ML Task 2

Anomaly Detection in OULAD

Faiza

BS Data Science

Group B

Table of Contents

Contents

Table of Contents **1**

1 Introduction **2**

2 Data Collection and Preprocessing **2**
2.1 Data Sources 2
2.2 Data Loading 2

3 Feature Engineering **2**
3.1 VLE Activity Aggregation 2
3.2 Merging Datasets 3
3.3 Additional Features 3

4 Cheater Detection **3**
4.1 Rule-Based Approach 3
4.2 Machine Learning Approach (Isolation Forest) 4

5 Comparison of Approaches **4**

6 Conclusion **5**

1 Introduction

Open University Learning Analytics Dataset (OULAD) provides student learning activity data. This task aims to identify cheating or anomalous behavior using click-based metrics and rule-based and machine learning (ML)-based approaches. Since the dataset is large, we utilize Dask for distributed data processing.

2 Data Collection and Preprocessing

2.1 Data Sources

The analysis used several datasets:

- Assessments data
- Courses information
- Student assessment results
- Student demographic information
- Student registration records
- Student VLE interactions
- VLE activity details

2.2 Data Loading

Due to the large size of the studentVLE dataset, it was loaded in chunks using Dask for efficient processing. Key steps included:

- Mounting Google Drive to access datasets
- Setting the correct file paths
- Installing and using Dask for large dataset handling
- Specifying appropriate data types for each column during loading

3 Feature Engineering

3.1 VLE Activity Aggregation

The studentVLE dataset was aggregated to create two key features per student:

- Total clicks: Sum of all interactions with the VLE

- Active days: Count of unique days with VLE activity

These aggregated results were saved to avoid recomputation.

3.2 Merging Datasets

Multiple datasets were merged to build a comprehensive student profile:

- Student assessment merged with assessment details
- Merged with student information
- Combined with student registration data
- Finally merged with the preprocessed VLE activity data

3.3 Additional Features

Several new features were created:

- Average assessment score per student
- Number of unique assessments attempted
- Days studied (difference between unregistration and registration dates)

4 Cheater Detection

4.1 Rule-Based Approach

Students were flagged as potential cheaters if they met all of:

- Average score ≥ 85
- Total clicks < 50
- Active days < 5
- Number of assessments < 3

This approach identified **38 suspected cheaters**.

Table 1: Sample of Suspected Cheaters (Rule-Based)

Student ID	Avg Score	Assessments	Total Clicks	Active Days	Days Studied	Final Result
25997	86.5	2	13.0	1.0	222.0	Withdrawn
380515	100.0	1	21.0	2.0	98.0	Withdrawn
384266	100.0	1	13.0	3.0	48.0	Withdrawn
386374	90.0	1	18.0	4.0	NaN	Fail
425314	100.0	1	12.0	1.0	95.0	Withdrawn

4.2 Machine Learning Approach (Isolation Forest)

An Isolation Forest model was trained to detect anomalies based on:

- Average score
- Number of assessments
- Total clicks
- Active days
- Days studied

The model identified **1253 potential cheaters** (anomalies).

Table 2: Sample of Suspected Cheaters (ML-Based)

Student ID	Avg Score	Assessments	Total Clicks	Active Days	Final Result
26211	89.25	12	15205.0	248.0	Pass
29639	83.77	13	10503.0	271.0	Pass
35355	75.22	9	3358.0	150.0	Withdrawn
42638	72.38	13	11845.0	252.0	Pass
42746	94.57	21	10673.0	219.0	Distinction

5 Comparison of Approaches

Table 3: Comparison of Cheater Detection Methods

Detection Type	**Number of Students**
Neither	23757
ML-Based Only	1252
Rule-Based Only	37
Both	1

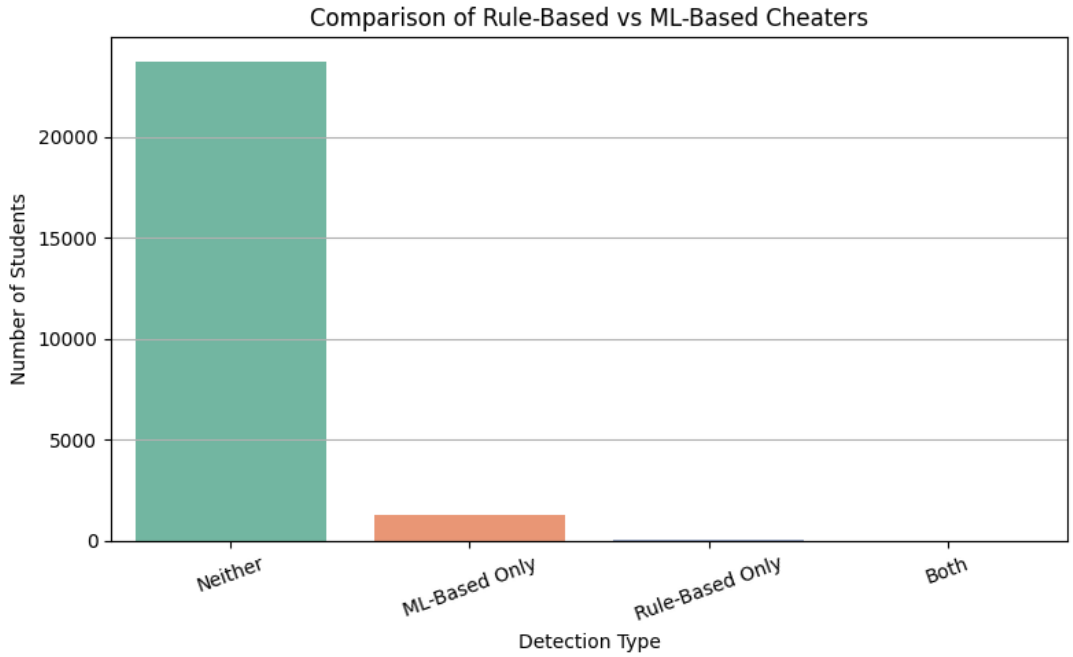

Figure 1: Visual Comparison of Detection Methods

6 Conclusion

The analysis revealed:

- The rule-based method identified students with high scores but low engagement
- The ML approach flagged more students showing unusual patterns
- Only 1 student was flagged by both methods
- Most students (23,757) were not flagged by either method

The Isolation Forest model proved more sensitive but less specific than the rule-based approach. The small overlap between methods suggests they detect different types of anomalies. Further investigation could combine both approaches for more robust cheating detection.