days student accessed the VLE before the exam .........51
Figure 18- Correlation of Final results with student activity on VLE .........................52
Figure 19- Distribution of number of students - Disability .........................................53
Figure 20- Correlation of final results with physical ability of students .......................54
Figure 21- Correlation of Final results with Highest Qualification ...............................55
Figure 22- Correlation of Withdrawal rates and highest qualifications .......................56
Figure 23- Correlation of withdrawal rates and imd_band ...........................................57
Figure 24- Correlation of withdrawal rates and code modules ....................................58
Figure 25- Correlation of Failure rates and imd_band ................................................59
Figure 26- Fail Results and highest educational qualification .....................................60
Figure 27- Fail results and code module .....................................................................61
Figure 28- Scores obtained by students per code module ..........................................62
Figure 29- Studied Credits .........................................................................................66
Abstract

Higher education institutions have access to a vast amount of student data that is recorded during the process of admissions and throughout the course of the program. After the outbreak of the COVID 19 pandemic most universities around the world had to either adopt a blended or a complete online mode of teaching. Students were required to access resources through a virtual learning platform and interaction with this environment also left an important trail of valuable information.

Data analysis and machine learning can provide insights and predictions about students’ performances, online activities and academic progress. Advanced analytics can provide necessary tools to decision makers to analyze and interpret the data appropriately and make informed decisions that improve the student academic outcomes and experience. The latest machine learning tools will not only provide insights that could be utilized to improve the quality of teaching and student performance but also build predictive models that could guide future decisions. Advanced Analytics can predict the future academic success of the students and completion of courses with a very high degree of accuracy. The higher education institutions will be able to leverage these tools to provide insights for other business decisions as well. In this capstone project I aim to explore data from a virtual learning environment and derive valuable conclusions through data exploration and build a predictive model by using advance analytics tools.

Keywords: higher education institutions, student data, advanced analytics, decisions, machine learning, predictive models, quality of teaching, student performance, completion.
Problem Statement

To analyze data gathered from a learning management software and utilize machine learning tools to create predictive analysis models to assist the stakeholders in higher educational institutions to make decisions to support student learning, curriculum and assessment planning and making future business decisions.

Background of the Problem

In recent times, the higher education sector is going through challenges of competition and adoption of the online mode of learning due to the COVID 19 pandemic. Almost all higher education institutions have integrated a hundred percent online or blended format of course delivery. Even though students are getting familiar with this new online delivery mode of education, higher education institutions are facing challenges in student retention and monitoring student progress. The increased number of higher educational institutions in UAE and other parts of the world leads to increased competition as well.

Higher education institutions rely on data gathered from student interaction with learning management software and information extracted from these sources can guide the decision makers to make important decisions. However, the data collected from all these sources is at times unstructured and huge. The data contains information about enrolment, student performance, disciplinary actions, interaction with the resources on the virtual learning platforms, demographic and course related information. It has become increasingly important to utilize the power of machine learning to derive useful conclusions from this data to make informed decisions to assist students and to make better decisions.

This model has various benefits to the stakeholders namely students, faculty, and the decision makers i.e. educational institutions.
Business Understanding of the problem-

- **Identification of ‘at-risk’ learners**

  We can build predictive models i.e. we can identify at-risk learners (risk of dropping out the course or students who need more learning support if their grades are falling. Faculty can use this information to identify students who are finding it difficult to cope with the demands or who are predicted to do so. The faculty can then intervene earlier during the course of the program and offer adequate support and guidance to the student. This will help in bringing the student back on track and also keep him enrolled in the program. The results from these interventions can also be monitored to gauge their effectiveness and these can be used for future students as well.

- **Changes in curriculum and assessment**

  Students can also be directed to better resources to acquire skills and knowledge required for their courses. This can be achieved by gathering data on which students are lagging behind or at risk of failing.

- **Potential Career paths and choices for students**

  Admissions can identify the best programs choices for students based on the insights and career counselors can give the right advice to students for their future career choices.

- **Assisting management decisions**

  These insights can also assist the management to make effective administrative decisions regarding the resources and facilities in the institutions.

- **Innovation**

  Faculty will be equipped with tools to customize their course delivery as per the student’s needs and organize their courses so that they can improve student participation and engagement.

- **Providing students with analysis on their own learning**

  This analysis can also give information to students about their own academic progress, their future growth and the suggested changes to achieve more success.
Project Definition and Goals

This project has been motivated by the intention of helping students and the higher education institutions as I have been part of this domain since many years and this is my way of giving back to this community of learners and educators.

The main aim of this capstone project is to leverage machine learning tools to create models to that analyze student data collected from a learning management software and to offer these valuable insights to higher educational institutions that will empower them to make right business decisions with regards for marketing, recruitment and admissions and enhancing student performance and future growth trajectory. The real time insights gained through these models will enable the institutions to gauge the student engagement, academic progress and identify areas of improvement and required support. Institutions will be able to use these insights to forecast graduation rates, drop-out and retention rates.

It is envisaged that education systems that do make the transition towards data-informed planning, decision making, and teaching and learning will hold significant competitive and quality advantages over those that do not. (Siemens et al., 2013, p. 2) Additionally, when used effectively,). More specifically, the Big Data approach can help researchers collect information in real time on students and their activities through the use of automated systems, tracking their interactions with the learning environment, peers and instructors, and provide other information that might be relevant to understand how a particular population of students or subgroup of students are performing in their academic program. Furthermore, if developed as an organizational capacity, the ongoing analysis of big data can provide insights into the design of learning environments and inform decisions about how to manage educational resources on all levels (Ifenthaler et al., 2015).

Goals of this Project

1. **Student profiling** - Exploring relationships with student performance and their background, demographics and profiles.

2. **Creating alerts for decision makers** - Segmenting students based on their performance and identifying students that have less engagement or displaying
undesirable student behaviors or less participation and supporting them with the right intervention

3. **Planning and Scheduling** - Providing visualization/ dashboards for student performance/ attrition rates/graduation and drop-out rates to decision makers/ higher management in higher education institutions to equip them with the right information to take necessary measures to improve admission rates/ student performance and reduce drop-out rates and increase the retention rates.

4. **Student modelling** – Customizing the courses based on the segmentation of students based on their demographic and academic choices and performance.

5. **Predicting student performance** - Creation of predictive models that can predict student’s future grades and choices of subjects based on historical data.

**Limitations of the Project**

1. The dataset is divided into 7 different datasets. Even though the datasets have similar attributes in most of them, understanding the datasets and merging them took a considerable amount of time. Domain knowledge was required to understand the datasets in more detail.

2. Some datasets did not have information on all students. For example, the VLE dataset had much lesser number of rows compared to the student info dataset. I had aimed at analyzing the student interaction with the VLE as a contributing factor to the overall student performance but due to this I was unable to utilize this information as this would have led to miscalculations.

3. The reasons for withdrawal are not recorded in any of the datasets. The reasons would have provided a greater insight into why the withdrawal rates were high.

4. There was a lack of qualitative data on the course content and the subjects so it was rather difficult to segment students based on this.
The contributions of Big Data and analytics to the landscape of higher education can be understood at fundamentally three—micro, meso and macro levels (Daniel 2015). Daniel explains that at micro level, Big Data and analytics will help institutions improve the quality of learning and teaching while streamlining processes and reducing administrative workload. Daniel explains that data analytics in the education sector can also be used to understand student behavior. When a student accesses the online learning environments they leave important trail of information that can be utilized to explore their feelings, their social connections and goals. “Researchers can use such data to examine patterns of student performance over time—from one semester to another or from one year to another—and develop rigorous data modeling and analysis to reveal the obstacles to student access and usability and to evaluate any attempts at intervention”. (Daniel, 2015). He describes in his paper that on the meso level, program performance can be monitored such as improvement in graduate rates and graduate satisfaction. “Further, with mounting pressure on institutions to constantly report to multiple stakeholders (government, public and others) on key performance indicators (KPIs), Big Data techniques can be used to harness available data to provide a broad macro view of institutional performances and accountability and identify areas that need particular attention. While the benefits of Big Data and analytics might appear obvious, the added value will come from the ability to develop and deploy robust models that can adequately capture and assess the present state of performance, as well as accurately predict future outcomes”. (Daniel, 2015).

Online environments allow the generation of large amounts of data related to learning/teaching processes, which offers the possibility of extracting valuable information that may be employed to improve students’ performance (Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Siemens, George and Baker in their paper explain how data analytics can assist the educators in understanding and analyzing ways in which students access and utilize the learning resources. They suggest that data analytics can then be utilized to develop a ‘competency-based learning mode’ in which every student can perform at their own pace and improve their academic success. “Their embrace of data analytics and predictive modelling will have a significant impact on the way in which courses and institutions are marketed, curricula are structured and students are monitored and supported. Growing interest in data and analytics in education,
teaching, and learning raises the priority for increased, high-quality research into the models, methods, technologies, and impact of analytics”. (Siemens, George & Baker, Ryan. (2012).

Agasisti and Bowers recommend that an education data scientist can play a key role to bridge the data analysts and educational practitioners in the “loop of data”, as the education data scientist “owns the technical skills to collect, analyze and use quantitative data, and at the same time the managerial and communication skills to interact with decision-makers and managers at the school level to individuate good ways of using the information in the practical way of improving practices and initiatives.

Pardo and Siemens explain in their paper that descriptive models can be created by analyzing the ‘transactional’ and the ‘interactional’ data which can assist in identifying trends of enrolments, graduation rates and patterns that need particular attention to improve student learning. They explain that the descriptive models alone are not sufficient but the institutions also need to examine the present performance and predict future results. “Predictive models provide institutions with the ability to uncover hidden relationships in data and predict future outcomes with a certain degree of accuracy. For instance, they enable institutions to identify students who are exhibiting risky behaviors during their academic program. Prescriptive models are actionable tools built based on insights gained from both descriptive and predictive models. They are intended to help institutions to accurately assess their current situation and make informed choices on alternative course of events based on valid and consistent prediction”. (Pardo & Siemens, 2014).

Agasisti and Bowers summarize four barriers that impede the use of data analytics in education and propose potential solutions to lower these barriers. The first concern is the potential threat to student privacy, as many tools are built upon tracking student information. The authors argue that “open code and open access standards must be used when data analytic or machine learning algorithms are used to inform evidence-based improvement cycles in schools, or to the extent that Learning Analytics algorithms make recommendations for content and instruction for student learning”. Second, to address the complexity of data, researchers are recommending using data warehouses for data reporting and analytics. Third, the creation of an adequate platform for data analysis can be costly. However, open access code publication will facilitate the sharing of code and reduce the cost to build such platforms. Fourth, it is challenging to develop methodologies that “present the results without excessive simplification (providing an
awareness of the complexities of the learning process) but with enough clarity to make the information understandable, and thus usable.

“In addition to student’s background and performance data, each action carried out (reading files, participating in forums, sending messages, or visiting recommended links, for example) leaves a digital fingerprint” (Calvet Liñán, L., & Juan Pérez, Á. A. (2015)) “Educational Data mining (EDM) is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn. EDM seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners”. (Romero, Ventura. (2010).

In the study published on Pattern Mining Approach to Categorization of Students’ Performance using Apriori Algorithm, (2015) Sushil Kumar et.al have used the Apriori Algorithm to categorize students based on their records and predict their future performance. Data mining is used in education to achieve quality education and to categorize the students’ performance through the analysis of educational data which reside or store in educational organization’s database (Sushil Kumar et.al.). In this paper, through association rule mining they developed a mechanism to identify the weaker students by comparing the performance in previous semesters. Yanbai He and Rui Che, et al. in their study Online At-Risk Student Identification using RNN-GRU Joint Neural Networks have aimed to predict student’s performance on the basis of historical data gathered from the learning management software. They have used three recurrent neural network (RNN) algorithms to extract features from the history of student interaction with the software. They have initially used a novel joint neural network model framework to identify at risk students based on their demographic information and stream data. They established that performance of the Gated recurrent unit (GRU) and simple RNN is better. Robert Peach, Sophia Yalirkaki et.al.in their paper aim to use a time series analysis of online learners’ interaction through which they use clustering of students that show similar behavior behaviors. A posteriori comparison against student performance shows that, whereas high performing learners are spread across clusters with diverse temporal engagement, low performers are located significantly in the massed learning cluster, and our unsupervised clustering identifies low performers more accurately than common machine learning classification methods trained on temporal statistics of the data. (Robert Peach, Sophia N. Yaliraki et.al, 2019).
Methodology

Data Analytics Model in Education

(1) Collection and acquisition

(2) Storage
(Datasets available in the institutions' registers)

(3) Cleaning
(Preparing datasets for their use)

(4) Integration
(multiple datasets)

(5) Analysis
(Techniques and applications)

(6) Representation and Visualization

(7) Action(s)
(Managerial and policy implications)

Environment

Main programming language – Python

Programming environment – Jupyter Notebook (An open source web application that you can use to create and share documents that contain live code)

Data Visualization tools – Matplotlib and Seaborn in Python

Data Mining – Random Forest Classifier, Logistic Regression and Support Vector Machines in Python

Hyperparameter Tuning - GridSearchCV in Python
I have used the CRISP-DM Methodology for this project and have followed the following steps –

- Data Understanding
- Data Preparation
- Data Modeling
- Evaluation
Sources of Data

Open University is a public British University that also has the highest number of undergraduate students in the UK. It is the largest academic institution in the United Kingdom (and one of the largest in Europe) with 2 Million enrolled students since it is established at 1969.

The Open University is one of the largest distance learning universities in the world with more than 150,000 students enrolled in various programs. All the reading materials and other resources and other course content are posted on the Virtual Learning Environment (VLE). Students’ interactions with the educational materials are recorded and stored in university data warehouse. The dataset contains the information about 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (10,655,280 entries). (Kuzilek, J., Hlosta, M. & Zdrahal, Z).

The dataset that is used for this project is the Open University Learning Analytics dataset (OULAD). This dataset consists of not only the assessment results but also information on student’s demographics and the student interaction with the Virtual Learning Environment. The dataset consists of information on 32593 students for a period of nine months between years 2014 and 2015.


It consists of 7 selected courses (mentioned as modules in the dataset). Different presentations indicated with letters "B" and "J" after year for semester 2 and semester 1 respectively.

Additionally, the dataset includes student demographics such as location, age group, disability, education level, gender etc. Student assessment marks, interactions with the Virtual Learning Environment (VLE) are also included.

Modules can be presented multiple times during the year. The names of the course modules end with the name of the month that is A for January, B for February etc. For example, presentations starting in January ends with A, in February with B and so on; so that ‘2013J’ means that the presentation started in October 2013. (Kuzilek, J., Hlosta, M. & Zdrahal, Z).
The grading scheme for the final exam is as follows – A fail grade is given to students who scored less than 40%. The dataset used for this project also lists the final result as ‘Withdrawn’ for those students who have withdrawn from the courses in the middle of the semesters or courses. This has been retained in the dataset to draw important conclusions from the data.

- Distinction
- Pass
- Fail
Analysis

A. Data Description –

The most important step for a data analyst is to understand the data that is being used. The OULAD dataset is downloaded as a dataset that consists of 7 different files. The following explain the attributes in each individual file briefly. During the data cleaning and reduction process, some attributes may be dropped as they may have null values and duplicates without affecting the quality and integrity of the data. Some of the files will be merged as well on the basis of similar attributes to create one single dataset for the purpose of data mining.

Datasets and attributes
Brief description of 7 datasets used for this project

University Learning Analytics dataset

The presentations are available on the VLE a few weeks before the start of the course module. During the presentation the students’ knowledge is evaluated in series of assessments, which defines the milestones in the module. At the end, there is usually the final exam. (Kuzilek, J., Hlosta, M. & Zdrahal, Z.)

1. Courses

<table>
<thead>
<tr>
<th>code_module</th>
<th>code_presentation</th>
<th>module.presentation_length</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>268</td>
</tr>
<tr>
<td>AAA</td>
<td>2014J</td>
<td>269</td>
</tr>
<tr>
<td>BBB</td>
<td>2013J</td>
<td>268</td>
</tr>
<tr>
<td>BBB</td>
<td>2014J</td>
<td>262</td>
</tr>
<tr>
<td>BBB</td>
<td>2013B</td>
<td>240</td>
</tr>
</tbody>
</table>

2. Assessments

This dataset consists of the name of the assessment, the weight and the expected deadline for each and the results of the submission.

<table>
<thead>
<tr>
<th>code_module</th>
<th>code_presentation</th>
<th>id_assessment</th>
<th>assessment_type</th>
<th>date</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>1752</td>
<td>TMA</td>
<td>19.0</td>
<td>10.0</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>1753</td>
<td>TMA</td>
<td>54.0</td>
<td>20.0</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>1754</td>
<td>TMA</td>
<td>117.0</td>
<td>20.0</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>1755</td>
<td>TMA</td>
<td>166.0</td>
<td>20.0</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>1756</td>
<td>TMA</td>
<td>215.0</td>
<td>30.0</td>
</tr>
</tbody>
</table>
3. **VLE**

Interaction of students with VLE is recorded as the number of times the student clicks on the system daily. The VLE system recorded the interaction of students and categorized them into 20 different classes and the medium of interaction i.e. URL clicks, completing quizzes, filling in questionnaires.

<table>
<thead>
<tr>
<th>id_site</th>
<th>code_module</th>
<th>code_presentation</th>
<th>activity_type</th>
<th>week_from</th>
<th>week_to</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>AAA</td>
<td>2013J</td>
<td>resource</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>1</td>
<td>AAA</td>
<td>2013J</td>
<td>oucontent</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>2</td>
<td>AAA</td>
<td>2013J</td>
<td>resource</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>3</td>
<td>AAA</td>
<td>2013J</td>
<td>url</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>4</td>
<td>AAA</td>
<td>2013J</td>
<td>resource</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

4. **Student Info**

This dataset contains information about the demographics and the final results of students.

5. **Student Registration**

This dataset consists of information recorded at the time of registration for the presentation. In case of students withdrawing from the presentation, this date is also recorded as date of unregistration.
6. Student Assessment

This file contains the results of students’ assessments. This file contains the following columns:

<table>
<thead>
<tr>
<th>id_assessment</th>
<th>id_student</th>
<th>date_submitted</th>
<th>is_banked</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1752</td>
<td>11391</td>
<td>18</td>
<td>78.0</td>
</tr>
<tr>
<td>1</td>
<td>1752</td>
<td>28400</td>
<td>0</td>
<td>70.0</td>
</tr>
<tr>
<td>2</td>
<td>1752</td>
<td>31604</td>
<td>0</td>
<td>72.0</td>
</tr>
<tr>
<td>3</td>
<td>1752</td>
<td>32885</td>
<td>0</td>
<td>69.0</td>
</tr>
<tr>
<td>4</td>
<td>1752</td>
<td>38053</td>
<td>0</td>
<td>79.0</td>
</tr>
</tbody>
</table>

7. studentVle.csv

The studentVle.csv file contains information on the number of clicks per day on the various materials in the VLE. This file contains the following columns:

<table>
<thead>
<tr>
<th>code_module</th>
<th>code_presentation</th>
<th>id_student</th>
<th>id_site</th>
<th>date</th>
<th>sum_click</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>28400</td>
<td>546652</td>
<td>-10</td>
<td>4</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>28400</td>
<td>546652</td>
<td>-10</td>
<td>1</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>28400</td>
<td>546652</td>
<td>-10</td>
<td>1</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>28400</td>
<td>546614</td>
<td>-10</td>
<td>11</td>
</tr>
<tr>
<td>AAA</td>
<td>2013J</td>
<td>28400</td>
<td>546714</td>
<td>-10</td>
<td>1</td>
</tr>
</tbody>
</table>
B. Data Cleaning and Reduction

After checking for null and NA values we found that some columns in various datasets had null values. I decided to perform the following operations as these would not significantly affect the dataset.

- **VLE file** –

  There were 5243 rows that had null values in week_from and week_to. These columns would not have had a significant effect on the analysis and hence both these columns were dropped from the dataset. The resulting dataset now looked as shown below-

<table>
<thead>
<tr>
<th>id_site</th>
<th>code_module</th>
<th>code_presentation</th>
<th>activity_type</th>
</tr>
</thead>
<tbody>
<tr>
<td>546943</td>
<td>AAA</td>
<td>2013J</td>
<td>resource</td>
</tr>
<tr>
<td>546712</td>
<td>AAA</td>
<td>2013J</td>
<td>oucontent</td>
</tr>
<tr>
<td>546998</td>
<td>AAA</td>
<td>2013J</td>
<td>resource</td>
</tr>
<tr>
<td>546888</td>
<td>AAA</td>
<td>2013J</td>
<td>url</td>
</tr>
<tr>
<td>547035</td>
<td>AAA</td>
<td>2013J</td>
<td>resource</td>
</tr>
</tbody>
</table>

- **Student Info and Student Registration**

  We merged these 2 data frames as the registration and withdrawn dates will be added to the student information file. The resulting data frame is as shown below-

```python
Int64Index: 32593 entries, 0 to 32592
Data columns (total 14 columns):
# Column        Non-Null Count   Dtype
0 code_module   32593 non-null   object
1 code_presentation 32593 non-null object
2 id_student    32593 non-null   int64
3 gender        32593 non-null   object
4 region        32593 non-null   object
5 highest_education 32593 non-null object
6 imd_band      32593 non-null   object
7 age_band      32593 non-null   object
8 num_of_prev_attempts 32593 non-null int64
9 studied_credits 32593 non-null int64
10 disability   32593 non-null   object
11 final_result 32593 non-null   object
12 date_registration 32593 non-null float64
13 date_unregistration 10072 non-null float64
dtypes: float64(2), int64(3), object(9)
memory usage: 3.7+ MB
```
I added a percentage sign to the values 10-20 as it was being read as Oct 20 in most of the columns as the original dataset did not have a percentage sign after the value. The resulting dataset had all rows that were consistent.
- **Student Info- IMD Band**

There were 1111 null values in the variable IMD band. This is the Index of Multiple Deprivation. The regions and IMD band as well did not show any correlation as they are distributed along every region as evident from the following chart-

**Correlation of imd_band and regions**

![Correlation of imd_band and regions](image)

Hence the rows with null values in IMD band were dropped from the dataset.

- **Student_registration- Date unregistration**

The Date Un-registration column shows the number of days after which the student withdrew from the course. There were 45 null values in the date registration column which shows the registered date for 45 students was not recorded. After filtering the data, it was found that most of these dates were null because there were withdrawn students. Out of these 45, 6 students did not have any dates recorded in either of these columns. Since this could interfere with the data interpretation I have considered this as an omission and removed these 6 rows from the final dataset. The date-registration column that shows the number of days that the student has
registered before the start of the module also is converted to integers. The final student_info.csv dataset has now the following attributes –

<table>
<thead>
<tr>
<th>id_student</th>
<th>num_of_prev_attempts</th>
<th>studied_credits</th>
<th>date_registration</th>
<th>date_unregistration</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>3.147600e+04</td>
<td>31476.000000</td>
<td>31476.000000</td>
<td>31437.000000</td>
</tr>
<tr>
<td>mean</td>
<td>7.066825e+05</td>
<td>0.164506</td>
<td>79.769666</td>
<td>-69.831981</td>
</tr>
<tr>
<td>std</td>
<td>5.501587e+05</td>
<td>0.482652</td>
<td>41.051985</td>
<td>49.199856</td>
</tr>
<tr>
<td>min</td>
<td>3.733000e+03</td>
<td>0.000000</td>
<td>30.000000</td>
<td>-322.000000</td>
</tr>
<tr>
<td>25%</td>
<td>5.075628e+05</td>
<td>0.000000</td>
<td>60.000000</td>
<td>-101.000000</td>
</tr>
<tr>
<td>50%</td>
<td>5.897960e+05</td>
<td>0.000000</td>
<td>60.000000</td>
<td>-57.000000</td>
</tr>
<tr>
<td>75%</td>
<td>6.441208e+05</td>
<td>0.000000</td>
<td>120.000000</td>
<td>-29.000000</td>
</tr>
<tr>
<td>max</td>
<td>2.718705e+06</td>
<td>6.000000</td>
<td>665.000000</td>
<td>167.000000</td>
</tr>
</tbody>
</table>

The above table shows us that out of 31476 students the dates of registration for 31437 students were recorded. This shows a difference of 39 students whose registration dates are not recorded. These students are shown as withdrawn as their date of ‘un-registration’ or withdrawn has been recorded.

Out of the 31437 students registered, 9837 students have unregistered or withdrawn from the courses. Let us further investigate in detail to know more about this dataset.

- **Student Assessments**

Score column had 173 null values. The submission column shows that the student has submitted the assessment. It could mean that the scores were not recorded in the dataset. Hence the null values were replaced with ‘0’ which effectively will mean that the student has failed the assessment.

- **Merging of student Assessments and assessments.csv**

I merged both these datasets by using the unique identifier of the assessment i.e. id assessment as the base. The resulting dataset contained the values of weight to denote the weight of each presentation relative to the total module. The dataset now had the following attributes-
Merging of studentVle.csv and vle_csv

Both datasets had similar variables except for the information on the activity type. It was found best to merge both these datasets to create one concise dataset on the basis of student IDs.

The header of the resulting dataset is shown below –

<table>
<thead>
<tr>
<th>code_module</th>
<th>code_presentation</th>
<th>id_student</th>
<th>id_site</th>
<th>date</th>
<th>sum_click</th>
<th>activity_type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 AAA</td>
<td>2013J</td>
<td>75091</td>
<td>546943</td>
<td>-10</td>
<td>1</td>
<td>resource</td>
</tr>
<tr>
<td>1 AAA</td>
<td>2013J</td>
<td>186149</td>
<td>546943</td>
<td>-10</td>
<td>1</td>
<td>resource</td>
</tr>
<tr>
<td>2 AAA</td>
<td>2013J</td>
<td>205350</td>
<td>546943</td>
<td>-10</td>
<td>2</td>
<td>resource</td>
</tr>
<tr>
<td>3 AAA</td>
<td>2013J</td>
<td>1626710</td>
<td>546943</td>
<td>-9</td>
<td>1</td>
<td>resource</td>
</tr>
<tr>
<td>4 AAA</td>
<td>2013J</td>
<td>2643002</td>
<td>546943</td>
<td>-8</td>
<td>1</td>
<td>resource</td>
</tr>
</tbody>
</table>

I have reduced the number of datasets from 7 to 3 after combining them based on similar attributes. The 3 combined datasets with their attributes are as follows –
While exploring the datasets as shown in the next section it was observed that merging of the 3 datasets will be required to train the model to conduct predictive analysis. In order to do this, I performed the following steps in Python.

- **Student Info** - This is used as the main dataset. The important features from the Student Assessments and Student VLE that were required for the final analysis and data mining were selected to be merged with the main dataset i.e. the Student Info dataset.

Before merging the datasets, I had to do the following steps –

Student Info dataset had rows per student ID but the Student VLE file had multiple rows for every activity done by each student and the sum of clicks per student ID.

The aim was to merge the date column i.e. the number of days before the presentation when the student accessed the VLE to each student ID. For this purpose, first I used the pivot function in Python and created a dataset which groups the columns per student ID and with the mean values of the sum click and the date columns.
The resulting file is shown below –

**Student VLE dataset**

<table>
<thead>
<tr>
<th>Unnamed: 0</th>
<th>id_student</th>
<th>activity_type</th>
<th>sum_click</th>
<th>id_site</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>6516.0</td>
<td>dataplus</td>
<td>5.250000</td>
<td>877421.750000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>6516.0</td>
<td>forumg</td>
<td>2.577143</td>
<td>877013.937143</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6516.0</td>
<td>homepage</td>
<td>3.145570</td>
<td>877030.000000</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>6516.0</td>
<td>oucontent</td>
<td>8.179348</td>
<td>877065.005435</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>6516.0</td>
<td>resource</td>
<td>1.631579</td>
<td>884216.052632</td>
</tr>
</tbody>
</table>

I then created another dataset by filtering only the columns for student ID and the mean values of the date. The other variables were irrelevant for the purpose of the final analysis as it only gave the details of the activity and the id site number. These were dropped from the final dataset.

The mean values of the date column were then merged with the Student. Info. dataset the description of the merged dataset is shown below – However, it was found that the number of rows in the Student VLE column when merged with the main dataset were much lesser and were only **4083 rows**. On further investigation it was found that the most of the student ID details were not recorded in the Student VLE dataset. Since this will not give us an accurate picture of student activity on VLE I decided to skip this part of merging the date column along with the main dataset.
**Studentassessments.csv** - There are 3 types of assessments that a student has to submit to achieve the final grade for the courses. They are listed as TMA, CMA and EMA i.e. Teacher marked assessments, Computer marked assessments and then EMA which is the Final Exam.

The scores for each of these are shown per exam for each student. The variables of Assessments (TMA, CMA and EMA) and their respective scores were merged with the Studentinfo.csv dataset on the basis of student ID.

The combined dataset showed some anomalies. Some of the student IDs were duplicated in the original dataset and these had resulted in the mismatch of information of the Assessment type and the respective scores. In order to further clean up the dataset I filtered the duplicate rows to find out which student IDs had duplicate rows in the original dataset. I found out that there were 3808 duplicate rows as per the student IDs. To further investigate this, I filtered out the duplicate rows and checked the details of each row. The following example shows that the merged dataset shows anomalies in the year column. These were corrected in all instances to clean the dataset.
Example 1 – Student Id (65002) below – It appears that this student had enrolled in 2013J and then withdrew and then he re-enrolled and then failed the course. He gave 2 TMA assessments in 2013J i.e October 2013 and got a mean score of 67 and then withdrew the course. However, he re-enrolled in the program in 2014J i.e. October 2014 but failed the course. His mean score however in the assessments i.e. in the TMAs was 67. The final exam grades as per the assessment.csv dataset i.e. Assessment Id– 1757 and 1763 were not recorded. Hence it is unclear what the student scored in the final exam. The dates were corrected in the final dataset.

<table>
<thead>
<tr>
<th>Studentassessments.csv</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 1752 65002 17 0 66 AAA 2013J TMA 19 10</td>
</tr>
<tr>
<td>515 1753 65002 51 0 68 AAA 2013J TMA 54 20</td>
</tr>
<tr>
<td>1768 1758 65002 -1 1 66 AAA 2014J TMA 19 10</td>
</tr>
<tr>
<td>1982 1759 65002 -1 1 68 AAA 2014J TMA 54 20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Studentinfofinal.csv</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA 2013J 65002 F AAA 2013J TMA 67</td>
</tr>
<tr>
<td>AAA 2013J 65002 F AAA 2014J TMA 67</td>
</tr>
<tr>
<td>AAA 2014J 65002 F AAA 2013J TMA 67</td>
</tr>
<tr>
<td>AAA 2014J 65002 F AAA 2014J TMA 67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessments.csv</th>
</tr>
</thead>
<tbody>
<tr>
<td>65002 AAA 2013J 0 60 N Withdrawn</td>
</tr>
<tr>
<td>65002 AAA 2014J 1 60 N Fail</td>
</tr>
<tr>
<td>AAA 2013J 1756 TMA 215 30</td>
</tr>
<tr>
<td>AAA 2014J 1763 Exam 100</td>
</tr>
</tbody>
</table>

Example 2

<table>
<thead>
<tr>
<th>Studentsassessment.csv</th>
</tr>
</thead>
<tbody>
<tr>
<td>1752 295741 18 0 38 AAA 2013J TMA 19 10</td>
</tr>
<tr>
<td>1753 295741 53 0 52 AAA 2013J TMA 54 20</td>
</tr>
<tr>
<td>1754 295741 115 0 52 AAA 2013J TMA 117 20</td>
</tr>
<tr>
<td>1755 295741 164 0 54 AAA 2013J TMA 166 20</td>
</tr>
<tr>
<td>1756 295741 220 0 55 AAA 2013J TMA 215 30</td>
</tr>
</tbody>
</table>
The details were checked against the studentassessments.csv file and corrected accordingly.

Again, it appears that the final exam scores are not recorded for this student as well. In the year 2013 he failed the exam after giving 3 different sets of TMAs with an average score of 50.2 and in 2014B he gave 1 set of CMAs and 1 set of TMAs and passed the final exam.

Since the dataset did not add give details for the scores of the final exams for most students and I couldn’t correlate the scores of the TMAs and CMAs to the final scores, the duplicate rows were dropped by filtering them per Student ID and Code Presentation columns. This gave us a final dataset with some students as shown in the example below who are duplicated because of their status i.e. if they have withdrawn and re-enrolled in the program.

Example – filtered dataset – shows duplicates for those students who have withdrawn in one term and re-enrolled in another. The other student IDs are all now unique in the final dataset.
After deleting the extra columns created while merging we now have a final dataset with the following attributes.

```python
import pandas as pd

data = pd.read_csv('dataset.csv')
data.head()
```

<table>
<thead>
<tr>
<th>id_student</th>
<th>num_of_prev_attempts</th>
<th>studied_credits</th>
<th>date_registration</th>
<th>date_unregistration</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>2.4905*10^04</td>
<td>24905.000000</td>
<td>24905.000000</td>
<td>24896.000000</td>
<td>4905.000000</td>
</tr>
<tr>
<td>mean</td>
<td>7.101334*10^05</td>
<td>0.165790</td>
<td>77.098173</td>
<td>-67.882592</td>
<td>98.218349</td>
</tr>
<tr>
<td>std</td>
<td>5.604140*10^05</td>
<td>0.481194</td>
<td>38.866749</td>
<td>48.457732</td>
<td>74.017930</td>
</tr>
<tr>
<td>min</td>
<td>6.516000*10^03</td>
<td>0.000000</td>
<td>30.000000</td>
<td>-320.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>25%</td>
<td>5.026970*10^05</td>
<td>0.000000</td>
<td>80.000000</td>
<td>-98.000000</td>
<td>45.000000</td>
</tr>
<tr>
<td>50%</td>
<td>5.873740*10^05</td>
<td>0.000000</td>
<td>60.000000</td>
<td>-54.000000</td>
<td>97.000000</td>
</tr>
<tr>
<td>75%</td>
<td>6.385150*10^05</td>
<td>0.000000</td>
<td>90.000000</td>
<td>-29.000000</td>
<td>156.000000</td>
</tr>
<tr>
<td>max</td>
<td>2.698586*10^06</td>
<td>6.000000</td>
<td>630.000000</td>
<td>167.000000</td>
<td>444.000000</td>
</tr>
</tbody>
</table>
```

Data columns (total 16 columns):

<table>
<thead>
<tr>
<th>#</th>
<th>Column</th>
<th>Non-Null Count</th>
<th>Dtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>code_module_x</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>1</td>
<td>code_presentation_x</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>2</td>
<td>id_student</td>
<td>24905</td>
<td>int64</td>
</tr>
<tr>
<td>3</td>
<td>gender</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>4</td>
<td>region</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>5</td>
<td>highest_education</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>6</td>
<td>imd_band</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>7</td>
<td>age_band</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>8</td>
<td>num_of_prev_attempts</td>
<td>24905</td>
<td>int64</td>
</tr>
<tr>
<td>9</td>
<td>studied_credits</td>
<td>24905</td>
<td>int64</td>
</tr>
<tr>
<td>10</td>
<td>disability</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>11</td>
<td>final_result</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>12</td>
<td>date_registration</td>
<td>24905</td>
<td>float64</td>
</tr>
<tr>
<td>13</td>
<td>date_unregistration</td>
<td>4905</td>
<td>float64</td>
</tr>
<tr>
<td>14</td>
<td>assessment_type</td>
<td>24905</td>
<td>object</td>
</tr>
<tr>
<td>15</td>
<td>score</td>
<td>24905</td>
<td>float64</td>
</tr>
</tbody>
</table>
C. Data Exploration

Exploring the 3 datasets above lead to various conclusions. I have used Matplotlib and Seaborn in Python to plot various types of graphs that will give us a detailed analysis on different attributes in the 3 different datasets.

Exploration of Demographic Data

The demographic data gave us various insights into the correlation between the student performance and their background. The Student Info dataset was used for this purpose.

<table>
<thead>
<tr>
<th>Gender Distribution – Students –</th>
<th>Age Distribution – Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are 17065 male students and 14411 female students</td>
<td>The majority of the students were aged under 35. There were 9007 students between the ages of 35-55. I have tried to find out if this age group was mostly enrolled in Graduate or Post graduate courses.</td>
</tr>
</tbody>
</table>

| M | 17065 |
| F | 14411 |

Name: gender, dtype: int64

| 0-35 | 22268 |
| 35-55 | 9007 |
| 55<= | 201 |

Name: age_band, dtype: int64

![Histogram of Gender Distribution](image1.png)

![Histogram of Age Distribution](image2.png)
**Geographic Distribution – Students**

Highest number of students belong to Scotland and the second highest region is the East Anglian Region.

**Age distribution per region**

We gather from the figure below that across all regions the students from age groups 0-35 are the highest in number. However, it is also observed that the number of students between the ages 35-55 are more from the South region, London, East Anglian Region and particularly high from Scotland. Most students from the ages 55 and above are from the South region as seen in the Figure below.
Students highest qualification was ‘A level or equivalent’ and the second highest was ‘Lower than A level’. The students who had possessed a Post graduate qualification were also enrolled in the University. The student numbers per highest education were as follows –

<table>
<thead>
<tr>
<th>Qualification</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Level or Equivalent</td>
<td>13759</td>
</tr>
<tr>
<td>Lower Than A Level</td>
<td>12759</td>
</tr>
<tr>
<td>HE Qualification</td>
<td>4444</td>
</tr>
<tr>
<td>No Formal quals</td>
<td>328</td>
</tr>
<tr>
<td>Post Graduate Qualification</td>
<td>186</td>
</tr>
</tbody>
</table>

Student ages distribution by highest qualification- It is observed from the figure below that students below the ages of 35 had qualifications that were mostly A level or equivalent and lower than A levels. The students in the age group of 35-55 also had qualifications similar to these but also the ratio of students between the ages of 35-55 having qualifications lower than A level were marginally higher. It was also observed that the few students i.e. 202 students who were 55 and above had qualifications titled as Higher Education and some also had Post-Graduate qualifications.
- Exploration of Demographic Data per Code module
  - Student age distribution by code module.

**Figure 1** illustrates that most of the students are enrolled in course modules between the ages of 0-35 and the most popular course modules is FFF, followed by BBB and DDD. Course module AAA has the least enrolment and almost equal number of students between the ages of less than 35 and 35-55 are enrolled in this module. Course module FFF and EEE are more popular among the students aged lesser than 35 than those between 35-55.

*Figure 1- Ages of students per code module*
**Fig. 2** below gives us a better view of the distribution of students per age group per module. Students between the ages of 0-35 have been studying course modules BBB and FFF. The most popular module for age groups 0-35 is BBB as well but the no of students for FFF between 0-35 is lesser.

*Figure 2- Ages of student per code module*
- Previous attempts by code module and grades

In Figure 3 it is observed that the number of students who have attempted it the second time are quite few across the modules and those that have attempted the modules more than three times are also not that many. There are students who have attempted these modules almost 5 or 6 times as well as per the counts below. We need to find out how many maximum attempts are allowed in a course module till the student is awarded a grade.

Figure 3 - Counts of previous attempts

<table>
<thead>
<tr>
<th>num_of_prev_attempts</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>27426</td>
</tr>
<tr>
<td>1</td>
<td>3195</td>
</tr>
<tr>
<td>2</td>
<td>659</td>
</tr>
<tr>
<td>3</td>
<td>140</td>
</tr>
<tr>
<td>4</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Name: num_of_prev_attempts, dtype: int64
Figure 4 shows that the frequency of students who have cleared the module at first attempt have achieved good grades. However, there are many students who have attempted the modules more than once and almost 3 times and yet withdrawn from the courses. We also observe that there are students who have attempted the module almost more than twice or thrice have also passed the course. However, the number of students who have attempted it more than twice have mostly failed the course or withdrawn. There is therefore a correlation between these 2 variables and this needs to be explored further to find out what the reasons for the withdrawal of students and for also whether there the students who failed the course did so because they found the course module difficult even after several attempts. We also notice that most student who have achieved Distinction have cleared the module at first attempt with a small percentage at second attempt and a minimum percentage at 3rd attempt. This also proves that there was no specific limit to the number of times a student could attempt the module.

Figure 4 - Correlation of Final results with previous attempts
Distribution by percentage of students’ final results

In Figure 5, we observe the percentages of Pass, Fail and Distinction grades awarded to students. We observe that almost 31.5% of the students have withdrawn from the courses. The percentage of students achieving a Distinction is also quite small compared to the Pass grades.

Figure 5- Percentage Distribution of Final results

```
percent3 = df1['final_result'].value_counts()
percent3
```

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass</td>
<td>11829</td>
</tr>
<tr>
<td>Withdrown</td>
<td>9920</td>
</tr>
<tr>
<td>Fail</td>
<td>6902</td>
</tr>
<tr>
<td>Distinction</td>
<td>2825</td>
</tr>
</tbody>
</table>

Name: final_result, dtype: int64
Correlation between Final results and IMD Band

Students came from regions with different IMD Band. I grouped them per IMD Band and it was observed that majority of the students belonged to the regions with IMD Bands of 20-30% and 30-40% and 10-20%. The students that came from IMD Bands 90-100% were the least in number.

Figure 6- Distribution of students per IMD Band
It was observed that the 2 variables i.e. the Final results and IMD Bands did not show a very direct correlation. Figure 7 shows that the number of students passing in the courses does not show a major difference. However, the students who withdrew from courses are a higher among the IMD Bands from 0-10%, 10-20% and 20-30%. Figure 7 also shows that even though the rate of students obtaining a Distinction does not vary too much across the IMD bands the rate does show that the number increases from the lower bands to the upper bands.

It is also observed that the number of students who failed the exams is lesser among the higher IMD Bands and higher in regions with lesser IMD Bands.

*Figure 7- Correlation of Final results and imd_band*
Correlation between Age and Final results

The ratio of students failing who range from the age groups of lesser than 35 years is higher than that of ages 35-55. The number of students withdrawing from courses is also quite high among this age group.

Figure 8- Correlation of Final results with Age of students
Correlation between Gender and Final results

The number of female students is lesser than the male students. The ratio of grades is almost equal except for the Fail grades and students who withdrew are higher among the male students.

Figure 9-Correlation of Final results with Gender
Correlation between Highest Qualifications and Final results

The students’ highest qualifications at the time of admissions was recorded in 5 categories. Since most of the students belonged to the UK curriculum, A level or Equivalent was the most frequent highest qualification. It was observed as shown in Fig 10. There were some outliers in the A level or Equivalent highest qualifications in the category of Pass and Withdrawn. Most PG Qualified students ranked Distinction and the number of students who had lower than A level were higher in number in the withdrawn students.

Figure 10-Correlation of Final results and Highest Qualification of Students
- **Code Presentations and Code Modules** - For this exploration, the Student assessments and the student info dataset were used accordingly.

It is observed from Figure 11 that 2014 i.e. the Code Presentation offered in Oct 2014 was the most popular among students. And the least number of students enrolled in 2013B i.e. in February 2013.

*Figure 11- Distribution of Code_presentations*
Figure 16 shows the distribution of code_modules per code_presentation. In 2013B i.e. February 2013 only 3 modules were offered i.e BBB, DDD and FFF. The highest number of modules offered was offered in 2014 October. Course module CCC was not offered in October 2013 and February 2013. This could imply that these were new modules which were introduced later. The number of modules in 2013 October were higher than 2013 February and the number of modules offered in 2014 October were higher than 2014 February.

Figure 12-Distribution of Code Modules
• **Final results based on course modules**

Fig 17 below shows that the pass percentages in the BBB and FFF modules are relatively higher than other course modules with students enrolled in BBB showing the highest ratio of pass percentages. The number of students who withdrew from BBB, FFF and DDD also are higher than other modules. Students who received the Distinction grade too were higher and almost equal in BBB and FFF. The least popular was the AAA module with very less students enrolling in it. The failure rate was extremely high in modules BBB and FFF compared other modules.

*Figure 13- Distribution of Final results per course modules*
Interaction of students with VLE platform- For this exploration, the Student VLE dataset has been used.

Figures 14 and 15 below indicate the level of activity and the frequency of visits to the various pages on the VLE platform. It is observed that Forum MNG and the homepage are the most visited pages. Students also accessed the online content but Dataplus and other pages were not so frequently visited by students.

Figure 14- Frequency of clicks per VLE page
Figure 15 - Number of Student Clicks per VLE resource
Figure 16- Average number of clicks per VLE resource
Figure 17 shows the average number of days a student has visited the VLE page after the presentation date. It is observed that students take maximum time to visit the Dataplus page and Oucontent but least number of days to access the OuCollaborate page. Quizzes are also visited quite frequently as this maybe a requirement for fulfilling the assessments.

Figure 17-Average number of days student accessed the VLE before the exam
• Correlation of interaction of students with VLE platform and Final results

It is observed that students who achieve Distinction and Pass grades have visited the VLE more frequently than those who have received a Fail grade or withdrawn.

Figure 18- Correlation of Final results with student activity on VLE
- **Correlation of Final results and Student Disability**

In order to establish whether student’s disability affected their performance, a graphical presentation as shown in figures below showed that the ratio of Pass, Fail and Distinction were not that high in students with disabilities. Instead the number of students withdrawing from courses was higher in students with no disability.

*Figure 19- Distribution of number of students - Disability*
Figure 20- Correlation of final results with physical ability of students

Figure 21 shows that there is correlation between the highest qualification of students and their final grades. Even though it could be argued that the number of students who have no formal qualifications is comparatively lesser than the rest, the students who possess an A level or equivalent qualification fare much better than those with HE qualification and lesser than O level as seen in Figure 21 below. The rate of students receiving a Distinction is also much higher among students with A level or equivalent compared to the rest of the students.

➢ Correlation of Final results and highest qualification
Withdrawals and highest qualification

An important part of this project is to investigate the reasons for students withdrawing during the course of the program. Since the reasons are not mentioned explicitly in any of the datasets, inferences need to be drawn from the data exploration.

To explore the reasons that could have an effect on withdrawal rates among the students I filtered the dataset to gather data only on students who have withdrawn.

Figure 22 shows that the withdrawal rates among students with Lower than O level qualifications is the highest. Interestingly it is quite high also among A level or equivalent qualified students.
Figure 22- Correlation of Withdrawal rates and highest qualifications
Withdrawals and IMD Band

Figure 23 below indicates that the number of withdrawals is highest among students who come from regions where the IMD Band is between 0-10%, 10-20% and 30%-40%. Withdrawal numbers in students who come from regions with higher IMD Band is much lesser.

Figure 23- Correlation of withdrawal rates and imd_band
To further explore whether IMD Band also have an effect on failure rates among the students I filtered the dataset to gather data only on students with Fail grade. It was observed that the failure rates was highest among students with lower IMD Bands.

➢ Failure rates and IMD Band
Figure 25- Correlation of Failure rates and imd_band
The filtered dataset for Fail grade students also gave us an insight into other possible factors. It was observed as shown in Figure 28 below that the failure rate was quite high among students whose highest qualification was lower than A level.

**Figure 26- Fail Results and highest educational qualification**
Failure rates per Code modules

Figure 27- Fail results and code module
Hypotheses –

After careful exploration of the dataset this project hypothesis the following –

1. IMD Band has a direct correlation with the failure and the withdrawal rates of students.
2. Age of the students affects the choice of subjects and the final results.
3. Highest education has a strong correlation with the withdrawal rates and also affect the overall performance and retention of students in the courses.
4. The scores of the assessments determine the final score in the exam and eventually the final results.
D. Data Mining

Machine learning is a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms. (Qiu, J., Wu, Q., Ding, G. et al). There are three types of machine learning techniques i.e. supervised, unsupervised and reinforcement learning. For the purpose of this project, I will be using the supervised learning techniques since we have labeled data which has the inputs and the desired output.

I have categorized the variables as follows –

**Independent variables** – IMD Band, Age Band, Highest Education, Region, Score

**Dependent variable** – Final Result

Supervised learning techniques- Classification consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes).

For the purpose of training our model for this project I have used the following algorithms in Python –

**Random Forest classifier** – This is ensemble learning method that is used for classification and regression tasks. This algorithm creates multiple decision trees while training the data.

**Logistical Regression model** - Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

**Support Vector machines** - A support vector machine is a supervised learning algorithm that sorts data into two categories. This algorithm determines which category a new data point belongs in.

I will be training the model with each of these algorithms and finally determining the accuracy of each of the algorithms.
Before we train the model, we need to prepare or transform the data in a form that the machine learning algorithm will understand. The following are the steps that I have undertaken to prepare the data for machine learning.

A. Data Transformation

- Outlier Detection

Before we train the dataset, we need to find out if there are any Outliers in any of the attributes as this could skew the entire result and the predictions will be inaccurate.

We checked for 2 variables –

**Score** – The minimum value of the score was 0 and the maximum value was 100. The Score column was showing that there were some values that were much lesser than the minimum pass value of 40. On checking the scores of these students, it was found that the majority of these students had either failed or withdrawn from the courses. There was a small percentage of students who had passed but had scored less than 40 in the TMAs but no score was recorded for these students for the final exams. It is assumed that they scored a higher grade in the EMAs and hence secured a Final result of Pass. There were no students who scored more than 100.
Studied Credits- In order to find any outliers in the Studied Credits column we checked if there are any outliers above 150, the maximum no of credits that could be obtained were 480 credits for an Integrated Master’s degree that typically takes 6-9 years of study. But we found one student who had done 630 credit hours. It was observed that the student had a disability and it is assumed that he may have received a special permission to complete his course over an extended period of time. It was decided to remove this row from the dataset so that it does not interfere in the results of the machine learning process.
It is difficult to assign the titles of the degrees as we have only the information on credits obtained/studied by students. From the above figure, it is understood that programs offering 60 credits had the most enrolments followed by 90 credits and then 30 credits. Some of the courses that needed 30, 60 and 90 credits were as follows –

- Open University certificates – 60 credits
• Undergraduate modules - 30 or 60 credits
• Postgraduate certificates – 60 credits
• Postgraduate modules - 15, 30, 60 or 120 credits

We can infer that these were the popular courses among students but we cannot assign values as per the degrees because this information is not provided with the dataset.

• The number of students who had completed 120 credits were also marginally high and these are the programs that needed 120 credits for completion.
• Certificate of Higher Education (Cert HE) - 120 credits
• Postgraduate diploma - 120 credits

Since students who had completed 120 and above credits were lower, it is assumed that the following courses were not very popular.

• Masters degrees - 180 credits
• Diploma of Higher Education - 240 credits
• Honours degrees – 360 credits
• Integrated masters degree – 480 credits

• Encoding

Most machine learning algorithms require that all the variables in the input dataset and the output dataset are numeric. Since our dataset has labels, it was important to convert/ encode these into numeric values.

• IMD Band – The column had IMD Band in percentage range and this were converted to the mean of each band for e.g. 0-10% to 5, 10-20% to 15 and so on.
• **Age Band** – The column had ages in ranges and I converted it as follows – 0-35 to 1, 35-55 to 2 and 55<= to 3

```
inputs.age_band=inputs.age_band.replace({"0-35":1,"35-55":2,"55<=":3})
```

inputs
The values in other columns were all converted to numeric data by using the LabelEncoder function.
• **NA values**

I found out that there were NA values in the Date Registration and Date Unregistration columns which will interfere in the working of the algorithm and hence these were filled with 0.

```python
code_module_x 0
code_presentation_x 0
id_student 0
gender 0
region 0
highest_education 0
imd_band 0
age_band 0
num_of_prev_attempts 0
studied_credits 0
disability 0
date_registration 9
date_unregistration 19999
assessment_type 0
score 0
dtype: int64

inputs['date_registration']=inputs['date_registration'].fillna(0)
```

```python
inputs.isna().sum()
```

<table>
<thead>
<tr>
<th>code_module_x</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>code_presentation_x</td>
<td>0</td>
</tr>
<tr>
<td>id_student</td>
<td>0</td>
</tr>
<tr>
<td>gender</td>
<td>0</td>
</tr>
<tr>
<td>region</td>
<td>0</td>
</tr>
<tr>
<td>highest_education</td>
<td>0</td>
</tr>
<tr>
<td>imd_band</td>
<td>0</td>
</tr>
<tr>
<td>age_band</td>
<td>0</td>
</tr>
<tr>
<td>num_of_prev_attempts</td>
<td>0</td>
</tr>
<tr>
<td>studied_credits</td>
<td>0</td>
</tr>
<tr>
<td>disability</td>
<td>0</td>
</tr>
<tr>
<td>date_registration</td>
<td>0</td>
</tr>
<tr>
<td>date_unregistration</td>
<td>0</td>
</tr>
<tr>
<td>assessment_type</td>
<td>0</td>
</tr>
<tr>
<td>score</td>
<td>0</td>
</tr>
</tbody>
</table>

dtype: int64

**B. Feature Importance**

**Feature importance** refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. (https://machinelearningmastery.com/calculate-feature-importance-with-python/).
Feature importance is an important step as it gives us insights about the model and acts as a foundation for dimensionality reduction and feature selection which will improve the accuracy of the predictive model.

Correlation matrix
The heatmap in figure below shows the correlation between the independent variables and the dependent variable.

To narrow it down to the most important features which from the above figure show most correlation to the dependent variable i.e. Final Result, the following heatmap was plotted to create a visual representation of the correlation.
It was observed that most variables had a negative correlation to the Final Result. But Highest Education showed the most correlation to the Final Result.

Age Band and IMD Band showed a maximum correlation to the variable ‘Score’.

To reduce the dimensionality of the dataset it is better to reduce the dataset to include only the features which have the most correlation to the dependent variable to reduce noise and achieve the highest accuracy of the model.

C. Sampling

The dataset was divided into training and test datasets.

```python
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(inputs, target1, test_size=0.2)
```

The dataset was divided in training and test datasets with 80% of the dataset for training the model and 20% of the dataset for testing.
I ran the Random Forest Classifier on the training dataset. I tried to increase the estimators in various instances and ran the model to check if the accuracy improved.

```python
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier()
model.fit(x_train, y_train.values.ravel())

model.score(x_train, y_train)
0.9999498067560106

model.score(x_test, y_test)
0.49427825737803655
```
With number of estimators increased to 25

```python
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier(n_estimators=25)
model.fit(x_train, y_train.values.ravel())

RandomForestClassifier(n_estimators=25)

model.score(x_train,y_train)

0.9988455553882447

model.score(x_test,y_test)

0.5282071873117847
```

With number of estimators increased to 30

```python
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier(n_estimators=30)
model.fit(x_train, y_train.values.ravel())

RandomForestClassifier(n_estimators=30)

model.score(x_train,y_train)

0.9994980675601064

model.score(x_test,y_test)

0.527604898614736
```
D. Confusion matrix

I tried to establish the model prediction accuracy by plotting the confusion matrix –

In order to find out whether the accuracy of the prediction model increases if the sample size of the test data was increased, I tried to divide the dataset to a training set of 70% and test set of 30% of the dataset.

```python
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(inputs, target1, test_size=0.3)
```
```python
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier()
model.fit(x_train, y_train.values.ravel())

RandomForestClassifier()

model.score(x_train,y_train)
1.0

model.score(x_test,y_test)
0.5331905781584583

from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier(n_estimators=10)
model.fit(x_train, y_train.values.ravel())

RandomForestClassifier(n_estimators=10)

model.score(x_train,y_train)
0.9858306562643414

model.score(x_test,y_test)
0.4969218415417559

from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier(n_estimators=20)
model.fit(x_train, y_train.values.ravel())

RandomForestClassifier(n_estimators=20)
```
Logistic Regression –

```python
model.score(x_train, y_train)

0.9985084901330886

model.score(x_test, y_test)

0.5108404710920771

lr=LogisticRegression()
lr.fit(x_train, y_train.values.ravel())
lr.score(x_train, y_train)

0.4662952366611454

lr=LogisticRegression()
lr.fit(x_train, y_train.values.ravel())
lr.score(x_test, y_test)

0.47038747239510137
```
Support Vector Machines

```python
svm=SVC()
svm.fit(x_train,y_train.values.ravel())
svm.score(x_train,y_train)
```

0.46624504341715606

```python
svm=SVC()
svm.fit(x_train,y_train.values.ravel())
svm.score(x_test,y_test)
```

0.47038747239510137

Since the algorithm selects a random sample at every iteration, the accuracy also changes. To avoid this, I used the K-Fold cross validation technique with a standard no of split value of 10.

```python
def kfold(n_splits=10):
kf

KFold(n_splits=10, random_state=None, shuffle=False)

for train_index, test_index in kf.split(inputs, target):
    print(train_index, test_index)
```

```
[2491 2492 2493 ... 24901 24902 24903]
[0 1 2 ... 2488 2489 2490]
[0 1 2 ... 24901 24902 24903]
[2491 2492 2493 ... 4979 4980 4981]
[0 1 2 ... 24901 24902 24903]
[4982 4983 4984 ... 7470 7471 7472]
[0 1 2 ... 24901 24902 24903]
[7473 7474 7475 ... 9961 9962 9963]
[0 1 2 ... 24901 24902 24903]
[9964 9965 9966 ... 12451 12452 12453]
[0 1 2 ... 24901 24902 24903]
[12454 12455 12456 ... 14941 14942 14943]
[0 1 2 ... 24901 24902 24903]
[14944 14945 14946 ... 17431 17432 17433]
[0 1 2 ... 24901 24902 24903]
[17434 17435 17436 ... 19921 19922 19923]
[0 1 2 ... 24901 24902 24903]
[19924 19925 19926 ... 22411 22412 22413]
[0 1 2 ... 22411 22412 22413]
[22414 22415 22416 ... 24901 24902 24903]
```
The average metrics for accuracy after training the model and running the prediction on the test dataset with 3 different algorithms is as follows.

<table>
<thead>
<tr>
<th>Train/Test Data</th>
<th>Random Forest Classifier</th>
<th>Support Vector Machines</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data</td>
<td>0.998</td>
<td>0.466</td>
<td>0.466</td>
</tr>
<tr>
<td>Test Data</td>
<td>0.534</td>
<td>0.460</td>
<td>0.465</td>
</tr>
</tbody>
</table>
Overfitting of Model-

As evident from the above table, I detected a problem of Overfitting of the model where in the accuracy was much higher in the training dataset but when the algorithm is introduced to the test data, the accuracy is much lower. To eliminate this problem, I tried to fit the model by implementing the following steps-

**Dimensionality reduction**

While analyzing the correlation of the different variables in my original dataset, it was found that not all features were important and only a few features had maximum correlation to the dependent variable. In order to reduce the noise in the dataset, I dropped some features and a new dataset was created with the following features.

```python
inputs1=inputs.drop(['Name: 0', 'code_module', 'code_presenation', 'id_student', 'gender', 'region', 'date_registration'], axis=columns)
```

The null values in the date_unregistration column were filled with 0. I retained this feature in the dataset as this shows after how many days the student has withdrawn from the course.

```python
inputs1['date_unregistration']=inputs['date_unregistration'].fillna(0)
inputs1.isna().sum()
```

We again transformed the target variable dataset i.e final_result into numeric form so that the algorithms will identify the variable.

```python
target['final_result']=le_final_result.fit_transform(target['final_result'])
target
```
I ran the Random Forest Classifier algorithm again on the training and test dataset that was derived after splitting the dataset by using the Kfold cross validation method.

```python
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier(n_estimators=10)
model.fit(x_train, y_train.values.ravel())

RandomForestClassifier(n_estimators=10)

model.score(x_train,y_train)
0.9452391708076093

model.score(x_test,y_test)
0.6611122264605501
```

```python
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier(n_estimators=25)
model.fit(x_train, y_train.values.ravel())

RandomForestClassifier(n_estimators=25)

model.score(x_train,y_train)
0.957636902072981

model.score(x_test,y_test)
0.6675366392290705
```

It was found that the accuracy improved from **0.534** to almost **0.66**. I also tried to increase the number of estimators to 25 and the results were almost slightly higher when the number of estimators was increased to 25. The highest accuracy on test data was achieved when the number of estimators was increased to 40.
rf=RandomForestClassifier(n_estimators=40)
rf.fit(x_train,y_train.values.ravel())
rf.score(x_train,y_train)

0.9603473372484064

rf=RandomForestClassifier(n_estimators=40)
rf.fit(x_train,y_train.values.ravel())
rf.score(x_test,y_test)

0.6705480827143144

y_pred=rf.predict(x_test)

from sklearn import metrics
print("Accuracy: ",metrics.accuracy_score(y_test,y_pred))

Accuracy: 0.6705480827143144
With the help of K-fold cross validation with number of folds set to 10, we tried to find out the accuracy of the 3 different algorithms -

**Logistic Regression** –

```python
get_score(logisticRegression(), x_train, x_test, y_train.values.ravel(), y_test.values.ravel())
```

0.6583015458743224

**Support Vector Machines**

```python
get_score(SVC(), x_train, x_test, y_train.values.ravel(), y_test.values.ravel())
```

0.674563486946396

```python
get_score(RandomForestClassifier(n_estimators=40), x_train, x_test, y_train.values.ravel(), y_test.values.ravel())
```

0.6731656293916884

**Hyperparameter Tuning** –

In order to reduce the time spent on finding the best parameters and combinations manually, I used GridSearchCV function from Python to derive the best parameter for the Random Forest Classifier.

```python
rf_Grid.best_params_
```

```
{'bootstrap': True,
'max_depth': 4,
'max_features': 'sqrt',
'min_samples_leaf': 2,
'min_samples_split': 5,
'n_estimators': 72}
```
After converting the parameters of the Random Forest Classifier algorithm to the above, with 3-fold cross validation, and fitting the model the following results for accuracy were derived-

```python
print(f'Train Accuracy - {rf_Grid.score(x_train,y_train):3f}')

Train Accuracy - 0.691111

print(f'Test Accuracy - {rf_Grid.score(x_test,y_test):3f}')

Test Accuracy - 0.697852
```

It was observed that the accuracy on the training and test data while using Random Forest Classifier Algorithm with 3-fold cross validation had improved to 0.69. It was also higher than the accuracy for Support Vector machines and Logistic Regression.
Results

After extensive data exploration the following results were derived-

1. **Number of previous attempts** - Students’ failure and withdrawal rates show correlation to number of previous attempts. This could mean that the students found the course content difficult and hence eventually failed or withdrew from the course. This warrants an intervention during the course of the program so that the students who find difficulty in understanding the course content can be provided with more resources or guidance from the instructors. This will help improve student retention and their overall academic performance.

2. **IMD Bands** - Withdrawal rates from regions with lower IMD bands was higher. IMD band stands for the Index of Multiple Deprivation. It considers the income of residents, their education level and crime levels in areas. The students gaining higher grades i.e. Distinction in the given dataset was higher among those that came from regions with a higher IMD band. Even though failure rates did not have a direct correlation, it can be argued that students who come from regions with better facilities, lower crime levels and better income maybe at an advantage to other students. It may be necessary for decision makers to consider the social and economic background of students through the demographic records to monitor their progress periodically and lend them adequate support if they exhibit poor academic performance. The advisors and student services should be provided with this data so that they can intervene early in the academic life of a student and offer them additional support if necessary to improve their grades and help them in their overall student life.

3. **Last qualification/ pre-requisites** - It was found that students who had their highest qualification at A levels or higher could cope better with the program and successfully completed it. Students who had completed less than A levels mostly withdrew from the course. We could infer that either the pre-requisites before enrolling the program should be clearly defined so that the students will be aware of the skills or knowledge required for the courses or a remedial/ bridge program must be offered to the students who do not satisfy the conditions at entry. This can substantially reduce withdrawal rates and high school students will be able to complete the courses with adequate support.
4. **Curriculum** – The dataset also revealed that some students preferred and excelled in particular course modules. This could lend valuable insights to decision makers to design the curricula or periodically assess the effectiveness and structure of the curriculum and modify it as needed. This will also increase the viability of programs offered by higher education institutions.

5. **Virtual Learning** - Students are getting more accustomed to access online academic resources mainly due to the COVID 19 pandemic where universities had to move to either blended or complete online mode of teaching. Students could not visit a physical library and had to procure these resources from an online source. It has been observed that students who frequently accessed these resources before the exams performed better than those students who accessed the resources later or for lesser number of times. This proves that students who read and access additional resources/ course material before exams or during the course of the program can do better. This trail of information gathered from the VLE is really important for instructors and advisors to intervene early during the course of the semester and encourage students to access the learning management software and refer to the resources or readings which will help them in their exams.

6. **Predictive model** – It was observed that Feature Engineering/ Importance had a major impact on the accuracy of the model. The model encountered an issue of overfitting which was reduced to some extent by feature importance and hyperparameter tuning. Eliminating noise from the data improved the accuracy marginally.
Conclusions and Future Work

The capstone project was aimed at analyzing data gathered from a learning management software and creating a predictive model that could assist higher educational institutions make better decisions and help students. I would have been satisfied with a much higher accuracy of the model even though I have tried various combinations and algorithms. I also wanted to draw conclusions from the frequency of interaction of students with the virtual learning environment and the final results but I had to drop this feature as the information was not recorded for all students. The dataset also lacked qualitative information as to the reasons for withdrawal which could have lent further insights. However, the dataset did provide useful results and gives a foundation for future related work in this area.

This study has drawn conclusions that the data gathered from the students both demographic and course related is essential to track their progress and assist them during their academic journey. The stakeholders can capitalize on this data to not only improve the academic progress of students but also improve the curriculum and the student retention and academic performance. If the study is supported by a bigger and more detailed data from a higher education institution by anonymizing the data, it will provide benefits to the Ministry of Education in UAE and to higher education institutions.
Bibliography


VI. Siemens, Dawson, & Lynch. (2013), Improving the quality and productivity of the higher education sector: policy and strategy for systems-level deployment of learning analytics


Websites –

https://analyse.kmi.open.ac.uk/open_dataset

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