**ANALYSIS OF OPEN UNIVERSITY LEARNING DATASET**

**Introduction**

The Open University (OU) is among the major universities in the UK, dedicated to distance education and the provision of educational opportunities to marginalized and varied social groups. A new Virtual Learning Environment has been implemented at OU with the goal of raising the bar for student learning and academic success. This paper analyzes the impact of the VLE on fluctuations in students' grades, utilizing engagement metrics with the system as predictors.

**Research Questions and Objectives**

The primary objectives of this study are to:

1. Assess whether the new VLE improves students’ grades.
2. Determine the extent to which students' grades can be predicted using their interaction data and other attributes.

To address these questions, the following hypotheses are formulated:

* **Null Hypothesis (H₀):** The interaction level with the VLE has no significant impact on students' grades.
* **Alternative Hypothesis (H₁):** Increased interaction with the VLE significantly improves students' grades.

**Methodology**

These methods involve the implementation of various machine learning and statistical techniques on the data from Open University's educational ecosystem. Among the several methodologies that can be employed, this research aims at implementing the following:

* Data cleaning and wrangling to ensure dataset integrity.
* EDA of module participation and scoring, and use of VLEs to identify emerging patterns and trends.
* Statistical hypothesis testing to establish the relationship between VLE interaction and academic performance.
* Decision tree classification to predict the success of students based on key features.

**Context and Dataset Overview**

The Open University puts a lot of emphasis on learning from home in all of its programs. Exams, Tutor Marked Assessments (TMA), and Computer Marked Assessments (CMA) are all types of assessments that count toward the final grade. The VLE makes these things easier by giving students resources and interactive tools to help them with their studies.

The dataset includes:

1. **Assessments**: Information on module assignments, their types, and weights.
2. **Courses**: Module details, durations, and presentations.
3. **Student Assessments**: Individual assessment scores and submission data.
4. **Student Info**: Demographics, educational background, and final results.
5. **Student Registration**: Registration times of student to module.
6. **Student VLE**: Interaction logs detailing student clicks and engagement with VLE materials.
7. **VLE**: Available materials in the VLE.

In this stage of the data exploration phase, key trends include the most popularly selected modules, trends in average scores in these modules, and interactions on a weekly basis. These then form the building blocks for both analysis and model building stages.

In the following sections, we take a deeper look into the data and methodologies, culminating in actionable recommendations that enhance the VLE's impact on student outcomes.

**2. Data Exploration**

**2.1 Data Cleaning and Wrangling**

The initial dataset was examined for completeness, consistency, and accuracy to ensure reliability in subsequent analyses. To clean and wrangle the data, it is preferred to analyse all of them one-by-one then merge them. Below are the steps and methods applied to clean and wrangle the data:

**2.1.1 Data Cleaning Steps**

1. **Missing Values:**
   * Columns with missing data searched and there is not any NULL values in the data.
   * For numeric features like score, non-numeric values such as '?' were replaced with NaN and removed.
   * Students who withdrawn the course were excluded from the data, since they do not complete the course, it makes irrelevant for the analysis.
2. **Outlier Detection and Removal:**
   * The VLE interactions, and number of previous attempts are one of the main features to investigate the outliers.
   * Outliers were detected using the Interquartile Range (IQR) method.
   * For sum\_click (total VLE interactions), values beyond the IQR were removed to ensure robust analysis.
3. **Duplicate Rows:**
   * To investigate duplicate entries, first it is required to find the unique identifiers. Hence based on some unique identifiers like id\_student, code\_module, code\_presentation for each student, it is detected that there are 128192 rows duplicated.
   * By keeping the first index method, those rows are eliminated from the datasets.
4. **Data Type Consistency:**
   * Categorical features like highest\_education and final\_result were mapped to numeric values for analysis and modeling.
5. **Inconsistent Cases:**
   * The dataset contains some inconsistent cases. Those cases can be grouped into two:
     1. diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü, çizgi içeren bir resim

        Açıklama otomatik olarak oluşturuldu**öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, metin, ekran görüntüsü içeren bir resim

        Açıklama otomatik olarak oluşturuldu**The students who have the average grade of 0 out of 100, but pass the courses. Therefore, it is required to detect those students before modelling the datasets. Hence the total of 12,893 students were passed their module even though they get 0, they are removed from the dataset. The difference between Figure 1 and Figure 2 is actually put the difference.

Figure 1 Students Score Distribution With Inconsistent Cases

Figure 2 Students Score Distribution Without Inconsistent Cases

* + 1. The dataset contains wrong information as well. According to domain knowledge, which is gathered from Open University Learning Analytics Dataset website, the students with an average score lower than 40, must fail the course otherwise they pass. For this reason, the “final\_result” column were edited accordingly.

**2.1.2 Data Wrangling**

1. **Merging Datasets:**
   * Relevant datasets merged. First, with a function, it is founded the mutual columns of datasets. Then based on the common oclumns, it is merged. For instance, student\_VLE data were merged with a student\_data by finding common columns such as 'code\_module', 'code\_presentation', 'id\_student'.
   * Key features such as avg\_score and sum\_click were calculated and added to the final dataset.
2. **Feature Engineering:**
   * Created a new column interaction\_level, which categorizes students into "High" or "Low" interaction groups based on the median of sum\_click.
   * Average scores for each student were calculated by weighting assessment scores based on the type of assessment (e.g., exams vs. TMAs).
     + - ekran görüntüsü, metin, renklilik, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

         Açıklama otomatik olarak oluşturulduAverage scores calculated based on student’s module. Since only the students who take the CCC and DDD module have an exam type assessment, as clearly stated in Figure 3, the algorithm works based on the assessment.

Figure 3 Average Score by Assessment Type and Code Module

1. **Final Dataset Summary:**
   * After cleaning and wrangling, the dataset was reduced to 17,191 rows with 14 relevant columns.
   * The cleaned dataset was verified to have no missing values or duplicates.

**2.2 Exploratory Data Analysis (EDA)**

The cleaned dataset was analyzed to uncover patterns and trends, providing insights into the relationship between VLE interactions, student demographics, and academic performance.

**2.2.1 Overview of the Dataset**

* **Modules:** The dataset covers seven modules across multiple presentations.
* **Students:** Data includes demographics, prior education, and engagement with the VLE.
* **Assessments:** Types include TMAs, CMAs, and exams, each with specific weights.

**2.2.2 Key Insights**

1. **Top 5 Modules Chosen by Students:**
   * Below, the figure 4 show the top 5 modules chosen by students. The top module is BBB with 4869 students, followed by FFF with 3652 students, then DDD with 3548 students, CCC having 2151 students, and lastly the total of 1931 students with EEE module.

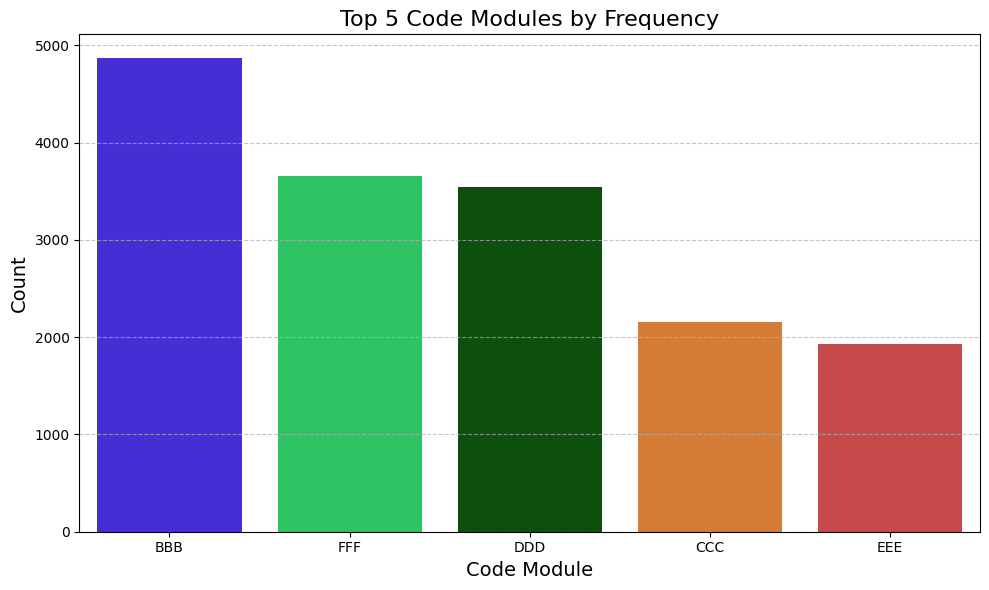


Figure 4 Top 5 Modules Chosen by Students

1. **Top 5 Modules with Highest and Lowest Average Scores:**
   * Highest average scores were observed in EEE, followed by FFF, BBB, AAA, and CCC.
   * ekran görüntüsü, metin, çizgi, diyagram içeren bir resim

     Açıklama otomatik olarak oluşturulduekran görüntüsü, çizgi, renklilik, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

     Açıklama otomatik olarak oluşturulduLowest average scores were recorded in GGG since they do not take any exam, therefore we can say DDD and CCC modules have the lowest average score.

Figure 6 Top 5 Modules by Lowest Average Scores

Figure 5 Top 5 Modules by Highest Average Score

1. **Modules with the Most Failures:**
   * metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

     Açıklama otomatik olarak oluşturulduDDD and CCC had the highest number of students failing their modules as it is shown in the Figure 7.

Figure 7 Top 5 Module with Most Failures

1. **Age Distribution of Students:**
   * Students were grouped into age bands: 0-35, 35-55, and 55<=, clearly showed in the figure 8.
   * ekran görüntüsü, metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

     Açıklama otomatik olarak oluşturulduA majority (72%) belonged to the 0-35 age group, followed by 35-55 (27%).

Figure 8 Age Band of Students

1. **Weekly VLE Interaction Patterns:**
   * öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, ekran görüntüsü, diyagram içeren bir resim

     Açıklama otomatik olarak oluşturulduFigure 9 explain the students’ VLE interaction, and it is interaction peaked during the first one and half months.

Figure 9 Wekkly VLE Interactions

**2.2.3 Summary Statistics**

* **Assessment Distribution:** The majority of assessments are TMAs, followed by CMAs and exams.
* **Clicks and Engagement:** Students with higher interaction levels (sum\_click) tended to have higher average scores.
* **Failures and Pass Rates:** Modules with low engagement often had higher failure rates.

**3. Methods**

This section outlines the methodological approach to analyze the impact of the Virtual Learning Environment (VLE) on students’ grades and to predict student performance. The methods were designed to explore hypotheses and develop predictive models.

**3.1 Hypothesis Testing**

**3.1.1 Research Hypotheses**

The hypotheses were developed to assess the correlation between VLE interaction levels and student performance on final outcomes.

* **Null Hypothesis (H₀):** The interaction levels with the VLE (sum\_click) have no significant effect on students' probability of passing a course. The mean sum\_click is consistent across all final\_result categories.
* **Alternative Hypothesis (H₁):** Increased interaction levels with the VLE (sum\_click) correlate positively with course passing rates among students. The mean sum\_click varies among final\_result categories.

**3.1.2 Choice of Statistical Test**

An ANOVA (Analysis of Variance) test was selected to compare the mean interaction levels (sum\_click) among the final\_result categories (Pass, Fail).

* Rationale for Employing ANOVA:
  + The dependent variable (sum\_click) is continuous. The independent variable (final\_result) is categorical, consisting of two groups: Pass and Fail.
  + ANOVA assesses the significance of differences in the means of sum\_click among various groups.

**3.1.3 Assumptions for ANOVA**

Each student's interaction with the VLE is independent of the interactions of others.   
The distribution of sum\_click is approximately normal within each final\_result group. The variance of sum\_click is consistent across groups.

**3.1.4 Steps in Hypothesis Testing**

1. Preprocessed the data by categorizing interaction\_level into "High" and "Low" based on the median of sum\_click.
2. Grouped the data by final\_result to compute mean and variance.
3. Performed ANOVA using the f\_oneway function from SciPy to compare the means.
4. Interpreted the p-value:
   * When p is less than 0.05, the null hypothesis should be rejected in favor of the alternative hypothesis.

**3.2 Statistical Models**

**3.2.1 Predictive Model**

The Decision Tree Classifier was selected as the predictive model to determine the pass or fail outcomes of the students in the course. This decision was based on the necessity to integrate various factors, including VLE interaction data denoted by sum\_click, demographic variables indicated by highest\_education, and academic performance metrics exemplified by code\_module and num\_of\_prev\_attempts. The Decision Tree algorithm partitions data into subsets according to feature values and constructs a tree-like framework of decisions and outcomes. The chosen model proved highly appropriate for this issue, as it adeptly managed both categorical and numerical variables, thereby aligning seamlessly with the provided dataset. By examining various splits at each decision node, the model identifies the most significant predictors of success or failure in student classification.

**3.2.2 Decision Tree - Justification**

The decision tree classifier was chosen because of the following reasons:

**Suitability for Categorical Data:**

* Decision trees do not necessitate preprocessing of features, including normalization or scaling, allowing them to directly accommodate both categorical and numerical features. These are thus suitable for this dataset, which includes predictors that vary from categorical variables such as highest\_education and code\_module to numerical variables like sum\_click and avg\_score.

**Interpretability:**

* Decision trees provide a visual method for prediction that is highly intuitive and straightforward. The imposition of conditions on features at each split facilitates straightforward forecasting by tracing a path from the root to the leaf. This interpretability enables a diverse group of stakeholders, such as administrators and educators, to understand the key factors affecting children's success or failure without requiring extensive technical knowledge.

**Non-Linearity Handling:**

* Decision trees can naturally capture nonlinear relationships between the features and the outcome. For instance, the relation between sum\_click and final\_result is definitely not linear (which also prevents to use Linear Regression algorithm, represented in Figure 10.), and the decision tree detects thresholds or other patterns—for example, specific ranges metin, ekran görüntüsü, açık mavi, çizgi içeren bir resim

  Açıklama otomatik olarak oluşturulduof sum\_click that predict Passing or Failing.

Figure 10 VLE Interactions and Average Score Relationship

* This flexibility ensures that the model can adapt to complex interactions among predictors, hence being robust for real-world educational data.

By capitalizing on these strengths, the decision tree classifier made a very pragmatic and efficient tool for predicting student performance based on their interaction with the VLE and demographic profile.

**3.2.3 Feature Selection**

The following features were considered for the model:

1. code\_module: Module code indicating the course.
2. highest\_education: Student's highest level of education.
3. num\_of\_prev\_attempts: Number of previous attempts at a module.
4. sum\_click: Total VLE interactions.

**3.2.4 Assumptions for Decision Tree**

* **Minimal Multicollinearity:** The features used in the model should not exhibit strong multicollinearity, ensuring each predictor contributes unique information to the decision-making process.
* **Adequate Data Volume:** The dataset must be sufficiently large to train the model effectively, minimizing the risk of overfitting and ensuring generalizability to unseen data.

**3.2.5 Model Calibration**

* **Hyperparameters:** The maximum depth of the tree was set to 5 to balance model complexity and interpretability.
* **Feature Encoding:** Categorical feature like code\_module were converted into numeric format using pandas get\_dummies function.
* **Data Splitting:** The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.

**3.2.6 Metrics for Evaluation**

* **Accuracy:** Percentage of correctly predicted outcomes.
* **Precision:** Proportion of true positives among predicted positives.
* **Recall:** Proportion of true positives among actual positives.
* **F1-Score:** Harmonic mean of precision and recall.
* **Confusion Matrix:** Visualization of true positives, false positives, true negatives, and false negatives.

**3.3 Feature Selection Techniques**

**3.3.1 Recursive Feature Elimination (RFE)**

* Used to rank features by recursively removing the least important feature and fitting the model on the remaining features.
* Output included accuracy scores for models with varying numbers of features, helping identify the optimal set of predictors.

**4. Results**

The results section presents the outcomes of the analysis, hypothesis testing, and predictive modeling. These findings are supported by statistical summaries and visualizations, emphasizing the relationship between Virtual Learning Environment (VLE) interactions and student performance.

**4.1 Hypothesis Testing Results**

**ANOVA Results**

The **ANOVA test** was conducted to determine whether students' interaction levels with the VLE (sum\_click) significantly impacted their likelihood of passing a course.

* **Null Hypothesis (H₀):** Interaction levels (sum\_click) do not significantly affect students’ likelihood of passing.
* **Alternative Hypothesis (H₁):** Higher interaction levels (sum\_click) increase the likelihood of passing.

**Test Outputs**

* **F-statistic:** 29424.06565252733
* **p-value:** 0.0 (less than 0.05)

**Conclusion:**

* This p-value reflects that there is a statistically significant difference in VLE interactions between students who passed versus those who failed.
* **Implication:** Students who used the VLE more were more likely to pass their courses.

**Additional Insights:**

* **Success Rate by Interaction Level:**
  + **High Interaction:** 79.09%
  + **Low Interaction:** 63.26%
* **Fail Rate by Interaction Level:**
  + **High Interaction:** 20.91%
  + **Low Interaction:** 36.74%
* Even though in the Figure 11, it is not too clear, the as the VLE Interactions of any students decreases, their chance to fail the class is increased by 43.09%.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu**Visualization:** A bar chart (Figure 11) was created to illustrate the relationship between interaction levels and final\_result, highlighting the difference in mean sum\_click values.

Figure 11 The VLE Interaction Levels by Final Result

**4.2 Predictive Model Results**

**4.2.1 Decision Tree Classifier**

The decision tree classifier was trained to predict whether a student would pass or fail based on features like sum\_click, num\_of\_prev\_attempts, code\_module, and highest\_education. The figure 12 showed how Decision Tree Classifier with a maximum depth of 5 plotted.

* **Training Accuracy:** 90%
* **Testing Accuracy:** 90.20%

ekran görüntüsü, çizgi, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 12 Decision Tree Graph

**4.2.2 Performance Metrics**

* **Precision:** 86.94%
* **Recall:** 91.67%
* **F1-Score:** 88.66%

**Confusion Matrix:**

The confusion matrix Figure 13 showed:

* True Positives: Students correctly predicted to pass.
* True Negatives: Students correctly predicted to fail.
* metin, ekran görüntüsü, diyagram içeren bir resim

  Açıklama otomatik olarak oluşturulduFalse Positives and False Negatives: Misclassified outcomes.

Figure13 Confusion Matrix

**Feature Importance:**

Recursive Feature Elimination (RFE) identified the following as the most significant predictors:

1. sum\_click
2. code\_module\_CCC
3. code\_module\_DDD
4. code\_module\_GGG

**4.4 Summary of Key Findings**

1. **VLE Interactions Improve Performance:**
   * Higher engagement with the VLE (sum\_click) is strongly associated with better academic performance.
   * Students with a high level of interaction had a 25.02% higher success rate compared to those with low interaction.
2. **Challenging Modules:**
   * Modules GGG, DDD, and CCC had the highest failure rates and lowest average scores, indicating the need for targeted interventions.
   * Additionally, an investigation is warranted as students in module GGG have all avg\_score values at 0.
3. **Success in Predictive Models:**
   * The decision tree classifier achieved 90.75% accuracy in predicting students' outcomes, with sum\_click along with CCC, DDD, and GGG code modules being the most impactful variable.

**5. Discussion**

This section contextualizes the previous results, examines limitations, and provides actionable recommendations. This section evaluates the implications of the findings for enhancing the Virtual Learning Environment (VLE) and student performance.

**5.1 Result Interpretation**

**5.1.1 Impact of VLE Interaction on Student Performance**

* The ANOVA test demonstrated a statistically significant correlation between sum\_click (VLE interaction) and final\_result (Pass/Fail), suggesting that increased VLE interaction enhances the probability of passing.
  + **Key Insight:** High engagement with the VLE, as indicated by sum\_click, serves as a significant predictor of success.
  + **Actionable Implication:** Encouraging active use of the VLE can serve as an effective strategy for improving student outcomes.

**5.1.2 Modules with High Failure Rates**

* Modules DDD and CCC, excluding GGG, which had an average score of 0 for all students, exhibited the lowest average scores and the highest failure rates.
  + **Key Insight:** These modules may present distinct challenges, including elevated difficulty levels, issues in content design, or insufficient student preparedness.
  + **Actionable Implication:** Redesigning and delivering these modules more effectively, along with providing additional support resources, can enhance outcomes.

**5.1.3 Predictive Model Insights**

* The decision tree classifier predicted student outcomes with 90% accuracy.
  + **Top Predictors:** sum\_click, code\_module\_CCC, code\_module\_DDD and code\_module\_GGG.
  + **Key Insight:** Engagement data, combined with demographic and academic background, can reliably predict performance.
  + **Actionable Implication:** Integrating predictive analytics can help identify at-risk students early, enabling timely interventions.

**5.2 Contextual Implications**

**5.2.1 VLE Benefits**

* The VLE serves as an essential instrument for course delivery and student engagement. The positive correlations between VLE interactions and academic success highlight its significance.
  + **Recommendation:** Improve the accessibility and usability of the VLE to enable more efficient navigation and interaction for students.

**5.2.2 Insights for Course Design**

* Courses with high failure rates can benefit from targeted improvements:
  + Adding interactive elements in DDD and CCC to enhance engagement.
  + Providing comprehensive guidelines for assessments, especially in modules featuring intricate CMAs, TMAs, and examinations.

**5.2.3 Predictive Tools for Student Success**

* Integrating the predictive model into the VLE can provide:
  + Early warnings for students with low engagement.
  + Personalized recommendations to enhance interaction and performance such as finding the least visited module weekly.

**5.3 Limitations of the Analysis**

1. **Data Scope:**
   * The dataset was confined to particular modules and presentations, potentially omitting variations in performance across other courses or academic years.
   * **Mitigation:** Future analyses should incorporate a wider array of modules and longitudinal data.
2. **Potential Confounding Variables:**
   * Factors such as student motivation and external commitments were not accounted for, potentially affecting outcomes.
   * **Mitigation:** Enhance analysis through supplementary surveys or qualitative data.
3. **Model Limitations:**
   * The decision tree classifier, although interpretable, may not effectively capture complex relationships compared to ensemble methods such as Random Forests.
   * **Mitigation:** Experiment with advanced models to improve predictive performance in future work even though the accuracy is quite high.

**5.4 Actionable Recommendations**

1. **Encourage VLE Engagement:**
   * Implement incentives for regular VLE participation, including gamified components or progress monitoring dashboards.
   * Organize workshops to educate students about VLE functionalities.
2. **Support for Challenging Modules:**
   * Provide additional resources such as tutorial videos, discussion forums, and one-on-one mentoring for DDD and CCC.
   * Assessments should be reorganized to better reflect students' skills.
3. **Implement Predictive Analytics:**
   * Utilize the decision tree model to identify at-risk students promptly and deliver targeted interventions, including personalized reminders or study strategies.
4. **Iterative VLE Improvement:**
   * Use student feedback to continuously enhance the VLE’s interface and content delivery mechanisms.

**6. Conclusion**

The purpose of this report was to investigate the influence that the Virtual Learning Environment (VLE) has on the academic performance of students attending the Open University and to develop a predictive model that can identify students who are at risk. Insights that can be put into action to improve student outcomes and refine the use of the virtual learning environment are provided by the findings.

**6.1 Summary of Findings**

**1. VLE Interaction and Performance:**

* A statistically significant relationship was found between VLE interactions (sum\_click) and student success. Students with higher interaction levels had an 25.02% higher success rate than those with lower interactions.
* **Conclusion:** Consistently encouraging meaningful engagement with the VLE can significantly enhance academic outcomes.

**2. Challenging Modules:**

* Modules DDD and CCC exhibited the highest failure rates and lowest average scores.
* **Conclusion:** These modules require targeted interventions, such as redesigning course content or enhancing support structures.

**3. Interesting Module:**

* Module GGG students get 0 on their average score.
* **Conclusion**: This module requires further investigations, to check why they all get 0.

**4. Predictive Model Success:**

* The decision tree classifier achieved 90% accuracy, reliably predicting student outcomes using features like VLE interactions, prior education, and module enrollment.
* **Conclusion:** Predictive analytics enables early identification of at-risk students, allowing timely and effective interventions.

**6.3 Future Work**

**1. Expand the Dataset:**

* Include data from more modules and presentations over multiple years for a comprehensive analysis.

**2. Advanced Modeling Techniques:**

* Experiment with ensemble methods, such as Random Forest and Gradient Boosting, to improve predictive accuracy.

**3. Add Qualitative Factors:**

* Collect additional data on students' motivation, external commitments, and learning preferences to enrich the analysis.

**6.4 Final Thought**

The findings highlight the key role that the virtual learning environment plays in facilitating academic achievement. Through the implementation of the actions that have been recommended, the Open University will be able to improve the educational opportunities it provides and better cater to the requirements of its diverse student body. The virtual learning environment (VLE) will continue to be an efficient and influential tool for distance learning if it is subjected to ongoing analysis and innovation.