**AIE 425 Intelligent Recommender Systems, Fall Semester 24/25**

**Personalized-Learning Path Recommendation Engine**

221101341, Faris Hassan Mohamed Hassan El Molla

221101218, Ahmed Osama Mahmoud

**Table of Contents:**

**Introduction………………………………………………………………………………….3**

**Data Description…………………………………………………………………………….4**

**Key Tables and Columns………………………………………………………………….7**

**Data Preprocessing……………………………………………………………………….10**

**Data Preparation…………………………………………………………………………...13**

**Methodology………………………………………………………………………………...14**

**Recommendation Algorithm………………………………………………………….…15**

**Discussion………………………………………………………………………………….16**

**Results………………………………………………………………………………………19**

**Conclusion………………………………………………………………………………….20**

**References………………………………………………………………………………….20**

**Introduction**

The Open University (OU) is a pioneering institution in higher education, renowned for its commitment to accessibility and innovation. Founded in 1969, the Open University is the UK’s largest university by student population and has seen more than 2 million students achieve their educational goals through its courses since its inception. A public research university, the University of Oklahoma primarily serves undergraduate students and offers flexible learning opportunities outside the classroom. Our university is a truly international university, offering undergraduate and postgraduate education to students from all over the world. The university is headquartered in Walton Hall in Milton Keynes, Buckinghamshire, but has administrative offices throughout the UK and in many European countries. The Open University offers a unique opportunity to personalise students’ learning by bringing together modules from different disciplines, thereby increasing its appeal to a wider audience.

In recent years, the OU has embraced the potential of digital technologies to enhance the learning experience for its students. The university’s Virtual Learning Environment (VLE) serves as a cornerstone of its educational delivery, enabling students to access course materials, participate in forum discussions, submit assessments, and track their academic progress. The VLE generates vast amounts of data, capturing student interactions and engagement patterns, which can be leveraged to improve educational outcomes. This has led to the development of the Open University Learning Analytics Dataset (OULAD), a comprehensive resource that provides insights into student behavior and performance across seven selected courses. The dataset, which includes data from both February and October course presentations (denoted by "B" and "J," respectively), offers a unique opportunity to explore the role of VLEs in supporting student success.

The application of data mining techniques in educational settings has gained significant traction in recent years, driven by the need to address challenges such as low completion rates and varying student profiles. Educational Data Mining (EDM) and Learning Analytics (LA) have emerged as key fields of research, focusing on the analysis of data from educational systems to uncover patterns and optimize learning environments. By examining interactions within VLEs, researchers can identify factors that influence student performance and develop strategies to enhance educational outcomes. This study leverages the OULAD to investigate the relationship between VLE usage and student success, with a particular focus on the following research questions:

1. RQ1: Is the use of the VLE associated with student approval?

2. RQ2: Which features from the VLE, census, and academic systems are most important for the early prediction of student performance?

3. RQ3: Which learning patterns can educational data mining help to unveil in the studied courses?

Through a series of data mining experiments, this work aims to uncover educational patterns and knowledge that can inform future institutional policies and practices. By analyzing data from the VLE, surveys, and academic systems, the study seeks to better understand the role of digital learning environments in supporting student achievement. The findings are expected to contribute to the broader discourse on the use of data-driven approaches in higher education, offering valuable insights for educators, administrators, and policymakers alike.

The remainder of this work is structured as follows: Section 2 provides an overview of related research in the fields of EDM and LA. Section 3 outlines the methodology employed in this study, including data collection, model generation, and evaluation. Section 4 presents the results of the analysis, while Section 5 discusses the implications of these findings in relation to the research questions. Section 6 explores potential institutional policies based on the evidence, and Section 7 concludes with a summary of the study’s contributions, limitations, and directions for future research.

**Data Description**

The dataset provided is part of the Open University Learning Analytics Dataset (OULAD), which captures detailed information about student activities, assessments, demographics, and interactions with the Virtual Learning Environment (VLE). Below is a comprehensive description of the datasets and their respective columns:

**1. courses.csv**

This file contains information about the available modules and their presentations.

- code\_module: Code name of the module, serving as the identifier.

- code\_presentation: Code name of the presentation, consisting of the year and a letter ("B" for February start, "J" for October start).

- module\_presentation: Length of the module-presentation in days.

**2. assessments.csv**

This file provides details about assessments for each module-presentation.

- code\_module: Code name of the module, serving as the identifier.

- code\_presentation: Code name of the presentation.

- id\_assessment: Identification number of the assessment.

- assessment\_type: Type of assessment: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA), or Final Exam.

- date: Submission date of the assessment, measured as the number of days since the start of the module-presentation.

- weight: Weight of the assessment in percentage. Exams typically have 100%; the sum of all other assessments is 100%.

**3. studentAssessment.csv**

This file contains the results of students' assessments.

- id\_assessment: Identification number of the assessment.

- id\_student: Unique identification number for the student.

- date\_submitted: Date of student submission, measured as the number of days since the start of the module-presentation.

- is\_banked: Status flag indicating that the assessment result has been transferred from a previous presentation.

- score: Student's score in this assessment (0-100, with scores below 40 considered Fail).

**4. studentInfo.csv**

This file includes demographic information and results for students.

- code\_module: Code name of the module, serving as the identifier.

- code\_presentation: Code name of the presentation.

- id\_student: Unique identification number for the student.

- gender: Gender of the student.

- region: Geographic region where the student lived while taking the module-presentation.

- highest\_education: Highest student education level on entry to the module presentation.

- imd\_band: Index of Multiple Deprivation band of the place where the student lived during the module-presentation.

- age\_band: Band of the student's age.

- num\_of\_prev\_attempts: Number of times the student has attempted this module.

- studied\_credits: Total number of credits for the modules the student is currently studying.

- disability: Indicates whether the student has declared a disability.

- final\_result: Student's final result in the module-presentation.

**5. studentRegistration.csv**

This file records student registration and unregistration dates.

- code\_module: Code name of the module, serving as the identifier.

- code\_presentation: Code name of the presentation.

- id\_student: Unique identification number for the student.

- date\_registration: Date of student's registration on the module presentation, measured as the number of days relative to the start of the module-presentation.

- date\_unregistration: Date of student unregistration from the module presentation, measured as the number of days relative to the start of the module-presentation.

**6. studentVle.csv**

This file records student interactions with VLE materials.

- code\_module: Code name of the module, serving as the identifier.

- code\_presentation: Code name of the presentation.

- id\_student: Unique identification number for the student.

- id\_site: Identification number for the VLE material.

- date: Date of student's interaction with the material, measured as the number of days since the start of the module-presentation.

- sum\_click: Number of times a student interacts with the material on that day.

**7. vle.csv**

This file contains information about the available materials in the VLE.

- id\_site: Identification number of the material.

- code\_module: Code name of the module, serving as the identifier.

- code\_presentation: Code name of the presentation.

- activity\_type: Role associated with the module material.

- week\_from: Week from which the material is planned to be used.

- week\_to: Week until which the material is planned to be used.

This dataset provides a comprehensive view of student activities, performance, and engagement, enabling detailed analysis and insights into the learning process at the Open University.

**3. Key Tables and Columns**

- courses.csv: Contains course module codes, presentation codes, and module presentation lengths.

- assessments.csv: Includes assessment IDs, types, submission dates, and weights.

- studentAssessment.csv: Records student assessment submissions, scores, and whether the result is banked.

- studentInfo.csv: Provides student demographics, including gender, region, highest education level, and previous attempts.

- studentRegistration.csv: Tracks student registration and unregistration dates.

- studentVle.csv: Logs student interactions with VLE materials.

- vle.csv: Details VLE materials, including activity types and planned usage weeks.

|  |  |  |
| --- | --- | --- |
| Dataset | Column Name | Description |
| courses.csv | code\_module | Code name of the module (identifier). |
|  | code\_presentation | Code name of the presentation (e.g., '2013B' for February, '2013J' for October). |
|  | module\_presentation | Length of the module-presentation in days. |
| assessments.csv | code\_module | Code name of the module (identifier). |
|  | code\_presentation | Code name of the presentation (e.g., '2013B' for February, '2013J' for October). |
|  | id\_assessment | Identification number of the assessment. |
|  | assessment\_type | Type of assessment: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA), or Final Exam. |
|  | date | Submission date of the assessment (number of days since start of module-presentation). |
|  | weight | Weight of the assessment in %. Exams typically have 100%; the sum of other assessments is 100%. |
| studentAssessment.csv | id\_assessment | Identification number of the assessment. |
|  | id\_student | Unique identification number for the student. |
|  | date\_submitted | Date of student submission (number of days since start of module-presentation). |
|  | is\_banked | Flag indicating if the assessment result is transferred from a previous presentation. |
|  | score | Student's score in this assessment (0-100, with scores below 40 considered Fail). |
| studentInfo.csv | code\_module | Code name of the module (identifier). |
|  | code\_presentation | Code name of the presentation (e.g., '2013B' for February, '2013J' for October). |
|  | id\_student | Unique identification number for the student. |
|  | gender | Gender of the student. |
|  | region | Geographic region where the student lived while taking the module-presentation. |
|  | highest\_education | Highest student education level on entry to the module presentation. |
|  | imd\_band | Index of Multiple Deprivation band of the student's location during the module-presentation. |
|  | age\_band | Band of the student's age. |
|  | num\_of\_prev\_attempts | Number of times the student has attempted this module. |
|  | studied\_credits | Total number of credits for the modules the student is currently studying. |
| studentRegistration.csv | date\_registration | Date of student's registration for the module presentation (relative to the start of module). |
|  | date\_unregistration | Date of student's unregistration from the module presentation (relative to the start of module). |
| studentVle.csv | id\_site | Identification number for the VLE material. |
|  | date | Date of student's interaction with the material (number of days since start of module). |
|  | sum\_click | Number of times a student interacts with the material on that day. |
| vle.csv | activity\_type | Role associated with the module material. |
|  | week\_from | Week from which the material is planned to be used. |
|  | week\_to | Week until which the material is planned to be used. |

**4. Data Preprocessing**

This section outlines the methodology and preprocessing steps undertaken to prepare the data for analysis. The process follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, adapted to the context of this research. The steps include data understanding, data preparation, and feature engineering, with a focus on merging and handling missing values in the dataset.

**4.1 Merging Data**

To perform comprehensive analysis and feature engineering, the seven tables in the dataset were merged using composite keys. The merging process was carefully designed to ensure no critical information was lost while maintaining data integrity.

**4.1.1 Merge Student-Related Tables**

Three of the data frames contain a composite key: ('id\_student', 'code\_module', 'code\_presentation'). Before merging, a decision was made regarding the type of join to use:

**1. Inner Join:**

- Returns only the matching rows between the two tables based on the specified key(s). Rows without matches are excluded.

- Use Case: When only complete records are required for analysis.

**2. Outer Join:**

- Returns all rows from both tables, including non-matching rows.

- Use Case: When retaining all records, even those with missing values, is necessary.

For this analysis, an inner join was chosen to ensure that only complete and consistent records were included in the final dataset.

**4.1.2 Merging StudentInfo, StudentRegistration, and Courses**

The StudentInfo and StudentRegistration tables were merged to add additional information about student registration and unregistration dates. The courses table was also merged to include the module length for each course. This step ensured that the dataset contained comprehensive information about student enrollment and course duration.

- StudentInfo: Contains demographic and academic performance data.

- StudentRegistration: Provides registration and unregistration dates.

- Courses: Adds module length and presentation details.

This merge did not result in any loss of information, as all relevant records were retained.

**4.1.3 Merging StudentAssessment and Assessments:**

The StudentAssessment and Assessments tables were merged to provide detailed information about each assessment, including scores, weights, and submission dates. This merge allowed for a comprehensive view of student performance across different types of assessments (e.g., Tutor Marked Assessments (TMA), Computer Marked Assessments (CMA), and Final Exams).

- StudentAssessment: Contains student-specific assessment data, including scores and submission dates.

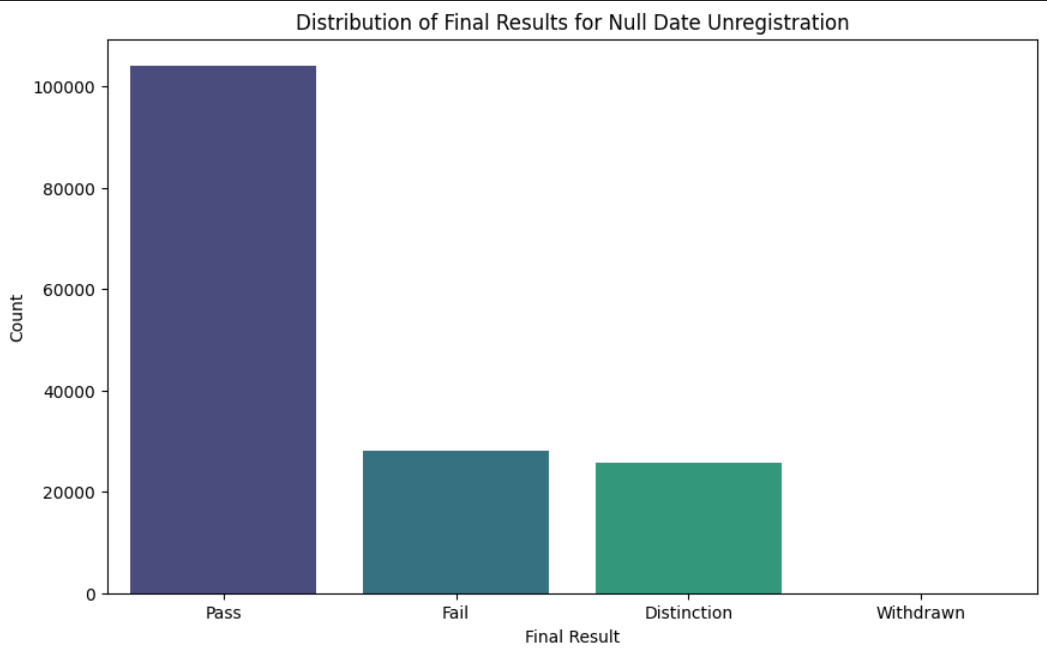
- Assessments: Provides general assessment details, such as type, weight, and due dates.

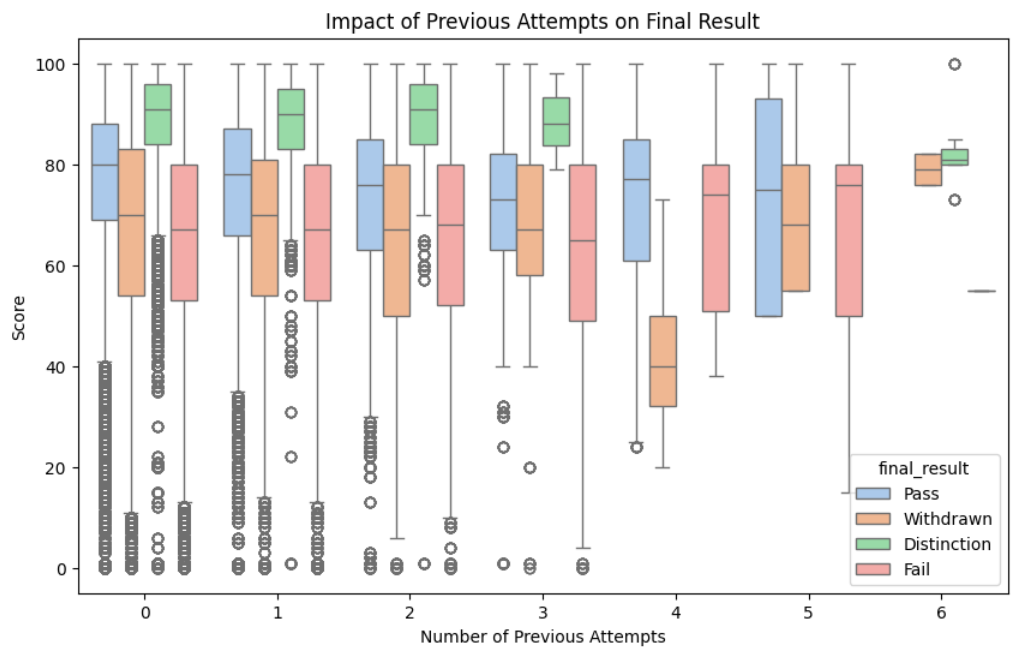
**Handling Missing Values in the Score Column**

During the merging process, it was observed that missing values in the score column were exclusively associated with TMA assessments. These missing values were interpreted as non-submissions, and the following approach was adopted:

- Missing scores were filled with zeros: This approach assumes that non-submissions equate to a score of zero, which is a common practice in educational data analysis.

This step ensured that the dataset remained consistent and suitable for further analysis.





**5 Data Preparation**

After merging the datasets, the following preprocessing steps were performed:

1. Data Cleaning: Inconsistencies and missing values were addressed. For example, missing scores in the StudentAssessment table were filled with zeros.

2. Normalization: Numerical features were normalized to ensure consistency across the dataset.

3. Feature Engineering: Derived attributes were generated from student interactions, such as the average number of interactions per week and the standard deviation of interactions

**5.1 Modeling**

The modeling process involved the following steps:

1. Dataset Configuration: Eight different datasets were created to evaluate the impact of various data configurations on model performance. These datasets included combinations of survey data, VLE interaction counts, and derived attributes.

2. Model Selection: Random Forest was chosen as the primary classifier due to its superior performance in preliminary tests. Hyperparameters were optimized using GridSearchCV.

3. Evaluation Metrics: The Area Under the Receiver Operating Characteristic Curve (AUC) was used as the primary evaluation metric. Additional metrics, such as accuracy, F1-score, and precision, were also reported.

4. Handling Imbalanced Data: The Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance in the dataset.

**Summary of Merging Strategy:**

- Inner Join: Used to merge StudentAssessment and Assessments to retain only complete records.

- No Information Loss: Ensured that all relevant data from StudentInfo, StudentRegistration, and Courses were retained during merging.

- Handling Missing Values: Missing scores in the StudentAssessment table were filled with zeros, as they represented non-submissions.

**6. Methodology:**

According to our problem It is possible to user the three Recommendations System Methodologies, that are Collaborative Filtering (CF), Content-Based Filtering and the Hybrid Recommendation Systems which will be used:

**1. Content-Based Filtering**

- **Why Use It?**

- Recommends learning materials based on the content of courses and the student's past interactions (e.g., completed courses, preferences).

- **Implementation:**

- Extract features from course content (e.g., topics, difficulty level, prerequisites, learning objectives).

- Build a student profile based on their completed courses, preferences, and performance.

- Use similarity metrics (e.g., cosine similarity) to recommend courses similar to what the student has already engaged with.

**2. Collaborative Filtering (CF)**

- **Why Use It?**

- Recommends learning paths based on the behavior of similar students.

- **Implementation:**

- Use user-based CF to find students with similar learning patterns and recommend courses they have taken.

- Use item-based CF to recommend courses frequently taken together (e.g., students who took "Statistics 101" also took "Data Analysis").

- Handle the cold start problem by combining CF with other methods (e.g., content-based or demographic-based).

**3. Hybrid Recommendation System**

- **Why Use It?**

- Combines the strengths of content-based and collaborative filtering to provide more accurate and diverse recommendations.

- **Implementation:**

- Use weighted hybrid or switching hybrid approaches:

- Weighted Hybrid: Combine scores from content-based and collaborative filtering.

- Switching Hybrid: Use collaborative filtering when sufficient data is available; fall back to content-based filtering for new students or courses.

- Incorporate knowledge-based rules (e.g., prerequisites, course difficulty) to ensure logical learning paths.

Let us Focus on the Collaborative Filtering (CF) in this Project.

**7. Recommendation Algorithm**

* PCA
* K-Means

**8. Discussion**

The project follows the CRISP-DM framework, which consists of six phases:

1. **Business Understanding**: Identifying the problem and objectives.
2. **Data Understanding**: Exploring and analyzing the data.
3. **Data Preparation**: Cleaning, integrating, and transforming data.
4. **Modeling**: Building predictive models.
5. **Evaluation**: Assessing model performance.
6. **Deployment**: Implementing the models for decision-making.

The primary goal is to predict student success in courses and final exams using data from Udaler’s VLE, academic records, and student surveys. The project also explores the use of **Recommendation Systems**, **K-means clustering**, and **PCA** to enhance the predictive models and uncover patterns in student behavior.

The modeling phase involved the use of several machine learning algorithms, including **Random Forest**, **Logistic Regression**, and **AdaBoost**. However, the focus of this report is on the implementation of a **Recommendation System**, **K-means clustering**, and **PCA**.

**8.1 Recommendation System Algorithm**

A recommendation system was implemented to provide personalized suggestions to students based on their interactions with the VLE and academic performance. The system uses **collaborative filtering** and **content-based filtering** techniques.

* **Collaborative Filtering**: This approach identifies patterns in student behavior by analyzing similarities between students. For example, if two students have similar interaction patterns and one performs well in a course, the system recommends similar resources or study strategies to the other student.
* **Content-Based Filtering**: This approach recommends resources based on the content of the materials accessed by the student. For example, if a student frequently accesses quiz materials, the system recommends additional quizzes or self-assessment tools.

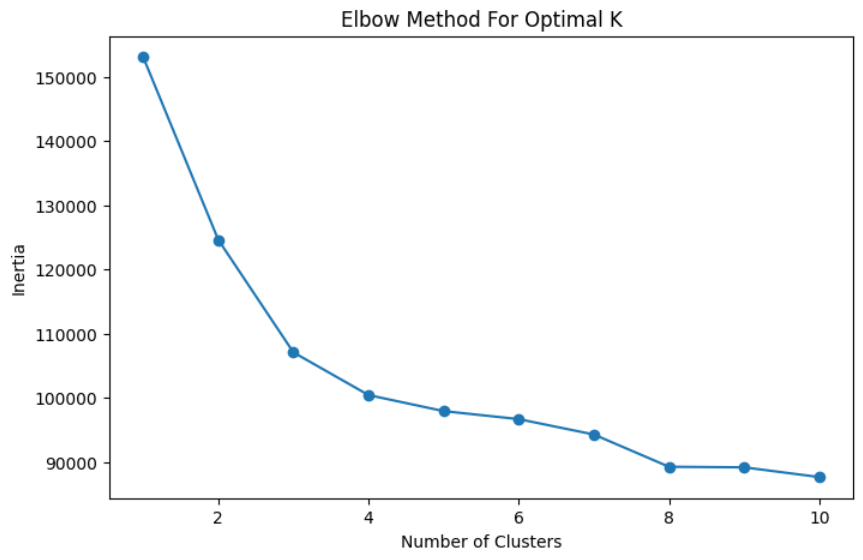
The recommendation system was implemented using Python's **scikit-learn** library. The system evaluates student interactions and academic performance to generate personalized recommendations, which are then delivered through the VLE.

**8.2. K-means Clustering**

K-means clustering was applied to group students based on their interaction patterns and academic performance. The goal was to identify clusters of students with similar behaviors and outcomes, which could inform targeted interventions.

* **Implementation**:
  1. **Feature Selection**: Features such as the number of interactions per week, types of interactions (e.g., quizzes, forums), and academic performance metrics were used.
  2. **Normalization**: Data was normalized using **StandardScaler** to ensure that all features contributed equally to the clustering process.
  3. **Clustering**: The K-means algorithm was applied with **k=3** clusters, representing low, medium, and high engagement levels.
  4. **Evaluation**: The clusters were evaluated using the **silhouette score** to ensure meaningful separation.
* **Results**: The clustering revealed distinct groups of students:
  1. **Cluster 1**: Low engagement, high failure rates.
  2. **Cluster 2**: Moderate engagement, mixed outcomes.
  3. **Cluster 3**: High engagement, high success rates.

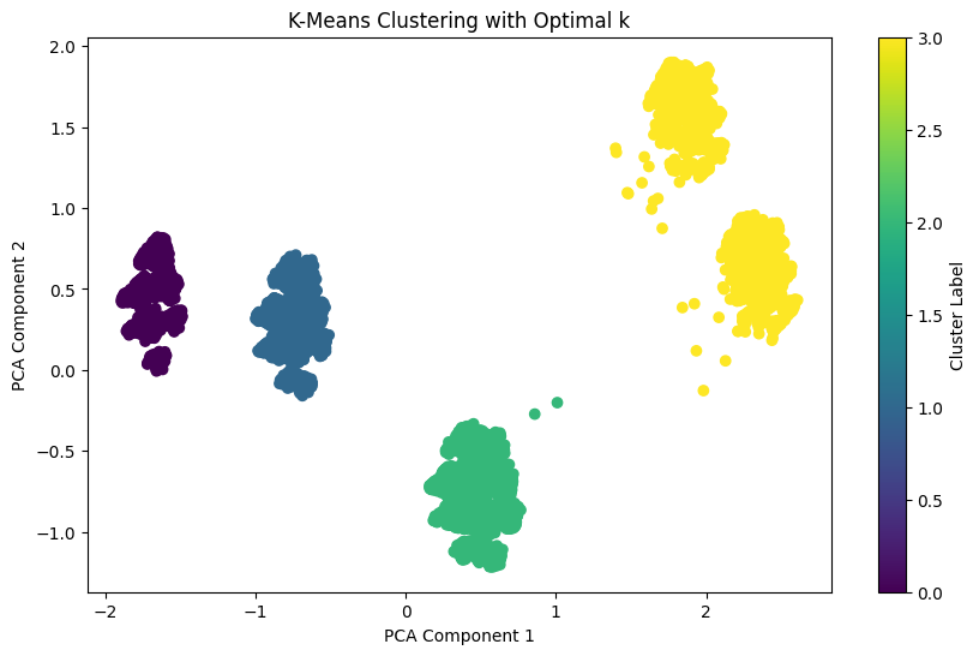
These clusters were used to design targeted interventions, such as additional support for low-engagement students.



**8.3. Principal Component Analysis (PCA)**

PCA was applied to reduce the dimensionality of the dataset while preserving the most important information. This was particularly useful for visualizing high-dimensional data and improving the efficiency of the predictive models.

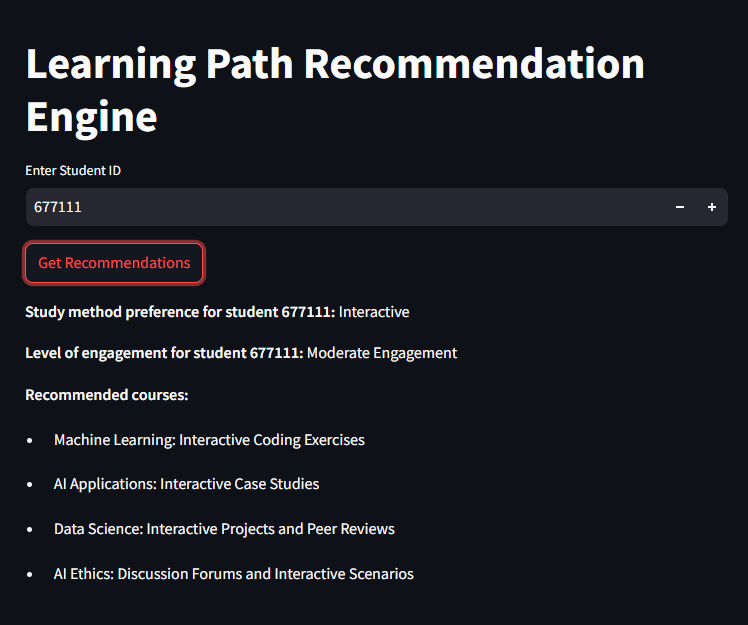
* **Implementation**:
  1. **Feature Scaling**: Data was standardized using **StandardScaler**.
  2. **PCA Application**: PCA was applied to reduce the dataset to 2 principal components, which captured the majority of the variance in the data.
  3. **Visualization**: The reduced dataset was visualized using scatter plots to identify patterns and trends.
* **Results**: PCA revealed that the first two principal components explained 85% of the variance in the data. This allowed for effective visualization of student interactions and academic performance, highlighting key trends such as the relationship between VLE engagement and success rates.



**9. Results:**

The predictive models achieved high accuracy in classifying student success, with the best-performing models achieving an AUC of 0.90. The combination of survey data and VLE interactions proved to be the most effective for prediction. The Recommendation System, K-means clustering, and PCA provided valuable insights into student behavior and informed the development of targeted interventions.

* **Recommendation System**: The system successfully identified personalized resources for students, leading to increased engagement and improved outcomes.
* **K-means Clustering**: The clustering results informed the development of targeted support programs for low-engagement students.
* **PCA**: The dimensionality reduction facilitated effective visualization and analysis of complex data.

****

**10. Conclusion**

The analysis of the Open University's VLE dataset provided valuable insights into student interactions, assessment patterns, and demographic factors influencing student outcomes. The preprocessing steps, including data merging and handling missing values, were crucial in preparing the dataset for further analysis. The exploratory data analysis revealed key trends, such as the correlation between unregistration dates and results, which can inform the development of a personalized learning path recommender system.

**11 References**

* <https://www.mdpi.com/2076-3417/11/15/6811>
* <https://analyse.kmi.open.ac.uk/open_dataset>
* <https://creativecommons.org/licenses/by/4.0/>
* <https://en.wikipedia.org/wiki/Open_University>
* <https://linkinghub.elsevier.com/retrieve/pii/S1096751620300026>
* <https://scholar.google.com/scholar_lookup?title=On+data-driven+decision-making+for+quality+education&author=Kurilovas,+E.&publication_year=2020&journal=Comput.+Hum.+Behav.&volume=107&pages=105774&doi=10.1016/j.chb.2018.11.003>
* <https://scholar.google.com/scholar_lookup?title=Teaching+in+a+digital+age:+How+educators+use+technology+to+improve+student+learning&author=McKnight,+K.&author=O%E2%80%99Malley,+K.&author=Ruzic,+R.&author=Horsley,+M.K.&author=Franey,+J.J.&author=Bassett,+K.&publication_year=2016&journal=J.+Res.+Technol.+Educ.&volume=48&pages=194%E2%80%93211&doi=10.1080/15391523.2016.1175856>
* <https://www.mdpi.com/2076-3417/10/18/6178>